Assessing Secondary Curriculums' Impact on Postsecondary First Year Academic Performance Using Data Science Techniques

A Dissertation submitted
to the Graduate School
Valdosta State University

in partial fulfillment of requirements
for the degree of

DOCTOR OF EDUCATION

in Leadership

in the Department of Leadership, Technology, and Workforce Development
of the Dewar College of Education and Human Services

March 2024

Barrie D. Fitzgerald

MPA, Valdosta State University, 2012
B.A. Political Science, Valdosta State University, 2008

This dissertation, "Assessing Secondary Curriculums' Impact on Postsecondary First-Year Academic Performance Using Data Science Techniques" by Barrie Dwain Fitzgerald, is approved by:

Dissertation Committee Chair

Lantry L. Brockmeier

Lantry L. Brockmeier, Ph.D.
Professor of Leadership, Technology, and Workforce Development

Committee Member

James Leon Pate, Ph.D.
Professor of Leadership, Technology, and Workforce Development

Committee Member

Kathy Nobles, Ed.D.
Assistant Professor of Leadership, Technology, and Workforce Development

Associate Provost for Graduate Studies and Research

Becky K. da Cruz

Becky K. da Cruz, Ph.D.
Professor of Criminal Justice

Defense Date

March 20, 2024

FAIR USE

DUPLICATION

I authorize the Head of Interlibrary Loan or the Head of Archives at the Odum Library at Valdosta State University to arrange for duplication of this dissertation for educational or scholarly purposes when so requested by a library user. The duplication shall be at the user's expense.

Signature

I refuse permission for this dissertation to be duplicated in whole or in part.

Signature

ABSTRACT

Regional comprehensive universities offer accessible and diverse undergraduate educational programs, while grappling with funding cuts and affordability. The study's first research question underscores the enduring importance of factors such as student characteristics, pre-college characteristics, and financial situations. The findings highlight high school GPA's (HS GPA) pivotal role in academic performance. Higher HS GPAs correlate with successful academic performance resulting in higher retaining likelihoods; conversely, lower HS GPAs are associated with academic struggles and increased departure likelihoods. HS curriculum variables also impact academic performance, notably in extreme gradient boosting (XGBoost) models.

The second research question centers on the algorithms' predictive power. XGBoost and random forest models consistently outperform the other models in predicting GPAs. Prioritizing area under the curve values for retention, both XGBoost and random forest models are statistically comparable for developing predictive algorithms, despite facing challenges with low specificity rates. Only slight enhancements in predictions were detected in the upsample ensemble learning models.

Implications for practice underscore the importance of targeted interventions through leveraging data science techniques and machine learning algorithms to identify and allocate support resources for at-risk students. This research significantly contributes to the discussion on student success in higher education by providing practical insights and guiding evidence-based practices. As education evolves, integrating data science into strategic planning becomes pivotal for shaping the trajectory of student success initiatives.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I find it challenging to fully articulate my gratitude towards Dr. Lantry Brockmeier. Throughout this journey, he has not only been a pillar of support but also a catalyst, pushing me to surpass my perceived limits. His enthusiasm for data analysis and R programming is contagious. He has provided invaluable guidance throughout my coursework and the completion of this dissertation. I am deeply thankful for his generous investment of time and insightful feedback.

I would also like to express my appreciation to my committee members, Dr. Leon Pate and Dr. Kathy Nobles. Their assistance in reading my overly complex work and helping to translate the "magic math" into words for others to understand has been indispensable.

To my parents, your unwavering love, support, and encouragement have been my constant motivation. I am truly fortunate to have you as my parents. To my grandparents, thank you for believing in me and for always recognizing my potential. My love for my parents and grandparents stretches to infinity and beyond!

My church family deserves special mention for their encouragement and for reminding me to maintain faith throughout this endeavor. I extend specific thanks to Jeff and Christy Scoggins, Chris Sanchez, Matt and Lindsay Faircloth, Divonne LaVan and Jeanine Boddie-LaVan, and the Northside Baptist Church's worship ministry for their encouraging words and patience with my endless discussions on this topic and programming. A special heartfelt thanks to those who took the time to review the initial 18-page document that was front and back in the early stages.

To my colleagues, both past and present, I am grateful for your role in shaping me into the higher education professional I am today. Individuals like Nijah Bryant, Sandra Jones, Carla Jordan, Dr. Vince Miller, Christy Croft, John Crawford, Keisha Roberts, Brian Haugabrook, and many others have made the workplace a source of joy and inspiration. Again, thank you for providing endless encouragement, sound advice, laughing with me throwing my pages up in the air, and at times making me laugh to get me through this journey.

Lastly, I would like to thank my Lord and Savior, Jesus Christ, and the hope found in His promises. Jeremiah 29:11 reminds us of His plan to give us hope and a future. The best is yet to come!

Chapter I

INTRODUCTION

Each May, students walk across a stage to receive a diploma, marking the end of their high school career and the beginning of their journey into early adulthood. Some of these students enter the workforce or enlist in the armed forces, while others pursue a postsecondary degree. Recent national data from Fall 2018 show an enrollment of nearly three million first-time freshmen pursuing a postsecondary degree, with 45.4% enrolled in a four-year public institution (National Center for Education Statistics [NCES], 2019). As students embark on their postsecondary education, administrators and policymakers monitor specific milestones to evaluate students' progress towards degree completion. Milestones, such as grade point average (GPA) and retention status, are critical indicators of academic performance examined by postsecondary institutions (Arnold, 1999; Offenstein & Shulock, 2010; Tai, 2020).

Nationally, governmental agencies and other organizations release academic performance metrics, including enrollment numbers and other relevant data, for public access. These published facts allow agencies and the public to understand the "full and complete... condition of postsecondary education" (NCES, n.d.c). A recent report by the National Student Clearinghouse Research Center (NSCRC) (2019) revealed 73.5% of Fall 2017 first-time, full-time freshmen (FTFTF) continued to the second year at their initial postsecondary institution. This retention rate indicates slightly more than a quarter of students do not remain at their first institution. For public institutions, the retention rate for FTFTF was 79.1%. Although public institutions have a retention rate more than 5% higher than the national average, they still slightly trail behind private, not-for-profit institutions (NSCRC, 2019).

With the availability of postsecondary data, journalists and news organizations have scrutinized and reported on the issue of inadequate preparation for academic performance. Despite changes in high school curriculum aimed at college preparation, a 2011 *Washington Post* article cited research from Johns Hopkins University and the University of Arizona, which found 40% of students were unprepared for postsecondary coursework and even for the modern-day workforce (de Vise, 2011). Furthermore, a *U.S. News & World Report* article highlighted of the 1.8 million students who took the ACT in 2013, only 26% were deemed college-ready in four subjects. Additionally, 27% of students were prepared in two or three subjects, and 16% were ready for one subject (Bidwell, 2013). Overall, Bidwell (2013) reported 33% of students were unprepared for postsecondary courses in at least one of the major core subject areas.

In attempting to establish a connection to the lack of preparedness stemming from high school curriculum, *The Chicago Tribune* examined 120,000 high school juniors' exam scores. The findings revealed a weak general curriculum and lack of advanced or rigorous courses contributed to 75% of students being classified as unprepared for postsecondary coursework in mathematics, social studies, or sciences, and 50% of students for postsecondary coursework in English (Rado, 2017). Butrymowicz (2017) and French (2016) revealed one out of five students who graduate from high school are considered unprepared to succeed within their first year of studies. Elaborating further, they indicated earning a high school diploma does not equate to being prepared or successfully performing in a postsecondary setting (Butrymowicz, 2017; French, 2016). From these published articles, it can be inferred the high school curriculum could be affecting students’ first-year academic performance in postsecondary institutions.

With increased awareness of the lack of preparedness and poor postsecondary academic performance, elected officials began to enact laws to strengthen the high school curriculum through accountability metrics for students' readiness. In 2002, the No Child Left Behind (NCLB) act made efforts to hold elementary and high schools accountable for the educational progress of all students (Lee, n.d.). Largely focusing on closing achievement gaps, annual testing became a significant component of accountability to assist in identifying existing educational gaps, yet schools faced financial penalties when making no progress (Lee, n.d.; U.S. Department of Education, n.d.). In 2015, the Every Student Succeeds Act (ESSA) replaced NCLB, charging all American high schools to teach a high standard to prepare students for postsecondary education and the modern-day workforce (United States Department of Education, n.d.). In addition to shifting the accountability systems back to the states’ governments, ESSA also removed financial penalties for struggling schools (Understood Team, n.d.). Under ESSA, the law provided each state the autonomy to develop and implement its accountability system. However, ESSA still requires specific academic factors to be included. The factors focused on curriculum, including the assessments of reading and mathematics scores and English proficiency levels. Additionally, the law required states to develop a college readiness metric (Understood Team, n.d.).

Within the State of Georgia, the Department of Education developed the accountability system called the College and Career Ready Performance Index (CCRPI).

According to the Georgia Department of Education (GaDOE) (2021c), the accountability system contains published information regarding the postsecondary and career readiness of each school. With states having the autonomy of the accountability system, the GaDOE has made continuous improvements to the readiness index over the years. The CCRPI is a score from 0 to 100 derived from five components for each high school. The components for high schools include content mastery, progress, closing gaps, readiness, and graduation rates. Assessment tests over the four main subject areas in high school are used to develop the content mastery score. Growth in proficiency levels provides the progress score. The closing of the gaps in the scores measures any progress in meeting the improvement targets. The graduation rates for each high school are included (GaDOE, 2018b).

For high schools, readiness comprises a mixture of data points, including a college and career readiness score. The data points consist of literacy of 9th grade English and Language Arts proficiencies; student attendance; the percentage of 12th graders enrolled in dual enrollment, Advanced Placement, or International Baccalaureate courses; the percentage of 12th graders completing advanced academic, CTAE, fine arts, or world language pathways; and a derived college and career readiness score. Components of the derived college and career readiness score incorporate the percentage of students who enroll in public postsecondary institutions without the need for any remedial coursework, admission test scores, completion of at least two or more advanced courses from high school, completion of a pathway assessment resulting in a credential, and work-based learning experience (GaDOE, 2018a).

Furthermore, the public postsecondary education systems in the State of Georgia have increasingly begun to focus on improving academic performance to build a better-educated workforce, as directed by the governor in mid-2011. This directive is called Complete College Georgia (CCG). As charged by the governor, one of the main areas of CCG (2021a) focuses on college readiness to repair and strengthen the level of preparedness for college-level coursework. With increased focus on the critical first year, a University System of Georgia's (USG) initiative, Momentum Year, was implemented to improve academic performance and increase the number of students who graduate on time. USG's administrators encourage institutions to use predictive analytics along with advising techniques to help students navigate their first-year of studies (CCG, 2021b).

Published facts about the first-time college students within USG (2018c, 2019e, 2021c, 2021b, 2021c, 2021d) institutions indicate a three-year average enrollment of 51,889, with a 75.5% three-year average one-year retention rate. Comparing the retention rates of the national public institutions, USG lags slightly behind the national rate for public institutions at 3.6%. Within the USG, the institutions are divided into four tiers: research universities, comprehensive universities, state universities, and state colleges. Research institutions typically admit high-performing high school graduates and are more appealing to out-of-state students. Two research institutions have received high rankings by *U.S. News & World Report* (2021a, 2021b): Georgia Institution of Technology ranked 35th nationally overall and 8th nationally for engineering institutions, and the University of Georgia ranked 47th nationally overall. The three-year average retention rate when excluding research institutions is 69.8%, which is 9.3% below the national average for public institutions (USG, 2021b, 2021c, 2021d).

Statement of Problem

Successful academic performance within the first year at a postsecondary institution aids in building momentum and establishing a good foundation for the remaining years of a student's pursuit towards degree attainment. Conversely, unsuccessful performance leads to compounding factors resulting in long-lasting effects on the student's life, family, and occupation. According to the literature, unsuccessful academic performance leads to students departing, whether voluntarily or involuntarily, from the institution (Astin, 1984, 1993; Bean, 1980; Spady, 1970, 1971; Tinto, 1975, 1993). Departing students who received loans are forced to budget their income as they begin paying off the debt, contributing to less money for personal and leisure expenses. Moreover, college dropouts are four times more likely to default on college loans than those earning a degree (Ezarik, 2020).

Exiting a postsecondary institution without earning a bachelor's degree contributes to fewer job opportunities and growth, as most future jobs require a college degree (Lee, 2019; Smith-Barrow, 2019). Similarly, dropout students earn three times less than those who earned a bachelor's degree (Lee, 2019). The current unemployment rate of college dropouts is 2.2% higher than those with at least a bachelor's degree, indicating individuals who drop out of college are more vulnerable to layoffs and job eliminations (Lee, 2019; United States Bureau of Labor Statistics, 2021). Other than earned income, individuals who earn at least a bachelor's degree have increased benefits in better health and life expectancies than students who drop out (Lee, 2019). While it is not impossible to return and obtain a degree, these students may face increased frustrations, such as institutions rejecting credit hours or challenges in balancing family, work, and school (Smith-Barrow, 2019). Identifying the specific factors in the early stages of students' academic journeys creates a never-ending problem for higher education in understanding unsuccessful performance at an institution.

Existing literature paints an overwhelming picture of factors influencing students' first-year success. However, very little research has been conducted to indicate whether high schools' curriculum prepares students for the rigor of postsecondary coursework. With the implementation of the ESSA requiring states to assess college readiness, postsecondary institutions can incorporate this information into existing theories. The incorporation of high school curriculum quality, along with the known factors influencing academic performance, may aid in identifying students who could be at risk of unsuccessful performance in the first year.

Purpose of the Study

Through the utilization of data science techniques, the purpose of this study was to identify factors impacting the first-year academic performance of students enrolled in regional comprehensive universities (RCUs) in the State of Georgia. The factors included student characteristics, precollege characteristics— including high school curriculum quality, financial situations, major or program of study, and institutional financial expenditures. An additional purpose of the study was to develop four data mining models to determine which of the four algorithms yields the most accurate model. The accuracy metrics involved the review of the root mean square error (RMSE) for the first-fall and first-year GPAs. The accuracy metrics for the one-year retention status included accuracy, sensitivity, specificity, f-measure scores, and area under the curve

(AUC) value. The final purpose of the study incorporated an ensemble learning model to determine if a higher accuracy rate could be produced than through a single model.

Research Questions

The following research questions for this study guided the examination of factors affecting significant milestones for first-year academic performance, as indicated by Arnold (1999), Offenstein and Shulock (2010), and Tai (2020). Examining the factors impacting earned GPA and retention status can assist institutions in the identification of at-risk students.

1. Are student characteristics, precollege characteristics (including high school curriculum quality), financial situations, major or program of study, and institutional financial expenditures significant predictors in first-time, full-time freshmen's academic performance in their first year?
    a. Are student characteristics (gender, race and ethnicity, family educational background, and locale), precollege characteristics (high school curriculum quality, high school GPA, and admissions test scores), financial situations (family financial situations and financial aid), major or program of study, and institutional financial expenditures significant predictors of first-time, full-time freshmen's first-fall GPA?
    b. Are student characteristics (gender, race and ethnicity, family educational background, and locale), precollege characteristics (high school curriculum quality, high school GPA, and admissions test scores), financial situations (family financial situations and financial aid), major or

program of study, and institutional financial expenditures significant predictors of first-time, full-time freshmen’s first-year GPA?

   c. Are student characteristics (gender, race and ethnicity, family educational background, and locale), precollege characteristics (high school curriculum quality, high school GPA, and admissions test scores), financial situations (family financial situations and financial aid), major or program of study, and institutional financial expenditures significant predictors of first-time, full-time freshmen’s one-year retention status?

2. Does one machine learning algorithm (regression, support vector machine, random forest, and extreme gradient boosting) or an ensemble learning algorithm produce a higher accuracy based on the evaluation metrics for accuracy in examination of first-year academic performance?
   a. Does one machine learning algorithm (linear regression, support vector machine, random forest, and extreme gradient boosting) or an ensemble learning algorithm produce a higher accuracy based on the evaluation metrics of the root mean squared error (RMSE) for first semester GPA?
   b. Does one machine learning algorithm (linear regression, support vector machine, random forest, and extreme gradient boosting) or an ensemble learning algorithm produce a higher accuracy based on the evaluation metrics of the RMSE for first-year GPA?
   c. Does one machine learning algorithm (logistic regression, support vector machine, random forest, and extreme gradient boosting) or an ensemble learning algorithm produce a higher accuracy based on the evaluation

metrics of accuracy, sensitivity, specificity, f measure scores, and AUC value for one-year retention status?

Research Methodology

This study was a nonexperimental, ex post facto, correlational research design. Research utilizing ex post facto allows for examining phenomena in which possible causes have already occurred (Bordens & Abbott, 2011). Institutional data collected for this study came from USG's Office of Research and Policy Analysis (RPA). RPA oversees the collection of census and other data files from each institution. Student enrollment data are collected twice a semester (USG, 2021b). Census files are data collected by the office from the information systems on specific dates to report facts to outside agencies (Milam & HigherEd.org, 2003). First-year academic performance was examined using these census files. Additional archival data was obtained from the GaDOE's College and Career Ready Performance Index, Governor's Office of Student Achievement, and NCES' websites. These files contained information on the high school quality of curriculum and the institutions' expenditures. As archival data prevent the independent variables from being manipulated, the nonexperimental, ex post facto research design was justified (Bordens & Abbott, 2011).

Within the study, the dependent variables consisted of three different measurements of academic performance. Two of the dependent variables were interval-level data, and one dependent variable was nominal-level data. The interval levels comprised the first-fall and first-year GPAs, while the nominal level comprised the one-year retention status. The independent variables consisted of 11 nominal variables and 25 interval measurement levels data. Nominal variables consisted of student characteristics (gender, race and ethnicity, family educational background, and locale), pre-college characteristics (five subject areas of the college preparatory curriculum requirements), financial status (Zell Miller recipient), and major or program of study. Interval level data consisted of precollege characteristics (high school GPA, admission test scores, and four advanced standing hours), high school curriculum (CCRPI content mastery, CCRPI readiness, EOC mean English and Language Arts, EOC mean Mathematics, EOC mean Science, and EOC mean Social Studies), financial situations (EFC, HOPE Scholarship, PELL Grant, federal subsidized and unsubsidized loans, and other loans), and institutional expenditures (instruction, research, public service, academic support, student services, institutional support, and other core expenses). In addition to the study's goal of identifying factors impacting first-year academic performance for three dependent variables, the research attempted to identify the best model in terms of accuracy in predictions utilizing data science techniques. The data science techniques involved analyzing accuracy metrics across four models along with an ensemble learning model to identify the best accuracy in predicting earned GPAs and retention status of students.

The target population enrolled in the four regional, comprehensive universities was the FTFTF pursuing a bachelor's degree who graduated from a public high school within the State of Georgia. The population obtained was for two fall semesters: Fall 2018 and Fall 2019. Based on published information from the system office, USG, the target population averaged 13,178 students per year, totaling 26,356 students (USG 2022a, 2022b). The system office does not publish a breakdown of students who graduated from a public high school and those who did not. For students to qualify for the study, the following three criteria were met:

1. Integrated Postsecondary Education Data System's (IPEDS) classification of first-time, full-time freshmen (determined by USG for IPEDS reporting)
2. Pursuing a bachelor's degree (determined by USG for IPEDS reporting)
3. Graduated from a Georgia public high school in 2018 or 2019 (calculated from the high school code in RPA census files)

A mixture of descriptive and inferential statistics was used in the study to analyze the data. For the interval-level data, the number of records, mean, median, standard deviation, minimum value, maximum value, skewness, and kurtosis were calculated. Frequencies and percentages were calculated for the nominal and ordinal variables. Within the study, statistical learning assisted in discovering which factors impact first-year academic performance. According to James, Witten, Hastie, and Tibshirani (2013), statistical learning incorporates tools to understand data involving either supervised or unsupervised learning techniques. Supervised learning is the most appropriate technique for the analysis of academic performance within the first year. Within supervised learning, developing predictive models involve at least one factor producing an estimate used to influence the outcome (James et al., 2013). Predictive algorithms are largely time-consuming to conduct by hand; however, with the advancement in technology, these algorithms can be conducted faster through automated processes. With the expansion of algorithms, testing multiple models allows researchers to find one with the best accuracy to provide insights to aid in the decision-making process. The insights provided from the predictive modeling with the best accuracy would allow administrators and other policymakers to effectively and efficiently allocate resources to support students identified as unsuccessful academic performers.

Within R programming, the tidymodels package were utilized to assist in data preparation, algorithm development, and the assessment of accuracy metrics (Kuhn & Silge, 2021). In addressing the first research question, predictors were assessed through variable importance analysis. While linear and logistic regression models produce coefficients to measure the impact, non-regression models do not produce equivalent values. Instead, each model's variable importance analyses were examined to measure the impact of the factors; additionally, the importance values were rescaled for comparison across models.

Before building the predictive models, it was crucial to review statistical considerations and assumptions to ensure the results draw valid conclusions on reality and produce meaningful research (Field, Miles, & Field, 2012; Garson, 2012). Depending on the statistical analysis, the considerations and assumptions may vary. Reviewing observation independence is important to ensure no individual presents a bias within the analysis (Heidel, 2022). Moreover, missing data is another consideration that needs to be reviewed. For this study, any missing data may be missing completely at random, missing at random, or missing not at random (Mack, Su, & Westreich, 2018). Missing completely at random data occurs when the missing data are independent regarding the observed and unobserved data, resulting in no systematic differences between individuals. Missing data related to the observed rather than the unobserved data are classified as missing at random, resulting in the introduction of bias in the analysis. Lastly, missing not at random data stems from data missing due to some factor can be accounted for by the researcher (Mack et al., 2018).

According to Osborne and Overbay (2004), data points beyond the norm are identified as outliers. The presence of outliers has the potential to introduce bias in the estimated parameters of a predictive model, which may result in Type I or Type II errors (Osborne & Overbay, 2004). With the usage of archival data, outliers largely will stem from data entry issues from being miskeyed into the information system. While archival data undergo data validations and cleanup processes, some errors go undetected. Outliers were reviewed through a combination of summary statistics, z-score examination, data visualization, and statistical analyses (Osborne & Overbay, 2004). Some analyses, like general linear models, require the data to be evenly distributed, as skewness creates undue influence and impacts the model's estimates (James et al., 2013; Sharma, 2019). Ensuring equal distribution, univariate and multivariate normality were examined through data visualization techniques and statistical methods such as Shapiro-Wilks and Royston's tests (Fife, 2019; Merler & Vannatta, 2002; Oppong & Agbedra, 2016). Linearity requires the data to have a linear relationship resembling a straight line with the dependent variable (Merler & Vannatta, 2002). Examination of linearity involved the review of the Pearson's R correlation coefficient and residual plots (Glen, 2022; Merler & Vannatta, 2002). The last assumption is the review of homoscedasticity. Using a statistical test such as Levenne's test allowed the assessment of the equal distribution of variance (Merler & Vannatta, 2002).

With the second research question, utilizing multiple predictive models and an ensemble learning algorithm to assess the accuracy of the predictions, accuracy metrics were examined to discover the appropriate accuracy and interpretability trade-off. According to Kuhn and Johnson (2013), predictive models should not sacrifice accuracy for interpretability. They further argued "the predictive models that are most powerful are usually the least interpretable" (Kuhn & Johnson, 2013, p. 50). Additionally, when comparing models for accuracy, unintended consequences may result from the selection of the wrong model (Kuhn & Silge, 2021). For regression analysis of the GPA dependent variables, accuracy metrics included the examination of the RMSE. Boehmke and Greenwell (2020) emphasized the importance of RMSE values in determining the accuracy of regression outputs over other metrics for regression outputs. For the retention-dependent variables, examination of classification accuracy of the models utilized different metrics. Classification models' accuracy metrics included overall accuracies, sensitivities, specificities, F-scores, and ROCs (Boehmke & Greenwell, 2020; Dey, 2021; Kuhn & Johnson, 2013). Dey (2021) stressed the importance of the AUC produced from the ROC graphs as the value used in comparing multiple classification models.

The evaluation of accuracy for the three dependent variables were used to assess the out-of-sample predictive power (Calvo & Santafé, 2016; Horthorn et al., 2005; James et al., 2013; Kuhn & Johnson, 2013; Kuhn & Johnson, 2019). In building the predictive algorithms, a 10-fold validation was applied to the training and testing data sets. The cross-validation of training data sets allowed for the review of the accuracy of the models for optimal performance before application to the testing or the unseen, real-world data set (Goyal, 2021; Soni, 2019). The 10-fold cross-validation methods was also applied to the testing data set. This method of cross-validation on the testing data set was used to examine how the models perform across several slices of the unseen real-world data.

According to Bose (2019), cross-validation allows researchers to review how a model may overfit to the training data set.

Significant of the Study

This study made a minimum of three contributions to the current literature. First, traditional models of examining academic performance often exclude the high school's curriculum quality. This study sought to explore whether there is a relationship between the high school's curriculum and academic performance while considering traditional factors. Findings from this study identified the impact of high school curriculum quality, if any, along with characteristics from previous literature, in successful first-year academic performance. Second, the study analyzed the performance of predictive algorithms by extending beyond traditional inferential statistical tests. By identifying which predictive model provided the best accuracy, postsecondary institutions would be able to gain insights into implementing additional analytical tools to detect students who are more likely to be unsuccessful within the first year of studies. Lastly, this study analyzed specific milestones within the first year to see how the identification of characteristics changes or enhances the accuracy of predictions.

With an increased focus on student success at the national and state levels, initiatives such as Complete College America (n.d.) and Complete College Georgia (2021a) have prompted postsecondary administrators and other policymakers to identify best practices and implement strategies to enhance academic performance. Furthermore, institutional researchers, student success data managers, and others studying academic performance may find the study results relevant. Their job duties may include assisting administrators and policymakers in assessing the impact of current and potential new strategies. The study aimed to assist postsecondary institutions in identifying characteristics affecting the academic performance of FTFTF students within the first year of their studies. Postsecondary employees could use the information derived from the findings to allocate expenditures and resources more effectively to improve academic performance.

Theoretical Framework of the Study

The theoretical framework for the study was the integration theory of student departure developed by Tinto (1975, 1993). As one of the most prominent theoretical models, Tinto argued previous models lacked attempts to provide an explanation for the student departure phenomenon. Initially, Tinto (1975) based the model on the utilization of Durkheim’s suicide theory. The initial model indicated students' departure decisions stemmed from a failure to integrate into the academic or social societies found within a postsecondary institution. Upon reflection and in response to negative criticism regarding the usage of the term "suicide," Tinto (1993) revised the model, arguing students' integration into the postsecondary institution's community resembles Van Gennep’s rites of passage theory, as displayed in Figure 1. According to Tinto (1993), students begin the initial steps of separation from familiar surroundings and norms when they enroll in a postsecondary institution. After the separation, students go through a transition phase as they interact with the new surroundings and norms. Tinto (1993) argued this process assists in facilitating their integration into the academic and social communities.

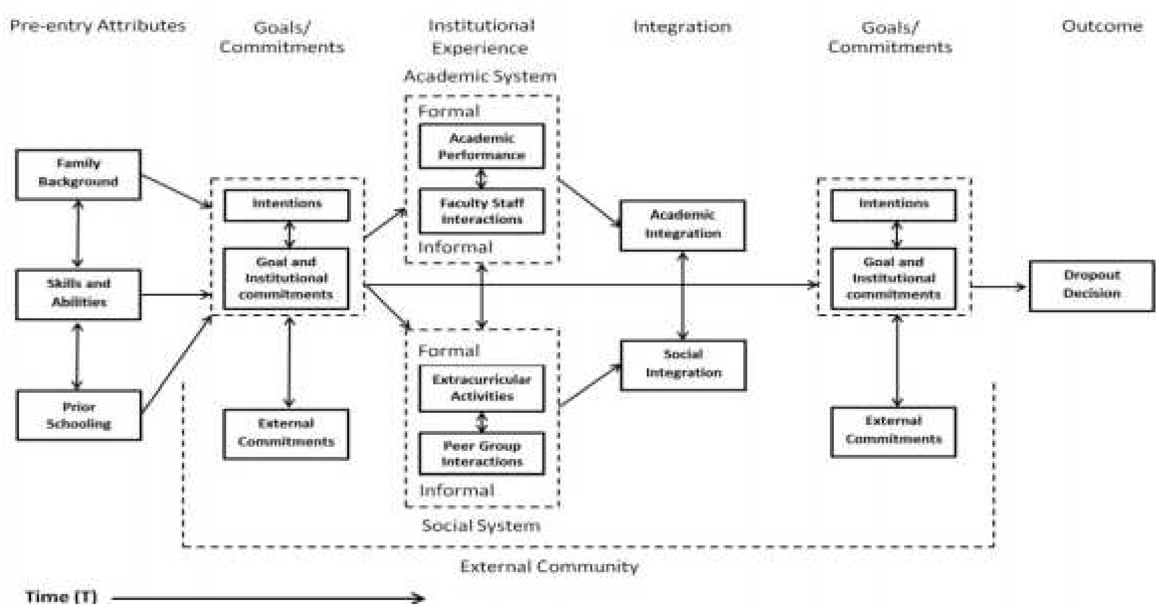


Figure 1. Tinto's integration model.

Tinto (1993) cautioned the failure to integrate stems from a void produced between the separation and incorporation stages. As students separate from their familiar norms, they may experience loneliness, hindering integration into the communities. Since the inception of his theoretical model, Tinto (1975, 1993) argued student attributes, such as family background, student abilities, and precollege schooling, factor into the students' goals and commitments. The goals and commitments of the students then influence their integration into the postsecondary communities. Noting the integration process does not occur within a vacuum, Tinto (1993) stated external commitments, such as family and job commitments, influence the departure decision.

Limitations of the Study

There were several limitations to the study. First, concerning the location of the population, the study was focused on students enrolled in four regional, comprehensive universities within USG. As a result of the selected institutions, other institutions may not be able to generalize the findings to their enrolled students. Another limitation pertained to the population examined. With the study focusing on the first-year academic performance of FTFTFs who recently graduated from high school, the findings would not be applicable to other undergraduate students—such as FTFTFs who delay enrollment and transfers—or beyond the FTFTFs' first year. This study did not include an exhaustive list of factors impacting academic performance. An analysis involving a more comprehensive list would alter the study's findings. Regarding the quality of the high school curriculum, using aggregated curriculum values of the schools to compare a subset of students posed a limitation since not all students from these schools enrolled in a postsecondary institution. Furthermore, the information collected on high school curriculum quality was based on pre-COVID-19 pandemic changes.

In studying the phenomenon of academic performance, the study had a limitation due to the utilization of ex post facto data. The data collected and utilized in examining the factors impacting first-year academic performance did not include students' motivation, as indicated by Tinto (2017) in his reflection on factors in student persistence. Tinto (2017) stated students' self-efficacy, sense of belonging, and perception of the curriculum provide further insights into students' decisions to persist at or depart from the institution. Factors related to student motivation are typically obtained through surveys and interviews rather than being stored in archival data. This study did not incorporate data collected from surveys regarding students' motivation attributes as they are beyond the scope of the study.

Lastly, the final limitation involved the record-keeping process. This study collected data from four different sources. Regarding the data, human error in the form of data entry may occurred. As data entry involves a person keying in information into a student information system, this may have resulted in someone accidentally entering the wrong data. For example, simple data errors could lead to the accidental entry of incorrect information for a student's high school GPA or admission test score during the processing of admissions applications.

Definitions of Terms

Within this study, the following terms or phrases are used and provide consistency:

- Area under the curve (AUC). The AUC is a diagnostic test to measure the overall usefulness of a model. A model with a no discrimination would have a value of .50, while a perfect model would have a value of 1 (Mandrekar, 2010).
- College and Career Ready Performance Index (CCRPI). As part of ESSA, GaDOE developed the CCRPI to communicate improvements and display accountability regarding each school's ability to promote readiness for life after graduating (GaDOE, 2021C).
- End-of-course (EOC) tests. Each high schools' rates of proficiency and above proficiency rates for the four main subject areas, given to students at the end of the semester (GOSA., n.d.).
- F-measure. The F-measure is a measurement of accuracy involving the harmonic mean of the recall and precision (Brownlee, 2020).
- False-negative rate. The false-negative rate is a measure of percentage of actual cases with the event wrongly predicted to not have the event occurring (Silipo & Widmann, 2019). The false-negative rate in this study was the percentage of students were predicted not to retain divided by total actual retained students.

- False-positive rate. The false-positive rate is a measure of percentage of actual cases within the event wrongly predicted to have the event occurring (Silipo & Widmann, 2019). The false-positive rate in this study was the percentage of students were predicted to retain divided by total actual not retained students.
- First-time, full-time freshmen (FTFTF). According to the Integrated Postsecondary Education Data System (IPEDS), an FTFTF are students pursing an undergraduate degree by matriculating into an institution for the first time. The student must enroll in the fall term or the proceeding summer term to qualify (National Center for Education Statistics [NCES], n.d.e). Furthermore, the student must be enrolled in at least 12 credit hours to be considered full-time (USG, 2020a).
- First-year academic performance. The first-year academic performance in this study was comprised of first-fall GPA, first year GPA, and one-year retention status.
- High school curriculum quality. In the study, the HS curriculum quality measured the schools' overall content mastery and readiness scores in additional the proficiency levels of the four main subject areas.
- Out-of-sample predictive power. The out-of-sample predictive power builds on the trustworthiness of the prediction: accuracy predicting the events (Kuhn & Johnson, 2013).
- Precision. The precision is a calculation measuring the accuracy actual positive cases within the total predicted positive cases (Silipo & Widmann, 2019).

- Predictive modeling/predictive algorithm. Predictive modeling and predictive algorithm may be used interchangeably through the study. Predictive modeling, as defined by Kuhn and Johnson, is "the process of developing a mathematical tool or model that generates an accurate prediction" (2013, p. 2).
- Receiver operating characteristics (ROC) curve. The ROC curve is a plot in which displays the true-positive and true negative rates. Within the plot, the point closest to the upper left corner is considered the most accurate (Kuhn & Johnson, 2013).
- R-square/Adjusted R-square. The r-square and adjusted r-square provides a measurement of correlation for ratio and interval dependent variables. The measurements assess how well the model fits the dependent variables (Kuhn & Johnson, 2013). These values measured the correlation between the predicted and actual GPAs.
- Regional, comprehensive university (RCU). A regional, comprehensive university is a four-year institution degree granting institution offering primarily bachelor's degrees to undergraduate students. These types of institutions have a different focus than the flagship and research institutions, as their primary goal is to have an impact on workforce in the region they are located. Regional, comprehensive universities not only provide affordable options for students, but according to Sandeen (2020) their primary goal is the success of the students.
- Retention status. The retention status is defined as whether a student returned to the initial postsecondary institution one-year later or graduated before the proceeding fall semester. (USG, 2020b).

- Root mean square error (RMSE). The RMSE is a measurement of accuracy of predicted values to the actual values for ratio or interval dependent variables (Kuhn & Johnson, 2013). In the study, RMSE measured the distance between the actual and the predicted GPA values.
- Sensitivity The true-positive rate, sensitivity, or recall evaluates the accuracy of the prediction within the sample with the event actually occurring. The percentage of the true predictive positive cases divided by the total actual true cases (Kuhn & Johnson, 2013). The true-positive rate in this study would be the percentage of students predicted to retained by the total actual students who retained.
- Specificity. The true-negative, specificity, or false alarm evaluates the accuracy of the predictions within the sample without the event actually occurring. The percent is the true predicted negatives divided by the total actual negative cases (Kuhn & Johnson, 2013; Silipo & Widmann, 2019). In this study, the true-negative rate is the percentage of students who were not predicted to retain divided by the total number of students who actually retained.
- Testing data set. The testing data set is the portion of the data set used in assessing the accuracy of the algorithm's predictions (Kuhn & Johnson, 2013).
- Training data set. The training data set is the portion of the data set used in developing the model (Kuhn & Johnson, 2013).
- University System of Georgia (USG). The USG (2021a) compromises of 26 public postsecondary institutions within the State of Georgia. Of the 26

institutions, four are research institutions; four are regional, comprehensive institutions; nine are state universities; and nine are state colleges.

Organization of the Study

The study is organized into five distinct chapters. Chapter 1 contains an introduction, statement of the problem, purpose statement, and significance for the study. Additionally, the chapter provides the research questions and methodology conducted in the study. Within Chapter 2, there is a review of relevant retention theories proposed by Spady (1970, 1971), Tinto (1975, 1993), Bean (1980), and Astin (1984, 1993), followed by a literature review of characteristics affecting academic performance within the first year at a postsecondary institution. Additionally, a review of RCUs and data science is also included in the literature review. Chapter 3 discusses the breakdown of the study to describe the research design, population, and methods of data collection. The third chapter explains the research methodology, containing a review of the data analysis involving descriptive and inferential statistics, along with the review of statistical considerations and assumptions. The fourth chapter contains the analysis and interpretation of the data to answer two research questions. As the conclusion of the study, the final chapter provides an overview of the study along with a discussion of the results and implications for future research.

Chapter II

LITERATURE REVIEW

In this study, students' academic performance in regional comprehensive universities in Georgia's public system is investigated to determine what independent variables, including student characteristics, precollege characteristics, graduating high school characteristics, major declaration, and institutional financial expenditures, have an impact within the first year. Retention theories proposed by Spady (1970, 1971), Tinto (1975, 1993), Bean (1980), and Astin (1984, 1993) provide the basis for examining academic performance at the comprehensive public post-secondary institutions. This chapter begins with an overview of regional comprehensive universities along with the growing concerns these institutions face. The next section has a review of the literature regarding the independent variables impacting first-year academic performance. The last section contains a review of data science and data mining techniques implemented in higher education.

Regional Comprehensive Universities

According to the American Association of State Colleges and Universities (AASCU) (2020), a regional comprehensive university (RCU) provides students with a high-quality education at affordable costs. Originally, RCUs were established to produce future educators, formed as night schools, and provide educational opportunities for veterans (AASCU, 2020; Orphan, 2018a, 2018b). Henderson (2009) stated public RCUs are "the People's University" due to the increased focus on educational opportunities and connection to the local economic area (p. 5). Unlike research universities, admission standards for RCUs are more relaxed requirements to which almost half of the students applying are accepted (Nietzel, 2019b; Orphan, 2018b). These institutions provide educational opportunities for a more diverse student population with relaxed admission requirements. Students attending RCUs overwhelmingly commute rather than live in residential dorms (Reis, n.d.). Within the student body, undergraduates come from diverse backgrounds and preparations. The diversity of the student population includes age, race and ethnicity, first-generation status, and individuals from low to middle socioeconomic statuses (Nietzel, 2019a). Frequently, these students come from a non-affluent background and enroll in an RCU because of the possibility to prepare them for a future or career advancement (Reis, n.d.).

While research institutions gain considerable media attention, RCUs provide a large share of enrolled students in the state, and administrators work to provide opportunities for success to these students. The focus of RCUs centers on undergraduate student programming in which faculty are praised for teaching (Nietzel, 2019b; Orphan, 2018b; Schleifer, Hagelskamp, & Riendhart, 2015). These institutions provide a wide range of undergraduate programs, along with a few master's degree programs (Henderson, 2009; Nietzel, 2019b). A few institutions have doctorate degrees primarily concentrating in the educational field (Henderson, 2009). RCUs and the local areas have a symbiotic relationship. RCUs depend on the areas as a large base for recruitment of new students, while the areas need RCUs to provide students who are career-ready (Orphan, 2018b). Additionally, RCUs educate over half of the teachers within the state's school system (Orphan, 2018a).

Largely, RCUs belong to the public sector of control and have experienced declining funding and often face budget cuts. Before the recession in 2008, public RCUs received 70% of their funding from state appropriations, only to see the funding percentage take around a 20% cut (Orphan, 2018b). In other words, RCUs have largely become reliant on funding from tuition generated based on student enrollment. With the reliance on tuition funding, RCUs have been the main driver of tuition increases, resulting in increased costs driving the public's growing disinvestment (AASCU, 2020). However, RCUs struggle with a delicate balance of affordability and funding. As tuition rates continue to rise, these institutions risk becoming counterproductive in providing affordable education. Yet without increasing tuition, educational quality may suffer from a lack of funds (Schleifer et al., 2015). A total of 27 states have linked academic performance to the available state appropriations for institutions, called performance-based funding (Schleifer et al., 2015).

Growing concerns for regional comprehensive universities. With RCUs considered the workhorse of the postsecondary institutions in producing degree-credentialed workforce for state and regional areas, these institutions need to consider any changes having a profound impact on the general population. While postsecondary institutions have a myriad of issues to deal with now, administrators and other policymakers will need to be aware of the current decline in the traditional age student population, changing demographics, and the public's perception of postsecondary education.

Decline in traditional-age student population. From December 2007 to June 2009, the United States of America experienced an economic recession. Considered the most significant recession since The Great Depression, Livingston and Cohn (2010) linked the recession to the birth rate decline. They reported the birth rate dropped from 69.9% to 68.8% (Livingston & Cohn, 2010). While the recession impacted states differently, an overall decline occurred (Livingston & Cohn, 2010). Along with a birth rate decline, high school graduates will naturally follow suit as children progress through the elementary, middle, and secondary school systems. In a *Hechinger Report*, Barshay (2018) wrote postsecondary enrollment should prepare for enrollment declines beginning in 2025. Furthermore, Barshay (2018) indicated some states would experience a greater decline than other states. While most postsecondary institutions are growing in student and employee size, Miller (2020) indicated some institutions have already begun to experience a downturn in enrollment.

In the beginning of 2020, COVID-19 presented significant challenges, and postsecondary institutions had to adapt to continue operating and educating students. As a result of the pandemic, Sedmak (2020) wrote a press release for NSCH indicating Fall 2020's decline was twice as severe as the decline in Fall 2019. As postsecondary institutions adapted to the new norms, some institutions modified the recruiting methods and admissions requirements. Due to several relaxed requirements, some institutions experienced relative increases in enrollment for the following fall semester (Sedmak, 2020). USG (2020c) experienced a 2.4% increase within the State of Georgia to achieve a system-wide all-time high in record enrollment for Fall 2020. For Spring 2021, Sedmak (2021) reported the national undergraduate enrollment in postsecondary institutions faced its sharpest decline of 5.9%. As national reports have begun to indicate the decline predicted to occur in 2025, the current trend along with the prediction does not provide a promising outlook for future enrollment growth of traditional-age students.

Changing demographics. In a *Pew Research* Report, Henderson (2016) stated more people within the nation had started to migrate from the Northeast to the South or West. He indicated the largest areas of growth occurred in the Sun-Belt region (Henderson, 2016). The migration has been large enough to impact how congressional house seats were distributed across the states (Henderson, 2016). Henderson (2016) reported the migration had an economic driver rather than weather or climate drivers. Continuing, he revealed two-thirds of the long-distance migrants were due to job opportunities and housing costs (Henderson, 2016). As migration within the nation occurred, this would result in population changes causing economic and population growth.

Individuals relocated for better opportunities, which sparked a population demographic change in the receiving regions. Additionally, projections have indicated the nation's demographics will be a majority minority by 2025. The change in demographics not only stems from within-the-nation migration, but the nation has reached an all-time high of migrations from other nations (Facing History & Ourselves, 2021). For postsecondary institutions, demographic changes have begun to be noticed. For the academic year 2017-2018, public postsecondary undergraduates were considered majority minority institutions as the percentage of White students fell below 50% (Miller, 2020).

Public's perception of postsecondary education. Within any election cycle, one can only look back at the 2020 election to see how candidates portrayed postsecondary education regarding affordability or indoctrination issues. According to Lederman (2019), the public nationally does not view postsecondary institutions the way politicians do. While tending to have a positive view of higher education, the public has indicated postsecondary institutions "must do a better job of educating students affordably and effectively" (Lederman, 2019). Further adding, Lederman (2019) stated the public still view obtaining a postsecondary degree forms the stepping stone to a successful career. While the positive view is held by many, they also believe postsecondary education may not be worth it as the costs continue to increase, especially for students who take on substantial debt (Lederman, 2019). Even though Lederman indicated an overall positive perception of postsecondary education, Nietzel (2019a) stated the positive belief of importance has begun to erode. He noted only 50.0% of respondents to a Gallup poll expressed belief in the positive importance (Nietzel, 2019a). In comparing the percentage from 2013 to 2019, this belief declined by 20.0% (Nietzel, 2019a).

Development of Attrition Theories

Academic performance metrics of successful first-year GPA and retention to the second year of studies continue to puzzle postsecondary administrators over the years. Students' academic performance within the first years is critical in developing a powerful momentum to propel them towards their eventual degree attainment. Until the 1970s, researchers analyzed data about students' characteristics and attributes to explain attrition (Aljohani, 2016). According to Spady (1970), he stated these early attempts lacked an "analytical-explanatory category" describing the attrition phenomenon (p. 64). Berger, Ramirez, and Lyon (2012) indicated the previous attempts used a psychological lens rather than a sociological lens in providing explanations regarding attrition. While

William first began using a sociological lens to develop a theory regarding retention, Spady's work introduced the term retention and explained how institutions have a shared responsibility in the phenomenon (Aljohani, 2016).

Spady's undergraduate dropout process model. Developed by Spady (1970, 1971), the model drew from Durkheim's suicide theory's sociological lens. This sociological lens sought to define and explain the relationship between students and institutions in a social construct (Aljohani, 2016). Using the stages of suicide, Spady (1970, 1971) argued student attrition occurred when students did not successfully integrate into the postsecondary institution environments. According to Spady (1970; 1971), attrition results when students fail to integrate into the communities, whether academic or social, resulting in postsecondary suicide. He explained postsecondary suicide described the student's departure from the academic or social communities (Spady, 1970, 1971). Spady solidified his model by analyzing data collected from students at the University of Chicago to discover how student characteristics, academic potential and ability, family background, and social support influenced retention decisions (Aljohani, 2016).

Tinto's institutional departure model. Tinto's model has become one of the most recognized and used models to explain attrition. In 1975, Tinto began examining and expanding Spady's departure theory. Like Spady, Tinto used Durkheim's suicide theory to examine students' integration into the institution's academic and social communities. After some reflection, he revisited his theoretical model to use Van Gennep's theory on the rites of passage in 1993 (Tinto, 1993). The rites of passage theory by Van Gennep (1960) focused on three distinct phases individuals experienced:

separation, transition, and incorporation. Tinto stated new students integrating into a postsecondary institution's communities first had to experience separation from their hometown or community's traditional and familiar norms. Students began to interact with the new traditions and norms in the postsecondary setting during the transition phase. Finally, Tinto argued students began to incorporate the new norms and traditions into their daily life. Over several years, the model has been revisited for revisions and expansions by multiple researchers (Cabrera et al., 1992; Cabrera et al., 1993; Pascarella & Terenzini, 1979, 1980, 1983; Terenzini et al., 1981; Tinto, 1988).

Later, Tinto (1993) added that students' integration into the academic and social communities is measurable through academic performance—grades earned and persistence—and student interactions with others—peers and faculty. Before entering a postsecondary institution, students possess attributes shaping and impacting their goals and dreams of eventual degree attainment. Students' characteristics and prior schooling are ever-present effects that "weaken or strengthen" their commitment to their academic journey (Alojanhi, 2016, p. 6). Also, Tinto (1975, 1993) added environmental factors, such as family and occupation, may affect students' academic performance. The impact on a student's decision on whether to retain or depart stems from these attributes.

Bean’s student attrition model. While aligning with the prior theories, Bean (1980) criticized Spady's and Tinto's initial use of Durkheim's Suicide Theory. Bean's (1980) main critique was that the prior theories only examined correlations rather than providing in-depth analytical explanations between students' attributes and post-secondary institutions. Purporting Price's (1978) employee turnover theory, Bean (1980) indicated that student attrition was similar in terms of satisfaction and departure decisions in the workforce. While students do not receive any monetary value for their academic journey, they do receive grades, GPAs, and intellectual development affecting their overall satisfaction with the institution, resulting in their decision to retain or depart.

Astin's student involvement theory. Like other theories examining students' departure decisions and the postsecondary institution, Astin (1984) suggested student retention is also related to the level of involvement in the institution. Like Tinto's (1975, 1993) theory, Astin's (1984) involvement theory examined the students' involvement or integration into the postsecondary environments. Expanding on the use of student backgrounds and precollege attributes, Astin (1984, 1993) included institutional factors such as resource expenditures impact and improvements in academic performance. Institutions expending dollars and allocating resources to assist students with their integration into the academic and social communities could relate to the eventual academic performance of students with a measure of devotion and intentionality towards their success.

Characteristics Impacting Academic Performance

The theories mentioned above suggest consensus on the overarching themes of factors impacting academic performance. There are five main themes: 1) student characteristics, 2) precollege characteristics, 3) financial situations, 4) major declaration, and 5) institutional characteristics. These factors often interact and influence academic performance, especially in the critical first years of postsecondary education. The following section will review the literature on these overarching themes of factors influencing academic performance.

Student characteristics.

Gender. Research indicates gender plays a key factor in determining successful academic performance indicating females outpace males in attendance and academic performance in postsecondary education. Jacob (2002) investigated the gender differences in postsecondary attendance and retention using a representative sample from the National Educational Longitudinal Study (NELS). The data Jacob obtained focused on eighth graders from 1988 to 1994. Within the data, observations had a corresponding survey regarding postsecondary and workforce events after high school graduation. Of the 10,925 students, he found females had a 4.7% higher postsecondary attendance than males (Jacob, 2002). Jacob (2002) noted the rates were higher than the national rates but still confirmed females were attending postsecondary institutions at higher rates than their counterparts. In examining the differences in enrollment rates, the NELS survey indicated males were more likely to dislike school, be employed in the workforce, or see no further need for education than females (Jacob, 2002). Also, he noted males were more likely to attend based on their family background, while females were more likely to attend based on their cognitive ability (Jacob, 2002). Additionally, Jacob (2002) mentioned females were more likely to return than males. He further reported around 90% of the gender gap is accounted for in non-cognitive abilities (Jacob, 2002). In reviewing the results of his study, Jacob (2002) suggested future research should include the characteristics of how the school's curriculum influences attendance and retention by gender.

In reviewing enrollment trends from the early 1960s to the mid-2000s, Buchmann and DiPrete (2006) confirmed females were outpacing males in postsecondary attendance and academic performance. Moreover, they alluded to the inequalities of postsecondary access in earlier years contributing to males attending at higher rates (Buchmann & DiPrete, 2006). Using the data from NELS of children born in 1973 or 1974, they conducted a logistic regression model and reported females were more likely to attend any postsecondary institution (*N* = 10,820, *B* = .219, $p \leq .01$) and to be successful within four years (*N* = 10,759, *B* = .234, $p \leq .01$) (Buchmann & DiPrete, 2006). Buchmann and DiPrete (2006) found no significant differences between females and males in enrolling in a four-year institution. In examining the gender gap, they found significant differences for earning a degree for a four-year institution (*N* = 6,014, *B* = .368, $p \leq .01$) and enrolling only in four-year institutions (*N* = 3,512, *B* = .454, $p \leq .01$) (Buchmann & DiPrete, 2006). Their findings suggest female students have higher likelihoods of successfully obtaining a degree within four years regardless if the female students initially matriculated or transferred into a four-year institution when compared to their counterparts.

Gender alone is not a sole factor in impacting academic performance in terms of GPA. In controlling for race and family background for 5,032 observations, Buchmann and DiPrete (2006) conducted an ordinary least squares regression and reported gender as a significant predictor (*B* = .263, $p \leq .01$) of earned GPA. Additionally, they found students identifying as White exhibited a strong predictor of academic performance compared to their underrepresented counterparts. Specifically, students identifying as Black or African American (*B* = -.443, $p \leq .01$) exhibited a significant predictor of unsuccessful performance or earning lower GPAs (Buchmann & DiPrete, 2006). However, Buchmann and DiPrete (2006) conducted a second model with 4,249 observations incorporating high school preparation to report a lesser negative effect for students identifying as Black or African American (B = -.254, $p \leq .01$) (Buchmann & DiPrete, 2006). When factoring in academic preparation, especially for Black or African American students, Buchmann and DiPrete (2006) stated students who had higher marks of preparedness earned higher college GPAs while students who were not as prepared earned lower GPAs. Students with at least one parent with some postsecondary education was a significant predictor (mother, B = .147, $p \leq$ .01and father, B = .149, $p \leq .01$), when controlling for race and family background (Buchmann & DiPrete, 2006). When including high school preparation, Buchmann and DiPrete (2006) reported only the father's postsecondary education was significant (B = .084, $p \leq .01$). While confirming females outperform their counterparts, Buchmann and DiPrete (2006) stated the gender gap largely stems from White female students.

In analyzing additional factors impacting the gender gap, Morales (2008) conducted a phenomenological study using a purposeful sample of 50 individuals. In his study, the findings revealed 93% of females having a conscience and intense connection between their degree program and their future occupation produced a strong motivation to perform successfully. Additionally, Morales (2008) reported 92% of females indicated they had clear professional goals compared to 30% of males. In the study, females largely reported resistance and harsh criticism from family members and others regarding pursuing a postsecondary degree. Specifically, underrepresented females reported more resistance than White females. Morales (2008) suggested this form of resistance is due to the cultural and traditional norms expected of females. Further examining the connectivity between the programs of study and future occupations, a qualitative study conducted by Kleinfield (2009) of 99 students confirmed females view postsecondary education as a pathway for their future occupation. Noting a lack of connection between major and occupation, Kleinfeld (2009) also recommended male students receive early preparation through dual enrollment courses in high school to help to form the connection to build momentum toward performing well in a postsecondary setting.

In terms of returning or departing from the institution, research has consistently indicated females retain at a higher rate than males. A study conducted by Stewart, Lim, and Kim (2015) analyzed the effect of remediation status, personal attributes, family background, prior schooling, and college academic performance impacted students' decisions on retaining or departing from the institution. Using a large public four-year institution, they obtained 3,212 student observations and analyzed the data with a factorial analysis of variance (Stewart et al., 2015). Stewart et al. (2015) conducted three factorial ANVOAs comparing separately how gender, race and ethnicity, and financial aid status along with remediation status impacted students' retention. The first factorial ANOVA study revealed only remediation status had a significant impact on persistence ($F(1, 3{,}212) = 3.948$, $p = .047$, $\eta^2 = .001$), while gender alone and the gender and remediation interaction had no impact (gender, $F(1, 3{,}212) = .399$, $p = .528$; gender and remediation, $F(1, 3{,}212) = 1.065$, $p = .302$) (Stewart et al., 2015). The second factorial ANOVA found only race and ethnicity ($F(4, 3{,}212) = 8.386$, $p < .01$, $\eta^2 = .010$) and the third factorial ANOVA found only financial aid status ($F(1, 3{,}212) = 12.825$, $p < .01$, $\eta^2 = .004$) influenced students retaining (Stewart et al., 2015). In both the second and third, remediation and the interaction with remediation were not found to be significant (Stewart et al., 2015). In the breakdown of the race and ethnicity, Stewart et al. (2015)

reported students' race and ethnicity and followed up with Games-Howell post hoc comparison. They reported Asian or Pacific Islander students (*M* = 4.97, *SD* = 1.394) had the highest and Black or African American students (*M* = 4.87, *SD* = 1.604) had the second highest likelihoods of persistence (Stewart et al., 2015). The findings from Stewart et al. (2015) indicated students' race and ethnicity and financial aid status contribute to a students' likelihood of persistence, while gender is not a contribution to the persistence likelihood. Yet, it is important to note the small effect sizes indicate very limited partial applications of the analyses.

Race and ethnicity. According to the Association of American Colleges and Universities (AACU) (2019), the student body in postsecondary institutions has become more diverse than in previous years. For underrepresented groups enrolled in a postsecondary degree, a 24.4% increase occurred from 1996 to 2016. The reports revealed most of the growth in the underrepresented groups resulted from students identifying as Hispanic (AACU, 2019). As the postsecondary student population is growing to become a majority minority, administrators and other policy-makers need to be aware of how race and ethnicity influence academic performance within the critical first years. Underrepresented groups of students may experience difficulties in performing successfully and integrating into the communities due to the lack of representation in the student body, faculty, and staff (Odell, Korgen, & Wang, 2005; Seidman, 2007).

Using data from the National Longitudinal Survey of Freshmen on student entering institutions in 1999, Fischer (2007) conducted a three-step survey of a stratified random sample of 3,924 first-time students to measure the influencing factors on earned

GPA and retention status. She noted people of color were oversampled in order to conduct within group analysis (Fischer, 2007). The first step involved collecting data on students' demographics, family background, and high school experience. The second and third surveys included the collection of students' adjustments to coursework, interaction with others on campus, integration into the campus communities, and experience with discrimination (Fischer, 2007). Within the collected data, Fischer (2007) reported the mean earned GPAs were the highest in Asian and White students (3.30 and 3.32, respectively). Additionally, she reported Hispanic and Black or African American students had significantly lower earned GPAs (3.08 and 2.95, respectively) (Fischer, 2007). Fischer (2007) conducted four ordinary least square models based on the students' race and ethnicity in order to determine how the factors impact students' GPA (White, $N = 891$, $R^2 = .209$; Asian, $N = 871$, $R^2 = .210$; Hispanic, $N = 820$, $R^2 = .216$; and Black or African American, $N = 885$, $R^2 = .225$). Fischer (2007) noted significant differences in GPA by race and ethnicity. For White, Hispanic, and Black or African American students, gender exhibited a significant negative influence for males on earned GPA (White, $B = -.07$, $p < .01$; Hispanic, $B = -.106$, $p < .05$; and Black or African American, $B = -.096$, $p < .001$) (Fischer, 2007). Furthermore, White ($B = -.134$, $p < .01$) and Hispanic ($B = -.147$, $p < .01$) first generational students exhibited a significant negative influence on the earned GPA (Fischer, 2007). Across the four different populations, high school GPA was the strongest factor impacting the earned GPA (White, $B = .483$, $p < .001$; Asian, $B = .610$, $p < .001$; Hispanic, $B = .506$, $p < .001$; and Black or African American, $B = .412$, $p < .001$). Fischer (2007) also noted only Black or African American students benefited from the number of Advanced Placement courses (B =.023, p < .05) (Fischer, 2007).

Like her analysis on GPA, Fischer (2007) conducted four logistic regression analyses based on students' race and ethnicity to determine how factors impacted students' decision to depart from the institution. Only Hispanic students experienced a significant contribution of high school GPA (B = -.221, p < .01) in their decision to depart from the institution (Fischer, 2007). Fischer's (2007) explained while Hispanic students have higher departure rates than other students, she mentioned there is a strong association between academic preparation and departing from the institution. She further explained Hispanic students with stronger academic preparation, in terms of better high school grades, contribute to the student actually not wanting to depart from the institution (Fischer, 2007). Having on campus connections to the students' own peer group contributed significantly to students' departure decision from the institution for all race and ethnicities (White, B = -.889, p < .05; Asian, B = -1.651, p < .001; Hispanic, B = -1.559, p < .01; and Black or African American, B = -.1.195, p < .01) (Fischer, 2007). The contributions of the findings from Fischer (2007) suggested students in all four race and ethnicity groupings having connections to their own social groups assisted in students not departing. Interestingly enough, Asian and Hispanic students' own peer group have stronger impact on departure decisions than Black or African American and White students (Fischer, 2007).

Flores and Park (2013) examined the type of institution underrepresented students had a preference to enroll in to obtain their degree in Texas. The institution preference examined whether underrepresented students would prefer to enroll in Historically Black

Colleges and Universities (HBCU), Hispanic Serving Institution (HSI), or a Predominately White Institution (PWI). Using the longitudinal data set from Texas agencies and supplemental national data, Flores and Park (2013) filtered the population to students graduating from high school graduates in 1997, 2000, 2002, 2006, and 2006. Utilizing logistic regression, Flores and Park (2013) reported consistent findings with previous literature indicating females were more likely than males to enroll in a postsecondary institution (males in 2017, B = -.233, p < .001; males in 2000, B = -.220, p < .001; males in 2002, B = -.174, p < .001; males in 2006, B = -.211, p < .001; and males in 2008, B = -.191, p < .001). Additionally, they reported Hispanic students are more likely to not enroll in any postsecondary institution (2017, B = -.185, p < .001; 2000, B = -.179, p < .001; 2002, B = -.328, p < .001; 2006, B = -.288, p < .001; and 2008, B = -.242, p < .001) (Flores & Park, 2013). For high school graduates identified as Black or African American, Flores and Park (2013) noted these students exhibited increasing odds of enrolling over the years (2017, B = -.061, p < .001; 2002, B = .049, p < .001; 2006, B = .247, p < .001; and 2008, B = .232, p < .001). Flores and Park (2013) next examined the types of postsecondary institutions students attend. They noted White students were more likely to enroll in PWI. Black or African American students are more likely than Hispanic or White students to enroll in one of the three types of institutions, while the odds are less for attending an HSI (Flores & Park, 2013). Interestingly, they noted Hispanic students were more likely to enroll in an HSI or HBCU than a PWI (Flores & Park, 2013). After examining the differences in persisting toward degree attainment, Flores and Park (2013) noted towards the end the academic journey, race and ethnicity no longer played a factor in determining successful performance. One aspect could be drawn from the success of students of color from the Flores and Park's (2013) study could be the students enrolling in minority-serving institutions experience self-identification to help integrate into postsecondary academic and social communities.

Stewart et al. (2015) conducted three factorial ANVOAs comparing separately how gender, race and ethnicity, and financial aid status along with remediation status impacted students' retention. In the factorial ANOVA examining the impact towards persistence using race and ethnicity and remediation status, only race and ethnicity was found to be significant, ($F(4, 3{,}212) = 8.386$, $p < .01$, $\eta^2 = .010$) (Stewart et al., 2015). Remediation status along with race and ethnicity was not found to be significant (Stewart et al., 2015). Stewart et al. (2015) reported the Games-Howell post hoc comparison resulted in Asian or Pacific Islanders ($M = 4.97$, $SD = 1.394$) had the highest and Black or African American students ($M = 4.87$, $SD = 1.604$) had the second highest likelihoods of persistence. Yet, it is important to note the small effect sizes indicate very limited partial applications of the analyses (Stewart et al., 2015). Stewart et al. (2015) also conducted two additional factorial ANOVAs analyzing the impact of gender and financial status along with remediation status had on retention. The gender and remediation status only revealed remediation had a significant impact, ($F(1, 3{,}212) = 3.948$, $p = .047$, $\eta^2 = .001$), while gender alone and gender and remediation status were not significant (gender, $F(1, 3{,}212) = .399$, $p = .528$; gender and remediation, $F(1, 3{,}212) = 1.065$, $p = .302$) (Stewart et al., 2015). The other factorial ANOVA found only financial aid status ($F(1, 3{,}212) = 12.825$, $p < .01$, $\eta^2 = .004$) influenced students retaining (Stewart et al., 2015). In both the second and third, remediation and the interaction with remediation were not found to be significant (Stewart et al., 2015). Like the race and ethnicity, the effect size was small in each of the additional findings indicating very little applicable implications (Stewart et al., 2015).

Family educational background. As post-secondary enrollment increased, enrollment of first-generation students also increased, providing different academic performance behaviors. While there are varying definitions, first-generation students are considered those individuals with parents or guardians who have not graduated college. Obtaining data from a four-year comprehensive university located in the Midwest, Ishitani (2003) examined the impact of first-generation status on retention. Within his study, the population focused on new students for Fall 1995. Ishitani (2003) reported first-generational students comprised 58% of the students. Using a survivor function, Ishitani (2003) noticed first generation students experienced a sharp decline compared to their counterparts. First-generational students after the first semester exhibited a .833 survival rate, while students with one parent exhibited .898 and those with both parents exhibited .913 (Ishitani, 2003). When comparing first generational students to the counterparts, the rate difference was -.065 for those with one parent and -.080 for those with both parents (Ishitani, 2003). The rate difference indicated first generational students were more likely to depart. Examining the survival rate beyond the first year, Ishitani (2003) reported the gap in the survival rates grew larger for first generation students. Further analyzing the data, Ishitani (2003) conducted a piecewise exponential model. He noted first generation students had higher odds of leaving the institution after the first year ($B = .534, p < .05$) (Ishitani, 2003). More importantly, he noted first generation students' odds of departing decrease as they continue to progress towards degree attainment (Ishitani, 2003). Factoring in additional variables, he reported race and ethnicity (*B* = -.557, *p* < .05), annual income of $25,000 or less (*B* = .400, *p* < .05), high school GPA (*B* = -.554, *p* < .05), and college GPA less than 2.00 (*B* = 1.356, *p* < .05) were significant factors in the first year of students (Ishitani, 2003). The departure odds of first generation students decline after the first year could be attributed to the student integrating into the academic and social communities. First generation students and their parents are unfamiliar with how to navigate through the communities which could provide hesitancy and anxiety into initially integrating into an institution's communities.

Conducting a second study on first-generational students, Ishitani (2006) used the NELS and Postsecondary Education Transcript Study data sets. The population consisted of a sample of students from the 1991, 1992, 1993, and 1994 new students from four-year postsecondary institutions. Ishitani (2006) used a two-tier definition of first-generation: parents with no college and parents with some but no degree. This population represented 14.7% and 34.8% of the sample (parents with no college and parents with some college, respectively). Ishitani (2006) found similar findings to the study conducted in 2003. He indicated students with parents without any college have the lowest survival rate, and the survival gap grows over the years compared to students with both parents with a degree (Ishitani, 2006). In conducting an exponential model, Ishitani (2006) found statistically significant higher odds of departing for students with no parents with college experience (*B* = .712, *p* < .05) and students with at least one parent with some experience but no degree (*B* = .739, *p* < .01) (Ishitani, 2006). He also stated family income was significant for students with family annual income of less than $35,000 ($0 to $19,999, *B* = 1.193, *p* < .01 and $20,000 to $34,999, *B* = .874, *p* < .01) (Ishitani, 2006). In terms of financial aid, students who received any grants (*B* = -.465, *p* < .01) and work study funding ($B = -.529, p < .05$) had lower odds of departing (Ishitani, 2006). Students low income families who did not receive any grant or work study funding would have higher odds of departing. These students would not be able to afford the cost of attendance. Ishitani (2006) also found preschooling academic preparation impacted decisions to depart. He found significant contributions of students with lower high school rank ($B = 1.337, p < .01$) and lower academic intensity (3rd quintile, $B = .599, p < .01$; 4th quintile, $B = .605, p < .01$; and 5th quintile, $B = .850, p < .01$) significantly contributed to students' departure (Ishitani, 2006).

A study conducted by Lohfink and Paulsen (2005) analyzed national data from the Beginning Postsecondary Student Longitudinal Survey to compare persistence of first generation and non-first generation students. First generation students exhibited a 5.63% persistence gap when compared to their counterparts. Using a logistic regression model, Lohfink and Paulsen (2005) found significant findings for first generational and continuing generational students who persisted. For the first generational student model, gender (*delta-p* = .094, $p < .01$), Hispanic students (*delta-p* = -.354, $p < .01$), family income (*delta-p* = .002, $p < .05$), living at home (*delta-p* = .183, $p < .01$), institution control (*delta-p* = -.124, $p < .05$), institution enrollment size (*delta-p* = .004, $p < .05$), first year GPA (*delta-p* = .128, $p < .01$), grant aid (*delta-p* = .0272, $p < .01$), and work study funding (*delta-p* = .064, $p < .05$) were found to have significant contributions to first generational students' likelihood of persisting (Lohfink & Paulsen, 2005).

Locale. With the enrollment increase, students from different regional areas began enrolling in postsecondary institutions due to the increased access. These regional areas are urban, suburban, and rural. Early research from Corley, Goodjoin, and York

(1991) indicated how underrepresented urban and rural students performed in their first year at a South Carolina postsecondary. A total of 760 freshmen across 21 course sections from 1988 and 1989 were surveyed. Urban students' mean SAT Verbal scores were 349.06 for 1988 and 329.64 for 1989, and mean SAT Math scores were 363.34 for 1988 and 355.16 for 1989 (Corley et al., 1991). For the rural students, the mean SAT Verbal scores were 327.61 for 1988 and 314.65 for 1989, and the mean SAT Math scores were 352.12 for 1988 and 345.83 for 1989 (Corley et al., 1991). Corley et al. (1991) noted rural students' SAT Verbal scores were significantly lower than their urban counterparts in 1988. Corley et al. (1991) also noted the earned GPA in high schools was significantly lower for rural students than for urban students. However, they reported the college GPA for rural students was "nearly two-tenths of a point" higher than their urban counterparts (Corley et al., 1991, p. 176). Corley et al. (1991) mentioned the lower admission test scores of rural students could be attributed to the availability of test preparation for the region. In their discussion, Corley et al. (1991) stated it appeared as if only the strongly motivated rural students would enroll in a postsecondary institution based on significantly lower test scores but higher earned GPAs. According to Corley et al. (1991), rural students will struggle to integrate into the social communities due to the disparities between the rural and institution environments.

While research from Corley et al. (1991) examined how underrepresented students from urban and rural, the impact extends to students of other races and ethnicities from like regions. Schultz (2004) conducted a phenomenological study on the first generation rural students enrolled at Mesa State College who came from agricultural families and their experiences in their first semester at a postsecondary institution.

Schultz (2004) reported students who had parents that supported the student's decision to attend college decided to pursue a postsecondary education easier than the other parents. He also noted when the support was split between the parents, it made the decision harder on whether to attend (Schultz, 2004). The father was the most reported parent who made the student's decision harder. The students reported the father either did not want the student to "leave the farm" or did not want to sign the required financial aid forms (Schultz, 2004, p. 48). While Schultz (2004) noted some students had positive views of moving from a small rural area to a larger, more populated area, most students reported they experienced anxiety about the move. Schultz (2004) concluded the assimilation process into the new community was counterproductive. With the horrible assimilation process into the campus community, Schultz (2004) reported the students were unaware of the reason they needed to form new relationships, e.g., friends and study peers, to help in adjusting to their new environment.

Fischer (2007) conducted a study using data from the National Longitudinal Survey of Freshmen matriculating into selective postsecondary institutions in 1999. He examined how college involvement and academic performance were impacted by the race and ethnicity of students (Fischer, 2007). Using a stratified sample, a total of 3,924 face-to-face interviews were conducted. The students underwent two phases of data collection. The first phase was conducted at the beginning of the first year, while the second occurred in the first spring semester. Fischer (2007) used logistic regression to produce estimates by race and ethnicity groups by the whether the student departed from the institution by the third year. She noted Black or African Americans from urban areas attending a postsecondary institution in a like area experienced a 60% reduction in the probability of departing from the institution when "compared to their counterparts on nonurban campus" (Fischer, 2007, p. 151).

Education in rural areas. In an NCES published report titled the Status of Education, Provasnik, KewalRamani, Coleman, Gilbertson, Herring, and Xie (2007) analyzed the impact on education in rural areas. Around 6% of rural students attended a private K-12 school in the 2003-2004 academic year, meaning most rural students attend a public school. These rural public schools receive less federal funding when compared to their urban counterparts (6% and 11%, respectively) (Provasnik et al., 2007). With the small amount of federal funding, the remaining portion of the funding would come from the state and local governments. However, around 45% of the public schools in rural areas are classified as moderate-to-high poverty schools (Prosvasnik et al., 2007). This would indicate rural area schools may not receive adequate funds to maintain updated education resources and to have competitive pay for qualified teachers. Rural schools pay teachers on average 1,000 to 2,000 less than urban and suburban teachers (Provasnik et al., 2007). With how technology has advanced to provide more available resources, rural students were reported to suffer lower rates of a computer with internet access than urban and suburban counterparts (Provasnik et al., 2007).

Based on the National Assessment of Educational Progression achievement marks, Provasnik et al. (2007) reported the percentage of students in reading, mathematics, and science for 2005 by regional areas. The areas were classified as city, suburban, town, and rural. Further adding to the complexity, they further defined rural as fringe, distant, and remote. The fringe areas were defined as a minimum of five miles, the distance was defined as more than five miles but less than or equal to 25 miles, and remote was more than 25 miles from an urbanized area. For distribution at a basic level or below, rural schools' distribution of students ranked third in reading, second in mathematics, and second in science in terms of the highest percentage. Nevertheless, the rankings differed when rural schools are further split by their fringe, distant, and remote status. Distant rural schools ranked the second highest in reading proficiency at basic or lower. Remote ranked first, and distant ranked third highest in mathematics proficiency levels at basic or lower. Remote ranked second and distant ranked third highest in science proficiency levels at basic or lower (Provasnik et al., 2007).

Provasnik et al. (2007) reported only 27% of rural students from ages 18 to 24 were found enrolled in any program in a postsecondary institution. Of the same age group, males from rural locations were reported at lower rates than females from rural locations (23.1% and 31.5%, respectively). Even though rural females were enrolling in postsecondary institutions, they lag females from the other locations (Provasnik et al., 2007). The low percentages of rural students attending postsecondary institutions could be impacted by the parents or family’s expectation of the student to pursue a degree. Students from rural areas are more likely to come from a household in which both parents do not have a bachelor’s degree (Grace et al., 2006; Provasnik et al., 2007). More specifically, Black or African American rural students' parents may not have even graduated from high school (Grace et al., 2006). Provasnik et al. (2007) reported 42% of parents from a rural expect their children to earn less than a bachelor’s degree. This expectation is lower when compared to their urban and suburban counterparts (30% and 25%, respectively) (Provasnik et al., 2007).

While conducting an epistemological qualitative study consisting of 21 students at the University of Louisville, Phillips (2015) reported rural students are more likely to be first generation students. Additional findings from his research indicated these students typically have very limited to no support from family and others back home. This is due to those back home being unfamiliar with life on a postsecondary campus. For rural students not receiving support from those back home, they made valuable connections on campus with faculty, staff, and peers. Using a logit model on a sample of 6,748 observations from the National Longitudinal Study of Youths, Velez (2014) reported a bleak outlook in postsecondary education for students from rural locations. She reported 91.9% of students from rural areas who never attended and 62.3% of rural students who dropped out of a four-year postsecondary institution had a probability of 50% or less to obtain a bachelor's degree (Velez, 2014). For students initially starting at a community college, Velez (2014) found 86.5% of rural students had a probability of earning a bachelor's degree at 50% or less.

In an article published by the Lumina Foundation (2019), the foundation reported students from rural areas were graduating from high school above the national rate but just under the students from the suburban regions. While students from rural regions were graduating from high schools above the national rate, the rate of matriculating into a postsecondary institution lags their urban and suburban counterparts (Lumina Foundation, 2019). The Lumina Foundation's (2019) review of the National Center for Educational Statistics (NCES) found only 59% of rural students who graduate matriculate into a postsecondary institution, which is 3% lower when compared to urban students and 8% lower when compared to urban and suburban, respectively, students. Furthermore, the Lumina Foundation reported, based on the most recent National Student Clearinghouse report, rural students are more likely to depart from a postsecondary institution.

Additionally, less than 20% of individuals living in rural communities have at least a bachelor's degree (Lumina Foundation, 2019). According to the Lumina Foundation (2019), historically, there have been multiple obstacles contributing to why rural students do not matriculate and eventually earn some form of postsecondary credential. The main overarching reason for the lack of wanting to attend and eventually earn a degree is the problem with access. This access comes in either the physical distance to the postsecondary institution or a lack of broadband internet (Lumina Foundation, 2019). In an *Inside Higher Ed* article, Fain (2019) stated only 14% of the postsecondary institutions are in rural areas, even though rural counties comprise 97% of the land in the U.S. This lack of institutions in rural areas has created what is known as ‘education deserts’ (Fain, 2019, Lumina Foundation, 2019). Fain (2019) indicated students located in the rural education deserts present a barrier to which students feel there is no possibility of obtaining a degree and thus “perpetuating the cycle of poverty.”

As most of the occupation in rural areas are blue-collar and requiring no further education beyond high school, the available job market within the area could contribute to the low postsecondary attendance and degree attainment rates (Lumina Foundation, 2019). Lastly, the Lumina Foundation (2019) reported impacting rural students' attendance and success in a postsecondary institution was the high school in the area. More specifically, the foundation mentioned the high school curriculum quality left them ill-prepared to handle the postsecondary coursework (Lumina Foundation, 2019). The

Lumina Foundation (2019) also mentioned the same reason impacting rural students is often the reason expressed by low-income urban student.

Precollege characteristics.

High school curriculum quality. Spady (1970, 1971) and Tinto (1975, 1988, 1993) mentioned precollege schooling characteristics are important factors in determining academic performance in a postsecondary institution. As Tinto (1993) mentioned incongruences in terms of mismatching attributes between students and postsecondary institutions, a connection between a weaker high school curriculum could cause students to feel as if they are mismatched, resulting in an inferior ability to perform at a higher rigor of postsecondary coursework. Alternatively, students from a higher high school curriculum rigor could also result in a mismatch to which the student departs as the postsecondary coursework rigor is not challenging enough. One of the most overlooked precollege schooling characteristics is the quality or characteristics of the high school. A multitude of research has only focused on the individual's ability in high school based on the HS GPA, *per se*, rather than including the overall quality of the high school's curriculum. In her book, McDonough (1997) used a qualitative approach to examine college choice using a sample of 24 females from four high schools. She noted explanations or reasons driving students to select a college are diverse and cannot fit neatly into one category. Furthermore in her book, McDonough (1997) noted educational settings are not on an equal playing field. She stated students attending more selective postsecondary institutions have higher chances of graduating and more opportunities after obtaining their degree than their counterparts. McDonough (1997) extended the same reasoning to the high schools, where she indicated more elite secondary schools lead to better opportunities down the road for attending postsecondary institutions. The attributing factors for the students who attend more elite secondary schools having better opportunities are the parents and the school's available resources (McDonough, 1997). The connection back to the high school would indicate the quality of the curriculum to prepare students could potentially limit the students' availability to postsecondary institutions and eventually be connected to the impact on the actual or perception of their ability to have successful academic performance while enrolled in the institution.

In their report for NCES, Horn and Kojaku (2001) published findings regarding the impact of high school curriculum had on persisting at a postsecondary institution. Using data from the Beginning Postsecondary Students Survey of the 1995 to 1996 cohort, they found only 8% of Black or African Americans and 9% of students with parents' education had high school education or less and took a more rigorous high school curriculum (Horn & Kojaku, 2001). Students taking the advanced high school curriculum were more likely to attend more selective institutions, while the students who took the mid-level high school curriculum or less were more likely to attend less selective institutions (Horn & Kojaku, 2001). Around 19.3% of students with the bare minimum high school curriculum were enrolled in remedial coursework, compared to only 2.7% of students with a rigorous curriculum. Horn and Kojake (2001) reported the average first year GPA was 2.53 for students with the basic curriculum, 2.67 for students with a mid-level curriculum, and 3.10 for students with a more rigorous curriculum. They noted the only measurable difference appeared when comparing the basic or mid-level to the more rigorous students. Students in the basic or mid-level curriculum were three times more likely to earn a lower first year GPA than students with a more rigorous high school curriculum. In examining the impact on first-year retention at the initial institution, Horn and Kojaku (2001) reported 71% of students from the more rigorous curriculum were retained as compared to 62% of the mid-level curriculum and 55% of the basic curriculum.

Choy (2002) produced a report regarding her findings examining 10 years of longitudinal data. From the National Education Longitudinal Study of eighth grade cohort data set in 1988, students who took more intense classes, especially Mathematics courses, in high school helped offset the students' likelihood of attending college, even when their parents did not (Choy, 2002). Examining the impact of advanced Mathematics courses, Choy (2002) reported 64% of students with parents with no college, 70% of students with parents with some college, and 85% of students with parents with at least a bachelor's degree enrolled in four-year postsecondary institution. Choy (2002) further indicated students with parents who did not attend college but took advanced Mathematics courses were twice as likely to attend college than their counterparts who only took up to Algebra II. Further elaborating, students who took the more rigorous Mathematics courses were academically prepared from the beginning rather than later in their high school academic journey (Choy, 2002). Using the Beginning Postsecondary Student Longitudinal study of the 1995 to 1996 cohort of students, Choy (2002) also found the curriculum contributed to students' persistence once attending a postsecondary institution. She reported after three years of matriculating into the institution, 87% of students who had a more rigorous high school curriculum were persisting at their initial institution or transferred to another institution (Choy, 2002). In comparison, only 62% of students who only had the basic high school curriculum were found to be persisting or transferred (Choy, 2002).

DeNicco, Harrington, and Fogg (2015) examined factors impacting student first year retention status of 1,800 students in a public state community college. One of the factors they examined was the high school characteristics and the impact they had on the students' academic performance. Due to the limitation of only examining the high schools from the same state as the attended institution, the sample size was reduced to 1,638. DeNicco et al. (2015) incorporated seven characteristics into their analysis. The characteristics were graduation rates, dropout rates, number of suspensions, attendance rates, and proficiency rates in Mathematics, English & Languages Arts, and Writing (DeNicco et al., 2015). DeNicco et al. (2015) indicated the proficiency rates were the school's overall rates due to their inability to have access to the individual student proficiency ratings. They indicated the result involved with these factors are attributed towards the high school's environment the students went to rather than to the students' abilities. In other words, the results could contribute back the environment in terms of the quality of education provided to all students attending the high school. When they averaged out the rates by students' retention status, DeNicco et al. (2015) noticed the rates were significantly different at the 1% threshold for graduation rates, dropout rates, attendance rates, and English & Language Arts and Mathematics proficiency rates. The average Writing proficiency rate was significant at the 5% threshold. English & Language Arts proficiency rate for retained students was 3.3% higher than students who did not retain, while the difference was 2.1% for Mathematics, and 1.9% for Writing proficiency rates (DeNicco et al., 2015). Using a logistic regression model for marginal effects, DeNicco et al. (2015) found only two proficiency rates to have significant marginal effects (English & Language Arts, marginal effect = .003, $p < .001$ and Mathematics, marginal effect = .003, $p < .001$).

High school GPA. In the formation of their theoretical models on student retention, high school GPA was a major component of academic preparation in the precollege schooling characteristics (Astin, 1975, 1993; Bean, 1980, 1983; Spady, 1970, 1971; Tinto, 1975, 1988, 1997). A study conducted by Bridgeman, Pollack, and Burton (2008) confirmed the relationship between high school GPA and first year GPA when examining a sample consisting of three different cohorts from 26 colleges totaling 110,468 students. In addition to the high school GPA, Bridgeman et al. (2008) included SAT test scores as predictors of first year GPA. In using regression analyses, they denoted a strong correlation between high school GPA and freshmen year GPA (B = .58, p < .05) (Bridgeman et al., 2008). The correlation for males was slightly lower than females (B = .56, $p < .05$ and B = .59, $p < .05$, respectively) (Bridgeman et al., 2008). White students exhibited the highest correlations (males, B = .57, $p < .05$ and females, B = .61, $p < .05$), while Black or African American (males, B = .41, $p < .05$ and females, B = .48, $p < .05$) and Hispanic (males, B = .51, $p < .05$ and females, B = .55, $p < .05$) students were the lowest (Bridgeman et al., 2008). The correlation of the freshmen's GPA strengthens when combining high school GPA and admission test scores (B = .65, $p < .05$) (Bridgeman et al., 2008).

In explaining the relationship of high school GPA to academic performance, Chen and St. John (2011) found it was a significant predictor for students with lower high school GPAs are more likely not to persist when compared to those with higher GPAs.

Chen and St. John (2011) obtained a sample of new students in 1996 from the Beginning Postsecondary Students database. They also obtained the students' data for the following five years after the matriculation year. Overall, the population for the study consisted of 6,383 students from 422 postsecondary institutions. Chen and St. John (2011) supplemented the data from additional national databases to incorporate financial indicators in the study. In analyzing the data, a hierarchical generalized linear model examined the impact of factors on students' persistence towards degree attainment (Chen & St. John, 2011). Only 12.0% of the students had a low high school GPA. Additionally, 46.0% of students were in a middle level high school GPA. Within their full model, only high levels of high school GPA were found to be significant (*OR* = 1.54, $p < .01$) (Chen & St. John, 2011). While Chen and St. John's (2011) examined persistence toward degree attainment, the study indicated high school GPA has a greater impact on the academic performance beyond the first year of study.

In 2012, the Governor's Office of Student Achievement (GOSA) (2012) in the State of Georgia analyzed the strength of the academics in high school impact on first year academic performance in an in-state public postsecondary institution. GOSA (2012) obtained data on new freshmen in 2006, 2007, and 2008 from USG. GOSA (2012) supplemented the sample data with high school characteristics. Splitting the data into two samples, GOSA (2012) conducted a linear regression model to examine the effects of (1) high school graduation test (HSGT) and (2) end-of-course test (EOCT) along with student characteristics on the first year GPA (HSGT, $N = 75,761$; EOCT, $N = 55,833$) (GOSA, 2012). The mean high school GPA of the two samples were the same (HSGT, $M = 3.20$, $SD = .513$; EOCT, $M = 3.19$, $SD = .516$) (GOSA, 2012). In both models, high school GPA contributed the strongest contributions to the first year GPA (HSGT, *B* = 1.305, *p* < .01; *EOCT*, B = 1.256, *p* < .01) (GOSA, 2012). GOSA (2012) indicated the marginal effect of high school GPA on first year GPA was .535 points. Additionally, GOSA's (2012) found female students were more likely to have higher first year GPA for both models (HSGT, *B* = .108, *p* < -.01; EOCT, B = .119, *p* < .01). Students identified as other race exhibited a negative significant factor in both models (HSGT, *B* = -114, *p* < .01; EOCT, *B* = -109, *p* < .01). While statistically significant in both models, students identified as Black or African American, disabled, and economically disadvantaged in addition to the admissions test scores and the high schools' percentage of Asian and Hispanic students exhibited a small impact on first year GPA.

With concerns about providing a more rigorous and college-preparatory curriculum, Allensworth and Clark (2020) confirmed high school GPA continues to be the strongest predictor of academic performance until degree completion in a postsecondary institution. Allensworth and Clark (2020) examined the impact of high school GPA on persistence towards degree attainment. The population utilized in the study was recent high school graduates from Chicago Public Schools and consisted of 17,753 students. In the first model, Allensworth and Clark (2020) examined the impact of grouped high school GPA on persistence until degree attainment. Without controlling for any demographics and institutional characteristics, students with at least a 3.25 had significant odds of persisting (HS GPA 3.25-3.50, *OR* = 1.15; HS GPA 3.50-3.75, *OR* = 3.65; and HS GPA 3.75-4.00, *OR* = 6.23) (Allensworth & Clark, 2020). When factoring in demographics and institutional characteristics, odd ratios for HS GPA 3.50-3.75 (*OR* = 2.56) and 3.75-4.00 (*OR* = 3.74) lessened (Allensworth & Clark, 2020). Elaborating on why high school GPAs were a more significant predictor of academic performance, Allensworth and Clark (2020) argued the high school GPA aggregates the performance across different schooling components, thus being a better indicator of students' ability to perform successfully in postsecondary coursework. They further indicated students who worked to improve their high school GPA exhibit signs of preparedness to successfully handle the rigor of postsecondary course work (Allensworth & Clark, 2020). While they did not factor in the impact of more rigorous courses, Allensworth and Clark (2020) suggested Advanced Placement and honors may impact high school GPA. They suggested the more rigorous the course in high school may result in lower high school GPA, yet students would be more prepared for postsecondary coursework (Allensworth & Clark, 2020).

Grade inflation concerns. According to a study conducted by Bowers (2011), he reported high school grading systems are based on teachers' perception of how well a student does at playing school and the lack of a consistent, systematic grading structure. He also included the grading system may incorporate how well the student plays or fits within the model of student behavior (Bowers, 2011). Furthermore, Bowers (2011) noted even some well-gifted students were given low grades due to failure to meet the perceived model of exemplary behavior. Bowers (2011) examined a multi-dimensional relationship between teachers assigning grades versus standardized testing using data from the Education Longitudinal Study. He reported high school grades and the English and Mathematics standardized test scores exhibited a strong, moderate correlation ($r(14,520) = .572$) (Bowers, 2011). This finding indicated teachers' grading assignments for core courses are related to the core concepts and knowledge. Using a multi-dimensional scaling, Bowers (2011) indicated the initial fit was well with a low stress value (.097), which explains 96.6% of the variation in the data. He stated grades from 9th and 10th grade clustered closer to the standardized test scores than the 11th and 12th grade grades. Bowers (2011) contributed this distinction related to the timing of testing and the spacing of when core courses are taken in the early high school grade levels compared to the later high school grade levels. In a deeper analysis of high school grading by subjects, Bowers (2011) examined the impact of grading in Mathematics, English, Science, Social Studies, Art, Foreign Language, and Physical Education courses. He reported the multi-dimensional scaling fit extremely well with a very low stress value (.012), accounting for 99.9% of the variance (Bowers, 2011). Bowers (2011) suggested the placement of the core subject being diagonal to the standard test could be associated with non-cognitive knowledge in grade assignment.

A concern of grade inflation is prevalent in any discussion of high school GPAs, especially when a state offers merit-based financial aid related to high school graduates' GPAs (e.g., Georgia's HOPE Scholarship). Studies have consistently confirmed high school GPAs have increased over time (Camara et al., 2004; Gershenson, 2018; Hiss & Franks, 2014; Hurwitz & Lee, 2018), but some of the research does not contribute to the rise in the GPAs to grade inflation (Pattison et al., 2013). While high school GPAs increased over the years, Camara et al. (2004) found no relative change in the admission test scores when analyzing data from SAT Student Descriptive Questionnaire data. Camara, Kimmel, Scheuneman, and Sawtell (2004) reported the mean high school GPA from 1981 to 2002 grew .31 points, with noticeable differences in gender, race and ethnicity, and parents' education. While Camara et al. (2004) noticed a sizable increase in

HS GPA, they compared it to the changes in SAT scores. They reported SAT verbal has not increased and SAT math has increased over the years. Hurwitz and Lee (2018) reported comparable results on the increase in GPA, indicating grade inflation has increased, but they reported test scores have made no significant changes during the same period. Gershenson (2018) analyzed grades and EOCT from high school students in 2005 and 2016 using the North Carolina Education Research Data Center. While confirming the existence of grade inflation affecting high school GPAs, Gerhsenson (2018) found the inflation was not evenly distributed amongst high schools as grade inflation was more noticeable in schools with more affluent students than their counterparts. He reported the GPA gap was .41 points (Gerhsen, 2018). According to Gershenson (2018), he concludes the educational achievement gaps could be attributed to the existing grade inflation.

Contrary to the findings of grade inflation, Pattison et al. (2013) disagreed with grade inflation as they found there was no support from their analysis in the high school and postsecondary settings. Pattison, Grodsky, and Muller (2013) examined data from four different databases: National Longitudinal Study of High School Class of 1972, High School and Beyond sophomore cohort, the National Educational Longitudinal Study of 1988, and the Educational Longitudinal Study of 2002. Pattison et al. (2013) indicated the mean high school GPA had been steadily increasing from 1982 to 2002, while the mean postsecondary GPA declined from 1972 to 1992 for students attending a four-year institution. Pattison et al. (2013) revealed the correlation analysis indicated a constant correlation between high school GPA and test scores. Hiss and Franks (2014) presented findings from a study analyzing the impact of admissions standings on students who did and did not submit test scores. Within the study, 20 private, six public, five minority-serving, and two art institutions participated, resulting in a population of 122,916 records (Hiss & Franks, 2014). Even if grade inflation occurs at the high school level, Hiss and Franks (2014) noted it had no impact on the students' postsecondary GPA. They noted moderate correlations between high school and postsecondary GPA. In further examination of the public institutions, Hiss and Franks (2014) stated admissions policies guaranteeing admission based on high school GPA experience significant contributions to the postsecondary institution.

Admission test scores. Another major component of students' academic preparedness used in the development of the theoretical models is the admissions test to measure the scholastic intellectual abilities (Astin, 1975, 1993; Bean, 1980, 1983; Spady, 1970, 1971; Tinto, 1975, 1988, 1997). These scholastic tests are the Scholastic Assessment Test—SAT—and American College Test—ACT. Research has been a bag of mixed results regarding the admission test scores having any impact on academic performance. Most of the impact found indicates that of a small impact when compared to high school GPA. Alternatively, Chen and DesJardins (2008) utilized a survival analysis on factors impacting students persisting to graduation and revealed admission test scores were not significant predictors of academic performance in retaining university students. A study by Stewart et al. (2015) confirmed admission test scores have no impact on students' persistence. Unlike Chen and DesJardins' (2008) study, Stewart et al. (2015) examined the impact on one year retention. They reported admission test scores exhibited a small correlation ($r(3{,}213) = .118$, $p < .01$) in retention status. Yet, in a stepwise regression analysis, they found admission test scores were not significant, while first semester GPA (*B* = .999, *p* < .01) and high school GPA (*B* = -.731, *p* < .01) were the two largest significant contributing factors (Stewart et al, 20015).

In 2020, Allensworth and Clark (2020) examined the impact of high school GPA and test scores admissions requirements on students' academic performance to persist until degree attainment. Using data collected from Chicago Public Schools graduates from 2006 to 2009 attending a four-year postsecondary institution, they analyzed the impact of persistence using hierarchical linear models. While the model examining high school GPA's impact on persistence indicated a strong contribution, they also examined the impact of ACT scores on persistence. Without controlling for student characteristics, the results indicated students in higher ACT bins have higher odds of persisting than those in lower ACT bins (ACT less than 14, *OR* = .20; ACT 30 or higher, *OR* = 4.86) (Allensworth & Clark, 2020). However, the odds ratios decreased significantly for higher ACT bins when controlling for student characteristics (ACT less than 14, *OR* = .39; ACT 30 or higher, *OR* = 1.66) (Allensworth & Clark, 2020). While the odds ratio of the highest bin of high school GPA also decreased when controlling for student characteristics, the high ACT bin experienced a more significant decrease in odds. The highest GPA bin experienced a 1.67 times likelihood decrease, while the same bin of ACT experienced a 2.93 times decrease (Allensworth & Clark, 2020). Allensworth and Clark (2020) conducted another model to examine the effects of high school GPA and ACT scores on persistence towards graduation. The findings from this model revealed the high school GPA (four-year model, *B* = .666, *p* < .001, *OR* = 1.95; six-year model, *B* = .768, *p* < .001, *OR* = 2.16) exhibited a higher contribution to the model in impacting persistence than ACT (four-year model, *B* = .107, *p* < .05, *OR* = 1.11). ACT scores were not found to be significant in the six-year model. Allensworth and Clark (2020) suggested "students with higher [ACT] scores are more likely to attend the types of colleges where students are more likely to graduate" to explain why ACT is not a strong predictor of persistence (p. 207). They continued to state students earning higher ACT scores are more likely to earn higher grades (Allensworth & Clark, 2020).

Typically, proponents of the inclusion of admission tests scores as predictors of academic success state the inclusion of the test in connection with high school GPA but still admit to the inferior predictive ability of the test scores (Bowen et al., 2009; Korbin et al., 2008; Lotkowski et al., 2004; Noble & Sawyer, 2002; Rothstein, 2004). Further confirming the insignificant of being a predictor, Korbin, Patterson, Shaw, Mattern, and Barbuti (2008) found changes to the admission test scores, specifically SATs, have not made the test more predictive of the first year academic performance. Krobin et al. (2008) analyzed a sample of 151,316 students from submissions to CollegeBoard for students entering in Fall 2006. Using correlation analysis, they reported high school GPA and SAT total scores exhibited a correlation (*r*(151,316) = .28, *r-adj*.= .53), but also suggested the two different admission criteria also measured different aspects of student ability to perform successfully (Korbin et al., 2008). Examination of the correlation to first year GPA revealed high school GPA (*r*(151,316) = .36, *r-adj*.= .54) has the strongest correlation than SAT scores (SAT math, *r*(151,316) = .26, *r-adj*.= .47; SAT critical reading (*r*(151,316) = .29, *r-adj*.= .48); SAT writing (*r*(151,316) = .33, *r-adj*.= .51); SAT total, (*r*(151,316) = .35, *r-adj*.= .53)) (Korbin et al., 2008). The correlation of the interaction of high school GPA and SAT total to first year GPA strengthened (*r*(151,316) = .46, *r-adj*.= .62) (Korbin et al., 2008). Korbin et al. (2008) argued the utilization of both high school GPA and SAT scores should be continued because of the strengthened correlation in its ability to measure the first year of academic performance. Major proponents of the continued utilization of admission test scores in the assessment of academic preparedness for postsecondary institutions argue the tests scores are a way:

1. to provide neutral measurement to guard against high school grade inflation (Buckley et al., 2018),
2. to provide an assessment on students for selected materials as compared to high schools' wide range of materials (Allensworth & Clark, 2020), and
3. to provide a supplemental predictor to counter the high school GPA's under-predictability of students with high test scores (Mattern et al., 2010a).

Financial situations.

Family financial income or situations. According to Tinto (1982), one most cited explanation and impacts of student departures are financial situations. Financial situations in terms of impact on students may be short-term, long-term, and indirect (Tinto, 1982). Specifically for students from a low socioeconomic status background, Tinto (1975, 1982) stated finances play a vital role in retaining or departing decisions. Students from lower socioeconomic status are more likely to depart, while students from higher status are more likely to retain (Tinto, 1975). Furthermore, financial situations are often cited as strong predictors of attrition in a student's early academic career rather than later (Tinto, 1982).

St. John, Paulsen, and Carter (2005) conducted a logistic regression analysis in sequential steps on the impact of the cost of attending college and the persistent rate differences between Black or African American and White students. The first model did not examine the impact of financial aid, the second model included the amount of financial aid, and the third model included the room and board costs (St. John et al., 2005). Using data obtained from the National Postsecondary Aid Survey of 1987, St. John et al. (2005) found more Black or African American students were in the lower to lower-middle income levels (36% and 37%, respectively) when compared to White counterparts (15% and 27%, respectively). In their model for Black or African American students, their family income levels were not significant in the three sequential logistic regression models (St. John et al., 2005). However, St. John et al. (2005) reported Black or African American students who were financially independent were less likely to persist in the first two models (*delta-p* = -.065, $p < .01$; *delta-p* = -.046, $p < .05$). They further explained the federal need analysis might not be appropriately factoring all the conditions contributing to the cost of attendance for independent students (St. John et al., 2005). For White students, family income levels and students' independent status were not found to be significant (St. John et al., 2005).

Using a sample of 6,733 students from the Beginning Postsecondary Student survey data set from the 1996 and 2001 cohorts, Chen and DesJardins (2008) examined the impact of income level had on students' risk of dropping out of a postsecondary institution. Of the students examined in the analysis, 19% were from an income level less than $25,000, 50% from an income greater than or equal to $25,000 but less than $75,000, 27% were from an income greater than or equal to $75,000, and 5% were missing income information (Chen & DesJardins, 2008). They also noted 72% received Pell grants, 60% received loans, and 21% received work-study aid for low-income students in their first year (Chen & DesJardins, 2008). Middle income students had 17%

receiving Pell grants, 58% receiving loans, and 20% receiving work-study aid, while the upper income students had 0.1% receiving Pell grants, 36% receiving loans, and 8% receiving work-study aid (Chen & DesJardins, 2008). They continued to note the low-income students continued to receive a higher percentage of aid throughout the six years in the analysis (Chen & DesJardins, 2008). Results of the survival analysis indicated the high level income students (*OR* = .611, *p* < .001) showed a significant difference when compared to the low income students (Chen & DesJardins, 2008). To measure what other factors may cause a student to drop out, they noted parents' education was a significant factor. Students with parents with at least a bachelor's degree had lower odds of dropping out (*OR* = .643, *p* < .001) than students with parents without a bachelor's degree (Chen & DesJarins, 2008). Chen and DesJardins (2008) predicted the dropout probability for low income students receiving Pell grants was .208 and those not receiving was .566. Also, they predicted the dropout probability for the middle income level students receiving Pell grants was .250 and for those not receiving was .153 (Chen & DesJardins, 2008). The assistance of the Pell grant was able to help lower the dropout probability by .358 points for the low income students. Pell grant assistance only slightly improved the dropout probability for middle income levels. More importantly, lower income students not receiving Pell have a .413 points higher dropout probability than their middle income counterparts (Chen & DesJardins, 2008).

Using a sample of 6,383 students from the Beginning Postsecondary Student data set of students from 1996 and 2001 and financial indicators of 49 states from the national higher education database, Chen and St. John (2011) examined the relationships between state-level financial policies and persistence to the institution. They further examined how the relationship is based on the students' socioeconomic status (SES) and racial and ethnic backgrounds. Student SES status was split into quartiles in which the first quartile was the lowest level SES, the second and third quartiles were combined to form the mid-level SES, and the fourth quartile was the highest level SES (Chen & St. John, 2011). In the study, persistence was defined as either graduating or still enrolled at the first institution by the sixth year. Approximately 70.6% of students from the highest SES level persisted, while 57.1% of the mid-level and 44.1% of the lowest level SES groups persisted (Chen & St. John, 2011). Using a hierarchical generalized linear model, Chen and St. John (2011) reported mid-level SES (OR = 1.32, $p < .01$) and high level SES were (OR = 1.55, $p < .001$) times more likely to persist to graduation than the lower SES students.

In her data analysis from the National Longitudinal Study of Youths from 1997, Velez (2014) analyzed a sample of 6,748 students using a logit model to predict students' degree attainment. Of the 1,062 students who dropped out of a four-year postsecondary institution, 27% were from a household income of less than $33,000 (Velez, 2014). From the logit model's prediction of students from low income households, 96.2% of students are predicted to never enroll in a four-year postsecondary institution, and for those who do enroll, 81.8% are predicted to drop out (Velez, 2014). She even found a high percentage (98.7%) of low-income students are predicted to not even attend a two-year institution (Velez, 2014). However, the predictions only indicated about 63.5% were predicted to drop out of the two-year institution (Velez, 2014).

Financial aid. After being accepted, the ability to attend a post-secondary institution with the continuous increasing expansion of a diverse set of students has greatly been assisted through the form of financial aid. According to the Common Data Set (2020) definitions, need-based aid involves grants, campus jobs, or loans resulting from the student's need for financial aid. In other words, need-based aid fills in the gaps in the student's ability to pay the tuition and fees to attend. Non-need-based aid, also called merit-based aid, involves a student qualifying to receive the aid (Common Data Set, 2020). Qualifying could range from scholastic ability, talent, or specific characteristics to be offered the aid. In 2020, approximately 86.4% of first-time, first-year undergraduate students received some form of financial aid, with an estimated growth of .09% each year (Hanson, 2020). Hanson (2020) also reported female students were 40% more likely to accept financial aid than their counterparts. When reporting the acceptance rate by race and ethnicity, he reported around 80% of Black or African American and 62% of Asian students accepted financial aid (Hanson, 2020). Across the nation, postsecondary students received an average loan of $8,285, with 40% of students between the ages of 23 or younger receiving loans (Hanson, 2020). In comparison, Hanson (2020) reported 46% of undergraduates accept financial loans with an average amount of $6,575 within the State of Georgia attending four-year institutions.

One of the most common forms of need-based aid is the federal PELL grant. According to the National Center for Education Statistics' (n.d.b) available data, approximately 34.0% of undergraduates receive the PELL grant. Bettinger (2004) analyzed data from the Ohio Board of Regents to examine the impact of the Pell grant on the 1999 incoming freshmen persistence in the states' public postsecondary institutions who completed the Free Application for Federal Student Aid (FAFSA). From the data Bettinger (2004) reported 35,233 initially filed a FAFSA form, with 12,143 did not file the second year. In the initial model, he examined the impact the dollar amount of financial had on students not returning to the institution. The amount of financial aid was significant, (*B* = .033, *p* = .002) (Bettinger, 2004). When adding student background and family income level, the amount of financial aid award impact lessened (*B* = .006, *p* = .003) (Bettinger, 2004). The third model included the first year GPA earned. While the first year GPA (*B* = -.138, *p* = .003) had a significant impact on a student's likelihood not to persist, the amount of financial aid awarded was reported to be still significant (*B* = .0002, *p* = .003) (Bettinger, 2004). Even though the financial aid award was significant in the third model, the strength of the impact was lessened. Bettinger (2004) indicated this could be related to the students' ability to do well in the postsecondary classroom.

St. John et al. (2005) conducted a logistic regression analysis in sequential steps on the impact of the cost of attending college and the persistent rate differences between Black or African American and White students. The data was obtained from the National Postsecondary Aid Survey of 1987. From the sample of 24,271 students, more financial aid had been awarded to the Black or African American students. Compared to White Students, Black or African American students received $837 more dollars in grants, $177 in loans, and $99 more dollars for the work study program (St. John et al., 2005). Additionally, St. John et al. (2005) reported more Black or African Americans were in the lower to lower-middle income levels (36.0% and 37.3%, respectively) when compared to White counterparts (15.0% and 27.3%, respectively). The first model did not examine the impact of financial aid, the second model included the amount of financial aid, and the third model included the room and board costs (St. John et al., 2005). In the second and third models for Black or African Americans, only the amount of grant dollars were found to be significant (second model, *delta-p* = -.033, *p* <= .01; third model, *delta-p* = -.0327, *p* <= .01) (St. John et al., 2005). In the second and third model for White students, grant dollars were found to be significant (second model, *delta-p* = -.0135, *p* < .01; third model, *delta-p* = -.0131, *p* <= .01) (St. John et al., 2005). Moreover, the second model for White students found loans to have a significant impact on persistence (*delta-p* = -.0089, *p* <= .01) (St. John et al., 2005). In other words, the models factoring in financial aid and living expenses found the amount of grant aid for African American and White students had a negative relationship in persisting. Moreover, the model not factoring in living expenses found the amount of loans had a negative relationship with persistence for White students. St. John et al. (2005) further explained the impact of for every $1,000 increase in the loan amount, the persistence probability decreased by 5%. In their model for Black or African Americans, students' family income levels were not significant in the three sequential logistic regression models (St. John et al., 2005). However, St. John et al. (2005) reported Black or African American students who were financially independent were less likely to persist in the first two models (first model, *delta-p* = -.065, *p* < .01; second model, *delta-p* = -.046, p < .05). They further explained the federal need analysis might not be appropriately factoring all the conditions contributing to the cost of attendance for independent students (St. John et al., 2005). For White students, family income levels and students' independent status were not found to be significant (St. John et al., 2005).

Using a sample of 6,733 students from the Beginning Postsecondary Student survey data set from the 1996 and 2001 cohorts, Chen and DesJardins (2008) examined the impact income level had on students' risk of dropping out of a postsecondary institution. Of the students examined in the analysis, 19% were from an income level less than $25,000, 50% from an income greater than or equal to $25,000 but less than $75,000, 27% were from an income greater than or equal to $75,000, and 5% were missing income (Chen & DesJardins, 2008). They also noted for the freshmen year of the low income students, 72% received Pell grants, 60% received loans, and 21% received work-study aid (Chen & DesJardins, 2008). Middle income students had 17% receiving Pell grants, 58% receiving loans, and 20% receiving work-study aid, while the upper income students had .1% receiving Pell grants, 36% receiving loans, and 8% receiving work-study aid (Chen & DesJardins, 2008). Chen and DesJardins' (2008) initial model indicated Pell grant recipients had a negative impact on the risk of dropping out, even though no significance was indicated. However, further examination of the interaction between income levels and Pell grant status revealed the interaction between middle income level and Pell grant to be significant (*OR* = 1.864, $p < .05$) (Chen & DesJardins, 2008). Chen and DesJardins (2008) indicated the likelihood of departing is moderated by the amount of Pell grant received. Chen and DesJardins (2008) predicted the probability of departing based on income level and the Pell grant recipient status. The departing probability was .208 for low income students receiving Pell and .566 for low income students not receiving Pell. Additionally, they predicted the probability of departing for the middle income level and Pell grant recipient status. Those receiving Pell had a predicted probability of .250, and those not receiving was .153 (Chen &

DesJardins, 2008). The assistance of the Pell grant was able to help lower the dropout probability by .358 points for the low income students (Chen & DesJardins, 2008). Pell grant assistance only slightly improved the dropout probability for the middle income level. More importantly, lower income students not receiving Pell have a .413 higher dropout probability than their middle income counterparts (Chen & DesJardins, 2008).

Stater (2009) analyzed 18,748 new students from 1994 to 1996 from three flagship institutions in Colorado, Indiana, and Oregon to measure how financial aid impacts the yearly earned GPA as a measure of academic integration. He also collected data on how well the students did after the first year in college (Stater, 2009). The mean GPA of the first year was 2.80 (*SD* = .6533) (Stater, 2009). Stater (2009) reported for the next three years the mean GPA experienced a significant increase from the prior years (year two, *M* = 2.90, *SD* = .5192; year three, *M* = 2.96, *SD* = .4753; and year four, *M* = 3.02, *SD* = .4510). Within the first year, 56.2% were female, 94.9% were White or Asian, and 65.8% lived in the institution's state (Stater, 2009). The average SAT total score was 1084, while the average high school GPA was 3.37 (Stater, 2009). Using an ordinary least squares model, the model not including the tuition amount was found to be significant in impacting first year earned GPA ($F(18{,}748) = 280.1$, $p < .01$, $R^2 = .2330$) (Stater, 2009). When controlling for the tuition amount, the model was also found to be significant ($F(18{,}748) = 268.1$, $p < .01$, $R^2 = .2330$) (Stater, 2009). The model found both need-based and merit-based aid were significant when tuition was not included, (need-based aid, $B = .0361$, $p < .1$; merit-based aid, $B = .2329$, $p < .001$) (Stater, 2009). However, when the amount of tuition was introduced to the model, only the amount of merit-based aid was found to be significant ($B = .2290$, $p < .001$) (Stater, 2009). Stater

(2009) reported an increase between .10 to .19 points in GPA was predicted per $1,000 increase in any financial aid. In both models, merit-based aid had a stronger contribution to predicting the GPA of the financial aid variables. From the contribution of the merit-based aid, the amount of financial aid could conclude the impact on successfully integrating into the academic communities as students must successfully perform in classes to continue receiving this aid. Stater (2009) further indicated merit-based aid is likely to foster institutional commitment as it typically is not transferable to another institution, unlike need-based aid.

In a multilevel event analysis, Chen (2012) examined the dropout risk using data from the Beginning Postsecondary Students from 1996 and 2001, with additional data coming from the Integrated Postsecondary Education Data Systems (IPEDS). The sample consisted of 5,762 students classified as first-time, full-time, degree-seeking freshmen attending a four-year institution. Chen (2012) reported around 40.7% of the students depart from their initial institution. On average, the students received $1,141 of Pell grant, $2,623 in subsidized loans, and $2,466 in unsubsidized loans (Chen, 2012). Additionally, an average of $3,309 were awarded in merit-based aid (Chen, 2012). From the analysis, four types of aid were reported to have significant contributions to students departing from the initial institution. Subsidized loans (*OR* = .92, *p* < .001), unsubsidized loans (*OR* = .95, *p* < .05), work-study programs (*OR* = .81, *p* < .01), and merit-based aid (*OR* = .94, *p* < .001) were the financial indicators found to be significant negative relationship in terms of dropping out (Chen, 2012). Chen (2012) further explained underrepresented students from lower socioeconomic status have a lower risk of departing from the institution when awarded higher financial aid amounts. In other words, the funding gap for these students is alleviated to allow them to focus on their studies.

Gross, Hossler, Zikin, and Berry (2015) analyzed a sample of data consisting of 12,301 first-time, degree-seeking students obtained from the Indian Commission for Higher Education’s longitudinal data system to measure the impact of institutional merit-based aid on student departure. A second sample of 4,254 was obtained using a coarsened exact matching method (Gross et al., 2015). While 52% qualified for merit-based financial aid, only 18% received the form of aid (Gross et al., 2015). Females and males had the same percentage of students receiving merit aid. They reported around 82% of the students receiving merit-aid were White. Based on family income, 54% of the students receiving merit-aid were from the highest family’s adjusted gross income level (Gross et al., 2015). Around 89% and 70% of the students receiving merit-based aid were from the high SAT total score grouping and the top quartile of the HS GPA (Gross et al., 2015). Using a discrete-time event history model, Gross et al. (2015) found merit-based aid recipients to have a lower risk of departing than those who did not receive the aid. In using the full sample, they reported for a $1,000 increase in merit-based aid, there was a 6.5% decrease in the departure odds, while need-based aid was a 6.0% decrease (Gross et al., 2015). More importantly, the matching sample results no longer indicated merit-based aid was significant while still indicating need-based aid was significant (Gross et al., 2015). The matched sample analysis indicated for a $1,000 increase in need-based aid, the departure likelihood was reduced by 5% (Gross et al., 2015). In explaining how other student characteristics are connected to the likelihood of dropping out, Gross et al. (2015) indicated less than 15% of the students who received merit-based aid are from the lower two income levels. These students were more likely to have received lower grades in high school and lower admissions test scores which would not qualify them to receive merit-based aid. Also, they noted, these students tended to be underrepresented or first-generational (Gross et al., 2015).

HOPE scholarship. One form of merit-based aid specific to Georgia is the HOPE scholarship. As the form of financial aid the been modified over recent years, the overarching requirement is a minimum 3.00 GPA from a high school within the state (Georgia Student Finance Commission, 2021a). Renewal of HOPE scholarship depends on students' academic performance of their cumulative earned grades at specific thresholds in their academic journey. Recipients of the HOPE scholarship must maintain a 3.00 GPA throughout their post-secondary journey (Georgia Student Finance Commission, 2021b).

Henry, Rubenstein, and Bugler (2004) analyzed data of Georgia high school graduates in 1995 to analyze the impact of the HOPE scholarship across four years enrolled in the public institutions in the state. Of the data, they limited the students they considered to be borderline HOPE scholarship recipients. In other words, students who barely had above a 3.00 high school GPA. The borderline HOPE recipients were a sample size of 1,915 students. Henry et al. (2004) reported these students earned a mean GPA of 2.44 across the core courses in the institutions. They next matched students who did not receive HOPE based on the GPA of the course courses to generate a total of 1,817 students. The second sample size consisted of 3,732. Around 48% of the students were female, 32% were Black or African American, and 53.5% were full-time students (Henry et al., 2004). Of those who received the scholarship, 47% were female, 28% were Black or African American, and 61% were enrolled full-time (Henry et al., 2004). Using a linear regression model, *adj.* $R^2 = .096$, Henry et al. (2004) analyzed the impact of the HOPE scholarship on the earned college GPA. They found HOPE scholarship to be a significant indicator ($B = .17, p < .01$) (Henry et al., 2004). Also, SAT total score ($B = .001, p < .001$), earning a college prep high school diploma ($B = .24, p < .05$), core high school GPA ($B = .40, p < .01$), and gender ($B = .14, p < .01$) were found to be significant (Henry, et al., 2004). Using a logistic regression model (Max-rescaled $R^2 = .11$), they analyzed the impact of HOPE scholarship on persisting to graduation (Henry et al., 2004). For students enrolled in the four-year institutions, the HOPE scholarship was found to be a significant contribution to the model ($B = .54, p < .01$) (Henry et al., 2004). Also, enrolled in remedial courses ($B = -1.03, p < .01$), core high school GPA ($B = 1.63, p < .01$), and gender ($B = .66, p < .01$) were reported as significant indicators for persisting to graduation (Henry et al., 2004).

In a Georgia Budget and Policy Institute report on higher education, Suggs (2016) wrote the HOPE scholarship and grant is the state's largest financial assistance for students enrolled in postsecondary institutions. In reviewing data from the University System of Georgia on the Fall 2013 undergraduate population, Suggs (2016) excluded students who were not eligible to receive the scholarship—out-of-state students, dual enrollment students, and post-baccalaureate students. She reported around 36% of students enrolled in the state's public institutions benefit from the scholarship. She also noted 64% of HOPE recipients are White students who only account for 54% of undergraduate enrollment within the states' institutions (Suggs, 2016). Continuing to note the disparities amongst underrepresented students, she noted only 20% of Black or

African American and 36% of Hispanic students receive the HOPE scholarship, while 46% of Asian and 45% of White students receive it (Suggs, 2016). She further analyzed the percentage using the HOPE scholarship by income level to determine the Pell Grant status. Approximately 30% of low income students benefited from the scholarship, which is around 12% less than their counterparts (Suggs, 2016). Even when low income students receive the HOPE scholarship, they continue to struggle financially and eventually drop out. Suggs (2016) reported 61% of low-income students using the HOPE scholarship persisted until graduation, while 75% of their counterparts persist until graduation. As the HOPE scholarship is a merit-based financial aid, students' academic performance impacts whether a student continues to receive the benefit of the scholarship. Suggs (2016) found low income students are more likely to lose HOPE scholarships than middle to high income students. Approximately 47% of low income students will lose the HOPE scholarship, while only 39% of middle to high income students lose HOPE (Suggs, 2016).

Major declaration and grouping. In an online *Forbes Magazine* article, Onink (2010) linked current workforce changes to students having a hard time making a major declaration. He indicated some of the "hot fields of study" cause students to stand in line at the unemployment office due to a lack of jobs (Onink, 2010). With the increasing number of undeclared majors, Onink (2010) indicated the lack of opportunities at the high school level, allowing students to have the possibility to explore possible majors before attending a postsecondary institution. At post-secondary institutions, majors or programs of studies are grouped into overarching fields of study, resulting in the programs being housed in specific colleges or departments. These groupings may be based on the type of degree, concentrations, or potential job-market opportunities after graduation.

Leppel (2001) examined the impact of grouping a major declaration by gender of the students exhibited on students' likelihood of persisting. Using data collected from the 1990 Beginning Postsecondary Student (BPS) database, Leppel (2001) analzyed a population of 2,426 males and 2,521 females. Descriptive analysis indicated males had higher persistence rates than females in business (males, 93.13%; females, 92.01%) and engineering (males, 93.43%; females, 92.31%) related majors (Leppel, 2001). Females exhibited higher rates in education (males, 88.72%; females, 94.39%) and health (males, 89.02%; females, 97.46%) related majors (Leppel, 2001). The gender gap was the same for art and sciences-related majors (males, 95.02%; females, 95.62%) and undeclared (males, 78.81%; females, 77.58%) (Leppel, 2001). Leppel (2001) conducted two logit analyses to examine the major declaration by gender. She reported both models to be statistically significant (males, $\chi^2(2,426) = 298.124$, $p < .01$; females, χ^2 $(2,521) = 734.911$, $p < .01$) (Leppel, 2001). For the model examining males, business ($B = .070$, $p < .1$), education ($B = -.542$, $p < .01$), and undeclared ($B = -.261$, $p < .01$) majors were found to be statistically significant (Leppel, 2001). Business-related majors were found to have a positive influence, while education and undeclared were found to exhibit a negative influence. The model examining major declaration for females found business ($B = -.303$, $p < .01$), education ($B = .084$, $p < .1$), health ($B = .635$, $p < .01$), and undeclared ($B = -.733$, $p < .01$) to be statistically significant (Leppel, 2001). Females declaring an education or health-related major have a positive likelihood of persistence, while business and undeclared have a negative likelihood. One interesting finding amongst the male and female undeclared majors, females are more likely to not persist than the males.

Additionally, Leppel (2001) examined the impact of a major declaration by gender on the first year earned GPA. Using linear regression analysis to examine the impact for males and females, both models were found to be statistically significant (males, *adj.* $R^2 = .149$, $F(2,426) = 7.493$, $p < .01$; females, *adj.* $R^2 = .167$, $F\ (2,521) = 7.722$, $p < .01$) (Leppel, 2001). Business-related major ($B = 14.856$, $p < .01$) exhibited a significant positive impact on the first year GPA, while undeclared majors ($B = -26.972$, $p < .01$) exhibited a significant negative impact for female students (Leppel, 2001). In terms of major grouping, only undeclared majors ($B = -43.285$, $p < .01$) negatively influenced the earned GPA (Leppel, 2001). Leppel's (2001) findings indicated males and females who are undeclared are negatively impacted regarding earned GPA and persistence and attributed the cause to possible low educational commitment. In other words, the findings from Leppel's (2001) study suggested there are natural tendencies based on major declarations, and students who fall within these tendencies are more likely to have successful academic performance.

In comparing science, technology, engineering, and mathematics (STEM) groupings to non-STEM, Gansemer-Topf, Kallasch, and Sun (2017) examined the effects these major groupings exhibited on first year academic performance. Gansemer-Topf et al. (2017) obtained data for Fall 2008 to Fall 2012 first time freshmen living on campus at a Midwestern research university. The population size consisted of 17,850 students. Gansemer-Topf et al. (2017) conducted a one-way analysis of variance (ANOVA) to determine if any statistical significance occurred between STEM and non-stem. The one-way ANOVA conducted did not find any statistical significances between STEM majors (*M* = 2.80, *SD* = .910) and non-STEM majors (*M* = 2.82, *SD* = .861), *F*(2, 17,850) = 1.818, *p* = .178 (Gansemer-Topf et al., 2017). For cumulative earned GPA at the end of the first spring, the one-way ANOVA found a statistically significance between STEM majors (*M* = 2.82, *SD* = .823) and non-STEM majors (*M* = 2.86, *SD* = .773), *F*(2, 17,850) = 9.405, *p* = .002 (Gansemer-Topf et al., 2017). Additionally, the one-way ANOVA found statistically significance for retention status between STEM majors (*M* = .884, *SD* = .320) and non-STEM majors (*M* = .874, *SD* = .332), *F*(2, 17,850) = 4.051, *p* = .012 (Gansemer-Topf et al., 2017). Their findings suggested STEM majors have a higher retention rate than non-STEM majors. Gansemer-Topf et al. (2017) alluded to students' commitment, aspirations, and academic preparedness may play an important role in the earned GPA at the end of the first spring and eventually the decision to return to the institution. However, Spight (2020) reported statistically no difference in the likelihood of a student persisting past the first year based on the major declaration. In his study, Spight (2020) sought to examine the relationship between a major declaration and academic performance using data collected from a Carnegie Doctoral/Research-Extensive institution. The population consisted of 4,489 first-time freshmen enrolled in Fall 2010. Spight (2020) first conducted an independent t-test on the number of terms enrolled by major declaration status. The test showed a statistically significant finding of undeclared majors (*M* = 11.71, *SD* = 2.800) tend to be enrolled for more terms than declared majors (*M* = 11.44, *SD* = 3.064), *t*(4,489) = 2.586, *p* < .01 (Spight, 2020). While the results indicated a significant difference, Spight (2020) indicated the difference of .27 in the mean terms enrolled has no practical difference. Additionally, Spight (2020) conducted a logistic regression analysis to determine the impact on first-year persistence. In the examination of the major declaration status, the logistic regression analysis was found to be statistically significant, $\chi^2(15) = 120.334$, Nagelkerke $R^2 = .079$, $p < .001$ (Spight, 2020). Spight (2020) also reported the Hosmer and Lemeshow's test for goodness of fit (χ^2 (8) = 11.589, $p = .171$) not to be significant, indicating the model was a good fit overall. He noted major declaration status did not contribute to a students' likelihood of persisting past the first year. Within the model student demographics and preschooling performance influenced students' persistence (HS GPA, $B = 1.177$, $p < .001$; in-state residency, $B = .820$, $p < .01$; female, $B = .260$, $p < .05$; and SAT composite score, $B = .001$, $p < .01$) (Spight, 2020).

Institutional financial expenditures. Most postsecondary institutions have some support services to assist students in academic or social communities. Services such as tutoring centers to assist students with classroom concepts and subjects, libraries for assistance in research projects, and student union centers for social engagement activities are examples of how institutions expend resources to assist with the students' integration into the communities. As institutions synchronize their expenditures and resources to their mission, studies have shown the institutions make improvements in service areas to increase academic performances.

Ryan (2004) examined the impact of expenditures on academic performance regarding students' retention towards degree attainment. Data was gathered from IPEDS and College Board for Carnegie Baccalaureate I or II institutions, which totaled 363 institutions in the sample. Using a regression analysis, Ryan (2004) reported the model was statistically significant ($R^2 = .725$, *adj.* $R^2 = .75$, $F(363) = 70.791$, $p < .001$). Of the expenditures included in the model, only instruction (B = .281, p < .001) and academic support (B = .119, p = .007) were found to have a significant contribution to students retaining until degree completion (Ryan, 2004). Ryan (2004) reported student services and instructional support did not significantly impact academic performance. He stated student services expenditures might be impacted by a few services (e.g., admissions and financial aid) included in the IPEDS expenditure classifications (Ryan, 2004). Ryan (2004) suggested instructional and academic support expenditures contribute significant influence to predicting academic performance. This could assist in integrating students into the institutions' communities.

Gansemer-Topf and Schuch (2006) analyzed the relationship of institutional expenditures on retention and graduation rates of various private institutions by selectivity type available from IPEDS. Using a multiple regression analysis, they reported significant findings on the expenditures and prediction of academic performance. Overall, the regression model found a statistically significant in predicting academic performance (retention, R^2 = .635, *adj.* R^2 = .629, F(6,369) = 107.02, p < .001) (Gansemer-Topf & Schuch, 2006). Continuing to examine impact of expenditures amounts, they noted instruction (B = .33, p < .001) and institutional grant (B = .22, p < .001) had positive interactions, while student service expenditures (B = -.13, p < .001) exhibited a negative impact on retention rates (Gansemer-Topf & Schuch, 2006). Gansemer-Topf and Schuch (2006) also conducted a multiple regression analysis on the percent of expenditures by classification exhibited on academic performance. This model also indicated statistically significant results for retention (R^2 = .588, *adj.* R^2 = .581, F(6,369) = 87.74, p < .001) (Gansemer-Topf & Schuch, 2006). Of the percentage of expenditures, student services (B = -.17, p < .001) exhibited a negative and institutional grant (B = .17, p < .001) exhibited a positive were ranked the highest in impact on retention rates (Gansemer-Topf & Schuch, 2006). Additionally, percentage of instruction (B = .13, p < .001) and academic support (B = .13, p < .001) exhibited a positive impact on academic performance (Gansemer-Topf & Schuch, 2006). The findings from Gansemer-Topf and Schuch (2006) suggested allocating resources and available funds to align with the institution's mission and strategic plan improve academic performance of enrolled students.

Webber and Ehrenberg (2009) obtained data from IPEDS consisting of 1,160 four-year institutions regarding institutional expenditures and the persistence and graduation rates of the first-time, full-time freshmen from the academic year 2002 to 2003 through 2005 to 2006. Using an econometric analysis, Webber and Ehrenberg (2009) found an increase of $500 in student services expenditures exhibited a .7% increase, while the same increase in academic and instruction expenditures only experienced a .3% increase of students persisting towards graduating. Due to the wide range of IPEDS definitions of student services expenditures, Webber and Ehrenberg (2009) cautioned postsecondary institution administrators to examine student service expenditures carefully.

Chen (2012) conducted a longitudinal study investigating the student dropout phenomenon within four-year postsecondary institutions. The study examined two research questions in which one examined how student characteristics impacted students retaining or departing, and the other examined institutional characteristics (Chen, 2012). Using data collected from Beginning Postsecondary Students (BPS96/01) and Integrated

Postsecondary Education Data System (IPEDS) 1995 to 2000, the population in the study consisted of 5,762 first-time, full-time freshmen from 400 four-year institutions from Fall 1995 and Fall 1996 (Chen, 2012). Chen (2012) utilized a multilevel event history analysis to examine the impact of student and institutional characteristics on student's attrition towards degree completion. Chen (2012) noted within the student characteristics the college GPA (*OR* = .59, *p* < .001) exhibited the largest impact dropping out. In other words, students who earned higher GPAs had lower odds of the students dropping out from the institution. Additionally, he reported certain types of financial aid had significant negative impacts on students' departing decisions (subsidized staff loans, *OR* = .92, *p* < .001; unsubsidized staff loans, *OR* = .95, *p* < .05; work-study aid, *OR* = .81, *p* < .01; and merit aid, *OR* = .94, *p* < .01) (Chen, 2012). Of the institutional expenditures analyzed, Chen (2012) found only student services expenditures (*OR* = .85, *p* < .05) had an impact on student dropout. The amount of student services expended reduced the odds of students departing. With only student services expenditures found to be significant, Chen (2012) suggested postsecondary administrators may need to look beyond the traditional educational structures to assist in retaining students; yet, he indicated the findings regarding student services do not provide strong justifications for a total funding change.

Data Science

A *Harvard Business Review* article mentioned one of the current century's sexist occupations had become a data scientist, stemming from the popularization of the buzzword data science (Daveport & Patil, 2012). While the explosion of the usage of data science has become new, the field of study has its roots reaching back to the early

1960s. In 1962, John Tukey predicted the use of data analysis to become a science produced by computing technological advancements (UW Data Science Team, 2017). As technology advanced, the time to conduct analyses was shortened as computations of statistical methods, and other algorithms became easier to execute. Until the 2000s, data science was not considered a field of study (UW Data Science Team, 2017). According to Conway (2014, 2015), data science is a field of study with a delicate balance of knowledge of mathematics and statistics, substantive expertise, and hacking skills, as displayed in Figure 2. In the hacking sphere, Conway (2014, 2015) argued data scientists are not using their skills to hack into companies but possess the knowledge of manipulating data files in an active and algorithmic line of thinking. He continued once the data files have been prepared, the knowledge of mathematics and statistics is utilized in building predictive models. Lastly, Conway (2014, 2015) stated data science has the investment into the data to explore and discover the findings while building modeling tools.

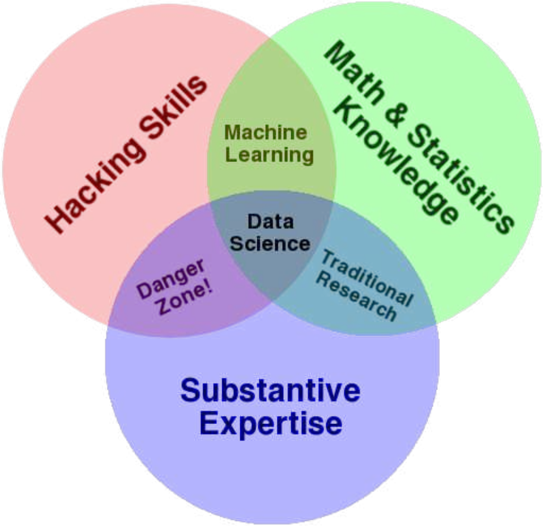


Figure 2. *Drew Conway's venn diagram of data science.*

As another explanation of data science, CEO of Metamarkers Mike Driscoll stated "data science is the civil engineering of data. Its acolytes possess a practical knowledge of tools and materials, coupled with a theoretical understanding of what's possible" (O'Neil & Schutt, 2014, p. 7). As a conference speaker, Driscoll (2013) explained data science incorporates social science aspects into the process and job duties. He stated, "data science, as a discipline, is fundamentally about human behavior" (Driscoll, 2013). Continuing about the inclusion of social science methods, he argued data science is never a black box method, but the tools need to be examined and explained to understand the action of predicted human behavior (Driscoll, 2013). In further a breakdown of his Venn diagram, Driscoll (2013) elaborated while, at times, 80% of the world may involve the hacking skills in obtaining and preparing the data sets, in reality, 80% of the hard work lies in the substantive expertise area. Within the substantive expertise area, he reiterated individuals spend half of their time asking questions about the data and figuring out which tools to use (Driscoll, 2013). The other half involves interpreting the results to know when the desired or predicted results are achieved (Driscoll, 2013).

Data mining and its application in higher education. Tracing the origins of the data mining process, individuals have been doing it since the late 1980s (Coenen, 2011). According to Coenen (2011), data mining, in a broad sense, is the extraction of information hidden. Within data science, data mining involves utilizing techniques to leverage findings previously unknown to produce valuable information (Tierney, 2014). In other words, data mining combs through large data sets to discover patterns hidden due to the sheer size of the information (Talend, n.d.). According to Coenen (2011), data mining is not comprised of scriptwriting. Data mining involved simple tabulations in the early stages as the processing time was large; however, because of technological advancements, data mining has grown into a discovery process of patterns and trends (Coenen, 2011; Tierney, 2014). While common data sets are maintained in structured files, data mining also allows individuals to discover patterns and trends found within unstructured data files, sound bites, and visual images (Attewell & Monaghan, 2015, Coenen, 2011). Overall, Tobin (2022) indicated there are five stages of data mining consisting of understanding the goals of the project, understanding where data is located, data preparation processes, analysis and model building, and sharing information discovered.

As data science and data mining techniques advanced, higher education began to adopt the discipline to assist in discovering trends and patterns for improvements regarding institutions' performance indicators, in addition to identifying prospective students, peer interaction, tracking health concerns, and alumni engagement (Belani, 2019; Data Science Degree Program, 2021; Matthews, 2018, 2019). As graduation rates within postsecondary institutions have become one of the most published and discussed facts, institutions incorporating data science and mining have the advantage of assisting students before academic performance issues arise (Data Science Degree Program, 2021). Using data mining techniques, administrators and policy-makers have been able to identify students who would be considered at-risk for not retaining and progressing towards degree attainment (Data Science Degree Program, 2021). Data science and mining techniques to predict at-risk students would assist postsecondary institutions in better utilizing academic and student support staff and resources (Matthews, 2018, 2019). Data science's application in higher education has also unlocked a great potential to assist in areas other than students' academic performance. Postsecondary institutions have begun to use data science in admissions offices to attract students and identify the available market of recruits and in health services to identify potential outbreaks concerning health issues (Belani, 2019; Data Science Degree Program, 2021; Matthews, 2018, 2019). DiMaggio (2021) argued data science goes beyond just predicting what will occur and involves the recommendations to optimize the potential impact. In using a coin analogy, he explained data sciences using both predictive analytics and prescriptive analytics to get the most out of the data (DiMaggio, 2021).

Cross-validation methods. Intending to build predictive models, data splitting into training and testing sets occurs with the goal of the testing data set resembling unseen real world events (Bose, 2019; Goyal, 2021; Soni, 2019). However, there is concern regarding predictive models to overfit or underfit (Attewell & Monaghan, 2015; Kuhn & Johnson, 2013; Tripathi, 2020). Tripathi (2020) indicated overfitting occurs when a model performs too well on the training data set but exhibits poor performance in the testing or unseen data. In overfitting, he stated the model focuses more on the noise than the true signals in the training data set (Tripathi, 2020). Drakos (2019) stated overfitting results in poor generalization of unseen data. Tripathi (2020) mentioned while underfitting occurs when the model performs poorly on the training data set, it is commonly not referred as with its identification. As a result of the easy identification of underfitting, it is recommended to try another model (Tripathi, 2020). However, overfitting involves more than building a new predictive model. The remedy to prevent overfitting is through cross-validation (Attewell & Monaghan, 2015; Drakos, 2019; Tripathi, 2020). Drakos (2019) stated cross-validation allows one to generate the accuracy of a model in practice in the training phase to limit problems of the predictive model, either underfitting or overfitting.

There are several cross-validation methods to test the model for overfitting during the model training phase. The most common cross-validation method is k-fold (Bose, 2019). Bose (2019) stressed the importance of cross-validation is performed on the training data set rather than the testing data set. Within k-fold cross-validation, the training data set is split into an equal number of smaller sets or folds, and the model's accuracy is tested on each fold (Bose, 2019). As a rule of thumb, folds are typically kept at either five or 10 as empirical results indicate these thresholds do not experience high bias or variance within the data (Drakos, 2019). Additionally, more folds in a cross-validation method result in the increased computational time to find the best model (Dantas, 2020; Drakos, 2019). The final accuracy of the k-fold cross-validation is reported as the average across the folds (Drakos, 2019).

Model evaluation methods. As multiple models are developed in data science, model evaluation becomes important to discern which model produces the highest accuracy. Evaluation metrics depend on the type of output or predictor variable produced. For regression models producing a continuous output variable, tests for goodness-of-fit and examination of the residuals historically were accuracy metrics (Boehmke & Greenwell, 2020). Nevertheless, Boehmke and Greenwell (2020) indicated these values could produce misleading conclusions about the accuracy of the models. Boehmke and Greenwell (2020) argued the evaluation of loss functions metrics produces a more accurate method of evaluating regression models. Loss functions evaluate errors between the predicted values to the actual values. One of the most popular metrics in regression models is the R^2 value. This value represents the degree of variance the model accounts for within the data set. Moreover, R^2 has several limitations. Boehmke & Greenwell (2020) mentioned two models for two data sets could have the same root mean squared error but two different R^2 values. The lower R^2 value will be produced from the less variability found within the data. Boehmke & Greenwell (2020) recommended one not emphasize the value of R^2 but continued to recommend additional metrics for evaluating regression models.

Root mean squared error (RMSE) measures the accuracy of predicted values to the actual values for ratio or interval dependent variables (Boehmke & Greenwell, 2020; Kuhn & Johnson, 2013). In squaring the errors between the predicted and actual values, the larger errors have greater penalties (Boehmke & Greenwell, 2020). The additional metric involves taking the square root of the MSE to produce the root mean squared error (RMSE), resulting in the same value as the response variable (Boehmke & Greenwell, 2020). Chugh (2020) stated the MSE examines the error between the predicted and actual values, while RMSE examines the standard deviation of the errors. When comparing the accuracy of multiple regression models, he further reiterated MSE and RMSE are better metrics for accuracy than the R^2 and R^2 adjusted values (Chugh, 2020).

Classification models produce different accuracy metrics than regression models. One of the most common classification metrics for evaluating a model's accuracy is the confusion matrix, as displayed in Table 1. The confusion matrix is a crosstabulation comparison of the predicted and actual outcomes. The matric reveals the performance of the models by revealing the true and false predictions. In the matrix, a true positive would indicate the model correctly predicted the event, while a false positive means the model incorrectly predicted the event. A true negative would mean the model correctly predicted the event not to occur, while a false negative would incorrectly predicted the event not to occur (Boehmke & Greenwell, 2020). From the matrix, multiple metrics can be calculated to evaluate the accuracy. The accuracy rate in which the truths or hits are displayed as a percentage of total events providing how the model performs overall at its maximum capacity (Boehmke & Greenwell, 2020).

Table 1

Classification Model's Confusion Matrix

Actual	Predicted	
	Event	Non-Event
Event	True Positive	False Negative
Non-Event	False Positive	True Negative

However, the accuracy rate does not consider metrics such as precision, sensitivity, and specificity. The model's precision calculates the accurate classification of the events occurring. As a ratio, the precision value compares the true positive to the false positive to evaluate the correct prediction of events that occurred. A model's sensitivity examines the true positives as a ratio to the false negatives to analyze the model's performance in correctly classifying the actual events (Boehmke & Greenwell, 2020). In other words, sensitivity produces a percentage of the true predictive positive cases divided by the total actual true cases (Kuhn & Johnson, 2013). In analyzing the performance of classifying the events not occurring, specificity provides a ratio of true negatives to false positives (Boehmke & Greenwell, 2020; Kuhn & Johnson, 2013). The last two metrics utilized in evaluating classification models is the receiver operating curve (ROC), in which the area under the curve (AUC) can be calculated. According to Dey (2021), the ROC and AUC provide a metric examining the "tradeoff between true positive rate and the predictive value." In other words, the ROC curve plot displays the true-positive and true negative rates (Boehmke & Greenwell, 2020). The point closest to the upper left corner is the most accurate within the plot (Kuhn & Johnson, 2013). Also known as the identity line, a diagonal line represents a model's accuracy equivalent to a coin flip of random guessing (Boehmke & Greenwell, 2020). AUC can be calculated from the ROC graph to determine how accurate the model can classify the events from either occurring or not occurring. Of the metrics used in measuring the accuracy of classification models, Dey (2021) recommended the AUC as the metric to compare and determine which model produces the best results.

Summary

As successful academic performance within the pivotal first year continues to puzzle postsecondary administrators and policy-makers, many retention and attrition models have been theorized to help understand the reason students retain and depart from an institution (Astin, 1984, 1993; Bean, 1980; Spady, 1970, 1971; and Tinto, 1975, 1993). While considered the workhorse for providing state and regional areas with a credentialed workforce, Regional comprehensive universities are unique as they do not get the attraction research universities get from the media but still offer affordable education to students to prepare students for the workforce. These institutions have and continue to face growing concerns impacting the overall enrollment headcounts in recent years. These concerns include the decline of the traditional-age student population coming out of high school, changing demographics of the regional areas, and growing public perception (Barshay, 2018; Henderson, 2016; Lederman, 2019; Livingston & Cohn, 2010; & Nietzel, 2019a).

Variable selection in developing and understanding the first year academic performance is important as it allows for an understanding of what impacts the students in their integration into the academic and social communities. In focusing on the RCUs' student body, variable selection focused on student characteristics, precollege characteristics, financial situations, major declaration, and institutional financial expenditures to identify which students would be at-risk of departing from the institution before a student registers for a class. In the precollege characteristics, incorporating high school characteristics of the curriculum and college readiness index could provide information regarding how well the high school prepares students for life after high school in a postsecondary setting. DeNicco et al. (2015) found English & Language Arts and Mathematics schools' proficiency rates were impactful in persisting to the second fall semester.

Additionally, this study incorporated data science and data mining techniques when previous studies may have only conducted one or two statistical method tests. Data science over the recent years has gained popularity in multiple fields, including higher education in the ability to use data mining techniques to unlock previously hidden information in large data sets. A black-box model approach or one statistical analysis no longer applies in building models to predict outcomes in data science. With building multiple models, accuracy metrics for continuous and dichotomous variables have become important to distinguish which model has the best performance.

Chapter III

METHODOLOGY

The research methods employed in this study are presented in this chapter, which comprises seven sections. The first section delves into the research design, encompassing the independent and dependent variables. The second section provides a description of the population. The third section elucidates the instrumentation and the data collection process. Additionally, within the third section, there is a discussion on the validity and reliability of the data collected. The fourth section outlines the procedures for data analysis concerning each research question. Statistical considerations and assumptions are addressed in the fifth section. The summary of the chapter is presented in the sixth section.

The following research questions guided this study:

1. Are student characteristics, precollege characteristics (including high school curriculum quality), financial situations, major or program of study, and institutional financial expenditures significant predictors in first-time, full-time freshmen's academic performance in their first year?
   a. Are student characteristics (gender, race and ethnicity, family educational background, and locale), precollege characteristics (high school curriculum quality, high school GPA, and admissions test scores), financial situations (family financial situations and financial aid), major or

program of study, and institutional financial expenditures significant predictors of first-time, full-time freshmen's first-fall GPA?

b. Are student characteristics (gender, race and ethnicity, family educational background, and locale), precollege characteristics (high school curriculum quality, high school GPA, and admissions test scores), financial situations (family financial situations and financial aid), major or program of study, and institutional financial expenditures significant predictors of first-time, full-time freshmen's first-year GPA?

c. Are student characteristics (gender, race and ethnicity, family educational background, and locale), precollege characteristics (high school curriculum quality, high school GPA, and admissions test scores), financial situations (family financial situations and financial aid), major or program of study, and institutional financial expenditures significant predictors of first-time, full-time freshmen's one-year retention status?

2. Does one machine learning algorithm (regression, support vector machine, random forest, and extreme gradient boosting) or an ensemble learning algorithm produce a higher accuracy based on the evaluation metrics for accuracy in examination of first-year academic performance?

    a. Does one machine learning algorithm (linear regression, support vector machine, random forest, and extreme gradient boosting) or an ensemble learning algorithm produce a higher accuracy based on the evaluation metrics of the root mean squared error (RMSE) for first semester GPA?

b. Does one machine learning algorithm (linear regression, support vector machine, random forest, and extreme gradient boosting) or an ensemble learning algorithm produce a higher accuracy based on the evaluation metrics of the RMSE for first-year GPA?

c. Does one machine learning algorithm (logistic regression, support vector machine, random forest, and extreme gradient boosting) or an ensemble learning algorithm produce a higher accuracy based on the evaluation metrics of accuracy, sensitivity, specificity, f measure scores, and AUC value for one-year retention status?

Research Design

As a nonexperimental, ex post facto, correlational research design, this study sought to analyze the effects on first year academic performance through the use of archival data obtained from USG, GaDOE, GOSA, and IPEDS. As the study involved assessing the correlations on first year academic performance, research design focused on analyzing the predictability of input variables influencing earned GPAs within the first year and one year retention status. This study was also considered a forecasting and classification study. The forecasting part examined the influence of the independent variables on the earned GPAs within the first year, while the classification part examined the influence of the independent variables on one year retention status.

Independent variables. In this study, 36 independent variables were classified into five distinct groups: student characteristics, precollege characteristics, financial situations, major or program of study, and institutional financial expenditures. Student characteristics variables included gender, race/ethnicity, family educational background, and locale. Precollege characteristics comprise admission test scores, high school GPA, AP advanced standing hours, IB advanced standing hours, CLEP advanced standing hours, and other advanced standing hours. The high school curriculum quality indicators comprised of CCRPI content mastery, CCRPI readiness, EOC mean English and Language Arts, EOC mean Mathematics, EOC mean Science, and EOC mean Social Studies. Financial situation variables encompassed expected family contribution, GA HOPE scholarship dollars, Zell Miller indicator, PELL grant dollars, federal subsidized and unsubsidized loans dollars, and other loans dollars. The major or program of study variable represented the student's primary major in the initial fall semester. Institutional expenditures included instruction, research, public service, academic support, student services, institutional support, and other core expenses.

Overall, the independent variables consisted of 11 nominal and 25 interval variables. Nominal variables consisted of student characteristics (gender, race and ethnicity, family educational background, and locale), pre-college characteristics (five subject areas of the college preparatory curriculum requirements), financial status (Zell Miller recipient), and major or program of study. Within the nominal variables, the race and ethnicity category was based on the required reporting to IPEDS, which combines students' responses to their race, ethnicity, and nationality (NCES, n.d.a). Additionally, race and ethnicity of the available population under 5% were grouped into an "Other" category to protect student anonymity within the data set. Family educational background responses were collected through the admissions process, to which first generation status is given to students indicating neither parent has at least a bachelor's degree (USG, 2023).

The interval level data consisted of precollege characteristics (high school GPA, admission test scores, four advanced standing hours), high school curriculum (CCRPI content mastery, CCRPI readiness, EOC mean English and Language Arts, EOC mean Mathematics, EOC mean Science, and EOC mean Social Studies), financial situations (EFC, HOPE Scholarship, PELL Grant, federal subsidized and unsubsidized loans, and other loans), and institutional expenditures (instruction, research, public service, academic support, student services, institutional support, and other core expenses).

The regional comprehensive universities (RCU) permit students to submit either SAT or ACT test scores. To facilitate this, ACT composite scores were converted to SAT total scores using a concordance crosswalk table. If a student submitted both SAT and ACT scores, the highest value was selected as the student's admissions test scores. HS GPA was measured as a weighted mean with two decimals. This calculation was based on quality points assigned to the grades earned in the classes and the course hours during a student's high school career.

The quality of the high school curriculum was determined by components from the CCRPI and the EOC mean scores. The CCRPI content mastery was measured on an interval scale ranging from 0 to 100 with one decimal place, assessing achievements in the four main high school curriculum subjects—English and Language Arts, Mathematics, Science, and Social Studies. Similarly, CCRPI readiness measured achievements preparing students for either postsecondary education or the workforce. The EOC mean scores were a derived variable from the mean of the total proficient and distinguished scores in corresponding subjects. Mean English was calculated from the mean of 9th-grade Literature and Composition and American Literature and Composition rates. Mean Mathematics was derived from the mean of Algebra I and Geometry rates. Mean Science was determined by the mean of Biology and Physical Science rates. Mean Social Science was computed from the mean of US History and Economics rates. Institutional expenditures were derived values obtained from IPEDS, indicating the amount of dollars rounded to the nearest dollar that the university expends per one full-time equivalent student.

Dependent variables. The dependent variables consisted of three different measurements of academic performance. These measurements were two interval and one nominal levels. The interval levels were comprised of the first-fall and first-year GPAs, while the nominal level comprised of the one-year retention status. The GPA variables were a weight mean calculated based on quality points assigned to the grades earned in the classes and the course hours. The retention status was the nominal variable in which a value of zero indicated a student retained and a value of one indicated a student did not retain to the second fall semester.

Participants

The target population comprised FTFTF cohorts pursuing bachelor's degrees from the four RCUs within USG. According to RPA, a total of 26,356 FTFTF students were enrolled in bachelor's degree programs during the Fall of 2018 and 2019. The gender distribution was 55.1% female and 44.9% male. Demographic breakdown indicated that among the cohorts, 0.2% were American Indian or Alaskan Native, 3.3% were Asian, 27.2% were Black or African American, 10.4% were Hispanic, 0.1% were Native Hawaiian or Other Pacific Islander, 0.8% had an unknown ethnicity, 4.8% identified as two or more races, and 53.1% were White (USG 2022a, 2022b). Reports on academic readiness revealed a mean high school GPA of 3.28 across the cohorts (USG, 2018b, 2019d), a mean SAT composite score of 1012 (USG, 2018a, 2019a), and a mean ACT Composite score of 22 (USG, 2019b, 2019c).

The accessible population enrolled in the four RCUs was the FTFTF pursuing a bachelor's degree who graduated from a public high school within the State of Georgia in 2018 or 2019. The following three criteria needed to be met for students to qualify:

1. IPEDS classification of first-time, full-time freshmen (determined by USG for IPEDS reporting)
2. Pursuing a bachelor's degree (determined by USG for IPEDS reporting)
3. Graduated from a Georgia public high school in 2018 or 2019 (calculated from high school code in RPA census files)

The accessible population included only recently high school graduates to investigate the quality of high school curriculum on first-year academic performance. From the data provided from USG, a total of 21,797 students were identified to have graduated from a GA public high school in 2018 or 2019 and enrolled in one of the four RCUs.

Instrumentation

The study focused on understanding how student characteristics, precollege factors, financial situations, academic engagement, social engagement, and institutional financial expenditures influence the first-year academic performance of FTFTF students enrolled in the University System of Georgia's RCUs. Ary et al. (2019) and Creswell (2014) emphasized the significance of ensuring the validity and reliability of data instrumentation to guarantee the accuracy of results. In this context, validity pertained to the precise measurement of the concept, while reliability refered to the internal consistency of the measured data (Ary et al., 2019; Creswell, 2014). After IRB approval, data were collected from four distinct sources. The first source involved a data request submitted to RPA. The second source wase obtained through the GaDOE's CCRPI website. The third source was acquired from the Governor's Office of Student Achievement's website, and the final source acquired through the IPEDS data center.

USG institutional data. Within Georgia's public sector institutions, the USG's RPA office is responsible for overseeing the collection of census and other data files from each institution. Student enrollment data are collected twice a semester. To maintain the consistency of the data, archival data for FTFTF students was retrieved from RPA, and these census files are stored in the USG's data warehouse. USG has established a reliable process for data collection, adhering to rigorous standards outlined in the data element dictionary (USG 2021c). These standards not only ensure consistency in the collected data for reporting but also facilitate robust data analysis. Furthermore, each institution is required to validate and certify the accuracy of the data submitted to the system-wide database, as per USG guidelines (USG 2021b). The data obtained from USG for FTFTFs enrolled in the RCUs included student-level variables, with a masked ID number and personally identifiable information excluded to preserve students' anonymity.

High schools' CCRPI data. The GaDOE provides information on the CCRPI for public consumption. According to the GaDOE's website, the CCRPI is a metric that the state utilizes to assess the level of college and career readiness of students in accordance with the ESSA law (GaDOE, 2021b). In a published report, the GaDOE's analysis of the validity for content mastery and post-high school readiness shows high levels of confidence (Georgia State University, 2016). The CCRPI scores for each public school are available as downloadable Microsoft Excel files, and these files provide a breakdown of the components that make up the CCRPI. Due to data safeguard procedures, some schools may not have information reported, especially if the student population is small (Data Quality Campaign, 2021). The downloadable data consist of derived or calculated variables for each public school, ensuring no personally identifiable information is contained in the spreadsheets.

High schools' EOC data. According to the GaDOE's Student Assessment Handbook, the EOC tests are considered valid measurements of student achievement for specific subjects. These assessments not only gauge the content learned but also factor into the test taker's final grade. The Official Code of Georgia Annotated emphasizes that all assessment tests must be verified for reliability and validity by a nationally recognized, research-based, third-party evaluator (GaDOE, 2021b, p. 149). In the most recent brief regarding the assessment and accountability of the Georgia Milestone assessment test, validity is examined in terms of "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (GaDOE, 2019, p. 1). The evaluation of the validity of the EOC test was conducted by edCount, LLC. In a recent report published by GaDOE, edCount, LLC, commended the department for the validity of the content measured in the assessments. The third-party company found GaDOE's assessments align with The Standards for Educational and Psychological Testing (Forte et al., 2017).

In the brief from the GaDOE, Cronbach's alpha was employed as a measure of reliability to assess the consistency of the EOC tests (GaDOE, 2019). Within this context, scores are evaluated as "the ratio of true score variance to observed total score variance" (GaDOE, 2019, p. 4). The mean reliability scores for the high school EOC tests from 2016-17 to 2018-19 are presented in Table 2. Across the three reporting years, the scores range from .88 to .93. Utilizing the industry standard rules of thumb for interpreting Cronbach's alpha, all of the scores fall within the range considered acceptable for reliability values (Glen, 2021).

Table 2

Coefficient Alpha Summary of ECOT Reliability Testing

High School EOC Test	2016-17	2017-18	2018-19
9th Grade Literature	.89	.90	.91
American Literature	.88	.90	.89
Algebra I	.90	.91	.91
Geometry	.91	.92	.92
Biology	.91	.92	.92
Physical Science	.88	.90	.91
US History	.92	.93	.92
Economics	.91	.91	.91

Note: Adapted from "An Assessment & Accountability Brief: 2016-17 Validity and Reliability for the Georgia Milestones Assessment System," by Georgia Department of Education, 2017, p. 4-5. Copyright 2017 by the Georgia Department of Education. Adapted from "An Assessment & Accountability Brief: 2017-18 Validity and Reliability for the Georgia Milestones Assessment System," by Georgia Department of Education, 2018c, p. 4-5. Copyright 2018 by the Georgia Department of Education. Adapted from "An Assessment & Accountability Brief: 2018-19 Validity and Reliability for the Georgia Milestones Assessment System," by Georgia Department of Education, 2019, p. 4-5. Copyright 2019 by the Georgia Department of Education.

Similar to the CCRPI data, the EOC data was obtained through downloadable Microsoft Excel spreadsheets hosted on the GOSA's website (n.d.). The EOC data contained in these spreadsheets include the total number of students and percentages for each achievement level. The spreadsheets do not contain any personally identifying information, as the data is aggregated to the school and subgroups within the school. Data protection procedures to safeguard students' information are incorporated into the spreadsheets (Data Quality Campaign, 2021; GOSA, n.d.). Schools and subgroups with fewer than 10 students taking the assessments do not have any numbers or percentages reported. For achievement levels with less than 10 students, only the percentages are reported, as the actual numbers are not disclosed (GOSA, n.d.).

Institutions' financial expenditures. Each year, the National Center for Education Statistics (NCES) collects data related to postsecondary education institutions. The data collected are aggregated numbers pertaining to 12 distinct aspects of the institutions. This three-part collection cycle covers survey components ranging from admissions, degrees conferred, financial information, graduation rates, human resources, library information, to student body information (NCES, n.d.c). Within the financial survey, institutions are required to submit expenses for the fiscal year by function areas. These areas include instruction, research, public service, academic support, student services, institutional support, scholarships and fellowship expenses, auxiliary, hospital services, independent operations, and other expenses (NCES, 2021). The NCES survey collection of financial data mandates institutions to submit information based on either the Governmental Accounting Standards Board (GASB) or Financial Accounting Standards Board (FASB) standards to ensure reliability in the collected data (NCES, 2021). Public institutions are obligated to report information based on the GASB standards, while private institutions must report using FASB standards. These standardized reporting practices allow institutions to submit data consistently with each other, facilitating reliable and comparable financial information across the higher education landscape.

Information collected by the National Center for Education Statistics (NCES) is available for public consumption and analysis. The data was downloaded in Microsoft

Excel spreadsheets from the IPEDS' data center. Within the IPEDS data center, the system provided derived or calculated variables related to expenses divided by the 12-month full-time equivalent headcount for seven functions. These functions included instruction, research, public service, academic support, student services, institutional support, and other core expenses (NCES, n.d.d). The data available for download contained aggregated derived variables, no personally identifiable information was collected from the IPEDS data center.

Data collection. Once IRB granted permission (Appendix A), a data request was made to RPA at USG (Appendix B). This request was to obtain data from the four RCUs regarding the FTFTF first-year academic performance for three cohorts. The data received from RPA had the personal identification removed, eliminating the need for informed consent. The provided data contained unmanipulated information regarding student variables in a Microsoft Excel file. A new scrambled identification variable was created to maintain students' anonymity. Data related to high schools were obtained through downloadable Microsoft Excel files hosted on the GaDOE and GOSA websites. Institutional expenditures was acquired through a downloadable Microsoft Excel file from the IPEDS' data center (NCES, n.d.d). To enhance security, the data was encrypted with a password and stored with multiple backups created for replication of the study. The same data was used for the study's research questions examining the first-year academic performance of FTFTFs enrolled in the USG RCUs. The saved data contained no personally identifiable information to ensure the maintenance of students' anonymity.

Data Analysis

Data analysis consisted of two phases based on the research questions, utilizing the current version of R, a statistical programming software, along with the current tidyverse and tidymodel packages (Korkmaz et al., 2014; Kuhn & Johnson, 2019; Kuhn & Wickham, 2020). The data analysis section had four components to address the research questions. The first section discussed data preparation, explaining how the four data sets was merged into one cohesive data set for analysis. The second section covered the utilization of descriptive statistics to provide an overview of the data. Following that, a discussion of the four predictive algorithms used in inferential statistics was presented. In this third section, each algorithm generateed feature importance to elucidate which input variables impact first-year academic performance for the first research question. Finally, the fourth section described the data science approach for the second research question, aiming to determine which predictive algorithm has the highest accuracy.

Data preparation. As the data for this study originates from four distinct sources, data set mergers were implemented to create a unified data source (Appendix C). One of the variables employed for merging the data sets is the high school code, also known as the College Entrance Examination Board (CEEB) codes. These CEEB codes facilitated the merging of the data obtained from the USG and high school curriculum data. The data sets collected from the GOSA and GaDOE utilize a combination of system and school identification numbers, which are unrelated to CEEB codes (GaDOE, 2021a). To integrate HS curriculum data, an unduplicated list of system and school ID numbers were compiled from the downloaded data sets. Downloadable list produced from WebAdMIT by Liaison (2021) and the NCES HS lookup tool were utilized to develop an up-to-date master list of CEEB codes. Connecting CEEB codes to the system and school ID numbers involved a manual process. After links between CEEB codes and HS data are established, the curriculum variables were merged into the USG data set based on the student's graduation year and CEEB codes. Data obtained from IPEDS was specific to each institution, resulting in a merger based on the institution's name for the corresponding year.

Descriptive statistics. The first part of the data analysis section included descriptive statistics to describe the data set. Descriptive statistics offered a representation or summary of the data rather than any generalization based on probability theory (Chaudhari, 2018; Sha, 2021). For interval variables, the generated statistics measured the central tendency and dispersion of the variables. Central tendency metrics, such as the mean, median, and mode, were employed to describe the symmetry of the data. Symmetrical data exhibited equal mean, median, and mode values (Chaudhari, 2018). Metrics for dispersion, including variance, standard deviation, range, quartiles, skewness, and kurtosis values (Chaudhari, 2018; Sha, 2021), were utilized to examine the spread of the data. These metrics contributed to a comprehensive understanding of the distribution characteristics. Descriptive statistics for categorical variables involved the frequencies and percentages of data within each category of the variable. Categorical descriptive statistics also encompassed the mode of the variable. The functions summary() and skim() was employed to generate descriptive statistics for subsequent review and analysis.

Inferential statistics. A total of four algorithms were employed to analyze the impact of the independent variables on the three dependent variables for the first research question (Appendix D). These algorithms included linear regression, logistic regression, support vector machines, random forest, and extreme gradient boosting. Additionally, the algorithms were integrated into an ensemble learning method. Overall, the models followed a supervised learning approach (James et al., 2013). According to James et al. (2013), supervised learning involves developing a model based on predictor variables and an outcome variable with the aim of accurately predicting the outcome and understanding the influence between the variables.

Linear regression, one of the most widely used statistical tools for modeling continuous variables, aims to investigate the influence of predictor variables on outcome variables (James et al., 2013). Similarly, logistic regression is employed to examine the impact of predictor variables on a dichotomous outcome based on conditional likelihood (James et al., 2013; Schmidt-Thieme, 2007). As a modeling tool, the support vector machine seeks to create margins of separations, known as hyperplanes, to discover the maximum margin for optimal explanation (Gandhi, 2018; James et al., 2013). The random forest algorithm develops a collection of decision trees over multiple subsets of the training data set through recursive partitioning (Richmond, 2016). The extreme gradient boosting (XGBoost) algorithm uses a combination of boosted trees and conditional random fields (Brownlee, 2016).

Linear regression. A multiple linear regression model was constructed using the linear_reg() function within the parsnip package. The model was configured for regression, and the engine was set to lm by default (Kuhn & Wickham, 2020). Utilizing the slope-intercept formula, the linear regression function generated a linear line of predictive outcomes.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

Before linear regression can be employed, considerations and assumptions were reviewed. Missing data and outliers were factors to be taken into account for linear regression. Addressing missing data were essential, as some data may be structurally missing due to safeguard protections to preserve individual anonymity (Bock, nd; Data Quality Campaign, 2021). Another consideration involved the examination of univariate and multivariate outliers, as they could introduce unnecessary bias to model parameters and estimates, potentially leading to Type I or Type II errors (Osborn & Overbay, 2004). Data visualizations, z-score review, and statistical tests were employed to identify univariate outliers. Grubbs' test is a statistical test used for identifying univariate outliers. For multivariate outliers, the Mahalanobis distance test, measuring the distance between two points based on the covariance of the data, was utilized (Cansiz, 2020). Values flagged as outliers underwent an individual case review to determine whether they are valid data points or outliers.

Linear regression relied on a total of four assumptions that must be met to draw meaningful research conclusions based on reality (Field et al., 2012; Garson, 2012). The observation independence assumption asserts there are no duplicated records for the event, as duplication could introduce bias toward observations appearing more than once (Heidel, 2022). Univariate normality, another assumption, pertains to the equal distribution of the data (Merler & Vannatta, 2002). This assumption was assessed through data visualizations such as histograms and Q-Q plots, as well as statistical tests like Shapiro-Wilks, Jarque-Bera, and Kolmogrovo-Smirnov tests (Kuhn & Johnson, 2019; Mishra et al., 2019). Multivariate normality was a crucial assumption checked by observing the combined effect of multiple variables on the distribution of the data. An elliptical shape in scatterplots would indicate multivariate normality (Fife, 2019; Oppong & Agbedra, 2016). Additionally, Royston's and Mardia's tests were employed to make a statistical decision regarding multivariate normality (Oppong & Agbedra, 2016). The linearity assumption asserts the relationship between input and output variables resembles a straight line (Merler & Vannatta, 2002). Pearson's R correlation coefficients were used to assess linearity between input and output variables (Glen, 2022). Multicollinearity were examined using the VIF test to determine whether independent variables are highly correlated, violating the multicollinearity assumption (James et al., 2013; Leung, 2021). VIF values exceeding the rule of thumb thresholds of five or 10 suggest multicollinearity violation (Bhandari, 2020; James et al., 2013). Variables exceeding the thresholds underwent a process to eliminate the multicollinearity, starting with the highest VIF score (Bhandari, 2020). The homoscedasticity assumption requires equal variance within the data. The Levene's test was employed to test for homoscedascity (Merler & Vannatta, 2002).

The linear regression model produced an F-statistic, R^2, R^2 adjusted, RMSE, and *p*-value. These values aided in assessing the accuracy of the model. The R^2 and R^2 adjusted were values accounting for the variance found within the data. The value ranged from zero to one. A value of zero indicates the model accounts for no variance, resulting in a poor model, while a perfect model would exhibit the value of one (James et al., 2013). While the R^2 is an easy interpretation of the models fit, a drawback occurs from it ability to not measure the fit of the predictions (Hiregoudar, 2020; James et al., 2013).

The RMSE values relates to the overall fit of the predicted outcomes compared to the actual outcomes. Values close to zero indicate an overall good fit (Hiregoudar, 2020).

For each of the independent variable, linear regression produced coefficients, standard error, *t*-statistics, and *p*-values (James et al., 2013). Predictor variables with a *p*-value less than or equal to the significant threshold were considered influential to the outcome variable. Moreover, significant independent variables' strength and direction of the relationship to the dependent variable were determined by the coefficients. The larger the value of the predictor indicated the more of an influence in the model. The value's positive or negative sign determined the direction of the relationship (James et al., 2013).

Logistic regression. Multiple logistic regression model were developed using the logistic_reg() function within the parsnip package. The model was set to classification and the engine was set to glm, which is the default (Kuhn & Wickham, 2020). The logistic regression algorithm modeled the probability of the student retaining to the second fall semester (James et al., 2013).

$$\mathrm{Pr}\ (\ \mathrm{default} = \mathrm{retained}|\mathrm{balance})$$

Using the logistic function, the outcome was produced the probability in an S-shaped curve of the student retaining ranging from zero to one in which student who would be classified as retained receiving the value of one (James et al., 2013).

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n}}$$

Generally, the decision threshold of the produced probabilities would indicate values less than or equal to .50 would be considered retaining, while values greater than .50 would be considered not retaining (James et al., 2013).

In the application of logistic regression, certain assumptions must be met. Logistic regression also required the assumptions of observation independence and outliers to not be violated. The independence of observation assumption stipulated there should be no duplication of records for the event (Leung, 2021). The presence of strong outliers, which can influence the estimates, were assessed using Grubb's test (James et al., 2013; Leung, 2021). The evaluation of outliers was conducted on a case-by-case basis, considering that values may genuinely reflect true data points. In addition to these shared assumptions, logistic regression had distinct requirements compared to linear regression. The outcome variable needed to be dichotomous (Leung, 2021), and the independent variables needed to exhibit a linear relationship with the log odds (James et al., 2013; Leung, 2021). The absence of multicollinearity was another crucial assumption for logistic regression, wherein independent variables should not be highly correlated with each other (Leung, 2021). This relationship were examined through VIF. The VIF tested for multicollinearity between the independent variables and the collinearity of the independent and dependent variables (James et al., 2013; Leung, 2021). Rule-of-thumb values of five or 10 are considered violations of the multicollinearity assumption (Bhandari, 2020; James et al., 2013). Variables exceeding these thresholds underwent a process to eliminate the multicollinearity, starting with the highest VIF score (Bhandari, 2020).

The logistic regression model produceed beta values, standard errors, *p*-values, χ^2 values, degrees of freedom, and odd ratios (James et al., 2013). These values indicated the significant factors impacting the student's probability of retaining or not retaining. The Hosmer and Lemeshow goodness-of-fit, Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), and the pseudo-R^2 outputs assisted in determining how the logistic regression model fits. As a goodness-of-fit, Hosmer-Lemeshow statistic provided a value indicating whether the model reflects the true outcomes within the data (Hosmer et al., 2013). Within the Hosmer-Lemeshow goodness-of-fit, the logistic regression model would not be a good fit if the *p*-value is significant (i.e., less than or equal to .05) (Hosmer et al., 2013). Like the Hosmer-Lemeshow, the AIC and BIC values provided a goodness-of-fit value, with BIC penalizing the more complex model. The larger the value of AIC and BIC indicated the model is not a good fit (Brownlee, 2019). Logistic regression also had multiple proposed pseudo-R^2 values to produce an equivalent to linear regression's R^2 value measuring the variance within the regression model. While there are multiple ways of calculating the pseudo-R^2, the consensus indicated the McFadden's pseudo-R^2 is the best value (Abhigyan, 2020).

Additionally, a confusion matrix was produced based on the predict outcomes compared to the actual outcomes. From the matrix, the overall accuracy, sensitivity, specificity, and F-score were produced (James et al., 2013; Narkhede, 2018). More accuracy metrics produced from the logistic regression model were the ROC and the resulting AUC value. The values of the AUC ranged from zero to one. A value of .50 was considered the events occurring by chance or a coin flip, while a value of 1.00 was considered a perfect model (James et al., 2013; Narkhede, 2018).

Support vector machine for regression and classification. Support vector machine (SVM) algorithms were adept at handling both continuous and categorical outcome variables (Yadav, 2018). Utilizing hyperplanes to establish decision criteria, SVM aimed to achieve optimal results by maximizing margins (Gandhi, 2018; James et al., 2013; Yadav, 2018). SVM gained popularity in assessing the impact of predictors on outcome variables due to its capability to use multiple kernels, enabling analysis of different relationship shapes—linear, nonlinear, and radial-based functions (Awasthi, 2020). Attewell and Monaghan (2015) recommended using multiple kernels to optimize SVM performance. The SVM algorithms were employ linear, polynomial, and radial-based functions. Using the parsnip package, three different functions were performed based on the type of SVM (Kuhn, 2019).

For the three dependent variables, SVM used the svm_linear() function for the linear model, svm_poly() function for the polynomial model, and svm_rbf() function for the radial-based function model (Kuhn & Wickham, 2020). Each model used the cost function to determine the optimal value in the accuracy of predictions (James et al., 2013). Unlike linear and logistic regression, SVM had no assumptions must be reviewed, but were vulnerable to large class imbalance for classification models (Batuwita & Palada, 2012; Dwivedi, 2020). Machine learning algorithms, like SVM, do not produce coefficient values akin to general linear models that measure the impact on the variable in the model. To gain insight into the factors influencing the model, a variable importance analysis was conducted. The scores from the variable importance analysis indicated the impact the factor contributes to the outcome variable (Shin, 2021). For the GPA dependent variables, accuracy metrics involved the calculation of the RMSE to indicate the overall fit of the model. Accuracy for the SVM models for the retention variable analyzed the confusion matrix, producing the overall accuracy rate, sensitivity, specificity, and F-score. Additional accuracy metrics involved the ROC and more specifically, the AUC (James et al., 2013; Narkhede, 2018).

Random forest. Random forest has emerged as one of the most popular predictive algorithms in contemporary data science. The random forest algorithm was a collection of decision trees developed to predict outcomes. Each tree within the algorithm was built using multiple subsets of the training data set, preventing overfitting for both classification and regression analyses. Unlike general linear models, random forest had no formal assumptions, making it capable of handling skewed and multi-modal data sets (Richmond, 2016). Due to its lack of formal assumptions, random forest algorithm required minimal data transformations (Ravindran, 2021).

The parsnip package's rand_forest() function from the tidymodel library utilized with the regression model for GPA predictions and the classification model for retention predictions. In the rand_forest() function, mtry referred to the number of random samples of variables, and trees refers to the number of trees developed. A variable importance analysis was conducted to understand the impact of features on the predicted outcome. Random forest accuracy was evaluated using two methods based on the type of dependent variable. For first-fall and first-year GPA dependent variables, accuracy was measured using the RMSE value, describing the overall fit of the predicted outcomes (Hiregoudar, 2020; James et al., 2013). Values close to or at zero indicated a good model. The random forest model for the dichotomous dependent variable—one-year retention status—were assessed using an analysis of the confusion matrix and the ROC. The confusion matrix provided overall accuracy, sensitivity, specificity, and F-score. Additionally, the ROC plot and AUC served as additional accuracy metrics (James et al., 2013; Narkhede, 2018).

Extreme gradient boosting. Like random forest, extreme gradient boosting (XGBoost) has gained popularity in recent years. The XGBoost algorithm was capable of handling both continuous and categorical outcome variables (Brownlee, 2016; Xgboost Developers, 2021). Developed by Tianqi Chen, XGBoost combines boosted trees and conditional random fields, showcasing improved computation speed and model performance. Pafka (2015) highlighted XGBoost's speed outperforming other random forest models across various statistical tools, and it has become a favored algorithm in data science competitions, often used by winners due to its accuracy (Brownlee, 2016).

Within the tidymodel parsnips package, the boost_tree() function were employed with the xgboost engine to create the model. The regression model was used for predicting the dependent variables related to the first-fall and first-year GPAs, while the classification model was applied for the retention status (Kuhn & Wickham, 2020). A variable importance analysis was conducted to understand the impact of features on the predicted outcome. Accuracy measurements for XGBoost mirrored those for random forest, involving two evaluations based on the type of dependent variable. For the two continuous variables—first-fall and first-year GPAs—accuracy was measured using RMSE, describing the overall fit of the predicted outcomes (Hiregoudar, 2020; James et al., 2013). Values close to or at zero indicated a good model. The XGBoost model for the dichotomous dependent variable—one-year retention status—was assessed using an analysis of the confusion matrix and the ROC. Within the confusion matrix, overall accuracy, sensitivity, specificity, and F-score was considered. Additionally, the ROC plot and AUC was used as additional accuracy metrics (James et al., 2013; Narkhede, 2018).

Ensemble learning. Ensemble learning involved the construction of multiple predictive algorithms to generate a single predictive outcome. Combining predictive outcomes from various algorithms offered several advantages. Ensemble learning algorithms often achieve higher accuracy compared to individual models, reduce bias and variance to avoid underfitting or overfitting, and demonstrate increased stability in predictive outcomes (Makhijani, 2020; Ravanshad, 2018). However, ensemble learning has drawbacks. It diminished the interpretability of the model, making it challenging to understand the factors influencing predictions. Although variable importance analysis helps discern impactful factors, Ravanshad (2018) noted ensemble learning often involves non-linear and interaction effects that variable importance analysis cannot fully explain. Additionally, ensemble learning increased computation time for prediction (Ravanshad, 2018). For regression analyses, a common ensemble method involved simple averaging of predicted outcomes. Simple averaging can also be used to average probabilities for classification outcomes. In classification analyses, the max voting method helps assign the classification based on the majority vote from predictive models with the same classification (Singh, 2018). The stacks package, developed by Couch and Kuhn (2022), was suitable for ensemble learning and integrated well with the tidymodels package to create a new model from the outputs of multiple models.

Data science approach. In the data science approach to data mining, the optimization for achieving the highest accuracy in addressing the second research question involved comparing the accuracy of different machine learning algorithms (Calvo & Santafé, 2016; Horthorn et al., 2005). In data science, the significance of having an accurate model lied in its out-of-sample predictive power (Kuhn & Johnson, 2013; Kuhn & Johnson, 2019). Calvo and Santafé (2016) noted some differences between models may not be obvious, emphasizing the importance of using assessment tools to determine the optimal model and avoid misleading conclusions. By comparing accuracy metrics and the ROC, each model was thoroughly assessed to identify the best-performing one (James et al., 2013).

The evaluation of regression and classification models was based on their performance on the test data set, which provided an unbiased assessment of how the models performed on unseen, simulated real-world data (Goyal, 2021; Shah, 2017; Soni, 2019). Ten-fold cross-validation was conducted on the training data set to assess potential accuracy before evaluating each model's real accuracy on the test data set. The same cross-validation was applied to the test data set, producing 10 accuracy metrics from which an average accuracy metric was derived. For the assessment of models predicting continuous dependent variables, the evaluation involved the RMSE (Boehmke & Greenwell, 2020; James et al., 2013). These metrics represented an overall measure of the difference between predicted and actual outcomes, with a value of zero indicating a perfect model that accurately predicts the outcome variable (James et al., 2013; Plagata, 2020). Boehmke and Greenwell (2020) highlighted the significance of RMSE values in determining the accuracy of regression outputs over other metrics. In contrast, classification models were evaluated using different accuracy metrics, including overall accuracies, sensitivities, specificities, F-scores, and AUCs to assess out-of-sample predictive power (Calvo & Santafé, 2016; Horthorn et al., 2005; James et al., 2013; Kuhn & Johnson, 2013; Kuhn & Johnson, 2019).

The accuracy of the models were examined through three statistical tests: Mann-Whitney test, Friedman's test, and Wilcoxon signed-ranked test (Demšar, 2006; Fernández-Delgado et al., 2014). The Mann-Whitney test measured the validity between differences in the training and testing data sets. Friedman's test identified statistically significant differences between the data sets' accuracy metrics. The Wilcoxon signed-ranked test, conducted as a pairwise comparison through an ad hoc test to determine the most accurate model (Demšar, 2006; Fernández-Delgado et al., 2014).

Statistical Considerations and Assumptions

During the data analysis, it was essential to thoroughly review statistical considerations and assumptions to ensure the validity and reliability of the results obtained from the predictive models. The nature of these considerations and assumptions can vary depending on the type of statistical analysis being conducted. For instance, some analyses, like general linear models, may have more rigid rules compared to others such as random forest. As stated by Garson (2012), violations of assumptions may not impact conclusions in some less stringent analyses, but in others, they can undermine meaningful research. Field et al. (2012) emphasized that assumption violations can hinder the ability to draw meaningful conclusions based on reality. To review considerations and assumptions, various avenues were explored, employing data exploration techniques that utilize data visualization tools and statistical tests (Appendix E). These methods aimed to uncover potential issues or violations that could affect the robustness of the models and their outcomes.

Considerations. Descriptive statistics played a crucial role in reviewing the data to identify any instances of missing data. The summary() function in R provided a quick overview of the data variables, allowing for a preliminary examination. Additionally, the skim() function from the skimr package were employed to present descriptive statistics, including histograms, alongside results for each variable, providing a comprehensive view of multiple data types (Medcalf, 2018). Both summary() and skim() functions reported the number of missing data points for each variable. In the context of archival educational data sets, missing data could have occurred due to various reasons, including structural missing data designed to protect individual anonymity and ensure effective data use for student support (Bock, nd; Data Quality Campaign, 2021). Other types of missing data may have been manifested as missing completely at random or missing at random in the data set provided by RPA. To identify missing data, a count of the rows and visualizations was examined. Imputation strategies were employed for handling missing data. For continuous variables, methods such as mean, median, and k-nearest neighbor were utilized to provide the best fit for missing data. Categorical missing variables were imputed using the "other" method. The primary goal of imputing missing data was to avoid skewing the central tendencies of the data. Observations containing missing values within the dependent variables were excluded from the data set to maintain data integrity and reliability in subsequent analyses.

Examining outliers within the data was another critical consideration, as the presence of outliers may lead to Type I or Type II errors, introducing biased influences on predictive models' parameters and estimates (Osborne & Overbay, 2004). Outliers can significantly impact the reliability of the analysis. The assessment for outliers encompassed various techniques, including the evaluation of summary statistics, distribution of z-scores, histograms, boxplots, and Q-Q plots (Korkmaz et al., 2014; Kuhn & Johnson, 2019; Soetewey, 2020). During the review of z-scores, any value less than -3.00 or greater than 3.00 was flagged as a potential outlier. While z-scores are effective, they can be sensitive to extreme values, impacting the mean (Alam, 2020). Moreover, z-scores were not suitable for evaluating potential outliers in categorical data. To complement this, visualizations allow for the identification of extreme values as they provided a method to examine predictor variables' data concerning the impact on the outcome variables (Silge, 2020).

To identify outliers in the continuous variables, the Grubbs' test was employed from the outliers package (Soetewey, 2020). From the Grubbs' test, values as potential outliers were identified (Soetewey, 2020). Notably, a univariate outlier may not necessarily signify an outlier, as it could be a multivariate outlier when compared to another variable. For detecting multivariate outliers, the Mahalanobis distance test was highly effective. The test measured the distance between two points based on the covariance of the data, calculating the number of standard deviations two points are away from each other (Cansiz, 2020). Utilizing the mahalanobis() function from the stats package, the analysis explored multivariate outliers. Any values flagged as outliers underwent individual case review to ascertain whether the value was a valid data point or an outlier.

Assumptions. Observation independence was a critical consideration in statistical analyses, and adherence to the assumption that no record occurs more than once in the data set was crucial (Heidel, 2022). Violations of this independence observation assumption can introduce bias in favor of duplicated observations, potentially distorting the results (Heidel, 2022). To assess compliance with this assumption in the data collected for each institution, a count of the masked identification numbers was examined. This count helped to determine whether any records have been duplicated, providing insight into the adherence to the observation independence assumption.

Beyond observation independence, three additional assumptions were vital for examining the data distribution to ensure no violations have occurred, enabling accurate conclusions. Skewness affected many statistical models significantly, as data with considerable skewness can disproportionately influence the model's estimates (James et al., 2013; Sharma, 2019). Normality pertained to the even distribution of data (Merler & Vannatta, 2002). Histograms and Q-Q plots were among the most common visualizations used to assess normal distribution (James et al., 2013; Kuhn & Johnson, 2019). Histograms allowed for the visualization of the distribution across the x-axis, enabling researchers to observe the spread and shape of the data. Q-Q plots compared the distribution of observed values (on the x-axis) with the expected values under a normal distribution (on the y-axis). A deviation from a straight line in a Q-Q plot indicated a violation of normality, suggesting the data may not be normally distributed (Merler & Vannatta, 2002).

Additionally, the descriptive statistics generated by the summary() function included skewness and kurtosis values. Skewness values measured the symmetry, while kurtosis values measured the height or peakedness of the distribution. Skewness and kurtosis values around zero typically indicated normality. A positive skewness value suggested a distribution with a long right tail, whereas a negative skewness value points to a long-left tail. A positive kurtosis value indicated a distribution with a high peak and short tails, whereas a negative kurtosis value suggested a flatter distribution with longer tails (Merler & Vannatta, 2002). Furthermore, even when visual inspections might not reveal skewness, statistical tests were able to detect it. The Shapiro-Wilk and Jarque-Bera tests were two statistical methods used for assessing normality. The Shapiro-Wilk test is grounded in frequentist statistics, whereas the Jarque-Bera test is based on moments (Tomšik, 2019). According to Tomšik (2019), among all normality tests, the Shapiro-Wilk test was the most effective at detecting normality, with the Jarque-Bera test being the second most powerful. A normality violation was indicated when the *p*-value is less than or equal to .05 (Merler & Vannatta, 2002; Mishra et al., 2019). Variables identified with skewness were subjected to various data transformations to determine the most suitable transformation method. These methods included the Box-Cox, Yeo-Johnson, and logarithmic transformations. For negatively skewed data, inverse transformations were required before applying Box-Cox and logarithm functions to function correctly. These data processing techniques aimed to transform the data into a symmetric distribution to minimize the undue influence of extreme values or distributions on the estimates (James et al., 2013; Kuhn & Johnson, 2019).

With the normality assumption, checks for multivariate normality were also a crucial assessment of the distribution. Like univariate normality, multivariate normality evaluated the skewness of the data; however, unlike univariate normality, multivariate normality examined the skewness by combining more than one variable (Wang, 2020). In essence, multivariate normality evaluated the combined effect multiple variables have on the distribution of the data (Sucky, 2020). One method to detect a violation of multivariate normality was through the analysis of scatterplots of the independent variables. An elliptical shape in the scatterplot distribution indicated multivariate normality (Fife, 2019; Oppong & Agbedra, 2016). Fife (2019) cautioned against confirmation bias when evaluating scatterplots and recommended the inclusion of either regression or lowess lines in the scatterplot. Additionally, Oppong and Agbedra (2016) suggested using Royston's and Mardia's tests to conclude any violation of the multivariate normality assumption. They explained that Royston's test was analogous to the Shapiro-Wilk test for univariate normality and Mardia's test calculates skewness and kurtosis values (Oppong & Agbedra, 2016). According to Oppong and Agbedra (2016), a violation of multivariate normality was indicated by a *p*-value less than or equal to .05 in Royston's test. For Mardia's test, skewness and kurtosis values at or near zero indicated a normal distribution (Merler & Vannatta, 2002).

Linearity, as a statistical assumption, referred to the relationship between the independent and dependent variables resembling a straight line (Merler & Vannatta, 2002). The presence of linearity in variables was crucial because a large portion of statistical analyses rely on the "linear combination of variables" (Merler & Vannatta, 2002, p. 32). One common method for assessing linearity was through the examination of the Pearson's R correlation coefficient, which determined the strength and direction of the relationship between the variables. In the psych package for R, the setCor() function assessed the correlation of the variables at each level (Revelle, 2020). Values at or close to zero indicated no linear relationship, thus violating the linearity assumption. Conversely, values of negative one or positive one indicated a perfect or very strong linear relationship (Glen, 2022). Additionally, evaluating linearity also involved examining the residuals or prediction errors. Residual plots that show no violation of linearity would have values evenly distributed along the zero line (Merler & Vannatta, 2002).

Finally, homoscedasticity refered to the variance within the data set to be equal. A violation of the equal variance would indicate the data is heteroscedastic (Merler & Vannatta, 2002). Testing for homogeneity of the variance was through using the Levenne's test. A violation of homoscedasticity occurred when the *p*-value is less than .05. This would mean the null hypothesis of equal variance was rejected (Merler & Vannatta, 2002). Merler and Vannetta (2002) indicated the analysis was not necessarily doomed when the Levene's test indicates heteroscedasticity.

Data imbalance. Data imbalance may be the result of small numbers of observations within one group of a classification study (Google Developers, 2021; Rocca, 2019). As a result of the data imbalance, predictive algorithms developed may automatically default to one classification over the other leading to a greater potential of a Type II error (Nallamuthu, 2020; Rocca, 2019). One sign of the greater potential for Type II error is identified from poor sensitivity found within the confusion matrix (Kuhn & Wickham, 2020; Sharma et al., 2009). Countering the imbalance is the need to have more representational data in the training data set to improve the accuracy of the algorithm (Rocca, 2019). In data science, the utilization of upsampling or downsampling are techniques to correct the dataset imbalance. The upsampling technique deals with the imbalance within the data set pertaining to the minority class, while downsampling deals with the imbalance in the majority class. Upsampling inserts new data into the minority class to the extent where both classifications are almost equal in size. In upsampling, one concern is the deliberate bias induced from adding more data in the minority class. Conversely, downsampling reduces the size of the majority class. Like upsampling, downsampling introduces bias towards the minority through a reduction of data in the majority class (Nallamuthu, 2020).

Data leakage. Within machine learning, the impact of data leakage becomes a major concern during data preprocessing. Soni (2019) defined data leakage as the sharing of information used in building the model between the training and testing data sets. If data preprocessing occurs before data splitting, it goes against the goal of having a testing data set that represents unseen or real-world data for the model to make predictions (Goyal, 2021; Soni, 2019). To split the data, the set.seed() function was used for the replication of the data split (James et al., 2013). In the tidymodel package, the initial_split() function was used to split the data into training and testing sets. The prop option in the initial_split() function was set to reserve 60% of the data for the training data set and the remaining 40% for the testing data set. The training() and testing() functions within tidymodels, called from the initial_split() result, allow the data splits to be saved as objects in the R environment (Kuhn & Wickham, 2020).

Data preparation. The tidymodel package's recipe() function was utilized to build the preprocessing method on the training data set. Additionally, the recipe() function preserved the methods used to ensure the same processes are applied to the testing data set. The initial step in constructing the recipe() function involved addressing any class imbalances resulting from the data splitting (Kuhn & Wickham, 2020; Silge, 2020).

For the retention data set, the step_upsample() function performed random upsampling or oversampling of the minority classes, with the default over_ratio setting of 1 to introduce more instances of the minority classes to equalize their numbers with the majority classes. Similarly, the step_downsample() function conducted random downsampling or undersampling of the majority classes, with the default under_ratio setting of 1 to reduce the instances of the majority classes to match the minority classes (Kuhn & Wickham, 2020; Silge, 2020). The correction of the imbalance was examined through both upsampling and downsampling methods.

Variables with missing data were either deleted or imputed. Only records with missing dependent variables were deleted using step_filter(), where the dependent variable was processed through the is.na() function. For missing data in categorical variables, step_impute_unknown() were used, while continuous variables with missing data were imputed using step_impute_mean(), step_impute_median(), and step_impute_knn(). These options allowed for the exploration of the best method for handling missing data in the data set. All predictor numerical values were normalized using the step_normalize() function (Kuhn & Wickham, 2020). According to Kuhn and Wickham (2020), this function normalized the data so the mean of the variable is zero, and the standard deviation is one. In other words, it standardized numerical values with respect to the variable's standard deviation and mean to prevent high values from exerting undue influence on the estimates (Browne-Anderson, 2016; Kuhn & Wickham, 2020; Roy, 2020). To address heterogeneity, step functions used in the recipe were built to correct the distribution. These transformation functions included step_BoxCox(), step_YeoJohnson(), and step_log() (Kuhn & Wickham, 2020).

To ensure optimal performance, resampling methods or cross-validation techniques were employed. One of the most common cross-validation techniques was the k-fold method. In this approach, the data was separated into specified equal partitions or folds, which were then used to assess the performance of the developed models. The function utilized for cross-validating the training and testing data sets was vfold_cv(), with the number of folds set to 10 (Kuhn & Wickham, 2020).

Summary

In this nonexperimental, correlational study, three research questions investigated the academic performance of two cohorts of FTFTF students enrolled in RCUs in the State of Georgia. The analysis involved a total of 36 independent variables, encompassing student characteristics, precollege attributes, financial situations, academic engagement, social engagement, and institutional financial expenditures. A data science approach were employed to examine the relationships and accuracy of each model concerning the dependent variables of first-year academic performance, including first-fall GPA, first-year GPA, and one-year retention status. The study's analysis was conducted using the R software and the tidymodel package.

The first research question in this study aimed to assess the predictability of 36 independent variables on three distinct first-year academic performance measures. These academic performance variables included first-fall GPA, first-year GPA, and one-year retention status. Four models, each tailored to a specific dependent variable, were constructed to investigate the influence of input variables on each output variable. Before developing the models, necessary considerations and assumptions were examined. The data underwent a 60/40 partitioning for training and testing data sets, employing a 10-fold cross-validation method. This cross-validation approach created 10 different data sets to enhance the validity of estimates within each model. The models were developed using the training data, and their accuracy was evaluated against the testing data set. The metrics obtained from the testing data set were used to assess the performance of the models, addressing the second research question.

The second research question employed a data science approach to analyze various predictive models, aiming to identify the most accurate model. For the dependent GPA variables, the model accuracy was assessed using metrics such as R^2 and RMSE. Evaluation of the one-year retention models utilized overall accuracy, sensitivity, specificity, F-score, and AUC. The comparison of accuracy metrics across models involved the use of significant tests and visual comparisons. This comprehensive analysis aided in determining which model exhibits the highest level of accuracy for the specified criteria.

Chapter IV

RESULTS

The chapter serves a three-fold purpose. First, the chapter aims to identify significant factors for first-year academic performance by examining several factors, such as student characteristics, pre-college characteristics, financial situations, major of study, and institutional expenditures. The second purpose is to identify which predictive algorithm exhibits the best accuracy amongst those utilized in the study. Lastly, the chapter aims to explore whether ensemble learning methods could enhance the predictive power of the algorithms. The research questions guiding this study are as follows:

1. Are student characteristics, precollege characteristics (including high school curriculum quality), financial situations, major or program of study, and institutional financial expenditures significant predictors in first-time, full-time freshmen's academic performance in their first year?
    a. Are student characteristics (gender, race and ethnicity, family educational background, and locale), precollege characteristics (high school curriculum quality, high school GPA, and admissions test scores), financial situations (family financial situations and financial aid), major or program of study, and institutional financial expenditures significant predictors of first-time, full-time freshmen's first-fall GPA?
    b. Are student characteristics (gender, race and ethnicity, family educational background, and locale), precollege characteristics (high school

    curriculum quality, high school GPA, and admissions test scores), financial situations (family financial situations and financial aid), major or program of study, and institutional financial expenditures significant predictors of first-time, full-time freshmen's first-year GPA?

    c. Are student characteristics (gender, race and ethnicity, family educational background, and locale), precollege characteristics (high school curriculum quality, high school GPA, and admissions test scores), financial situations (family financial situations and financial aid), major or program of study, and institutional financial expenditures significant predictors of first-time, full-time freshmen's one-year retention status?

2. Does one machine learning algorithm (regression, support vector machine, random forest, and extreme gradient boosting) or an ensemble learning algorithm produce a higher accuracy based on the evaluation metrics for accuracy in examination of first-year academic performance?
    a. Does one machine learning algorithm (linear regression, support vector machine, random forest, and extreme gradient boosting) or an ensemble learning algorithm produce a higher accuracy based on the evaluation metrics of the root mean squared error (RMSE) for first semester GPA?
    b. Does one machine learning algorithm (linear regression, support vector machine, random forest, and extreme gradient boosting) or an ensemble learning algorithm produce a higher accuracy based on the evaluation metrics of the RMSE for first-year GPA?

c. Does one machine learning algorithm (logistic regression, support vector machine, random forest, and extreme gradient boosting) or an ensemble learning algorithm produce a higher accuracy based on the evaluation metrics of accuracy, sensitivity, specificity, f measure scores, and AUC value for one-year retention status?

The data utilized for the study were from four different sources. The primary data source originated from the USG, encompassing the Fall 2018 and Fall 2019 FTFTF bachelor's degree seeking cohorts from the system's four RCUs. The second data set came from the GaDOE's website containing information about the CCRPI for 2018 and 2019. The third data set came from the Georgia GOSA's website containing information on the end of course's (EOC) subject proficiency levels for 2018 and 2019. The final data set was acquired from the IPEDS data center, which stores institutional expenditures per FTE for FY2018 and FY2019. After processing the four data sets, they were consolidated into a single data set for the analysis of the study.

The chapter contains the results of the data analysis for each research question and is comprised of six sections. The characteristics of the population section are comprised of demographic and descriptive statistics of the entire FTFTF population, and the population identified to have graduated from a GA public high school. Data splitting and imbalance correction methods are addressed in the second section. The next section contains the critical review of the preliminary considerations and assumptions for statistical analysis to ensure the validity and reliability of the statistical inferences produced in the findings. The next two sections encompass the analyses of two research questions on the significant factors impacting academic performance—first-fall GPA, first-year GPA, and one-year retention status—and the predictive power of the predictive algorithms. The summary of the results is in the last section.

Population Characteristics

GA public high school represented. The HS curriculum data were collected through combining the CCRPI content mastery and readiness rates and EOC subject proficiency levels together in one data set. Table 3 displays the rates of the GA public HS. Moreover, the schools were filtered down to the GA public HS represented in the four RCUs within USG. Of the GA public HS represented in the RCUs for the Fall 2018 and Fall 2019 FTFTF students, the mean content mastery (*M* = .639, *SD* = .185) was slightly higher than the overall (*M* = .600, *SD* = .222). Likewise, the readiness scores were higher for the students enrolled in an RCU (*M* = .726, *SD* = .112) than the overall (*M* = .707, *SD* = .169). The difference in the EOC proficiency levels for the four subjects ranged between .027 to .030 higher for students enrolled in an RCU. The English proficiency levels (*M* = .467, *SD* = .179) were the highest, while mathematics proficiency levels (*M* = .345,

SD = .198) were the lowest.

Table 3

Descriptive Statistics on GA Public High Schools

	All GA Public High Schools			Represented in RCUs		
	N	*M*	*SD*	*N*	*M*	*SD*
CCRPI						
Content Mastery	471	.600	.222	398	.639	.185
Readiness	497	.707	.169	398	.726	.112
EOC Proficiency Levels						
English	464	.439	.205	398	.467	.179
Mathematics	459	.318	.210	398	.345	.198
Science	454	.391	.208	398	.421	.188
Social Studies	453	.406	.206	396	.436	.185

Note. CCRPI = college and career ready performance index. EOC = end of course.

Institutional expenditures per FTE. The IPEDS data center provides data for public consumption on information nationwide for postsecondary institutions. One aspect of the data center provides derived variables. The derived variables utilized in the data analysis were the institutional expenditures per FTE for each RCUs within USG. The expenditures data is provided by fiscal year, which spans from July 1 of the prior year to June 30 of the current year. In relation to academic terms, fiscal year (FY) 2019 was for Fall 2018, while FY2020 was for Fall 2019. Table 4 displays the expenditures per FTE by area and institution for each FYs, along with the mean and standard deviation for each variable. The University of West Georgia expended the most for academic support (*M* = 2,594.00, *SD* = 157.6). Valdosta State University expended the most for institutional support (*M* = 2,732.50, *SD* = 26.16). For instruction, the University of West Georgia expended the most (*M* = 7,692.50, *SD* = 340.12). For research, Georgia Southern University expended the most (*M* = 551.50, *SD* = 28.99). Valdosta State University expended the most on student support (*M* = 2,092.50, *SD* = 229.81).

Table 4

Institutional Expenditures per Full-time Equivalency by Institution

Institution Name	FY2019	FY2020	*M*	*SD*
Georgia Southern University				
Academic Support	2,255.00	2,265.00	2,260.00	7.071
All Other	2,179.00	3,124.00	2,651.50	668.216
Institution Support	1,580.00	2,411.00	1,995.50	587.606
Instruction	6,671.00	6,628.00	6,649.50	30.406
Public Service	111.00	114.00	112.50	2.121
Research	572.00	531.00	551.50	28.991
Student Services Support	1,586.00	1,546.00	1,566.00	28.284
Kennesaw State University				
Academic Support	2,588.00	2,311.00	2,449.50	195.869
All Other	2,572.00	3,368.00	2,970.00	562.857
Institution Support	1,878.00	2,375.00	2,126.50	351.432
Instruction	5,449.00	5,844.00	5,646.50	279.307
Public Service	450.00	363.00	406.50	61.518
Research	67.00	79.00	73.00	8.485
Student Services Support	1,559.00	1,283.00	1,421.00	195.162
University of West Georgia				
Academic Support	2,483.00	2,706.00	2,594.50	157.685
All Other	1,467.00	2,710.00	2,088.50	878.934
Institution Support	2,439.00	3,303.00	2,871.00	610.940
Instruction	7,452.00	7,933.00	7,692.50	340.118
Public Service	28.00	33.00	30.50	3.536
Research	187.00	213.00	200.00	18.385
Student Services Support	1,668.00	1,683.00	1,675.50	10.607
Valdosta State University				
Academic Support	1,682.00	1,359.00	1,520.50	228.396
All Other	1,442.00	4,694.00	3,068.00	2,299.511
Institution Support	2,714.00	2,751.00	2,732.50	26.163
Instruction	6,845.00	5,315.00	6,080.00	1,081.873
Public Service	162.00	158.00	160.00	2.828
Research	24.00	23.00	23.50	0.707
Student Services Support	2,255.00	1,930.00	2,092.50	229.810

Demographic characteristics. Demographics of the students enrolled in one of the four RCUs who graduated from a GA public HS in 2018 or 2019 are displayed in Table 5. The table displays the information for students in each cohort in addition to the total. For the Fall 2018 cohort, 10,441 students fit the study's profile, and 11,356 students fit the profile for the Fall 2019 cohort. In total, the population consisted of 21,797 students. Overall, females represented 55.5% of the population, while their male counterparts represented 44.5%. Across the four RCUs for the two cohorts, 11,306 (51.9%) students were identified as White, 6,130 (28.1%) students were identified as Black or African American, 2,313 (10.6%) were identified as Hispanic, and the remaining 2,048 (9.4%) students were underrepresented groups. For the two cohorts, 13.1% of the students were identified as first generational students, while the remaining 86.9% students were non-first generational students. A total of 11,606 (53.2%) students graduated from a GA public HS classified in a suburban location, as identified by the National Center for Education Statistics. Approximately, 27.0% or 5,887 students graduated from a rural public HS. In comparison, 11.3% of the students graduated from a public HS in a city, and 8.5% of the students from a public HS in a town.

Table 5

Demographic Characteristics of Students in Fall 2018 and Fall 2019 Cohorts

	Fall 2018		Fall 2019		Total	
	N	%	N	%	N	%
Gender						
Female	5,752	55.1%	6,342	55.8%	12,094	55.5%
Male	4,689	44.9%	5,014	44.2%	9,703	44.5%
Race/Ethnicity						
Black or African American	2,934	28.1%	3,196	30.6%	6,130	28.1%
Hispanic	1,014	9.7%	1,299	12.4%	2,313	10.6%
Other	936	9.0%	1,112	10.7%	2,048	9.4%
White	5,557	53.2%	5,749	55.1%	11,306	51.9%
First Generation Status						
No	9,106	87.2%	9,831	86.6%	18,937	86.9%
Yes	1,335	12.8%	1,525	13.4%	2,860	13.1%
High School Locale						
City	1,168	11.2%	1,293	11.4%	2,461	11.3%
Rural	2,926	28.0%	2,961	26.1%	5,887	27.0%
Suburb	5,448	52.2%	6,158	54.2%	11,606	53.2%
Town	899	8.6%	944	8.3%	1,843	8.5%
Received Zell Miller						
No	9,367	89.7%	10,113	89.1%	19,480	89.4%
Yes	1,074	10.3%	1,243	10.9%	2,317	10.6%

Note. Interdisciplinary studies also contain students who were undeclared majors as their CIP codes are classified in the interdisciplinary studies classification.

Table 5 (continued)

Demographic Characteristics of Students in Fall 2018 and Fall 2019 Cohorts

	Fall 2018		Fall 2019		Total	
	N	%	N	%	N	%
One-year Retention Status						
Retained	8,052	77.1%	8,860	78.0%	16,912	77.6%
Not Retained	2,389	22.9%	2,496	22.0%	4,885	22.4%
Major Grouping						
Business	1,427	13.7%	1,718	15.1%	3,145	14.4%
Education	511	4.9%	550	4.8%	1,061	4.9%
Fine Arts	835	8.0%	969	8.5%	1,804	8.3%
Interdisciplinary Studies	1,470	14.1%	1,068	9.4%	2,538	11.6%
Healthcare	1,547	14.8%	1,846	16.3%	3,393	15.6%
Human Services	167	1.6%	153	1.3%	320	1.5%
Humanities	218	2.1%	219	1.9%	437	2.0%
Social Sciences	922	8.8%	1,182	10.4%	2,104	9.7%
STEM	3,344	32.0%	3,651	32.2%	6,995	32.1%

Note. Interdisciplinary studies also contain students who were undeclared majors as their CIP codes are classified in the interdisciplinary studies classification.

Around 10.6% of students who graduated from a GA public HS were awarded the second tier of the HOPE scholarship, known as the Zell Miller Scholarship. From the two cohorts, 77.6% of students were retained for the next fall semester, while 22.4% of students did not retain. The STEM majors, at 32.1% of students, were the most popular major grouping from both cohorts, with healthcare (15.6% of students) and business (14.4% of students) being the second and third popular major grouping. Human services (1.5% of students) and humanities (2.0% of students) were the two least popular majors.

Figure 3 illustrates the characteristics of the students' reported satisfaction of the college preparatory curriculum (CPC) for English, foreign language, mathematics, science, and social science. Of the CPC variables, a total of 20,355 (93.4%) students satisfied the English requirements, 20,939 (96.1%) students satisfied the foreign language requirements, 20,385 (93.5%) students satisfied the mathematics requirements, 20,595 (94.5%) satisfied the science requirements, and 20,887 (95.8%) students satisfied the social science requirements. A small percentage of students were reported with a deficiency in English (3.3%), foreign language (0.6%), mathematics (3.1%), science (2.2%), and social studies (0.8%).

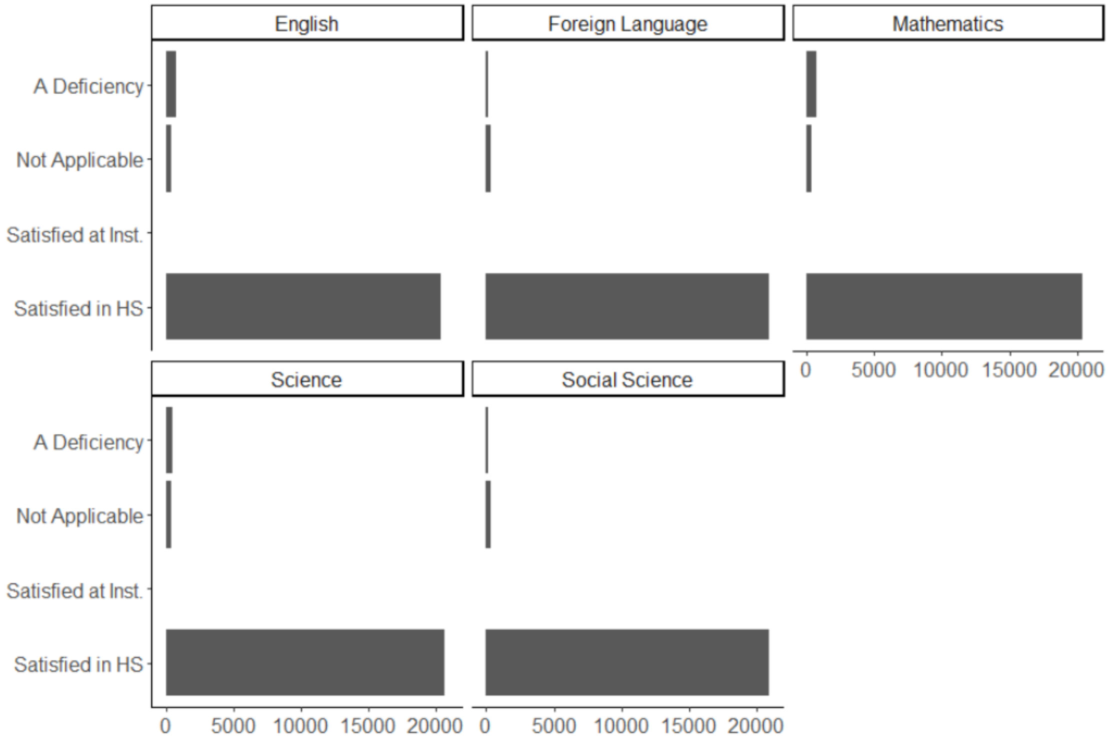


Figure 3. Bar chart of college preparatory curriculum satisfaction. This bar chart displays the number of students who did and did not satisfy the admissions requirements. Satisfied at Inst. = satisfied at institution. Satisfied in HS = satisfied in high school.

Descriptive statistics. Table 6 displays the initial descriptive statistics of the continuous independent and the two dependent GPA variables for Fall 2018 and Fall 2019 cohorts who graduated from a GA public HS. The table displays the number of observations, mean, standard deviation, median, and range. The initial skewness and kurtosis values are also displayed in the table. For the pre-college characteristics, the mean HS GPA was 3.34 (*SD* = 0.42), and the mean admissions test score was 1126.75 (*SD* = 116.54) for the two cohorts. On average, students earned 9.35 AP hours (*SD* =

7.89), 5.70 CLEP hours (*SD* = 4.33), 9.08 IB hours (*SD* = 7.51), and 4.38 other advanced standing hours (*SD* = 2.86).

Table 6

Descriptive Statistics of Continuous Independent and Dependent Variables

Variable	*N*	*M*	*SD*	*Mdn*
GPA Dependent Variables				
First-Fall GPA	21,782	2.80	1.02	3.05
First-Year GPA	21,709	2.82	0.97	3.06
Pre-college Characteristics				
HS GPA	21,678	3.34	0.42	3.36
Test Scores	21,761	1,126.75	116.54	1,110.00
AP Hours	4,843	9.35	7.89	6.00
CLEP Hours	63	5.70	4.33	4.00
IB Hours	212	9.08	7.51	6.00
Other Hours	52	4.38	2.86	3.00
Financial Situations				
EFC	20,727	20,724.57	47,643.25	7,451.00
Federal Subsidized Loans	9,048	1,681.39	278.07	1,732.00
Federal Unsubsidized Loans	10,251	1,644.55	896.64	990.00
HOPE	16,612	2,367.02	235.11	2,430.00
Other Loans	1,162	4,797.76	2,523.51	3,950.00
PELL Grant	9,060	2,520.88	806.01	3,048.00

Note. HS = high school. GPA = grade point average. EFC = expected family contribution.

Table 6 (continued)

Descriptive Statistics of Continuous Independent and Dependent Variables

Variable	Min	Max	Skew	Kurtosis
GPA Dependent Variables				
First-fall GPA	0.00	4.00	-1.07	0.54
First-year GPA	0.00	4.00	-1.13	0.76
Pre-college Characteristics				
HS GPA	1.41	4.00	-0.34	2.40
Test Scores	630.00	1,590.00	0.53	3.08
AP Hours	0.00	57.00	1.85	7.25
CLEP Hours	0.00	24.00	1.58	6.46
IB Hours	0.00	34.00	1.07	3.56
Other Hours	0.00	12.00	1.51	4.53
Financial Situations				
EFC	0.00	999,999.00	10.20	166.09
Federal Subsidized Loans	3.00	3,464.00	-2.66	18.98
Federal Unsubsidized Loans	28.00	5,938.00	0.80	2.41
HOPE Scholarship	86.84	2,781.00	-2.07	15.06
Other Loans	197.00	14,352.00	0.69	2.99
PELL Grant	0.00	3,098.00	-1.28	3.33

Note. HS = high school. GPA = grade point average. EFC = expected family contribution.

For the two cohorts, the mean HOPE scholarship awarded was $2,367.02 (*SD* = 235.11), and the mean PELL Grant awarded wase $2,520.88 (*SD* = 806.01). The average amount of loans taken out was $1,681.39 (*SD* = 278.07) for federal subsidized loans, $1,644.55 (*SD* = 896.64) for federal unsubsidized loans, and $4,797.76 (*SD* = 2,523.51) for all other loans. For the calculated expected family contribution to the student's education, the mean amount was $20,724.57 (*SD* = 47,643.25).

Data Splitting and Imbalance

Prior to the development of the predictive algorithms, the data set was partitioned into two data sets. The initial set, known as the training data set, encompassed 60% of the data, while the remaining 40% constituted the testing data set. Both sets underwent a 10-fold cross-validation process, facilitated by the set.seed() function for the replication of the results. The cross-validation data sets were employed to evaluate the predictive performance of the algorithms. The optimal predictive algorithms were derived using the training data sets. After tuning the algorithms, the predictive performance was assessed in both cross-validation of both the training and testing data sets. The mean accuracy metric was analyzed to measure predictive power, enabling the selection of the most effective model.

For the one-year retention dependent variable, additional sampling techniques were utilized. Downsampling and upsampling techniques were utilized to address data imbalances. These methods were attempts in improving the model's predictive power, given the disparity in data distribution between the majority and minority classes. Downsampling involved reducing the instances of the majority class, whereas upsampling increased the occurrences of the minority class. These approaches were aimed in attempts to counteract biases of the models from defaulting to the majority class, to ensure a more balanced and accurate prediction.

Preliminary Considerations and Assumptions

Preliminary considerations and assumptions, which may vary across different models, were rigorously examined before the algorithm development process. These considerations and assumptions were reviewed on the training data set to prevent any data leakage from the testing data set. Violations of these assumptions were identified and rectified during the data cleanup process. The training data set consisted of 13,078 observations. Addressing any consideration and assumption violations is imperative for producing meaningful research results, as it allows for the drawing of valid conclusions about the real-world phenomena (Field et al., 2012; Garson, 2012). To ensure consistency in handling and rectifying data consideration and assumption violations, the recipe() function provided the systematic means to apply the data cleanup process to both the training and testing data sets.

Considerations. Two preliminary considerations were reviewed. The first consideration was the missing data, and the second was the review for outliers.

Missing data. Before the HS curriculum data was joined to the data set collected from USG, the missing data for the HS were reviewed on the CCRPI and EOC variables used in the study. Figure 4 illustrates the number of missing observations within the GA public HS represented in the four RCUs. Specifically, for the CCRPI variables, both content mastery and readiness variables were each missing seven observations. In the case of EOC subjects' proficiency levels, English had eight, mathematics had nine, science had seven, and social studies had nine missing observations.

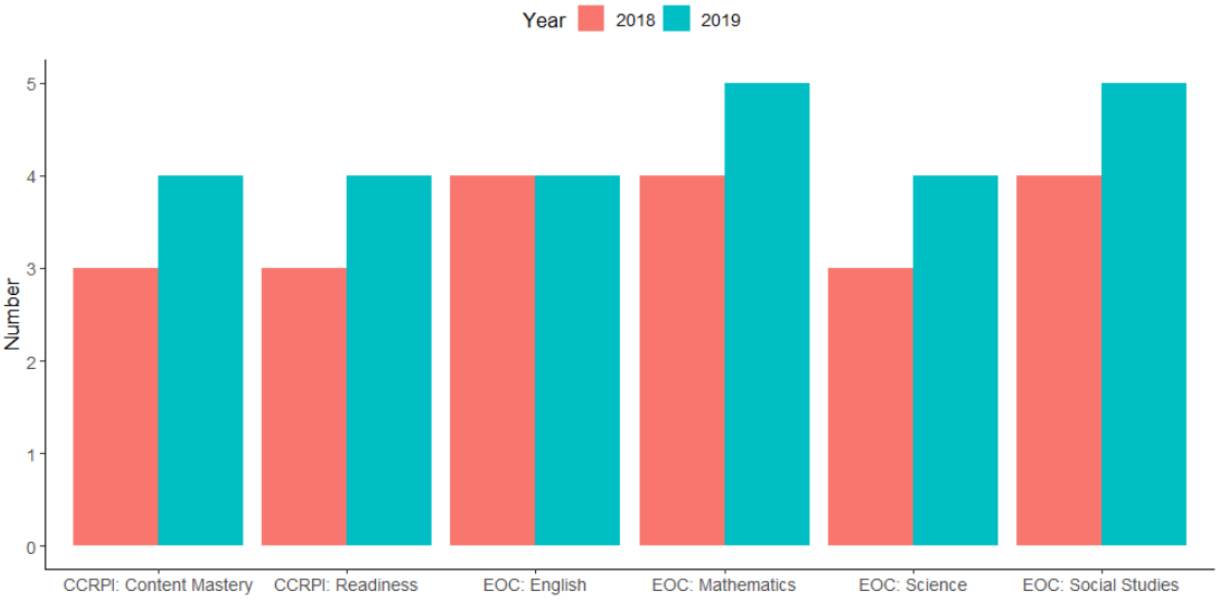


Figure 4. HS curriculum quality variables with missing observations. The bar chart illustrates the number of missing observations within the HS curriculum variables. CCRPI = college and career ready performance index. EOC = end-of-course.

Missing HS curriculum observations were imputed before being integrated into the complete data set. This approach was selected to prevent bias or distortion within the imputation process, ensuring no duplication of the HS influencing the imputation method. In Table 7, the impact of three imputation methods—zero, mean, and median—on HS curriculum data summary statistics is presented. Among these methods, zero imputation had the most significant effect on the distribution. Mean imputation did not alter the minimum, mean, or maximum values; however, it influenced the first quartile, median, and third quartile. Although minor noticeable differences were observed in the first and third quartiles, as well as the mean, for the median imputation, its minimal impact on the distribution led to its utilization to impute the missing data.

Table 7

Summary Statistics of HS Curriculum Data with Zero, Mean, and Median Imputations

	Min	Q1	*Mdn*	*M*	Q3	Max
Before Imputation						
Content Mastery	0.073	0.512	0.633	0.639	0.770	1.000
Readiness	0.293	0.664	0.732	0.726	0.800	0.994
English	0.000	0.344	0.447	0.467	0.578	0.995
Mathematics	0.000	0.199	0.316	0.345	0.463	0.997
Science	0.000	0.285	0.406	0.421	0.543	1.000
Social Studies	0.011	0.308	0.421	0.436	0.571	0.989
Zero Imputation						
Content Mastery	0.000	0.508	0.632	0.633	0.765	1.000
Readiness	0.000	0.660	0.731	0.720	0.799	0.994
English	0.000	0.338	0.443	0.463	0.578	0.995
Mathematics	0.000	0.197	0.314	0.341	0.461	0.997
Science	0.000	0.283	0.402	0.417	0.539	1.000
Social Studies	0.000	0.306	0.420	0.431	0.571	0.989
Mean Imputation						
Content Mastery	0.073	0.513	0.635	0.639	0.765	1.000
Readiness	0.293	0.664	0.731	0.726	0.799	0.994
English	0.000	0.348	0.450	0.467	0.578	0.995
Mathematics	0.000	0.200	0.320	0.345	0.461	0.997
Science	0.000	0.288	0.407	0.421	0.539	1.000
Social Studies	0.011	0.309	0.425	0.436	0.571	0.989
Median Imputation						
Content Mastery	0.073	0.513	0.633	0.639	0.765	1.000
Readiness	0.293	0.664	0.732	0.726	0.799	0.994
English	0.000	0.348	0.447	0.467	0.579	0.995
Mathematics	0.000	0.200	0.316	0.344	0.461	0.997
Science	0.000	0.286	0.406	0.421	0.539	1.000
Social Studies	0.011	0.309	0.421	0.436	0.571	0.989

After combining the HS curriculum data into the data set, the remaining missing observations were analyzed. Figure 5 provides a visual representation of the number of missing data points by variable. Among the dependent variables, seven observations were missing for first-fall GPA, and 62 observations were missing for first-year GPA. Since these are dependent variables, these records would be removed from both the training and testing data sets. Additionally, the advanced standing hours variables exhibited a substantial number of missing observations. Specifically, AP hours had 10,188, CLEP hours had 13,034, IB hours had 12,953, and other hours had 13,054 missing observations. Given these variables represent the number of earned advanced standing hours before matriculating into the institution, the selection of zero imputation was the most suitable approach. This decision was influenced by the data collection process employed by USG, which only collects records with specific identifying codes for each earned advanced hour (USG, 2023).

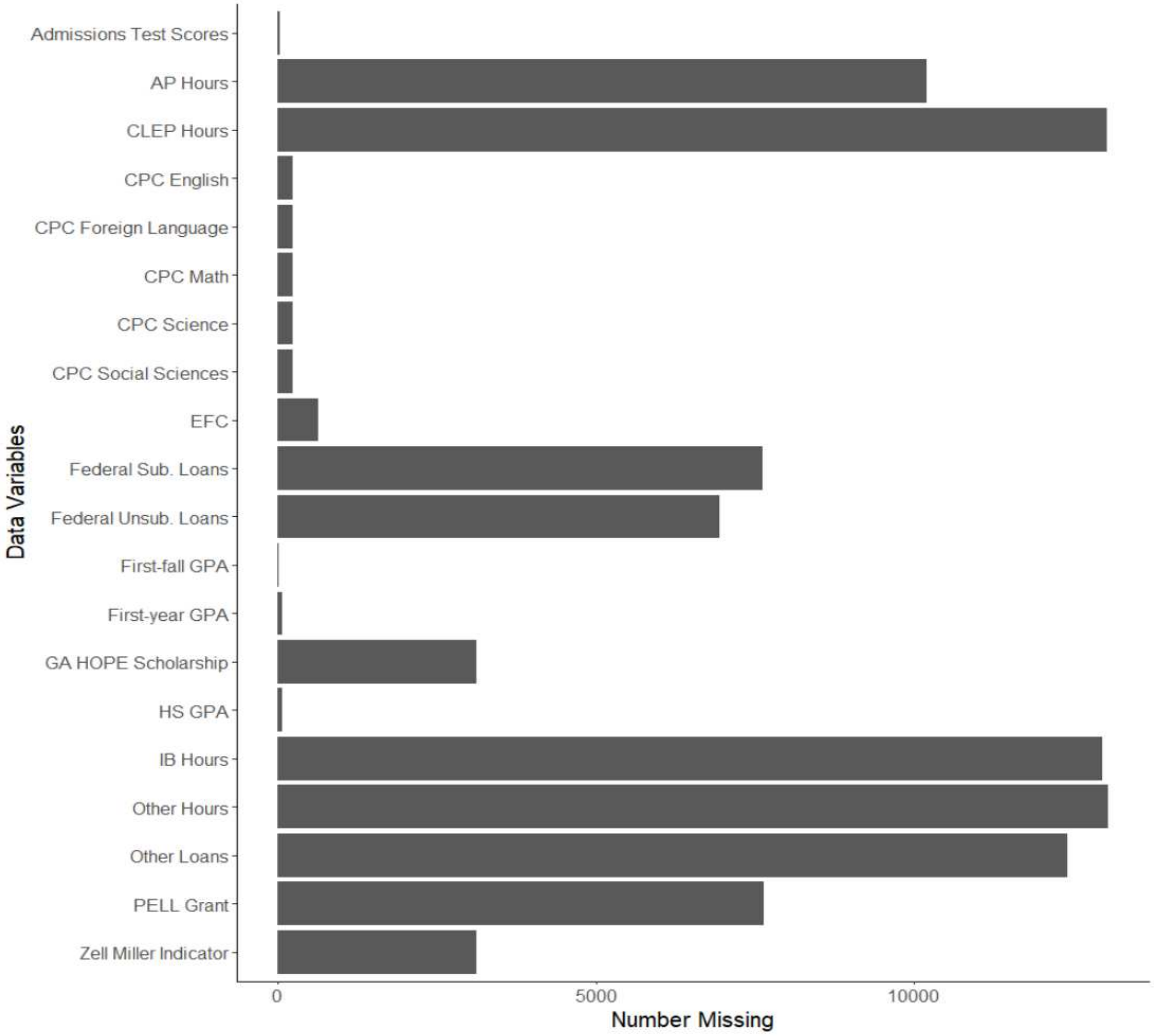


Figure 5. Distribution of missing variables within training data set. The bar chart illustrates the number of missing observations within the training data set. CCRPI = college and career ready performance index. EOC = end-of-course. EFC = expected family contribution. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. GPA = grade point average. HS = high school.

Each of the five CPC variables exhibited 231 missing observations. These missing data points were imputed with a new variable to indicate the observation was unknown. The Zell Miller indicator had 3,112 observations missing an indicator. These observations were coded to indicate the students did not receive the second tier of the GA HOPE scholarship as the data collection submitted to USG was like the advanced standing hours (USG, 2023). The GA HOPE scholarship (3,112 observations), PELL Grant (7,636 observations), federal subsidized loans (7,622 observations), federal unsubsidized loans (6,937 observations), and other loans (12,398 observations) variables exhibited missing observations. The financial aid missing observations were imputed with a zero as it indicated the student did not receive any dollars of these types of aid based on the USG data collection process (USG, 2021e).

The admissions test scores (24 observations), HS GPA (74 observations), and expected family contribution (639 observations) underwent an examination of data imputation methods for the missing observations, even though the USG data collection process was like the prior variables. These three variables were subject to three different imputation methods: mean, median, and k-nearest neighbor (KNN). The KNN imputation neighbors was set 10. Table 8 displays the summary of the statistics derived from the imputation methods for admissions test scores, HS GPA, and expected family contributions. Comparing the impact of the three imputation methods on admissions test scores and HS GPA, no significant alterations to the variable distribution were detectable. The unnoticeable impact was due to the small number of missing observations for the variables. Conversely, the three imputation methods exhibited a noticeable impact on the EFC's distribution. To determine the most suitable imputation method, skewness and kurtosis values for each approach were examined. The KNN method exhibited skewness (10.09) and kurtosis (171.47) values closely resembled the distribution before any imputation (skewness = 9.99; kurtosis = 166.20) and was selected for imputation for the variable. For consistency, the KNN imputation method was also selected in imputing missing values for HS GPA and admissions test scores variables.

Table 8

Summary Statistics of HS Curriculum Data with Mean, Median, and KNN Imputations

	Imputation Methods			
	Before	Mean	Median	KNN
Admissions Test Scores				
Min	630	630	630	630
Q1	1040	1040	1040	1040
Mdn	1110	1110	1110	1110
M	1127	1127	1127	1127
Q3	1210	1210	1210	1210
Max	1550	1550	1550	1550
Skewness	0.52	0.52	0.53	0.53
Kurtosis	0.06	0.07	0.07	0.07
HS GPA				
Min	1.43	1.43	1.43	1.43
Q1	3.04	3.04	3.04	3.04
Mdn	3.36	3.36	3.36	3.36
M	3.34	3.34	3.34	3.34
Q3	3.69	3.69	3.69	3.69
Max	4.00	4.00	4.00	4.00
Skewness	-0.32	-0.32	-0.32	-0.32
Kurtosis	-0.65	-0.63	-0.63	-0.64
EFC				
Min	0.00	0.00	0.00	0.00
Q1	69.50	246.20	246.20	246.20
Mdn	7,475.00	8,738.50	7,475.00	8,724.50
M	20,152.40	20,152.40	19,532.90	20,754.90
Q3	24,578.50	23,134.50	23,134.50	26,473.60
Max	999,999.00	999,999.00	999,999.00	999,999.00
Skewness	9.99	10.25	10.23	10.09
Kurtosis	166.20	174.90	174.06	171.47

Note. HS = high school. GPA = grade point average. EFC = expected family contribution. KNN = k-nearest neighbor.

Outliers. Univariate outliers were identified by examining histograms and density curves of the z-scores for each variable. Analyzing the graphs of the z-scores facilitated the easier detection of outliers, especially given the number of observations. Figure 6 illustrates the distribution of z-scores for each continuous variable. Upon examination, noticeable outliers were observed in variables such as English proficiency levels, federal subsidized loans, GA HOPE scholarship dollars, institutional support expenditures, instruction expenditures, mathematics proficiency levels, PELL grant dollars, public service expenditures, research expenditures, science proficiency levels, and social studies proficiency levels. While the conventional thresholds for identifying outliers are typically set at -3 and +3 z-scores, adjustments are allowed based on the specific variable, as suggested by Merler and Vannatta (2002).

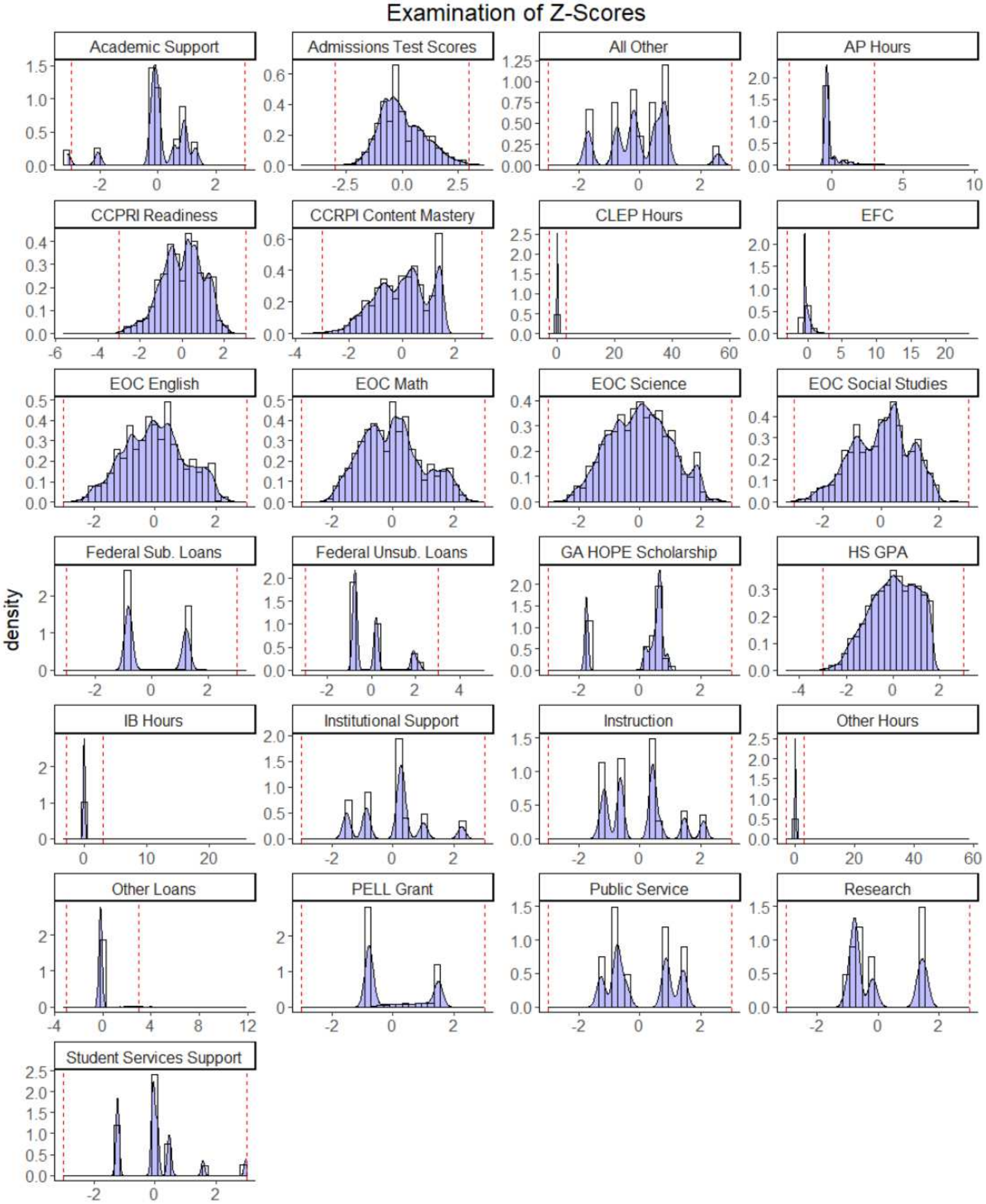


Figure 6. Histograms and density curves of continuous variables' z-scores. These histograms and density curves illustrate the distribution of the z-scores for outlier detection. The graphs also include the conventional outlier thresholds of -3 and +3. CCRPI = college and career ready performance index. EOC = end-of-course. EFC = expected family contribution. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. GPA = grade point average. HS = high school.

Grubbs' statistical test for outliers was conducted to identify each variable's outliers, as detailed in Table 9. Among the pre-college characteristics, the four advanced standing hours were found to be significant (AP hours, $G = 9.846$, $p < .001$; CLEP hours, $G = 58.856$, $p < .001$; IB hours, $G = 25.537$, $p < .001$; and other hours, $G = 56.771$, $p < .001$). Upon reviewing the flagged values for outliers, values were determined to be acceptable, as students had earned the hours before matriculating into the institution. Two financial situation variables were found to be significant, indicating potential outliers. Expected family contribution was significant, $G = 23.150$, $p < .001$, with values of 0 and 999,999 flagged as outliers. The values were deemed acceptable as they fell within the expected ranges resulting from the calculations (The Scholarship System, 2023). Similarly, the other loans variable was found to be significant, $G = 11.923$, $p < .001$, with 0 and 14,352 flagged as outliers. These values were considered acceptable, reflecting scenarios where some students do not need to take out any dollars in other loans, while others may need $14,352 to cover the full cost of attending an institution. No institutional expenditure variables were identified as having outliers by Grubb's test.

Table 9

Grubb's Test for Univariate Outliers

					Outliers	
	G	*U*	*p*		Value 1	Value 2
Pre-college Characteristics						
Admissions Test Scores	7.898	.998	1.000		630	1550
HS GPA	6.113	.998	1.000		1.43	4.00
AP Hours	9.846	.993	< .001	***	0	54
CLEP Hours	58.856	.736	< .001	***	0	24
IB Hours	25.537	.950	< .001	***	0	29
Other Hours	56.771	.754	< .001	***	0	12
CCRPI Content Mastery	5.129	.999	1.000		0.129	1.000
CCRPI Readiness	7.751	.997	1.000		0.293	0.994
EOC English	5.333	.999	1.000		0.084	0.995
EOC Math	4.904	.999	1.000		0.004	0.997
EOC Science	5.571	.999	1.000		0.000	1.000
EOC Social Studies	5.601	.999	1.000		0.020	0.989
Financial Situations						
EFC	23.150	.961	< .001	***	0	999,999
GA HOPE Scholarship	2.701	1.000	1.000		0	2,781
PELL Grant	2.300	1.000	1.000		0	3,098
Federal Sub. Loans	4.085	.999	1.000		0	3,464
Federal Unsub. Loans	5.772	.998	1.000		0	5,938
Other Loans	11.923	.990	< .001	***	0	14,352
Institutional Expenditures						
Academic Support	4.434	.999	1.000		1,359	2,706
All Other	4.275	.999	1.000		1,442	4,694
Institutional Support	3.770	.999	1.000		1,580	3,303
Instruction	3.412	1.000	1.000		5,315	7,933
Public Service	2.701	1.000	1.000		28	450
Research	2.530	1.000	1.000		23	572
Student Support	4.201	1.000	1.000		1,283	2,255

Note. *** p < .001. ** p < .01. * p < .05. CCRPI = college and career ready performance index. EOC = end-of-course. EFC = expected family contribution. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. GPA = grade point average. HS = high school.

To identify multivariate outliers, a Mahalanobis test was conducted on the continuous variables. Figure 5 depicts the distribution of row level p-values, revealing 4,326 (33.1%) observations were identified as having multivariate outliers. Upon careful examination, the observations were determined to acceptable value ranges. Consequently, no elimination or capping procedures were deemed necessary.

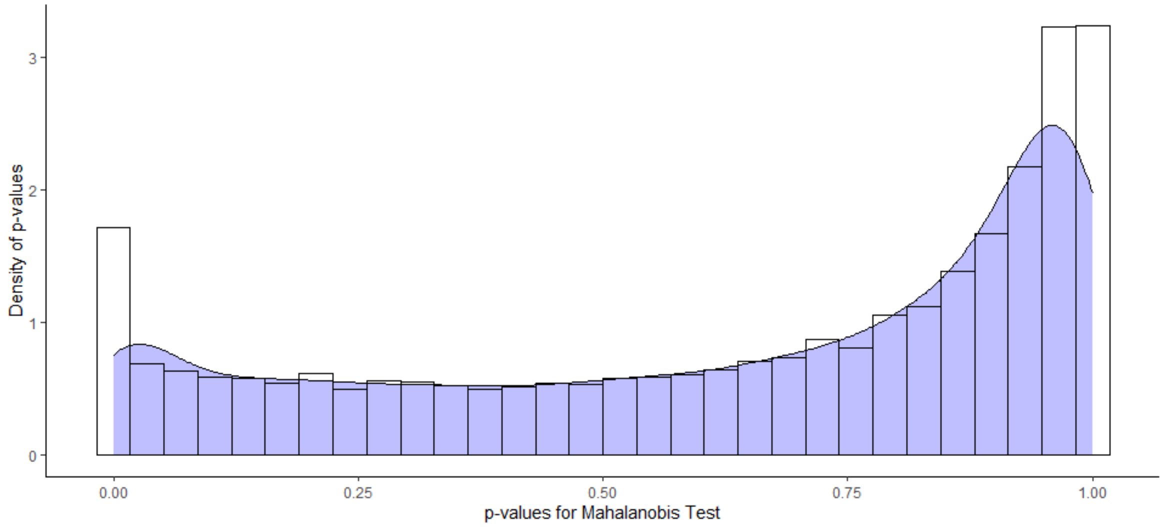


Figure 7. Histogram and density curve of the Mahalanobis test of the continuous independent variables observation's p-values.

Assumptions. Four fundamental assumptions were reviewed to ensure the production meaningful research outcomes to establish reliability and validity of the findings (Field et al., 2012; Garson, 2012). Like the considerations, the assumptions were examined using the training data set to prevent any data leakage. The examined assumptions encompassed observation independence, linearity and collinearity, univariate and multivariate normality, and homogeneity of variance.

Observation independence. Observation independence implies no duplicated records existed within the data file. To validate this, unique identifiers for each institution and cohort year were reviewed. The analysis revealed no student records were duplicated. This finding aligns with the definition of FTFTF, where a student is exclusively classified as a first-time, full-time freshman at the initial matriculation into any institution. Consequently, students were not duplicated across institutions within the file, ensuring the integrity of the data set.

Linearity. A correlation analysis was conducted on the training data set to examine the linear relationship between the independent and three dependent variables (See Appendix F). For the retention dependent variable, 18 independent variables were significant. While the relationship was very weak, 12 variables have a negative relationship and 6 have a positive relationship (GA HOPE Scholarship, *r(*13,078) = -.183, *p* < .001; HS GPA, *r(*13,078) = -.165, *p* < .001; gender, *r(*13,078) = .080, *p* < .001; instruction expenditures, *r(*13,078) = .057, *p* < .001; institutional support expenditures, *r(*13,078) = .054, *p* < .001; student services expenditures, *r(*13,078) = .054, *p* < .001; public service expenditures, *r(*13,078) = -.052, *p* < .001; federal subsidized loans, *r(*13,078) = .045, *p* < .001; Zell Miller, *r(*13,078) = -.045, *p* < .001; content mastery, *r(*13,078) = -.040, *p* = .001; PELL grant, *r(*13,078) = .037, *p* = .006; science proficiency levels, *r(*13,078) = -.037, *p* = .007; all other expenditures, *r(*13,078) = -.036, *p* =.010; readiness, *r(*13,078) = -.035, *p* =.016; social studies proficiency levels, *r(*13,078) = -.034, *p* = .025; English proficiency levels, *r(*13,078) = -.033, *p* = .039; and admissions test scores, *r(*13,078) = -.031, *p* < .001).

The first-fall GPA dependent variable exhibited 29 significant relationships with the independent variables. In the student characteristics, gender, *r(*13,078) = -.149, *p* < .001, exhibited a slight negative relationship, race and ethnicity, *r(*13,078) = -.019, *p* = 0.32, exhibited a very slight negative relationship, and first generation status, *r(*13,078) = -.032, *p* < .001, exhibited a very slight negative relationship with the first-fall GPA. The students' HS GPA, *r(*13,078) = .488, *p* < .001, exhibited a moderate positive relationship on the first-fall GPA, while the admissions test scores, *r(*13,078) = .257, *p* < .001, a small positive relationship. The number of advanced standing AP hours, *r(*13,078) = .219, *p* <

.001, had a small positive relationship with the earned GPA, while the IB hours (*r(*13,078) = .039, *p* < .001), and CLEP hours (*r(*13,078) = .033, *p* < .001), had a very slight positive relationship. Of the CPC, social science, *r(*13,078) = .020, *p* = .025, exhibited a very slight positive relationship with the fall GPA. Of the graduating HS CCRPI scores, both the content mastery (*r(*13,078) = .114, *p* < .001), and the readiness (*r(*13,078) = .096, *p* < .001), scores exhibited a slight positive relationship on the fall GPA. The four EOC subject proficiency levels had slight relationships with the fall GPA. All proficiency levels but science had a positive relationship (social studies proficiency levels, *r(*13,078) = .113, *p* < .001; mathematics proficiency levels, *r(*13,078) = .109, *p* < .001; science proficiency levels, *r(*13,078) = -.102, *p* < .001; and English proficiency levels, *r(*13,078) = .100, *p* < .001).

Students who were awarded GA HOPE scholarship dollars, *r(*13,078) = .430, *p* < .001, had a moderate positive relationship, and those who earned the Zell Miller, *r(*13,078) = .281, *p* < .001, had a small positive relationship on the first-fall GPA. The other financial situation variables had a slight negative relationship on the first-fall GPA (federal subsidized loans, *r(*13,078) = -.135, *p* < .001; federal unsubsidized loans, *r(*13,078) = -.101, *p* < .001; PELL grant, *r(*13,078) = -.097, *p* < .001; and other loans, *r(*13,078) = -.066, *p* < .001). Academic support expenditures (*r(*13,078) = .061, *p* < .001), all other expenditures (*r(*13,078) = .035, *p* < .001), and public service expenditures (*r(*13,078) = .122, *p* < .001) had slight positive relationships on the first-fall GPA, while institutional support expenditures (*r(*13,078) = -.046, *p* < .001), instruction expenditures (*r(*13,078) = -.102, *p* < .001), research expenditures (*r(*13,078) = -.065, *p* < .001), and student services expenditures ($r(13{,}078) = -.061$, $p < .001$) had slight negative relationships.

For the first-year GPA, only 28 independent variables were found to have significant relationships with the dependent variable. Gender ($r(13{,}078) = -.163$, $p < .001$), race and ethnicity ($r(13{,}078) = -.022$, $p = .011$) and first generation status ($r(13{,}078) = -.039$, $p < .001$) were the only three student characteristics variables with a significant correlational relationship on the first-year GPA. These variables had a slight negative relationship with the dependent variable. Students HS GPA ($r(13{,}078) = .507$, $p < .001$) and admissions test scores ($r(13{,}078) = .254$, $p < .001$) were significant. HS GPA exhibited a solid, moderate positive relationship to the first-year GPA and was the variable with the highest correlational strength on the dependent variable. Admissions test scores also had a low positive relationship with the dependent variable. In examining students who enter in with advanced standing hours, the number of AP hours ($r(13{,}078) = .215$, $p < .001$), CLEP hours ($r(13{,}078) = .033$, $p < .001$), and IB hours ($r(13{,}078) = .040$, $p < .001$), were found to be significant. The number of AP hours exhibited a low positive relationship, while CLEP and IB hours had slight positive relationships. Examining the HS Curriculum scores, both the CCRPI content mastery ($r(13{,}078) = .120$, $p < .001$), and readiness ($r(13{,}078) = .102$, $p < .001$), scores had a slight positive relationship. All four subject areas in the EOC proficiency levels were found to have a slight positive relationship (English proficiency levels, $r(13{,}078) = .110$, $p < .001$; mathematics proficiency levels, $r(13{,}078) = .114$, $p < .001$; social studies proficiency levels, $r(13{,}078) =.114$, $p < .001$; and science proficiency levels, $r(13{,}078) = .108$, $p < .001$).

Of the financial situations, the GA HOPE scholarship exhibited the highest correlation with the first-year GPA, *r(*13,078) = .452, *p* < .001. The relationship is a moderately positive one. The Zell Miller indicator, *r(*13,078) = .285, *p* < .001, was found to have a slight positive significant relationship with the dependent variable. The expected family contribution, *r(*13,078) = .070, *p* < .001, had a slight positive relationship. The remaining financial situation variables exhibited slight negative relationships with the dependent variable (PELL Grant, *r(*13,078) = -.109, *p* < .001; federal subsidized loans, *r(*13,078) = -.148, *p* < .001; federal unsubsidized loans, *r(*13,078) = -.101, *p* < .001; and other loans, *r(*13,078) = -.061, *p* < .001). Of the seven expenditure variables, three exhibited a significant slight positive relationship. The remaining four had slight negative relationships (academic support expenditures, *r(*13,078) = .027, *p* = .002, all other expenditures, *r(*13,078) = .089, *p* < .001; public service expenditures, *r(*13,078) =.106, *p* < .001; institutional support expenditures, *r(*13,078) = -.021, *p* = .019; research expenditures, *r(*13,078) = -.042, *p* < .001; student services expenditures, *r(*13,078) = -.084, *p* < .001; and instruction expenditures, *r(*13,078) = -.105, *p* < .001).

In a review of the correlation matrix, several independent variables exhibited strong correlations amongst each other. These correlating independent variables were identified as having potential multicollinearity amongst the variables. These three areas were the CPC variables, the CCRPI and EOC variables, and the institutional expenditures variables. Results of the VIF analysis of the independent variables on the three dependent variables was conducted (See Appendix G). None of the student characteristics, financial situations, and major group variables were above the VIF thresholds of five or ten. In the pre-college characteristics, CPC English (VIF = 44.737), CPC foreign language (VIF =24.689), CPC mathematics (VIF = 41.182), CPC science (VIF = 20.702), CPC social sciences (VIF = 28.274), content mastery (VIF = 33.242), English proficiency levels (VIF = 13.956), mathematics proficiency levels (VIF = 9.207), science proficiency levels (VIF = 8.408) and social studies proficiency levels (VIF = 9.648) were above the thresholds. Of the institutional expenditures, all other expenditures (VIF = 19.205), institutional support (VIF = 20.543), instruction (VIF = 65.942), public service (VIF = 28.914), research (VIF = 7.9934), and student support (VIF = 5.203) were above the thresholds. These 16 independent variables exhibited multicollinearity within the training data set.

With the variables identified exhibiting multicollinearity in the data set, methods were derived to eliminate or lessen the effect. For the five CPC variables, the observations were recoded to zero for unsatisfied and one for satisfied, which the variables were added together to produce a single variable with values ranging from zero to five. The CCRPI content mastery and readiness scores were averaged together. Since EOC subject areas are components of the CCRPI calculations, the difference between the EOC four subject areas proficiency levels and the content mastery and readiness mean were calculated. Academic and institutional support were added together for the institutional expenditures, and likewise for public service and research factors.

After data manipulation to remove multicollinearity, the Pearson's correlation analysis was conducted on the revised data set (See Appendix H). For the one-year retention variable, 15 variables were found to be significant with 7 exhibited a weak positive correlation and 8 exhibited a weak negative correlation (GA HOPE Scholarship, *r(*13,078) = -.183, *p* < .001; HS GPA, *r(*13,078) = -.165, *p* < .001; gender, *r(*13,078) = .080, *p* < .001; public service and research expenditures, *r(*13,078) = -.071, *p* < .001; instruction expenditures, *r(*13,078) = .057, *p* < .001; student services expenditures, *r(*13,078) = .054, *p* < .001; expected family contribution, *r(*13,078) = -.043, *p* < .001; content mastery and readiness mean, *r(*13,078) = -.040, *p* < .001; academic and institutional support expenditures, *r(*13,078) = .039, *p* < .001; federal subsidized loans, *r(*13,078) = .045, *p* < .001; Zell Mill indicator, *r(*13,078) = -.045, *p* < .001; admissions test scores, *r(*13,078) = -.031, *p* < .001; PELL grant, *r(*13,078) = .037, *p* = .005; all other expenditures, *r(*13,078) = -.036, *p* = .008; and science proficiency levels difference from content mastery and readiness mean, *r(*13,078) = -.020, *p* = .019).

A total of 24 independent variables were found to have a significant correlation with the first-fall GPA dependent variable. HS GPA (*r(*13,078) = .488, *p* < .001) and GA HOPE scholarship (*r(*13,078) = .430, *p* < .001) were both weak positive correlations and were the two highest independent variables with significant correlations to the first-fall GPA. The remaining significant independent variables exhibited a significant but barely noticeable correlation to the first-fall GPA (Zell Miller indicator, *r(*13,078) = .281, *p* < .001; admissions test scores, *r(*13,078) = .257, *p* < .001; AP hours, *r(*13,078) = .219, *p* < .001; gender, *r(*13,078) = -.149, *p* < .001; federal subsidized loans, *r(*13,078) = -.135, *p* < .001; content mastery and readiness mean, *r(*13,078) = .111, *p* < .001; instruction expenditures, *r(*13,078) = -.102, *p* < .001; federal unsubsidized loans, *r(*13,078) = -.101, *p* < .001; PELL grant, *r(*13,078) = -.097, *p* < .001; math proficiency levels difference from the content mastery and readiness mean, *r(*13,078) = .083, *p* < .001; social studies proficiency levels difference from the content mastery and readiness mean, *r(*13,078) =

.074, *p* < .001; other loans, *r(*13,078) = -.066, *p* < .001; student services support expenditures, *r(*13,078) = -.061, *p* < .001; expected family contribution, *r(*13,078) = .055, *p* < .001; science proficiency levels difference from the content mastery and readiness mean, *r(*13,078) = .052, *p* < .001; English proficiency levels difference from the content mastery and readiness mean, *r(*13,078) = .047, *p* < .001; IB hours, *r(*13,078) = .039, *p* < .001; all other expenditures, *r(*13,078) = .035, *p* < .001; CLEP hours, *r(*13,078) = .033, *p* < .001; first generation status, *r(*13,078) = -.032, *p* < .001; public service and research expenditures, *r(*13,078) = .029, *p* < .001, and race and ethnicity, *r(*13,078) = -.019, *p* = .032).

The first-year GPA dependent variable had 24 significant correlations with 15 positive correlations and nine negative correlations. The HS GPA (*r(*13,078) = .507, *p* < .001) variable exhibited a moderate positive correlation and was the strongest independent variable. The second strongest variable was the GA HOPE Scholarship (*r(*13,078) = .452, *p* < .001). The remaining 21 variables were found to be significant, but exhibited weak correlations to the dependent variable (Zell Miller indicator, *r*(13,078) = .285, *p* < .001; admissions test scores, *r*(13,078) = .254, *p* < .001; A*P* hours, *r*(13,078) = .215, *p* < .001; gender, *r*(13,078) = -.163, *p* < .001; federal subsidized loans, *r*(13,078) = -.148, *p* < .001; content mastery and readiness mean, *r*(13,078) = .117, *p* < .001; PELL grant, *r*(13,078) = -.109, *p* < .001; instruction expenditures, *r*(13,078) = -.105, *p* < .001; federal unsubsidized loans, *r*(13,078) = -.101, *p* < .001; all other expenditures, *r*(13,078) = .089, *p* < .001; math proficiency levels difference from the content mastery and readiness mean, *r*(13,078) = .085, *p* < .001; student service support expenditures, *r*(13,078) = -.084, *p* < .001; expected family contribution, *r*(13,078) = .07, *p* < .001; social studies proficiency levels difference from the content mastery and readiness mean, *r*(13,078) = .067, *p* < .001; English proficiency levels difference from the content mastery and readiness mean, *r*(13,078) = .062, *p* < .001; other loans, *r*(13,078) = -.061, *p* < .001; science proficiency levels difference from the content mastery and readiness mean, *r*(13,078) = .057, *p* < .001; public service and research expenditures, *r*(13,078) = .044, *p* < .001; IB hours, *r*(13,078) = .04, *p* < .001; first generation status, *r*(13,078) = -.039, *p* < .001; and CLEP hours, *r*(13,078) = .033, *p* < .001; and race and ethnicity, *r(*13,078) = -.022, *p* = .011). Additionally, the VIF analysis for multicollinearity after data manipulation resulted in all independent variables exhibiting VIF values below five (See Appendix I).

Normality. After addressing multicollinearity through data manipulation, univariate and multivariate normality were assessed using the training data set. The Q-Q plots of continuous independent variables are illustrated in Figure 8. Upon review of the plots, all variables except the admissions test scores appeared to have violated the univariate normality assumption. Very few independent variables followed a straight line except for the tails, indicating deviations from normal distribution.

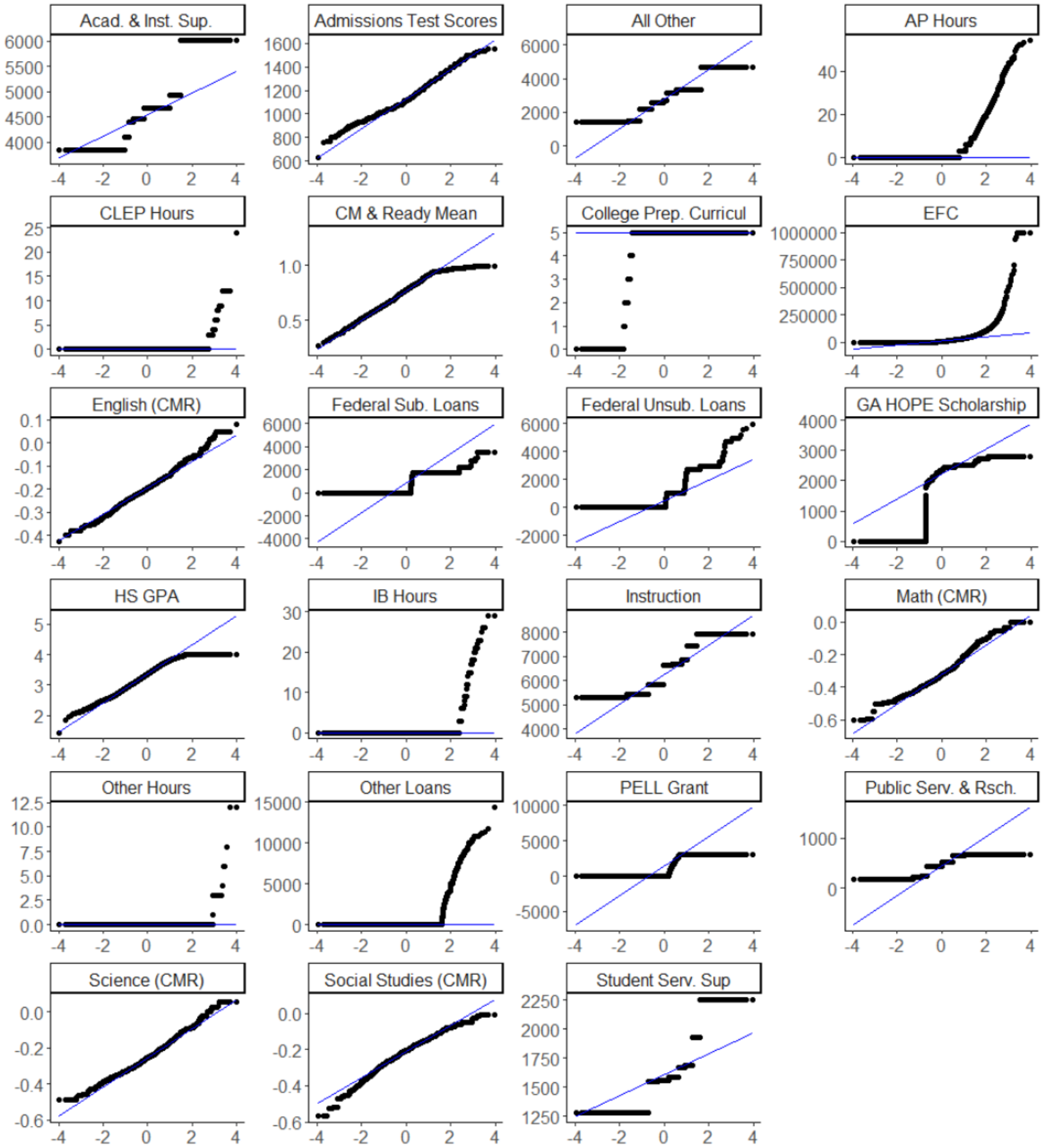


Figure 8. Q-Q plots of the continuous variables. The figure displays the Q-Q plots of the continuous variables to assess normal distribution within the data set. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. Acad. & Inst. Sup. = academic and institutional support expenditures. College Prep. Curricul = college preparatory curriculum. EFC = expected family contribution. Public Serv. & Rsch. = public service and research expenditures. Student Serv. Sup. = student services support expenditures.

Table 10 displays the results of the Shapiro-Wilks and Jarque-Bera statistical tests for univariate normality before any data transformations. Due to the test's limitations, the first 5,000 observations were sampled for the Shapiro-Wilks. The results of these two statistical tests collectively indicated the continuous independent variables did not exhibit a normal distribution. Furthermore, multivariate normality was examined using Mardia's test. The results indicated a lack of multivariate normal distribution, as evidenced by significant skewness (*M*(13,078) = 3,093.726, *p* < .001) and kurtosis (*M*(13,078) = 4,354.193, *p* < .001).

Table 10

Results of Shapiro-Wilks and Jarque-Bera Univariate Normality Test Before Transformations

	Shapiro-Wilks			Jarque-Bera			
	W	*p*		χ^2	*df*	*p*	
Pre-college Characteristics							
HS GPA	0.972	< .001	***	448	2	< .001	***
Admissions Test Scores	0.974	< .001	***	606	2	< .001	***
AP Hours	0.452	< .001	***	211,512	2	< .001	***
CLEP Hours	0.025	< .001	***	946,726,989	2	< .001	***
IB Hours	0.046	< .001	***	47,077,347	2	< .001	***
Other Hours	0.017	< .001	***	1,962,050,112	2	< .001	***
College Prep. Curriculum	0.264	< .001	***	157,390	2	< .001	***
CM & Ready Mean	0.977	< .001	***	335	2	< .001	***
English (CMR)	0.994	< .001	***	141	2	< .001	***
Math (CMR)	0.977	< .001	***	652	2	< .001	***
Science (CMR)	0.988	< .001	***	340	2	< .001	***
Social Studies (CMR)	0.989	< .001	***	724	2	< .001	***
Financial Situations							
EFC	0.436	< .001	***	16,247,837	2	< .001	***
GA HOPE Scholarship	0.672	< .001	***	2,903	2	< .001	***
PELL Grant	0.689	< .001	***	2,035	2	< .001	***
Fed Sub. Loans	0.654	< .001	***	2,003	2	< .001	***
Fed Unsub. Loans	0.733	< .001	***	3,398	2	< .001	***
Other Loans	0.211	< .001	***	730,643	2	< .001	***

Note. *** *p* < .001. ** *p* < .01. * *p* < .05. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. Acad. & Inst. Sup. = academic and institutional support expenditures. EFC = expected family contribution. Public Serv. & Rsch. = public service and research expenditures. Student Serv. Sup. = student services support expenditures.

Table 10 (continued)

Results of Shapiro-Wilks and Jarque-Bera Univariate Normality Test Before Transformations

	Shapiro-Wilks			Jarque-Bera			
	W	*p*		χ^2	*df*	*p*	
Institutional Expenditures							
Acad. & Inst. Support	0.811	< .001	***	4,834	2	< .001	***
All Other	0.912	< .001	***	126	2	< .001	***
Instruction	0.892	< .001	***	809	2	< .001	***
Public & Research	0.875	< .001	***	978	2	< .001	***
Student Support	0.809	< .001	***	5,353	2	< .001	***

Note. *** *p* < .001. ** *p* < .01. * *p* < .05. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. Acad. & Inst. Sup. = academic and institutional support expenditures. EFC = expected family contribution. Public Serv. & Rsch. = public service and research expenditures. Student Serv. Sup. = student services support expenditures.

Various transformations, including Yeo-Johnson, logarithmic, and Box-Cox, were explored to address the skewed nature of certain variables within the data set. Initial examination revealed six variables were negatively skewed. These variables were HS GPA (skewness = -0.32), CPC (skewness = -4.01), content mastery and readiness mean (skewness = -0.30), social studies difference from content mastery and readiness mean (skewness = -0.51), GA HOPE Scholarship (skewness = -1.11), and public and research expenditures (skewness = -0.36). For the logarithmic and Box-Cox transformations to function properly, these negatively skewed variables underwent the step_inverse() function first. Logarithmic transformations were ineffective for variables with zero values in the data set, whereas Box-Cox and Yeo Johnson transformations proved successful. The Yeo-Johnson transformation was selected due to its effectiveness in addressing normality assumption violations in some factors. Table 9 displays the results of the Shapiro-Wilks and Jarque-Bera normality test on the continuous variables after Yeo Johnson transformations and normalization were performed. The Jarque-Bera test resulted in three variables exhibiting normal distribution (admissions test scores, $\chi^2(2) = 4.15$, $p = .125$; social studies proficiency difference from content mastery and readiness mean, χ^2 (2) = 4.22, $p = .121$; and all other expenditures, χ^2 (2) = 5.032, $p = .081$); however, the Shapiro-Wilks test indicated these variables violated normality assumption. All other variables violated the normality assumption after transformations and normalization occurred for both tests.

Table 11

Results of Shapiro-Wilks and Jarque-Bera Test for Normality After Data Transformation and Normalization

	Shapiro-Wilks			Jarque-Bera			
	W	*p*		χ^2	*df*	*p*	
Pre-college Characteristics							
HS GPA	0.976	< .001	***	447.14	2	< .001	***
Admissions Test Scores	0.991	< .001	***	4.15	2	.125	
AP Hours	0.525	< .001	***	4,062.20	2	< .001	***
CLEP Hours	0.025	< .001	***	946,726,989.00	2	< .001	***
IB Hours	0.046	< .001	***	47,077,347.00	2	< .001	***
Other Hours	0.017	< .001	***	1,962,050,112.00	2	< .001	***
College Prep. Curriculum	0.264	< .001	***	15,739.00	2	< .001	***
CM & Ready Mean	0.979	< .001	***	337.42	2	< .001	***
English (CMR)	0.995	< .001	***	54.10	2	< .001	***
Math (CMR)	0.994	< .001	***	51.74	2	< .001	***
Science (CMR)	0.997	< .001	***	8.83	2	.012	*
Social Studies (CMR)	0.998	< .001	***	4.22	2	.121	
Financial Situations							
EFC	0.903	< .001	***	747.31	2	< .001	***
GA HOPE Scholarship	0.605	< .001	***	3,257.30	2	< .001	***
PELL Grant	0.643	< .001	***	2,170.70	2	< .001	***
Fed Sub. Loans	0.632	< .001	***	2,179.60	2	< .001	***
Fed Unsub. Loans	0.675	< .001	***	2,126.70	2	< .001	***
Other Loans	0.228	< .001	***	146,731.00	2	< .001	***
Institutional Expenditures							
Acad. & Inst. Support	0.856	< .001	***	137.74	2	< .001	***
All Other	0.913	< .001	***	5.03	2	.081	
Instruction	0.903	< .001	***	779.07	2	< .001	***
Public & Research	0.880	< .001	***	900.54	2	< .001	***
Student Support	0.847	< .001	***	13.49	2	.001	**

Note. *** $p < .001$. ** $p < .01$. * $p < .05$. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. Acad. & Inst. Sup. = academic and institutional support expenditures. EFC = expected family contribution. Public Serv. & Rsch. = public service and research expenditures. Student Serv. Sup. = student services support expenditures.

Homogeneity of variance. The final preliminary assumption examined is the homogeneity of variance. The review for the homogeneity of variance underwent a review of visualizations (Figure 9 and Figure 10) and the utilization of statistical tests (See Appendix J). In Figure 9, the distribution of the population by retention factors for the categorical factors are displayed. The graphs were confirmed by the statistical tests to indicate gender, first generation status, HS locale, and major groupings were found to have heterogeneity within the distribution.

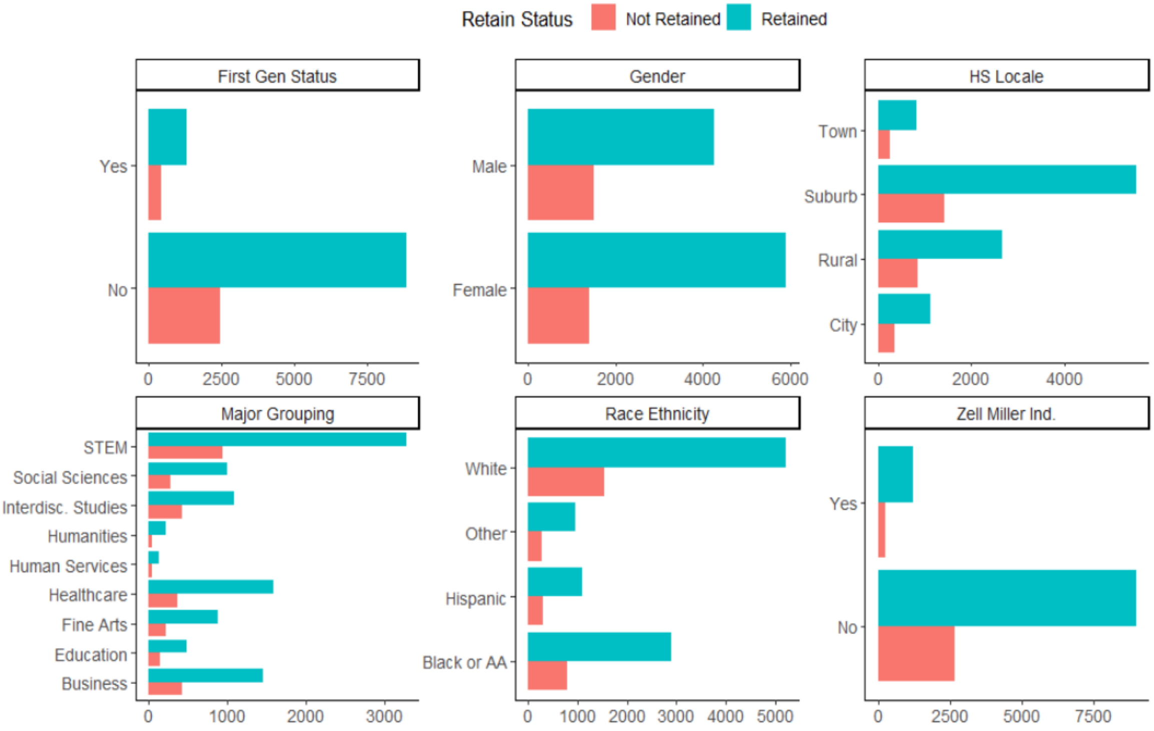


Figure 9. Distribution of categorical variables by retention status. The figure displays a bar chart of the distribution of the population by the retention status for the categorical variables. First Gen Status = first generation status. HS = high school. Ind. = indicator.

From Figure 10, the distribution of the population for the continuous variable is illustrated. Utilizing the review of the boxplots and the statistical tests, HS GPA, admissions test scores, AP and IB hours, EFC, GA HOPE scholarship, federal subsidized and unsubsidized loans, other loans, and all expenditures were found to have heterogeneity of variances.

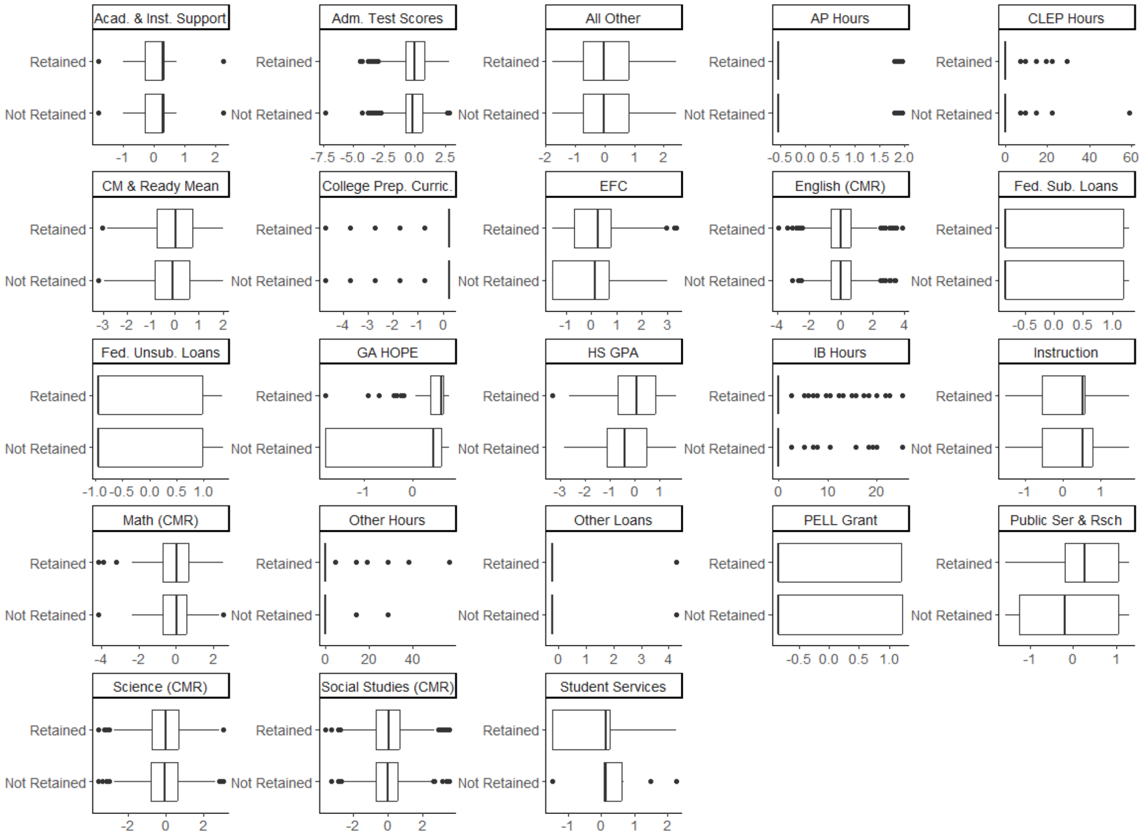


Figure 10. Boxplots of continuous variables by retention status. The figure displays the population's boxplots by retention status for the continuous variables. Adm. Tests Score = admissions test scores. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. Acad. & Inst. Sup. = academic and institutional support expenditures. EFC = expected family contribution. Public Ser & Rsch. = public service and research expenditures. Student Serv. Sup. = student services support expenditures.

In reviewing the training data set, numerous factors influencing the dependent variables continued to maintain existing relationships, even after adjustments to eliminate interrelationships between independent factors. Despite attempts to transform the data to conform to normality assumptions, the factors persisted in their non-normal distribution. Additionally, a few factors exhibited variations in the distributions. The presence of nonnormality and heterogeneity of variance in the data set raises concerns about the reliability and validity of the results. The violation of the normality assumption has the potential to impact the accuracy of tests and predictions, introducing biased estimates due to skewed data and the presence of outliers. Similarly, factors displaying heterogeneity of variance may affect the precision of the analysis, influencing the reliability of the results. Recognizing and addressing these issues is essential for ensuring the robustness and interpretability of the findings.

First Research Question

The following is the first research question:

1. Are student characteristics, precollege characteristics (including high school curriculum quality), financial situations, major or program of study, and institutional financial expenditures significant predictors in first-time, full-time freshmen's academic performance in their first year?
    a. Are student characteristics (gender, race and ethnicity, family educational background, and locale), precollege characteristics (high school curriculum quality, high school GPA, and admissions test scores), financial situations (family financial situations and financial aid), major or program of study, and institutional financial expenditures significant predictors of first-time, full-time freshmen's first-fall GPA?
    b. Are student characteristics (gender, race and ethnicity, family educational background, and locale), precollege characteristics (high school curriculum quality, high school GPA, and admissions test scores), financial situations (family financial situations and financial aid), major or

program of study, and institutional financial expenditures significant predictors of first-time, full-time freshmen's first-year GPA?

c. Are student characteristics (gender, race and ethnicity, family educational background, and locale), precollege characteristics (high school curriculum quality, high school GPA, and admissions test scores), financial situations (family financial situations and financial aid), major or program of study, and institutional financial expenditures significant predictors of first-time, full-time freshmen's one-year retention status?

First-fall GPA. The data set used for model development was the training data set, comprising 13,078 observations, with seven observations removed due to missing data in the dependent variable. The revised number of observations was 13,071. Six predictive algorithms were employed for data analysis, which included linear regression, three support vector machines (SVM), random forest, and extreme gradient boosting (XGBoost). The SVM models utilized three different kernels: linear, polynomial, and radial basis function. The linear regression model did not require tuning. The three SVM, random forest, and XGBoost models underwent tuning using a grid of 20 models based on a 10-fold cross-validation samples derived from the training data set. To ensure reproducibility, the set.seed() function was utilized. Within the tuning process, various model parameters were examined to determine the best optimal performance. The best model was selected based on the lowest RMSE value. Following the development of predictive algorithms, all six models analyzed the training and testing data sets to identify the factors influencing the first-fall GPA, with the emphasis placed on the results from the testing data set.

Linear regression. Linear regression is used to analyze the relationship between multiple independent variables and a continuous dependent variable. Moreover, linear regression can be used to predict the outcome between the independent and dependent variables. The linear regression model was built using the linear_reg() function with the engine set to lm and the mode set to regression. The results of the model on the training data set are displayed in Table 12. The linear regression model proved to be significant (R^2 = .303, *adj* R^2 = .302, $F(29, 13{,}041) = 195.6$, $p < .001$), explaining 30.3% of the variance in the data set. The regression model displayed a small effect size. The model's RMSE was 0.848. Of the 29 independent variables, 18 factors were found to be significant. Of the student characteristics, gender ($B = -.187$, $\beta = -0.091$, $t = -11.914$, $p < .001$), race and ethnicity ($B = .026$, $\beta = 0.025$, $t = 3.271$ $p < .001$), and HS locale ($B = -.040$, $\beta = -0.039$, $t = -5.073$, $p < .001$) were found to be significant factors. From the student characteristics, gender of a student was the strongest contribution to the first-fall GPA earned. Based on the results, male students tended to earn slightly lower first-fall GPA compared to their female counterparts. Although race and ethnicity, along with high school locale, were statistically significant, their practical contribution on the earned GPA was very small.

Table 12

Results of Linear Regression on Training Data Set for First-Fall GPA Dependent Variable

	B	β	*SE*	*t*	*p*	
Intercept	2.995		0.033	91.009	< .001	***
Student Characteristics						
Gender	-.187	-0.091	0.016	-11.914	< .001	***
Race/Ethnicity	.026	0.025	0.008	3.271	.001	**
First Generation Status	-.017	-0.006	0.023	-0.752	.452	
HS Locale	-.040	-0.039	0.008	-5.073	< .001	***
Pre-college Characteristics						
HS GPA	.350	0.344	0.012	28.865	< .001	***
Admissions Test Scores	-.012	-0.011	0.010	-1.126	.260	
AP Hours	.076	0.075	0.009	8.731	< .001	***
CLEP Hours	.010	0.010	0.007	1.390	.165	
IB Hours	.020	0.020	0.007	2.676	.007	**
Other Hours	-.001	-0.001	0.007	-0.070	.944	
College Prep. Curriculum	-.004	-0.004	0.008	-0.528	.597	
CM & Ready Mean	.070	0.069	0.012	5.886	< .001	***
English (CMR)	.039	0.039	0.010	3.864	< .001	***
Math (CMR)	.001	0.001	0.010	0.106	.915	
Science (CMR)	-.008	-0.008	0.009	-0.911	.362	
Social Studies (CMR)	.044	0.044	0.009	5.125	< .001	***
Financial Situations						
EFC	.043	0.042	0.015	2.936	.003	**
GA HOPE Scholarship	.156	0.153	0.010	15.076	< .001	***
Zell Miller Indicator	.201	0.062	0.029	6.910	< .001	***
PELL Grant	.010	0.010	0.014	0.729	.466	
Federal Sub. Loans	-.006	-0.006	0.011	-0.561	.575	
Federal Unsub. Loans	-.018	-0.018	0.010	-1.857	.063	
Other Loans	-.016	-0.015	0.008	-2.067	.039	*
Major Groupings	-.015	-0.030	0.004	-4.077	< .001	***
Institutional Expenditures						
Academic & Institutional Support	.064	0.063	0.011	5.638	< .001	***
All Others	-.049	-0.048	0.010	-4.778	< .001	***
Instruction	-.069	-0.068	0.011	-6.262	< .001	***
Public Service & Research	-.004	-0.004	0.010	-0.400	.689	
Student Service Support	.051	0.051	0.010	5.082	< .001	***

Note. $R^2 = .303$, *adj* $R^2 = .302$, $F(29, 13{,}041) = 195.6$, $p < .001$. *** $p < .001$. ** $p < .01$. * $p < .05$. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

Six pre-college characteristics were identified as significant predictors, with three variables linked to the HS curriculum variables. HS GPA (*B* = .350, β = 0.344, *t* = 28.865, *p* < .001) exhibited the most influential factor. The strength of the factor indicated students with higher HS GPAs were more likely to achieve higher first-fall GPAs when compared to their peers with lower HS GPAs. Regarding the advanced standing hours, AP hours (*B* = .076, β = 0.075, *t* = 8.731, *p* < .001) and IB hours (*B* = .020, β = 0.020, *t* = 2.676, *p* = .007) were found to be significant. However, the variables' practical impact on the first-fall GPA was minimal. Among the significant HS curriculum variables, content mastery and readiness mean (*B* = .070, β = 0.069, *t* = 5.886, *p* < .001), English proficiency levels difference from content mastery and readiness mean (*B* = .039, β = 0.039, *t* = 3.864, *p* < .001), and social studies proficiency levels difference from content mastery and readiness mean (*B* = .044, β = 0.044, *t* = 5.125, *p* < .001) were found to be significant. .

Of the financial factors, GA HOPE scholarship (*B* = .156, β = 0.153, *t* = 15.076, *p* < .001) and Zell Miller indicator (*B* = .201, β = 0.062, *t* = 6.910, *p* < .001) exhibited the highest contributions, showing strong positive relationships with the first-fall GPA. While the variables' contributions were small, expected family contribution (*B* = .043, β = 0.042, *t* = 2.936, *p* = .003) and other loans (*B* = -.016, β = -0.015, *t* = -2.067, *p* = .039) were found to be significant. In terms of major groupings of the programs of study, a slight negative relationship was found to be significant, *B* = -.015, β = -0.030, *t* = -4.077, *p* < .001. For the expenditure factors, academic and institutional support (*B* = .064, β = 0.063, *t* = 5.639, *p* < .001), instruction (*B* = -.069, β = -0.068, *t* = -4.778, *p* < .001), student services support (*B* = .05, β = 0.051, *t* = 5.082, *p* < .001), and all others (*B* = -

.049, $\beta = -0.048$, $t = -4.778$, $p < .001$) exhibited a small significant impact on the first-fall GPA.

Assumptions for linear regression. The first assumption examined for linear regression was the presence of a linear relationship between the independent variables and the dependent variable. Among the 29 independent variables, 23 demonstrated a significant correlational relationship with the first-fall GPA. The two variables with the strongest relationships with the first-fall GPA were HS GPA ($r(13{,}071) = .487$, $p < .001$) and GA HOPE scholarship ($r(13{,}071) = .419$, $p < .001$), both exhibiting moderate positive relationships with the dependent variable. The Zell Miller indicator ($r(13{,}071) = .281$, $p < .001$), admissions test scores ($r(13{,}071) = .258$, $p < .001$), and AP Hours ($r(13{,}071) = .251$, $p < .001$) exhibited small or low positive relationships with the first-fall GPA. The remaining variables had either very weak or no relationship with the dependent variable. Low levels of multicollinearity were identified through VIF analysis, and any pre-existing multicollinearity issues were addressed during the data preprocessing stage.

The assumption regarding the normality of errors was evaluated using three statistical tests: Kolmogorov-Smirnov ($D = .098$, $p < .001$), Jarque-Bera ($\chi^2(2) = 3{,}431.3$, $p < .001$), and the Shapiro-Wilks ($W = .939$, $p < .001$) test conducted on a sample of the first 5,000 observations. All three tests indicated a violation of the normality assumption, indicating the errors were not normally distributed. Both the standardized and studentized residuals had a mean approximately zero indicating the assumption of the mean of the errors equal to zero was not violated. A Durbin-Watson value of 1.981 ($p = .290$) indicated no autocorrelation, demonstrating the errors were independent. The assumption of homogeneity of variance was violated due to significant results from the Breusch-Pagan or non-constant variance test ($\chi^2(1) = 846.334$, $p < .001$).

The results of the linear regression model on the testing data set are displayed in Table 13. The linear regression model proved to be significant ($R^2 = .283$ *adj* $R^2 = .281$, $F(29, 8{,}681) = 118.4$, $p < .001$), explaining 28.3% of the variance in the data set. The regression model displayed a small effect size. From the model on the training data set, the variance accounted for decreased two percentage points. The model's RMSE was 0.864, which is an increase of 0.016 points. Of the 29 independent variables, 17 were found to be significant. While other loans variable was significant from the training data set, the factor was found not to be significant in the testing data set.

Table 13

Results of Linear Regression on Testing Data Set for First-Fall GPA Dependent Variable

	B	β	*SE*	*t*	*p*	
Intercept	2.958		0.041	72.302	< .001	***
Student Characteristics						
Gender	-.152	-0.074	0.020	-7.714	< .001	***
Race/Ethnicity	.025	0.024	0.010	2.465	.014	*
First Generation Status	-.008	-0.002	0.029	-0.260	.795	
HS Locale	-.032	-0.032	0.010	-3.317	.001	**
Pre-college Characteristics						
HS GPA	.341	0.335	0.015	22.908	< .001	***
Admissions Test Scores	-.015	-0.015	0.013	-1.157	.247	
AP Hours	.062	0.061	0.011	5.729	< .001	***
CLEP Hours	-.006	-0.005	0.011	-0.551	.581	
IB Hours	.022	0.022	0.009	2.417	.016	*
Other Hours	-.002	-0.003	0.006	-0.321	.749	
College Prep. Curriculum	-.015	-0.015	0.010	-1.589	.112	
CM & Ready Mean	.045	0.043	0.015	2.992	.003	**
English (CMR)	.039	0.038	0.013	3.085	.002	**
Math (CMR)	-.013	-0.012	0.013	-0.990	.322	
Science (CMR)	.002	0.002	0.011	0.185	.854	
Social Studies (CMR)	.067	0.065	0.011	6.064	< .001	***
Financial Situations						
EFC	.040	0.039	0.018	2.226	.026	*
GA HOPE Scholarship	.162	0.158	0.013	12.545	< .001	***
Zell Miller Indicator	.214	0.063	0.037	5.745	< .001	***
PELL Grant	.008	0.008	0.017	0.479	.632	
Federal Sub. Loans	.002	0.002	0.013	0.137	.891	
Federal Unsub. Loans	-.012	-0.012	0.012	-0.957	.339	
Other Loans	.001	0.001	0.009	0.060	.952	
Major Groupings	-.016	-0.032	0.005	-3.488	< .001	***
Institutional Expenditures						
Academic & Institutional Support	.061	0.059	0.014	4.295	< .001	***
All Others	-.046	-0.046	0.012	-3.675	< .001	***
Instruction	-.088	-0.086	0.014	-6.300	< .001	***
Public Service & Research	.000	0.000	0.013	0.022	.982	
Student Service Support	.026	0.026	0.013	2.080	.038	*

Note. $R^2 = .283$ *adj* $R^2 = .281$, $F(29, 8{,}681) = 118.4$, $p < .001$. *** $p < .001$. ** $p < .01$. * $p < .05$. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

A variable importance analysis, as illustrated in Figure 11, was conducted on the training and testing data sets, and the importance values were rescaled to 100 for comparison across models. For the linear regression model, factors were color-coded to indicate the type of impact, ranging from negative to positive. According to the analysis on the training data set, HS GPA (importance = 28.865, rescaled importance = 100.000) had the most significant influence on the first-fall GPA, with a positive impact. The influence indicated students with higher HS GPAs were strongly associated with higher first-fall GPAs, and students with lower HS GPAs were strongly associated with lower first-fall GPAs. The GA HOPE scholarship (importance = 15.076, rescaled importance = 52.230) showed a similar but less substantial, positive impact on first-fall GPAs. In contrast, the gender of the student (importance = 11.914, rescaled importance = 41.274) exhibited a notable negative impact. This impact indicated male students were more likely to earn lower first-fall GPAs. Surprisingly, none of the five HS curriculum variables were among the top five factors influencing first-fall GPA within the training data set.

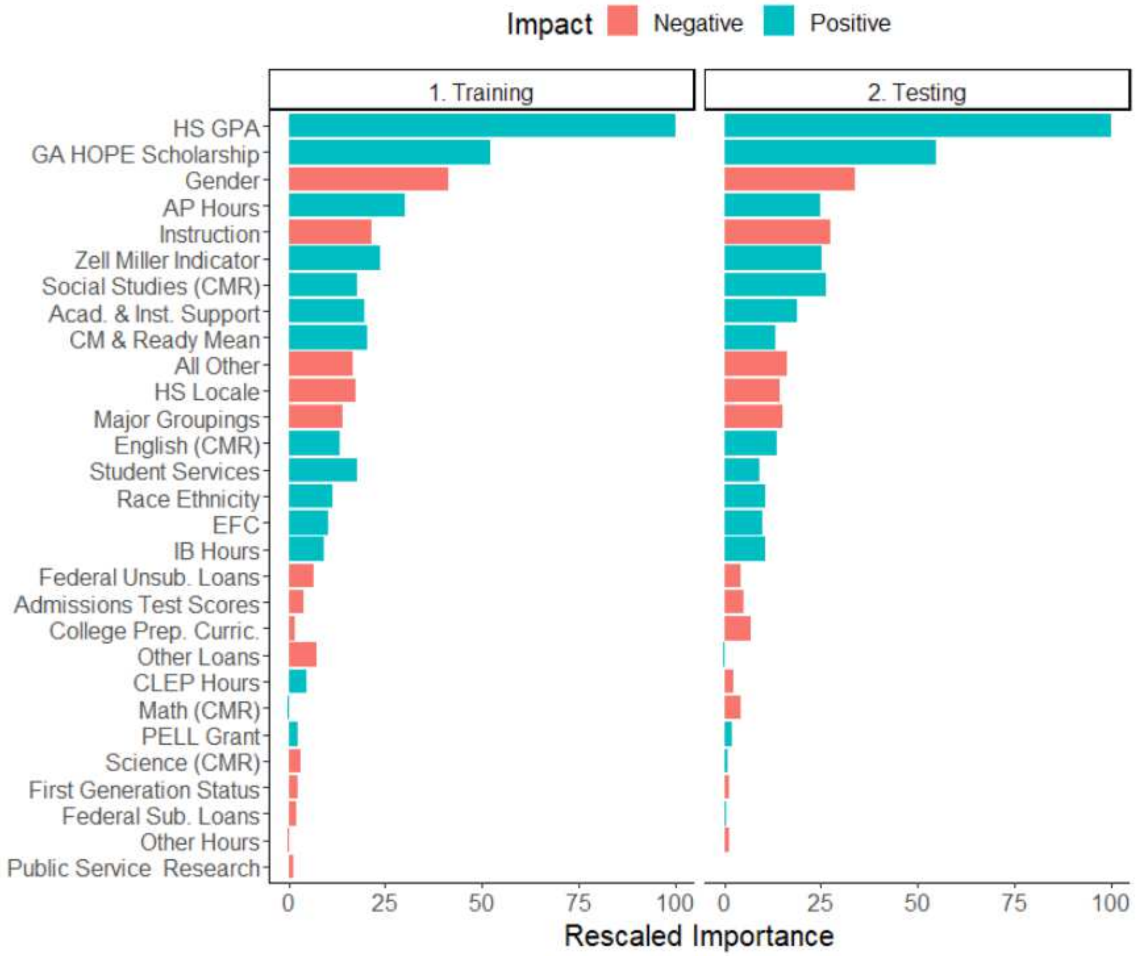


Figure 11. First-fall GPA variable importance plot for the linear regression model using the training and testing data sets. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph. The color of the bar indicates whether the impact is negative or positive on the first-fall GPA. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

Examining the variable importance analysis of the testing data set, HS GPA (importance = 22.908, rescaled importance = 100.000) remained the most influential factor. The GA HOPE Scholarship (importance = 12.545, rescaled importance = 54.763) retained its position as the second most influential factor, and like the analysis of the training data set, the factor's influence is far less than that of HS GPA. As the top negative factor, gender remained the third influential factor. While AP hours and Zell

Miller indicator were placed fourth and fifth in the training data set, instructional expenditures (importance = 6.300, rescaled importance = 27.502) and social studies proficiency levels difference from content mastery and readiness mean (importance = 6.064, rescaled importance = 26.472) replaced them in the testing data set.

Support vector machine with linear kernel. The SVM algorithm using linear kernel model was built using the svm_linear() function with the engine set to kernlab and the model set to regression. The cost and margin components in the model were tuned across a grid of 20 models using the training data, with the set.seed() function replication purposes. Despite the tuning process resulting in 20 models with similar performance, the optimal model achieved an RMSE value of 0.866 and an R^2 value of .299. While this optimal model demonstrated the lowest RMSE, it only explained 29.9% of the variance in the data set. In this model, the cost was set to 0.304, and the margin was set to 0.194.

Figure 12 presents the results of the variable importance analysis on the training and testing data sets, where the importance values were rescaled to 100 for comparison across models. Unlike the linear regression model, this analysis did not calculate the type of impact of the variables. HS GPA (importance = 0.135, rescaled importance = 100.000) exerted the greatest influence on the first-fall GPA in the training data set. This influence suggests students with a higher HS GPA correspond to a higher first-fall GPA earned. Likewise, students with a lower HS GPA correspond to a lower first-fall GPA. Additionally, the GA HOPE scholarship (importance = 0.025, rescaled importance = 18.790) and gender (importance = 0.010, rescaled importance = 7.451) were the next two most impactful factors on first-fall GPA. The remaining variables had a very small influence on the dependent variable within the training data set.

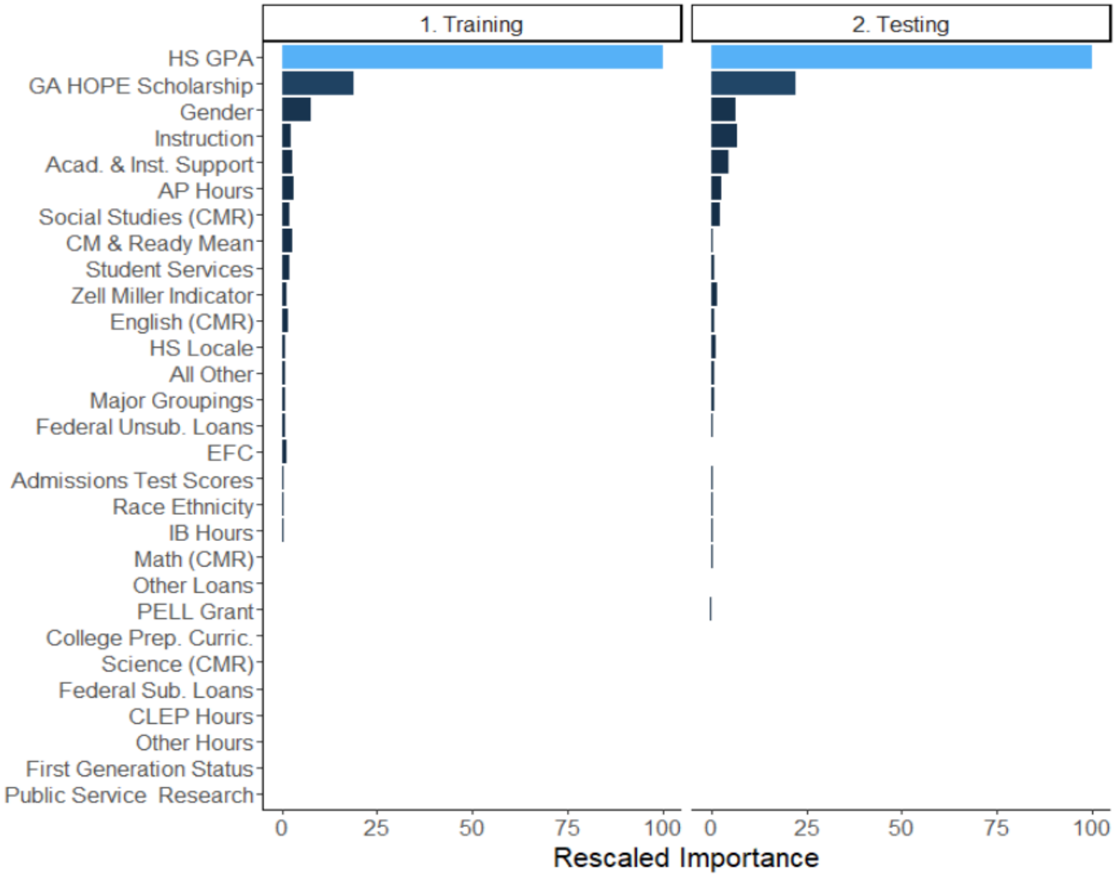


Figure 12. First-fall GPA variable importance plot for the SVM model using a linear kernel on training and testing data set. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

In the testing data set, HS GPA (importance = 0.129, rescaled importance = 100.000) and GA HOPE scholarship (importance = 0.028, rescaled importance = 21.979) remained the top two factors impacting the first-fall GPA. Instructional expenditures (importance = 0.009, rescaled importance = 6.948) replaced gender as the third factor, and gender (importance = 0.008, rescaled importance = 6.529) fell to the fourth factor. Academic and institutional expenditures (importance = 0.006, rescaled importance = 4.464) were the fifth top factors in the testing data set. The remaining variables within the testing data set had a very small influence on the dependent variable.

Support vector machine with polynomial kernel. Using a polynomial kernel, another SVM algorithm was tuned employing the svm_poly() function with the kernlab engine and regression as the model type. The cost, degree, scale_factor, and margin parameters in the svm_poly() function were tuned across 20 models using the training data, with the set.seed() function for replication purposes. The resulting optimal model achieved an RMSE value of 0.865 and an R^2 value of .300, explaining 30.0% of the variance within the data set. The tuned model had a cost of 14.782, a degree of 3, a scale factor of 0.0001, and a margin of 0.188.

Figure 13 presents the variable importance analysis on the training and testing data sets, where the importance values were rescaled to 100 for comparison across models. HS GPA (importance = 0.131, rescaled importance = 100.000) emerged as the factor with the most substantial impact on the first-fall GPA within the training data set. Like the previous models, students with a higher HS GPA are associated with a higher first-fall GPA, while students with a lower HS GPA are associated with a lower first-fall GPA. The GA HOPE scholarship (importance = 0.026, rescaled importance = 20.030) and gender (importance = 0.010, rescaled importance = 7.653) were the next most influential factors on first-fall GPA. The impact of the remaining variables was very small in the training data set.

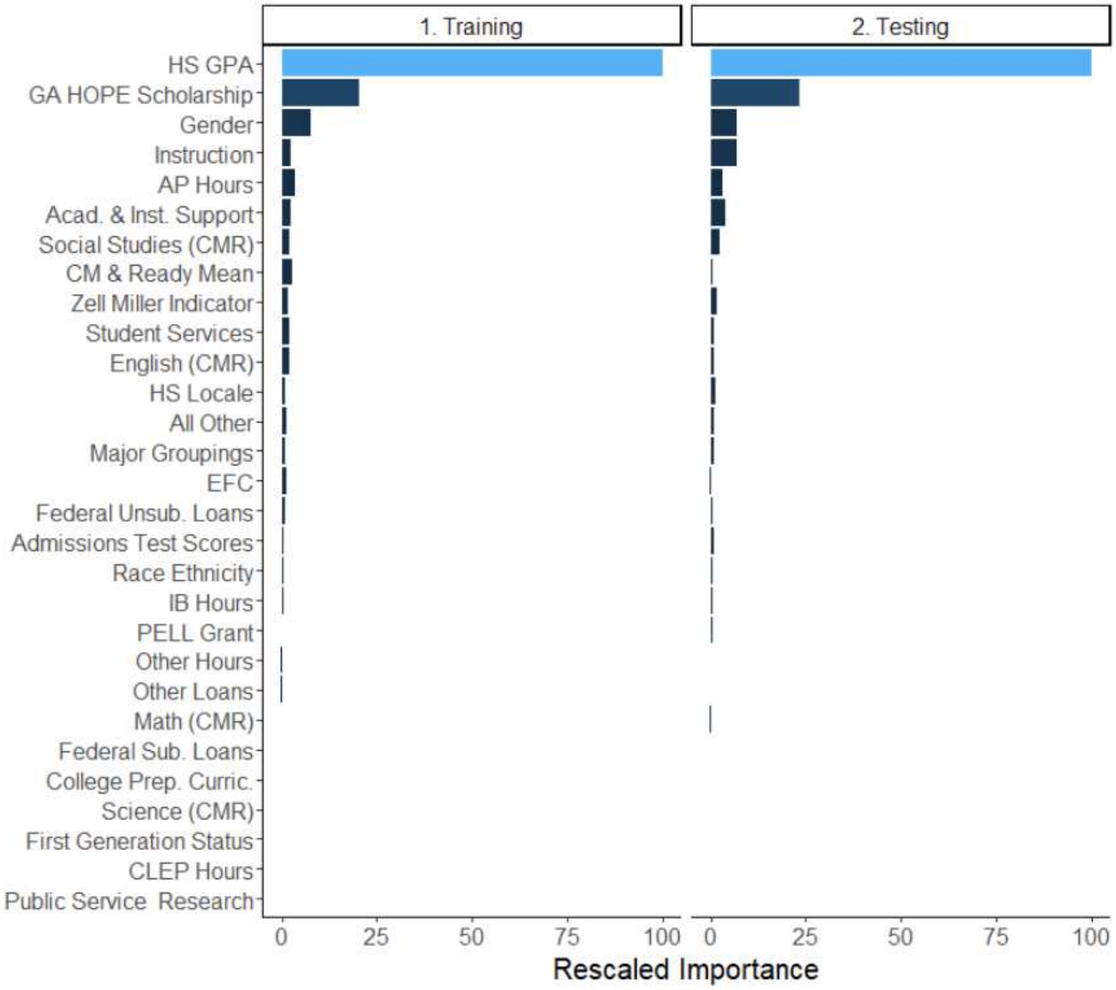


Figure 13. First-fall GPA variable importance plot for the SVM model using a polynomial kernel on the training and testing data sets. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

The review of the variable importance analysis on the testing data set reveals HS GPA (importance = 0.125, rescaled importance = 100.000), GA HOPE scholarship (importance = 0.029, rescaled importance = 23.252), and gender (importance = 0.009, rescaled importance = 6.958) retained their positions as the top three influential factors. Instruction expenditures (importance = 0.009, rescaled importance = 6.809) and academic and institutional support expenditures (importance = 0.005, rescaled importance = 4.035) were found in the testing data set to have a small impact on the first-fall GPA. The remaining variables within the testing data set were found to have a very small impact on the dependent variable.

Support vector machine with radial basis function kernel. The final SVM was built using the radial basis function kernel, with the engine set to kernlab and the model set to regression. Through a grid of 20 models developed on training data set, the cost, radial basis function sigma, and margin were tuned. The set.seed() function was utilized for replication. The optimal model achieved an RMSE value of 0.863 and an R^2 value of .305, accounting for approximately 30.5% of the variance. The tuned features of the model included a cost of 19.460, sigma of 0.0005, and a margin of 0.123.

In Figure 14, the variable importance analysis results on the training and testing data sets are displayed, with the results rescaled to 100 for comparison across models. Within the training data set, HS GPA (importance = 0.133, rescaled importance = 100.000) emerged as the most influential factor on the first-fall GPA. Additionally, the GA HOPE scholarship (importance = 0.026, rescaled importance = 19.855), gender (importance = 0.010, rescaled importance = 7.745), and AP hours (importance = 0.007, rescaled importance = 5.538) were the next most impactful variables on first-fall GPA. The impact of the remaining variables was too low to significantly affect the first-fall GPA in the training data set. Reviewing the importance of the factors on the testing data set, HS GPA (importance = 0.128, rescaled importance = 100.000) and GA HOPE scholarship (importance = 0.029, rescaled importance = 23.002) remained the top two factors. Expenditures on instruction (importance = 0.012, rescaled importance = 9.472) was the third influential factor, while gender (importance = 0.009, rescaled importance =

7.242) fell to the fourth influential factor spot. The contributions of the remaining factors from the testing data set were too small.

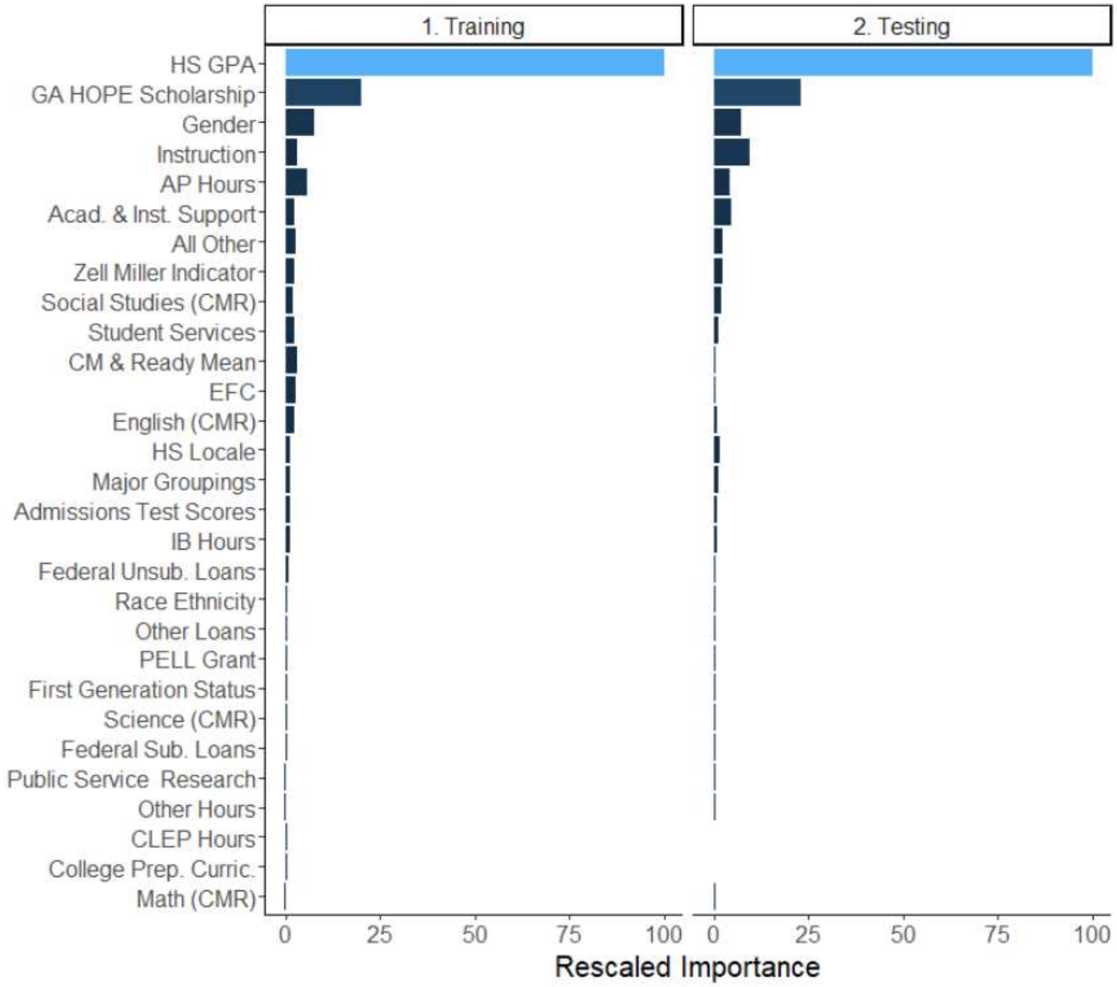


Figure 14. First-fall GPA variable importance plot for the SVM model using a radial basis function kernel on the training and testing data sets. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

Random forest. The random forest algorithm was built using the rand_forest() function with the engine set to ranger and the mode set to regression. In the rand_forest() function, the mtry, trees, and min_n options were tuned to find the values for the optimal model. The model was tuned through a grid of 20 models using the training data set, with the set.seed() function for replication purposes. The best model was selected based on the lowest RMSE value, and the optimal model exhibited an RMSE value of 0.832. The R^2 value for this model was .330, indicating the model accounts for 33.0% of the variance within the data set. The optimal model included 1,580 trees with a mtry of 9 and a minimum number of observations of 37. The optimal model's mtry of 9 is close to being one-third of the independent variables as recommended by James et al. (2013) and Kuhn and Johnson (2013).

Figure 15 illustrates the variable importance analysis results on the training and testing data sets, with the importance values rescaled to 100 for comparison across models. HS GPA (importance = 0.293, rescaled importance = 100.000) and GA HOPE scholarship (importance = 0.266, rescaled importance = 90.704) were the two major factors influencing the first-fall GPA in the training data set. Content mastery and readiness mean (importance = 0.048, rescaled importance = 16.531) and Zell Miller indicator (importance = 0.043, rescaled importance = 14.823) were the third and fourth factors affecting the first-fall GPA. The content mastery and readiness mean and Zell Miller indicator variables exhibited a small impact. From the testing data set, HS GPA (importance = 0.286, rescaled importance = 100.000), GA HOPE scholarship (importance = 0.255, rescaled importance = 89.142), and content mastery and readiness mean (importance = 0.036, rescaled importance = 12.458) remained the top three most influential factors. The fourth and fifth top factors from the testing data set were the EFC (importance = 0.034, rescaled importance = 11.942) and the English proficiency levels difference from content mastery and readiness mean (importance = 0.031, rescaled importance = 10.921).

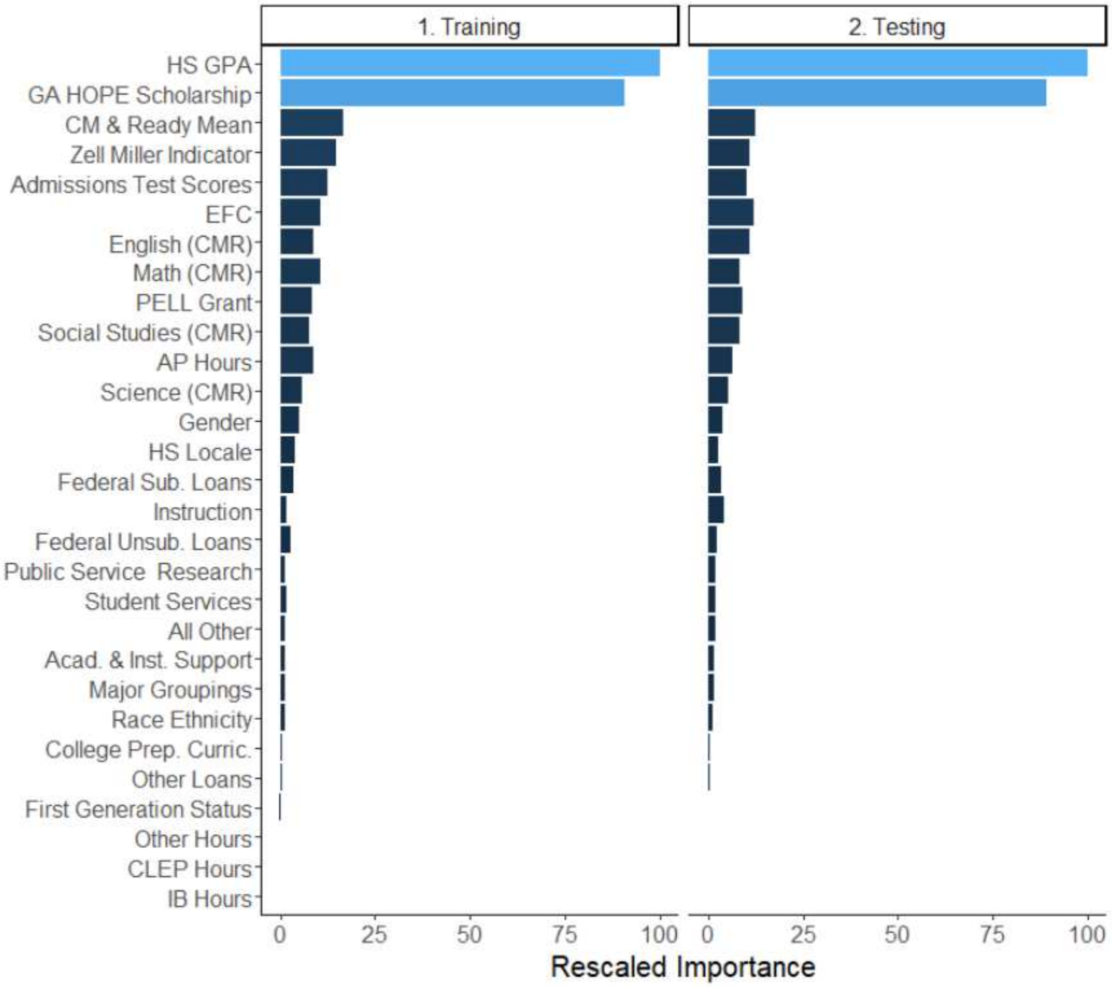


Figure 15. First-fall GPA variable importance plot for the random forest model on the training and testing data sets. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

Extreme gradient boosting. The XGBoost model was constructed using the boost_tree() function with the engine set to xgboost and the mode set to regression. To develop the optimal model, the trees, tree depth, min_n, loss reduction, sample size, mtry, and learn rate were tuned through a grid of 20 models using the training data set, with the set.seed() function for replication purposes. The best model was selected based on the lowest RMSE value. The optimal model exhibited an RMSE value of 0.832 and an R^2

value of 0.330, accounting for 33.0% of the variance within the data set. The optimal model consisted of 1,804 trees with an mtry of 10, minimum observations of 3, and a tree depth of 8. The model's learn rate was set to 0.005, loss reduction was 4.797, and sample size was 0.106.

Figure 16 displays the variable importance analysis of the training and testing data sets, with the importance values rescaled to 100 for comparison purposes across models. HS GPA (importance = 0.189, rescaled importance = 100.00) and GA HOPE scholarship (importance = 0.106, rescaled importance = 55.990) were the top two factors impacting the first-fall GPA in the training data set. The five HS curriculum variables were ranked third through seventh in terms of importance within the training data set. HS GPA (importance = 0.182, rescaled importance = 100.000) and GA HOPE scholarship (importance = 0.096, rescaled importance = 52.706) remained the top two influential factors in the testing data set. The social studies proficiency levels difference from content mastery and readiness mean (importance = 0.071, rescaled importance = 39.063) was the highest HS curriculum factor, placing third. The remaining four HS curriculum factors remained within the top 10 factors. Admissions test scores (importance = 0.068, rescaled importance = 37.373) split the HS factors by placing sixth.

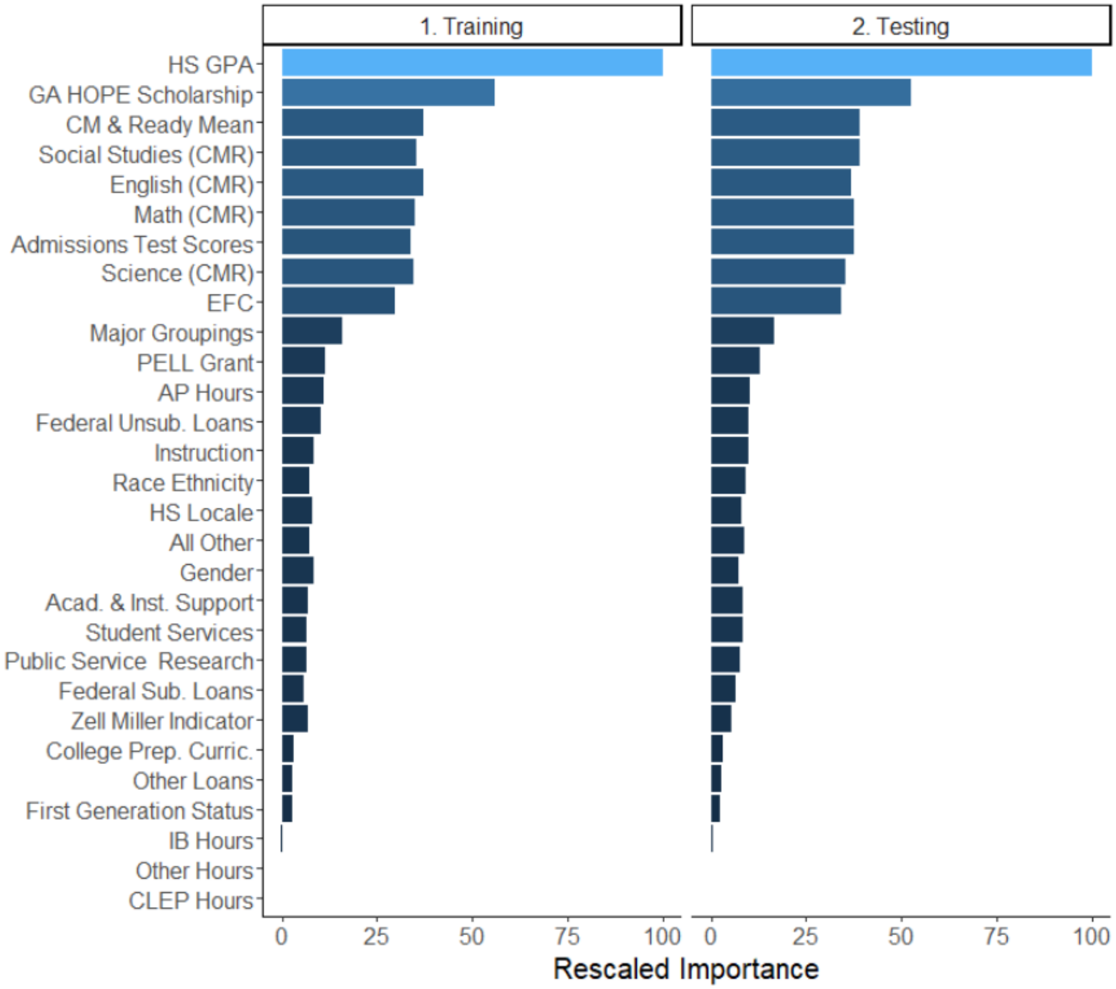


Figure 16. First-fall GPA variable importance plot for the XGBoost model on the training and testing data sets. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

Variable importance comparison for first-fall GPA. While patterns emerged from the results of the variables' importance analysis from the training data set, the results from the testing data set exhibited more important patterns to assess which factors impact the first-fall GPA, as displayed in Figure 17. The HS GPA factor demonstrated to be the dominant factor overall across the models. The second most influential factor was the GA HOPE scholarship, with the highest impact found in the random forest model. No additional factors were consistent in their rankings. The linear regression and SVM using a polynomial kernel found gender as the third-ranked factor, while the SVMs using linear and radial basis function kernels found expenditures on instruction to be the third most influential factor. The random forest indicated the content mastery and readiness mean was ranked third, while the XGBoost model found the social studies proficiency levels difference from content mastery and readiness mean to be the third most influential factor.

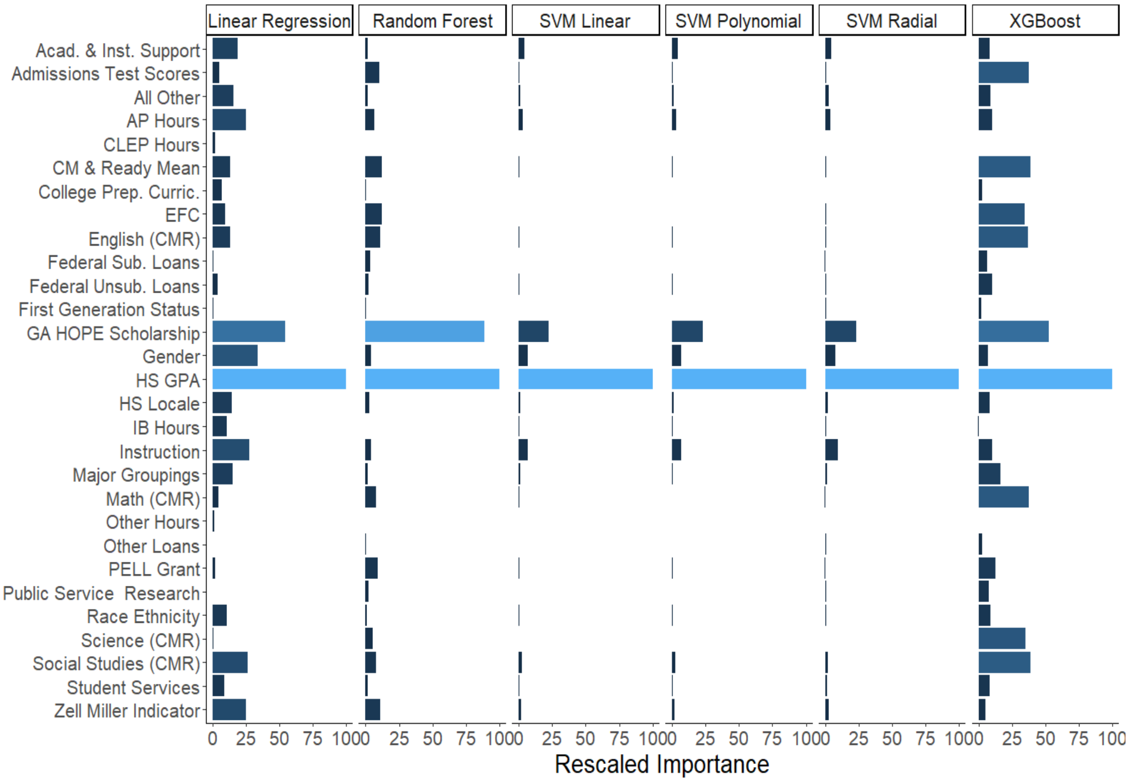


Figure 17. Comparison of variable importance results on testing data set for first-fall GPA. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

Reviewing the impact of the five HS curriculum factors in importance analysis across the models, these variables overall exhibited no consistency in rankings. Only the XGBoost model indicated the five factors exhibited moderate influences on the first-fall GPA. Of the subject proficiency levels in the XGBoost model, the social studies factor was ranked the highest, while science was ranked the least. Other than the random forest model, the remaining models ranked the social studies proficiency levels difference from content mastery and readiness mean above the other HS curriculum factors. The random forest model indicated the content mastery and readiness mean was higher than the subject proficiency levels. Furthermore, the number of AP hours was found to exhibit influence in the linear regression model only, while the other advanced standing hours exhibited very little influence on the first-fall GPA. Admissions test scores exhibited very little influence on the first-fall GPA when excluding the XGBoost model. The XGBoost model ranked admissions test scores in sixth place. Lastly, the number of satisfied college preparatory curriculum requirements exhibited relatively no influence on the dependent variable.

While the GA HOPE scholarship consistently ranked in second place, no other financial situation variables from the analyses on the testing data set exhibited consistency across the models. From the linear regression and SVM models, the Zell Miller scholarship factor was the second most influential financial situations factor, with the strongest contribution found in the linear regression model. Both the random forest and XGBoost models found the EFC to be the second most influential factor among the financial situation variables. The remaining financial factors exhibited very little influence across the models. Likewise, the selected major and remaining expenditures exhibited very little influence on the first-fall GPA.

First-year GPA. Predictive models for the first-year GPA dependent variable were developed using the initial training data set, which initially contained 13,078 observations. However, 62 observations were removed from this data set due to missing observations for the first-year GPA, resulting in a revised training data set with 13,016 records. A total of six statistical models were developed for data analysis, including linear regression, three SVM, random forest, and XGBoost. The linear regression model did not require any tuning. The SVM models employed linear, polynomial, and radial basis function kernels. To optimize the models, three SVM models, random forest, and XGBoost were tuned by exploring a grid of 20 models based on 10-fold cross-validation samples from the training data. To ensure replicability, the set.seed() function was used. Model parameters were tuned using the grid, and the best-performing model was selected based on the lowest RMSE value. Following the development of predictive algorithms, all six models analyzed the training and testing data sets to identify the factors influencing the first-year GPA, with the emphasis placed on the results from the testing data set.

Linear regression. A linear regression was conducted on the training data set, and the results are displayed in Table 14. The linear regression model was significant, $R^2 = .325$, adj $R^2 = .324$, $F(29, 12{,}986) = 215.90$, $p < .001$. The regression model accounts for 32.5% of the variance found within the data. The regression models exhibited a small effect size. A total of 17 out of the 29 independent variables were found to be significant. The model's RMSE was 0.795. For the student characteristics variables, gender (B = -

.201, β = -0.103, *t* = -13.591, *p* < .001), race and ethnicity (*B* = .027, β = 0.022, *t* = 3.618, *p* < .001), and HS locale (*B* = -.041, β = -0.043, *t* = -5.624, *p* < .001) were found to be significant. Within the student characteristics, gender exhibited the strongest influence on the first-year GPA. Male students are more likely to earn slightly lower first-year GPA than their female counterparts. Although race and ethnicity and the graduating HS locale were found to be significant, the factors contribution to the first-fall GPA were very small.

Overall HS GPA, *B* = .341, β = 0.352, t = 29.948, *p* < .001, exhibited not only the strongest contribution to the first-year GPA, but also within the pre-college characteristics. This factor influencing the earned GPA would indicate students with higher HS GPA are more likely to earn a higher GPA at the end of their first year. While admissions test scores, *B* = -.020, β = -0.021, *t* = -2.108, *p* = .035, the practical contribution of the scores are little. The advanced standing AP hours, *B* = .066, β = 0.068, *t* = 8.101, *p* < .001, and IB hours, *B* = .02, β = 0.02, *t* = 2.766, *p* = .006, were also found to be significant. Three of the five HS curriculum variables were found to be significant (content mastery and readiness mean, *B* = .067, β = 0.069, *t* = 5.977, *p* < .001; English proficiency levels difference from the content mastery and readiness mean, *B* = .041, β = 0.042, *t* = 4.251, *p* < .001; and social studies proficiency levels difference from the content mastery and readiness mean, *B* = .040, β = 0.041, *t* = 4.878, *p* < .001); yet, the variables' impacts were very small.

Table 14

Results of Linear Regression on Training Data Set for First-Year GPA Dependent Variable

	B	β	*SE*	*t*	*p*	
Intercept	3.009		0.031	97.284	< .001	***
Student Characteristics						
Gender	-.201	-0.103	0.015	-13.591	< .001	***
Race/Ethnicity	.027	0.028	0.008	3.618	< .001	***
First Generation Status	-.029	-0.010	0.022	-1.354	.176	
HS Locale	-.041	-0.043	0.007	-5.624	< .001	***
Pre-college Characteristics						
HS GPA	.341	0.352	0.011	29.948	< .001	***
Admissions Test Scores	-.020	-0.021	0.010	-2.108	.035	*
AP Hours	.066	0.068	0.008	8.101	< .001	***
CLEP Hours	.010	0.010	0.007	1.448	.148	
IB Hours	.019	0.020	0.007	2.766	.006	**
Other Hours	-.002	-0.002	0.007	-0.219	.827	
College Prep. Curriculum	.003	0.003	0.007	0.440	.660	
CM & Ready Mean	.067	0.069	0.011	5.977	< .001	***
English (CMR)	.041	0.042	0.010	4.251	< .001	***
Math (CMR)	-.002	-0.002	0.010	-0.172	.863	
Science (CMR)	-.001	-0.001	0.008	-0.150	.881	
Social Studies (CMR)	.040	0.041	0.008	4.878	< .001	***
Financial Situations						
EFC	.060	0.061	0.014	4.313	< .001	***
GA HOPE Scholarship	.165	0.171	0.010	17.044	< .001	***
Zell Miller Indicator	.192	0.062	0.027	7.051	< .001	***
PELL Grant	.015	0.015	0.013	1.137	.255	
Federal Sub. Loans	-.014	-0.015	0.010	-1.428	.153	
Federal Unsub. Loans	-.014	-0.015	0.009	-1.573	.116	
Other Loans	-.010	-0.010	0.007	-1.388	.165	
Major Groupings	-.012	-0.026	0.004	-3.535	< .001	***
Institutional Expenditures						
Academic & Institutional Support	.034	0.035	0.011	3.163	.002	**
All Others	.030	0.031	0.010	3.129	.002	**
Instruction	-.007	-0.008	0.010	-0.717	.474	
Public Service & Research	-.008	-0.009	0.010	-0.863	.388	
Student Service Support	.034	0.035	0.009	3.605	< .001	***

Note. $R^2 = .325$, adj $R^2 = .324$, $F(29, 12{,}986) = 215.90, p < .001$. *** $p < .001$. ** $p < .01$. * $p < .05$. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

The analysis revealed among the financial factors the expected family contribution (*B* = .060, β = 0.061, *t* = 4.313, *p* < .001), GA HOPE scholarship (*B* = .165, β = 0.171, *t* = 17.044, *p* < .001), and Zell Miller indicator (*B* = .192, β = 0.062, *t* = 7.051, *p* < .001) were the only significant variables influencing the first-year GPA. Of these three variables, the GA HOPE scholarship and Zell Miller indicator exhibited the highest impacts on the first-fall GPA. Specifically, students receiving the GA HOPE scholarship, but not at the Zell Miller level, tended to achieve higher first-year GPAs compared to those not receiving the scholarship. Furthermore, students who received the GA HOPE scholarship at the Zell Miller level demonstrated the highest first-year GPA among the three groups analyzed. Other variables related to financial aid situations did not show significant effects on first-fall GPA. The major grouping of the declared major (*B* = -.012, β = -0.026, *t* = -3.535, *p* < .001), academic & institutional support (*B* = .034, β = 0.035, *t* = 3.163, *p* =.002), student services support (*B* = .035, β = 0.036, *t* = 3.605, *p* < .001), and all others (*B* = .030, β = 0.031, *t* = 3.129, *p* = .002) were found to be significant factors impacting the first-year GPA.

Assumptions for linear regression. The first assumption examined for linear regression was the presence of a linear relationship between the independent variables and the dependent variable. Among the 29 independent variables, a total of 23 exhibited a significant correlational relationship with the first-year GPA. The two variables with the strongest relationship with the first-year GPA were HS GPA (*r*(13,016) = .506, *p* < .001) and GA HOPE scholarship (*r*(13,016) = .441, *p* < .001), with the relationships being moderate positive ones. The Zell Miller indicator (*r*(13,016) = .285, *p* < .001), admissions test scores (*r*(13,016) = 0.254, *p* < .001), and AP Hours (*r*(13,016) = .2471, *p*

$< .001$) exhibited small or low positive relationships with the first-year GPA. The remaining variables had either very weak or no relationship with the dependent variable. Low levels of multicollinearity were identified through VIF analysis, and any pre-existing multicollinearity issues were addressed during the data preprocessing stage.

The assumption regarding the normality of errors was evaluated using three statistical tests: Kolmogorov-Smirnov ($D = .104$, $p < .001$), Jarque-Bera ($\chi^2(2) = 4{,}678.9$, $p < .001$), and the Shapiro-Wilks ($W = .930$, $p < .001$) test conducted on a sample of the first 5,000 observations. All three tests indicated a violation of the normality assumption, indicating the errors were not normally distributed. Both the standardized and studentized residuals had a mean approximately equal to zero indicate the assumption of the mean of the errors equal to zero was not violated. A Durbin-Watson value of 1.994 ($p = .728$) indicated no autocorrelation, demonstrating the errors were independent. The assumption of homogeneity of variance was violated due to significant results from the Breusch-Pagan or non-constant variance test ($\chi^2(1) = 871.076$, $p < .001$).

The results of the linear regression model on the testing data set are displayed in Table 15. The linear regression model proved to be significant ($R^2 = .312$, $adj\ R^2 = .310$, $F(29, 8{,}663) = 118.40$, $p < .001$), explaining 31.2% of the variance in the data set. The regression model displayed a small effect size. From the model on the training data set, the variance accounted for decreased two percentage points. The model's RMSE was 0.807, which is an increase of 0.012 points. Of the 29 independent variables, 15 factors were found to be significant. Of the 17 factors from the training data sets found to be significant, 14 of them in the testing data set were significant. From the testing data set, one factor—instruction expenditures ($B = -.026$, $\beta = -0.027$, $t = -1.96$, $p = .047$)—was found to be significant when it was not significant in the training data set. While admissions test scores, IB hours, and student service support expenditures were found to be significant in the training data set, these factors were not found to be significant in the testing data set.

As presented in Figure 18, a variable importance analysis was conducted on the training and testing data sets, with the importance values rescaled to 100 for comparison purposes across the models. Additionally, the impact of each factor was indicated by color, ranging from negative to positive impact. Based on the importance analysis from the training data set, HS GPA (importance = 29.948, rescaled importance = 100.000) emerged as the most influential factor on first-year GPA. This impact was positive, signifying a higher HS GPA is strongly associated with a higher first-year GPA, while a lower HS GPA is strongly associated with a lower first-year GPA. The GA HOPE scholarship (importance = 17.044, rescaled importance = 56.913) also exerted a significant influence on first-year GPA, although to a lesser extent compared to HS GPA. This finding implies students receiving the scholarship are more likely to achieve a higher GPA than those without it. Conversely, the gender of a student (importance = 13.591, rescaled importance = 45.383) had a negative impact, indicating male students are more likely to have a lower first-year GPA than their female counterparts. Interestingly, the five HS curriculum variables did not rank among the top five factors within influencing first-year GPA according to the analysis from the training data set.

Table 15

Results of Linear Regression on Testing Data Set for First-Year GPA Dependent Variable

	B	β	*SE*	*t*	*p*	
Intercept	2.994		0.038	78.332	< .001	***
Student Characteristics						
Gender	-.179	-0.091	0.018	-9.707	< .001	***
Race/Ethnicity	.024	0.024	0.009	2.555	.011	*
First Generation Status	-.025	-0.009	0.027	-0.928	.353	
HS Locale	-.039	-0.041	0.009	-4.335	< .001	***
Pre-college Characteristics						
HS GPA	.337	0.347	0.014	24.260	< .001	***
Admissions Test Scores	-.018	-0.018	0.012	-1.488	.137	
AP Hours	.058	0.060	0.010	5.706	< .001	***
CLEP Hours	-.007	-0.006	0.010	-0.646	.518	
IB Hours	.016	0.017	0.008	1.876	.061	
Other Hours	-.004	-0.006	0.006	-0.606	.544	
College Prep. Curriculum	-.008	-0.009	0.009	-0.932	.351	
CM & Ready Mean	.054	0.055	0.014	3.884	< .001	***
English (CMR)	.039	0.040	0.012	3.261	.001	**
Math (CMR)	-.012	-0.012	0.012	-0.988	.323	
Science (CMR)	.002	0.002	0.010	0.176	.860	
Social Studies (CMR)	.054	0.055	0.010	5.243	< .001	***
Financial Situations						
EFC	.053	0.055	0.017	3.194	.001	**
GA HOPE Scholarship	.160	0.165	0.012	13.348	< .001	***
Zell Miller Indicator	.204	0.063	0.035	5.887	< .001	***
PELL Grant	.009	0.009	0.016	0.538	.591	
Federal Sub. Loans	.001	0.001	0.012	0.049	.961	
Federal Unsub. Loans	-.021	-0.021	0.012	-1.801	.072	
Other Loans	-.001	-0.001	0.009	-0.058	.953	
Major Groupings	-.013	-0.027	0.004	-3.052	.002	**
Institutional Expenditures						
Academic & Institutional Support	.042	0.042	0.013	3.147	.002	**
All Others	.033	0.034	0.012	2.810	.005	**
Instruction	-.026	-0.027	0.013	-1.986	.047	*
Public Service & Research	.003	0.003	0.012	0.246	.806	
Student Service Support	.020	0.021	0.018	1.702	.089	

Note. $R^2 = .312$, *adj* $R^2 = .310$, $F(29, 8{,}663) = 118.40$, $p < .001$. *** $p < .001$. ** $p < .01$. * $p < .05$. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

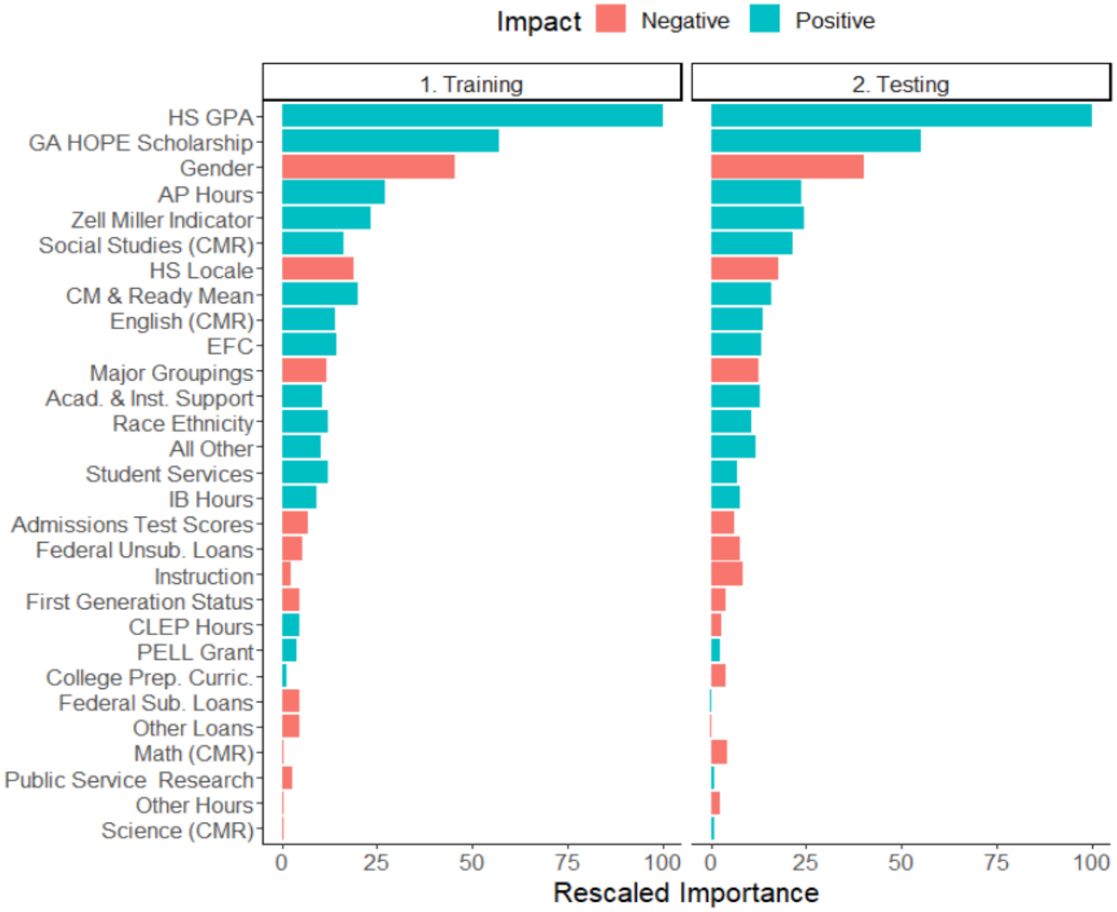


Figure 18. First-year GPA variable importance plot for the linear regression model on the training and testing data sets. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph. The color of the bar indicates whether the impact is negative or positive on the first-fall GPA. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

Examining the results of the variable importance analysis on the testing data set confirmed the three dominant influential factors were HS GPA (importance = 24.260, rescaled importance = 100.000), GA HOPE scholarship (importance = 13.348, rescaled importance = 55.021), and gender (importance = 9.707, rescaled importance = 40.014). While remaining a negative influential factor, the strength of gender’s influence on the first-year GPA lessened in the testing data set. The fourth and fifth factors changed in the testing data set (Zell Miller indicator, importance = 5.887, rescaled importance = 24.266;

AP hours, importance = 5.706, rescaled importance = 23.521). The remaining factors' contribution was very small from the analysis in the testing data set.

Support vector machine with linear kernel. The SVM algorithm using linear kernel model was built using the svm_linear() function with the engine set to kernlab and the model set to regression. The cost and margin components in the model were tuned across a grid of 20 models using a training data set. The set.seed() function was utilized for replication purposes. Despite the tuning process resulting in 20 models with similar performance, the optimal model achieved an RMSE value of 0.812 and an R^2 value of .321. Although this optimal model demonstrated the lowest RMSE, it only explained 32.1% of the variance in the data set. In this model, the cost was set to 0.304, and the margin was set to 0.194.

Figure 19 presents the results of the variable importance analysis on the training and testing data sets, where the importance values were rescaled to 100. Unlike the linear regression model, this analysis did not calculate the type of impact of the variables. HS GPA (importance = 0.129, rescaled importance = 100.000) exerted the greatest influence on the first-year GPA in the training data set. This influence suggests students with higher HS GPA correspond to a higher first-year GPA earned. Likewise, students with lower HS GPA correspond to a lower first-year GPA. Additionally, the GA HOPE scholarship (importance = 0.032, rescaled importance = 24.536) and gender (importance = 0.012, rescaled importance = 9.058) were the next most impactful factors on first-year GPA. Though with a very small influence, the content mastery and readiness mean (importance = 0.003, rescaled importance = 2.464) was ranked fifth. The remaining variables had a very small influence on the dependent variable in the training data set.

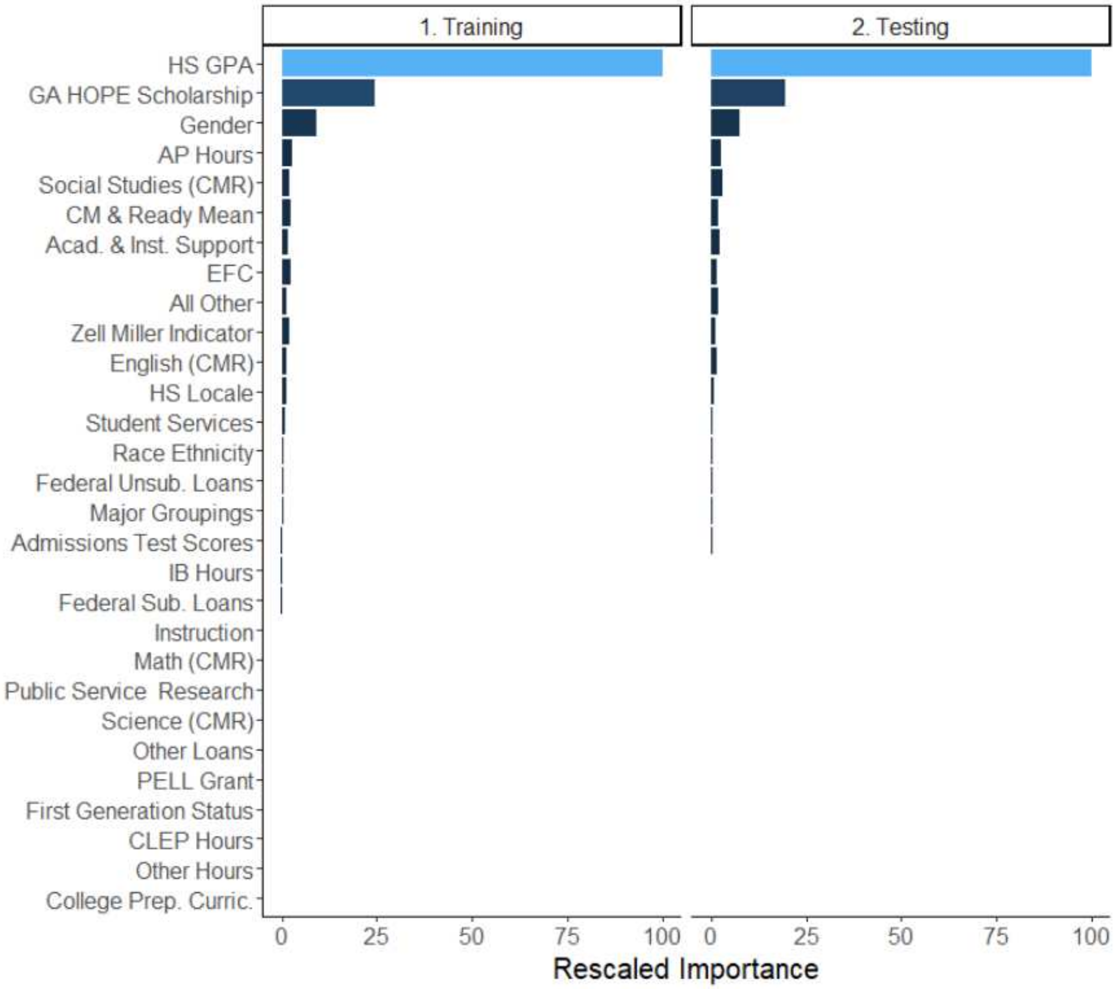


Figure 19. First-year GPA variable importance plot for the SVM model using a linear kernel on the training and testing data sets. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

For the analysis on the testing data set, HS GPA (importance = 0.133, rescaled importance = 100.000), GA HOPE scholarship (importance = 0.026, rescaled importance = 19.363), and gender (importance = 0.010, rescaled importance = 7.357) were found to be the three most influential factors. The social studies proficiency levels difference from the content mastery and readiness mean (importance = 0.004, rescaled importance = 3.038) was ranked fourth, exhibiting very small influences on the GPA.

Support vector machine with polynomial kernel. Using a polynomial kernel, another SVM algorithm was tuned employing the svm_poly() function with the kernlab engine and regression as the model type. The cost, degree, scale_factor, and margin parameters in the svm_poly() function were tuned across 20 models using training data sets, with the set.seed() function for replication purposes. The resulting optimal model achieved an RMSE value of 0.812 and an R^2 value of 0.322, explaining 32.2% of the variance within the data set. The tuned model had a cost of 14.782, a degree of 3, a scale factor of 0.001, and a margin of 0.188.

Figure 20 presents the variable importance analysis on the training and testing data sets, where the importance values were rescaled to 100 for comparison across models. From the analysis of the training data set, HS GPA (importance = 0.127, rescaled importance = 100.000) emerged as the factor with the most substantial impact on the first-year GPA. Like the previous models, students with higher HS GPA are associated with a higher first-year GPA, while students with lower HS GPA are associated with a lower first-year GPA. The GA HOPE scholarship (importance = 0.032, rescaled importance = 25.518) and gender (importance = 0.012, rescaled importance = 9.307) were the next most influential factors on first-year GPA. While exhibiting a very small influence, content mastery and readiness mean (importance = 0.003, rescaled importance = 2.607) was ranked fifth. The impact of the remaining variables was very small from the training data set.

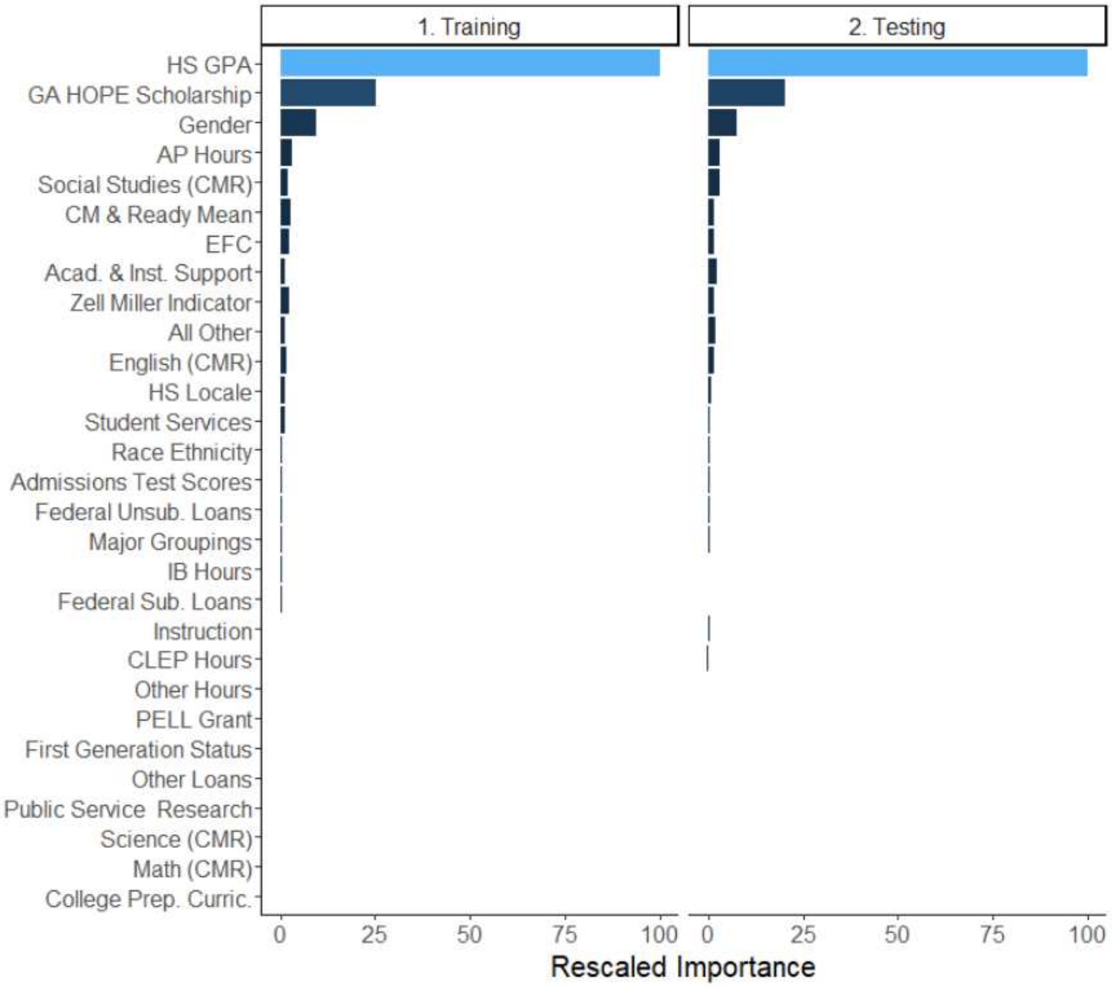


Figure 20. First-year GPA variable importance plot for the SVM model using a polynomial kernel on training and testing data sets. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

Within the testing data set, HS GPA (importance = 0.130, rescaled importance = 100.000), GA HOPE scholarship (importance = 0.026, rescaled importance = 20.248), and gender (importance = 0.010, rescaled importance = 7.693) remained the top three most influential factors. The social studies proficiency levels differences from the content mastery and readiness mean (importance = 0.004, rescaled importance = 3.037) was ranked fourth, even though the influence on the GPA is very small. Content mastery and readiness mean (importance = 0.002, rescaled importance = 1.677) fell to rank eight, exhibiting a very small influence on the first-year GPA in the testing data set.

Support vector machine with radial basis function kernel. The final SVM was built using the radial basis function kernel, with the engine set to kernlab and the model set to regression. Through a grid of 20 models developed on training data set, the cost, radial basis function sigma, and margin were tuned. The set.seed() function was utilized for replication purposes. The optimal model achieved an RMSE value of 0.812 and an R^2 value of .325, accounting for approximately 32.5% of the variance. The tuned features of the model included a cost of 19.46, sigma of 0.0005, and a margin of 0.123.

In Figure 21, the variable importance analysis results on the training and testing data sets are displayed, with the results rescaled to 100 for comparison across models. HS GPA (importance = 0.128, rescaled importance = 100.000) emerged as the most influential factor on the first-year GPA within the training data set. Additionally, the GA HOPE scholarship (importance = 0.033, rescaled importance = 24.933) and gender (importance = 0.012, rescaled importance = 9.149) were the next most impactful variables on first-year GPA. While exhibiting a small contribution, AP hours (importance = 0.006, rescaled importance = 5.639) was ranked fourth. The impact of the remaining variables within the training data set was too low to significantly affect the first-year GPA. For the results on the testing data set, HS GPA (importance = 0.133, rescaled importance = 100.000), GA HOPE scholarship (importance = 0.026, rescaled importance = 22.276), and gender (importance = 0.011, rescaled importance = 7.501) maintained their top three most influential factors. While AP hours (importance = 0.004, rescaled importance = 2.644) fell in ranking due to academic and institutional expenditures (importance = 0.004, rescaled importance = 2.880), these two factors' contribution to the GPA was very small. The remaining factors exhibited very little influence on the first-year GPA in the testing data set.

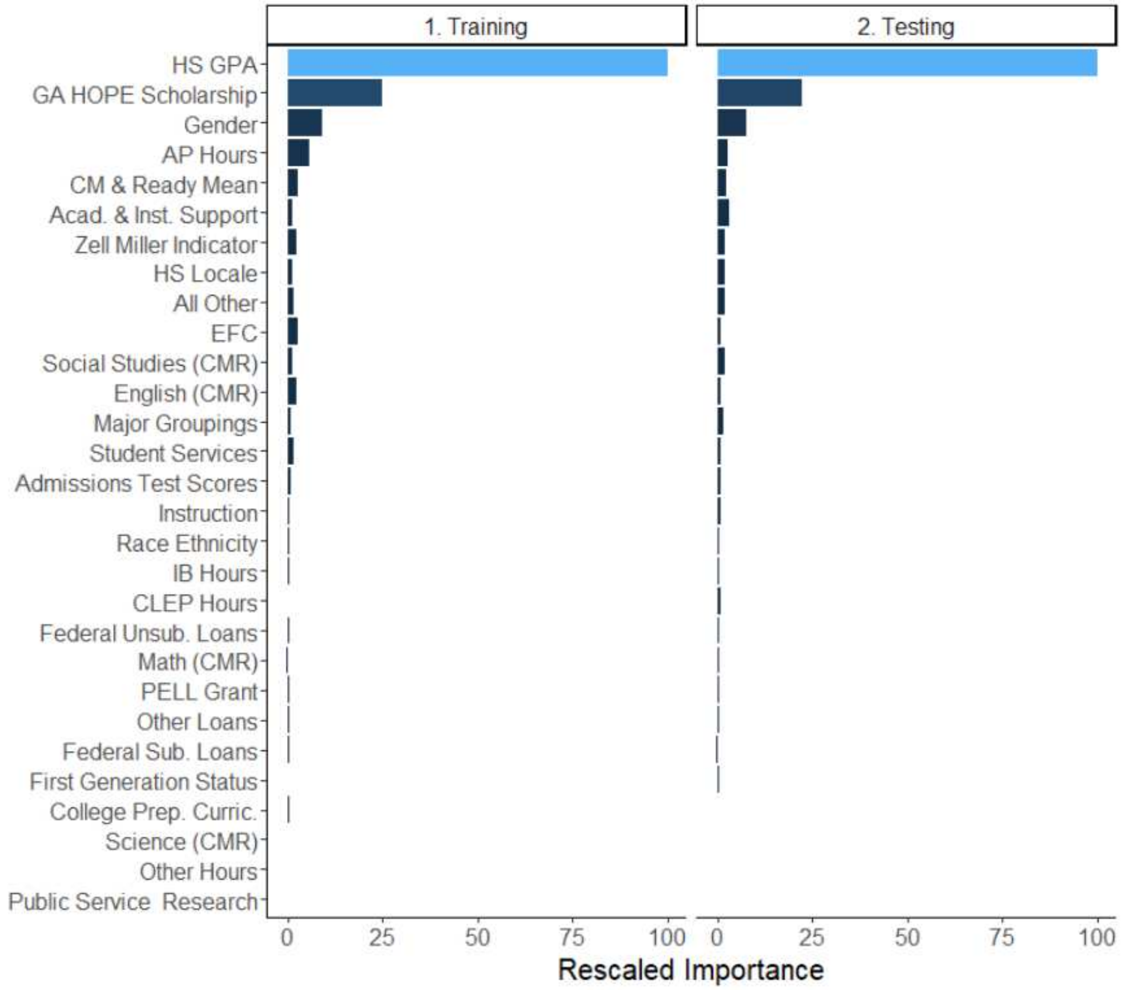


Figure 21. First-year GPA variable importance plot for the SVM model using a radial basis function kernel on the training and testing data sets. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

Random forest. The random forest algorithm was built using the rand_forest() function with the engine set to ranger and the mode set to regression. In the rand_forest() function, the mtry, trees, and min_n options were tuned to find the values for the optimal model. The model was tuned through a grid of 20 models using the training data set, with the set.seed() function for replication purposes. The best model was chosen based on the lowest RMSE value, and the optimal model exhibited an RMSE value of 0.787. The R^2 value for this model was .340, indicating the model accounts for 34.0% of the variance within the data set. The optimal model included 1,580 trees with a mtry of 9 and a minimum number of observations of 37. The optimal model's mtry of 9 is close to being one-third of the independent variables as recommended by James et al. (2013) and Kuhn and Johnson (2013).

Figure 22 illustrates the variable importance analysis results on the training and testing data sets, with the importance values rescaled to 100 for comparison across models. HS GPA (importance = 0.278, rescaled importance = 100.000) and GA HOPE scholarship (importance = 0.264, rescaled importance = 95.020) were the two major factors influencing the first-year GPA in the training data set. Content mastery and readiness mean (importance = 0.043, rescaled importance = 15.493), Zell Miller indicator (importance = 0.037, rescaled importance = 13.302), admissions test scores (importance = 0.037, rescaled importance = 13.274), and EFC (importance = 0.031, rescaled importance = 11.320) were the third through sixth factors affecting the first-year GPA. These variables exhibited a small impact, while the remaining variables in the training data set had very low impacts on the first-year GPA.

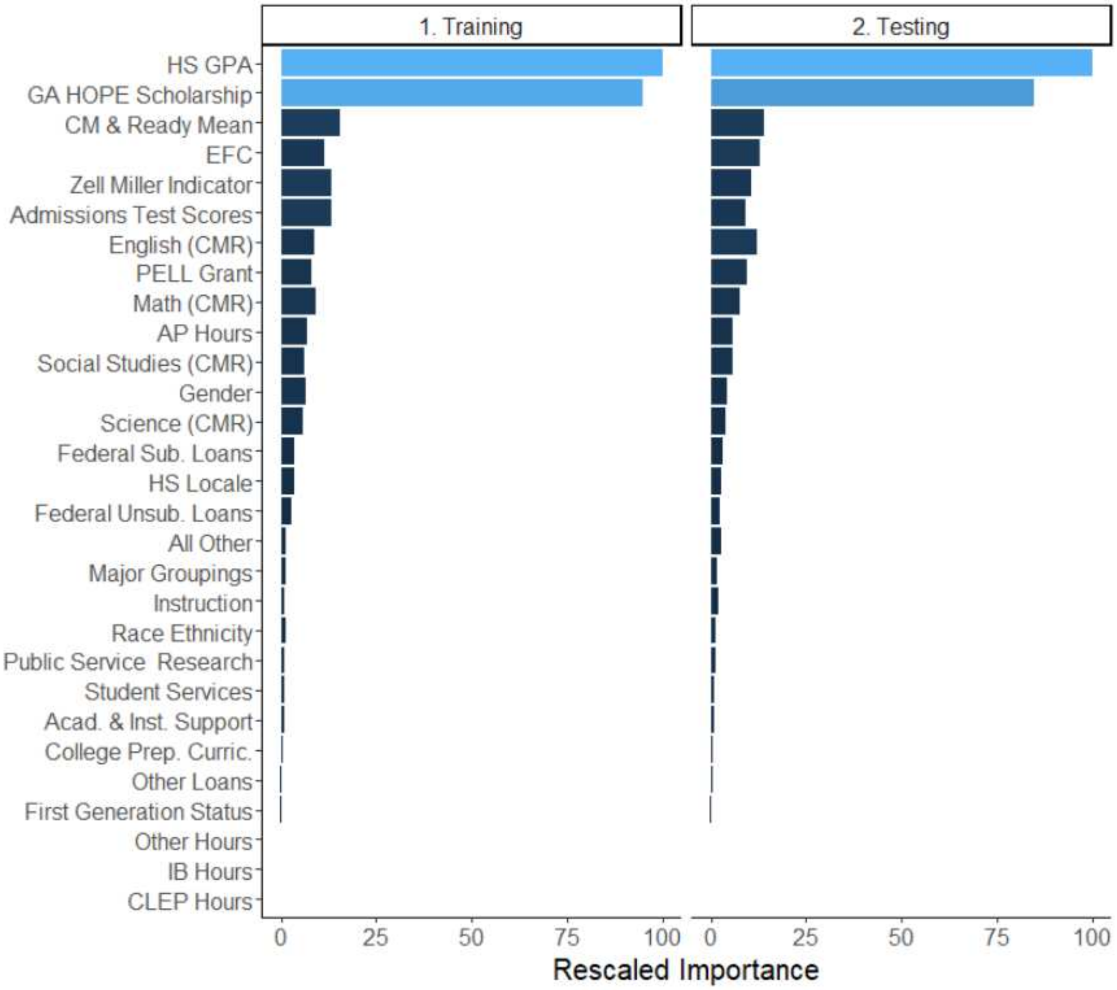


Figure 22. First-year GPA variable importance plot for the random forest model on the training and testing data sets. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

From the analysis on the testing data set, HS GPA (importance = 0.284, rescaled importance = 100.000), GA HOPE scholarship (importance = 0.241, rescaled importance = 84.674), and content mastery (importance = 0.039, rescaled importance = 13.796) remained the top three most influential factors on the first-year GPA. HS GPA and GA HOPE scholarship factors' contribution to the dependent variable was very large. While originally ranked sixth, EFC (importance = 0.037, rescaled importance = 13.022) in the testing data set was ranked fourth. The English proficiency level difference from the content mastery and readiness mean (importance = 0.035, rescaled importance = 12.276) was ranked fifth, and Zell Miller indicator (importance = 0.030, rescaled importance = 10.420) fell to the sixth ranked factor. The remaining variables' contribution to the GPA in the testing data set was very small.

Extreme gradient boosting. The XGBoost model was constructed using the boost_tree() function with the engine set to xgboost and the mode set to regression. To develop the optimal model, the trees, tree depth, min_n, loss reduction, sample size, mtry, and learn rate were tuned through a grid of 20 models using the training data set. The set.seed() function was utilized for replication purposes. The best model was selected based on the lowest RMSE value. The optimal model from the tuning exhibited an RMSE value of 0.779 and an R^2 value of 0.352, accounting for 35.2% of the variance within the data set. The optimal model consisted of 1,264 trees with a mtry of 22 and the minimum observations of 20 with a tree depth of 5. The model’s learn rate was .007, loss reduction was .004, and sample size .285.

Figure 23 displays the variable importance analysis on the training and testing data sets, with the importance values rescaled to 100 for comparison purposes across models. HS GPA (importance = 0.25, rescaled importance = 100.00) and GA HOPE scholarship (importance = 0.223, rescaled importance = 87.827) within the training data set were the top two factors impacting the first-year GPA. Ranking third through seventh, the five HS curriculum variables had a moderate influence on the GPA (English proficiency level difference, importance = 0.053, rescaled importance = 20.859; content mastery and readiness mean, importance = 0.052, rescaled importance = 20.396; social studies proficiency level difference, importance = 0.051, rescaled importance = 20.183; science proficiency level difference, importance = 0.044, rescaled importance = 17.429; and math proficiency level difference, importance = 0.044, rescaled importance = 17.150). Additionally, EFC (importance = 0.042, rescaled importance = 16.500) and admissions test scores (importance = 0.037, rescaled importance = 14.533) contribute a low to moderate contribution to the first-fall GPA.

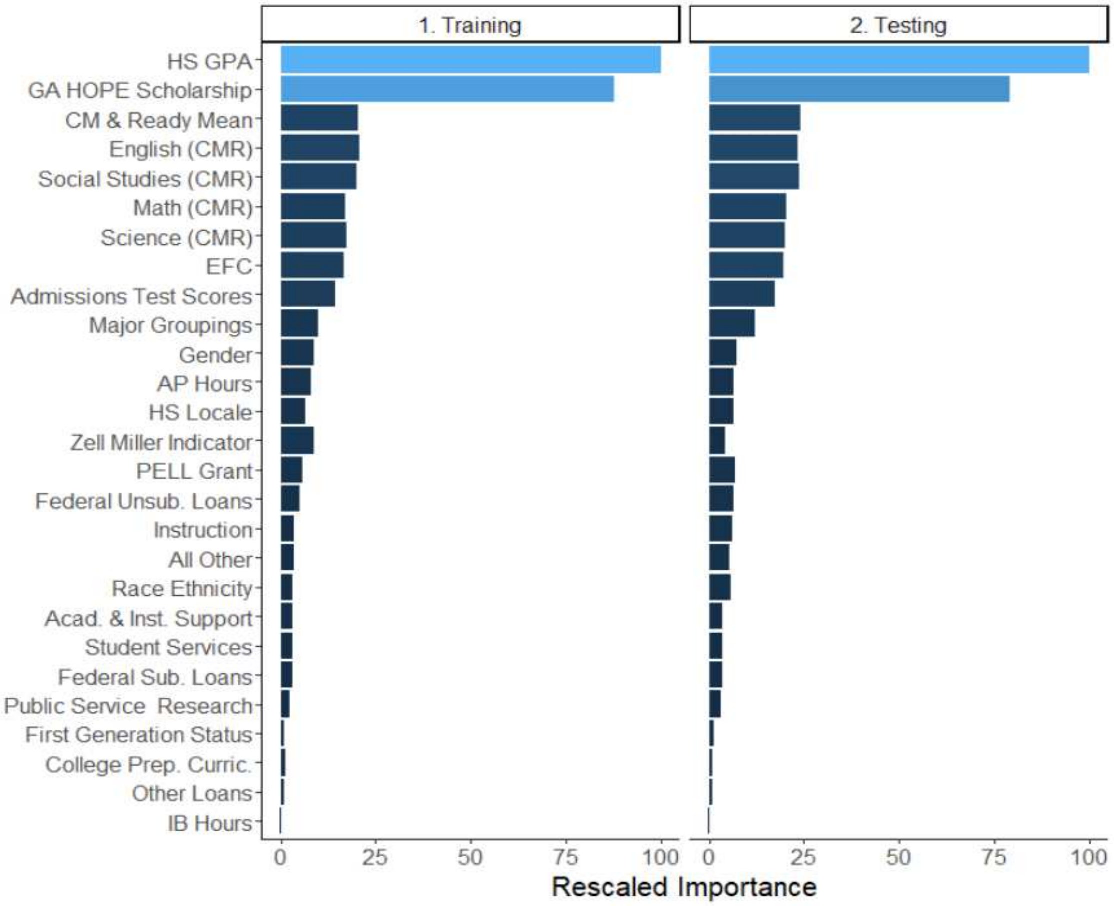


Figure 23. First-year GPA variable importance plot for the XGBoost model on the training and testing data sets. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

For the review of the results from the testing data set, HS GPA (importance = 0.243, rescaled importance = 100.000) and GA HOPE scholarship (importance = 0.193, rescaled importance = 79.220) remained the top two most influential factors on the first-year GPA. While maintaining the third through seventh ranked factors, the order of the HS curriculum variables changed. The content mastery and readiness mean (importance = 0.059, rescaled importance = 24.132) was the highest ranked HS curriculum variables. Social studies proficiency levels (importance = 0.058, rescaled importance = 23.885) was ranked fourth, English proficiency levels (importance = 0.056, rescaled importance = 23.225) was ranked fifth, math proficiency levels (importance = 0.049, rescaled importance = 20.362) was ranked sixth, and science proficiency levels (importance = 0.048, rescaled importance = 19.852) was ranked seventh. Furthermore, EFC (importance = 0.048, rescaled importance = 19.685) and admissions tests scores (importance = 0.042, rescaled importance = 17.327) held their rank, while major groupings' (importance = 0.030, rescaled importance = 12.173) contribution increased as the tenth ranked factor.

Variable importance comparison for first-year GPA. The patterns emerged in the training data set changed slightly when the results of the variable importance analysis were conducted on the testing data set. Figure 24 illustrates a comparison of the variable importance analysis on the testing data set across each model. The HS GPA and GA HOPE scholarship factors remained as the top two factors influencing the first-year GPA. HS GPA in each model remained the predominant factor. GA HOPE scholarship's largest influence was found to be in the random forest model, with the XGBoost model having just a slightly lesser effect. No additional factors were consistent in their rankings.

Gender in the linear regression and SVMs found gender to be the third rank factor, while content mastery and readiness mean was the third rank factor in the random forest and XGBoost models.

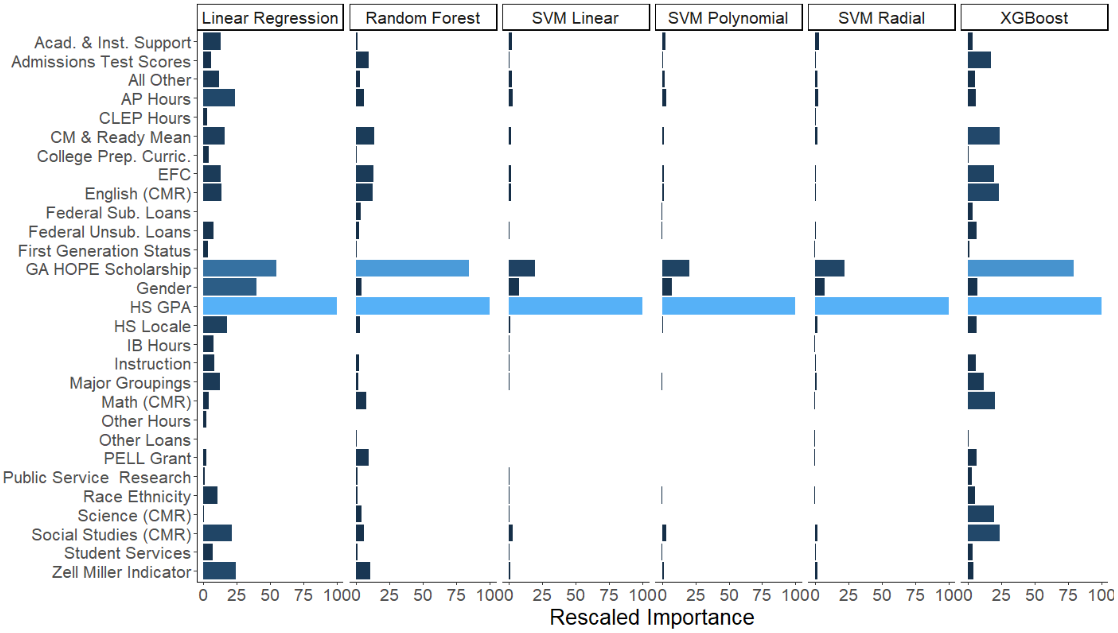


Figure 24. Comparison of variable importance results on the testing data set for first-year GPA. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

For the five HS curriculum factors, the content mastery and readiness mean factor was the highest of the five in the random forest, XGBoost, and the SVM using a radial basis function. The social studies proficiency level difference from the content mastery and readiness mean was ranked the top HS curriculum factor for the linear regression, SVM using a linear kernel, and SVM using a polynomial kernel. Of the non-HS curriculum variables, the number of AP hours was ranked over the admissions test scores in the linear regression and the three SVM models. Random forest and XGBoost models ranked admissions test scores as the top non-HS curriculum factors for the pre-college characteristics. The remaining advanced standing hours and the number of satisfactory college preparatory curriculum requirements exhibited very little contribution to the first-year GPA.

Excluding the GA HOPE scholarship, the EFC was the top financial situation factor in the random forest and XGBoost models, while the Zell Miller factor was the top for the linear regression and the three SVM models. PELL grant factor was also one of the top factors within the financial situations in the random forest and XGBoost models. The remaining financial situation factors exhibited little contribution to the first-year GPA. Likewise, the selected major and remaining expenditures exhibited very little influence on the GPA.

One-year retention. The training data set, comprising 13,078 observations, was utilized to develop predictive algorithms for assessing the likelihood of a student not retaining at the initial institution. To address potential data imbalances, two additional methods were employed: downsampling the majority class and upsampling the minority class. The downsampling and upsampling techniques aimed to avoid automatic bias toward the majority class caused by an imbalance in observations. Six statistical models were developed to analyze the data. The models consisted of logistic regression, three SVM models, random forest, and XGBoost. The kernels utilized in the SVM models were the linear, polynomial, and radial basis function kernels. The logistic regression algorithm did not need to be tuned. The three SVM models, random forest, and XGBoost models were tuned using a grid of 20 models based on 10-fold cross-validation of the training data set. To ensure the replicability of the algorithms, the set.seed() functions were utilized for the models. A tuning grid allowed the models' parameters to be optimized for the best performance based on the training data set. The area under the curve (AUC) served as the ultimate metric for selecting the optimal model in predicting the probability of not being retained at the institution, as recommended by Dey (2021). The highest AUC value indicated the optimal model. After the predictive algorithms were developed, all six models analyzed the training and testing data sets to identify the factors influencing the one-year retention status, with the emphasis placed on the results from the testing data set.

Logistic regression. Under the umbrella of GLM, both linear regression and logistic regression serve as essential tools. However, logistic regression distinguishes itself from linear regression by handling dichotomous dependent variables. The logistic regression model was constructed using the logistic_reg() function, utilizing the glm engine and classification mode. No tuning was required to achieve an optimal model. The model's outcomes are presented in Table 16. The Hosmer-Lemeshow's goodness-of-fit test yielded, $\chi^2(8) = 12.189$, $p = .143$, was found to be not significant, indicating the model was a good fit. The AIC value was 13,270 and BIC value was 13,494. McFadden's pseudo-R^2 was calculated at .047, indicating the model accounted for 4.7% of the variance in the data set, while Nagelkerke's pseudo-R^2 stood at .074, signifying 7.4% of the variance was explained by the model.

Table 16

Logistic Regression Results on Training Data Set for One-year Retention Dependent Variable

							95% CI	
	B	*SE*	*z*	*p*		*OR*	*LL*	*UL*
Intercept	-1.701	0.098	-17.356	< .001	***	0.182	0.150	0.221
Student Characteristics								
Gender	.281	0.045	6.206	< .001	***	1.325	1.212	1.448
Race/Ethnicity	-.070	0.024	-2.888	.004	**	0.932	0.889	0.978
First Generation Status	.101	0.065	1.547	.122		1.106	0.973	1.257
HS Locale	.104	0.023	4.587	< .001	***	1.109	1.061	1.160
Pre-college Characteristics								
HS GPA	-.243	0.036	-6.751	< .001	***	0.784	0.731	0.842
Admissions Test Scores	.169	0.029	5.730	< .001	***	1.184	1.118	1.254
AP Hours	-.057	0.027	-2.149	.032	*	0.945	0.897	0.995
CLEP Hours	.011	0.021	0.523	.601		1.011	0.964	1.051
IB Hours	-.011	0.025	-0.445	.656		0.989	0.938	1.034
Other Hours	-.018	0.029	-0.629	.529		0.982	0.911	1.028
College Prep. Curriculum	-.045	0.023	-1.988	.047	*	0.956	0.914	1.000
CM & Ready Mean	-.106	0.034	-3.098	.002	**	0.900	0.842	0.962
English (CMR)	-.006	0.030	-0.186	.852		0.995	0.938	1.054
Math (CMR)	.052	0.030	1.737	.082		1.053	0.994	1.116
Science (CMR)	-.015	0.026	-0.599	.549		0.985	0.936	1.036
Social Studies (CMR)	.014	0.025	0.538	.590		1.014	0.965	1.065
Financial Situations								
EFC	-.158	0.043	-3.707	< .001	***	0.854	0.786	0.928
GA HOPE Scholarship	-.252	0.029	-8.795	< .001	***	0.777	0.735	0.822
Zell Miller Indicator	.070	0.090	0.775	.438		1.072	0.898	1.279
PELL Grant	-.062	0.041	-1.521	.128		0.940	0.868	1.018
Federal Sub. Loans	.014	0.031	0.444	.657		1.014	0.953	1.079
Federal Unsub. Loans	-.003	0.029	-0.093	.926		0.997	0.942	1.055
Other Loans	-.016	0.022	-0.723	.470		0.984	0.943	1.027
Major Groupings	.020	0.011	1.775	.076		1.020	0.998	1.042
Institutional Expenditures								
Academic & Inst. Support	.019	0.033	0.564	.573		1.019	0.955	1.087
All Others	-.002	0.029	-0.086	.931		0.998	0.942	1.056
Instruction	.042	0.033	1.270	.204		1.043	0.978	1.112
Public Service & Research	-.121	0.030	-3.955	< .001	***	0.886	0.835	0.941
Student Service Support	.012	0.030	0.420	.674		1.012	0.956	1.073

Note. AIC = 13,270, BIC = 13,494, McFadden pseudo-R^2 = .047, and Nagelkerke pseudo-R^2 = .074. *** $p < .001$. *** $p < .001$. ** $p < .01$. * $p < .05$. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution. Inst. = institutional.

A total of 11 of the 29 independent variables were found to be significant in predicting the student's likelihood of not retaining. Of the student characteristics, gender (*B* = .281, *z* = 6.206, *p* < .001, *OR* = 1.325, 95% CI [1.212, 1.448]), race/ethnicity (*B* = -.070, *z* = -2.888, *p* = 0.004, *OR* = 0.932, 95% CI [0.889, 0.978]), and HS locale (*B* = .104, *z* = 4.587, *p* < .001, *OR* = 1.109, 95% CI [1.061, 1.160]) were found to be significant. Male students exhibited a departure likelihood 1.325 times higher than their female counterparts. Furthermore, graduates from HS located in towns or rural areas had a departure likelihood 1.109 times greater than those from urban or suburban high schools. Intriguingly, students identified as Hispanic or placed into the Other grouping showed a decreased likelihood of departure in comparison to their White or Black and African American peers (*OR* = 0.932). However, it is noteworthy the odds of locale of the HS and race and ethnicity on non-retention was nearly equal to one, suggesting these two variables did not significantly impact the decision to depart.

Five of the pre-college characteristics were found to be significant (HS GPA, *B* = -.243, *z* = -6.751, *p* < .001, *OR* = 0.784, 95% CI [0.731, 0.842]; admissions test scores, *B* = .169, *z* = 5.730, *p* < .001, *OR* = 1.184, 95% CI [1.118, 1.254]; AP Hours, *B* = -.057, *z* = -2.149, *p* = .032, *OR* = 0.945, 95% CI [0.897, 0.995]; college preparatory curriculum, *B* = -.045, *z* = -1.988, *p* = .047, *OR* = 0.956, 95% CI [0.914, 1.000]; and content mastery and readiness mean, *B* = -.106, *z* = -3.098, *p* = .002, *OR* = 0.900, 95% CI [0.842, 0.963]). Students with higher HS GPAs were 0.784 times less likely to depart from the institution compared to their peers with lower HS GPAs. This finding implies for every unit increase in HS GPA, the odds of departing decreased by 21.6%. In practical terms, students who performed well academically in HS had a significantly reduced likelihood of departing from the institution. In contrast, students with higher admission test scores were 1.184 times more likely to depart than those with lower test scores. This means for every unit increase in admission test scores, the odds of departing increased by 18.4%. Surprisingly, higher test scores did not correlate with greater retention; instead, students with comparatively lower test scores exhibited a higher likelihood of staying enrolled. It is notable the odds ratio for the test scores was close to one, suggesting the variable's impact was limited. Additionally, students who earned AP hours (*OR* = 0.945) or successfully fulfilled all aspects of the CPC (*OR* = 0.956) exhibited slightly lower odds of departing from the institution.

The expected family contribution (*B* = -.158, *z* = -3.707, *p* < .001, *OR* = 0.854, 95% CI [0.786, 0.928]) and GA HOPE scholarship (*B* = -.252, *z* = -8.795, *p* < .001, *OR* = 0.777, 95% CI [0.735, 0.822]) were the only financial situation variables found to be significant. Students with higher expected family contributions were found to be 0.854 times less likely to depart from the institution. This finding means for every unit increase in the expected family contribution the odds of departure decreased by 14.6%. This finding suggests students from families with higher financial resources might have greater stability and support, reducing the chances of departure. Similarly, students who received the GA HOPE scholarship demonstrated a departure likelihood 0.777 times lower than their counterparts. In essence, students benefiting from the scholarship program had a 22.3% reduction in odds of departing from the institution compared to students without the scholarship. The GA HOPE scholarship, serving as a significant financial aid initiative, played a substantial role in retaining students within the academic environment. Only the public and research expenditures, public and research support, *B*

= -.121, z = -3.955, p < .001, OR = 0.886, 95% CI [0.835, 0.941], were found to be significant.

Assumptions for logistic regression. The first assumption of logistic regression necessitates the classification of the dependent variable as dichotomous, with two distinct values: zero denoting retained students and one indicating non-retained students. During the preprocessing stage, a thorough examination of univariate and multivariate outliers was conducted. Although certain variables were initially flagged as outliers, upon closer examination, the values were deemed acceptable. The correlation values between independent variables and the log odds of non-retention probabilities were generated and displayed Figure 25. Surprisingly, among the 29 independent variables, four exhibited no linear relationship with the log odds of the predicted probabilities, indicating a lack influence. GA HOPE scholarship (r(13,078) = -.762, p < .001) and HS GPA (r(13,078) = -.708, p < .001) have the two highest correlations to the log odds of the predicted probabilities. Both variables have a very strong negative relationship, in which the higher the value of the independent factor the more likely the student will retain rather than depart from the institution. Gender, r(13,078) = .375, p < .001, exhibited a positive negative relationship with the log odds of the predicted values. Public and research expenditures (r(13,078) = -.308, p < .001), AP hours (r(13,078) = -.274, p < .001), expected family contribution (r(13,078) = -.264, p < .001), EFC (r(13,078) = -.262, p < .001), instruction expenditures (r(13,078) = -.224, p < .001), and student support service expenditures (r(13,078) = -.205, p < .001) have weak negative relationships with the log odds of the predicted values. The remaining variables either exhibited a very weak or no relationship with the log odds of the predicted probabilities. Additionally, multicollinearity was investigated, and all variables demonstrated minimal levels of multicollinearity, adhering to the specified threshold value of five. The low levels multicollinearity was the result of the data cleanup during the preprocessing stage, effectively rectifying any pre-existing multicollinearity issues.

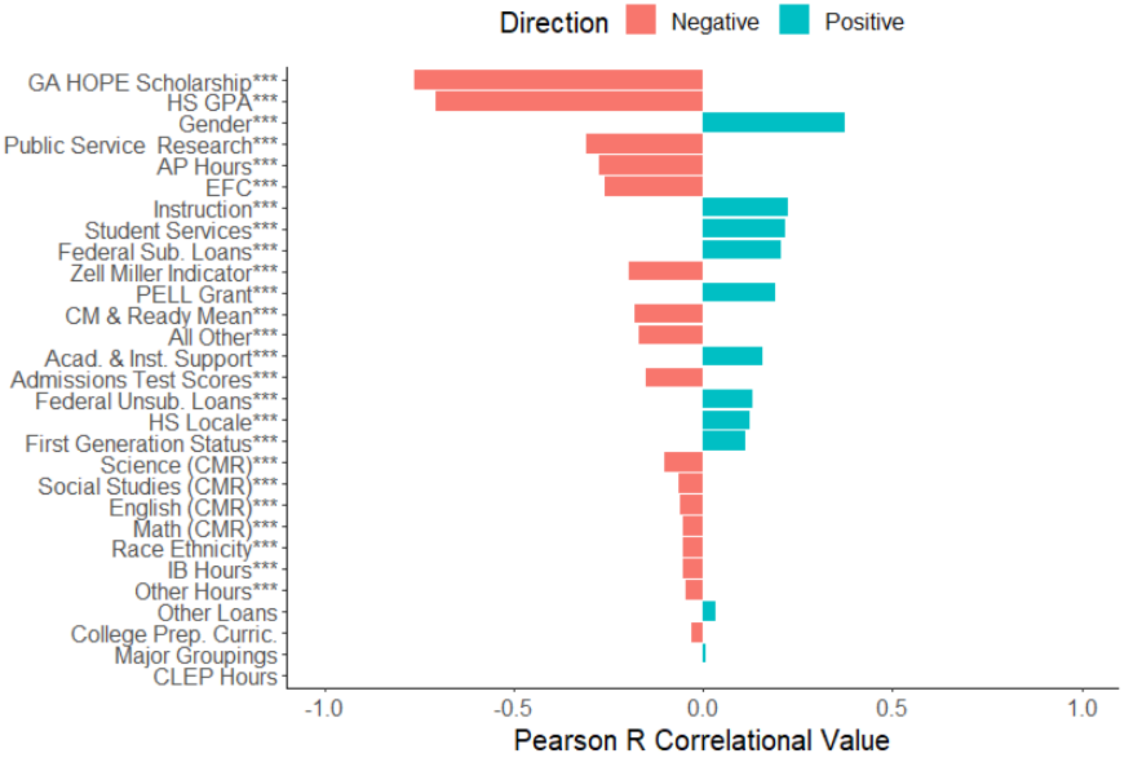


Figure 25. Correlational analysis to the log odds of the predicted probabilities of not retaining. The plot displays the correlation of the independent variables with the log odds of the predicted probabilities of not retaining. The color of the bar indicates whether the impact is negative or positive on the retention status. *** p < .001. *** p < .001. ** p < .01. * p < .05. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

Sampling modifications. The training data set underwent a downsampling techniques facilitated by the step_downsample() function, which led to a reduction in the majority class. This downsizing aimed to mitigate any significant impact of class imbalances to prevent bias towards the majority class. The original data set, comprising

13,078 observations, was reduced to 5,812 observations. Subsequently, a logistic regression model was constructed on the downsample training data set. The model was developed using the logistic_reg() function, with the engine set to glm and the mode configured for classification. The model did not require any tuning to achieve its optimal state. The outcomes of this logistic regression model are presented in Table 17. The Hosmer-Lemeshow's goodness-of-fit, $\chi^2(8) = 6.985$, $p = .538$, was found to be not significant indicating the downsample model was a good fit. The downsample model's AIC is 7,707 and BIC is 7,907. The McFadden's pseudo-R^2 was .051 and Nagelkerke's pseudo-R^2 was .091. McFadden's pseudo-R^2 value indicated 5.1% and Nagelkerke's pseudo-R^2 indicated 9.1% of the variance was accounted for within the data set.

Table 17

Logistic Regression Results on Training Data Set with Downsampling Techniques for One-year Retention Dependent Variable

							95% CI	
	B	*SE*	*z*	*p*		*OR*	*LL*	*UL*
Intercept	-.450	0.123	-3.664	< .001	***	0.638	0.501	0.811
Student Characteristics								
Gender	.285	0.057	4.964	< .001	***	1.330	1.188	1.489
Race/Ethnicity	-.090	0.029	-3.064	.002	**	0.914	0.862	0.968
First Generation Status	.056	0.083	0.674	.500		1.058	0.899	1.245
HS Locale	.100	0.028	3.530	< .001	***	1.106	1.046	1.169
Pre-college Characteristics								
HS GPA	-.295	0.045	-6.638	< .001	***	0.744	0.682	0.812
Admissions Test Scores	.155	0.037	4.179	< .001	***	1.168	1.086	1.256
AP Hours	-.033	0.033	-1.005	.315		0.968	0.907	1.032
CLEP Hours	.000	0.023	0.006	.995		1.000	0.950	1.048
IB Hours	-.007	0.030	-0.244	.807		0.993	0.934	1.052
Other Hours	-.031	0.036	-0.868	.385		0.970	0.890	1.028
College Prep. Curriculum	-.069	0.030	-2.294	.022	*	0.933	0.879	0.990
CM & Ready Mean	-.086	0.043	-2.020	.043	*	0.917	0.844	0.997
English (CMR)	-.065	0.037	-1.748	.080		0.937	0.870	1.008
Math (CMR)	.069	0.037	1.857	.063		1.071	0.996	1.152
Science (CMR)	.011	0.032	0.340	.734		1.011	0.949	1.077
Social Studies (CMR)	.012	0.032	0.384	.701		1.012	0.951	1.077
Financial Situations								
EFC	-.175	0.053	-3.278	.001	**	0.839	0.756	0.932
GA HOPE Scholarship	-.226	0.036	-6.261	< .001	***	0.797	0.743	0.856
Zell Miller Indicator	.131	0.109	1.205	.228		1.140	0.921	1.411
PELL Grant	-.065	0.051	-1.284	.199		0.937	0.848	1.035
Federal Sub. Loans	.030	0.039	0.763	.445		1.030	0.954	1.113
Federal Unsub. Loans	-.015	0.036	-0.408	.683		0.985	0.918	1.058
Other Loans	-.020	0.028	-0.708	.479		0.980	0.928	1.036
Major Groupings	.029	0.014	2.062	.039	*	1.029	1.001	1.058
Institutional Expenditures								
Academic & Inst. Support	.051	0.042	1.209	.227		1.052	0.969	1.142
All Others	.020	0.037	0.555	.579		1.021	0.950	1.097
Instruction	.059	0.041	1.437	.151		1.061	0.979	1.150
Public Service & Research	-.079	0.038	-2.074	.038	*	0.924	0.857	0.996
Student Service Support	.005	0.037	0.130	.897		1.005	0.934	1.080

Note. AIC = 7,707, BIC = 7,907, McFadden pseudo-R^2 = .051, and Nagelkerke pseudo-R^2 = .091. *** $p < .001$. *** $p < .001$. ** $p < .01$. * $p < .05$. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution. Inst. = institutional.

The results of the downsample model indicated 11 of the 29 independent variables were significant. From the student characteristic variables, gender (B = .285, z = 4.964, p

< .001, *OR* = 1.330, 95% CI [1.188, 1.489]), race and ethnicity (*B* = -.090, *z* = -3.064, *p* = 0.002, *OR* = 0.914, 95% CI [0.862, 0.968]), and HS locale (*B* = .100, *z* = 3.530, *p* < .001, *OR* = 1.106, 95% CI [1.046, 1.169]) were found to be significant. From the downsample model, male students were 1.330 times more likely to not retain than their female counterparts. While significant, race and ethnicity played a small factor in the departure decisions as White or Black and African American students were more likely to depart than their counterparts identified as Hispanic or other underrepresented minorities. Only HS GPA (*B* = -.295, *z* = -6.638, *p* < .001, OR = 0.744, 95% CI [0.682, 0.812]), admissions test scores (*B* = .155, *z* = 4.179, *p* < .001, OR = 1.1.168, 95% CI [1.086, 1.256]), college preparatory curriculum (*B* = -.069, *z* = -2.294, *p* = 0.022, OR = 0.933, 95% CI [0.879, 0.990]), and content mastery and readiness mean (*B* = -.086, *z* = --2.020, *p* = .043, OR = 0.917, 95% CI [0.844, 0.997]) were found to be significant from the pre-college characteristics. The downsample model also indicated the higher the students HS GPA were less likely not to depart than those with lower HS GPA. Students with higher HS GPA were 0.744 times less likely to depart, and with every unit increase in the HS GPA the odds of departing decreased by 25.6%. Higher test scores did not correlate with greater retention as the odds ratio was 1.168, meaning with every unit increase of the admissions test scores the odds of departing increased by 16.8%. Alternatively, students with comparatively lower test scores exhibited a higher likelihood of staying enrolled.

Two of the financial situation variables were found to be significant (expected family contribution, *B* = -.175, *z* = -3.278, *p* = .001, *OR* = 0.839, 95% CI [0.756, 0.932]; and GA HOPE scholarship, *B* = -.226, *z* = -6.261, *p* < .001, *OR* = 0.797, 95% CI [0.743, 0.856]). Students with higher expected family contributions were found to be 0.839

times less likely to depart from the institution. This finding means for every unit increase in the expected family contribution, the odds of departure decreased by 16.1%. Similarly, students who received the GA HOPE scholarship demonstrated a departure likelihood 0.798 times lower than their counterparts. Students benefiting from the scholarship program had a 20.3% reduction in odds of departing from the institution compared to students without the scholarship. The major grouping of the programs of study, *B* = .029, *z* = 2.062, *p* = .039, *OR* = 1.029, 95% CI [1.001, 1.058], was found to be significant. Only the public and research expenditures, public service and research expenditures, *B* = -.079, *z* = -2.074, *p* = .038, *OR* = 0.924, 95% CI [0.857, 0.996], were found to be significant.

Lastly, the training data set underwent an upsampling technique facilitated by the step_upsample() function, which led to an increase in the minority class. The upsizing aimed to mitigate any significant impact of class imbalances to prevent bias towards the majority class. The original 13,078 observations were increased to 20,344 observations. The logistic regression model using the upsample training data set was built utilizing the logistic_reg() function with the engine set to glm and the mode set to classification. No retuning was needed to achieve the optimal model. The results of the logistic regression model utilizing the upsample are displayed in Table 18. The Hosmer-Lemeshow's goodness-of-fit, $\chi^2(8) = 17.555$, p = .025, was found to be significant indicating the upsample model was poor fit. The upsample model's AIC was 26,860 and BIC was 27,097. The AIC and BIC values for the upsample were higher than the regular model indicating a poor fit. The McFadden's pseudo-R^2 was .050 and Nagelkerke's pseudo-R^2

was .089. McFadden's pseudo-R^2 value indicated only 5.0% and Nagelkerke's pseudo-R^2 indicated only 8.9% of the variance was accounted for by the upsample model.

Table 18

Logistic Regression Results on Data Set with Upsampling Techniques for One-year Retention Dependent Variable

							95% CI	
	B	*SE*	*z*	*p*		*OR*	*LL*	*UL*
Intercept	-.463	0.065	-7.125	< .001	***	0.629	0.554	0.715
Student Characteristics								
Gender	.319	0.030	10.465	< .001	***	1.376	1.296	1.461
Race/Ethnicity	-.069	0.016	-4.360	< .001	***	0.934	0.905	0.963
First Generation Status	.078	0.044	1.758	.079		1.081	0.991	1.180
HS Locale	.099	0.015	6.527	< .001	***	1.104	1.072	1.137
Pre-college Characteristics								
HS GPA	-.250	0.024	-10.558	< .001	***	0.778	0.743	0.815
Admissions Test Scores	.145	0.020	7.365	< .001	***	1.156	1.112	1.202
AP Hours	-.069	0.018	-3.920	< .001	***	0.933	0.902	0.966
CLEP Hours	.005	0.013	0.415	.679		1.005	0.979	1.032
IB Hours	-.028	0.017	-1.665	.096		0.973	0.940	1.004
Other Hours	-.032	0.021	-1.541	.123		0.968	0.923	1.004
College Prep. Curriculum	-.041	0.016	-2.636	.008	**	0.960	0.931	0.990
CM & Ready Mean	-.096	0.023	-4.158	< .001	***	0.909	0.868	0.951
English (CMR)	-.010	0.020	-0.486	.627		0.990	0.952	1.030
Math (CMR)	.036	0.020	1.808	.071		1.037	0.997	1.078
Science (CMR)	-.006	0.017	-0.335	.738		0.994	0.961	1.029
Social Studies (CMR)	.011	0.017	0.649	.517		1.011	0.978	1.045
Financial Situations								
EFC	-.165	0.029	-5.694	< .001	***	0.848	0.802	0.898
GA HOPE Scholarship	-.236	0.019	-12.190	< .001	***	0.790	0.760	0.820
Zell Miller Indicator	.114	0.059	1.945	.052		1.121	0.999	1.257
PELL Grant	-.073	0.027	-2.675	.007	**	0.930	0.881	0.981
Federal Sub. Loans	.035	0.021	1.690	.091		1.036	0.994	1.079
Federal Unsub. Loans	-.023	0.019	-1.192	.233		0.977	0.941	1.015
Other Loans	-.013	0.015	-0.863	.388		0.987	0.959	1.016
Major Groupings	.021	0.007	2.819	.005	**	1.021	1.006	1.036
Institutional Expenditures								
Academic & Inst. Support	.040	0.022	1.771	.077		1.040	0.996	1.087
All Others	-.017	0.020	-0.849	.396		0.983	0.946	1.022
Instruction	.026	0.022	1.193	.233		1.026	0.983	1.071
Public Service & Research	-.113	0.020	-5.526	< .001	***	0.893	0.858	0.930
Student Service Support	.000	0.020	0.004	.997		1.000	0.962	1.040

Note. AIC = 26,793, BIC = 27,030, McFadden pseudo-R^2 = .052, and Nagelkerke pseudo-R^2 = .093. *** $p < .001$. ** $p < .01$. * $p < .05$. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution. Inst. = institutional.

A total of 13 variables were found to be significant out of 29 factors. Three of the four student characteristic variables were found to be significant (gender, B =.319, z = 10.465, p < .001, OR = 1.376, 95% CI [1.296, 1.461]; race and ethnicity, B = -.069, z = -4.360, p < .001, OR = 0.934, 95% CI [0.905, 0.963]; and HS locale, B = .099, z = 6.527, p < .001, OR = 1.104, 95% CI [1.072, 1.137]). Of the student characteristics for the upsample mode, student's gender was the factor with the most contribution to the departure decision. Male students were 1.376 times more likely to depart than their female counterparts. While found to be significant, students' race and ethnicity and HS locale the odds ratios were close to one indicating these three factors were not major contributors to the departure decision.

Within the pre-college characteristics, the HS GPA (B = -.250, z = -10.558, p < .001, OR = 0.778, 95% CI [0.743, 0.815]) and admissions test scores (B = .145, z = 7.365, p < .001, OR = 1.156, 95% CI [1.112, 1.202]) were found to be significant. The AP hours, B = -069, z = -3.920, p < .001, OR = 0.933, 95% CI [0.902, 0.966] variable was the only advance hours found to be significant. The number of college preparatory curriculum satisfied, B = -041, z = -2.636, p = .008, OR = 0.960, 95% CI [0.931, 0.990], was found to be significant. The last pre-college characteristics found to be significant was the content mastery and readiness mean, B = -096, z = -4.158, p < .001, OR = 0.909, 95% CI [0.868, 0.951], variable. HS GPA was the variable with the highest contribution in the departure decisions. Students with a higher HS GPA are 0.778 times less likely to departure than those with a lower HS GPA, and with every unit increase in the HS GPA the odds of departing decreased by 22.2%. The odds ratio of the admissions test scores indicated students with higher scores were 1.156 times more likely to depart, meaning with every unit increase of the admissions test scores the odds of departing increased by 15.6%. The expected family contribution (*B* = -.165, *z* = -5694, *p* < .001, *OR* = 0.848, 95% CI [0.802, 0.898]), GA HOPE scholarship (*B* = -.236, *z* = -12.1980, *p* < .001, *OR* = 0.790, 95% CI [0.760, 0.80]), and PELL grant (*B* = -.073, *z* = -2.675, *p* = .007, *OR* = 0.930, 95% CI [0.881, 0.981]) were found to be significant. Students with higher expected family contributions were found to be 0.848 times less likely to depart from the institution. This finding means for every unit increase in the EFC the odds of departure decreased by 15.2%. Similarly, students who received the GA HOPE scholarship demonstrated a departure likelihood 0.790 times lower than their counterparts. Students benefiting from the scholarship program had a 21.0% reduction in odds of departing from the institution compared to students without the scholarship. Unlike the original and downsample model, PELL grant was found to be significant, and the odds were .933 times less likely to depart than those who do not receive the grant. Students benefiting from the grant had a 7.0% reduction in the odds with every unit increase in the PELL grant. The major grouping of the program of study (*B* = .021, *z* = 2.819, *p* = .005, *OR* = 1.021, 95% CI [1.006, 1.036]) and public service and research expenditures (*B* = -.113, *z* = -5.526, *p* < .001, *OR* = 0.893, 95% CI [0.858, 0.930]) were found to be significant.

Testing data set. The results of the logistic regression model on the testing data set are displayed in Table 19. The Hosmer-Lemeshow's goodness-of-fit test yielded, $\chi^2(8) = 11.158$, *p* = .193, was found to be not significant, indicating the model was a good fit. The AIC value was 8,948 and BIC value was 9,160. McFadden's pseudo-R^2 was calculated at .048, indicating the model accounted for 4.8% of the variance in the data set, while Nagelkerke's pseudo-R^2 stood at .077, signifying 7.7% of the variance was explained by the model. Ten of the 29 variables were found to be significant.

Table 19

Logistic Regression Results on Testing Data Set for One-year Retention Dependent Variable

							95% CI	
	B	*SE*	*z*	*p*		*OR*	*LL*	*UL*
Intercept	-1.400	0.117	-11.968	< .001	***	0.247	0.196	0.310
Student Characteristics								
Gender	.185	0.056	3.327	.001	**	1.204	1.079	1.342
Race/Ethnicity	-.069	0.029	-2.342	.019	*	0.933	0.881	0.988
First Generation Status	.199	0.080	2.496	.013	*	1.221	1.043	1.426
HS Locale	.037	0.027	1.349	.177		1.038	0.983	1.095
Pre-college Characteristics								
HS GPA	-.228	0.043	-5.351	< .001	***	0.796	0.732	0.865
Admissions Test Scores	.191	0.036	5.357	< .001	***	1.211	1.129	1.299
AP Hours	-.081	0.033	-2.500	.012	*	0.922	0.865	0.982
CLEP Hours	.036	0.029	1.232	.218		1.036	0.976	1.096
IB Hours	-.061	0.037	-1.664	.096		0.941	0.866	1.002
Other Hours	.028	0.016	1.764	.078		1.029	0.995	1.062
College Prep. Curriculum	.002	0.028	0.069	.945		1.002	0.950	1.059
CM & Ready Mean	-.011	0.042	-0.273	.785		0.989	0.911	1.074
English (CMR)	-.045	0.036	-1.243	.214		0.956	0.891	1.026
Math (CMR)	.024	0.036	0.662	.508		1.024	0.954	1.100
Science (CMR)	-.055	0.031	-1.760	.078		0.947	0.891	1.006
Social Studies (CMR)	-.044	0.031	-1.413	.158		0.957	0.900	1.017
Financial Situations								
EFC	-.187	0.051	-3.666	< .001	***	0.829	0.750	0.916
GA HOPE Scholarship	-.270	0.035	-7.805	< .001	***	0.763	0.713	0.817
Zell Miller Indicator	-.178	0.118	-1.508	.132		0.837	0.663	1.052
PELL Grant	-.165	0.049	-3.355	.001	**	0.848	0.769	0.933
Federal Sub. Loans	.038	0.038	1.008	.313		1.039	0.964	1.119
Federal Unsub. Loans	-.011	0.035	-0.315	.753		0.989	0.923	1.060
Other Loans	-.032	0.026	-1.227	.220		0.969	0.920	1.019
Major Groupings	.007	0.013	0.544	.586		1.007	0.981	1.034
Institutional Expenditures								
Academic & Inst. Support	.030	0.040	0.740	.459		1.030	0.952	1.115
All Others	.020	0.035	0.587	.557		1.020	0.954	1.092
Instruction	.051	0.040	1.261	.207		1.052	0.973	1.139
Public Service & Research	-.082	0.036	-2.238	.025	*	0.922	0.858	0.990
Student Service Support	.049	0.036	1.357	.175		1.050	0.979	1.127

Note. AIC = 8,948, BIC = 9,160, McFadden pseudo-R^2 = .048, and Nagelkerke pseudo-R^2 = .077. *** $p < .001$. *** $p < .001$. ** $p < .01$. * $p < .05$. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution. Inst. = institutional.

A total of 10 of the 29 factors were found to be significant. In comparing the results from the testing data set to the results from three training data sets, gender, race and ethnicity, HS GPA, admission test scores, AP hours, EFC, GA HOPE scholarship, PELL grant, and public service and research expenditures remained significant. First generation status of students (*B* = .199, *z* = 2.496, *p* = .013, *OR* = 1.221, 95% CI [1.043, 1.426]) was identified from the testing data set to be significant. From the findings, first generation students are 1.221 times more likely to depart than non-first generation students. College preparatory curriculum, content mastery and readiness mean, and major groupings were not found to be significant from the testing data set when the factors were found to be significant in the three models on the training data set.

Figure 26 illustrates the variable importance analysis across the three models from the training data sets in addition to the results from the testing data set. From the results of the three training data set models, GA HOPE scholarship, HS GPA, and gender are the top three most influential factors, exhibiting a strong negative impact on the departure of a student. Unlike the GA HOPE scholarship and HS GPA, gender exhibited a positive impact, indicating male students are more likely than female students to depart. In the three models on the training data sets, admissions test scores and HS locale placed in the fourth and fifth spots, exhibiting a moderate positive influence on departing from the institution. For each of the three models on the training data sets, none of the five HS curriculum variables were ranked in the top five spots. Content mastery and readiness mean ranked in the top 10 only in the no modification training model.

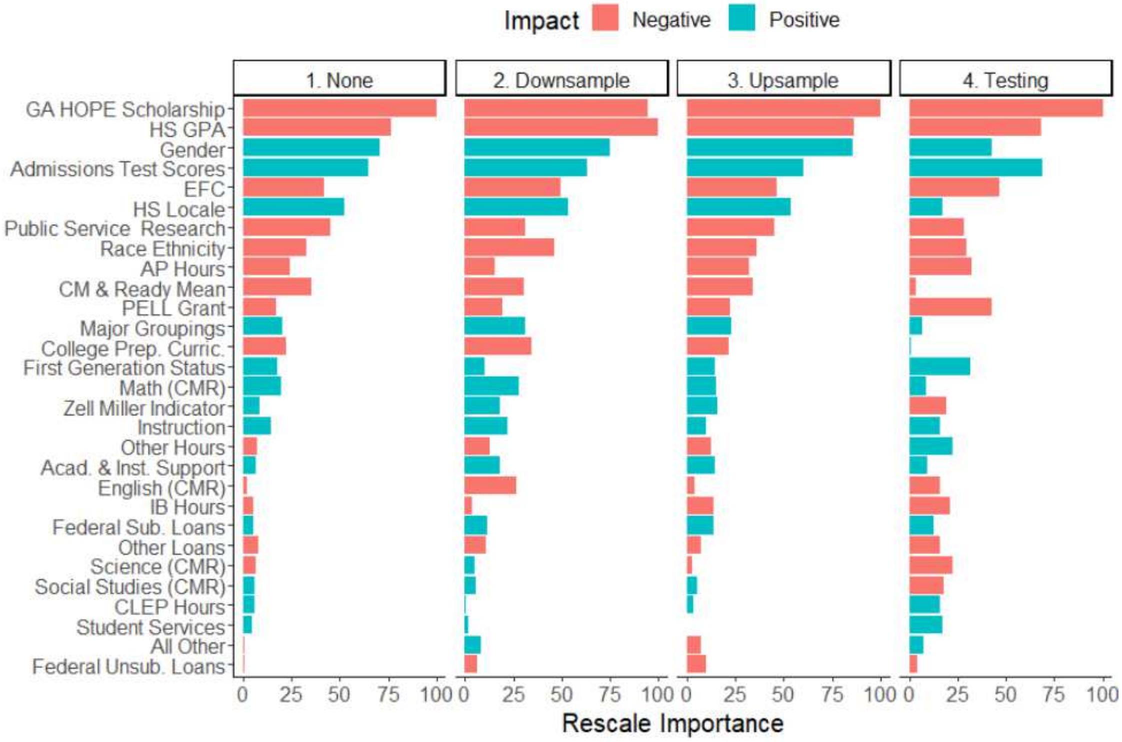


Figure 26. One-year retention variable importance plot for the logistic regression models. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph. The color of the bar indicates whether the impact is negative or positive on the one-year retention status. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

The variable importance results from the testing data revealed a separate set of influential factors, in which the top five relate to the student's academic preparation and ability to pay for the continued cost of attending. GA HOPE scholarship (importance = 7.805, rescaled importance = 100.000) maintained the top influential factor with a negative influence. Admissions test scores (importance = 5.357, rescaled importance = 68.635) barely surpassed HS GPA (importance = 5.351, rescaled importance = 68.565) for the second-place factor. Admissions test scores exhibited a moderate positive influence, while HS GPA exhibited a moderate negative influence. These findings indicate students with higher admissions test scores are more likely to depart, and students with higher HS GPA are less likely to depart. Surprisingly, EFC (importance = 3.666, rescaled importance = 46.972) and PELL Grant (importance = 3.355, rescaled importance = 42.989) were ranked fourth and fifth with negative moderate influence on the decision to depart.

Support vector machine with linear kernel. The SVM using a linear kernel was implemented via the svm_linear() function, employing the kernlab engine and setting the mode to classification. The cost and margin components underwent refinement through a grid of 20 models utilizing the training data set. A set.seed() function was employed for replication purposes. As highlighted by Batuwita and Palada (2012), SVM models are sensitive to data imbalances, potentially resulting in suboptimal models. Following the tuning process, the best model was identified based on the highest AUC value (0.554). This optimized model also exhibited an accuracy rate of 0.778, a sensitivity rate of 1.000, and a specificity rate of 0.001. These initial accuracy metrics suggest a predisposition for the model to favor the majority class in one-year retention predictions. The optimal configuration for the model included a cost value of 0.008 and a margin of 0.120.

The SVM model with a linear kernel underwent retuning for both the downsampled and upsampled data sets. The retuning process for the two models involved a tuning grid of 20 models to determine the optimal cost and margin for each respective model. The best-tuned downsampled model achieved an AUC value of 0.644. This optimized model exhibited an accuracy rate of 0.588, a sensitivity rate of 0.797, and a specificity rate of 0.379. The optimal configuration for the downsampled model included a cost of 0.002 and a margin of 0.122. The best-tuned upsampled model achieved an AUC value of 0.648. This optimized upsampled model displayed an accuracy rate of 0.592, a sensitivity rate of 0.803, and a specificity rate of 0.382. The optimal configuration for this model included a cost of 0.001 and a margin of 0.067. Both the downsampled and upsampled models exhibited substantial improvements in performance compared to the original model. These results indicated noticeable enhancements in the accuracy metrics, with no evidence of defaulting to majority classification. The models utilizing the downsample and upsample techniques appeared to be advancements over the original model.

A variable importance analysis was performed on the three SVM models utilizing a linear kernel, applied to both the training and testing data sets. In the model developed without sampling modifications, no variable was recognized as influential within the training and testing data sets. In the downsampled model, only the GA HOPE scholarship factor was identified as influential (importance = 0.083), while in the upsampled model, GA HOPE scholarship also played a role (importance = 0.091) on the training data sets. However, the downsampled and upsampled models on the testing data set did not reveal any variable influencing the decision to depart from the institution.

Support vector machine with polynomial kernel. The SVM model using a polynomial kernel was created through the svm_poly() function, employing the kernlab engine and setting the mode to classification. The cost, degree, scale factor, and margin components underwent fine-tuning using a grid of 20 models with the training data set. A set.seed() function was employed for replication. As previously mentioned, Batuwita and Palada (2012) highlighted the susceptibility of SVM models to data imbalances, leading to the development of suboptimal models. The best-tuned model exhibited the highest AUC value of 0.625. This tuned model also demonstrated a 0.778 accuracy rate, a sensitivity rate of 1.000, and a specificity rate of 0.001. These accuracy metrics suggest a potential tendency for the model to default to the majority class in one-year retention predictions. The optimal model configuration included a cost of 0.115, a degree of 1, a scale factor of 0.001, and a margin of 0.034.

The SVM model with a polynomial kernel underwent retuning for both the downsampled and upsampled data sets. The retuning process for the two models utilized a tuning grid of 20 models to determine the optimal cost and margin for each respective model. The best-tuned downsampled model achieved an AUC value of 0.644. This optimized model displayed an accuracy rate of 0.599, a sensitivity rate of 0.729, and a specificity rate of 0.471. The optimal configuration for the downsampled model included a cost of 0.004, a degree of 3, a scale factor of 0.076, and a margin of 0.089. The best-tuned upsampled model achieved an AUC value of 0.687. This optimized upsampled model exhibited an accuracy rate of 0.631, a sensitivity rate of 0.693, and a specificity rate of 0.568. The optimal configuration for this model included a cost of 0.004, a degree of 3, a scale factor of 0.076, and a margin of 0.089. Both the downsampled and upsampled models demonstrated substantial improvements in performance compared to the original model. These results indicate marked enhancements over the original model, with no evidence of defaulting to majority classification.

A variable importance analysis using a polynomial kernel SVM was conducted on three models, and Figure 27 displays the rescaled importance values. The analysis was performed on both training and testing data sets. The non-modified training data model is excluded from the figure due the inability to produce factors impacting the retention status. For the downsampled model, GA HOPE scholarship (importance = 0.040, rescaled importance = 100.000) emerged as the primary factor influencing departure, followed by AP hours (importance = 0.010, rescaled importance = 25.106) and student gender (importance = 0.009, rescaled importance = 23.40) within the training data set. In the testing data set, other advanced standing hours (importance = 0.0008, rescaled importance = 100.000) and CLEP hours (importance = 0.0008, rescaled importance = 96.029) were identified as dominant factors influencing departure. Following closely were HS locale, public service and research expenditures, and first-generation status as the third to fifth ranked variables with moderate influence. Admissions test scores, social studies proficiency level differences, and other loans contributed minimally to the dependent variable.

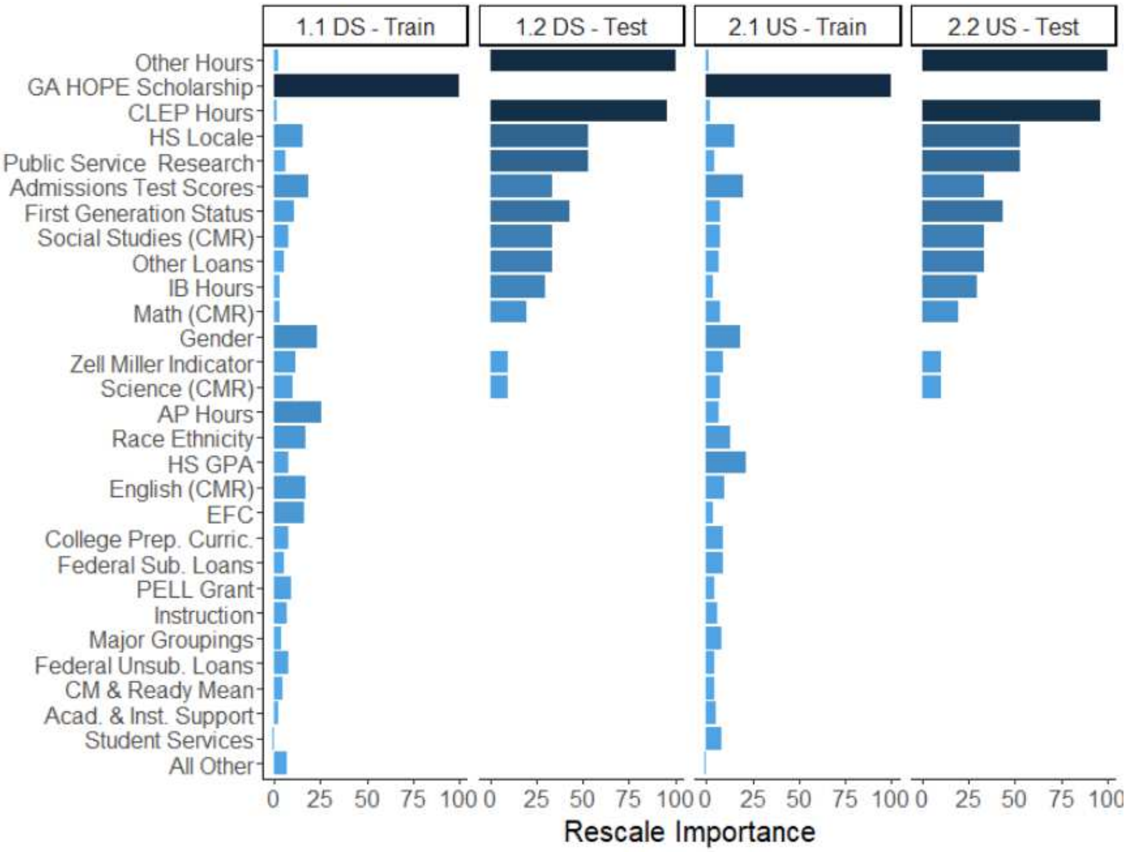


Figure 27. One-year retention variable importance plot for the three SVM models using a polynomial kernel on the training and testing data sets. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph. The model developed on no sampling modifications resulted in no distinguishable factors influencing the dependent variable and was excluded from the figure. DS = downsample. US = upsample. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

From the upsampled model's variable importance on the training data set, the GA HOPE scholarship (importance = 0.058, rescaled importance = 100.000) emerged as the dominant influential factor on the dependent variable, with HS GPA (importance = 0.013, rescaled importance = 21.783) as the second influential factor exhibiting a smaller influence. The factor influences from the testing data set changed, with other advanced standing hours (importance = 0.0008, rescaled importance = 100.000) and CLEP hours (importance = 0.0008, rescaled importance = 96.029) being the dominant impactful variables. The level of influence experienced a steep decline from the third to fifth ranked variables, including HS locale, public service and research expenditures, and first-generation status.

Support vector machine with radial basis function kernel. The SVM using a radial basis function kernel was developed through the svm_rbf() function with the engine set to kernlab and mode set to classification. The cost, rbf sigma, and margin components were fine-tunded using a grid of 20 models using the training data set. A set.seed() function was utilized for replication. Due to data imbalances, the SVM model could be susceptible in developing a non-optimal model (Batuwita & Palada, 2012). The best tuned model exhibited the highest AUC value of .626. This tuned model also exhibited a .778 accuracy rate, 1.000 sensitivity rate, and 0.000 specificity rate. These accuracy metrics indicate the model potentially defaulting to the majority class in the one-year retention predictions. The optimal model exhibited a cost of 0.024, radial basis function sigma of 0.001, and a margin of 0.048.

The SVM model with a radial basis function kernel was retuned for both the downsampled and upsampled data sets. The retuning of the two models utilized a tuning grid of 20 models to select the optimal cost and margin for each respective model. The best-tuned downsampled model achieved an AUC value of .644. This optimized model displayed an accuracy rate of .588, a sensitivity rate of .797, and a specificity rate of .379. The optimal configuration for the downsample exhibited a cost of 19.463, radial basis function sigma of 0.001, and a margin of 0.123. The best-tuned upsample model achieved an AUC value of .950. This optimized upsampled model displayed an accuracy rate of .493, a sensitivity rate of .400, and a specificity rate of .600. The optimal configuration for this model included a cost of 0.019, radial basis function sigma of 0.467, and a margin of 0.013. Both the downsampled and upsampled models demonstrated substantial improvements in performance when compared to the original model. These results indicated marked enhancements, with no indication of defaulting to majority classification, on the original model.

A variable importance analysis using a radial basis function kernel SVM was conducted on three models, and Figure 28 displays the rescaled importance values. The analysis was performed on both training and testing data sets, rescaling the importance values for easy comparison across models. The results from the non-modified training data model on the training and testing data sets are excluded from the figure due to no factors being identified as exhibiting influence on the dependent variable. From the analysis of the training data sets, the downsampled model only identified GA HOPE scholarship as the sole dominant factor impacting the departure decision. However, the number of CLEP and other advanced standing hours were identified as the sole dominant factors at the same level in the analysis of the downsampled model applied to the testing data set. The upsampled model applied to the training data set indicated numerous factors impacting departure. The top five factors were federal subsidized and unsubsidized loans, PELL grant, gender, and GA HOPE. Interestingly, the analysis on the testing data set yielded no variable influencing the dependent variable.

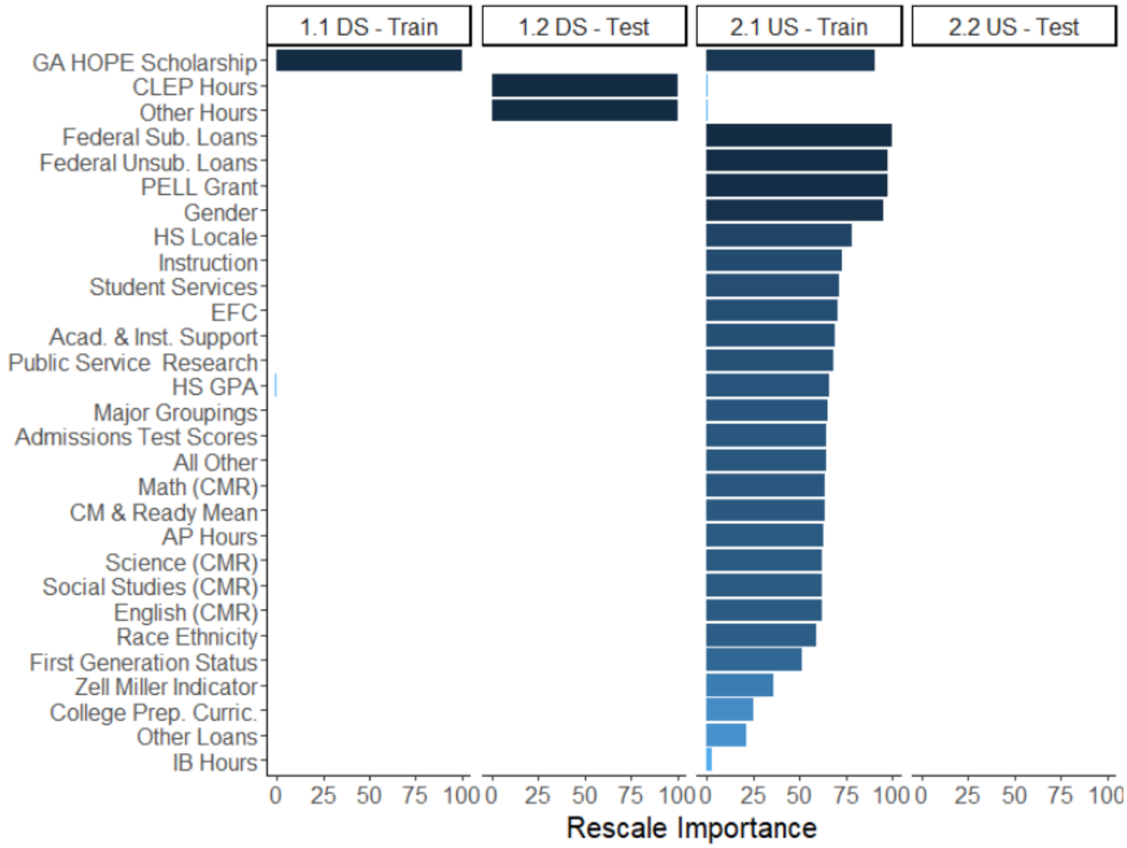


Figure 28. One-year retention variable importance plot for the three SVM models using a radial basis function kernel on the training and testing data sets. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph. The model developed on no sampling modifications resulted in no distinguishable factors influencing the dependent variable and was excluded from the figure. DS = downsample. US = upsample. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

Random forest. The random forest predictive algorithm was developed using the rand_forest() function, where the engine was configured to ranger and the mode set for classification. To optimize the model, the mtry, trees, and min_n parameters were tuned through a grid comprising 20 models utilizing training resampled data. The set.seed() function was utilized for replication purposes.

The best-tuned model exhibited the highest AUC value of .655, with accuracy rate of .778, a sensitivity rate of .999, and a specificity rate of .004. The optimal model configuration included 1,710 trees, an mtry value of 2, and a minimum value of 29. More importantly, the model's early accuracy metrics suggested a possible tendency to default to the majority class.

The random forest predictive model was subjected to retuning using both the downsampled and upsampled data sets. For the downsampled data, the mtry, trees, and min_n options were retuned. The best-tuned downsampled model achieved an AUC value of .650. This retuned model demonstrated a .611 accuracy rate, a .626 sensitivity rate, and a .596 specificity rate. The optimal configuration for this model included 1,306 trees, an mtry value of 3, and a minimum value of 24. Similarly, the random forest algorithm was retuned using the upsampled data set. Through the evaluation of a grid comprising 20 models, the mtry, trees, and min_n options were optimized, leading to the identification of the best-tuned upsampled model with an AUC value of .984. This retuned model exhibited a .940 accuracy rate, a .913 sensitivity rate, and a .966 specificity rate. The optimal configuration for the upsampled model comprised 731 trees, an mtry value of 24, and a minimum value of 3. Both the downsampled and upsampled models exhibited substantial improvements over the original model, specifically in the specificity and sensitivity rates. These enhancements suggest a significant increase in the predictive accuracy and reliability of the random forest models when trained on the modified data sets.

A variable importance analysis was conducted on the three random forest models, and Figure 29 displays the rescaled importance values. The analysis was performed on both training and testing data sets, with the rescaling of the importance values for easy comparison across models. Consistently, the top two factors across the three variable importance analyses of the training data set, GA HOPE scholarship (no modifications, importance = 0.012, rescaled importance = 100.000; downsample, importance = 0.015, rescaled importance = 100.00; upsample, importance = 0.173, rescaled importance = 100.000) and HS GPA (no modifications, importance = 0.009, rescaled importance = 74.676; downsample, importance = 0.009, rescaled importance = 64.190; upsample, importance = 0.171, rescaled importance = 100.000) were the top two influential factors.

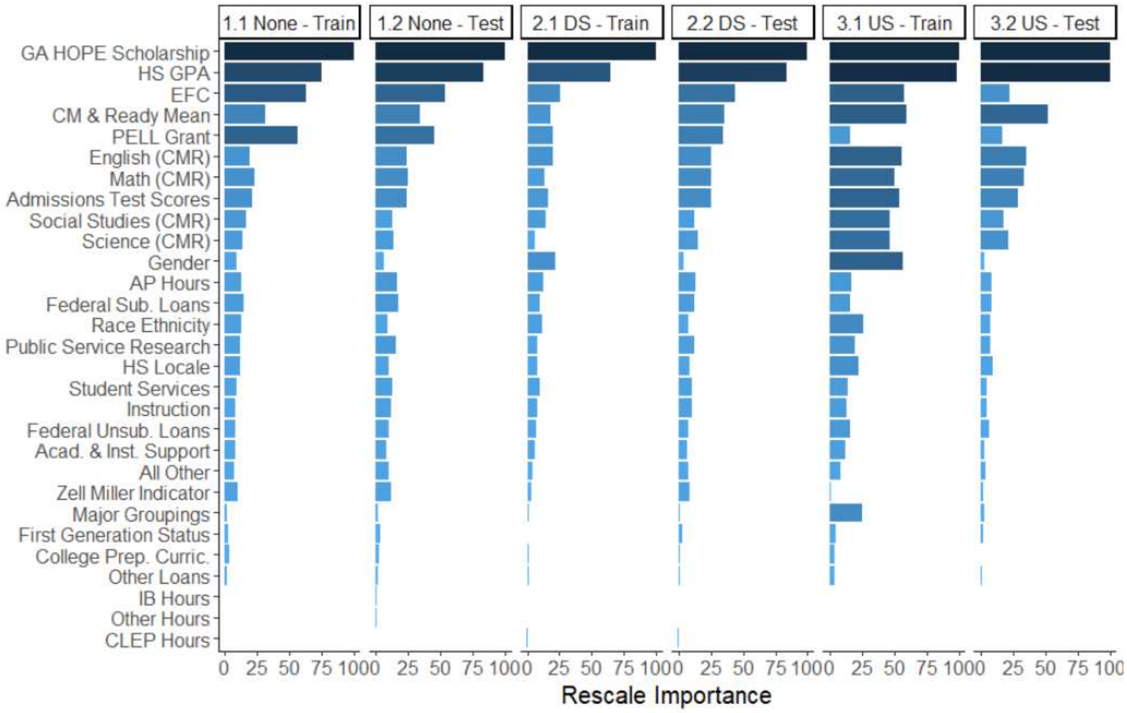


Figure 29. One-year retention variable importance plot for the three random forest models on the training and testing data sets. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph. None = no sampling modifications. DS = downsample. US = upsample. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution

Within the training data sets, the HS GPA factor exhibited its strongest influence within the upsample and weakest influence in the downsample model. EFC was identified as the third influential factor in the no modification and downsample models, while in the upsample model, the factor was the fourth most influential variable (no modifications, importance = 0.008, rescaled importance = 62.863; downsample, importance = 0.004, rescaled importance = 44.169; upsample, importance = 0.010, rescaled importance = 57.407). In the upsample model, the content mastery and readiness mean (importance = 0.102, rescaled importance = 59.094) variable ranked third and exhibited a moderate influence on the departure decision.

The variable importance analyses for each of the models were rerun on the testing data set. For each of the three models, GA HOPE scholarship and HS GPA remained the two dominant influential factors. In the upsample model, HS GPA was the top factor, while GA HOPE scholarship was the top factor for the no modification and downsample models. EFC factor was consistently the third-ranked factor with a moderate impact in the no modifications and downsample models, while within the upsample model, it was ranked seventh with a small impact. With a moderate impact, the content mastery and readiness mean factor was the third-ranked variable in the upsample model.

Extreme gradient boosting. An XGBoost model was constructed using the boost_tree() function, configuring the engine to xgboost and setting the mode for classification. Through the tuning process involving a grid of 20 models, the trees, tree depth, min_n, loss reduction, sample size, mtry, and learn rate parameters were optimized, with the set.seed() function for replication. The best tuned XGBoost model exhibited with an AUC value of .663. This optimal model also demonstrated an accuracy rate of .779, a sensitivity rate of .988, and a specificity rate of .045. In optimal configuration, the XGBoost model comprised 1,264 trees with an mtry of 22 and a minimum value of 20. The tree depth was set at 5, the learn rate at 0.007, the loss reduction at 0.004, and the sample size at 0.285.

Using the downsampled data set, the XGBoost predictive algorithm was retuned, resulting in an AUC value of .658. This optimized model exhibited an accuracy rate of .617, a sensitivity rate of .644, and a specificity rate of .590. Utilizing the training data set with the applied downsampling techniques, the model developed comprised 123 trees with an mtry of 7 and a minimum value of 17. The tree depth was set to 9, the learn rate to 0.020, the loss reduction to 0.006, and the sample size to 0.370. Additionally, the XGBoost predictive algorithm was retuned using the upsampled data set, with the best-tuned model exhibiting an AUC value of .976. This optimal model's other accuracy metrics were an accuracy rate of .916, a sensitivity rate of .863, and a specificity rate of .969. The returned XGBoost comprised of 1,686 trees with an mtry of 29 and a minimum value of 9. The tree depth was set to 12, the learn rate to .058, the loss reduction to .116, and the sample size to .879. Both the downsample and upsample models exhibited substantial improvements over the original model, specifically in the specificity and sensitivity rates.

A variable importance analysis was conducted on the three XGB models, and Figure 30 displays the rescaled importance values. The analysis was performed on both training and testing data sets, with the rescaling of the importance values to enable easy comparison across models. The top factor was not consistent among the three models on the training data set. GA HOPE scholarship (importance = 0.111, rescaled importance =

100.000) was the top factor for the model with no sampling modifications, HS GPA (importance = 0.161, rescaled importance = 100.000) was the top for the downsampled model, and EFC (importance = 0.115, rescaled importance = 100.000) was the top for the upsampled model. The second factor in each of the three models on the training data set exhibited a very strong influence on the retention decision, and consistency was exhibited in the no sampling modification model and the upsampled. HS GPA (no modification, importance = 0.108, rescaled importance = 97.505; upsample, importance = 0.113, rescaled importance = 98.484) was the second influential factor in the no sampling modification and upsampled models, while GA HOPE scholarship (importance = 0.146, rescaled importance = 90.902) was the second factor for the downsampled model.

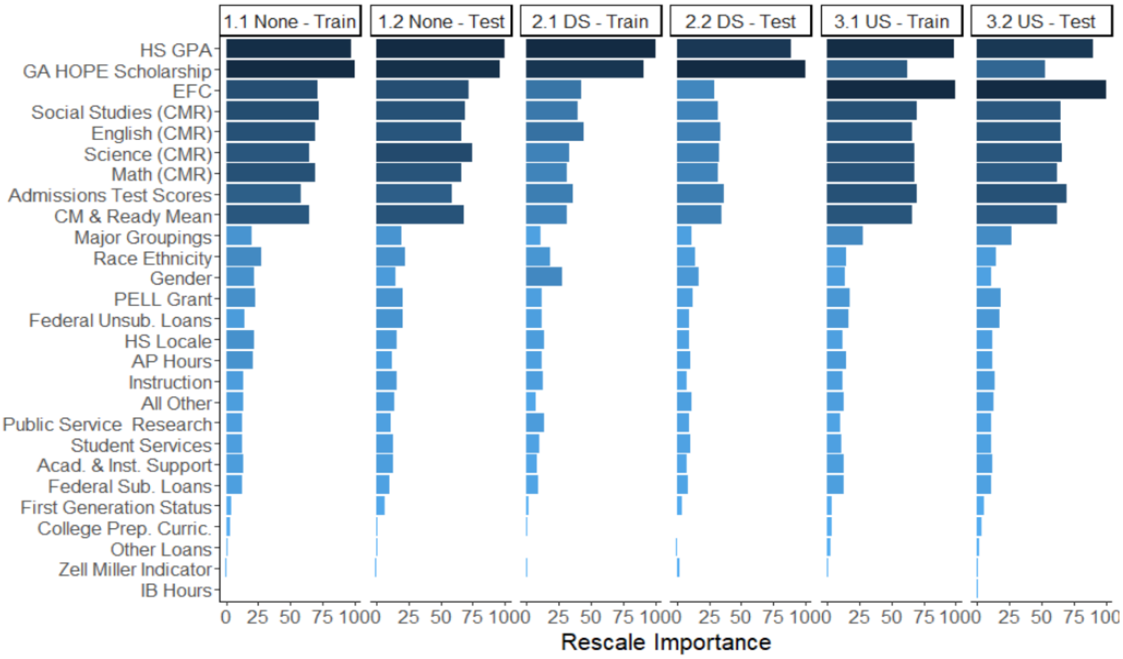


Figure 30. One-year retention variable importance plot for the three XGB models on the training and testing data sets. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph. None = no sampling modifications. DS = downsample. US = upsample. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

The five HS curriculum factors were ranked in the top 10 factors across all three models in the training data set. In comparing the results produced from the testing data set, the factors maintained their rankings. Across all three models in the testing data set, the top 10 factors are variables representing students' ability to perform academically, students' ability to pay, and the HS curriculum factors.

Variable importance comparison for one-year retention status. In comparing the key factors across the predictive algorithms, the results from the testing data set were utilized. Figure 31 contains the results of the logistic regression model along with the other models trained without any sampling modifications. Due to the imbalance of the classes in the dependent variable, the analysis of the three SVM models did not produce any factors exhibiting influence on the retention decision. From the remaining models, the analysis of the testing data set indicated GA HOPE scholarship was the top variable in the logistic regression and random forest models, while the factor ranked second in the XGBoost model. For the XGBoost model, HS GPA was the top influential factor, while it was ranked third in the logistic regression and second in the random forest models. Interestingly, the admissions test scores factor was ranked second in the logistic regression model, with just a slight advantage over HS GPA. The EFC factor was consistently ranked in the top five factors across the three algorithms.

Regarding the five HS curriculum variables, the random forest model found the content mastery and readiness mean and the differences in English and math proficiency levels from the content mastery and readiness mean to be in the top 10 factors with moderate influence. Within the XGBoost model, all five HS curriculum variables were ranked in the top 10 factors, with science and social studies proficiency levels differences from the content mastery and readiness mean exhibiting strong influences on the retention decision.

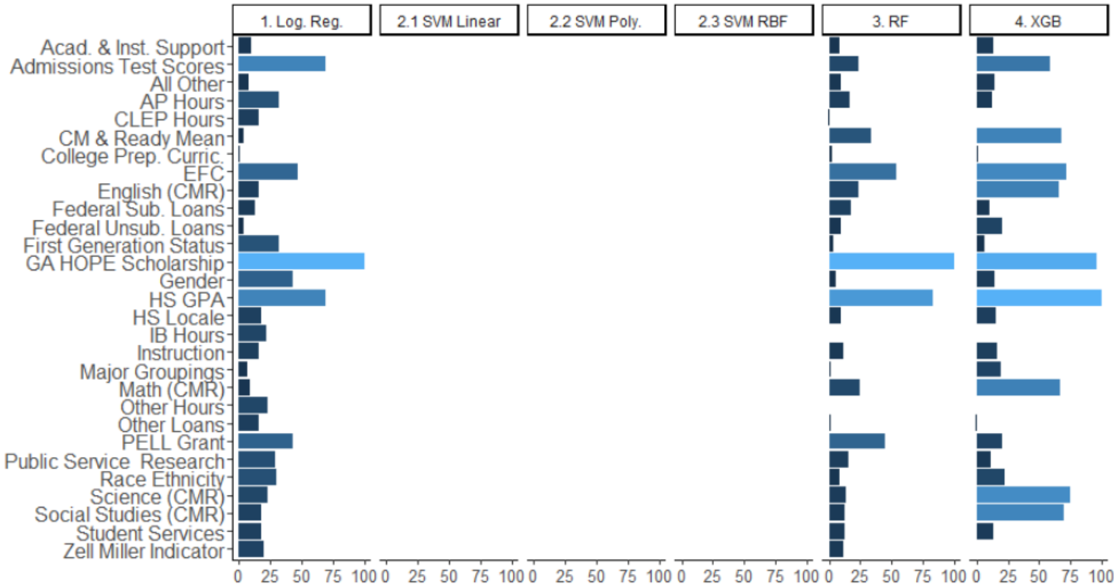


Figure 31. One-year retention variable importance plot comparison from predictive models utilizing the testing data set with no sampling modifications. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph. Log. Reg. = logistic regression, SVM linear = support vector machine using linear kernel. SVM Poly. = support vector machine using polynomial kernel. SVM RBF = support vector machine using radial basis function. RF = random forest. XGB = extreme gradient boost. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

Figure 32 displays the results of the variable importance analysis on the testing data set of the logistic regression model along with the models tuned utilizing the downsample techniques. The logistic regression model exhibited the same results, as it did not need retuning, resulting in the same outcome as shown in Figure 29. All the SVM models except the model using the linear kernel found factors influencing the dependent variable. However, the SVM models using the polynomial and radial basis kernels identified the number of CLEP and other advanced hours as the top two influential factors.

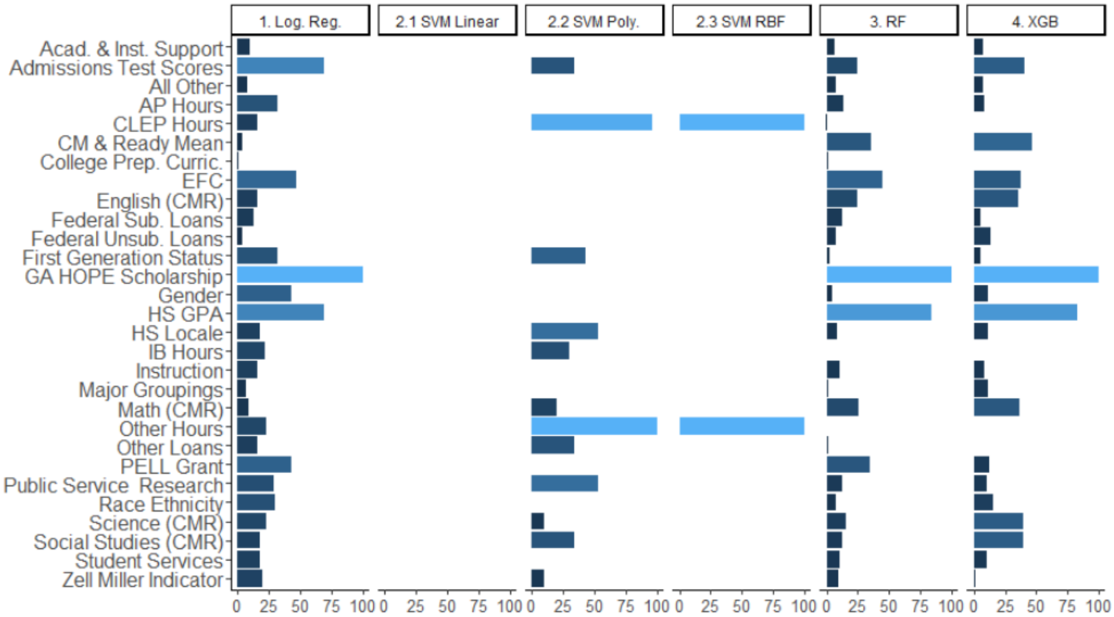


Figure 32. One-year retention variable importance plot comparison from predictive models utilizing the testing data set with downsampling techniques applied. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph. Log. Reg. = logistic regression, SVM linear = support vector machine using linear kernel. SVM Poly. = support vector machine using polynomial kernel. SVM RBF = support vector machine using radial basis function. RF = random forest. XGB = extreme gradient boost. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

In both the random forest and XGBoost downsample models, GA HOPE scholarship was identified as the top influential factor, with HS GPA ranking second. The EFC factor was ranked third in the random forest model, with a moderate impact; however, it was ranked seventh in the XGBoost model. All the HS curriculum factors, except for the social studies proficiency levels difference from the content mastery and readiness mean, were ranked in the top 10 factors, exhibiting moderate impacts. Content mastery and readiness mean exhibited a moderate impact and ranked as the third most influential factor in the XGBoost downsample model. Additionally, the proficiency levels differences from the content mastery and readiness mean factors in the four subject areas were ranked in the top 10 with moderate influence in the XGBoost model.

Figure 31 displays the variable importance analysis on the testing data set from the logistic regression model and the tuned upsample models. The logistic regression model exhibited the same results, as it did not need retuning, resulting in the same outcome as shown in Figure 31. Only the SVM model using the polynomial kernel produced influential factors, and the model was consistent with the downsample model in identifying the number of CLEP and other advanced standing hours as the top two factors. The top factors between the random forest and XGBoost models were not consistent. The analysis on the logistic regression model yielded HS GPA for the random forest model and EFC for the XGBoost model as the top influential factor. The GA HOPE scholarship factor was ranked second in the random forest model but ninth in the XGBoost model, exhibiting a much weaker influence in the XGBoost model. The admissions test scores factor was ranked third with a moderate influence in the XGBoost model, while ranking sixth with a small influence in the random forest model. While ranked in the top 10 factors, the five HS curriculum factors in the XGBoost model still exhibited moderate influences. The five HS curriculum factors were also ranked in the top 10 factors, but only content mastery and readiness mean exhibited a moderate influence in the random forest model.

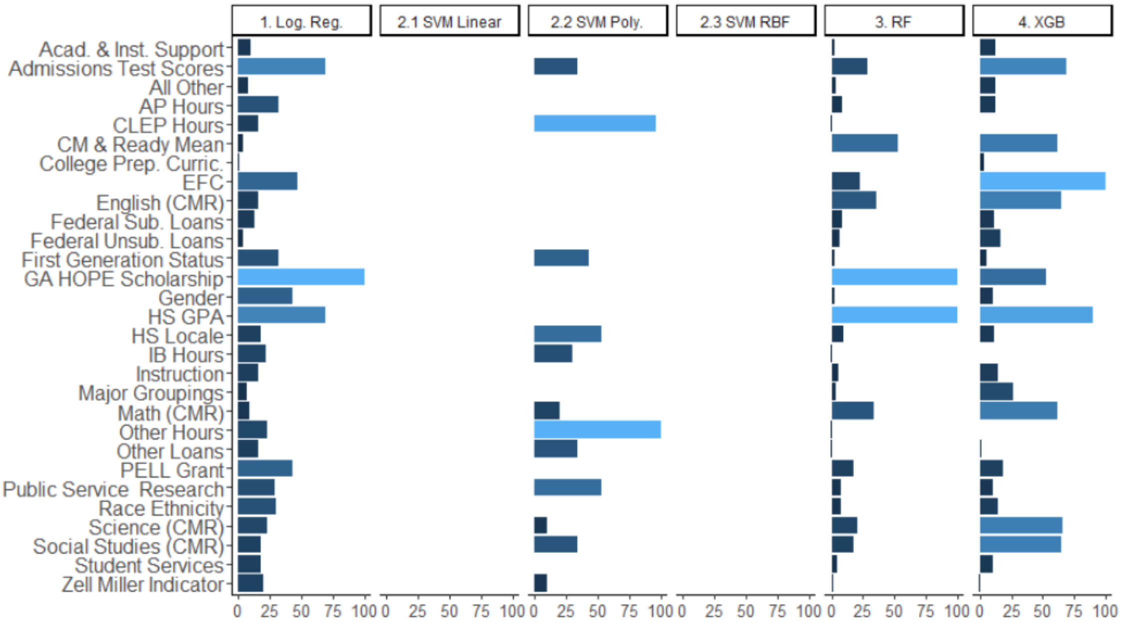


Figure 33. One-year retention variable importance plot comparison from predictive models utilizing the testing data set with upsampling techniques applied. The plot displays the variables in order of impact from highest to lowest with the names of the variables located on the y-axis of the graph. Log. Reg. = logistic regression, SVM linear = support vector machine using linear kernel. SVM Poly. = support vector machine using polynomial kernel. SVM RBF = support vector machine using radial basis function. RF = random forest. XGB = extreme gradient boost. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

Second Research Question

The following is the second research question:

2. Does one machine learning algorithm (linear regression, logistic regression, support vector machine, random forest, and extreme gradient boosting) or an ensemble learning algorithm produce a higher accuracy based on the evaluation metrics for accuracy in examination of first-year academic performance?
   a. Does one machine learning algorithm (linear regression, support vector machine, random forest, and extreme gradient boosting) or an ensemble

learning algorithm produce a higher accuracy based on the evaluation metrics of the root mean squared error (RMSE) for first-fall GPA?

b. Does one machine learning algorithm (linear regression, support vector machine, random forest, and extreme gradient boosting) or an ensemble learning algorithm produce a higher accuracy based on the evaluation metrics of the RMSE for first-year GPA?

c. Does one machine learning algorithm (logistic regression, support vector machine, random forest, and extreme gradient boosting) or an ensemble learning algorithm produce a higher accuracy based on the evaluation metrics of accuracy, sensitivity, specificity, f measure scores, and AUC value for one-year retention status?

First-fall GPA. To assess the predictive performance of the linear regression model and the five optimal models, two sets of cross-validation data were utilized in conjunction with the training and testing data sets. The cross-validation method applied to the training data involved a 10-fold approach, providing an initial assessment of predictive performance by calculating the mean performance across the folds. This method allowed for a preliminary measurement of accuracy before applying the models to unseen or testing data sets. The second set of cross-validation was applied to the testing data, which underwent the same preprocessing procedures as the training data. This assessment aimed to evaluate the models' predictive power on unseen data, providing insights into their generalization capabilities. In addition to the cross-validations, an ensemble learning approach was employed to enhance predictions. Two ensemble methods were used: one involved calculating the mean across the predictions, and the other utilized a blended technique with the stacking package. The blended method involved the use of functions such as blend_predictions() and fit_members() from the stack package, resulting in the generation of penalty and mixture values. These values were integrated into the linear_reg() function to blend the predictions together, creating more robust and accurate predictions.

By employing these cross-validation techniques and ensemble methods, a comprehensive evaluation of the models' performance was conducted, facilitating the selection of the best-performing model based on predictive accuracy and generalization capabilities. Examining the R^2 value of the training data set, the best value was obtained from using the XGBoost model ($R^2 = .329$), accounting for 32.9% of the variance within the data set. The random forest model ($R^2 = .318$) came in second. Of the three SVM models, the radial basis function kernel ($R^2 = .305$) exhibited the best value to account for the variance within the data set. To assess the overall predictive power of the models, the RMSE value was used as a metric to determine accuracy in predictions, as displayed in Figure 34. For the training data set, XGBoost exhibited the best RMSE value of 0.832, indicating the best predictive accuracy. The random forest model followed closely with an RMSE of 0.839, signifying its strong performance as well. Of the three SVMs, the SVM using a radial basis function exhibited the best RMSE value (0.863).

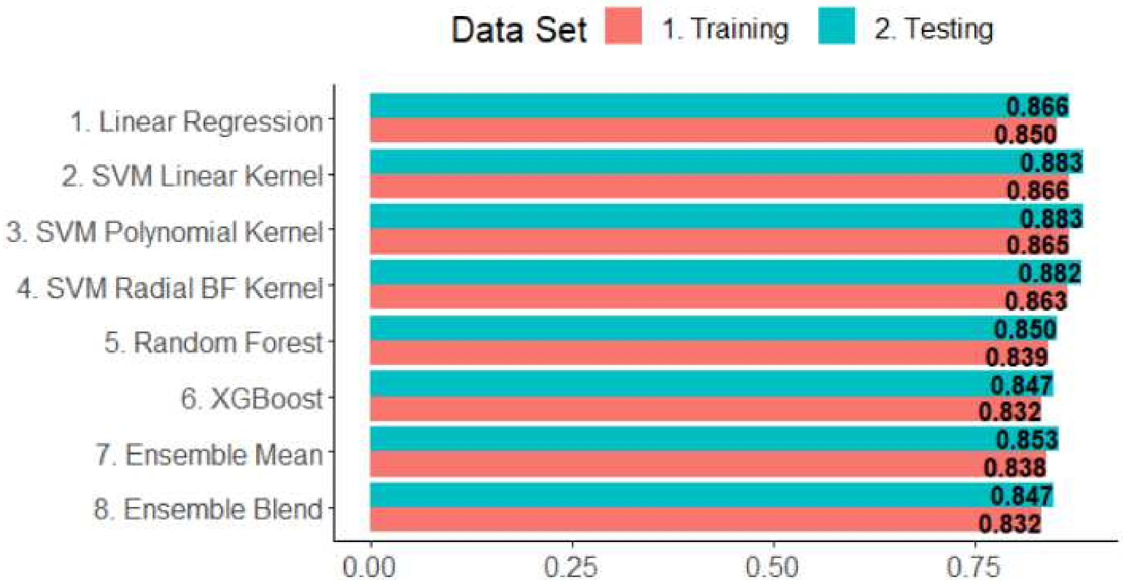


Figure 34. First-fall GPA predictive algorithms' RMSE values of the training and testing data sets. The figure also includes the RMSE produced from the two ensemble learning methods techniques.

In the testing data set, XGBoost maintained the best RMSE value (.847) with the random forest model performing well (RMSE of .850) too. Among the SVM models, the radial basis function kernel exhibited the best performance with an RMSE of .863 in the training data set and .882 in the testing data set. In evaluating the ensemble learning methods, the SVM models with linear and polynomial kernels were excluded, as the radial basis function kernel model proved to be the most effective among the three SVM models. Of the two ensemble learning methods, the blended method was the best when comparing the training and testing data sets. However, the XGBoost model was the best predictive model for the testing data set, with the random forest model coming in at a close second.

When comparing the distribution of RMSE values between the training and testing data sets, the RMSE values within the training data set were closely clustered together, as illustrated in Figure 35. The ensemble models were excluded from this comparison, as their inclusion did not lead to improvements in predictive power. Across the 10 folds in the training data set, the XGBoost model exhibited the most favorable distribution, with the random forest model following closely behind. The RMSE values for each model across the 10 folds were consistent, indicating stable performance across different subsets of the data. However, when examining the predictive performance across the 10 folds of the testing data set, a different distribution pattern emerged. Overall, XGBoost still exhibited the most favorable distribution of RMSE values with random forest coming in second, indicating consistent accuracy across different folds. To further evaluate these observations, a Mann-Whitney test was conducted to assess whether the RMSE values of the 10 folds differed significantly between the training and testing data sets. This statistical analysis aimed to provide a robust evaluation of the models' performance consistency across various data subsets. The linear regression (W = 23, p = .043) model was found to be significant between the median RMSE value of the training and testing data sets. This would indicate the models' predictive power on unseen data would differ from the training data set. The three SVM models (linear, W = 25, p = .063; polynomial, W = 28, p = .105; and radial basis function, W = 24, p = .052), random forest (W = 29, p = .123), and XGBoost (W = 27, p = .089) were found not to be significant. While determined to not have significant differences in the median RMSE values, these models' performance could be considered reliable in their performance between training and testing data sets.

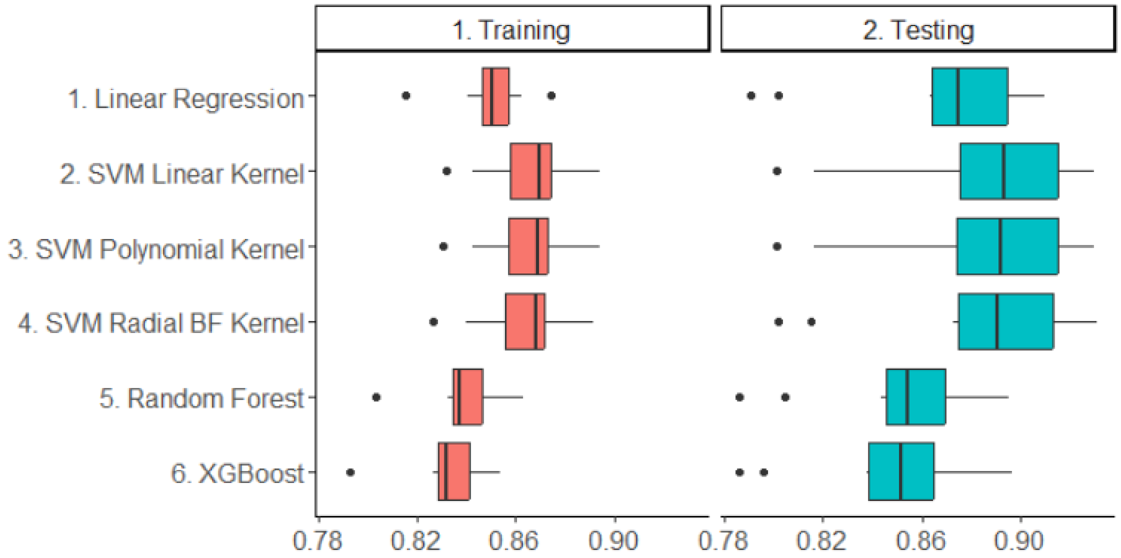


Figure 35. Boxplot of the first-fall GPA's predictive algorithms' RMSE values from the cross-validation folds of the training and testing data sets.

The Friedman's test was used to determine if one model's predictive performance is more significant than the others. Examining the RMSE values of the training data, the Friedman's test was found to be significant with a large effect size, $\chi^2(5) = 48.60$, $p <$ .001, $W = .971$. The Wilcoxon signed-rank with a Bonferroni multiple testing correction was utilized to determine which model was more significant. With the XGBoost model (*Mdn* = .832) performing the best in terms of the lowest RMSE value, the model was found to be significantly different from all of the five remaining models: (linear regression, *Mdn* = .850, *p* = .002; SVM using linear kernel, *Mdn* = .870, *p* = .002; SVM using polynomial kernel, *Mdn* = .869, *p* = .002, SVM using radial basis function kernel, *Mdn* = .868, *p* = .002; and random forest, *Mdn* = .838, *p* = .002). The Friedman's test was also conducted on the RMSE values produced by the testing data set. The test was found to be significant with a large effect size, $\chi^2(5) = 42.90$, $p < .001$, $W = .858$. The Wilcoxon signed-rank with a Bonferroni multiple testing correction was utilized to determine which model was more significant to use. With the XGBoost model (*Mdn* =

.852) performing the best in terms of the lowest RMSE value, the model was found to be significantly different from four of the five remaining models: (linear regression, *Mdn* = .875, *p* = .002; SVM using linear kernel, *Mdn* = .893, *p* = .002; SVM using polynomial kernel, *Mdn* = .892, *p* = .002, and SVM using radial basis function kernel, *Mdn* = .890, *p* = .002). The difference between random forest and XGBoost was found not to be significant in analyzing the testing data RMSE values. This finding could suggest the difference in the predictive power from either the random forest model or XGBoost model would not be sufficiently different from each other.

First-year GPA. To determine the model with the best predictive performance, two sets of cross-validation data were employed in conjunction with the linear regression model and the five optimal models. The cross-validation method applied to the training and testing data sets was a 10-fold. The cross-validation of the training data set provided an initial assessment of the predictive performance by calculating the mean performance across multiple folds. This method allowed for preliminary measurement of accuracy before the application of the models to unseen or testing data sets. The second set of cross-validation was applied to the testing data, which underwent the same preprocessing procedures as the training data. This assessment aimed to evaluate the models' predictive power on unseen data, providing insights into their generalization capabilities. In addition to the cross-validations, an ensemble learning approach was employed to enhance the predictions. Two ensemble methods consistent of calculating the mean across the predictions and the utilization of the blended technique with the stacking package. The blended method involved the use of functions like blend_predictions() and fit_members() from the stack package, resulting in the generation of a penalty and mixture value. These values were then integrated into the linear_reg() function to blend the predictions together, creating a more robust and accurate predictive model. By employing these cross-validation techniques and ensemble methods, a comprehensive evaluation of the models' performance was conducted, facilitating the selection of the best-performing model based on predictive accuracy and generalization capabilities.

Examining the R^2 value of the training data set, the best value was obtained from using the XGBoost model (R^2 = .352), accounting for 35.2% of the variance within the data set. The random forest model (R^2 = .340) came in second. Of the three SVM models, the radial basis function kernel (R^2 = .325) exhibited the best value to account for the variance within the data set. In assessing the overall predictive power of the models, the RMSE value was utilized to determine accuracy in the predictions as illustrated in Figure 36. For the training data set, XGBoost exhibited the best (.779), and random forest model exhibited the second best (.787) RMSE value. The RMSE of the testing data set indicated XGBoost still had the best RMSE value (.794) with random forest (.798) exhibiting the second best. The radial basis function kernel (training data set = .811; testing data set = .825) exhibited the best of the three SVM models.

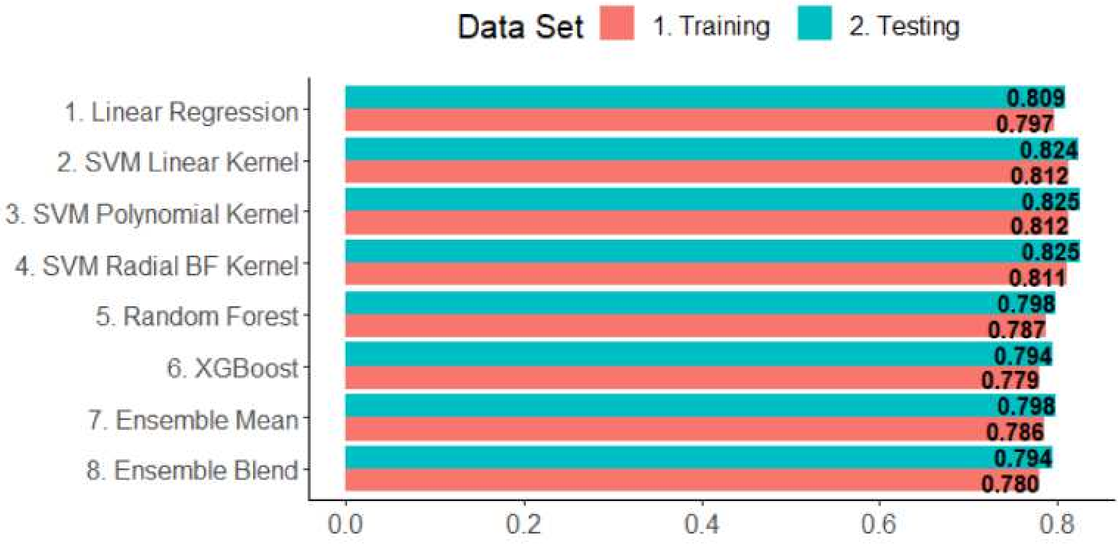


Figure 36. First-year GPA predictive algorithms' RMSE values of the training and testing data sets. The figure also includes the RMSE produced from the two ensemble learning methods techniques.

For the two ensemble learning methods, the SVM linear and polynomial kernel models were excluded as the radial basis function kernel was the best of the three. Of the two ensemble learning methods, the blended method was the best when comparing the training and testing data sets. However, the XGBoost model was the best predictive model for the testing data set, with the random forest model coming in at a close second.

When comparing the distribution of RMSE values between the training and testing data sets, the RMSE values within the training data set were closely clustered together, as illustrated in Figure 37. The ensemble models were excluded from this comparison, as their inclusion did not lead to improvements in predictive power. Across the 10 folds in the training data set, the XGBoost model exhibited the most favorable distribution, with the random forest model following closely behind. The RMSE values for each model across the 10 folds were consistent, indicating stable performance across different subsets of the data. However, when examining the predictive performance across the 10 folds of the testing data set, a different distribution pattern emerged.

Overall, XGBoost still exhibited the best distribution. To further evaluate these observations, a Mann-Whitney test was conducted to assess whether the RMSE values of the 10 folds differed significantly between the training and testing data sets. This statistical analysis aimed to provide an evaluation of the models' performance consistency across various data sets. All of the models were found to not be significant (linear regression, $W = 39$, $p = .436$; SVM linear, $W = 38$, $p = .393$; SVM polynomial, $W = 38$, $p = .393$; SVM radial basis function, $W = 39$, $p = .436$; random forest, $W = 41$, $p = .529$; and XGBoost, $W = 36$, $p = .315$), indicating the models performance between the training and testing data sets were fairly similar. While determined to not have significant differences in the median RMSE values, these models could be dependable in the performance on unseen data.

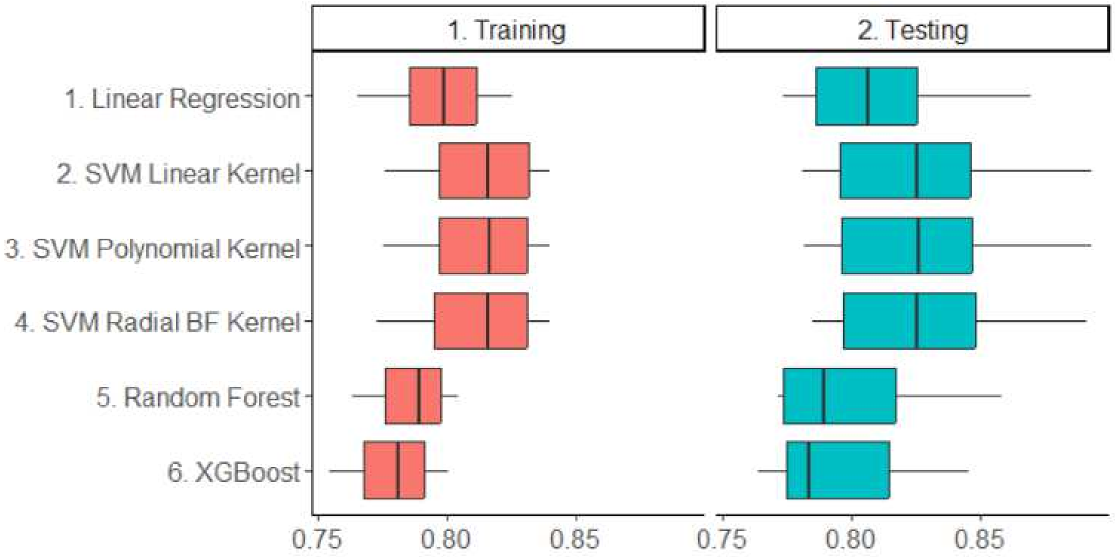


Figure 37. Boxplot of the first-year GPA's predictive algorithms' RMSE values from the cross-validation folds of the training and testing data sets.

The Friedman's test was used to determine if one model's predictive performance is more significant than the others. Examining the RMSE values of the training data, the Friedman's test was to be significant with a large effect size, $\chi^2(5) = 42.7$, $p < .001$, $W = .854$. The Wilcoxon signed-rank with a Bonferroni multiple testing correction was utilized to determine which model was more significant to use. With the XGBoost model (*Mdn* = .781) performing the best in terms of the lowest RMSE value, the model was found to be significantly different from all of the five remaining models: (linear regression, *Mdn* = .799, *p* = .029; SVM using linear kernel, *Mdn* = .816, *p* = .029; SVM using polynomial kernel, *Mdn* = .816, *p* = .029, SVM using radial basis function kernel, *Mdn* = .816, *p* = .029; and random forest, *Mdn* = .789, *p* = .029). The Friedman's test was also conducted on the RMSE values produced by the testing data set. The test was found to be significant with a large effect size, $\chi^2(5) = 43.00$, $p < .001$, $W = .859$. The Wilcoxon signed-rank with a Bonferroni multiple testing correction was utilized to determine which model was more significant to use. With the XGBoost model (*Mdn* = .784) performing the best in terms of the lowest RMSE value, the model was found to be significantly different from four of the five models: (linear regression, *Mdn* = .806, *p* = .029; SVM using linear kernel, *Mdn* = .825, *p* = .029; SVM using polynomial kernel, *Mdn* = .826, *p* = .029, and SVM using radial basis function kernel, *Mdn* = .825, *p* = .029). The difference between random forest and XGBoost was not found to be significant when analyzing the testing data RMSE values. This finding could suggest the difference in the predictive ability from either the random forest model or XGBoost model would not be sufficiently different from each other.

One-year retention status. To determine the model with the best predictive performance, two sets of cross-validation data were employed in conjunction with the linear regression model and the five optimal models. A 10-fold cross-validation method was applied to both the training and testing data sets. The cross-validation of the training data set provided an initial assessment of predictive performance by calculating the mean performance across multiple folds. This method allowed for a preliminary measurement of accuracy before applying the models to unseen or testing data sets. The second set of cross-validation was applied to the testing data, which underwent the same preprocessing procedures as the training data. This assessment aimed to evaluate the models' predictive power on unseen data, providing insights into their generalization capabilities.

In addition to the cross-validations, an ensemble learning approach was employed to further enhance the predictions. Two ensemble methods consisted of calculating the mean across the predictions and utilizing the blended technique with the stacking package. The blended method involved the use of functions like blend_predictions() and fit_members() from the stack package, resulting in the generation of a penalty and mixture value. These values were integrated into the linear_reg() function to blend the predictions together, creating a more robust and accurate predictive model. By employing these cross-validation techniques and ensemble methods, a comprehensive evaluation of the models' performance was conducted, facilitating the selection of the best-performing model based on predictive accuracy and generalization capabilities. From the training and testing data sets, the predicted class were compared to the actual class displayed in a two-by-two grid known as a confusion matrix. The accuracy metrics of overall accuracy, sensitivity, specificity, and f-measure scores were produced from the matrix. Additionally, the AUC value was calculated to assess the accuracy of the model. The best model was selected based on the best AUC value, as Dey (2021) indicated this metric was the best to determine accuracy.

Original model. Table 30 displays the confusion matrices of the predictive algorithms' performance on both the training and testing data sets without utilizing data rebalancing techniques. The receiver operating characteristic (ROC) curves are illustrated in Figure 38. Both the table and the figure include the results from the two ensemble learning methods. Across the models, the overall accuracy on the training data set was similar, showing a slight decrease when assessing the performance on the testing data set. A notable observation was the high bias towards the majority class within both the training and testing data sets. This bias was a consequence of the class imbalance in the retention status within the data sets. Upon comparing the predicted values to the actual values, all six models, along with the two ensemble learning models, exhibited a high bias towards the majority class. This imbalance led to extremely low values of specificity. For example, in the training data set, the logistic regression model had specificity values of 0.023, while the SVM models exhibited a specificity of 0.000. Similarly, the random forest and XGBoost models showed specificity values of 0.003 and 0.046, respectively, in the training data set. These low specificity values indicated a high number of false positives, meaning a sizable proportion of students predicted to retain were, in fact, not retained. This observation highlights the challenges posed by class imbalance and underscores the need to investigate techniques to address this issue and improve any model's predictive performance.

Table 20

Confusion Matrices Results of the Predictive Algorithms on the Training and Testing Data Sets without Data Rebalancing Procedures

	Training Data Set		Testing Data Set	
	Actual Retained	Actual Not Retained	Actual Retained	Actual Not Retained
Logistic Regression				
Predicted Retained	10,114	2,839	6,683	1,917
Predicted Not Retained	58	67	57	62
SVM Linear				
Predicted Retained	10,172	2,906	6,740	1,979
Predicted Not Retained	0	0	0	0
SVM Polynomial				
Predicted Retained	10,172	2,906	6,740	1,979
Predicted Not Retained	0	0	0	0
SVM Radial Basis Function				
Predicted Retained	10,172	2,906	6,740	1,979
Predicted Not Retained	0	0	0	0
Random Forest				
Predicted Retained	10,166	2,897	6,735	1,972
Predicted Not Retained	6	9	5	7
XGBoost				
Predicted Retained	10,054	2,772	6,662	1,904
Predicted Not Retained	118	134	78	75
Ensemble Learning Mean				
Predicted Retained	10,140	2,861	6,717	1,960
Predicted Not Retained	32	45	23	19
Ensemble Learning Blended				
Predicted Retained	10,008	2,745	6,651	1,896
Predicted Not Retained	164	161	89	83

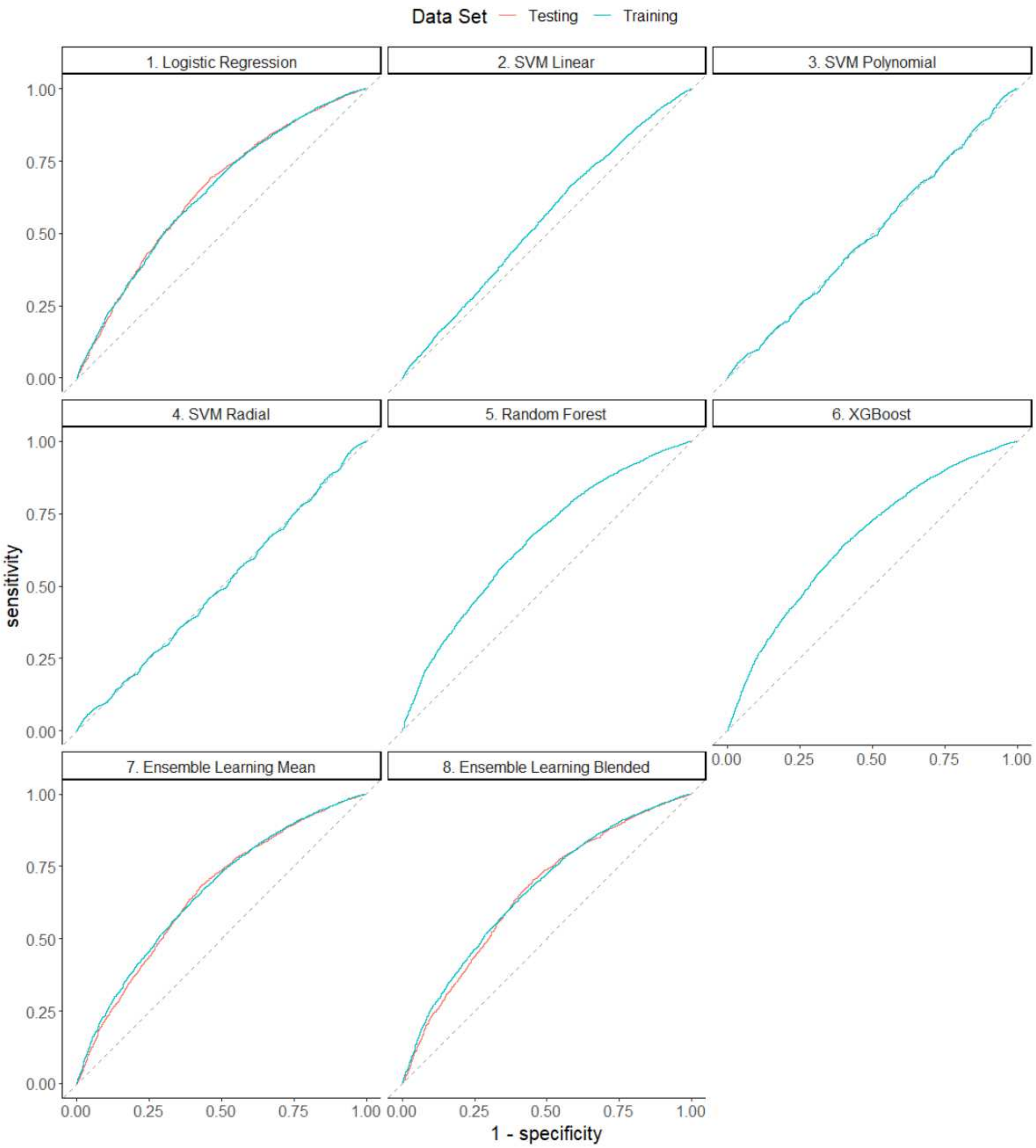


Figure 38. One-year retention ROC curves of the predictive algorithms developed on the training without any sampling techniques. The image includes the ROC curves produced from the training and testing data sets.

Following Dey's recommendation (2021), the preferred metric to determine the best model is the AUC metric. In the evaluation of AUC values across different models, the logistic regression model exhibited an AUC value of .640 for the training data set and .643 for the testing data set. The SVM model using the linear kernel demonstrated an

AUC of .554 for the training data set and .503 for the testing data set. There was an improvement in the AUC values for the SVM model using the polynomial kernel (.625 for training and .623 for testing data sets) and the radial basis function kernel (.626 for training and .623 for testing data sets). Among the models, the random forest exhibited AUC values of .655 for the training data set and .649 for the testing data set, whereas XGBoost achieved the highest performance with AUC values of .663 for the training data set and .653 for the testing data set. Despite attempts to enhance performance through ensemble learning methods, these methods did not significantly improve predictive performance due to the models' susceptibility to exhibiting bias towards the majority class. The AUC metric confirmed both the random forest and XGBoost algorithms outperformed others.

When comparing the distribution of AUC values across folds between the training and testing data sets without any sampling techniques, the AUC values within the training data were closely clustered together, as depicted in Figure 39. The ensemble models were excluded from this comparison as they did not enhance predictions. Across the 10 folds in the training data set, the XGBoost model exhibited the most consistent distribution, with the random forest model coming in second. In contrast, the distributions of AUC values for logistic regression and SVM using the linear kernel were more dispersed compared to the other models. When comparing the AUC distributions in the testing data sets, all models, except the SVM using the linear kernel, closely resembled the distribution observed in the training data set. Overall, XGBoost performed marginally better than the random forest model in the testing data set. To evaluate these observations, a Mann-Whitney test was conducted to assess whether the AUC values of the 10 folds differed significantly between the training and testing data sets. This statistical analysis aimed to provide an evaluation of the consistency in the models' performance. The SVM using linear kernel (*W* = 83, *p* = .012) was found to be significantly different. The result indicated the SVM using a linear kernel model did not exhibit consistent performance between the training and testing data sets. All of the remaining models were found to not be significant (logistic regression, *W* = 44, *p* = .6842; SVM polynomial, *W* = 55, *p* = .739, SVM radial basis function, *W* = 59, *p* = .529; random forest, *W* = 56, *p* = .684; and XGBoost, *W* = 63, *p* = .353), indicating the models' performances between the training and testing data sets were fairly similar. While determined to not have significant differences in the median AUC values, these four models could be considered to be dependable in the performance on unseen data.

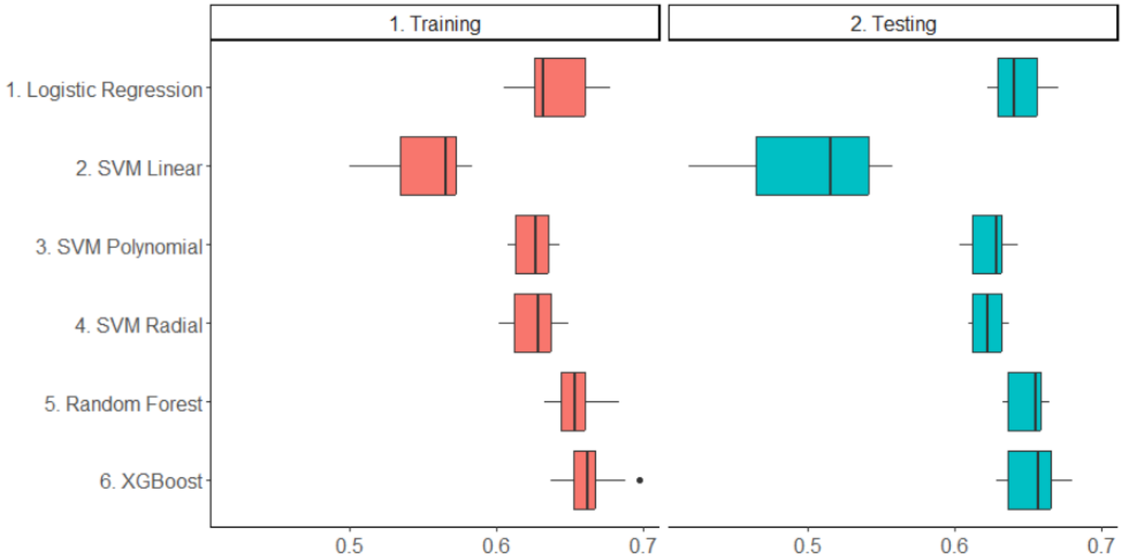


Figure 39. Boxplot of the one-year retention's predictive algorithms' AUC values from the cross-validation folds of the training and testing data sets.

The Friedman's test was used to determine if one model's performance is more significant than the others. Examining the AUC values of the training data, the Friedman's test was to be significant with a large effect size, $\chi^2(5) = 44.00$, $p < .001$, *W* =

.880. The Wilcoxon signed-rank with a Bonferroni multiple testing correction was utilized to determine which model was more significant to use. With the XGBoost model (*Mdn* = .662) performing the best in terms of the highest AUC value, the model was found to be significantly different from four of the five remaining models: (logistic regression, *Mdn* = .632, *p* = .029; SVM using linear kernel, *Mdn* = .566, *p* = .029; SVM using polynomial kernel, *Mdn* = .627, *p* = .029, SVM using radial basis function kernel, *Mdn* = .629, *p* = .029). The difference between XGBoost (.662) and random forest (.653) model was not statistically significant within the training data set. The Friedman's test was also conducted on the AUC values produced by the testing data set. The test was found to be significant with a large effect size, $\chi^2(5) = 40.90$, $p < .001$, $W = .817$. The Wilcoxon signed-rank with a Bonferroni multiple testing correction was utilized to determine which model was more significant to use. With the XGBoost model (*Mdn* = .657) performing the best in terms of the highest AUC value, the model was found to be significantly different from four of the five models: (logistic regression, *Mdn* = .640, *p* = .029; SVM using linear kernel, *Mdn* = .516, *p* = .029; SVM using polynomial kernel, *Mdn* = .628, *p* = .029, and SVM using radial basis function kernel, *Mdn* = .623, *p* = .029). The difference between random forest (*Mdn* = .655) and XGBoost was not found to be significant analyzing the testing data AUC values. This finding could suggest the difference in the predictive ability from either the random forest model or XGBoost model would not sufficiently differ from each other.

Downsample model. Table 21 presents the confusion matrices of the six algorithms developed using the training data set with the downsampling technique applied. The table also includes the two ensemble methods. Upon examining the distribution of predicted and actual values in the training data set, the bias towards the majority class present in the original data set has been corrected, as also evidenced by increased specificity rates (logistic regression, .543; SVM using linear kernel, .379; SVM using polynomial kernel, .471; SVM using radial basis function, .379; random forest, .602; and XGBoost, .622). While the downsampling technique resulted in improved specificity within the training data set, the improvement came at the cost of decreased overall accuracy and sensitivity when compared to the original data set. However, despite the correction of the bias in the training data set, the performance on the testing data set did not show significant improvements, resulting in a continued bias towards the majority class. Additionally, the ROC curves are displayed in Figure 40. The AUC values for the logistic regression model improved from .638 in the training data set to .643 in the testing data set. The three SVM models exhibited AUC values of .644 in the training data set, but the values dropped in the testing data set (linear kernel, .514; polynomial kernel, .548; and radial basis function, .550). The random forest (training data set, .650; and testing data set, .649) and XGBoost (training data set, .658; and testing data set, .652) algorithms displayed the highest AUC values among the models. The AUC values for the random forest were similar between the training and testing data sets, differing by only .001. No improvements in the predictions were exhibited in the ensemble learning methods. The AUC values for the ensemble learning methods were as follows: mean method (training data set, .656; and testing data set, .656) and blended method (training data set, .658; and testing data set, .655).

Table 21

Confusion Matrices Results of the Predictive Algorithms on the Training and Testing Data Sets with Downsampling Techniques

	Training Data Set		Testing Data Set	
	Actual Retained	Actual Not Retained	Actual Retained	Actual Not Retained
Logistic Regression				
Predicted Retained	1,919	1,327	6,683	1,917
Predicted Not Retained	987	1,579	57	62
SVM Linear				
Predicted Retained	2,317	1,803	6,740	1,979
Predicted Not Retained	589	1,103	0	0
SVM Polynomial				
Predicted Retained	2,115	1,538	6,736	1,977
Predicted Not Retained	791	1,368	4	2
SVM Radial Basis Function				
Predicted Retained	2,316	1,805	6,740	1,979
Predicted Not Retained	590	1,101	0	0
Random Forest				
Predicted Retained	1,841	1,158	6,735	1,972
Predicted Not Retained	1,065	1,748	5	7
XGBoost				
Predicted Retained	1,808	1,155	6,663	1,913
Predicted Not Retained	1,098	1,751	77	66
Ensemble Learning Mean				
Predicted Retained	1,894	1,220	6,718	1,958
Predicted Not Retained	1,012	1,686	22	21
Ensemble Learning Blended				
Predicted Retained	1,839	1,180	6,650	1,896
Predicted Not Retained	1,067	1,726	90	83

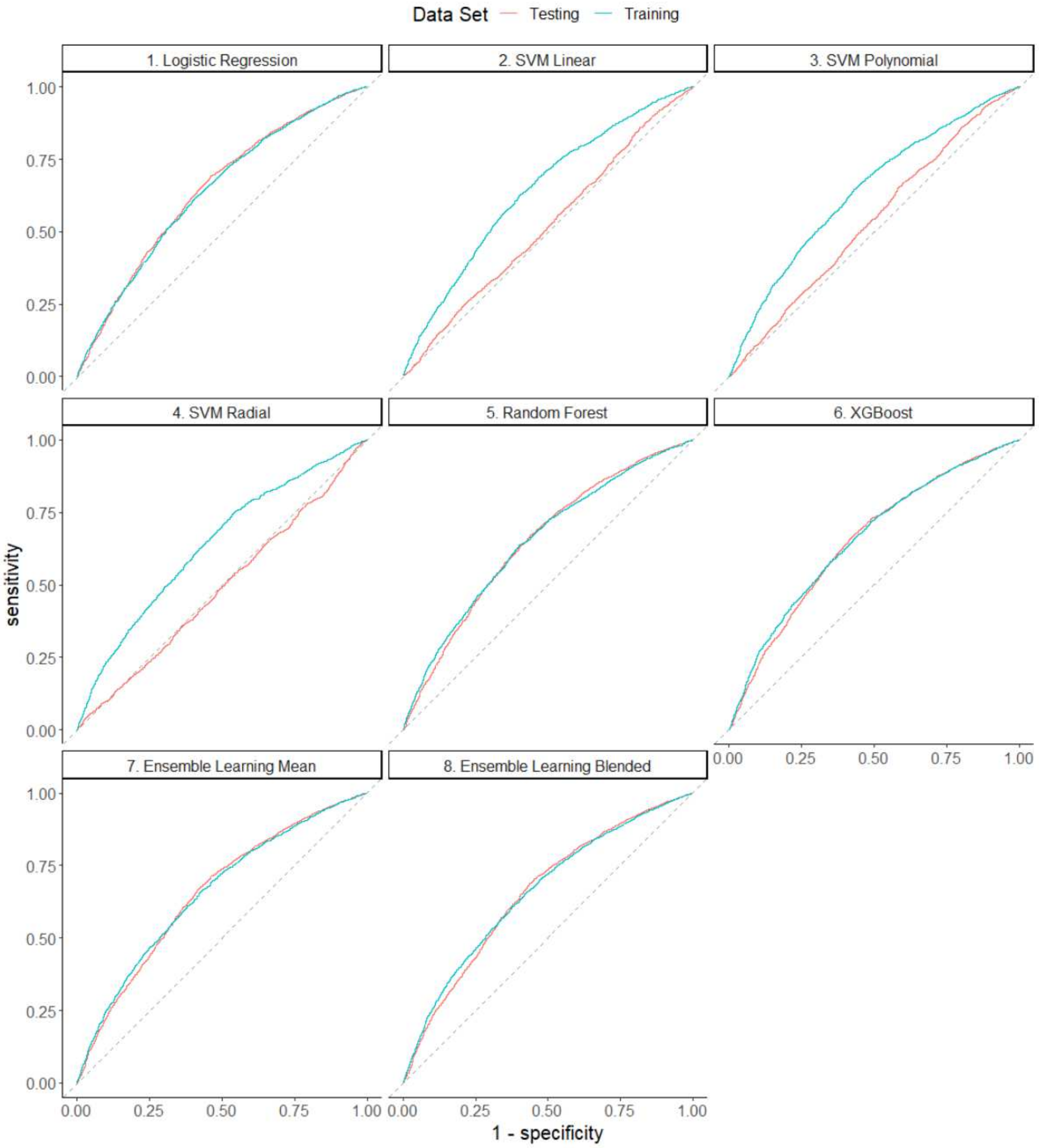


Figure 40. One-year retention ROC curves of the predictive algorithms developed on the training with downsampling techniques. The image includes the ROC curves produced from the training and testing data sets.

When comparing the distribution of AUC values across folds between the training and testing data sets with downsampling techniques applied, the AUC values within the training data were closely clustered together, as depicted in Figure 41. The ensemble models were excluded from this comparison as they did not enhance predictive power. Across the 10 folds in the training data set, XGBoost model exhibited the best distribution with random forest model coming second. The logistic regression and SVM using the linear or the radial basis function kernels models' distributions of the AUC values were more dispersed than the other models. When comparing the testing data sets' AUC distribution, all but the SVM models were close to the training data set distribution. Overall, XGBoost was only slightly higher than the random forest.

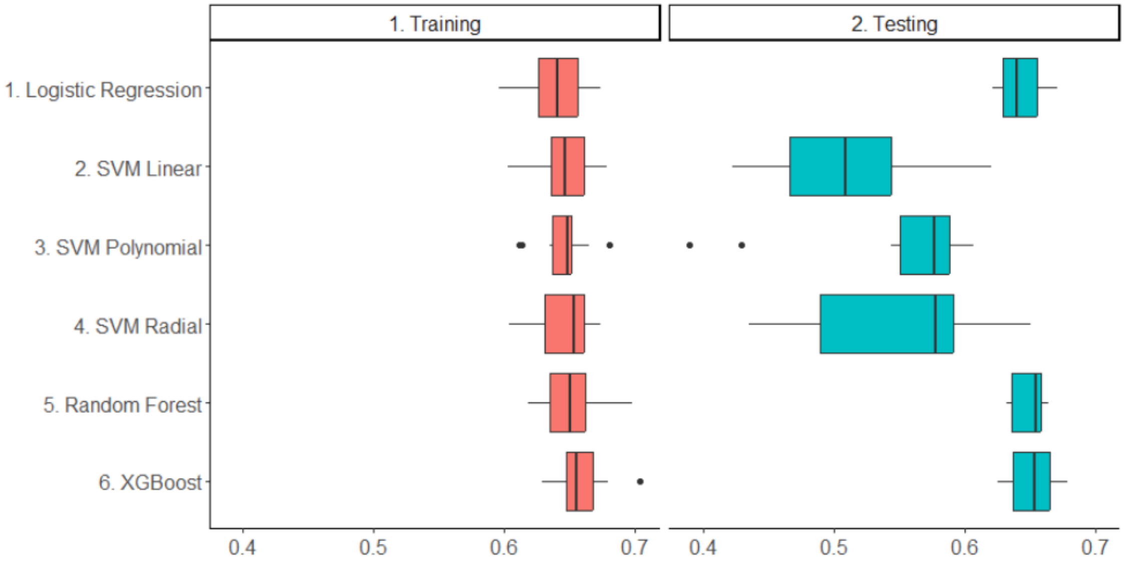


Figure 41. Boxplot of the one-year retention's predictive algorithms' AUC values from the cross-validation folds of the training and testing data sets.

Examining the predictive performance of each algorithm, a Mann-Whitney test was performed to evaluate the AUC values of the 10 folds between the training and testing data sets. The three SVM models were found to be significantly different (linear kernel, $W = 98$, $p < .001$, polynomial kernel, $W = 100$, $p < .001$; and radial basis function, $W = 96$, $p < .001$). The result indicated the three SVM models would not exhibit consistent performance between the training and testing data sets. The remaining models were found to not be significant (logistic regression, $W = 47$, $p = .853$; random forest, $W = 45$, $p = .739$; and XGBoost, $W = 55$, $p = .739$), indicating the model's performance between the training and testing data sets were fairly similar. While determined to not have significant differences in the median AUC values, these other models' performance would be dependable.

The Friedman's test was used to determine if one model is more significant than the other. Examining the AUC values of the training data, the Friedman's test was to be significant with a large effect size, $\chi^2(5) = 26.4$, $p < .001$, $W = .526$. The Wilcoxon signed-rank with a Bonferroni multiple testing correction was utilized to determine which model was more significant to use. The only statistically significant difference occurred from the logistic regression (*Mdn* = .640, *p* =.029) model to the XGBoost (*Mdn* = .655). Based on the performance of the models developed on the downsample training data set, the models, excluding the XGBoost model, would be expected to perform like each other. While the accuracy metrics on the training data set indicated the models would perform statistically similar, a different story is revealed when conducting the Friedmen's test on the performance of the testing data sets. The test was found to be significant with a large effect size, $\chi^2(5) = 40.70$, $p < .001$, $W = .815$. The Wilcoxon signed-rank with a Bonferroni multiple testing correction was utilized to determine which model was more significant to use. The random forest model (*Mdn* = .655) was the model with the highest AUC value and was found to be statistically different from the three SVM models (linear kernel, *Mdn* = .509, *p* = .029; polynomial kernel, *Mdn* = .577, *p* = .029; and radial basis function, *Mdn* = .579, *p* = .029). Random forest model was not statistically different from the logistic regression model (*Mdn* = .640, *p* = 1.000) and the XGBoost model (*Mdn* = .654, *p* = 1.000). This finding could suggest the difference in the predictive ability from the logistic regression, random forest model or XGBoost models would not be sufficiently different from each other.

Upsample model. The upsample data set confusion matrices results are displayed in Table 22, which includes the distribution of the predicted and actual retention status of the ensemble learning methods. Furthermore, the ROC for the training and testing data sets are illustrated in Figure 42. The distribution of the predicted and actual retention statuses for the training data set exhibits no signs of bias towards the majority class. For the upsample data set, the overall accuracy rates of the training data set were noticeably higher in the random forest (.939) and XGBoost (.915) than the accuracy rates were in the original and downsample data sets. However, the distribution of the predicted and actual status and the overall accuracy rates returned to the rates experienced within the original data set when examining the performance on the testing data set. Additionally, the sensitivity and specificity within the training data set metrics indicated higher accuracy in predicting the true positives and true negatives, but the models' performance reverted to the bias of the majority classification, resulting in a high number of false negatives or low rate of specificity. The review of the ROC graphs produced model are exhibited in Figure 42. These ROC graphs are based on the algorithms developed utilizing the upsample training data set. Based on the figure, the random forest and XGBoost models' ROC produced from the training data set revealed a near perfect accuracy rate (random forest's AUC .984, and XGBoost's AUC .976); however, the AUC values of these two models were lowered on the testing data set (random forest's AUC .633, and XGBoost's AUC .607). Random forest algorithm was the highest AUC value in the training data set, with XGBoost coming in at a close second. Yet, the logistic regression model comes in at the top performing model regarding the AUC values of the testing data set (AUC .643). The ensemble learning methods exhibited slight improvements to the accuracy metrics in the both the training and testing data sets.

Table 22

Confusion Matrices Results of the Predictive Algorithms on the Training and Testing Data Sets with Upsampling Techniques

	Training Data Set		Testing Data Set	
	Actual Retained	Actual Not Retained	Actual Retained	Actual Not Retained
Logistic Regression				
Predicted Retained	6,712	4,558	6,683	1,917
Predicted Not Retained	3,460	5,14	57	62
SVM Linear				
Predicted Retained	8,148	6,335	6,740	1,979
Predicted Not Retained	2,024	3,837	0	0
SVM Polynomial				
Predicted Retained	7,051	4,392	6,736	1,977
Predicted Not Retained	3,121	5,780	4	2
SVM Radial Basis Function				
Predicted Retained	4,001	4,137	6,740	1,979
Predicted Not Retained	6,171	6,035	0	0
Random Forest				
Predicted Retained	9,270	349	6,517	1,801
Predicted Not Retained	902	9,823	223	178
XGBoost				
Predicted Retained	8,756	314	6,136	1,616
Predicted Not Retained	1,416	9,858	604	363
Ensemble Learning Mean				
Predicted Retained	8,850	316	6,494	1,780
Predicted Not Retained	1,322	9,856	246	199
Ensemble Learning Blended				
Predicted Retained	9,887	397	6,652	1,889
Predicted Not Retained	285	9,775	88	90

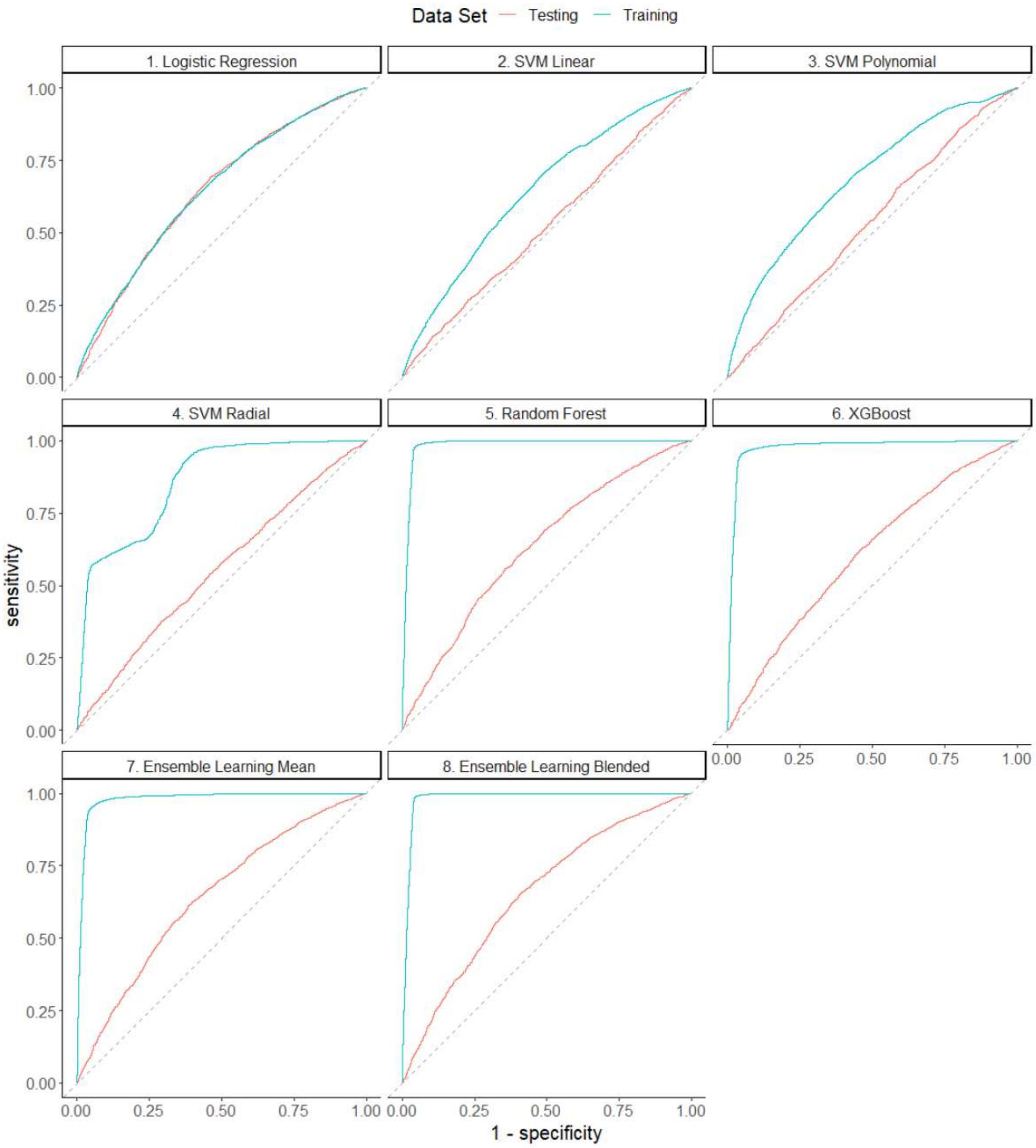


Figure 42. One-year retention ROC curves of the predictive algorithms developed on the training with downsampling techniques. The image includes the ROC curves produced from the training and testing data sets.

In comparing the distribution of the AUC values of the folds between the training and testing data sets using the upsample technique, the distribution of the values within the training values are clustered close together, as displayed in Figure 43. The ensemble models were excluded from the analysis. Across the 10 folds in the training data set, random forest model exhibited the best distribution with XGBoost model coming second and SVM using a radial basis function coming in third. Yet, comparing the distribution of the folds' AUC values on the testing data set, the random forest, XGBoost, and SVM using radial basis function exhibited AUC values closer to the other three models. Examining the predictive performance of each algorithm, a Mann-Whitney test was performed to evaluate the AUC values of the 10 folds differed between the training and testing data sets. Except for the logistic regression model, all models were found to exhibit statistically significant AUC values between each model's training and testing data sets (SVM using linear kernel, $W = 100$, $p < .001$; SVM using polynomial kernel, $W = 100$, $p < .001$, SVM using radial basis function, $W = 100$, $p < .001$; random forest, $W = 100$, $p < .001$; and XGBoost, $W = 100$, $p < .001$). These findings indicate the logistic regression model's performance between the training and testing data sets were fairly like each other, while the other five models' performance could be considered inconsistent.

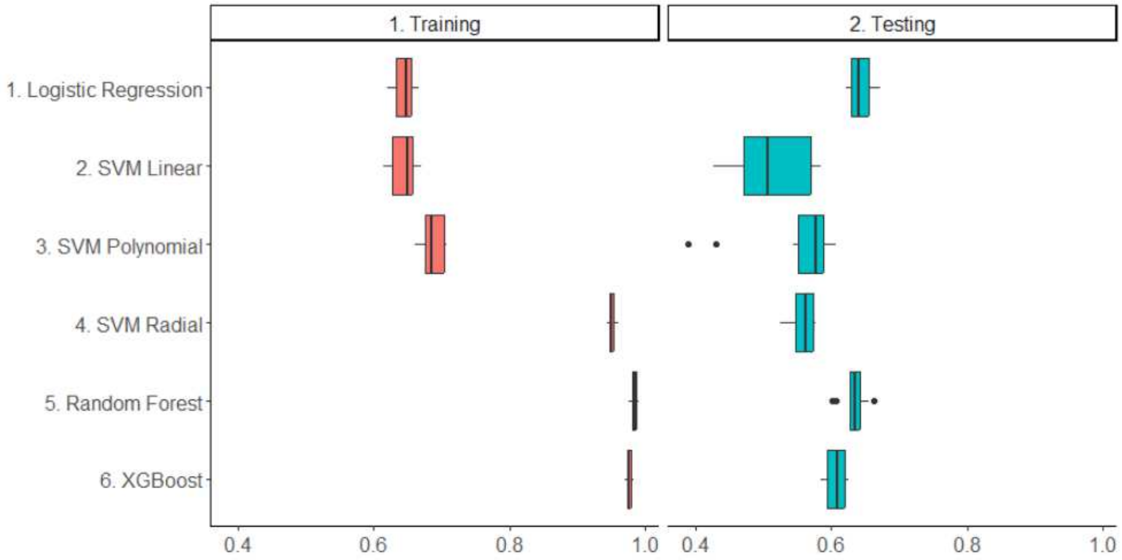


Figure 43. Boxplot of the one-year retention's predictive algorithms' AUC values from the cross-validation folds of the training and testing data sets.

The Friedman's test was used to determine if one model is more significant than the other. Examining the AUC values of the training data, the Friedman's test was to be significant with a large effect size, $\chi^2(5) = 48.60$, $p < .001$, $W = .971$. The Wilcoxon signed-rank with a Bonferroni multiple testing correction was utilized to determine which model was more significant to use. For the AUC values of the training data set, the random forest model (*Mdn* = .984) was the highest performing model and was found to be statistically significantly different from the five other models (logistic regression, *Mdn* = .647, *p* = .029; SVM using linear kernel, *Mdn* = .650, *p* = .029; SVM using polynomial kernel, *Mdn* = .685, *p* = .029; SVM using radial basis function kernel, *Mdn* = .949, *p* = .029; and XGBoost, *Mdn* = .975, *p* = .029). XGBoost model was found to be significantly different from the logistic regression model and the three SVM models. SVM using radial basis function kernel was found to be significantly different from the logistic regression model and the two other SVM models. SVM using polynomial kernel was found to be significantly different from the logistic regression and SVM using linear kernel. Using the training data set, the best model to predict retention would be the random forest.

Yet, a different story is revealed when examining the performance of the models on the testing or unseen data set. The Friedmen's test on the performance of the testing data set was conducted. The test was found to be significant with a large effect size, $\chi^2(5) = 45.20$, $p < .001$, $W = .904$. The Wilcoxon signed-rank with a Bonferroni multiple testing correction was utilized to determine which model was more significant to use. The logistic regression model (*Mdn* = .640) was the model with the highest AUC value and was found to be statistically different from the three SVM models (linear kernel, *Mdn*

= .507, *p* = .029; polynomial kernel, *Mdn* = .577, *p* = .029; and radial basis function, *Mdn* = .562, *p* = .029), the random forest model (*Mdn* = .635, *p* = .029) and the XGBoost model (*Mdn* = .608, *p* = .029). The random forest model was the model with the second highest AUC value and was found to be statistically different from the three SVM models and the XGBoost model. This finding could suggest the best performing model would be the logistic regression model.

Summary

The study involved the development of four types of predictive algorithms to analyze the significant influencing factors on three academic performance metrics within the first year. These algorithms included the linear regression, logistic regression, support vector machine, random forest, and XGBoost. To assess differences in SVM model's performance, three different kernels—linear, polynomial, and radial basis function—were utilized. The three academic performance metrics considered for the first year were first-fall GPA, first-year GPA, and one-year retention status. The data set was divided into training and testing sets. In the training data set used for building the retention algorithms, downsampling and upsampling techniques were applied in attempts to prevent the algorithms from defaulting to the majority class due to class imbalance issues within the data.

Both the training and testing data sets were divided into 10 folds for cross-validation purposes. Each predictive algorithm was developed using the training data set. Initially, tuning of the SVM, random forest, and XGBoost algorithms was conducted using resampled training data sets. Variable importance analyses were performed on the training and testing data sets, with a preference for the testing data sets to determine the factors impacting the academic performance metrics. The evaluation of models' accuracy depended on the nature of the dependent variables. For the two GPA variables, the RMSE values were calculated to assess accuracy. For the retention variable, a comprehensive evaluation of multiple accuracy metrics such as overall accuracy, F-measure, sensitivity, specificity, and AUC was utilized. Statistical analyses were conducted to assess the performance between the training and testing data sets and to determine which model showed significant differences. The RMSE values were used for the two GPA dependent variables, while AUC values were employed for the retention status in the analysis. These evaluations ensured a thorough understanding of the model performances across various metrics and data sets. Finally, statistical analyses were conducted to compare the models against each other and determine if one model performed better than the others. This step involved a robust comparison to identify the most effective algorithm among the ones developed.

Across the models developed for first-fall GPA, HS GPA consistently emerged as the most influential factor from the testing data set. The second most influential factor was the GA HOPE scholarship, except in the SVM models. Among student characteristics, no variable consistently influenced first-fall GPA across the models. Apart from GA HOPE scholarship, no other financial situation variables consistently proved influential. Neither the major grouping of the program of study nor institutional expenditures impacted first-fall GPA consistently across the models. Table 23 presents the RMSE values of the training and testing data sets for six predictive algorithms and two ensemble learning methods.

Table 23

First-Fall GPA Predictive Algorithm's RMSE Values on Training and Testing Data Set

	Training	Testing
Linear Regression	0.850	0.866
SVM Linear	0.866	0.883
SVM Polynomial	0.865	0.883
SVM Radial	0.863	0.882
Random Forest	0.839	0.850
XGBoost	0.832	0.847
Ensemble Learning Mean	0.838	0.853
Ensemble Learning Blended	0.832	0.847

Comparing the results of the training and testing data sets, XGBoost was the best-performing algorithm, with the random forest model coming in second. Neither ensemble learning method led to improvements in predictions. Utilizing Mann-Whitney's test to assess the RMSE values within cross-validation training and testing data sets, all models except linear regression consistently performed well. Furthermore, Friedman's test and Wilcoxon signed-rank test indicated no statistically significant difference between XGBoost and random forest algorithms in both data sets.

Similar to the first-fall GPA, HS GPA emerged as the most impactful factor on first-year GPA from the testing data set. The GA HOPE scholarship was the second most impactful factor outside of the SVM models. No student characteristics, major grouping, or institutional expenditures had a consistent influential impact on first-year GPA. Content mastery and readiness mean showed consistency in the SVM model using the linear kernel, random forest, and XGBoost models. Table 24 presents the RMSE values of the training and testing data sets for six predictive algorithms and two ensemble learning methods. Comparing the results of the training and testing data sets, XGBoost was the best-performing algorithm, with the random forest model coming in second. Neither ensemble learning method resulted in improvements in predictions. Utilizing

Mann-Whitney's test to assess the RMSE values within cross-validation training and testing data sets, all models were determined to perform consistently. Furthermore, Friedman's test and Wilcoxon signed-rank test indicated no statistically significant difference between XGBoost and random forest algorithms in both data sets.

Table 24

First-Year GPA Predictive Algorithm's RMSE Values on Training and Testing Data Set

	Training	Testing
Linear Regression	0.797	0.809
SVM Linear	0.812	0.824
SVM Polynomial	0.812	0.825
SVM Radial	0.811	0.824
Random Forest	0.787	0.798
XGBoost	0.779	0.794
Ensemble Learning Mean	0.786	0.798
Ensemble Learning Blended	0.780	0.794

In conducting the variable importance analysis on the testing data utilizing the models developed with no sampling modifications, the SVM models were unable to produce any variable importance due to the data imbalance within the classes of the retention status. The GA HOPE scholarship was the most influential factor, with HS GPA being the second most influential factor for the logistic regression and random forest models. The XGBoost model's results yielded the reverse order of importance. EFC was consistently ranked in the top five across the models, with the SVM model excluded. Within the random forest and XGBoost models, the five HS curriculum factors were consistently in the top 10 factors with at least moderate influences. Student characteristics, major groupings, and institutional expenditures were not found to be consistent factors impacting the retention decision across the models. No additional pre-college characteristics and financial situations were found to be influential factors across the models. Table 25 displays the accuracy metrics used to evaluate the performance of the algorithms. Based on the recommendation of Dey (2021), the AUC value was examined to assess the overall performance of the algorithms. Comparing the AUC values of the training and testing data sets, the random forest model was the top-performing algorithm, with the XGBoost model coming in second. Neither ensemble learning method resulted in any improvements in the predictions. Utilizing the Mann-Whitney test to assess the performance of the AUC values within the cross-validation training and testing data sets, all models were determined to perform consistently. Furthermore, the Friedman test and Wilcoxon signed-rank test indicated there was no statistically significant difference between the random forest and XGBoost algorithms in both data sets.

Table 25

One-Year Retention Predictive Algorithm's Evaluation Metrics on Training and Testing Data Set without Sampling Techniques

	Accuracy	Sensitivity	Specificity	F-measure	AUC
Logistic Regression					
Training	.778	.994	.023	.875	.640
Testing	.774	.992	.031	.871	.643
SVM Linear					
Training	.778	1.000	.000	.875	.554
Testing	.773	1.000	.000	.872	.503
SVM Polynomial					
Training	.778	1.000	.000	.875	.625
Testing	.773	1.000	.000	.872	.623
SVM Radial Basis Function					
Training	.778	1.000	.000	.875	.626
Testing	.773	1.000	.000	.872	.623
Random Forest					
Training	.778	.999	.003	.875	.655
Testing	.773	.999	.004	.872	.649
XGBoost					
Training	.779	.988	.046	.874	.663
Testing	.773	.988	.038	.870	.653
Ensemble Learning Mean					
Training	.779	.997	.016	.875	.662
Testing	.773	.997	.010	.871	.656
Ensemble Learning Blended					
Training	.778	.984	.055	.873	.663
Testing	.772	.987	.042	.870	.656

Except for the SVM model using the linear kernel, the remaining two SVM models were able to analyze the factors of the testing data set for the downsample models. Among the influencing factors, the GA HOPE scholarship emerged as the most significant across the models, with HS GPA being the second most influential factor. EFC in the random forest and XGBoost models were ranked third. Several HS curriculum variables were consistently ranked in the top 10 factors in the random forest and XGBoost models. No student characteristics, other pre-college characteristics, financial situations, major groupings, and institutional expenditures exhibited consistency in the impact on retention decisions. Table 26 displays the accuracy metrics used to assess algorithm performance with the downsampling technique.

Table 26

One-Year Retention Predictive Algorithm's Evaluation Metrics on Training and Testing Data Set Utilizing Downsampling Techniques

	Accuracy	Sensitivity	Specificity	F-measure	AUC
Logistic Regression					
Training	.602	.661	.543	.624	.638
Testing	.774	.992	.031	.871	.643
SVM Linear					
Training	.588	.797	.379	.659	.644
Testing	.773	1.000	.000	.872	.514
SVM Polynomial					
Training	.599	.729	.471	.645	.644
Testing	.773	.999	.001	.872	.548
SVM Radial Basis Function					
Training	.588	.797	.379	.659	.644
Testing	.773	1.000	.000	.872	.550
Random Forest					
Training	.618	.634	.602	.623	.650
Testing	.773	.999	.004	.872	.649
XGBoost					
Training	.612	.622	.603	.615	.658
Testing	.772	.989	.033	.870	.652
Ensemble Learning Mean					
Training	.616	.652	.580	.629	.656
Testing	.773	.997	.011	.872	.656
Ensemble Learning Blended					
Training	.613	.633	.594	.621	.658
Testing	.772	.987	.042	.870	.655

When comparing AUC values between training and testing data sets, the XGBoost model demonstrated the highest performance, with the random forest model following closely. The ensemble learning methods did not enhance the predictions. The three SVM models displayed inconsistent performance in cross-validation data sets according to the Mann-Whitney's test. Moreover, both Friedman's test and Wilcoxon signed-rank test revealed no statistically significant differences between the random forest and XGBoost algorithms in the training data set. In the testing data set, the logistic regression model also showed no significant difference from the random forest and XGBoost models.

Of the SVM upsample models, only the one using the polynomial kernel yielded results. There was no consistency in the influential factors across the model developed from the upsampling technique. Table 27 illustrates the accuracy metrics used for evaluating algorithms developed with the upsampling technique. Upon comparing AUC values between training and testing data sets, the XGBoost model highlighted superior performance, with the random forest model as the second-best performer. The ensemble learning method utilizing the blended technique made a slight improvements in the evaluation metrics for the training and testing data sets. When utilizing Mann-Whitney's test to assess AUC values in cross-validation training and testing data sets, the logistic regression model exhibited consistent performance across both sets. Furthermore, Friedman's test ($\chi^2(5) = 48.60$, $p < .001$, $W = .971$) and Wilcoxon signed-rank test indicated the logistic regression model displayed a statistically significant difference from the remaining five models in the testing data sets (logistic regression, *Mdn* = .647, *p* = .029; SVM using linear kernel, *Mdn* = .650, *p* = .029; SVM using polynomial kernel, *Mdn* = .685, *p* = .029; SVM using radial basis function kernel, *Mdn* = .949, *p* = .029; and XGBoost, *Mdn* = .975, *p* = .029).

Table 27

One-Year Retention Predictive Algorithm's Evaluation Metrics on Training and Testing Data Set Utilizing Upsampling Techniques

	Accuracy	Sensitivity	Specificity	F-measure	AUC
Logistic Regression					
Training	.606	.660	.552	.626	.645
Testing	.774	.992	.031	.871	.643
SVM Linear					
Training	.587	.801	.373	.660	.644
Testing	.773	1.000	.000	.872	.513
SVM Polynomial					
Training	.631	.693	.568	.652	.687
Testing	.773	.999	.001	.872	.548
SVM Radial Basis Function					
Training	.493	.400	.600	.659	.950
Testing	.773	1.000	.000	.872	.558
Random Forest					
Training	.939	.911	.966	.937	.984
Testing	.768	.967	.090	.866	.633
XGBoost					
Training	.915	.861	.969	.910	.976
Testing	.745	.910	.184	.847	.607
Ensemble Learning Mean					
Training	.919	.870	.969	.915	.976
Testing	.768	.964	.101	.865	.642
Ensemble Learning Blended					
Training	.966	.972	.961	.967	.983
Testing	.773	.987	.046	.871	.655

Chapter V

SUMMARY, DISCUSSION, AND CONCLUSIONS

Regional comprehensive universities (RCUs) will continue to play a vital role in educating the local workforce. Each fall, these institutions welcome new students, who are primarily recent HS graduates. University administrators and researchers consistently evaluate key student success metrics to gauge the overall academic performance of the student body (Arnold, 1999; Offensten & Shulock, 2010; Tai, 2020). Given the persistent challenges within higher education and the anticipated decline in the traditional-age population, optimizing student success provides a crucial foundation to stabilize enrollment at these institutions. Developing models to identify at-risk students early is essential, enabling universities to provide the necessary resources and support. Identifying at-risk students involves recognizing those likely to earn poor grades, potentially leading to probation, as well as those who may leave the institution within a year, either voluntarily or involuntarily

With technological advancements, the field of data science has expanded significantly (UW Data Science Team, 2017). In addition to traditional statistical methods, emerging tools such as random forest and XGBoost algorithms have expanded the scope of statistical analysis. Higher education institutions are increasingly utilizing data science techniques to enhance operational efficiency and improve student success metrics. The identification of at-risk students and the efficient resource allocations have played a pivotal role in optimizing academic support, influencing various aspects of higher education (Matthews, 2018).

Overview of the Study

To enhance student success, institutions must effectively identify and support at-risk students. This study evaluates the impact of secondary curriculum on the first year at RCUs. The study's population includes FTFTFs who graduated from Georgia public HS in 2018 and 2019 and attended one of four RCUs within the USG in Fall 2018 or Fall 2019. Using four predictive algorithms, the research analyzes the impact of HS curriculum factors and early factors on first-year academic performance. Variables align with Tinto's student integration model, encompassing student characteristics, pre-college characteristics, financial situations, major of study, and institutional expenditures. The study concludes by assessing the predictive accuracy of linear regression, logistic regression, support vector machine, random forest, and extreme gradient boosting algorithms, selected for commonality and popularity.

Related Literature

Researchers continuously investigate student success and attrition to improve outcomes and comprehend the factors contributing to student departure. Various frameworks, as identified in existing studies (Aljohani, 2016; Astin, 1993; Berger et al., 2012; Tinto, 1993), have been developed to identify influential factors. Among these, Tinto's (1993) integration model stands out as one of the most widely recognized and employed. The model explains students' integration into an institution and its impact on their decision to leave. Tinto (1993) outlined students' experiences around three distinct phases—separation, transition, and incorporation—during their postsecondary journey.

Failure to integrate into academic or social communities within the institution can lead to a student's departure (Tinto, 1993).

Data science. Data science has become a buzzword, experiencing a surge in popularity due to technological advancements reducing computational time (UW Data Science Team, 2017). Conway (2014, 2015) noted data science requires a delicate balance of mathematics, substantive expertise, and hacking skills to uncover hidden patterns in data sets. Driscoll (2013) highlighted a similar balance, stating "data science incorporates social science methods into its processes and job duties." Focusing on the social science aspects, Driscoll (2013) continued to argue data science goes beyond a black box analysis, emphasizing the time spent analyzing patterns in data and selecting appropriate analysis tools for the development of predictive algorithms. Optimal models are chosen after evaluating the accuracy rates and subtle differences among multiple algorithms (Calvo & Santafé, 2016; Horthorn et al., 2005). Examining whether models underfit or overfit the data is done through cross-validation techniques, such as a 10-fold method, which involves assessing the accuracy rates (Attewell & Monaghan, 2015; Bose, 2019; Drakos, 2019; Goyal, 2021; Shah, 2017; Soni, 2019; Tripathi, 2020).

Linear regression and logistic regression are two commonly used methods, serving as standard tools for analyzing relationships between independent and dependent variables (James et al., 2013). Linear regression assesses the influence of predictor variables on continuous dependent variables, while logistic regression focuses on dichotomous dependent variables (James et al., 2013; Schmidt-Thieme, 2007). Support vector machines, as an algorithm, create margins or hyperplanes to identify maximum margin patterns within the data, utilizing kernels of different shapes for optimization

(Attewell & Mongahan, 2015; Awasthi, 2020; Gandhi, 2018; James et al., 2013). Through the development of multiple decision trees, random forest algorithms are widely adopted for their flexibility with skewed data and for producing accurate results without the need for data transformation (Fernandz-Delgado et al., 2014; James et al., 2013; Ravindran, 2001; Richmond, 2016). Like random forest algorithms, the XGBoost algorithm has gained popularity for its speed and accuracy and its capability in handling both continuous and categorical variables through boosted trees and conditional random fields (Brownlee, 2016; Pafka, 2015; Xgboost Developers, 2021).

Predictive factors. Students bring specific attributes, including characteristics and prior educational experiences, to postsecondary institutions, influencing their goals and degree aspirations (Alojanhi, 2016; Tinto, 1993). Tinto (1975, 1993) emphasized the significance of environmental factors, such as family and finances, in shaping students' academic performance. Astin (1984, 1993) further proposed institutional factors, including expenditures, impacting students' integration into academic and social communities. These frameworks offer valuable insights into the complex factors affecting students' experiences and decisions within higher education institutions.

Methodology

As a nonexperimental, ex post facto, correlational research design, the study utilized data science techniques to construct forecasting and classification models predicting first-year academic performance at an RCU within the USG. Four machine learning algorithms—linear regression, logistic regression, support vector machine, random forest, and XGBoost—were employed to identify potential factors influencing first-year academic performance. The algorithms, along with an ensemble learning method, underwent evaluation based on accuracy metrics (RMSE for the GPA dependent variables and AUC values for retention dependent variable) and inferential statistics to determine the optimal predictive model.

Participants

The target population of the study consisted of the FTFTF population pursuing a bachelor's degree who graduated from a Georgia public HS in 2018 or 2019. These students subsequently enrolled in one of the four RCUs within the USG. The cohorts in the study comprised students enrolled in the Fall of 2018 (10,441 students, 47.9%) and the Fall of 2019 (11,356 students, 52.1%), totaling 21,797 students. Among the Georgia public HS, a minimum of 396 schools were represented by FTFTF students enrolled in one of the four RCUs.

Variables Studied

The first research question aimed to identify factors influencing students' first-year academic performance, encompassing student characteristics, pre-college characteristics, financial situations, program of study, and institutional expenditures. Student characteristics included gender, race and ethnicity, family educational background, and HS locale. Pre-college characteristics comprised HS GPA, admission test scores, EOC subject areas' proficiency rates, and CCRPI content mastery and readiness mean score. Financial situations covered EFC, GA HOPE scholarship dollars, PELL grant dollars, and loan dollars. Institutional expenditures incorporated instruction, research, public service, academic support, student services, institutional support, and other core expenditures.

Student characteristics. The student characteristics considered in the study were gender, race and ethnicity, family educational background, and the locale of graduating HS. Existing research on gender suggested females are more likely to enroll in postsecondary institutions and tend to academically outperform their male counterparts (Buchmann & DiPrete, 2006; Flores & Park, 2013; Jacob, 2002; Lohfink & Paulsen, 2005; Tinto, 1975, 1993). With the increasing diversity of the postsecondary population (AACU, 2019), underrepresented minority students were more likely to underperform when compared to White and Asian students (Fischer, 2007; Odell et al., 2005; Seidman, 2007; Stewart et al., 2015; Tinto, 1975; Tinto, 1993). Students who are the first in their family to pursue education beyond HS tended to lag academically compared to those with at least one parent holding a bachelor's degree (Ishitani, 2003, 2006; Lohfink & Paulsen, 2005). While students from rural areas exhibit more hesitation to attend postsecondary institutions, they often exhibited a lower academic foundation, resulting in lower academic performance in postsecondary settings (Corely et al., 1991; Fischer, 2007; Lumina Foundation, 2019; Provasnik et al., 2007; Schultz, 2004; Velez, 2014).

Pre-college characteristics. In addition to HS GPA and admissions test scores, the study considered satisfaction of college preparatory curriculum requirements, the number of advanced standing hours, and high school curriculum factors, which encompassed proficiency level rates of the four HS subject areas and content mastery and readiness mean score. High school GPA and admissions test scores were widely acknowledged as significant factors influencing postsecondary academic performance. Consistent research has emphasized HS GPA as the primary factor affecting academic performance within the first year and beyond to graduation (Allensworth & Clark, 2020;

Bridgeman et al., 2008; Chen & St. John, 2011; Spady, 1970, 1971; Tinto, 1975, 1988, 1997). However, research on the influence of admissions test scores on academic performance has produced mixed results (Allensworth & Clark, 2020; Bowen et al., 2009; Korbin et al., 2008; Lotkowski et al., 2004; Noble & Sawyer, 2002; Rothstein, 2004; Stewart et al., 2015). The impact of HS curriculum, including the availability of more rigorous coursework, has been consistently shown to prepare students for postsecondary education (Choy, 2002; DeNicco et al., 2015; Horn & Kojaku, 2001; McDonough, 1997; Provasnik et al., 2007; Tinto, 1993).

Financial situations. Financial situation factors in the study included expected family contribution, GA HOPE and Zell Miller scholarships, PELL grant, federal subsidized and unsubsidized loans, and other loans. The research indicated the financial burden of paying for postsecondary education, particularly for students from lower socioeconomic backgrounds, was associated with lower academic performance and a higher risk of departure (Chen & DesJardins, 2008; Chen & St. John, 2011; St. John et al., 2005; Tinto, 1975, 1982; Velez, 2014). Financial aid, ranging from merit-based aid to need-based aid, helps alleviate this burden. Merit-based aid, such as the GA HOPE scholarship, exhibits positive influence on academic performance and persistence until graduation, mainly due to a commitment to acquire and maintain the aid type (Chen, 2012; Georgia Student Finance Commission, 2021a, 2021b; Gross et al., 2015; Henry et al., 2004; Suggs, 2016; Stater, 2009). Need-based aid, like the PELL grant, slightly reduces the likelihood of underperforming and not persisting (Chen & DesJardins, 2008; Chen, 2012; Gross et al., 2015). As an additional form of financial aid, taking out loans to cover the cost of attendance, has mixed effects on academic performance and persistence, as indicated by research (Bettinger, 2004; Gross et al., 2015; Hanson, 2020; St. John et al., 2005).

Program of study. Programs of study were combined into natural classification groupings based on the Classification of Instructional Programs (CIP) codes. The selected major or program of study exhibited different influences on academic performance. Students aligning with the inherent tendencies of their major were more likely to succeed academically and exhibit higher retention rates (Leppel, 2001). Research on students majoring in STEM programs indicated these students were more likely to earn slightly lower GPAs due to the courseload rigor but persist at higher rates than non-STEM majors (Gansemer-Topf et al., 2017).

Institutional expenditures. Institutional expenditures per full-time equivalent variables were collected from the National Center for Education Statistics' Integrated Postsecondary Education Data System (IPEDS) website. Expenditures included academic support, institutional support, instruction, student services support, public service, research, and all other expenditure categories. The research outcomes were found to be varied regarding which type of expenditure demonstrated an impact on first-year academic performance (Chen, 2012; Gansemer-Topf & Schuch, 2006; Ryan, 2004; Webber & Ehrenberg, 2009).

Procedures

After separately analyzing each FTFTF cohort to reveal their similarities and differences, the cohorts were combined to address the research questions. The data set was split to allocate 60% (*N* = 13,078) of the data as the training data set and the remaining 40% (*N* = 8,719) as the testing data set. Statistical considerations and assumptions were assessed for the training data set, with necessary corrections made for violations. Techniques, such as variable modifications to reduce multicollinearity and Yeo-Johnson transformation for normality corrections, were applied. To address data imbalances in the retention models, additional sampling techniques were applied to the training data set. Models for each dependent variable were developed and evaluated on the training and testing data sets. To answer the first research question regarding the influential factors on first-year academic performance, emphasis was placed on the results produced from the testing data set. For the second research question, 10-fold cross-validations were applied to the training and testing data sets to analyze predictive power. Statistical analyses were conducted to compare the predictive performance of each model between the training and testing data sets and to identify any significant differences between the algorithms.

Summary of Findings

The study's aim was to identify factors influencing first-year academic performance and determine the optimal predictive algorithm for two research questions. The findings offer valuable insights for postsecondary administrators and policymakers by facilitating the early identification of at-risk students. This critical information enables the effective modification and reallocation of academic and student support services, ensuring targeted assistance for identified at-risk students to succeed in their first year. The selected predictors were tailored to the characteristics of traditional-age students starting their postsecondary educational journey.

First research question. Are student characteristics, precollege characteristics—including high school curriculum quality, financial situations, major or program of study, and institutional financial expenditures significant predictors in first-time, full-time freshmen's first-fall GPA, first-year GPA, and one-year retention status?

First-fall GPA. Across the models, no consistent patterns on the testing data set emerged regarding the impact of student characteristics on first-fall GPA. With 55.5% of students being females, gender was identified as a significant influencer in the linear regression model; however, in the remaining models, the gender variable exhibited weak importance, with values close to 0. Approximately 51.9% of students were identified as White, and 48.1% were categorized as underrepresented minorities. Interestingly, the models did not consistently identify race and ethnicity as an influential factor to first-fall GPA. Specifically, the SVM models did not recognize these as influential variables, whereas the linear regression and XGBoost models indicated weak influences. First-generation status was only a minimally influential factor in the XGBoost model. Among student characteristics, HS locale was identified as having a weak influence on the dependent variable in the linear regression models.

For each algorithm, HS GPA emerged as the most influential factor affecting first-fall GPA. The linear regression model did not attribute any significance to first-fall GPA from admissions test scores. Conversely, the variable was identified to have a medium influence in the XGBoost model. The number of advanced standing AP hours exhibited a medium impact and IB hours demonstrated a slight influence only in the linear regression model. Other advanced standing hours were not identified as factors impacting first-fall GPA. The college preparatory curriculum variable did not contribute significantly to first-fall GPA. Regarding HS curriculum variables, the content mastery and readiness mean held significance with a small influence in the linear regression and random forest models, while exhibiting a medium influence in the XGBoost model. Proficiency levels differences in all four subject areas from the content mastery and readiness mean were found to have a medium impact on first-fall GPA in the XGBoost model. The social studies proficiency levels difference exhibited a medium influence in the linear regression model, while the English proficiency levels difference exhibited a small impact. Also, the English proficiency levels difference exhibited a small influence in the random forest model.

Among the financial factors considered, the GA HOPE scholarship consistently emerged as the influential factor across all models affecting first-fall GPA. The SVM models identified the contribution as the weakest among the models, with the strength of impact relatively the same across the three kernels. Related to the GA HOPE scholarship, the Zell Miller indicator was deemed a significant factor with a medium influence in the linear regression model and a small influence in the random forest model. Excluding the SVM models, the EFC demonstrated small influences on first-fall GPA in the linear regression and random forest model but a medium influence in the XGBoost models. The XGBoost models identified the PELL grant as exhibiting a small influence. The other financial considerations did not exhibit any influence on the first-term GPA.

The major groupings variable emerged as significant solely in the linear regression and XGBoost models, contributing a small influence on first-fall GPA. Outside the linear regression model, expenditures did not contribute any impact on the first-fall GPA.

Instruction expenditures, along with academic and institutional support expenditures, exhibited a small influence in the linear regression model.

First-year GPA. Similar to the outcomes for the first-fall GPA, no consistent patterns emerged from the testing data set for the first-year GPA regarding student characteristics. Only the linear regression algorithm identified gender as a strong influencing factor on the first-year GPA. The race and ethnicity and first-generation variables were not identified as factors impacting first-year GPA across all models. The locale of the graduating HS was only found to exhibit a small influencing factor in the linear regression model.

Within the pre-college characteristics, the dominant influential factor on the first-year GPA was the HS GPA across all models. Admissions test scores were identified to exhibit small influences in the XGBoost model. The number of advanced AP hours contributed close to a moderate influence in the linear regression model but exhibited barely any influence in the remaining models. Neither the remaining advanced standing hours nor the number of college preparatory curriculum satisfactions were found to have a consistent influence across the models. The mean content mastery and readiness variable was found to exhibit a small influence in the linear regression, random forest, and XGBoost models. The content mastery and readiness mean exhibited small influences in the linear regression, random forest, and XGBoost models. While all four subject area proficiency levels differences were found to exhibit a small impact in the XGBoost model, only the social studies and English proficiency levels differences from the content mastery and readiness mean were found to exhibit a small influence in the linear regression model.

All models identified the GA HOPE scholarship as the second most influential factor on the first-year GPA, with random forest and XGBoost indicating the strength of the variable was close to the HS GPA. Only the linear regression model identified the Zell Miller indicator variable exhibited a small influence on the first-year GPA. The EFC was found to exhibit a small influence on the first-year GPA in all models except the SVM models. The remaining financial aid variables did not influence the first-year GPA. For the first-year GPA, the major groupings variable emerged as significant solely in the linear regression and XGBoost models, contributing a small influence. Exhibiting a small influence only in the linear regression, academic and institutional support was the only expenditure impacting the first-year GPA.

One-year retention status. The SVM models applied to the testing data set failed to identify significant factors, even with corrections for class imbalance during model development. Notably, the SVM model utilizing polynomial kernels produced an insightful analysis highlighting some variables influencing the retention decision. Interestingly, both analyses from the SVM model identified the number of CLEP credits and other advanced standing hours as the top two variables impacting retention. However, given the inconsistency in generating influential factors across the SVM models, they were excluded from the final results pertaining to the factors influencing retention status.

Upon analyzing the testing data set, the gender factor emerged with a moderate influence in the linear regression model, whereas it demonstrated a slight impact in the three XGBoost models. Markedly, the linear regression and XGBoost models without modifications revealed race and ethnicity exhibited a moderate influence, contrasting with the XGBoost downsampled and upsampled models' findings indicating only a minor impact. The first generation factor displayed a moderate influence solely in the linear regression model. Furthermore, the locale of the graduating HS was identified as having a minor impact on the retention decision in the linear regression model and across the three XGBoost models.

Across the various models, HS GPA emerged as a consistently potent determinant of the retention decision. Specifically, in the linear regression model, admissions test scores were marginally more influential than HS GPA, a contrast observed in other algorithms where test scores exerted a lesser impact compared to HS GPA. In the linear regression model, all advanced standing hours demonstrated minor influences. Whereas in the random forest and XGBoost no modification models, as well as the downsampled and upsampled XGBoost models, only AP hours exhibited a modest impact. The number of satisfied college preparatory requirements consistently showed no influence on the retention decision. Within the XGBoost models, all five HS curriculum variables demonstrated an impact on the retention decision, with a strong effect in the no modification and upsample models and a moderate effect in the downsampled model. Conversely, the random forest models attributed small influences to the five HS curriculum variables. In the logistic regression model, only the proficiency differences in science, social studies, and English proficiency levels were identified as exerting minor influences on the retention factor.

The GA HOPE scholarship, identified as a pivotal financial factor, emerged as one of the top two influencers in all models, except the XGBoost upsampled model. In the linear regression model, the Zell Miller factor displayed only a minor influence.

Conversely, the EFC, while ranked as the top factor in the XGBoost upsampled model, consistently demonstrated a moderate impact on the retention factor across all models. The PELL grant exhibited a medium influence in the linear regression, random forest with no modification, and downsampled models. However, in the XGBoost three models and the random forest upsampled model, the PELL grant factor was identified as having a minor impact on the dependent variable. The remaining financial situation variables exhibited nothing beyond a slightly small impact, highlighting the nuanced significance of specific financial factors in predicting retention outcomes.

With the exception of the XGBoost upsampled model, the major grouping variable did not exhibit any discernible influence on the retention decision. In the XGBoost upsampled model, all expenditures demonstrated an impact on the retention variable. In models excluding XGBoost upsampled, public service and research expenditures, student support services, and instruction expenditures were identified as having small influences on the retention decision.

Second research question. Does one machine learning algorithm (linear or logistic regression, support vector machine, random forest, and extreme gradient boosting) or an ensemble learning algorithm produce a higher accuracy based on the evaluation metrics for accuracy in examination of first-fall GPA, first-year GPA, and one-year retention status?

First-fall GPA. Six models and two ensemble learning methods were assessed for accuracy metrics through 10-fold cross-validations on training and testing data sets. Statistical analysis was conducted to assess performance differences across the data sets and a comparison amongst the models after visual inspection. The linear regression model, implemented without tuning for optimal performance using the lm engine, demonstrated significant results with a small effect size (R^2 = .303, *adj* R^2 = .302, *F*(29, 13,041) = 195.600, $p < 0.001$), explaining 30.3% of variance with an RMSE value of 0.848 of the training data set. For the testing data set, the linear regression model demonstrated significant results (R^2 = .283, *adj* R^2 = .281, *F*(29, 8,681) = 118.400, $p < 0.001$), explaining 28.3% of variance with an RMSE value of 0.864. The evaluation of the 10-folds indicated an RMSE of 0.850 for the training data set and 0.866 for the testing data set. Mann-Whitney's test indicated statistically significant differences in the model's accuracy metrics performance between the training and testing data sets.

The SVM model with a linear kernel was developed using the kernlab engine, tuned with a cost of 0.304 and a margin of 0.194. It explained 29.9% of the variance, yielding an RMSE of 0.866. Across 10-folds, performance metrics were consistent with values of 0.866 for the training set and 0.883 for the testing set, based on the Mann-Whitney's test. The SVM model was retuned to use a polynomial kernel with optimal parameters of a cost of 14.782, degree of 3, scale factor of 0.0001, and margin of 0.188, explaining 30.0% of the variance with an RMSE value of 0.865. The Mann-Whitney's test indicated the consistency of the model (training's RMSE = 0.865, testing's RMSE = 0.883). The radial basis function kernel was then applied, producing an optimal model with a cost of 19.460, sigma of 0.0005, and margin of 0.123, explaining 30.5% of the variance with an RMSE value of 0.863. Performance metrics were consistent across 10-folds for both training (RMSE = 0.863) and testing (RMSE = 0.882) data sets, confirmed by the Mann-Whitney's test.

Developed with the ranger engine, the random forest model was tuned for mtry, trees, and min_n. The optimal model achieved an RMSE of 0.832, explaining 33.0% of the variance. The tuned model comprised 1,580 trees, mtry of 9, and min_n of 37. Across 10-folds, RMSE values were consistent based on the Mann-Whitney's test (training = 0.839, testing = 0.851). For the final model, XGBoost utilized the xgboost engine with tuned parameters of 1,804 trees, mtry of 10, min_n of 3, tree depth of 8, learn rate of 0.005, loss reduction of 4.797, and sample size of 0.106. The optimal XGBoost model achieved an RMSE of 0.832, with consistent performance metrics across 10-folds (training = 0.832, testing = 0.847), based on the Mann-Whitney's test. Friedman's test and Wilcoxon signed-ranked test favored XGBoost as the best model on the training set, though no statistically significant differences were found between XGBoost and the random forest models in the testing data set.

The ensemble learning methods employing the mean approach exhibited an RMSE of 0.838 for the training set and 0.853 for the testing set. The blended approach yielded an RMSE of 0.832 for the training set and 0.847 for the testing set. Neither ensemble methods resulted in any enhancements to the predictions.

First-year GPA. Six models and two ensemble learning methods were assessed for accuracy metrics through 10-fold cross-validations on training and testing data sets. Statistical analysis was conducted to assess performance differences across the data sets and a comparison amongst the models after visual inspection. The linear regression model, implemented without tuning for optimal performance using the lm engine, demonstrated significant results with a small effect size (R^2 = .325, *adj* R^2 = .324, F(29, 12,986) = 215.900, p < 0.001), explaining 32.5% of variance with an RMSE value of

0.795 of the training data set. For the testing data set, the linear regression model demonstrated significant results ($R^2 = .312$, *adj* $R^2 = .310$, $F(29, 8{,}663) = 118.4$, $p < 0.001$), explaining 31.2% of variance with an RMSE value of 0.807. The evaluation of the 10-folds indicated an RMSE of 0.850 for the training data set and 0.866 for the testing data set. Mann-Whitney's test indicated statistically significant differences in the model's accuracy metrics performance between the training and testing data sets.

The SVM model with a linear kernel was developed using the kernlab engine, tuned with a cost of 0.304 and a margin of 0.194. It explained 29.9% of the variance, yielding an RMSE of 0.866. Across 10-folds, performance metrics were consistent with values of 0.866 for the training set and 0.883 for the testing set, based on the Mann-Whitney's test. The SVM model was retuned to use a polynomial kernel with optimal parameters of a cost of 14.782, degree of 3, scale factor of 0.0001, and margin of 0.188, explaining 30.0% of the variance with an RMSE value of 0.865. The Mann-Whitney's test indicated the consistency of the model (training's RMSE = 0.865, testing's RMSE = 0.883). The radial basis function kernel was then applied, producing an optimal model with a cost of 19.460, sigma of 0.0005, and margin of 0.123, explaining 30.5% of the variance with an RMSE value of 0.863. Performance metrics were consistent across 10-folds for both training (RMSE = 0.863) and testing (RMSE = 0.882) data sets, confirmed by the Mann-Whitney's test.

Developed with the ranger engine, the random forest model was tuned for mtry, trees, and min_n. The optimal model achieved an RMSE of 0.832, explaining 33.0% of the variance. The tuned model comprised 1,580 trees, mtry of 9, and min_n of 37. Across 10-folds, RMSE values were consistent based on the Mann-Whitney's test (training = 0.839, testing = 0.851). For the final model, XGBoost utilized the xgboost engine with tuned parameters of 1,804 trees, mtry of 10, min_n of 3, tree depth of 8, learn rate of 0.005, loss reduction of 4.797, and sample size of 0.106. The optimal XGBoost model achieved an RMSE of 0.832, with consistent performance metrics across 10-folds (training = 0.832, testing = 0.847), based on the Mann-Whitney's test. Friedman's test and Wilcoxon signed-ranked test favored XGBoost as the best model on the training set, though no statistically significant differences were found between XGBoost and the random forest models in the testing data set.

After tuning the cost and margin for the SVM model using a linear kernel, the optimal model exhibited a cost of 0.304 and a margin of 0.194, explaining 32.1% of the variance with an RMSE of 0.812. In the 10-fold cross-validation, the mean RMSE was 0.812 for the training set and 0.825 for the testing set. Mann-Whitney's test revealed no significant difference in the model's performance between the two data sets. The SVM model with a polynomial kernel was tuned to optimal parameters, resulting in a cost of 14.782, degree of 3, scale factor of 0.0001, and margin of 0.188. The optimal model explained 32.2% of the variance with an RMSE of 0.812. Mann-Whitney's test indicated no significant difference in the evaluation metrics between the training (0.812) and testing (0.825) data sets. The tuned SVM model with a radial basis function kernel exhibited an optimal cost of 19.46, sigma of 0.0005, and a margin of 0.123. The optimal model accounted for 32.5% of the variance with an RMSE of 0.812. Mean RMSE values of 0.812 for the training set and 0.824 for the testing set indicated no significant difference between the data sets, based on the Mann-Whitney's test.

Developing the optimal random forest model involved tuning parameters for mtry, trees, and min_n. The optimal model, with 1,580 trees, mtry of 9, and min_n of 37, explained 34.0% of the variance with an RMSE value of 0.787. Performance metrics indicated mean RMSE values of 0.787 for the training set and 0.798 for the testing set were consistent with no significant differences, based on the Mann-Whitney's test. For the final algorithm, XGBoost optimal model consisted of 1,264 trees, mtry of 22, min_n of 20, tree depth of 5, learn rate of 0.007, loss reduction of 0.004, and sample size of 0.285. The optimal model accounted for 35.2% of the variance with an RMSE of 0.779. Mean RMSE values for the training and testing sets were 0.779 and 0.794, respectively, with no statistical difference according to Mann-Whitney's test. On the training set, XGBoost was found to be the optimal based on Friedman's test and Wilcoxon signed-ranked test, while for the testing data set no statistical significance was found between XGBoost and the random forest model in their overall performance.

The ensemble learning methods employing the mean approach exhibited an RMSE of 0.786 for the training set and 0.789 for the testing set. The blended approach yielded an RMSE of 0.780 for the training set and 0.794 for the testing set. Neither ensemble methods resulted in any enhancements to the predictions.

One-year retention status. A total of six algorithms with two ensemble learning methods were initially developed on the training data without addressing data imbalance. Subsequently, the algorithms were redeveloped with the data sets undergoing both downsampling and upsampling techniques. Performance metrics, including overall accuracy, sensitivity, specificity, f-measure score, and AUC value, were evaluated for each algorithm across the two data sets. Following Dey's (2021) recommendation, the

AUC value was analyzed to assess the models' performance across the training and testing data sets. The performance metrics of the retention models are displayed in Table 28 to Table 30.

The initial logistic regression model, developed with the glm engine, did not demonstrate significance based on the Hosmer-Lemeshow's goodness-of-fit test ($\chi^2(8)$ = 12.189, p = .143), indicating that the model was a good fit. The model accounted for 4.7% and 7.4% of the variance with McFadden's pseudo-R^2 and Nagelkerke's pseudo-R^2, respectively. The logistic regression model developed from the downsampled training data was also not found to be significant (Hosmer-Lemeshow's $\chi^2(8)$ = 6.985, p = .538) and explained 5.1% and 9.1% of the variance with McFadden's pseudo-R^2 and Nagelkerke's pseudo-R^2. The logistic regression model from the upsampled training data was found to be significant (Hosmer-Lemeshow's, $\chi^2(8)$ = 17.555, p = .025) and explained 5.0% and 8.9% of the variance with McFadden's pseudo-R^2 and Nagelkerke's pseudo-R^2. Based on the testing data set, the model was not significant (Hosmer-Lemeshow's, $\chi^2(8)$ = 11.158, p = .193) and explained 4.8% and 7.7% of the variance with McFadden's pseudo-R^2 and Nagelkerke's pseudo-R^2.

Table 28

One-Year Retention Predictive Algorithm's Evaluation Metrics on Training and Testing Data Set without Sampling Techniques

	Accuracy	Sensitivity	Specificity	F-measure	AUC
Logistic Regression					
Training	.778	.994	.023	.875	.640
Testing	.774	.992	.031	.871	.643
SVM Linear					
Training	.778	1.000	.000	.875	.554
Testing	.773	1.000	.000	.872	.503
SVM Polynomial					
Training	.778	1.000	.000	.875	.625
Testing	.773	1.000	.000	.872	.623
SVM Radial Basis Function					
Training	.778	1.000	.000	.875	.626
Testing	.773	1.000	.000	.872	.623
Random Forest					
Training	.778	.999	.003	.875	.655
Testing	.773	.999	.004	.872	.649
XGBoost					
Training	.779	.988	.046	.874	.663
Testing	.773	.988	.038	.870	.653
Ensemble Learning Mean					
Training	.779	.997	.016	.875	.662
Testing	.773	.997	.010	.871	.656
Ensemble Learning Blended					
Training	.778	.984	.055	.873	.663
Testing	.772	.987	.042	.870	.656

The Mann-Whitney's test revealed consistency in the AUC values with no significant difference found within the model developed with no sampling modifications. The downsampled model was found to be consistent with no significant differences between the training and testing data sets per the Mann-Whitney's test. Additionally, the downsampled AUC across 10-fold cross-validations suggested similar performance to all models except XGBoost on the training data set; yet logistic regression model's performance was found to be like the random forest and XGBoost on the testing data set. AUC values were consistent between training and testing data sets, based on the Mann-Whitney's test from the upsampled model. The Friedman's test and Wilcoxon signed-ranked test identified logistic regression as statistically the best algorithm for AUC performance on the testing data set.

The SVM model with a linear kernel was initially developed and resulted in an optimal model with a cost of 0.008 and a margin of 0.120. The SVM model was retuned using the downsampled data set, with the optimal model exhibiting a cost of 0.002 and a margin of 0.122. The SVM model was retuned using the upsampled data set, resulting in an optimal model with a cost of 0.001 and a margin of 0.067. Examining the AUC values across the 10-folds, the Mann-Whitney’s test revealed that the initial, downsampled, and upsampled SVM models using the linear kernel did not consistently perform across the training and testing data sets.

The SVM model using the polynomial kernel was initially developed, with the optimal model exhibiting a cost of 0.115, a degree of 1, a scale factor of 0.001, and a margin of 0.034. The SVM model was retuned using the downsampled data set, where the optimal model exhibited a cost of 0.004, a degree of 3, a scale factor of 0.076, and a margin of 0.089. The SVM model was retuned using the upsampled data set, with the optimal model exhibiting a cost of 0.004, a degree of 3, a scale factor of 0.076, and a margin of 0.089. In examining the AUC values across the 10-folds, the Mann-Whitney’s test revealed the initial SVM model's performance was not significantly different, indicating consistency between the training and testing data sets. However, the performance of the downsampled and upsampled SVM models using the polynomial kernel did not exhibit consistent performance across the training and testing data sets based on the Mann-Whitney’s test.

Table 29

One-Year Retention Predictive Algorithm's Evaluation Metrics on Training and Testing Data Set Utilizing Downsampling Techniques

	Accuracy	Sensitivity	Specificity	F-measure	AUC
Logistic Regression					
Training	.602	.661	.543	.624	.638
Testing	.774	.992	.031	.871	.643
SVM Linear					
Training	.588	.797	.379	.659	.644
Testing	.773	1.000	.000	.872	.514
SVM Polynomial					
Training	.599	.729	.471	.645	.644
Testing	.773	.999	.001	.872	.548
SVM Radial Basis Function					
Training	.588	.797	.379	.659	.644
Testing	.773	1.000	.000	.872	.550
Random Forest					
Training	.618	.634	.602	.623	.650
Testing	.773	.999	.004	.872	.649
XGBoost					
Training	.612	.622	.603	.615	.658
Testing	.772	.989	.033	.870	.652
Ensemble Learning Mean					
Training	.616	.652	.580	.629	.656
Testing	.773	.997	.011	.872	.656
Ensemble Learning Blended					
Training	.613	.633	.594	.621	.658
Testing	.772	.987	.042	.870	.655

The SVM model using the polynomial kernel was initially developed, with the optimal model exhibiting a cost of 0.115, a degree of 1, a scale factor of 0.000, and a margin of 0.034. The SVM model was retuned using the downsampled data set, where the optimal model exhibited a cost of 0.004, a degree of 3, a scale factor of 0.076, and a margin of 0.089. The SVM model was retuned using the upsampled data set, with the optimal model exhibiting a cost of 0.004, a degree of 3, a scale factor of 0.076, and a margin of 0.089. In examining the AUC values across the 10-folds, the Mann-Whitney's test revealed the initial SVM model's performance was not significantly different, indicating consistency between the training and testing data sets. However, the performance of the downsampled and upsampled SVM models using the polynomial kernel did not exhibit consistent performance across the training and testing data sets based on the Mann-Whitney's test.

The SVM model using the radial basis function kernel was developed with the optimal model exhibiting a cost of 0.024, radial basis function sigma of 0.001, and a margin of 0.048. The SVM model was retuned using the downsampled data set, in which the optimal model exhibited a cost of 19.463, radial basis function sigma of 0.000, and a margin of 0.123. The SVM model was retuned using the upsampled data set, in which the optimal model exhibited a cost of 0.019, radial basis function sigma of 0.467, and a margin of 0.013. In examining the AUC values of the 10-folds, the Mann-Whitney's test revealed the initial SVM model's performance was not significantly different, indicating consistent performance between the training and testing data sets. However, the performance of the downsampled and upsampled SVM models using the radial basis function kernel did not exhibit consistent performance across the training and testing data sets based on the Mann-Whitney's test.

Table 30

One-Year Retention Predictive Algorithm's Evaluation Metrics on Training and Testing Data Set Utilizing Upsampling Techniques

	Accuracy	Sensitivity	Specificity	F-measure	AUC
Logistic Regression					
Training	.606	.660	.552	.626	.645
Testing	.774	.992	.031	.871	.643
SVM Linear					
Training	.587	.801	.373	.660	.644
Testing	.773	1.000	.000	.872	.513
SVM Polynomial					
Training	.631	.693	.568	.652	.687
Testing	.773	.999	.001	.872	.548
SVM Radial Basis Function					
Training	.493	.400	.600	.659	.950
Testing	.773	1.000	.000	.872	.558
Random Forest					
Training	.939	.911	.966	.937	.984
Testing	.768	.967	.090	.866	.633
XGBoost					
Training	.915	.861	.969	.910	.976
Testing	.745	.910	.184	.847	.607
Ensemble Learning Mean					
Training	.919	.870	.969	.915	.976
Testing	.768	.964	.101	.865	.642
Ensemble Learning Blended					
Training	.966	.972	.961	.967	.983
Testing	.773	.987	.046	.871	.655

For the initial random forest algorithm, the optimal model was tuned using the ranger engine, resulting in 1,710 trees, an mtry value of 2, and a minimum value of 29. The random forest was retuned utilizing the downsampled training data set and exhibited 1,306 trees, an mtry value of 3, and a minimum value of 24. The final random forest model, retuned using the upsampled training data set, comprised 731 trees, an mtry value of 24, and a minimum value of 3. The Mann-Whitney's test indicated only the performance of the random forest tuned using the upsampled data set was statistically different, suggesting inconsistency between the training and testing data sets. The Friedman's test and Wilcoxon signed-ranked test identified the initial random forest model's performance did not differ from the XGBoost model on both the training and testing data sets. For the model tuned with the downsampled training data set, the performance of the algorithm was found not to be significantly different from the logistic regression and XGBoost models based on the Friedman's test and the Wilcoxon signed-ranked test. The upsampled model's performance was not consistent between the training and testing data sets based on the Mann-Whitney's test. Based on the training data set, the Friedman's test and Wilcoxon signed-ranked test indicated the random forest was the best algorithm; however, the analysis on the testing data set did not reveal the same results.

Using the xgboost engine to tune to optimal performance, the initial XGBoost algorithm exhibited 1,264 trees with an mtry of 22, minimum observations of 20, tree depth of 5, learning rate of 0.007, loss reduction of 0.004, and sample size of 0.285. After retuning with the downsample, the revised XGBoost algorithm exhibited 123 trees with an mtry of 7, minimum observations of 17, tree depth of 9, learning rate of 0.020, loss reduction of 0.006, and sample size of 0.370. For the upsample model, the revised XGBoost algorithm exhibited 1,686 trees with an mtry of 29, minimum observations of 9, tree depth of 12, learning rate of 0.058, loss reduction of 0.116, and sample size of 0.876. The Mann-Whitney's test only found the upsample model to be statistically different, indicating inconsistent performance between the training and testing data sets. The Friedman's test and Wilcoxon signed-ranked test found the XGBoost model to be the best-performing model on the initial data set; yet the performance was not significantly different from the random forest on the training and testing data sets. For the downsample model, the XGBoost model's performance was only found to be similar to the logistic regression and random forest when conducting the Friedman's test and Wilcoxon signed-ranked test on the testing data set.

Utilizing the initial models, the ensemble learning methods exhibited varied accuracy metrics. The mean method results between the training and testing data sets without sampling modifications exhibited similar rates, while the blended method exhibited slight differences. While the results in the training data set in the downsampled models indicated slight improvements, both methods of the ensemble learning resulted in the testing data set exhibiting very limited improvements in the predictive power. The ensemble learning method from the upsampled models exhibited very noticeable improvements in the training data set's predictive power but reverted back to bias toward the majority class in the testing data set. The blended method exhibited slight improvements in the accuracy metrics in the training testing data sets.

Discussions of Findings

The study aimed to identify key factors influencing the first-year academic performance of FTFTFs enrolled in GA's RCUs. The effectiveness of predictive algorithms in forecasting academic outcomes was also evaluated. Additionally, the research incorporated the utilization of ensemble learning methods to enhance the predictions. Significant factors influencing first-year academic performance were identified, with at least one predictive algorithm demonstrating superior performance compared to others.

First research question. Continuing to play a crucial role, factors such as student characteristics, pre-college characteristics, and financial situations significantly influence first-year academic performance. Tinto's (1993) integration model highlights the impact of internal and external factors across three phases as students integrate into the academic and social communities of their postsecondary institution. Particularly, HS GPA emerges as the most influential factor in first-fall and first-year GPA, with a notable impact on one-year retention status. Higher HS GPA correlates with students' ability to successfully manage multiple subjects in HS, potentially translating the needed skills for continued academic success and integration into the postsecondary community, resulting in higher retention rates. Conversely, lower HS GPA is associated with academic struggles, lower first-year grades, and a higher likelihood of departure. Except for the SVM models, the inclusion of HS curriculum variables affected first-fall GPA, first-year GPA, and one-year retention status depending on the model. The mean content mastery and readiness scores consistently influenced these academic performance metrics with the highest impact noted in the XGBoost models. Although not as potent as HS GPA, students from HS with higher content mastery and readiness rates are more likely to be academically prepared, improving their chances of success. Conversely, students from HS with lower content mastery and readiness rates may face challenges adjusting to the rigor of postsecondary coursework, potentially impacting their academic integration. Proficiency levels in the four subject areas and their contributions to students' performance and retention largely depended upon the predictive algorithm, with the largest impact found in the XGBoost models.

The Georgia HOPE scholarship emerged as the second most influential factor in first-fall and first-year GPAs and overwhelmingly the strongest factor in retention decisions. Despite being linked to an HS GPA requirement, recipients of the merit-based financial aid were more likely to succeed academically and retain, eliminating the burden of the full cost of attendance on the student. In contrast, non-recipients, even if academically successful in the first year, were less likely to remain at the institution, potentially due to the financial burden of covering attendance costs. The amount of the PELL grant and EFC exhibited moderate impacts on first-year academic performance, excluding SVM models. These variables reflect the capacity to cover the full cost of attendance, with lower family contributions potentially leading to one-year departures and lower academic performance. Even with a high HS GPA and the GA HOPE scholarship, students from families with lower expected contributions might leave the institution due to the financial burden of continued attendance costs.

Second research question. The study assessed the performance metrics of predictive algorithms for three academic variables. Two ensemble learning methods were also employed to explore potential enhancements in predictive power. Additionally, statistical inference tests were conducted to evaluate the consistency of algorithm performance across cross-validated training and testing data sets and determine the most effective algorithm. For first-fall and first-year GPA, the evaluation focused on RMSE values generated by each model on training and testing data sets. One-year retention algorithm performance was assessed based on overall accuracy, sensitivity, specificity, F-measure, and AUC values from training and testing data sets, with AUC values prioritized, aligning with recommendations by Dey (2021).

The examination of RMSE values visualized for first-fall GPA algorithms indicated the XGBoost model demonstrated the highest accuracy, closely followed by the random forest algorithm. Statistically, the predictive power of both the random forest and XGBoost models showed consistency, with no significant difference detected, resulting in two viable options for developing a predictive algorithm. Likewise, for first-year GPA predictions, the XGBoost model outperformed others, with the random forest model as the second-best performing algorithm. Predictive performances of the random forest and XGBoost algorithms were consistent and statistically indistinguishable, resulting in two options for developing a predictive algorithm. Particularly, the evaluation of the two ensemble learning methods revealed no enhancements in predictive power for first-fall and first-year GPA. For administrators looking to utilize predictive analytics to identify at-risk students based on projected first-fall and first-year GPAs, the XGBoost or random forest algorithms are recommended given the consistent performance between data sets and exhibited similar predictive power of all the algorithms.

Derived for the training and testing data sets, the analysis of accuracy metrics for one-year retention predictive algorithms revealed a bias towards the majority class with low specificity rates across various models. The algorithms developed from both the downsample and upsample techniques showed improvements in specificity rates during the evaluation of the training data set. However, when assessing the testing data set, the high bias towards the majority class was observed. In the initial model, XGBoost exhibited the highest AUC value, with the random forest model ranking second. Statistically, there was no significant difference between the random forest and XGBoost models, resulting in two viable options for developing predictive algorithms without sampling techniques. For the downsample algorithms, logistic regression, random forest, and XGBoost were statistically equivalent, while the logistic regression model emerged as the best-performing model for the upsample algorithms. Notably, the evaluation of the two ensemble learning methods revealed no enhancements in the predictive power of one-year retention algorithms, except for slight improvements in the upsample models. The recommended application of algorithms to identify at-risk student of departing from the institution, administrators have some choices. For developing algorithms without any sampling modifications, the recommended models are the XGBoost and random forest because of the consistent performance between data sets and exhibiting similar predictive power. Administrators can select from logistic regression, random forest, and XGBoost models when looking to implement algorithms utilizing downsample modification. These three models will exhibit consistent performance between data sets with similar predictive power. In upsample modifications algorithms, administrators could implement a logistic regression model or an ensemble learning model utilizing the blended method.

Limitations of the Study

The primary aim of this study was to evaluate the impact of the secondary curriculum on academic performance within postsecondary institutions. Yet, several limitations restrict the generalizability of the findings. Firstly, the study focused solely on four public RCUs in the southeastern United States, limiting its applicability to institutions in other geographical regions or various institutional types. The study specifically examined FTFTFs who entered their initial institution after graduating from a public Georgia HS in 2018 or 2019. Therefore, the results cannot be generalized to individuals who graduated before 2018 or after 2019, those from private Georgia HS, students graduating from out-of-state HS, or homeschooled students. Additionally, it is important to note the integration of CCRPI and EOC variables involved aggregated school-level data rather than individual student-level data within RCUs. Furthermore, the data's validity relies on the absence of modifications to the CCRPI and EOC variables by the GaDOE in 2018 and 2019, and the study does not account for potential changes due to the COVID-19 pandemic. Any alterations to these variables could impact the study's generalizability to the broader population.

The independent variables incorporated into this study were chosen to specifically assess academic performance before students enrolled in their classes. However, it is crucial to acknowledge this selection does not represent an exhaustive list of variables identified in previous research. The inclusion of additional elements, such as class types, online course participation, employment status while attending college, and housing arrangements, has the potential to influence the study's outcomes. Furthermore, the current study did not incorporate survey-based assessments of student mindset or grit, both of which could measure students' mental preparedness to adapt to the demands of college life. Additionally, the study did not explore student engagement within academic or social communities. The inclusion of engagement factors introduces a potential avenue for the study’s refinement, as their inclusion could potentially lead to altered findings.

The selection of four predictive algorithms provides a diverse set of models for analyzing factors impacting academic performance and assessing their predictive power. Yet, these algorithms represent only a small fraction of the various models available in data science techniques. The utilization of different models has the potential to influence the analysis and reveal distinct factors influencing academic performance. Prior to developing the models, thorough evaluations of preliminary considerations and assumptions were conducted on the training data set. While random forest and XGBoost algorithms demonstrated tolerance towards deviations from these considerations and assumptions, a detailed review was undertaken for linear regression and logistic regression due to their sensitivity to such violations. Additionally, the support vector machine algorithm is sensitive to extreme values and extreme class imbalances. These considerations included examining missing data points and reviewing univariate and multivariate outliers. The assumptions involved a comprehensive assessment of observation independence, linearity and collinearity, univariate and multivariate normality, and homogeneity of variance. The class imbalance correction utilized downsampling and upsampling techniques.

Missing data were addressed through various methods. The small number of missing observations in EOC and CCRPI variables underwent median imputation, while student-level data were set to zero for certain variables based on the USG data collection method. The Zell Miller indicator used "N" for missing values, while the missing values for the four college preparatory curriculum areas were marked as "U". K-Nearest Neighbors imputation was employed for missing HS GPA, admissions test scores, and expected family contribution. Visual inspection of histograms and Q-Q plots were used for outlier detection, and Grubb's statistical test identified outliers in certain variables. Mahalanobis' test for multivariate outliers revealed violations in approximately one-third of observations, but no outlier capping was applied. Observation independence and linearity assumptions were not violated. However, during multicollinearity examination, three clusters of independent variables displayed linearity. To mitigate this, college preparatory curriculum variables were consolidated, CCRPI and EOC variables were adjusted, and institutional expenditure variables were combined. Most variables did not conform to a univariate normal distribution, confirmed by Q-Q plots, Shapiro-Wilks, and

Jarque-Bera tests. Mardia's test indicated multivariate normality violations, despite Yeo-Johnson transformation method being applied with only minimal improvements. Furthermore, the analysis in this study did not include interactions among the variables. The inclusion of interactions among the independent variables could potentially have an impact on the findings related to the factors influencing first-year academic performance. Several assumptions for the linear regression and logistic regression models were violated. In the models predicting first-fall GPA and first-year GPA, eight variables lacked identifiable correlation with the dependent variables. Normal distribution assessments, using Kolmogorov-Smirnov, Jarque-Bera, and Shapiro-Wilk's statistical tests, revealed violations. Additionally, for the two GPA models, the assumption of homogeneity of variance was also violated. In the retention status model, five variables showed no correlation with the log odds of non-retention probabilities. These violations may impact both evaluation metrics and variable importance within the models.

The data set was split into a training set comprising 60% of the total observations (*N* = 13,078) and a testing set containing the remaining 40% (*N* = 8,719) to prevent overfitting through data leakage. Additionally, a 10-fold cross-validation approach was applied to both sets, providing a robust assessment of predictive performance and helping identify overfitting issues. The seed used for data splitting remained consistent for replication purposes. The development of models employed the tidymodels and tidyverse packages, widely recognized tools in data science, which streamlined data preprocessing and facilitated algorithm development. In the tuning process for SVMs, random forest, and XGBoost algorithms, RMSE was used for the GPA models, while AUC was employed for the retention variable models. The choice of accuracy metrics potentially limited the models by not displaying each model's complete accuracy metrics.

Implications for Future Research

The aim of the research was to identify significant factors impacting FTFTFs' academic performance in the first year, even before students enrolled in courses. To enhance the study, exploring interaction effects among existing factors and incorporating typical first-year course selections could be beneficial. Factors such as enrollment in first-year English and mathematics courses, credit hour load, course modality, living arrangements, employment status, and participation in engagement events could provide a more comprehensive understanding of students' integration into academic and social communities. Additionally, expanding models by incorporating surveys' assessment of students' grit and resilience mindsets could offer insights into their integration into postsecondary communities in the face of setbacks experienced. After the first-fall semester has concluded, incorporating first-fall GPA could enhance predictions for first-year GPA and one-year retention, aiding in identifying potential at-risk students who may need support.

To expand the study's impact regarding the HS curriculum, modifying the EOC subject by not combining the two similar subjects into a single variable and incorporating all components of the CCRPI variables could be considered. Two comparison analyses could be conducted to assess any changes in the study. First, a comparison involving pre-pandemic, mid-pandemic, and post-pandemic students could provide insights into the evolving impact of COVID-19 on education and help identify at-risk students requiring support post-pandemic adjustment. The second comparison would involve changes to

EOC or CCRPI variables over time. As the GaDOE may modify these variables, a comparison analysis could determine whether these variables gained or lost any significance within the study.

Advancing the study could involve exploring different predictive algorithms, especially considering technological advancements introduce new models. Different or new algorithms might uncover additional influential factors in first-year academic performance. The use of diverse ensemble learning methods could further enhance predictions. Developing optimal algorithms with varied accuracy metrics, such as R^2 values or specificity, or different engines might yield different insights into significant factors and predictive powers.

Recommendations for Practice

Postsecondary institutions stand to benefit greatly from leveraging the power of predictive analytics to identify at an early stage students who exhibit signs of at-risk tendencies of unsuccessful academic performance. From the forecast outcomes, such as GPA or the likelihood of departing, these students can be categorized into distinct risk levels, ranging from minimal to high risk. This stratification enables tailored strategies and plans to be developed and implemented. Specifically, students identified as high risk would receive a comprehensive, hands-on set of strategies and support plans designed to address their unique challenges and to provide them a launching pad towards success. This intensive support mechanism is aimed at equipping these students with the essential skills required for academic and personal success, both within the classroom environment and beyond. Conversely, students deemed to be at minimal risk would benefit from a more flexible, as-needed support system, allowing for intervention when necessary but otherwise permitting them to proceed with their studies unencumbered. This differentiated approach not only ensures resources are allocated efficiently but also fosters an environment where all students have the opportunity to succeed and thrive while at the institution.

Expanding wraparound services represents a pivotal strategy in bolstering support for at-risk students within educational institutions. Creating a comprehensive support team comprising of faculty mentors, academic advisors, and peer mentors holds potential in providing tailored assistance for students facing academic challenges. These support teams serve as invaluable resources for at-risk students, offering guidance on navigating both the academic and social communities of the institution. Furthermore, reallocating staff members' responsibilities to facilitate targeted outreach efforts for high-risk students, particularly those enrolled in courses with historically gateway courses, can be highly effective. Individuals tasked with outreach can identify when students are struggling in their courses and connect them with resources, such as tutoring services, to address their academic difficulties promptly and to mitigate the risk of academic setbacks. Another essential wraparound service is the implementation of mandatory study hall hours for high-risk students. These structured study sessions offer a dedicated environment for at-risk students to develop crucial study skills and habits essential for academic success. By fostering a culture of targeted academic support, such initiatives have the potential to yield long-term benefits beyond the students' first year, equipping the students with the tools necessary for sustained academic achievement.

Conclusions

Given ongoing challenges and an anticipated decline in the traditional-age student population, identifying at-risk students and providing support during their integration into postsecondary communities is crucial for sustaining healthy enrollment levels. In the development of at-risk models, HS GPA continues to emerge as a primary factor influencing students' success in coursework and subsequently achieving successful GPAs in the first year. The majority of predictive models highlight the significance of mean content mastery and readiness scores, along with proficiency levels in selected subject areas from the graduating high school, in shaping academic performance during the first year. These variables signify the overall rates of preparedness from high school as a supplementary factor influencing students' ability to manage postsecondary coursework.

Furthermore, students' capacity to afford the cost of attendance remains a primary factor of whether they retain or depart after the first year. In Georgia, the HOPE scholarship significantly contributes to students' ability to retain by alleviating some financial burden through reducing the total attendance cost. Certain predictive algorithms indicated expected family contributions and the PELL grant also impacted one-year retention, making these variables vital factors. These variables may be particularly of interest for administrators and policymakers to understand the impact of students from families with lower expected contributions, despite having a successful HS GPA, might opt to leave the institution due to the ongoing financial burden of continued attendance. There were no consistent patterns across predictive algorithms for many other factors influencing academic performance. For student characteristic and financial situation variables, including them in models can assist postsecondary institutions in comprehending these factors.

In comparing predictive models, both XGBoost and random forest models emerged as the recommended viable options for forecasting students' first-fall and first-year GPAs. These two models consistently performed well, with no statistically significant differences in predictive power on unseen data. Despite being a common method for GPA analysis, the linear regression model did not outperform XGBoost and random forest. In the retention models' comparison, models applied to the testing data set exhibited bias towards the majority class, even with class imbalance corrections. Among models not utilizing imbalance sampling techniques, random forest and XGBoost are recommended algorithms given consistency between data sets and showed no statistical differences in predictive power. Logistic regression, random forest, and XGBoost models utilizing downsample modifications are recommended models because of exhibiting consistency between data sets and no statistical differences in the predictive power. Logistic regression and employing an ensemble learning method utilizing the blended method are recommended as viable options with upsampling techniques. The underperformance of linear regression and logistic regression without imbalance corrections may be attributed to assumption violations. The SVM model, using three different kernels, consistently performed the worst. Overall, the utilization of ensemble learning methods did not provide enhancements to the predictions. The only improvement was exhibited in the blended method for the upsample models for one-year retention projections.

REFERENCES

Abhigyan. (2020). *Understanding logistic regression*. https://medium. com/analytics-vidhya/understanding-logistic-regression-b3c672deac04

Alam, M. (2020). Z-score for anomaly detection. *Towards Data Science*. https://towardsdatascience.com/z-score-for-anomaly-detection-d98b0006f510

Aljohani, O. (2016). A comprehensive review of major studies and theoretical models of student retention in higher education. *Higher Education Studies, 6*(2), 1-18.

Allensworth, E. & Clark, K. (2020). High school GPAs and ACT scores as predictors of college completion: Examining assumptions about consistency across high schools. *Educational Researcher, 49*(3), 198-211.

American Association of State Colleges and Universities. (2020). *Public policy agenda: Principles and policies*. Washington, D.C.

Arnold, A. (1999, March). Retention and persistence in postsecondary education: A summation of research studies. *Texas Guaranteed Student Loan Corporation*. http://www.trelliscompany.org/wp-content/uploads/2017/02/ persistence.pdf

Ary, D., Jacobs, L. C., Irvine, C. K. S., & Walker, D. A. (2019). *Introduction to research in education* (10th ed.). Boston, MA: Cengage Learning, Inc.

Association of American Colleges & Universities. (2019, March). *Facts & figures: College students are more diverse than ever. Faculty and administrators are not.* https://www.aacu.org/aacu-news/newsletter/2019/march/facts-figures#:~:text= According%20to%20a%20new%20report,to%2045.2%20percent %20in%202016

Astin, A. (1975). *Preventing students from dropping out*. San Francisco: Jossey-Bass.

Astin, A. (1984). Student involvement: A developmental theory for higher education. *Journal of College Student Development, 25*(4), 297-308.

Astin, A. (1993). *What matters in college: Four critical years revisited*. San Francisco: Jossey-Bass.

Attewell, P., & Monaghan, D. (2015). *Data mining for the social sciences: An introduction.* Oakland: University of California Press.

Awasthi, S. (2020). *Seven most popular SVM kernels*. https://dataaspirant.com/svm-kernels/

Barshay, J. (2018, September, 10). College students predicted to fall by more than 15% after the year 2025. *The Hechinger Report*. https://hechingerreport.org/college-students-predicted-to-fall-by-more-than-15-after-the-year-2025/

Batuwita, R. & Palade, V. (2012). Class imbalance learning methods for support vector machine in He, H. & Ma, Y. (Eds), *Imbalanced learning: Foundations, algorithms, and applications* (pp. 1-20). John Wiley & Sons, Inc. https://www.cs.ox.ac.uk/people/vasile.palade/papers/Class-Imbalance-SVM.pdf

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education, 12*(2), 155-187.

Bean, J. P. (1983). The application of a model of turnover in work organizations to the student attrition process. *Review of Higher Education, 6*(2), 129-148.

Belani, G. (2019). How data science is playing a big role in higher education? *Data Science Central*. https://www.datasciencecentral.com/how-data-science-is-playing-a-big-role-in-higher-education/

Berger, J., Ramirez, G. B., & Lyon, S. (2012). Past to present: A historical look at retention. *College student retention: Formula for student success,* (pp. 7-34). Plymouth: Rowman & Littlefield Publishers.

Bettinger, E. (2004). How financial aid affects persistence. In Hoxby, C. (Ed.), *College choices: The economics of where to go, when to go, and how to pay for it.* (pp. 207-237). University of Chicago Press, Chicago.

Bhandari, A. (2020). *What is multicollinearity? Here's everything you need to know*. https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/

Bidwell, A. (2013, August 21). High school graduates still struggle with college readiness. *U.S. News & World Report*. https://www.usnews.com/news/articles/ 2013/08/21/high-school-graduates-still-struggle-with-college-readiness

Bock, T. (n.d.) *What are the different types of missing data*? https://www.displayr. com/different-types-of-missing-data/

Boehmke, R. & Greenwell, B. (2020). *Hands-on machine learning with R*. New York, NY: Chapman and Hall/CRC Press. https://bradleyboehmke.github.io/HOML/

Bordens, K. S. & Abbott, B. B. (2011). *Research designs and methods: A process approach* (8th ed.). New York, NY: McGraw-Hill.

Bose, A. (2019). Cross validation—Why & how. *Towards Data Science*. https://towardsdatascience.com/cross-validation-430d9a5fee22#:~:text=Cross%20 validation%20is%20a%20technique,complementary%20subset%20of%20the%20 data

Bowen, W., Chingos, M., & McPherson, M. (2009). *Crossing the finish line: Completing college at America's public universities.* Princeton University Press.

Bowers, A. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research and Evaluation, 17*(3), 141–159.

Bridgeman, B., Pollack, J., & Burton, N. (2008). *Predicting grades in different types of college courses* (College Board Research Report No. 2008-1). New York, NY: The College Board.

Browne-Anderson, H. (2016). Preprocessing in data science (part 1): Centering, scaling, and KNN. *Datacamp*. https://www.datacamp.com/community/ tutorials/preprocessing-in- data-science-part-1-centering-scaling-and-knn

Brownlee, J. (2016). A gentle introduction to XGBoost for applied machine learning. *Machine Learning Mastery*. https://machinelearningmastery. com/gentle-introduction-xgboost-applied-machine-learning/

Brownlee, J. (2019). Probabilistic model selection with AIC, BIC, and MDL. *Machine Learning Mastery*. https://machinelearningmastery.com/ probabilistic-model-selection-measures/

Brownlee, J. (2020). How to calculate precision, recall, and f-measure for imbalanced classification. *Machine Learning Mastery*. https://machinelearning mastery.com/precision-recall-and-f-measure-for-imbalanced-classification/

Buchmann, C., & DiPrete, T. (2006). The growing female advantage in college completion: The role of family background and academic achievement. *American Sociological Review, 71*(4), 515-541.

Buckley, J., Letukas, L., & Wildavsky, B. (Eds). (2018). *Measuring success: Testing, grades, and the future of college admissions.* Johns Hopkins University Press.

Butrymowicz, S. (2017, January 30). Most colleges enroll many students who aren't prepared for higher education. *Hechinger Report*. https://hechingerreport.org/ colleges-enroll- students-arent-prepared-higher-education/

Cabrera, A., Castaneda, M., Nora, A., & Hengstler, D. (1992). The convergence between two theories of college persistence. *The Journal of Higher Education, 63*(2), 143-164.

Cabrera, A., Nora, A., & Castaneda, M. (1993). College Persistence: Structural Equations Modeling Test of an Integrated Model of Student Retention. *The Journal of Higher Education, 64*(2), 123-139.

Calvo, B. & Santafé, G. (2016). Scmamp: Statistical comparison of multiple algorithms in multiple problems. *The R Journal, 8*(1), 248-256

Camara, W., Kimmel, E., Scheuneman, J., & Sawtell, E.A. (2004). *Whose grades are inflated?* (Research Report No. 2003-4). College Entrance Examination Board.

Cansiz, S. (2020). Mahalanobis distance and multivariate outlier detection in R. *Towards Data Science*. https://towardsdatascience.com/mahalonobis-distance-and-outlier-detection-in-r-cb9c37576d7d

Chaudhari, S. (2018). Descriptive statistics. *Towards Data Science*. https://towardsdata science.com/descriptive-statistics-f2beeaf7a8df

Chen, R. (2012). Institutional characteristics and college student dropout risks: A multilevel event history analysis. *Research in Higher Education, 53*(5), 487-505.

Chen, R., & DesJardins, S. L. (2008). Exploring the effects of financial aid on the gap in student dropout risks by income level. *Research Higher Education, 49*, 1-18.

Chen, R., & St. John, E. P. (2011). State financial policies and college student persistence: A national study. *The Journal of Higher Education, 82*(5), 629-660.

Choy, S. P. (2002). *Access & persistence: Findings from 10 years of longitudinal research on students.* Washington, DC: American Council on Education.

Chugh, A. (2020). *MAE, MSE, RMSE, coefficient of determination, adjusted R squared—Which metric is better?* https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

Coenen, F. (2011). Data mining: Past, present, and future. *The Knowledge Engineering Review, 26*(1), 25-29. https://www.researchgate.net/publication/220254364_Data_mining_Past_present_and_future

Common Data Set. (2020). *CDS definitions*. https://commondataset.org/

Complete College America. (n.d.). *About us*. https://completecollege.org/our-work/

Complete College Georgia. (2021a). *About complete college Georgia.* https://completega.org/content/about-complete-college-georgia

Complete College Georgia. (2021b). *What is a momentum year?* https://complete georgia.org/what-momentum-year

Conway, D. (2014). *Data science through the lens of social science* [video file]. http://videolectures.net/kdd2014_conway_social_science/

Conway, D. (2015). *The data science venn diagram*. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Corley, R. E., Goodjoin, R., & York, S. (1991). Differences in grades and SAT scores among minority college students from urban and rural environments. *The High School Journal, 74*(3), 173–177.

Couch, S. & Kuhn, M. (2022) *Stacks: Tidy model stacking*. https://stacks.tidymodels.org/

Creswell, J. (2014). *Research design: Qualitative, quantitative and mixed methods approaches* (4th ed.). Thousand Oak, CA: SAGE Publications, Inc.

Dantas, J. (2020). The importance of k-fold cross validation for model prediction in machine learning. *Towards Data Science*. https://towardsdatascience.com/the-importance-of-k-fold-cross-validation-for-model-prediction-in-machine-learning-4709d3fed2ef

Data Quality Campaign. (2021). *Safeguarding data*. https://dataqualitycampaign.org/topic/safeguarding-data/

Data Science Degree Program. (2021). *Data science in higher education*. https://www.datasciencedegreeprograms.net/industries/higher-education/

Daveport, T. H. & Patil, D. J. (2012, October). Data scientist: The sexiest job of the 21st centry. *Harvard Business Review*. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

de Vise, D. (2011, December 12). Study: Two-fifths of high school graduates are unprepared for college or the workforce. *The Washington Post*. https://www.washingtonpost.com/blogs/college-inc/post/study-two-fifths-of-high-school-graduates-are-unprepared/2011/12/12/gIQArZKnpO_blog.html

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research, 7*(1), 1-30

DeNicco, J., Harrington, P., & Fogg, N. (2015). Factors of one-year college retention in a public state college system. *Research for Higher Education Journal, 27*, 1-13.

Dey, V. (2021). *Understanding the AUC-ROC curve in machine learning classification.*

https://analyticsindiamag.com/understanding-the-auc-roc-curve-in-machine-learning-classification/

DiMaggio, B. (2021). *The future of data analytics in higher education is prescriptive analytics*. https://www.othot.com/blog/2021-the-future-of-data-analytics-in-higher-education-is-prescriptive-analytics

Drakos, G. (2019). *Importance of cross-validation*. https://gdcoder.com/ importance-of-cross-validation/

Driscoll, M. (2013, December). *NYC data business meetup* [Video]. YouTube. https://www.youtube.com/watch?v=_S5mDcpxX1A

Dwivedi, R. (2020). How does support vector machine (SVM) algorithm works in machine learning? *Analytics Steps*. https://www.analyticssteps. com/blogs/how-does-support-vector-machine-algorithm-works-machine-learning

Ezarik, M. (2020). New analysis: Student loan borrowers with no degree. *University Business*. https://universitybusiness.com/new-analysis-student-loan-borrowers-with-no-degree/

Facing History & Ourselves. (2021). *Shifting demographics in the United States.* https://www.facinghistory.org/resource-library/my-part-story/shifting-demographics-united-states

Fain, P. (2019, Sept.). Race, geography and degree attainment. *Inside Higher Ed.* https://www.insidehighered.com/news/2019/06/27/rural-areas-lag-degree-attainment-while-urban-areas-feature-big-racial-gaps

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of*

Machine Learning Research, 15(1), 3,133-3,181

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Thousand Oaks, CA: SAGE Publications, Inc.

Fife, D. (2019). Bivariate visualizations, part 1: Interpreting scatterplots. [Video]. YouTube. https://www.youtube.com/watch?v=c2PYhJMiza8&list=PL8F480DgtpW8WFhHFRzos7iUK2r-MhmKw

Fischer, M. (2007). Settling into campus life: Differences by race/ethnicity in college involvement and outcomes. *The Journal of Higher Education, 78*(2), 125-161.

Flores, St. & Park, T. (2013). Race, ethnicity, and college success: Examining the continued significance of the minority-serving institution. *Educational Researcher, 42*(3), 115-128.

Forte, E., Towles, E., Greninger, E., Buchanan, E., & Deters, L. (2017). *Evaluation of the quality of the Georgia milestones assessment system in ELA, mathematics, science, and social studies*. https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Documents/Milestones/Georgia_Milestones_Alignment_Evaluation_Executive_Summary.pdf

French, R. (2016, September 3). Schools try to make grads ready for college. *Atlanta Journal Constitution*. https://www.ajc.com/news/local-education/schools-try-make-more-grads-ready-for-college/Nu02dZSb4qAGEgjTcDbLXO/

Gandhi, R. (2018). Support vector machine—introduction to machine learning algorithims. *Towards Data Science*. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Gansemer-Topf, A. & Schuh, J. (2006). Institutional selectivity and institutional expenditures: Examining organizational factors that contribute to retention and graduation. *Research in Higher Education, 47*(6), 613-642.

Gansemer-Topf, A., Kollasch, A. & Sun, J. (2017). A house divided? Examining persistence for on-campus STEM and non-STEM students. *Journal of College Student Retention: Research, Theory & Practice, 19*(2), 199-223.

Garson, G. D. (2012). *Testing statistical assumptions*. Asheboro, NC: Statistical Publishing Associates.

Georgia Department of Education [GaDOE]. (2017). *Validity and reliability for 2016-2017 Georgia milestones assessment system*. https://www.gadoe. org/Curriculum-Instruction-and-Assessment/Assessment/Documents/ Milestones/2016-17_Georgia_Milestones_Validity _and_Reliability_Brief.pdf

Georgia Department of Education [GaDOE]. (2018a). Redesigned college and career ready performance index: College and career ready performance index (CCRPI) indicators. https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Accountability/Documents/Resdesigned%20CCRPI%20Support%20 Documents/Redesigned%20CCRPI%20Indicators%20011918.pdf

Georgia Department of Education [GaDOE]. (2018b). *Redesigned college and career ready performance index: Side-by-side comparison of the previous and redesigned CCRPI.* https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Accountability/Documents/Resdesigned%20CCRPI %20Support%20Documents/Redesigned%20CCRPI%20Side%20by%20Side%20 013118.pdf.

Georgia Department of Education [GaDOE]. (2018c). *Validity and reliability for 2017-2018 Georgia milestones assessment system*. https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Documents/Milestones/2017-18_Georgia_Milestones_Validity_and_Reliability_Brief.pdf

Georgia Department of Education [GaDOE]. (2019). *Validity and reliability for 2018-2019 Georgia milestones assessment system*. https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Documents/Milestones/2018-19_Georgia_Milestones_Validity_and_Reliability_Brief.pdf

Georgia Department of Education [GaDOE]. (2021a). *School codes*. https://www.gadoe.org/Finance-and-Business-Operations/Facilities-Services/Pages/School-Codes.aspx

Georgia Department of Education [GaDOE]. (2021b). *2021-2022 student assessment handbook*. https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/ Documents/For%20Educators/2021-2022_Student_Assessment_Handbook.pdf

Georgia Department of Education [GaDOE]. (2021c). *College and career ready performance index.* https://www.gadoe.org/CCRPI/Pages/default.aspx

Georgia State University. (2016). An analysis of Georgia's school-accountability measures. *The Center for State and Local Finance*. https://www.metroatlantachamber.com/ assets/analysis-of-georgias-current-school-accountability-measures_20161202_ npVnblz.pdf

Georgia Student Finance Commission. (2021a). *Eligibility for the HOPE scholarship.* https://www.gafutures.org/hope-state-aid-programs/hope-zell-miller-scholarships/hope-scholarship/eligibility/

Georgia Student Finance Commission. (2021b). *Maintaining eligibility for the HOPE scholarship*. https://www.gafutures.org/hope-state-aid-programs/hope-zell-miller-scholarships/hope-scholarship/maintaining-eligibility-for-the-hope-scholarship/

Gershenson, S. (2018). *Grade inflation in high schools (2005–2016).* Thomas B. Fordham Institute.

Glen, S. (2021). *Cronbach's alpha: Simple definition, use and interpretation*. https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/cronbachs-alpha-spss/

Glen, S. (2022). *Correlation coefficient: Simple definition, formula, easy steps*. https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/

Google Developers. (2021). *Imbalanced data*. https://developers.google.com/ machine-learning/data-prep/construct/sampling-splitting/imbalanced-data

Governor's Office of Student Achievement. (2012). The effects of academic rigor in high school on academic performance in college. https://gosa.georgia.gov/document /publication/full-report-1/download

Governor's Office of Student Achievement. (n.d.). *Downloadable data*. https://gosa.georgia.gov/dashboards-data-report-card/downloadable-data

Goyal, C. (2021). *Data leakage and its effect on the performance of an ml model*. https://www.analyticsvidhya.com/blog/2021/07/data-leakage-and-its-effect-on-the-performance-of-an-ml-model/

Grace, C., Shores, E. F., Martha, Z., Brown, B., Aufsesser, D., & Bell, L. (2006). Rural disparities in baseline data of the early childhood longitudinal study: A chartbook. *National Center for Rural Early Childhood Learning Initiatives, Mississippi State University Early Childhood Institute.* https://files.eric.ed.gov/fulltext/ED495855.pdf

Gross, J., Hossler, D., Ziskin, M., & Berry, M. (2015). Institutional merit-based aid and student departure: A longitudinal analysis. *The Review of Higher Education, 38*(2), 221-250.

Hanson, M. (2020). *Financial aid statistics.* Educationdata.org. https://educationdata.org/financial-aid-statistics

Heidel, E. (2022). *Independence of observation*. https://www.scalestatistics.com/independence-of-observations.html

Henderson, B. (2009). Introduction: The work of the people's university. *Teacher-Scholar: The Journal of State Comprehensive University, 1*(2) 5-26. https://scholars. fhsu.edu/cgi/viewcontent.cgi?article=1001&context=ts

Henderson, T. (2016). *Americans are moving south, west again*. Pew Research Center. https://www.pewtrusts.org/en/research-and-analysis/blogs/stateline/2016/01/08/americans-are-moving-south-west-again

Henry, G., Rubenstein, R. & Bugler, D. (2004). Is HOPE enough? Impacts of receiving and losing merit-based financial aid. *Educational Policy, 18*(5), 686-709.

Hiregoudar, S. (2020). Ways to evaluate regression model. *Towards Data Science*. https://towardsdatascience.com/ways-to-evaluate-regression-models-77a3ff45ba70

Hiss, W. & Franks, V. (2014). *Defining promise: Optional standardized testing policies in*

American college and university admissions. The National Association for College Admission Counseling.

Horn, L. J., & Kojaku, L. K. (2001). *High school academic curriculum and the persistence path through college: persistence and transfer behavior of undergraduates 3 years after entering 4-year institutions.* Washington, DC: National Center for Education Statistics, Department of Education

Horthorn, T., Leisch, F., Zeileis, A., & Hornik, K. (2005). The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics, 14*(3), 675-699.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons, Inc.

Hurwitz, M., & Lee, J. (2018). Grade inflation and the role of standardized testing. In J. Buckley, L. Letukas, & B. Wildavsky (Eds.), *Measuring success: Testing, grades, and the future of college admissions* (pp. 64–93). Johns Hopkins University Press.

Ishitani, T. (2003). A longitudinal approach to assessing attrition behavior among first-generation students: Time-varying effects of pre-college characteristics. *Research in Higher Education, 44*(4), 433-449.

Ishitani, T. (2006). Studying attrition and degree completion behavior among first-generation college students in the United States. *The Journal of Higher Education, 77*(5), 861-885.

Jacob, B. A. (2002). Where the boys aren’t: Non-cognitive skills, returns to school and the gender gap in higher education. *Economics of Education Review, 21*, 589-598.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York: Springer.

Kleinfeld, J. (2009). No map to manhood: Male and female mindsets behind the college gender gap. *Gender Issues, 26*, 171-182.

Kobrin, J., Patterson, B., Shaw, E., Mattern, K., & Barbuti, S. (2008). *Validity of the SAT® for predicting first-year college grade point average* (Research Report No. 2008-5). College Entrance Examination Board.

Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An R package for assessing multivariate normality. *The R Journal, 6*(2), 151-162

Kuhn, M. (2019). *The caret package*. https://topepo.github.io/caret/index.html

Kuhn, M. & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.

Kuhn, M. & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models.* Boca Raton; CRC Press

Kuhn, M. & Silge, J. (2021). *Tidy modeling with R.* https://www.tmwr.org/

Kuhn M., & Wickham H. (2020). *Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles*. https://www.tidymodels.org

Lederman, D. (2019, June 17). The public's support for (and doubts about) higher ed. *Inside Higher Ed*. https://www.insidehighered.com/news/2019/06/17/survey-shows-publics-support-and-qualms-about-higher-education

Lee, A. M. I. (n.d.). *No child left behind (NCLB): What you need to know.* https://www.understood.org/en/school-learning/your-childs-rights/basics-about-childs-rights/no-child-left-behind-nclb-what-you-need-to-know.

Lee, J. (2019). 2019 Georgia higher education data book. *Georgia Budget & Policy*

Institute. https://gbpi.org/georgia-higher-education-data-book-2019/

Leppel, K. (2001). The impact of major on college persistence among freshmen. *Higher Education, 41*(3), 327–342.

Leung, K. (2021). Assumptions of logistic regression, clearly explained. *Towards Data Science*. https://towardsdatascience.com/assumptions-of-logistic-regression-clearly-explained-44d85a22b290

Livingston, G. & Cohn, D. (2010). *U.S. birth rate decline linked to recession*. Pew Research Center, Washington, D.C. https://www.pewresearch.org/social-trends/2010/04/06/us-birth-rate-decline-linked-to-recession/

Lohfink, M. M., & Paulsen, M. B. (2005). Comparing the determinants of persistence for first-generation and continuing-generation students. *Journal of College Student Development, 46*(4), 409-428.

Lotkowski, V., Robbins, S., & Noeth, R. (2004). *The Role of Academic and NonAcademic Factors in Improving College Retention.* Iowa, IA: ACT Policy Report.

Lumina Foundation. (2019, Fall). In rural America, too few roads lead to colleges success. *Focus*. https://focus.luminafoundation.org/in-rural-america-too-few-roads-lead-to-college-success/

Mack, C.M., Su, Z., & Westreich, D. (2018). Managing missing data in patient registries [White paper]. *Agency for Healthcare Research and Quality*. https://www.ncbi.nlm.nih.gov/books/NBK493611/pdf/Bookshelf_NBK493611.pdf

Makhijani, C. (2020). Advanced ensemble learning techniques. *Towards Data Science*. https://towardsdatascience.com/advanced-ensemble-learning-techniques-bf755e38cbfb

Mandrekar, J. (2010). Receiver operating characteristics curve in diagnostic test assessment. *Journal of Thoracic Oncology, 5*(9), 1315-1316. https://www.sciencedirect.com/science/article/pii/S1556086415306043?via%3Dihub

Mattern, K., Shaw, E. & Kobrin, J. (2010a, May 3). *A case for not going SAT-optional: Students with discrepant SAT and HSGPA performance* [Conference Presentation]. American Educational Research Association, Devener, CO, United States.

Matthews, K. (2018). *How data science is improving higher education.* https://www.kdnuggets.com/2018/11/data-science-improving-higher-education.html

Matthews, K. (2019). 5 ways data science is improving higher education. *Towards Data Science*. https://towardsdatascience.com/5-ways-data-science-is-improving-higher-education-b5bf402d50c4

McDonough, P. (1997). *Choosing colleges: How social class and schools structure opportunity.* Albany: State University of New York Press.

Medcalf, A. (2018). My favorite R package for: Summarizing data. *Dabbling with Data*. https://dabblingwithdata.wordpress.com/2018/01/02/my-favourite-r-package-for-summarising-data/

Merler, C & Vannatta, R. (2002). *Advanced and multivariate statistical methods: Practical application and interpretation* (2nd ed.). Pyrczak Publishing: Los Angelas, CA.

Milam, J. H. & HigherEd.org, Inc. (2003). Using national datasets for postsecondary education research. In W. E. Knight, *The primer of institutional research* (pp. 123-

149). Association for Institutional Research.

Miller, B. (2020, September 28). It's time to worry about college enrollment declines among black students. *Center for American Progress*. https://www.americanprogress.org/issues/education-postsecondary/reports/2020/09/28/490838/time-worry-college-enrollment-declines-among-black-students/.

Mishra, P., Pandey, C. M., Uttam, S., Gupta, A., Sahu, C. & Keshir, A. (2019). Descriptive statistics and normality tests for statistical data. *Ann Card Anaesth, 22*(1): 67-72. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6350423/#:~:text=The%20Shapiro%E2%80%93Wilk%20test%20is, taken%20from%20normal%20distributed%20population

Morales, E. E. (2008). Exceptional female students of color: Academic resilience and gender in higher education. *Innovative Higher Education, 33*(3), 197-213.

Nallamuthu, N. (2020). *Handling imbalanced data—machine learning, computer vision and NLP*. https://www.analyticsvidhya.com/blog/2020/11/handling-imbalanced-data-machine-learning-computer-vision-and-nlp/

Narkhede, S. (2018). Understanding confusion matrix. *Towards Data Science*. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

National Center for Education Statistics [NCES]. (2019). Total fall enrollment of first-time degree/certificate-seeking students in degree-granting postsecondary institutions, by attendance status, sex of student, and level and control of institution: 1960 through 2029. https://nces.ed.gov/programs/digest/d19/tables/dt19_305.10.asp

National Center for Education Statistics [NCES]. (2021). *Finance for degree-granting* *public institutions using GASB*. https://surveys.nces.ed.gov/ipeds/public/survey-materials/instructions?instructionid=30068

National Center for Education Statistics [NCES]. (n.d.a). *Changes to race/ethnicity reporting to IPEDS.* https://nces.ed.gov/ipeds/report-your-data/race-ethnicity-reporting-changes.

National Center for Education Statistics [NCES]. (n.d.b). Financial aid: What is the percent of undergraduates students awarded pell grants. https://nces.ed.gov/ipeds/TrendGenerator/app/answer/8/35

National Center for Education Statistics [NCES]. (n.d.c). *IPEDS survey components*. https://nces.ed.gov/ipeds/use-the-data/survey-components

National Center for Education Statistics [NCES]. (n.d.d). *Use the data: Compare Institution*. https://nces.ed.gov/ipeds/use-the-data

National Center for Education Statistics [NCES]. (n.d.e.) *View glossary.* https://surveys.nces.ed.gov/ipeds/public/glossary

National Student Clearinghouse Research Center. (2019). *Persistence & retention—2019.* https://nscresearchcenter.org/snapshotreport35-first-year-persistence-and-retention/#:~:text=Overall%20Persistence%20and%20Retention%20Rates,retained%20at%20their%20starting%20institution

Nietzel, M. (2019a, December 12). American's view of higher education takes a major drop. *Forbes*. https://www.forbes.com/sites/michaeltnietzel/2019/12/12/americans-view-of-higher-education-takes-a-major-drop/?sh=6bcf41e05ba5

Nietzel, M. (2019b, April 8). New evidence for the success of comprehensive universities. *Forbes*. https://www.forbes.com/sites/michaelt nietzel/2019/04/08/new-evidence-for-the-success-of-comprehensive-universities/?sh=7f5a80a1e56b

Noble, J., & Sawyer, R. (2002). *Predicting different levels of academic success in college using high school GPA and ACT composite score.* ACT, Inc.

O'Neil, C. & Schutt, R. (2014). *Doing data science: Straight talk from the frontline*. Sebastopol, CA: O'Reily Media, Inc.

Odell, P., Korgen, K., & Wang, G. (2005). Cross-racial friendships and social distance between racial groups on a college campus. *Innovative Higher Education, 29*(4), 291–305.

Offenstein, J., & Shulock, N. (2010). *Taking the next step: The promise of intermediate measures for meeting postsecondary completion goals*. Boston, MA: Jobs for the Future.

Onink, T. (2010, December 16). Bad college advice – the undeclared major. *Forbes Magazine.* http://www.forbes.com/sites/troyonink/2010/12/16/bad-collegeadvice-the-undeclared-major/#16fdb4441e88.

Oppong, F. B. & Agbedra, S. Y. (2016). Assessing univariate and multivariate normality, a guide for non-statisticians. *Mathematical Theory and Modeling, 6*(2). https://core.ac.uk/download/pdf/234680353.pdf

Orphan, C. (2018a). Public purpose under pressure: Examining the effects of neoliberal public policy on the missions of regional comprehensive universities. *Journal of Higher Education Outreach and Engagement, 22*(2) 89-101

Orphan, C. (2018b). Why regional comprehensive universities are vital parts of U.S. higher education. *Scholars Strategy Network*. https://scholars.org/brief/why-regional-comprehensive-universities-are-vital-parts-us-higher-education#:~:text=Regional%20comprehensive%20universities%20have%20been,and%20civic%20and%20cultural%20life

Osborn, J. W. & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation, 9*(6). https://scholarworks.umass.edu/pare/vol9/iss1/6/

Pafka, S. (2015). *Benchmarking random forest implementation*. http://datascience.la/benchmarking-random-forest-implementations/

Pascarella, E. T., & Terenzini, P. T. (1979). Interaction Effects in Spady and Tinto's Conceptual Models of College Attrition. *Sociology of Education, 52*(4), 197-210.

Pascarella, E. T., & Terenzini, P. T. (1980). Predicting Freshman Persistence and Voluntary Dropout Decisions from a Theoretical Model. *The Journal of Higher Education, 51*(1), 60-75.

Pascarella, E. T., & Terenzini, P. T. (1983). Predicting voluntary freshman year persistence/withdrawal behavior in a residential university: A path analytic validation of Tinto's model. *Journal of Educational Psychology, 75*(2), 215.

Pattison, E., Grodsky, E., & Muller, C. (2013). Is the sky falling? Grade inflation and the signaling power of grades. *Educational Researcher, 42*(5), 259–265.

Phillips, A. (2015). I had to learn that on my own: Successful first-generation, low-income college students from rural areas at an urban institution (Doctoral dissertation, University of Louisville). doi:10.18297/etd/2003

Plagata, T. (2020). *Interpreting the root mean square error of a linear regression*. https://tiaplagata.medium.com/interpreting-the-root-mean-squared-error-of-a-linear-regression-model-5166e6b10db8

Price, J. L. (1978). The study of turnover. *Administrative Science Quarterly, 23*(2), 351-353. https://www.jstor.org/stable/2392571

Provasnik, S., KewalRamani, A., Coleman, M. M., Gilbertson, L., Herring, W., & Xie, Q. (2007). *Status of education in rural America* (NCES 2007-040). National Center for Education Statistics, Institute for Education Sciences, U.S. Department of Education. Washington, DC. https://nces.ed.gov/pubs2007/2007040.pdf

Rado, D. (2017, May 19). Tribune analysis: College prep courses not preparing kids for college. *Chicago Tribune*. https://www.chicagotribune.com/news/breaking/ct-illinois-high-school-courses-met-20170519-story.html

Ravanshad, A. (2018). *Ensemble methods*. https://medium.com/@aravanshad/ensemble-methods-95533944783f

Ravindran, S. K. (2021). Random forest is simple English: Why is it so popular. *Towards Data Science*. https://towardsdatascience.com/ random-forest-in-simple-english-why-is-it-so-popular-3ba04d0374d

Reis, R. (n.d.). Re: Regional public universities—where the action is. *Standard Tomorrow's Professor Postings*. https://tomorrowsprofessor.sites.stanford.edu/posting/1090.

Revelle, W. (2020). An introduction to the psych package: Part II scale construction and psychometrics. https://mran.revolutionanalytics.com/snapshot/2021-03-20/web/packages/psych/vignettes/overview.pdf

Richmond, S. (2016). *Algorithms exposed: Random forest*. https://bccvl.org.au/algorithms-exposed-random-forest/

Rocca, B. (2019). Handling imbalanced datatsets in machine learning. *Towards Data Science*. https://towardsdatascience.com/ handling-imbalanced-datasets-in-machine-learning-7a0e84220f28

Rothstein, J. M. (2004). College performance predictions and the SAT. *Journal of Econometrics, 121*(1), 297–317.

Roy, B. (2020) All about scaling. *Towards Data Science*. https://towards datascience.com/all-about-feature-scaling-bcc0ad75cb35

Ryan, J. F. (2004). The relationship between institutional expenditures and degree attainment at baccalaureate college. *Research in Higher Education 45*(2), 97–114.

Sandeen, C. (2020, October 30). Public regional comprehensives' time to shine. *University Business.* https://universitybusiness.com/public-regional-comprehensives-time-to-shine/

Schimdt-Thieme, L. (2007). *Machine learning: Logistic regression and LDA*. https://www.ismll.uni-hildesheim.de/lehre/ml-07w/skript/ml-2up-02-logisticregression.pdf

Schleifer, D., Hagelskamp C., & Riendhart, C. (2015). *A difficult balance: Trustees speak about the challenges facing comprehensive universities*. San Francisco: Public Agenda.

Schultz, P. (2004). Upon entering college: First semester experiences of first-generation, rural students from agricultural families. *Rural Educator, 26*, 28-51.

Sedmak, T. (2020, December 17). Fall 2020 college enrollment declines 2.5%: Nearly twice the rate of decline of fall 2019. *National Student Clearinghouse.* https://www.studentclearinghouse.org/blog/fall-2020-college-enrollment-declines-

2-5-nearly-twice-the-rate-of-decline-of-fall-2019/

Sedmak, T. (2021, April 29). Spring undergraduate enrollment down 5.9%: Steepest decline so far since the pandemic. *National Student Clearinghouse*. https://www.studentclearinghouse.org/blog/spring-undergraduate-enrollment-down-5-9-steepest-decline-so-far-since-the-pandemic/

Seidman, A. (2007). Minority student retention: The best of the Journal of College Student Retention. *NASPA Journal, 45*(2). Amityville, NY: Baywood.

Sha, I. (2021). *Descriptive statistics: A beginners guide!* https://www.analyticsvidhya.com/blog/2021/06/descriptive-statistics-a-beginners-guide/

Shah, T. (2017). About train, validation and test sets in machine learning. *Towards Data Science*. https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

Sharma, D., Yadav, U. B., & Sharma, P. (2009). The concept of sensitivity and specificity in relation to two types of errors and its application in medical research. *Journal of Reliability and Statistical Studies, 2*(2), 53-58. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.301.2735&rep=rep1&type=pdf

Sharma, R. (2019). Skewed data: A problem to your statistical model. *Towards Data Science*. https://towardsdatascience.com/skewed-data-a-problem-to-your-statistical-model-9a6b5bb74e37

Shin, T. (2021). Understanding feature importance and how to implement it in python. *Towards Data Science*. https://towardsdatascience.com/understanding-feature-importance-and-how-to-implement-it-in-python-ff0287b20285

Silge, J. (2020). Modeling hotel bookings in R using tidymodels and recipes [video].

https://www.youtube.com/watch?v=dbXDkEEuvCU

Silipo, R. & Widmann, M. (2019). Confusion matrix and class statistics. https://towardsdatascience.com/confusion-matrix-and-class-statistics-68b79f4f510b

Singh, A. (2018). *A comprehensive guide to ensemble learning (with python codes)*. https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/

Smith-Barrow, D. (2019, November 8). *More students are leaving college without a degree*. https://hechingerreport.org/more-students-are-leaving-college-without-a-degree/

Soeteway, A. (2020). *Outliers detection in R*. https://statsandr.com/blog/outliers-detection-in-r/

Soni, D. (2019). Data leakage in machine learning. *Towards Data Science*. https://towardsdatascience.com/data-leakage-in-machine-learning-10bdd3eec742

Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange, 1*(1), 64-85.

Spady, W. G. (1971). Dropouts from higher education: Toward an empirical model. *Interchange*, *2*(3), 38-62.

Spight, D. (2020). Early declaration of a college major and its relationship to persistence. *NACADA Journal, 40*(1), 94-109.

St. John, E. P., Paulsen, M. B., & Carter, D. F. (2005). Diversity, college costs, and postsecondary opportunity: An examination of the financial nexus between college choice and persistence for African Americans and whites. *The Journal of*

Higher Education, 76(5), 545-569.

Stater, M. (2009). The impact of financial aid on college GPA at three flagship public institutions. *American Educational Research Journal, 46*(3), 782-815.

Stewart, S., Lim, D. H., & Kim, J. (2015). Factors influencing college persistence for first-time students. *Journal of Development Education, 38*(3), 572-593. https://files.eric.ed.gov/fulltext/EJ1092649.pdf

Sucky, R. N. (2020). Univariate and multivariate Gaussian distribution: Clear understanding with visuals. *Towards Data Science*. https://towardsdatascience. com/ univariate-and-multivariate-gaussian-distribution-clear-understanding-with-visuals-5b85e53ea76

Suggs, C. (2016). Troubling Gaps in HOPE point to need-based aid solutions. *Georgia Budget and Policy Institute.* https://gbpi.org/gaps-in-hope-point-to-need-based-aid/.

Tai, J. (2020, October 19). Do college grades predict future success? *Forbes*. https://www.forbes.com/sites/theyec/2020/10/19/do-college-grades-predict-future-success/?sh=7ab469bb5af6

Talend. (n.d.). *Business intelligence vs. data mining*. https://www.talend.com/resources/ business-intelligence-data-mining/

Terenzini, P. T., Lorang, W., & Pascarella, E. T. (1981). Predicting freshman persistence and voluntary dropout decisions: A replication. *Research in Higher Education, 15*(2), 109-127.

The Scholarship System. (2023). *What is EFC? And how to use the FAFSA EFC calculator*. https://thescholarshipsystem.com/blog-for-students-families/what-is-efc-and-how-to-use-the-fafsa-efc-calculator/#:~:text=EFC%20999%2C999%20is %20the%20highest,depending%20on%20certain%20other%20factors.

Tierney, B. (2014). *Predictive analytics using oracle data miner: Develop &use data mining models in oracle data miner, SQL & PL/SQL*. New York: McGraw-Hill Education.

Tinto, V. (1975). Dropout for higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45*(1), 89-125.

Tinto, V. (1982). Limits of theory and practice in student attrition. *The Journal of Higher Education, 53*(6), 687-700.

Tinto, V. (1988). Stages of student departure: Reflections on the longitudinal character of student leaving. *The Journal of Higher Education, 59*(4), 438-455.

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition research.* (2nd ed.). Chicago, IL: University of Chicago Press.

Tinto, V. (1997). Classrooms as communities: Exploring the educational character of student persistence. *The Journal of Higher Education, 68*(6), 599-623.

Tinto, V. (2017). Reflections on student persistence. *Student Success, 8*(2), 1-8.

Tobin, D. (2022). *3 Data mining & business intelligence cases.* https://www.integrate.io/ blog/real-life-applications-of-data-mining-and-business-intelligence/

Tomšik, R. (2019). Power comparisons of Shapiro-Wilks, Kolmogorov-Smirnov, and Jarque-Berta tests. *Scholars Journal of Research in Mathematics and Computer Science, 3*(3), p. 238-243.

Tripathi, M. (2020). Underfitting and overfitting in machine learning. *Data Science Foundation*. https://datascience.foundation/sciencewhitepaper/

underfitting-and-overfitting-in-machine-learning

U.S. News & World Report. (2021a). *2021 best national university rankings.* https://www.usnews.com/best-colleges/rankings/national-universities

U.S. News & World Report. (2021b). *2022 best engineering schools.* https://www.usnews.com/best-graduate-schools/top-engineering-schools/eng-rankings

Understood Team. (n.d.). The difference between the every student succeeds act and no child left behind. https://www.understood.org/en/school-learning/your-childs-rights/basics-about-childs-rights/the-difference-between-the-every-student-succeeds-act-and-no-child-left-behind

United States Bureau of Labor Statistics. (2021). Unemployment rate 3.7 percent for college grads, 6.7 percent for high school grades in March 2021. https://www.bls.gov/opub/ ted/2021/unemployment-rate-3-7-percent-for-college-grads-6-7-percent-for-high-school-grads-in-march-2021.htm

United States Department of Education. (n.d.). *Every student succeeds act (ESSA).* https://www.ed.gov/essa?src=rn

University System of Georgia [USG]. (2018a). *First-time freshmen – IPEDS definition who matriculated in fall 2018: SAT scores*. https://www.usg.edu/ research/ assets/research/ documents/freshmen/SRPT301_-_First-Time_Freshmen_SAT_Scores_ IPEDS_Fall18.pdf

University System of Georgia [USG]. (2018b). *High school GPA for first-time freshmen IPEDS definition: Fall 2018.* https://www.usg.edu/ research/assets/research/ documents/freshmen/SRPT_304_IPEDS_Fall_2018.pdf

University System of Georgia [USG]. (2018c). *Semester enrollment report: Fall 2018.* https://www.usg.edu/research/assets/research/documents/enrollment_reports/SER_Fall_18_Final_11072018.pdf

University System of Georgia [USG]. (2019a). *First-time freshmen – IPEDS definition who matriculated in fall 2019: SAT scores*. https://www.usg.edu/research/assets/research/ documents/freshmen/SRPT301_-_First-Time_Freshmen_SAT_Scores_ IPEDS_FALL19.pdf

University System of Georgia [USG]. (2019b). *First-time freshmen new ACT scores by degree level: Fall 2018*. https://www.usg.edu/research/assets/research/documents/freshmen/SRPT9650_First-Time_Freshmen_New_ACT_Scores_by_Degree_Level_Fall_2018_FTF_IPEDS.pdf

University System of Georgia [USG]. (2019c). *First-time freshmen new ACT scores by degree level: Fall 2019*. https://www.usg.edu/research/assets/research/documents/freshmen/ SRPT9650_First-Time_Freshmen_New_ACT_Scores_by_Degree_Level_IPEDS_FALL19.pdf

University System of Georgia [USG]. (2019d). *High school GPA for first-time freshmen IPEDS definition: Fall 2019.* https://www.usg.edu/research/assets/research/documents/freshmen/SRPT304_%E2%80%93_High_School_GPA_Fall_2019_IPEDS.pdf

University System of Georgia [USG]. (2019e). *Semester enrollment report: Fall 2019.* https://www.usg.edu/research/assets/research/documents/enrollment_reports/SER_Fall_19_Final.pdf

University System of Georgia [USG]. (2020a). *Fall headcount enrollment: Definition of* *variables.* https://www.usg.edu/research/assets/research/documents/ Enrollment_Definitions_April_2020.pdf

University System of Georgia [USG]. (2020b). *Retention rates: Definition of variables.* https://www.usg.edu/research/assets/research/documents/Retention_ Rates_ Definitions_January_2020.pdf

University System of Georgia [USG]. (2020c, November 11). *University system of Georgia enrollment increases to record high*. https://www.usg.edu/news/ release/university_system_of_georgia_enrollment_increases_to_record_high.

University System of Georgia [USG]. (2021a). *About us*. https://www.usg.edu/news/ usgfacts

University System of Georgia [USG]. (2021b). *Academic data collection certification resource guide.* https://www.usg.edu/research/assets/research/documents/ Academic_Data_Collection_Certification_Resource_Guide_for_USG_Institutions .pdf

University System of Georgia [USG]. (2021c). *Reporting resources*. https://www.usg. edu/research/reporting_resources

University System of Georgia [USG]. (2021d). *Semester enrollment report: Fall 2020.* https://www.usg.edu/research/assets/research/documents/enrollment_reports/ SER_Fall_2020_Update(3).pdf

University System of Georgia [USG]. (2021e). *Semester enrollment report data element dictionary: Financial aid data collection.* https://www.usg.edu/research/ assets/research/documents/FADC_DED_2021.pdf

University System of Georgia [USG]. (2022a). *University system of Georgia retention* *rate report: All degree type, one-year rate, first-time full-time freshmen fall 2018 cohort.* https://www.usg.edu/research/usgbythenumbers/

University System of Georgia [USG]. (2022b). *University system of Georgia retention rate report: All degree type, one-year rate, first-time full-time freshmen fall 2019 cohort*. https://www.usg.edu/research/usgbythenumbers/

University System of Georgia [USG]. (2023) *Data element dictionary: Academic data collection*. https://www.usg.edu/research/assets/research/documents/ADC_DED.pdf

UW Data Science Team. (2017). *A modern history of data science*. https://datasciencedegree.wisconsin.edu/blog/history-of-data-science/#:~:text=Pushing%20Rewind%20on%20Data%20Science&text=A%20trip%20into%20the%20history,analysis%20as%20an%20empirical%20science

Van Gennep, A. (1960). *The Rites of Passage* (M. Vizedine & G. Caffee, Trans.). Chicago: University of Chicago Press.

Velez, E. (2014). *America's college drop-out epidemic: Understanding the college drop-out population* (Working Paper 109). National Center for Analysis of Longitudinal Data in Education Research. https://caldercenter.org/sites/default/files/WP-109-Final.pdf

Wang, S. (2020). Multivariate normal distribution. *Towards Data Science.* https://towardsdatascience.com/multivariate-normal-distribution-562b28ec0fe0

WebAdMIT by Liaison. (2021). *Master college code list*. https://help.liaisonedu.com/WebAdMIT_Help_Center/Documents_and_Reference_Guides/Master_College_Code_List

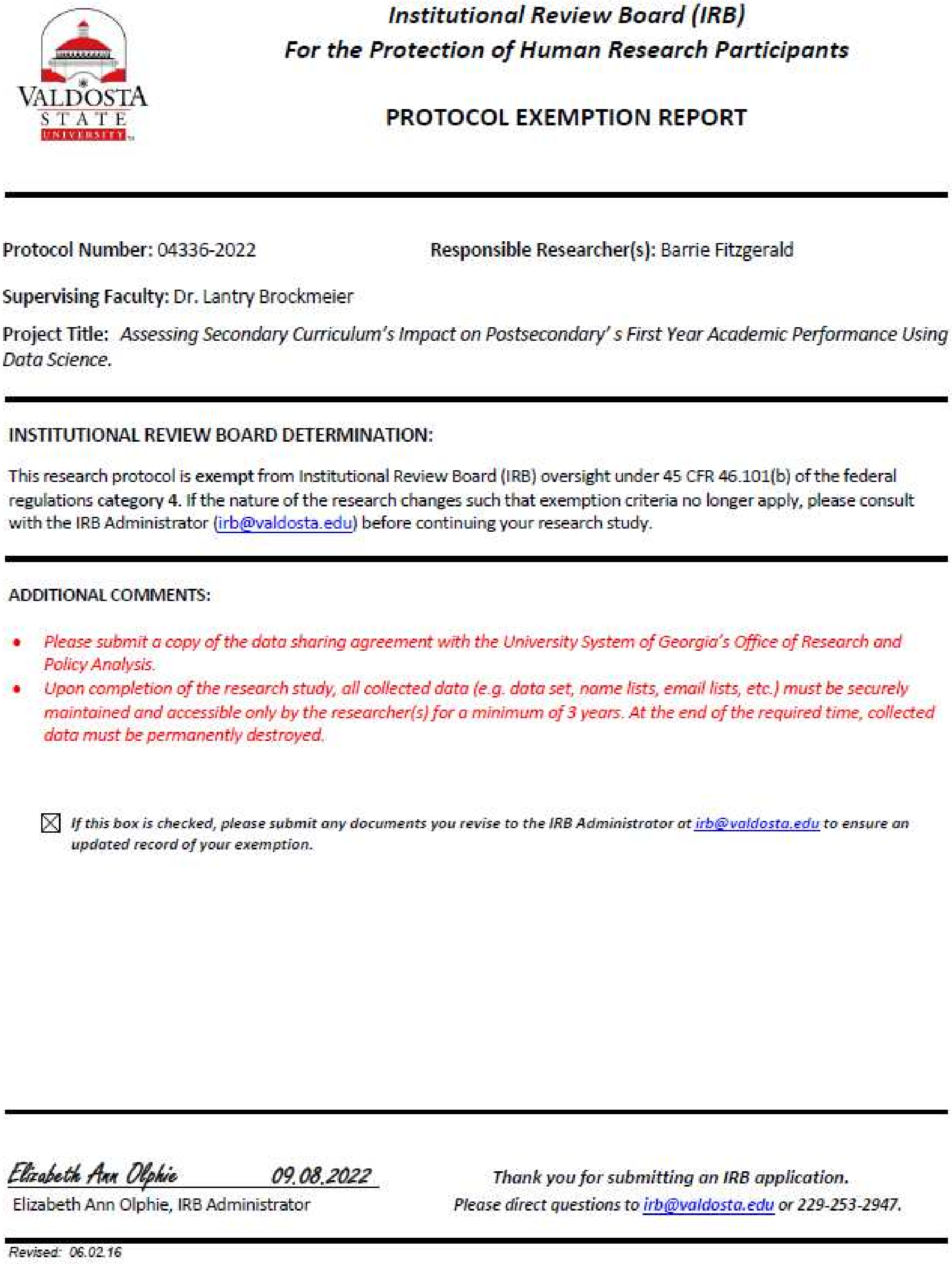
Webber, D. A. & Ehrenberg, R. (2009). Do expenditures other than instructional expenditures affect graduation and persistence rates in American higher education? Working Paper. *Cornell Higher Education Research Institute (CHERI).*

XGBoost Developers. (2021). *XGBoost documentation*. https://xgboost. readthedocs.io/en/stable/R-package/xgboostPresentation.html

Yadav, A. (2018). Support vector machine (SVM). *Towards Data Science*. https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589

APPENDIX A:

Institutional Review Board Protocol Exemption Report



Institutional Review Board (IRB)
For the Protection of Human Research Participants

PROTOCOL EXEMPTION REPORT

Protocol Number: 04336-2022 **Responsible Researcher(s):** Barrie Fitzgerald

Supervising Faculty: Dr. Lantry Brockmeier

Project Title: *Assessing Secondary Curriculum's Impact on Postsecondary' s First Year Academic Performance Using Data Science.*

INSTITUTIONAL REVIEW BOARD DETERMINATION:

This research protocol is **exempt** from Institutional Review Board (IRB) oversight under 45 CFR 46.101(b) of the federal regulations **category 4**. If the nature of the research changes such that exemption criteria no longer apply, please consult with the IRB Administrator (irb@valdosta.edu) before continuing your research study.

ADDITIONAL COMMENTS:

- *Please submit a copy of the data sharing agreement with the University System of Georgia's Office of Research and Policy Analysis.*
- *Upon completion of the research study, all collected data (e.g. data set, name lists, email lists, etc.) must be securely maintained and accessible only by the researcher(s) for a minimum of 3 years. At the end of the required time, collected data must be permanently destroyed.*

☒ ***If this box is checked, please submit any documents you revise to the IRB Administrator at irb@valdosta.edu to ensure an updated record of your exemption.***

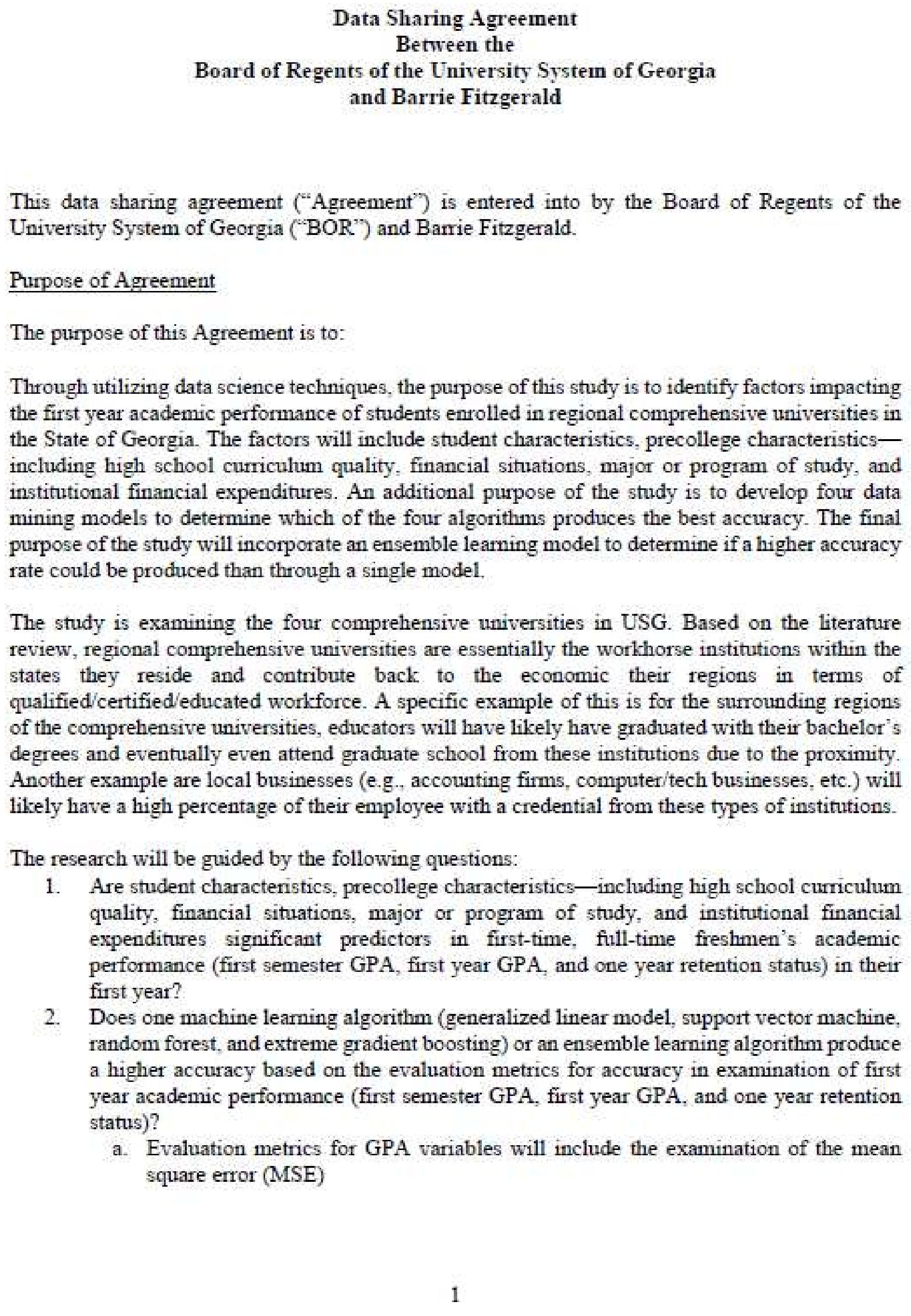
Elizabeth Ann Olphie 09.08.2022
Elizabeth Ann Olphie, IRB Administrator

Thank you for submitting an IRB application.
Please direct questions to irb@valdosta.edu or 229-253-2947.

Revised: 06.02.16

APPENDIX B:

Data Sharing Agreement



Data Sharing Agreement
Between the
Board of Regents of the University System of Georgia
and Barrie Fitzgerald

This data sharing agreement ("Agreement") is entered into by the Board of Regents of the University System of Georgia ("BOR") and Barrie Fitzgerald.

Purpose of Agreement

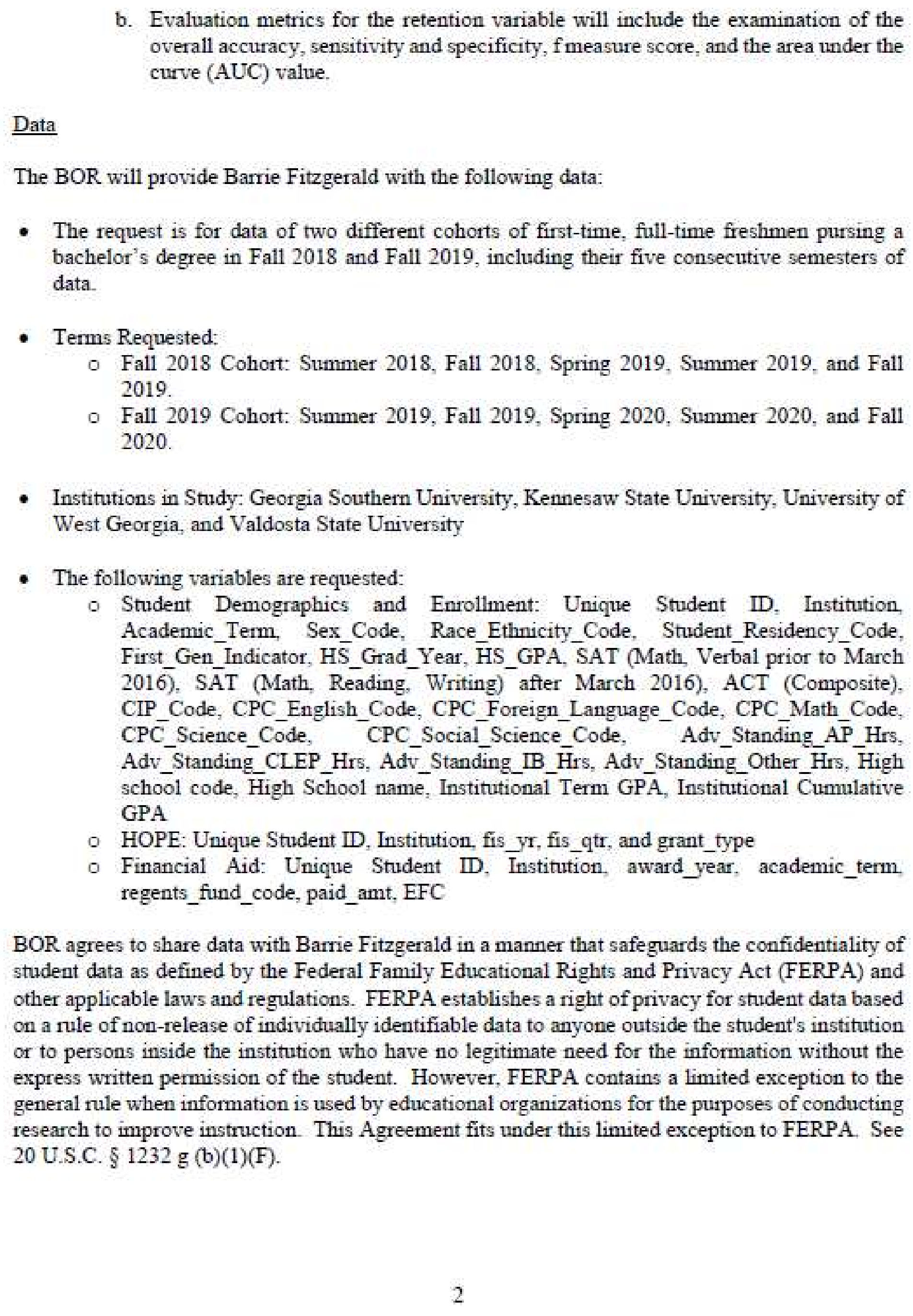
The purpose of this Agreement is to:

Through utilizing data science techniques, the purpose of this study is to identify factors impacting the first year academic performance of students enrolled in regional comprehensive universities in the State of Georgia. The factors will include student characteristics, precollege characteristics—including high school curriculum quality, financial situations, major or program of study, and institutional financial expenditures. An additional purpose of the study is to develop four data mining models to determine which of the four algorithms produces the best accuracy. The final purpose of the study will incorporate an ensemble learning model to determine if a higher accuracy rate could be produced than through a single model.

The study is examining the four comprehensive universities in USG. Based on the literature review, regional comprehensive universities are essentially the workhorse institutions within the states they reside and contribute back to the economic their regions in terms of qualified/certified/educated workforce. A specific example of this is for the surrounding regions of the comprehensive universities, educators will have likely have graduated with their bachelor's degrees and eventually even attend graduate school from these institutions due to the proximity. Another example are local businesses (e.g., accounting firms, computer/tech businesses, etc.) will likely have a high percentage of their employee with a credential from these types of institutions.

The research will be guided by the following questions:

1. Are student characteristics, precollege characteristics—including high school curriculum quality, financial situations, major or program of study, and institutional financial expenditures significant predictors in first-time, full-time freshmen's academic performance (first semester GPA, first year GPA, and one year retention status) in their first year?
2. Does one machine learning algorithm (generalized linear model, support vector machine, random forest, and extreme gradient boosting) or an ensemble learning algorithm produce a higher accuracy based on the evaluation metrics for accuracy in examination of first year academic performance (first semester GPA, first year GPA, and one year retention status)?
    a. Evaluation metrics for GPA variables will include the examination of the mean square error (MSE)



b. Evaluation metrics for the retention variable will include the examination of the overall accuracy, sensitivity and specificity, f measure score, and the area under the curve (AUC) value.

Data

The BOR will provide Barrie Fitzgerald with the following data:

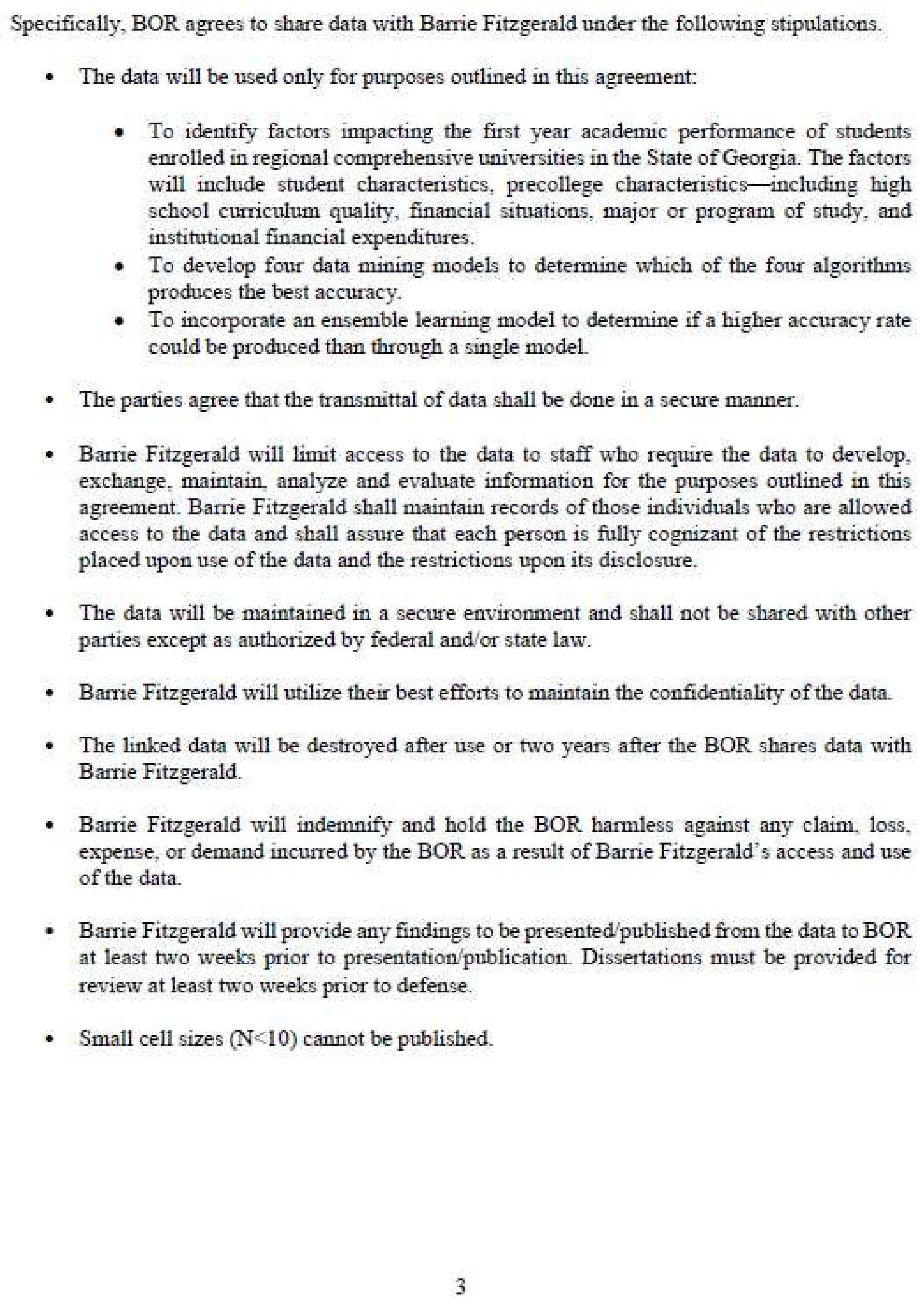
- The request is for data of two different cohorts of first-time, full-time freshmen pursing a bachelor's degree in Fall 2018 and Fall 2019, including their five consecutive semesters of data.

- Terms Requested:
  - Fall 2018 Cohort: Summer 2018, Fall 2018, Spring 2019, Summer 2019, and Fall 2019.
  - Fall 2019 Cohort: Summer 2019, Fall 2019, Spring 2020, Summer 2020, and Fall 2020.

- Institutions in Study: Georgia Southern University, Kennesaw State University, University of West Georgia, and Valdosta State University

- The following variables are requested:
  - Student Demographics and Enrollment: Unique Student ID, Institution, Academic_Term, Sex_Code, Race_Ethnicity_Code, Student_Residency_Code, First_Gen_Indicator, HS_Grad_Year, HS_GPA, SAT (Math, Verbal prior to March 2016), SAT (Math, Reading, Writing) after March 2016), ACT (Composite), CIP_Code, CPC_English_Code, CPC_Foreign_Language_Code, CPC_Math_Code, CPC_Science_Code, CPC_Social_Science_Code, Adv_Standing_AP_Hrs, Adv_Standing_CLEP_Hrs, Adv_Standing_IB_Hrs, Adv_Standing_Other_Hrs, High school code, High School name, Institutional Term GPA, Institutional Cumulative GPA
  - HOPE: Unique Student ID, Institution, fis_yr, fis_qtr, and grant_type
  - Financial Aid: Unique Student ID, Institution, award_year, academic_term, regents_fund_code, paid_amt, EFC

BOR agrees to share data with Barrie Fitzgerald in a manner that safeguards the confidentiality of student data as defined by the Federal Family Educational Rights and Privacy Act (FERPA) and other applicable laws and regulations. FERPA establishes a right of privacy for student data based on a rule of non-release of individually identifiable data to anyone outside the student's institution or to persons inside the institution who have no legitimate need for the information without the express written permission of the student. However, FERPA contains a limited exception to the general rule when information is used by educational organizations for the purposes of conducting research to improve instruction. This Agreement fits under this limited exception to FERPA. See 20 U.S.C. § 1232 g (b)(1)(F).



Specifically, BOR agrees to share data with Barrie Fitzgerald under the following stipulations.

- The data will be used only for purposes outlined in this agreement:
  - To identify factors impacting the first year academic performance of students enrolled in regional comprehensive universities in the State of Georgia. The factors will include student characteristics, precollege characteristics—including high school curriculum quality, financial situations, major or program of study, and institutional financial expenditures.
  - To develop four data mining models to determine which of the four algorithms produces the best accuracy.
  - To incorporate an ensemble learning model to determine if a higher accuracy rate could be produced than through a single model.

- The parties agree that the transmittal of data shall be done in a secure manner.

- Barrie Fitzgerald will limit access to the data to staff who require the data to develop, exchange, maintain, analyze and evaluate information for the purposes outlined in this agreement. Barrie Fitzgerald shall maintain records of those individuals who are allowed access to the data and shall assure that each person is fully cognizant of the restrictions placed upon use of the data and the restrictions upon its disclosure.

- The data will be maintained in a secure environment and shall not be shared with other parties except as authorized by federal and/or state law.

- Barrie Fitzgerald will utilize their best efforts to maintain the confidentiality of the data.

- The linked data will be destroyed after use or two years after the BOR shares data with Barrie Fitzgerald.

- Barrie Fitzgerald will indemnify and hold the BOR harmless against any claim, loss, expense, or demand incurred by the BOR as a result of Barrie Fitzgerald's access and use of the data.

- Barrie Fitzgerald will provide any findings to be presented/published from the data to BOR at least two weeks prior to presentation/publication. Dissertations must be provided for review at least two weeks prior to defense.

- Small cell sizes (N<10) cannot be published.

Termination

This Agreement shall take effect upon completion of signatures and remain in effect for one year or until terminated. This Agreement may be terminated by either BOR or Barrie Fitzgerald upon notice to the other party. BOR may terminate this Agreement with or without cause at any time by providing written notice to [party] thirty (30) calendar days prior to the termination date. Upon termination, all projects using the linked data must be immediately discontinued.

Board of Regents of the University System of Georgia

Name: Angela Bell

Title: Vice Chancellor for Research and Policy Analysis

Organization: Board of Regents of the University System of Georgia

DocuSigned by:

Signature: Angela Bell

Date: January 12, 2023

Name: Barrie Fitzgerald

Title: Doctoral Candidate

Organization: Valdosta State University

DocuSigned by:

Signature: Barrie O Fitzgerald

Date: January 12, 2023

APPENDIX C:

Data Clean Up R Code Source Files

```
################################
## USG Data Clean Up Source File ##
################################

## libraries and options utilized
library(tidyverse); library(readxl)
options(scipen=999)
select <- dplyr::select

## import data
usg2018 <- read_excel('C:/Users/bdfitzgerald/Desktop/Dissertation/USG
Data/Term_Enrollment_data_20192.xlsx')
usg2019 <- read_excel('C:/Users/bdfitzgerald/Desktop/Dissertation/USG
Data/Term_Enrollment_data_20202.xlsx')
efc2018 <- read_excel('C:/Users/bdfitzgerald/Desktop/Dissertation/USG
Data/EFC_data_20192.xlsx')
efc2019 <- read_excel('C:/Users/bdfitzgerald/Desktop/Dissertation/USG
Data/EFC_data_20202.xlsx')
hope2018 <- read_excel('C:/Users/bdfitzgerald/Desktop/Dissertation/USG
Data/HOPE_data_20192.xlsx')
hope2019 <- read_excel('C:/Users/bdfitzgerald/Desktop/Dissertation/USG
Data/HOPE_data_20202.xlsx')
pell2018 <- read_excel('C:/Users/bdfitzgerald/Desktop/Dissertation/USG
Data/Pell_data_20192.xlsx')
pell2019 <- read_excel('C:/Users/bdfitzgerald/Desktop/Dissertation/USG
Data/Pell_data_20202.xlsx')
loans2018 <- read_excel('C:/Users/bdfitzgerald/Desktop/Dissertation/USG
Data/Loan_data_20192 (2).xlsx')
loans2019 <- read_excel('C:/Users/bdfitzgerald/Desktop/Dissertation/USG
Data/Loan_data_20202 (1).xlsx')
ga_hs <- read_excel('C:/Users/bdfitzgerald/Desktop/Dissertation/ga_public_private_
highschools.xlsx')
cip_groupings <- read_excel('C:/Users/bdfitzgerald/Desktop/Dissertation/cip_codes_
groupings.xlsx')
old_sat_xwalk <- read_excel('C:/Users/bdfitzgerald/Desktop/Dissertation/USG
Data/sat_act_xwalk.xlsx', sheet = 'sat_old_new')
act_sat_xwalk <- read_excel('C:/Users/bdfitzgerald/Desktop/Dissertation/USG
Data/sat_act_xwalk.xlsx', sheet = 'act_to_new_sat')

## adjusting column names
colnames(usg2018) <- tolower(colnames(usg2018))
colnames(usg2019) <- tolower(colnames(usg2019))
colnames(efc2018) <- tolower(colnames(efc2018))
colnames(efc2019) <- tolower(colnames(efc2019))
colnames(hope2018) <- tolower(colnames(hope2018))
colnames(hope2019) <- tolower(colnames(hope2019))
colnames(pell2018) <- tolower(colnames(pell2018))
colnames(pell2019) <- tolower(colnames(pell2019))
colnames(loans2018) <- tolower(colnames(loans2018))
colnames(loans2019) <- tolower(colnames(loans2019))
colnames(ga_hs) <- tolower(colnames(ga_hs))
colnames(cip_groupings) <- tolower(colnames(cip_groupings))
colnames(old_sat_xwalk) <- tolower(colnames(old_sat_xwalk))
colnames(act_sat_xwalk) <- tolower(colnames(act_sat_xwalk))

## adjusting terms to common terms
term_transpose <- usg2018 %>% select(academic_term) %>%
 rbind(usg2019 %>% select(academic_term)) %>%
 filter(!duplicated(academic_term)) %>%  arrange(academic_term) %>%
 mutate(academic_term = as.integer(academic_term),
       term_enrolled = paste0(ifelse(as.integer(substring(academic_term, 5, 5)) == 4,
                            as.integer(substring(academic_term, 1, 4)),
                            as.integer(substring(academic_term, 1, 4)) - 1),
                       ifelse(as.integer(substring(academic_term, 5, 5)) == 1, '05',
                            ifelse(as.integer(substring(academic_term, 5, 5)) == 2,
                                 '08', '02'))))

## grouping the data together
usg <- usg2018 %>% mutate(cohort_term = '201808') %>%
 rbind(usg2019 %>%mutate(cohort_term = '201908')) %>%
 left_join(term_transpose)

## unique_student_list
unique_stus <- usg %>% filter(cohort_term == term_enrolled) %>%
 select(cohort_term, uniqueid, setid_consol, term_enrolled)

## developing the dependent variables
initial_fall <- usg %>% filter(cohort_term == term_enrolled) %>%
 ## creating the one-year retention variable DV
 left_join(usg %>% filter(as.integer(cohort_term) + 100 ==
                          as.integer(term_enrolled)) %>%
         select(cohort_term, uniqueid) %>% mutate(next_fall = 1)) %>%
 mutate(dv_next_fall = ifelse(is.na(next_fall), 0, next_fall)) %>%
 select(-next_fall) %>%
## labeling the first fall GPA DV
 rename(dv_first_fall_gpa = usg_cum_gpa) %>%
 ## creating the end of first year GPA DV
 left_join(usg %>%
         select(cohort_term, term_enrolled, uniqueid) %>%
         filter(as.integer(term_enrolled) >= as.integer(cohort_term) &
               as.integer(term_enrolled) < (as.integer(cohort_term) + 100)) %>%
         group_by(cohort_term, uniqueid) %>%
         summarise(term_enrolled = max(term_enrolled), .groups = 'drop') %>%
         left_join(usg %>% select(cohort_term, uniqueid, term_enrolled,
                                  usg_cum_gpa) %>%
               rename(dv_first_yr_gpa = usg_cum_gpa)) %>%
         select(-term_enrolled)) %>%
 ## pulling in the expected family contribution
 left_join(unique_stus %>%
         inner_join(efc2018 %>%
                 mutate(cohort_term = paste0('20', substr(award_year, 1, 2),
                                         '08')) %>%
                 select(uniqueid, setid_consol, cohort_term,
                       expected_family_contribution) %>%
                 rbind(efc2019 %>%
                       mutate(cohort_term = paste0('20', substr(award_year, 1, 2),
                                         '08')) %>%
                       select(uniqueid, setid_consol, cohort_term,
                            expected_family_contribution)))) %>%
 ## pulling in the hope/zell scholarship
 left_join(unique_stus %>%
         inner_join(hope2018 %>%
                 select(uniqueid, setid_consol, grant_type, total_award,
                       hope_academic_term) %>%
                 rename(academic_term = hope_academic_term) %>%
                 left_join(term_transpose) %>% select(-academic_term) %>%
                 rbind(hope2019 %>%
                       select(uniqueid, setid_consol, grant_type, total_award,
                            hope_academic_term) %>%
                       rename(academic_term = hope_academic_term) %>%
                       left_join(term_transpose) %>% select(-academic_term)) %>%
                 mutate(grant_type = substr(grant_type, 1, 1),
                       grant_type = ifelse(grant_type == 'H', 'hope',
                                   ifelse(grant_type== 'Z', 'zell', 'check')))) %>%
         group_by(uniqueid, setid_consol, grant_type, term_enrolled) %>%
         summarise(dollars = sum(total_award), .groups = 'drop') %>%
         spread(grant_type, dollars) %>%
         mutate(ga_hope = ifelse(is.na(hope), 0, hope) + ifelse(is.na(zell), 0, zell),
             zell_ind = ifelse(is.na(zell), 'N', 'Y')) %>% select(-hope, -zell)) %>%
 ## pulling in pell amount
 left_join(unique_stus %>%
         inner_join(pell2018 %>%
                 select(uniqueid, setid_consol, academic_term, regents_fund_code,
                       paid_amt) %>%
                 rbind(pell2019 %>%
                       select(uniqueid, setid_consol, academic_term, regents_fund_code,
                            paid_amt)) %>%
                 left_join(term_transpose)) %>%
            mutate(regents_fund_code = tolower(regents_fund_code)) %>%
            spread(regents_fund_code,  paid_amt)) %>%
 ## pulling in loan amounts
 left_join(
  unique_stus %>%
   inner_join(
    loans2018 %>%
     select(uniqueid, setid_consol, academic_term, regents_fund_code, paid_amt) %>%
     mutate(regents_fund_code = tolower(regents_fund_code),
            regents_fund_code = case_when(regents_fund_code %in% c('loand',
                                                  'stlnd') ~  'oth_loans', TRUE ~ regents_fund_code)) %>%
     group_by(uniqueid, setid_consol, academic_term, regents_fund_code) %>%
     summarise(paid_amt = sum(paid_amt, na.rm = TRUE), .groups = 'drop') %>%
     filter(paid_amt != 0) %>% spread(regents_fund_code, paid_amt) %>%
     left_join(term_transpose) %>%
     rbind(loans2019 %>%
         select(uniqueid, setid_consol, academic_term,  regents_fund_code, paid_amt) %>%
            mutate(regents_fund_code = tolower(regents_fund_code),
             regents_fund_code = case_when(regents_fund_code %in% c('loand',
            'stlnd') ~ 'oth_loans', TRUE ~ regents_fund_code)) %>%
            group_by(uniqueid, setid_consol, academic_term, regents_fund_code) %>%
            summarise(paid_amt = sum(paid_amt, na.rm = TRUE),  .groups = 'drop') %>%
            filter(paid_amt != 0) %>% spread(regents_fund_code, paid_amt) %>%
            left_join(term_transpose) ) %>%
     rename(fed_sub_loans = direct, fed_unsub_loans = dluns) %>%
     select(-academic_term))) %>%
 ## majors to the groupings
 left_join(cip_groupings) %>% select(-enrollment_cip_code_enrolled) %>%
 ## sorting variable identification variables
select(uniqueid, cohort_term,
        ## dependent variables
        dv_first_fall_gpa, dv_first_yr_gpa,dv_next_fall,
        ## will need to be removed as only used for validation of the stu population
        ## and retention rates
        enrollment_institution_name, setid_consol,
        ## used to filter down to recent hs graduates and public hs then removed afterwards
        hs_grad_year, hs_code,
        ## student characteristics
        gender_descr, admit_first_gen_ind, ipeds_race_ethnicity_descr,
        ## pre-college preparation
        hs_gpa, act_composite, sat_math, sat_verbal, sat_2016_total_score,
        ## extra pre-college preparation
        adv_standing_ap_hrs, adv_standing_clep_hrs, adv_standing_ib_hrs,
        adv_standing_other_hrs, cpc_english_code, cpc_foreign_language_code,
        cpc_math_code, cpc_science_code, cpc_social_science_code,
        ## major/program of student
```

```
      cip_categories,
      ## finanical situations
      expected_family_contribution, ga_hope, zell_ind, pell, fed_sub_loans,
      fed_unsub_loans, oth_loans)

## filter down to students who graduated in 2018 and 2019
recent_hs_grads <- initial_fall %>% filter(hs_grad_year %in% c(2018, 2019))

## filtering down to identified public hs codes
recent_ga_public_hs <- recent_hs_grads %>%
 inner_join(ga_hs %>% filter(substr(state_school_id, 1, 2) == 'GA' &
                                    ceeb_code != 'NA') %>%
          select(ceeb_code) %>% filter(!duplicated(ceeb_code)) %>%
          rename(hs_code = ceeb_code))

## developing the admissions test scores
## sat and act crosswalk
recent_ga_public_hs <- recent_ga_public_hs %>% left_join(act_sat_xwalk) %>%
 mutate(old_sat_total_score = sat_verbal + sat_math) %>%
 left_join(old_sat_xwalk) %>%
 mutate(adm_test_score = ifelse(is.na(sat_2016_total_score) &
                         !is.na(new_sat_total), new_sat_total, NA),
      adm_test_score = ifelse(is.na(adm_test_score) & !is.na(sat_2016_total_score) &
                         is.na(new_sat_total), sat_2016_total_score, adm_test_score),
      adm_test_score = ifelse(is.na(adm_test_score) & !is.na(sat_2016_total_score) &
                         !is.na(new_sat_total),  pmax(sat_2016_total_score, new_sat_total),
                        adm_test_score),
      adm_test_score_rv = ifelse(is.na(adm_test_score) & !is.na(new_sat_total_score),
                          new_sat_total_score, NA),
      adm_test_score_rv = ifelse(is.na(adm_test_score_rv) & !is.na(adm_test_score) &
                           is.na(new_sat_total_score), adm_test_score, adm_test_score_rv),
      adm_test_score_rv = ifelse(is.na(adm_test_score_rv) & !is.na(adm_test_score) &
                           !is.na(new_sat_total_score),
                          pmax(adm_test_score, new_sat_total_score),
                          adm_test_score_rv)) %>%
 select(-act_composite, -sat_verbal, -sat_math, -old_sat_total_score,
        -sat_2016_total_score, -new_sat_total, -new_sat_total_score,
        -adm_test_score) %>% rename(adm_test_score = adm_test_score_rv)

## high schools file
ga_hs <- ga_hs %>% filter(substr(state_school_id, 1, 2) == 'GA' &
                           ceeb_code != 'NA') %>%
 select(ceeb_code, state_school_id, locale_code, locale) %>%
 filter(!duplicated(ceeb_code)) %>% rename(hs_code = ceeb_code)

## removing unnecessary objects
```

```
rm(usg2018, usg2019, efc2018, efc2019, hope2018, hope2019, pell2018, pell2019,
   loans2018, loans2019,  unique_stus, term_transpose, usg, cip_groupings,
   initial_fall, recent_hs_grads, act_sat_xwalk, old_sat_xwalk)

####################################
## CCRPI Data Clean Up Source File ##
####################################

## libraries utilized
library(rio)

## import data files from web
ccrpi18 <- rio::import('https://www.gadoe.org/CCRPI/Documents/2018/2018%20
CCRPI%20Scoring%20by%20Component_12_14_18.xlsx')
ccrpi19 <- rio::import('https://www.gadoe.org/CCRPI/Documents/2019/2019%20
CCRPI%20Scoring%20by%20Component_04_01_20.xls')

## fixing column names
dat_colnames <- as.data.frame(colnames(ccrpi18)) %>%
 cbind(as.data.frame(colnames(ccrpi19)))

## since the names matched decided to select ccrpi19 names
## for easer modifications
colnames(ccrpi18) <- gsub(x = tolower(colnames(ccrpi19)),
                          pattern = ' ', replacement = '_')
colnames(ccrpi19) <- gsub(x = tolower(colnames(ccrpi19)),
                          pattern = ' ', replacement = '_')

## combining the ccrpi data together
ccrpi <- ccrpi18 %>% rbind(ccrpi19) %>%
 filter(school_id != 'ALL' & grade_cluster == 'H') %>%
 select(school_year, system_id, system_name, school_id, school_name,
           content_mastery, readiness) %>%
 mutate(content_mastery = as.numeric(content_mastery),
        readiness = as.numeric(readiness),
        school_code = paste('GA', system_id, school_id, sep = '-'))

rm(dat_colnames , ccrpi18, ccrpi19)

##################################
## EOC Data Clean Up Source File ##
##################################

## libraries utilized
library(tidyverse)
```

```
## import data
eoc18 <- readr::read_csv('https://download.gosa.ga.gov/2018/EOC_2018_By_
Grad_FEB_24_2020.csv')
eoc19 <- readr::read_csv('https://download.gosa.ga.gov/2019/EOC_2019_By_
Grad_FEB_24_2020.csv')

colnames(eoc18) <- tolower(colnames(eoc18))
colnames(eoc19) <- tolower(colnames(eoc19))

## rbinding the data together
eoc <- eoc18 %>%
 rbind(eoc19) %>%
 ## only selecting the 9th through 12th grade schools
 filter(subgroup_name == 'All Students' & instn_number != 'ALL' &
        acdmc_lvl >= 9) %>%
 mutate(instn_number = as.character(instn_number),
       instn_number = if_else(nchar(instn_number) == 3, paste0('0',
                                              instn_number), instn_number),
       school_code = paste('GA', school_distrct_cd, instn_number, sep = '-'),
       test_type = ifelse((test_cmpnt_typ_nm == '9th Grade Literature and Composition' |
                          test_cmpnt_typ_nm == 'American Literature and Composition'),
                         'english', NA),
       test_type = ifelse((test_cmpnt_typ_nm == 'Algebra I' |
                          test_cmpnt_typ_nm == 'Geometry' |
                          test_cmpnt_typ_nm == 'Analytic Geometry' |
                          test_cmpnt_typ_nm == 'Coordinate Algebra'), 'math', test_type),
       test_type = ifelse((test_cmpnt_typ_nm == 'Physical Science' |
                          test_cmpnt_typ_nm == 'Biology' ),  'science', test_type),
       test_type = ifelse((test_cmpnt_typ_nm == 'US History' |
                          test_cmpnt_typ_nm == 'Economics/Business/Free Enterprise' ),
                         'social_studies', test_type),
       year = paste0('20',substring(long_school_year, 6, 7)),
       numb_tested = ifelse(num_tested_cnt == 'TFS', 0, as.numeric(num_tested_cnt)),
       prof_numb = ifelse(as.numeric(proficient_cnt) >= 0,
                         as.numeric(proficient_cnt ), NA),
       prof_numb = ifelse(is.na(prof_numb) & !is.na(proficient_pct / 100),
                         round(numb_tested * (proficient_pct / 100), 0), prof_numb),
       prof_numb = ifelse(is.na(prof_numb), 0, prof_numb),
       dist_numb = ifelse(as.numeric(distinguished_cnt) >= 0,
                         as.numeric(distinguished_cnt ), NA),
       dist_numb = ifelse(is.na(dist_numb) & !is.na(distinguished_pct / 100),
                         round(numb_tested * (distinguished_pct / 100), 0), dist_numb),
       dist_numb = ifelse(is.na(dist_numb), 0, dist_numb)) %>%
 select(-test_cmpnt_typ_nm)

## distinct high schools
```

```
hs <- eoc18 %>% rbind(eoc19) %>%
 ## only selecting the 9th through 12th grade schools
 filter(subgroup_name == 'All Students' & instn_number != 'ALL' &
        acdmc_lvl >= 9) %>%
 mutate(instn_number = as.character(instn_number),
        instn_number = if_else(nchar(instn_number) == 3, paste0('0', instn_number),
                               instn_number),
        school_code = paste('GA', school_distrct_cd, instn_number, sep = '-'),
        test_type = ifelse((test_cmpnt_typ_nm == '9th Grade Literature and Composition' |
                            test_cmpnt_typ_nm == 'American Literature and Composition'),
                           'english', NA),
        test_type = ifelse((test_cmpnt_typ_nm == 'Algebra I' |
                            test_cmpnt_typ_nm == 'Geometry' |
                            test_cmpnt_typ_nm == 'Analytic Geometry' |
                            test_cmpnt_typ_nm == 'Coordinate Algebra'), 'math', test_type),
        test_type = ifelse((test_cmpnt_typ_nm == 'Physical Science' |
                            test_cmpnt_typ_nm == 'Biology' ),  'science', test_type),
        test_type = ifelse((test_cmpnt_typ_nm == 'US History' |
                            test_cmpnt_typ_nm == 'Economics/Business/Free Enterprise' ),
                           'social_studies', test_type),
        year = paste0('20',substring(long_school_year, 6, 7))) %>%
 select(year, school_dstrct_nm, school_code, instn_name, test_type) %>%
 group_by(school_code, school_dstrct_nm, instn_name, test_type) %>%
 summarise(count = n_distinct(year), .groups = 'drop') %>%
 spread(test_type, count) %>%
 mutate(english = ifelse(is.na(english), 0, english),
        math = ifelse(is.na(math), 0, math),
        science = ifelse(is.na(science), 0, science),
        social_studies = ifelse(is.na(social_studies), 0, social_studies),
        missing = ifelse((english + math + science + social_studies) / 4 == 2, 'N', 'Y'))

## high schools to investigate if the campus is like a 9th grade campus
hs <- hs %>%
 mutate(school_code_rv = ifelse(school_code == 'GA-634-0308', 'GA-634-0195',
                                NA),
        school_code_rv = ifelse(school_code == 'GA-635-3052', 'GA-635-1554',
                                school_code_rv),
        school_code_rv = ifelse(school_code == 'GA-642-0109', 'GA-642-0198',
                                school_code_rv),
        school_code_rv = ifelse(school_code == 'GA-668-0106', 'GA-668-2052',
                                school_code_rv),
        school_code_rv = ifelse(school_code == 'GA-688-0209', 'GA-688-0193',
                                school_code_rv),
        school_code_rv = ifelse(school_code == 'GA-729-0205', 'GA-729-0105',
                                school_code_rv),
        school_code_rv = ifelse(school_code == 'GA-737-3052', 'GA-737-0199',
                        school_code_rv),
        school_code_rv = ifelse(school_code == 'GA-754-0204', 'GA-754-0105',
                        school_code_rv),
        school_code_rv = ifelse(is.na(school_code_rv), school_code,
                        school_code_rv))

## data cleaned up
eoc_prep <- eoc %>% left_join(hs %>% select(school_code, school_code_rv)) %>%
 select(school_code_rv, year, test_type, numb_tested, prof_numb, dist_numb) %>%
 group_by(school_code_rv, year, test_type) %>%
 summarise(numb_tested = sum(numb_tested),
           prof_numb = sum(prof_numb),
           dist_numb = sum(dist_numb),
           .groups = 'drop') %>%
 mutate(prof_dist = prof_numb + dist_numb) %>%
 select(-prof_numb, -dist_numb) %>%
 mutate(prof_dist_rate = ifelse(numb_tested == 0, NA, prof_dist / numb_tested)) %>%
 select(-numb_tested, -prof_dist) %>% spread(test_type, prof_dist_rate) %>%
 left_join(eoc %>% left_join(hs %>% select(school_code, school_code_rv)) %>%
           select(year, school_code_rv, school_dstrct_nm, instn_name) %>%
           filter(!duplicated(paste0(year, school_code_rv))))

rm(eoc, eoc18, eoc19, hs)

####################################
## IPEDS Data Clean Up Source File ##
####################################

## libraries utilized
library(tidyverse)

## import data
ipeds <- read.csv('C:/Users/bdfitzgerald.VSU/Desktop/ipeds_expend_fte.csv')

## ipeds data clean up
ipeds.clean <- ipeds %>%
 gather(var_type, var_results, -UnitID, -Institution.Name) %>%
 mutate(fy = as.numeric(substr(substr(var_type, nchar(var_type) - 7,
                              nchar(var_type)), 1, 4)) - 1,
        var_type = ifelse(str_detect(var_type, 'Instruction.expenses.per.FTE'), 'instr_exp',
                   var_type),
        var_type = ifelse(str_detect(var_type, 'Research.expenses.per.FTE'), 'rsch_exp',
                   var_type),
        var_type = ifelse(str_detect(var_type, 'Public.service.expenses.per.FTE'),
                   'public_serv_exp', var_type),
        var_type = ifelse(str_detect(var_type, 'Academic.support.expenses.per.FTE'),
                    'acay_sup_exp', var_type),
      var_type = ifelse(str_detect(var_type, 'Student.service.expenses.per.FTE'),
                    'stu_serv_exp', var_type),
      var_type = ifelse(str_detect(var_type, 'Institutional.support.expenses.per.FTE'),
                   'inst_sup_exp', var_type),
      var_type = ifelse(str_detect(var_type, 'All.other.core.expenses.per.FTE'),
                   'all_other_exp', var_type)) %>%
 spread(var_type, var_results)

rm(ipeds)
```

APPENDIX D:

R Code for Analysis

```
###############
## ANALYSIS ##
###############

## libraries utilized
library(tidyverse); library(tidymodels); library(lsr); library(regclass)
library(car); library(DescTools) (vip); library(pdp); library(psych); library(tseries)
library(doParallel); library(xgboost); library(kernlab); library(stacks)
library(rstatix)

select <- dplyr::select

## setting working directory
setwd(' C:/Users/bdfitzgerald/Desktop/Dissertation/')

## data clean up source files
## USG Data
source('./dissertation_scripts/01.0 USG Data Clean Up.R')
## CCRPI Data
source('./dissertation_scripts/02.0 CCRPI Data Clean Up.R')
## EOC Data
source('./dissertation_scripts/03.0 EOC Data Clean Up.R')
## IPEDS Data
source('./dissertation_scripts/04.0 IPEDS Data Clean Up.R')

######################
## COMBING DATA   ##
######################

## USG data to IPEDS expenditures
dat <- recent_ga_public_hs %>%mutate(fy = as.integer(substr(cohort_term, 1, 4))) %>%
 left_join(ipeds.clean %>%
          rename(enrollment_institution_name = Institution.Name)) %>%
 select(-fy, -UnitID)

## ga public high schools data
## EOC and CCRPI to the distinct high schools represented
## in the four RCUs
hs_curriculum <- dat %>% select(hs_code, hs_grad_year) %>%
 filter(!duplicated(paste0(hs_grad_year, hs_code))) %>%
 left_join(ga_hs) %>%
 left_join(ccrpi %>%
          rename(state_school_id = school_code,
              hs_grad_year = school_year) %>%
          mutate(content_mastery = content_mastery / 100,
              readiness = readiness / 100) %>%
          select(hs_grad_year, content_mastery, readiness, state_school_id)) %>%
 left_join(eoc_prep %>%
          select(-school_dstrct_nm, -instn_name) %>%
          mutate(year = as.integer(year)) %>%
          rename(hs_grad_year = year, state_school_id = school_code_rv))

## removing unnecessary objects
rm(ccrpi, eoc_prep, ga_hs, ipeds.clean, recent_ga_public_hs)

## modifying USG data
dat <- dat %>% mutate(dv_next_fall = case_when(dv_next_fall == 1 ~ 0, TRUE ~ 1),
       race_eth = case_when(ipeds_race_ethnicity_descr %in%
                                                 c('White', 'Black or African American',
                                                    'Hispanic or Latino') ~
                                                  ipeds_race_ethnicity_descr, TRUE ~ 'Other'),
       unique_identifer = paste(uniqueid, cohort_term, enrollment_institution_name,
                           setid_consol, sep = '.')) %>%
 select(-ipeds_race_ethnicity_descr, -uniqueid, -cohort_term,
       -enrollment_institution_name, -setid_consol)

##########################
## DATA PARTITIONING ##
##########################

## data splitting
set.seed(51823)
dat_split <- initial_split(data = dat, prop = .6)
dat_train <- training(dat_split); dat_test <- testing(dat_split)

############################
## DEPENDENT VARAIBLE ##
## FIRST-FALL GPA            ##
############################

## data clean up recipe
fsgpa_rec <- recipe(dv_first_fall_gpa ~ ., data = dat_train) %>%
 update_role(unique_identifer,  new_role = 'id variable') %>%
 step_filter(!is.na(dv_first_fall_gpa)) %>%
 step_mutate_at(c(adv_standing_ap_hrs:adv_standing_other_hrs,
               ga_hope,  pell:oth_loans),  fn = ~ replace_na(., 0)) %>%
 step_mutate_at(zell_ind,  fn = ~ replace_na(., 'N')) %>%
 step_novel(c(cpc_english_code:cpc_social_science_code)) %>%
 step_unknown(c(cpc_english_code:cpc_social_science_code), new_level = 'U') %>%
 step_mutate(gender_descr = case_when(gender_descr == 'Male' ~ 1, TRUE ~ 0),
     admit_first_gen_ind = case_when(admit_first_gen_ind == 'Y' ~ 1,  TRUE ~ 0),
     college_prep = case_when(cpc_english_code == 'S' ~ 1,
                         cpc_english_code == 'X' ~ 1,  TRUE ~ 0) +
     case_when(cpc_foreign_language_code == 'S' ~ 1,
                cpc_foreign_language_code == 'X' ~ 1, TRUE ~ 0) +
     case_when(cpc_math_code == 'S' ~ 1, cpc_math_code == 'X' ~ 1, TRUE ~ 0)  +
     case_when(cpc_science_code == 'S' ~ 1, cpc_science_code == 'X' ~ 1, TRUE ~ 0) +
     case_when(cpc_social_science_code == 'S' ~ 1,
                cpc_social_science_code == 'X' ~ 1, TRUE ~ 0),
     acay_inst_sup_exp = (acay_sup_exp + inst_sup_exp),
     public_rsch_exp = (public_serv_exp + rsch_exp),
     cm_ready = (content_mastery + readiness) / 2,  english_cm = english - cm_ready,
     math_cm = math - cm_ready, science_cm = science - cm_ready,
     social_studies_cm = social_studies - cm_ready,
     cip_categories = case_when(cip_categories == 'Social Sciences' ~ 1,
           cip_categories == 'Fine Arts' ~ 2, cip_categories == 'Human Services' ~ 3,
           cip_categories == 'Business' ~ 4, cip_categories == 'STEM' ~ 5,
           cip_categories == 'General/Interdisciplinary Studies' ~ 6,
           cip_categories == 'Healthcare' ~ 7, cip_categories == 'Education' ~ 8,
           TRUE ~ 9),
     zell_ind = case_when(zell_ind == 'Y' ~ 1, TRUE ~ 0),
     locale_group = case_when(locale_group == 'City' ~ 1, locale_group == 'Suburb' ~ 2,
                         locale_group == 'Town' ~ 3, TRUE ~ 4),
     race_eth = case_when(race_eth == 'White' ~ 1,
                      race_eth == 'Black or African American' ~ 2,
                      race_eth == 'Hispanic or Latino' ~ 3, TRUE ~ 4),
     adv_standing_ib_hrs = adv_standing_ib_hrs,
     adv_standing_clep_hrs = adv_standing_clep_hrs,
     adv_standing_other_hrs = adv_standing_other_hrs) %>%
 step_rm(cpc_english_code, cpc_foreign_language_code,
     cpc_math_code, cpc_science_code, cpc_social_science_code, acay_sup_exp,
     inst_sup_exp, public_serv_exp, rsch_exp, english, math, science,
     social_studies, dv_next_fall, content_mastery, readiness,
     dv_first_yr_gpa, hs_code, hs_grad_year, state_school_id, locale_code, locale) %>%
 step_impute_knn(c(hs_gpa, adm_test_score, expected_family_contribution),
     neighbors = 10) %>%
 step_rename(hsgpa_knn = hs_gpa,  ats_knn = adm_test_score,
     efc_knn = expected_family_contribution) %>%
 step_YeoJohnson(hsgpa_knn, ats_knn, college_prep, adv_standing_ap_hrs,
     adv_standing_clep_hrs, adv_standing_ib_hrs, adv_standing_other_hrs, cm_ready,
     english_cm, math_cm, science_cm, social_studies_cm, efc_knn, ga_hope, pell,
     fed_sub_loans, fed_unsub_loans, oth_loans, acay_inst_sup_exp,  all_other_exp,
     instr_exp, stu_serv_exp, public_rsch_exp) %>%
 step_normalize(hsgpa_knn,  ats_knn, college_prep, adv_standing_ap_hrs,
     adv_standing_clep_hrs, adv_standing_ib_hrs, adv_standing_other_hrs, cm_ready,
     college_prep, english_cm, math_cm, science_cm, social_studies_cm, efc_knn,
     ga_hope, pell, fed_sub_loans, fed_unsub_loans, oth_loans, acay_inst_sup_exp,
     all_other_exp, instr_exp, stu_serv_exp, public_rsch_exp)
```

```
#######################
## CLEAN DATA SETS ##
#######################

## produce clean training data set
fs_gpa <- fsgpa_rec %>% prep() %>% juice()
fs_gpa_rec <- recipe(dv_first_fall_gpa ~ .,  data = fs_gpa %>% select(-unique_identifer))
## processing testing data
fs_gpa_test <- fsgpa_rec %>% prep() %>% bake(dat_test) %>%
 filter(!is.na(dv_first_fall_gpa))

#########################
##CROSS-VALIDATIONS##
#########################

## training cross validation
set.seed(51823)
fs_gpa_cv <- vfold_cv(fs_gpa %>% select(-unique_identifer), v = 10)
## testing cross validation
set.seed(51823)
fs_gpa_cv_t <- vfold_cv(fs_gpa_test %>% select(-unique_identifer), v = 10)

#########################
## TUNING CONTROLS ##
#########################

## control grid set up
ctrl_grid <- control_grid(save_pred = TRUE, save_workflow = TRUE)
## model_metrics
model_metrics <- metric_set(rmse, rsq)

##########################
##LINEAR REGRESSION ##
##########################

## training data set
gpa_ln <- lm(formula = dv_first_fall_gpa ~ .,
            data = fs_gpa %>% select(-unique_identifer))
## model summary
gpa_ln %>% summary()
## rmse
data.frame(predicted = predict(object = gpa_ln, fs_gpa),
                 actuals = fs_gpa$dv_first_fall_gpa) %>%
 mutate(diff = (predicted - actuals)^2) %>%
 select(diff) %>% summarise(rmse = sqrt(mean(diff)))
```

```
## coefficients aka beta weights
gpa_ln %>% coefficients()
## confidence intervals of coefficients
gpa_ln %>% confint()
## standardized regression coefficients
gpa_ln %>% lm.beta()
## standardized and unstandardized coefficients
gpa_ln %>% standardCoefs()

## linear regression assumptions
## correlation with dv
fs_gpa_corr <- fs_gpa %>% select(-unique_identifer) %>%
 corr.test(use = 'pairwise', method = 'pearson', adjust = 'holm', alpha = .05)
## VIF
gpa_ln %>% VIF() %>% data.frame()
## means of errors
gpa_ln %>% rstandard() %>% mean()
gpa_ln %>% rstudent() %>% mean()
##correlation amongst the residuals
gpa_ln %>% durbinWatsonTest()
## homogeneity of variance
gpa_ln %>% ncvTest()
## normality of residuals
gpa_ln %>% rstudent() %>% LillieTest()
(gpa_ln %>% rstudent())[0:5000] %>% shapiro.test()
gpa_ln %>% rstudent() %>% jarque.bera.test()

## results of the model on testing data set
gpa_ln_test <- lm(formula = dv_first_fall_gpa ~ .,
                  data = fs_gpa_test %>% select(-unique_identifer))
## model summary
gpa_ln_test %>% summary()
## rmse
data.frame(predicted = predict(object = gpa_ln_test, fs_gpa_test),
           actuals = fs_gpa_test$dv_first_fall_gpa) %>%
 mutate(diff = (predicted - actuals)^2) %>%  select(diff) %>%
 summarise(rmse = sqrt(mean(diff)))

## variable importance analysis
## creating the tidymodels linear regression
ln_reg <- linear_reg() %>% set_engine(engine = 'lm') %>% set_mode('regression')
## tidymodels workflow
ln_reg_wf <- workflow() %>% add_model(ln_reg) %>% add_recipe(fs_gpa_rec)
lin_reg_vi <- fit(ln_reg_wf, fs_gpa) %>% extract_fit_parsnip() %>%  vi() %>%
 left_join(fit(ln_reg_wf, fs_gpa) %>% extract_fit_parsnip() %>%
           vi(scale = TRUE) %>% select(-Sign) %>%
        rename(rescaled_importance = Importance))
lin_reg_vi_test <- fit(ln_reg_wf, fs_gpa_test) %>% extract_fit_parsnip() %>%  vi() %>%
 left_join(fit(ln_reg_wf, fs_gpa_test) %>% extract_fit_parsnip() %>%
        vi(scale = TRUE) %>% select(-Sign) %>%
        rename(rescaled_importance = Importance))

lin_reg_vi %>% mutate(data_set = '1. Training') %>%
 rbind(lin_reg_vi_test %>% mutate(data_set = '2. Testing')) %>%
 rename(Impact = Sign) %>%
 mutate(Impact = case_when(Impact == 'NEG' ~ 'Negative', TRUE ~ 'Positive'),
      Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                   Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                   Variable == 'hsgpa_knn' ~ 'HS GPA',
                   Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                   Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                   Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                   Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                   Variable == 'cip_categories' ~ 'Major Groupings',
                   Variable == 'efc_knn' ~ 'EFC',
                   Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                   Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                   Variable == 'pell' ~ 'PELL Grant',
                   Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                   Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                   Variable == 'oth_loans' ~ 'Other Loans',
                   Variable == 'ats_knn' ~ 'Admissions Test Scores',
                   Variable == 'all_other_exp' ~ 'All Other',
                   Variable == 'instr_exp' ~ 'Instruction',
                   Variable == 'stu_serv_exp' ~ 'Student Services',
                   Variable == 'race_eth' ~ 'Race Ethnicity',
                   Variable == 'cm_ready' ~ 'CM & Ready Mean',
                   Variable == 'locale_group' ~ 'HS Locale',
                   Variable == 'college_prep' ~ 'College Prep. Curric.',
                   Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                   Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                   Variable == 'english_cm' ~ 'English (CMR)',
                   Variable == 'math_cm' ~ 'Math (CMR)',
                   Variable == 'science_cm' ~ 'Science (CMR)',
                   Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                   TRUE ~ 'CHECK')) %>%
 ggplot(aes(x = reorder(Variable, abs(rescaled_importance)), y = rescaled_importance,
        fill = Impact)) + geom_bar(aes(fill = Impact), stat = 'identity') +
 theme_classic() + ylab("Rescaled Importance") +
 theme(legend.position = 'top', axis.title.y = element_blank(),
     text = element_text(size = 15)) + coord_flip() + facet_wrap(.~ data_set)
```

```
## predictive power
## assessing training data set
ln_reg_wf_eval <- ln_reg_wf %>%
 fit_resamples(dv_first_fall_gpa ~., resamples = fs_gpa_cv,
              metrics = model_metrics, control = ctrl_grid)
ln_reg_wf_eval %>% collect_metrics
## assessing testing data set
ln_reg_wf_eval_t <- ln_reg_wf %>%
 fit_resamples(dv_first_fall_gpa ~., resamples = fs_gpa_cv_t,
              metrics = model_metrics, control = ctrl_grid)
ln_reg_wf_eval_t %>% collect_metrics

########################################
##Support Vector Machine-Linear Kernel ##
########################################

## model specifications
svm_l_spec <- svm_linear(cost = tune(),margin = tune()) %>%
 set_mode('regression') %>%set_engine('kernlab')
## model workflow
svm_l_wf <- workflow() %>% add_model(svm_l_spec) %>% add_recipe(fs_gpa_rec)
## tuning
set.seed(52323)
svm_l_tune <- tune_grid(svm_l_wf,  resamples = fs_gpa_cv,  grid = 20)
## selecting best model
svm_l_final <- select_best(svm_l_tune, 'rmse')
## model fixed to the best outcome model
svml_tune_final <- finalize_model(svm_l_spec, svm_l_final)

## variable importance analysis
## training data set
set.seed(51923)
svml_fit <- workflow() %>% add_model(svml_tune_final)  %>%
 add_recipe(fs_gpa_rec) %>% fit(fs_gpa %>% select(-unique_identifer))
set.seed(51923)
svml_vi <- svml_fit %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = FALSE,  pred_wrapper = kernlab::predict,
    metric = 'rmse', target = 'dv_first_fall_gpa',
    train = fs_gpa %>% select(-unique_identifer))
set.seed(51923)
svml_vi_rs <- svml_fit %>%extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE, pred_wrapper = kernlab::predict,
    metric = 'rmse', target = 'dv_first_fall_gpa',
    train = fs_gpa %>% select(-unique_identifer)) %>%
 rename(rescaled_importance = Importance)
```

```
## testing data set
## vip of the xgb
set.seed(51923)
svml_fit_test <- workflow() %>%  add_model(svml_tune_final)  %>%
 add_recipe(fs_gpa_rec) %>%  fit(fs_gpa_test %>%select(-unique_identifer))
set.seed(51923)
svml_vi_test <- svml_fit_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = FALSE,  pred_wrapper = kernlab::predict,
    metric = 'rmse', target = 'dv_first_fall_gpa',
    train = fs_gpa_test %>% select(-unique_identifer))
set.seed(51923)
svml_vi_rs_test <- svml_fit_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE,  pred_wrapper = kernlab::predict,
    metric = 'rmse', target = 'dv_first_fall_gpa',
    train = fs_gpa_test %>% select(-unique_identifer)) %>%
 rename(rescaled_importance = Importance)

svml_vi_rs %>% mutate(data_set = '1. Training') %>%
 rbind(svml_vi_rs_test %>% mutate(data_set = '2. Testing')) %>%
 mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                             Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                             Variable == 'hsgpa_knn' ~ 'HS GPA',
                             Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                             Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                             Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                             Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                             Variable == 'cip_categories' ~ 'Major Groupings',
                             Variable == 'efc_knn' ~ 'EFC',
                             Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                             Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                             Variable == 'pell' ~ 'PELL Grant',
                             Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                             Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                             Variable == 'oth_loans' ~ 'Other Loans',
                             Variable == 'ats_knn' ~ 'Admissions Test Scores',
                             Variable == 'all_other_exp' ~ 'All Other',
                             Variable == 'instr_exp' ~ 'Instruction',
                             Variable == 'stu_serv_exp' ~ 'Student Services',
                             Variable == 'race_eth' ~ 'Race Ethnicity',
                             Variable == 'cm_ready' ~ 'CM & Ready Mean',
                             Variable == 'locale_group' ~ 'HS Locale',
                             Variable == 'college_prep' ~ 'College Prep. Curric.',
                             Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                             Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                             Variable == 'english_cm' ~ 'English (CMR)',
                             Variable == 'math_cm' ~ 'Math (CMR)',
                        Variable == 'science_cm' ~ 'Science (CMR)',
                        Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                        TRUE ~ 'CHECK')) %>%
 ggplot(aes(x = reorder(Variable, rescaled_importance), y = rescaled_importance)) +
 geom_bar(aes(fill = rescaled_importance/10),  stat = 'identity') +
 theme_classic() + theme(legend.position = 'none',  axis.title.y = element_blank(),
      text = element_text(size = 15)) +
 ylab('Rescaled Importance') + coord_flip() + facet_wrap(. ~ data_set)

## predictive power
## assessing training data
doParallel::registerDoParallel()
set.seed(51923)
svm_l_cv <- svml_tune_final %>%
 fit_resamples(dv_first_fall_gpa ~., resamples = fs_gpa_cv,
              metrics = model_metrics,  control = ctrl_grid)
svm_l_cv %>% collect_metrics()

## assessing testing data
doParallel::registerDoParallel()
set.seed(51923)
svm_l_cv_t <- svml_tune_final %>%
 fit_resamples(dv_first_fall_gpa ~., resamples = fs_gpa_cv_t,
              metrics = model_metrics, control = ctrl_grid)
svm_l_cv_t %>% collect_metrics()

##########################################
##Support Vector Machine-Polynomial Kernel ##
##########################################

## model specifications
svm_p_spec <- svm_poly(cost = tune(), degree = tune(), scale_factor = tune(),
                    margin = tune()) %>% set_mode('regression') %>% set_engine('kernlab')
## model workflow
svm_p_wf <- workflow() %>% add_model(svm_p_spec) %>% add_recipe(fs_gpa_rec)
## tuning model
set.seed(52323)
svm_p_tune <- tune_grid(svm_p_wf, resamples = fs_gpa_cv, grid = 20)
## selecting best model
svm_p_final <- select_best(svm_p_tune, 'rmse')
## model fixed to the best outcome model
svmp_tune_final <- finalize_model(svm_p_spec, svm_p_final)

## variable importance analysis
## training data set
set.seed(51923)
```

```
svmp_fit <- workflow() %>% add_model(svmp_tune_final)  %>%
 add_recipe(fs_gpa_rec) %>% fit(fs_gpa %>% select(-unique_identifer))
set.seed(51923)
svmp_vi <- svmp_fit %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = FALSE,  pred_wrapper = kernlab::predict,
    metric = 'rmse', target = 'dv_first_fall_gpa',
    train = fs_gpa %>% select(-unique_identifer))
set.seed(51923)
svmp_vi_rs <- svmp_fit %>%extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE,  pred_wrapper = kernlab::predict,
    metric = 'rmse', target = 'dv_first_fall_gpa',
    train = fs_gpa %>% select(-unique_identifer)) %>%
 rename(rescaled_importance = Importance)

## testing data set
set.seed(51923)
svmp_fit_test <- workflow() %>%add_model(svmp_tune_final)  %>%
 add_recipe(fs_gpa_rec) %>% fit(fs_gpa_test %>% select(-unique_identifer))
set.seed(51923)
svmp_vi_test <- svmp_fit_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = FALSE,  pred_wrapper = kernlab::predict,
    metric = 'rmse', target = 'dv_first_fall_gpa',
    train = fs_gpa_test %>% select(-unique_identifer))
set.seed(51923)
svmp_vi_rs_test <- svmp_fit_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE,  pred_wrapper = kernlab::predict,
    metric = 'rmse', target = 'dv_first_fall_gpa',
    train = fs_gpa_test %>% select(-unique_identifer)) %>%
 rename(rescaled_importance = Importance)

svmp_vi_rs %>% mutate(data_set = '1. Training') %>%
 rbind(svmp_vi_rs_test %>% mutate(data_set = '2. Testing')) %>%
 mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                             Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                             Variable == 'hsgpa_knn' ~ 'HS GPA',
                             Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                             Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                             Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                             Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                             Variable == 'cip_categories' ~ 'Major Groupings',
                             Variable == 'efc_knn' ~ 'EFC',
                             Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                             Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                             Variable == 'pell' ~ 'PELL Grant',
                             Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                             Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                             Variable == 'oth_loans' ~ 'Other Loans',
                             Variable == 'ats_knn' ~ 'Admissions Test Scores',
                             Variable == 'all_other_exp' ~ 'All Other',
                             Variable == 'instr_exp' ~ 'Instruction',
                             Variable == 'stu_serv_exp' ~ 'Student Services',
                             Variable == 'race_eth' ~ 'Race Ethnicity',
                             Variable == 'cm_ready' ~ 'CM & Ready Mean',
                             Variable == 'locale_group' ~ 'HS Locale',
                             Variable == 'college_prep' ~ 'College Prep. Curric.',
                             Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                             Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                             Variable == 'english_cm' ~ 'English (CMR)',
                             Variable == 'math_cm' ~ 'Math (CMR)',
                             Variable == 'science_cm' ~ 'Science (CMR)',
                             Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                             TRUE ~ 'CHECK')) %>%
 ggplot(aes(x = reorder(Variable, rescaled_importance), y = rescaled_importance)) +
 geom_bar(aes(fill = rescaled_importance/10), stat = 'identity') + theme_classic() +
 theme(legend.position = 'none', axis.title.y = element_blank(),
     text = element_text(size = 15)) +
 ylab('Rescaled Importance') + coord_flip() +  facet_wrap(. ~ data_set)

## predictive power
## assessing training data set
doParallel::registerDoParallel()
set.seed(51923)
svm_p_cv <- svmp_tune_final %>%
 fit_resamples(dv_first_fall_gpa ~., resamples = fs_gpa_cv,
          metrics = model_metrics, control = ctrl_grid)
svm_p_cv %>% collect_metrics()

## assessing testing data set
doParallel::registerDoParallel()
set.seed(51923)
svm_p_cv_t <- svmp_tune_final %>%
 fit_resamples(dv_first_fall_gpa ~., resamples = fs_gpa_cv_t,
          metrics = model_metrics, control = ctrl_grid)
svm_p_cv_t %>% collect_metrics()

####################################################
##Support Vector Machine-Radial Basis Function Kernel ##
####################################################

## model specifications
svm_r_spec <- svm_rbf(cost = tune(), rbf_sigma = tune(),margin = tune()) %>%
 set_mode('regression') %>% set_engine('kernlab')
```

```
## model workflow
svm_r_wf <- workflow() %>% add_model(svm_r_spec) %>% add_recipe(fs_gpa_rec)
## tuning model
set.seed(52323)
svm_r_tune <- tune_grid(svm_r_wf, resamples = fs_gpa_cv, grid = 20)
## selecting best model
svm_r_final <- select_best(svm_r_tune, 'rmse')
## model fixed to the best outcome model
svmr_tune_final <- finalize_model(svm_r_spec, svm_r_final)

## variable importance analysis
## training data set
set.seed(51923)
svmr_fit <- workflow() %>% add_model(svmr_tune_final)  %>%
 add_recipe(fs_gpa_rec) %>% fit(fs_gpa %>% select(-unique_identifer))
set.seed(51923)
svmr_vi <- svmr_fit %>%extract_fit_parsnip() %>%
 vi(method = 'permute', scale = FALSE, pred_wrapper = kernlab::predict,
    metric = 'rmse', target = 'dv_first_fall_gpa',
    train = fs_gpa %>%select(-unique_identifer))
set.seed(51923)
svmr_vi_rs <- svmr_fit %>%extract_fit_parsnip() %>%
 vi(method = 'permute', scale = TRUE, pred_wrapper = kernlab::predict,
    metric = 'rmse', target = 'dv_first_fall_gpa',
    train = fs_gpa %>% select(-unique_identifer)) %>%
 rename(rescaled_importance = Importance)

# testing data set
set.seed(51923)
svmr_fit_test <- workflow() %>% add_model(svmr_tune_final)  %>%
 add_recipe(fs_gpa_rec) %>% fit(fs_gpa_test %>% select(-unique_identifer))
set.seed(51923)
svmr_vi_test <- svmr_fit_test %>%extract_fit_parsnip() %>%
 vi(method = 'permute', scale = FALSE, pred_wrapper = kernlab::predict,
    metric = 'rmse', target = 'dv_first_fall_gpa',
    train = fs_gpa_test %>%select(-unique_identifer))
set.seed(51923)
svmr_vi_rs_test <- svmr_fit_test %>%extract_fit_parsnip() %>%
 vi(method = 'permute', scale = TRUE, pred_wrapper = kernlab::predict,
    metric = 'rmse', target = 'dv_first_fall_gpa',
    train = fs_gpa_test %>%select(-unique_identifer)) %>%
 rename(rescaled_importance = Importance)

svmr_vi_rs %>% mutate(data_set = '1. Training') %>%
 rbind(svmr_vi_rs_test %>% mutate(data_set = '2. Testing')) %>%
 mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                        Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                        Variable == 'hsgpa_knn' ~ 'HS GPA',
                        Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                        Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                        Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                        Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                        Variable == 'cip_categories' ~ 'Major Groupings',
                        Variable == 'efc_knn' ~ 'EFC',
                        Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                        Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                        Variable == 'pell' ~ 'PELL Grant',
                        Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                        Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                        Variable == 'oth_loans' ~ 'Other Loans',
                        Variable == 'ats_knn' ~ 'Admissions Test Scores',
                        Variable == 'all_other_exp' ~ 'All Other',
                        Variable == 'instr_exp' ~ 'Instruction',
                        Variable == 'stu_serv_exp' ~ 'Student Services',
                        Variable == 'race_eth' ~ 'Race Ethnicity',
                        Variable == 'cm_ready' ~ 'CM & Ready Mean',
                        Variable == 'locale_group' ~ 'HS Locale',
                        Variable == 'college_prep' ~ 'College Prep. Curric.',
                        Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                        Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                        Variable == 'english_cm' ~ 'English (CMR)',
                        Variable == 'math_cm' ~ 'Math (CMR)',
                        Variable == 'science_cm' ~ 'Science (CMR)',
                        Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                        TRUE ~ 'CHECK')) %>%
 ggplot(aes(x = reorder(Variable, rescaled_importance), y = rescaled_importance)) +
 geom_bar(aes(fill = rescaled_importance/10),  stat = 'identity') + theme_classic() +
 theme(legend.position = 'none',  axis.title.y = element_blank(),
     text = element_text(size = 15)) +
 ylab('Rescaled Importance') + coord_flip() + facet_wrap(. ~ data_set)

## predictive power
## assessing training data set
doParallel::registerDoParallel()
set.seed(51923)
svm_r_cv <- svmr_tune_final %>%
 fit_resamples(dv_first_fall_gpa ~., resamples = fs_gpa_cv,
          metrics = model_metrics,  control = ctrl_grid)
svm_r_cv %>% collect_metrics()

## assessing testing data set
doParallel::registerDoParallel()
```

```
set.seed(51923)
svm_r_cv_t <- svmr_tune_final %>%
 fit_resamples(dv_first_fall_gpa ~., resamples = fs_gpa_cv_t,
             metrics = model_metrics,  control = ctrl_grid)
svm_r_cv_t %>% collect_metrics()

######################
##RANDOM FOREST ##
######################

## model specifications
rf_spec <- rand_forest(  mtry = tune(), trees = tune(),min_n = tune()) %>%
 set_mode("regression") %>% set_engine("ranger")
## workflow
rf_wf <- workflow() %>% add_model(rf_spec) %>% add_recipe(fs_gpa_rec)
## tuning model
doParallel::registerDoParallel()
set.seed(51823)
rf_wf_tune <- tune_grid(rf_wf, resamples = fs_gpa_cv, grid = 20)
## best model
rf_tune_best <- select_best(rf_wf_tune, 'rmse')
## model fixed to the best outcome model
rf_tune_final <- finalize_model(rf_spec, rf_tune_best)

## variable importance analysis
## training data set
set.seed(511923)
rf_vi <- rf_tune_final %>% set_engine('ranger', importance = 'permutation') %>%
 fit(dv_first_fall_gpa ~ .,  data = fs_gpa %>% select(-unique_identifer)) %>% vi()
set.seed(511923)
rf_vi_rs <- rf_tune_final %>% set_engine('ranger', importance = 'permutation') %>%
 fit(dv_first_fall_gpa ~ ., data = fs_gpa %>% select(-unique_identifer)) %>%
 vi(scale = TRUE) %>% rename(rescaled_importance = Importance)

## testing data set
set.seed(511923)
rf_vi_test <- rf_tune_final %>% set_engine('ranger', importance = 'permutation') %>%
 fit(dv_first_fall_gpa ~ .,  data = fs_gpa_test %>% select(-unique_identifer)) %>% vi()
set.seed(511923)
rf_vi_rs_test <- rf_tune_final %>% set_engine('ranger', importance = 'permutation') %>%
 fit(dv_first_fall_gpa ~ .,  data = fs_gpa_test %>% select(-unique_identifer)) %>%
 vi(scale = TRUE) %>% rename(rescaled_importance = Importance)

rf_vi_rs %>% mutate(data_set = '1. Training') %>%
 rbind(rf_vi_rs_test %>% mutate(data_set = '2. Testing')) %>%
 mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                    Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                    Variable == 'hsgpa_knn' ~ 'HS GPA',
                    Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                    Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                    Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                    Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                    Variable == 'cip_categories' ~ 'Major Groupings',
                    Variable == 'efc_knn' ~ 'EFC',
                    Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                    Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                    Variable == 'pell' ~ 'PELL Grant',
                    Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                    Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                    Variable == 'oth_loans' ~ 'Other Loans',
                    Variable == 'ats_knn' ~ 'Admissions Test Scores',
                    Variable == 'all_other_exp' ~ 'All Other',
                    Variable == 'instr_exp' ~ 'Instruction',
                    Variable == 'stu_serv_exp' ~ 'Student Services',
                    Variable == 'race_eth' ~ 'Race Ethnicity',
                    Variable == 'cm_ready' ~ 'CM & Ready Mean',
                    Variable == 'locale_group' ~ 'HS Locale',
                    Variable == 'college_prep' ~ 'College Prep. Curric.',
                    Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                    Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                    Variable == 'english_cm' ~ 'English (CMR)',
                    Variable == 'math_cm' ~ 'Math (CMR)',
                    Variable == 'science_cm' ~ 'Science (CMR)',
                    Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                    TRUE ~ 'CHECK')) %>%
 ggplot(aes(x = reorder(Variable, rescaled_importance), y = rescaled_importance)) +
 geom_bar(aes(fill = rescaled_importance/10),  stat = 'identity') +
 theme_classic() + theme(legend.position = 'none',  axis.title.y = element_blank(),
     text = element_text(size = 15)) + ylab('Rescaled Importance') +
 coord_flip() + facet_wrap(.~ data_set)

## predictive power
rf_wf_final <- workflow() %>% add_model(rf_tune_final) %>% add_recipe(fs_gpa_rec)
## assessing training data set
doParallel::registerDoParallel()
set.seed(51923)
rf_train <- rf_wf_final %>%
 fit_resamples(dv_first_fall_gpa ~., resamples = fs_gpa_cv,
          metrics = model_metrics, control = ctrl_grid)
## model performance
rf_train %>% collect_metrics()
```

```
## assessing testing data set
doParallel::registerDoParallel()
set.seed(51923)
rf_test <- rf_wf_final %>%
 fit_resamples(dv_first_fall_gpa ~.,  resamples = fs_gpa_cv_t,
              metrics = model_metrics,  control = ctrl_grid)
rf_test %>% collect_metrics()

######################################
##EXTREME GRADIENT BOOSTING ##
######################################

## building out model specs
xgb <- boost_tree(trees = tune(), tree_depth = tune(),min_n = tune(),
 loss_reduction = tune(), sample_size = tune(),  mtry = tune(),  learn_rate = tune()) %>%
 set_engine('xgboost') %>% set_mode('regression')
xgb_wf <- workflow() %>% add_model(xgb) %>% add_recipe(fs_gpa_rec)
## tuning model
doParallel::registerDoParallel()
set.seed(51923)
xgb_wf_tune <- tune_grid(xgb_wf, resamples = fs_gpa_cv, grid = 20)
## best model
xgb_tune_best <- select_best(xgb_wf_tune, 'rmse')
## model fixed to the best outcome model
xgb_tune_final <- finalize_model(xgb, xgb_tune_best)

## variable importance analysis
## training data set
set.seed(51923)
xgb_vi <- xgb_tune_final %>% set_engine('xgboost') %>%
 fit(dv_first_fall_gpa ~ .,  data = fs_gpa %>% select(-unique_identifer)) %>% vi()
set.seed(51923)
xgb_vi_rs <- xgb_tune_final %>% set_engine('xgboost') %>%
 fit(dv_first_fall_gpa ~ ., data = fs_gpa %>%select(-unique_identifer)) %>%
 vi(scale = TRUE) %>% rename(rescaled_importance = Importance)

## testing data set
set.seed(51923)
xgb_vi_test <- xgb_tune_final %>% set_engine('xgboost') %>%
 fit(dv_first_fall_gpa ~ .,  data = fs_gpa_test %>% select(-unique_identifer)) %>% vi()
set.seed(51923)
xgb_vi_rs_test <- xgb_tune_final %>% set_engine('xgboost') %>%
 fit(dv_first_fall_gpa ~ .,
    data = fs_gpa_test %>% select(-unique_identifer)) %>% vi(scale = TRUE) %>%
 rename(rescaled_importance = Importance)
```

```
xgb_vi_rs %>% mutate(data_set = '1. Training') %>%
 rbind(xgb_vi_rs_test %>% mutate(data_set = '2. Testing')) %>%
 mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                             Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                             Variable == 'hsgpa_knn' ~ 'HS GPA',
                             Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                             Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                             Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                             Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                             Variable == 'cip_categories' ~ 'Major Groupings',
                             Variable == 'efc_knn' ~ 'EFC',
                             Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                             Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                             Variable == 'pell' ~ 'PELL Grant',
                             Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                             Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                             Variable == 'oth_loans' ~ 'Other Loans',
                             Variable == 'ats_knn' ~ 'Admissions Test Scores',
                             Variable == 'all_other_exp' ~ 'All Other',
                             Variable == 'instr_exp' ~ 'Instruction',
                             Variable == 'stu_serv_exp' ~ 'Student Services',
                             Variable == 'race_eth' ~ 'Race Ethnicity',
                             Variable == 'cm_ready' ~ 'CM & Ready Mean',
                             Variable == 'locale_group' ~ 'HS Locale',
                             Variable == 'college_prep' ~ 'College Prep. Curric.',
                             Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                             Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                             Variable == 'english_cm' ~ 'English (CMR)',
                             Variable == 'math_cm' ~ 'Math (CMR)',
                             Variable == 'science_cm' ~ 'Science (CMR)',
                             Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                             TRUE ~ 'CHECK')) %>%
 ggplot(aes(x = reorder(Variable, rescaled_importance), y = rescaled_importance)) +
 geom_bar(aes(fill = rescaled_importance/10),  stat = 'identity') + theme_classic() +
 theme(legend.position = 'none',  axis.title.y = element_blank(),
      text = element_text(size = 15)) + ylab('Rescaled Importance') +
 coord_flip() + facet_wrap(. ~ data_set)
```

```
## predictive power
xgb_wf_final <- workflow() %>% add_model(xgb_tune_final) %>%
 add_recipe(fs_gpa_rec)
## assessing training data
doParallel::registerDoParallel()
set.seed(51923)
xgb_train <- xgb_wf_final %>%
 fit_resamples(dv_first_fall_gpa ~., resamples = fs_gpa_cv,
             metrics = model_metrics, control = ctrl_grid)
xgb_train %>%collect_metrics()

## assessing testing data
doParallel::registerDoParallel()
set.seed(51923)
xgb_test <- xgb_wf_final %>%
 fit_resamples(dv_first_fall_gpa ~., resamples = fs_gpa_cv_t,
             metrics = model_metrics,  control = ctrl_grid)
xgb_test %>% collect_metrics()

########################################################
##VARIABLE COMPARISON OF TESTING DATA SETS ##
########################################################

lin_reg_vi_test %>% select(Variable, rescaled_importance) %>%
 mutate(type = 'Linear Regression') %>%
 rbind(svml_vi_rs_test %>% mutate(type = 'SVM Linear')) %>%
 rbind(svmp_vi_rs_test %>% mutate(type = 'SVM Polynomial')) %>%
 rbind(svmr_vi_rs_test %>% mutate(type = 'SVM Radial')) %>%
 rbind(rf_vi_rs_test %>% mutate(type = 'Random Forest')) %>%
 rbind(xgb_vi_rs_test %>% mutate(type = 'XGBoost')) %>%
 mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                        Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                        Variable == 'hsgpa_knn' ~ 'HS GPA',
                        Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                        Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                        Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                        Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                        Variable == 'cip_categories' ~ 'Major Groupings',
                        Variable == 'efc_knn' ~ 'EFC',
                        Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                        Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                        Variable == 'pell' ~ 'PELL Grant',
                        Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                        Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                        Variable == 'oth_loans' ~ 'Other Loans',
                        Variable == 'ats_knn' ~ 'Admissions Test Scores',
                        Variable == 'all_other_exp' ~ 'All Other',
                        Variable == 'instr_exp' ~ 'Instruction',
                        Variable == 'stu_serv_exp' ~ 'Student Services',
                        Variable == 'race_eth' ~ 'Race Ethnicity',
                        Variable == 'cm_ready' ~ 'CM & Ready Mean',
                        Variable == 'locale_group' ~ 'HS Locale',
                        Variable == 'college_prep' ~ 'College Prep. Curric.',
                        Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                        Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                        Variable == 'english_cm' ~ 'English (CMR)',
                        Variable == 'math_cm' ~ 'Math (CMR)',
                        Variable == 'science_cm' ~ 'Science (CMR)',
                        Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                        TRUE ~ 'CHECK')) %>%
 ggplot(aes(x = reorder(Variable, desc(Variable)), y = rescaled_importance)) +
 geom_bar(aes(fill = rescaled_importance/10),  stat = 'identity') + theme_classic() +
 theme(legend.position = 'none',  axis.title.y = element_blank(),
       text = element_text(size = 20)) +
 ylab('Rescaled Importance') + coord_flip()+ facet_wrap(. ~ type,  ncol = 6)

###########################
##ENSEMBLE LEARNING ##
###########################

## pulling out predictions from training data set
train_pred <- ln_reg_wf_eval %>%collect_predictions() %>%
 select(dv_first_fall_gpa, .pred) %>% rename(linear_reg = .pred) %>%
 cbind(svm_r_cv %>% collect_predictions() %>%
       rename(svmrbf = .pred) %>%select(svmrbf)) %>%
 cbind(rf_train %>% collect_predictions() %>%
       rename(rf = .pred) %>% select(rf)) %>%
 cbind(xgb_train %>% collect_predictions() %>%
       rename(xgb = .pred) %>% select(xgb))

## pulling out the predictions from testing data set
test_pred <- ln_reg_wf_eval_t %>%collect_predictions() %>%
 select(dv_first_fall_gpa, .pred) %>% rename(linear_reg = .pred) %>%
 cbind(svm_r_cv_t %>% collect_predictions() %>%
       rename(svmrbf = .pred) %>% select(svmrbf)) %>%
 cbind(rf_test %>% collect_predictions() %>%
       rename(rf = .pred) %>% select(rf)) %>%
 cbind(xgb_test %>% collect_predictions() %>%
       rename(xgb = .pred) %>% select(xgb))

## mean method
## training data set
```

```
train_metrics_rv <- train_metrics %>% filter(.metric == 'rmse') %>%
 select(-.metric) %>%
 rbind(train_pred %>%
       mutate(mean_pred = (linear_reg + svmrbf + rf + xgb) / 4,
             diff = mean_pred - dv_first_fall_gpa,
             diff = diff^2) %>% select(diff) %>%
       summarise(mean = mean(diff),  mean = sqrt(mean)) %>%
       mutate(`Data Set` = '1. Training',  model = '7. Ensemble Mean'))

## testing data set
test_metrics_rv <- test_metrics %>% filter(.metric == 'rmse') %>%
 select(-.metric) %>%
 rbind(test_pred %>%
       mutate(mean_pred = (linear_reg + svmrbf + rf + xgb) / 4,
             diff = mean_pred - dv_first_fall_gpa,
             diff = diff^2) %>% select(diff) %>%
       summarise(mean = mean(diff), mean = sqrt(mean)) %>%
       mutate(`Data Set` = '2. Testing', model = '7. Ensemble Mean'))

## blended method
org_stack <- stacks() %>% add_candidates(ln_reg_wf_eval) %>%
 add_candidates(svm_r_cv) %>% add_candidates(rf_train) %>%
 add_candidates(xgb_train)

set.seed(7423)
org_stack_fit <- org_stack %>% blend_predictions() %>% fit_members()

## training data set
blend_train_lr <- linear_reg(penalty = (org_stack_fit$penalty)$penalty,
                             mixture = (org_stack_fit$penalty)$mixture) %>%
 set_engine('lm') %>% set_mode('regression') %>%
 fit(dv_first_fall_gpa ~ ., data = train_pred)

train_metrics_rv <- train_metrics_rv %>%
 rbind(train_pred %>% cbind(blend_train_lr %>% predict(train_pred)) %>%
       mutate(diff = .pred - dv_first_fall_gpa, diff = diff^2) %>% select(diff) %>%
       summarise(mean = mean(diff), mean = sqrt(mean)) %>%
       mutate(`Data Set` = '1. Training', model = '8. Ensemble Blend'))

## testing data set
test_metrics_rv <- test_metrics_rv %>%
 rbind(test_pred %>% cbind(blend_train_lr %>% predict(test_pred)) %>%
       mutate(diff = .pred - dv_first_fall_gpa,  diff = diff^2) %>%
       select(diff) %>% summarise(mean = mean(diff),  mean = sqrt(mean)) %>%
       mutate(`Data Set` = '2. Testing',  model = '8. Ensemble Blend'))
```

```
train_metrics_rv %>% rbind(test_metrics_rv) %>% ggplot() +
 geom_bar(aes(x = reorder(model, desc(model)), y = mean,
            fill = `Data Set`),  stat = 'identity', position = 'dodge') +
 geom_text(aes(x = reorder(model, desc(model)),  y = mean,
             group = `Data Set`, label = format(round(mean, 3), nsmall = 3)),
          position = position_dodge(width = 1),  hjust = 1,  fontface = 'bold',
          size = 4) + theme_classic() +
 theme(legend.position = 'top',  axis.title = element_blank(),
      text = element_text(size = 15)) + coord_flip()

########################
## RMSE COMPARISON ##
########################

rmse_rs <- ln_reg_wf_eval[[3]]  %>% as.data.frame() %>%
 filter(.metric == 'rmse') %>% gather(var_type, train) %>%
 filter(var_type %like% '%estimate%') %>%
 mutate(fold = 1:10,  model = '1. Linear Regression') %>%
 select(model, fold, train) %>%
 left_join(ln_reg_wf_eval_t[[3]]  %>% as.data.frame() %>%
          filter(.metric == 'rmse') %>% gather(var_type,  test) %>%
          filter(var_type %like% '%estimate%') %>%
          mutate(fold = 1:10,  model = '1. Linear Regression') %>%
          select(model, fold, test)) %>%
 rbind(svm_l_cv[[3]]  %>% as.data.frame() %>%
      filter(.metric == 'rmse') %>% gather(var_type,  train) %>%
      filter(var_type %like% '%estimate%') %>%
      mutate(fold = 1:10,  model = '2. SVM Linear Kernel') %>%
      select(model, fold, train) %>%
      left_join(svm_l_cv_t[[3]]  %>% as.data.frame() %>%
             filter(.metric == 'rmse') %>% gather(var_type,  test) %>%
             filter(var_type %like% '%estimate%') %>%
             mutate(fold = 1:10,  model = '2. SVM Linear Kernel') %>%
             select(model, fold, test))) %>%
 rbind(svm_p_cv[[3]]  %>% as.data.frame() %>%
      filter(.metric == 'rmse') %>% gather(var_type,  train) %>%
      filter(var_type %like% '%estimate%') %>%
      mutate(fold = 1:10,  model = '3. SVM Polynomial Kernel') %>%
      select(model, fold, train) %>%
      left_join(svm_p_cv_t[[3]]  %>% as.data.frame() %>%
             filter(.metric == 'rmse') %>% gather(var_type,  test) %>%
             filter(var_type %like% '%estimate%') %>%
             mutate(fold = 1:10,  model = '3. SVM Polynomial Kernel') %>%
             select(model, fold, test))) %>%
 rbind(svm_r_cv[[3]]  %>% as.data.frame() %>%
      filter(.metric == 'rmse') %>% gather(var_type,  train) %>%
      filter(var_type %like% '%estimate%') %>%
      mutate(fold = 1:10,  model = '4. SVM Radial BF Kernel') %>%
      select(model, fold, train) %>%
      left_join(svm_r_cv_t[[3]]  %>% as.data.frame() %>%
             filter(.metric == 'rmse') %>% gather(var_type,  test) %>%
             filter(var_type %like% '%estimate%') %>%
             mutate(fold = 1:10,  model = '4. SVM Radial BF Kernel') %>%
             select(model, fold, test))) %>%
 rbind(rf_train[[3]]  %>% as.data.frame() %>%
      filter(.metric == 'rmse') %>% gather(var_type,  train) %>%
      filter(var_type %like% '%estimate%') %>%
      mutate(fold = 1:10,  model = '5. Random Forest') %>%
      select(model, fold, train) %>%
      left_join(rf_test[[3]]  %>% as.data.frame() %>%
             filter(.metric == 'rmse') %>% gather(var_type,  test) %>%
             filter(var_type %like% '%estimate%') %>%
             mutate(fold = 1:10,  model = '5. Random Forest') %>%
             select(model, fold, test))) %>%
 rbind(xgb_train[[3]]  %>% as.data.frame() %>%
      filter(.metric == 'rmse') %>% gather(var_type,  train) %>%
      filter(var_type %like% '%estimate%') %>%
      mutate(fold = 1:10, model = '6. XGBoost') %>%
      select(model, fold, train) %>%
      left_join(xgb_test[[3]]  %>% as.data.frame() %>%
             filter(.metric == 'rmse') %>% gather(var_type,  test) %>%
             filter(var_type %like% '%estimate%') %>%
             mutate(fold = 1:10,  model = '6. XGBoost') %>%
             select(model, fold, test))) %>%
 mutate(train = as.numeric(train), test = as.numeric(test))

rmse_rs %>%  gather(data_set, values, -model, -fold) %>%
 mutate(values = as.numeric(values),
      data_set = case_when(data_set == 'test' ~ '2. Testing',  TRUE ~ '1. Training')) %>%
 rename(`Data Set` = data_set) %>%
 ggplot(aes(x = reorder(model, desc(model)), y = values, fill = `Data Set`)) +
 geom_boxplot() + theme_classic() +
 theme(legend.position = 'none',  text = element_text(size = 15),
      axis.title = element_blank()) + coord_flip() + facet_wrap(. ~ `Data Set`)

## inferential statistics on algorithms
## wilcox test between the training and testing dataset
wilcox.test(as.numeric((rmse_rs %>%filter(model == '1. Linear Regression'))$train),
          as.numeric((rmse_rs %>% filter(model == '1. Linear Regression'))$test),
          paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
          conf.level = 0.95)
wilcox.test(as.numeric((rmse_rs %>% filter(model == '2. SVM Linear Kernel'))$train),
          as.numeric((rmse_rs %>% filter(model == '2. SVM Linear Kernel'))$test),
          paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
          conf.level = 0.95)
wilcox.test(as.numeric((rmse_rs %>%
                filter(model == '3. SVM Polynomial Kernel'))$train),
          as.numeric((rmse_rs %>%filter(model == '3. SVM Polynomial Kernel'))$test),
          paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
          conf.level = 0.95)
wilcox.test(as.numeric((rmse_rs %>%
                filter(model == '4. SVM Radial BF Kernel'))$train),
          as.numeric((rmse_rs %>%filter(model == '4. SVM Radial BF Kernel'))$test),
          paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
          conf.level = 0.95)
wilcox.test(as.numeric((rmse_rs %>%filter(model == '5. Random Forest'))$train),
          as.numeric((rmse_rs %>%filter(model == '5. Random Forest'))$test),
          paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
          conf.level = 0.95)
wilcox.test(as.numeric((rmse_rs %>%filter(model == '6. XGBoost'))$train),
          as.numeric((rmse_rs %>%filter(model == '6. XGBoost'))$test),
          paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
          conf.level = 0.95)

## friedmen test of the best model
## training data set
rmse_rs %>% friedman_test(train ~ model|fold)
rmse_rs %>% friedman_effsize(train ~ model|fold)
rmse_rs %>% wilcox_test(train ~ model, paired = TRUE, p.adjust.method = 'bonferroni')

## testing data set
rmse_rs %>% friedman_test(test ~ model|fold)
rmse_rs %>% friedman_effsize(test ~ model|fold)
rmse_rs %>% wilcox_test(test ~ model, paired = TRUE, p.adjust.method = 'bonferroni')

## median rmse values
rmse_rs %>% select(-fold) %>% group_by(model) %>%
 summarise(train = median(train), test = median(test),.groups = 'drop')

############################
## DEPENDENT VARAIBLE ##
## FIRST-YEAR GPA          ##
############################

## data clean up recipe
fygpa_rec <- recipe(dv_first_yr_gpa ~ ., data = dat_train) %>%
 update_role(unique_identifer, new_role = 'id variable') %>%
 step_filter(!is.na(dv_first_yr_gpa)) %>%
 step_mutate_at(c(adv_standing_ap_hrs:adv_standing_other_hrs, ga_hope,
            pell:oth_loans), fn = ~ replace_na(., 0)) %>%
 step_mutate_at(zell_ind, fn = ~ replace_na(., 'N')) %>%
 step_novel(c(cpc_english_code:cpc_social_science_code)) %>%
 step_unknown(c(cpc_english_code:cpc_social_science_code), new_level = 'U') %>%
 step_mutate(gender_descr = case_when(gender_descr == 'Male' ~ 1, TRUE ~ 0),
    admit_first_gen_ind = case_when(admit_first_gen_ind == 'Y' ~ 1,  TRUE ~ 0),
    college_prep = case_when(cpc_english_code == 'S' ~ 1,
                      cpc_english_code == 'X' ~ 1,  TRUE ~ 0) +
    case_when(cpc_foreign_language_code == 'S' ~ 1,
              cpc_foreign_language_code == 'X' ~ 1, TRUE ~ 0) +
    case_when(cpc_math_code == 'S' ~ 1, cpc_math_code == 'X' ~ 1, TRUE ~ 0)  +
    case_when(cpc_science_code == 'S' ~ 1, cpc_science_code == 'X' ~ 1, TRUE ~ 0) +
    case_when(cpc_social_science_code == 'S' ~ 1,
              cpc_social_science_code == 'X' ~ 1, TRUE ~ 0),
    acay_inst_sup_exp = (acay_sup_exp + inst_sup_exp),
    public_rsch_exp = (public_serv_exp + rsch_exp),
    cm_ready = (content_mastery + readiness) / 2,  english_cm = english - cm_ready,
    math_cm = math - cm_ready, science_cm = science - cm_ready,
    social_studies_cm = social_studies - cm_ready,
    cip_categories = case_when(cip_categories == 'Social Sciences' ~ 1,
           cip_categories == 'Fine Arts' ~ 2, cip_categories == 'Human Services' ~ 3,
           cip_categories == 'Business' ~ 4, cip_categories == 'STEM' ~ 5,
           cip_categories == 'General/Interdisciplinary Studies' ~ 6,
           cip_categories == 'Healthcare' ~ 7, cip_categories == 'Education' ~ 8,
           TRUE ~ 9),
    zell_ind = case_when(zell_ind == 'Y' ~ 1, TRUE ~ 0),
    locale_group = case_when(locale_group == 'City' ~ 1, locale_group == 'Suburb' ~ 2,
                      locale_group == 'Town' ~ 3, TRUE ~ 4),
    race_eth = case_when(race_eth == 'White' ~ 1,
                   race_eth == 'Black or African American' ~ 2,
                   race_eth == 'Hispanic or Latino' ~ 3, TRUE ~ 4),
    adv_standing_ib_hrs = adv_standing_ib_hrs,
    adv_standing_clep_hrs = adv_standing_clep_hrs,
    adv_standing_other_hrs = adv_standing_other_hrs) %>%
step_rm(cpc_english_code, cpc_foreign_language_code,
    cpc_math_code, cpc_science_code, cpc_social_science_code, acay_sup_exp,
    inst_sup_exp, public_serv_exp, rsch_exp, english, math, science,
    social_studies, dv_next_fall, content_mastery, readiness,
    dv_first_fall_gpa, hs_code, hs_grad_year, state_school_id, locale_code, locale) %>%
 step_impute_knn(c(hs_gpa, adm_test_score, expected_family_contribution),
    neighbors = 10) %>%
 step_rename(hsgpa_knn = hs_gpa,  ats_knn = adm_test_score,
    efc_knn = expected_family_contribution) %>%
 step_YeoJohnson(hsgpa_knn, ats_knn, college_prep, adv_standing_ap_hrs,
    adv_standing_clep_hrs, adv_standing_ib_hrs, adv_standing_other_hrs, cm_ready,
      english_cm, math_cm, science_cm, social_studies_cm, efc_knn, ga_hope, pell,
      fed_sub_loans, fed_unsub_loans, oth_loans, acay_inst_sup_exp,  all_other_exp,
      instr_exp, stu_serv_exp, public_rsch_exp) %>%
 step_normalize(hsgpa_knn,  ats_knn, college_prep, adv_standing_ap_hrs,
      adv_standing_clep_hrs, adv_standing_ib_hrs, adv_standing_other_hrs, cm_ready,
      college_prep, english_cm, math_cm, science_cm, social_studies_cm, efc_knn,
      ga_hope, pell, fed_sub_loans, fed_unsub_loans, oth_loans, acay_inst_sup_exp,
      all_other_exp, instr_exp, stu_serv_exp, public_rsch_exp)

#######################
## CLEAN DATA SETS ##
#######################

## produce clean training data set
fy_gpa <- fygpa_rec %>% prep() %>% juice()
fy_gpa_rec <- recipe(dv_first_yr_gpa ~ ., data = fy_gpa %>% select(-unique_identifer))
## processing testing data
fy_gpa_test <- fygpa_rec %>% prep() %>% bake(dat_test) %>%
 filter(!is.na(dv_first_yr_gpa))

##########################
## CROSS-VALIDATIONS ##
##########################

## training cross validation
set.seed(51823)
fy_gpa_cv <- vfold_cv(fy_gpa %>% select(-unique_identifer), v = 10)
## testing cross validation
set.seed(51823)
fy_gpa_cv_t <- vfold_cv(fy_gpa_test %>% select(-unique_identifer), v = 10)

########################
##TUNING CONTROLS ##
########################

## control grid set up
ctrl_grid <- control_grid(save_pred = TRUE, save_workflow = TRUE)
## model_metrics
model_metrics <- metric_set(rmse, rsq)

##########################
##LINEAR REGRESSION ##
##########################

## training data set
gpa_ln <- lm(formula = dv_first_yr_gpa ~ .,
           data = fy_gpa %>% select(-unique_identifer))
## model summary
gpa_ln %>% summary()
## rmse
data.frame(predicted = predict(object = gpa_ln, fy_gpa),
               actuals = fy_gpa$dv_first_yr_gpa) %>%
 mutate(diff = (predicted - actuals)^2) %>%
 select(diff) %>% summarise(rmse = sqrt(mean(diff)))
## coefficients aka beta weights
gpa_ln %>% coefficients()
## confidence intervals of coefficients
gpa_ln %>% confint()
## standardized regression coefficients
gpa_ln %>% lm.beta()
## standardized and unstandardized coefficients
gpa_ln %>% standardCoefs()

## linear regression assumptions
## correlation with dv
fy_gpa_corr <- fy_gpa %>% select(-unique_identifer) %>%
 corr.test(use = 'pairwise', method = 'pearson', adjust = 'holm', alpha = .05)
## VIF
gpa_ln %>% VIF() %>% data.frame()
## means of errors
gpa_ln %>% rstandard() %>%mean()
gpa_ln %>% rstudent() %>% mean()
##correlation amongst the residuals
gpa_ln %>% durbinWatsonTest()
## homogeneity of variance
gpa_ln %>% ncvTest()
## normality of residuals
gpa_ln %>% rstudent() %>% LillieTest();
(gpa_ln %>% rstudent())[0:5000] %>% shapiro.test()
gpa_ln %>% rstudent() %>% jarque.bera.test()

## results of the model on testing data set
gpa_ln_test <- lm(formula = dv_first_yr_gpa ~ .,
               data = fy_gpa_test %>% select(-unique_identifer))
## model summary
gpa_ln_test %>% summary()
## rmse
data.frame(predicted = predict(object = gpa_ln_test, fy_gpa_test),
          actuals = fy_gpa_test$dv_first_yr_gpa) %>%
 mutate(diff = (predicted - actuals)^2) %>%  select(diff) %>%
 summarise(rmse = sqrt(mean(diff)))
```

```
## variable importance analysis
## creating the tidymodels linear regression
ln_reg <- linear_reg() %>% set_engine(engine = 'lm') %>% set_mode('regression')
## tidymodels workflow
ln_reg_wf <- workflow() %>% add_model(ln_reg) %>% add_recipe(fy_gpa_rec)
lin_reg_vi <- fit(ln_reg_wf, fy_gpa) %>% extract_fit_parsnip() %>%  vi() %>%
 left_join(fit(ln_reg_wf, fy_gpa) %>% extract_fit_parsnip() %>%
          vi(scale = TRUE) %>% select(-Sign) %>%
          rename(rescaled_importance = Importance))
lin_reg_vi_test <- fit(ln_reg_wf, fy_gpa_test) %>% extract_fit_parsnip() %>% vi() %>%
 left_join(fit(ln_reg_wf, fy_gpa_test) %>% extract_fit_parsnip() %>%
          vi(scale = TRUE) %>% select(-Sign) %>%
          rename(rescaled_importance = Importance))

lin_reg_vi %>% mutate(data_set = '1. Training') %>%
 rbind(lin_reg_vi_test %>% mutate(data_set = '2. Testing')) %>%
 rename(Impact = Sign) %>%
 mutate(Impact = case_when(Impact == 'NEG' ~ 'Negative', TRUE ~ 'Positive'),
       Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                       Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                       Variable == 'hsgpa_knn' ~ 'HS GPA',
                       Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                       Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                       Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                       Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                       Variable == 'cip_categories' ~ 'Major Groupings',
                       Variable == 'efc_knn' ~ 'EFC',
                       Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                       Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                       Variable == 'pell' ~ 'PELL Grant',
                       Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                       Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                       Variable == 'oth_loans' ~ 'Other Loans',
                       Variable == 'ats_knn' ~ 'Admissions Test Scores',
                       Variable == 'all_other_exp' ~ 'All Other',
                       Variable == 'instr_exp' ~ 'Instruction',
                       Variable == 'stu_serv_exp' ~ 'Student Services',
                       Variable == 'race_eth' ~ 'Race Ethnicity',
                       Variable == 'cm_ready' ~ 'CM & Ready Mean',
                       Variable == 'locale_group' ~ 'HS Locale',
                       Variable == 'college_prep' ~ 'College Prep. Curric.',
                       Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                       Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                       Variable == 'english_cm' ~ 'English (CMR)',
                       Variable == 'math_cm' ~ 'Math (CMR)',
                       Variable == 'science_cm' ~ 'Science (CMR)',
                        Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                        TRUE ~ 'CHECK')) %>%
 ggplot(aes(x = reorder(Variable, abs(rescaled_importance)), y = rescaled_importance,
          fill = Impact)) + geom_bar(aes(fill = Impact), stat = 'identity') +
 theme_classic() + ylab("Rescaled Importance") +
 theme(legend.position = 'top', axis.title.y = element_blank(),
      text = element_text(size = 15)) + coord_flip() + facet_wrap(.~ data_set)

## predictive power
## assessing training data set
ln_reg_wf_eval <- ln_reg_wf %>%
 fit_resamples(dv_first_yr_gpa ~., resamples = fy_gpa_cv,
             metrics = model_metrics, control = ctrl_grid)
ln_reg_wf_eval %>% collect_metrics

## assessing testing data set
ln_reg_wf_eval_t <- ln_reg_wf %>%
 fit_resamples(dv_first_yr_gpa ~., resamples = fy_gpa_cv_t,
             metrics = model_metrics, control = ctrl_grid)
ln_reg_wf_eval_t %>% collect_metrics

#########################################
##Support Vector Machine-Linear Kernel ##
#########################################

## model specifications
svm_l_spec <- svm_linear(cost = tune(),margin = tune()) %>%
 set_mode('regression') %>%set_engine('kernlab')
## model workflow
svm_l_wf <- workflow() %>% add_model(svm_l_spec) %>% add_recipe(fy_gpa_rec)
## tuning
set.seed(52323)
svm_l_tune <- tune_grid(svm_l_wf,  resamples = fy_gpa_cv,  grid = 20)
## selecting best model
svm_l_final <- select_best(svm_l_tune, 'rmse')
## model fixed to the best outcome model
svml_tune_final <- finalize_model(svm_l_spec, svm_l_final)

## variable importance analysis
## training data set
set.seed(51923)
svml_fit <- workflow() %>% add_model(svml_tune_final)  %>%
 add_recipe(fy_gpa_rec) %>% fit(fy_gpa %>% select(-unique_identifer))
set.seed(51923)
svml_vi <- svml_fit %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = FALSE,  pred_wrapper = kernlab::predict,
   metric = 'rmse', target = 'dv_first_yr_gpa',
   train = fy_gpa %>% select(-unique_identifer))
set.seed(51923)
svml_vi_rs <- svml_fit %>%extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE, pred_wrapper = kernlab::predict,
   metric = 'rmse', target = 'dv_first_yr_gpa',
   train = fy_gpa %>% select(-unique_identifer)) %>%
 rename(rescaled_importance = Importance)

## testing data set
## vip of the xgb
set.seed(51923)
svml_fit_test <- workflow() %>%  add_model(svml_tune_final)  %>%
 add_recipe(fy_gpa_rec) %>%  fit(fy_gpa_test %>%select(-unique_identifer))
set.seed(51923)
svml_vi_test <- svml_fit_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = FALSE,  pred_wrapper = kernlab::predict,
   metric = 'rmse', target = 'dv_first_yr_gpa',
   train = fy_gpa_test %>% select(-unique_identifer))
set.seed(51923)
svml_vi_rs_test <- svml_fit_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE,  pred_wrapper = kernlab::predict,
   metric = 'rmse', target = 'dv_first_yr_gpa',
   train = fy_gpa_test %>% select(-unique_identifer)) %>%
 rename(rescaled_importance = Importance)

svml_vi_rs %>% mutate(data_set = '1. Training') %>%
 rbind(svml_vi_rs_test %>% mutate(data_set = '2. Testing')) %>%
 mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                      Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                      Variable == 'hsgpa_knn' ~ 'HS GPA',
                      Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                      Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                      Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                      Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                      Variable == 'cip_categories' ~ 'Major Groupings',
                      Variable == 'efc_knn' ~ 'EFC',
                      Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                      Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                      Variable == 'pell' ~ 'PELL Grant',
                      Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                      Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                      Variable == 'oth_loans' ~ 'Other Loans',
                      Variable == 'ats_knn' ~ 'Admissions Test Scores',
                      Variable == 'all_other_exp' ~ 'All Other',
                      Variable == 'instr_exp' ~ 'Instruction',
                      Variable == 'stu_serv_exp' ~ 'Student Services',
                      Variable == 'race_eth' ~ 'Race Ethnicity',
                      Variable == 'cm_ready' ~ 'CM & Ready Mean',
                      Variable == 'locale_group' ~ 'HS Locale',
                      Variable == 'college_prep' ~ 'College Prep. Curric.',
                      Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                      Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                      Variable == 'english_cm' ~ 'English (CMR)',
                      Variable == 'math_cm' ~ 'Math (CMR)',
                      Variable == 'science_cm' ~ 'Science (CMR)',
                      Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                      TRUE ~ 'CHECK')) %>%
 ggplot(aes(x = reorder(Variable, rescaled_importance), y = rescaled_importance)) +
 geom_bar(aes(fill = rescaled_importance/10),  stat = 'identity') + theme_classic() +
 theme(legend.position = 'none',  axis.title.y = element_blank(),
     text = element_text(size = 15)) +
 ylab('Rescaled Importance') + coord_flip() + facet_wrap(. ~ data_set)

## predictive power
## assessing training data
doParallel::registerDoParallel()
set.seed(51923)
svm_l_cv <- svml_tune_final %>%
 fit_resamples(dv_first_yr_gpa ~., resamples = fy_gpa_cv,
          metrics = model_metrics,  control = ctrl_grid)
svm_l_cv %>% collect_metrics()

## assessing testing data
doParallel::registerDoParallel()
set.seed(51923)
svm_l_cv_t <- svml_tune_final %>%
 fit_resamples(dv_first_yr_gpa ~., resamples = fy_gpa_cv_t,
          metrics = model_metrics, control = ctrl_grid)
svm_l_cv_t %>% collect_metrics()

##############################################
##Support Vector Machine-Polynomial Kernel ##
##############################################

## model specifications
svm_p_spec <- svm_poly(cost = tune(), degree = tune(), scale_factor = tune(),
               margin = tune()) %>% set_mode('regression') %>% set_engine('kernlab')
## model workflow
svm_p_wf <- workflow() %>% add_model(svm_p_spec) %>% add_recipe(fy_gpa_rec)
## tuning model
set.seed(52323)
```

```
svm_p_tune <- tune_grid(svm_p_wf, resamples = fy_gpa_cv, grid = 20)
## selecting best model
svm_p_final <- select_best(svm_p_tune, 'rmse')
## model fixed to the best outcome model
svmp_tune_final <- finalize_model(svm_p_spec, svm_p_final)

## variable importance analysis
## training data set
set.seed(51923)
svmp_fit <- workflow() %>% add_model(svmp_tune_final)  %>%
 add_recipe(fy_gpa_rec) %>% fit(fy_gpa %>% select(-unique_identifer))
set.seed(51923)
svmp_vi <- svmp_fit %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = FALSE,  pred_wrapper = kernlab::predict,
   metric = 'rmse', target = 'dv_first_yr_gpa',
   train = fy_gpa %>% select(-unique_identifer))
set.seed(51923)
svmp_vi_rs <- svmp_fit %>%extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE,  pred_wrapper = kernlab::predict,
   metric = 'rmse', target = 'dv_first_yr_gpa',
   train = fy_gpa %>% select(-unique_identifer)) %>%
 rename(rescaled_importance = Importance)

## testing data set
set.seed(51923)
svmp_fit_test <- workflow() %>%add_model(svmp_tune_final)  %>%
 add_recipe(fy_gpa_rec) %>% fit(fy_gpa_test %>% select(-unique_identifer))
set.seed(51923)
svmp_vi_test <- svmp_fit_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = FALSE,  pred_wrapper = kernlab::predict,
   metric = 'rmse', target = 'dv_first_yr_gpa',
   train = fy_gpa_test %>% select(-unique_identifer))

set.seed(51923)
svmp_vi_rs_test <- svmp_fit_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE,  pred_wrapper = kernlab::predict,
   metric = 'rmse', target = 'dv_first_yr_gpa',
   train = fy_gpa_test %>% select(-unique_identifer)) %>%
 rename(rescaled_importance = Importance)

svmp_vi_rs %>% mutate(data_set = '1. Training') %>%
 rbind(svmp_vi_rs_test %>% mutate(data_set = '2. Testing')) %>%
 mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                         Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                         Variable == 'hsgpa_knn' ~ 'HS GPA',
                         Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                         Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                         Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                         Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                         Variable == 'cip_categories' ~ 'Major Groupings',
                         Variable == 'efc_knn' ~ 'EFC',
                         Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                         Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                         Variable == 'pell' ~ 'PELL Grant',
                         Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                         Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                         Variable == 'oth_loans' ~ 'Other Loans',
                         Variable == 'ats_knn' ~ 'Admissions Test Scores',
                         Variable == 'all_other_exp' ~ 'All Other',
                         Variable == 'instr_exp' ~ 'Instruction',
                         Variable == 'stu_serv_exp' ~ 'Student Services',
                         Variable == 'race_eth' ~ 'Race Ethnicity',
                         Variable == 'cm_ready' ~ 'CM & Ready Mean',
                         Variable == 'locale_group' ~ 'HS Locale',
                         Variable == 'college_prep' ~ 'College Prep. Curric.',
                         Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                         Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                         Variable == 'english_cm' ~ 'English (CMR)',
                         Variable == 'math_cm' ~ 'Math (CMR)',
                         Variable == 'science_cm' ~ 'Science (CMR)',
                         Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                         TRUE ~ 'CHECK')) %>%
 ggplot(aes(x = reorder(Variable, rescaled_importance), y = rescaled_importance)) +
 geom_bar(aes(fill = rescaled_importance/10), stat = 'identity') + theme_classic() +
 theme(legend.position = 'none', axis.title.y = element_blank(),
       text = element_text(size = 15)) +
 ylab('Rescaled Importance') + coord_flip() +  facet_wrap(. ~ data_set)

## predictive power
## assessing training data set
doParallel::registerDoParallel()
set.seed(51923)
svm_p_cv <- svmp_tune_final %>%
 fit_resamples(dv_first_yr_gpa ~., resamples = fy_gpa_cv,
              metrics = model_metrics, control = ctrl_grid)
svm_p_cv %>% collect_metrics()

## assessing testing data set
doParallel::registerDoParallel()
set.seed(51923)
svm_p_cv_t <- svmp_tune_final %>%
 fit_resamples(dv_first_yr_gpa ~., resamples = fy_gpa_cv_t,
              metrics = model_metrics, control = ctrl_grid)
svm_p_cv_t %>% collect_metrics()

#####################################################
##Support Vector Machine-Radial Basis Function Kernel ##
#####################################################

## model specifications
svm_r_spec <- svm_rbf(cost = tune(), rbf_sigma = tune(),margin = tune()) %>%
 set_mode('regression') %>% set_engine('kernlab')
## model workflow
svm_r_wf <- workflow() %>% add_model(svm_r_spec) %>% add_recipe(fy_gpa_rec)
## tuning model
set.seed(52323)
svm_r_tune <- tune_grid(svm_r_wf, resamples = fy_gpa_cv, grid = 20)
## selecting best model
svm_r_final <- select_best(svm_r_tune, 'rmse')
## model fixed to the best outcome model
svmr_tune_final <- finalize_model(svm_r_spec, svm_r_final)

## variable importance analysis
## training data set
set.seed(51923)
svmr_fit <- workflow() %>% add_model(svmr_tune_final)  %>%
 add_recipe(fy_gpa_rec) %>% fit(fy_gpa %>% select(-unique_identifer))
set.seed(51923)
svmr_vi <- svmr_fit %>%extract_fit_parsnip() %>%
 vi(method = 'permute', scale = FALSE, pred_wrapper = kernlab::predict,
    metric = 'rmse', target = 'dv_first_yr_gpa',
    train = fy_gpa %>%select(-unique_identifer))
set.seed(51923)
svmr_vi_rs <- svmr_fit %>%extract_fit_parsnip() %>%
 vi(method = 'permute', scale = TRUE, pred_wrapper = kernlab::predict,
    metric = 'rmse', target = 'dv_first_yr_gpa',
    train = fy_gpa %>% select(-unique_identifer)) %>%
 rename(rescaled_importance = Importance)

# testing data set
set.seed(51923)
svmr_fit_test <- workflow() %>% add_model(svmr_tune_final)  %>%
 add_recipe(fy_gpa_rec) %>% fit(fy_gpa_test %>% select(-unique_identifer))
set.seed(51923)
svmr_vi_test <- svmr_fit_test %>%extract_fit_parsnip() %>%
 vi(method = 'permute', scale = FALSE, pred_wrapper = kernlab::predict,
    metric = 'rmse', target = 'dv_first_yr_gpa',
    train = fy_gpa_test %>%select(-unique_identifer))
```

```
set.seed(51923)
svmr_vi_rs_test <- svmr_fit_test %>%extract_fit_parsnip() %>%
 vi(method = 'permute', scale = TRUE, pred_wrapper = kernlab::predict,
    metric = 'rmse', target = 'dv_first_yr_gpa',
    train = fy_gpa_test %>%select(-unique_identifer)) %>%
 rename(rescaled_importance = Importance)

svmr_vi_rs %>% mutate(data_set = '1. Training') %>%
 rbind(svmr_vi_rs_test %>% mutate(data_set = '2. Testing')) %>%
 mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                             Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                             Variable == 'hsgpa_knn' ~ 'HS GPA',
                             Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                             Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                             Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                             Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                             Variable == 'cip_categories' ~ 'Major Groupings',
                             Variable == 'efc_knn' ~ 'EFC',
                             Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                             Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                             Variable == 'pell' ~ 'PELL Grant',
                             Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                             Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                             Variable == 'oth_loans' ~ 'Other Loans',
                             Variable == 'ats_knn' ~ 'Admissions Test Scores',
                             Variable == 'all_other_exp' ~ 'All Other',
                             Variable == 'instr_exp' ~ 'Instruction',
                             Variable == 'stu_serv_exp' ~ 'Student Services',
                             Variable == 'race_eth' ~ 'Race Ethnicity',
                             Variable == 'cm_ready' ~ 'CM & Ready Mean',
                             Variable == 'locale_group' ~ 'HS Locale',
                             Variable == 'college_prep' ~ 'College Prep. Curric.',
                             Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                             Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                             Variable == 'english_cm' ~ 'English (CMR)',
                             Variable == 'math_cm' ~ 'Math (CMR)',
                             Variable == 'science_cm' ~ 'Science (CMR)',
                             Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                             TRUE ~ 'CHECK')) %>%
 ggplot(aes(x = reorder(Variable, rescaled_importance), y = rescaled_importance)) +
 geom_bar(aes(fill = rescaled_importance/10),  stat = 'identity') + theme_classic() +
 theme(legend.position = 'none',  axis.title.y = element_blank(),
      text = element_text(size = 15)) +
 ylab('Rescaled Importance') + coord_flip() + facet_wrap(. ~ data_set)
```

```
## predictive power
## assessing training data set
doParallel::registerDoParallel()
set.seed(51923)
svm_r_cv <- svmr_tune_final %>%
 fit_resamples(dv_first_yr_gpa ~., resamples = fy_gpa_cv,
             metrics = model_metrics,  control = ctrl_grid)
svm_r_cv %>% collect_metrics()

## assessing testing data set
doParallel::registerDoParallel()
set.seed(51923)
svm_r_cv_t <- svmr_tune_final %>%
 fit_resamples(dv_first_yr_gpa ~., resamples = fy_gpa_cv_t,
             metrics = model_metrics,  control = ctrl_grid)
svm_r_cv_t %>% collect_metrics()

######################
##RANDOM FOREST ##
######################

## model specifications
rf_spec <- rand_forest(  mtry = tune(), trees = tune(),min_n = tune()) %>%
 set_mode("regression") %>% set_engine("ranger")
## workflow
rf_wf <- workflow() %>% add_model(rf_spec) %>% add_recipe(fy_gpa_rec)
## tuning model
doParallel::registerDoParallel()
set.seed(51823)
rf_wf_tune <- tune_grid(rf_wf, resamples = fy_gpa_cv, grid = 20)
## best model
rf_tune_best <- select_best(rf_wf_tune, 'rmse')
## model fixed to the best outcome model
rf_tune_final <- finalize_model(rf_spec, rf_tune_best)

## variable importance analysis
## training data set
set.seed(511923)
rf_vi <- rf_tune_final %>% set_engine('ranger', importance = 'permutation') %>%
 fit(dv_first_yr_gpa ~ .,  data = fy_gpa %>% select(-unique_identifer)) %>% vi()
set.seed(511923)
rf_vi_rs <- rf_tune_final %>% set_engine('ranger', importance = 'permutation') %>%
 fit(dv_first_yr_gpa ~ ., data = fy_gpa %>% select(-unique_identifer)) %>%
 vi(scale = TRUE) %>% rename(rescaled_importance = Importance)
```

```
## testing data set
set.seed(511923)
rf_vi_test <- rf_tune_final %>% set_engine('ranger', importance = 'permutation') %>%
 fit(dv_first_yr_gpa ~ .,  data = fy_gpa_test %>% select(-unique_identifer)) %>% vi()
set.seed(511923)
rf_vi_rs_test <- rf_tune_final %>% set_engine('ranger', importance = 'permutation') %>%
 fit(dv_first_yr_gpa ~ .,  data = fy_gpa_test %>% select(-unique_identifer)) %>%
 vi(scale = TRUE) %>% rename(rescaled_importance = Importance)

rf_vi_rs %>% mutate(data_set = '1. Training') %>%
 rbind(rf_vi_rs_test %>% mutate(data_set = '2. Testing')) %>%
 mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                          Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                          Variable == 'hsgpa_knn' ~ 'HS GPA',
                          Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                          Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                          Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                          Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                          Variable == 'cip_categories' ~ 'Major Groupings',
                          Variable == 'efc_knn' ~ 'EFC',
                          Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                          Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                          Variable == 'pell' ~ 'PELL Grant',
                          Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                          Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                          Variable == 'oth_loans' ~ 'Other Loans',
                          Variable == 'ats_knn' ~ 'Admissions Test Scores',
                          Variable == 'all_other_exp' ~ 'All Other',
                          Variable == 'instr_exp' ~ 'Instruction',
                          Variable == 'stu_serv_exp' ~ 'Student Services',
                          Variable == 'race_eth' ~ 'Race Ethnicity',
                          Variable == 'cm_ready' ~ 'CM & Ready Mean',
                          Variable == 'locale_group' ~ 'HS Locale',
                          Variable == 'college_prep' ~ 'College Prep. Curric.',
                          Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                          Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                          Variable == 'english_cm' ~ 'English (CMR)',
                          Variable == 'math_cm' ~ 'Math (CMR)',
                          Variable == 'science_cm' ~ 'Science (CMR)',
                          Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                          TRUE ~ 'CHECK')) %>%
 ggplot(aes(x = reorder(Variable, rescaled_importance), y = rescaled_importance)) +
 geom_bar(aes(fill = rescaled_importance/10),  stat = 'identity') +
 ylab('Rescaled Importance') + theme_classic() +
 theme(legend.position = 'none',  axis.title.y = element_blank(),
      text = element_text(size = 15)) + coord_flip() + facet_wrap(.~ data_set)
```

```
## predictive power
rf_wf_final <- workflow() %>% add_model(rf_tune_final) %>% add_recipe(fy_gpa_rec)
## assessing training data set
doParallel::registerDoParallel()
set.seed(51923)
rf_train <- rf_wf_final %>%
 fit_resamples(dv_first_yr_gpa ~., resamples = fy_gpa_cv,
             metrics = model_metrics, control = ctrl_grid)
rf_train %>% collect_metrics()

## assessing testing data set
doParallel::registerDoParallel()
set.seed(51923)
rf_test <- rf_wf_final %>%
 fit_resamples(dv_first_yr_gpa ~.,  resamples = fy_gpa_cv_t,
             metrics = model_metrics,  control = ctrl_grid)
rf_test %>% collect_metrics()

######################################
##EXTREME GRADIENT BOOSTING ##
######################################

## building out model specs
xgb <- boost_tree(trees = tune(), tree_depth = tune(),min_n = tune(),
 loss_reduction = tune(), sample_size = tune(),  mtry = tune(),  learn_rate = tune()) %>%
 set_engine('xgboost') %>% set_mode('regression')
xgb_wf <- workflow() %>% add_model(xgb) %>% add_recipe(fy_gpa_rec)
## tuning model
doParallel::registerDoParallel()
set.seed(51923)
xgb_wf_tune <- tune_grid(xgb_wf, resamples = fy_gpa_cv, grid = 20)
## best model
xgb_tune_best <- select_best(xgb_wf_tune, 'rmse')
## model fixed to the best outcome model
xgb_tune_final <- finalize_model(xgb, xgb_tune_best)

## variable importance analysis
## training data set
set.seed(51923)
xgb_vi <- xgb_tune_final %>% set_engine('xgboost') %>%
 fit(dv_first_yr_gpa ~ .,  data = fy_gpa %>% select(-unique_identifer)) %>% vi()
set.seed(51923)
xgb_vi_rs <- xgb_tune_final %>% set_engine('xgboost') %>%
 fit(dv_first_yr_gpa ~ ., data = fy_gpa %>%select(-unique_identifer)) %>%
 vi(scale = TRUE) %>% rename(rescaled_importance = Importance)
```

```
## testing data set
set.seed(51923)
xgb_vi_test <- xgb_tune_final %>% set_engine('xgboost') %>%
 fit(dv_first_yr_gpa ~ .,  data = fy_gpa_test %>% select(-unique_identifer)) %>% vi()
set.seed(51923)
xgb_vi_rs_test <- xgb_tune_final %>% set_engine('xgboost') %>%
 fit(dv_first_yr_gpa ~ .,
     data = fy_gpa_test %>% select(-unique_identifer)) %>% vi(scale = TRUE) %>%
 rename(rescaled_importance = Importance)

xgb_vi_rs %>% mutate(data_set = '1. Training') %>%
 rbind(xgb_vi_rs_test %>% mutate(data_set = '2. Testing')) %>%
 mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                          Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                          Variable == 'hsgpa_knn' ~ 'HS GPA',
                          Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                          Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                          Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                          Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                          Variable == 'cip_categories' ~ 'Major Groupings',
                          Variable == 'efc_knn' ~ 'EFC',
                          Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                          Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                          Variable == 'pell' ~ 'PELL Grant',
                          Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                          Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                          Variable == 'oth_loans' ~ 'Other Loans',
                          Variable == 'ats_knn' ~ 'Admissions Test Scores',
                          Variable == 'all_other_exp' ~ 'All Other',
                          Variable == 'instr_exp' ~ 'Instruction',
                          Variable == 'stu_serv_exp' ~ 'Student Services',
                          Variable == 'race_eth' ~ 'Race Ethnicity',
                          Variable == 'cm_ready' ~ 'CM & Ready Mean',
                          Variable == 'locale_group' ~ 'HS Locale',
                          Variable == 'college_prep' ~ 'College Prep. Curric.',
                          Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                          Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                          Variable == 'english_cm' ~ 'English (CMR)',
                          Variable == 'math_cm' ~ 'Math (CMR)',
                          Variable == 'science_cm' ~ 'Science (CMR)',
                          Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                          TRUE ~ 'CHECK')) %>%
 ggplot(aes(x = reorder(Variable, rescaled_importance), y = rescaled_importance)) +
 geom_bar(aes(fill = rescaled_importance/10),  stat = 'identity') +
 ylab('Rescaled Importance') + theme_classic() +
 theme(legend.position = 'none',  axis.title.y = element_blank(),
      text = element_text(size = 15)) + coord_flip() + facet_wrap(. ~ data_set)

## predictive power
xgb_wf_final <- workflow() %>% add_model(xgb_tune_final) %>%
 add_recipe(fy_gpa_rec)
## assessing training data
doParallel::registerDoParallel()
set.seed(51923)
xgb_train <- xgb_wf_final %>%
 fit_resamples(dv_first_yr_gpa ~., resamples = fy_gpa_cv,
              metrics = model_metrics, control = ctrl_grid)
xgb_train %>%collect_metrics()

## assessing testing data
doParallel::registerDoParallel()
set.seed(51923)
xgb_test <- xgb_wf_final %>%
 fit_resamples(dv_first_yr_gpa ~., resamples = fy_gpa_cv_t,
              metrics = model_metrics,  control = ctrl_grid)
xgb_test %>% collect_metrics()

########################################################
## VARIABLE COMPARISON OF TESTING DATA SETS ##
########################################################

lin_reg_vi_test %>% select(Variable, rescaled_importance) %>%
 mutate(type = 'Linear Regression') %>%
 rbind(svml_vi_rs_test %>% mutate(type = 'SVM Linear')) %>%
 rbind(svmp_vi_rs_test %>% mutate(type = 'SVM Polynomial')) %>%
 rbind(svmr_vi_rs_test %>% mutate(type = 'SVM Radial')) %>%
 rbind(rf_vi_rs_test %>% mutate(type = 'Random Forest')) %>%
 rbind(xgb_vi_rs_test %>% mutate(type = 'XGBoost')) %>%
 mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                         Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                         Variable == 'hsgpa_knn' ~ 'HS GPA',
                         Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                         Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                         Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                         Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                         Variable == 'cip_categories' ~ 'Major Groupings',
                         Variable == 'efc_knn' ~ 'EFC',
                         Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                         Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                         Variable == 'pell' ~ 'PELL Grant',
                         Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                         Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                         Variable == 'oth_loans' ~ 'Other Loans',
                         Variable == 'ats_knn' ~ 'Admissions Test Scores',
                         Variable == 'all_other_exp' ~ 'All Other',
                         Variable == 'instr_exp' ~ 'Instruction',
                         Variable == 'stu_serv_exp' ~ 'Student Services',
                         Variable == 'race_eth' ~ 'Race Ethnicity',
                         Variable == 'cm_ready' ~ 'CM & Ready Mean',
                         Variable == 'locale_group' ~ 'HS Locale',
                         Variable == 'college_prep' ~ 'College Prep. Curric.',
                         Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                         Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                         Variable == 'english_cm' ~ 'English (CMR)',
                         Variable == 'math_cm' ~ 'Math (CMR)',
                         Variable == 'science_cm' ~ 'Science (CMR)',
                         Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                         TRUE ~ 'CHECK')) %>%
 ggplot(aes(x = reorder(Variable, desc(Variable)), y = rescaled_importance)) +
 geom_bar(aes(fill = rescaled_importance/10),  stat = 'identity') + theme_classic() +
 theme(legend.position = 'none',  axis.title.y = element_blank(),
      text = element_text(size = 20)) +
 ylab('Rescaled Importance') + coord_flip()+ facet_wrap(. ~ type,  ncol = 6)

############################
##ENSEMBLE LEARNING ##
############################

## pulling out predictions from training data set
train_pred <- ln_reg_wf_eval %>%collect_predictions() %>%
 select(dv_first_yr_gpa, .pred) %>% rename(linear_reg = .pred) %>%
 cbind(svm_r_cv %>% collect_predictions() %>%
       rename(svmrbf = .pred) %>%select(svmrbf)) %>%
 cbind(rf_train %>% collect_predictions() %>%
       rename(rf = .pred) %>% select(rf)) %>%
 cbind(xgb_train %>% collect_predictions() %>%
       rename(xgb = .pred) %>% select(xgb))
## pulling out the predictions from testing data set
test_pred <- ln_reg_wf_eval_t %>%collect_predictions() %>%
 select(dv_first_yr_gpa, .pred) %>% rename(linear_reg = .pred) %>%
 cbind(svm_r_cv_t %>% collect_predictions() %>%
       rename(svmrbf = .pred) %>% select(svmrbf)) %>%
 cbind(rf_test %>% collect_predictions() %>%
       rename(rf = .pred) %>% select(rf)) %>%
 cbind(xgb_test %>% collect_predictions() %>%
       rename(xgb = .pred) %>% select(xgb))
```

```
## mean method
## training data set
train_metrics_rv <- train_metrics %>% filter(.metric == 'rmse') %>%
 select(-.metric) %>%
 rbind(train_pred %>%
       mutate(mean_pred = (linear_reg + svmrbf + rf + xgb) / 4,
              diff = mean_pred - dv_first_yr_gpa,
              diff = diff^2) %>% select(diff) %>%
       summarise(mean = mean(diff),  mean = sqrt(mean)) %>%
       mutate(`Data Set` = '1. Training',  model = '7. Ensemble Mean'))

## testing data set
test_metrics_rv <- test_metrics %>% filter(.metric == 'rmse') %>%
 select(-.metric) %>%
 rbind(test_pred %>%
       mutate(mean_pred = (linear_reg + svmrbf + rf + xgb) / 4,
              diff = mean_pred - dv_first_yr_gpa,
              diff = diff^2) %>% select(diff) %>%
       summarise(mean = mean(diff), mean = sqrt(mean)) %>%
       mutate(`Data Set` = '2. Testing', model = '7. Ensemble Mean'))

## blended method
org_stack <- stacks() %>% add_candidates(ln_reg_wf_eval) %>%
 add_candidates(svm_r_cv) %>% add_candidates(rf_train) %>%
 add_candidates(xgb_train)
set.seed(7423)
org_stack_fit <- org_stack %>% blend_predictions() %>%
 fit_members()

## training data set
blend_train_lr <- linear_reg(penalty = (org_stack_fit$penalty)$penalty,
                  mixture = (org_stack_fit$penalty)$mixture) %>%
 set_engine('lm') %>% set_mode('regression') %>%
 fit(dv_first_yr_gpa ~ ., data = train_pred)
train_metrics_rv <- train_metrics_rv %>%
 rbind(train_pred %>% cbind(blend_train_lr %>% predict(train_pred)) %>%
       mutate(diff = .pred - dv_first_yr_gpa, diff = diff^2) %>% select(diff) %>%
       summarise(mean = mean(diff), mean = sqrt(mean)) %>%
       mutate(`Data Set` = '1. Training', model = '8. Ensemble Blend'))

## testing data set
test_metrics_rv <- test_metrics_rv %>%
 rbind(test_pred %>% cbind(blend_train_lr %>% predict(test_pred)) %>%
       mutate(diff = .pred - dv_first_yr_gpa,  diff = diff^2) %>%
       select(diff) %>% summarise(mean = mean(diff),  mean = sqrt(mean)) %>%
       mutate(`Data Set` = '2. Testing',  model = '8. Ensemble Blend'))
```

```
train_metrics_rv %>% rbind(test_metrics_rv) %>% ggplot() +
 geom_bar(aes(x = reorder(model, desc(model)), y = mean,
            fill = `Data Set`),  stat = 'identity', position = 'dodge') +
 geom_text(aes(x = reorder(model, desc(model)),  y = mean,
             group = `Data Set`, label = format(round(mean, 3), nsmall = 3)),
          position = position_dodge(width = 1),  hjust = 1,  fontface = 'bold',
          size = 4) +
theme_classic() + theme(legend.position = 'top',  axis.title = element_blank(),
      text = element_text(size = 15)) + coord_flip()

########################
## RMSE COMPARISON ##
########################

rmse_rs <- ln_reg_wf_eval[[3]]  %>% as.data.frame() %>%
 filter(.metric == 'rmse') %>% gather(var_type, train) %>%
 filter(var_type %like% '%estimate%') %>%
 mutate(fold = 1:10,  model = '1. Linear Regression') %>%
 select(model, fold, train) %>%
 left_join(ln_reg_wf_eval_t[[3]]  %>% as.data.frame() %>%
          filter(.metric == 'rmse') %>% gather(var_type,  test) %>%
          filter(var_type %like% '%estimate%') %>%
          mutate(fold = 1:10,  model = '1. Linear Regression') %>%
          select(model, fold, test)) %>%
 rbind(svm_l_cv[[3]]  %>% as.data.frame() %>%
       filter(.metric == 'rmse') %>% gather(var_type,  train) %>%
       filter(var_type %like% '%estimate%') %>%
       mutate(fold = 1:10,  model = '2. SVM Linear Kernel') %>%
       select(model, fold, train) %>%
       left_join(svm_l_cv_t[[3]]  %>% as.data.frame() %>%
              filter(.metric == 'rmse') %>% gather(var_type,  test) %>%
              filter(var_type %like% '%estimate%') %>%
              mutate(fold = 1:10,  model = '2. SVM Linear Kernel') %>%
              select(model, fold, test))) %>%
 rbind(svm_p_cv[[3]]  %>% as.data.frame() %>%
       filter(.metric == 'rmse') %>% gather(var_type,  train) %>%
       filter(var_type %like% '%estimate%') %>%
       mutate(fold = 1:10,  model = '3. SVM Polynomial Kernel') %>%
       select(model, fold, train) %>%
       left_join(svm_p_cv_t[[3]]  %>% as.data.frame() %>%
              filter(.metric == 'rmse') %>% gather(var_type,  test) %>%
              filter(var_type %like% '%estimate%') %>%
              mutate(fold = 1:10,  model = '3. SVM Polynomial Kernel') %>%
              select(model, fold, test))) %>%
 rbind(svm_r_cv[[3]]  %>% as.data.frame() %>%
       filter(.metric == 'rmse') %>% gather(var_type,  train) %>%
       filter(var_type %like% '%estimate%') %>%
       mutate(fold = 1:10,  model = '4. SVM Radial BF Kernel') %>%
       select(model, fold, train) %>%
       left_join(svm_r_cv_t[[3]]  %>% as.data.frame() %>%
                filter(.metric == 'rmse') %>% gather(var_type,  test) %>%
                filter(var_type %like% '%estimate%') %>%
                mutate(fold = 1:10,  model = '4. SVM Radial BF Kernel') %>%
                select(model, fold, test))) %>%
 rbind(rf_train[[3]]  %>% as.data.frame() %>%
       filter(.metric == 'rmse') %>% gather(var_type,  train) %>%
       filter(var_type %like% '%estimate%') %>%
       mutate(fold = 1:10,  model = '5. Random Forest') %>%
       select(model, fold, train) %>%
       left_join(rf_test[[3]]  %>% as.data.frame() %>%
                filter(.metric == 'rmse') %>% gather(var_type,  test) %>%
                filter(var_type %like% '%estimate%') %>%
                mutate(fold = 1:10,  model = '5. Random Forest') %>%
                select(model, fold, test))) %>%
 rbind(xgb_train[[3]]  %>% as.data.frame() %>%
       filter(.metric == 'rmse') %>% gather(var_type,  train) %>%
       filter(var_type %like% '%estimate%') %>%
       mutate(fold = 1:10, model = '6. XGBoost') %>%
       select(model, fold, train) %>%
       left_join(xgb_test[[3]]  %>% as.data.frame() %>%
                filter(.metric == 'rmse') %>% gather(var_type,  test) %>%
                filter(var_type %like% '%estimate%') %>%
                mutate(fold = 1:10,  model = '6. XGBoost') %>%
                select(model, fold, test))) %>%
 mutate(train = as.numeric(train), test = as.numeric(test))

rmse_rs %>%  gather(data_set, values, -model, -fold) %>%
 mutate(values = as.numeric(values),
       data_set = case_when(data_set == 'test' ~ '2. Testing',  TRUE ~ '1. Training')) %>%
 rename(`Data Set` = data_set) %>%
 ggplot(aes(x = reorder(model, desc(model)), y = values, fill = `Data Set`)) +
 geom_boxplot() + theme_classic() +
 theme(legend.position = 'none',  text = element_text(size = 15),
      axis.title = element_blank()) + coord_flip() + facet_wrap(. ~ `Data Set`)

## inferential statistics on algorithms
## wilcox test between the training and testing dataset
wilcox.test(as.numeric((rmse_rs %>%filter(model == '1. Linear Regression'))$train),
           as.numeric((rmse_rs %>% filter(model == '1. Linear Regression'))$test),
           paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
           conf.level = 0.95)
```

```
wilcox.test(as.numeric((rmse_rs %>% filter(model == '2. SVM Linear Kernel'))$train),
        as.numeric((rmse_rs %>% filter(model == '2. SVM Linear Kernel'))$test),
        paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
        conf.level = 0.95)

wilcox.test(as.numeric((rmse_rs %>%
                filter(model == '3. SVM Polynomial Kernel'))$train),
        as.numeric((rmse_rs %>%filter(model == '3. SVM Polynomial Kernel'))$test),
        paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
        conf.level = 0.95)

wilcox.test(as.numeric((rmse_rs %>%
                filter(model == '4. SVM Radial BF Kernel'))$train),
        as.numeric((rmse_rs %>%filter(model == '4. SVM Radial BF Kernel'))$test),
        paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
        conf.level = 0.95)

wilcox.test(as.numeric((rmse_rs %>%filter(model == '5. Random Forest'))$train),
        as.numeric((rmse_rs %>%filter(model == '5. Random Forest'))$test),
        paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
        conf.level = 0.95)

wilcox.test(as.numeric((rmse_rs %>%filter(model == '6. XGBoost'))$train),
        as.numeric((rmse_rs %>%filter(model == '6. XGBoost'))$test),
        paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
        conf.level = 0.95)

## friedmen test of the best model
## training data set
rmse_rs %>% friedman_test(train ~ model|fold)
rmse_rs %>% friedman_effsize(train ~ model|fold)
rmse_rs %>% wilcox_test(train ~ model, paired = TRUE, p.adjust.method = 'bonferroni')

## testing data set
rmse_rs %>% friedman_test(test ~ model|fold)
rmse_rs %>% friedman_effsize(test ~ model|fold)
rmse_rs %>% wilcox_test(test ~ model, paired = TRUE, p.adjust.method = 'bonferroni')

## median rmse values
rmse_rs %>% select(-fold) %>% group_by(model) %>%
 summarise(train = median(train), test = median(test),.groups = 'drop')

############################
## DEPENDENT VARAIBLE ##
## ONE-YEAR RETENTION ##
############################
```

```
## data clean up recipe
retain_rec <- recipe(dv_next_fall ~ .,  data = dat_train) %>%
 update_role(unique_identifer,  new_role = 'id variable') %>%
 step_mutate_at(c(adv_standing_ap_hrs:adv_standing_other_hrs,
               ga_hope, pell:oth_loans),  fn = ~ replace_na(., 0)) %>%
 step_mutate_at(zell_ind,  fn = ~ replace_na(., 'N')) %>%
 step_novel(c(cpc_english_code:cpc_social_science_code)) %>%
 step_unknown(c(cpc_english_code:cpc_social_science_code),  new_level = 'U') %>%
 step_mutate(gender_descr = case_when(gender_descr == 'Male' ~ 1, TRUE ~ 0),
     admit_first_gen_ind = case_when(admit_first_gen_ind == 'Y' ~ 1,  TRUE ~ 0),
     college_prep = case_when(cpc_english_code == 'S' ~ 1,
                           cpc_english_code == 'X' ~ 1,  TRUE ~ 0) +
     case_when(cpc_foreign_language_code == 'S' ~ 1,
                    cpc_foreign_language_code == 'X' ~ 1, TRUE ~ 0) +
     case_when(cpc_math_code == 'S' ~ 1, cpc_math_code == 'X' ~ 1, TRUE ~ 0)  +
     case_when(cpc_science_code == 'S' ~ 1, cpc_science_code == 'X' ~ 1, TRUE ~ 0) +
     case_when(cpc_social_science_code == 'S' ~ 1,
                    cpc_social_science_code == 'X' ~ 1, TRUE ~ 0),
     acay_inst_sup_exp = (acay_sup_exp + inst_sup_exp),
     public_rsch_exp = (public_serv_exp + rsch_exp),
     cm_ready = (content_mastery + readiness) / 2,  english_cm = english - cm_ready,
     math_cm = math - cm_ready, science_cm = science - cm_ready,
     social_studies_cm = social_studies - cm_ready,
     cip_categories = case_when(cip_categories == 'Social Sciences' ~ 1,
            cip_categories == 'Fine Arts' ~ 2, cip_categories == 'Human Services' ~ 3,
            cip_categories == 'Business' ~ 4, cip_categories == 'STEM' ~ 5,
            cip_categories == 'General/Interdisciplinary Studies' ~ 6,
            cip_categories == 'Healthcare' ~ 7, cip_categories == 'Education' ~ 8,
            TRUE ~ 9),
     zell_ind = case_when(zell_ind == 'Y' ~ 1, TRUE ~ 0),
     locale_group = case_when(locale_group == 'City' ~ 1, locale_group == 'Suburb' ~ 2,
                           locale_group == 'Town' ~ 3, TRUE ~ 4),
     race_eth = case_when(race_eth == 'White' ~ 1,
                        race_eth == 'Black or African American' ~ 2,
                        race_eth == 'Hispanic or Latino' ~ 3, TRUE ~ 4),
     adv_standing_ib_hrs = adv_standing_ib_hrs,
     adv_standing_clep_hrs = adv_standing_clep_hrs,
     adv_standing_other_hrs = adv_standing_other_hrs) %>%
 step_rm(cpc_english_code, cpc_foreign_language_code,
     cpc_math_code, cpc_science_code,  cpc_social_science_code, acay_sup_exp,
     inst_sup_exp, public_serv_exp,  rsch_exp, english, math, science,
     social_studies, dv_first_yr_gpa,  content_mastery, readiness,
     dv_first_fall_gpa, hs_code,  hs_grad_year, state_school_id, locale_code, locale) %>%
 step_impute_knn(c(hs_gpa, adm_test_score, expected_family_contribution),
     neighbors = 10) %>%
 step_rename(hsgpa_knn = hs_gpa,  ats_knn = adm_test_score,
     efc_knn = expected_family_contribution) %>%
 step_YeoJohnson(hsgpa_knn, ats_knn, college_prep, adv_standing_ap_hrs,
     adv_standing_clep_hrs, adv_standing_ib_hrs, adv_standing_other_hrs, cm_ready,
     english_cm, math_cm, science_cm, social_studies_cm, efc_knn, ga_hope, pell,
     fed_sub_loans, fed_unsub_loans, oth_loans, acay_inst_sup_exp,  all_other_exp,
     instr_exp, stu_serv_exp, public_rsch_exp) %>%
 step_normalize(hsgpa_knn,  ats_knn, college_prep, adv_standing_ap_hrs,
     adv_standing_clep_hrs, adv_standing_ib_hrs, adv_standing_other_hrs, cm_ready,
     college_prep, english_cm, math_cm, science_cm, social_studies_cm, efc_knn,
     ga_hope, pell, fed_sub_loans, fed_unsub_loans, oth_loans, acay_inst_sup_exp,
     all_other_exp, instr_exp, stu_serv_exp, public_rsch_exp)
#######################
## CLEAN DATA SETS ##
#######################

## no modifications
retain <- retain_rec %>% step_rm(unique_identifer) %>%
 step_mutate(dv_next_fall = as.factor(dv_next_fall)) %>% prep() %>% juice()
retain_rec_rv <- recipe(dv_next_fall ~ ., data = retain)
## downsample technique
retain_ds <- retain_rec %>% step_rm(unique_identifer) %>%
 step_mutate(dv_next_fall = as.factor(dv_next_fall)) %>%
 step_downsample(dv_next_fall) %>% prep() %>% juice()
retain_rec_ds <- recipe(dv_next_fall ~ .,  data = retain_ds)
## upsample technique
retain_us <- retain_rec %>% step_rm(unique_identifer) %>%
 step_mutate(dv_next_fall = as.factor(dv_next_fall)) %>%
 step_upsample(dv_next_fall) %>% prep() %>% juice()
retain_rec_us <- recipe(dv_next_fall ~ .,  data = retain_us)
## testing data set
retain_test <- retain_rec %>% step_rm(unique_identifer) %>%
 step_mutate(dv_next_fall = as.factor(dv_next_fall)) %>% prep() %>% bake(dat_test)
retain_test_rec_rv <- recipe(dv_next_fall ~ ., data = retain_test)

##########################
## CROSS-VALIDATIONS ##
##########################

## splitting the no modifications
set.seed(51823)
retain_cv <- vfold_cv(retain, v = 10)
## splitting the downsample modifications
set.seed(51823)
retain_cv_ds <- vfold_cv(retain_ds, v = 10)
## splitting the upsample modifications
```

```
set.seed(51823)
retain_cv_us <- vfold_cv(retain_us, v = 10)
## splitting the testing data set
set.seed(7323)
retain_test_cv <- vfold_cv(retain_test, v = 10)

########################
## TUNING CONTROLS ##
########################

## control grid set up
ctrl_grid <- control_grid(save_pred = TRUE, save_workflow = TRUE)
## model_metrics
model_metrics <- metric_set(roc_auc, accuracy, spec, f_meas, sens)

###########################
##LOGISTIC REGRESSION ##
###########################

## no modifications--training data set
retain_lr <- retain %>% glm(formula = dv_next_fall ~ .,  family = 'binomial')
retain_lr %>% summary()
## odds ratios
exp(coef(retain_ lr)); exp(confint(retain_ lr))
## model stats
retain_lr %>% blr_model_fit_stats()
retain_lr %>% blr_test_hosmer_lemeshow()

## logistic regression assumptions
## linear relationship with log odds
retain_probs <- predict(retain_lr, type = "response")
retain_cor <- retain %>% cbind(retain_probs) %>%
 mutate(logit = log(retain_probs / (1 - retain_probs))) %>%
 select(-dv_next_fall, -retain_probs)
logit_linear <- as.data.frame((retain_cor %>%
                        corr.test(use = 'pairwise',method = 'pearson',
                                adjust = 'holm', alpha = .05))$r) %>% select(logit) %>%
 cbind(as.data.frame((retain_cor %>%
                 corr.test(use = 'pairwise', method = 'pearson',  adjust = 'holm',
                         alpha = .05))$p) %>%
      rename(logit_p = logit) %>% select(logit_p)) %>% mutate(logit = round(logit, 5))
logit_linear$Variable <- logit_linear %>% row.names()
logit_linear %>%
 mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                      Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                      Variable == 'hsgpa_knn' ~ 'HS GPA',
                      Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                      Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                      Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                      Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                      Variable == 'cip_categories' ~ 'Major Groupings',
                      Variable == 'efc_knn' ~ 'EFC',
                      Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                      Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                      Variable == 'pell' ~ 'PELL Grant',
                      Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                      Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                      Variable == 'oth_loans' ~ 'Other Loans',
                      Variable == 'ats_knn' ~ 'Admissions Test Scores',
                      Variable == 'all_other_exp' ~ 'All Other',
                      Variable == 'instr_exp' ~ 'Instruction',
                      Variable == 'stu_serv_exp' ~ 'Student Services',
                      Variable == 'race_eth' ~ 'Race Ethnicity',
                      Variable == 'cm_ready' ~ 'CM & Ready Mean',
                      Variable == 'locale_group' ~ 'HS Locale',
                      Variable == 'college_prep' ~ 'College Prep. Curric.',
                      Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                      Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                      Variable == 'english_cm' ~ 'English (CMR)',
                      Variable == 'math_cm' ~ 'Math (CMR)',
                      Variable == 'science_cm' ~ 'Science (CMR)',
                      Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                      TRUE ~ 'Logit of Probs')) %>% arrange(abs(logit)) %>% View()
filter(Variable != 'Logit of Probs') %>%
 mutate(Direction = case_when(logit < 0 ~ 'Negative', TRUE ~ 'Positive'),
      prob = case_when(logit_p < .001 ~ '***', logit_p < .01 ~ '**',
                    logit_p < .05 ~ '*', TRUE ~ '')) %>%  ggplot() +
 geom_bar(aes(x = reorder(paste0(Variable, prob), abs(logit)),
            y = logit,  fill = Direction), stat = 'identity') +
 ylim(-1, 1) + ylab('Pearson R Correlational Value') + theme_classic() +
 theme(legend.position = 'top',  axis.title.y = element_blank(),
      text = element_text(size = 15)) + coord_flip()
## multicollinearity
cbind(retain_lr %>% vif() )

## downsample modification
retain_lr_ds <- retain_ds %>% glm(formula = dv_next_fall ~ ., family = 'binomial')
retain_lr_ds %>% summary()
retain_lr_ds %>% summary() %>% coef()
## odds ratio
exp(coef(retain_lr_ds)); exp(confint(retain_lr_ds))
```

```
## model stats
retain_lr_ds %>% blr_model_fit_stats()
retain_lr_ds %>% blr_test_hosmer_lemeshow()

## upsample modification
retain_lr_us <- retain_us %>% glm(formula = dv_next_fall ~ .,  family = 'binomial')
retain_lr_us %>% summary()
## odds ratio
exp(coef(retain_lr_us)); exp(confint(retain_lr_us))
## model stats
retain_lr_us %>% blr_model_fit_stats()
retain_lr_us %>% blr_test_hosmer_lemeshow()

## testing data set
retain_test_lr <- retain_test %>% glm(formula = dv_next_fall ~ ., family = 'binomial')
retain_test_lr %>% summary()
## odds ratio
exp(coef(retain_test_lr)); exp(confint(retain_test_lr))
## model stats
retain_test_lr %>% blr_model_fit_stats()
retain_test_lr %>% blr_test_hosmer_lemeshow()

## variable importance analysis
## Logistic Regression Specs
log_reg_spec <- logistic_reg() %>% set_engine('glm') %>% set_mode('classification')
## workflows
## no modifications
log_reg_wf <- workflow() %>% add_model(log_reg_spec) %>%
 add_recipe(retain_rec_rv)
## downsample modifications
log_reg_ds_wf <- workflow() %>% add_model(log_reg_spec) %>%
 add_recipe(retain_rec_ds)
## upsample modifications
log_reg_us_wf <- workflow() %>% add_model(log_reg_spec) %>%
 add_recipe(retain_rec_us)

## joining the three sample VI information together
# no modifications
log_reg_vi <- fit(log_reg_wf, retain) %>% extract_fit_parsnip() %>%  vi() %>%
 left_join(fit(log_reg_wf,  retain) %>% extract_fit_parsnip() %>%
           vi(scale = TRUE) %>%rename(rescale_importance = Importance) %>%
           select(-Sign)) %>% mutate(log_reg_type = '1. None') %>%
 rbind(fit(log_reg_wf, retain_test) %>% extract_fit_parsnip() %>%  vi() %>%
       left_join(fit(log_reg_wf,  retain_test) %>% extract_fit_parsnip() %>%
              vi(scale = TRUE) %>% rename(rescale_importance = Importance) %>%
              select(-Sign)) %>% mutate(log_reg_type = '4. Testing')) %>%
 rbind(fit(log_reg_ds_wf,  retain_ds) %>% extract_fit_parsnip()  %>%  vi() %>%
      left_join(fit(log_reg_ds_wf, retain_ds) %>% extract_fit_parsnip() %>%
              vi(scale = TRUE) %>% rename(rescale_importance = Importance) %>%
              select(-Sign)) %>% mutate(log_reg_type = '2. Downsample')) %>%
 rbind(fit(log_reg_us_wf,  retain_us) %>% extract_fit_parsnip()  %>% vi() %>%
      left_join(fit(log_reg_us_wf, retain_us) %>% extract_fit_parsnip() %>%
              vi(scale = TRUE) %>% rename(rescale_importance = Importance) %>%
              select(-Sign)) %>% mutate(log_reg_type = '3. Upsample'))

log_reg_vi %>% rename(Impact = Sign) %>%
 mutate(Impact = case_when(Impact == 'NEG' ~ 'Negative', TRUE ~ 'Positive'),
      Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                     Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                     Variable == 'hsgpa_knn' ~ 'HS GPA',
                     Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                     Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                     Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                     Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                     Variable == 'cip_categories' ~ 'Major Groupings',
                     Variable == 'efc_knn' ~ 'EFC',
                     Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                     Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                     Variable == 'pell' ~ 'PELL Grant',
                     Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                     Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                     Variable == 'oth_loans' ~ 'Other Loans',
                     Variable == 'ats_knn' ~ 'Admissions Test Scores',
                     Variable == 'all_other_exp' ~ 'All Other',
                     Variable == 'instr_exp' ~ 'Instruction',
                     Variable == 'stu_serv_exp' ~ 'Student Services',
                     Variable == 'race_eth' ~ 'Race Ethnicity',
                     Variable == 'cm_ready' ~ 'CM & Ready Mean',
                     Variable == 'locale_group' ~ 'HS Locale',
                     Variable == 'college_prep' ~ 'College Prep. Curric.',
                     Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                     Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                     Variable == 'english_cm' ~ 'English (CMR)',
                     Variable == 'math_cm' ~ 'Math (CMR)',
                     Variable == 'science_cm' ~ 'Science (CMR)',
                     Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                     TRUE ~ 'CHECK')) %>%
 ggplot() + geom_bar(aes(x = reorder(Variable, rescale_importance),
          y = rescale_importance,  fill = Impact),  stat = 'identity') + theme_classic() +
 theme(legend.position = 'top', axis.title.y = element_blank(),
     text = element_text(size = 15)) + ylab('Rescale Importance') + coord_flip() +
 facet_wrap(. ~ log_reg_type,  ncol = 4)
```

```
## predictive power
## no modifications--training data set
log_reg_cv_train <- log_reg_wf %>% fit_resamples(dv_next_fall ~.,
            resamples = retain_cv,  metrics = model_metrics, control = ctrl_grid)
log_reg_cv_train %>% collect_predictions() %>% conf_mat(dv_next_fall,  .pred_class)
log_reg_cv_train %>% collect_metrics()

## testing data set
log_reg_cv_test <- log_reg_wf %>% update_recipe(retain_test_rec_rv) %>%
 fit_resamples(as.factor(dv_next_fall) ~.,  resamples = retain_test_cv,
            metrics = model_metrics, control = ctrl_grid)
log_reg_cv_test %>% collect_predictions() %>% conf_mat(dv_next_fall,  .pred_class)
log_reg_cv_test %>% collect_metrics()

## downsample modification--training data set
log_reg_cv_train_ds <- log_reg_ds_wf %>% fit_resamples(dv_next_fall ~.,
            resamples = retain_cv_ds,  metrics = model_metrics, control = ctrl_grid)
log_reg_cv_train_ds %>% collect_predictions() %>% conf_mat(dv_next_fall,
        .pred_class)
log_reg_cv_train_ds %>% collect_metrics()

## testing data set
log_reg_cv_test_ds <- log_reg_ds_wf %>% update_recipe(retain_test_rec_rv) %>%
 fit_resamples(as.factor(dv_next_fall) ~.,  resamples = retain_test_cv,
            metrics = model_metrics, control = ctrl_grid)
log_reg_cv_test_ds %>% collect_predictions() %>% conf_mat(dv_next_fall,
        .pred_class)
log_reg_cv_test_ds %>%  collect_metrics()

## upsample modification--training data set
log_reg_cv_train_us <- log_reg_us_wf %>% fit_resamples(dv_next_fall ~.,
            resamples = retain_cv_us,  metrics = model_metrics, control = ctrl_grid)
log_reg_cv_train_us %>% collect_predictions() %>% conf_mat(dv_next_fall,
        .pred_class)
log_reg_cv_train_us %>%collect_metrics()

## testing data set
log_reg_cv_test_us <- log_reg_us_wf %>% update_recipe(retain_test_rec_rv) %>%
 fit_resamples(as.factor(dv_next_fall) ~.,  resamples = retain_test_cv,
            metrics = model_metrics, control = ctrl_grid)
log_reg_cv_test_us %>% collect_predictions() %>% conf_mat(dv_next_fall,
        .pred_class)
log_reg_cv_test_us %>% collect_metrics()
```

```
## roc values
log_roc <- log_reg_cv_train %>% collect_predictions() %>%
 roc_curve(truth = dv_next_fall, `.pred_0`) %>%
 mutate(model = '1. None',  type = 'Training') %>%
 rbind(log_reg_cv_test %>% collect_predictions() %>%
        roc_curve(truth = dv_next_fall,  `.pred_0`) %>%
        mutate(model = '1. None',  type = 'Testing')) %>%
 rbind(log_reg_cv_train_ds %>% collect_predictions() %>%
        roc_curve(truth = dv_next_fall,  `.pred_0`) %>%
        mutate(model = '2. Downsample', type = 'Training')) %>%
 rbind(log_reg_cv_test_ds %>% collect_predictions() %>%
        roc_curve(truth = dv_next_fall,  `.pred_0`) %>%
        mutate(model = '2. Downsample', type = 'Testing')) %>%
 rbind(log_reg_cv_train_us %>% collect_predictions() %>%
        roc_curve(truth = dv_next_fall, `.pred_0`) %>%
        mutate(model = '3. Upsample', type = 'Training')) %>%
 rbind(log_reg_cv_test_us %>%collect_predictions() %>%
        roc_curve(truth = dv_next_fall, `.pred_0`) %>%
        mutate(model = '3. Upsample', type = 'Testing')) %>%
 rename(`Data Set` = type) %>%  mutate(Algorithm = '1. Logistic Regression')

##########################################
## Support Vector Machine-Linear Kernel ##
##########################################

## model specifications
svm_l_spec <- svm_linear(cost = tune(),margin = tune()) %>%
 set_mode('classification') %>%set_engine('kernlab')
## model workflow
## no modifications
svm_l_wf <- workflow() %>% add_model(svm_l_spec) %>%
 add_recipe(retain_rec_rv)
## downsample modifications
svm_l_wf_ds <- workflow() %>% add_model(svm_l_spec) %>%
 add_recipe(retain_rec_ds)
## upsample modifications
svm_l_wf_us <- workflow() %>% add_model(svm_l_spec) %>%
 add_recipe(retain_rec_us)

## tuning
## no modifications
set.seed(52323)
doParallel::registerDoParallel()
svm_l_tune <- tune_grid(svm_l_wf, resamples = retain_cv,
                    metrics = model_metrics, control = ctrl_grid, grid = 20)
## best model
```

```
svml_tune_best <- select_best(svm_l_tune, 'roc_auc')
## model fixed to the best outcome model
svml_tune_final <- finalize_model(svm_l_spec, svml_tune_best)
## downsample modifications
set.seed(52323)
doParallel::registerDoParallel()
svm_l_tune_ds <- tune_grid(svm_l_wf_ds, resamples = retain_cv_ds,
                           metrics = model_metrics, control = ctrl_grid, grid = 20)
## best model
svml_tune_best_ds <- select_best(svm_l_tune_ds, 'roc_auc')
## model fixed to the best outcome model
svml_tune_final_ds <- finalize_model(svm_l_spec, svml_tune_best_ds)
set.seed(52323)
doParallel::registerDoParallel()
svm_l_tune_us <- tune_grid(svm_l_wf_us,  resamples = retain_cv_us,
                           metrics = model_metrics,  control = ctrl_grid, grid = 20)
## best model
svml_tune_best_us <- select_best(svm_l_tune_us, 'roc_auc')
## model fixed to the best outcome model
svml_tune_final_us <- finalize_model(svm_l_spec, svml_tune_best_us)
```

## variable importance analysis

```
## no modifications
set.seed(51923)
svml_fit <- workflow() %>% add_model(svml_tune_final)  %>%
 add_recipe(retain_rec_rv) %>% fit(retain)
## training data set
set.seed(51923)
svml_vi <- svml_fit %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  pred_wrapper = kernlab::predict, reference_class = '1',
    metric = 'auc', target = 'dv_next_fall', train = retain)
set.seed(51923)
svml_vi_rs <- svml_fit %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE, pred_wrapper = kernlab::predict,
    reference_class = '1', metric = 'auc', target = 'dv_next_fall', train = retain)
## testing data set
set.seed(51923)
svml_fit_test <- workflow() %>% add_model(svml_tune_final)  %>%
 add_recipe(retain_rec_rv) %>% fit(retain_test)
set.seed(51923)
svml_vi_test <- svml_fit_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  pred_wrapper = kernlab::predict, reference_class = '1',
    metric = 'auc', target = 'dv_next_fall', train = retain_test)
set.seed(51923)
svml_vi_rs_test <- svml_fit_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE, pred_wrapper = kernlab::predict,
    reference_class = '1', metric = 'auc', target = 'dv_next_fall', train = retain_test)

## downsample modification--training data set
set.seed(51923)
svml_fit_ds <- workflow() %>% add_model(svml_tune_final_ds)  %>%
 add_recipe(retain_rec_ds) %>%
 fit(retain_ds)
set.seed(51923)
svml_vi_ds <- svml_fit_ds %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  pred_wrapper = kernlab::predict, reference_class = '1',
    metric = 'auc', target = 'dv_next_fall', train = retain_ds)
set.seed(51923)
svml_vi_ds_rs <- svml_fit_ds %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE, pred_wrapper = kernlab::predict,
    reference_class = '1', metric = 'auc', target = 'dv_next_fall', train = retain_ds) %>%
 rename(rescaled_importance = Importance)
## testing data set
set.seed(51923)
svml_fit_ds_test <- workflow() %>% add_model(svml_tune_final_ds)  %>%
 add_recipe(retain_rec_ds) %>% fit(retain_test
set.seed(51923)
svml_vi_ds_test <- svml_fit_ds_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  pred_wrapper = kernlab::predict, reference_class = '1',
    metric = 'auc', target = 'dv_next_fall', train = retain_test)
set.seed(51923)
svml_vi_ds_rs_test <- svml_fit_ds_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE, pred_wrapper = kernlab::predict,
    reference_class = '1', metric = 'auc', target = 'dv_next_fall', train = retain_test) %>%
 rename(rescaled_importance = Importance)

## upsample modification--training data set
set.seed(51923)
svml_fit_us <- workflow() %>% add_model(svml_tune_final_us)  %>%
 add_recipe(retain_rec_us) %>% fit(retain_us)
set.seed(51923)
svml_vi_us <- svml_fit_us %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  pred_wrapper = kernlab::predict, reference_class = '1',
    metric = 'auc', target = 'dv_next_fall', train = retain_us)
set.seed(51923)
svml_vi_us_rs <- svml_fit_us %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE, pred_wrapper = kernlab::predict,
    reference_class = '1', metric = 'auc', target = 'dv_next_fall', train = retain_us) %>%
 rename(rescaled_importance = Importance)
## testing data set
set.seed(51923)
svml_fit_us_test <- workflow() %>% add_model(svml_tune_final_us)  %>%
 add_recipe(retain_rec_us) %>% fit(retain_test)
set.seed(51923)
svml_vi_us_test <- svml_fit_us_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  pred_wrapper = kernlab::predict, reference_class = '1',
    metric = 'auc', target = 'dv_next_fall', train = retain_test)
set.seed(51923)
svml_vi_us_rs_test <- svml_fit_us_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE, pred_wrapper = kernlab::predict,
    reference_class = '1', metric = 'auc', target = 'dv_next_fall', train = retain_test) %>%
 rename(rescaled_importance = Importance)

## joining the results together
svml_vi_compar <- svml_vi %>% select(-Importance) %>%
 mutate(rescaled_importance = 0,  type = '1. None') %>%
 rbind(svml_vi_ds_rs %>% mutate(type = '2. Downsample')) %>%
 rbind(svml_vi_us_rs %>% mutate(type = '3. Upsample'))

svml_vi_compar %>%
 mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                             Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                             Variable == 'hsgpa_knn' ~ 'HS GPA',
                             Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                             Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                             Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                             Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                             Variable == 'cip_categories' ~ 'Major Groupings',
                             Variable == 'efc_knn' ~ 'EFC',
                             Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                             Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                             Variable == 'pell' ~ 'PELL Grant',
                             Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                             Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                             Variable == 'oth_loans' ~ 'Other Loans',
                             Variable == 'ats_knn' ~ 'Admissions Test Scores',
                             Variable == 'all_other_exp' ~ 'All Other',
                             Variable == 'instr_exp' ~ 'Instruction',
                             Variable == 'stu_serv_exp' ~ 'Student Services',
                             Variable == 'race_eth' ~ 'Race Ethnicity',
                             Variable == 'cm_ready' ~ 'CM & Ready Mean',
                             Variable == 'locale_group' ~ 'HS Locale',
                             Variable == 'college_prep' ~ 'College Prep. Curric.',
                             Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                             Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                             Variable == 'english_cm' ~ 'English (CMR)',
                             Variable == 'math_cm' ~ 'Math (CMR)',
                             Variable == 'science_cm' ~ 'Science (CMR)',
                        Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                        TRUE ~ 'CHECK')) %>%
 ggplot() + geom_bar(aes(x = reorder(Variable, rescaled_importance),
            y = rescaled_importance,  fill = desc(rescaled_importance/100)),
         stat = 'identity') + theme_classic() +
 theme(legend.position = 'none',  axis.title.y = element_blank(),
      text = element_text(size = 15)) + ylab('Rescale Importance') +
 coord_flip() + facet_wrap(. ~ type)
```

## predictive power

```
## no modifications--training data set
set.seed(7323)
svml_cv_train <- svml_tune_final %>% fit_resamples(dv_next_fall ~.,
             resamples = retain_cv, metrics = model_metrics, control = ctrl_grid)
svml_cv_train %>% collect_predictions() %>% conf_mat(dv_next_fall,  .pred_class)
svml_cv_train %>% collect_metrics()
## testing data set
set.seed(7323)
svml_cv_test <- svm_l_wf %>% update_recipe(retain_test_rec_rv) %>%
 update_model(svml_tune_final) %>% fit_resamples(as.factor(dv_next_fall) ~.,
             resamples = retain_test_cv,  metrics = model_metrics, control = ctrl_grid)
svml_cv_test %>% collect_predictions() %>% conf_mat(dv_next_fall,  .pred_class)
svml_cv_test %>% collect_metrics()

## downsample modification--training data set
set.seed(7323)
svml_cv_train_ds <- svml_tune_final_ds %>%fit_resamples(dv_next_fall ~.,
             resamples = retain_cv_ds, metrics = model_metrics, control = ctrl_grid)
svml_cv_train_ds %>% collect_predictions() %>% conf_mat(dv_next_fall,  .pred_class)
svml_cv_train_ds %>% collect_metrics()
## testing data set
set.seed(7323)
svml_cv_test_ds <- svm_l_wf_ds %>% update_recipe(retain_test_rec_rv) %>%
 update_model(svml_tune_final_ds) %>% fit_resamples(as.factor(dv_next_fall) ~.,
             resamples = retain_test_cv,  metrics = model_metrics, control = ctrl_grid)
svml_cv_test_ds %>% collect_predictions() %>% conf_mat(dv_next_fall,  .pred_class)
svml_cv_test_ds %>%collect_metrics()

## upsample modification--training data set
set.seed(7323)
svml_cv_train_us <- svml_tune_final_us %>%fit_resamples(dv_next_fall ~.,
             resamples = retain_cv_us, metrics = model_metrics, control = ctrl_grid)
svml_cv_train_us %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
svml_cv_train_us %>% collect_metrics()
## testing data set
set.seed(7323)
```

```
svml_cv_test_us <- svm_l_wf_us %>% update_recipe(retain_test_rec_rv) %>%
 update_model(svml_tune_final_us) %>% fit_resamples(as.factor(dv_next_fall) ~.,
             resamples = retain_test_cv,  metrics = model_metrics, control = ctrl_grid)
svml_cv_test_us %>% collect_predictions() %>% conf_mat(dv_next_fall,  .pred_class)
svml_cv_test_us %>%collect_metrics()

## roc values
svml_roc <- svml_cv_train %>% collect_predictions() %>%
 roc_curve(truth = dv_next_fall,  .pred_0) %>%
 mutate(model = '1. None',  type = 'Training') %>%
 rbind(svml_cv_train %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall,  .pred_0) %>%
       mutate(model = '1. None',  type = 'Testing')) %>%
 rbind(svml_cv_train_ds %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall,  .pred_0) %>%
       mutate(model = '2. Downsample',  type = 'Training')) %>%
 rbind(svml_cv_test_ds %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall,  .pred_0) %>%
       mutate(model = '2. Downsample', type = 'Testing')) %>%
 rbind(svml_cv_train_us %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall, pred_0) %>%
       mutate(model = '3. Upsample', type = 'Training')) %>%
 rbind(svml_cv_test_us %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall, .pred_0) %>%
       mutate(model = '3. Upsample', type = 'Testing')) %>%
 rename(`Data Set` = type) %>% mutate(Algorithm = '2. SVM Linear')

###########################################
##Support Vector Machine-Polynomial Kernel ##
###########################################

## model specifications
svm_p_spec <- svm_poly(cost = tune(), degree = tune(), scale_factor = tune(),
                  margin = tune()) %>% set_mode('classification') %>%
 set_engine('kernlab')
## model workflow
## no modifications
svm_p_wf <- workflow() %>% add_model(svm_p_spec) %>%
 add_recipe(retain_rec_rv)
## downsample modifications
svm_p_wf_ds <- workflow() %>% add_model(svm_p_spec) %>%
 add_recipe(retain_rec_ds)
## upsample modifications
svm_p_wf_us <- workflow() %>% add_model(svm_p_spec) %>%
 add_recipe(retain_rec_us)
```

```
## tuning
## no modifications
set.seed(52323)
doParallel::registerDoParallel()
svm_p_tune <- tune_grid(svm_p_wf, resamples = retain_cv,
                        metrics = model_metrics, control = ctrl_grid, grid = 20)
## best model
svmp_tune_best <- select_best(svm_p_tune, 'roc_auc')
## model fixed to the best outcome model
svmp_tune_final <- finalize_model(svm_p_spec, svmp_tune_best)
## downsample modifications
set.seed(52323)
doParallel::registerDoParallel()
svm_p_tune_ds <- tune_grid(svm_p_wf_ds, resamples = retain_cv_ds,
                           metrics = model_metrics, control = ctrl_grid, grid = 20)
## best model
svmp_tune_best_ds <- select_best(svm_p_tune_ds, 'roc_auc')
## model fixed to the best outcome model
svmp_tune_final_ds <- finalize_model(svm_p_spec, svmp_tune_best_ds)
## upsample modifications
set.seed(52323)
doParallel::registerDoParallel()
svm_p_tune_us <- tune_grid(svm_p_wf_us, resamples = retain_cv_us,
                           metrics = model_metrics, control = ctrl_grid, grid = 20)
## best model
svmp_tune_best_us <- select_best(svm_p_tune_us, 'roc_auc')
## model fixed to the best outcome model
svmp_tune_final_us <- finalize_model(svm_p_spec, svmp_tune_best_us)

## variable importance analysis
## no modifications
set.seed(51923)
svmp_fit <- workflow() %>% add_model(svmp_tune_final)  %>%
 add_recipe(retain_rec_rv) %>% fit(retain)
## training data set
set.seed(51923)
svmp_vi <- svmp_fit %>%extract_fit_parsnip() %>%
 vi(method = 'permute',  pred_wrapper = kernlab::predict, reference_class = '1',
   metric = 'auc', target = 'dv_next_fall', train = retain)
set.seed(51923)
svmp_vi_rs <- svmp_fit %>%extract_fit_parsnip() %>%
 vi(method = 'permute', scale = TRUE, pred_wrapper = kernlab::predict,
   reference_class = '1', metric = 'auc', target = 'dv_next_fall', train = retain) %>%
 rename(rescaled_importance = Importance)
## testing data set
set.seed(51923)
```

```
svmp_fit_test <- workflow() %>% add_model(svmp_tune_final)  %>%
 add_recipe(retain_rec_rv) %>% fit(retain_test)
set.seed(51923)
svmp_vi_test <- svmp_fit_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  pred_wrapper = kernlab::predict, reference_class = '1',
    metric = 'auc', target = 'dv_next_fall', train = retain_test)
set.seed(51923)
svmp_vi_rs_test <- svmp_fit_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE, pred_wrapper = kernlab::predict,
    reference_class = '1', metric = 'auc', target = 'dv_next_fall', train = retain_test) %>%
 rename(rescaled_importance = Importance)

## downsample modifications
set.seed(51923)
svmp_fit_ds <- workflow() %>% add_model(svmp_tune_final_ds)  %>%
 add_recipe(retain_rec_ds) %>% fit(retain_ds)
## training data set
set.seed(51923)
svmp_vi_ds <- svmp_fit_ds %>% extract_fit_parsnip() %>%
 vi(method = 'permute', pred_wrapper = kernlab::predict, reference_class = '1',
    metric = 'auc', target = 'dv_next_fall', train = retain_ds)
set.seed(51923)
svmp_vi_ds_rs <- svmp_fit_ds %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE, pred_wrapper = kernlab::predict,
    reference_class = '1', metric = 'auc', target = 'dv_next_fall',
    train = retain_ds) %>% rename(rescaled_importance = Importance)
## testing data set
set.seed(51923)
svmp_fit_ds_test <- workflow() %>% add_model(svmp_tune_final_ds)  %>%
 add_recipe(retain_rec_ds) %>% fit(retain_test)
set.seed(51923)
svmp_vi_ds_test <- svmp_fit_ds_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  pred_wrapper = kernlab::predict, reference_class = '1',
    metric = 'auc', target = 'dv_next_fall', train = retain_test)
set.seed(51923)
svmp_vi_ds_rs_test <- svmp_fit_ds_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE, pred_wrapper = kernlab::predict,
    reference_class = '1', metric = 'auc', target = 'dv_next_fall', train = retain_test) %>%
 rename(rescaled_importance = Importance)

## upsample modification--training data set
set.seed(51923)
svmp_fit_us <- workflow() %>% add_model(svmp_tune_final_us)  %>%
 add_recipe(retain_rec_us) %>% fit(retain_us)
set.seed(51923)
svmp_vi_us <- svmp_fit_us %>% extract_fit_parsnip() %>%
 vi(method = 'permute', pred_wrapper = kernlab::predict, reference_class = '1',
   metric = 'auc', target = 'dv_next_fall', train = retain_us)
set.seed(51923)
svmp_vi_us_rs <- svmp_fit_us %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE, pred_wrapper = kernlab::predict,
   reference_class = '1', metric = 'auc', target = 'dv_next_fall', train = retain_us) %>%
 rename(rescaled_importance = Importance)
## testing data set
set.seed(51923)
svmp_fit_us_test <- workflow() %>% add_model(svmp_tune_final_us)  %>%
 add_recipe(retain_rec_us) %>% fit(retain_test)
set.seed(51923)
svmp_vi_us_test <- svmp_fit_us_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  pred_wrapper = kernlab::predict, reference_class = '1',
   metric = 'auc', target = 'dv_next_fall', train = retain_test)
set.seed(51923)
svmp_vi_us_rs_test <- svmp_fit_us_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE, pred_wrapper = kernlab::predict,
   reference_class = '1', metric = 'auc', target = 'dv_next_fall', train = retain_test) %>%
 rename(rescaled_importance = Importance)

## joining the results together
svmp_vi_compar <- svmp_vi %>% select(-Importance) %>%
 mutate(rescaled_importance = 0, type = '1.1 None - Train') %>%
 rbind(svmp_vi_rs_test %>%
        mutate(rescaled_importance = 0, type = '1.2 None - Test')) %>%
 rbind(svmp_vi_ds_rs %>% mutate(type = '1.1 DS - Train')) %>%
 rbind(svmp_vi_ds_rs_test %>% mutate(type = '1.2 DS - Test')) %>%
 rbind(svmp_vi_us_rs %>% mutate(type = '2.1 US - Train')) %>%
 rbind(svmp_vi_us_rs_test %>% mutate(type = '2.2 US - Test'))

svmp_vi_compar %>% filter(!type %in% c('1.1 None - Train',  '1.2 None - Test') ) %>%
 mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                         Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                         Variable == 'hsgpa_knn' ~ 'HS GPA',
                         Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                         Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                         Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                         Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                         Variable == 'cip_categories' ~ 'Major Groupings',
                         Variable == 'efc_knn' ~ 'EFC',
                         Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                         Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                         Variable == 'pell' ~ 'PELL Grant',
                         Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                         Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                         Variable == 'oth_loans' ~ 'Other Loans',
                         Variable == 'ats_knn' ~ 'Admissions Test Scores',
                         Variable == 'all_other_exp' ~ 'All Other',
                         Variable == 'instr_exp' ~ 'Instruction',
                         Variable == 'stu_serv_exp' ~ 'Student Services',
                         Variable == 'race_eth' ~ 'Race Ethnicity',
                         Variable == 'cm_ready' ~ 'CM & Ready Mean',
                         Variable == 'locale_group' ~ 'HS Locale',
                         Variable == 'college_prep' ~ 'College Prep. Curric.',
                         Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                         Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                         Variable == 'english_cm' ~ 'English (CMR)',
                         Variable == 'math_cm' ~ 'Math (CMR)',
                         Variable == 'science_cm' ~ 'Science (CMR)',
                         Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                         TRUE ~ 'CHECK')) %>%ggplot() +
 geom_bar(aes(x = reorder(Variable, rescaled_importance), y = rescaled_importance,
      fill = desc(rescaled_importance/100)),  stat = 'identity') + theme_classic() +
 theme(legend.position = 'none',  axis.title.y = element_blank(),
   text = element_text(size = 15)) + ylab('Rescale Importance') + coord_flip() +
 facet_wrap(. ~ type,  ncol = 4)

## predictive power
## no modifications--training data set
set.seed(7323)
svmp_cv_train <- svmp_tune_final %>% fit_resamples(dv_next_fall ~.,
      resamples = retain_cv, metrics = model_metrics, control = ctrl_grid)
svmp_cv_train %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
svmp_cv_train %>% collect_metrics()
## testing data set
set.seed(7323)
svmp_cv_test <- svm_p_wf %>% update_recipe(retain_test_rec_rv) %>%
 update_model(svmp_tune_final) %>% fit_resamples(as.factor(dv_next_fall) ~.,
      resamples = retain_test_cv,  metrics = model_metrics, control = ctrl_grid)
svmp_cv_test %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
svmp_cv_test %>% collect_metrics()

## downsample modifications--training data set
set.seed(7323)
svmp_cv_train_ds <- svmp_tune_final_ds %>% fit_resamples(dv_next_fall ~.,
      resamples = retain_cv_ds, metrics = model_metrics, control = ctrl_grid)

svmp_cv_train_ds %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
svmp_cv_train_ds %>% collect_metrics()
## testing data set
set.seed(7323)
```

```
svmp_cv_test_ds <- svm_p_wf_ds %>% update_recipe(retain_test_rec_rv) %>%
 update_model(svmp_tune_final_ds) %>% fit_resamples(as.factor(dv_next_fall) ~.,
            resamples = retain_test_cv,  metrics = model_metrics, control = ctrl_grid)
svmp_cv_test_ds %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
svmp_cv_test_ds %>% collect_metrics()

## upsample modifications--training data set
set.seed(7323)
svmp_cv_train_us <- svmp_tune_final_us %>% fit_resamples(dv_next_fall ~.,
            resamples = retain_cv_us, metrics = model_metrics, control = ctrl_grid)
svmp_cv_train_us %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
svmp_cv_train_us %>% collect_metrics()
## testing data set
set.seed(7323)
svmp_cv_test_us <- svm_p_wf_us %>% update_recipe(retain_test_rec_rv) %>%
 update_model(svmp_tune_final_us) %>% fit_resamples(as.factor(dv_next_fall) ~.,
            resamples = retain_test_cv,  metrics = model_metrics, control = ctrl_grid)
svmp_cv_test_us %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
svmp_cv_test_us %>% collect_metrics()

## roc values
svmp_roc <- svmp_cv_train %>% collect_predictions() %>%
 roc_curve(truth = dv_next_fall, .pred_0) %>%
 mutate(model = '1. None', type = 'Training') %>%
 rbind(svmp_cv_train %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall,  .pred_0) %>%
       mutate(model = '1. None',  type = 'Testing')) %>%
 rbind(svmp_cv_train_ds %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall,  .pred_0) %>%
       mutate(model = '2. Downsample', type = 'Training')) %>%
 rbind(svmp_cv_test_ds %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall, .pred_0) %>%
       mutate(model = '2. Downsample',  type = 'Testing')) %>%
 rbind(svmp_cv_train_us %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall, .pred_0) %>%
       mutate(model = '3. Upsample',  type = 'Training')) %>%
 rbind(svmp_cv_test_us %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall, .pred_0) %>%
       mutate(model = '3. Upsample',  type = 'Testing')) %>%
 rename(`Data Set` = type) %>% mutate(Algorithm = '3. SVM Polynomial')

######################################################
##Support Vector Machine-Radial Basis Function Kernel ##
######################################################

## model specifications
```

```
svm_r_spec <- svm_rbf(cost = tune(),  rbf_sigma = tune(), margin = tune()) %>%
 set_mode('classification') %>% set_engine('kernlab')
## model workflow
## no modifications
svm_r_wf <- workflow() %>% add_model(svm_r_spec) %>%
 add_recipe(retain_rec_rv)
## downsample modifications
svm_r_wf_ds <- workflow() %>%  add_model(svm_r_spec) %>%
 add_recipe(retain_rec_ds)
## upsample modifications
svm_r_wf_us <- workflow() %>% add_model(svm_r_spec) %>%
 add_recipe(retain_rec_us)

## tuning
## no modifications
set.seed(52323)
doParallel::registerDoParallel()
svm_r_tune <- tune_grid(svm_r_wf, resamples = retain_cv, metrics = model_metrics,
              control = ctrl_grid, grid = 20)
## best model
svmr_tune_best <- select_best(svm_r_tune, 'roc_auc')
## model fixed to the best outcome model
svmr_tune_final <- finalize_model(svm_r_spec, svmr_tune_best)
## downsample modifications
set.seed(52323)
doParallel::registerDoParallel()
svm_r_tune_ds <- tune_grid(svm_r_wf_ds, resamples = retain_cv_ds,
                 metrics = model_metrics, control = ctrl_grid, grid = 20)
## best model
svmr_tune_best_ds <- select_best(svm_r_tune_ds, 'roc_auc')
## model fixed to the best outcome model
svmr_tune_final_ds <- finalize_model(svm_r_spec, svmr_tune_best_ds)
## upsample
set.seed(52323)
doParallel::registerDoParallel()
svm_r_tune_us <- tune_grid(svm_r_wf_us, resamples = retain_cv_us,
                 metrics = model_metrics, control = ctrl_grid, grid = 20)
## best model
svmr_tune_best_us <- select_best(svm_r_tune_us, 'roc_auc')
## model fixed to the best outcome model
svmr_tune_final_us <- finalize_model(svm_r_spec, svmr_tune_best_us)
## variable importance analysis
## no modifications--training data set
set.seed(51923)
svmr_fit <- workflow() %>%  add_model(svmr_tune_final)  %>%
 add_recipe(retain_rec_rv) %>%fit(retain)
```

```
set.seed(51923)
svmr_vi <- svmr_fit %>% extract_fit_parsnip() %>%
 vi(method = 'permute', pred_wrapper = kernlab::predict, reference_class = '1',
    metric = 'auc', target = 'dv_next_fall', train = retain) %>%
 mutate(rescaled_importance = 0)
## testing data set
set.seed(51923)
svmr_fit_test <- workflow() %>%add_model(svmr_tune_final)  %>%
 add_recipe(retain_rec_rv) %>%fit(retain_test)
set.seed(51923)
svmr_vi_test <- svmr_fit_test %>%extract_fit_parsnip() %>%
 vi(method = 'permute', pred_wrapper = kernlab::predict, reference_class = '1',
    metric = 'auc', target = 'dv_next_fall', train = retain_test) %>%
 mutate(rescaled_importance = 0)

## downsample modification
set.seed(51923)
svmr_fit_ds <- workflow() %>% add_model(svmr_tune_final_ds)  %>%
 add_recipe(retain_rec_ds) %>% fit(retain_ds)
set.seed(51923)
svmr_vi_ds <- svmr_fit_ds %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  pred_wrapper = kernlab::predict, reference_class = '1',
    metric = 'auc', target = 'dv_next_fall', train = retain_ds)
set.seed(51923)
svmr_vi_ds_rs <- svmr_fit_ds %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE, pred_wrapper = kernlab::predict,
    reference_class = '1', metric = 'auc', target = 'dv_next_fall', train = retain_ds) %>%
 rename(rescaled_importance = Importance)
## testing data set
set.seed(51923)
svmr_fit_ds_test <- workflow() %>% add_model(svmr_tune_final_ds)  %>%
 add_recipe(retain_rec_ds) %>% fit(retain_test)
set.seed(51923)
svmr_vi_ds_test <- svmr_fit_ds_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  pred_wrapper = kernlab::predict, reference_class = '1',
    metric = 'auc', target = 'dv_next_fall', train = retain_test)
set.seed(51923)
svmr_vi_ds_rs_test <- svmr_fit_ds_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  scale = TRUE, pred_wrapper = kernlab::predict,
    reference_class = '1', metric = 'auc', target = 'dv_next_fall', train = retain_test) %>%
 rename(rescaled_importance = Importance)
## upsample modifications
set.seed(51923)
svmr_fit_us <- workflow() %>% add_model(svmr_tune_final_us)  %>%
 add_recipe(retain_rec_us) %>% fit(retain_us)
set.seed(51923)
```

```
svmr_vi_us <- svmr_fit_us %>% extract_fit_parsnip() %>%
 vi(method = 'permute',  pred_wrapper = kernlab::predict, reference_class = '1',
   metric = 'auc', target = 'dv_next_fall', train = retain_us)
set.seed(51923)
svmr_vi_us_rs <- svmr_fit_us %>%extract_fit_parsnip() %>%
 vi(method = 'permute', scale = TRUE, pred_wrapper = kernlab::predict,
   reference_class = '1', metric = 'auc', target = 'dv_next_fall', train = retain_us) %>%
 rename(rescaled_importance = Importance)
## testing data set
set.seed(51923)
svmr_fit_us_test <- workflow() %>% add_model(svmr_tune_final_us)  %>%
 add_recipe(retain_rec_us) %>%fit(retain_test)
set.seed(51923)
svmr_vi_us_test <- svmr_fit_us_test %>% extract_fit_parsnip() %>%
 vi(method = 'permute', pred_wrapper = kernlab::predict, reference_class = '1',
   metric = 'auc', target = 'dv_next_fall', train = retain_test)
set.seed(51923)
svmr_vi_us_rs_test <- svmr_fit_us_test %>%extract_fit_parsnip() %>%
 vi(method = 'permute', scale = TRUE, pred_wrapper = kernlab::predict,
   reference_class = '1', metric = 'auc', target = 'dv_next_fall', train = retain_test) %>%
 rename(rescaled_importance = Importance)

## joining the results together
svmr_vi_compar <- svmr_vi %>% select(-Importance) %>%
 mutate(rescaled_importance = 0, type = '1.1 None - Train') %>%
 rbind(svmr_vi_test %>% select(-Importance) %>%
        mutate(type = '1.2 None - Test')) %>%
 rbind(svmr_vi_ds_rs %>% mutate(type = '1.1 DS - Train')) %>%
 rbind(svmr_vi_ds_rs_test %>% mutate(type = '1.2 DS - Test')) %>%
 rbind(svmr_vi_us_rs %>% mutate(type = '2.1 US - Train')) %>%
 rbind(svmr_vi_us_rs_test %>% mutate(rescaled_importance = 0,  type = '2.2 US - Test'))

svmr_vi_compar %>%  filter(!type %in% c('1.1 None - Train',  '1.2 None - Test')) %>%
 mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                     Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                     Variable == 'hsgpa_knn' ~ 'HS GPA',
                     Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                     Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                     Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                     Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                     Variable == 'cip_categories' ~ 'Major Groupings',
                     Variable == 'efc_knn' ~ 'EFC',
                     Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                     Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                     Variable == 'pell' ~ 'PELL Grant',
                     Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                     Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                     Variable == 'oth_loans' ~ 'Other Loans',
                     Variable == 'ats_knn' ~ 'Admissions Test Scores',
                     Variable == 'all_other_exp' ~ 'All Other',
                     Variable == 'instr_exp' ~ 'Instruction',
                     Variable == 'stu_serv_exp' ~ 'Student Services',
                     Variable == 'race_eth' ~ 'Race Ethnicity',
                     Variable == 'cm_ready' ~ 'CM & Ready Mean',
                     Variable == 'locale_group' ~ 'HS Locale',
                     Variable == 'college_prep' ~ 'College Prep. Curric.',
                     Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                     Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                     Variable == 'english_cm' ~ 'English (CMR)',
                     Variable == 'math_cm' ~ 'Math (CMR)',
                     Variable == 'science_cm' ~ 'Science (CMR)',
                     Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                     TRUE ~ 'CHECK')) %>%ggplot() +
 geom_bar(aes(x = reorder(Variable, rescaled_importance), y = rescaled_importance,
      fill = desc(rescaled_importance/100)), stat = 'identity') + theme_classic() +
 theme(legend.position = 'none', axis.title.y = element_blank(),
   text = element_text(size = 15)) + ylab('Rescale Importance') + coord_flip() +
 facet_wrap(. ~ type, ncol = 4)

## predictive power
## no modifications--training data set
set.seed(7323)
svmr_cv_train <- svmr_tune_final %>% fit_resamples(dv_next_fall ~.,
      resamples = retain_cv, metrics = model_metrics, control = ctrl_grid)
svmr_cv_train %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
svmr_cv_train %>% collect_metrics()
## testing data set
set.seed(7323)
svmr_cv_test <- svm_r_wf %>% update_recipe(retain_test_rec_rv) %>%
 update_model(svmr_tune_final) %>% fit_resamples(as.factor(dv_next_fall) ~.,
      resamples = retain_test_cv, metrics = model_metrics, control = ctrl_grid)
svmr_cv_test %>% collect_predictions() %>% conf_mat(dv_next_fall,  .pred_class)
svmr_cv_test %>% collect_metrics()

## downsample modification--training data set
set.seed(7323)
svmr_cv_train_ds <- svmr_tune_final_ds %>% fit_resamples(dv_next_fall ~.,
      resamples = retain_cv_ds,  metrics = model_metrics, control = ctrl_grid)
svmr_cv_train_ds %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
svmr_cv_train_ds %>% collect_metrics()
## testing data set
set.seed(7323)
```

```
svmr_cv_test_ds <- svm_r_wf_ds %>% update_recipe(retain_test_rec_rv) %>%
 update_model(svmr_tune_final_ds) %>% fit_resamples(as.factor(dv_next_fall) ~.,
            resamples = retain_test_cv, metrics = model_metrics, control = ctrl_grid)
svmr_cv_test_ds %>% collect_predictions() %>% conf_mat(dv_next_fall,
        .pred_class)
svmr_cv_test_ds %>% collect_metrics()

## upsample modifications--training data set
set.seed(7323)
svmr_cv_train_us <- svmr_tune_final_us %>% fit_resamples(dv_next_fall ~.,
            resamples = retain_cv_us, metrics = model_metrics, control = ctrl_grid)
svmr_cv_train_us %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
svmr_cv_train_us %>% collect_metrics()
## testing data set
set.seed(7323)
svmr_cv_test_us <- svm_r_wf_us %>% update_recipe(retain_test_rec_rv) %>%
 update_model(svmr_tune_final_us) %>% fit_resamples(as.factor(dv_next_fall) ~.,
            resamples = retain_test_cv,  metrics = model_metrics, control = ctrl_grid)
svmr_cv_test_us %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
svmr_cv_test_us %>% collect_metrics()

## roc values
svmr_roc <- svmr_cv_train %>% collect_predictions() %>%
 roc_curve(truth = dv_next_fall,  .pred_0) %>%
 mutate(model = '1. None',  type = 'Training') %>%
 rbind(svmr_cv_train %>% collect_predictions() %>%
      roc_curve(truth = dv_next_fall,  .pred_0) %>%
      mutate(model = '1. None',  type = 'Testing')) %>%
 rbind(svmr_cv_train_ds %>% collect_predictions() %>%
      roc_curve(truth = dv_next_fall,  .pred_0) %>%
      mutate(model = '2. Downsample', type = 'Training')) %>%
 rbind(svmr_cv_test_ds %>% collect_predictions() %>%
      roc_curve(truth = dv_next_fall,  .pred_0) %>%
      mutate(model = '2. Downsample',  type = 'Testing')) %>%
 rbind(svmr_cv_train_us %>% collect_predictions() %>%
      roc_curve(truth = dv_next_fall,  .pred_0) %>%
      mutate(model = '3. Upsample', type = 'Training')) %>%
 rbind(svmr_cv_test_us %>% collect_predictions() %>%
      roc_curve(truth = dv_next_fall,.pred_0) %>%
      mutate(model = '3. Upsample', type = 'Testing')) %>%
 rename(`Data Set` = type) %>% mutate(Algorithm = '4. SVM Radial')

####################
##RANDOM FOREST ##
####################
```

```
## model specifications
rf_spec <- rand_forest(mtry = tune(), trees = tune(), min_n = tune()) %>%
 set_mode("classification") %>% set_engine("ranger")
## workflows
## no modifications
rf_wf <- workflow() %>%  add_model(rf_spec) %>% add_recipe(retain_rec_rv)
## downsample modifications
rf_wf_ds <- workflow() %>% add_model(rf_spec) %>% add_recipe(retain_rec_ds)
## upsample modifications
rf_wf_us <- workflow() %>% add_model(rf_spec) %>% add_recipe(retain_rec_us)

##tuning
## no modifications
doParallel::registerDoParallel()
set.seed(51823)
rf_wf_tune <- tune_grid(rf_wf,resamples = retain_cv,
                    metrics = model_metrics, control = ctrl_grid, grid = 20)
## best model for no sample
rf_tune_best <- select_best(rf_wf_tune, 'roc_auc')

# finalized model for no sample
rf_tune_final <- finalize_model(rf_spec, rf_tune_best)
## downsample modifications
doParallel::registerDoParallel()
set.seed(51823)
rf_wf_tune_ds <- tune_grid(rf_wf_ds, resamples = retain_cv_ds,
                      metrics = model_metrics, control = ctrl_grid, grid = 20)
## best model for downsample
rf_tune_best_ds <- select_best(rf_wf_tune_ds, 'roc_auc')
## finalized model for downsample
rf_tune_final_ds <- finalize_model(rf_spec, rf_tune_best_ds)
## upsample modifications
doParallel::registerDoParallel()
set.seed(51823)
rf_wf_tune_us <- tune_grid(rf_wf_us,resamples = retain_cv_us,
                      metrics = model_metrics, control = ctrl_grid, grid = 20)
## best model for upsample
rf_tune_best_us <- select_best(rf_wf_tune_us, 'roc_auc')
## finalized model for upsample
rf_tune_final_us <- finalize_model(rf_spec, rf_tune_best_us)
```

```
## variable importance analysis
## no modifications--training data set
set.seed(511923)
rf_vi <- rf_tune_final %>% set_engine('ranger', importance = 'permutation') %>%
 fit(dv_next_fall ~ ., data = retain) %>%  vi()
set.seed(511923)
rf_vi_rs <- rf_tune_final %>% set_engine('ranger', importance = 'permutation') %>%
 fit(dv_next_fall ~ ., data = retain) %>% vi(scale = TRUE) %>%
 rename(rescaled_importance = Importance) %>% mutate(type = '1.1 None - Train')
## testing data set
set.seed(511923)
rf_vi_test <- rf_tune_final %>% set_engine('ranger', importance = 'permutation') %>%
 fit(dv_next_fall ~ ., data = retain_test) %>% vi()
set.seed(511923)
rf_vi_rs_test <- rf_tune_final %>% set_engine('ranger', importance = 'permutation') %>%
 fit(dv_next_fall ~ ., data = retain_test) %>% vi(scale = TRUE) %>%
 rename(rescaled_importance = Importance) %>% mutate(type = '1.2 None - Test')

## downsample modification--training data set
set.seed(511923)
rf_vi_ds <- rf_tune_final_ds %>% set_engine('ranger', importance = 'permutation') %>%
 fit(dv_next_fall ~ ., data = retain_ds) %>%vi()
set.seed(511923)
rf_vi_ds_rs <- rf_tune_final_ds %>% set_engine('ranger',
          importance = 'permutation') %>% fit(dv_next_fall ~ ., data = retain_ds) %>%
 vi(scale = TRUE) %>% rename(rescaled_importance = Importance) %>%
 mutate(type = '2.1 DS - Train')
## testing data set
set.seed(511923)
rf_vi_ds_test <- rf_tune_final_ds %>% set_engine('ranger',
          importance = 'permutation') %>% fit(dv_next_fall ~ .,
    data = retain_test) %>% vi()
set.seed(511923)
rf_vi_ds_rs_test <- rf_tune_final_ds %>%set_engine('ranger',
          importance = 'permutation') %>% fit(dv_next_fall ~ .,
    data = retain_test) %>% vi(scale = TRUE) %>%
 rename(rescaled_importance = Importance) %>% mutate(type = '2.2 DS - Test')

## upsample modifications--training data set
set.seed(511923)
rf_vi_us <- rf_tune_final_us %>% set_engine('ranger',
          importance = 'permutation') %>% fit(dv_next_fall ~ ., data = retain_us) %>% vi()
set.seed(511923)
rf_vi_us_rs <- rf_tune_final_us %>% set_engine('ranger',
          importance = 'permutation') %>% fit(dv_next_fall ~ ., data = retain_us) %>%
 vi(scale = TRUE) %>% rename(rescaled_importance = Importance) %>%
 mutate(type = '3.1 US - Train')
## testing data set
set.seed(511923)
rf_vi_us_test <- rf_tune_final_us %>% set_engine('ranger',
          importance = 'permutation') %>% fit(dv_next_fall ~ .,
     data = retain_test) %>% vi()
set.seed(511923)
rf_vi_us_rs_test <- rf_tune_final_us %>% set_engine('ranger',
          importance = 'permutation') %>% fit(dv_next_fall ~ .,
     data = retain_test) %>% vi(scale = TRUE) %>%
 rename(rescaled_importance = Importance) %>% mutate(type = '3.2 US - Test')

## joining result together
rf_vi_rs %>% rbind(rf_vi_rs_test) %>%
 rbind(rf_vi_ds_rs) %>%  rbind(rf_vi_ds_rs_test) %>%
 rbind(rf_vi_us_rs) %>% rbind(rf_vi_us_rs_test) %>%
 mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                         Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                         Variable == 'hsgpa_knn' ~ 'HS GPA',
                         Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                         Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                         Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                         Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                         Variable == 'cip_categories' ~ 'Major Groupings',
                         Variable == 'efc_knn' ~ 'EFC',
                         Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                         Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                         Variable == 'pell' ~ 'PELL Grant',
                         Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                         Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                         Variable == 'oth_loans' ~ 'Other Loans',
                         Variable == 'ats_knn' ~ 'Admissions Test Scores',
                         Variable == 'all_other_exp' ~ 'All Other',
                         Variable == 'instr_exp' ~ 'Instruction',
                         Variable == 'stu_serv_exp' ~ 'Student Services',
                         Variable == 'race_eth' ~ 'Race Ethnicity',
                         Variable == 'cm_ready' ~ 'CM & Ready Mean',
                         Variable == 'locale_group' ~ 'HS Locale',
                         Variable == 'college_prep' ~ 'College Prep. Curric.',
                         Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                         Variable == 'public_rsch_exp' ~ 'Public Service Research',
                         Variable == 'english_cm' ~ 'English (CMR)',
                         Variable == 'math_cm' ~ 'Math (CMR)',
                         Variable == 'science_cm' ~ 'Science (CMR)',
                         Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                         TRUE ~ 'CHECK')) %>%ggplot() +
 geom_bar(aes(x = reorder(Variable, rescaled_importance), y = rescaled_importance,
            fill = desc(rescaled_importance/100)), stat = 'identity') + theme_classic() +
 theme(legend.position = 'none',  axis.title.y = element_blank(),
       text = element_text(size = 15)) + ylab('Rescale Importance') + coord_flip() +
 facet_wrap(. ~ type,  ncol = 6)

## predictive power
## no modifications--training data set
set.seed(7323)
rf_cv_train <- rf_tune_final %>% fit_resamples(dv_next_fall ~.,
             resamples = retain_cv, metrics = model_metrics, control = ctrl_grid)
rf_cv_train %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
rf_cv_train %>%collect_metrics()
## testing data set
set.seed(7323)
rf_cv_test <- rf_wf %>% update_recipe(retain_test_rec_rv) %>%
 update_model(rf_tune_final) %>% fit_resamples(as.factor(dv_next_fall) ~.,
             resamples = retain_test_cv, metrics = model_metrics, control = ctrl_grid)
rf_cv_test %>% collect_predictions() %>% conf_mat(dv_next_fall,.pred_class)
rf_cv_test %>% collect_metrics()

## downsample modifications--training data set
set.seed(7323)
rf_cv_train_ds <- rf_tune_final_ds %>%fit_resamples(dv_next_fall ~.,
             resamples = retain_cv_ds, metrics = model_metrics, control = ctrl_grid)
rf_cv_train_ds %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
rf_cv_train_ds %>% collect_metrics()
## testing data set
set.seed(7323)
rf_cv_test_ds <- rf_wf_ds %>% update_recipe(retain_test_rec_rv) %>%
 update_model(rf_tune_final_ds) %>% fit_resamples(as.factor(dv_next_fall) ~.,
             resamples = retain_test_cv, metrics = model_metrics, control = ctrl_grid)
rf_cv_test_ds %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
rf_cv_test_ds %>% collect_metrics()

## upsample modifications--training data set
set.seed(7323)
rf_cv_train_us <- rf_tune_final_us %>%fit_resamples(dv_next_fall ~.,
             resamples = retain_cv_us, metrics = model_metrics, control = ctrl_grid)
rf_cv_train_us %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
rf_cv_train_us %>% collect_metrics()
## testing data set
set.seed(7323)
rf_cv_test_us <- rf_wf_us %>% update_recipe(retain_test_rec_rv) %>%
 update_model(rf_tune_final_us) %>% fit_resamples(as.factor(dv_next_fall) ~.,
             resamples = retain_test_cv, metrics = model_metrics, control = ctrl_grid)
```

```
rf_cv_test_us %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
rf_cv_test_us %>% collect_metrics()

## roc values
rf_roc <- rf_cv_train %>% collect_predictions() %>%
 roc_curve(truth = dv_next_fall,.pred_0) %>%
 mutate(model = '1. None', type = 'Training') %>%
 rbind(rf_cv_train %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall, .pred_0) %>%
       mutate(model = '1. None',  type = 'Testing')) %>%
 rbind(rf_cv_train_ds %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall, .pred_0) %>%
       mutate(model = '2. Downsample', type = 'Training')) %>%
 rbind(rf_cv_test_ds %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall,  .pred_0) %>%
       mutate(model = '2. Downsample', type = 'Testing')) %>%
 rbind(rf_cv_train_us %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall, .pred_0) %>%
       mutate(model = '3. Upsample', type = 'Training')) %>%
 rbind(rf_cv_test_us %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall, .pred_0) %>%
       mutate(model = '3. Upsample', type = 'Testing')) %>%
 rename(`Data Set` = type) %>% mutate(Algorithm = '5. Random Forest')

######################################
##EXTREME GRADIENT BOOSTING ##
######################################

## model specifications
xgb <- boost_tree(trees = tune(), tree_depth = tune(),  min_n = tune(),
 loss_reduction = tune(), sample_size = tune(), mtry = tune(),
 learn_rate = tune()) %>% set_engine('xgboost') %>% set_mode('classification')
## workflows
## no modifications
xgb_wf <- workflow() %>% add_model(xgb) %>% add_recipe(retain_rec_rv)
## downsample modifications
xgb_wf_ds <- workflow() %>% add_model(xgb) %>% add_recipe(retain_rec_ds)
## upsample modifications
xgb_wf_us <- workflow() %>% add_model(xgb) %>% add_recipe(retain_rec_us)

## tuning
## no modifications
doParallel::registerDoParallel()
set.seed(51923)
xgb_wf_tune <- tune_grid(xgb_wf, resamples = retain_cv, metrics = model_metrics,
                    control = ctrl_grid, grid = 20)
```

```
## best model
xgb_tune_best <- select_best(xgb_wf_tune, 'roc_auc')
## model fixed to the best outcome model
xgb_tune_final <- finalize_model(xgb, xgb_tune_best)
## downsample
doParallel::registerDoParallel()
set.seed(51923)
xgb_wf_tune_ds <- tune_grid(xgb_wf_ds, resamples = retain_cv_ds,
                          metrics = model_metrics, control = ctrl_grid, grid = 20)
## best model
xgb_tune_best_ds <- select_best(xgb_wf_tune_ds, 'roc_auc')
## model fixed to the best outcome model
xgb_tune_final_ds <- finalize_model(xgb, xgb_tune_best_ds)
## upsample
doParallel::registerDoParallel()
set.seed(51923)
xgb_wf_tune_us <- tune_grid(xgb_wf_us, resamples = retain_cv_us,
                          metrics = model_metrics, control = ctrl_grid, grid = 20)
## best model
xgb_tune_best_us <- select_best(xgb_wf_tune_us, 'roc_auc')
## model fixed to the best outcome model
xgb_tune_final_us <- finalize_model(xgb, xgb_tune_best_us)

## variable importance analysis
## no modifications--training data set
set.seed(511923)
xgb_vi <- xgb_tune_final %>% set_engine('xgboost') %>%fit(dv_next_fall ~ .,
    data = retain ) %>% vi()
set.seed(511923)
xgb_vi_rs <- xgb_tune_final %>% set_engine('xgboost') %>%
 fit(dv_next_fall ~ ., data = retain) %>% vi(scale = TRUE) %>%
 rename(rescaled_importance = Importance) %>% mutate(type = '1.1 None - Train')
## testing data set
set.seed(511923)
xgb_vi_test <- xgb_tune_final %>% set_engine('xgboost') %>%
 fit(dv_next_fall ~ .,  data = retain_test ) %>% vi()
set.seed(511923)
xgb_vi_rs_test <- xgb_tune_final %>% set_engine('xgboost') %>%
 fit(dv_next_fall ~ ., data = retain_test) %>% vi(scale = TRUE) %>%
 rename(rescaled_importance = Importance) %>% mutate(type = '1.2 None - Test')

## downsample modifications--training data set
set.seed(511923)
xgb_vi_ds <- xgb_tune_final_ds %>% set_engine('xgboost') %>%
 fit(dv_next_fall ~ .,  data = retain_ds) %>%vi()
set.seed(511923)
```

```
xgb_vi_ds_rs <- xgb_tune_final_ds %>% set_engine('xgboost') %>%
 fit(dv_next_fall ~ ., data = retain_ds) %>% vi(scale = TRUE) %>%
 rename(rescaled_importance = Importance) %>% mutate(type = '2.1 DS - Train')
## testing data set
set.seed(511923)
xgb_vi_ds_test <- xgb_tune_final_ds %>% set_engine('xgboost') %>%
 fit(dv_next_fall ~ ., data = retain_test) %>% vi()
set.seed(511923)
xgb_vi_ds_rs_test <- xgb_tune_final_ds %>% set_engine('xgboost') %>%
 fit(dv_next_fall ~ ., data = retain_test) %>% vi(scale = TRUE) %>%
 rename(rescaled_importance = Importance) %>% mutate(type = '2.2 DS - Test')

## upsample modifications--training data set
set.seed(511923)
xgb_vi_us <- xgb_tune_final_us %>% set_engine('xgboost') %>%
 fit(dv_next_fall ~ .,  data = retain_us) %>% vi()
set.seed(511923)
xgb_vi_us_rs <- xgb_tune_final_us %>% set_engine('xgboost') %>%
 fit(dv_next_fall ~ ., data = retain_us) %>% vi(scale = TRUE) %>%
 rename(rescaled_importance = Importance) %>% mutate(type = '3.1 US - Train')
## testing data set
set.seed(511923)
xgb_vi_us_test <- xgb_tune_final_us %>%set_engine('xgboost') %>%
 fit(dv_next_fall ~ ., data = retain_test) %>% vi()
set.seed(511923)
xgb_vi_us_rs_test <- xgb_tune_final_us %>% set_engine('xgboost') %>%
 fit(dv_next_fall ~ ., data = retain_test) %>% vi(scale = TRUE) %>%
 rename(rescaled_importance = Importance) %>% mutate(type = '3.2 US - Test')

## joining results together
xgb_vi_rs %>%
 rbind(xgb_vi_rs_test) %>%
 rbind(xgb_vi_ds_rs) %>%
 rbind(xgb_vi_ds_rs_test) %>%
 rbind(xgb_vi_us_rs) %>%
 rbind(xgb_vi_us_rs_test) %>%
 mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                             Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                             Variable == 'hsgpa_knn' ~ 'HS GPA',
                             Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                             Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                             Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                             Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                             Variable == 'cip_categories' ~ 'Major Groupings',
                             Variable == 'efc_knn' ~ 'EFC',
                             Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                             Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                             Variable == 'pell' ~ 'PELL Grant',
                             Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                             Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                             Variable == 'oth_loans' ~ 'Other Loans',
                             Variable == 'ats_knn' ~ 'Admissions Test Scores',
                             Variable == 'all_other_exp' ~ 'All Other',
                             Variable == 'instr_exp' ~ 'Instruction',
                             Variable == 'stu_serv_exp' ~ 'Student Services',
                             Variable == 'race_eth' ~ 'Race Ethnicity',
                             Variable == 'cm_ready' ~ 'CM & Ready Mean',
                             Variable == 'locale_group' ~ 'HS Locale',
                             Variable == 'college_prep' ~ 'College Prep. Curric.',
                             Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                             Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                             Variable == 'english_cm' ~ 'English (CMR)',
                             Variable == 'math_cm' ~ 'Math (CMR)',
                             Variable == 'science_cm' ~ 'Science (CMR)',
                             Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                             TRUE ~ 'CHECK')) %>%  ggplot() +
 geom_bar(aes(x = reorder(Variable, rescaled_importance), y = rescaled_importance,
        fill = desc(rescaled_importance/100)),  stat = 'identity') + theme_classic() +
 theme(legend.position = 'none',  axis.title.y = element_blank(),
    text = element_text(size = 15)) + ylab('Rescale Importance') + coord_flip() +
 facet_wrap(. ~ type,  ncol = 6)

## predictive power
## no modifications--training data set
set.seed(7323)
xgb_cv_train <- xgb_tune_final %>% fit_resamples(dv_next_fall ~.,
        resamples = retain_cv,  metrics = model_metrics, control = ctrl_grid)
xgb_cv_train %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
xgb_cv_train %>% collect_metrics()
## testing data set
set.seed(7323)
xgb_cv_test <- xgb_wf %>% update_recipe(retain_test_rec_rv) %>%
 update_model(xgb_tune_final) %>% fit_resamples(as.factor(dv_next_fall) ~.,
        resamples = retain_test_cv, metrics = model_metrics, control = ctrl_grid)
xgb_cv_test %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
xgb_cv_test %>% collect_metrics()

## downsample modifications--training data set
set.seed(7323)
xgb_cv_train_ds <- xgb_tune_final_ds %>% fit_resamples(dv_next_fall ~.,
        resamples = retain_cv_ds, metrics = model_metrics, control = ctrl_grid)
xgb_cv_train_ds %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
xgb_cv_train_ds %>%collect_metrics()
## testing data set
set.seed(7323)
xgb_cv_test_ds <- xgb_wf_ds %>% update_recipe(retain_test_rec_rv) %>%
 update_model(xgb_tune_final_ds) %>% fit_resamples(as.factor(dv_next_fall) ~.,
              resamples = retain_test_cv, metrics = model_metrics, control = ctrl_grid)
xgb_cv_test_ds %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
xgb_cv_test_ds %>% collect_metrics()

## upsample modifications--training data set
set.seed(7323)
xgb_cv_train_us <- xgb_tune_final_us %>% fit_resamples(dv_next_fall ~.,
              resamples = retain_cv_us, metrics = model_metrics, control = ctrl_grid)
xgb_cv_train_us %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
xgb_cv_train_us %>% collect_metrics()
## testing data set
set.seed(7323)
xgb_cv_test_us <- xgb_wf_us %>% update_recipe(retain_test_rec_rv) %>%
 update_model(xgb_tune_final_us) %>% fit_resamples(as.factor(dv_next_fall) ~.,
              resamples = retain_test_cv, metrics = model_metrics, control = ctrl_grid)
xgb_cv_test_us %>% collect_predictions() %>% conf_mat(dv_next_fall, .pred_class)
xgb_cv_test_us %>% collect_metrics()

## roc values
xgb_roc <- xgb_cv_train %>% collect_predictions() %>%
 roc_curve(truth = dv_next_fall,.pred_0) %>%
 mutate(model = '1. None', type = 'Training') %>%
 rbind(xgb_cv_train %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall,.pred_0) %>%
       mutate(model = '1. None', type = 'Testing')) %>%
 rbind(xgb_cv_train_ds %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall,.pred_0) %>%
       mutate(model = '2. Downsample', type = 'Training')) %>%
 rbind(xgb_cv_test_ds %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall,.pred_0) %>%
       mutate(model = '2. Downsample', type = 'Testing')) %>%
 rbind(xgb_cv_train_us %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall,  .pred_0) %>%
       mutate(model = '3. Upsample', type = 'Training')) %>%
 rbind(xgb_cv_test_us %>% collect_predictions() %>%
       roc_curve(truth = dv_next_fall,.pred_0) %>%
       mutate(model = '3. Upsample', type = 'Testing')) %>%
 rename(`Data Set` = type) %>% mutate(Algorithm = '6. XGBoost')
```

```
######################################################
## VARIABLE COMPARISON OF TESTING DATA SETS ##
######################################################

## no modifications
log_reg_vi %>% filter(log_reg_type == '4. Testing') %>%
  mutate(type = '1. Log. Reg.') %>%
  select(Variable, rescale_importance, type) %>%
  rbind(svml_vi_rs_test %>% mutate(Importance = 0, type = '2.1 SVM Linear') %>%
        rename(rescale_importance = Importance)) %>%
  rbind(svmp_vi_rs_test %>% mutate(rescaled_importance = 0,
              type = '2.2 SVM Poly.') %>%
        rename(rescale_importance = rescaled_importance)) %>%
  rbind(svmr_vi_test %>% mutate(type = '2.3 SVM RBF') %>%
        rename(rescale_importance = rescaled_importance) %>%
        select(Variable, rescale_importance, type)) %>%
  rbind(rf_vi_rs_test %>% rename(rescale_importance = rescaled_importance) %>%
        mutate(type = '3. RF') %>%
        select(Variable, rescale_importance, type)) %>%
  rbind(xgb_vi_rs_test %>% rename(rescale_importance = rescaled_importance) %>%
        mutate(type = '4. XGB') %>%
        select(Variable, rescale_importance, type)) %>%
 mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                         Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                         Variable == 'hsgpa_knn' ~ 'HS GPA',
                         Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                         Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                         Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                         Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                         Variable == 'cip_categories' ~ 'Major Groupings',
                         Variable == 'efc_knn' ~ 'EFC',
                         Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                         Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                         Variable == 'pell' ~ 'PELL Grant',
                         Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                         Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                         Variable == 'oth_loans' ~ 'Other Loans',
                         Variable == 'ats_knn' ~ 'Admissions Test Scores',
                         Variable == 'all_other_exp' ~ 'All Other',
                         Variable == 'instr_exp' ~ 'Instruction',
                         Variable == 'stu_serv_exp' ~ 'Student Services',
                         Variable == 'race_eth' ~ 'Race Ethnicity',
                         Variable == 'cm_ready' ~ 'CM & Ready Mean',
                         Variable == 'locale_group' ~ 'HS Locale',
                         Variable == 'college_prep' ~ 'College Prep. Curric.',
                         Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                            Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                            Variable == 'english_cm' ~ 'English (CMR)',
                            Variable == 'math_cm' ~ 'Math (CMR)',
                            Variable == 'science_cm' ~ 'Science (CMR)',
                            Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                            TRUE ~ 'CHECK')) %>%
 ggplot(aes(x = reorder(Variable, desc(Variable)), y = rescale_importance)) +
 geom_bar(aes(fill = rescale_importance/10),  stat = 'identity') + theme_classic() +
 theme(legend.position = 'none',  axis.title = element_blank(),
      axis.text.x = element_text(size = 10), axis.text.y = element_text(size = 12)) +
 coord_flip() + facet_wrap(. ~ type, ncol = 6)

## downsample modifications
log_reg_vi %>%   filter(log_reg_type == '4. Testing') %>%
 mutate(type = '1. Log. Reg.') %>%
 select(Variable, rescale_importance, type) %>%
 rbind(svml_vi_ds_rs_test %>%
       mutate(rescaled_importance = 0, type = '2.1 SVM Linear') %>%
       rename(rescale_importance = rescaled_importance)) %>%
 rbind(svmp_vi_ds_rs_test %>% mutate(type = '2.2 SVM Poly.') %>%
       rename(rescale_importance = rescaled_importance)) %>%
 rbind(svmr_vi_ds_rs_test %>% mutate(type = '2.3 SVM RBF') %>%
       rename(rescale_importance = rescaled_importance) %>%
       select(Variable,  rescale_importance,  type)) %>%
 rbind(rf_vi_ds_rs_test %>% rename(rescale_importance = rescaled_importance) %>%
       mutate(type = '3. RF') %>%
       select(Variable, rescale_importance, type)) %>%
 rbind(xgb_vi_ds_rs_test %>%
       rename(rescale_importance = rescaled_importance) %>%
       mutate(type = '4. XGB') %>%
       select(Variable, rescale_importance, type)) %>%
mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                            Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                            Variable == 'hsgpa_knn' ~ 'HS GPA',
                            Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                            Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                            Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                            Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                            Variable == 'cip_categories' ~ 'Major Groupings',
                            Variable == 'efc_knn' ~ 'EFC',
                            Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                            Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                            Variable == 'pell' ~ 'PELL Grant',
                            Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                            Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                            Variable == 'oth_loans' ~ 'Other Loans',
                            Variable == 'ats_knn' ~ 'Admissions Test Scores',
                            Variable == 'all_other_exp' ~ 'All Other',
                            Variable == 'instr_exp' ~ 'Instruction',
                            Variable == 'stu_serv_exp' ~ 'Student Services',
                            Variable == 'race_eth' ~ 'Race Ethnicity',
                            Variable == 'cm_ready' ~ 'CM & Ready Mean',
                            Variable == 'locale_group' ~ 'HS Locale',
                            Variable == 'college_prep' ~ 'College Prep. Curric.',
                            Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                            Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                            Variable == 'english_cm' ~ 'English (CMR)',
                            Variable == 'math_cm' ~ 'Math (CMR)',
                            Variable == 'science_cm' ~ 'Science (CMR)',
                            Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                            TRUE ~ 'CHECK')) %>%
 ggplot(aes(x = reorder(Variable, desc(Variable)), y = rescale_importance)) +
 geom_bar(aes(fill = rescale_importance/10), stat = 'identity') + theme_classic() +
 theme(legend.position = 'none', axis.title = element_blank(),
    axis.text.x = element_text(size = 10), axis.text.y = element_text(size = 12)) +
 coord_flip()+ facet_wrap(. ~ type, ncol = 6)

## upsample modifications
log_reg_vi %>%  filter(log_reg_type == '4. Testing') %>%
 mutate(type = '1. Log. Reg.') %>%
 select(Variable,  rescale_importance,  type) %>%
 rbind(svml_vi_us_rs_test %>%
       mutate(rescaled_importance = 0,  type = '2.1 SVM Linear') %>%
       rename(rescale_importance = rescaled_importance)) %>%
 rbind(svmp_vi_us_rs_test %>%  mutate(type = '2.2 SVM Poly.') %>%
       rename(rescale_importance = rescaled_importance)) %>%
 rbind(svmr_vi_us_rs_test %>% mutate(type = '2.3 SVM RBF') %>%
       rename(rescale_importance = rescaled_importance) %>%
       select(Variable, rescale_importance, type)) %>%
 rbind(rf_vi_us_rs_test %>%
       rename(rescale_importance = rescaled_importance) %>%
       mutate(type = '3. RF') %>%
       select(Variable, rescale_importance, type)) %>%
 rbind(xgb_vi_us_rs_test %>%
       rename(rescale_importance = rescaled_importance) %>%
       mutate(type = '4. XGB') %>%
       select(Variable, rescale_importance, type)) %>%
mutate(Variable = case_when(Variable == 'gender_descr' ~ 'Gender',
                            Variable == 'admit_first_gen_ind' ~ 'First Generation Status',
                            Variable == 'hsgpa_knn' ~ 'HS GPA',
                            Variable == 'adv_standing_ap_hrs' ~ 'AP Hours',
                            Variable == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                            Variable == 'adv_standing_ib_hrs' ~ 'IB Hours',
                            Variable == 'adv_standing_other_hrs' ~ 'Other Hours',
                            Variable == 'cip_categories' ~ 'Major Groupings',
                            Variable == 'efc_knn' ~ 'EFC',
                            Variable == 'ga_hope' ~ 'GA HOPE Scholarship',
                            Variable == 'zell_ind' ~ 'Zell Miller Indicator',
                            Variable == 'pell' ~ 'PELL Grant',
                            Variable == 'fed_sub_loans' ~ 'Federal Sub. Loans',
                            Variable == 'fed_unsub_loans' ~ 'Federal Unsub. Loans',
                            Variable == 'oth_loans' ~ 'Other Loans',
                            Variable == 'ats_knn' ~ 'Admissions Test Scores',
                            Variable == 'all_other_exp' ~ 'All Other',
                            Variable == 'instr_exp' ~ 'Instruction',
                            Variable == 'stu_serv_exp' ~ 'Student Services',
                            Variable == 'race_eth' ~ 'Race Ethnicity',
                            Variable == 'cm_ready' ~ 'CM & Ready Mean',
                            Variable == 'locale_group' ~ 'HS Locale',
                            Variable == 'college_prep' ~ 'College Prep. Curric.',
                            Variable == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                            Variable == 'public_rsch_exp' ~ 'Public Service  Research',
                            Variable == 'english_cm' ~ 'English (CMR)',
                            Variable == 'math_cm' ~ 'Math (CMR)',
                            Variable == 'science_cm' ~ 'Science (CMR)',
                            Variable == 'social_studies_cm' ~ 'Social Studies (CMR)',
                            TRUE ~ 'CHECK')) %>%
          ggplot(aes(x = reorder(Variable, desc(Variable)), y = rescale_importance)) +
          geom_bar(aes(fill = rescale_importance/10), stat = 'identity') + theme_classic() +
          theme(legend.position = 'none',  axis.title = element_blank(),
              axis.text.x = element_text(size = 10), axis.text.y = element_text(size = 12)) +
          coord_flip() + facet_wrap(. ~ type, ncol = 6)

        ###########################
        ##ENSEMBLE LEARNING ##
        ###########################

        ## pulling out prediction
        ## Logistic Regression
        ## no modifications
        lr_train_preds <- log_reg_cv_train %>%collect_predictions()
        lr_test_preds <- log_reg_cv_test %>% collect_predictions()
        ## downsample modifications
        lr_train_preds_ds <- log_reg_cv_train_ds %>% collect_predictions()
        lr_test_preds_ds <- log_reg_cv_test_ds %>% collect_predictions()
        ## upsample modifications
        lr_train_preds_us <- log_reg_cv_train_us %>% collect_predictions()
        lr_test_preds_us <- log_reg_cv_test_us %>% collect_predictions()
```

```
## Random Forest
## no modifications
rf_train_pred <- rf_cv_train %>% collect_predictions()
rf_test_pred <- rf_cv_test %>% collect_predictions()
## downsample modifications
rf_train_pred_ds <- rf_cv_train_ds %>% collect_predictions()
rf_test_pred_ds <- rf_cv_test_ds %>% collect_predictions()
## upsample modifications
rf_train_pred_us <- rf_cv_train_us %>%collect_predictions()
rf_test_pred_us <- rf_cv_test_us %>% collect_predictions()

## Extreme Gradient Boosting
## no modifications
xgb_train_pred <- xgb_cv_train %>% collect_predictions()
xgb_test_pred <- xgb_cv_test %>% collect_predictions()
## downsample modifications
xgb_train_pred_ds <- xgb_cv_train_ds %>% collect_predictions()
xgb_test_pred_ds <- xgb_cv_test_ds %>% collect_predictions()
## upsample modifications
xgb_train_pred_us <- xgb_cv_train_us %>% collect_predictions()
xgb_test_pred_us <- xgb_cv_test_us %>% collect_predictions()
## no modifications--training data set
blend_train <- lr_train_preds %>% select(.pred_0, dv_next_fall) %>%
 rename(log_reg = .pred_0) %>%
 cbind(rf_train_pred %>% rename(rand_for = .pred_0) %>% select(rand_for)) %>%
 cbind(xgb_train_pred %>% rename(xgb = .pred_0) %>% select(xgb))

## mean of the predictions
## no  modifications--training data set
mean_train <- lr_train_preds %>% select(.pred_0,  dv_next_fall) %>%
 rename(log_reg = .pred_0) %>%
 cbind(rf_train_pred %>% rename(rand_for = .pred_0) %>% select(rand_for)) %>%
 cbind(xgb_train_pred %>% rename(xgb = .pred_0) %>% select(xgb)) %>%
 mutate(mean_prob = (log_reg + rand_for + xgb) / 3,
        mean_pred_class = as.factor(case_when(mean_prob >= .5 ~ 0, TRUE ~ 1)),
        mean_prob_1 = 1 - mean_prob) %>%
 select(dv_next_fall,  mean_prob,  mean_prob_1, mean_pred_class)
mean_train %>%
 mutate(dv_next_fall = as.factor(dv_next_fall),
        mean_pred_class = as.factor(mean_pred_class)) %>%
 conf_mat(dv_next_fall, mean_pred_class) %>% summary()
## auc value
mean_train %>%  select(mean_prob,  dv_next_fall) %>%
 roc_auc(mean_prob, truth = dv_next_fall)
## testing data set
```

```
mean_test <- lr_test_preds %>% select(.pred_0, dv_next_fall) %>%
 rename(log_reg = .pred_0) %>%
 cbind(rf_test_pred %>% rename(rand_for = .pred_0) %>% select(rand_for)) %>%
 cbind(xgb_test_pred %>% rename(xgb = .pred_0) %>% select(xgb)) %>%
 mutate(mean_prob = (log_reg + rand_for + xgb) / 3,
        mean_pred_class = as.factor(case_when(mean_prob >= .5 ~ 0, TRUE ~ 1)),
        mean_prob_1 = 1 - mean_prob) %>%
 select(dv_next_fall, mean_prob, mean_prob_1, mean_pred_class)
mean_test %>%
 mutate(dv_next_fall = as.factor(dv_next_fall),
        mean_pred_class = as.factor(mean_pred_class)) %>%
 conf_mat(dv_next_fall, mean_pred_class) %>% summary()
## auc value
mean_test %>% select(mean_prob, dv_next_fall) %>%
 roc_auc(mean_prob, truth = dv_next_fall)

## downsample modifications—training data set
mean_train_ds <- lr_train_preds_ds %>% select(.pred_0,  dv_next_fall) %>%
 rename(log_reg = .pred_0) %>%
 cbind(rf_train_pred_ds %>% rename(rand_for = .pred_0) %>% select(rand_for)) %>%
 cbind(xgb_train_pred_ds %>% rename(xgb = .pred_0) %>% select(xgb)) %>%
 mutate(mean_prob = (log_reg + rand_for + xgb) / 3,
        mean_pred_class = as.factor(case_when(mean_prob >= .5 ~ 0,  TRUE ~ 1)),
        mean_prob_1 = 1 - mean_prob) %>%
 select(dv_next_fall, mean_prob, mean_prob_1,
        mean_pred_class)
mean_train_ds %>%
 mutate(dv_next_fall = as.factor(dv_next_fall),
        mean_pred_class = as.factor(mean_pred_class)) %>%
 conf_mat(dv_next_fall, mean_pred_class) %>%summary()
## auc value
mean_train_ds %>%  select(mean_prob, dv_next_fall) %>%
 roc_auc(mean_prob, truth = dv_next_fall)
## testing data set
mean_test_ds <- lr_test_preds_ds %>% select(.pred_0, dv_next_fall) %>%
 rename(log_reg = .pred_0) %>%
 cbind(rf_test_pred_ds %>% rename(rand_for = .pred_0) %>% select(rand_for)) %>%
 cbind(xgb_test_pred_ds %>% rename(xgb = .pred_0) %>% select(xgb)) %>%
 mutate(mean_prob = (log_reg + rand_for + xgb) / 3,
        mean_pred_class = as.factor(case_when(mean_prob >= .5 ~ 0,  TRUE ~ 1)),
        mean_prob_1 = 1 - mean_prob) %>%
 select(dv_next_fall, mean_prob, mean_prob_1, mean_pred_class)
mean_test_ds %>%
 mutate(dv_next_fall = as.factor(dv_next_fall),
        mean_pred_class = as.factor(mean_pred_class)) %>%
 conf_mat(dv_next_fall, mean_pred_class) %>% summary()
```

```
## auc value
mean_test_ds %>% select(mean_prob, dv_next_fall) %>%
 roc_auc(mean_prob, truth = dv_next_fall)

## upsample modifications--training data set
mean_train_us <- lr_train_preds_us %>% select(.pred_0, dv_next_fall) %>%
 rename(log_reg = .pred_0) %>%
 cbind(rf_train_pred_us %>% rename(rand_for = .pred_0) %>% select(rand_for)) %>%
 cbind(xgb_train_pred_us %>% rename(xgb = .pred_0) %>% select(xgb)) %>%
 mutate(mean_prob = (log_reg + rand_for + xgb) / 3,
        mean_pred_class = as.factor(case_when(mean_prob >= .5 ~ 0, TRUE ~ 1)),
        mean_prob_1 = 1 - mean_prob) %>%
 select(dv_next_fall, mean_prob, mean_prob_1, mean_pred_class)
mean_train_us %>%
 mutate(dv_next_fall = as.factor(dv_next_fall),
        mean_pred_class = as.factor(mean_pred_class)) %>%
 conf_mat(dv_next_fall,  mean_pred_class) %>% summary()
mean_train_us %>% select(mean_prob, dv_next_fall) %>%
 roc_auc(mean_prob, truth = dv_next_fall)
## testing data set
mean_test_us <- lr_test_preds_us %>% select(.pred_0,  dv_next_fall) %>%
 rename(log_reg = .pred_0) %>%
 cbind(rf_test_pred_us %>% rename(rand_for = .pred_0) %>%select(rand_for)) %>%
 cbind(xgb_test_pred_us %>% rename(xgb = .pred_0) %>% select(xgb)) %>%
 mutate(mean_prob = (log_reg + rand_for + xgb) / 3,
        mean_pred_class = as.factor(case_when(mean_prob >= .5 ~ 0, TRUE ~ 1)),
        mean_prob_1 = 1 - mean_prob) %>%
 select(dv_next_fall, mean_prob, mean_prob_1, mean_pred_class)
mean_test_us %>%
 mutate(dv_next_fall = as.factor(dv_next_fall),
        mean_pred_class = as.factor(mean_pred_class)) %>%
 conf_mat(dv_next_fall, mean_pred_class) %>% summary()
mean_test_us %>% select(mean_prob, dv_next_fall) %>%
 roc_auc(mean_prob, truth = dv_next_fall)

## roc values
esl_mean_roc <- mean_train %>% roc_curve(truth = dv_next_fall, mean_prob) %>%
 mutate(model = '1. None', type = 'Training') %>%
 rbind(mean_test %>% roc_curve(truth = dv_next_fall, mean_prob) %>%
        mutate(model = '1. None', type = 'Testing')) %>%
 rbind(mean_train_ds %>% roc_curve(truth = dv_next_fall, mean_prob) %>%
        mutate(model = '2. Downsample', type = 'Training')) %>%
 rbind(mean_test_ds %>% roc_curve(truth = dv_next_fall, mean_prob) %>%
        mutate(model = '2. Downsample', type = 'Testing')) %>%
 rbind(mean_train_us %>% roc_curve(truth = dv_next_fall, mean_prob) %>%
        mutate(model = '3. Upsample', type = 'Training')) %>%
 rbind(mean_test_us %>% roc_curve(truth = dv_next_fall, mean_prob) %>%
       mutate(model = '3. Upsample', type = 'Testing')) %>%
 rename(`Data Set` = type) %>% mutate(Algorithm = '7. Ensemble Learning Mean')

## blended method
## no modifications
blend_train <- lr_train_preds %>% select(.pred_0,  dv_next_fall) %>%
 rename(log_reg = .pred_0) %>%
 cbind(rf_train_pred %>% rename(rand_for = .pred_0) %>% select(rand_for)) %>%
 cbind(xgb_train_pred %>% rename(xgb = .pred_0) %>% select(xgb))
org_stack <- stacks() %>% add_candidates(log_reg_cv_train) %>%
 add_candidates(rf_cv_train) %>% add_candidates(xgb_cv_train)
set.seed(7423)
org_stack_fit <- org_stack %>% blend_predictions() %>% fit_members()
blend_train_lr <- logistic_reg(penalty = (org_stack_fit$penalty)$penalty,
                          mixture = (org_stack_fit$penalty)$mixture) %>%
 set_engine('glm') %>% set_mode('classification') %>%
 fit(dv_next_fall ~ .,  data = blend_train)
blend_train %>% cbind(blend_train_lr %>% predict(blend_train, type = 'prob')) %>%
 cbind(blend_train_lr %>% predict(blend_train, type = 'class')) %>%
 conf_mat(dv_next_fall, .pred_class) %>% summary()
## testing data set
blend_test <- lr_test_preds %>% select(.pred_0, dv_next_fall) %>%
 rename(log_reg = .pred_0) %>%
 cbind(rf_test_pred %>% rename(rand_for = .pred_0) %>% select(rand_for)) %>%
 cbind(xgb_test_pred %>% rename(xgb = .pred_0) %>% select(xgb))

blend_test_lr <- logistic_reg(penalty = (org_stack_fit$penalty)$penalty,
                         mixture = (org_stack_fit$penalty)$mixture) %>%
 set_engine('glm') %>% set_mode('classification') %>%
 fit(dv_next_fall ~ ., data = blend_test)
blend_test %>% cbind(blend_test_lr %>% predict(blend_test, type = 'prob')) %>%
 cbind(blend_test_lr %>% predict(blend_test,  type = 'class')) %>%
 conf_mat(dv_next_fall, .pred_class) %>% summary()

## downsample modifications--training data set
blend_train_ds <- lr_train_preds_ds %>% select(.pred_0, dv_next_fall) %>%
 rename(log_reg = .pred_0) %>%
 cbind(rf_train_pred_ds %>% rename(rand_for = .pred_0) %>% select(rand_for)) %>%
 cbind(xgb_train_pred_ds %>% rename(xgb = .pred_0) %>% select(xgb))
org_stack_ds <- stacks() %>% add_candidates(log_reg_cv_train_ds) %>%
 add_candidates(rf_cv_train_ds) %>% add_candidates(xgb_cv_train_ds)
set.seed(7423)
org_stack_fit_ds <- org_stack %>% blend_predictions() %>% fit_members()
blend_train_lr_ds <- logistic_reg(penalty = (org_stack_fit_ds$penalty)$penalty,
                             mixture = (org_stack_fit_ds$penalty)$mixture) %>%
 set_engine('glm') %>% set_mode('classification') %>%
 fit(dv_next_fall ~ .,  data = blend_train_ds)
blend_train_ds %>%
 cbind(blend_train_lr_ds %>% predict(blend_train_ds, type = 'prob')) %>%
 cbind(blend_train_lr_ds %>% predict(blend_train_ds, type = 'class')) %>%
 conf_mat(dv_next_fall, .pred_class) %>% summary()
## testing data set
blend_test_ds <- lr_test_preds_ds %>% select(.pred_0, dv_next_fall) %>%
 rename(log_reg = .pred_0) %>%
 cbind(rf_test_pred_ds %>% rename(rand_for = .pred_0) %>% select(rand_for)) %>%
 cbind(xgb_test_pred_ds %>% rename(xgb = .pred_0) %>% select(xgb))
blend_test_lr_ds <- logistic_reg(penalty = (org_stack_fit_ds$penalty)$penalty,
                                  mixture = (org_stack_fit_ds$penalty)$mixture) %>%
 set_engine('glm') %>% set_mode('classification') %>%
 fit(dv_next_fall ~ .,  data = blend_test_ds)
blend_test_ds %>%
 cbind(blend_test_lr_ds %>% predict(blend_test_ds, type = 'prob')) %>%
 cbind(blend_test_lr_ds %>% predict(blend_test_ds, type = 'class')) %>%
 conf_mat(dv_next_fall,.pred_class) %>% summary()

## upsample modifications--training data set
blend_train_us <- lr_train_preds_us %>% select(.pred_0, dv_next_fall) %>%
 rename(log_reg = .pred_0) %>%
 cbind(rf_train_pred_us %>% rename(rand_for = .pred_0) %>% select(rand_for)) %>%
 cbind(xgb_train_pred_us %>% rename(xgb = .pred_0) %>% select(xgb))
org_stack_us <- stacks() %>% add_candidates(log_reg_cv_train_us) %>%
 add_candidates(rf_cv_train_us) %>% add_candidates(xgb_cv_train_us)
set.seed(7423)
org_stack_fit_us <- org_stack %>% blend_predictions() %>% fit_members()
blend_train_lr_us <- logistic_reg(penalty = (org_stack_fit_us$penalty)$penalty,
                                   mixture = (org_stack_fit_us$penalty)$mixture) %>%
 set_engine('glm') %>% set_mode('classification') %>%
 fit(dv_next_fall ~ .,  data = blend_train_us)
blend_train_us %>%
 cbind(blend_train_lr_us %>% predict(blend_train_us,  type = 'prob')) %>%
 cbind(blend_train_lr_us %>% predict(blend_train_us,  type = 'class')) %>%
 conf_mat(dv_next_fall,  .pred_class) %>% summary()
## testing data set
blend_test_us <- lr_test_preds_us %>% select(.pred_0, dv_next_fall) %>%
 rename(log_reg = .pred_0) %>%
 cbind(rf_test_pred_us %>% rename(rand_for = .pred_0) %>% select(rand_for)) %>%
 cbind(xgb_test_pred_us %>% rename(xgb = .pred_0) %>% select(xgb))
blend_test_lr_us <- logistic_reg(penalty = (org_stack_fit_us$penalty)$penalty,
                                  mixture = (org_stack_fit_us$penalty)$mixture) %>%
 set_engine('glm') %>% set_mode('classification') %>%
 fit(dv_next_fall ~ ., data = blend_test_us)
```

```
blend_test_us %>%
 cbind(blend_test_lr_us %>% predict(blend_test_us, type = 'prob')) %>%
 cbind(blend_test_lr_us %>% predict(blend_test_us, type = 'class')) %>%
 conf_mat(dv_next_fall, .pred_class) %>% summary()

## roc values
esl_blend_roc <- blend_train %>%
 cbind(blend_train_lr %>% predict(blend_train,  type = 'prob')) %>%
 roc_curve(truth = dv_next_fall, .pred_0) %>%
 mutate(model = '1. None', type = 'Training') %>%
 rbind(blend_test %>% cbind(blend_test_lr %>% predict(blend_test, type = 'prob')) %>%
       roc_curve(truth = dv_next_fall,  .pred_0) %>%
       mutate(model = '1. None',  type = 'Testing')) %>%
 rbind(blend_train_ds %>% cbind(blend_train_lr_ds %>%
             predict(blend_train_ds,  type = 'prob')) %>%
       roc_curve(truth = dv_next_fall,.pred_0) %>%
       mutate(model = '2. Downsample',  type = 'Training')) %>%
 rbind(blend_test_ds %>%
       cbind(blend_test_lr_ds %>% predict(blend_test_ds, type = 'prob')) %>%
       roc_curve(truth = dv_next_fall,  .pred_0) %>%
       mutate(model = '2. Downsample', type = 'Testing')) %>%
 rbind(blend_train_us %>%
       cbind(blend_train_lr_us %>% predict(blend_train_us, type = 'prob')) %>%
       roc_curve(truth = dv_next_fall,.pred_0) %>%
       mutate(model = '3. Upsample', type = 'Training')) %>%
 rbind(blend_test_us %>%
       cbind(blend_test_lr_us %>% predict(blend_test_us, type = 'prob')) %>%
       roc_curve(truth = dv_next_fall,  .pred_0) %>%
       mutate(model = '3. Upsample', type = 'Testing')) %>%
 rename(`Data Set` = type) %>% mutate(Algorithm = '8. Ensemble Learning Blended')
## training data set
blend_train %>% cbind(blend_train_lr %>% predict(blend_train, type = 'prob')) %>%
          roc_auc(truth = dv_next_fall, .pred_0)
blend_train_ds %>% cbind(blend_train_lr_ds %>% predict(blend_train_ds,
             type = 'prob')) %>% roc_auc(truth = dv_next_fall, .pred_0)
blend_train_us %>% cbind(blend_train_lr_us %>% predict(blend_train_us,
             type = 'prob')) %>% roc_auc(truth = dv_next_fall,.pred_0)
## testing data set
blend_test %>% cbind(blend_test_lr %>% predict(blend_test,  type = 'prob')) %>%
     roc_auc(truth = dv_next_fall,  .pred_0)
blend_test_ds %>% cbind(blend_test_lr_ds %>% predict(blend_test_ds,
             type = 'prob')) %>% roc_auc(truth = dv_next_fall,  .pred_0)
blend_test_us %>% cbind(blend_test_lr_us %>% predict(blend_test_us,
             type = 'prob')) %>% roc_auc(truth = dv_next_fall,  .pred_0)
```

```
#################################
##ROC AND AUC COMPARISONS##
#################################

## roc graphs
combined_roc <- log_roc %>%  rbind(rf_roc) %>% rbind(xgb_roc) %>%
  rbind(svml_roc) %>% rbind(svmp_roc) %>% rbind(svmr_roc) %>%
  rbind(esl_mean_roc) %>% rbind(esl_blend_roc)
 combined_roc%>%
  ## remove comment tags to change between modifications
  ## filter(model == '1. None') %>%
  ## filter(model == '3. Downsample') %>%
  filter(model == '3. Upsample') %>%
  ggplot(aes(x = 1- specificity, y = sensitivity,  color = `Data Set`)) +
  geom_abline(slope = 1, color = "gray50", lty = 2, alpha = 0.8) +
  geom_path() + ##size = 1.5, alpha = 0.7) +
  theme_classic() + theme(legend.position = 'top', text = element_text(size = 15)) +
  facet_wrap(. ~ Algorithm,  ncol = 3) +  coord_fixed()

 ## no modifications
 auc_retain <- log_reg_cv_train[[3]] %>% as.data.frame() %>%
  filter(.metric == 'roc_auc') %>% gather(var_type, train) %>%
  filter(var_type %like% '%estimate%') %>%
  mutate(fold = 1:10,  model = '1. Logistic Regression') %>%
  select(model, fold, train) %>%
  left_join(log_reg_cv_test[[3]]  %>% as.data.frame() %>%
            filter(.metric == 'roc_auc') %>% gather(var_type,  test) %>%
            filter(var_type %like% '%estimate%') %>%
            mutate(fold = 1:10,  model = '1. Logistic Regression')  %>%
            select(model, fold, test)) %>%
  rbind(svml_cv_train[[3]] %>% as.data.frame() %>%
        filter(.metric == 'roc_auc') %>% gather(var_type, train) %>%
        filter(var_type %like% '%estimate%') %>%
        mutate(fold = 1:10, model = '2. SVM Linear') %>%
        select(model, fold, train) %>%
        left_join(svml_cv_test[[3]]  %>% as.data.frame() %>%
                  filter(.metric == 'roc_auc') %>% gather(var_type, test) %>%
                  filter(var_type %like% '%estimate%') %>%
                  mutate(fold = 1:10, model = '2. SVM Linear')  %>%
                  select(model, fold, test))) %>%
  rbind(svmp_cv_train[[3]] %>% as.data.frame() %>%
        filter(.metric == 'roc_auc') %>% gather(var_type,  train) %>%
        filter(var_type %like% '%estimate%') %>%
        mutate(fold = 1:10,  model = '3. SVM Polynomial') %>%
        select(model, fold, train) %>%
        left_join(svmp_cv_test[[3]]  %>% as.data.frame() %>%
              filter(.metric == 'roc_auc') %>% gather(var_type, test) %>%
              filter(var_type %like% '%estimate%') %>%
              mutate(fold = 1:10, model = '3. SVM Polynomial')  %>%
              select(model, fold, test))) %>%
 rbind(svmr_cv_train[[3]] %>% as.data.frame() %>%
       filter(.metric == 'roc_auc') %>% gather(var_type, train) %>%
       filter(var_type %like% '%estimate%') %>%
       mutate(fold = 1:10, model = '4. SVM Radial') %>%
       select(model, fold, train) %>%
       left_join(svmr_cv_test[[3]]  %>% as.data.frame() %>%
              filter(.metric == 'roc_auc') %>% gather(var_type, test) %>%
              filter(var_type %like% '%estimate%') %>%
              mutate(fold = 1:10,  model = '4. SVM Radial')  %>%
              select(model, fold, test))) %>%
 rbind(rf_cv_train[[3]] %>% as.data.frame() %>% filter(.metric == 'roc_auc') %>%
       gather(var_type, train) %>%
       filter(var_type %like% '%estimate%') %>%
       mutate(fold = 1:10, model = '5. Random Forest') %>%
       select(model, fold, train) %>%
       left_join(rf_cv_test[[3]]  %>% as.data.frame() %>%
              filter(.metric == 'roc_auc') %>% gather(var_type, test) %>%
              filter(var_type %like% '%estimate%') %>%
              mutate(fold = 1:10, model = '5. Random Forest')  %>%
              select(model, fold, test))) %>%
 rbind(xgb_cv_train[[3]] %>% as.data.frame() %>%
       filter(.metric == 'roc_auc') %>% gather(var_type, train) %>%
       filter(var_type %like% '%estimate%') %>%
       mutate(fold = 1:10, model = '6. XGBoost') %>%
       select(model, fold, train) %>%
       left_join(xgb_cv_test[[3]]  %>%as.data.frame() %>%
              filter(.metric == 'roc_auc') %>% gather(var_type,  test) %>%
              filter(var_type %like% '%estimate%') %>%
              mutate(fold = 1:10, model = '6. XGBoost')  %>%
              select(model, fold, test)))
## auc distributions
auc_retain %>% gather(data_set, values, -model, -fold) %>%
 mutate(values = as.numeric(values),
       data_set = case_when(data_set == 'test' ~ '2. Testing', TRUE ~ '1. Training')) %>%
 rename(`Data Set` = data_set) %>%
 ggplot(aes(x = reorder(model, desc(model)), y = values, fill = `Data Set`)) +
 geom_boxplot() + theme_classic() +
 theme(legend.position = 'none',  text = element_text(size = 15),
      axis.title = element_blank()) + coord_flip() + facet_wrap(. ~ `Data Set`)
```

```
## inferential statistics on algorithms no modifications
 ## wilcox test between the training and testing dataset
 wilcox.test(as.numeric((auc_retain %>%filter(model == '1. Logistic Regression'))$train),
          as.numeric((auc_retain %>% filter(model == '1. Logistic Regression'))$test),
          paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
          conf.level = 0.95)
 wilcox.test(as.numeric((auc_retain %>% filter(model == '2. SVM Linear'))$train),
          as.numeric((auc_retain %>% filter(model == '2. SVM Linear'))$test),
          paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
          conf.level = 0.95)
  wilcox.test(as.numeric((auc_retain %>% filter(model == '3. SVM Polynomial'))$train),
          as.numeric((auc_retain %>% filter(model == '3. SVM Polynomial'))$test),
          paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
          conf.level = 0.95)
  wilcox.test(as.numeric((auc_retain %>% filter(model == '4. SVM Radial'))$train),
          as.numeric((auc_retain %>% filter(model == '4. SVM Radial'))$test),
          paired = FALSE,  exact = TRUE, correct = TRUE, conf.int = TRUE,
          conf.level = 0.95)
  wilcox.test(as.numeric((auc_retain %>% filter(model == '5. Random Forest'))$train),
          as.numeric((auc_retain %>% filter(model == '5. Random Forest'))$test),
          paired = FALSE,  exact = TRUE, correct = TRUE, conf.int = TRUE,
          conf.level = 0.95)
  wilcox.test(as.numeric((auc_retain %>% filter(model == '6. XGBoost'))$train),
          as.numeric((auc_retain %>% filter(model == '6. XGBoost'))$test),
          paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
          conf.level = 0.95)

## friedmen test of the best model on no modifications
## training data set
 auc_retain %>% friedman_test(train ~ model|fold)
 auc_retain %>% friedman_effsize(train ~ model|fold)
 auc_retain %>%   mutate(train = as.numeric(train)) %>%
         wilcox_test(train ~ model, paired = TRUE, p.adjust.method = 'bonferroni')
 ## testing data set
 auc_retain %>% friedman_test(test ~ model|fold)
 auc_retain %>% friedman_effsize(test ~ model|fold)
 auc_retain %>%  mutate(test = as.numeric(test)) %>%
 wilcox_test(test ~ model, paired = TRUE, p.adjust.method = 'bonferroni')
## median auc values
 auc_retain %>%  select(-fold) %>%
  mutate(train = as.numeric(train), test = as.numeric(test)) %>%
  group_by(model) %>% summarise(train = median(train),
         test = median(test),.groups = 'drop')

 ## downsample modifications
 auc_retain_us <- log_reg_cv_train_ds[[3]] %>% as.data.frame() %>%
filter(.metric == 'roc_auc') %>% gather(var_type,  train) %>%
filter(var_type %like% '%estimate%') %>%
mutate(fold = 1:10,  model = '1. Logistic Regression') %>%
select(model, fold, train) %>%
left_join(log_reg_cv_test_ds[[3]]  %>% as.data.frame() %>%
      filter(.metric == 'roc_auc') %>% gather(var_type, test) %>%
      filter(var_type %like% '%estimate%') %>%
      mutate(fold = 1:10, model = '1. Logistic Regression')  %>%
      select(model, fold, test)) %>%
rbind(svml_cv_train_ds[[3]] %>% as.data.frame() %>%
    filter(.metric == 'roc_auc') %>% gather(var_type, train) %>%
    filter(var_type %like% '%estimate%') %>%
    mutate(fold = 1:10, model = '2. SVM Linear') %>%
    select(model, fold, train) %>%
    left_join(svml_cv_test_ds[[3]]  %>% as.data.frame() %>%
          filter(.metric == 'roc_auc') %>% gather(var_type, test) %>%
          filter(var_type %like% '%estimate%') %>%
          mutate(fold = 1:10, model = '2. SVM Linear')  %>%
          select(model, fold, test))) %>%
rbind(svmp_cv_train_ds[[3]] %>% as.data.frame() %>%
    filter(.metric == 'roc_auc') %>% gather(var_type, train) %>%
    filter(var_type %like% '%estimate%') %>%
    mutate(fold = 1:10, model = '3. SVM Polynomial') %>%
    select(model, fold, train) %>%
    left_join(svmp_cv_test_ds[[3]]  %>% as.data.frame() %>%
          filter(.metric == 'roc_auc') %>% gather(var_type, test) %>%
          filter(var_type %like% '%estimate%') %>%
          mutate(fold = 1:10,  model = '3. SVM Polynomial')  %>%
          select(model, fold, test))) %>%
rbind(svmr_cv_train_ds[[3]] %>% as.data.frame() %>%
    filter(.metric == 'roc_auc') %>% gather(var_type, train) %>%
    filter(var_type %like% '%estimate%') %>%
    mutate(fold = 1:10, model = '4. SVM Radial') %>%
    select(model, fold, train) %>%
    left_join(svmr_cv_test_ds[[3]]  %>% as.data.frame() %>%
          filter(.metric == 'roc_auc') %>% gather(var_type, test) %>%
          filter(var_type %like% '%estimate%') %>%
          mutate(fold = 1:10, model = '4. SVM Radial')  %>%
          select(model, fold, test))) %>%
rbind(rf_cv_train_ds[[3]] %>% as.data.frame() %>%
    filter(.metric == 'roc_auc') %>% gather(var_type, train) %>%
    filter(var_type %like% '%estimate%') %>%
    mutate(fold = 1:10,  model = '5. Random Forest') %>%
    select(model, fold, train) %>%
    left_join(rf_cv_test_ds[[3]]  %>% as.data.frame() %>%
          filter(.metric == 'roc_auc') %>% gather(var_type, test) %>%
              filter(var_type %like% '%estimate%') %>%
              mutate(fold = 1:10, model = '5. Random Forest')  %>%
              select(model, fold, test))) %>%
  rbind(xgb_cv_train_ds[[3]] %>% as.data.frame() %>%
        filter(.metric == 'roc_auc') %>% gather(var_type, train) %>%
        filter(var_type %like% '%estimate%') %>%
        mutate(fold = 1:10, model = '6. XGBoost') %>%
        select(model, fold, train) %>%
        left_join(xgb_cv_test_ds[[3]]  %>% as.data.frame() %>%
              filter(.metric == 'roc_auc') %>% gather(var_type,  test) %>%
              filter(var_type %like% '%estimate%') %>%
              mutate(fold = 1:10,  model = '6. XGBoost')  %>%
              select(model, fold, test)))
  ## auc distributions
 auc_retain_ds %>% gather(data_set, values, -model, -fold) %>%
  mutate(values = as.numeric(values),
        data_set = case_when(data_set == 'test' ~ '2. Testing',  TRUE ~ '1. Training')) %>%
  rename(`Data Set` = data_set) %>%
  ggplot(aes(x = reorder(model, desc(model)), y = values, fill = `Data Set`)) +
  geom_boxplot() + theme_classic() +
  theme(legend.position = 'none',  text = element_text(size = 15),
       axis.title = element_blank()) + coord_flip() + facet_wrap(. ~ `Data Set`)

## inferential statistics on algorithms on downsample modifications
## wilcox test between the training and testing dataset
 wilcox.test(as.numeric((auc_retain_ds %>%
                 filter(model == '1. Logistic Regression'))$train),
         as.numeric((auc_retain_ds %>%filter(model == '1. Logistic Regression'))$test),
         paired = FALSE,  exact = TRUE, correct = TRUE, conf.int = TRUE,
         conf.level = 0.95)
 wilcox.test(as.numeric((auc_retain_ds %>% filter(model == '2. SVM Linear'))$train),
         as.numeric((auc_retain_ds %>% filter(model == '2. SVM Linear'))$test),
         paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
         conf.level = 0.95)
 wilcox.test(as.numeric((auc_retain_ds%>%filter(model== '3. SVM Polynomial'))$train),
         as.numeric((auc_retain_ds %>% filter(model == '3. SVM Polynomial'))$test),
         paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
         conf.level = 0.95)
 wilcox.test(as.numeric((auc_retain_ds %>% filter(model == '4. SVM Radial'))$train),
         as.numeric((auc_retain_ds %>% filter(model == '4. SVM Radial'))$test),
         paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
         conf.level = 0.95)
 wilcox.test(as.numeric((auc_retain_ds %>% filter(model == '5. Random Forest'))$train),
         as.numeric((auc_retain_ds %>% filter(model == '5. Random Forest'))$test),
         paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
         conf.level = 0.95)
 wilcox.test(as.numeric((auc_retain_ds %>% filter(model == '6. XGBoost'))$train),
            as.numeric((auc_retain_ds %>% filter(model == '6. XGBoost'))$test),
            paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
            conf.level = 0.95)

## friedmen test of the best model on downsample modifications
## training data set
auc_retain_ds %>% friedman_test(train ~ model|fold)
auc_retain_ds %>% friedman_effsize(train ~ model|fold)
auc_retain_ds %>%  mutate(train = as.numeric(train)) %>%
  wilcox_test(train ~ model, paired = TRUE, p.adjust.method = 'bonferroni')
## testing data set
auc_retain_ds %>% friedman_test(test ~ model|fold)
auc_retain_ds %>% friedman_effsize(test ~ model|fold)
auc_retain_ds %>%  mutate(test = as.numeric(test)) %>%
  wilcox_test(test ~ model, paired = TRUE, p.adjust.method = 'bonferroni')
## median auc values
auc_retain_ds %>% select(-fold) %>%
  mutate(train = as.numeric(train), test = as.numeric(test)) %>%
  group_by(model) %>% summarise(train = median(train),
           test = median(test),.groups = 'drop')

 ## upsample modifications
 auc_retain_us <- log_reg_cv_train_us[[3]] %>% as.data.frame() %>%
  filter(.metric == 'roc_auc') %>% gather(var_type, train) %>%
  filter(var_type %like% '%estimate%') %>%
  mutate(fold = 1:10, model = '1. Logistic Regression') %>%
  select(model, fold, train) %>%
  left_join(log_reg_cv_test_us[[3]]  %>% as.data.frame() %>%
            filter(.metric == 'roc_auc') %>% gather(var_type,  test) %>%
            filter(var_type %like% '%estimate%') %>%
            mutate(fold = 1:10, model = '1. Logistic Regression')  %>%
            select(model, fold, test)) %>%
  rbind(svml_cv_train_us[[3]] %>% as.data.frame() %>%
        filter(.metric == 'roc_auc') %>% gather(var_type, train) %>%
        filter(var_type %like% '%estimate%') %>%
        mutate(fold = 1:10, model = '2. SVM Linear') %>%
        select(model, fold, train) %>%
        left_join(svml_cv_test_us[[3]]  %>% as.data.frame() %>%
                  filter(.metric == 'roc_auc') %>% gather(var_type, test) %>%
                  filter(var_type %like% '%estimate%') %>%
                  mutate(fold = 1:10, model = '2. SVM Linear')  %>%
                  select(model, fold, test))) %>%
  rbind(svmp_cv_train_us[[3]] %>% as.data.frame() %>%
        filter(.metric == 'roc_auc') %>% gather(var_type, train) %>%
        filter(var_type %like% '%estimate%') %>%
        mutate(fold = 1:10, model = '3. SVM Polynomial') %>%
        select(model, fold, train) %>%
        left_join(svmp_cv_test_us[[3]]  %>% as.data.frame() %>%
                  filter(.metric == 'roc_auc') %>% gather(var_type, test) %>%
                  filter(var_type %like% '%estimate%') %>%
                  mutate(fold = 1:10, model = '3. SVM Polynomial')  %>%
                  select(model, fold, test))) %>%
  rbind(svmr_cv_train_us[[3]] %>% as.data.frame() %>%
        filter(.metric == 'roc_auc') %>% gather(var_type,  train) %>%
        filter(var_type %like% '%estimate%') %>%
        mutate(fold = 1:10,  model = '4. SVM Radial') %>%
        select(model, fold, train) %>%
        left_join(svmr_cv_test_us[[3]]  %>% as.data.frame() %>%
                  filter(.metric == 'roc_auc') %>% gather(var_type, test) %>%
                  filter(var_type %like% '%estimate%') %>%
                  mutate(fold = 1:10, model = '4. SVM Radial')  %>%
                  select(model, fold, test))) %>%
  rbind(rf_cv_train_us[[3]] %>% as.data.frame() %>%
        filter(.metric == 'roc_auc') %>% gather(var_type, train) %>%
        filter(var_type %like% '%estimate%') %>%
        mutate(fold = 1:10, model = '5. Random Forest') %>%
        select(model, fold, train) %>%
        left_join(rf_cv_test_us[[3]]  %>% as.data.frame() %>%
                  filter(.metric == 'roc_auc') %>% gather(var_type, test) %>%
                  filter(var_type %like% '%estimate%') %>%
                  mutate(fold = 1:10, model = '5. Random Forest')  %>%
                  select(model, fold, test))) %>%
  rbind(xgb_cv_train_us[[3]] %>% as.data.frame() %>%
        filter(.metric == 'roc_auc') %>% gather(var_type, train) %>%
        filter(var_type %like% '%estimate%') %>%
        mutate(fold = 1:10, model = '6. XGBoost') %>%
        select(model, fold, train) %>%
        left_join(xgb_cv_test_us[[3]]  %>% as.data.frame() %>%
                  filter(.metric == 'roc_auc') %>% gather(var_type, test) %>%
                  filter(var_type %like% '%estimate%') %>%
                  mutate(fold = 1:10, model = '6. XGBoost')  %>%
                  select(model, fold, test)))
 ## auc distribution
 auc_retain_us %>% gather(data_set, values, -model, -fold) %>%
  mutate(values = as.numeric(values),
        data_set = case_when(data_set == 'test' ~ '2. Testing',  TRUE ~ '1. Training')) %>%
  rename(`Data Set` = data_set) %>%
  ggplot(aes(x = reorder(model, desc(model)), y = values, fill = `Data Set`)) +
  geom_boxplot() + theme_classic() +
  theme(legend.position = 'none',  text = element_text(size = 15),
     axis.title = element_blank()) + coord_flip() + facet_wrap(. ~ `Data Set`)

 ## inferential statistics on algorithms on upsample modifications
 ## wilcox test between the training and testing dataset
 wilcox.test(as.numeric((auc_retain_us %>%
                 filter(model == '1. Logistic Regression'))$train),
         as.numeric((auc_retain_us %>%filter(model == '1. Logistic Regression'))$test),
         paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
         conf.level = 0.95)
 wilcox.test(as.numeric((auc_retain_us %>% filter(model == '2. SVM Linear'))$train),
         as.numeric((auc_retain_us %>% filter(model == '2. SVM Linear'))$test),
         paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
         conf.level = 0.95)
 wilcox.test(as.numeric((auc_retain_us%>%filter(model== '3. SVM Polynomial'))$train),
         as.numeric((auc_retain_us %>% filter(model == '3. SVM Polynomial'))$test),
         paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
         conf.level = 0.95)
 wilcox.test(as.numeric((auc_retain_us %>%filter(model == '4. SVM Radial'))$train),
         as.numeric((auc_retain_us %>%filter(model == '4. SVM Radial'))$test),
         paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
         conf.level = 0.95)
 wilcox.test(as.numeric((auc_retain_us %>%filter(model == '5. Random Forest'))$train),
         as.numeric((auc_retain_us %>% filter(model == '5. Random Forest'))$test),
         paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
         conf.level = 0.95)
 wilcox.test(as.numeric((auc_retain_us %>%filter(model == '6. XGBoost'))$train),
         as.numeric((auc_retain_us %>% filter(model == '6. XGBoost'))$test),
         paired = FALSE, exact = TRUE, correct = TRUE, conf.int = TRUE,
         conf.level = 0.95)

 ## friedmen test of the best model on upsample modifications
 ## training data set
 auc_retain_us %>% friedman_test(train ~ model|fold)
 auc_retain_us %>% friedman_effsize(train ~ model|fold)
 auc_retain_us %>% mutate(train = as.numeric(train)) %>%
   wilcox_test(train ~ model, paired = TRUE, p.adjust.method = 'bonferroni')
 ## testing data set
 auc_retain_us %>% friedman_test(test ~ model|fold)
 auc_retain_us %>% friedman_effsize(test ~ model|fold)
 auc_retain_us %>%  mutate(test = as.numeric(test)) %>%
   wilcox_test(test ~ model,  paired = TRUE, p.adjust.method = 'bonferroni')
 ## median auc values
 auc_retain_us %>% select(-fold) %>%
   mutate(train = as.numeric(train),  test = as.numeric(test)) %>%
   group_by(model) %>%  summarise(train = median(train),
         test = median(test), groups = 'drop')
```

APPENDIX E:

Considerations and Assumptions Review R Code

```
########################################
## Considerations and Assumptions Review ##
########################################

## libraries utilized
library(tidyverse); library(tidymodels); library(themis); library(blorr)
library(DAAG); library(lsr); library(regclass); library(car); library(DescTools);
library(psych); library(vip); library(pdp); library(doParallel); library(xgboost)
library(kernlab); library(stacks); library(rstatix)

select <- dplyr::select

## setting working directory
setwd(' C:/Users/bdfitzgerald/Desktop/Dissertation/')

## data clean up source files
## USG Data
source('./dissertation_scripts/01.0 USG Data Clean Up.R')
## CCRPI Data
source('./dissertation_scripts/02.0 CCRPI Data Clean Up.R')
## EOC Data
source('./dissertation_scripts/03.0 EOC Data Clean Up.R')
## IPEDS Data
source('./dissertation_scripts/04.0 IPEDS Data Clean Up.R')

######################
## COMBING DATA   ##
######################

## USG data to IPEDS expenditures
dat <- recent_ga_public_hs %>%mutate(fy = as.integer(substr(cohort_term, 1, 4))) %>%
 left_join(ipeds.clean %>%
          rename(enrollment_institution_name = Institution.Name)) %>%
 select(-fy, -UnitID)

## ga public high schools data
## EOC and CCRPI to the distinct high schools represented
## in the four RCUs
hs_curriculum <- dat %>% select(hs_code, hs_grad_year) %>%
 filter(!duplicated(paste0(hs_grad_year, hs_code))) %>%
 left_join(ga_hs) %>%
 left_join(ccrpi %>% rename(state_school_id = school_code,
                hs_grad_year = school_year) %>%
          mutate(content_mastery = content_mastery / 100,
                readiness = readiness / 100) %>%
          select(hs_grad_year, content_mastery, readiness, state_school_id)) %>%
 left_join(eoc_prep %>% select(-school_dstrct_nm, -instn_name) %>%
           mutate(year = as.integer(year)) %>%
           rename(hs_grad_year = year, state_school_id = school_code_rv))

## removing unnecessary objects
rm(ccrpi, eoc_prep, ga_hs, ipeds.clean, recent_ga_public_hs)

## modifying USG data
dat <- dat %>%
 mutate(dv_next_fall = case_when(dv_next_fall == 1 ~ 0, TRUE ~ 1),
        race_eth = case_when(ipeds_race_ethnicity_descr %in%
                             c('White',  'Black or African American',
                               'Hispanic or Latino') ~ ipeds_race_ethnicity_descr, TRUE ~ 'Other'),
        unique_identifer = paste(uniqueid, cohort_term, enrollment_institution_name,
                                 setid_consol, sep = '.')) %>%
 select(-ipeds_race_ethnicity_descr, -uniqueid, -cohort_term,
        -enrollment_institution_name, -setid_consol)

##########################
## DATA PARTITIONING ##
##########################

set.seed(51823)
dat_split <- initial_split(data = dat, prop = .6)
dat_train <- training(dat_split); ## dat_test <- testing(dat_split)

####################################
## REVIEW OF MISSING DATA FOR  ##
## HS CURRICULUM VARIABLES      ##
####################################

## summary of the high school
hs_curriculum %>% summary()

## plot of missing data points
## by high school year
hs_curriculum %>% select(-hs_code,  -state_school_id) %>%
 gather(var_name,  var_results, -hs_grad_year) %>%
 filter(is.na(var_results)) %>% select(-var_results) %>% group_by_all() %>%
 summarise(n.missing = n(), .groups = 'drop') %>%
 mutate(hs_grad_year = as.character(hs_grad_year),
       var_name = case_when(var_name == 'content_mastery' ~ 'CCRPI: Content Mastery',
                            var_name == 'readiness' ~ 'CCRPI: Readiness',
                            var_name == 'english' ~ 'EOC: English',
                            var_name == 'math' ~ 'EOC: Mathematics',
                            var_name == 'science' ~ 'EOC: Science',
                  var_name == 'social_studies' ~ 'EOC: Social Studies',
                  TRUE ~ 'OTHER')) %>%
 rename(Year = hs_grad_year) %>%
 ggplot() + geom_bar(aes(x = var_name, y = n.missing, fill = Year), stat = 'identity',
       position = 'dodge') + ylab('Number') + theme_classic() +
 theme(plot.title = element_text(hjust = .5),  axis.title.x = element_blank(),
     text = element_text(size = 13), legend.position = 'top')

## baseline
hs_curriculum %>% select(-hs_code, -hs_grad_year, -state_school_id,
      -locale_code, -locale) %>%summary()
## with zero
hs_curriculum %>% select(-hs_code, -hs_grad_year, -state_school_id,
      -locale_code, -locale) %>% mutate_if(is.numeric, ~ replace_na(., 0)) %>%
 summary()
## with mean
hs_curriculum %>% select(-hs_code, -hs_grad_year, -state_school_id,
      -locale_code, -locale) %>%
 mutate_if(is.numeric, ~ replace_na(., mean(., na.rm = TRUE))) %>%
 summary()
## with median
hs_curriculum %>% select(-hs_code, -state_school_id, -hs_grad_year,
      -locale_code, -locale) %>%
 mutate_if(is.numeric, ~ replace_na(., median(., na.rm = TRUE))) %>%
 summary()

## selecting the median imputation
hs_curriculum <- hs_curriculum %>%
 mutate_if(is.numeric, ~ replace_na(., median(., na.rm = TRUE))) %>%
mutate(locale_group = case_when(locale_code %in% c('11', '12', '13') ~ 'City',
                       locale_code %in% c('21', '22', '23') ~ 'Suburb',
                       locale_code %in% c('31', '32', '33') ~ 'Town',  TRUE ~ 'Rural'))

## joining hs curriculum data with the USG data
dat_train_rv <- dat_train %>% left_join(hs_curriculum) %>%
 select(-hs_code, -hs_grad_year, -state_school_id, -locale_code, -locale)

rm(hs_curriculum)

########################################
## INITIAL RECIPE TO EXPLORE DATA ##
## TRAINING DATA SET ONLY              ##
#########################################

retain_train <- recipe(formula = dv_next_fall ~ ., data = dat_train_rv) %>%
 update_role(unique_identifer, new_role = 'id variable')
```

```
## data to explore
retain_exp <- juice(prep(retain_train))

#######################
## CONSIDERATIONS ##
#######################

###################################
## MISSING DATA OBSERVATIONS ##
###################################

## bar chart of missing observations
retain_exp %>% select(-unique_identifer) %>%
 rename(`First-fall GPA` = dv_first_fall_gpa, `First-year GPA` = dv_first_yr_gpa,
       Gender = gender_descr, `First Generation Status` = admit_first_gen_ind,
       `HS GPA` = hs_gpa, `AP Hours` = adv_standing_ap_hrs,
       `CLEP Hours` = adv_standing_clep_hrs, `IB Hours` = adv_standing_ib_hrs,
       `Other Hours` = adv_standing_other_hrs, `CPC English` = cpc_english_code,
       `CPC Foreign Language` = cpc_foreign_language_code,
       `CPC Math` = cpc_math_code, `CPC Science` = cpc_science_code,
       `CPC Social Sciences` = cpc_social_science_code,
       `Major Groupings` = cip_categories, `EFC` = expected_family_contribution,
       `GA HOPE Scholarship` = ga_hope, `Zell Miller Indicator` = zell_ind,
       `PELL Grant` = pell, `Federal Sub. Loans` = fed_sub_loans,
       `Federal Unsub. Loans` = fed_unsub_loans, `Other Loans` = oth_loans,
       `Admissions Test Scores` = adm_test_score, `Academic Support` = acay_sup_exp,
       `All Other` = all_other_exp, `Institutional Support` = inst_sup_exp,
       `Instruction` = instr_exp, `Public Service` = public_serv_exp,
       `Research` = rsch_exp, `Student Services Support` = stu_serv_exp,
       `Race/Ethnicity` = race_eth, `CCRPI Content Mastery` = content_mastery,
       `CCPRI Readiness` = readiness, `EOC English` = english,
       `EOC Math` = math, `EOC Science` = science,
       `EOC Socail Studies` = social_studies, `HS Locale` = locale_group,
       Retained = dv_next_fall) %>%
 gather(var_name, var_results) %>% filter(is.na(var_results)) %>%
 select(-var_results) %>% group_by_all() %>%
 summarise(hc = n(),  .groups = 'drop') %>% ggplot() +
 geom_bar(aes(x = reorder(var_name,  desc(var_name)), y = hc), stat = 'identity') +
 xlab('Data Variables') + ylab('Number Missing') +
 theme_classic() + theme(text = element_text(size = 15)) + coord_flip()

## recipe update
retain_train <- retain_train  %>%
step_mutate_at(c(adv_standing_ap_hrs:adv_standing_other_hrs, ga_hope,
             pell:oth_loans), fn = ~ replace_na(., 0)) %>%
step_mutate_at(zell_ind, fn = ~ replace_na(., 'N')) %>%
step_unknown(c(cpc_english_code:cpc_social_science_code), new_level = 'U')  %>%
step_mutate(hsgpa_mean = hs_gpa, hsgpa_median = hs_gpa, hsgpa_knn = hs_gpa,
            ats_mean = adm_test_score, ats_median = adm_test_score,
            ats_knn = adm_test_score, efc_mean = expected_family_contribution,
            efc_median = expected_family_contribution,  efc_knn =
expected_family_contribution) %>%
 step_impute_mean(hsgpa_mean, ats_mean, efc_mean) %>%
 step_impute_median(hsgpa_median, ats_median, efc_median) %>%
 step_impute_knn(c(hsgpa_knn, ats_knn, efc_knn), neighbors = 10)

## data to explore
retain_exp <- juice(prep(retain_train))

## examining the imputation impact on HS GPA
retain_exp %>% select(hs_gpa, hsgpa_mean, hsgpa_median, hsgpa_knn) %>%
 summary()
(retain_exp %>% select(hs_gpa, hsgpa_mean, hsgpa_median, hsgpa_knn) %>%
  describe()) %>% select(skew, kurtosis)

## examining the imputation impact on admission test scores
retain_exp %>% select(adm_test_score, ats_mean, ats_median, ats_knn) %>%
 summary()
(retain_exp %>% select(adm_test_score, ats_mean, ats_median, ats_knn) %>%
  describe()) %>%select(skew, kurtosis)

## examining the imputation impact on expected family contributions
retain_exp %>% select(expected_family_contribution, efc_mean, efc_median,
           efc_knn) %>%summary()
(retain_exp %>%
  select(expected_family_contribution, efc_mean, efc_median, efc_knn) %>%
  describe()) %>% select(skew, kurtosis)

## update recipe
## removing exploratory missing data variables (original, mean, median)
retain_train <- retain_train %>%
step_rm(hs_gpa, hsgpa_mean,hsgpa_median, adm_test_score,ats_mean,
        ats_median, expected_family_contribution, efc_mean, efc_median)
## data to explore
retain_exp <- juice(prep(retain_train))

######################
## OUTLIER REVIEW ##
######################

## outliers by z-scores
retain_train_zs <- retain_train %>%
 step_normalize(c(adv_standing_ap_hrs:adv_standing_other_hrs,
                 ga_hope, pell:stu_serv_exp, content_mastery:social_studies,
                 hsgpa_knn, ats_knn, efc_knn))
retain_exp_zs <- juice(prep(retain_train_zs))

## review of z-scores distributions
retain_exp_zs %>%
 select(c(adv_standing_ap_hrs:adv_standing_other_hrs, ga_hope, pell:stu_serv_exp,
         content_mastery:social_studies, hsgpa_knn, ats_knn, efc_knn)) %>%
 rename(`HS GPA` = hsgpa_knn, `AP Hours` = adv_standing_ap_hrs,
       `CLEP Hours` = adv_standing_clep_hrs, `IB Hours` = adv_standing_ib_hrs,
       `Other Hours` = adv_standing_other_hrs, `EFC` = efc_knn,
       `GA HOPE Scholarship` = ga_hope, `PELL Grant` = pell,
       `Federal Sub. Loans` = fed_sub_loans, `Federal Unsub. Loans` = fed_unsub_loans,
       `Other Loans` = oth_loans, `Admissions Test Scores` = ats_knn,
       `Academic Support` = acay_sup_exp, `All Other` = all_other_exp,
       `Institutional Support` = inst_sup_exp, `Instruction` = instr_exp,
       `Public Service` = public_serv_exp, `Research` = rsch_exp,
       `Student Services Support` = stu_serv_exp,
       `CCRPI Content Mastery` = content_mastery,
       `CCPRI Readiness` = readiness, `EOC English` = english,
       `EOC Math` = math, `EOC Science` = science,
       `EOC Social Studies` = social_studies) %>%
 gather(var_name, var_results) %>% ggplot(aes(x = var_results)) +
 geom_density(aes(x = var_results), fill = 'blue', alpha = 0.25) +
 geom_histogram(aes(y = ..density..), fill = 'NA', color = 'black') +
 geom_vline(xintercept = -3,  color = 'red', linewidth = 0.5,  linetype = 'dashed') +
 geom_vline(xintercept = 3,  color = 'red', linewidth = 0.5,  linetype = 'dashed') +
 xlab('z-scores') + ggtitle('Examination of Z-Scores') + theme_classic() +
 theme(plot.title = element_text(hjust = 0.5),  text = element_text(size = 15)) +
 facet_wrap(var_name ~ ., scales = 'free',  ncol = 4)

vars_outlier_test <- retain_exp %>%
 select(c(adv_standing_ap_hrs:adv_standing_other_hrs, ga_hope, pell:stu_serv_exp,
         content_mastery:social_studies, hsgpa_knn,ats_knn, efc_knn))

## grubb's test for univariate outliers
grubbs.test(vars_outlier_test$adv_standing_ap_hrs, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$adv_standing_clep_hrs, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$adv_standing_ib_hrs,  type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$adv_standing_other_hrs, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$efc_knn, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$ga_hope, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$pell, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$fed_sub_loans, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$fed_unsub_loans, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$oth_loans, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$acay_sup_exp, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$all_other_exp, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$inst_sup_exp, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$instr_exp, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$public_serv_exp, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$rsch_exp, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$stu_serv_exp, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$content_mastery, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$readiness, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$english, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$math, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$science, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$social_studies, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$hsgpa_knn, type = 11, opposite = TRUE)
grubbs.test(vars_outlier_test$ats_knn, type = 11, opposite = TRUE)

## multivariate outliers
vars_outlier_test$mahalanobis <- mahalanobis(x = vars_outlier_test,
                              colMeans(vars_outlier_test),  cov(vars_outlier_test))
vars_outlier_test$p_value <- pchisq(vars_outlier_test$mahalanobis,
                        df = 24, lower.tail = FALSE)

## distribution of p-values
vars_outlier_test %>% select(p_value) %>%
 mutate(p_value = round(p_value, 3)) %>%
 ggplot(aes(x = p_value)) + geom_density(aes(x = p_value),  fill = 'blue', alpha = .25) +
 geom_histogram(aes(y = ..density..),  fill = 'NA', color = 'black') +
 theme_classic() + ylab('Density of p-values') + xlab('p-values for Mahalanobis Test')

vars_outlier_test %>%
 mutate(sig =  case_when(round(p_value, 3) <= 0.5 ~ 'Sig',  TRUE ~ 'Not Sig')) %>%
 select(sig) %>% group_by_all() %>% summarise(hc = n(), .groups = 'drop')

###################
## ASSUMPTIONS ##
###################

###################################
## OBSERVATION INDEPENDENCE ##
###################################

## observations that are more than one
retain_exp %>%  select(unique_identifer) %>% group_by_all() %>%
summarise(hc = n(), .groups = 'drop') %>% filter(hc > 1)
```

```
#################
## LINEARITY ##
#################

## univarite linearity
## recipe update for cor matrix
retain_train_cor_max <- retain_train %>%
 step_mutate(gender_descr = case_when(gender_descr == 'Male' ~ 1, TRUE ~ 0),
            admit_first_gen_ind = case_when(admit_first_gen_ind == 'Y' ~ 1, TRUE ~ 0),
            cpc_english_code = case_when(cpc_english_code == 'S' ~ 1,
                                      cpc_english_code == 'R' ~ 2,  cpc_english_code == 'X' ~ 3,
                                      cpc_english_code == 'N' ~ 4, TRUE ~ 5),
            cpc_foreign_language_code = case_when(cpc_foreign_language_code == 'S' ~ 1,
                                              cpc_foreign_language_code == 'R' ~ 2,
                                              cpc_foreign_language_code == 'N' ~ 3,  TRUE ~ 4),
            cpc_math_code = case_when(cpc_math_code == 'S' ~ 1,
                                    cpc_math_code == 'R' ~ 2, cpc_math_code == 'X' ~ 3,
                                    cpc_math_code == 'N' ~ 4,  TRUE ~ 5),
            cpc_science_code = case_when(cpc_science_code == 'S' ~ 1,
                                      cpc_science_code == 'R' ~ 2, cpc_science_code == 'N' ~ 3,
                                      TRUE ~ 4),
            cpc_social_science_code = case_when(cpc_social_science_code == 'S' ~ 1,
                                            cpc_social_science_code == 'R' ~ 2,
                                            cpc_social_science_code == 'N' ~ 3, TRUE ~ 4),
            cip_categories = case_when(cip_categories == 'Social Sciences' ~ 1,
                                     cip_categories == 'Fine Arts' ~ 2,
                                     cip_categories == 'Human Services' ~ 3,
                                     cip_categories == 'Business' ~ 4, cip_categories == 'STEM' ~ 5,
                                     cip_categories == 'General/Interdisciplinary Studies' ~ 6,
                                     cip_categories == 'Healthcare' ~ 7,
                                     cip_categories == 'Education' ~ 8, TRUE ~ 9),
            zell_ind = case_when(zell_ind == 'Y' ~ 1, TRUE ~ 0),
            locale_group = case_when(locale_group == 'City' ~ 1,
                                   locale_group == 'Suburb' ~ 2, locale_group == 'Town' ~ 3,
                                   TRUE ~ 4),
            race_eth = case_when(race_eth == 'White' ~ 1,
                               race_eth == 'Black or African American' ~ 2,
                               race_eth == 'Hispanic or Latino' ~ 3, TRUE ~ 4))
retain_exp_cor <- juice(prep(retain_train_cor_max))

## correlational analysis
as.data.frame((retain_exp_cor %>%
              select(-unique_identifer) %>%
              corr.test(use = 'pairwise', method = 'pearson',
                    adjust = 'holm', alpha = .05))$r) %>%
 select( ## dv_next_fall,
        ## dv_first_fall_gpa,
        dv_first_yr_gpa ) %>%
 cbind(as.data.frame((retain_exp_cor %>%
                select(-unique_identifer) %>%
                corr.test(use = 'pairwise', method = 'pearson',
                       adjust = 'holm', alpha = .05))$p) %>%
     mutate(dv_next_fall_p = round(dv_next_fall, 3),
          dv_first_fall_gpa_p = round(dv_first_fall_gpa, 3),
          dv_first_yr_gpa_p = round(dv_first_yr_gpa, 3)) %>%
     select(## dv_next_fall_p,
           ## dv_first_fall_gpa_p,
           dv_first_yr_gpa_p)) %>%
 filter(## dv_next_fall_p <= .05,
       ## dv_first_fall_gpa_p <= .05,
       dv_first_yr_gpa_p <= .05) %>%
 arrange(## abs(dv_next_fall),
        ## abs(dv_first_fall_gpa),
        abs(dv_first_yr_gpa))

## correlational heat map
cor.plot(retain_exp_cor %>%
      select(dv_next_fall, dv_first_fall_gpa, dv_first_yr_gpa,
          gender_descr, race_eth, admit_first_gen_ind,
          locale_group, hsgpa_knn, ats_knn, adv_standing_ap_hrs,
          adv_standing_clep_hrs, adv_standing_ib_hrs, adv_standing_other_hrs,
          cpc_english_code, cpc_foreign_language_code, cpc_math_code,
          cpc_science_code, cpc_social_science_code, content_mastery,
          readiness, english, math, science, social_studies, efc_knn,
          ga_hope, zell_ind, pell, fed_sub_loans, fed_unsub_loans,
          oth_loans, cip_categories, acay_sup_exp, all_other_exp,
          inst_sup_exp, instr_exp, public_serv_exp, rsch_exp,
          stu_serv_exp) %>%
      rename(Retain = dv_next_fall, `Fall GPA` = dv_first_fall_gpa,
          `Year GPA` = dv_first_yr_gpa, Gender = gender_descr,
          `Race Eth` = race_eth, `First Gen` = admit_first_gen_ind,
          `HS Locale` = locale_group, `HS GPA` = hsgpa_knn,
          `Test Scores` = ats_knn, `AP Hours` = adv_standing_ap_hrs,
          `CLEP Hours` = adv_standing_clep_hrs, `IB Hours` = adv_standing_ib_hrs,
          `Other Hours` = adv_standing_other_hrs, `CPC English` = cpc_english_code,
          `CPC Fore. Lang.` = cpc_foreign_language_code,
          `CPC Math` = cpc_math_code, `CPC Science` = cpc_science_code,
          `CPC Social Sci` = cpc_social_science_code,
          `Content Mastery` = content_mastery, `Readiness` = readiness,
          `English PL` = english, `Math PL` = math,
          `Science PL` = science, `Social Studies PL` = social_studies,
          `Exp. Fam. Contrib.` = efc_knn, `GA HOPE` = ga_hope,
          `Zell Miller` = zell_ind, `PELL Grant` = pell,
          `Fed Sub Loans` = fed_sub_loans, `Fed Unsub Loans` = fed_unsub_loans,
          `Other Loans` = oth_loans, `Major Groupings` = cip_categories,
          `Academic Sup` = acay_sup_exp, `All Other` = all_other_exp,
          `Inst. Sup` = inst_sup_exp, `Instruction` = instr_exp,
          `Public Service` = public_serv_exp, `Research` = rsch_exp,
          `Student Serv` = stu_serv_exp),
      xlas = 2, stars = TRUE, show.legend = FALSE)

## vif for multicollinearity
retain_vif <- setCor(dv_next_fall ~ gender_descr + admit_first_gen_ind +
                adv_standing_ap_hrs + adv_standing_clep_hrs + adv_standing_ib_hrs +
                adv_standing_other_hrs + cpc_english_code +
                cpc_foreign_language_code +
                cpc_math_code + cpc_science_code + cpc_social_science_code +
                cip_categories + ga_hope + zell_ind + pell + fed_sub_loans +
                fed_unsub_loans + oth_loans + acay_sup_exp + all_other_exp +
                inst_sup_exp + instr_exp + public_serv_exp + rsch_exp +
                stu_serv_exp + race_eth + content_mastery + readiness + english + math +
                science + social_studies + locale_group + hsgpa_knn + ats_knn + efc_knn,
               data = retain_exp_cor, us = 'pairwise',plot = FALSE)
fall_gpa <- setCor(dv_first_fall_gpa ~ gender_descr + admit_first_gen_ind +
               adv_standing_ap_hrs + adv_standing_clep_hrs + adv_standing_ib_hrs +
               adv_standing_other_hrs + cpc_english_code + cpc_foreign_language_code +
               cpc_math_code + cpc_science_code + cpc_social_science_code +
               cip_categories + ga_hope + zell_ind + pell + fed_sub_loans +
               fed_unsub_loans + oth_loans + acay_sup_exp + all_other_exp +
               inst_sup_exp + instr_exp + public_serv_exp + rsch_exp +
               stu_serv_exp + race_eth + content_mastery + readiness + english + math +
               science + social_studies + locale_group + hsgpa_knn + ats_knn + efc_knn,
              data = retain_exp_cor, us = 'pairwise', plot = FALSE)
year_gpa <- setCor(dv_first_yr_gpa ~ gender_descr + admit_first_gen_ind +
               adv_standing_ap_hrs + adv_standing_clep_hrs + adv_standing_ib_hrs +
               adv_standing_other_hrs + cpc_english_code + cpc_foreign_language_code +
               cpc_math_code + cpc_science_code + cpc_social_science_code +
               cip_categories + ga_hope + zell_ind + pell + fed_sub_loans +
               fed_unsub_loans + oth_loans + acay_sup_exp + all_other_exp +
               inst_sup_exp + instr_exp + public_serv_exp + rsch_exp +
               stu_serv_exp + race_eth + content_mastery + readiness + english + math +
               science + social_studies + locale_group + hsgpa_knn + ats_knn + efc_knn,
              data = retain_exp_cor, us = 'pairwise', plot = FALSE)

########################################
## ELIMINATION OR REDUCTION OF ##
## MULTICOLLINEARITY                ##
########################################
```

```
## recipe update
retain_train_cor_max2 <- retain_train %>%
 step_mutate(gender_descr = case_when(gender_descr == 'Male' ~ 1, TRUE ~ 0),
         admit_first_gen_ind = case_when(admit_first_gen_ind == 'Y' ~ 1,  TRUE ~ 0),
         college_prep = case_when(cpc_english_code == 'S' ~ 1,
                          cpc_english_code == 'X' ~ 1, TRUE ~ 0) +
          case_when(cpc_foreign_language_code == 'S' ~ 1,
                cpc_foreign_language_code == 'X' ~ 1, TRUE ~ 0) +
          case_when(cpc_math_code == 'S' ~ 1,  cpc_math_code == 'X' ~ 1, TRUE ~ 0)  +
          case_when(cpc_science_code == 'S' ~ 1,  cpc_science_code == 'X' ~ 1,
                TRUE ~ 0) +
          case_when(cpc_social_science_code == 'S' ~ 1,
                cpc_social_science_code == 'X' ~ 1,  TRUE ~ 0),
         acay_inst_sup_exp = (acay_sup_exp + inst_sup_exp),
         public_rsch_exp = (public_serv_exp + rsch_exp),
         cm_ready = (readiness + content_mastery) / 2,
         english_cm = english - cm_ready,
         math_cm = math - cm_ready,
         science_cm = science - cm_ready,
         social_studies_cm = social_studies - cm_ready,
         cip_categories = case_when(cip_categories == 'Social Sciences' ~ 1,
                           cip_categories == 'Fine Arts' ~ 2,
                           cip_categories == 'Human Services' ~ 3,
                           cip_categories == 'Business' ~ 4,
                           cip_categories == 'STEM' ~ 5,
                           cip_categories == 'General/Interdisciplinary Studies' ~ 6,
                           cip_categories == 'Healthcare' ~ 7,
                           cip_categories == 'Education' ~ 8,
                           TRUE ~ 9),
         zell_ind = case_when(zell_ind == 'Y' ~ 1, TRUE ~ 0),
         locale_group = case_when(locale_group == 'City' ~ 1,
                          locale_group == 'Suburb' ~ 2, locale_group == 'Town' ~ 3,
                          TRUE ~ 4),
         race_eth = case_when(race_eth == 'White' ~ 1,
                       race_eth == 'Black or African American' ~ 2,
                       race_eth == 'Hispanic or Latino' ~ 3, TRUE ~ 4)) %>%
 step_rm(cpc_english_code, cpc_foreign_language_code,  cpc_math_code,
      cpc_science_code, cpc_social_science_code,  acay_sup_exp, inst_sup_exp,
      public_serv_exp, rsch_exp,  english, math,  science,social_studies,
      readiness,  content_mastery)
retain_exp_cor2 <- juice(prep(retain_train_cor_max2))

## examining the results of the correlation after
## fixing the multicollinearity violations
cor.plot(retain_exp_cor2 %>%
        select(dv_next_fall, dv_first_fall_gpa, dv_first_yr_gpa,
            gender_descr, race_eth, admit_first_gen_ind, locale_group,
            hsgpa_knn, ats_knn, adv_standing_ap_hrs, adv_standing_clep_hrs,
            adv_standing_ib_hrs, adv_standing_other_hrs, college_prep,
            cm_ready, english_cm, math_cm, science_cm, social_studies_cm,
            efc_knn, ga_hope, zell_ind, pell, fed_sub_loans, fed_unsub_loans,
            oth_loans, cip_categories, acay_inst_sup_exp, all_other_exp, instr_exp,
            stu_serv_exp, public_rsch_exp) %>%
        rename(Retain = dv_next_fall, `Fall GPA` = dv_first_fall_gpa,
            `Year GPA` = dv_first_yr_gpa, Gender = gender_descr,
            `Race Eth` = race_eth, `First Gen` = admit_first_gen_ind,
            `HS Locale` = locale_group, `HS GPA` = hsgpa_knn,
            `Test Scores` = ats_knn, `AP Hours` = adv_standing_ap_hrs,
            `CLEP Hours` = adv_standing_clep_hrs,
            `IB Hours` = adv_standing_ib_hrs, `Other Hours` = adv_standing_other_hrs,
            `College Prep` = college_prep, `Content Mastery/Readiness` = cm_ready,
            `English CM` = english_cm,`Math CM` = math_cm,
            `Science CM` = science_cm, `Social Studies CM` = social_studies_cm,
            `Exp. Fam. Contrib.` = efc_knn, `GA HOPE` = ga_hope,
            `Zell Miller` = zell_ind, `PELL Grant` = pell,
            `Fed Sub Loans` = fed_sub_loans, `Fed Unsub Loans` = fed_unsub_loans,
            `Other Loans` = oth_loans, `Major Groupings` = cip_categories,
            `Acad. Inst. Sup.` = acay_inst_sup_exp, `All Other` = all_other_exp,
            Instruction = instr_exp, `Student Serv` = stu_serv_exp,
            `Public Rsch` = public_rsch_exp),
       xlas = 2, stars = TRUE, show.legend = FALSE)

## vif after fixing multicollinearity
retain_2 <- setCor(dv_next_fall ~ gender_descr + admit_first_gen_ind +
                adv_standing_ap_hrs + adv_standing_clep_hrs +
                adv_standing_ib_hrs + adv_standing_other_hrs +
                cip_categories + ga_hope + zell_ind + pell + fed_sub_loans +
                fed_unsub_loans + oth_loans + all_other_exp + instr_exp + stu_serv_exp +
                race_eth + cm_ready + locale_group + hsgpa_knn + ats_knn +  efc_knn +
                college_prep + acay_inst_sup_exp + public_rsch_exp + english_cm +
                math_cm + science_cm + social_studies_cm,
               data = retain_exp_cor2, us = 'pairwise', plot = FALSE)
fall_gpa_2 <- setCor(dv_first_fall_gpa ~ gender_descr + admit_first_gen_ind +
                  adv_standing_ap_hrs +  adv_standing_clep_hrs + adv_standing_ib_hrs +
                  adv_standing_other_hrs +  cip_categories + ga_hope + zell_ind + pell +
                  fed_sub_loans +  fed_unsub_loans + oth_loans + all_other_exp +
                  instr_exp + stu_serv_exp + race_eth + cm_ready + locale_group +
                  hsgpa_knn + ats_knn + efc_knn + college_prep + acay_inst_sup_exp +
                  public_rsch_exp + english_cm + math_cm + science_cm +
                  social_studies_cm, data = retain_exp_cor2, us = 'pairwise', plot = FALSE)
year_gpa_2 <- setCor(dv_first_yr_gpa ~ gender_descr + admit_first_gen_ind +
                adv_standing_ap_hrs +  adv_standing_clep_hrs + adv_standing_ib_hrs +
                adv_standing_other_hrs +  cip_categories + ga_hope + zell_ind + pell +
                fed_sub_loans +  fed_unsub_loans + oth_loans + all_other_exp +
                instr_exp + stu_serv_exp + race_eth + cm_ready + locale_group +
                hsgpa_knn + ats_knn + efc_knn + college_prep + acay_inst_sup_exp +
                public_rsch_exp + english_cm + math_cm + science_cm +
                social_studies_cm, data = retain_exp_cor2, us = 'pairwise', plot = FALSE)

##################
## NORMALITY ##
##################

## univarity normality
retain_exp_cor2 %>%
 select(hsgpa_knn, ats_knn, adv_standing_ap_hrs, adv_standing_clep_hrs,
       adv_standing_ib_hrs, adv_standing_other_hrs, college_prep,
       cm_ready, english_cm, math_cm, science_cm, social_studies_cm,
       efc_knn, ga_hope, pell, fed_sub_loans, fed_unsub_loans, oth_loans,
       acay_inst_sup_exp, all_other_exp, instr_exp, stu_serv_exp,
       public_rsch_exp) %>%
 rename(`HS GPA` = hsgpa_knn, `Admissions Test Scores` = ats_knn,
       `AP Hours` = adv_standing_ap_hrs, `CLEP Hours` = adv_standing_clep_hrs,
       `IB Hours` = adv_standing_ib_hrs, `Other Hours` = adv_standing_other_hrs,
       `College Prep. Curricul` = college_prep,
       `CM & Ready Mean` = cm_ready, `English (CMR)` = english_cm,
       `Math (CMR)` = math_cm, `Science (CMR)` = science_cm,
       `Social Studies (CMR)` = social_studies_cm, `EFC` = efc_knn,
       `GA HOPE Scholarship` = ga_hope, `PELL Grant` = pell,
       `Federal Sub. Loans` = fed_sub_loans, `Federal Unsub. Loans` = fed_unsub_loans,
       `Other Loans` = oth_loans, `Acad. & Inst. Sup.` = acay_inst_sup_exp,
       `All Other` = all_other_exp, Instruction = instr_exp,
       `Student Serv. Sup` = stu_serv_exp,
       `Public Serv. & Rsch.` = public_rsch_exp) %>%
 gather(var_name, var_results) %>%
 ggplot(aes(sample = var_results)) + stat_qq() + stat_qq_line(color = 'blue') +
 facet_wrap(var_name ~ ., scales = 'free', ncol = 4) +
 theme_classic() + theme(text = element_text(size = 15))

## normality shapiro wilks
## sample of the first 5,000 observations
shapiro.test(retain_exp_cor2$adv_standing_ap_hrs[0:5000])
shapiro.test(retain_exp_cor2$adv_standing_clep_hrs[0:5000])
shapiro.test(retain_exp_cor2$adv_standing_ib_hrs[0:5000])
shapiro.test(retain_exp_cor2$adv_standing_other_hrs[0:5000])
shapiro.test(retain_exp_cor2$ga_hope[0:5000])
shapiro.test(retain_exp_cor2$pell[0:5000])
shapiro.test(retain_exp_cor2$college_prep[0:5000])
shapiro.test(retain_exp_cor2$fed_sub_loans[0:5000])
shapiro.test(retain_exp_cor2$fed_unsub_loans[0:5000])
shapiro.test(retain_exp_cor2$oth_loans[0:5000])
shapiro.test(retain_exp_cor2$acay_inst_sup_exp[0:5000])
shapiro.test(retain_exp_cor2$all_other_exp[0:5000])
shapiro.test(retain_exp_cor2$instr_exp[0:5000])
shapiro.test(retain_exp_cor2$public_rsch_exp[0:5000])
shapiro.test(retain_exp_cor2$stu_serv_exp[0:5000])
shapiro.test(retain_exp_cor2$cm_ready[0:5000])
shapiro.test(retain_exp_cor2$english_cm[0:5000])
shapiro.test(retain_exp_cor2$math_cm[0:5000])
shapiro.test(retain_exp_cor2$science_cm[0:5000])
shapiro.test(retain_exp_cor2$social_studies_cm[0:5000])
shapiro.test(retain_exp_cor2$hsgpa_knn[0:5000])
shapiro.test(retain_exp_cor2$ats_knn[0:5000])
shapiro.test(retain_exp_cor2$efc_knn[0:5000])

## normality jarque bera
jarque.bera.test(retain_exp_cor2$adv_standing_ap_hrs)
jarque.bera.test(retain_exp_cor2$adv_standing_clep_hrs)
jarque.bera.test(retain_exp_cor2$adv_standing_ib_hrs)
jarque.bera.test(retain_exp_cor2$adv_standing_other_hrs)
jarque.bera.test(retain_exp_cor2$college_prep)
jarque.bera.test(retain_exp_cor2$ga_hope)
jarque.bera.test(retain_exp_cor2$pell)
jarque.bera.test(retain_exp_cor2$fed_sub_loans)
jarque.bera.test(retain_exp_cor2$fed_unsub_loans)
jarque.bera.test(retain_exp_cor2$oth_loans)
jarque.bera.test(retain_exp_cor2$acay_inst_sup_exp)
jarque.bera.test(retain_exp_cor2$all_other_exp)
jarque.bera.test(retain_exp_cor2$instr_exp)
jarque.bera.test(retain_exp_cor2$public_rsch_exp)
jarque.bera.test(retain_exp_cor2$stu_serv_exp)
jarque.bera.test(retain_exp_cor2$cm_ready)
jarque.bera.test(retain_exp_cor2$english_cm)
jarque.bera.test(retain_exp_cor2$math_cm)
jarque.bera.test(retain_exp_cor2$science_cm)
jarque.bera.test(retain_exp_cor2$social_studies_cm)
jarque.bera.test(retain_exp_cor2$hsgpa_knn)
jarque.bera.test(retain_exp_cor2$ats_knn)
jarque.bera.test(retain_exp_cor2$efc_knn)

## multivaraite normality
## mardia's test
multivar_norm <- mult.norm(retain_exp_cor2 %>%
                          select(hsgpa_knn, ats_knn, adv_standing_ap_hrs,
                               adv_standing_clep_hrs, adv_standing_ib_hrs,
                               adv_standing_other_hrs, college_prep,
                               cm_ready, english_cm, math_cm, science_cm,
                               social_studies_cm,
                               efc_knn, ga_hope, pell, fed_sub_loans, fed_unsub_loans,
                               oth_loans,
                               acay_inst_sup_exp, all_other_exp, instr_exp, stu_serv_exp,
                               public_rsch_exp))
multivar_norm$mult.test

## negative skewed factors
(retain_exp_cor2 %>%
  select(hsgpa_knn,  ats_knn, college_prep,
        adv_standing_ap_hrs:adv_standing_other_hrs,
        cm_ready, english_cm:social_studies_cm, efc_knn,
        ga_hope, pell:oth_loans, acay_inst_sup_exp, all_other_exp:stu_serv_exp,
        public_rsch_exp) %>%
  describe()) %>% select(min, skew, kurtosis) %>% filter(skew < 0)

## data transformation to fix normality violations
## updating recipe for exploration
retain_xform <- retain_train_cor_max2 %>%
 step_normalize(hsgpa_knn, ats_knn, adv_standing_ap_hrs, adv_standing_clep_hrs,
              adv_standing_ib_hrs, adv_standing_other_hrs, cm_ready, english_cm,
              math_cm, science_cm, social_studies_cm, efc_knn, ga_hope, pell,
              fed_sub_loans, fed_unsub_loans, oth_loans, acay_inst_sup_exp,
              all_other_exp, instr_exp, public_rsch_exp, stu_serv_exp) %>%
 ## step_inverse(hsgpa_knn, cm_ready, college_prep, social_studies_cm,
 ##            ga_hope, public_rsch_exp) %>%
 ## testing out different transformation methods
 step_YeoJohnson(
  #step_log(
  #step_BoxCox(
  hsgpa_knn, ats_knn, adv_standing_ap_hrs, adv_standing_clep_hrs,
  adv_standing_ib_hrs, adv_standing_other_hrs, cm_ready,
  english_cm, math_cm, science_cm, social_studies_cm, efc_knn,
  ga_hope, pell, fed_sub_loans, fed_unsub_loans, oth_loans,
  acay_inst_sup_exp, all_other_exp, instr_exp, public_rsch_exp, stu_serv_exp)
retain_exp_xf <- juice(prep(retain_xform))

## evaluating the changes in skewness values
(retain_exp_xf %>%
  select(hsgpa_knn, ats_knn, adv_standing_ap_hrs, adv_standing_clep_hrs,
        adv_standing_ib_hrs, adv_standing_other_hrs, cm_ready, english_cm,
        math_cm, science_cm, social_studies_cm, efc_knn, ga_hope, pell, fed_sub_loans,
        fed_unsub_loans, oth_loans, acay_inst_sup_exp, all_other_exp,
        instr_exp, public_rsch_exp, stu_serv_exp) %>% describe())

## revised recipe
retain_xform_rv <- retain_train_cor_max2 %>%
step_YeoJohnson(hsgpa_knn, ats_knn, college_prep, adv_standing_ap_hrs,
                adv_standing_clep_hrs, adv_standing_ib_hrs, adv_standing_other_hrs,
                cm_ready, english_cm, math_cm, science_cm,
                social_studies_cm, efc_knn, ga_hope, pell, fed_sub_loans,
                fed_unsub_loans, oth_loans, acay_inst_sup_exp, all_other_exp,
                instr_exp, stu_serv_exp, public_rsch_exp) %>%
 step_normalize(hsgpa_knn, ats_knn,college_prep, adv_standing_ap_hrs,
               adv_standing_clep_hrs, adv_standing_ib_hrs, adv_standing_other_hrs,
               cm_ready, english_cm, math_cm, science_cm, social_studies_cm, efc_knn,
               ga_hope, pell, fed_sub_loans, fed_unsub_loans, oth_loans, acay_inst_sup_exp,
               all_other_exp, instr_exp, stu_serv_exp, public_rsch_exp)
retain_xf_dat <- juice(prep(retain_xform_rv))

(retain_xf_dat %>%
  select(hsgpa_knn, ats_knn, adv_standing_ap_hrs, adv_standing_clep_hrs,
        adv_standing_ib_hrs, adv_standing_other_hrs, cm_ready, english_cm,
        math_cm, science_cm, social_studies_cm, efc_knn, ga_hope, pell, fed_sub_loans,
        fed_unsub_loans, oth_loans, acay_inst_sup_exp, all_other_exp,
        instr_exp, public_rsch_exp,  stu_serv_exp) %>%
  describe()) %>%select(skew, kurtosis)

## reassessing normality
## normality shapiro wilks--sample of the first 5,000 observations
shapiro.test(retain_xf_dat$hsgpa_knn[0:5000])
shapiro.test(retain_xf_dat$ats_knn[0:5000])
shapiro.test(retain_xf_dat$adv_standing_ap_hrs[0:5000])
shapiro.test(retain_xf_dat$adv_standing_clep_hrs[0:5000])
shapiro.test(retain_xf_dat$adv_standing_ib_hrs[0:5000])
shapiro.test(retain_xf_dat$adv_standing_other_hrs[0:5000])
shapiro.test(retain_xf_dat$cm_ready[0:5000])
shapiro.test(retain_xf_dat$college_prep[0:5000])
shapiro.test(retain_xf_dat$english_cm[0:5000])
shapiro.test(retain_xf_dat$math_cm[0:5000])
shapiro.test(retain_xf_dat$science_cm[0:5000])
shapiro.test(retain_xf_dat$social_studies_cm[0:5000])
shapiro.test(retain_xf_dat$efc_knn[0:5000])
shapiro.test(retain_xf_dat$ga_hope[0:5000])
shapiro.test(retain_xf_dat$pell[0:5000])
shapiro.test(retain_xf_dat$fed_sub_loans[0:5000])
shapiro.test(retain_xf_dat$fed_unsub_loans[0:5000])
shapiro.test(retain_xf_dat$oth_loans[0:5000])
shapiro.test(retain_xf_dat$acay_inst_sup_exp[0:5000])
shapiro.test(retain_xf_dat$all_other_exp[0:5000])
shapiro.test(retain_xf_dat$instr_exp[0:5000])
shapiro.test(retain_xf_dat$public_rsch_exp[0:5000])
shapiro.test(retain_xf_dat$stu_serv_exp[0:5000])

## normality jarque bera
jarque.bera.test(retain_xf_dat$hsgpa_knn)
jarque.bera.test(retain_xf_dat$ats_knn)
jarque.bera.test(retain_xf_dat$adv_standing_ap_hrs)
jarque.bera.test(retain_xf_dat$adv_standing_clep_hrs)
jarque.bera.test(retain_xf_dat$adv_standing_ib_hrs)
jarque.bera.test(retain_xf_dat$adv_standing_other_hrs)
jarque.bera.test(retain_xf_dat$cm_ready)
jarque.bera.test(retain_xf_dat$college_prep)
jarque.bera.test(retain_xf_dat$english_cm)
jarque.bera.test(retain_xf_dat$math_cm)
jarque.bera.test(retain_xf_dat$science_cm)
jarque.bera.test(retain_xf_dat$social_studies_cm)
jarque.bera.test(retain_xf_dat$efc_knn)
jarque.bera.test(retain_xf_dat$ga_hope)
jarque.bera.test(retain_xf_dat$pell)
jarque.bera.test(retain_xf_dat$fed_sub_loans)
jarque.bera.test(retain_xf_dat$fed_unsub_loans)
jarque.bera.test(retain_xf_dat$oth_loans)
jarque.bera.test(retain_xf_dat$acay_inst_sup_exp)
jarque.bera.test(retain_xf_dat$all_other_exp)
jarque.bera.test(retain_xf_dat$instr_exp)
jarque.bera.test(retain_xf_dat$public_rsch_exp)
jarque.bera.test(retain_xf_dat$stu_serv_exp)

####################################
## HOMOGENEITY OF VARIANCE ##
####################################

## levene test center mean
leveneTest(retain_xf_dat$gender_descr ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$race_eth ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$admit_first_gen_ind ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$locale_group ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$hsgpa_knn ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
```

```
leveneTest(retain_xf_dat$ats_knn ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$adv_standing_ap_hrs ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$adv_standing_clep_hrs ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$adv_standing_ib_hrs ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$adv_standing_other_hrs ~
as.factor(retain_xf_dat$dv_next_fall), center = mean)
leveneTest(retain_xf_dat$college_prep ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$cm_ready ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$english_cm ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$math_cm ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$science_cm ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$social_studies_cm ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$efc_knn ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$ga_hope ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$zell_ind ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$pell ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$fed_sub_loans ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$fed_unsub_loans ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$oth_loans ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$cip_categories ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$acay_inst_sup_exp ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$all_other_exp ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
leveneTest(retain_xf_dat$instr_exp ~ as.factor(retain_xf_dat$dv_next_fall), center =
mean)
leveneTest(retain_xf_dat$public_rsch_exp ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)
```

```
leveneTest(retain_xf_dat$stu_serv_exp ~ as.factor(retain_xf_dat$dv_next_fall),
center = mean)

## bartlett test
bartlett.test(retain_xf_dat$gender_descr ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$race_eth ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$admit_first_gen_ind ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$locale_group ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$hsgpa_knn ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$ats_knn ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$adv_standing_ap_hrs ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$adv_standing_clep_hrs ~
as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$adv_standing_ib_hrs ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$adv_standing_other_hrs ~
as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$college_prep ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$cm_ready ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$english_cm ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$math_cm ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$science_cm ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$social_studies_cm ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$efc_knn ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$ga_hope ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$zell_ind ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$pell ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$fed_sub_loans ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$fed_unsub_loans ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$oth_loans ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$cip_categories ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$acay_inst_sup_exp ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$all_other_exp ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$instr_exp ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$public_rsch_exp ~ as.factor(retain_xf_dat$dv_next_fall))
bartlett.test(retain_xf_dat$stu_serv_exp ~ as.factor(retain_xf_dat$dv_next_fall))

## homogeneity graphs
retain_xf_dat %>%
 select(dv_next_fall, gender_descr, race_eth, admit_first_gen_ind, zell_ind,
        locale_group, cip_categories) %>%
 mutate(dv_next_fall = case_when(dv_next_fall == 1 ~ 'Not Retained',
                           TRUE ~ 'Retained'),
        gender_descr = case_when(gender_descr == 1 ~ 'Male', TRUE ~ 'Female'),
        admit_first_gen_ind = case_when(admit_first_gen_ind == 1 ~ 'Yes', TRUE ~ 'No'),
        zell_ind =  case_when(zell_ind == 1 ~ 'Yes', TRUE ~ 'No'),
        race_eth = case_when(race_eth == 1 ~ 'White', race_eth == 2 ~ 'Black or AA',
                    race_eth == 3 ~ 'Hispanic', TRUE ~ 'Other'),
      locale_group = case_when(locale_group == 1 ~ 'City', locale_group == 2 ~ 'Suburb',
                       locale_group == 3 ~ 'Town', TRUE ~ 'Rural'),
      cip_categories = case_when(cip_categories == 1 ~ 'Social Sciences',
                        cip_categories == 2 ~ 'Fine Arts',
                        cip_categories == 3 ~ 'Human Services',
                        cip_categories == 4 ~ 'Business',
                        cip_categories == 5 ~ 'STEM',
                        cip_categories == 6 ~ 'Interdisc. Studies',
                        cip_categories == 7 ~ 'Healthcare',
                        cip_categories == 8 ~ 'Education', TRUE ~ 'Humanities')) %>%
 rename(`Retain Status` = dv_next_fall, Gender = gender_descr,
      `First Gen Status` = admit_first_gen_ind, `Race Ethnicity` = race_eth,
      `Zell Miller Ind.` = zell_ind, `HS Locale` = locale_group,
      `Major Grouping` = cip_categories) %>%
 gather(var_type, var_results, -`Retain Status`) %>% group_by_all() %>%
 summarise(hc = n(),.groups = 'drop') %>%
 ggplot() + geom_bar(aes(x = as.character(var_results), y = hc, fill = `Retain Status`),
       stat = 'identity', position = 'dodge') + theme_classic() +
 theme(plot.title = element_text(hjust = .5), axis.title = element_blank(),
     text = element_text(size = 13), legend.position = 'top') +
 facet_wrap(var_type ~ ., scales = 'free') + coord_flip()

retain_xf_dat %>%
 select(dv_next_fall, adv_standing_ap_hrs:adv_standing_other_hrs,
      ga_hope:stu_serv_exp, -zell_ind, hsgpa_knn:social_studies_cm) %>%
 mutate(dv_next_fall = case_when(dv_next_fall == 1 ~ 'Not Retained',
                       TRUE ~ 'Retained')) %>%
 gather(var_type, var_results, dv_next_fall) %>%
 mutate(var_type = case_when(var_type == 'hsgpa_knn' ~ 'HS GPA',
                    var_type == 'adv_standing_ap_hrs' ~ 'AP Hours',
                    var_type == 'adv_standing_clep_hrs' ~ 'CLEP Hours',
                    var_type == 'adv_standing_ib_hrs' ~ 'IB Hours',
                    var_type == 'adv_standing_other_hrs' ~ 'Other Hours',
                    var_type == 'efc_knn' ~ 'EFC',
                    var_type == 'ga_hope' ~ 'GA HOPE',
                    var_type == 'pell' ~ 'PELL Grant',
                    var_type == 'fed_sub_loans' ~ 'Fed. Sub. Loans',
                    var_type == 'fed_unsub_loans' ~ 'Fed. Unsub. Loans',
                    var_type == 'oth_loans' ~ 'Other Loans',
                    var_type == 'ats_knn' ~ 'Adm. Test Scores',
                    var_type == 'all_other_exp' ~ 'All Other',
                    var_type == 'instr_exp' ~ 'Instruction',
                    var_type == 'stu_serv_exp' ~ 'Student Services',
                    var_type == 'cm_ready' ~ 'CM & Ready Mean',
                    var_type == 'college_prep' ~ 'College Prep. Curric.',
                    var_type == 'acay_inst_sup_exp' ~ 'Acad. & Inst. Support',
                    var_type == 'public_rsch_exp' ~ 'Public Ser & Rsch',
                    var_type == 'english_cm' ~ 'English (CMR)',
                    var_type == 'math_cm' ~ 'Math (CMR)',
                    var_type == 'science_cm' ~ 'Science (CMR)',
                    var_type == 'social_studies_cm' ~ 'Social Studies (CMR)',
                    TRUE ~ var_type)) %>%
 ggplot() + geom_boxplot(aes(x = dv_next_fall, y = var_results)) + theme_classic() +
 theme(plot.title = element_text(hjust = .5), axis.title = element_blank(),
     text = element_text(size = 13), legend.position = 'top') +
 facet_wrap(var_type ~ ., scales = 'free') + coord_flip()
```

APPENDIX F:

Pearson Correlation Matrix Before Data Manipulation

Table 31

Pearson Correlation Matrix Before Data Manipulation

	Variable	1		2		3		4		5	
1	First-fall GPA	1.000	***								
2	First-year GPA	.927	***	1.000	***						
3	One-year Retention Status	-.386	***	-.441	***	1.000	***				
4	Gender	-.149	***	-.163	***	.080	***	1.000	***		
5	First Generation Status	-.032	***	-.039	***	.024		-.033	***	1.000	***
6	AP Hours	.219	***	.215	***	-.007		.075	***	-.043	***
7	CLEP Hours	.033	***	.033	***	.000		.002		-.006	
8	IB Hours	.039	***	.040	***	-.011		-.005		-.004	
9	Other Hours	.006		.007		-.008		-.012		.022	*
10	CPC English	.016		.010		.004		-.031	***	-.009	
11	CPC Foreign Language	.014		.009		.008		-.037	***	-.002	
12	CPC Math	.015		.010		.004		-.031	***	-.007	
13	CPC Science	.013		.007		.006		-.029	***	-.011	
14	CPC Social Sciences	.020	*	.014		.006		-.039	***	-.004	
15	Major Grouping	.002		.008		.000		-.047	***	.019	*
16	GA HOPE Scholarship	.430	***	.452	***	-.183	***	-.111	***	-.010	
17	Zell Miller Indicator	.281	***	.285	***	-.045	***	.001		-.044	***
18	PELL Grant	-.097	***	-.109	***	.037	**	-.082	***	.262	***
19	Federal Sub. Loans	-.135	***	-.148	***	.045	***	-.080	***	.071	***
20	Federal Unsub. Loans	-.101	***	-.101	***	.027		-.037	***	-.045	***
21	Other Loans	-.066	***	-.061	***	.008		.003		-.012	
22	Academic Support	.061	***	.027	**	-.015		.057	***	-.023	**
23	All Other	.035	***	.089	***	-.036	*	.025	**	.003	
24	Institutional Support	-.046	***	-.021	*	.054	***	-.075	***	.058	***
25	Instruction	-.102	***	-.105	***	.057	***	-.080	***	.032	***
26	Public Service	.122	***	.106	***	-.052	***	.097	***	-.031	***
27	Research	-.065	***	-.042	***	-.019		-.014		-.027	**
28	Student Services	-.061	***	-.084	***	.054	***	-.094	***	.028	**
29	Race/Ethnicity	-.019	*	-.022	*	-.009		-.018	*	.159	***
30	CCRPI Content Mastery	.114	***	.120	***	-.040	**	.073	***	-.159	***
31	CCPRI Readiness	.096	***	.102	***	-.035	*	.066	***	-.149	***
32	EOC English	.100	***	.110	***	-.033	*	.074	***	-.158	***
33	EOC Mathematics	.109	***	.114	***	-.031		.075	***	-.144	***
34	EOC Science	.102	***	.108	***	-.037	**	.069	***	-.138	***
35	EOC Social Studies	.113	***	.114	***	-.034	*	.070	***	-.158	***
36	Graduating HS Locale	-.001		-.001		.023		-.014		.036	***
37	HS GPA	.488	***	.507	***	-.165	***	-.152	***	.007	
38	Admissions Test Scores	.257	***	.254	***	-.031	***	.156	***	-.107	***
39	EFC	.055	***	.070	***	-.043	***	.048	***	-.119	***

Note. *** $p < .001$. ** $p < .01$. * $p < .05$. $df = 13{,}078$. CPC = college preparatory curriculum. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. CCRPI = college and career ready performance index. EOC = end-of-course. HS = high school. GPA = grade point average. EFC = expected family contributions.

Table 31 (continued)

Pearson Correlation Matrix Before Data Manipulation

	Variable	6		7		8		9		10	
1	First-fall GPA										
2	First-year GPA										
3	One-year Retention Status										
4	Gender										
5	First Generation Status										
6	AP Hours	1.000	***								
7	CLEP Hours	.048	***	1.000	***						
8	IB Hours	.000		.010		1.000	***				
9	Other Hours	.004		-.002		-.003		1.000	***		
10	CPC English	-.010		.004		-.006		.002		1.000	***
11	CPC Foreign Language	-.008		.008		-.006		.003		.959	***
12	CPC Math	-.013		.003		-.006		.002		.986	***
13	CPC Science	-.014		.002		-.006		.001		.971	***
14	CPC Social Sciences	-.007		.003		-.006		.002		.966	***
15	Major Grouping	-.013		-.001		-.014		.002		.017	
16	GA HOPE Scholarship	.181	***	.020	*	.038	***	.015		.038	***
17	Zell Miller Indicator	.356	***	.028	**	.027	**	.013		.046	***
18	PELL Grant	-.121	***	-.014		.005		.006		-.010	
19	Federal Sub. Loans	-.131	***	-.023	**	-.016		.012		-.016	
20	Federal Unsub. Loans	-.094	***	-.017		-.031	***	-.001		.005	
21	Other Loans	-.038	***	-.009		-.005		-.001		.004	
22	Academic Support	.050	***	.005		-.002		.005		-.008	
23	All Other	.080	***	.010		.030	***	.020	*	-.143	***
24	Institutional Support	-.072	***	-.008		-.002		.013		-.178	***
25	Instruction	-.158	***	-.024	**	-.037	***	-.004		.128	***
26	Public Service	.181	***	.027	**	.043	***	.007		-.164	***
27	Research	-.081	***	-.017	*	-.032	***	-.001		.224	***
28	Student Services	-.133	***	-.010		-.037	***	-.022	*	.040	***
29	Race/Ethnicity	.014		.013		.049	***	.009		-.028	**
30	CCRPI Content Mastery	.210	***	.019	*	-.039	***	.000		-.055	***
31	CCPRI Readiness	.178	***	.017		-.025	**	-.003		-.029	***
32	EOC English	.209	***	.028	**	-.028	**	-.001		-.067	***
33	EOC Mathematics	.204	***	.027	**	-.041	***	-.006		-.052	***
34	EOC Science	.204	***	.024	**	-.037	***	.000		-.049	***
35	EOC Social Studies	.213	***	.019	*	-.039	***	-.001		-.052	***
36	Graduating HS Locale	-.072	***	-.005		-.051	***	.004		.032	***
37	HS GPA	.249	***	.027	**	.036	***	.013		.073	***
38	Admissions Test Scores	.523	***	.055	***	.062	***	.006		.023	**
39	EFC	.088	***	.008		-.004		-.008		-.007	

Note. *** $p < .001$. ** $p < .01$. * $p < .05$. $df = 13{,}078$. CPC = college preparatory curriculum. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. CCRPI = college and career ready performance index. EOC = end-of-course. HS = high school. GPA = grade point average. EFC = expected family contributions.

Table 31 (continued)

Pearson Correlation Matrix Before Data Manipulation

	Variable	11	12	13	14	15
1	First-fall GPA					
2	First-year GPA					
3	One-year Retention Status					
4	Gender					
5	First Generation Status					
6	AP Hours					
7	CLEP Hours					
8	IB Hours					
9	Other Hours					
10	CPC English					
11	CPC Foreign Language	1.000 ***				
12	CPC Math	.959 ***	1.000 ***			
13	CPC Science	.952 ***	.969 ***	1.000 ***		
14	CPC Social Sciences	.976 ***	.964 ***	.956 ***	1.000 ***	
15	Major Grouping	.018 *	.016	.014	.019 *	1.000 ***
16	GA HOPE Scholarship	.035 ***	.036 ***	.037 ***	.039 ***	.049 ***
17	Zell Miller Indicator	.044 ***	.043 ***	.039 ***	.042 ***	.014
18	PELL Grant	.000	-.009	-.005	-.006	-.013
19	Federal Sub. Loans	-.012	-.017	-.014	-.015	-.004
20	Federal Unsub. Loans	.004	.004	.005	.002	-.005
21	Other Loans	.005	.004	.002	-.001	.005
22	Academic Support	-.006	-.011	-.008	-.005	-.042 ***
23	All Other	-.115 ***	-.141 ***	-.139 ***	-.122 ***	-.023 **
24	Institutional Support	-.129 ***	-.174 ***	-.176 ***	-.142 ***	-.027 **
25	Instruction	.108 ***	.126 ***	.122 ***	.112 ***	.002
26	Public Service	-.136 ***	-.162 ***	-.157 ***	-.141 ***	-.023 **
27	Research	.172 ***	.220 ***	.219 ***	.185 ***	.041 ***
28	Student Services	.037 ***	.041 ***	.037 ***	.036 ***	.051 ***
29	Race/Ethnicity	-.017 *	-.028 **	-.027 **	-.023 **	-.032 ***
30	CCRPI Content Mastery	-.055 ***	-.054 ***	-.059 ***	-.047 ***	-.015
31	CCPRI Readiness	-.033 ***	-.028 **	-.035 ***	-.023 **	.003
32	EOC English	-.065 ***	-.066 ***	-.071 ***	-.057 ***	-.020 *
33	EOC Mathematics	-.050 ***	-.052 ***	-.054 ***	-.043 ***	-.010
34	EOC Science	-.051 ***	-.048 ***	-.054 ***	-.042 ***	-.011
35	EOC Social Studies	-.054 ***	-.051 ***	-.056 ***	-.045 ***	-.020 *
36	Graduating HS Locale	.027 **	.033 ***	.030 ***	.028 **	.047 ***
37	HS GPA	.073 ***	.071 ***	.071 ***	.074 ***	.081 ***
38	Admissions Test Scores	.018 *	.019 *	.018 *	.021 *	-.016
39	EFC	-.010	-.007	-.010	-.008	-.009

Note. *** *p* < .001. ** *p* < .01. * *p* < .05. *df* = 13,078. CPC = college preparatory curriculum. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. CCRPI = college and career ready performance index. EOC = end-of-course. HS = high school. GPA = grade point average. EFC = expected family contributions.

Table 31 (continued)

Pearson Correlation Matrix Before Data Manipulation

	Variable	16		17		18		19		20	
1	First-fall GPA										
2	First-year GPA										
3	One-year Retention Status										
4	Gender										
5	First Generation Status										
6	AP Hours										
7	CLEP Hours										
8	IB Hours										
9	Other Hours										
10	CPC English										
11	CPC Foreign Language										
12	CPC Math										
13	CPC Science										
14	CPC Social Sciences										
15	Major Grouping										
16	GA HOPE Scholarship	1.000	***								
17	Zell Miller Indicator	.270	***	1.000	***						
18	PELL Grant	-.077	***	-.101	***	1.000	***				
19	Federal Sub. Loans	-.133	***	-.121	***	.391	***	1.000	***		
20	Federal Unsub. Loans	-.096	***	-.085	***	.007		.281	***	1.000	***
21	Other Loans	-.093	***	-.027	**	-.064	***	.065	***	.095	***
22	Academic Support	.021	*	.018	*	-.084	***	-.055	***	-.040	***
23	All Other	.074	***	.031	***	-.034	***	-.062	***	-.059	***
24	Institutional Support	-.072	***	-.052	***	.112	***	.065	***	.048	***
25	Instruction	-.118	***	-.072	***	.108	***	.104	***	.116	***
26	Public Service	.113	***	.075	***	-.129	***	-.118	***	-.135	***
27	Research	-.002		-.011		-.006		.021	*	.067	***
28	Student Services	-.090	***	-.062	***	.131	***	.115	***	.100	***
29	Race/Ethnicity	-.063	***	-.063	***	.239	***	.113	***	-.011	
30	CCRPI Content Mastery	.007		.033	***	-.318	***	-.251	***	-.098	***
31	CCPRI Readiness	.007		.027	**	-.280	***	-.218	***	-.082	***
32	EOC English	-.013		.022	*	-.302	***	-.235	***	-.089	***
33	EOC Mathematics	.005		.031	***	-.298	***	-.248	***	-.097	***
34	EOC Science	.003		.022	*	-.275	***	-.220	***	-.086	***
35	EOC Social Studies	-.001		.025	**	-.310	***	-.244	***	-.098	***
36	Graduating HS Locale	.098	***	.059	***	-.024	**	.001		.013	
37	HS GPA	.708	***	.445	***	-.089	***	-.154	***	-.135	***
38	Admissions Test Scores	.274	***	.475	***	-.241	***	-.228	***	-.138	***
39	EFC	.050	***	.035	***	-.364	***	-.313	***	-.027	**

Note. *** $p < .001$. ** $p < .01$. * $p < .05$. $df = 13{,}078$. CPC = college preparatory curriculum. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. CCRPI = college and career ready performance index. EOC = end-of-course. HS = high school. GPA = grade point average. EFC = expected family contributions.

Table 31 (continued)

Pearson Correlation Matrix Before Data Manipulation

	Variable	21		22		23		24		25	
1	First-fall GPA										
2	First-year GPA										
3	One-year Retention Status										
4	Gender										
5	First Generation Status										
6	AP Hours										
7	CLEP Hours										
8	IB Hours										
9	Other Hours										
10	CPC English										
11	CPC Foreign Language										
12	CPC Math										
13	CPC Science										
14	CPC Social Sciences										
15	Major Grouping										
16	GA HOPE Scholarship										
17	Zell Miller Indicator										
18	PELL Grant										
19	Federal Sub. Loans										
20	Federal Unsub. Loans										
21	Other Loans	1.000	***								
22	Academic Support	-.024	**	1.000	***						
23	All Other	.020	*	-.296	***	1.000	***				
24	Institutional Support	-.016		-.142	***	.256	***	1.000	***		
25	Instruction	-.010		.167	***	-.534	***	.420	***	1.000	***
26	Public Service	-.003		.214	***	.315	***	-.326	***	-.869	***
27	Research	.030	***	.023	**	-.182	***	-.336	***	.450	***
28	Student Services	-.012		-.498	***	-.414	***	.240	***	.341	***
29	Race/Ethnicity	-.018	*	-.003		.025	**	.043	***	-.039	***
30	CCRPI Content Mastery	.002		.131	***	.149	***	-.113	***	-.235	***
31	CCPRI Readiness	.003		.061	***	.115	***	-.105	***	-.175	***
32	EOC English	.004		.129	***	.192	***	-.069	***	-.219	***
33	EOC Mathematics	.000		.105	***	.122	***	-.104	***	-.211	***
34	EOC Science	.004		.113	***	.125	***	-.112	***	-.220	***
35	EOC Social Studies	-.006		.153	***	.104	***	-.148	***	-.228	***
36	Graduating HS Locale	.011		-.060	***	-.030	***	.006		.068	***
37	HS GPA	-.068	***	.003		.076	***	-.101	***	-.134	***
38	Admissions Test Scores	-.022	*	.088	***	.096	***	-.223	***	-.283	***
39	EFC	-.012		.042	***	.044	***	-.054	***	-.076	***

Note. *** $p < .001$. ** $p < .01$. * $p < .05$. $df = 13{,}078$. CPC = college preparatory curriculum. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. CCRPI = college and career ready performance index. EOC = end-of-course. HS = high school. GPA = grade point average. EFC = expected family contributions.

Table 31 (continued)

Pearson Correlation Matrix Before Data Manipulation

	Variable	26		27		28		29		30	
1	First-fall GPA										
2	First-year GPA										
3	One-year Retention Status										
4	Gender										
5	First Generation Status										
6	AP Hours										
7	CLEP Hours										
8	IB Hours										
9	Other Hours										
10	CPC English										
11	CPC Foreign Language										
12	CPC Math										
13	CPC Science										
14	CPC Social Sciences										
15	Major Grouping										
16	GA HOPE Scholarship										
17	Zell Miller Indicator										
18	PELL Grant										
19	Federal Sub. Loans										
20	Federal Unsub. Loans										
21	Other Loans										
22	Academic Support										
23	All Other										
24	Institutional Support										
25	Instruction										
26	Public Service	1.000	***								
27	Research	-.617	***	1.000	***						
28	Student Services	-.464	***	-.014		1.000	***				
29	Race/Ethnicity	.058	***	-.097	***	-.018	*	1.000	***		
30	CCRPI Content Mastery	.277	***	-.082	***	-.268	***	-.104	***	1.000	***
31	CCPRI Readiness	.183	***	-.016		-.177	***	-.105	***	.869	***
32	EOC English	.258	***	-.074	***	-.290	***	-.088	***	.953	***
33	EOC Mathematics	.247	***	-.079	***	-.219	***	-.112	***	.938	***
34	EOC Science	.258	***	-.079	***	-.226	***	-.084	***	.934	***
35	EOC Social Studies	.282	***	-.081	***	-.270	***	-.089	***	.942	***
36	Graduating HS Locale	-.104	***	.079	***	.088	***	-.128	***	-.143	***
37	HS GPA	.119	***	.014		-.093	***	-.067	***	-.056	***
38	Admissions Test Scores	.298	***	-.049	***	-.208	***	-.107	***	.297	***
39	EFC	.088	***	-.004		-.097	***	-.139	***	.210	***

Note. *** $p < .001$. ** $p < .01$. * $p < .05$. $df = 13{,}078$. CPC = college preparatory curriculum. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. CCRPI = college and career ready performance index. EOC = end-of-course. HS = high school. GPA = grade point average. EFC = expected family contributions.

Table 31 (continued)

Pearson Correlation Matrix Before Data Manipulation

	Variable	31		32		33		34		35	
1	First-fall GPA										
2	First-year GPA										
3	One-year Retention Status										
4	Gender										
5	First Generation Status										
6	AP Hours										
7	CLEP Hours										
8	IB Hours										
9	Other Hours										
10	CPC English										
11	CPC Foreign Language										
12	CPC Math										
13	CPC Science										
14	CPC Social Sciences										
15	Major Grouping										
16	GA HOPE Scholarship										
17	Zell Miller Indicator										
18	PELL Grant										
19	Federal Sub. Loans										
20	Federal Unsub. Loans										
21	Other Loans										
22	Academic Support										
23	All Other										
24	Institutional Support										
25	Instruction										
26	Public Service										
27	Research										
28	Student Services										
29	Race/Ethnicity										
30	CCRPI Content Mastery										
31	CCPRI Readiness	1.000	***								
32	EOC English	.871	***	1.000	***						
33	EOC Mathematics	.840	***	.917	***	1.000	***				
34	EOC Science	.812	***	.900	***	.891	***	1.000	***		
35	EOC Social Studies	.809	***	.894	***	.871	***	.862	***	1.000	***
36	Graduating HS Locale	-.070	***	-.164	***	-.142	***	-.172	***	-.165	***
37	HS GPA	-.034	***	-.089	***	-.053	***	-.058	***	-.074	***
38	Admissions Test Scores	.247	***	.287	***	.291	***	.278	***	.287	***
39	EFC	.184	***	.207	***	.210	***	.188	***	.199	***

Note. *** $p < .001$. ** $p < .01$. * $p < .05$. $df = 13{,}078$. CPC = college preparatory curriculum. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. CCRPI = college and career ready performance index. EOC = end-of-course. HS = high school. GPA = grade point average. EFC = expected family contributions.

Table 31 (continued)

Pearson Correlation Matrix Before Data Manipulation

	Variable	36	37	38	39
1	First-fall GPA				
2	First-year GPA				
3	One-year Retention Status				
4	Gender				
5	First Generation Status				
6	AP Hours				
7	CLEP Hours				
8	IB Hours				
9	Other Hours				
10	CPC English				
11	CPC Foreign Language				
12	CPC Math				
13	CPC Science				
14	CPC Social Sciences				
15	Major Grouping				
16	GA HOPE Scholarship				
17	Zell Miller Indicator				
18	PELL Grant				
19	Federal Sub. Loans				
20	Federal Unsub. Loans				
21	Other Loans				
22	Academic Support				
23	All Other				
24	Institutional Support				
25	Instruction				
26	Public Service				
27	Research				
28	Student Services				
29	Race/Ethnicity				
30	CCRPI Content Mastery				
31	CCPRI Readiness				
32	EOC English				
33	EOC Mathematics				
34	EOC Science				
35	EOC Social Studies				
36	Graduating HS Locale	1.000 ***			
37	HS GPA	.153 ***	1.000 ***		
38	Admissions Test Scores	-.035 ***	.382 ***	1.000 ***	
39	EFC	-.014	.034 ***	.150 ***	1.000 ***

Note. *** $p < .001$. ** $p < .01$. * $p < .05$. df = 13,078. CPC = college preparatory curriculum. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. CCRPI = college and career ready performance index. EOC = end-of-course. HS = high school. GPA = grade point average. EFC = expected family contributions.

APPENDIX G:

VIF Analysis for Multicollinearity Before Data Manipulation

Table 32

VIF Analysis for Multicollinearity Before Data Manipulation

	VIF		
	First-fall GPA	First-year GPA	One-year Retention
Student Characteristics			
Gender	1.107	1.107	1.107
Race/Ethnicity	1.128	1.128	1.128
First Generation Status	1.108	1.108	1.108
HS Locale Group	1.127	1.127	1.127
Pre-college Characteristics			
HS GPA	2.682	2.682	2.682
Admissions Test Scores	2.031	2.031	2.031
AP Hours	1.453	1.453	1.453
CLEP Hours	1.007	1.007	1.007
IB Hours	1.021	1.021	1.021
Other Hours	1.003	1.003	1.003
CPC English	44.737	44.737	44.737
CPC Foreign Language	24.689	24.689	24.689
CPC Math	41.182	41.182	41.182
CPC Science	20.702	20.702	20.702
CPC Social Sciences	28.274	28.274	28.274
CCRPI Content Mastery	33.242	33.242	33.242
CCRPI Readiness	4.846	4.846	4.846
EOC English	13.956	13.956	13.956
EOC Mathematics	9.207	9.207	9.207
EOC Science	8.408	8.408	8.408
EOC Social Studies	9.648	9.648	9.648
Financial Situations			
EFC	1.222	1.222	1.222
GA HOPE Scholarship	2.052	2.052	2.052
Zell Miller Indicator	1.506	1.506	1.506
PELL Grant	1.500	1.500	1.500
Fed Sub. Loans	1.403	1.403	1.403
Fed Unsub. Loans	1.141	1.141	1.141
Other Loans	1.033	1.033	1.033
Major Grouping	1.020	1.020	1.020
Institutional Expenditures			
Academic Support	4.027	4.027	4.027
All Other	19.205	19.205	19.205
Institutional Support	20.543	20.543	20.543
Instruction	65.942	65.942	65.942
Public Service	28.914	28.914	28.914
Research	7.993	7.993	7.993
Student Support	5.203	5.203	5.203

Note. CPC = college preparatory curriculum. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. CCRPI = college and career ready performance index. EOC = end-of-course. HS = high school. GPA = grade point average. EFC = expected family contributions.

APPENDIX H:

Pearson Correlation Matrix after Data Manipulation

Table 33

Pearson Correlation Matrix after Data Manipulation

		1		2		3		4		5	
1	First-fall GPA	1.000	***								
2	First-year GPA	.927	***	1.000	***						
3	One-year Retention Status	-.386	***	-.441	***	1.000	***				
4	Gender	-.149	***	-.163	***	.080	***	1.000	***		
5	First Generation Status	-.032	***	-.039	***	.024		-.033	***	1.000	***
6	AP Hours	.219	***	.215	***	-.007		.075	***	-.043	***
7	CLEP Hours	.033	***	.033	***	.000		.002		-.006	
8	IB Hours	.039	***	.040	***	-.011		-.005		-.004	
9	Other Hours	.006		.007		-.008		-.012		.022	*
10	Major Grouping	.002		.008		.000		-.047	***	.019	*
11	GA HOPE Scholarship	.430	***	.452	***	-.183	***	-.111	***	-.010	
12	Zell Miller Indicator	.281	***	.285	***	-.045	***	.001		-.044	***
13	PELL Grant	-.097	***	-.109	***	.037	**	-.082	***	.262	***
14	Federal Sub. Loans	-.135	***	-.148	***	.045	***	-.080	***	.071	***
15	Federal Unsub. Loans	-.101	***	-.101	***	.027		-.037	***	-.045	***
16	Other Loans	-.066	***	-.061	***	.008		.003		-.012	
17	All Other	.035	***	.089	***	-.036	**	.025	**	.003	
18	Instruction	-.102	***	-.105	***	.057	***	-.080	***	.032	***
19	Student Services	-.061	***	-.084	***	.054	***	-.094	***	.028	**
20	Race/Ethnicity	-.019	*	-.022	*	-.009		-.018	*	.159	***
21	Graduating HS Locale	-.001		-.001		.023		-.014		.036	***
22	HS GPA	.488	***	.507	***	-.165	***	-.152	***	.007	
23	Admissions Test Scores	.257	***	.254	***	-.031	***	.156	***	-.107	***
24	EFC	.055	***	.070	***	-.043	***	.048	***	-.119	***
25	College Prep. Curriculum	-.005		.000		-.006		.016		.011	
26	Acad. & Inst. Support	-.005		-.002		.039	***	-.033	***	.038	***
27	Public Service & Research	.029	***	.044	***	-.071	***	.070	***	-.062	***
28	CM & Readiness Mean	.111	***	.117	***	-.040	***	.073	***	-.160	***
29	English (CMR)	.047	***	.062	***	-.011		.053	***	-.106	***
30	Math (CMR)	.083	***	.085	***	-.014		.062	***	-.093	***
31	Science (CMR)	.052	***	.057	***	-.020	*	.040	***	-.057	***
32	Social Studies (CMR)	.074	***	.067	***	-.012		.038	***	-.095	***

Note. *** *p* < .001. *** *p* < .001. ** *p* < .01. * *p* < .05. CMR Readiness Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

Table 33 (continued)

Pearson Correlation Matrix After Data Manipulation

		6		7		8		9		10	
1	First-fall GPA										
2	First-year GPA										
3	One-year Retention Status										
4	Gender										
5	First Generation Status										
6	AP Hours	1.000	***								
7	CLEP Hours	.048	***	1.000	***						
8	IB Hours	.000		.010		1.000	***				
9	Other Hours	.004		-.002		-.003		1.000	***		
10	Major Grouping	-.013		-.001		-.014		.002		1.000	***
11	GA HOPE Scholarship	.181	***	.020	*	.038	***	.015		.049	***
12	Zell Miller Indicator	.356	***	.028	**	.027	**	.013		.014	
13	PELL Grant	-.121	***	-.014		.005		.006		-.013	
14	Federal Sub. Loans	-.131	***	-.023	**	-.016		.012		-.004	
15	Federal Unsub. Loans	-.094	***	-.017		-.031	***	-.001		-.005	
16	Other Loans	-.038	***	-.009		-.005		-.001		.005	
17	All Other	.080	***	.010		.030	***	.020	*	-.023	**
18	Instruction	-.158	***	-.024	**	-.037	***	-.004		.002	
19	Student Services	-.133	***	-.010		-.037	***	-.022	*	.051	***
20	Race/Ethnicity	.014		.013		.049	***	.009		-.032	***
21	Graduating HS Locale	-.072	***	-.005		-.051	***	.004		.047	***
22	HS GPA	.249	***	.027	**	.036	***	.013		.081	***
23	Admissions Test Scores	.523	***	.055	***	.062	***	.006		-.016	
24	EFC	.088	***	.008		-.004		-.008		-.009	
25	College Prep. Curriculum	.022	*	-.005		.009		-.002		-.011	
26	Acad. & Inst. Support	-.035	***	-.004		-.003		.014		-.049	***
27	Public Service & Research	.063	***	.003		-.001		.005		.031	***
28	CM & Readiness Mean	.205	***	.019	*	-.036	***	-.001		-.009	
29	English (CMR)	.156	***	.039	***	-.004		.000		-.036	***
30	Math (CMR)	.161	***	.033	***	-.040	***	-.011		-.010	
31	Science (CMR)	.132	***	.024	**	-.027	**	.001		-.011	
32	Social Studies (CMR)	.149	***	.011		-.031	***	.000		-.031	***

Note. *** *p* < .001. *** *p* < .001. ** *p* < .01. * *p* < .05. CMR Readiness Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

Table 33 (continued)

Pearson Correlation Matrix After Data Manipulation

		11		12		13		14		15	
1	First-fall GPA										
2	First-year GPA										
3	One-year Retention Status										
4	Gender										
5	First Generation Status										
6	AP Hours										
7	CLEP Hours										
8	IB Hours										
9	Other Hours										
10	Major Grouping										
11	GA HOPE Scholarship	1.000	***								
12	Zell Miller Indicator	.270	***	1.000	***						
13	PELL Grant	-.077	***	-.101	***	1.000	***				
14	Federal Sub. Loans	-.133	***	-.121	***	.391	***	1.000	***		
15	Federal Unsub. Loans	-.096	***	-.085	***	.007		.281	***	1.000	***
16	Other Loans	-.093	***	-.027	**	-.064	***	.065	***	.095	***
17	All Other	.074	***	.031	***	-.034	***	-.062	***	-.059	***
18	Instruction	-.118	***	-.072	***	.108	***	.104	***	.116	***
19	Student Services	-.090	***	-.062	***	.131	***	.115	***	.100	***
20	Race/Ethnicity	-.063	***	-.063	***	.239	***	.113	***	-.011	
21	Graduating HS Locale	.098	***	.059	***	-.024	**	.001		.013	
22	HS GPA	.708	***	.445	***	-.089	***	-.154	***	-.135	***
23	Admissions Test Scores	.274	***	.475	***	-.241	***	-.228	***	-.138	***
24	EFC	.050	***	.035	***	-.364	***	-.313	***	-.027	**
25	College Prep. Curriculum	-.028	**	-.040	***	.014		.019	*	-.006	
26	Acad. & Inst. Support	-.052	***	-.036	***	.050	***	.025	**	.019	*
27	Public Service & Research	.099	***	.055	***	-.124	***	-.080	***	-.038	***
28	CM & Readiness Mean	.007		.032	***	-.314	***	-.247	***	-.096	***
29	English (CMR)	-.049	***	-.004		-.189	***	-.142	***	-.049	***
30	Math (CMR)	.000		.023	**	-.214	***	-.198	***	-.079	***
31	Science (CMR)	-.004		-.002		-.120	***	-.102	***	-.042	***
32	Social Studies (CMR)	-.014		.005		-.188	***	-.149	***	-.065	***

Note. *** *p* < .001. *** *p* < .001. ** *p* < .01. * *p* < .05. CMR Readiness Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

Table 33 (continued)

Pearson Correlation Matrix After Data Manipulation

		16		17		18		19		20	
1	First-fall GPA										
2	First-year GPA										
3	One-year Retention Status										
4	Gender										
5	First Generation Status										
6	AP Hours										
7	CLEP Hours										
8	IB Hours										
9	Other Hours										
10	Major Grouping										
11	GA HOPE Scholarship										
12	Zell Miller Indicator										
13	PELL Grant										
14	Federal Sub. Loans										
15	Federal Unsub. Loans										
16	Other Loans	1.000	***								
17	All Other	.020	*	1.000	***						
18	Instruction	-.010		-.534	***	1.000	***				
19	Student Services	-.012		-.414	***	.341	***	1.000	***		
20	Race/Ethnicity	-.018	*	.025	**	-.039	***	-.018	*	1.000	***
21	Graduating HS Locale	.011		-.030	***	.068	***	.088	***	-.128	***
22	HS GPA	-.068	***	.076	***	-.134	***	-.093	***	-.067	***
23	Admissions Test Scores	-.022	*	.096	***	-.283	***	-.208	***	-.107	***
24	EFC	-.012		.044	***	-.076	***	-.097	***	-.139	***
25	College Prep. Curriculum	-.005		.163	***	-.142	***	-.046	***	.038	***
26	Acad. & Inst. Support	-.028	**	.053	***	.474	***	-.081	***	.036	***
27	Public Service & Research	.034	***	.057	***	-.222	***	-.439	***	-.070	***
28	CM & Readiness Mean	.002		.142	***	-.221	***	-.244	***	-.107	***
29	English (CMR)	.005		.236	***	-.149	***	-.296	***	-.023	**
30	Math (CMR)	-.003		.071	***	-.153	***	-.140	***	-.096	***
31	Science (CMR)	.005		.057	***	-.145	***	-.122	***	-.020	*
32	Social Studies (CMR)	-.017		.001		-.156	***	-.214	***	-.025	**

Note. *** *p* < .001. *** *p* < .001. ** *p* < .01. * *p* < .05. CMR Readiness Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

Table 33 (continued)

Pearson Correlation Matrix After Data Manipulation

		21		22		23		24		25	
1	First-fall GPA										
2	First-year GPA										
3	One-year Retention Status										
4	Gender										
5	First Generation Status										
6	AP Hours										
7	CLEP Hours										
8	IB Hours										
9	Other Hours										
10	Major Grouping										
11	GA HOPE Scholarship										
12	Zell Miller Indicator										
13	PELL Grant										
14	Federal Sub. Loans										
15	Federal Unsub. Loans										
16	Other Loans										
17	All Other										
18	Instruction										
19	Student Services										
20	Race/Ethnicity										
21	Graduating HS Locale	1.000	***								
22	HS GPA	.153	***	1.000	***						
23	Admissions Test Scores	-.035	***	.382	***	1.000	***				
24	EFC	-.014		.034	***	.150	***	1.000	***		
25	College Prep. Curriculum	-.032	***	-.058	***	-.017		.007		1.000	***
26	Acad. & Inst. Support	-.030	***	-.089	***	-.147	***	-.023	**	.196	***
27	Public Service & Research	.005		.125	***	.209	***	.074	***	-.159	***
28	CM & Readiness Mean	-.121	***	-.050	***	.288	***	.208	***	.052	***
29	English (CMR)	-.201	***	-.140	***	.199	***	.146	***	.088	***
30	Math (CMR)	-.140	***	-.046	***	.233	***	.170	***	.054	***
31	Science (CMR)	-.193	***	-.052	***	.167	***	.094	***	.033	***
32	Social Studies (CMR)	-.178	***	-.086	***	.179	***	.110	***	.044	***

Note. *** $p < .001$. *** $p < .001$. ** $p < .01$. * $p < .05$. CMR Readiness Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

Table 33 (continued)

Pearson Correlation Matrix After Data Manipulation

		26		27		28		29		30	
1	First-fall GPA										
2	First-year GPA										
3	One-year Retention Status										
4	Gender										
5	First Generation Status										
6	AP Hours										
7	CLEP Hours										
8	IB Hours										
9	Other Hours										
10	Major Grouping										
11	GA HOPE Scholarship										
12	Zell Miller Indicator										
13	PELL Grant										
14	Federal Sub. Loans										
15	Federal Unsub. Loans										
16	Other Loans										
17	All Other										
18	Instruction										
19	Student Services										
20	Race/Ethnicity										
21	Graduating HS Locale										
22	HS GPA										
23	Admissions Test Scores										
24	EFC										
25	College Prep. Curriculum										
26	Acad. & Inst. Support	1.000	***								
27	Public Service & Research	-.510	***	1.000	***						
28	CM & Readiness Mean	-.037	***	.152	***	1.000	***				
29	English (CMR)	.112	***	.078	***	.577	***	1.000	***		
30	Math (CMR)	-.015		.063	***	.648	***	.536	***	1.000	***
31	Science (CMR)	-.015		.063	***	.481	***	.424	***	.471	***
32	Social Studies (CMR)	-.033	***	.101	***	.452	***	.348	***	.344	***

Note. *** *p* < .001. *** *p* < .001. ** *p* < .01. * *p* < .05. CMR Readiness Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

Table 33 (continued)

Pearson Correlation Matrix After Data Manipulation

		31		32	
1	First-fall GPA				
2	First-year GPA				
3	One-year Retention Status				
4	Gender				
5	First Generation Status				
6	AP Hours				
7	CLEP Hours				
8	IB Hours				
9	Other Hours				
10	Major Grouping				
11	GA HOPE Scholarship				
12	Zell Miller Indicator				
13	PELL Grant				
14	Federal Sub. Loans				
15	Federal Unsub. Loans				
16	Other Loans				
17	All Other				
18	Instruction				
19	Student Services				
20	Race/Ethnicity				
21	Graduating HS Locale				
22	HS GPA				
23	Admissions Test Scores				
24	EFC				
25	College Prep. Curriculum				
26	Acad. & Inst. Support				
27	Public Service & Research				
28	CM & Readiness Mean				
29	English (CMR)				
30	Math (CMR)				
31	Science (CMR)	1.000	***		
32	Social Studies (CMR)	.285	***	1.000	***

Note. *** *p* < .001. *** *p* < .001. ** *p* < .01. * *p* < .05. CMR Readiness Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

APPENDIX I:

VIF Analysis for Multicollinearity After Data Manipulation

Table 34

VIF Analysis for Multicollinearity After Data Manipulation

	VIF		
	First-fall GPA	First-year GPA	One-year Retention
Student Characteristics			
Gender	1.104	1.104	1.104
Race/Ethnicity	1.127	1.127	1.127
First Generation Status	1.107	1.107	1.107
HS Locale Group	1.126	1.126	1.126
Pre-college Characteristics			
HS GPA	2.675	2.675	2.675
Admissions Test Scores	2.028	2.028	2.028
AP Hours	1.451	1.451	1.451
CLEP Hours	1.006	1.006	1.006
IB Hours	1.020	1.020	1.020
Other Hours	1.002	1.002	1.002
College Prep Curriculum	1.144	1.144	1.144
CM & Readiness Mean	2.397	2.397	2.397
English (CMR)	1.902	1.902	1.902
Math (CMR)	1.971	1.971	1.971
Science (CMR)	1.451	1.451	1.451
Social Studies (CMR)	1.377	1.377	1.377
Financial Situations			
EFC	1.221	1.221	1.221
GA HOPE Scholarship	2.050	2.050	2.050
Zell Miller Indicator	1.504	1.504	1.504
PELL Grant	1.497	1.497	1.497
Federal Sub. Loans	1.403	1.403	1.403
Federal Unsub. Loans	1.141	1.141	1.141
Other Loans	1.031	1.031	1.031
Major Grouping	1.019	1.019	1.019
Institutional Expenditures			
Acad. & Inst. Support	3.008	3.008	3.008
All Other	1.885	1.885	1.885
Instruction	2.962	2.962	2.962
Public Service & Research	2.336	2.336	2.336
Student Services Support	2.223	2.223	2.223

Note. CMR Readiness Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. EFC = expected family contribution.

APPENDIX J:

Levene’s Test and Bartlett’s Test for Homogeneity of Variance

Table 35

Levene's Test and Bartlett's Test for Homogeneity of Variance

	Levene's Test				Bartlett's Test			
	F	*df*	*p*		K^2	*df*	*p*	
Student Characteristics								
Gender	73.71	1, 13,076	< .001	***	0.79	1	.375	
Race/Ethnicity	0.01	1, 13,076	.909		0.04	1	.850	
First Generation Status	29.33	1, 13,076	< .001	***	16.69	1	< .001	***
HS Locale Group	28.52	1, 13,076	< .001	***	6.69	1	.010	*
Pre-college Characteristics								
HS GPA	10.68	1, 13,076	.001	**	10.93	1	< .001	***
Admissions Test Scores	5.01	1, 13,076	.025	***	14.78	1	< .001	***
AP Hours	201.99	1, 13,076	< .001	***	48.59	1	< .001	***
CLEP Hours	0.01	1, 13,076	.921		443.22	1	< .001	***
IB Hours	6.34	1, 13,076	.012	***	56.79	1	< .001	***
Other Hours	3.61	1, 13,076	.573		980.30	1	< .001	***
College Prep Curriculum	1.33	1, 13,076	.249		1.50	1	.220	
CM & Ready Mean	0.18	1, 13,076	.671		0.95	1	.331	
English (CMR)	1.73	1, 13,076	.188		1.55	1	.214	
Math (CMR)	2.27	1, 13,076	.132		1.99	1	.159	
Science (CMR)	0.22	1, 13,076	.639		0.56	1	.454	
Social (CMR)	0.36	1, 13,076	.551		0.53	1	.465	
Financial Situations								
EFC	13.77	1, 13,076	< .001	***	0.89	1	.346	
GA HOPE Scholarship	962.56	1, 13,076	< .001	***	161.56	1	< .001	***
Zell Miller	110.29	1, 13,076	< .001	***	94.94	1	< .001	***
PELL Grant	52.96	1, 13,076	< .001	***	0.85	1	.358	
Fed Sub. Loans	64.94	1, 13,076	< .001	***	1.02	1	.312	
Fed Unsub. Loans	7.14	1, 13,076	.008	**	0.05	1	.823	
Other Loans	3.50	1, 13,076	.061		6.97	1	.008	**
Major Groupings	6.43	1, 13,076	.011	*	5.91	1	.015	*
Institutional Expenditures								
Acad. & Inst. Support	3.82	1, 13,076	.051		5.51	1	.019	*
All Other	31.29	1, 13,076	< .001	***	38.88	1	< .001	***
Instruction	16.03	1, 13,076	< .001	***	9.42	1	.002	**
Public Serv & Research	26.13	1, 13,076	< .001	***	14.17	1	< .001	***
Student Service Support	0.87	1, 13,076	.351		6.20	1	.013	*

Note. *** $p < .001$. ** $p < .01$. * $p < .05$. CM & Ready Mean = mean value of the CCRPI content mastery and readiness scores. Federal Sub. Loans = federal subsidized loans. Federal Unsub. Loans = federal unsubsidized loans. HS = high school. GPA = grade point average. CMR = mean value of the CCRPI content mastery and readiness scores. Acad. & Inst. Sup. = academic and institutional support expenditures. EFC = expected family contribution. Public Serv. & Rsch. = public service and research expenditures. Student Serv. Sup. = student services support expenditures.