The Predictability of Nontraditional Student Retention in the Technical College System of Georgia

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 Curriculum, Leadership, and Technology
of the Dewar College of Education and Human Services

March 2021

Brandy D. Taylor

M.B.A., Georgia Southern University, 2001
B.B.A., Georgia Southern University, 1996

This dissertation, "The Predictability of Nontraditional Student Retention in the Technical College System of Georgia," by Brandy D. Taylor, is approved by:

Dissertation Committee Chair

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

Committee Member

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

Committee Member

Michael J. Bochenko, Ed.D.
Assistant Professor of Leadership, Technology, and Workforce Development

Associate Provost for Graduate Studies and Research

Becky K. de Cruz, Ph.D., J.D.
Professor of Criminal Justice

Defense Date

3/15/21

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

State and federal governments regularly focus on improving student retention and completion in higher education as a means of increasing the skills of the workforce to better meet the challenges of a global economy. The findings of this research present a statewide picture of retention for nontraditional students in the Technical College System of Georgia and generalizations could be used to specifically improve processes and procedures on how colleges recruit and respond to this growing and diverse student population. With a specific focus on nontraditional students in diploma and certificate programs, the outcomes of this research will allow decision-makers to consider how student factors, and the relationship between those factors, influence nontraditional student progression in order to make informed decisions on how to better serve the needs of this specific student population.

The purpose of this nonexperimental, ex post facto, correlational study was to examine the predictability of academic factors (student GPA and program type), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and environmental factors (Pell eligibility, single parent status, displaced homemaker status) on the retention of nontraditional students enrolled in diploma and certificate programs in the Technical College System of Georgia. To do so, this study addressed which prediction model, out of two data modeling approaches (logistic regression and linear discriminant analysis) and three data mining approaches (classification tree, random forest, and support vector machine models), best predicts whether a student was retained or not retained.

The predictor variables GPA, programs related to Transportation and Logistics, female students, Black students, and Pell eligibility were influential in students being retained. Being out of high school for five years or more and being enrolled in Cyber, Engineer, or Healthcare programs or Industrial Technology programs were influential predictors of students not being retained. The support vector machine will generate an accurate classification model based on the goal of correctly identifying students who will not be retained so adequate assistance and resources can be provided to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

Thank you to the members of my dissertation committee for their assistance through the research, writing, and review of this dissertation and for their genuine interest in this study. I especially want to thank Dr. Brockmeier for his thoughtful insights, honest feedback, wisdom, advice, and words of encouragement as the chair of my dissertation committee. Thank you to each of my professors for laying the foundation of this work and instilling in me the confidence to strive for excellence in my education.

To Dr. Futch, thank you for constantly supporting me by serving as my personal counselor and therapist as well as my number one cheerleader. To Dr. Foley, thank you for always supporting me and being a sounding board on the good days and the bad. You rock! To my mom who probably didn't understand why I set out on this journey, I owe my work ethic to you and only you. You taught me fierce independence and determination and I will forever be grateful for your love and support.

To Chad, thank you for your patience, humor, and support. There is absolutely no way I could have completed this journey without you. Thank you for loving me through my doubts and fears. To my beautiful daughters, Addison and Avery, you are blessings from God and I will forever cherish the inspiration you have provided me with. You both supported and loved me in your own unique way and for that I am grateful. I pray that you will never stop learning. All the honor and glory for this accomplishment goes to my Lord and Savior, Jesus Christ.

Chapter I

INTRODUCTION

The mission of the Technical College System of Georgia (TCSG) is to build a well-educated workforce for Georgia. The Technical College System of Georgia has multiple partnerships with the Georgia Department of Economic Development, like the High Demand Career initiative, the Trade Five program (formerly Go Build Georgia), and the Complete College Georgia initiative which support its mission (Wilson, Epps, Tanner, Gordon, & Sig, 2014). The Georgia State Workforce Investment Board (2013) indicates these initiatives are critical in Georgia where high growth is projected in key strategic industries across the state over the next several years. State-funded award programs like the HOPE Grant and the HOPE Career Grant (formerly known as the Strategic Industries Workforce Development Grant) have been specifically designated for in-demand diploma and certificate programs to create a pipeline of skilled workers for Georgia employers (Georgia Student Finance Commission, n.d.).

The Technical College System of Georgia along with the University System of Georgia (USG) aims to develop an educated workforce while at the same time focusing on college retention and completion (Complete College Georgia, 2011). The attainment goals set by state and national leaders cannot be met unless significantly more adults and other nontraditional students return to higher education and complete a degree or credential (Complete College America, n.d.). Conventional retention strategies aimed at traditional students may not work with today's college students.

Community colleges provide a path to postsecondary education for a diverse student population made up mostly of students characterized as nontraditional (American Association of Community Colleges, 2015; Kim, 2002). The majority of college students today are part-time students and full-time providers (Lumina Foundation, 2015). These students are older, busier, more diverse, and more financially strained. The National Center for Education Statistics (1996) broadly defines nontraditional students by seven characteristics: delayed enrollment in postsecondary education from high school, financial independence, full-time employment, enrolled part-time, has dependents, is a single parent, and earned a General Educational Development (GED®) diploma instead of a high school diploma. Several other definitions of this student population exist which adds to the challenge of concisely labeling this group. Bean and Metzner (1985) noted that attrition models typically share the assumption that postsecondary students are young (24 years old or younger), reside on campus, and take coursework full-time. Based on this assumption, the researchers defined nontraditional students based on age, residence, employment, and being enrolled in non-degree occupational programs. Jones and Watson (1990) defined nontraditional students as being women, minorities, adults, and enrolled part-time in their study on high risk students.

With the diversity of nontraditional students' demographic and socioeconomic characteristics, the Advisory Committee on Student Financial Assistance (ACSFA) indicates this population consists of many subgroups, each with unique circumstances, educational needs, and goals (Advisory Committee on Student Financial Assistance, 2012). Pelletier (2010) reports nontraditional is the new traditional based on the demographics of current postsecondary students. In fall 2015, almost 6.3 million students

were enrolled in public, two-year colleges (Ginder, Kelly-Reid, & Mann, 2017a). Of those, 2.3 million students were full-time students and almost 4 million students were part-time (Ginder et al., 2017a). In Georgia, 70% of technical college students are enrolled part-time and students 25 years of age or older made up 40% of technical college enrollment in 2016 (Lee, 2017). Over time, the characteristics of these students have changed and will likely continue to change (Peters, Hyun, Taylor, & Varney, 2010).

Many of these students have external demands unlike their traditional counterparts (Shapiro et al., 2016). A nontraditional student maybe a younger, single parent with a full-time job or a 45-year-old attending college for the first time (Peters et al., 2010). A nontraditional student is less likely to persist and complete degree programs than a full-time traditional student (Advisory Committee on Student Financial Assistance, 2012). The majority of postsecondary students are no longer enrolling in college full-time immediately after high school (Petty, 2014; Reeves, Miller, & Rouse, 2011; Shapiro et al., 2016). Although enrollment shifts may be occurring, Reeves et al. (2011) argued nontraditional college students consistently represent the majority of undergraduates at postsecondary educational institutions. The reality, however, is higher education is not structured to serve this population adequately (Advisory Committee on Student Financial Assistance, 2012). Therefore, the views we have of nontraditional students and the decisions we make as academic administrators must frequently be revisited to retain them.

Although there are established models for retention and attrition of traditional students providing concepts and understandings broadly applied to nontraditional students, few studies specifically address the demographics shifts of this population or

their needs (Monroe, 2006). As a result, conventional postsecondary measures of student achievement, such as retention rates for first-time, full-time degree-seeking cohorts, are not enough to understand the specific opportunities and risks which define nontraditional students' academic careers (Shapiro et al., 2016). Nontraditional students bring with them significant life experiences and are often motivated learners with strong opinions and perspectives (Chen, 2017). Conversely, their diverse characteristics, many times seen as strengths, can represent challenges and risks. Chen (2017) states many nontraditional students are isolated and alienated by the traditional youth-centric environments of colleges and universities. Many times these students are torn between their employee and student identity (Keith, 2007). Chen (2017) calls this the competing nature of life roles which accompany adulthood.

One pervasive issue in understanding retention at community colleges is the lack of consistency in how student retention is defined. According to Wild and Ebbers (2002), most research in this area is based on traditional-age students in the residential settings of universities, which does little for community colleges. According to the National Center for Education Statistics (1996), retention measures the rate at which students persist in their educational program at an institution. For two-year institutions, this is the percentage of first-time degree or certificate-seeking students from the previous fall who either reenrolled or graduated by the current fall (NCES, 1996). Within Georgia, there is little surprise the University System of Georgia and the Technical College System of Georgia differ in the definition of retention as well. For USG institutions (both two and four years), a student is considered to be retained if enrolled in a USG institution in the same academic term one year later (University System of Georgia, n.d.). Although the

definition of the cohort may vary according to the subject of interest, the most common cohorts studied are first-time, full-time, degree-seeking freshman students (University System of Georgia, n.d.). Within the Technical College System of Georgia, the definition of retention used to compare colleges within the system is a beginning fall cohort student from the previous year. The number retained is defined as those who graduated from the same college or a different college, or were still enrolled in the same college or a different college.

A combined focus on increasing postsecondary education attainment and improving college completion rates comes from federal agencies, policymakers, and higher education. Monroe (2006) asserted the complex, dynamic nature of nontraditional students requires continuous examination and refinement of our understanding of this population's changing demographics concerning attrition. If colleges do nothing to improve the odds of retention for nontraditional students, a large segment of our population and the majority of college students will continue on the path to failure (Chen, 2017). The changing characteristics of nontraditional students need to be understood before retention efforts in the community and technical colleges are effective (Ashar & Skenes, 1993).

Statement of the Problem

Understanding the shifting characteristics of college students is critical to curriculum, program, and policy design (Reeves et al., 2011). Higher education and most financial aid programs are not structured to serve the nontraditional student population adequately (Advisory Committee on Student Financial Assistance, 2012). Many nontraditional students bring a wealth of life experiences into learning situations which

may enhance or prevent learning (Chen, 2017). These students bring with them different expectations and different needs (Ross-Gordon, 2011). Failure to track these expectations, nontraditional trends, and to provide accurate information may result in educational administrators misunderstanding the needs of 21st-century undergraduates and/or misappropriating educational resources (Reeves et al., 2011). Nationally representative data which tracks nontraditional college enrollment and persistence does not exist (Advisory Committee on Student Financial Assistance, 2012). Although several studies focus predominantly on traditional students in associate or bachelor's degree programs, we do not have an understanding of factors related to college retention for nontraditional students seeking only a diploma or certificate. The Technical College System of Georgia does not monitor the retention of this student population.

The future of Georgia's workforce depends on the diversity, adaptability, and broad-based talents and skills students acquire through quality higher education. By 2020, 65% of Georgia's jobs will require some level of postsecondary education and 22% will require a bachelor's degree (Complete College Georgia, 2011). In 2015, Georgia produced fewer adults (ages 25–64) with postsecondary credentials than needed, leaving a gap of 189,000 workers with some college education, an associate's degree, or certificate (Lee, 2017). There are simply not enough high school and traditional college students to create the educated workforce required for the 21st-century economy (Pingel, Parker, & Sisneros, 2016). Research of nontraditional student retention from the first to second year is specifically needed for diploma and certificate students as workforce initiatives continue to promote skilled trade education across the state. Ambitious college completion goals call for equally ambitious policies which go beyond a focus on

traditional students (Pingel et al., 2016). A careful review of retention models and theories through the lens of nontraditional students can not only help colleges develop policies and procedures to facilitate student retention, but can align Georgia's nontraditional students with Georgia's workforce needs and requirements.

Purpose of the Study

Although there are established models for retention and attrition of traditional students providing concepts and understandings broadly applied to nontraditional students, few studies have specifically addressed the demographic shifts of this population or their needs (Monroe, 2006). While there is prolific literature on the challenges and struggles facing nontraditional students, very little literature focuses on how the student's unique characteristics contribute to retention specific to the community and technical college environment. The purpose of this study was to examine the predictability of academic, background, and environmental factors such as Pell eligibility, single parent status, displaced homemaker status, age, race or ethnicity, gender, high school diploma type, high school graduation date, student grade point average (GPA), and program type on the retention of nontraditional students enrolled in diploma and certificate programs in the Technical College System of Georgia. To do so, this study focused on multiple prediction models used to predict whether a student was retained or not retained.

Research Questions

Each of the following research questions focuses on the predictability of academic, background, and environmental factors on the retention of nontraditional

students enrolled in diploma and certificate programs in the Technical College System of Georgia.

1. Are environmental factors, background factors, and academic integration components significant predictors of nontraditional student retention for certificates or diplomas?
    a. Are environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) significant predictors of nontraditional student retention for certificates 9–17 credit hours in length?
    b. Are environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) significant predictors of nontraditional student retention for certificates 18–36 credit hours in length?
    c. Are environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) significant predictors of nontraditional student retention for diplomas 37–48 credit hours in length?

d. Are environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) significant predictors of nontraditional student retention for diplomas 49–59 credit hours in length?

2. Does one of the selected statistical procedures generate a more accurate classification model based on Cohen's Kappa, ROC curves, and sensitivity and specificity by certificate or diploma type?

Research Methodology

A nonexperimental, ex post facto, correlational research design was used in this study. In ex post facto research, the researcher predicts the possible causes behind an effect which has already occurred (Ary, Jacobs, Sorensen, & Razavieh, 2006). Archival data obtained from the Technical College System of Georgia were retrospectively analyzed to measure first-year retention. The use of archival data makes the manipulation of the variables unlikely and unethical (Bordens & Abbott, 2011). Therefore, a nonexperimental, ex post facto research design was more appropriate for this study as the independent predictor variables will not be manipulated. There were two continuous predictor variables representing background and academic factors (age and GPA). There were five dichotomous variables (gender, code for high school graduation date, single parent indicator, displaced homemaker indicator, Pell eligibility indicator), two nominal variables (race and HOPE program of study), and one ordinal variable (high school diploma type) representing background, environmental, and academic factors. Because

the goal was to predict values on a binary outcome variable, the researcher attempted to identify which prediction model, out of two data modeling approaches and three data mining approaches, best predicts whether a student was retained or not retained.

The target population included students identified as nontraditional at each of the 22 technical colleges in Georgia. The accessible population included first-time students identified as nontraditional at each of the 22 technical colleges in Georgia who were enrolled in one of 17 program areas defined by the HOPE Career Grant Program. The expected cohort size was approximately 8,000 students per cohort for a total of 16,000 students. Program areas were subdivided into four distinct groups of certificates with 9–17 credit hours, certificates with 18–36 credit hours, diplomas with 37–48 credit hours, and diplomas with 49–59 credit hours. Students were classified as nontraditional if they meet all three of the following criteria:

- First-Time - Beginning student (queried from Banner field *Student Type*)
- Age - 25 years old or older (calculated from Banner field Date of Birth)
- Enrollment Status - Part-time (calculated from Banner field Earned Hours)

A combination of descriptive and inferential statistics was used in the analysis of the data. Descriptive statistics such as the mean, median, and standard deviation (*SD*) were calculated for the continuous variables of age and GPA. To identify nontraditional student characteristics which best predict first-year retention, a statistical learning approach was applied in this study. Statistical learning refers to tools and techniques for understanding data (James, Witten, Hastie, & Tibshirani, 2013). Supervised statistical learning involves building a statistical model for predicting, or estimating, an output based on one or more inputs (James et al., 2013). By testing multiple statistical models to

illustrate the classification power of these models, researchers are better equipped to provide timely data and information to key decision-makers (Knowles, 2014). Instead of identifying one single best model, the researcher evaluated several models to identify the most accurate predictions on future student cohorts.

To address Research Question 1, the coefficients, standard error, odds ratios, p-values, and confidence intervals were evaluated for each predictor. Predictors were considered statistically significant at the .05 level. Model-specific procedures were employed iteratively to arrive at the final model of significant predictor variables. Because there are many different metrics available to evaluate prediction models, the researcher utilized three common metrics to evaluate binary classification datasets. Knowles (2014) stated because of the complexity of the model building process, every aspect of the modeling process is crucial in balancing the tradeoff between accuracy and complexity. Therefore, the accuracy metric, Cohen's Kappa statistic, the receiver operating characteristics (ROC) curve metric (area under the curve), sensitivity, and specificity were used in various R packages including tidyverse and tidymodels packages to evaluate the accuracy of each model and to address Research Question 2. The tidyverse, a collection of R packages designed for data preparation and data analysis, contains a subset of packages specifically focused on data modeling (Kuhn & Silge, 2021). Likewise, the tidymodels framework is a collection of packages for modeling and machine learning using tidyverse principles (Kuhn & Silge, 2021).

Significance of the Study

State and federal governments have focused on improving student retention and completion in higher education as a means of increasing the skills of the workforce to

better meet the challenges of a global economy (Hirschy, Bremer, & Castellano, 2011). These efforts reflect a shift toward acknowledging the distinctive nature of students in the community college setting (Hirschy et al., 2011). Community colleges should focus on the unique skills and abilities of students and their commitment to complete a program (Monroe, 2006). A specific understanding of the influences and characteristics nontraditional students bring with them to the classroom will give decision-makers a better understanding of what is needed in their colleges to support and retain this population. Whether or not these characteristics of nontraditional students are barriers or opportunities, understanding this student population is critical (Reeves et al., 2011).

This data presented a statewide picture of retention for nontraditional students in TCSG and generalizations can be used to specifically improve processes and procedures on how colleges recruit and respond to this growing and diverse student population. With a specific focus on nontraditional students in diploma and certificate programs, the outcomes of this research will allow decision-makers to consider how student factors, and the relationship between those factors, influence nontraditional student progression from year 1 to year 2 to make informed decisions on how to better serve the needs of this specific student population. Results of this study will support and enhance statewide initiatives such as the High Demand Career initiative, the Trade Five program (formerly Go Build Georgia), and the Complete College Georgia initiative thus, ultimately leading to a more educated and trained workforce geared toward industries in Georgia where there are more jobs available than there are skilled workers to fill them (Wilson et al., 2014).

Conceptual Framework of the Study

The guiding conceptual models for this study were Bean and Metzner's (1985) Model of Nontraditional Undergraduate Student Attrition and Hirschy, Bremer, and Castellano's (2011) Conceptual Model for Student Success in Community College Occupational Programs. Unlike previous models which addressed students more generally or focused on four-year degree students, Bean and Metzner's model was the first conceptual model to specifically address the nontraditional student experience in higher education (Bean & Metzner, 1985). Students' social integration into the college community was a key aspect of other theoretical models of that time. Bean and Metzner felt another model was needed since most nontraditional students were not often socially integrated into the college (Bean & Metzner, 1985; Hirschy et al., 2011). Since a large number of technical college students are nontraditional under Bean and Metzner's definition, this model was relevant to this proposed study.

Similar to traditional student models, Bean and Metzner's model of attrition is concerned with student institution fit, that is, students' academic and social integration into the institution (Monroe, 2006). The Bean and Metzner (1985) model proposed four sets of variables affecting the dropout decision: academic performance, intent, background and defining variables (e.g., age, gender, race or ethnicity), and environmental variables not controlled by the institution (e.g., finances, outside encouragement). The variables identified in the Bean and Metzner model (academic, background, and environmental) were used in part to guide the selection of variables for this study. Specifically, the variables selected for this study were identified using Bean and Metzner's model as described in Figure 1. Bean and Metzner (1985) suggest the

structure of the model was meant to be flexible and future researchers were encouraged to include factors not included in the original model, as well as concentrate their efforts on specific parts of the model.

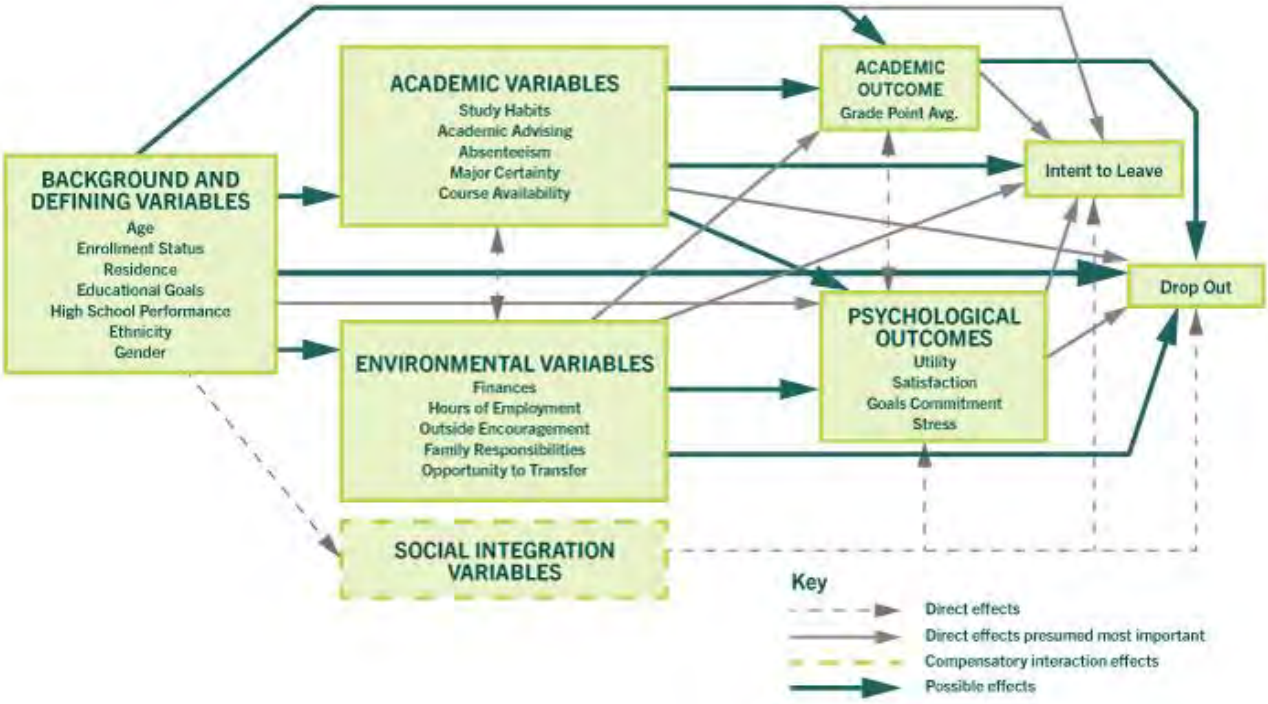

Figure 1. Bean and Metzner's model of nontraditional undergraduate student attrition (1985).

Unlike previous models, Hirschy et al.'s (2011) model is focused specifically on career and technical education (CTE) students and suggests students pursuing occupational associate's degrees or certificates differ from those students seeking academic majors at two-year institutions. The model, as described in Figure 2, has four sets of interrelated constructs: student characteristics, college environment, local community environment, and student success outcomes. Hirschy et al. (2011) assert that student characteristics influence and are influenced by the ways individuals interact with the college and local communities. Therefore, student characteristics both directly and

indirectly influence student success. The model acknowledges community college students are members of multiple communities—on and off campus—which affect their educational goals and experiences (Hirschy et al., 2011). The authors suggest the introduction of a career integration variable, promoted the collection and tracking of student educational goals, and expanding traditional student success measures to better reflect the experiences of CTE students (Hirschy et al., 2011).

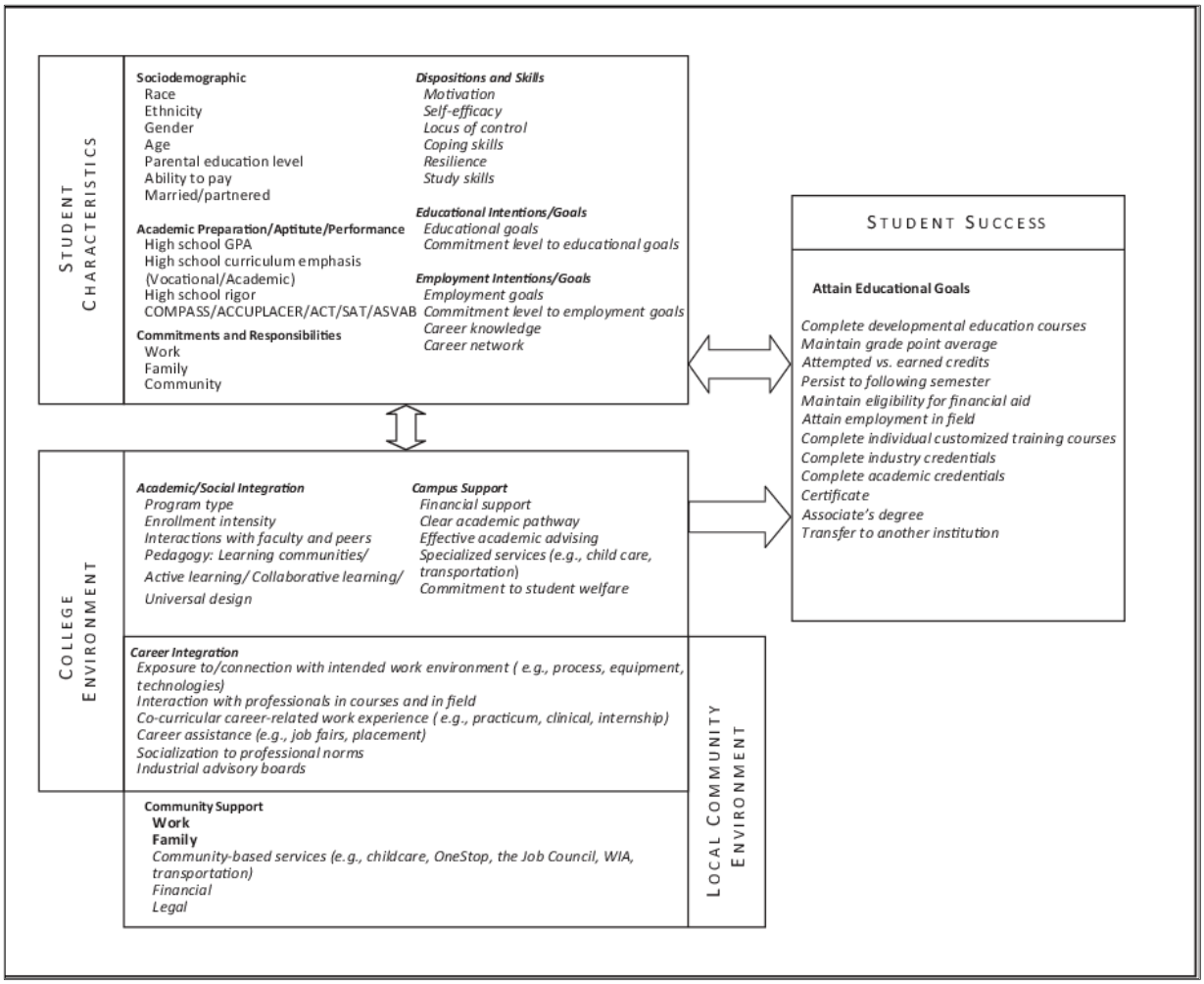

Figure 2. Hirschy, Bremer, and Castellano's conceptual model for student success in community college occupational programs (2011).

Together Bean and Metzner's (1985) nontraditional model and Hirschy, Bremer, and Castellano's community college model (2011) provided a solid basis for this study with a

combined focus on nontraditional students and occupational programs at community colleges.

Limitations of the Study

The data for this research study was not collected to answer the researcher's specific research questions. By only using historical student-level data, this study was limited to variables only available through the Technical College System of Georgia Data Center. Additional variables identified in the literature review were not available for analysis, and therefore not included in the study. These variables included financial independence, employment status, marital status, and having dependents. Additional variables may provide better results to assist colleges in developing policies and procedures to facilitate nontraditional student retention. Based on the results of this study, future researchers could create a more comprehensive model of all the factors influencing nontraditional student retention.

The accuracy of data extracted from each college-level student information system was not guaranteed. Simonton (2003) suggested historical data may sometimes contain errors and are not always as reliable as more conventional data sources. A Banner data custodian is ultimately responsible for the integrity and reliability of the data contained in their functional area. Banner security ensures only authorized users can view and/or update specific data, forms, tables, processes, and reports as required by the user's role. While the majority of data errors in files extracted from a Banner database can be attributed to human error, many errors are mitigated through the design of the Banner user interface. Meaning many data fields used in this study only accept specific values, thereby decreasing the chance of data entry error.

The cohort for this study was limited to nontraditional students who were enrolled for the first time at any of the technical colleges in Georgia and were not high school students. First-time students identified as special admit or learning support were not included in this study as they cannot receive federal financial aid. Two independent variables, single parent and displaced homemaker, were self-reported by students. A limitation of self-reported data is the accuracy of responses cannot be determined.

Definition of Terms

For this study, it was necessary to define specific terms to provide clarification, to further define the scope and focus of the study, and to avoid confusion.

- *Beginning Student* - First-time, first-year student: A student attending any institution for the first time at the undergraduate level. This includes students enrolled in the fall term who attended college for the first time in the prior summer term. It also includes students who were previously coded as an H, but are attending a technical college for the first time as non-high school students (Knowledge Management System, 2014).
- *CIP Codes (Classification of Instructional Programs)* - Classification of Instructional Programs is the accepted federal government statistical standard on instructional program classifications and is used in a variety of education information surveys and databases (Knowledge Management System, 2014).
- *Cohen's Kappa Statistic* - A statistic which takes into account the accuracy generated simply by chance using an observed accuracy and an expected accuracy based on the marginal totals of a confusion matrix (Kuhn & Johnson, 2013).

- *Collinearity* - Refers to the situation in which two or more predictor variables are closely related to one another (James et al., 2013).
- *Community Colleges* - A postsecondary institution which offers programs of at least two but less than four years' duration. Community colleges can include occupational and technical programs and academic programs of less than four years, but do not include bachelor's degree-granting institutions. Based on this definition, and for this research, colleges within the Technical College System of Georgia are considered community colleges.
- *Diploma* - The range of semester credit hours required for graduation with a diploma is typically 37 to 59. The models shall require diploma programs to be organized in general education and occupational courses. Diploma programs shall be composed of courses listed in the system-wide Catalog of Courses. Only general education courses numbered 1000 or above shall be credited toward diploma requirements (Technical College System of Georgia, 2018).
- *Displaced Homemaker* - An adult, who is divorced, widowed, separated, or has a disabled spouse and is unemployed or underemployed. The displaced homemaker is also one who has worked primarily without pay to care for a home and family and has diminished marketable skills (Southeastern Technical College, n.d.).
- *Economically Disadvantaged* - A student is reported as economically disadvantaged if the student is a needs-based financial aid recipient (Pell or TANF) (Knowledge Management System, 2014).

- *Full-Time Student* - A student is defined as full-time for a semester if they enrolled in 12 or more credit hours within the semester. A student is defined as full-time for a fiscal year if they were full-time for at least one semester within the year (Knowledge Management System, 2014).
- *Grade Point Average* - Semester or term GPA is computed by dividing the number of credit hours into the courses attempted for the semester into the number of quality points earned on those hours scheduled for the semester (Knowledge Management System, 2014).
- *Graduate* - A graduate is a student who received at least one award (Technical Certificate of Credit, Diploma, and/or Associate Degree). This is used to report an unduplicated count of graduates for the college, counting each student who received an award once; regardless of how many awards they received (Knowledge Management System, 2014).
- *Leaver* - A student who was enrolled in a major of study and not coded as special admit or transient, did not graduate from that major, and is no longer enrolled in the major for two consecutive terms. There are two exceptions: students who enrolled in the summer, did not enroll in fall, and returned in the spring with the same major are not considered leavers; students enrolled in the fall did not enroll in spring, and returned in the following summer (of the next fiscal year) with the same major, are not considered leavers (Knowledge Management System, 2014).
- *Near Zero Variance* - Variables with extremely low variances because they usually consist of a single unique value (Kuhn & Johnson, 2013).

- *Nontraditional Student* - First-time students who are 25 years of age or older and are enrolled part-time (Advisory Committee on Student Financial Assistance, 2012; Bean & Metzner, 1985; Cleveland-Innes, 1994; Hirschy et al., 2011; Hurtado, Kurotsuchi, & Sharp, 1996; NCES, 1996; Nora, Barlow, & Crisp, 2005).
- *Part-Time Student* - A student is defined as part-time for a semester if they enrolled in less than 12 credit hours within the semester. A student is defined as part-time for a fiscal year if they were part-time in each semester they were enrolled within the year (i.e. they were never full-time) (Knowledge Management System, 2014).
- *Pell Financial Aid* - The Higher Education Act of 1965, Title IV, Part A, Subpart I, as amended. Provides grant assistance to help meet education expenses to eligible undergraduate postsecondary students with a demonstrated financial need (Knowledge Management System, 2014).
- *Persistence* - The continuation of a student's postsecondary education which leads to graduation (National Student Clearinghouse, 2017a).
- *Provisional Admit* - Students admitted provisionally require no more than one learning support course in each area of deficiency (English and/or math and/or reading); students may begin taking occupational courses concurrently (Knowledge Management System, 2014).
- *Receiver Operating Characteristics (ROC) Curve Metric* - Provides a graphical representation of possible cut points (predictions) and computed false positive (1-specificity) and true positive rates (sensitivity) for a range of

values. Estimates of the area under the curve (AUC) indicate the overall performance of a classifier summarized over all possible thresholds (James et al., 2013).

- *Regular Admit* - A student is granted regular admission to a specific program if they have met the minimum admissions requirements for the program and its award level (Knowledge Management System, 2014).
- *Retention* - Measured by the student returning to the institution they attended the previous year (National Student Clearinghouse, 2017a).
- *Sensitivity* - Measures the percentage of cases in which retention is predicted correctly (James et al., 2013).
- *Single Parent* - An individual who is unmarried or legally separated from a spouse and has a minor child or children for whom the parent has either custody or joint custody (Southeastern Technical College, n.d.).
- *Specificity* - The percentage of cases in which not being retained or attrition is predicted correctly (James et al., 2013).
- *Technical Certificate of Credit (TCC)* - The range of semester credit hours required for graduation is 9-36. The technical certificate may be used to provide programs in areas of specialization which do not require study of sufficient length to award a diploma or degree or to add on areas of specialization after the completion of a diploma or degree (Technical College System of Georgia, 2018).

Organization of the Study

Chapter 1 highlighted the importance of community colleges understanding the shifting characteristics of today's college students. The chapter conveyed the purpose and significance of studying nontraditional student retention in alignment with Georgia's workforce needs and requirements. Chapter 2 will provide a review of community colleges, community college funding, nontraditional students in today's community colleges, nontraditional student retention, and relevant student retention theories, models, and frameworks. Chapter 3 will present the research design, the data collection procedures, variables, instruments, and the data analysis procedures to be used in this study. Data screening and preprocessing approaches will be discussed in Chapter 4. Chapter 4 also describes the model training and the statistical significance of the variables in each model, as well as the accuracy of the various classification models. Chapter 5 will contain a discussion of the results and limitations of this study. In conclusion, Chapter 5 will offer suggestions for future research and the implications of the study.

Chapter II

LITERATURE REVIEW

Colleges and universities craft the state's future workers, entrepreneurs, and leaders (Lee, 2017). In particular, community colleges play an integral role in expanding postsecondary education opportunities (National Student Clearinghouse, 2017b). The American Association of Community Colleges (AACC) (2015) defines a community college as a two-year, associate degree-granting institution. The National Center for Education Statistics (NCES) defines a two-year institution as a postsecondary institution which offers programs of at least two but less than four years' duration (Ginder et al., 2017a). The missions, philosophies, and student populations of community colleges differ from those of four-year institutions (Seidman, 1993). Community colleges can include occupational and technical programs and academic programs of less than four years, but do not include bachelor's degree-granting institutions where the baccalaureate program can be completed in three years (Ginder et al., 2017a). Community college students can pursue career and technical education (CTE) in health care, manufacturing, and personal and consumer services; academic education such as liberal arts; or STEM programs which include both occupational and academic subjects, such as math, science, and computer and information technology (Horn & Li, 2009). Many states in the U.S. have combined community and technical college systems (Kentucky, Louisiana, North Carolina, Washington, and West Virginia). Other states structure technical and community colleges as technical divisions. Georgia has the University System of Georgia

and the Technical College System of Georgia. Within the University System of Georgia, there are research universities, comprehensive universities, state universities, and state colleges. Of the state colleges in Georgia, Atlanta Metropolitan State College is the only college that meets the definition of a community college. Based on these definitions, and for this research, colleges within the Technical College System of Georgia are considered community colleges.

Community colleges serve as an access point into postsecondary education for many traditional and nontraditional students (Brooks-Leonard, 1991; Fain, 2012; Wyner, 2014). The occupational programs at community colleges hold the promise of a better life for many students, including those directly out of high school and those who are returning to school from the workforce (Hirschy et al., 2011). Hirschy et al. (2011) suggested employers identify community colleges as the primary institutions for licensure and certification as well as imparting soft skills like critical thinking and problem-solving. Carnevale and Desrochers (as cited by Hirschy et al., 2011, p. 300) reported as more complex or specialized occupations develop, community colleges are looked upon to provide the certification training. Community colleges provide additional education and job skills training to those impacted by unemployment during times of economic hardship (National Student Clearinghouse, 2017b).

Community college students have characteristics and needs distinctive from traditional, residential students enrolling in four-year universities (Crisp & Mina, 2012; Fike & Fike, 2008). The reasons students attend community colleges vary from academic transferability to workforce or technical training (Kim, 2002). In fall 2015, almost 6.3 million students were enrolled in public, two-year colleges (Ginder et al., 2017a). Of those, 2.3 million students were full-time students and almost 4 million students were

part-time (Ginder et al., 2017a). Juszkiewicz (2017) noted approximately 50% of all African American, Native American, and Hispanic college students are enrolled at community colleges. This student population consists primarily of commuter students, where 35% of first-time enrollees work full time in contrast to 11% in public four-year institutions (Juszkiewicz, 2017). The average age of a community college student is 28 years old, 17% are single parents, and 34% receive federal Pell grants (Radwin et al., 2018). Seventy percent of technical college students in Georgia are enrolled part-time and students 25 years of age or older made up 40% of technical college enrollment in 2016 (Lee, 2017). In contrast to the University System of Georgia, the Technical College System of Georgia provides workforce focused instruction in addition to adult education and continuing education training. It includes 22 technical colleges with 85 campuses throughout the state (Technical College System of Georgia, 2017). Total Georgia technical school enrollment reached 133,455 students in 2016, higher than in Georgia's research universities (Lee, 2017). Technical colleges serve students who are diverse in race, ethnicity, age, and income. Although almost all technical colleges serve both rural and urban areas of Georgia, a third of Georgia's technical colleges serve predominantly rural areas (Lee, 2017). Technical colleges in Georgia serve a larger population of adult and low-income students than the university system (Lee, 2017).

On average, 28% of community college revenues come from tuition and 33% come from state agencies (Baum, Ma, Pender, & Welch, 2017). The remainder of the revenues come from federal or local sources. In Georgia, the state's 2018 budget allots $3.4 billion for higher education, with $322 million designated to the Technical College System of Georgia (Lee, 2017). Technical college enrollment, like community college

enrollment, rises and falls with the economy. The recession which began in 2007 led to a dramatic spike in community college enrollment (Juszkiewicz, 2017). TCSG enrollment was impacted as state funding fell and enrollment increased because displaced workers were trying to upgrade their skills in technical certificate and diploma programs (Lee, 2017). From 2006 to 2015, core revenues from state funding decreased from 55% to 37% (Lee, 2017). Georgia's higher education institutions are funded directly by the state and indirectly through student financial aid, including federal loans, the Pell grant, and HOPE (Georgia Student Finance Commission, n.d.). Georgia appropriates money from lottery proceeds for HOPE scholarships and grants each year. The major HOPE programs are (Georgia Student Finance Commission, n.d.):

- HOPE Scholarship – Partial tuition for bachelor's or associate's degree programs at public and private colleges and universities.
- Zell Miller Scholarship – Full tuition for bachelor's or associate's degree programs at public colleges and universities, partial tuition at private colleges.
- HOPE Grant – Partial tuition for certificate or diploma programs in technical colleges.
- Zell Miller Grant – Full tuition for certificate or diploma programs in technical colleges.
- HOPE Career Grant – Partial tuition for specific certificate or diploma programs.

In the Technical College System of Georgia, students can receive the HOPE or Zell Miller Grant and the HOPE Career Grant, depending on their program of study (Lee, 2017). From 2004 to 2009, $3 billion was spent by state and local governments to community colleges to help pay for the education of students who did not return for a

second year (Schneider & Yin, 2011). As the environment for higher education has changed from adequate resources to diminishing resources, there has been a heightened focus by colleges, universities, and state governments to increase the rate at which students persist and graduate from both two- and four-year colleges and universities (Tinto, 2006).

State governments and the federal government have focused on improving student retention and completion in all forms of higher education as a means of increasing the skills of the workforce to better meet the challenges of a global economy (Hirschy et al., 2011). This is evident in recent efforts to establish additional or alternative measures of student success through the Higher Education Opportunity Act of 2008 and the American Association of Community Colleges (Hirschy et al., 2011). The Committee on Measures of Student Success suggested the federal government expand the range of completion and graduation data which degree-granting institutions are required to report on to reflect the diversity of community college campuses (U. S. Department of Education, 2011). While efforts have been made to collect data from community colleges on alternative measures of success such as student learning and employment after college, there is no consistency in the way data are gathered or reported on by each college (U. S. Department of Education, 2011).

The Committee on Measures of Student Success recommended publicly disclosing the information to give potential students, parents, and policymakers easier access to student achievement information at two-year institutions (U. S. Department of Education, 2011). Traditional IPEDS (Integrated Postsecondary Education Data System) measures for community colleges limited the cohort to students who enroll in college for

the first time, take a full course load, and calculated the percentage of students who graduated within three years of enrollment (American Association of Community Colleges, 2018). Beginning with the 2015-2016 collection cycle, IPEDS outcome measures were updated to include three new cohorts: First-time, part-time students; full-time students that are not first-time students; and part-time students that are not first-time students (Ginder, Kelly-Reid, & Mann, 2017b). Additional measures were added to the 2017-2018 collection cycle to include data on whether students are Pell grant recipients, the type of award earned (certificate, associate degree, or bachelor's degree), and the status of students at four years after enrollment (Ginder et al., 2017b). Despite these changes, the American Association of Community Colleges (2018) claimed the Voluntary Framework of Accountability (VFA) metrics are a better measure of community college student success than IPEDS metrics. The VFA metrics look at all entering students, calculate the graduation rate within six years of enrollment, and use nine separate outcomes to determine student success (American Association of Community Colleges, 2018). The American Association of Community Colleges maintained that data on subpopulations is needed to fully understand what is happening in community colleges. These collective efforts reflect a shift toward acknowledging the distinctive nature of students in the community college setting (Hirschy et al., 2011).

Nontraditional Students at Community Colleges

Community colleges provide a path to postsecondary education for a diverse student population (American Association of Community Colleges, 2015; Chen, 2017; Kim, 2002). Kim explained (2002) community colleges provide an avenue to higher education to a larger range of students than those found at most four-year institutions.

This diverse, yet increasingly familiar population is made up mostly of students characterized as nontraditional and has become the norm in postsecondary education (Carnevale, Smith, Melton, & Price, 2015; Chen, 2017; Westervelt, 2016). Markle (2015) stated college entry by students age 25 years and older is expected to increase by up to 28% by 2019. The term nontraditional, as well as adult learner or post-traditional, is used across research to cover a variety of characteristics which make nontraditional students different from the traditional population such as age, ethnicity, residence, disability status, and gender (Monroe, 2006; Watt & Wagner, 2016). These students are predominantly older female students who have not been enrolled in school in at least a year, have children, and are working full time (Copper, 2017).

In addition to demographic characteristics, nontraditional students are differentiated based on life experiences and choices (Watt & Wagner, 2016). With them, these students bring life experiences, self-awareness, and a great value to both higher education and the economy (Watt & Wagner, 2016). Mezirow's (1997) transformative learning theory states adult learners carry with them frames of reference acquired from life experiences, associations, concepts, values, feelings, and conditioned responses. The process of transformative learning is effecting change based on those various frames of reference (Mezirow, 1997). That is, this theory is based on how adults make sense of their life experiences. Students' opinions, points of view, and reasoning likely stem from their life experiences and frames of reference. Mezirow differentiates between types of meaning structures, including frame of reference, habits of mind, and points of view (Merriam, Cafarella, & Baumgartner, 2007). Mezirow indicates a frame of reference is the structure of assumptions and expectations through which students filter impressions

and it provides the context for making meaning (Merriam et al., 2007). This learning is accelerated within a social context as issues related to race, class, and gender enter the learning process and understanding of experience (Cranton & Taylor, 2012). The key component of this theory is that nontraditional students' experiences and narratives are critical to their learning (Chen, 2014).

Of the 2011-2012 undergraduates, 74% had at least one nontraditional characteristic, such as being over age 25, having dependents of their own, not entering postsecondary education immediately after high school, or working while enrolled in school (Radford, Cominole, & Skomsvold, 2015). Additionally, 55% of the same undergraduate population included students with two or more nontraditional characteristics (Radford et al., 2015). A key characteristic distinguishing nontraditionals from other college students is the likelihood this population juggles multiple life roles while attending school, including being an employee, a spouse or partner, a parent, a caregiver, and a community member (Monroe, 2006; Ross-Gordon, 2011). Carey (2017) stated nontraditional students often return to school after years away, have full-time jobs, and take longer to graduate than the three years the U.S. Department of Education used to gauge the success of people pursuing two-year degrees. Based on these characteristics, the majority of students in undergraduate programs can be classified as nontraditional (Choy, 2002; MacDonald, 2018). Choy (2002) claimed traditional students, who are enrolled full time and live on campus, are now the exception rather than the rule even though traditional students receive the majority of attention and resources from colleges and universities. Hittepole (n.d.) agreed colleges and universities look to supply the needs of traditional students first and foremost despite the growing presence of nontraditional

students in higher education. Only 58% of institutions participating in the 2014 National Association of Student Personnel Administrators (NASPA) Student Affairs Census offer nontraditional student services (Hittepole, n.d.).

Because community college students often meet more than one definition of nontraditional, it is important to have an understanding of the definitions used by researchers in studying nontraditional community college students (Kim, 2002). Most often age (especially being over the age of 25 years) has been the defining characteristic for the nontraditional population (Bean & Metzner, 1985; Cleveland-Innes, 1994; Hurtado et al., 1996). When using age as a key identifier for nontraditional students in the community college, research is limiting as students who are under 25 years old may share characteristics of students who are over age 25 (Kim, 2002). For example, Hamilton (1998) reported students age 25 years or less at the time of enrollment who entered college within four years of receiving their GED®, required more developmental courses, and had lower GPAs and one-year persistence rates. Although these students were the same age as their classmates, different high school experiences triggered different college experiences and outcomes (Hamilton, 1998). Ely (1997) estimated students age 25 years and older must find a balance between college, job, family, and financial responsibilities, making the reach to their educational goals and objectives harder. Because these students spend most on-campus time in the classroom, flexible schedules are needed to improve their basic academic, study, decision making, and stress management skills (Ely, 1997). Bean and Metzner (1985) used race and gender to define nontraditional students. Additional variables typically used to characterize nontraditional students are residence

(i.e., not living on campus), level of employment (especially working full time), and being enrolled in non-degree occupational programs (Jones & Watson, 1990).

In a statistical analysis report by the National Center for Education Statistics (1996), instead of focusing on age or other background characteristics, nontraditional students were identified using criteria which revolved around choices and behaviors which may increase a students' risk of attrition. The criteria used to identify nontraditional students in the NCES (1996) report were enrollment patterns, financial and family status, and high school graduation status. The report assumed traditional-age enrollment in postsecondary education was defined as immediate enrollment after high school and attending full time (NCES, 1996). Therefore, those students who chose to delay enrollment in postsecondary education by a year or more after high school or who attended part-time were considered nontraditional (NCES, 1996). Additional qualifiers used to identify family responsibilities and financial constraints of nontraditional students included having dependents in addition to a spouse, being a single parent, working full-time while enrolled, or being financially independent of parents (NCES, 1996). Students who did not earn a standard high school diploma but instead earned some type of certificate of completion (GED® recipients) were also considered nontraditional (NCES, 1996). Based on these criteria, the term nontraditional was broadly defined by seven characteristics, including delayed enrollment in postsecondary education from high school, financial independence, full-time employment, enrolled part-time, has dependents, is a single parent, and earned a GED® diploma instead of a high school diploma. The NCES classified nontraditional students as minimally (one factor), moderately (two or three factors), or highly (four or more factors) nontraditional (NCES,

1996). The NCES report described nearly 75% of beginning undergraduates as at least minimally nontraditional (NCES, 1996). Kim (2002) stated nontraditional students in public two-year institutions are more likely to have two or more risk factors compared to public four-year institution students as these characteristics are likely to change over a student's educational life.

Ely (1997) indicated social integration is important to nontraditional students in the community college setting. Bean and Metzner (1985) suggested nontraditional students are more likely to invest time in enhancing the learning experience, while traditional students make time for involvement in social activities, including club sports and Greek life. Because nontraditionals have more of a business mindset in regards to their educational experience, their social interests tend to develop at a slower rate and only as time in their schedules permits (Grabowski, Rush, Ragen, Fayard, & Watkins-Lewis, 2016). Levine (1993) stated this business mindset includes an expectation of customer-oriented services from colleges and universities. In contrast to their traditional counterparts, nontraditional students expect efficient educational experiences and will look for colleges which can save them money and maximize learning outcomes (Levine, 1993).

Retention and Nontraditional Students

Student retention is critical to community colleges (Wild & Ebbers, 2002). Kenner and Weinerman (2011) attributed higher attrition rates for nontraditional students than traditional college students with the challenges of immersing themselves in the academic environment. The challenge for institutional leaders is how to engage the different student populations like nontraditional students (Wyatt, 2011). The retention of

nontraditional students is particularly important as this population is exposed and somewhat vulnerable to the college environment as it relates to interaction with peers, the classroom, and the campus environment (Wyatt, 2011). Radford et al. (2015) indicated 67% of nontraditional students drop out of college before receiving a degree. Taniguchi and Kaufman (2005) examined factors related to nontraditional student attrition and completion and confirmed characteristics such as full-time student status, part-time employment, positive interactions with instructors, and a supportive family environment increased the likelihood of completion. The researchers revealed part-time enrollment, childcare issues, and being divorced adversely affected completion (Taniguchi & Kaufman, 2005).

Wolf (2011) studied the influence of external factors of family support systems on the persistence of underserved college students. Wolf (2011) discovered financial support, how needs were prioritized, and how they valued ambition, openness, and communication skills were common themes impacting the persistence of this student population. Goncalves and Trunk (2014) determined feelings of isolation, inattention to nontraditional student needs, and lack of resources were obstacles to nontraditional student persistence.

Conversely, Oden (2011) studied factors which affect the persistence of nontraditional students in two-year colleges but did not consider students who are parents. Using Bean and Metzner's model (1985), Oden (2011) focused on external influences impacting nontraditional students and found a greater need for improved quality of life and how these students must consider the responsibilities of work and family. In contrast, the study found traditional students have a greater need for

engagement within the college or university (Oden, 2011; Pascarella & Terenzini, 1991; Pascarella & Terenzini, 2005; Tinto, 1993). Oden (2011) and Wolf (2011) agreed in their findings that the external factors were more impactful on persistence than the social integration for these students. Persistence was enhanced by family support and self-determination (Oden, 2011; Wolf, 2011).

Because of the inherent difference in missions and student populations between community colleges and four-year institutions, different criteria and methodologies for judging institutional effectiveness, including retention, are warranted (Seidman, 1993). The National Center for Education Statistics (1996) states that retention measures the rate at which students persist in their educational program at an institution. For two-year institutions, this is the percentage of first-time degree, diploma, or certificate-seeking students from the previous fall who either re-enrolled or graduated by the current fall (NCES, 1996). One pervasive issue in understanding retention at community colleges is the lack of consistency in how student retention is defined. A specific challenge in developing a common definition is many of the definitions used today in academia were developed for retention considerations in university settings (Wild & Ebbers, 2002). Wild and Ebbers (2002) suggested most research in this area is based on traditional-age students in the residential settings of universities which does little for community colleges. For example, Wyman (1997), whose definition was specific to community colleges, defined retention as a percentage of students either graduating or persisting in their studies at an institution.

Within Georgia, there is little surprise the University System of Georgia and the Technical College System of Georgia differ in the definition of retention as well. For

USG institutions (both two- and four-year), a student is considered to be retained if enrolled in a USG institution in the same academic term one year later (University System of Georgia, n.d.). Although the definition of the cohort may vary according to the subject of interest, the most common cohorts studied are first-time, full-time, degree-seeking freshman students (University System of Georgia, n.d.). Within the Technical College System of Georgia, the definition of retention used to compare colleges within the system is a beginning fall cohort student from the previous year. The number retained is based on those who graduated from the same college or a different college, or were still enrolled in the same college or a different college. Within TCSG's institutional effectiveness system, known as the Performance Accountability System (PAS), retention is defined differently. In PAS the number retained is any student from the fall cohort who graduated that fall term or any subsequent term that year or the following year, from any program at any TCSG or USG college, or was enrolled during any term the following year at any TCSG or USG college (Technical College System of Georgia, 2016). In TCSG a fall cohort may be any full- or part-time, first-time students at the college, regularly admitted student from all major code levels (certificate, diploma, or degree) except for high school and transient students.

Student Retention Theories, Models, and Frameworks

Demetriou and Schmitz-Sciborski (2011) suggested the first studies of undergraduate retention began to develop in the 1930s. McNeely's (1937) study on student demographics, social engagement, and reasons for departure became the precursor for many studies in the 1960s (Berger & Lyon, 2005; Demetriou & Schmitz-Sciborski, 2011). Large-scale studies encouraged a comprehensive examination of

student attrition which in part focused on student characteristics (Barnett & Lewis, 1963; Berger & Lyon, 2005; Demetriou & Schmitz-Sciborski, 2011; Panos & Astin, 1968). By the end of the 1960s decade, retention was a common concern, and college and university campuses began to develop research activities specific to understanding and supporting retention (Demetriou & Schmitz-Sciborski, 2011).

The decade of the 1970s brought several theoretical models of student retention. Spady's (1970) sociological model of student dropout, based on the experiences of traditional students in four-year, residential institutions, was the first widely recognized model in retention study (Berger & Lyon, 2005; Demetriou & Schmitz-Sciborski, 2011; Hirschy et al., 2011). Spady (1970) suggested five variables (academic potential, normative congruence, grade performance, intellectual development, and friendship support) grounded in social integration and indirectly linked to a student's decision to drop out of school through the intervening variables of satisfaction and commitment (Berger & Lyon, 2005; Demetriou & Schmitz-Sciborski, 2011; Hirschy et al., 2011). Spady's (1970) research revealed the goals, interests, skills, and attitudes of the student, along with family, cultural, and institutional characteristics must be consistent for retention. In Spady's model, grades and student learning represented the academic systems, and friendships and involvement with others at the institution represented the social system (Spady, 1970). Spady's research (1970) found students who did not integrate socially and intellectually with their institution were more likely to drop out. In subsequent testing Spady (1971) determined the primary factor influencing attrition was academic performance, while social integration and institutional commitment were secondary.

Tinto's (1975) student integration model identified the factor of persistence as being how well the student integrated into college (Berger & Lyon, 2005; Demetriou & Schmitz-Sciborski, 2011; Hirschy et al., 2011; Wild & Ebbers, 2002). This model suggested the interaction between the student and the academic and social systems of the college are necessary for the student to connect to and persist through college (Demetriou & Schmitz-Sciborski, 2011; Tinto, 1975; Wild & Ebbers, 2002). Tinto (1975) hypothesized the quality of academic and social interactions significantly influenced the person-environment fit, which is the level of involvement the student has with the institution and how it influences retention. Students who are not involved in college activities or who do not feel integrated into the culture of the college do not persist (Tinto, 1975; Tinto, 1993). His research focused on traditional students at traditional four-year institutions and did not take into account the experiences of the student before entering the college or outside the environment of the college once enrolled (Tinto, 1975). Subsequent work by Tinto suggested the student's financial resources and communities, such as family and work, play a key part in the students' departure decisions (Hirschy et al., 2011; Tinto, 1975; Tinto, 1986; Tinto, 1993). Tinto (1993) modified his original model to account for external factors in the student's decision to leave. Because Tinto's model relies heavily on social integration and academic integration outside of the classroom, the application of it through the lens of two-year institutions remains an open research question (Alfonso, Bailey, & Scott, 2005; Braxton, Hirschy, & McClendon, 2004; Hirschy et al., 2011).

During the 1980s, retention became the focus of many institutions' strategic planning processes (Demetriou & Schmitz-Sciborski, 2011). Bean (1980) theorized

background characteristics, such as prior academic performance and socioeconomic status, influenced a student's departure from an institution (Demetriou & Schmitz-Sciborski, 2011). The mid-1980s saw the development of a critical theory and a defining model. Astin's (1984) student involvement theory focused on the motivation and behavior of the student. The core of Astin's theory involved input-environment-output (I-E-O) categories where inputs are characteristics students bring with them to college (e.g., gender and academic preparation), the environment is the student's actual experiences while in college, and outputs are the student's educational outcomes (e.g., persistence, educational goal attainment, and degree completion) (Astin, 1984; Hirschy et al., 2011).

Bean and Metzner's (1985) student attrition model, which was informed by Bean's earlier work (1980), indicated institutional experiences and other non-institutional factors shape beliefs, which in turn impact persistence (Hirschy et al., 2011). Bean and Metzner's (1985) research indicated environmental variables are presumed to be more important for nontraditional students than academic variables. Based on this assumption, the model suggested three scenarios. First, students are likely to remain in school when both academic and environmental variables are good but would possibly leave school when both variables are poor (Bean & Metzner, 1985). Second, students are more likely to leave school when academic variables are good but environmental variables are poor (Bean & Metzner, 1985). Third, students are more likely to remain in school when environmental support is good and academic support is poor (Bean & Metzner, 1985). That is, the environmental support will compensate for low scores on the academic variables (Bean & Metzner, 1985). For example, despite strong academic support, a student will not remain in school if their child care arrangements are inadequate or their

work schedules interfere with classes (Bean & Metzner, 1985). However, a student with good environmental support such as encouragement to stay in school by family and employers will likely remain in school despite poor academic support (Bean & Metzner, 1985).

Tinto's (1993) continuation of his student integration model in the 1990s focused on specific student groups, such as students from low-income families, adult students, and transfer students, requiring dedicated interventions and policies (Demetriou & Schmitz-Sciborski, 2011). As quality support services became the focus across campuses, Swail's 1995 framework for student retention recommended collaboration between both academic and student services (Demetriou & Schmitz-Sciborski, 2011; Swail, 1995; Swail, 2004). In a combination of several previous theories and models, Wyckoff (1998) posited students are influenced to remain at an institution based on their interactions with all members of the institutional environment (other students, faculty, staff, and administrators) (Demetriou & Schmitz-Sciborski, 2011). Towards the end of the 1990s, counseling and advising were being emphasized throughout colleges and universities. Houland, Crockett, McGuire, and Anderson's (1997) work focused on academic advising as a motivator and stimulator for students, as it helped them work towards a meaningful goal, thus remaining in school (Demetriou & Schmitz-Sciborski, 2011).

The idea all members of the campus environment impacted retention carried over to the 2000s. Bean and Eaton (2000) suggested as students interact with the college environment, their characteristics impact attitudes, motivations, and behaviors (Hirschy et al., 2011). These researchers offered a model integrating four psychological theories: attitude-behavior theory, coping behavior theory, self-efficacy, and attribution theory

(Bean & Eaton, 2000; Hirschy et al., 2011). Swail, Redd, and Perna (2003) used a force-field approach which, instead of an input-process-outcomes framework like Astin's, represented the interaction of positive and negative effects on student outcomes. The student persistence and achievement model involved a triangle (representing the student experience) which included cognitive factors (e.g., academic, rigor, aptitude, study skills, time management), social factors (e.g., financial issues, maturity, cultural values, goal commitment), and institutional factors (e.g., financial aid, student services, curriculum, and instruction) (Swail et al., 2003). Habley's (2004) work supported previous theories in which the interactions students have with individuals on campus (other students, advisors, faculty, staff, and administrators) directly influence their retention (Demetriou & Schmitz-Sciborski, 2011). In response to Tinto's original (1975) model being supported at commuter institutions, Braxton, et al. (2004) developed a model which included student entry characteristics that could influence the student's initial commitment to the institution, which in turn could influence a student's internal campus environment and external environment (e.g., work, family, community) (Braxton & Hirschy, 2005). Braxton and Hirschy (2005) suggested each of these components could influence the student's commitment to the institution and decision to persist.

Soon after, Nora et al. (2005) completed a different study using a model of student-institution engagement to examine factors affecting the persistence of students who had already completed their first year of college (Hirschy et al., 2011). The study was based on a single public, commuter institution which shared similar characteristics to many community colleges (Hirschy et al., 2011; Nora, et al., 2005). The findings of the study indicated attrition each year and eventual graduation was related to high school

performance, SAT scores, early performance in college, educational costs and financial aid, enrollment status, course-taking patterns, and demographic characteristics such as gender, race, and ethnicity (Hirschy et al., 2011; Nora, et al., 2005). Terenzini and Reason (2005) offered a conceptual framework which included additional influences on college success than earlier models (Hirschy et al., 2011; Reason, 2009). Terenzini and Reason (2005) argued the influence of precollege characteristics and experiences, the organization or organizational context of the institution, the student peer environment, and individual student experiences lead to a better understanding of student persistence (Hirschy et al., 2011; Reason, 2009).

The previous theories, models, and frameworks were not based on research regarding student retention in a community college setting. Previous studies were typically focused on traditional-age students in universities (Wild & Ebbers, 2002). Hirschy et al.'s (2011) model focused specifically on career and technical education (CTE) students and suggested students pursuing occupational associate's degrees or certificates differ from those students seeking academic majors at two-year institutions. The authors suggested the introduction of a career integration variable, promoted the collection and tracking of student educational goals, and expanding traditional student success measures to better reflect the experiences of CTE students (Hirschy et al., 2011). However, although there are established models for retention and attrition of traditional students which do provide concepts and understandings which may be broadly applied to nontraditional students, few studies have specifically addressed the demographics shifts of this population or their needs (Monroe, 2006). Monroe (2006) asserted the complex, dynamic nature of nontraditional students requires continuous examination and

refinement of our understanding of this population's changing demographics concerning attrition.

Factors Related to Student Retention

The previously discussed retention theories, models, and frameworks have been tested in various studies. Depending on the applicability of the model, specific variables instead of all variables identified in the original model were used to study their impact or influence on various outcomes such as retention. The following is a thorough explanation of variables included in past retention studies and this study. Although not all variables found within the literature apply to this research, the variables were found to be consistent within the literature and can be applied to nontraditional students attending technical colleges.

Pell Eligibility. Astin (1975) found financial difficulty is commonly reported by students to be a primary reason for leaving an institution. Swail et al. (2003) suggested that because attending college has both direct and indirect costs, students make financial decisions which have both short- and long-term effects on college persistence. Since 1972, the Pell grant program has been used by the federal government to help students attend college (Soares, Gagliardi, & Nellum, 2017). Congress approved the restoration of year-round Pell to better assist students who depend on federal financial aid (Kreighbaum, 2017). Perna (1998) discovered a statistically significant relationship between work-study aid and degree completion *(*$\chi^2(1, N = 3186) = 10.6, p < .001$), as well as receiving grants only and degree completion ($\chi^2(4, N = 3{,}186) = 30.0, p < .001$). Perna (1998) revealed both receiving work-study aid ($\beta = .04, p < .05$) and grant aid only ($\beta = .04, p < .05$) had positive direct effects on persistence. Perna (1998) used descriptive

statistics, chi-square tests, ANOVA tests, and path analysis to determine the highest degree completion rates were associated with aid packages limited to grants only (56%) and packages comprised of grants, loans, and work-study aid (59%). In contrast, Perna (1998) found completion rates were lower for aid recipients who received loans (45%) than for other aid recipients (53%). A comparison of the total effects indicated grants are more effective in promoting persistence than loans (Nora, 1990; Perna, 1998).

In a recent study, Turk and Chen (2017), in trying to understanding how, when, and why community college students transfer to four-year colleges and universities, found receiving federal financial aid significantly impacts the likelihood of retention. Using a nationally representative data source and a multilevel model, the researchers used logistic regression to test a series of academic, demographic, social, and institutional-level characteristics to determine what impact they have on community college students' likelihood of upward transfer. Although marginally significant, receiving a Pell grant was associated with a 28% reduction in the chances of transfer (β = -.33, p = .06, odds ratio = 0.72) (Turk & Chen, 2017). However, students who received a federal student loan were more than four times as likely to transfer to a four-year institution as students who did not receive a federal loan (β = 1.52, p < .001, odds ratio = 4.56) (Turk & Chen, 2017). Turk and Chen (2017) recommended federal funding increases should keep pace with inflation to help nontraditional students afford postsecondary education. Based on these studies, the Pell grant and federal aid, in general, cannot be dismissed as their impact on retention is both significant and relevant to the current study. The research designs and data modeling approaches used in these studies are relevant and important to the current study.

Single Parent or Displaced Homemaker Status. The Bean and Metzner (1985) student attrition model posited family responsibility such as being married, caring for dependents, or being a single parent negatively affected retention. The researchers indicated environmental variables such as finances, hours of employment, and family responsibilities have a greater influence on the decisions of adult students to leave than academic variables such as study habits and academic advising (Bean & Metzner, 1985). When Metzner and Bean (1987) tested a similar but slightly different model on part-time students at a commuter university using the number of dependents as the measure of family responsibilities, the researchers did not observe any direct effects on retention. Metzner and Bean (1987) used ordinary least squares multiple regression in a path analysis framework to estimate the 1985 theoretical model. The overall model fit was R^2 = .29 (adjusted R^2 = .26) accounting for 29% of the variance in the dropout rate (Metzner & Bean, 1987). This fit was consistent with other studies of student attrition at the time (Metzner & Bean, 1987). The results indicated the number of dependents was not a statistically significant predictor of dropout at the alpha level of .05 and had one of the smallest effect coefficients (β = -.01, n.s.) (Metzner & Bean, 1987).

Research of both student-level and institution-level data was conducted by Titus (2004) to determine which student characteristics, experiences, attitudes, and environment pull variables influence student persistence at a four-year college or university. The Titus (2004) study used hierarchical generalized linear modeling (HGLM) and the sample used was limited to first-time, full-time, degree-seeking undergraduate students. The results showed after taking other variables into account, a one standard deviation increase in a student's financial need related to a 2% increase in

student persistence ($\beta = .109$, $p < .05$, $\Delta p = 2.01$) and a one standard deviation increase in average hours worked per week related to a 3% increase in student persistence ($\beta = .186$, $p < .001$, $\Delta p = 3.37$) (Titus, 2004).

Grabowski et al. (2016) agreed additional stress and emotional strain are compounded on degree completion by the pressure of balancing work, family responsibilities, and other life circumstances. Adult female learners with dependents are especially impacted by these additional stressors (Grabowski et al., 2016).

In an NCES three-year persistence and attainment report by Berkner, Horn, and Clune (2000), results indicated having dependent children had a negative association with student retention. An article by Swift, Colvin, and Mills (1987) characterized the displaced homemaker as 27 years old or older, primarily a homemaker before her enrollment, and married, with at least one child. Many of the displaced female homemakers reported a change in lifestyle such as divorce, separation, or death of a spouse to be the precipitating factor in their enrollment (Swift et al., 1987). Bozick and DeLuca (2005), using a nationally representative high school cohort sample, sought to determine which young adults are most likely to delay postsecondary enrollment, what effect that delay has on degree completion, and if institutional type impacts delayed enrollment. The researchers found when students delay the transition to college, they substantially decrease their chances of completing a degree, and the most extensive delays were those who married or had children before entering college (Bozick & DeLuca, 2005). Bozick and DeLuca (2005) found those who were married either before entering college (odds ratio = .48, $p < .05$) or once enrolled in college (odds ratio = .87, $p < .05$) have lower chances of degree completion than those who were not married.

Likewise, the researchers found students who have children before (odds ratio = .47, $p <$ .10) or during college (odds ratio = .55, $p < .01$) have lower odds of degree completion than those who do not have children while in college (Bozick & DeLuca, 2005). Following these studies, the current study will use variables representing family responsibilities such as being married, caring for dependents, or being a single parent as these characteristics represent a large proportion of nontraditional students.

Age. Historically, age was not typically included in the research on retention because most research focused on traditional-age students (Cochran, Campbell, Baker, & Leeds, 2013). For studies using age as a potential explanatory variable, the results were contradictory. Pascarella, Duby, Miller, and Rasher (1981) found age to be a moderate predictor of student persistence using Tinto's student integration model on 853 students at a commuter four-year college. The researchers used a longitudinal study using the ACE (American Council on Education) Cooperative Institutional Research Program survey and data collected on all incoming students, such as high school rank and college entrance test scores (Pascarella et al., 1981). Three-group discriminant function analysis was used for freshman to sophomore persisters, freshman stopouts, and first-quarter freshman withdrawals (Pascarella et al., 1981). The first stage of analysis included all pre-enrollment characteristics (high school academic performance, age, perceived likelihood of dropping out, perceived likelihood of transfer, and perceived need for remediation), and only those variables contributing to group discrimination significant at $p < .10$ were used in the second stage of the stepwise discriminant analysis (Pascarella et al., 1981). The results indicated pre-enrollment variables like age, along with first-quarter GPA, significantly differentiate between freshman year persisters and early withdrawals

(Pascarella et al., 1981). The classification analysis based on the six-variable equation correctly identified 72% of the early withdrawals and 74% of the persisters (Pascarella et al., 1981). The findings revealed a significant main effect for the age variable on persisters and withdrawals ($F(1, 847) = 7.12$, $p < .01$) (Pascarella et al., 1981).

Over a decade later, Feldman's (1993) study of one-year retention of first-time students at a community college used chi-square analysis for univariate comparisons and logistic regression to select and order the factors which contributed to retention. She found age had a significant impact ($\chi^2(1) = 26.13$, $p < .001$) on retention using both univariate and multivariate analysis (Feldman, 1993). The odds of students age 20-24 years old dropping out was 1.77 times that of students aged 19 or younger and the 20-24 age range was the most significant predictor age range according to the Wald statistic ($\chi^2(1) = 7.37$, $p < .001$) (Feldman, 1993).

The Nakajima, Dembo, and Mossler (2012) study of 427 community college students looked at the influence of background variables, financial variables, and academic variables on students' persistence in community college education. Nakajima et al. (2012) questioned if academic integration and psychosocial variables influence student persistence by using a 63-item survey assessing psychosocial variables, academic integration, and various background variables. Among the background variables, the study used t-tests to reveal age and high school graduation year influenced student persistence in community college students (Nakajima et al., 2012). Those who persisted were younger ($M = 24.12$, $SD = 8.19$) compared to those who did not persist ($M = 26.23$, $SD = 8.48$) ($t(370) = 2.13$; $p < .05$), but these effects diminished once multiple variables were entered into the analysis (Nakajima et al., 2012). Nakajima et al. (2012) also found

students who graduated from high school in 2004 or earlier had the most nonpersisting rate compared to students who graduated in 2005 or later ($\chi^2(5, N = 381) = 17.13, p < .01$).

Other studies contradicted these findings. Metzner and Bean's (1987) study with 624 nontraditional students did not reveal age as a significant predictor of student persistence. Mohammadi's (1994) longitudinal study of 3,843 first-time community college students was designed to explain retention and attrition. The quantitative ex post facto study used exploratory data analysis and logistic regression to determine age was not a significant predictor of persistence for fall to fall retention (Mohammadi, 1994).

Fike and Fike (2008), who collected data from a Texas public urban community college, found age was a weak predictor of retention after controlling for covariates. The researchers quantitative, retrospective study assessed predictors of student retention for first-time students in a community college using chi-square analysis, calculated correlation coefficients, and multivariate logistic regression (Fike & Fike, 2008). The bivariate correlation of student age with retention was negative for fall to spring retention ($r(9,194) = -.08, p < .001$) and for fall to fall retention ($r(9,194) = -.10, p < .001$). Although it was a positive predictor of fall to spring retention in the logistic regression model, the contribution of student age was weak ($\beta = .01, p < .001$, odds ratio = 1.01, 95% CI = 1.01 to 1.02) (Fike & Fike, 2008). The relevance of these findings to the current study is the data modeling approaches used and the variable age will be used to define which students fall into the nontraditional category.

Race or Ethnicity. Ethnicity differences are factors in some retention studies. Singell and Waddell's (2010) research, which used an empirical model developed by

Singell (2004), centered on whether the University of Oregon could effectively identify students who might be retention risks early in their college careers using accessible data. The researchers combined logistic regression and hazard modeling approaches of prior work and used existing student-level data to estimate a predicted retention probability based on gender, race, high school GPA, and SAT scores (Singell & Waddell, 2010). Singell and Waddell (2010) estimated separate prediction models for residents and nonresidents supported by a likelihood ratio test which rejects the restriction of equal coefficients by residential status at the 99% level. Singell and Waddell (2010) claimed, absent of other attributes, African American ($\beta = .06$, $p < .01$) and Asian ($\beta = .04$, $p < .01$) students are more likely to be retained than White students in the fall term of their second year. This research found Hispanic, Native American, and other non-White students do not differ in their retention probabilities from White students (Singell & Waddell, 2010). In addition to providing context between race, ethnicity, and retention, Singell and Waddell's (2010) research found students at risk of dropping out can be identified using accessible statistical models and information available at the time a student enrolls and monitoring students as they matriculate improves the model's ability to predict retention. This implies a trade-off between early identification and intervention and the information gained by including additional data which becomes available as the student progresses through their program of study.

Fike and Fike (2008), who collected data from a Texas public urban community college, found student ethnicity was not a significant predictor of retention. This quantitative, retrospective study assessed predictors of student retention for first-time students in a community college using chi-square analysis, calculated correlation

coefficients, and multivariate logistic regression (Fike & Fike, 2008). The bivariate correlation found student ethnicity was not consistently associated with student retention for Hispanic students ($r(8,947) = -.01$, $p = .511$), White students ($r(8,947) = .01$, $p = .226$), or other ($r(8,947) = -.01$, $p = .287$) (Fike & Fike, 2008). In the logistic regression model, student ethnicity was not statistically significant after controlling for covariates (Fike & Fike, 2008).

These findings complement Murtaugh, Burns, and Schuster's (1999) research which claimed results are explained by the variables contained within the model. This research used a university student database and focused on demographic and academic variables which were available in the first term of enrollment (Murtaugh et al., 1999). While the research was limited to the information contained in the student database, similar to the current study, the researchers were confident the variables summarize many of the important influences on student retention (Murtaugh et al., 1999). For univariate analyses, estimated retention probabilities used the Kaplan-Meier method and for multiple variable analyses, the Cox proportional hazards regression model was used (Murtaugh et al., 1999). Hazard ratios, factors by which a student's withdrawal is multiplied by a unit increase in the predictor, were calculated for both the univariate model and the final multiple variable model at 95% confidence intervals (Murtaugh et al., 1999). When only race was considered in the model, African Americans (hazard ratio = 1.38), Hispanics (hazard ratio = 1.37), and American Indians (hazard ratio = 1.45) were at greater risk of withdrawing from college than White students (Murtaugh et al., 1999). When multiple variables like age and GPA were included in the model, the differences for American Indian (hazard ratio = 1.14) students and Hispanic (hazard ratio = 0.95)

students decrease and Black (hazard ratio = 0.68) students have a reduced risk compared to White students (Murtaugh et al., 1999).

More recent research of first-time students obtained a statistically significant main effect for race or ethnicity on persistence ($p < .01$) using Tinto's (1993) longitudinal model of institutional departure (Stewart, Doo, & Kim, 2015), which was consistent with findings from Terenzini and Pascarella's (1978) original research on students' precollege characteristics. Terenzini and Pascarella's (1978) longitudinal, ex post facto study was completed at Syracuse University and used the Adjective Rating Scale to measure students' expectations. The purpose of the study was to determine the influence of students' pre-college characteristics on attrition, the experiences of the freshman year, and the interaction of certain student traits with their institutional experiences (Terenzini & Pascarella, 1978). The overall multiple regression, using all 528 respondents and all variables and interaction vectors, produced a multiple $R = .51$, with an $R^2 = .26$, $F(76, 451) = 2.05$, $p < .001$ (Terenzini & Pascarella, 1978). All variables and interaction vectors made a significant contribution ($p < .05$) to the prediction of attrition status, and the overall set of interactions explained 10% of the variance which warranted the investigation of individual interactions (Terenzini & Pascarella, 1978). The researchers used stepwise multiple regressions to determine the interaction between race or ethnic origin and the affective appeal students have for the academic program was statistically significant ($F(1, 451) = 6.58$, $p < .05$) and the interaction between race, ethnic origin, and intellectual development and progress ($F(1, 451) = 5.00$, $p < .05$) (Terenzini & Pascarella, 1978). The researchers were able to highlight race or ethnic origin was

involved in two significant and unique interactions related to the probability of dropping out voluntarily (Terenzini & Pascarella, 1978).

Stewart et al. (2015) conducted a study using an ex post facto design to examine what demographic, family characteristics, pre-college, and college academic performance factors predict persistence between students placed in remedial courses and students not placed in remedial courses at a four-year public research institution. In addition to descriptive statistics, inferential statistics used to answer each research question included factorial analysis of variance (ANOVA), Pearson's product-moment correlations, and multiple regression analysis (Stewart et al., 2015). Among the ethnic groups, the overall group means revealed the Asian/Pacific-Islander students were most likely to persist (M = 4.97, SD = 1.39), followed by African-American/non-Hispanic (M = 4.87, SD = 1.60), White/non-Hispanic (M = 4.69, SD = 1.53), Hispanic (M = 4.54, SD = 1.54), and American Indian/Alaska Native (M = 4.13, SD = 1.73) (Stewart et al., 2015). Because there was no significant interaction between ethnicity and remediation, the Stewart et al. (2015) study recommended academic affairs and student affairs administrators should ensure special population groups continue to have access and are encouraged to utilize support services like advising and counseling to foster student success and increase student persistence. The previous studies influence the current study based on both the research methods utilized and the varying results. The impact of race or ethnicity of nontraditional students in technical education in Georgia is both significant and relevant to the current study.

Gender. Existing literature reveals varying results about the effects of gender differences on persistence. Mohammadi (1994) found men more likely to persist than

women. Chen and Thomas (2001) and Halpin (1990) found women more likely to persist than men. Horn, Peter, and Rooney's 2002 NCES report indicated no influence by gender on persistence (Horn, Peter, & Rooney, 2002). Although Pritchard and Wilson's (2003) research of 218 undergraduate students from a private Midwestern university focused on student's emotional and social factors, the researchers also investigated the influence of traditional demographic variables like gender and found gender did not influence persistence. Pritchard and Wilson's (2003) study was designed to identify the relationship between student emotional and social health and academic success and retention. Multiple regressions were used to assess the influence of demographic variables, the effect of emotional health, and the effect of social health on GPA and retention (Pritchard & Wilson, 2003). While the combined influence of all the demographic variables in the study had a significant effect on GPA ($R^2 = .22$, $F(7, 109) = 4.17$, $p < .001$), they had no effect on the intent to drop out ($R^2 = .02$, $F(7, 182) = 1.00$, $p = .80$). (Pritchard & Wilson, 2003).

While gender was related to higher cumulative GPAs in females in Craig and Ward's (2008) study, gender alone was not a predictor of having greater or lesser chances of success in college. This study focused on a cohort of first-time, full-time students at a public community college in New England which were analyzed using analysis of variance and logistic regression analysis (Craig & Ward, 2008). The analysis of variance showed a main effect of gender on cumulative GPA, $F(1, 1727) = 10.16$, $p < .001$ (Craig & Ward, 2008). Gender, age, race, and ethnicity were removed as nonsignificant factors in the logistic regression analysis (statistics were not provided for nonsignificant factors) (Craig & Ward, 2008).

In general, sociodemographic variables, such as race, ethnicity, and gender, are difficult to interpret and higher education researchers have difficulty in finding actionable implications from these studies with these variables (Reason, 2009). However, sociodemographic variables are important to include in this study to provide a greater understanding of the conditional effects of interventions aimed at increasing student persistence (Pascarella & Terenzini, 1991). Reason (2009) suggested one cannot assume a single intervention is effective for all students, or assume interventions influence students the same way or to the same magnitude.

High School Diploma Type. Among variables previously discussed like age, gender and, ethnicity, high school grades and standardized test scores have been consistently found to be strong predictors of degree attainment for undergraduates (Astin & Oseguera, 2005; Titus, 2004). The Titus (2004) study used student-level data from the NCES and institution-level data from IPEDS to determine which student characteristics, experiences, attitudes, and environment pull variables influence student persistence at a four-year college or university. Ahierarchical generalized linear modeling (HGLM) approach was utilized and the sample used was limited to first-time, full-time, degree-seeking undergraduate students (Titus, 2004). Student background characteristics included academic ability where ability was measured by a composite based on standardized high school grade point average and standardized SAT scores (Titus, 2004). The results showed after taking other variables into account, a one standard deviation increase in a student's ability is related to a 2% increase in student persistence ($\beta = .13$, $p < .05$, $\Delta p = 2.44$) (Titus, 2004).

As it relates to the current study, a high school student's GPA, SAT, ACT, and level of coursework have shown to be essential predictors of how well students perform during their first year of college (Geiser & Santelices, 2007; Hodara & Lewis, 2017). Given 75% of students usually drop out of college in the first two years, and 57% of students leave their first college without graduating (Tinto, 1993), it is not surprising the attributes and characteristics students bring with them to college greatly influence their first-year grades. Stewart et al.'s (2015) study, among other things, examined the relationship between high school GPA and first-semester college GPA. Although weak, a statistically significant positive correlation existed between high school GPA and persistence ($r = .18$, $p < .01$) and a moderately stronger significant positive correlation existed between the first semester college cumulative GPA and persistence ($r = .42$, $p < .01$) (Stewart et al., 2015). Using stepwise regression, the results showed first semester college GPA and high school GPA had a significant contribution on persistence, accounting for 26% of the variance and demonstrating a strong correlation coefficient value ($R = .51$, $R^2 = .26$, Adjusted $R^2 = .26$, $p < .01$) (Stewart et al., 2015).

High School Graduation Date. Longitudinal NCES reports have shown students who delay enrollment in college are at substantial risk of not completing a postsecondary credential when compared to their peers who enroll immediately after high school graduation (Berkner, Cuccaro-Alamin, & McCormick, 1996; Berkner, He, & Cataldi, 2002; Carroll, 1989; Horn, 1996; Tuma & Geis, 1995). Those who delay postsecondary enrollment are more likely than immediate enrollees to have family and educational experiences which place them at greater risk for dropping out of high school (Horn, Cataldi, & Sikora, 2005). Students who do not enroll in college immediately after high

school are more likely to have lackluster study habits and have lost some content knowledge, especially in mathematics and science (Bozick & DeLuca, 2005; Horn et al., 2005; Peltier, Laden, & Matranga, 1999). Horn et al. (2005) explained some students may not be academically prepared to attend or have the financial resources necessary to enroll in college, while others may enroll in the military, find a job, or start a family before enrolling. Horn et al.'s (2005) research found the longer the delay, the less likely students enrolled in bachelor's degree programs but were more likely to enroll in technical certificate programs. These findings are relevant to the current study which will use the student high school graduation date to determine how subsequent or delayed enrollment in college impacted their retention.

Bozick and DeLuca (2005), using a nationally representative high school cohort sample, sought to determine which young adults are most likely to delay postsecondary enrollment, what effect that delay has on degree completion, and if institutional type impacts delayed enrollment. The researchers found (in analyses not shown) when students delay the transition to college, they substantially decrease their chances of completing a degree (Bozick & DeLuca, 2005). The results indicated taking a year off after high school reduces the likelihood of degree completion by 64% with all other factors being equal (Bozick & DeLuca, 2005). Descriptive statistics revealed on average, White students began postsecondary enrollment eight months after completing high school, compared to the 10-month delay of Hispanic students and 11-month delay of Black students (Bozick & DeLuca, 2005). A larger proportion of females enroll in college on time while a greater proportion of males delay their enrollment or do not

enroll at all and delayers are more concentrated in the South and suburban areas (Bozick & DeLuca, 2005).

Grade Point Average. Tinto's (1997) study of 287 first-year community college students set out to determine the degree to which learning communities and the adoption of collaborative learning strategies impacted persistence. Tinto (1997) used stepwise logit regression analysis to predict second-year persistence using both qualitative and quantitative methods. Five variables proved to be significant predictors of persistence using an alpha level of .10 among students at Seattle Central Community College (participation in the Coordinated Studies Program, college grade point average, hours studied per week, perceptions of faculty, and a factor score on involvement with other students). That same year, McGrath and Braunstein (1997) completed a study to identify the predictors of attrition for freshmen who voluntarily withdrew by studying the relationship between attrition and certain demographic, academic, financial, and social factors. Specifically, McGrath and Braunstein (1997) looked at which factors differentiate between those freshmen who were retained and those who were not retained. The researchers used the College Student Inventory to assess predispositions, pre-college experiences, and attributes which may influence retention for full-time freshmen at Iona College in New York (McGrath & Braunstein, 1997). Because there were additional data used from students' academic, demographic, and financial records, a preliminary analysis of t-tests was used to reduce the number of variables for use in a logistic regression (McGrath & Braunstein, 1997). A significant difference was found between the groups when McGrath and Braunstein (1997) used a t-test on the first semester GPAs for freshmen who were retained ($M = 2.67$, $SD = .64$) and those who were not retained ($M =$

1.76, SD = 1.17), $t(297) = 8.9$, $p < .001$, $d = .96$. Independent variables which were statistically significant at the .05 level were entered into a stepwise logistic regression (McGrath & Braunstein, 1997). The results indicated first-semester college GPA (β = 1.15, $p < .001$, $R = .34$) as the strongest variable in predicting persistence between the first and second years (McGrath & Braunstein, 1997). McGrath and Braunstein (1997) used logistic regression to predict the probability of freshmen returning for their sophomore year by assigning students to a “retained” group if the predicted probability of retention was greater than 50%; otherwise, students were assigned to the "non-retained" group. The researchers applied these criteria to the final sample of 322 freshmen, and along with students' impressions of other students, were able to make correct predictions in approximately 80% of the analyzed cases (McGrath & Braunstein, 1997).

In Craig and Ward’s (2008) study of 1,729 first-time, full-time community college students, the researchers found GPA was a significant indicator of student retention using logistic regression analysis. On average, students not retained had a cumulative GPA of 1.68 and had earned only 16.8 credit hours compared to 2.29 for retained students (Craig & Ward, 2008). Of the student academic characteristics, cumulative GPA ($\beta = .73$, $\chi^2(1, N = 1729) = 91.44$, $p < .001$) was most strongly related to student success with a 2.04 odds ratio (Craig & Ward, 2008). Second semester GPA (β = .32, $\chi^2(1, N = 1729) = 44.14$, $p < .001$) and attempted but unearned credits ($\beta = -.03$, $\chi^2(1, N = 1729) = 38.36$, $p < .001$) were also significant (Craig & Ward, 2008). Second semester GPA had a positive association with student success with an odds ratio of 1.38, but attempted but unearned credits had a negative association with an odds ratio of 0.97 (Craig & Ward, 2008).

Titus (2006) performed another study which relates college GPA to student persistence. Titus (2006) conducted a study using hierarchical generalized linear modeling on 4,951 first-time, full-time students using a national database of four-year institutions. He found GPA significantly increased the odds for persistence ($\beta = .48$, odds ratio = 1.61; $p < .001$) (Titus, 2006).

In the study conducted by Nakajima et al. (2012) where student retention was measured through college enrollment the following semester at one institution, cumulative GPA was found to be the strongest predictor of student persistence ($t(365) = -2.56$; $p < .05$). Students who had higher cumulative GPAs were twice as likely to stay in college and this effect did not diminish when other variables were entered into the model (Nakajima et al., 2012).

Likewise, Gifford, Briceno-Perriott, and Mianzo (2006) found freshmen retained to their sophomore year demonstrated a statistically significant higher GPA than those who were not retained. Gifford et al. (2006) used the Adult Nowicki-Strickland Internal External Control Scale (ANS-IE) on 3,066 first-time freshmen in two cohorts at a large public state university to determine if students who are retained their sophomore year would have a higher cumulative GPA than students who were not retained to their sophomore year. Using a t-test, the results indicated freshmen retained to their sophomore year ($M = 2.67$, $SD = .86$) demonstrated a statistically significant higher GPA ($t(3064) = 15.05$, $p < .05$) than those students who were not retained ($M = 2.11$, $SD = 1.10$) (Gifford et al., 2006). The data modeling approaches used and the results of these studies, especially those related to prediction in regards to college GPA, are both significant and relevant to the current study.

Program Type. Retention based on a program of study or major may be tracked by specific colleges or universities but is not nationally tracked and remains difficult to measure (Seidman, 2005). Program-specific issues, which may influence retention, vary by delivery (Craig & Ward, 2008). In Craig and Ward's (2008) study, which looked at a cohort of first-time, full-time students at a public community college in New England, initial program major was a significant predictor of success or failure in their logistic regression analysis. Students majoring in engineering or chemistry ($\beta = 1.54$, $\chi^2(1, N = 1729) = 12.85$, $p < .001$), business administration ($\beta = .73$, $\chi^2(1, N = 1729) = 4.27$, $p < .05$), and legal studies ($\beta = .75$, $\chi^2(1, N = 1729) = 4.18$, $p < .05$) had some of the lowest grade point averages, but resulted in student success as defined as being awarded a degree, a certificate, or transferring to another institution (Craig & Ward, 2008). Of the initial programs, engineering or chemistry majors had the highest odds ratio at 4.67. Business administration and legal studies both had a positive association with student success with odds ratios of 2.07 and 2.12 respectively (Craig & Ward, 2008).

Daempfle's (2003) article on first-year college majors highlighted lower enrollment, higher transfers to other disciplines, and lower retention rates were more prevalent among students majoring in mathematics, science, or engineering. St. John, Hu, Simmons, Carter, and Weber's (2004) logistic regression study indicated student major influences persistence decisions. This study, using the Indiana Commission for Higher Education's Student Information System, found White freshmen who major in social sciences ($\beta = -.82$, $p < .05$) or those who were undecided ($\beta = -.66$, $p < .01$) had a lower probability of persisting than other White students, although African American freshmen in the undecided majors were not significantly different from other African American

students in persistence (St. John et al., 2004). St. John et al. (2004) also found three distinct programs of study Health ($\beta = 1.09$, $p < .05$), Business ($\beta = 1.10$, $p < .01$), and Engineering or Computer Science ($\beta = 1.20$, $p < .05$) had positive associations with the persistence of African American sophomores, implying the economic potential of a major field had a substantial impact on the student's persistence. These studies are relevant to the current study in terms of the HOPE Career Grant which is specifically designed for in-demand diploma and certificate programs. The accessible population of the current study will include students enrolled in one of 17 program areas defined by the HOPE Career Grant Program. Program areas will be subdivided into four distinct groups of certificates with 9–17 credit hours, certificates with 18–36 credit hours, diplomas with 37–48 credit hours, and diplomas with 49–59 credit hours. Within these groupings, the impact of program type on retention will be significant to the current study.

Data Modeling and Data Mining Approaches Related to Student Retention

In addition to retention theories and variables used within past retention studies, a thorough review of data modeling and data mining approaches in regards to predictors of nontraditional student retention are necessary to identify which statistical procedures generate a more effective classification model. The following is an explanation of retention studies specifically focused on data modeling and data mining approaches.

Morris, Wu, and Finnegan (2005) investigated the accuracy and classification of students in online courses using a parametric method called linear discriminant analysis. The researchers identified the most important variables concerning predictive accuracy applying a linear classification rule to students enrolled in eCore® courses in the University System of Georgia (Morris et al., 2005). A subset of seven predictor variables

focused on students' demographic and academic information (gender, age, verbal ability, mathematic ability, current credit hours, high school achievement GPA, and college achievement GPA) (Morris et al., 2005). The objective was to determine how well a student can be correctly classified into dropout and completion based on his or her scores on the seven predictors (Morris et al., 2005). The number of students correctly predicted, that is the prediction accuracy of the model, was called the hit rate (Morris et al., 2005).

A two-group predictive discriminant analysis (PDA) was able to classify student dropout with an accuracy of 52.6% and completion with 66.1%, and the overall hit rate was 62.8% (Morris et al., 2005). Test statistics indicated actual classification results were better than chance ($z = 2.26$; $p < .05$) (Morris et al., 2005). To determine which predictors were the most important in terms of the predictive power of accuracy, the researchers performed seven analyses leaving one variable out each time (Morris et al., 2005). The results indicated high school GPA (0.48) and SAT math score (0.56) were the most important predictors because their leave-one-out hit rate decreased the most (Morris et al., 2005). High school GPA and SAT math score were considered to be the most important predictors (Morris et al., 2005).

Although retention has been thoroughly studied using parametric methods, few studies on retention take advantage of the strong predictive power associated with data mining tools (Herzog, 2006). Yu, DiGangi, Jannasch-Pennell, and Kaprolet (2010) explored three data mining techniques, classification trees, multivariate adaptive regression splines (MARS), and neural networks using transferred hours, residency, and ethnicity as factors to retention. Yu et al. (2010) tracked the continuous enrollment or withdrawal of 6,690 sophomore students enrolled at Arizona State University using

demographic, pre-college academic performance indicators, and online class hours. A classification tree was used to rank order the factors which affect retention by dividing the original group of data into pairs of subgroups (Yu et al., 2010). The resulting *G2* value was the likelihood ratio for testing the independence of the outcome and predictor variables (Yu et al., 2010). A larger *G2* value indicated a more significant split (Yu et al., 2010). The results showed ethnicity was the third largest *G2* value demonstrating significant splits (*G2* = 4,326.84) (Yu et al., 2010).

Mendez, Buskirk, Lohr, and Haag (2013) investigated how classification trees and random forests could be used to identify factors associated with persistence to a science or engineering degree not found by logistic regression. The study, which looked at freshman students who were STEM majors from 1999 to 2000, used institutional data which included 18 demographic, cognitive, and non-cognitive variables (including work-study status, number of courses, and financial aid support) (Mendez et al., 2013). Using a classification tree, the data were initially split on high school GPA and no additional predictors of persistence in engineering were evident when student high school GPA was below 3.59 (Mendez et al., 2013). The Gini index, the proportion of students in the node who persist, for terminal node 1 was 0.22 which suggested most of the students in the node were in one category or the other (Mendez et al., 2013). In comparison, the results of the classification tree were consistent with the logistic regression, but the importance scores from the classification tree model provided additional information on the results of the logistic regression (Mendez et al., 2013). Mendez et al. (2013) found while high school GPA was a strong predictor of persistence in the logistic regression model ($\chi^2(1) = 15.16$, $p < .001$) with an odds ratio of 4.81, the classification tree indicated GPAs below

3.59 were a risk factor for non-persistence. Although citizenship status did not appear in the classification tree, citizenship and ethnicity had a concordance measure of 89.2 which meant if the ethnicity of a freshman was unknown, the citizenship status of the student could be used in the node decision (Mendez et al., 2013).

A traditional random forest is a model which consists of multiple classification trees where variables are randomly sampled as candidates at each split (Kuhn & Johnson, 2013). Mendez, et al. (2013) used variable importance scores from the random forest to find the optimal subset of variables to build a single classification tree. The random forest method can list variables in order of predictive ability or importance giving researchers the ability to reduce a large set of variables to a working subset without making any model assumptions (Mendez et al., 2013). Of the importance scores for all 18 variables, the highest score belonged to cumulative GPA (Mendez et al., 2013). Except for high school GPA, the variables for the random forest are the same as those using stepwise selection in logistic regression (Mendez et al., 2013). According to the logistics regression results, high school GPA was masked by the other predictors so that cumulative GPA was moderately correlated with high school GPA ($r = .49$, $p < .001$), therefore high school GPA was not included in the model as statistically significant (Mendez et al., 2013). In contrast, the random forest method identified the importance of high school GPA and ranked it second in predicting STEM persistence (Mendez et al., 2013). In summary, the researchers found classification trees and random forests identified factors and complex relationships not found by other statistical methods (Mendez et al., 2013).

Jia and Mareboyana (2013) explored the effectiveness of machine learning techniques to determine factors influencing student retention at Historically Black Colleges and Universities (HBCU) and to create retention predictive models. Based on full-time, first-time undergraduate students which were tracked for six years, a support vector machine (SVM) algorithm resulted in cumulative GPA and total credit hours as impacting retention (Jia & Mareboyana, 2013). SVM creates separate hyperplanes with a maximum margin separator or parameter (Kuhn & Johnson, 2013). The SVM function attempts to separate the classes into either side of the plane by a specified margin (Kuhn & Johnson, 2013). Using the nonlinear SVM boundary, the researchers mapped the data into a new *z* space using a Kernel function and changed the curve to a line, and created a retention regression (Jia & Mareboyana, 2013). This improved the model's accuracy to 94% (Jia & Mareboyana, 2013).

Summary

While there is prolific literature on the challenges and struggles facing nontraditional students, very little literature focuses on how the student's unique characteristics contribute to retention specific to the community and technical college environment. Institutional leaders are challenged with how to engage the different student populations like nontraditional students (Wyatt, 2011). Because community college students often meet more than one definition of nontraditional, a thorough understanding of the definitions used by researchers in studying nontraditional community college students is necessary (Kim, 2002). Likewise, as this population is exposed to the college environment, the retention of nontraditional students is notably important as it relates to interaction with peers, the classroom, and the campus environment (Wyatt, 2011). In

addition to understanding the changing demographics of nontraditionals concerning retention, it is imperative to create a better understanding of this population in the Technical College System of Georgia. To that end, various theories, models, and frameworks were investigated, focusing on both traditional-age students in the university setting and students in the community college or career and technical education setting.

Monroe (2006) asserted the complex, dynamic nature of nontraditional students requires continuous examination and refinement of our understanding of this population's changing demographics concerning attrition. A thorough explanation of variables, included in past retention studies and this study, was provided. In addition to retention theories and variables used within past retention studies, a thorough review of data modeling and data mining approaches in regards to predictors of nontraditional student retention were necessary to identify which statistical procedures generate a more effective classification model. In summary, this review of the current literature provided the issues and difficulties most often associated with nontraditional students and explored best practices for this population.

Chapter III

METHODOLOGY

This chapter contains a description of the research methodology, design, and procedures used to answer the two research questions in this study. The first section describes the quantitative research design, the rationale for its use, and the variables used in the study. The second section details the population of interest, while the third and fourth sections explain the data collection and analysis procedures. Also, statistical considerations and assumptions for each model are discussed.

The following research questions guide the proposed study:

1. Are environmental factors, background factors, and academic integration components significant predictors of nontraditional student retention for certificates or diplomas?
    a. Are environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) significant predictors of nontraditional student retention for certificates 9–17 credit hours in length?
    b. Are environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation

date), and academic integration components (student GPA and program type) significant predictors of nontraditional student retention for certificates 18–36 credit hours in length?

c. Are environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) significant predictors of nontraditional student retention for diplomas 37–48 credit hours in length?

d. Are environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) significant predictors of nontraditional student retention for diplomas 49–59 credit hours in length?

2. Does one of the selected statistical procedures generate a more accurate classification model based on Cohen's Kappa, ROC curves, and sensitivity and specificity by certificate or diploma type?

Research Design

A nonexperimental, ex post facto, correlational research design was used in this study. Archival data obtained from the Technical College System of Georgia were retrospectively analyzed to measure first-year retention. Therefore, a nonexperimental, ex post facto research design was more appropriate for this study as the independent

predictor variables were not manipulated. Because the goal was to predict values on a binary outcome variable, the researcher attempted to identify which prediction model, out of two data modeling approaches and three data mining approaches, best predicts whether a student will be retained or not retained. Supervised statistical learning involves building a statistical model for predicting, or estimating, an output based on one or more inputs (James et al., 2013). By testing multiple statistical models to illustrate the classification power of these models, researchers are better equipped to provide timely data and information to key decision-makers (Knowles, 2014).

Independent variables were aligned with the academic, background, and environmental factors described in Bean and Metzner's model. There were two continuous predictor variables representing background and academic factors (age and GPA). There were five dichotomous variables (gender, high school graduation date, single parent indicator, displaced homemaker indicator, Pell eligibility indicator), two nominal variables (race and HOPE program of study), and one ordinal variable (high school diploma type) representing background, environmental, and academic factors. Table 1 describes the categories for the dichotomous, nominal, and ordinal independent variables of this study.

Table 1

Categories and Codes for Dichotomous, Nominal, and Ordinal Independent Variables

Variable	Description	Categories and Codes
Race or Ethnicity (racecode)	Race or ethnicity of the student.	1 = White 2 = Black 3 = Hispanic 4 = Other
Gender (gencode)	Gender of the student.	0 = Male 1 = Female
High School Diploma Type (hsdipcode)	Type of high school diploma the student graduated with.	1 = Certificate of Attendance; Certificate of Performance; Special Needs Certificate 2 = General Educational Development Diploma 3 = Home School Diploma; Foreign Diploma; Vocational; Tech Prep; College Prep Diploma
High School Graduation Date (gradcode)	Date of student's high school graduation. Variable was coded 0 if the student has been out of high school for four years or less, coded 1 if the student has been out of high school for at least five years or more.	0 = Four years or less 1 = Five years or more
Single Parent Status (sparcode)	If a student self-identified as a single parent, the variable was coded 1. Otherwise, the variable was coded 0.	0 = No 1 = Yes
Displaced Homemaker Status (dhomcode)	If a student self-identified as a displaced homemaker, the variable was coded 1. Otherwise, the variable was coded 0.	0 = No 1 = Yes
Pell Eligibility (pellcode)	If a student did not receive the Pell Grant, the variable was coded 0. If the student did not receive the Pell Grant, the variable was coded 1.	0 = No 1 = Yes
HOPE Career Grant Program of Study (hopepos)	Based on a list of 426 HOPE Career Grant major codes, this variable was coded to one of 17 program areas as defined by the Georgia Student Finance Commission. This variable was further collapsed into seven generalized industry/occupational areas which were consistent across all 22 technical colleges in Georgia.	1 = Not HOPE Career Grant Program 2 = Cyber and Related; Engineering; Film; Healthcare and Public Service Technologies 3 = Industrial Technologies; Manufacturing; Welding and Joining 4 = Transportation and Logistics

College student retention was the one dependent variable for this study. Retention was a dichotomous variable coded as 0 for not retained and 1 for retained. First-time fall cohort students who are still enrolled the following fall were considered retained.

Population

In academic year 2016, 131,644 students were enrolled in credit courses in technical colleges across Georgia, and in academic year 2017, 133,081 students were enrolled. Among the total population of 264,725 students, the percentage of females accounted for 63% (*N* = 166,777), whereas males accounted for 37% (*N* = 97,948). The overall ethnic proportion of this total population consisted of predominantly White (47.5%) and Black (39.3%) students, whereas 7.4% were Hispanic and 2.0% were Asian. The number of students who received the Pell grant was 140,569 representing 53.1% of this population.

The target population included students identified as nontraditional at each of the 22 technical colleges in Georgia. The accessible population included first-time students identified as nontraditional at each of the 22 technical colleges in Georgia who were enrolled in one of 17 program areas defined by the HOPE Career Grant Program. The expected cohort size was approximately 8,000 students per cohort for a total of 16,000 students. Program areas were subdivided into four distinct groups of certificates with 9–17 credit hours, certificates with 18–36 credit hours, diplomas with 37–48 credit hours, and diplomas with 49–59 credit hours. Students were classified as nontraditional if they meet all three of the following criteria:

- First-Time - Beginning student (queried from Banner field *Student Type*)
- Age - 25 years old or older (calculated from Banner field *Date of Birth*)

- Enrollment Status - Part-time (calculated from Banner field *Earned Hours*)

The cohort consisted of nontraditional students who were enrolled for the first time at any of the technical colleges in Georgia and were not high school students. First-time students identified as special admit or learning support were not included in this study as they cannot receive federal financial aid. The cohort period of fall to fall was used to measure first-year retention.

Data Collection

Once the Institutional Review Board (IRB) granted permission, the IRB approval as shown in Appendix A was submitted to the Office of Accountability and Institutional Effectiveness at the Technical College System of Georgia. Once final approval was acknowledged from TCSG, a request for data was submitted to TCSG's Division of Data, Planning, and Research. Specifically, the Banner student information system maintained by TCSG's Knowledge Management System (KMS) was utilized for this study. All 22 technical colleges in Georgia utilize Banner as their student information system. Banner is a comprehensive, integrated information and management system which allows financial and student data to be shared by multiple users through relational databases of information (Ellucian, n.d.). With the assistance of the Data Compliance Manager and the Reporting Manager, a data script was created which was used to query the database.

Data was collected from each of the 22 technical colleges in Georgia for first-time students identified as nontraditional who were enrolled in diploma and certificate programs for academic year 2017-2018 and academic year 2018-2019. Based on state standards, diploma programs vary in length from 37 to 59 hours and certificate programs vary in length from 9 to 36 semester credit hours. Year 1 to year 2 enrollment was

measured from first-time enrollment during the fall term to the following fall term. Two separate data files were requested from the Data Compliance Manager and the Reporting Manager at TCSG. The Data Compliance Manager recoded the student identification (ID) number field to a new nominal number before releasing the data files to maintain student confidentiality. Variables needed to address each research question were recoded based on data type and measurement. No interaction or intervention with students was necessary for the collection of data used in this study. All data was stored securely and backup copies were created before any modifications were made.

Data Analysis

The data analysis section first describes the descriptive statistics used in the study. To address Research Question 1 and 2, the statistical considerations and assumptions for the study were described followed by a discussion of the inferential statistics for each statistical model.

Descriptive Statistics. Data analysis was conducted using the statistical software package R. Datasets were loaded and the recoding of data types and factors were evaluated. A combination of descriptive and inferential statistics was used in the analysis of the data. Descriptive statistics such as the mean, median, and standard deviation (*SD*) were calculated for the continuous variables of age and GPA. Each dichotomous variable (gender, code for high school graduation date, single parent indicator, displaced homemaker indicator, Pell eligibility indicator), nominal variable (race and HOPE program of study), and ordinal variable (high school diploma type) were summarized with percentages and frequency data.

Statistical Considerations and Assumptions. Data for two academic years (2017-2018 and 2018-2019) were partitioned into a training data set and a test data set to be used to implement two data modeling approaches (logistic regression and linear discriminant analysis) and three data mining approaches (classification tree, random forest, and support vector machine models). The training data set was used to build the model and the test data set was used to estimate the model's predictive performance (Kuhn & Johnson, 2013). Data were examined for missing values, special values, corrupt data, and outliers. The specific type of data cleaning was determined by the predictor variables and type of model being used (Kuhn & Johnson, 2013). Multicollinearity and near zero variance were considered in preprocessing the data. All recoding, deletions, or data transformations were documented. Logistic regression requires observations to be independent of each other and requires little or no multicollinearity among the independent variables (James et al., 2013). Logistic regression requires a large sample size and assumes linearity of the independent variables and the log odds (James et al., 2013). Similar to logistic regression, linear discriminant analysis assumes little or no multicollinearity (Tabachnick & Fidell, 2013). The independent variables in linear discriminant analysis (LDA) are assumed to have a multivariate normal (Gaussian) distribution (James et al., 2013). Violating this assumption is normally acceptable as long as the sample size is large enough (James et al., 2013). However, because of this assumption, LDA does not discriminate among categories using a mix of continuous and categorical variables. The population variances and covariances for all independent variables are required to be equal across the dependent variable groups (Spicer, 2005). Stated differently, the values of each variable vary around the mean by the same amount

on average (Spicer, 2005). This is known as the homogeneity of variance-covariance matrices assumption (Spicer, 2005).

The classification and regression tree (CART) methodology requires no distributional assumptions for predictor variables and is resistant to outliers, multicollinearity, and heteroscedasticity (Breiman, Friedman, Olshen, & Stone, 1984). Like classification trees, there are no formal distributional assumptions with random forests (Breiman et al., 1984). Random forests are nonparametric and can tolerate skewed data as well as categorical data which are ordinal or nonordinal (Breiman et al., 1984). Assumptions for support vector machines are the margin should be as large as possible and the support vectors are the most useful because they are the data points most likely to be incorrectly classified (Kuhn & Johnson, 2013).

Inferential Statistics. To identify nontraditional student characteristics which best predict first-year retention, a statistical learning approach was applied in this study. Statistical learning refers to tools and techniques for understanding data (James et al., 2013). More specifically, supervised statistical learning involves building a statistical model for predicting, or estimating, an output based on one or more inputs (James et al., 2013). By testing multiple statistical models to illustrate the classification power of these models, researchers are better equipped to provide timely data and information to key decision-makers (Knowles, 2014). Instead of identifying one single best model, the researcher evaluated several models to identify the most accurate predictions on future student cohorts. To address Research Question 1, the coefficients, standard error, odds ratios, p-values, and confidence intervals were evaluated for each predictor. Predictors were considered statistically significant at the .05 level. Model-specific procedures were

employed iteratively to arrive at the final model of significant predictor variables. Because there are many different metrics available to evaluate prediction models, the researcher utilized three common metrics used to evaluate binary classification datasets. Knowles (2014) stated because of the complexity of the model building process, every aspect of the modeling process is crucial in balancing the tradeoff between accuracy and complexity. Therefore, the accuracy metric, Cohen's Kappa statistic, the ROC curve metric (area under the curve), sensitivity, and specificity were used in various R packages including tidyverse and tidymodels packages to evaluate the accuracy of each model and to address Research Question 2. The overall accuracy metric reflects the agreement between the observed and predicted classes (true positives and true negatives) (Kuhn & Johnson, 2013).

The Kappa statistic is a measure of how well the classifier performed as compared to how well it would have performed simply by chance (Kuhn & Johnson, 2013). The Kappa statistic can take on values between −1 and 1 where a value of 0 means there is no agreement between the observed and predicted classes, and a value of 1 indicates perfect agreement between the model prediction and the observed classes (Kuhn & Johnson, 2013). A confusion matrix was used to describe the performance of each model. The confusion matrix included true positives, true negatives, false positives (type I error), and false negatives (type II error) (Kuhn & Johnson, 2013). From the confusion matrix, several rates were calculated. The accuracy rate is calculated by adding true positives and true negatives then dividing by the total (Kuhn & Johnson, 2013). The true positive rate (sensitivity) is calculated by dividing the true positives (students correctly predicted as retained) by the actual number of retained students (Kuhn & Johnson, 2013). Likewise,

the true negative rate (specificity) is calculated by dividing the true negatives (students correctly predicted as not retained) by the actual number of students not retained (Kuhn & Johnson, 2013). The ROC curve metric was used to summarize the performance of the classifier over varying thresholds by plotting the true positive rate against the false positive rate (Kuhn & Johnson, 2013).

Logistic Regression. Logistic regression analysis was used to determine if a binary outcome variable illustrates predictive differences between 10 independent predictor variables. In instances where the dependent variable is dichotomous and the independent variables are categorical or a mix of continuous and categorical, logistic regression is appropriate (Burns & Burns, 2008). Cross-validation was used on a sample of the training data to obtain additional information about the fitted model. An important consideration is model fit as adding independent variables will increase the amount of variance explained in the log odds (James et al., 2013). Cross-validation can be used to both estimate the test error to evaluate performance or to select the appropriate level of flexibility (James et al., 2013). The sample training data was divided into 10 folds. James et al. (2013) explained this value has been shown empirically to yield test error rate estimates which have low bias and low variance. The first fold will be treated as a validation set and the statistical method will be fit on the remaining folds. The misclassified observations were used to quantify the test error. After fitting the model, the overall model fit was tested using the McFadden pseudo R^2 index. The logistic_reg() function used the glm engine to express the coefficients, standard errors, the z-statistic (Wald statistic), and the associated p-values. The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable

(James et al., 2013). Then the predicted probability was calculated followed by the plotting of a ROC curve and calculating the AUC (area under the curve) to assess the performance of the model. In addition to the ROC metric, which includes sensitivity and specificity, the accuracy metric, which includes Kappa, was evaluated.

***Linear Discriminant Analysis*.** Similar to logistic regression, linear discriminant analysis is a common multivariate statistical method used to analyze categorical outcome variables (James et al., 2013). Linear discriminant analysis focuses on determining which variable discriminates between two or more classes and is used to develop a classification model for predicting the group membership of new observations (Spicer, 2005). It does this by maximizing the distance between the means of each class and minimizing the variation (scatter) within each class (Kuhn & Johnson, 2013). LDA was used to make predictions by estimating the probability that a new set of inputs belongs to each class. The model used Bayes Theorem to estimate the probabilities and the class which gets the highest probability will be the output class. As in logistic regression, cross-validation was used to estimate how accurate the LDA predictive model may be in actual practice. The discrim_linear() function used the MASS engine to specify the model. The model described the probability of randomly selecting an observation from each of the classes from the training data, the mean value for each of the independent variables for each class, the coefficients of the linear discriminants, which defines the coefficient of the linear equation which is used to classify the response classes. For this study, there were only two response classes, therefore there was only one set of coefficients. The training data was verified using the predict() command and the same command was used to run the test data against the model to determine its accuracy. The

classification accuracy and error were computed by comparing the observed classes in the test data against the predicted classes based on the model. A confusion matrix was used to display how the observations were assigned in the actual group and the predicted group, and the resulting misclassifications. The confusion matrix included the model accuracy, Kappa, sensitivity, and specificity. A ROC curve was computed and plotted to understand the impact on sensitivity and specificity as the threshold for the classifier is changed.

Classification Tree. Two tree-based methods were explored: classification trees and random forests. A classification tree was used to predict that each observation belongs to the most commonly occurring class of training observations in the node or region to which it belongs (James et al., 2013). A decision tree creates separations between groups and subgroups, partitioning the data into smaller, homogeneous groups. Kuhn and Johnson (2013) define pure homogeneity in classification by maximizing accuracy or minimizing misclassification error. The CART methodology was used to identify the subgroups by selecting the best possible variable (primary splitter) to partition the parent node into two child nodes. Breiman et al. (1984) define CART as a nonparametric methodology which can be used to classify data involving either outcome or predictor variables which are categorical, ordinal, or continuous. The rpart.plot() function in R was used to fit the classification tree while evaluating the relative error for different splits. This process evaluated the number of terminal nodes and the misclassification error rate. The Gini Impurity Index was used throughout the partitioning process to select the best split among the values of the predictor which results in the lowest impurity measure. James et al. (2013) state when building a classification tree, the

Gini index is typically used to evaluate the quality of a particular split being this approach is more sensitive to node purity than is the classification error rate. The tree was pruned and refitted using a training set and a test set. The training set tree was used to make predations using the test set. Cross-validation was used to prune the tree optimally and control for variance. As in previous models, a confusion matrix included the model accuracy, Kappa, sensitivity, and specificity. A ROC curve was used to visualize model performance.

***Random Forests*.** While decision trees are easy to interpret and tolerate different types of predictors and missing data, they can suffer from model instability and may not produce optimal predictive performance (Breiman et al., 1984; Doyle & Donovan, 2014). The stability and predictive performance of decision trees can be substantially improved by aggregating using the random forests method (James et al., 2013). Random forests refer to a model of the entire system of random decision trees which are essential in predictive modeling for regression, classification, and analyses, which function by forming an array of classification trees at test time and releasing the group which appears most frequently of the groups or average forecast (regression) of the particular trees. At each split in the tree, a random sample of predictors is chosen as split candidates from the full set of predictors (James et al., 2013). This differs from bagging where all of the original predictors are considered at every split (Kuhn & Johnson, 2013). James et al. (2013) refer to this as decorrelating making the average of the resulting trees less variable and more reliable in random forests. The rand_forest() function in R used the random forest engine to make splits using both a training dataset and a test dataset. The random forest was fit on the training dataset using default parameters and evaluated the

percentage of variance explained based on the out-of-bag estimates (error estimates). The random forest was tuned by adjusting the node size (min_n) and the number of variables randomly sampled at each stage (mtry). Predictions were made and compared on both the training dataset and the test dataset.

***Support Vector Machine*.** The final model explored was a support vector machine which uses a hyperplane to separate two categories of data (Cortes & Vapnik, 1995; James et al., 2013; Marbouti, Diefes-Dux & Madhavan, 2016). A support vector machine model was used to find the margin, which is the distance between the classification boundary and the closest training set data point (Kaiser, Meyers, Morrison, & Skelton, 2016; Kuhn & Johnson, 2013). In essence, the margin, defined by these data points, can be quantified and used to evaluate the performance of the model (Kuhn & Johnson, 2013). SVM is sensitive to the training set samples which are closest to the boundary (Marbouti et al., 2016). Since the prediction equation is supported by training set data points only, the maximum margin classifier is usually called the support vector machine (Kuhn & Johnson, 2013). Like previous models, the data was split into a training set and a test set and then fit the model on the training data. Before the model was trained, repeated k-fold cross-validation was used to determine the overall accuracy estimate of the trained model. Predictions were made on the test data and then predictions were plotted. A confusion matrix was used to evaluate and compare prediction, accuracy, p-value, Kappa, sensitivity, and specificity. The choice of the kernel function and its parameters along with the cost value was used to control the complexity of the model to avoid over-fitting the training data. The tune() function was used to test parameters and identified which value produces the best fitting model. A cross-validation of the true

versus predicted values was computed and the performance of the training set and testing set methods were computed by comparing the accuracy rates and the Kappa coefficients.

Summary

This chapter outlined the research design and methodology for studying the ability of multiple prediction models to predict whether a student will be retained or not retained. A nonexperimental, ex post facto, correlational research design was used to retrospectively analyze first-year retention in nontraditional students in the Technical College System of Georgia. Specifically, the predictability of academic, background, and environmental factors such as GPA, program choice, high school diploma type, high school graduation date, financial aid eligibility, and disadvantaged student status on the retention of nontraditional students enrolled in diploma and certificate programs was examined. The population was identified as nontraditional students who are 25 years of age or older and were enrolled part-time. Further, the accessible population was first-time nontraditional students at each of the 22 technical colleges in Georgia who were enrolled in one of 17 program areas defined by the HOPE Career Grant Program. Data were obtained from the Technical College System of Georgia. Two data modeling approaches (logistic regression and linear discriminant analysis) and three data mining approaches (classification tree, random forest, and support vector machine models) were used to best predict first-year retention.

Chapter IV

RESULTS

The purpose of this study was to examine the predictability of academic, background, and environmental factors such as Pell eligibility, single parent status, displaced homemaker status, age, race or ethnicity, gender, high school diploma type, high school graduation date, student grade point average, and program type on the retention of nontraditional students enrolled in diploma and certificate programs in the Technical College System of Georgia. To examine the predictability of academic, background, and environmental factors on the retention of nontraditional students in the Technical College System of Georgia, archival data were retrospectively analyzed to measure first-year retention. Multiple prediction models were examined to predict whether a student would be retained or not retained. The analysis specifically focused on two cohorts of diploma and certificate-seeking students who began their enrollment in fall 2017 and fall 2018. The cohorts consisted of nontraditional students who were enrolled for the first time at any of the technical colleges in Georgia and were not high school students. A statistical learning approach was used to evaluate several models to identify the most accurate predictions on future student cohorts.

This chapter presents the results of analyses conducted to answer the following research questions for this study.

1. Are environmental factors, background factors, and academic integration components significant predictors of nontraditional student retention for certificates or diplomas?
    a. Are environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) significant predictors of nontraditional student retention for certificates 9–17 credit hours in length?
    b. Are environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) significant predictors of nontraditional student retention for certificates 18–36 credit hours in length?
    c. Are environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) significant predictors of nontraditional student retention for diplomas 37–48 credit hours in length?
    d. Are environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or

ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) significant predictors of nontraditional student retention for diplomas 49–59 credit hours in length?

2. Does one of the selected statistical procedures generate a more accurate classification model based on Cohen's Kappa, ROC curves, and sensitivity and specificity by certificate or diploma type?

This chapter is organized into four sections. The first section presents data screening and descriptive statistics for each of the eight data files. The second section of the chapter presents data preprocessing and feature engineering related to missing data, normality, outliers, multicollinearity, and linearity to meet the statistical assumptions required for each data mining and data modeling approach. The third section of the chapter addressed model training and the statistical significance of the variables in each model, while the fourth section addressed the accuracy of the various classification models. The chapter concludes with a summary highlighting the main findings.

Data Screening and Descriptive Statistics

The 2017-2018 and 2018-2019 datasets provided by TCSG originally contained 37,177 total records (18,866 for the 2017 data and 18,311 for the 2018 data) which represented 14,448 unique student IDs. TCSG's Data Compliance Manager recoded the student ID number field to a dummy ID number before releasing the data files to maintain student confidentiality. For 134 records, student major was coded as "special admits", "learning support", and "institutionally accepted." To be included in the analysis, the major must be a valid program. For this reason, these records were removed

leaving 14,314 records. Records with major codes above the diploma level were identified as associate degree-seeking students (1,328 for the 2017 data and 1,846 for the 2018 data). Because associate degree analysis falls outside of the scope of this research, these records were removed leaving 11,140 records. Term data were provided for any first-time student in academic years 2017-2018 and 2018-2019 who were 25 years of age or older and initially enrolled part-time. If a student was enrolled, term data could have included Fall 2016, Spring 2017, Summer 2017, Fall 2017, Spring 2018, Summer 2018, Fall 2018, Spring 2019, and Summer 2019. The following data were included in each dataset: dummy student ID, term of enrollment, student enrollment type, major code, major, award level, age, race, gender, hours enrolled, student enrollment status, high school diploma type, high school graduation date, GPA, and indicators for single parent status, displaced homemaker, and Pell grant eligibility. Students who were 24 years old were included in each cohort if they were identified as 25 years old during one or more terms when they were coded as a beginning student. Term GPA was provided for each student, not cumulative GPA. Therefore, the term GPA for the student's last semester of enrollment was used in the analysis. Program length in credit hours and HOPE program of study were additional fields (from TCSG's Knowledge Management System) used to determine classifications for student retention and whether or not a major was an approved HOPE Career Grant program. The 2017-2018 data were used as the training data and the 2018-2019 data were used as the test data.

Before analysis, several items were recoded or renamed in preparation for data cleaning. Items used for analysis were coded as continuous (age and GPA), dichotomous (gender, code for high school graduation date, single parent indicator, displaced

homemaker indicator, Pell eligibility indicator, and retained), nominal (race and HOPE program of study), and ordinal (high school diploma type). Enrollment term, program length in credit hours, and hours enrolled were used to derive the variable, "retained." Students were labeled as "retained" if one of the following conditions were met:

- if enrolled in Fall 2016 and student completed or was still enrolled in Fall 2017;
- if enrolled in Fall 2017 and student completed or was still enrolled in Fall 2018;
- if enrolled in Spring 2017 and student completed in Spring 2017; or
- if enrolled in Spring 2018 and student completed in Spring 2018.

Table 2 contains the demographics of beginning students enrolled in technical certificates 9 to 17 credit hours in length and 18 to 36 credit hours in length in both cohorts of 2017 (representing the 2017-2018 academic year) and 2018 (representing the 2018-2019 academic year). The 2017 dataset for certificates 9 to 17 credit hours totaled 1,277 records, contained more males (66.6%) than females (33.4%) and the majority of students did not receive the Pell grant (97.7%) compared to those who did (2.3%). The high percentage of students not receiving the Pell grant was likely due to most smaller certificates not qualifying for the Pell grant. White students accounted for 53.3% of the dataset, Black students 41.8%, Hispanic students 3.0%, and all other races represented 1.9%. The majority of the dataset did not self-identify as a single parent (93.0%) and 77.1% did not identify as a displaced homemaker. Students who enrolled in Transportation and Logistics programs accounted for 23.1% of the dataset, while the majority of students (54.3%) were not enrolled in a HOPE Career Grant major. The

variables for high school diploma type and how long a student had been out of high school were both missing 32.6% of the values. Of the 1,277 valid values, 70.8% were classified as having graduated with a GED®. Students who had been out of high school for four years or less accounted for 56.2% of the valid values compared to those who had been out of high school for at least five years or more (43.8%).

The 2018 dataset for certificates 9 to 17 credit hours was similar to the 2017 cohort. For 1,264 records, there were more males (63.3%) than females (36.7%) and the majority of students did not receive the Pell grant (97.8%) compared to those who did (2.2%). Similar to students in the 2017 cohort, White students represented 52.0% of the population while Black students represented 42.9% of the cohort. Hispanic students accounted for 3.1% and all other races represented 2.1%. The majority of the dataset did not self-identify as a single parent (94.3%) and 84.1% did not identify as a displaced homemaker. Students not enrolled in a HOPE Career Grant major represented 55.3% of the cohort, while those enrolled in Transportation and Logistics programs accounted for 19.1%. The variables for high school diploma type and how long a student had been out of high school were both missing 35.3% of the values. Of the 1,264 valid values, 64.6% were classified as having graduated with a GED®. Students who had been out of high school for four years or less accounted for 52.3% of the valid values compared to those who had been out of high school for at least five years or more (47.7%).

The 2017 and 2018 cohorts for certificates 18 to 36 credit hours totaled 1,710 and 1,183 records, respectively, and shared similar demographics as the previous cohorts but also included key differences. Unlike the previous cohorts, both datasets for certificates 18 to 36 credit hours contained more females than males. The 2017 cohort was 77.6%

female and 22.4% male, and the 2018 cohort was 73.6% female and 26.4% male. In contrast to the previous cohorts, the majority of students (64.6% in the 2017 cohort) received the Pell grant (56.2% in the 2018 cohort). The race or ethnicity of students enrolled in certificates 18 to 36 credit hours was similar to the previous cohorts except for the race or ethnicity outside of Black or White students. For the 2017 cohort, Black students accounted for 46.5% of the dataset with White students representing 41.3%. In the 2018 cohort, 46.6% of the population were Black students while White students accounted for 40.6%. In both cohorts, the percentages for Hispanic students and all other races were similar with Hispanic students representing 6.0% and 6.8% respectively, and all other races representing 6.2% and 6.0% respectively. The majority of the dataset did not self-identify as a single parent (85.7% and 87.2%) and only a small percentage (4.5% and 5.6%) identified as a displaced homemaker. In the 2017 cohort, the majority of students were not enrolled in a HOPE Career Grant major and the next highest percentage was students enrolled in Cyber, Engineering, or Healthcare fields (39.1%). But those percentages are reversed in the 2018 cohort with Cyber, Engineering, or Healthcare programs represented 54.8%, while 30.8% were enrolled in programs that were not considered HOPE Career Grant programs. The variables for high school diploma type and how long a student had been out of high school were both missing a small percentage of values. Of the 1,644 valid values in the 2017 cohort, 49.6% were classified as having graduated with a college prep or tech prep high school diploma (47.9% in the 2018 cohort), differing from the previous cohorts where the majority of students had a GED®. Another difference was 81.2% and 80.4% represented students who had been out of high

school for five years or more compared to those who had been out of high school for four years or less.

Table 2

Demographics for Students Enrolled in Certificate Programs

	9 to 17 credit hours		18 to 36 credit hours	
	2017	2018	2017	2018
Race or Ethnicity				
White	681 (53.3%)	657 (52.0%)	706 (41.3%)	480 (40.6%)
Black	534 (41.8%)	542 (42.9%)	795 (46.5%)	551 (46.6%)
Hispanic	38 (3.0%)	39 (3.1%)	103 (6.0%)	81 (6.8%)
Other	24 (1.9%)	26 (2.1%)	106 (6.2%)	71 (6.0%)
Gender				
Male	851 (66.6%)	800 (63.3%)	383 (22.4%)	312 (26.4%)
Female	426 (33.4%)	464 (36.7%)	1,327 (77.6%)	871 (73.6%)
High School Diploma Type				
Certificate of Attendance	55 (4.3%)	64 (5.1%)	265 (16.1%)	185 (16.4%)
GED®	904 (70.8%)	817 (64.6%)	563 (34.2%)	401 (35.6%)
College Prep/Tech Prep	318 (24.9%)	383 (30.3%)	816 (49.6%)	539 (47.9%)
High School Graduation				
Four years or less	718 (56.2%)	661 (52.3%)	309 (18.8%)	221 (19.6%)
Five years or more	559 (43.8%)	603 (47.7%)	1,335 (81.2%)	904 (80.4%)
Single Parent				
No	1,188 (93.0%)	1,192 (94.3%)	1,466 (85.7%)	1,032 (87.2%)
Yes	89 (7.0%)	72 (5.7%)	244 (14.3%)	151 (12.8%)
Displaced Homemaker				
No	984 (77.1%)	1,063 (84.1%)	1,633 (95.5%)	1,117 (94.4%)
Yes	293 (22.9%)	201 (15.9%)	77 (4.5%)	66 (5.6%)
Pell Eligibility				
No	1,248 (97.7%)	1,236 (97.8%)	606 (35.4%)	518 (43.8%)
Yes	29 (2.3%)	28 (2.2%)	1,104 (64.6%)	665 (56.2%)
HOPE Program				
Not HOPE Career Grant	693 (54.3%)	699 (55.3%)	893 (52.2%)	364 (30.8%)
Cyber, Eng., Healthcare	99 (7.8%)	102 (8.1%)	669 (39.1%)	648 (54.8%)
Industrial Technologies	190 (14.9%)	221 (17.5%)	49 (2.9%)	59 (5.0%)
Transportation/Logistics	295 (23.1%)	242 (19.1%)	99 (5.8%)	112 (9.5%)
Total	1,277	1,264	1,710	1,183

Table 3 and Table 4 include the descriptive statistics for both certificate cohorts. Table 3 contains the descriptive statistics for the 2017 dataset (N = 1,277) and the 2018 dataset (N = 1,264) for certificates 9 to 17 credit hours. Descriptive statistics indicated the overall mean age as M = 38.3 (SD = 9.7) with a range from 25 to 77 and average GPA as M = 2.80 (SD = 1.4) for the 2017 cohort. The 2018 cohort had similar statistics with a mean age of M = 38.1 (SD = 9.9) with a range from 25 to 76 and an average GPA of M = 2.85 (SD = 1.4). In both cohorts, age is moderately skewed right, and GPA is substantially skewed left with higher concentrations in the lower tail and upper tail.

Table 3

Descriptive Statistics for Continuous Variables in Certificates 9 to 17 Credit Hours

		N	*M*	*Mdn*	*SD*	Min. value	Max. value	Skew	Kurtosis
2017	age	1,277	38.26	37	9.67	25	77	0.69	0
	gpa	1,277	2.80	3.33	1.42	0	4	-1.17	-0.14
2018	age	1,264	38.06	37	9.89	25	76	0.70	-0.14
	gpa	1,264	2.85	3.40	1.38	0	4	-1.22	0.07

Table 4 contains the descriptive statistics for the 2017 dataset (N = 1,710) and the 2018 dataset (N = 1,183) for certificates 18 to 36 credit hours. Descriptive statistics indicated the overall mean age as M = 34.6 (SD = 8.9) with a range from 25 to 71 and average GPA as M = 2.41 (SD = 1.6) for the 2017 cohort. The 2018 cohort had similar statistics with a mean age of M = 34.7 (SD = 9.0) with a range from 25 to 72 and an average GPA of M = 2.48 (SD = 1.5). In both cohorts, age is substantially skewed right and GPA is moderately skewed left with higher concentrations in the lower tail and upper tail.

Table 4

Descriptive Statistics for Continuous Variables in Certificates 18 to 36 Credit Hours

		N	*M*	*Mdn*	*SD*	Min. value	Max. value	Skew	Kurtosis
2017	age	1,710	34.63	32	8.95	25	71	1.16	0.96
	gpa	1,710	2.41	3	1.56	0	4	-0.58	-1.25
2018	age	1,183	34.74	32	9.04	25	72	1.09	0.67
	gpa	1,183	2.48	3	1.49	0	4	-0.7	-1.03

Table 5 contains the demographics of beginning students enrolled in diplomas 37 to 48 credit hours in length and 49 to 59 credit hours in length in both cohorts of 2017 (representing the 2017-2018 academic year) and 2018 (representing the 2018-2019 academic year). The 2017 dataset for diplomas 37 to 48 credit hours totaled 635 records, contained a balanced proportion of females (52.0%) and males (48.0%) and the majority of students did receive the Pell grant (73.5%) compared to those who did not (26.5%). Black students accounted for 54.2% of the dataset followed by 37.5% of White students, 5.5% of Hispanic students, and 2.8% for all other races. As the case in previous cohorts, the majority of students did not self-identify as a single parent (89.5%) and 96.4% did not identify as a displaced homemaker. The majority of students in this cohort (56.2%) were not enrolled in a HOPE Career Grant major, while 23.2% were enrolled in Industrial Technology programs. The variables for high school diploma type and how long a student had been out of high school were both missing 3.15% of the values in the 2017 cohort. Of the 615 valid values, 49.9% were classified as having graduated with a college prep or tech prep high school diploma. Students who had been out of high school for five

years or more accounted for 82.3% of the valid values compared to those who had been out of high school four years or less (17.7%).

The 2018 dataset for diplomas 37 to 48 credit hours was similar to the 2017 cohort across all variables. For 670 records, there were essentially an equal proportion of males (50.9%) to females (49.1%) and the majority of students did the Pell grant (72.8%) compared to those who did not (27.2%). Black students accounted for 53.4% of the dataset followed by 35.5% of White students, 6.9% of Hispanic students, and 4.2% for all other races. The population of Hispanic students for this cohort was the largest for this race or ethnicity across all eight datasets. The majority of the dataset did not self-identify as a single parent (90.1%) and 96.3% did not identify as a displaced homemaker. Similar to the 2017 cohort, students not enrolled in a HOPE Career Grant major represented 48.1% of the cohort, while those enrolled in Industrial Technology programs accounted for 25.1%. The variables for high school diploma type and how long a student had been out of high school were both missing 3.73% of the values. Of the 645 valid values, 50.1% were classified as having graduated with a college prep or tech prep high school diploma. Students who had been out of high school for five years or more accounted for 83.4% of the valid values compared to those who had been out of high school four years or less (16.6%).

The 2017 and 2018 cohorts for diplomas 49 to 59 credit hours totaled 1,636 and 1,456 records, respectively, and shared similar demographics as the previous diploma cohorts with a few notable differences. Unlike the previous cohorts, both datasets for diplomas 49 to 59 credit hours contained slightly more females than males. The 2017 cohort was 58.4% female and 41.6% male, and the 2018 cohort was 62.0% female and

38.0% male. The majority of students (74.4% in the 2017 cohort) received the Pell grant (76.2% in the 2018 cohort). Black students accounted for the same percentage in both the 2017 and 2018 cohort (60.4%) followed by White students representing 31.5% in the 2017 cohort and 29.7% in the 2018 cohort. The majority of students did not self-identify as a single parent (87.5% and 89.2%) or self-identify as a displaced homemaker (96.0% and 97.2%). One notable difference from the previous diploma cohorts, yet similar to the 2018 cohort of certificates 18 to 36 credit hours, was most students were enrolled in Cyber, Engineering, or Healthcare programs. Other students were either not enrolled in any HOPE Career Grant programs or enrolled in Industrial Technology programs. The variables for high school diploma type and how long a student had been out of high school were both missing a small percentage of values. Of the 1,576 valid values of the 2017 cohort, 53.3% were classified as having graduated with a college prep or tech prep high school diploma. Of the 1,396 valid values in the 2018 cohort, the same was true with 54.7% graduating with a college prep or tech prep high school diploma. In both cohorts, the vast majority represented students who had been out of high school for five years or more compared to those who had been out of high school for four years or less.

Table 5

Demographics for Students Enrolled in Diploma Programs

	37 to 48 credit hours		49 to 59 credit hours	
	2017	2018	2017	2018
Race or Ethnicity				
White	238 (37.5%)	238 (35.5%)	515 (31.5%)	433 (29.7%)
Black	344 (54.2%)	358 (53.4%)	988 (60.4%)	879 (60.4%)
Hispanic	35 (5.5%)	46 (6.9%)	66 (4.0%)	68 (4.7%)
Other	18 (2.8%)	28 (4.2%)	67 (4.1%)	76 (5.2%)
Gender				
Male	305 (48.0%)	341 (50.9%)	680 (41.6%)	553 (38.0%)
Female	330 (52.0%)	329 (49.1%)	956 (58.4%)	903 (62.0%)
High School Diploma Type				
Certificate of Attendance	113 (18.4%)	115 (17.8%)	261 (16.6%)	229 (16.4%)
GED®	195 (31.7%)	207 (32.1%)	475 (30.1%)	403 (28.9%)
College Prep/Tech Prep	307 (49.9%)	323 (50.1%)	840 (53.3%)	764 (54.7%)
High School Graduation				
Four years or less	109 (17.7%)	107 (16.6%)	265 (16.8%)	237 (17.0%)
Five years or more	506 (82.3%)	538 (83.4%)	1,311 (83.2%)	1,159 (83.0%)
Single Parent				
No	568 (89.5%)	604 (90.1%)	1,431 (87.5%)	1,299 (89.2%)
Yes	67 (10.5%)	66 (9.8%)	205 (12.5%)	157 (10.8%)
Displaced Homemaker				
No	612 (96.4%)	645 (96.3%)	1,570 (96.0%)	1,415 (97.2%)
Yes	23 (3.6%)	25 (3.7%)	66 (4.0%)	41 (2.8%)
Pell Eligibility				
No	168 (26.5%)	182 (27.2%)	419 (25.6%)	346 (23.8%)
Yes	467 (73.5%)	488 (72.8%)	1,217 (74.4%)	1,110 (76.2%)
HOPE Program				
Not HOPE Career Grant	357 (56.2%)	322 (48.1%)	388 (23.7%)	350 (24.0%)
Cyber, Engineer, Healthcare	84 (13.2%)	132 (19.7%)	803 (49.1%)	748 (51.4%)
Industrial Technologies	147 (23.2%)	168 (25.1%)	356 (21.8%)	275 (18.9%)
Transportation and Logistics	47 (7.4%)	48 (7.2%)	89 (5.4%)	83 (5.7%)
Total	635	670	1,636	1,456

Table 6 and Table 7 include the descriptive statistics for both diploma cohorts. Table 6 contains the descriptive statistics for the 2017 dataset ($N = 635$) and the 2018 dataset ($N = 670$) for diplomas 37 to 48 credit hours. Descriptive statistics indicated the overall mean age as $M = 35.1$ ($SD = 9.7$) with a range from 25 to 85 and average GPA as $M = 2.33$ ($SD = 1.5$) for the 2017 cohort. The 2018 cohort had similar statistics with a mean age of $M = 34.7$ ($SD = 9.8$) with a range from 25 to 77 and an average GPA of $M = 2.46$ ($SD = 1.5$). In both cohorts, age is substantially skewed right and GPA is moderately skewed left with higher concentrations in the lower tail and upper tail.

Table 6

Descriptive Statistics for Continuous Variables in Diplomas 37 to 48 Credit Hours

		N	*M*	*Mdn*	*SD*	Min. value	Max. value	Skew	Kurtosis
2017	age	635	35.14	32	9.71	25	85	1.22	1.35
	gpa	635	2.33	3	1.5	0	4	-0.5	-1.26
2018	age	670	34.65	31	9.83	25	77	1.28	1.18
	gpa	670	2.46	3	1.45	0	4	-0.66	-1.00

Table 7 contains the descriptive statistics for the 2017 dataset ($N = 1{,}636$) and the 2018 dataset ($N = 1{,}456$) for diplomas 49 to 59 credit hours. Descriptive statistics indicated the overall mean age as $M = 34.6$ ($SD = 9.4$) with a range from 24 to 75 and average GPA as $M = 2.42$ ($SD = 1.5$) for the 2017 cohort. The 2018 cohort had similar statistics with a mean age of $M = 35.0$ ($SD = 9.5$) with a range from 25 to 71 and an average GPA of $M = 2.45$ ($SD = 1.5$). In both cohorts, age is substantially skewed right and GPA is moderately skewed left with higher concentrations in the lower tail and upper tail.

Table 7

Descriptive Statistics for Continuous Variables in Diplomas 49 to 59 Credit Hours

		N	*M*	*Mdn*	*SD*	Min. value	Max. value	Skew	Kurtosis
2017	age	1,636	34.64	32	9.37	24	75	1.27	1.28
	gpa	1,636	2.42	3	1.49	0	4	-0.6	-1.14
2018	age	1,456	34.97	32	9.53	25	71	1.13	0.76
	gpa	1,456	2.45	3	1.48	0	4	-0.64	-1.08

Data Preprocessing and Feature Engineering

Data analysis was conducted using the statistical software package R as shown in Appendix B. Data were examined for missing data, normality, outliers, multicollinearity, and linearity. Also, model-specific statistical assumptions were checked. A review of missing data including summaries of the data frames and analysis of the missing values indicated all variables except two (high school diploma type and high school graduation date) had zero missing values in each of the eight data files. Table 8 displays the percentages of missing data by cohort for the respective independent variables in this study. Given the percentage of missing values was less than 5%, data were imputed to address the missing values in six of the eight data files. Imputation via bagged trees and k-nearest neighbors produced similar distributions in each of the two variables. After imputation, chi-square tests were used for significance testing between the original data and the imputed data. In the data file representing certificates 9 to 17 credit hours, high school diploma type, and high school graduation date were missing 619 of 1,896 records (32.65%). Despite multiple attempts at imputation, a large disparity in the percentages between the complete dataset and the imputed dataset existed. The difference in high school diploma type from the original data to the imputed data was found to be

statistically significant, $\chi 2(2) = 36.68$, $p < 0.001$, as well as high school graduation date at $\chi 2(1) = 53.99$, $p < 0.001$. The 2018 cohort produced similar results where 690 of 1,954 records (35.31%) were missing. The difference in high school diploma type from the original data to the imputed data was found to be statistically significant, $\chi 2(2) = 25.82$, $p < 0.001$, as well as high school graduation date at $\chi 2(1) = 36.70$, $p < 0.001$. Therefore, the 619 identified records were removed from the dataset resulting in 1,277 records.

Table 8

Percentage of Missing Data by Cohort and Variable

	2017		2018	
Independent Variable	*N*	% of Students	*N*	% of Students
TCC's 9 to 17 Credit Hours				
High School Diploma Type	619	32.7	690	35.3
High School Graduation Date	619	32.7	690	35.3
TCC's 18 to 36 Credit Hours				
High School Diploma Type	66	3.9	58	4.9
High School Graduation Date	66	3.9	58	4.9
Diploma's 37 to 48 Credit Hours				
High School Diploma Type	20	3.2	25	3.7
High School Graduation Date	20	3.2	25	3.7
Diploma's 49 to 59 Credit Hours				
High School Diploma Type	60	3.7	60	4.1
High School Graduation Date	60	3.7	60	4.1

The independent variables in linear discriminant analysis are assumed to have a multivariate normal (Gaussian) distribution (James et al., 2013). Because of this assumption, LDA does not discriminate among categories using a mix of continuous and categorical variables. However, violating this assumption is normally acceptable as long as the sample size is large enough (James et al., 2013). Also, when the objective is only prediction or classification, these assumptions are less constraining. As previously stated,

a review of skewness and kurtosis values and histograms indicated the continuous variables age and GPA were both moderately to substantially skewed. Therefore, the assumption of normality was not met. Being neither of the continuous variables followed a normal distribution which violates an assumption for linear discriminant analysis, a Yeo-Johnson transformation was applied. Yeo-Johnson transformation is similar to the Box-Cox transformation but does not require the input variables to be strictly positive. Once the Yeo-Johnson transformation was applied, the assumption of normality was met.

Before transformation, each continuous variable was evaluated for outliers by inspecting boxplots and z-scores. A standard boxplot and adjusted boxplot for skewed distributions were used initially to identify potential outliers. The standard boxplot identified four outliers for age (72, 74, 75, and 77) and one outlier for GPA (0). However, the adjusted box plot for skewed distributions did not identify outliers in either the age or GPA variable. Both variables were converted to z-scores to mathematically assess for outliers. The z-scores were analyzed for outliers using a cutoff z-score of 4.0 and -4.0. There were no outliers for GPA and one outlier for age (77). It was determined this data point would remain in the data set.

Many, but not all, underlying model calculations require predictor values to be encoded as numbers. All predictors except age and GPA were converted from nominal data (e.g., factors) into one or more numeric binary variables representing specific factor level values. The default approach was to create dummy variables using the "reference cell" parameterization. This means, if there are C levels of the factor, there will be C - 1 dummy variables created and all but the first factor level will be made into new columns. Dummy variables are described in Table 9.

Table 9

Categories and Codes for Dummy Variables

Variable	Dummy Variables
Race or Ethnicity (racecode)	racecode_X1 = White racecode_X2 = Black racecode_X3 = Hispanic racecode_X4 = Other
Gender (gencode)	gencode_X0 = Male gencode_X1 = Female
High School Diploma Type (hsdipcode)	hsdipcode_X1 = Certificate of Attendance hsdipcode_X2 = GED® hsdipcode_X3 = College Prep/Tech Prep
High School Graduation Date (gradcode)	gradcode_X0 = Four years or less gradcode_X1 = Five years or more
Single Parent Status (sparcode)	sparcode_X0 = No sparcode_X1 = Yes
Displaced Homemaker Status (dhomcode)	dhomcode_X0 = No dhomcode_X1 = Yes
Pell Eligibility (pellcode)	pellcode_X0 = No pellcode_X1 = Yes
HOPE Career Grant Program of Study (hopepos)	hopepos_X1 = Not HOPE Career Grant Program hopepos_X2 = Cyber, Engineer, Healthcare hopepos_X3 = Industrial Technologies hopepos_X4 = Transportation and Logistics

All numeric variables were centered and scaled. The most straightforward and common data transformation is to center scale the predictor variables. To center a predictor variable, the average predictor value is subtracted from all the values. As a result of centering, the predictor has a zero mean. Similarly, to scale the data, each value of the predictor variable is divided by its standard deviation. Scaling the data coerce the

values to have a common standard deviation of one. These manipulations are generally used to improve the numerical stability of some calculations.

Logistic regression and linear discriminant analysis require little or no multicollinearity among the independent variables (James et al., 2013). Multicollinearity was assessed in two ways: by examining the variance inflation factor (VIF) computed for each predictor and generating correlation coefficients between variables. The absence of multicollinearity has a VIF value of one. Typically, a VIF value which exceeds five to 10 indicates a problematic amount of collinearity, and the troubled variable should be removed. In each cohort of the eight data files variance inflation factors ranged from 1.01 to 2.16 indicating no issues with multicollinearity. Also, a correlation matrix was generated for each cohort of the eight data files to measure the strength of the correlations among variables. In Tables 8 through 15, Pearson correlations were identified for the continuous variables, polychoric correlations for polytomous variables, and tetrachoric correlations for the dichotomous variables. Each correlation was assessed for its significance as well as its strength.

For the 2017 cohort of certificates 9 to 17 credit hours in length, there was a strong correlation between the HOPE program of study and the graduation code, $r(1,277) = .86$, $p < .001$. There was also a strong correlation between Pell eligibility and displaced homemaker, $r(1,277) = .84$, $p < .01$. For the 2018 cohort of certificates 9 to 17 credit hours in length, there was a strong correlation between the HOPE program of study and the graduation code, $r(1,264) = .81$, $p < .001$. The correlation coefficient values for the 2017 and 2018 cohorts of certificates 9 to 17 credit hours in length, shown in Table 10 and Table 11, show no correlation values greater than, or equal to .90 indicating no items

exhibited extreme collinearity, and, therefore, all items could be included in the model (Kline, 2011). The assumption of little or no multicollinearity was met.

Table 10

Correlations among Variables in Certificates 9 to 17 Credit Hours (2017 Cohort)

	1	2	3	4	5
1 age	-				
2 gpa	-0.01	-			
3 racecode	-0.08*	-0.07*	-		
4 hsdipcode	0.19***	0.09**	0.19***	-	
5 hopepos	0.05	0.06	0.31***	0.41***	-
6 gencode	-0.09**	0.06	-0.35***	-0.10*	-0.48***
7 gradcode	0.06	0.08*	0.27***	0.46***	0.86***
8 sparcode	-0.09	0.10	-0.05	-0.08	-0.24***
9 dhomcode	-0.01	-0.12**	-0.33***	-0.33***	-0.41***
10 pellcode	-0.21**	-0.12*	0.43***	0.01	0.29**
11 retained	0.00	0.13**	0.01	0.03	0.12*

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

Table 10 (continued)

Correlations among Variables in Certificates 9 to 17 Credit Hours (2017 Cohort)

	6	7	8	9	10	11
6 gencode	-					
7 gradcode	-0.34***	-				
8 sparcode	0.53***	-0.18**	-			
9 dhomcode	0.31***	-0.44***	0.56***	-		
10 pellcode	0.01	0.31**	-0.03	-0.84**	-	
11 retained	-0.12*	-0.07	-0.16*	-0.12*	0.01	-

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

Table 11

Correlations among Variables in Certificates 9 to 17 Credit Hours (2018 Cohort)

	1	2	3	4	5
1 age	-				
2 gpa	-0.02	-			
3 racecode	-0.12***	-0.08*	-		
4 hsdipcode	0.19***	0.06	0.06*	-	
5 hopepos	0.03	0.03	0.30***	0.22***	-
6 gencode	-0.14***	0.03	-0.37***	-0.04	-0.46***
7 gradcode	0.09**	0.09**	0.14***	0.37***	0.81***
8 sparcode	-0.12*	0.00	-0.10	-0.21**	-0.23***
9 dhomcode	-0.02	0.08	-0.41***	-0.18***	-0.51***
10 pellcode	-0.27***	-0.12*	0.25**	0.15	-0.07
11 retained	-0.07	0.26***	-0.03	0.05	-0.06

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

Table 11 (continued)

Correlations among Variables in Certificates 9 to 17 Credit Hours (2018 Cohort)

	6	7	8	9	10	11
6 gencode	-					
7 gradcode	-0.26***	-				
8 sparcode	0.61***	-0.16*	-			
9 dhomcode	0.53***	-0.39***	0.73***	-		
10 pellcode	0.29**	0.38***	0.15	-0.31	-	
11 retained	-0.05	-0.21***	-0.03	0.00	-0.39***	-

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

For the 2017 cohort of certificates 18 to 36 credit hours in length, there was a moderate correlation between gender and Pell eligibility, $r(1,710) = .60$, $p < .001$. The same variables, gender and Pell eligibility, had a slightly higher moderate correlation in the 2018 cohort, $r(1,183) = .72$, $p < .001$. The correlation coefficient values for the 2017 and 2018 cohorts of certificates 18 to 36 credit hours in length, shown in Table 12 and

Table 13, show no correlation values greater than, or equal to .90 indicating no items exhibited extreme collinearity, and, therefore, all items could be included in the model (Kline, 2011). The assumption of little or no multicollinearity was met.

Table 12

Correlations among Variables in Certificates 18 to 36 Credit Hours (2017 Cohort)

	1	2	3	4	5
1 age	-				
2 gpa	0.12***	-			
3 racecode	-0.03	-0.12***	-		
4 hsdipcode	0.28***	-0.01	0.02	-	
5 hopepos	0.11***	0.11***	-0.08**	-0.09**	-
6 gencode	-0.05	-0.11***	0.05	0.09*	-0.34***
7 gradcode	0.07*	-0.09**	-0.08*	0.22***	-0.23***
8 sparcode	-0.07	-0.09*	0.03	-0.07	-0.14***
9 dhomcode	0.19***	0.10	-0.14*	-0.10	0.37***
10 pellcode	-0.23***	-0.18***	0.07	-0.03	-0.30***
11 retained	0.07*	0.32***	-0.02	0.01	0.25***

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

Table 12 (continued)

Correlations among Variables in Certificates 18 to 36 Credit Hours (2017 Cohort)

	6	7	8	9	10	11
6 gencode	-					
7 gradcode	0.07	-				
8 sparcode	0.54***	0.09	-			
9 dhomcode	-0.25***	-0.22***	0.33***	-		
10 pellcode	0.60***	0.09*	0.43***	-0.35***	-	
11 retained	-0.20***	-0.15***	-0.11*	0.26***	-0.11**	-

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

Table 13

Correlations among Variables in Certificates 18 to 36 Credit Hours (2018 Cohort)

	1	2	3	4	5
1 age	-				
2 gpa	0.06*	-			
3 racecode	-0.07*	-0.03	-		
4 hsdipcode	0.23***	-0.03	0.06	-	
5 hopepos	0.04	-0.03	-0.08**	-0.11**	-
6 gencode	-0.11**	0.02	0.11*	0.10	-0.49***
7 gradcode	0.17***	0.06	-0.11*	0.21***	-0.09**
8 sparcode	-0.05	-0.04	-0.06	-0.09	-0.10**
9 dhomcode	-0.01	-0.15*	-0.13*	0.02	0.50***
10 pellcode	-0.17***	-0.04	0.04	-0.01	-0.18***
11 retained	-0.04	0.40***	-0.01	-0.01	-0.01

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

Table 13 (continued)

Correlations among Variables in Certificates 18 to 36 Credit Hours (2018 Cohort)

	6	7	8	9	10	11
6 gencode	-					
7 gradcode	0.10	-				
8 sparcode	0.45***	0.00	-			
9 dhomcode	-0.31***	-0.40***	0.25**	-		
10 pellcode	0.72***	0.04	0.37***	-0.32***	-	
11 retained	-0.03	-0.18***	0.04	0.16*	0.02	-

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

For the 2017 cohort of diplomas 37 to 48 credit hours in length, there was a moderate correlation between gender and HOPE program of study, $r(635) = .64$, $p < .001$. The same variables, gender and HOPE program of study, had a similar, moderate correlation in the 2018 cohort, $r(670) = .65$, $p < .001$. The correlation coefficient values for the 2017 and 2018 cohorts of diplomas 37 to 48 credit hours in length, shown in Table

14 and Table 15, show no correlation values greater than, or equal to .90 indicating no items exhibited extreme collinearity, and, therefore, all items could be included in the model (Kline, 2011). The assumption of little or no multicollinearity was met.

Table 14

Correlations among Variables in Diplomas 37 to 48 Credit Hours (2017 Cohort)

	1	2	3	4	5
1 age	-				
2 gpa	0.09*	-			
3 racecode	-0.01	-0.18***	-		
4 hsdipcode	0.29***	0.03	-0.06	-	
5 hopepos	-0.13**	0.23***	-0.13*	-0.08	-
6 gencode	0.12*	-0.11*	0.17**	0.16**	-0.64***
7 gradcode	0.15**	0.13*	-0.06	0.05	0.00
8 sparcode	0.00	-0.05	0.05	0.01	-0.09
9 dhomcode	0.08	-0.05	-0.01	-0.10*	-0.18
10 pellcode	-0.08	-0.17**	0.12*	-0.05	-0.23***
11 retained	0.06	0.38***	-0.04	-0.02	0.08

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

Table 14 (continued)

Correlations among Variables in Diplomas 37 to 48 Credit (2017 Cohort)

	6	7	8	9	10	11
6 gencode	-					
7 gradcode	-0.10*	-				
8 sparcode	0.46***	-0.18*	-			
9 dhomcode	0.32*	-0.19*	0.55***	-		
10 pellcode	0.35***	-0.19*	0.51***	0.43*	-	
11 retained	-0.05	0.09	0.10*	0.08	0.08	-

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

Table 15

Correlations among Variables in Diplomas 37 to 48 Credit (2018 Cohort)

	1	2	3	4	5
1 age	-				
2 gpa	0.08*	-			
3 racecode	-0.06	-0.06	-		
4 hsdipcode	0.25***	-0.05	-0.05	-	
5 hopepos	-0.12**	0.16***	-0.07	-0.05	-
6 gencode	0.09	-0.10*	0.16*	0.05	-0.65***
7 gradcode	0.18**	0.08	-0.14*	0.15	-0.05
8 sparcode	-0.05	0.04	-0.01	-0.01	-0.19**
9 dhomcode	0.04	-0.13*	0.03	-0.07	-0.10*
10 pellcode	-0.10*	-0.20***	0.16*	-0.14*	-0.25***
11 retained	0.09	0.33***	-0.07	0.05	-0.03

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

Table 15 (continued)

Correlations among Variables in Diplomas 37 to 48 Credit (2018 Cohort)

	6	7	8	9	10	11
6 gencode	-					
7 gradcode	0.03	-				
8 sparcode	0.34***	-0.09	-			
9 dhomcode	0.17*	-0.06	0.57***	-		
10 pellcode	0.38***	-0.23**	0.33**	0.11*	-	
11 retained	0.12*	-0.01	-0.01	-0.14*	0.00	-

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

For the 2017 cohort of diplomas 49 to 59 credit hours in length, there was a moderate correlation between gender and HOPE program of study, $r(1,636) = .66$, $p < .001$. The same variables, gender and HOPE program of study, had a moderate correlation in the 2018 cohort, $r(1,456) = .63$, $p < .001$. The correlation coefficient values for the 2017 and 2018 cohorts of diplomas 49 to 59 credit hours in length, shown in Table

16 and Table 17, show no correlation values greater than, or equal to .90 indicating no items exhibited extreme collinearity, and, therefore, all items could be included in the model (Kline, 2011). The assumption of little or no multicollinearity was met.

Table 16

Correlations among Variables in Diplomas 49 to 59 Credit Hours (2017 Cohort)

	1	2	3	4	5
1 age	-				
2 gpa	0.10***	-			
3 racecode	-0.01	-0.08**	-		
4 hsdipcode	0.24***	0.04	0.12**	-	
5 hopepos	-0.05	0.10***	-0.10*	-0.07*	-
6 gencode	-0.08*	-0.10**	0.12**	-0.02	-0.66***
7 gradcode	0.18***	-0.01	-0.06	0.19***	0.07
8 sparcode	-0.09*	-0.06	0.00	0.00	-0.16***
9 dhomcode	0.10*	-0.01	0.01	-0.02	-0.09*
10 pellcode	-0.21***	-0.15***	0.11*	-0.10*	-0.22***
11 retained	0.01	0.28***	0.07*	0.01	0.00

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

Table 16 (continued)

Correlations among Variables in Diplomas 49 to 59 Credit Hours (2017 Cohort)

	6	7	8	9	10	11
6 gencode	-					
7 gradcode	-0.11*	-				
8 sparcode	0.41***	-0.04	-			
9 dhomcode	0.16*	0.06	0.47***	-		
10 pellcode	0.43***	-0.15**	0.31***	0.23*	-	
11 retained	0.06	-0.07	-0.04	-0.03	0.14**	-

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

Table 17

Correlations among Variables in Diplomas 49 to 59 Credit Hours (2018 Cohort)

	1	2	3	4	5
1 age	-				
2 gpa	0.10***	-			
3 racecode	-0.02	-0.07*	-		
4 hsdipcode	0.26***	0.04	0.05	-	
5 hopepos	-0.07*	0.05	0.00	0.03	-
6 gencode	0.04	-0.05	0.03	-0.01	-0.63***
7 gradcode	0.16***	0.07	-0.09*	0.21***	0.04
8 sparcode	-0.09*	-0.03	-0.05	-0.04	-0.14**
9 dhomcode	0.11*	0.03	0.01	-0.02	-0.27***
10 pellcode	-0.19***	-0.12***	0.07	-0.08	-0.22***
11 retained	0.05	0.41***	-0.02	0.04	0.00

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

Table 17 (continued)

Correlations among Variables in Diplomas 49 to 59 Credit Hours (2018 Cohort)

	6	7	8	9	10	11
6 gencode	-					
7 gradcode	-0.10*	-				
8 sparcode	0.36***	0.00	-			
9 dhomcode	0.04	-0.04	0.40***	-		
10 pellcode	0.39***	-0.11*	0.30***	0.14*	-	
11 retained	0.03	-0.09*	0.05	-0.05	0.10*	-

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

Also, logistic regression assumes linearity of continuous predictors and the log odds (James et al., 2013). Linearity in the logit was assessed by constructing component-plus-residual plots of the residuals of each continuous predictor against the dependent variable. The assumption of linearity of the independent variables age and GPA and the

log odds was not met. Once the Yeo-Johnson transformation was applied, the assumption of linearity was met.

In general, binary logistic regression requires a large sample, a binomial distribution, and observations independent of each other. Based on Peduzzi, Concato, Kemper, Holford, and Feinstein's research (1996), the cases in this study exceeded the guidelines for minimum sample size. With p as the smallest of the proportions of cases in the population and k the number of covariates (the number of independent variables), the minimum number of cases to include is $N = 10\ k / p$. The assumption of having a response variable which follows a binomial distribution was met because the dependent variable, retention, was a dichotomous variable coded as 0 for not retained and 1 for retained. Thus, the dichotomous variable retention had mutually exclusive and exhaustive categories. The assumption of observations being independent of each other was met as data from the 2017-2018 and 2018-2019 datasets represented 14,448 unique student IDs.

The population variances and covariances for all independent variables are required to be equal across the dependent variable groups in linear discriminant analysis (Spicer, 2005). Stated differently, the values of each variable vary around the mean by the same amount on average (Spicer, 2005). This is known as the homogeneity of variance-covariance matrices assumption (Spicer, 2005). A Levene's Test found the assumption of homogeneity of variance was met for the age variable ($p = .84$), but not for the variable GPA where $p < 0.001$.

The classification and regression tree (CART) methodology requires no distributional assumptions for predictor variables and is resistant to outliers, multicollinearity, and heteroscedasticity (Breiman, Friedman, Olshen, & Stone, 1984).

Like classification trees, there are no formal distributional assumptions with random forests (Breiman et al., 1984). Random forests are nonparametric and can tolerate skewed data as well as categorical data which are ordinal or nonordinal (Breiman et al., 1984). Assumptions for support vector machines are the margin should be as large as possible and the support vectors are the most useful because they are the data points most likely to be incorrectly classified (Kuhn & Johnson, 2013).

Model Training and Significant Predictors

Research Question 1 was subdivided into four sections to identify which environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) were significant predictors of nontraditional student retention for certificate or diploma programs. Five statistical analyses were utilized to answer each of the four subsections representing certificates 9–17 credit hours in length, certificates 18–36 credit hours in length, diplomas 37–48 credit hours in length, and diplomas 49–59 credit hours in length. The training data set was used to build the model and the test data set was used to estimate the model's predictive performance (Kuhn & Johnson, 2013). The 2017-2018 data were used as the training data and the 2018-2019 data were used as the test data. During model training, upsampling was used to mitigate the effects of class imbalance in the outcome variable retained. An imbalance exists across each data file in the variable retained as one class has very low proportions in the training data as compared to the other class. Upsampling simulates or imputes additional data points to improve balance across classes. In both the 2017 and 2018 cohorts of certificates 9–17 credit hours in

length, the class imbalance was the greatest with the rate of students not retained only accounting for 17.93% and 17.72% respectively. Each model was evaluated using 10-fold cross-validation repeated five times with stratification.

Research Question 1A. Are environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) significant predictors of nontraditional student retention for certificates 9–17 credit hours in length?

Two data modeling approaches (logistic regression and linear discriminant analysis) and three data mining approaches (classification tree, random forest, and support vector machine models) were used to answer this research question. Multivariate logistic regression analysis was the first of two data modeling approaches performed to answer this research question. Logistic regression analysis is used to predict a discrete outcome from various types of predictor variables. In instances where the dependent variable is dichotomous and the independent variables are categorical or a mix of continuous and categorical, logistic regression is appropriate (Burns & Burns, 2008). The Hosmer-Lemeshow goodness of fit test examined whether the observed proportion of students retained is similar to or differs from the expected frequencies of retained students using a Pearson chi-square statistic ($\chi 2(14) = 71.59$, $p < 0.001$). Small values with large p-values indicate a good fit to the data while large values with p-values below 0.05 indicate a poor fit. Also, the McFadden pseudo R^2 value was calculated as 0.11 indicating the model can account for 11% of the retained variable.

Table 18 shows the results from the logistic regression analysis. The overall model was found to be statistically significant, $\chi 2(15) = 334.09$, $p < 0.001$, and the model resulted in a training error rate of 0.34. Of the 15 predictor variables, 11 were statistically significant as predictors of student retention: age, GPA, race (Black), race (Hispanic), gender (female), graduation date (out of high school at least five years or more), single parents, Pell eligibility, Cyber, Engineer, or Healthcare programs, Industrial Technology programs, and Transportation and Logistics programs. By enrolling in a Transportation and Logistics program of study, the odds of a student being retained increases by a factor of 1.95 (odds ratio = 1.95), given all other variables are unchanged. If a student enrolls in Industrial Technology programs, the odds of those students being retained decreases by 24.7% (odds ratio = 0.753 – 1), keeping other variables constant. The predictor variables which were not statistically significant included race (other), high school diploma (GED®), high school diploma (college prep or tech prep), and displaced homemakers.

To determine the most influential predictors of retention, variable importance was measured using the odds ratio. The variables of Transportation and Logistics programs (OR = 1.951, 95% CI = 1.669 to 2.290) and GPA (OR = 1.382, 95% CI = 1.261 to 1.517) were the strongest predictors of being retained. The weakest predictors of being retained were graduation date (out of high school at least five years or more) (OR = 0.759, 95% CI = 0.657 to 0.876) and Industrial Technology programs (OR = 0.753, 95% CI = 0.673 to 0.842).

Table 18

Variables Used to Predict Retention Utilizing Logistic Regression (Training Data)

Predictor	Log Odds	*SE*	*Z*	Pr(>|z|)		*OR*	*95% Confidence Interval* *Lower*	*Upper*
(Intercept)	0.231	0.050	4.606	$p < .001$	***	1.259	1.142	1.390
age	0.101	0.050	2.027	0.043	*	1.106	1.004	1.219
gpa	0.324	0.047	6.863	$p < .001$	***	1.382	1.261	1.517
racecode_X2	-0.118	0.054	-2.161	0.031	*	0.889	0.799	0.989
racecode_X3	-0.098	0.047	-2.092	0.036	*	0.906	0.824	0.992
racecode_X4	0.027	0.050	0.538	0.590		1.027	0.931	1.133
gencode_X1	-0.140	0.058	-2.425	0.015	*	0.869	0.776	0.974
hsdipcode_X2	0.213	0.124	1.717	0.086		1.237	0.971	1.579
hsdipcode_X3	0.161	0.113	1.419	0.156		1.174	0.941	1.468
gradcode_X1	-0.275	0.073	-3.769	$p < .001$	***	0.759	0.657	0.876
sparcode_X1	-0.135	0.048	-2.783	0.005	**	0.874	0.794	0.960
dhomcode_X1	-0.060	0.054	-1.117	0.264		0.942	0.847	1.046
pellcode_X1	0.118	0.053	2.228	0.026	*	1.125	1.014	1.249
hopepos_X2	-0.129	0.057	-2.253	0.024	*	0.879	0.785	0.983
hopepos_X3	-0.284	0.057	-4.945	$p < .001$	***	0.753	0.673	0.842
hopepos_X4	0.668	0.081	8.288	$p < .001$	***	1.951	1.669	2.290

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

Additionally, the finalized logistic regression model was applied to the test data for comparison. Table 19 shows the results from the logistic regression analysis. The overall model was found to be statistically significant, $\chi 2(15) = 530.63$, $p < 0.001$, and the model resulted in an error rate of 0.34. Also, the McFadden pseudo R^2 value was calculated as 0.18 indicating the model can account for 18% of the retained variable. Of the 15 predictor variables, nine were statistically significant as predictors of student retention: age, GPA, race (Hispanic), gender (female), graduation date (out of high school at least five years or more), Pell eligibility, Cyber, Engineer, or Healthcare programs, Industrial Technology programs, and Transportation and Logistics programs. Given all

other variables are unchanged, the odds of a student being retained increases by a factor of 1.59 (odds ratio = 1.59) as GPA increases. Likewise, enrollment in Transportation and Logistics programs increases the odds of a student being retained by a factor of 1.50 (odds ratio = 1.50) given all other variables remain unchanged. If a student enrolls in Industrial Technology programs, the odds of those students being retained decreases by 49% (odds ratio = 0.510 – 1), keeping other variables constant. The predictor variables which were not statistically significant included race (Black), race (other), high school diploma (GED®), high school diploma (college prep or tech prep), single parents, and displaced homemakers.

Table 19

Variables Used to Predict Retention Utilizing Logistic Regression (Test Data)

Predictor	Log Odds	*SE*	*Z*	Pr(>|z|)		*OR*	*95% Confidence Interval*	
							Lower	*Upper*
(Intercept)	0.353	0.053	6.597	$p < .001$	***	1.423	1.282	1.581
Age	-0.138	0.054	-2.578	0.010	**	0.871	0.784	0.967
Gpa	0.463	0.049	9.373	$p < .001$	***	1.589	1.444	1.753
racecode_X2	0.014	0.056	0.241	0.809		1.014	0.908	1.131
racecode_X3	-0.169	0.047	-3.607	$p < .001$	***	0.844	0.768	0.924
racecode_X4	-0.057	0.055	-1.041	0.298		0.945	0.848	1.052
gencode_X1	-0.289	0.062	-4.651	$p < .001$	***	0.749	0.663	0.846
hsdipcode_X2	-0.096	0.128	-0.748	0.455		0.909	0.708	1.169
hsdipcode_X3	0.064	0.118	0.540	0.589		1.066	0.846	1.346
gradcode_X1	-0.265	0.070	-3.789	$p < .001$	***	0.767	0.669	0.880
sparcode_X1	-0.036	0.053	-0.689	0.491		0.964	0.869	1.070
dhomcode_X1	0.054	0.054	1.000	0.318		1.056	0.949	1.175
pellcode_X1	-0.154	0.047	-3.256	$p < .001$	***	0.857	0.777	0.937
hopepos_X2	-0.271	0.055	-4.937	$p < .001$	***	0.763	0.684	0.849
hopepos_X3	-0.673	0.059	-11.397	$p < .001$	***	0.510	0.454	0.572
hopepos_X4	0.403	0.081	4.964	$p < .001$	***	1.496	1.279	1.758

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

The variables of GPA (OR = 1.589, 95% CI = 1.444 to 1.753) and Transportation and Logistics programs (OR = 1.496, 95% CI = 1.279 to 1.758) were the strongest predictors of being retained in the test data. With every one point increase in GPA, the odds of being retained increases 59% (odds ratio = 1.59), and compared to students enrolled in other HOPE Career Grant programs, students enrolled in a program related to Transportation and Logistics were more likely to be retained. The weakest predictors of being retained were Cyber, Engineer, or Healthcare programs (OR = 0.763, 95% CI = 0.684 to 0.849), females (OR = 0.749, 95% CI = 0.663 to 0.846), and Industrial Technology programs (OR = 0.510, 95% CI = 0.454 to 0.572). The variable importance plot including all 15 predictor variables is shown in Figure 3. Because variable importance for logistic regression is based on the absolute values of the z-statistic, both the most influential and the least influential predictors may be displayed at the top of the plot.

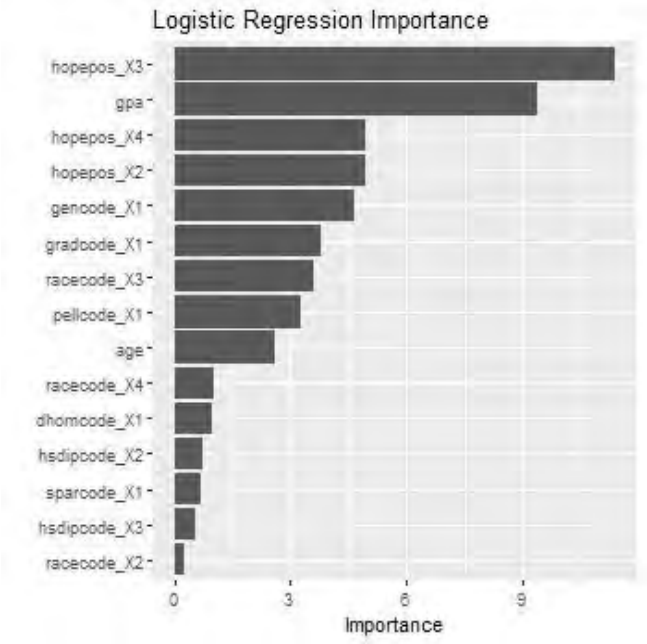

Figure 3. Logistic regression variable importance plot for certificates 9 to 17 credit hours in length.

The second data modeling approach used to address Research Question 1A was linear discriminant analysis. Similar to logistic regression, linear discriminant analysis is a common multivariate statistical method used to analyze categorical outcome variables (James et al., 2013). Linear discriminant analysis focuses on determining which variable discriminates between two or more classes and is used to develop a classification model for predicting the group membership of new observations (Spicer, 2005). It does this by maximizing the distance between the means of each class and minimizing the variation (scatter) within each class (Kuhn & Johnson, 2013).

The LDA model fit resulted in a training error rate of 0.34 and the results were similar to those of the logistic regression. As shown in Table 20, coefficients with the strongest associated weights included GPA (0.389), high school diploma (GED®) (0.210), graduation date (out of high school at least five years or more) (-0.339), Industrial Technology programs (-0.368), and Transportation and Logistics programs (0.780). The larger the coefficient of a predictor in the standardized discriminant function, the more important its role in the discriminant function. Transportation and Logistics programs were the strongest predictors of being retained or not with a coefficient of 0.780 and GPA was the second most influential with a coefficient of 0.389. Both race (other) (0.025) and displaced homemakers (-0.070) had the least influential coefficients of being retained or not. Although the variable age had one of the lowest weighted coefficients, it also had the largest mean difference within the group means. The group mean is the average of each predictor within each class. The variable age might have a slightly greater influence (negative) on students not being retained (-8.925) than on students being retained (6.832).

Table 20

Variables Used to Predict Retention Utilizing LDA (Training Data)

Independent Variable	Not Retained Mean	Retained Mean	Coefficients of Linear Discriminants: LD1
age	-8.925	6.832	0.115
gpa	-0.263	0.038	0.389
racecode_X2	-0.078	0.021	-0.127
racecode_X3	0.167	-0.029	-0.102
racecode_X4	-0.005	0.002	0.025
gencode_X1	0.110	-0.030	-0.174
hsdipcode_X2	-0.033	-0.004	0.210
hsdipcode_X3	-0.020	0.007	0.156
gradcode_X1	0.127	-0.019	-0.339
sparcode_X1	0.123	-0.030	-0.159
dhomcode_X1	0.074	-0.031	-0.070
pellcode_X1	-0.018	0.001	0.130
hopepos_X2	0.263	-0.058	-0.167
hopepos_X3	0.367	-0.083	-0.368
hopepos_X4	-0.371	0.090	0.780

Note. Prior probabilities of groups: not retained: 0.5, retained: 0.5.

In comparison, Table 21 includes the coefficients of linear discriminants for the test data. The model resulted in an error rate of 0.34. The coefficients with the strongest associated weights included GPA (0.420), females (-0.265), Cyber, Engineer, or Healthcare programs (-0.289), Industrial Technology programs (-0.666), and Transportation and Logistics programs (0.368). The strongest predictor of being retained or not based on test data was Industrial Technology programs with a negative coefficient of -0.666, GPA with a coefficient of 0.420, and Transportation and Logistics programs with a coefficient of 0.368. The variables race (other) (-0.042), single parents (-0.042), and race (Black) (0.022) had the least influential coefficients of being retained or not. The variable Industrial Technology programs had the largest variance within the group means.

This variable has a greater influence on students not being retained (0.605) than on students being retained (-0.126). The variable importance plot including all 15 predictor variables is shown in Figure 4.

Table 21

Variables Used to Predict Retention Utilizing LDA (Test Data)

Independent Variable	Not Retained Mean	Retained Mean	Coefficients of Linear Discriminants: LD1
age	0.142	-0.021	-0.119
gpa	-0.295	0.067	0.420
racecode_X2	-0.122	0.018	0.022
racecode_X3	0.194	-0.034	-0.155
racecode_X4	0.004	-0.009	-0.042
gencode_X1	0.056	-0.012	-0.265
hsdipcode_X2	0.036	-0.008	-0.085
hsdipcode_X3	-0.053	0.012	0.051
gradcode_X1	0.227	-0.052	-0.259
sparcode_X1	0.016	-0.005	-0.042
dhomcode_X1	0.015	-0.001	0.046
pellcode_X1	0.294	-0.059	-0.126
hopepos_X2	0.279	-0.053	-0.289
hopepos_X3	0.605	-0.126	-0.666
hopepos_X4	-0.357	0.073	0.368

Note. Prior probabilities of groups: not retained: 0.5, retained: 0.5.

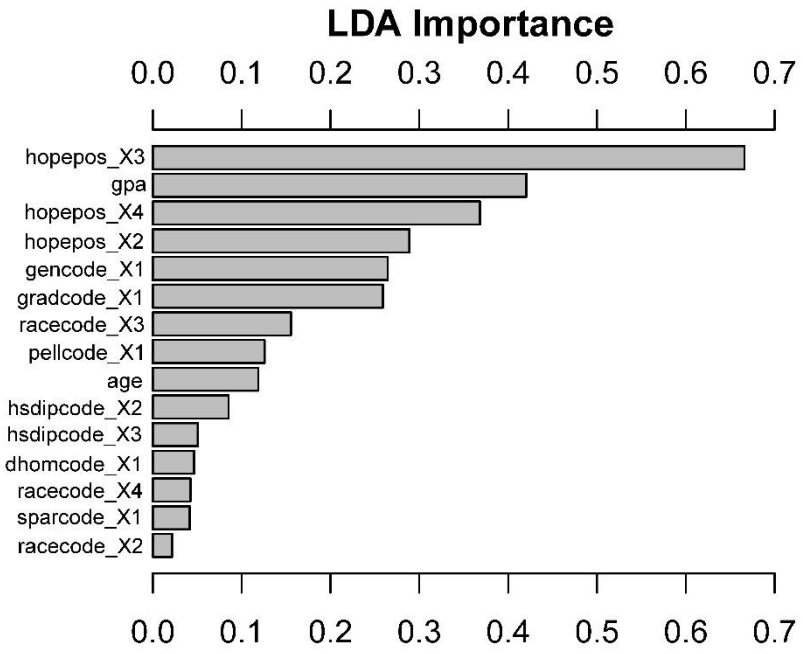

Figure 4. Linear discriminant analysis variable importance plot for certificates 9 to 17 credit hours in length.

A classification tree, the first of three data mining approaches performed to answer Research Question 1A, will be used to predict that each observation belongs to the most commonly occurring class of training data observations in the node or region to which it belongs (James et al., 2013). A decision tree creates separations between groups and subgroups, partitioning the data into smaller, homogeneous groups. The Gini Impurity Index was used throughout the partitioning process to select the best split among the values of the predictor which results in the lowest impurity measure. As a measure of node purity, a smaller value indicates a node contains observations primarily from a single class.

All predictor variables were allowed to enter the model. The prior probabilities were specified as 0.50 for predicted class 1, students who were retained, and 0.50 for predicted class 0, students who were not retained. The classification tree model was

evaluated using 10-fold cross-validation repeated five times using stratification and tuned for the parameters complexity and tree depth. The best parameters, based on the largest AUC metric (area under the ROC curve), were used to select the optimal model. The optimal complexity factor (0.000474) and maximum tree depth (7) were used to update the model and refit the training data. The 10-fold cross-validation was used to obtain a cross-validated error rate where the lowest rate indicated the tree that best fit the data. The resulting model had 25 total splits and a cross-validated error rate of 0.54. The overall training error rate for the model was 0.29.

To determine the strongest predictors of retention, variable importance was measured as the sum of the goodness of split measures (Gini index). Because a variable may appear in the tree many times, either as a primary or a surrogate variable, the overall measure of variable importance is the sum of the goodness of split measures for each split for which it was the primary variable, plus the adjusted agreement for all splits in which it was a surrogate variable:

$$I_G(\theta) = \sum_T \sum_\tau \Delta i_\theta(\tau, T)$$

The variables of GPA ($I_G = 135.99$), age ($I_G = 83.19$), and Transportation and Logistics programs ($I_G = 69.79$) were the strongest predictors of being retained or not. The weakest predictors of being retained or not were single parents ($I_G = 10.28$), race (Hispanic) ($I_G = 5.80$), race (other) ($I_G = 3.33$), race (Black) ($I_G = 1.81$), and Pell eligibility ($I_G = 0.24$).

In comparison, the model applied to test data resulted in 26 total splits, a cross-validated error rate of 0.43, and an overall error rate of 0.38. The variables of GPA ($I_G = 149.53$), Industrial Technology programs ($I_G = 102.86$), females ($I_G = 70.21$), and age ($I_G = 69.76$) were the strongest predictors of being retained or not. The weakest predictors of

being retained or not were single parents ($I_G = 11.54$), Pell eligibility ($I_G = 10.52$), race (Hispanic) ($I_G = 9.29$), and race (other) ($I_G = 2.18$). The variable importance plot including all 15 predictor variables is shown in Figure 5.

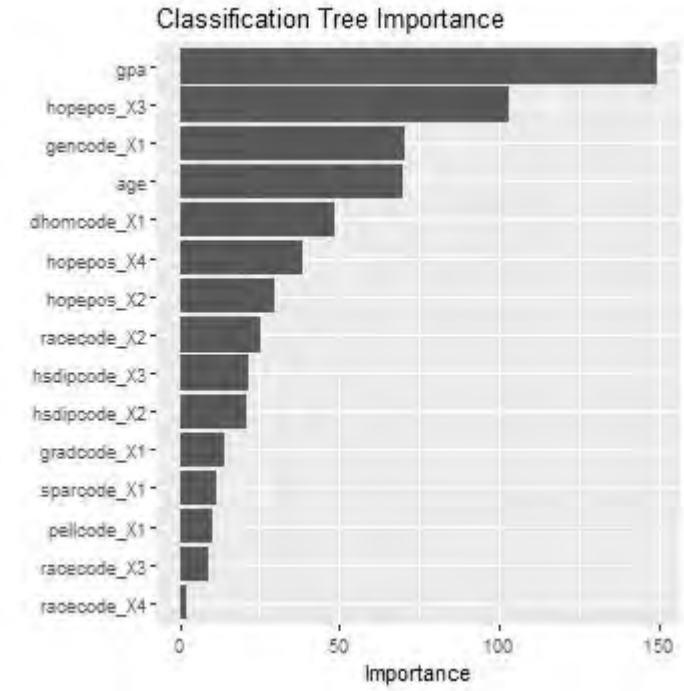

Figure 5. Classification tree variable importance plot for certificates 9 to 17 credit hours in length.

The random forest model was the second of three data mining approaches performed to answer Research Question 1A. Random forests refer to a model of the entire system of random decision trees which are essential in predictive modeling for regression, classification, and analyses, which function by forming an array of classification trees at test time and releasing the group which appears most frequently of the groups or average forecast (regression) of the particular trees. At each split in the tree, a random sample of predictors is chosen as split candidates from the full set of predictors (James et al., 2013).

Similar to the classification tree model, the prior probabilities were specified as 0.50 for predicted class 1, students who were retained, and 0.50 for predicted class 0, students who were not retained. The random forests model was evaluated using 10-fold cross-validation repeated five times using stratification. The two arguments tuned for the model were mtry and node size. The mtry argument is the number of predictors which were randomly sampled at each split when the tree models were created. By default, mtry is the square root of the number of predictors for classification. The node size argument is the minimum number of data points in a node required for the node to be split further. The default node size is 1 for classification. Tuned parameters, based on the largest AUC metric, were used to select the optimal model. The optimal mtry parameter (6) and minimum node size (25) were used to update the model and refit the training data. The 10-fold cross-validation was used to obtain a cross-validated error rate where the lowest rate indicated the tree which best fit the data. The resulting model had 500 trees with an out-of-bag (OOB) error rate of 17.65%, an error rate of 9.92% for class 0 (not retained), and an error rate of 25.38% for class 1 (retained). The overall training error rate for the model at 0.17 was close to the OOB error rate. For each random sample of predictors taken from the training data, some samples are not included called the out-of-bag samples. The OOB error rate is the average error for each of these OOB samples. Although the OOB error is used frequently for error estimation within random forests, it has been shown to overestimate in settings that include an equal number of observations from all response classes (balanced samples) (Janitza & Hornung, 2018).

The mean decrease in Gini was used to measure how important each variable was for estimating the value of the target variable across all of the trees which made up the

forest. The mean decrease in Gini is the average (mean) of the variable's total decrease in node impurity, weighted by the proportion of samples reaching that node in each decision tree in the random forest. The most important variables to the model result in the largest mean decrease in Gini value. The variables of GPA (137.47), age (132.06), and Transportation and Logistics programs (54.76) were the strongest predictors of being retained. The weakest predictors of retention were single parents (9.66), race (Hispanic) (6.36), race (other) (5.07), and Pell eligibility (4.30).

In comparison, the model applied to test data resulted in 500 trees with an out-of-bag (OOB) error rate of 14.86%, an error rate of 9.23% for class 0 (not retained), and an error rate of 20.48% for class 1 (retained). The overall error rate for the model was 0.34. The most important variables to the model result in the largest mean decrease in Gini value. The variables of GPA (162.68), age (104.84), and Industrial Technology programs (94.76) were the most influential predictors of being retained or not. The weakest predictors of retention were race (other) (10.71), high school diploma (GED®) (9.86), Pell eligibility (8.33), and single parents (6.69). The variable importance plot including all 15 predictor variables is shown in Figure 6.

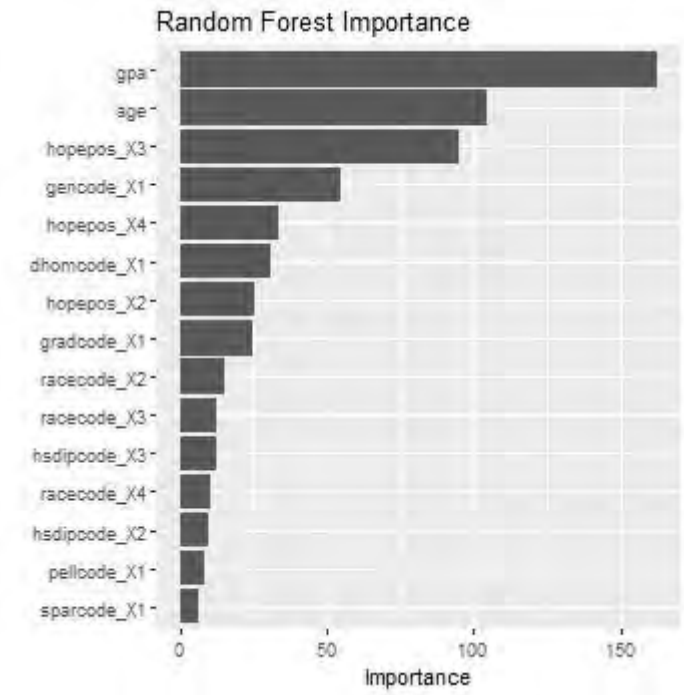

Figure 6. Random forests variable importance plot for certificates 9 to 17 credit hours in length.

The support vector machine model was the final data mining approach performed to answer Research Question 1A. A support vector machine model will be used to find the margin, which is the distance between the classification boundary and the closest training set data point (Kaiser et al., 2016; Kuhn & Johnson, 2013). In essence, the margin, defined by these data points, can be quantified and used to evaluate the performance of the model (Kuhn & Johnson, 2013).

All predictor variables were allowed to enter the model. The prior probabilities were specified as 0.50 for predicted class 1, students who were retained, and 0.50 for predicted class 0, students who were not retained. The SVM model was evaluated using 10-fold cross-validation repeated five times using stratification and tuned for the parameters cost and rbf_sigma. For the training data set, the optimal cost (based on the largest AUC metric) was calculated to be 0.1 and the optimal rbf_sigma was calculated to

be 0.1. These tuned parameters were used to update the model and refit the training data. The best model resulted in 1,635 support vectors, an objective function value of -142.55, and an error rate of 0.28. The overall training error rate for the model was 0.37.

Permutation-based variable importance scores were computed for each predictor in the SVM model. If a variable is important, the model's performance (based on the AUC metric) should change after permuting or rearranging the values of the variable. A larger change in the performance will indicate a more important variable. The strongest predictors of being retained were GPA (0.090), Industrial Technology programs (0.070), displaced homemakers (0.065), and Transportation and Logistics programs (0.062). The weakest predictors of retention were race (Black) (0.003), Pell eligibility (0.002), and race (other) (0.000).

In comparison, the model applied to test data resulted in 1,489 support vectors, an objective function value of -125.01, and an error rate of 0.23. The overall error rate for the model was 0.45. The most influential predictors of being retained or not were Industrial Technology programs (0.059), GPA (0.049), and Cyber, Engineer, or Healthcare programs (0.024). The least influential predictors of being retained or not were females (-0.004), displaced homemakers (-0.010), and graduation date (out of high school at least five years or more) (-0.011). The variable importance plot including all 15 predictor variables is shown in Figure 7.

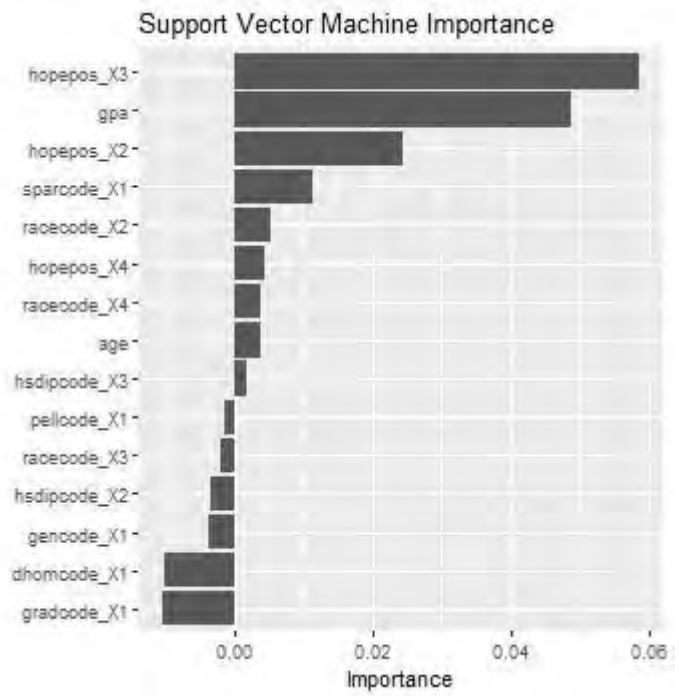

Figure 7. Support vector machine variable importance plot for certificates 9 to 17 credit hours in length.

For certificates 9–17 credit hours in length, environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) were analyzed to determine which, if any, were significant predictors of nontraditional student retention. Of the five statistical models evaluated using data for certificates 9 to 17 credit hours in length, the random forest model, the logistic regression model, and the linear discriminant model shared the same error rate using test data (0.34). The highest error rate using test data was the support vector machine model at 0.45. Both data modeling approaches, logistic regression and linear discriminant analysis, shared similar results in terms of variable importance. The predictor variables GPA and programs related to Transportation and Logistics were the most influential in students being retained, both of which represented

academic integration components. With every one-point increase in GPA, the odds of being retained increases 59% (odds ratio = 1.59), and compared to students enrolled in other HOPE Career Grant programs, students enrolled in a program related to Transportation and Logistics were more likely to be retained. Both the logistic regression and linear discriminant analysis models indicated Industrial Technology programs as the most influential predictor of students not being retained. Each of the three data mining approaches (classification trees, random forests, and support vector machines) identified similar predictor variables. The most common, influential predictors were GPA, age, and either Transportation and Logistics programs or Industrial Technology programs. However, the support vector machines model did not identify age as one of the top four predictors.

Research Question 1B. Are environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) significant predictors of nontraditional student retention for certificates 18–36 credit hours in length?

Two data modeling approaches (logistic regression and linear discriminant analysis) and three data mining approaches (classification tree, random forest, and support vector machine models) were used to answer this research question. Multivariate logistic regression analysis was the first of two data modeling approaches performed to answer this research question. The Hosmer-Lemeshow goodness of fit test examined whether the observed proportion of students retained is similar to or differs from the expected frequencies of retained students using a Pearson chi-square statistic ($\chi 2(14) =$

16.63, $p = 0.276$). Small values with large p-values indicate a good fit to the data while large values with p-values below 0.05 indicate a poor fit. In addition, the McFadden pseudo R^2 value was calculated as 0.09 indicating the model can account for 9% of the retained variable.

Table 22 shows the results from the logistic regression analysis. The overall model was found to be statistically significant, $\chi 2(15) = 157.25$, $p < 0.001$, and the model resulted in a training error rate of 0.37. Of the 15 predictor variables, three were statistically significant as predictors of student retention: GPA, Industrial Technology programs, and Transportation and Logistics programs. Given all other variables are unchanged, the odds of a student being retained increases by a factor of 1.77 (odds ratio = 1.77) as GPA increases. The odds of a student being retained when enrolled in a Transportation and Logistics program of study in certificates 18 to 36 credit hours increases by a factor of 1.44 (odds ratio = 1.44), given all other variables are unchanged. If a student self-identified as a single parent, the odds of those students being retained decreases by 8.3% (odds ratio = 0.917 – 1), keeping other variables constant.

To determine the most influential predictors of retention, variable importance was measured using the odds ratio. The variables of GPA (OR = 1.766, 95% CI = 1.561 to 2.001) and Transportation and Logistics programs (OR = 1.439, 95% CI = 1.238 to 1.689) were the strongest predictors of being retained. The weakest predictors of being retained were graduation date (out of high school at least five years or more) (OR = 0.936, 95% CI = 0.821 to 1.068) and single parents (OR = 0.917, 95% CI = 0.812 to 1.033).

Table 22

Variables Used to Predict Retention Utilizing Logistic Regression (Training Data)

Predictor	Log Odds	*SE*	*Z*	Pr(>|z|)		*OR*	*95% Confidence Interval* *Lower*	*Upper*
(Intercept)	-0.042	0.059	-0.711	0.477		0.959	0.855	1.076
age	0.053	0.063	0.842	0.400		1.054	0.932	1.192
gpa	0.568	0.063	8.983	$p < .001$	***	1.766	1.561	2.001
racecode_X2	0.053	0.065	0.810	0.418		1.054	0.928	1.197
racecode_X3	-0.025	0.060	-0.411	0.681		0.976	0.866	1.098
racecode_X4	0.029	0.060	0.481	0.631		1.029	0.914	1.159
gencode_X1	-0.064	0.069	-0.930	0.353		0.938	0.819	1.074
hsdipcode_X2	-0.016	0.097	-0.170	0.865		0.984	0.814	1.189
hsdipcode_X3	0.046	0.089	0.514	0.607		1.047	0.879	1.248
gradcode_X1	-0.066	0.067	-0.982	0.326		0.936	0.821	1.068
sparcode_X1	-0.087	0.061	-1.417	0.157		0.917	0.812	1.033
dhomcode_X1	0.120	0.065	1.843	0.065		1.127	0.994	1.285
pellcode_X1	0.087	0.066	1.315	0.188		1.091	0.958	1.244
hopepos_X2	0.077	0.061	1.267	0.205		1.080	0.959	1.216
hopepos_X3	0.122	0.062	1.953	0.051	*	1.130	1.003	1.283
hopepos_X4	0.364	0.079	4.613	$p < .001$	***	1.439	1.238	1.689

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

Additionally, the finalized logistic regression model was applied to the test data for comparison. Table 23 shows the results from the logistic regression analysis. The overall model was found to be statistically significant, $\chi 2(15) = 149.27$, $p < 0.001$, and the model resulted in an error rate of 0.39. In addition, the McFadden pseudo R^2 value was calculated as 0.11 indicating the model can account for 11% of the retained variable. Of the 15 predictor variables, five were statistically significant as predictors of student retention: GPA, graduation date (out of high school at least five years or more), Pell eligibility, Cyber, Engineer, or Healthcare programs, and Transportation and Logistics programs. Given all other variables are unchanged, the odds of a student being retained

increases by a factor of 2.21 (odds ratio = 2.21) as GPA increases. If a student enrolls in Cyber, Engineer, or Healthcare programs, the odds of those students being retained decreases by 21% (odds ratio = 0.792 – 1), keeping other variables constant.

Table 23

Variables Used to Predict Retention Utilizing Logistic Regression (Test Data)

Predictor	Log Odds	*SE*	*Z*	Pr(>|z|)		*OR*	*95% Confidence Interval*	
							Lower	*Upper*
(Intercept)	-0.088	0.071	-1.239	0.216		0.915	0.796	1.052
age	-0.089	0.075	-1.181	0.238		0.915	0.789	1.060
gpa	0.792	0.080	9.844	$p < .001$	***	2.209	1.892	2.594
racecode_X2	0.022	0.077	0.281	0.779		1.022	0.878	1.189
racecode_X3	0.053	0.075	0.708	0.479		1.054	0.911	1.223
racecode_X4	-0.063	0.072	-0.878	0.380		0.939	0.815	1.081
gencode_X1	0.094	0.095	0.997	0.319		1.099	0.913	1.325
hsdipcode_X2	-0.004	0.115	-0.035	0.972		0.996	0.795	1.247
hsdipcode_X3	0.093	0.104	0.899	0.369		1.098	0.896	1.347
gradcode_X1	-0.202	0.082	-2.472	0.013	*	0.817	0.696	0.959
sparcode_X1	0.096	0.074	1.292	0.196		1.101	0.952	1.275
dhomcode_X1	0.096	0.077	1.236	0.216		1.100	0.948	1.285
pellcode_X1	0.178	0.088	2.016	0.044	*	1.195	1.006	1.422
hopepos_X2	-0.233	0.084	-2.777	0.005	**	0.792	0.672	0.933
hopepos_X3	0.037	0.079	0.471	0.637		1.038	0.889	1.215
hopepos_X4	0.249	0.093	2.688	0.007	**	1.282	1.071	1.541

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

The variables of GPA (OR = 2.209, 95% CI = 1.892 to 2.594) and Transportation and Logistics programs (OR = 1.282, 95% CI = 1.071 to 1.541) were the strongest predictors of being retained in the test data. The weakest predictors of being retained were graduation date (out of high school at least five years or more) (OR = 0.817, 95% CI = 0.696 to 0.959) and Cyber, Engineer, or Healthcare programs (OR = 0.792, 95% CI = 1.071 to 1.541). The variable importance plot including all 15 predictor variables is

shown in Figure 8. Because variable importance for logistic regression is based on the absolute values of the z-statistic, both the most influential and the least influential predictors may be displayed at the top of the plot.

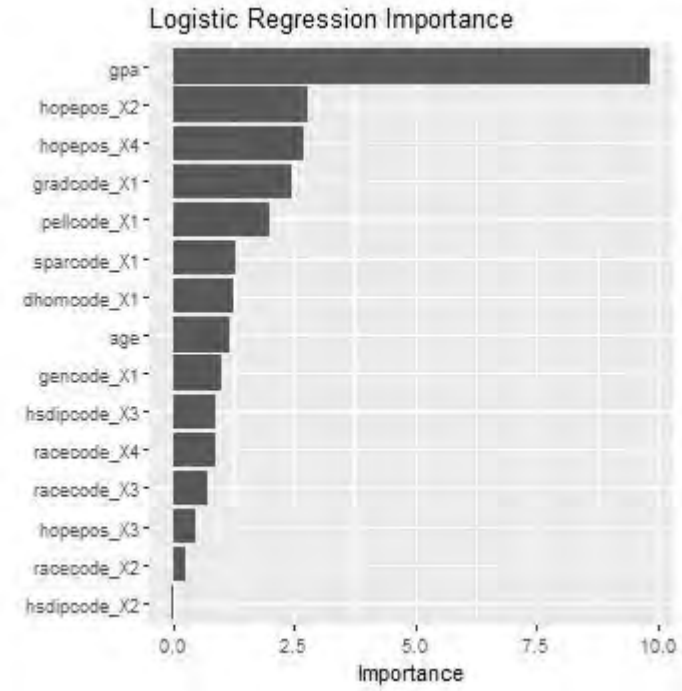

Figure 8. Logistic regression variable importance plot for certificates 18 to 36 credit hours in length.

The second data modeling approach used to address Research Question 1B was linear discriminant analysis. The LDA model fit resulted in a training error rate of 0.37 and the results were similar to those of the logistic regression. As shown in Table 24, coefficients with the strongest associated weights included GPA (0.830), displaced homemakers (0.143), Industrial Technology programs (0.178), and Transportation and Logistics programs (0.455). The larger the coefficient of a predictor in the standardized discriminant function, the more important its role in the discriminant function. Similar to certificates 9 to 17 credit hours in length, GPA was the strongest predictor of retention with a coefficient of 0.830, and Transportation and Logistics programs were the second

most influential with a coefficient of 0.455. Both high school diploma (GED®) (-0.026) and race (Hispanic) (-0.034) had the least influential coefficients indicating they are not significant predictors of retention.

Table 24

Variables Used to Predict Retention Utilizing LDA (Training Data)

Independent Variable	Not Retained Mean	Retained Mean	Coefficients of Linear Discriminants: LD1
age	-0.038	0.082	0.074
gpa	-0.220	0.314	0.830
racecode_X2	0.028	-0.020	0.070
racecode_X3	0.025	-0.020	-0.034
racecode_X4	-0.002	0.011	0.038
gencode_X1	0.094	-0.143	-0.095
hsdipcode_X2	-0.022	0.086	-0.026
hsdipcode_X3	0.013	-0.035	0.074
gradcode_X1	0.044	-0.114	-0.102
sparcode_X1	0.064	-0.069	-0.119
dhomcode_X1	-0.072	0.123	0.143
pellcode_X1	0.069	-0.085	0.125
hopepos_X2	0.011	-0.005	0.112
hopepos_X3	-0.064	0.107	0.178
hopepos_X4	-0.145	0.228	0.455

Note. Prior probabilities of groups: not retained: 0.5, retained: 0.5.

In comparison, Table 25 includes the coefficients of linear discriminants for the test data. The overall error rate for the model was 0.39 and the coefficients with the strongest associated weights included GPA (0.978), graduation date (out of high school at least five years or more) (-0.245), Pell eligibility (0.215), Cyber, Engineer, or Healthcare programs (-0.287), and Transportation and Logistics programs (0.305). The strongest predictor of being retained or not based on test data was GPA with a coefficient of 0.978 and Transportation and Logistics programs with a coefficient of 0.305. The variables race

(Black) (0.020) and high school diploma (GED®) (-0.021) had the least influential coefficients of being retained or not. The variable GPA had the largest variance within the group means. This variable has a greater influence on students not being retained (0.605) than on students being retained (-0.126). The variable importance plot including all 15 predictor variables is shown in Figure 9.

Table 25

Variables Used to Predict Retention Utilizing LDA (Test Data)

Independent Variable	Not Retained Mean	Retained Mean	Coefficients of Linear Discriminants: LD1
age	0.026	-0.022	-0.102
gpa	-0.256	0.383	0.978
racecode_X2	0.005	-0.062	0.020
racecode_X3	-0.022	0.037	0.062
racecode_X4	0.030	0.013	-0.072
gencode_X1	-0.007	-0.021	0.109
hsdipcode_X2	-0.056	0.027	-0.021
hsdipcode_X3	0.034	-0.017	0.107
gradcode_X1	0.054	-0.121	-0.245
sparcode_X1	-0.049	0.027	0.113
dhomcode_X1	-0.069	0.077	0.117
pellcode_X1	-0.002	0.015	0.215
hopepos_X2	0.090	-0.163	-0.287
hopepos_X3	-0.017	0.031	0.040
hopepos_X4	-0.087	0.100	0.305

Note. Prior probabilities of groups: not retained: 0.5, retained: 0.5.

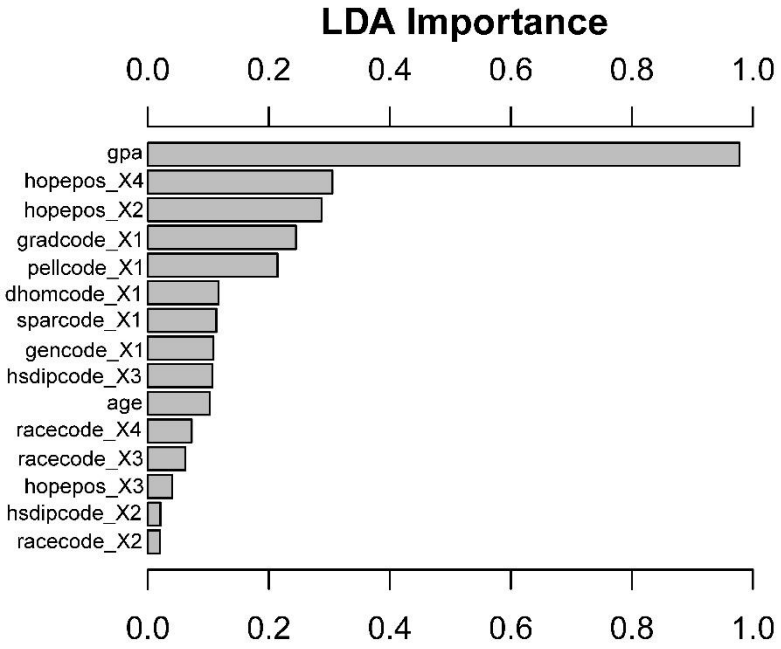

Figure 9. Linear discriminant analysis variable importance plot for certificates 18 to 36 credit hours in length.

The first of the three data mining approaches, a classification tree, was used to answer Research Question 1B where all predictor variables were allowed to enter the model. The prior probabilities were specified as 0.50 for predicted class 1, students who were retained, and 0.50 for predicted class 0, students who were not retained. The classification tree model was evaluated using 10-fold cross-validation repeated five times using stratification and tuned for the parameters complexity and tree depth. The best parameters, based on the largest AUC metric, were used to select the optimal model. The optimal complexity factor (0.0000793) and maximum tree depth (4) were used to update the model and refit the training data. The 10-fold cross-validation was used to obtain a cross-validated error rate where the lowest rate indicated the tree which best fit the data. The resulting model had 4 total splits and a cross-validated error rate of 0.74. The overall training error rate for the model was 0.41.

To determine the strongest predictors of retention, variable importance was measured as the sum of the goodness of split measures (Gini index). The variables of GPA ($I_G = 54.30$) and Transportation and Logistics programs ($I_G = 21.95$) were the strongest predictors of being retained or not. The weakest predictors of being retained or not were race (other) ($I_G = 0.51$), graduation date (out of high school at least five years or more) ($I_G = 0.35$), single parents ($I_G = 0.16$), and Cyber, Engineer, or Healthcare programs ($I_G = 0.05$).

In comparison, the model applied to test data resulted in 5 total splits, a cross-validated error rate of 0.77, and an overall error rate for the model of 0.41. GPA ($I_G = 57.06$) was the most influential predictor of being retained or not, with age coming in a distant second with $I_G = 8.65$. The weakest predictors of being retained or not were race (Black) ($I_G = 0.40$), race (other) ($I_G = 0.33$), high school diploma (college prep or tech prep) ($I_G = 0.19$), and single parents ($I_G = 0.19$). The variable importance plot including 13 of the 15 predictor variables is shown in Figure 10. Two predictor variables, race (Hispanic) and Industrial Technology programs, had importance scores below zero.

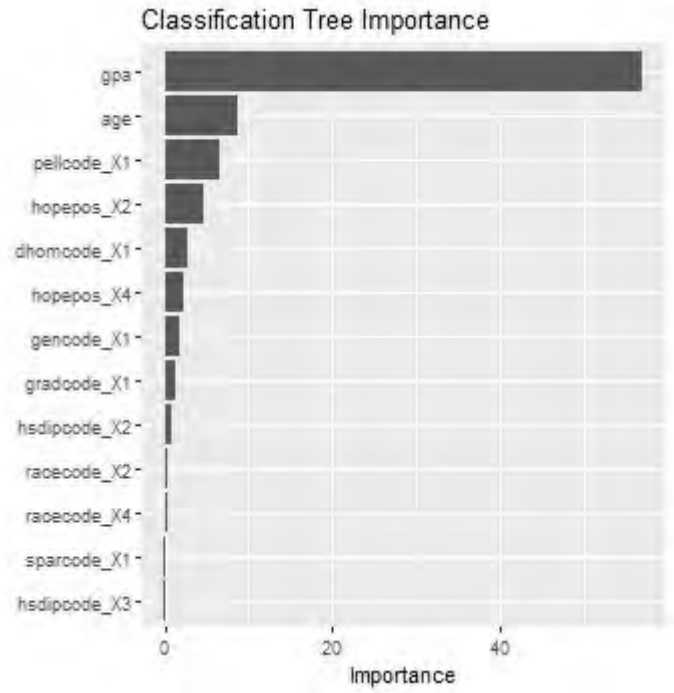

Figure 10. Classification tree variable importance plot for certificates 18 to 36 credit hours in length.

The random forest model was the second of three data mining approaches performed to answer Research Question 1B. The prior probabilities were specified as 0.50 for predicted class 1, students who were retained, and 0.50 for predicted class 0, students who were not retained. The random forests model was evaluated using 10-fold cross-validation repeated five times using stratification. Tuned parameters, based on the largest AUC metric, were used to select the optimal model. The optimal mtry parameter (6) and minimum node size (39) were used to update the model and refit the training data. The 10-fold cross-validation was used to obtain a cross-validated error rate where the lowest rate indicated the tree which best fit the data. The resulting model had 500 trees with an out-of-bag (OOB) error rate of 38.36%, an error rate of 43.69% for class 0 (not retained), and an error rate of 33.03% for class 1 (retained). The overall training error rate for the model was 0.31.

The mean decrease in Gini was used to measure how important each variable was for estimating the value of the target variable across all of the trees that made up the forest. The mean decrease in Gini is the average (mean) of the variable's total decrease in node impurity, weighted by the proportion of samples reaching that node in each decision tree in the random forest. The most important variables to the model result in the largest mean decrease in Gini value. The variables of GPA (72.12), age (33.33), and Transportation and Logistics programs (14.76) were the strongest predictors of being retained. The weakest predictors of retention were high school diploma (GED®) (3.40), race (other) (2.94), and race (Hispanic) (2.83).

In comparison, the model applied to test data resulted in 500 trees with an out-of-bag (OOB) error rate of 36.55%, an error rate of 46.85% for class 0 (not retained), and an error rate of 26.26% for class 1 (retained). The overall error rate for the model was 0.39. The most important variables to the model result in the largest mean decrease in Gini value. The variables of GPA (68.60) and age (23.79) were the most influential predictors of being retained or not. The weakest predictors of retention were high school diploma (GED®) (2.11), Industrial Technology programs (1.52), and race (other) (1.47). The variable importance plot including all 15 predictor variables is shown in Figure 11.

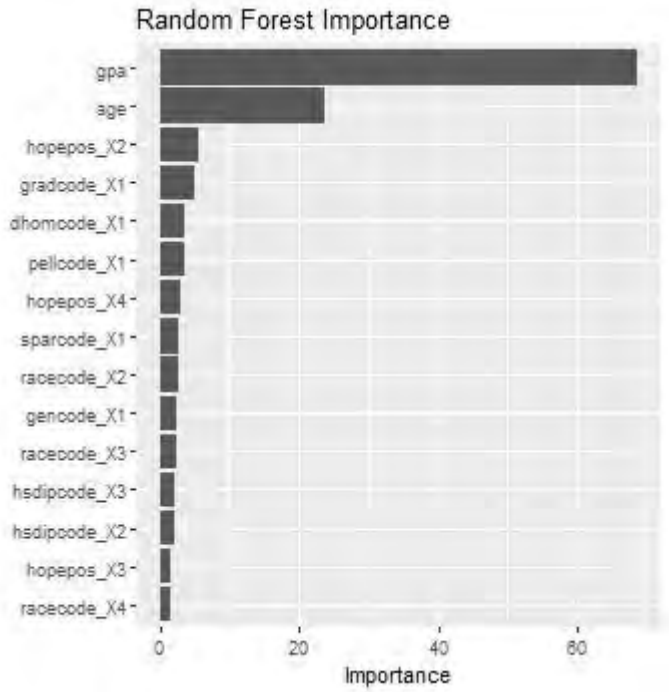

Figure 11. Random forests variable importance plot for certificates 18 to 36 credit hours in length.

The support vector machine model was the final data mining approach performed to answer Research Question 1B. All predictor variables were allowed to enter the model. The prior probabilities were specified as 0.50 for predicted class 1, students who were retained, and 0.50 for predicted class 0, students who were not retained. The SVM model was evaluated using 10-fold cross-validation repeated five times using stratification and tuned for the parameters cost and rbf_sigma. For the training data set, the optimal cost (based on the largest AUC metric) was calculated to be 0.1 and the optimal rbf_sigma was calculated to be 0.1. These tuned parameters were used to update the model and refit the training data. The best model resulted in 1,238 support vectors, an objective function value of -113.06, and an error rate of 0.35. The overall training error rate for the model was 0.40.

Permutation-based variable importance scores were computed for each predictor in the SVM model. If a variable is important, the model's performance (based on the AUC metric) should change after permuting or rearranging the values of the variable. A larger change in the performance will indicate a more important variable. For the SVM model, the strongest predictors of being retained were GPA (0.107), Transportation and Logistics programs (0.012), and single parents (0.011). The weakest predictors of retention were graduation date (out of high school at least five years or more) (0.002), displaced homemaker (0.002), and high school diploma (GED®) (0.001).

In comparison, the model applied to test data resulted in 901 support vectors, an objective function value of -80.91, and an error rate of 0.36. The overall error rate for the model was 0.43. The most influential predictors of being retained or not were GPA (0.071), Transportation and Logistics programs (0.017), and race (Black) (0.004). The least influential predictors of being retained or not were age, race (Hispanic), and displaced homemaker, all sharing the same importance score of -0.006. The variable importance plot including all 15 predictor variables is shown in Figure 12.

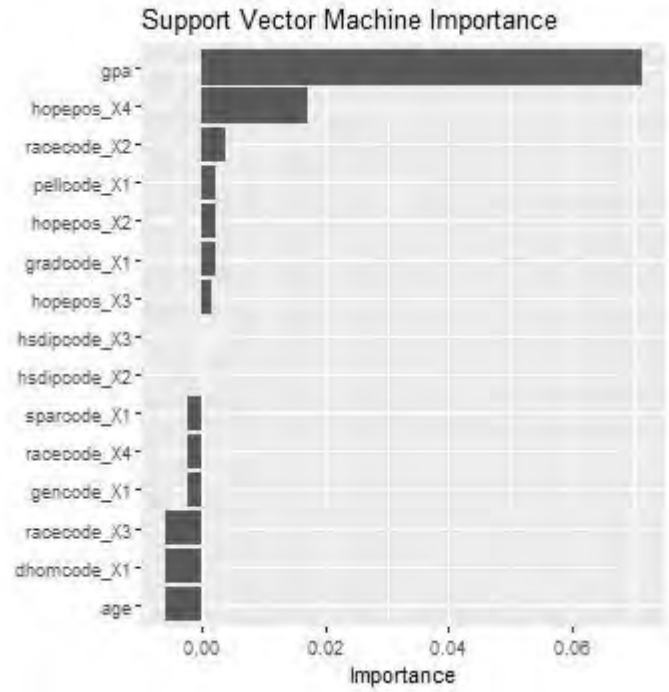

Figure 12. Support vector machine variable importance plot for certificates 18 to 36 credit hours in length.

For certificates 18–36 credit hours in length, environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) were analyzed to determine which, if any, were significant predictors of nontraditional student retention. Of the five statistical models using test data, the random forest model, the logistic regression model, and the linear discriminant model all shared the lowest error rate of 0.39. The highest error rate using test data was the support vector machine model at 0.45. In terms of variable importance, the results for this group were similar to those of certificates 9–17 credit hours in length. Both data modeling approaches, logistic regression, and linear discriminant analysis indicated GPA and programs related to Transportation and Logistics were the most influential in students being retained. Both

the logistic regression and linear discriminant analysis models indicated that being out of high school for five years or more and being enrolled in a Cyber, Engineer, or Healthcare program are influential predictors of students not being retained. For example, if a student enrolls in a Cyber, Engineer, or Healthcare program, the odds of that student being retained decreases by 21% (odds ratio = 0.792 – 1), keeping other variables constant. Each of the three data mining approaches (classification trees, random forests, and support vector machines) identified similar predictor variables. The most common, influential predictors were GPA and age. However, similar to certificates 9–17 credit hours in length, the support vector machines model did not identify age as one of the top predictors.

Research Question 1C. Are environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) significant predictors of nontraditional student retention for diplomas 37–48 credit hours in length?

Two data modeling approaches (logistic regression and linear discriminant analysis) and three data mining approaches (classification tree, random forest, and support vector machine models) were used to answer this research question. Multivariate logistic regression analysis was the first of two data modeling approaches performed to answer this research question. The Hosmer-Lemeshow goodness of fit test examined whether the observed proportion of students retained is similar to or differs from the expected frequencies of retained students using a Pearson chi-square statistic ($\chi 2(14) = 32.59$, $p < 0.01$). Small values with large p-values indicate a good fit to the data while

large values with p-values below 0.05 indicate a poor fit. In addition, the McFadden pseudo R^2 value was calculated as 0.09 indicating the model can account for 9% of the retained variable.

Table 26 shows the results from the logistic regression analysis. The overall model was found to be statistically significant, $\chi 2(15) = 112.66$, $p < 0.001$, and the model resulted in a training error rate of 0.40. Of the 15 predictor variables, three were statistically significant as predictors of student retention: age, GPA, and Pell eligibility. The largest odds ratio indicates, given all other variables are unchanged, the odds of a student being retained increases by a factor of 2.01 as GPA increases. Also, the odds of a student being retained when they receive the Pell grant increases by a factor of 1.28 (odds ratio = 1.28), given all other variables are unchanged.

To determine the strongest predictors of retention, variable importance was measured using the odds ratio. The variables of GPA (OR = 2.012, 95% CI = 1.705 to 2.387) and Pell eligibility (OR = 1.283, 95% CI = 1.095 to 1.507) were the strongest predictors of being retained. The weakest predictors of retention were high school diploma (college prep or tech prep) (OR = 0.930, 95% CI = 0.755 to 1.143) and high school diploma (GED®) (OR = 0.875, 95% CI = 0.702 to 1.089).

Table 26

Variables Used to Predict Retention Utilizing Logistic Regression (Training Data)

Predictor	Log Odds	*SE*	*Z*	Pr(>|z|)		*OR*	*95% Confidence Interval* *Lower*	*Upper*
(Intercept)	-0.125	0.075	-1.668	0.095		0.883	0.762	1.021
age	0.165	0.081	2.047	0.041	*	1.179	1.007	1.382
gpa	0.699	0.086	8.148	$p < .001$	***	2.012	1.705	2.387
racecode_X2	0.001	0.083	0.010	0.992		1.001	0.852	1.177
racecode_X3	0.112	0.074	1.523	0.128		1.119	0.969	1.295
racecode_X4	0.001	0.078	0.015	0.988		1.001	0.857	1.166
gencode_X1	-0.019	0.096	-0.202	0.840		0.981	0.813	1.184
hsdipcode_X2	-0.133	0.112	-1.193	0.233		0.875	0.702	1.089
hsdipcode_X3	-0.073	0.106	-0.689	0.491		0.930	0.755	1.143
gradcode_X1	0.132	0.086	1.527	0.127		1.141	0.964	1.353
sparcode_X1	0.095	0.079	1.196	0.232		1.099	0.942	1.285
dhomcode_X1	-0.036	0.083	-0.436	0.663		0.964	0.817	1.135
pellcode_X1	0.249	0.081	3.069	0.002	**	1.283	1.095	1.507
hopepos_X2	0.054	0.077	0.707	0.480		1.056	0.908	1.228
hopepos_X3	0.127	0.092	1.378	0.168		1.136	0.948	1.362
hopepos_X4	0.016	0.082	0.201	0.841		1.017	0.866	1.194

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

Additionally, the finalized logistic regression model was applied to the test data for comparison. Table 27 shows the results from the logistic regression analysis. The overall model was found to be statistically significant, $\chi 2(15) = 162.89$, $p < 0.001$, and the model resulted in an error rate of 0.43. In addition, the McFadden pseudo R^2 value was calculated as 0.12 indicating the model can account for 12% of the retained variable. Of the 15 predictor variables, six were statistically significant as predictors of student retention: GPA, race (Black), race (other), gender (female), high school diploma (college prep or tech prep), and Cyber, Engineer, or Healthcare programs. Given all other variables are unchanged, the odds of a student being retained increases by a factor of 2.21

(odds ratio = 2.21) as GPA increases. If a student enrolls in Industrial Technology programs, the odds of those students being retained decreases by 49% (odds ratio = 0.510 – 1), keeping other variables constant.

Table 27

Variables Used to Predict Retention Utilizing Logistic Regression (Test Data)

Predictor	Log Odds	*SE*	*Z*	Pr(>|z|)		*OR*	*95% Confidence Interval* *Lower*	*Upper*
(Intercept)	-0.221	0.075	-2.927	0.003	**	0.802	0.691	0.929
age	0.092	0.074	1.242	0.214		1.097	0.948	1.270
gpa	0.793	0.086	9.190	$p < .001$	***	2.211	1.872	2.627
racecode_X2	0.270	0.087	3.109	0.002	**	1.310	1.106	1.554
racecode_X3	-0.145	0.083	-1.733	0.083		0.865	0.733	1.018
racecode_X4	-0.185	0.088	-2.108	0.035	*	0.831	0.694	0.982
gencode_X1	0.285	0.093	3.064	0.002	**	1.330	1.109	1.598
hsdipcode_X2	-0.042	0.111	-0.378	0.706		0.959	0.772	1.191
hsdipcode_X3	0.208	0.104	1.996	0.046	*	1.231	1.004	1.512
gradcode_X1	-0.117	0.080	-1.453	0.146		0.890	0.760	1.041
sparcode_X1	-0.089	0.080	-1.114	0.265		0.914	0.781	1.071
dhomcode_X1	-0.122	0.093	-1.317	0.188		0.885	0.733	1.058
pellcode_X1	0.136	0.079	1.735	0.083		1.146	0.983	1.338
hopepos_X2	-0.328	0.086	-3.808	$p < .001$	***	0.720	0.607	0.851
hopepos_X3	0.156	0.090	1.727	0.084		1.169	0.980	1.397
hopepos_X4	0.041	0.083	0.491	0.623		1.041	0.885	1.225

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

The variables of GPA (OR = 2.211, 95% CI = 1.872 to 2.627), females (OR = 1.330, 95% CI = 1.109 to 1.598), and race (Black) (OR = 1.310, 95% CI = 1.106 to 1.554) were the most influential predictors of being retained or not in the test data. The weakest predictors of being retained or not were race (other) (OR = 0.831, 95% CI = 0.694 to 0.982) and Cyber, Engineer, or Healthcare programs (OR = 0.720, 95% CI = 0.607 to 0.851). The variable importance plot including all 15 predictor variables is

shown in Figure 13. Because variable importance for logistic regression is based on the absolute values of the z-statistic, both the most influential and the least influential predictors may be displayed at the top of the plot.

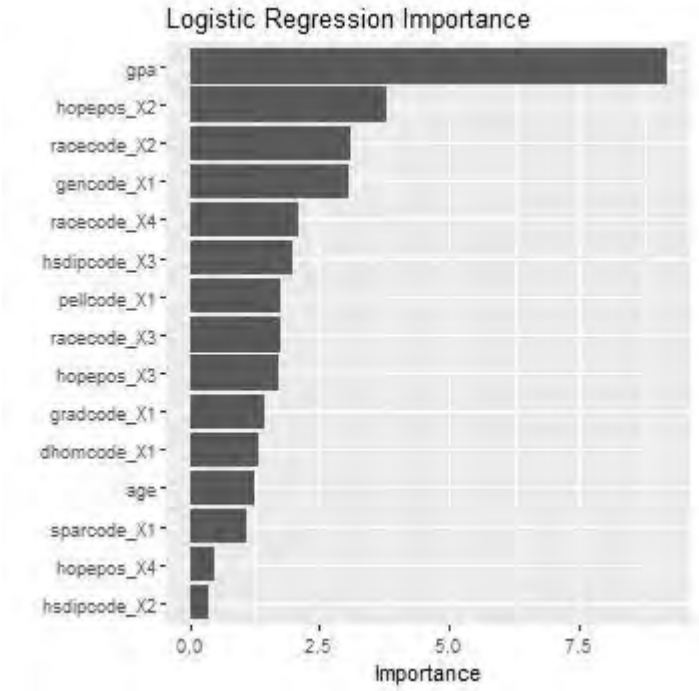

Figure 13. Logistic regression variable importance plot for diplomas 37 to 48 credit hours in length.

The second data modeling approach used to address Research Question 1C was linear discriminant analysis. The LDA model fit resulted in a training error rate of 0.40 and the results were similar to those of the logistic regression. As shown in Table 28, coefficients with the strongest associated weights included age (0.219), GPA (0.957), high school diploma (GED®) (-0.179), graduation date (out of high school at least five years or more) (0.175), and Pell eligibility (0.330). The larger the coefficient of a predictor in the standardized discriminant function, the more important its role in the discriminant function. Similar to both certificate data files, GPA was the strongest predictor of being retained or not with a coefficient of 0.957. However, instead of

Transportation and Logistics programs ranking as an influential predictor, Pell eligibility served as the second strongest predictor of student retention with a coefficient of 0.330. Both high school diploma (college prep or tech prep) (-0.095) and high school diploma (GED®) (-0.179) had the least influential coefficients of being retained or not.

Table 28

Variables Used to Predict Retention Utilizing LDA (Training Data)

Independent Variable	Not Retained Mean	Retained Mean	Coefficients of Linear Discriminants: LD1
age	-0.038	0.132	0.219
gpa	-0.197	0.409	0.957
racecode_X2	0.033	-0.102	0.000
racecode_X3	-0.007	0.064	0.150
racecode_X4	-0.003	-0.031	-0.004
gencode_X1	0.021	-0.053	-0.029
hsdipcode_X2	0.047	-0.104	-0.179
hsdipcode_X3	-0.020	0.049	-0.095
gradcode_X1	-0.032	0.165	0.175
sparcode_X1	-0.033	0.043	0.122
dhomcode_X1	-0.020	-0.044	-0.044
pellcode_X1	-0.033	0.057	0.330
hopepos_X2	0.008	-0.006	0.070
hopepos_X3	-0.036	0.140	0.168
hopepos_X4	-0.016	-0.025	0.014

Note. Prior probabilities of groups: not retained: 0.5, retained: 0.5.

In comparison, Table 29 includes the coefficients of linear discriminants for the test data. The overall error rate for the model was 0.44 and the coefficients with the strongest associated weights included GPA (0.915), race (Black) (0.296), females (0.327), and Cyber, Engineer, or Healthcare programs (-0.359). The most influential predictors of being retained or not based on test data were GPA with a coefficient of 0.915, Cyber, Engineer, or Healthcare programs with a negative coefficient of -0.359, and

females with a coefficient of 0.327. The variables Transportation and Logistics programs (0.039) and high school diploma (GED®) (-0.025) had the least influential coefficients of being retained or not. The variable GPA had the largest variance within the group means. This variable has a greater influence on students not being retained (0.605) than on students being retained (-0.126). The variable importance plot including all 15 predictor variables is shown in Figure 14.

Table 29

Variables Used to Predict Retention Utilizing LDA (Test Data)

Independent Variable	Not Retained Mean	Retained Mean	Coefficients of Linear Discriminants: LD1
age	-0.031	0.114	0.104
gpa	-0.160	0.381	0.915
racecode_X2	-0.031	0.143	0.296
racecode_X3	0.030	-0.079	-0.157
racecode_X4	0.045	-0.113	-0.201
gencode_X1	-0.046	0.103	0.327
hsdipcode_X2	0.038	-0.126	-0.025
hsdipcode_X3	-0.032	0.150	0.243
gradcode_X1	0.002	-0.004	-0.137
sparcode_X1	0.004	-0.060	-0.110
dhomcode_X1	0.027	-0.096	-0.095
pellcode_X1	0.001	0.077	0.146
hopepos_X2	0.059	-0.218	-0.359
hopepos_X3	-0.041	0.086	0.175
hopepos_X4	0.026	-0.040	0.039

Note. Prior probabilities of groups: not retained: 0.5, retained: 0.5.

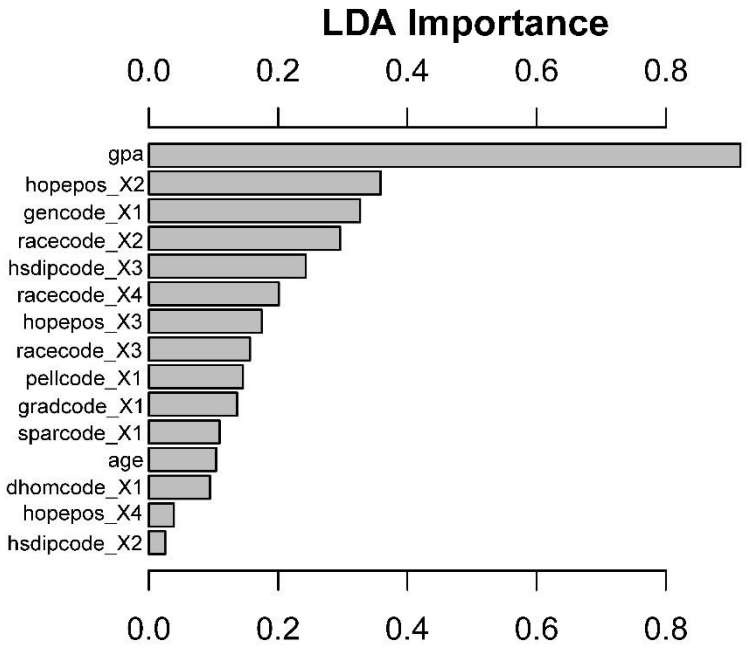

Figure 14. Linear discriminant analysis variable importance plot for diplomas 37 to 48 credit hours in length.

The first of the three data mining approaches, a classification tree, was used to answer Research Question 1C where all predictor variables were allowed to enter the model. The prior probabilities were specified as 0.50 for predicted class 1, students who were retained, and 0.50 for predicted class 0, students who were not retained. The classification tree model was evaluated using 10-fold cross-validation repeated five times using stratification and tuned for the parameters complexity and tree depth. The best parameters, based on the largest AUC metric, were used to select the optimal model. The optimal complexity factor (0.0000793) and maximum tree depth (4) were used to update the model and refit the training data. The 10-fold cross-validation was used to obtain a cross-validated error rate where the lowest rate indicated the tree which best fit the data. The resulting model had 4 total splits and a cross-validated error rate of 0.69. The overall training error rate for the model was 0.38.

To determine the strongest predictors of retention, variable importance was measured as the sum of the goodness of split measures (Gini index). The variable GPA was by far the strongest predictor of being retained or not with a Gini index of 61.03. The variable race (Hispanic) was a distant second with a Gini index of 7.42. The weakest predictors of retention were Cyber, Engineer, or Healthcare programs ($I_G = 0.72$), displaced homemakers ($I_G = 0.33$), and Transportation and Logistics programs ($I_G = 0.17$).

In comparison, the model applied to test data resulted in 7 total splits, a cross-validated error rate of 0.69, and an overall error rate for the model of 0.50. The variables of GPA ($I_G = 46.21$) and age ($I_G = 21.13$) were the strongest predictors of being retained or not. In contrast, the weakest predictors of being retained or not were females ($I_G = 0.75$), Transportation and Logistics programs ($I_G = 0.41$), and displaced homemakers ($I_G = 0.38$). The variable importance plot including 13 of the 15 predictor variables is shown in Figure 15. Two predictor variables, race (other) and single parents, had importance scores below zero.

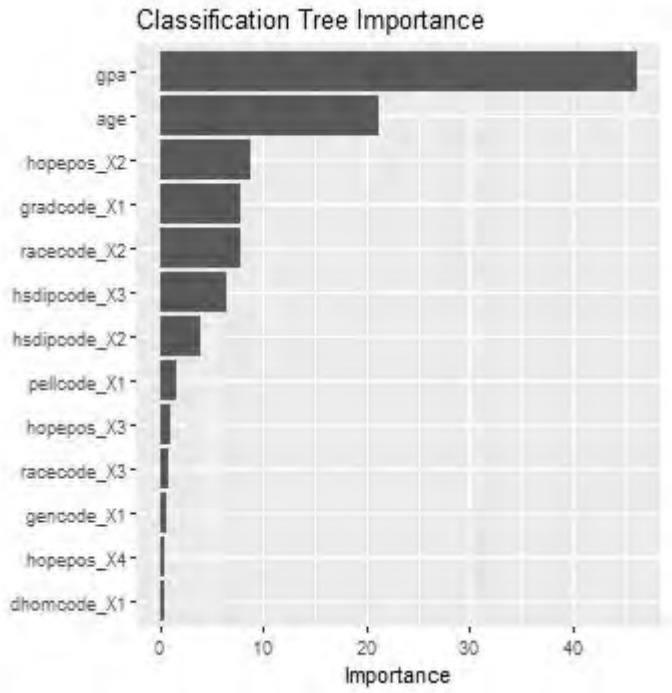

Figure 15. Classification tree variable importance plot for diplomas 37 to 48 credit hours in length.

The random forest model was the second of three data mining approaches performed to answer Research Question 1C. The prior probabilities were specified as 0.50 for predicted class 1, students who were retained, and 0.50 for predicted class 0, students who were not retained. The random forests model was evaluated using 10-fold cross-validation repeated five times using stratification. Tuned parameters, based on the largest AUC metric, were used to select the optimal model. The optimal mtry parameter (10) and minimum node size (35) were used to update the model and refit the training data. The 10-fold cross-validation was used to obtain a cross-validated error rate where the lowest rate indicated the tree which best fit the data. The resulting model had 500 trees with an out-of-bag (OOB) error rate of 29.07%, an error rate of 38.84% for class 0 (not retained), and an error rate of 19.30% for class 1 (retained). The overall training error rate for the model was 0.24.

The mean decrease in Gini was used to measure how important each variable was for estimating the value of the target variable across all of the trees that made up the forest. The mean decrease in Gini is the average (mean) of the variable's total decrease in node impurity, weighted by the proportion of samples reaching that node in each decision tree in the random forest. The most important variables to the model result in the largest mean decrease in Gini value. The variables of GPA (83.73), age (33.44), and high school diploma (college prep or tech prep) (7.33) were the strongest predictors of being retained. The weakest predictors of retention were Transportation and Logistics programs (2.06), displaced homemakers (1.79), and race (other) (0.93).

In comparison, the model applied to test data resulted in 500 trees with an out-of-bag (OOB) error rate of 26.06%, an error rate of 33.69% for class 0 (not retained), and an error rate of 18.43% for class 1 (retained). The overall error rate for the model was 0.44. The most important variables to the model result in the largest mean decrease in Gini value. The variables of GPA (73.20), age (49.98), and Cyber, Engineer, or Healthcare programs (11.47) were the most influential predictors of being retained or not. The weakest predictors of retention were single parents (2.23), race (other) (2.11), and displaced homemakers (1.08). The variable importance plot including all 15 predictor variables is shown in Figure 16.

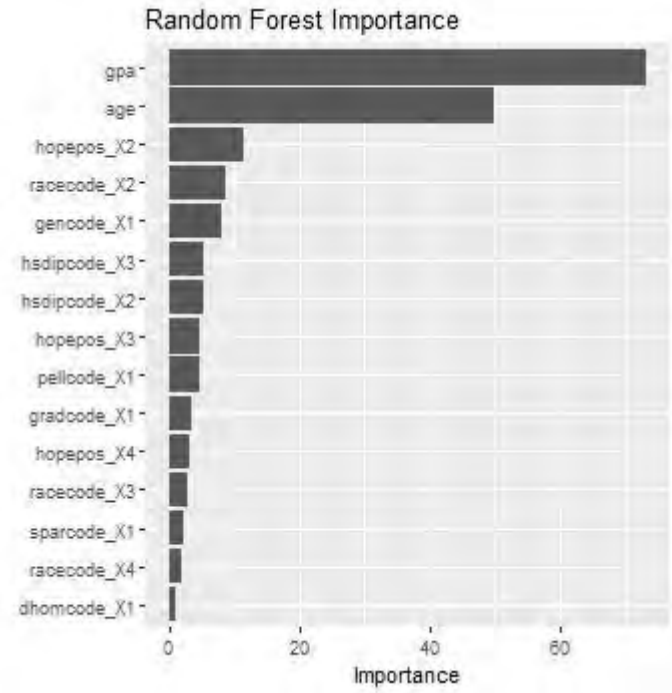

Figure 16. Random forests variable importance plot for diplomas 37 to 48 credit hours in length.

The support vector machine model was the final data mining approach performed to answer Research Question 1C. All predictor variables were allowed to enter the model. The prior probabilities were specified as 0.50 for predicted class 1, students who were retained, and 0.50 for predicted class 0, students who were not retained. The SVM model was evaluated using 10-fold cross-validation repeated five times using stratification and tuned for the parameters cost and rbf_sigma. For the training data set, the optimal cost (based on the largest AUC metric) was calculated to be 0.1 and the optimal rbf_sigma was calculated to be 0.1. These tuned parameters were used to update the model and refit the training data. The best model resulted in 812 support vectors, an objective function value of -71.36, and an error rate of 0.29. The overall training error rate for the model was 0.35.

Permutation-based variable importance scores were computed for each predictor in the SVM model. If a variable is important, the model's performance (based on the AUC metric) should change after permuting or rearranging the values of the variable. A larger change in the performance will indicate a more important variable. For the SVM model, the strongest predictors of being retained or not were GPA (0.119), high school diploma (GED®) (0.034), and graduation date (out of high school at least five years or more) (0.023). The weakest predictors of retention were race (Black) (0.007), Industrial Technology programs (0.007), and race (Hispanic) (0.001).

In comparison, the model applied to test data resulted in 880 support vectors, an objective function value of -76.35, and an error rate of 0.32. The overall error rate for the model was 0.42. The most influential predictors of being retained or not were GPA (0.061), age (0.019), and race (Black) (0.015). The least influential predictors of being retained or not were displaced homemakers (-0.007), race (Hispanic) (-0.009), and single parents (-0.009). The variable importance plot including all 15 predictor variables is shown in Figure 17.

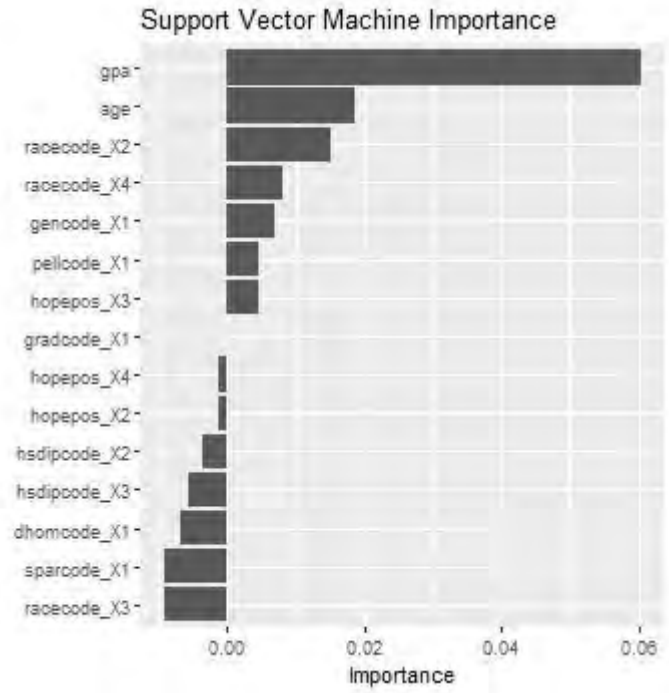

Figure 17. Support vector machine variable importance plot for diplomas 37 to 48 credit hours in length.

For diplomas 37–48 credit hours in length, environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) were analyzed to determine which, if any, were significant predictors of nontraditional student retention. Of the five statistical models evaluated using data for diplomas 37 to 48 credit hours in length, the support vector machine model produced the lowest error rate using test data (0.42). The highest error rate using test data was a dismal 0.50 which belonged to the classification tree model. Both data modeling approaches, logistic regression and linear discriminant analysis, shared similar results in regards to variable importance. The predictor variables GPA, female students, and Black students were the most influential in students being retained, representing background factors (gender and race) and an academic integration

component (GPA). Both the logistic regression and linear discriminant analysis models using the test data indicated Cyber, Engineer, or Healthcare programs as the most influential predictor of students not being retained. Each of the three data mining approaches (classification trees, random forests, and support vector machines) identified similar predictor variables. The most common, influential predictors were GPA (an academic integration component) and age (a background factor). Unlike the previous results for certificate programs, the support vector machines model did identify age as one of the most influential predictors of student retention, but only in the test data.

Research Question 1D. Are environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) significant predictors of nontraditional student retention for diplomas 49–59 credit hours in length?

Two data modeling approaches (logistic regression and linear discriminant analysis) and three data mining approaches (classification tree, random forest, and support vector machine models) were used to answer this research question. Multivariate logistic regression analysis was the first of two data modeling approaches performed to answer this research question. The Hosmer-Lemeshow goodness of fit test examined whether the observed proportion of students retained is similar to or differs from the expected frequencies of retained students using a Pearson chi-square statistic ($\chi 2(14) = 37.61, p < 0.001$). Small values with large p-values indicate a good fit to the data while large values with p-values below 0.05 indicate a poor fit. In addition, the McFadden

pseudo R^2 value was calculated as 0.08 indicating the model can account for 8% of the retained variable.

Table 30 shows the results from the logistic regression analysis. The overall model was found to be statistically significant, $\chi 2(15) = 252.27$, $p < 0.001$, and the model resulted in a training error rate of 0.39. Of the 15 predictor variables, seven were statistically significant as predictors of student retention: GPA, race (other), high school diploma (GED®), graduation date (out of high school at least five years or more), Pell eligibility, Industrial Technology programs, and Transportation and Logistics programs. Given all other variables are unchanged, the odds of a student being retained increases by a factor of 1.79 (odds ratio = 1.79) as GPA increases. Also, the odds of a student being retained when they receive the Pell grant increases by a factor of 1.27 (odds ratio = 1.27), given all other variables are unchanged.

To determine the strongest predictors of retention, variable importance was measured using the odds ratio. The variables of GPA (OR = 1.791, 95% CI = 1.627 to 1.975) and Pell eligibility (OR = 1.269, 95% CI = 1.158 to 1.392) were the most influential predictors of being retained or not. The weakest predictors of being retained or not were Industrial Technology programs (OR = 0.847, 95% CI = 0.746 to 0.962) and high school diploma (GED®) (OR = 0.789, 95% CI = 0.694 to 0.896).

Table 30

Variables Used to Predict Retention Utilizing Logistic Regression (Training Data)

Predictor	Log Odds	*SE*	*Z*	Pr(>|z|)		*OR*	*95% Confidence Interval* *Lower*	*Upper*
(Intercept)	-0.116	0.044	-2.642	0.008	**	0.890	0.817	0.970
age	0.075	0.047	1.618	0.106		1.078	0.984	1.182
gpa	0.583	0.049	11.791	$p < .001$	***	1.791	1.627	1.975
racecode_X2	-0.016	0.049	-0.325	0.745		0.984	0.895	1.083
racecode_X3	0.081	0.042	1.916	0.055		1.084	0.999	1.179
racecode_X4	0.119	0.043	2.741	0.006	**	1.126	1.036	1.228
gencode_X1	0.015	0.057	0.256	0.798		1.015	0.907	1.136
hsdipcode_X2	-0.237	0.065	-3.650	$p < .001$	***	0.789	0.694	0.896
hsdipcode_X3	-0.085	0.062	-1.369	0.171		0.919	0.814	1.037
gradcode_X1	-0.113	0.045	-2.498	0.012	*	0.893	0.817	0.976
sparcode_X1	-0.010	0.044	-0.233	0.816		0.990	0.908	1.079
dhomcode_X1	-0.074	0.047	-1.585	0.113		0.929	0.847	1.017
pellcode_X1	0.238	0.047	5.095	$p < .001$	***	1.269	1.158	1.392
hopepos_X2	0.020	0.054	0.372	0.710		1.020	0.918	1.133
hopepos_X3	-0.166	0.065	-2.557	0.011	*	0.847	0.746	0.962
hopepos_X4	0.129	0.048	2.688	0.007	**	1.138	1.036	1.251

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

Additionally, the finalized logistic regression model was applied to the test data for comparison. Table 31 shows the results from the logistic regression analysis. The overall model was found to be statistically significant, $\chi 2(15) = 355.54$, $p < 0.001$, and the model resulted in an error rate of 0.39. In addition, the McFadden pseudo R^2 value was calculated as 0.12 indicating the model can account for 12% of the retained variable. Of the 15 predictor variables, three were statistically significant as predictors of student retention: GPA, graduation date (out of high school at least five years or more), and Pell eligibility. Given all other variables are unchanged, the odds of a student being retained increases by a factor of 2.50 (odds ratio = 2.50) as GPA increases. If a student enrolls in

Industrial Technology programs, the odds of those students being retained decreases by 49% (odds ratio = 0.510 – 1), keeping other variables constant.

Table 31

Variables Used to Predict Retention Utilizing Logistic Regression (Test Data)

Predictor	Log Odds	*SE*	*Z*	Pr(>|z|)		*OR*	*95% Confidence Interval* *Lower*	*Upper*
(Intercept)	-0.204	0.050	-4.070	$p < .001$	***	0.816	0.739	0.899
age	0.047	0.051	0.939	0.348		1.049	0.950	1.158
gpa	0.918	0.059	15.503	$p < .001$	***	2.503	2.233	2.816
racecode_X2	-0.037	0.055	-0.661	0.509		0.964	0.865	1.075
racecode_X3	-0.029	0.050	-0.585	0.558		0.971	0.880	1.072
racecode_X4	0.080	0.048	1.662	0.097		1.083	0.987	1.192
gencode_X1	-0.053	0.062	-0.861	0.389		0.948	0.839	1.070
hsdipcode_X2	0.070	0.073	0.956	0.339		1.073	0.929	1.239
hsdipcode_X3	0.090	0.072	1.241	0.215		1.094	0.949	1.261
gradcode_X1	-0.187	0.051	-3.692	$p < .001$	***	0.829	0.750	0.916
sparcode_X1	0.057	0.049	1.163	0.245		1.058	0.962	1.165
dhomcode_X1	-0.062	0.048	-1.273	0.203		0.940	0.855	1.034
pellcode_X1	0.164	0.051	3.211	$p < .001$	***	1.179	1.066	1.303
hopepos_X2	0.033	0.060	0.553	0.580		1.034	0.919	1.163
hopepos_X3	-0.066	0.070	-0.944	0.345		0.936	0.816	1.074
hopepos_X4	-0.039	0.056	-0.692	0.489		0.962	0.862	1.073

Note. $p < 0.001$ '***', $p < 0.01$ '**', $p < 0.05$ '*'.

The variables of GPA (OR = 2.503, 95% CI = 2.233 to 2.816) and Pell eligibility (OR = 1.179, 95% CI = 1.066 to 1.303) were the most influential predictors of being retained or not in the test data. The weakest predictors of being retained or not were Industrial Technology programs (OR = 0.936, 95% CI = 0.816 to 1.074) and graduation date (out of high school at least five years or more) (OR = 0.829, 95% CI = 0.750 to 0.916). The variable importance plot including all 15 predictor variables is shown in Figure 18. Because variable importance for logistic regression is based on the absolute

values of the z-statistic, both the most influential and the least influential predictors may be displayed at the top of the plot.

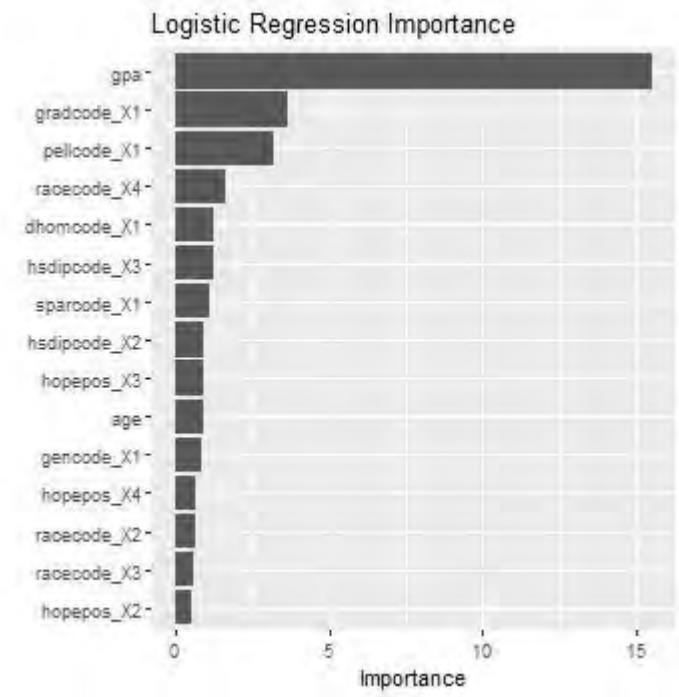

Figure 18. Logistic regression variable importance plot for diplomas 49 to 59 credit hours in length.

The second data modeling approach used to address Research Question 1D was linear discriminant analysis. The LDA model fit resulted in a training error rate of 0.39 and the results were similar to those of the logistic regression. As shown in Table 32, coefficients with the strongest associated weights included GPA (0.889), high school diploma (GED®) (-0.359), Pell eligibility (0.359), and Industrial Technology programs (-0.250). The larger the coefficient of a predictor in the standardized discriminant function, the more important its role in the discriminant function. Similar to all other data files, GPA was the strongest predictor of retention with a coefficient of 0.889 and Pell eligibility (0.359) and high school diploma (GED®) (-0.359) tied as the second most

influential predictor. Both females (0.021) and single parents (-0.017) had the least influential coefficients.

Table 32

Variables Used to Predict Retention Utilizing LDA (Training Data)

Independent Variable	Not Retained Mean	Retained Mean	Coefficients of Linear Discriminants: LD1
age	-0.018	0.011	0.114
gpa	-0.121	0.350	0.889
racecode_X2	0.023	-0.094	-0.025
racecode_X3	-0.019	0.082	0.123
racecode_X4	-0.047	0.137	0.174
gencode_X1	-0.020	0.057	0.021
hsdipcode_X2	0.030	-0.095	-0.359
hsdipcode_X3	-0.011	0.057	-0.126
gradcode_X1	0.025	-0.063	-0.173
sparcode_X1	0.011	0.013	-0.017
dhomcode_X1	0.006	-0.044	-0.108
pellcode_X1	-0.044	0.132	0.359
hopepos_X2	-0.026	0.071	0.030
hopepos_X3	0.045	-0.134	-0.250
hopepos_X4	-0.035	0.112	0.194

Note. Prior probabilities of groups: not retained: 0.5, retained: 0.5.

In comparison, Table 33 includes the coefficients of linear discriminants for the test data. The overall error rate for the model was 0.39 and the coefficients with the strongest associated weights included GPA (1.092), graduation date (out of high school at least five years or more) (-0.212), and Pell eligibility (0.189). The strongest predictor of being retained or not based on test data was GPA with a coefficient of 1.092 and graduation date (out of high school at least five years or more) with a negative coefficient of -0.212. The variables race (Hispanic) (-0.036) and Cyber, Engineer, or Healthcare programs (0.041) had the least influential coefficients of being retained or not. The

variable GPA had the largest variance within the group means. This variable has a greater influence on students not being retained (0.605) than on students being retained (-0.126). The variable importance plot including all 15 predictor variables is shown in Figure 19.

Table 33

Variables Used to Predict Retention Utilizing LDA (Test Data)

Independent Variable	Not Retained Mean	Retained Mean	Coefficients of Linear Discriminants: LD1
age	-0.027	0.087	0.060
gpa	-0.189	0.518	1.092
racecode_X2	0.063	-0.149	-0.046
racecode_X3	-0.023	0.004	-0.036
racecode_X4	-0.034	0.120	0.092
gencode_X1	-0.009	-0.004	-0.064
hsdipcode_X2	-0.002	0.093	0.069
hsdipcode_X3	-0.005	-0.017	0.077
gradcode_X1	0.034	-0.126	-0.212
sparcode_X1	-0.013	0.036	0.069
dhomcode_X1	0.008	0.008	-0.073
pellcode_X1	-0.032	0.044	0.189
hopepos_X2	-0.009	0.052	0.041
hopepos_X3	0.001	-0.026	-0.082
hopepos_X4	0.009	-0.012	-0.046

Note. Prior probabilities of groups: not retained: 0.5, retained: 0.5.

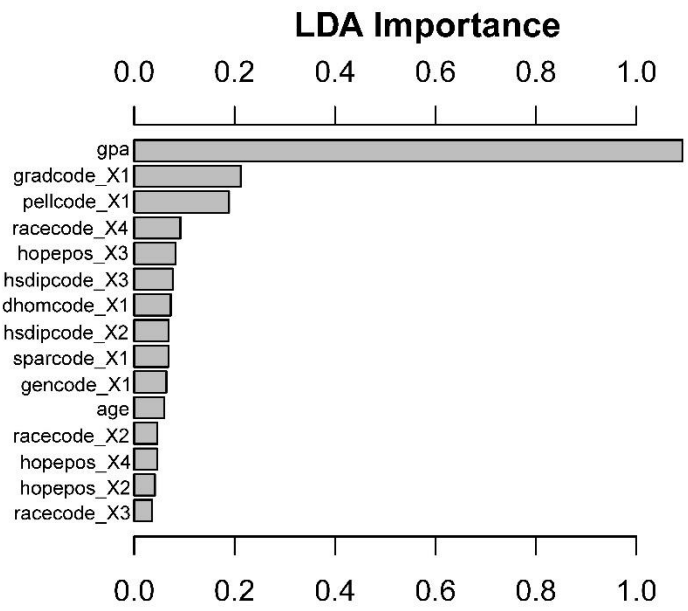

Figure 19. Linear discriminant analysis variable importance plot for diplomas 49 to 59 credit hours in length.

The first of the three data mining approaches, a classification tree, was used to answer Research Question 1D where all predictor variables were allowed to enter the model. The prior probabilities were specified as 0.50 for predicted class 1, students who were retained, and 0.50 for predicted class 0, students who were not retained. The classification tree model was evaluated using 10-fold cross-validation repeated five times using stratification and tuned for the parameters complexity and tree depth. The best parameters, based on the largest AUC metric, were used to select the optimal model. The optimal complexity factor (0.00233) and maximum tree depth (3) were used to update the model and refit the training data. The 10-fold cross-validation was used to obtain a cross-validated error rate where the lowest rate indicated the tree which best fit the data. The resulting model had 3 total splits and a cross-validated error rate of 0.73. The overall training error rate for the model was 0.45.

To determine the strongest predictors of retention, variable importance was measured as the sum of the goodness of split measures (Gini index). The variable GPA was by far the strongest predictor of being retained with a Gini index of 97.58. The variable Pell eligibility was a distant second with a Gini index of 14.68. The weakest predictors of retention were single parents ($I_G = 0.30$), graduation date (out of high school at least five years or more) ($I_G = 0.25$), and age ($I_G = 0.09$).

In comparison, the model applied to test data resulted in one split, a cross-validated error rate of 0.66, and an overall error rate of 0.45. The variable GPA ($I_G = 168.93$) was the most influential predictor of being retained or not. The variables graduation date (out of high school at least five years or more) ($I_G = 12.65$) and age ($I_G = 0.09$) were a distant second with age being the least influential predictor. The variable importance plot including three of the 15 predictor variables is shown in Figure 20. The remaining 12 predictor variables had importance scores below zero.

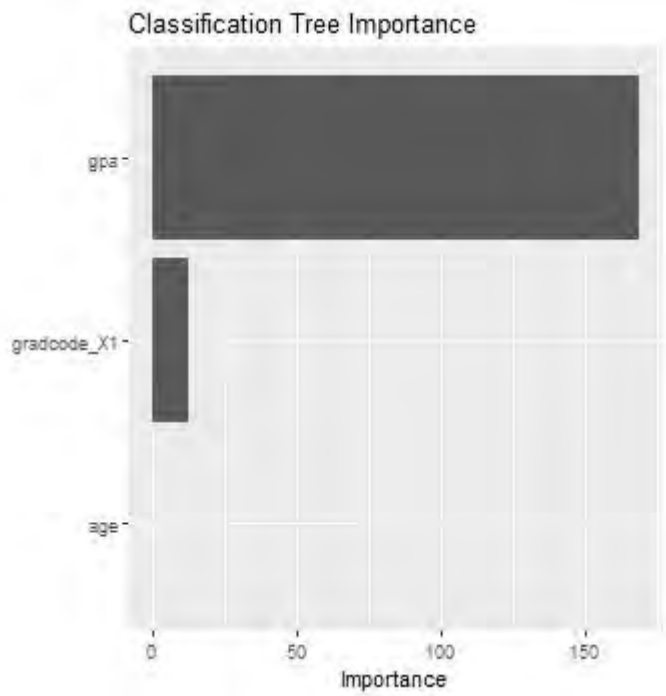

Figure 20. Classification tree variable importance plot for diplomas 49 to 59 credit hours in length.

The random forest model was the second of three data mining approaches performed to answer Research Question 1D. The prior probabilities were specified as 0.50 for predicted class 1, students who were retained, and 0.50 for predicted class 0, students who were not retained. The random forests model was evaluated using 10-fold cross-validation repeated five times using stratification. Tuned parameters, based on the largest AUC metric, were used to select the optimal model. The optimal mtry parameter (2) and minimum node size (31) were used to update the model and refit the training data. The 10-fold cross-validation was used to obtain a cross-validated error rate where the lowest rate indicated the tree which best fit the data. The resulting model had 500 trees with an out-of-bag (OOB) error rate of 33.75%, an error rate of 48.38% for class 0 (not retained), and an error rate of 19.12% for class 1 (retained). The overall training error rate for the model was 0.38.

The mean decrease in Gini was used to measure how important each variable was for estimating the value of the target variable across all of the trees that made up the forest. The mean decrease in Gini is the average (mean) of the variable's total decrease in node impurity, weighted by the proportion of samples reaching that node in each decision tree in the random forest. The most important variables to the model result in the largest mean decrease in Gini value. The variables of GPA (83.77), age (35.32), and Pell eligibility (11.22) were the strongest predictors of being retained. The weakest predictors of retention were single parents (4.85), race (Hispanic) (4.27), and displaced homemakers (3.34).

In comparison, the model applied to test data resulted in 500 trees with an out-of-bag (OOB) error rate of 31.11%, an error rate of 48.05% for class 0 (not retained), and an error rate of 14.18% for class 1 (retained). The overall error rate for the model was 0.44. The most important variables to the model result in the largest mean decrease in Gini value. The variables of GPA (119.51) and age (27.74) were the most influential predictors of being retained or not. The weakest predictors of retention were displaced homemakers (3.07) and Transportation and Logistics programs (2.64). The variable importance plot including all 15 predictor variables is shown in Figure 21.

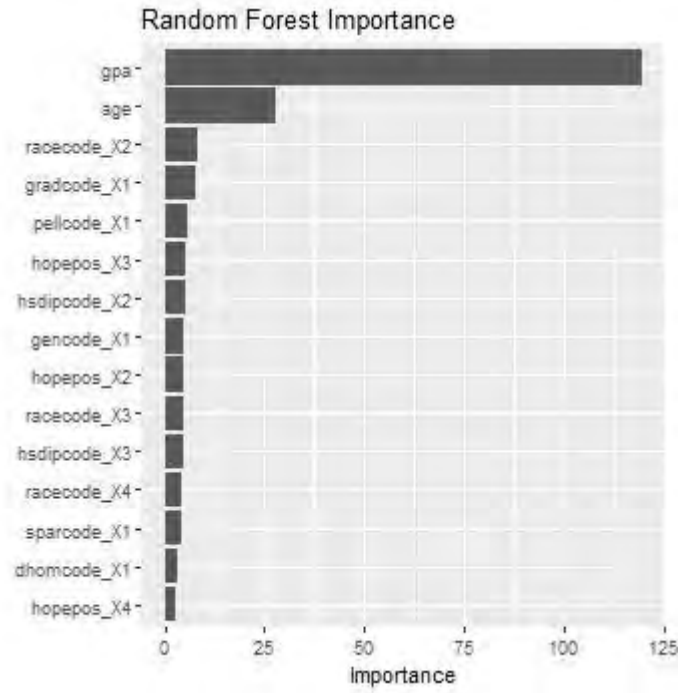

Figure 21. Random forests variable importance plot for diplomas 49 to 59 credit hours in length.

The support vector machine model was the final data mining approach performed to answer Research Question 1D. All predictor variables were allowed to enter the model. The prior probabilities were specified as 0.50 for predicted class 1, students who were retained, and 0.50 for predicted class 0, students who were not retained. The SVM model was evaluated using 10-fold cross-validation repeated five times using stratification and tuned for the parameters cost and rbf_sigma. For the training data set, the optimal cost (based on the largest AUC metric) was calculated to be 0.1 and the optimal rbf_sigma was calculated to be 0.1. These tuned parameters were used to update the model and refit the training data. The best model resulted in 2,195 support vectors, an objective function value of -198.18, and an error rate of 0.35. The overall training error rate for the model was 0.41.

Permutation-based variable importance scores were computed for each predictor in the SVM model. If a variable is important, the model's performance (based on the AUC metric) should change after permuting or rearranging the values of the variable. A larger change in the performance will indicate a more important variable. For the SVM model, the strongest predictors of being retained were GPA (0.072), Pell eligibility (0.031), and Transportation and Logistics programs (0.012). The weakest predictors of retention were high school diploma (college prep or tech prep) (0.003), race (Hispanic) (0.002), and high school diploma (GED®) (0.001).

In comparison, the model applied to test data resulted in 1,837 support vectors, an objective function value of -164.49, and an error rate of 0.30. The overall error rate for the model was 0.44 and the most influential predictors of being retained or not were GPA (0.080), Industrial Technology programs (0.010), and high school diploma (college prep or tech prep) (0.006). The least influential predictors of being retained or not were age (-0.003), race (Hispanic) (-0.008), and Cyber, Engineer, or Healthcare programs (-0.009). The variable importance plot including all 15 predictor variables is shown in Figure 22.

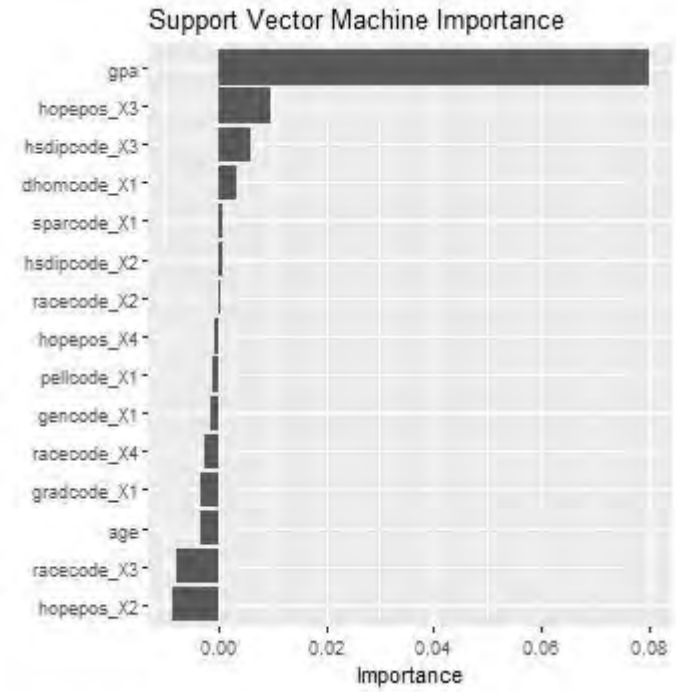

Figure 22. Support vector machine variable importance plot for diplomas 49 to 59 credit hours in length.

For diplomas 49–59 credit hours in length, environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) were analyzed to determine which, if any, were significant predictors of nontraditional student retention. Using test data, the logistic regression model and the linear discriminant model shared the lowest error rate of 0.39. The highest error rate using test data was the classification tree model at 0.45. Both data modeling approaches, logistic regression and linear discriminant analysis, shared similar results in regards to variable importance. The predictor variables GPA (an academic integration component) and Pell eligibility (an environmental factor) were the most influential in students being retained. Both the logistic regression and linear discriminant analysis models indicated that being out of high school for five years

or more is an influential indicator of students not being retained. Each of the three data mining approaches (classification trees, random forests, and support vector machines) identified similar predictor variables. The variables GPA, age, and Industrial Technology programs were the most influential predictors of being retained. The weakest predictors of retention were displaced homemakers, Transportation and Logistics programs, and Cyber, Engineer, or Healthcare programs.

Model Comparisons for Research Question 1

Five statistical models were utilized to answer each of the four subsections of Research Question 1 representing certificates 9–17 credit hours in length, certificates 18–36 credit hours in length, diplomas 37–48 credit hours in length, and diplomas 49–59 credit hours in length. The 2017-2018 data were used as the training data and the 2018-2019 data were used as the test data. During model training, upsampling was used to mitigate the effects of class imbalance in the outcome variable retained. In both the 2017 and 2018 cohorts of certificates 9–17 credit hours in length, the class imbalance was the greatest with the rate of students not retained only accounting for 17.93% and 17.72% respectively. Each model was evaluated using 10-fold cross-validation repeated five times with stratification.

Once the models were applied to the test data, the logistic regression and linear discriminant analysis models produced the lowest error rates in three of the four data files. The random forest model matched the logistic regression and linear discriminant analysis models in the two data files representing certificate programs. For the logistic regression and linear discriminant analysis models, the predictor variables GPA and programs related to Transportation and Logistics were the most influential in students

being retained across both data files representing certificate programs. Both of these variables represented academic integration components. However, results for diploma programs differed slightly. For diplomas 37–48 credit hours in length, the predictor variables GPA, female students, and Black students were the most influential in students being retained, representing background factors (gender and race) and an academic integration component (GPA). This was true across both data modeling approaches, logistic regression, and linear discriminant analysis. For diplomas 49–59 credit hours in length, GPA (an academic integration component) and Pell eligibility (an environmental factor) were the most influential in students being retained. Across each of the four data files, the logistic regression and linear discriminant analysis models shared similar results for the most influential predictors of students not being retained. Being out of high school for five years or more and being enrolled in a Cyber, Engineer, or Healthcare program are influential predictors of students not being retained. One of the certificate data files indicated Industrial Technology programs as influential predictors of students not being retained.

GPA (an academic integration component) was the most influential predictor across each data file in each of the three data mining approaches (classification trees, random forests, and support vector machines). Results for other influential predictors were mixed. Age (a background factor) was the second most influential predictor. However, in both certificate data files, the support vector machines model did not identify age as one of the top predictors. Other influential predictors included Transportation and Logistics programs or Industrial Technology programs. Overall, GPA

(an academic integration component) was the most influential predictor across each data file and each data model.

Accuracy of the Classification Models

Research Question 2. Does one of the selected statistical procedures generate a more accurate classification model based on Cohen's Kappa, ROC curves, and sensitivity and specificity by certificate or diploma type?

For each statistical procedure Cohen's Kappa, ROC curves, sensitivity, and specificity were used to identify the most accurate classification model at predicting nontraditional student retention. The effectiveness of machine learning lies in its ability to make good predictions on unknown data by learned data models. Thus, the goal of predictive modeling is to create a model which performs best with new unknown data. To take advantage of the generalizing power of the model, data are partitioned into training and test sets. The training data is used to build the model and the test data is used to estimate the model's predictive performance. A confusion matrix produces a table of actual and predicted values for the test data and associated statistics which represent the model's predictive performance. As described in Table 34, the actual and predicted values are classified as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Table 34

Confusion Matrix

	Actual Not Retained	Actual Retained
Predicted Not Retained	TN	FN
Predicted Retained	FP	TP

True positives include cases where the actual value was true and the model also predicted it as true. Similarly, a true negative is where the actual value was false and the model also predicted it as false. False negatives include cases where the actual value was true (retained) and the model predicted it as false (not retained). A false positive is where the actual value was false (not retained) and the model predicted it as true (retained). With a specific focus on the retention of nontraditional students in diploma and certificate programs, the outcome of this research will help colleges develop policies and procedures to facilitate student retention, and also align Georgia's nontraditional students with Georgia's workforce needs. That is, the goal would be to correctly identify students who will not be retained so adequate assistance and resources can be provided to them. Therefore, higher true negatives rates will rank higher and false positives will be considered costlier and should be minimized if possible. However, it's often necessary to consider multiple evaluation metrics when comparing various models.

The accuracy of a model measures how many observations, both positive and negative, were correctly classified. It is calculated as the ratio of the sum of true positives and true negatives to the total number of predictions. Classification accuracy can easily be turned into a misclassification rate or error rate by inverting the accuracy value. When data has an uneven number of classes, the accuracy metric can be misleading. However, because each data file used to train each model was upsampled to mitigate the effects of class imbalance, the accuracy of the model and its equivalent error rate are important metrics to evaluate and consider. Cohen's Kappa is a measure of agreement or inter-rater reliability for categorical variables where there are two raters. Kappa takes into account the accuracy generated simply by chance using an observed accuracy and an expected

accuracy based on the marginal totals of a confusion matrix (Kuhn & Johnson, 2013). The Kappa statistic can take on values between −1 and 1 where a value of 0 means there is no agreement between the observed and predicted classes, and a value of 1 indicates perfect agreement between the model prediction and the observed classes. A common interpretation of the Kappa statistic is as follows: values ≤ 0 indicate no agreement and 0.01–0.20 as none to poor, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement. Sensitivity measures the percentage of cases in which retention is predicted correctly. This metric describes how sensitive the model is when predicting positive cases. Sensitivity is the true positive rate, also called the recall, and is calculated as the ratio of true positives (students correctly predicted as retained) to the sum of true positives and false negatives (the actual number of retained students). Conversely, specificity refers to the percentage of cases in which not being retained or attrition is predicted correctly. It describes how accurate the model is when predicting negative cases. Specificity is also called the true negative rate and is calculated as the ratio of true negatives (students correctly predicted as not retained) to the sum of true negatives and false positives (the actual number of students not retained). The F1 score is the calculated mean of the model's precision (the ratio of the students correctly predicted as retained to all students predicted as retained) and recall. It is also known as F-measure or balanced F-score. The F1 score can be interpreted where an F1 score reaches its best value at 1 and worst value at 0.

In addition to the performance metrics of the confusion matrix, the ROC curve is a commonly used method to visualize the performance of a binary classier for different thresholds. ROC curves summarize the trade-off between the true positive rate and false

positive rate for a predictive model using different probability thresholds. For each threshold, the resulting true positive rate (sensitivity) and the false positive rate (1-specificity) are plotted against each other. The optimal model should be shifted towards the upper left corner of the plot. Alternatively, the model with the largest area under the ROC curve (AUC) would be the most effective. Estimates of the AUC indicate the overall performance of a classifier summarized over all possible thresholds (James et al., 2013).

The confusion matrix for certificates 9 to 17 credit hours used to predict retention using logistic regression is shown in Table 35. The true positive rate (sensitivity) was 0.63 and the false positive rate (1 – specificity) was 0.23. The true negative rate (specificity) was 0.77 and the false negative rate (type II error) was 0.37 with an overall error rate of 0.34.

Table 35

Confusion Matrix for Variables Used to Predict Retention Utilizing Logistic Regression (Certificates 9 to 17 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	172	380
Predicted Retained	52	660

The confusion matrix for certificates 9 to 17 credit hours used to predict retention using linear discriminant analysis is shown in Table 36. The results were similar to those of the logistic regression. The true positive rate (sensitivity) was 0.64 and the false positive rate (1 – specificity) was 0.24. The true negative rate (specificity) was 0.76 and the false negative rate (type II error) was 0.36 with an overall error rate of 0.34.

Table 36

Confusion Matrix for Variables Used to Predict Retention Utilizing Linear Discriminant Analysis (Certificates 9 to 17 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	171	379
Predicted Retained	53	661

The confusion matrix for certificates 9 to 17 credit hours used to predict retention using a classification tree is shown in Table 37. The true positive rate (sensitivity) was 0.62 and the false positive rate (1 – specificity) was 0.37. The true negative rate (specificity) was 0.63 and the false negative rate (type II error) was 0.38 with an overall error rate of 0.38.

Table 37

Confusion Matrix for Variables Used to Predict Retention Utilizing a Classification Tree (Certificates 9 to 17 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	141	399
Predicted Retained	83	641

The confusion matrix for certificates 9 to 17 credit hours used to predict retention using random forests is shown in Table 38. The true positive rate (sensitivity) was 0.68 and the false positive rate (1 – specificity) was 0.43. The true negative rate (specificity) was 0.57 and the false negative rate (type II error) was 0.32 with an overall error rate of 0.34.

Table 38

Confusion Matrix for Variables Used to Predict Retention Utilizing Random Forests (Certificates 9 to 17 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	127	332
Predicted Retained	97	708

The confusion matrix for certificates 9 to 17 credit hours used to predict retention using a support vector machine is shown in Table 39. The true positive rate (sensitivity) was 0.49 and the false positive rate (1 – specificity) was 0.19. The true negative rate (specificity) was 0.81 and the false negative rate (type II error) was 0.51 with an overall error rate of 0.45.

Table 39

Confusion Matrix for Variables Used to Predict Retention Utilizing a Support Vector Machine (Certificates 9 to 17 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	182	527
Predicted Retained	42	513

The plot of the ROC curve for each of the five models using test data for certificates 9 to 17 credit hours in length is shown in Figure 23 followed by Table 40 which includes key metrics from the confusion matrix associated with each classification model. The logistic regression and the linear discriminant analysis models had the highest area under the ROC curve. The linear discriminant analysis model was slightly higher with a fair AUC of 0.747. Of the five models, four performed similarly in terms of accuracy, error rate, and F1 score. The logistic regression, linear discriminant analysis,

classification tree, and random forest models produced good F1 scores within 0.73 to 0.77. The logistic regression, linear discriminant analysis, and support vector machine had higher specificity than sensitivity and the classification tree had almost identical specificity and sensitivity rates. The random forest model was the only model in this cohort with a higher true positive rate (0.68) as compared to the true negative rate (0.57). However, of the four test data sets, this data set had more retained cases (82.28%) than not retained cases (17.72%) indicating sensitivity was estimated with greater precision than the specificity. Precision, the ratio of the students correctly predicted as retained to all students predicted as retained, ranged from 0.88 to 0.93 for this cohort. Kappa coefficients ranged from poor to fair with the logistic regression and linear discriminant analysis having the largest values at 0.26 and 0.25, respectively. The support vector machine model had the lowest accuracy and F1 score, and the second lowest Kappa coefficient at 0.17, but had the largest specificity rate of all the models in this cohort (0.81). The higher specificity is likely a result of the class imbalance in the data. Overall, logistic regression and linear discriminant analysis performed well across metrics and may generate a more accurate classification model. Between the two models, most classification metrics were identical or virtually identical with sensitivity and specificity rates having the most variance. However, with its high true negative rate and low false positive rate, the support vector machine will generate an equally accurate classification model based on the goal of correctly identifying students who will not be retained. Therefore, of the five classification models for certificates 9 to 17 credit hours in length, the support vector machine model will generate a more accurate classification model based on specificity.

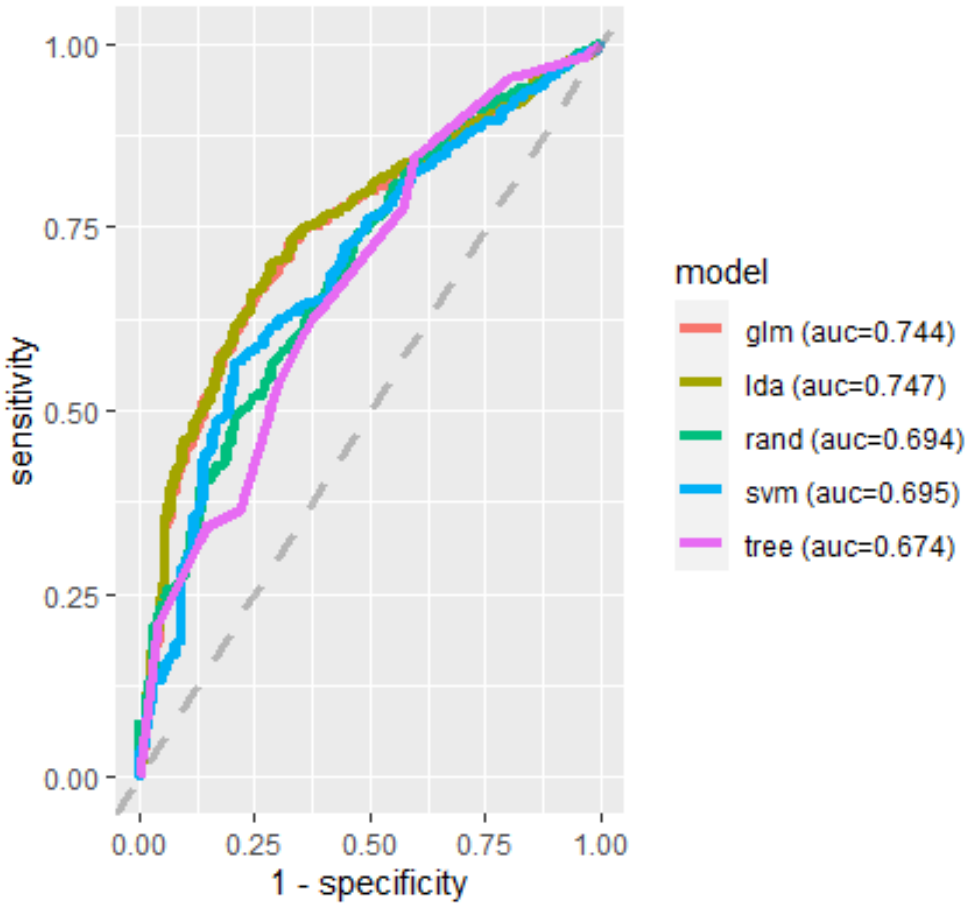

Figure 23. ROC curve results for certificates 9 to 17 credit hours in length used to predict retention utilizing five data models.

Table 40

Prediction Models for Certificates 9 to 17 Credit Hours Using Test Data

Model	Accuracy	Error Rate	Cohen's Kappa	Sensitivity	Specificity	F1
GLM	.66	.34	.26	.63	.77	.75
LDA	.66	.34	.25	.64	.76	.75
Tree	.62	.38	.16	.62	.63	.73
Rand	.66	.34	.18	.68	.57	.77
SVM	.55	.45	.17	.49	.81	.64

The confusion matrix for certificates 18 to 36 credit hours used to predict retention using logistic regression is shown in Table 41. The true positive rate (sensitivity) was 0.74 and the false positive rate (1 – specificity) was 0.48. The true negative rate (specificity) was 0.52 and the false negative rate (type II error) was 0.26 with an overall error rate of 0.39.

Table 41

Confusion Matrix for Variables Used to Predict Retention Utilizing Logistic Regression (Certificates 18 to 36 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	367	126
Predicted Retained	340	350

The confusion matrix for certificates 18 to 36 credit hours used to predict retention using linear discriminant analysis is shown in Table 42. The true positive rate (sensitivity) was 0.74 and the false positive rate (1 – specificity) was 0.49. The true negative rate (specificity) was 0.51 and the false negative rate (type II error) was 0.26 with an overall error rate of 0.39.

Table 42

Confusion Matrix for Variables Used to Predict Retention Utilizing Linear Discriminant Analysis (Certificates 18 to 36 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	364	124
Predicted Retained	343	352

The confusion matrix for certificates 18 to 36 credit hours used to predict retention using a classification tree is shown in Table 43. The true positive rate

(sensitivity) was 0.90 and the false positive rate (1 – specificity) was 0.62. The true negative rate (specificity) was 0.38 and the false negative rate (type II error) was 0.11 with an overall error rate of 0.41.

Table 43

Confusion Matrix for Variables Used to Predict Retention Utilizing a Classification Tree (Certificates 18 to 36 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	269	50
Predicted Retained	438	426

The confusion matrix for certificates 18 to 36 credit hours used to predict retention using random forests is shown in Table 44. The true positive rate (sensitivity) was 0.73 and the false positive rate (1 – specificity) was 0.47. The true negative rate (specificity) was 0.53 and the false negative rate (type II error) was 0.27 with an overall error rate of 0.39.

Table 44

Confusion Matrix for Variables Used to Predict Retention Utilizing Random Forests (Certificates 18 to 36 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	372	128
Predicted Retained	335	348

The confusion matrix for certificates 18 to 36 credit hours used to predict retention using a support vector machine is shown in Table 45. The true positive rate (sensitivity) was 0.86 and the false positive rate (1 – specificity) was 0.62. The true

negative rate (specificity) was 0.38 and the false negative rate (type II error) was 0.14 with an overall error rate of 0.43.

Table 45

Confusion Matrix for Variables Used to Predict Retention Utilizing a Support Vector Machine (Certificates 18 to 36 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	267	66
Predicted Retained	440	410

The plot of the ROC curve for each of the five models using test data for certificates 18 to 36 credit hours in length is shown in Figure 24 followed by Table 46 which includes key metrics from the confusion matrix associated with each classification model. The logistic regression and the linear discriminant analysis models had the highest area under the ROC curve. Although it was a relatively poor AUC, the linear discriminant analysis model had a slightly higher AUC between the two models with 0.674. For this cohort, all models performed similarly in terms of accuracy, error rate, Kappa, and F1 score. Kappa coefficients indicated fair agreement ranging from 0.21 to 0.24 while F1 scores ranged from 0.60 to 0.64. Similar to the previous cohort of certificates 9 to 17 credit hours, the support vector machine model had the lowest accuracy (0.57) and lowest Kappa (0.21). All five models had higher sensitivity than specificity with the greatest difference being the classification tree model with a sensitivity of 0.90 and specificity of 0.38 which produced the highest F1 score of 0.64. The second highest F1 score of 0.62 belonged to the support vector machine model which also had the second highest sensitivity of 0.86. Although the logistic regression and linear discriminant analysis had the highest AUC, the random forest had the highest specificity. Each of these three

models had the same accuracy. Therefore, the random forest will generate a slightly more accurate classification model based on specificity for certificates 18 to 36 credit hours in length.

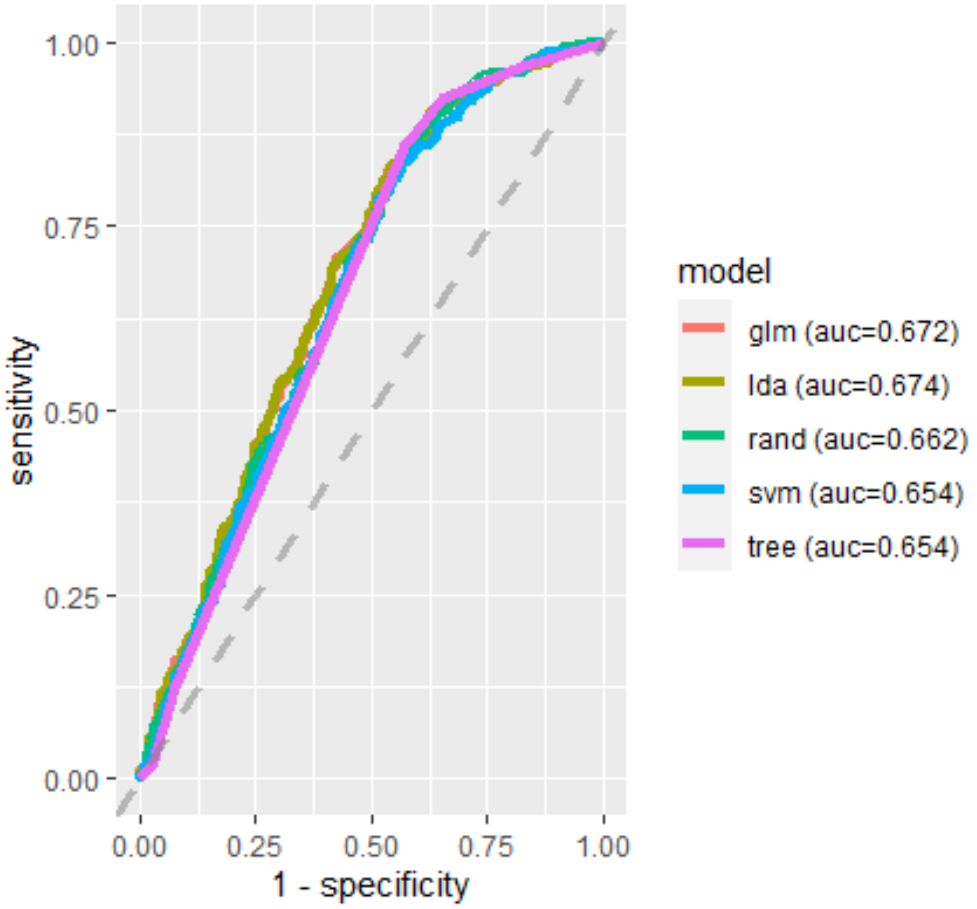

Figure 24. ROC curve results for certificates 18 to 36 credit hours in length used to predict retention utilizing five data models.

Table 46

Prediction Models for Certificates 18 to 36 Credit Hours Using Test Data

Model	Accuracy	Error Rate	Cohen's Kappa	Sensitivity	Specificity	F1
GLM	.61	.39	.24	.74	.52	.60
LDA	.61	.39	.24	.74	.51	.60
Tree	.59	.41	.24	.90	.38	.64
Rand	.61	.39	.24	.73	.53	.60
SVM	.57	.43	.21	.86	.38	.62

The confusion matrix for diplomas 37 to 48 credit hours used to predict retention using logistic regression is shown in Table 47. The true positive rate (sensitivity) was 0.66 and the false positive rate (1 – specificity) was 0.47. The true negative rate (specificity) was 0.53 and the false negative rate (type II error) was 0.34 with an overall error rate of 0.43.

Table 47

Confusion Matrix for Variables Used to Predict Retention Utilizing Logistic Regression (Diplomas 37 to 48 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	248	67
Predicted Retained	224	131

The confusion matrix for diplomas 37 to 48 credit hours used to predict retention using linear discriminant analysis is shown in Table 48. The true positive rate (sensitivity) was 0.68 and the false positive rate (1 – specificity) was 0.48. The true negative rate (specificity) was 0.52 and the false negative rate (type II error) was 0.32 with an overall error rate of 0.44.

Table 48

Confusion Matrix for Variables Used to Predict Retention Utilizing Linear Discriminant Analysis (Diplomas 37 to 48 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	244	64
Predicted Retained	228	134

The confusion matrix for diplomas 37 to 48 credit hours used to predict retention using a classification tree is shown in Table 49. The true positive rate (sensitivity) was 0.75 and the false positive rate (1 – specificity) was 0.60. The true negative rate (specificity) was 0.40 and the false negative rate (type II error) was 0.25 with an overall error rate of 0.50.

Table 49

Confusion Matrix for Variables Used to Predict Retention Utilizing a Classification Tree (Diplomas 37 to 48 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	190	50
Predicted Retained	282	148

The confusion matrix for diplomas 37 to 48 credit hours used to predict retention using random forests is shown in Table 50. The true positive rate (sensitivity) was 0.54 and the false positive rate (1 – specificity) was 0.44. The true negative rate (specificity) was 0.56 and the false negative rate (type II error) was 0.46 with an overall error rate of 0.44.

Table 50

Confusion Matrix for Variables Used to Predict Retention Utilizing Random Forests (Diplomas 37 to 48 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	265	91
Predicted Retained	207	107

The confusion matrix for diplomas 37 to 48 credit hours used to predict retention using a support vector machine is shown in Table 51. The true positive rate (sensitivity) was 0.62 and the false positive rate (1 – specificity) was 0.43. The true negative rate (specificity) was 0.57 and the false negative rate (type II error) was 0.38 with an overall error rate of 0.42.

Table 51

Confusion Matrix for Variables Used to Predict Retention Utilizing a Support Vector Machine (Diplomas 37 to 48 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	267	76
Predicted Retained	205	122

The plot of the ROC curve for each of the five models using test data for diplomas 37 to 48 credit hours in length is shown in Figure 25 followed by Table 52 which includes key metrics from the confusion matrix associated with each classification model. Of the five models, three had the highest AUCs: logistic regression, linear discriminant analysis, and support vector machine. Although a poor AUC, the linear discriminant analysis model had a slightly higher AUC of 0.634. For this cohort, all models performed poorly. One possible reason may be the small sample size (670 records) with the retained

class representing 29.55% (198). While this meets the minimum sample size based on $N = 10 k / p$ where p is the smallest of the proportions of cases in the population and k the number of covariates (the number of independent variables), the small size of the test set may not have sufficient power or precision to make reasonable judgments between the two classes. The accuracy of each model ranged from 0.50 to 0.58 and the Kappa coefficients were dismal and indicated poor agreement ranging from 0.09 to 0.16. F1 scores for all models were below 0.50 ranging from 0.42 to 0.48 indicating poor precision and poor recall. The classification tree had the lowest accuracy (0.50) and second lowest Kappa (0.11). Four of the five models had slightly higher sensitivity rates compared to the specificity rates. The classification tree was the exception with a sensitivity of 0.75 and specificity of 0.40. However, the classification tree's high sensitivity was not enough to compensate for the poor accuracy and Kappa coefficient. Overall, the logistic regression, linear discriminant analysis, and support vector machine each performed almost identical across all performance metrics. Of the three, the linear discriminant analysis model had a slightly higher F1 score and AUC. However, the support vector machine had the highest accuracy and specificity. Therefore, the support vector machine will generate a more accurate classification model based on specificity for diplomas 37 to 48 credit hours in length.

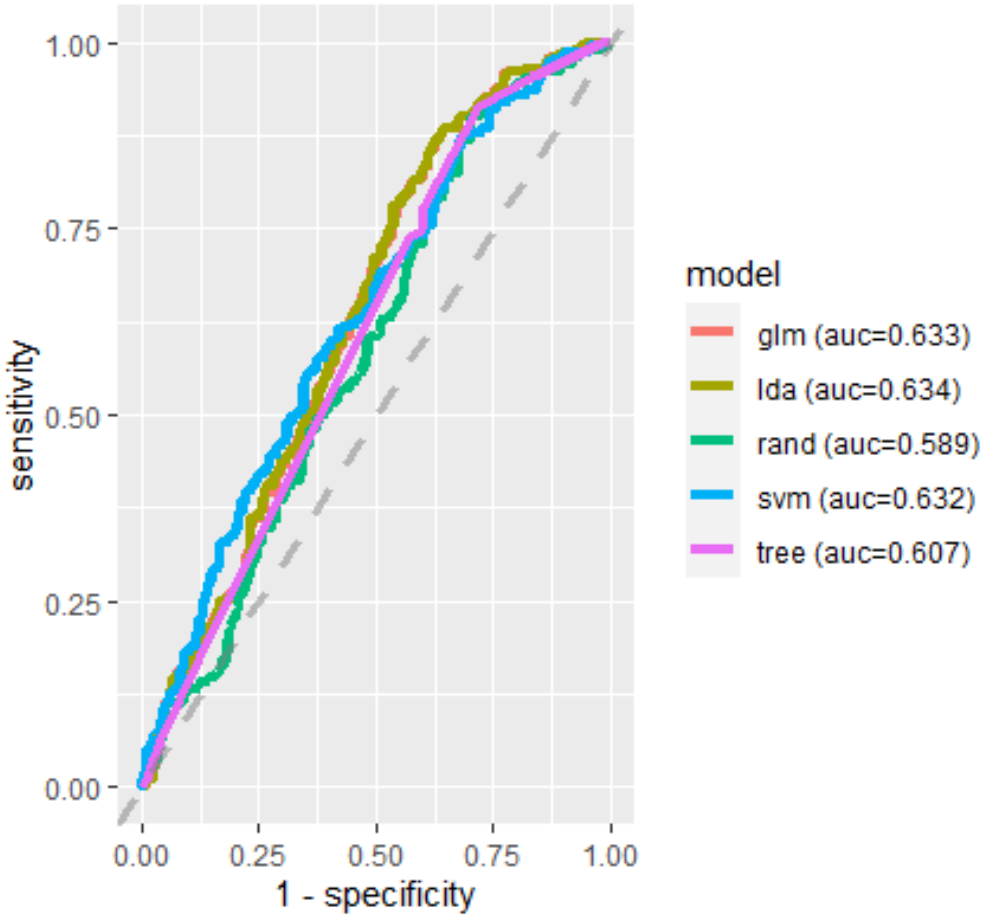

Figure 25. ROC curve results for diplomas 37 to 48 credit hours in length used to predict retention utilizing five data models.

Table 52

Prediction Models for Diplomas 37 to 48 Credit Hours Using Test Data

Model	Accuracy	Error Rate	Cohen's Kappa	Sensitivity	Specificity	F1
GLM	.57	.43	.15	.66	.53	.47
LDA	.56	.44	.16	.68	.52	.48
Tree	.50	.50	.11	.75	.40	.47
Rand	.56	.44	.09	.54	.56	.42
SVM	.58	.42	.15	.62	.57	.46

The confusion matrix for diplomas 49 to 59 credit hours used to predict retention using logistic regression is shown in Table 53. The true positive rate (sensitivity) was

0.72 and the false positive rate (1 – specificity) was 0.43. The true negative rate (specificity) was 0.57 and the false negative rate (type II error) was 0.28 with an overall error rate of 0.39.

Table 53

Confusion Matrix for Variables Used to Predict Retention Utilizing Logistic Regression (Diplomas 49 to 59 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	600	114
Predicted Retained	451	291

The confusion matrix for diplomas 49 to 59 credit hours used to predict retention using linear discriminant analysis is shown in Table 54. The true positive rate (sensitivity) was 0.73 and the false positive rate (1 – specificity) was 0.44. The true negative rate (specificity) was 0.56 and the false negative rate (type II error) was 0.27 with an overall error rate of 0.39.

Table 54

Confusion Matrix for Variables Used to Predict Retention Utilizing Linear Discriminant Analysis (Diplomas 49 to 59 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	593	110
Predicted Retained	458	295

The confusion matrix for diplomas 49 to 59 credit hours used to predict retention using a classification tree is shown in Table 55. The true positive rate (sensitivity) was 0.83 and the false positive rate (1 – specificity) was 0.56. The true negative rate

(specificity) was 0.44 and the false negative rate (type II error) was 0.17 with an overall error rate of 0.45.

Table 55

Confusion Matrix for Variables Used to Predict Retention Utilizing a Classification Tree (Diplomas 49 to 59 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	462	67
Predicted Retained	589	338

The confusion matrix for diplomas 49 to 59 credit hours used to predict retention using random forests is shown in Table 56. The true positive rate (sensitivity) was 0.73 and the false positive rate (1 – specificity) was 0.51. The true negative rate (specificity) was 0.49 and the false negative rate (type II error) was 0.27 with an overall error rate of 0.44.

Table 56

Confusion Matrix for Variables Used to Predict Retention Utilizing Random Forests (Diplomas 49 to 59 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	513	108
Predicted Retained	538	297

The confusion matrix for diplomas 49 to 59 credit hours used to predict retention using a support vector machine is shown in Table 57 and the results are similar to those of the random forests. The true positive rate (sensitivity) was 0.74 and the false positive rate (1 – specificity) was 0.51. The true negative rate (specificity) was 0.49 and the false negative rate (type II error) was 0.26 with an overall error rate of 0.44.

Table 57

Confusion Matrix for Variables Used to Predict Retention Utilizing a Support Vector Machine (Diplomas 49 to 59 Credit Hours)

	Actual Not Retained	Actual Retained
Predicted Not Retained	513	104
Predicted Retained	538	301

The plot of the ROC curve for each of the five models using test data for diplomas 49 to 59 credit hours in length is shown in Figure 26 followed by Table 58 which includes key metrics from the confusion matrix associated with each classification model. Similar to previous cohorts, the models for logistic regression and linear discriminant analysis had the highest AUC both at a poor 0.690. Of the five models, three performed similarly in terms of accuracy, Kappa, and F1 score. The logistic regression, linear discriminant analysis, and classification tree models produced poor accuracy rates within 0.55 to 0.61 and fair Kappa coefficients ranging from 0.20 to 0.23. All three of these models had the same F1 score of 0.51. All five models in this cohort had higher sensitivity rates with the classification tree having the highest sensitivity at 0.83. Although the random forest model and the support vector machine model had a slightly higher accuracy (0.56) than the classification tree (0.55), both had poor Kappa coefficients (Kappa = 0.17) and F1 scores (0.48). Similar to the cohort for diplomas 37 to 48 credit hours, all models for diplomas 49 to 59 credit hours performed poorly. Overall, logistic regression and linear discriminant analysis performed the best based on accuracy and AUC. Between the two models, almost all classification metrics were identical with sensitivity and specificity rates having a slight variance. However, with its higher

specificity, the logistic regression model may generate a slightly more accurate classification model for diplomas 49 to 59 credit hours in length.

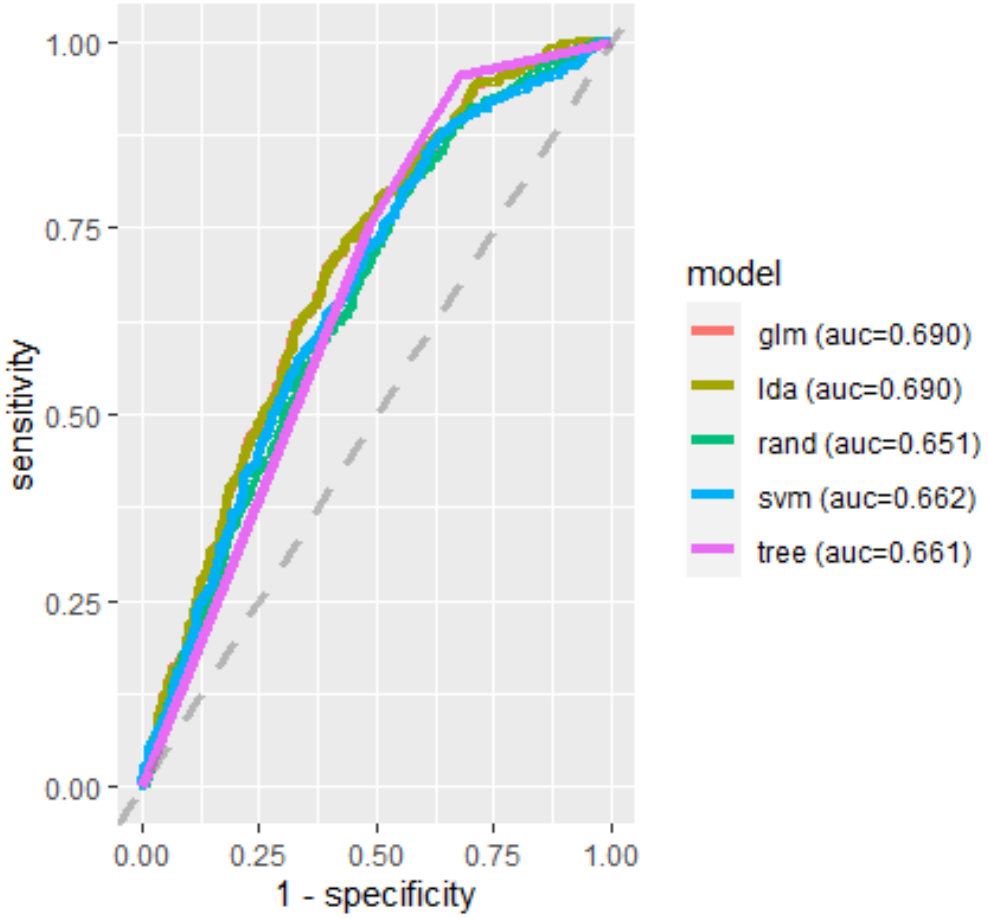

Figure 26. ROC curve results for diplomas 49 to 59 credit hours in length used to predict retention utilizing five data models.

Table 58

Prediction Models for Diplomas 49 to 59 Credit Hours Using Test Data

Model	Accuracy	Error Rate	Cohen's Kappa	Sensitivity	Specificity	F1
GLM	.61	.39	.23	.72	.57	.51
LDA	.61	.39	.23	.73	.56	.51
Tree	.55	.45	.20	.83	.44	.51
Rand	.56	.44	.17	.73	.49	.48
SVM	.56	.44	.17	.74	.49	.48

Model Comparisons for Research Question 2

Five statistical models were utilized to answer Research Question 2 representing certificates 9–17 credit hours in length, certificates 18–36 credit hours in length, diplomas 37–48 credit hours in length, and diplomas 49–59 credit hours in length. For each statistical procedure Cohen's Kappa, ROC curves, sensitivity, and specificity were used to identify the most accurate classification model at predicting nontraditional student retention. The effectiveness of machine learning lies in its ability to make good predictions on unknown data by learned data models. Thus, the goal of predictive modeling is to create a model which performs best with new unknown data. To take advantage of the generalizing power of the model, data were partitioned into training and test sets. The training data was used to build the model and the test data was used to estimate the model's predictive performance. A confusion matrix produces a table of actual and predicted values for the test data and associated statistics which represent the model's predictive performance.

For certificates 9 to 17 credit hours in length, the model with the lowest false positive rate was the support vector machine at 0.19. The logistic regression and the linear discriminant analysis models had the highest area under the ROC curve. Of the five models, four performed similarly in terms of accuracy, error rate, and F1 score. The logistic regression, linear discriminant analysis, and support vector machine had higher specificity than sensitivity and the classification tree had almost identical specificity and sensitivity rates. The random forest model was the only model in this cohort with a higher true positive rate (0.68) as compared to the true negative rate (0.57). The support vector machine model had the lowest accuracy and F1 score, and the second lowest Kappa

coefficient at 0.17, but had the largest specificity rate of all the models in this cohort (0.81). Overall, logistic regression and linear discriminant analysis performed well across metrics. However, the support vector machine will generate an accurate classification model based on the goal of correctly identifying students who will not be retained. Of the five classification models for certificates 9 to 17 credit hours in length, the support vector machine model will generate a more accurate classification model based on specificity.

For certificates 18 to 36 credit hours in length the logistic regression and the linear discriminant analysis models had the highest area under the ROC curve. All models performed similarly in terms of accuracy, error rate, Kappa, and F1 score. Similar to the previous cohort of certificates 9 to 17 credit hours, the support vector machine model had the lowest accuracy (0.57) and lowest Kappa (0.21). All five models had higher sensitivity than specificity with the greatest difference being the classification tree model with a sensitivity of 0.90 and specificity of 0.38 which produced the highest F1 score of 0.64. Although the logistic regression and linear discriminant analysis had the highest AUC, the random forest had the highest specificity. Each of these three models had the same accuracy. Therefore, the random forest model will generate a more accurate classification model based on specificity for certificates 18 to 36 credit hours in length.

For diplomas 37 to 48 credit hours in length, all models performed poorly. Of the five models, three had the highest AUCs: logistic regression, linear discriminant analysis, and support vector machine. The accuracy of each model ranged from 0.50 to 0.58 and the Kappa coefficients were dismal and indicated poor agreement ranging from 0.09 to 0.16. Four of the five models had slightly higher sensitivity rates compared to the specificity rates. Overall, the logistic regression, linear discriminant analysis, and support

vector machine each performed almost identical across all performance metrics. Of the three, the linear discriminant analysis model had a slightly higher F1 score and AUC. However, the support vector machine had the highest accuracy and specificity. Therefore, the support vector machine will generate a more accurate classification model based on specificity for diplomas 37 to 48 credit hours in length.

For diplomas 49 to 59 credit hours, the models for logistic regression and linear discriminant analysis had the highest AUC both at a poor 0.690. Of the five models, three performed similarly in terms of accuracy, Kappa, and F1 score. All five models in this cohort had higher sensitivity rates with the classification tree having the highest sensitivity at 0.83. Although the random forest model and the support vector machine model had a slightly higher accuracy (0.56) than the classification tree (0.55), both had poor Kappa coefficients (Kappa = 0.17) and F1 scores (0.48). Overall, logistic regression and linear discriminant analysis performed the best based on accuracy and AUC. Between the two models, almost all classification metrics were identical with sensitivity and specificity rates having a slight variance. However, with its higher specificity, the logistic regression model may generate a slightly more accurate classification model for diplomas 49 to 59 credit hours in length.

Overall, for each of the four cohorts and each of the five classification models, the support vector machine would generate the most accurate classification model based on the goal of correctly identifying students who will not be retained so adequate assistance and resources can be provided to them.

Summary

The purpose of the study was to identify the significant predictors of nontraditional student retention for certificates 9–17 and 18–36 credit hours in length and diplomas 37–48 and 49–59 credit hours in length. Statistical procedures generating a more accurate classification model were identified based on Cohen's Kappa, ROC curves, and sensitivity and specificity by certificate or diploma type. The results of this study were framed by a thorough explanation of the data screening and preprocessing necessary for the two academic years of data used in the study. The results addressed model training, variable importance, and the accuracy of two data modeling approaches (logistic regression and linear discriminant analysis) and three data mining approaches (classification tree, random forest, and support vector machine models).

For certificates 9–17 credit hours in length, environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) were analyzed to determine which, if any, were significant predictors of nontraditional student retention. Both data modeling approaches, logistic regression and linear discriminant analysis, shared similar results using both training data and test data. The predictor variables GPA and programs related to Transportation and Logistics were the most influential in students being retained, both of which represented academic integration components. With every one-point increase in GPA, the odds of being retained increases 59% (odds ratio = 1.59), and compared to students enrolled in other HOPE Career Grant programs, students enrolled in a program related to Transportation and Logistics were more likely to be retained.

Both the logistic regression and linear discriminant analysis models indicated Industrial Technology programs as the most influential predictor of students not being retained. Each of the three data mining approaches (classification trees, random forests, and support vector machines) identified similar predictor variables. The most common, influential predictors were GPA, age, and either Transportation and Logistics programs or Industrial Technology programs. However, the support vector machines model did not identify age as one of the top four predictors.

Of the five models for certificate 9 to 17 credit hours in length, four performed similarly in terms of accuracy, error rate, and F1 score. The logistic regression, linear discriminant analysis, classification tree, and random forest models produced similar results. The random forest model was the only model in this cohort with a higher true positive rate (0.68) as compared to the true negative rate (0.57). The support vector machine model had the lowest accuracy and F1 score, and the second lowest Kappa coefficient at 0.17, but had the largest specificity rate of all the models in this cohort (0.81). Overall, logistic regression and linear discriminant analysis performed well across metrics. However, the support vector machine will generate an accurate classification model based on the goal of correctly identifying students who will not be retained. Of the five classification models for certificates 9 to 17 credit hours in length, the support vector machine model will generate a more accurate classification model based on specificity.

For certificates 18–36 credit hours in length, environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) were analyzed to

determine which, if any, were significant predictors of nontraditional student retention. The results for this group were similar to those of certificates 9–17 credit hours in length. Both data modeling approaches, logistic regression, and linear discriminant analysis indicated GPA and programs related to Transportation and Logistics were the most influential in students being retained. Both the logistic regression and linear discriminant analysis models indicated that being out of high school for five years or more and being enrolled in a Cyber, Engineer, or Healthcare program are influential predictors of students not being retained. For example, if a student enrolls in a Cyber, Engineer, or Healthcare program, the odds of that student being retained decreases by 21% (odds ratio = 0.792 – 1), keeping other variables constant. Each of the three data mining approaches (classification trees, random forests, and support vector machines) identified similar predictor variables. The most common, influential predictors were GPA and age. However, similar to certificates 9–17 credit hours in length, the support vector machines model did not identify age as one of the top predictors.

For the cohort for certificates 18–36 credit hours in length, the logistic regression, and the linear discriminant analysis models had the highest area under the ROC curve. Although all models performed similarly in terms of accuracy, error rate, Kappa, and F1 score. All five models had higher sensitivity than specificity with the greatest difference being the classification tree model with a sensitivity of 0.90 and specificity of 0.38 which produced the highest F1 score of 0.64. Although the logistic regression and linear discriminant analysis had the highest AUC, the random forest had the highest specificity. Each of these three models had the same accuracy. Therefore, the random forest model

will generate a more accurate classification model based on specificity for certificates 18 to 36 credit hours in length.

For diplomas 37–48 credit hours in length, environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) were analyzed to determine which, if any, were significant predictors of nontraditional student retention. Both data modeling approaches, logistic regression and linear discriminant analysis, indicated predictor variables GPA, female students, and Black students were the most influential in students being retained. These represented background factors (gender and race) and an academic integration component (GPA). Both the logistic regression and linear discriminant analysis models using the test data indicated Cyber, Engineer, or Healthcare programs as the most influential predictor of students not being retained. Each of the three data mining approaches (classification trees, random forests, and support vector machines) identified similar predictor variables. The most common, influential predictors were GPA (an academic integration component) and age (a background factor). Unlike the previous results for certificate programs, the support vector machines model did identify age as one of the most influential predictors of student retention, but only in the test data.

In terms of model accuracy for diplomas 37 to 48 credit hours in length, all models performed poorly. Four of the five models had slightly higher sensitivity rates compared to the specificity rates. Overall, the logistic regression, linear discriminant analysis, and support vector machine each performed almost identical across all

performance metrics. Of the three, the linear discriminant analysis model had a slightly higher F1 score and AUC. However, the support vector machine had the highest accuracy and specificity. Therefore, the support vector machine will generate a more accurate classification model based on specificity for diplomas 37 to 48 credit hours in length.

For diplomas 49–59 credit hours in length, environmental factors (Pell eligibility, single parent status, displaced homemaker status), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and academic integration components (student GPA and program type) were analyzed to determine which, if any, were significant predictors of nontraditional student retention. Both data modeling approaches, logistic regression and linear discriminant analysis, shared similar results. The predictor variables GPA (an academic integration component) and Pell eligibility (an environmental factor) were the most influential in students being retained. Both the logistic regression and linear discriminant analysis models indicated that being out of high school for five years or more is an influential indicator of students not being retained. Each of the three data mining approaches (classification trees, random forests, and support vector machines) identified similar predictor variables. The variables GPA, age, and Industrial Technology programs were the most influential predictors of being retained.

Similar to the cohort for diplomas 37 to 48 credit hours, all models for diplomas 49 to 59 credit hours performed poorly. All five models in this cohort had higher sensitivity rates with the classification tree having the highest sensitivity at 0.83. Overall, logistic regression and linear discriminant analysis performed the best based on accuracy and AUC. Between the two models, almost all classification metrics were identical with

sensitivity and specificity rates having a slight variance. However, with its higher specificity, the logistic regression model may generate a slightly more accurate classification model for diplomas 49 to 59 credit hours in length.

Overall, for each of the four cohorts and each of the five classification models, the most significant predictor of nontraditional student retention for certificates or diplomas was GPA, an academic integration component. Based on the goal of correctly identifying students who will not be retained so adequate assistance and resources can be provided to them, the support vector machine will generate a more accurate classification model based on specificity.

Chapter V

SUMMARY, CONCLUSIONS, AND IMPLICATIONS

This chapter contains a discussion of the results of this study. Initially, a summary of the study is presented and then the purpose of the study is reviewed. Then, an overview of the methods and procedures for data analysis is discussed. A detailed discussion of the findings is organized by each research question and the limitations of the study are described. In conclusion, the suggestions for future research are offered and the conceptual and practical implications of the study are discussed.

The Technical College System of Georgia, whose mission is to build a well-educated workforce for Georgia, has multiple partnerships across the state specifically designated for in-demand diploma and certificate programs to create a pipeline of skilled workers for Georgia employers. However, the attainment goals set by state and national leaders cannot be met unless significantly more adults and other nontraditional students return to higher education and complete a degree or credential (Complete College America, n.d.). Conventional retention strategies aimed at traditional students may not work with today's college students. Therefore, the changing characteristics of nontraditional students need to be understood before retention efforts in the community and technical colleges are effective (Ashar & Skenes, 1993).

The future of Georgia's workforce depends on the diversity, adaptability, and broad-based talents and skills students acquire through quality higher education. There are simply not enough high school and traditional college students to create the educated

workforce required for the 21st-century economy (Pingel et al., 2016). Monroe (2006) asserted the complex, dynamic nature of nontraditional students requires continuous examination and refinement of our understanding of this population's changing demographics concerning attrition. A careful review of retention models and theories through the lens of nontraditional students can not only help colleges develop policies and procedures to facilitate student retention, but can align Georgia's nontraditional students with Georgia's workforce needs and requirements.

Although several studies focus predominantly on traditional students in associate or bachelor's degree programs, we do not have an understanding of factors related to college retention for nontraditional students seeking only a diploma or certificate. While there is prolific literature on the challenges and struggles facing nontraditional students, very little literature focuses on how the student's unique characteristics contribute to retention specific to the community and technical college environment. The purpose of this study was to examine the predictability of academic factors (student GPA and program type), background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and environmental factors (Pell eligibility, single parent status, displaced homemaker status) on the retention of nontraditional students enrolled in diploma and certificate programs in the Technical College System of Georgia. To do so, the researcher examined multiple prediction models to identify which statistical procedure generates the most accurate classification model.

A nonexperimental, ex post facto, correlational research design was used in this study. Archival data obtained from the Technical College System of Georgia were retrospectively analyzed to measure first-year retention. The use of archival data makes

the manipulation of the variables unlikely and unethical (Bordens & Abbott, 2011). Therefore, a nonexperimental, ex post facto research design was appropriate for this study as the independent predictor variables were not manipulated. Because the goal was to predict values on a binary outcome variable, the researcher identified which prediction model, out of two data modeling approaches and three data mining approaches, best predicts whether a student will be retained or not retained.

The target population included students identified as nontraditional at each of the 22 technical colleges in Georgia. The accessible population included first-time students identified as nontraditional at each of the 22 technical colleges in Georgia who were enrolled in one of 17 program areas defined by the HOPE Career Grant Program. Students were classified as nontraditional if they met the following criteria: first-time or beginning student, 25 years of age or older, and enrolled part-time. Program areas were subdivided into four distinct groups of certificates with 9–17 credit hours, certificates with 18–36 credit hours, diplomas with 37–48 credit hours, and diplomas with 49–59 credit hours.

Summary of Findings

The purpose of this study was to examine the predictability of academic, background, and environmental factors such as Pell eligibility, single parent status, displaced homemaker status, age, race or ethnicity, gender, high school diploma type, high school graduation date, student grade point average, and program type on the retention of nontraditional students enrolled in diploma and certificate programs in the Technical College System of Georgia. This study presented significant predictors of nontraditional student retention for certificates 9–17 and 18–36 credit hours in length,

and diplomas 37–48 and 49–59 credit hours in length. Statistical procedures generating a more accurate classification model were identified based on Cohen's Kappa, ROC curves, and sensitivity and specificity by certificate or diploma type. The analysis specifically focused on two cohorts of diploma and certificate-seeking students who began their enrollment in fall 2017 and fall 2018. The cohorts consisted of nontraditional students enrolled for the first time at any of the technical colleges in Georgia and were not high school students. A statistical learning approach was used to evaluate several models to identify the most accurate predictions on future student cohorts.

Various feature engineering techniques were used to transform the original data most suitable for the data modeling and data mining techniques being used. A review of missing data indicated all variables except two (high school diploma type and high school graduation date) had zero missing values in each of the eight data files. Given the percentage of missing values was less than 5%, data were imputed to address the missing values in six of the eight data files. Imputation via bagged trees and k-nearest neighbors produced similar distributions in each of the two variables. After imputation, chi-square tests were used for significance testing between the original data and the imputed data. In the 2017 data file representing certificates 9 to 17 credit hours, high school diploma type, and high school graduation date were missing 619 of 1,896 records (32.65%). Despite multiple attempts at imputation, a large disparity in the percentages between the complete dataset and the imputed dataset existed. Therefore, the 619 identified records were removed from the dataset resulting in 1,277 records.

Each continuous variable, age and GPA, were evaluated for outliers by inspecting boxplots and z-scores. A standard boxplot and adjusted boxplot for skewed distributions

were used initially to identify potential outliers. Both variables were converted to z-scores to mathematically assess for outliers. Neither of the continuous variables followed a normal distribution which violates an assumption for linear discriminant analysis. Therefore, a Yeo-Johnson transformation was applied. All predictors except age and GPA were converted from nominal data (e.g., factors) into one or more numeric binary variables representing specific factor level values. All numeric variables were centered and scaled.

Data for two academic years (2017-2018 and 2018-2019) were partitioned into a training data set and a test data set to be used to implement two data modeling approaches (logistic regression and linear discriminant analysis) and three data mining approaches (classification tree, random forest, and support vector machine models). During model training, upsampling was used to mitigate the effects of class imbalance in the outcome variable retained. In both the 2017 and 2018 cohorts of certificates 9–17 credit hours in length, the class imbalance was the greatest with the rate of students not retained only accounting for 17.93% and 17.72% respectively. Each model was evaluated using 10-fold cross-validation repeated five times with stratification and all predictor variables were allowed to enter each model. The prior probabilities were specified as 0.50 for predicted class 1, students who were retained, and 0.50 for predicted class 0, students who were not retained.

Conclusions for Research Question 1A

The logistic regression model was found to be statistically significant and resulted in an error rate of 0.34. Of the 15 predictor variables, nine were statistically significant as predictors of student retention: age, GPA, race (Hispanic), gender (female), graduation

date (out of high school at least five years or more), Pell eligibility, Cyber, Engineer, or Healthcare programs, Industrial Technology programs, and Transportation and Logistics programs. Given all other variables are unchanged, the odds of a student being retained increases by a factor of 1.59 (odds ratio = 1.59) as GPA increases. Likewise, enrollment in Transportation and Logistics programs increases the odds of a student being retained by a factor of 1.50 (odds ratio = 1.50) given all other variables remain unchanged. If a student enrolls in Industrial Technology programs, the odds of those students being retained decreases by 49% (odds ratio = 0.510 – 1), keeping other variables constant.

The predictor variables which were not statistically significant included race (Black), race (other), high school diploma (GED®), high school diploma (college prep or tech prep), single parents, and displaced homemakers. The variables of GPA and Transportation and Logistics programs were the strongest predictors of being retained. With every one-point increase in GPA, the odds of being retained increases 59% (odds ratio = 1.59), and compared to students enrolled in other HOPE Career Grant programs, students enrolled in a program related to Transportation and Logistics were more likely to be retained. The weakest predictors of being retained were Cyber, Engineer, or Healthcare programs, females, and Industrial Technology programs.

The linear discriminant analysis model resulted in an error rate of 0.34. The coefficients with the strongest associated weights included GPA, females, Cyber, Engineer, or Healthcare programs, Industrial Technology programs, and Transportation and Logistics programs. The strongest predictor of being retained or not was Industrial Technology programs with a negative coefficient of -0.666, GPA with a coefficient of 0.420, and Transportation and Logistics programs with a coefficient of 0.368. The

variables race (other), single parents, and race (Black) had the least influential coefficients of being retained or not.

The classification tree model was tuned for the parameters complexity and tree depth. The resulting model had 26 total splits and a cross-validated error rate of 0.43. The overall training error rate for the model was 0.38. To determine the strongest predictors of retention, variable importance was measured as the sum of the goodness of split measures (Gini index). The variables of GPA, Industrial Technology programs, females, and age were the strongest predictors of being retained or not. The weakest predictors of being retained or not were single parents, Pell eligibility, race (Hispanic), and race (other).

The two arguments tuned for the random forest model were mtry and node size. The resulting model had 500 trees with an out-of-bag (OOB) error rate of 14.86%, an error rate of 9.23% for class 0 (not retained), and an error rate of 20.48% for class 1 (retained). The overall error rate for the model was 0.34. The most important variables to the model result in the largest mean decrease in Gini value. The variables of GPA, age, and Industrial Technology programs were the most influential predictors of being retained or not. The weakest predictors of retention were race (other), high school diploma (GED®), Pell eligibility, and single parents.

The support vector machine model was tuned for the parameters cost and rbf_sigma. The best model resulted in 1,489 support vectors, an objective function value of -125.01, and an error rate of 0.23. The overall error rate for the model was 0.45. Permutation-based variable importance scores were computed for each predictor in the SVM model. The most influential predictors of being retained or not were Industrial

Technology programs, GPA, and Cyber, Engineer, or Healthcare programs. The least influential predictors of being retained or not were females, displaced homemakers, and graduation date (out of high school at least five years or more).

Both data modeling approaches, logistic regression and linear discriminant analysis, shared similar results in terms of variable importance. The predictor variables GPA and programs related to Transportation and Logistics were the most influential in students being retained, both of which represented academic integration components. Both the logistic regression and linear discriminant analysis models indicated Industrial Technology programs as the most influential predictor of students not being retained. Each of the three data mining approaches (classification trees, random forests, and support vector machines) identified similar predictor variables. The most common, influential predictors were GPA, age, and either Transportation and Logistics programs or Industrial Technology programs. However, the support vector machines model did not identify age as one of the top four predictors.

For certificates 9 to 17 credit hours in length, the most influential predictors of nontraditional student retention included one background factor (age) and two academic integration components (student GPA and program type). These predictors were also included in past retention studies where various retention theories, models, and frameworks were tested. Tinto's (1997) study of 287 first-year community college students set out to determine the degree to which learning communities and the adoption of collaborative learning strategies impacted persistence. Tinto (1997) used stepwise logit regression analysis to predict second-year persistence using both qualitative and quantitative methods. Five variables proved to be significant predictors of persistence using an alpha level of .10 among students at Seattle Central Community College (participation in the

Coordinated Studies Program, college grade point average, hours studied per week, perceptions of faculty, and a factor score on involvement with other students).

Retention based on a program of study or major may be tracked by specific colleges or universities but is not nationally tracked and remains difficult to measure (Seidman, 2005). Program-specific issues, which may influence retention, vary by delivery (Craig & Ward, 2008). In Craig and Ward's (2008) study, which looked at a cohort of first-time, full-time students at a public community college in New England, initial program major was a significant predictor of success or failure in their logistic regression analysis. Students majoring in engineering or chemistry *(*$\beta = 1.54$, $\chi^2(1, N = 1{,}729) = 12.85$, $p < .001$), business administration *(*$\beta = .73$, $\chi^2(1, N = 1{,}729) = 4.27$, $p < .05$), and legal studies *(*$\beta = .75$, $\chi^2(1, N = 1{,}729) = 4.18$, $p < .05$) had some of the lowest grade point averages, but resulted in student success as defined as being awarded a degree, a certificate, or transferring to another institution (Craig & Ward, 2008). Of the initial programs, engineering or chemistry majors had the highest odds ratio at 4.67. Business administration and legal studies both had a positive association with student success with odds ratios of 2.07 and 2.12 respectively (Craig & Ward, 2008).

Historically, age was not typically included in the research on retention because most research focused on traditional-age students (Cochran et al., 2013). For studies using age as a potential explanatory variable, the results were contradictory. Pascarella et al. (1981) found age to be a moderate predictor of student persistence using Tinto's student integration model on 853 students at a commuter four-year college. The researchers used a longitudinal study using the ACE (American Council on Education) Cooperative Institutional Research Program survey and data collected on all incoming

students, such as high school rank and college entrance test scores (Pascarella et al., 1981). Three-group discriminant function analysis was used for the freshman to sophomore persisters, freshman stop-outs, and first-quarter freshman withdrawals (Pascarella et al., 1981). The first stage of analysis included all pre-enrollment characteristics (high school academic performance, age, perceived likelihood of dropping out, perceived likelihood of transfer, and perceived need for remediation), and only those variables contributing to group discrimination significant at $p < .10$ were used in the second stage of the stepwise discriminant analysis (Pascarella et al., 1981). The results indicated pre-enrollment variables like age, along with first-quarter GPA, significantly differentiate between freshman year persisters and early withdrawals (Pascarella et al., 1981). The classification analysis based on the six-variable equation correctly identified 72% of the early withdrawals and 74% of the persisters (Pascarella et al., 1981). The findings revealed a significant main effect for the age variable on persisters and withdrawals ($F(1, 847) = 7.12$, $p < .01$) (Pascarella et al., 1981).

Conclusions for Research Question 1B

The logistic regression model was found to be statistically significant and the model resulted in an error rate of 0.39. Of the 15 predictor variables, five were statistically significant as predictors of student retention: GPA, graduation date (out of high school at least five years or more), Pell eligibility, Cyber, Engineer, or Healthcare programs, and Transportation and Logistics programs. Given all other variables are unchanged, the odds of a student being retained increases by a factor of 2.21(odds ratio = 2.21) as GPA increases. If a student enrolls in Cyber, Engineer, or Healthcare programs, the odds of those students being retained decreases by 21% (odds ratio = 0.792 – 1),

keeping other variables constant. The variables of GPA and Transportation and Logistics programs were the strongest predictors of being retained. The weakest predictors of being retained were graduation date (out of high school at least five years or more) and Cyber, Engineer, or Healthcare programs.

The overall error rate for the linear discriminant analysis model was 0.39 and the coefficients with the strongest associated weights included GPA, graduation date (out of high school at least five years or more), Pell eligibility, Cyber, Engineer, or Healthcare programs, and Transportation and Logistics programs. The strongest predictor of being retained or not was GPA with a coefficient of 0.978 and Transportation and Logistics programs with a coefficient of 0.305. The variables race (Black) and high school diploma (GED®) had the least influential coefficients of being retained or not.

The classification tree model was tuned for the optimal complexity factor (0.0000793) and maximum tree depth (4). The resulting model had 5 total splits, a cross-validated error rate of 0.77, and an overall error rate for the model of 0.41. GPA was the most influential predictor of being retained or not, with age coming in a distant second. The weakest predictors of being retained or not were race (Black), race (other), high school diploma (college prep or tech prep), and single parents.

The optimal mtry parameter (6) and minimum node size (39) were used to update the random forest model. The resulting model had 500 trees with an out-of-bag (OOB) error rate of 36.55%, an error rate of 46.85% for class 0 (not retained), and an error rate of 26.26% for class 1 (retained). The overall error rate for the model was 0.39. The variables of GPA and age were the most influential predictors of being retained or not.

The weakest predictors of retention were high school diploma (GED®), Industrial Technology programs, and race (other).

The support vector machine model was tuned for the parameters cost and rbf_sigma. The best model resulted in 901 support vectors, an objective function value of -80.91, and an error rate of 0.36. The overall error rate for the model was 0.43. The most influential predictors of being retained or not were GPA, Transportation and Logistics programs, and race (Black). The least influential predictors of being retained or not were age, race (Hispanic), and displaced homemaker.

In terms of variable importance, both data modeling approaches indicated GPA and programs related to Transportation and Logistics were the most influential in students being retained. Both the logistic regression and linear discriminant analysis models indicated that being out of high school for five years or more and being enrolled in a Cyber, Engineer, or Healthcare program are influential predictors of students not being retained. Each of the three data mining approaches (classification trees, random forests, and support vector machines) identified similar predictor variables. The most common, influential predictors were GPA and age. However, similar to certificates 9–17 credit hours in length, the support vector machines model did not identify age as one of the top predictors.

For certificates 18 to 36 credit hours in length, the most influential predictors of nontraditional student retention included one background factor (age) and two academic integration components (student GPA and program type). McGrath and Braunstein's (1997) completed a study to identify the predictors of attrition for freshmen who voluntarily withdrew by studying the relationship between attrition and certain

demographic, academic, financial, and social factors. Specifically, McGrath and Braunstein (1997) looked at which factors differentiate between those freshmen who were retained and those who were not retained. The researchers used the College Student Inventory to assess predispositions, pre-college experiences, and attributes which may influence retention for full-time freshmen at Iona College in New York (McGrath & Braunstein, 1997). Because there were additional data used from students' academic, demographic, and financial records, a preliminary analysis of t-tests was used to reduce the number of variables for use in a logistic regression (McGrath & Braunstein, 1997). A significant difference was found between the groups when McGrath and Braunstein (1997) used a t-test on the first semester GPAs for freshmen who were retained (M = 2.67, SD = .64) and those who were not retained (M = 1.76, SD = 1.17), $t(297) = 8.9$, $p < .001$, $d = .96$. Independent variables which were statistically significant at the .05 level were entered into a stepwise logistic regression (McGrath & Braunstein, 1997). The results indicated first-semester college GPA ($\beta = 1.15$, $p < .001$, $R = .34$) as the strongest variable in predicting persistence between the first and second years (McGrath & Braunstein, 1997). McGrath and Braunstein (1997) used logistic regression to predict the probability of freshmen returning for their sophomore year by assigning students to a “retained” group if the predicted probability of retention was greater than 50%; otherwise, students were assigned to the "non-retained" group. The researchers applied these criteria to the final sample of 322 freshmen, and along with students' impressions of other students, were able to make correct predictions in approximately 80% of the analyzed cases (McGrath & Braunstein, 1997).

Daempfle's (2003) article on first-year college majors highlighted lower enrollment, higher transfers to other disciplines, and lower retention rates were more prevalent among students majoring in mathematics, science, or engineering. The St. John et al. (2004) logistic regression study indicated student major influences persistence decisions. This study, using the Indiana Commission for Higher Education's Student Information System, found White freshmen who major in social sciences ($\beta = -.82$, $p < .05$) or those who were undecided ($\beta = -.66$, $p < .01$) had a lower probability of persisting than other White students, although African American freshmen in the undecided majors were not significantly different from other African American students in persistence (St. John et al., 2004). St. John et al. (2004) also found three distinct programs of study Health ($\beta = 1.09$, $p < .05$), Business ($\beta = 1.10$, $p < .01$), and Engineering or Computer Science ($\beta = 1.20$, $p < .05$) had positive associations with the persistence of African American sophomores, implying the economic potential of a major field had a substantial impact on the student's persistence.

Feldman's (1993) study of one-year retention of first-time students at a community college used chi-square analysis for univariate comparisons and logistic regression to select and order the factors which contributed to retention. She found age had a significant impact ($\chi^2(1) = 26.13$, $p < .001$) on retention using both univariate and multivariate analysis (Feldman, 1993). The odds of students age 20-24 years old dropping out was 1.77 times that of students aged 19 or younger and the 20-24 age range was the most significant predictor age range according to the Wald statistic ($\chi^2(1) = 7.37$, $p < .001$) (Feldman, 1993).

Conclusions for Research Question 1C

The logistic regression model was found to be statistically significant and the model resulted in an error rate of 0.43. Of the 15 predictor variables, six were statistically significant as predictors of student retention: GPA, race (Black), race (other), gender (female), high school diploma (college prep or tech prep), and Cyber, Engineer, or Healthcare programs. Given all other variables are unchanged, the odds of a student being retained increases by a factor of 2.21 (odds ratio = 2.21) as GPA increases. If a student enrolls in Industrial Technology programs, the odds of those students being retained decreases by 49% (odds ratio = 0.510 – 1), keeping other variables constant. The variables of GPA, females, and race (Black) were the most influential predictors of being retained or not. The weakest predictors of being retained or not were race (other) and Cyber, Engineer, or Healthcare programs.

The overall error rate for the linear discriminant analysis model was 0.44 and the coefficients with the strongest associated weights included GPA, race (Black), females, and Cyber, Engineer, or Healthcare programs. The most influential predictors of being retained or not were GPA with a coefficient of 0.915, Cyber, Engineer, or Healthcare programs with a negative coefficient of -0.359, and females with a coefficient of 0.327. The variables Transportation and Logistics programs and high school diploma (GED®) had the least influential coefficients of being retained or not.

The classification tree model was tuned for the optimal complexity factor (0.0000793) and maximum tree depth (4). The resulting model had 7 total splits, a cross-validated error rate of 0.69, and an overall error rate for the model of 0.50. The variables of GPA and age were the strongest predictors of being retained or not. In contrast, the

weakest predictors of being retained or not were females, Transportation and Logistics programs, and displaced homemakers.

The optimal mtry parameter (10) and minimum node size (35) were used to update the random forest model. The resulting model had 500 trees with an out-of-bag (OOB) error rate of 26.06%, an error rate of 33.69% for class 0 (not retained), and an error rate of 18.43% for class 1 (retained). The overall error rate for the model was 0.44. The most important variables to the model result in the largest mean decrease in Gini value. The variables of GPA, age, and Cyber, Engineer, or Healthcare programs were the most influential predictors of being retained or not. The weakest predictors of retention were single parents, race (other), and displaced homemakers.

The support vector machine model was tuned for the parameters cost and rbf_sigma. The best model resulted in 880 support vectors, an objective function value of -76.35, and an error rate of 0.32. The overall error rate for the model was 0.42. The most influential predictors of being retained or not were GPA, age, and race (Black). The least influential predictors of being retained or not were displaced homemakers, race (Hispanic), and single parents.

Both data modeling approaches, logistic regression and linear discriminant analysis, shared similar results in regards to variable importance. The predictor variables GPA, female students, and Black students were the most influential in students being retained, representing background factors (gender and race) and an academic integration component (GPA). Both the logistic regression and linear discriminant analysis models indicated Cyber, Engineer, or Healthcare programs as the most influential predictor of students not being retained. Each of the three data mining approaches (classification

trees, random forests, and support vector machines) identified similar predictor variables. The most common, influential predictors were GPA (an academic integration component) and age (a background factor). Unlike the previous results for certificate programs, the support vector machines model did identify age as one of the most influential predictors of student retention.

For diplomas 37 to 48 credit hours in length, the most influential predictors of nontraditional student retention included three background factors (gender, race or ethnicity, and age) and one academic integration component (student GPA). These predictors were also included in past retention studies where various retention theories, models, and frameworks were tested. Existing literature reveals varying results about the effects of gender differences on persistence. Mohammadi (1994) found men more likely to persist than women. Chen and Thomas (2001) and Halpin (1990) found women more likely to persist than men. The Horn et al. (2002) NCES report indicated no influence by gender on persistence (Horn et al., 2002). Although Pritchard and Wilson's (2003) research of 218 undergraduate students from a private Midwestern university focused on student's emotional and social factors, the researchers also investigated the influence of traditional demographic variables like gender and found gender did not influence persistence. Pritchard and Wilson's (2003) study was designed to identify the relationship between student emotional and social health and academic success and retention. Multiple regressions were used to assess the influence of demographic variables, the effect of emotional health, and the effect of social health on GPA and retention (Pritchard & Wilson, 2003). While the combined influence of all the demographic variables in the study had a significant effect on GPA ($R^2 = .22$, $F(7, 109) = 4.17$, $p < .001$), they had no

effect on the intent to drop out (R^2 = .02, F(7, 182) = 1.00, p = .80). (Pritchard & Wilson, 2003).

Ethnicity differences are factors in some retention studies. Singell and Waddell's (2010) research, which used an empirical model developed by Singell (2004), centered on whether the University of Oregon could effectively identify students who might be retention risks early in their college careers using accessible data. The researchers combined logistic regression and hazard modeling approaches of prior work and used existing student-level data to estimate a predicted retention probability based on gender, race, high school GPA, and SAT scores (Singell & Waddell, 2010). Singell and Waddell (2010) estimated separate prediction models for residents and nonresidents supported by a likelihood ratio test that rejects the restriction of equal coefficients by residential status at the 99% level. Singell and Waddell (2010) claimed, absent of other attributes, African American (β = .06, p < .01) and Asian (β = .04, p < .01) students are more likely to be retained than White students in the fall term of their second year. This research found Hispanic, Native American, and other non-White students do not differ in their retention probabilities from White students (Singell & Waddell, 2010). In addition to providing context between race, ethnicity, and retention, Singell and Waddell's (2010) research found students at risk of dropping out can be identified using accessible statistical models and information available at the time a student enrolls and monitoring students as they matriculate improves the model's ability to predict retention. This implies a trade-off between early identification and intervention and the information gained by including additional data which becomes available as the student progresses through their program of study.

Conclusions for Research Question 1D

The overall logistic regression model was found to be statistically significant and the model resulted in an error rate of 0.39. Of the 15 predictor variables, three were statistically significant as predictors of student retention: GPA, graduation date (out of high school at least five years or more), and Pell eligibility. Given all other variables are unchanged, the odds of a student being retained increases by a factor of 2.50 (odds ratio = 2.50) as GPA increases. If a student enrolls in Industrial Technology programs, the odds of those students being retained decreases by 49% (odds ratio = 0.510 – 1), keeping other variables constant. The variables of GPA and Pell eligibility were the most influential predictors of being retained or not. The weakest predictors of being retained or not were Industrial Technology programs and graduation date (out of high school at least five years or more).

The overall error rate for the linear discriminant analysis model was 0.39 and the coefficients with the strongest associated weights included GPA, graduation date (out of high school at least five years or more), and Pell eligibility. The strongest predictor of being retained or not was GPA with a coefficient of 1.092 and graduation date (out of high school at least five years or more) with a negative coefficient of -0.212. The variables race (Hispanic) and Cyber, Engineer, or Healthcare programs had the least influential coefficients of being retained or not.

The classification tree model was tuned for the optimal complexity factor (0.00233) and maximum tree depth (3). The resulting model had one split, a cross-validated error rate of 0.66, and an overall error rate of 0.45. The variable GPA was the most influential predictor of being retained or not. The variables graduation date (out of

high school at least five years or more) and age were a distant second with age being the least influential predictor.

The tuned parameters for the random forest model were mtry (2) and node size (31). The resulting model had 500 trees with an out-of-bag (OOB) error rate of 31.11%, an error rate of 48.05% for class 0 (not retained), and an error rate of 14.18% for class 1 (retained). The overall error rate for the model was 0.44. The most important variables to the model result in the largest mean decrease in Gini value. The variables GPA and age were the most influential predictors of being retained or not. The weakest predictors of retention were displaced homemakers and Transportation and Logistics programs.

The support vector machine model was tuned for the parameters cost and rbf_sigma. The best model resulted in 1,837 support vectors, an objective function value of -164.49, and an error rate of 0.30. The overall error rate for the model was 0.44 and the most influential predictors of being retained or not were GPA, Industrial Technology programs, and high school diploma (college prep or tech prep). The least influential predictors of being retained or not were age, race (Hispanic), and Cyber, Engineer, or Healthcare programs.

In both data modeling approaches (logistic regression and linear discriminant analysis), the predictor variables GPA (an academic integration component) and Pell eligibility (an environmental factor) were the most influential in students being retained. The same models indicated that being out of high school for five years or more is an influential indicator of students not being retained. Each of the three data mining approaches (classification trees, random forests, and support vector machines) identified similar predictor variables. The variables GPA, age, and Industrial Technology programs

were the most influential predictors of being retained. The weakest predictors of retention were displaced homemakers, Transportation and Logistics programs, and Cyber, Engineer, or Healthcare programs.

For diplomas 49 to 59 credit hours in length, the most influential predictors of nontraditional student retention included one environmental factor (Pell eligibility), one background factor (age), and two academic integration components (student GPA and program type). These predictors were also included in past retention studies where various retention theories, models, and frameworks were tested. In Craig and Ward's (2008) study of 1,729 first-time, full-time community college students, the researchers found GPA was a significant indicator of student retention using logistic regression analysis. On average, students not retained had a cumulative GPA of 1.68 and had earned only 16.8 credit hours compared to 2.29 for retained students (Craig & Ward, 2008). Of the student academic characteristics, cumulative GPA ($\beta = .73$, $\chi^2(1, N = 1{,}729) = 91.44$, $p < .001$) was most strongly related to student success with a 2.04 odds ratio (Craig & Ward, 2008). Second semester GPA ($\beta = .32$, $\chi^2(1, N = 1{,}729) = 44.14$, $p < .001$) and attempted but unearned credits ($\beta = -.03$, $\chi^2(1, N = 1{,}729) = 38.36$, $p < .001$) were also significant (Craig & Ward, 2008). Second semester GPA had a positive association with student success with an odds ratio of 1.38 but attempted but unearned credits had a negative association with an odds ratio of 0.97 (Craig & Ward, 2008). Titus (2006) performed another study that relates college GPA to student persistence. Titus (2006) conducted a study using hierarchical generalized linear modeling on 4,951 first-time, full-time students using a national database of four-year institutions. He found GPA significantly increased the odds for persistence ($\beta = .48$, odds ratio = 1.61; $p < .001$) (Titus, 2006).

The Nakajima et al. (2012) study of 427 community college students looked at the influence of background variables, financial variables, and academic variables on students' persistence in community college education. Nakajima et al. (2012) questioned if academic integration and psychosocial variables influence student persistence by using a 63-item survey assessing psychosocial variables, academic integration, and various background variables. Among the background variables, the study used t-tests to reveal age and high school graduation year influenced student persistence in community college students (Nakajima et al., 2012). Those who persisted were younger (M = 24.12, SD = 8.19) compared to those who did not persist (M = 26.23, SD = 8.48) ($t(370) = 2.13$; $p <$ .05), but these effects diminished once multiple variables were entered into the analysis (Nakajima et al., 2012).

In a recent study, Turk and Chen (2017), in trying to understanding how, when, and why community college students transfer to four-year colleges and universities, found receiving federal financial aid significantly impacts the likelihood of retention. Using a nationally representative data source and a multilevel model, the researchers used logistic regression to test a series of academic, demographic, social, and institutional-level characteristics to determine what impact they have on community college students' likelihood of upward transfer. Although marginally significant, receiving a Pell grant was associated with a 28% reduction in the chances of transfer (β = -.33, p = .06, odds ratio = 0.72) (Turk & Chen, 2017). However, students who received a federal student loan were more than four times as likely to transfer to a four-year institution as students who did not receive a federal loan (β = 1.52, $p < .001$, odds ratio = 4.56) (Turk & Chen, 2017). Turk

and Chen (2017) recommended federal funding increases should keep pace with inflation to help nontraditional students afford postsecondary education.

Conclusions for Research Question 2

For each statistical procedure Cohen's Kappa, ROC curves, sensitivity, and specificity were used to identify the most accurate classification model at predicting nontraditional student retention. Of the five statistical models evaluated using data for certificates 9 to 17 credit hours in length, the random forest model, the logistic regression model, and the linear discriminant model shared the same error rate (0.34). The highest error rate was the support vector machine model at 0.45. The random forest model had the highest sensitivity and F1 score, while the logistic regression and linear discriminant analysis had the best accuracy and AUC. The support vector machine model had the lowest accuracy and F1 score, and the second-lowest Kappa coefficient at 0.17, but had the largest specificity rate of all the models in this cohort (0.81). Overall, logistic regression and linear discriminant analysis performed well across metrics. However, the support vector machine will generate an accurate classification model based on the goal of correctly identifying students who will not be retained. Of the five classification models for certificates 9 to 17 credit hours in length, the support vector machine model will generate a more accurate classification model based on specificity.

Of the five statistical models for certificates 18 to 36 credit hours in length, the random forest model, the logistic regression model, and the linear discriminant model all shared the lowest error rate of 0.39. The highest error rate was the support vector machine model at 0.45. All models performed similarly in terms of accuracy, error rate, Kappa, and F1 score. All five models had higher sensitivity than specificity (indicting

higher false positive rates) with the greatest difference being the classification tree model with a sensitivity of 0.90 and specificity of 0.38 which produced the highest F1 score of 0.64. Although the logistic regression and linear discriminant analysis had the highest AUC, the random forest had the highest specificity. Each of these three models had the same accuracy. Therefore, the random forest model will generate a more accurate classification model based on specificity for certificates 18 to 36 credit hours in length.

Of the five statistical models evaluated using data for diplomas 37 to 48 credit hours in length, the support vector machine model produced the lowest error rate (0.42). The highest error rate was a dismal 0.50 which belonged to the classification tree model. For diplomas 37 to 48 credit hours in length, all models performed poorly. The accuracy of each model ranged from 0.50 to 0.58 and the Kappa coefficients were dismal and indicated poor agreement ranging from 0.09 to 0.16. Four of the five models had slightly higher sensitivity rates compared to the specificity rates. Overall, the logistic regression, linear discriminant analysis, and support vector machine each performed almost identical across all performance metrics. Of the three, the linear discriminant analysis model had a slightly higher F1 score and AUC. However, the support vector machine had the highest accuracy and specificity. Therefore, the support vector machine will generate a more accurate classification model based on specificity for diplomas 37 to 48 credit hours in length.

Similar to the cohort for diplomas 37 to 48 credit hours, all models for diplomas 49 to 59 credit hours performed poorly. The logistic regression model and the linear discriminant model shared the lowest error rate of 0.39. The highest error rate was the classification tree model at 0.45. All five models in this cohort had higher sensitivity rates

with the classification tree having the highest sensitivity at 0.83. Although the random forest model and the support vector machine model had a slightly higher accuracy (0.56) than the classification tree (0.55), both had poor Kappa coefficients (Kappa = 0.17) and F1 scores (0.48). Overall, logistic regression and linear discriminant analysis performed the best based on accuracy and AUC. Between the two models, almost all classification metrics were identical with sensitivity and specificity rates having a slight variance. However, with its higher specificity, the logistic regression model may generate a slightly more accurate classification model for diplomas 49 to 59 credit hours in length.

Limitations

The data for this research study was not collected to answer the researcher's specific research questions. By only using historical student-level data, this study was limited to variables only available through the Technical College System of Georgia Data Center. Additional variables identified in the literature review were not available for analysis, and therefore not included in the study. These variables included financial independence, employment status, marital status, and having dependents. Also, the accuracy of data extracted from each college-level student information system was not guaranteed free of errors. The majority of data errors in files extracted from a Banner database can be attributed to human error. However, many errors were mitigated through the design of the Banner user interface. Meaning many Banner data fields used in this study only accept specific values, thereby decreasing the chance of data entry error.

The cohort for this study was limited to nontraditional students who were enrolled for the first time at any of the technical colleges in Georgia and were not high school students. First-time students identified as special admit or learning support was not

included in this study as they cannot receive federal financial aid. Two independent variables, single parent and displaced homemaker, were self-reported by students. A limitation of self-reported data is the accuracy of responses could not be determined.

Additionally, not all assumptions for each of the five models were met. The independent variables in the linear discriminant analysis are assumed to have a multivariate normal (Gaussian) distribution (James et al., 2013). However, violating this assumption is normally acceptable as long as the sample size is large enough (James et al., 2013). Also, when the objective is only prediction or classification, these assumptions are less constraining. A review of skewness and kurtosis values and histograms indicated the continuous variables age and GPA were both moderately to substantially skewed. Therefore, the assumption of normality was not met. Being neither of the continuous variables followed a normal distribution which violates an assumption for linear discriminant analysis, a Yeo-Johnson transformation was applied. Yeo-Johnson transformation is similar to the Box-Cox transformation but does not require the input variables to be strictly positive. Once the Yeo-Johnson transformation was applied, the assumption of normality was met.

Also, logistic regression assumes the linearity of continuous predictors and the log odds (James et al., 2013). Linearity in the logit was assessed by constructing component-plus-residual plots of the residuals of each continuous predictor against the dependent variable. The assumption of linearity of the independent variables age and GPA and the log odds was not met. Once the Yeo-Johnson transformation was applied, the assumption of linearity was met.

Implications

Conceptual Implications. The guiding conceptual models for this study were Bean and Metzner's (1985) Model of Nontraditional Undergraduate Student Attrition and Hirschy et al.'s (2011) Conceptual Model for Student Success in Community College Occupational Programs. Bean and Metzner's (1985) model, the first model to specifically address the nontraditional student experience in higher education, proposed four sets of variables affecting the dropout decision: academic performance, intent, background, and environmental variables. Hirschy et al.'s (2011) model focused specifically on career and technical education students and suggests students pursuing occupational associate's degrees or certificates differ from those students seeking academic majors at two-year institutions. Independent variables for this study were aligned with the academic, background, and environmental factors described in both conceptual models. There were two academic factors (student GPA and program type), five background factors (age, race or ethnicity, gender, high school diploma type, high school graduation date), and three environmental factors (Pell eligibility, single parent status, displaced homemaker status). Since a large number of technical college students are nontraditional under Bean and Metzner's definition, the conceptual models for this study were suitable. This research did support both conceptual models. However, based on the predictability of program type this study may suggest modifications to these models to encompass only certificates and/or diplomas in technical colleges.

Practical Implications. The findings in this research study provide insight and understanding on how factors influence nontraditional student retention. Mindful of the complexities of today's technical college students, decision-makers must be aware of the

issues related to nontraditional students and be prepared to make informed decisions on how to better serve the needs of this specific student population. If colleges do nothing to improve the odds of retention for nontraditional students, a large segment of our population and the majority of college students will continue on the path to failure (Chen, 2017).

Many nontraditional students bring with them different expectations and different needs (Ross-Gordon, 2011). Failure to track these expectations, nontraditional trends, and to provide accurate information may result in educational administrators misunderstanding the needs of 21st-century undergraduates and/or misappropriating educational resources (Reeves et al., 2011). Administrators must intentionally track nontraditional student trends by examining institutional success measures such as enrollment and retention. Although institutions cannot specifically change the academic, environmental, or background factors related to nontraditional students, administrators can cultivate a supportive environment and develop processes and procedures which will benefit the retention of these students. For example, the needs of these students go beyond the simple registration process and call for more effective and efficient methods of academic advising (Hunter & White, 2004). Academic advising is integral to student success, persistence, and retention. Advising is not only linked to student retention and completion but student employability as well (Council for the Advancement of Standards in Higher Education, 2014). Based on this research study, it is recommended to train advisors in the use of proactive advising which rests on the premise that colleges should not wait for students to fall into academic difficulty before making contact with them (Varney, 2013). Advisors should play a key role in motivating students to utilize campus

services and resources before they face a crisis. Therefore, it is recommended that advising no longer be approached as a singular event, but as an ongoing process which assists students throughout their academic careers. For example, nontraditional students could meet with advisors at scheduled checkpoints throughout the semester. Checkpoints could include an advisement session with advisors at the beginning of the semester, an advisement session with advisors after the first six weeks, and an advisement session with advisors to advise for the next semester (approximately 10 weeks).

Also, community college students are more likely to persist if they are, not only advised about what courses to take but also helped in setting academic goals and creating a plan for achieving those goals (Center for Community College Student Engagement, 2015). College students are more likely to complete a degree in a timely fashion if they choose a program and develop an academic plan early, have a clear roadmap of the courses they need to take to complete a credential, and receive guidance and support to help them stay on a plan (Bailey & Smith Jaggars, 2015). The academic curricula in catalogs and on websites do not guide students on how to develop an academic plan. Based on the findings of this research, it is recommended beginning nontraditional students develop an academic plan with their academic advisor during their first semester. Advisors could develop an academic plan based on students attending full-time (three semesters), part-time (three semesters), full-time (fall and spring), and part-time (fall and spring). The academic plan could include financial costs for books and equipment for each semester and any courses offered only once during the academic year. This comprehensive academic plan could be designed to explain and communicate to

students their program requirements from start to finish and provide them with an advisor who will support them throughout their program of study.

Recommendations for Future Research

As previously noted, one of the guiding conceptual models for this study was Bean and Metzner's (1985) Model of Nontraditional Undergraduate Student Attrition. The variables identified in the Bean and Metzner model (academic, background, and environmental) were used in part to guide the selection of variables for this study. Bean and Metzner (1985) suggest the structure of the student attrition model was meant to be flexible and future researchers were encouraged to include factors not included in the original model, as well as concentrate their efforts on specific parts of the model. Additional variables may provide better results to assist colleges in developing policies and procedures to facilitate nontraditional student retention. Based on the results of this study, future researchers could create a more comprehensive model of all the factors influencing nontraditional student retention. The following variables identified in the literature review could be acquired through most college's student information system and/or through state data systems such as GA AWARDS, Georgia's Pre-K through workforce longitudinal data system: residency, high school GPA, SAT scores, early performance in college, course-taking patterns, and course withdrawal patterns. Other variables identified in the literature review such as employment, finances, family responsibilities, childcare issues, outside encouragement, educational goals, motivation, study skills, and time management could be acquired through surveys or personal interviews.

In addition, all data analysis for this research was conducted with software based on the R programming language. The tidyverse, a collection of R packages designed for data preparation and data analysis, contains a subset of packages specifically focused on data modeling (Kuhn & Silge, 2021). The tidymodels framework is a collection of packages for modeling and machine learning using tidyverse principles (Kuhn & Silge, 2021). The tidymodels packages, in terms of a software development lifecycle, are relatively new and continue to be tested and integrated (Kuhn & Silge, 2021). Therefore, future research may produce more accurate models as components of the tidyverse system are further developed and documented. Additionally, other classification models such as neural networks, K-nearest neighbors, and C5.0 could be considered in future research.

Further, the recipes package defines data preprocessing and feature engineering steps which are then applied to models being evaluated (Wickham et al., 2019). Feature engineering includes activities that reformat predictor values to make them easier for a model to use effectively (i.e. dummy variables, imputation, and normalization) (Wickham et al., 2019). Future studies could take the feature engineering transformations further to improve model performance. Additional transformations might include engineering new features/encodings or creating interaction effects. Creating new features is critical when considering which variables are required when building a predictive model, versus focusing on the available variables. Feature extraction includes principal component analysis (PCA), cluster analysis, and text analytics. Interaction effects, where two or more predictor variables are working together, can create a variation in the response variable according to Kuhn and Johnson (2020). Predictors which interact can be

included in a model to help explain additional variation in the response variable and improve the predictive ability of the model (Kuhn & Johnson, 2020). For example, Terenzini and Pascarella (1978) used stepwise multiple regressions to determine the interaction between race or ethnic origin. The interaction between race, ethnic origin, and intellectual development and progress was statistically significant ($F(1, 451) = 5.00$, $p < .05$) (Terenzini & Pascarella, 1978). The researchers were able to highlight race or ethnic origin was involved in two significant and unique interactions related to the probability of dropping out voluntarily (Terenzini & Pascarella, 1978).

Finally, many models, especially complex predictive or machine learning models, can work well on the training data but may fail when exposed to new data. Often, this issue is due to decisions made during the training of the models. Specific parameters for the classification tree, random forests, and support vector machine models were tuned. Future studies could train baseline models without tuning and subsequently introduce other tuning parameters such as the number of trees or the minimum number of data points in each node split to compare the predictive performance of the models.

Conclusion

Findings from this study confirm previous studies which show several academic, environmental, and background factors as significant predictors of student retention. Overall, GPA (an academic integration component) was the most influential predictor across each data file and each data model. The predictor variables GPA, programs related to Transportation and Logistics, female students, Black students, and Pell eligibility were influential in students being retained, representing academic integration (GPA and program type), background factors (gender and race), and environmental factors (Pell).

Being out of high school for five years or more and being enrolled in Cyber, Engineer, or Healthcare programs or Industrial Technology programs were influential predictors of students not being retained. Overall, for each of the four cohorts and each of the five classification models, the logistic regression and linear discriminant analysis performed the most consistently in terms of accuracy and AUC. However, if the goal is to correctly identify students who will not be retained so adequate assistance and resources can be provided to them, one should consider the support vector machine to generate a more accurate classification model based on specificity.

The findings of this research present a statewide picture of retention for nontraditional students in the Technical College System of Georgia and generalizations could be used to specifically improve processes and procedures on how colleges recruit and respond to this growing and diverse student population. With a specific focus on nontraditional students in diploma and certificate programs, the outcomes of this research allow decision-makers to consider how student factors influence nontraditional student progression from year 1 to year 2 to make informed decisions on how to better serve the needs of this specific student population.

REFERENCES

Advisory Committee on Student Financial Assistance [ACSFA]. (2012). *Pathways to success: Integrating learning with life and work to increase national college completion*. Washington, DC: US Department of Education.

Alfonso, M., Bailey, T. R., & Scott, M. (2005). The educational outcomes of occupational sub-baccalaureate students: Evidence from the 1990s. *Economics of Education Review, 24*(2), 197-212.

American Association of Community Colleges. (2015). *Community college completion*. Washington, DC: Author.

American Association of Community Colleges. (2018). *Fast facts 2018* [Data file].

Ary, D., Jacobs, L. C., Sorensen, C., & Razavieh, A. (2006). *Introduction to research in education* (8th ed.). Belmont, CA: Wadsworth.

Ashar, H., & Skenes, R. (1993). Can Tinto's student departure model be applied to nontraditional students? *Adult Education Quarterly, 43*(2), 90-100.

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

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

Astin, A. W., & Oseguera, L. (2005). *Degree attainment rates at American colleges and universities.* Los Angeles, CA: Higher Education Research Institute.

Bailey, T. R., & Smith Jaggars, S. (2015). *Redesigning America's community colleges: A clearer path to student success*. Cambridge, MA: Harvard University Press.

Barnett, V. & Lewis, T. (1963). A study of the relationship between GCE and Degree results. *Journal of the Royal Statistical Society, 126*, 187-216.

Baum, S., Ma, J., Pender, M., & Welch, M. (2017). *Trends in student aid 2017*. New York, NY: The College Board.

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

Bean, J. P., & Eaton, S. (2000). A psychological model of college student retention. In J. M. Braxton (Ed.), *Reworking the departure puzzle: New theory and research on college student retention* (pp. 48-61). Nashville, TN: University of Vanderbilt Press.

Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research, 55*(4), 485-540.

Berger, J. B., & Lyon, S. C. (2005). Past to present: A historical look at retention. In A. Seidman (Ed.), *College student retention: Formula for student success* (pp. 1-30). Westport, CT: Praeger Publishers.

Berkner, L. K., Cuccaro-Alamin, S., & McCormick, A. C. (1996). *Descriptive summary of 1989–90 beginning postsecondary students five years later: With an essay on postsecondary persistence and attainment* (NCES 96–155). Washington, DC: National Center for Education Statistics.

Berkner, L. K., He, S., & Cataldi, E. F. (2002). *Descriptive summary of 1995–96 beginning postsecondary students six years later* (NCES 2003–151). Washington, DC: National Center for Education Statistics.

Berkner, L. K., Horn, L., & Clune, M. (2000). *Descriptive summary of 1995–96 beginning postsecondary students: Three years later* (NCES 2000-154). Washington, DC: National Center for Education Statistics.

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

Bozick, R., & DeLuca, S. (2005). Better late than never? Delayed enrollment in the high school to college transition. *Social Forces, 84*, 527-550.

Braxton, J. M., & Hirschy, A. S. (2005). Theoretical developments in the study of college student departure. In A. Seidman (Ed.), *College student retention: Formula for student success* (pp. 61-87). Westport, CT: Greenwood.

Braxton, J. M., Hirschy, A. S., & McClendon, S. A. (2004). *Understanding and reducing college student departure* (ASHE-ERIC Higher Education Research Report Series, Vol. 30, No. 3). San Francisco, CA: Jossey-Bass.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Monterey, CA: Wadsworth.

Brooks-Leonard, C. (1991). Demographic and academic factors associated with first-to-second-term retention in a two-year college. *Community/Junior College Quarterly of Research and Practice*, *15*(1), 57-69.

Burns, R. B., & Burns, R. A. (2008). *Business research methods and statistics using SPSS*. London: Sage Publications.

Carey, K. (2017, October 31). Revised data shows community colleges have been underappreciated. *The New York Times*.

Carnevale, A. P., Smith, N., Melton, M., & Price, E. W. (2015). *Learning while earning: The new normal*. Washington, DC: Center on Education and the Workforce.

Carroll, C. D. (1989). *College persistence and degree attainment for 1980 high school graduates: Hazards for transfers, stopouts, and part-timers* (NCES 89–302). Washington, DC: National Center for Education Statistics.

Center for Community College Student Engagement. (2015). *Engagement rising: A decade of CCSSE data shows improvements across the board.* Austin, TX: The University of Texas at Austin, Program in Higher Education Leadership.

Chen, J. C. (2014). Teaching nontraditional adult students: Adult learning theories in practice. *Teaching in Higher Education, 19*, 406-418.

Chen, J. C. (2017). Nontraditional adult learners: The neglected diversity in postsecondary education. *SAGE Open, 7*(1), 1-12. doi:10.1177/2158244017697161

Chen, S., & Thomas, H. (2001). Constructing vocational and technical college student persistence models. *Journal of Vocational Education Research*, *26*(1), 26-55.

Choy, S. (2002). *Findings from the condition of education 2002: Nontraditional undergraduates*. Washington, DC: National Center for Education Statistics.

Cleveland-Innes, M. (1994). Adult student dropout at postsecondary institutions. *Review of Higher Education, 17*, 423-445.

Cochran, J. D., Campbell, S. M., Baker, H. M., & Leeds, E. M. (2013). The role of student characteristics in predicting retention in online courses. *Research in Higher Education, 55*(1), 27-48.

Complete College America (n.d.). A better deal for returning adults. Retrieved from https://completecollege.org/strategy/adult-learners-strategy/

Complete College Georgia. (2011). *Georgia's higher education completion plan 2012*. Atlanta, GA: The Governor's Office of Student Achievement.

Copper, B. (2017). Changing demographics: Why nontraditional students should matter to enrollment managers and what they can do to attract them. *AACRAO Consulting*. Retrieved from http://consulting.aacrao.org/publications-events/publications/changing-demographics-why-nontraditional-students-should-matter-to-enrollment-managers-and-what-they-can-do-to-attract-them/

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297.

Council for the Advancement of Standards in Higher Education. (2014). CAS Professional Standards. Retrieved from http://standards.cas.edu/getpdf.cfm?PDF=E864D2C4-D655-8F74-2E647CDECD29B7D0

Craig, A. J., & Ward, C. L. (2008). Retention of community college students: Related student and institutional characteristics. *Journal of College Student Retention: Research, Theory and Practice, 9*(4), 505-517.

Cranton, P., & Taylor, E. W. (2012). Transformative learning theory: Seeking a more unified theory. In E. W. Taylor & P. Cranton (Eds.), *The handbook of transformative learning: Theory, research, and practice* (pp. 3-20). San Francisco, CA: Jossey-Bass.

Crisp, G., & Mina, L. (2012). The community college: Retention trends and issues. In A. Seidman (Ed.), *College student retention: Formula for student success* (2nd ed., pp.147-165). Lanham, MD: Rowman & Littlefield Publishers, Inc.

Daempfle, P. A. (2003). An analysis of the high attrition rates among first year science, math, and engineering majors. *Journal of College Student Retention Research, Theory, and Practice, 5*(1), 37-52.

Demetriou, C. & Schmitz-Sciborski, A. (2011). Integration, motivation, strengths and optimism: Retention theories past, present and future. In R. Hayes (Ed.), *Proceedings of the 7th National Symposium on Student Retention, 2011, Charleston* (pp. 300-312). Norman, OK: The University of Oklahoma.

Doyle, S. R. & Donovan, D. M. (2014). Applying an ensemble classification tree approach to the prediction of completion of a 12-step facilitation intervention with stimulant abusers. *Psychology of Addictive Behaviors, 28*(4), 1127-1143. doi:10.1037/a0037235

Ellucian. (n.d.). Student Information Systems. Retrieved from https://www.ellucian.com/Software/Student-Information-Systems/

Ely, E. E. (1997). The non-traditional student. Paper presented at the American Association of Community Colleges Annual Conference, Anaheim, CA.

Fain, P. (2012). *Overkill on remediation*. Retrieved from https://www.insidehighered.com/news/2012/06/19/complete-college-america-declares-war-remediation

Feldman, M. (1993). Factors associated with one-year retention in a community college. *Research in Higher Education*, *34*(4), 503-512.

Fike, D. S., & Fike, R. (2008). Predictors of first-year student retention in the community college. *Community College Review*, *36*(2), 68-88. doi:10.1177/00915522108320222

Geiser, S., & Santelices, M. V. (2007). *Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes*. University of California, Berkeley Center for Studies in Higher Education Research & Occasional Paper Series: CSHE.6.07.

Georgia State Workforce Investment Board. (2013). *Georgia integrated state plan*. Atlanta, GA, Author.

Georgia Student Finance Commission. (n.d.). Programs and regulations. Retrieved from https://gsfc.georgia.gov/hope#field_related_links-576-5

Gifford, D. D., Briceno-Perriott, J., & Mianzo, F. (2006). Locus of control: Academic achievement and retention in a sample university first-year students. *Journal of College Admission, 191*,18-25.

Ginder, S. A., Kelly-Reid, J. E., & Mann, F. B. (2017a). *Enrollment and employees in postsecondary institutions, fall 2015; and financial statistics and academic libraries, fiscal year 2015: First look* (NCES 2017-024). Washington, DC: National Center for Education Statistics.

Ginder, S. A., Kelly-Reid, J. E., & Mann, F. B. (2017b). *Graduation rates for selected cohorts, 2008–13; outcome measures for cohort year 2008; student financial aid, academic year 2015–16; and admissions in postsecondary institutions, fall 2016:*

First look (provisional data) (NCES 2017-150rev). Washington, DC: National Center for Education Statistics.

Goncalves, S. A., & Trunk, D. (2014). Obstacles to success for the nontraditional student in higher education. *Psi Chi Journal of Psychological Research*, *19*(4), 164-172.

Grabowski, C., Rush, M., Ragen, K., Fayard, V., & Watkins-Lewis, K. (2016). Today's non-traditional student: Challenges to academic success and degree completion. *Inquiries Journal, 8*(3), 1-2.

Habley, W. R. (Ed.). (2004). *The status of academic advising: Findings from the ACT sixth national survey* (Monograph No. 10). Manhattan, KS: National Academic Advising Association.

Halpin, R. L. (1990). An application of the Tinto model to the analysis of freshman persistence in a community college. *Community College Review*, *17*(4), 22-32.

Hamilton, J. (1998). *First-time students entering a two-year public college with a GED, fall 1991 to fall 1996*. Gainesville, GA: Gainesville College. Retrieved from ERIC database. (ED415938)

Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-a-vis regression. *New Directions for Institutional Research, 131*, 17-33.

Hirschy, A., Bremer, C., & Castellano, M. (2011). Career and technical education (CTE) student success in community colleges: A conceptual model. *Community College Review, 39*(3), 296-318.

Hittepole, C. (n.d.). *Nontraditional students: Supporting changing student populations*. Retrieved from https://www.naspa.org/images/uploads/main/Hittepole_NASPA_Memo.pdf

Hodara, M., & Lewis, K. (2017). *How well does high school grade point average predict college performance by student urbanicity and timing of college entry?* (REL 2017-250). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northwest. Retrieved from http://ies.ed.gov/ncee/edlabs

Horn, L. J. (1996). *Nontraditional undergraduates: Trends in enrollment from 1986 to 1992 and persistence and attainment among 1989–90 beginning postsecondary students* (NCES 97–578). Washington, DC: National Center for Education Statistics.

Horn, L. J., Cataldi, E. F., & Sikora, A. (2005). *Waiting to attend college: Undergraduates who delay their postsecondary enrollment* (NCES 2005–152). Washington, DC: National Center for Education Statistics.

Horn, L. J., & Li, X. (2009). *Changes in postsecondary awards below the bachelor's degree: 1997 to 2007* (NCES 2010-167). Washington, DC: National Center for Education Statistics.

Horn, L., Peter, K., & Rooney, K. (2002). *Profile of undergraduates in U.S. postsecondary institutions: 1999-2000* (NCES 2002-168). Washington, DC: National Center for Education Statistics.

Houland, M., Crockett, D., McGuire, W., & Anderson, E. C. (1997). *Academic advising for student success and retention*. Iowa City, IA: Noel-Levitz.

Hunter, M. S., & White, E. R. (2004). Could fixing academic advising fix higher education? *About Campus, 9*(1), 20-25.

Hurtado, S., Kurotsuchi, K., & Sharp, S. (1996). *College entry by age groups: Path of traditional delayed-entry, and nontraditional students*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

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

Janitza, S., & Hornung, R. (2018). On the overestimation of random forest's out-of-bag error. *PLoS ONE, 13*(8). doi:10.1371/journal.pone.0201904

Jia, J., & Mareboyana, M. (2013). Machine learning algorithms and predictive models for undergraduate student retention. *World Congress on Engineering and Computer Science, 1*, 23-25.

Jones, D. J., & Watson, B. C. (1990). *"High risk" students and higher education: Future trends* (ASHE-ERIC Higher Education Report No.3). Washington, DC: George Washington University, School of Education and Human Development.

Juszkiewicz, J. (2017). *Trends in community college enrollment and completion data, 2017.* Washington, DC: American Association of Community Colleges.

Kaiser, L., Meyers, J., Morrison, D., & Skelton, A. (2016). *Machine learning to predict student retention*. Retrieved from http://websites.uwlax.edu/schen/resources/Student-Retention-Project.pdf

Keith, P. M. (2007). Barriers and nontraditional students' use of academic and social services. *College Student Journal, 41*(4), 1123-1127.

Kenner, C., & Weinerman, J. (2011). Adult learning theory: Applications to nontraditional college students. *Journal of College Reading and Learning, 41*, 87-96.

Kim, K. (2002). ERIC Review: Exploring the meaning of "Nontraditional" at the community college. *Community College Review, 30*(1), 74-89.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Knowledge Management System. (2014). KMS report descriptions. Retrieved from https://kms.tcsg.edu/DPR/DPR/Default.aspx

Knowles, J. E. (2014). *Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin*. Madison, WI: Wisconsin Department of Public Instruction.

Kreighbaum, A. (2017, June 20). *Year-round Pell grants available July 1*. Retrieved from https://www.insidehighered.com/quicktakes/2017/06/20/yearround-pell-grants-available-july-1

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

Kuhn, M., & Johnson, K. (2020). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Boca Raton, FL: Chapman & Hall/CRC Press.

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

Lee, J. (2017). *Georgia higher education data book*. Atlanta, GA: Georgia Budget and

Policy Institute.

Levine, A. (1993). Student expectations of college [Editorial]. *Change: The Magazine of Higher Learning, 25*(5), 4. doi:10.1080/00091383.1993.9939896

Lumina Foundation. (2015). *Who is today's student?* Indianapolis, IN: Author.

MacDonald, K. (2018). A review of the literature: The needs of nontraditional students in postsecondary education. *Strategic Enrollment Management Quarterly, 5*(4), 159-164. doi:10.1002/sem3.20115

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016) Models for early prediction of at-risk students in a course using standards-based grading, *Computer & Education, 103,* 1-15.

Markle, G. (2015). Factors influencing persistence among nontraditional university students. *Adult Education Quarterly*, *65*(3), 267-285. doi:10.1177/0741713615583085

McGrath, M. M., & Braunstein, A. (1997). The prediction of freshmen attrition: An examination of the importance of certain demographic, academic, financial, and social factors. *College Student Journal*, *31*(3), 1-11.

McNeely, J. H. (1937). *College student mortality* (U.S. Office of Education, Bulletin 1937, No. 11). Washington, DC: U.S. Government Printing Office.

Mendez, G., Buskirk, T. D., Lohr, S., & Haag, S. (2013). Factors associated with persistence in science and engineering majors: An exploratory study using classification trees and random forests. *The Research Journal for Engineering Education, 97*(1), 57-70. doi:10.1002/j.2168-9830.2008.tb00954

Merriam, B. S., Cafarella, S. R., & Baumgartner, M. L. (2007). *Learning in adulthood: A comprehensive guide*. San Francisco, CA: John Wiley & Sons.

Metzner, B. S., & Bean, J. P. (1987). The estimation of a conceptual model of nontraditional undergraduate student attrition. *Research in Higher Education, 27*(1), 15-38.

Mezirow, J. (1997). Transformative learning: Theory to practice. In P. Cranton (Ed.), *Transformative learning in action: Insights from practice. New directions in adult and continuing education* (pp. 5-12). San Francisco, CA: Jossey-Bass.

Mohammadi, J. (1994). *Exploring retention and attrition in a two-year public community college*. Martinsville, VA: Patrick Henry Community College. Retrieved from ERIC database. (ED382257)

Monroe, A. (2006). Non-traditional transfer student attrition. *The Community College Enterprise, 12*(2), 33-54.

Morris, L. V., Wu, S., & Finnegan, C. L. (2005) Predicting retention in online general education courses. *The American Journal of Distance Education, 19*(1), 23-26.

Murtaugh, P., Burns, L., & Schuster, J. (1999). Predicting the retention of university students. *Research in Higher Education, 40*(3), 355-371.

Nakajima, M. A., Dembo, M. H., & Mossler, R. (2012). Student persistence in community colleges. *Community College Journal of Research and Practice*, *36*(8), 591-613. doi:10.1080/10668920903054931

National Center for Education Statistics [NCES]. (1996). Nontraditional undergraduates/definitions and data. Retrieved from https://nces.ed.gov/pubs/web/97578e.asp

National Student Clearinghouse [NSC]. (2017a). Snapshot report: First-year persistence and retention. Retrieved from https://nscresearchcenter.org/wp-content/uploads/SnapshotReport28a.pdf

National Student Clearinghouse [NSC]. (2017b). The role of community colleges in postsecondary success. Retrieved from https://studentclearinghouse.info/onestop/wpcontent /uploads/Comm-Colleges-Outcomes-Report.pdf

Nora, A. (1990). Campus-based aid programs as determinants of retention among Hispanic community college students. *The Journal of Higher Education, 61*(3), 312-331. doi:10.2307/1982133

Nora, A., Barlow, L., & Crisp, G. (2005). Student persistence and degree attainment beyond the first year in college. In A. Seidman (Ed.), *College student retention: Formula for success* (pp. 129-153). Westport, CT: Praeger.

Oden, L. M. (2011). *Factors affecting persistence of nontraditional students enrolled in two year colleges* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3465700)

Panos, R. J., & Astin, A. W. (1968). Attrition among college students. *American Educational Research Journal, 1*, 57-72.

Pascarella, E. T., Duby, P. B., Miller, V. A., & Rasher, S. P. (1981). Preenrollment variables and academic performance as predictors of freshman year persistence,

early withdrawal, and stopout behavior in an urban, nonresidential university. *Research in Higher Education*, *15*(4), 329-349. doi:10.1007/BF00973513

Pascarella, E. T., & Terenzini, P. T. (1991). *How college affects students: Findings and insights from twenty years of research*. San Francisco, CA: Jossey-Bass.

Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: a third decade of research*. San Francisco, CA: Jossey-Bass.

Peduzzi P., Concato J., Kemper E., Holford T. R., Feinstein A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology, 49*, 1373-1379.

Pelletier, S. G. (2010). Success for adult students. *Public Purpose*. Retrieved from https://www.aascu.org/uploadedFiles/AASCU/Content/Root/MediaAndPublications/PublicPurposeMagazines/Issue/10fall_adultstudents.pdf

Peltier, G. L., Laden, R., & Matranga, M. (1999). Student persistence in college: A review of research. *Journal of College Student Retention, 1*(4), 357-375.

Perna, L. W. (1998). The contribution of financial aid to undergraduate persistence. *Journal of Student Financial Aid, 28*(3), 25-40.

Peters, L., Hyun, M., Taylor, S., & Varney, J. (2010). Advising non-traditional students: Beyond class schedules and degree requirements. *Academic Advising Today, 33*(3).

Petty, T. (2014). Motivating first-generation students to academic success and college completion. *College Student Journal, 48*(2), 257-264.

Pingel, S., Parker, E., & Sisneros, L. (2016). *Free community college: An approach to increase adult student success in postsecondary education*. Denver, CO: Education Commission of the States.

Pritchard, M. E., & Wilson, G. S. (2003). Using emotional and social factors to predict student success. *Journal of College Student Development*, *44*(1), 18-28.

Radford, A., Cominole, M., & Skomsvold, P. (2015). *Demographic and enrollment characteristics of nontraditional undergraduates: 2011-12* (NCES no. 2015-25). Washington, DC: National Center for Education Statistics.

Radwin, D., Conzelmann, J. G., Nunnery, A., Lacy, T. A., Wu, J., Lew, S., Wine, J., & Siegel, P. (2018). *2015–16 National Postsecondary Student Aid Study (NPSAS:16): Student Financial Aid Estimates for 2015–16* (NCES 2018-466). Washington, DC: National Center for Education Statistics.

Reason, R. D. (2009). An examination of persistence research through the lens of a comprehensive conceptual framework. *Journal of College Student Development, 50*(6), 659-682. doi:10.1353/csd.0.0098

Reeves, T. J., Miller, L. A., & Rouse, R. A. (2011). *Reality check: A vital update to the landmark 2002 NCES study of nontraditional college students*. Phoenix, AZ: Apollo Research Institute.

Ross-Gordon, J. M. (2011). Research on adult learners: Supporting the needs of a student population that is no longer nontraditional. *Peer Review, 13*, 26-29.

Schneider, M., & Yin, L. (2011). *The hidden costs of community colleges*. Washington, DC: American Institutes for Research.

Seidman, A. (1993). Needed: A research methodology to assess community college effectiveness. *Community College Journal, 63*(5), 36-40.

Seidman, A. (Ed.). (2005). *College student retention: formula for student success*. Westport, CT: ACE/Praeger.

Shapiro, D., Dundar, A., Wakhungu, P. K., Yuan, X., Nathan, A., & Hwang, Y. (2016). *Completing college: A national view of student attainment rates-fall 2010 cohort* (Signature Report No. 12). Herndon, VA: National Student Clearinghouse Research Center.

Simonton, D. K. (2003). Qualitative and quantitative analyses of historical data. *Annual Review of Psychology, 54*(1), 617–640.

Singell, L. D. (2004). Come and stay a while: Does financial aid effect retention conditioned on enrollment at a large public university? *Economics of Education Review, 23*(5), 459-471.

Singell, L. D., & Waddell, G. R. (2010). Modeling retention at a large public university: Can at-risk students be identified early enough to treat? *Research in Higher Education, 51*(6), 546-572.

Soares, L., Gagliardi, J. S., & Nellum, C. J. (2017). *The post-traditional learners manifesto revisited: Aligning postsecondary education with real life for adult student success*. Retrieved from http://www.acenet.edu/news-room/Documents/The-Post-Traditional-Learners-Manifesto-Revisited.pdf

Southeastern Technical College. (n.d.). Special populations. Retrieved from http://www.southeasterntech.edu/student-affairs/special-populations.php

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.

Spicer, J. (2005). *Making sense of multivariate data analysis*. Thousand Oaks, CA: Sage.

St. John, E. P., Hu, S., Simmons, A., Carter, D. F., & Weber, J. (2004). What difference does a major make? The influence of college major field on persistence by African American and White students. *Research in Higher Education, 45*(3), 209-232. doi:10.1023/B:RIHE.0000019587.46953.9d

Stewart, S., Doo, H. L., & Kim, J. (2015). Factors influencing college persistence for first-time students. *Journal of Developmental Education, 38*(3), 12-20.

Swail, W. S. (1995). *A conceptual framework for student retention in science, engineering, and mathematics* (Unpublished doctoral dissertation). George Washington University, Washington, DC.

Swail, W. S. (2004). *The art of student retention: A handbook for practitioners and administrators*. Austin, TX: Educational Policy Institute.

Swail, W. S., Redd, K. E., & Perna, L. W. (2003). *Retaining minority students in higher education: A framework for success* (ASHE-ERIC Higher Education Research Report Series, Vol. 30, No. 2). San Francisco, CA: Jossey-Bass.

Swift, J. S., Colvin, C., & Mills, D. (1987). Displaced homemakers: Adults returning to college with different characteristics and needs. *Journal of College Student Personnel, 28*(4), 343-349.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

Taniguchi, H., & Kaufman, G. (2005). Degree completion among nontraditional college students. *Social Science Quarterly, 86*(4), 912-927.

Technical College System of Georgia. (2016). *Performance accountability system manual*. Atlanta, GA: Author.

Technical College System of Georgia. (2017). *End-of-year enrollment report 2016*. Atlanta, GA: Author.

Technical College System of Georgia. (2018). *State Board Policies and TCSG Procedures Manual*. Atlanta, GA: Author.

Terenzini, P. T., & Pascarella, E. T. (1978). The relation of students' pre-college characteristics and freshman year experience to voluntary attrition. *Research in Higher Education*, *9*(4), 347-366.

Terenzini, P. T., & Reason, R. D. (2005). *Parsing the first year of college: Rethinking the effects of college on students*. Paper presented at the Annual Conference of the Association for the Study of Higher Education, Philadelphia, PA.

Tinto, V. (1975). Dropouts from higher education: A theoretical synthesis of recent literature. *A Review of Educational Research, 45*, 89-125.

Tinto, V. (1986). Theories of student departure revisited. In J. Smart (Ed.), *Higher education: A handbook of theory and research* (pp. 359-384). New York, NY: Agathon.

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition* (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. (2006). Research and practice of student retention: What next? *Journal of College Student Retention, 8*(1), 1-19.

Titus, M. A. (2004). An examination of the influence of institutional context on student persistence at 4-year colleges and universities: A multilevel approach. *Research in Higher Education*, *45*(7), 673-699.

Titus, M. A. (2006). Understanding the influence of the financial context of institutions on student persistence at four-year colleges and universities. *The Journal of Higher Education*, *77*(2), 353-375.

Tuma, J., & Geis, S. (1995). *Educational attainment of 1980 high school sophomores by 1992* (NCES 95–304). Washington, DC: National Center for Education Statistics.

Turk, J. M., & Chen, W. (2017). *Improving the odds: An empirical look at the factors that influence upward transfer*. Washington, DC: American Council on Education.

University System of Georgia (n.d.). *USG by the numbers*. Retrieved from http://www.info.usg.edu/

U. S. Department of Education. (2011). *Committee on measures of student success: A report to secretary of education Arne Duncan*. Washington, DC: Author.

Varney, J. (2013). Proactive advising. In J. K. Drake, P. Jordan, & M. A. Miller (Eds.), *Academic advising approaches: Strategies that teach students to make the most of college* (pp. 137-154). Manhattan, KS: NACADA: The Global Community for Academic Advising.

Watt, C., & Wagner, E. (2016). *Improving post-traditional student success* [White paper]. Retrieved from https://www.hobsons.com/res/Whitepapers/PostTraditionalStudents_ParFramework_February2016.pdf

Westervelt, E. (2016). Shaken by economic change, 'non-traditional' students are becoming the new normal. *National Public Radio*. Retrieved from http://www.npr.org/sections/ed/2016/09/25/495188445/shaken-by-economic-change-non-traditional-students-are-becoming-thenew-normal

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R.,…Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. doi:10.21105/joss.01686

Wild, L., & Ebbers, L. (2002). Rethinking student retention in community colleges. *Community College Journal of Research and Practice*, *26*, 503-519.

Wilson, G., Epps, D., Tanner, D., Gordon, R., & Sig, T. J. (2014). *Governor's high demand career initiative report*. Athens, GA: Carl Vinson Institute of Government.

Wolf, D. S. (2011). *Uncovering the complexity of student-family support systems and their subsequent influence on the persistence of underserved college students* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3452117)

Wyatt, L. (2011). Nontraditional student engagement: Increasing adult student success and retention. *The Journal of Continuing Higher Education, 59*(1), 10-20. doi:10.1080/07377363.2011.544977

Wyckoff, S. (1998). Retention theories in higher education: Implications for institutional practice. *Recruitment and Retention in Higher Education, 12*(2), 2-7.

Wyman, F. J. (1997). A predictive model of retention rate at regional two-year colleges. *Community College Review, 25*(1), 29-58.

Wyner, J. S. (2014). *What excellent community colleges do: Preparing all students for success*. Cambridge, MA: Harvard Education Press.

Yu, C. H., DiGangi, S., Jannasch-Pennell, A., & Kaprolet, C. (2010). A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science, 8*, 307-325.

APPENDIX A:

Institutional Review Board Protocol Exemption Report

Institutional Review Board (IRB)
For the Protection of Human Research Participants

PROTOCOL EXEMPTION REPORT

Protocol Number: 03798-2019

Responsible Researchers: Brandy Taylor

Supervising Faculty: Dr. Lantry Brockmeier

Project Title: *The Predictability of Nontraditional Student Retention in the Technical College System of Georgia.*

INSTITUTIONAL REVIEW BOARD DETERMINATION:

This research protocol is **Exempt** from Institutional Review Board (IRB) oversight under Exemption **Category 2**. Your research study may begin immediately. If the nature of the research project changes such that exemption criteria may no longer apply, please consult with the IRB Administrator (irb@valdosta.edu) before continuing your research.

ADDITIONAL COMMENTS:

- *Upon completion of the research study all data (data list, email correspondence, etc.) must be securely maintained (locked file cabinet, password protected computer, etc.) and accessible only by the researchers for a minimum of 3 years.*

☒ ***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 *04.01.2019*
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:

R Code for Data Analysis

R is a powerful programming language used for statistical computing and data analysis. R is an open source language and code development is ongoing. In particular, the tidyverse packages have experienced tremendous popularity and advanced packages are continuously being developed. Because R is an ever-evolving language, the following code was included to both document the code used at the time of this study as well as assist future research using R for data analysis.

```
library(car)
library(devtools)
library(modeldata)
library(recipes)
library(themis)
library(caret)
library(caTools)
library(corrplot)
library(FactoMineR)
library(Hmisc)
library(lsr)
library(mctest)
library(psych)
library(summarytools)
library(readxl)
library(kernlab)
library(rminer)
library(rsample)
library(skimr)
library(tibble)
library(VIM)
library(discrim)
library(vip)
library(vctrs)
library(tidyverse)
library(tidymodels)
library(themis)
library(DescTools)
library(jtools)
library(MASS)
library(ResourceSelection)
library(MKmisc)
library(usdm)
library(rpart.plot)
library(performance)
```

```
#import data/descriptive statistics
TCC_18_to_36_2017 <- read_excel("C:/Users/btaylor/OneDrive/VSU/Fall 2020/TCC18to36/TCC 18 to 36 - 2017.xlsx")
View(TCC_18_to_36_2017)
TCC_18_to_36_2018 <- read_excel("C:/Users/btaylor/OneDrive/VSU/Fall 2020/TCC18to36/TCC 18 to 36 - 2018.xlsx")
View(TCC_18_to_36_2018)
TCC18to36_2017 <- TCC_18_to_36_2017
TCC18to36_2018 <- TCC_18_to_36_2018
skimr::skim(TCC18to36_2017)
skimr::skim(TCC18to36_2018)
dfSummary(TCC18to36_2017)
dfSummary(TCC18to36_2018)
describe(TCC18to36_2017$age)
describe(TCC18to36_2017$gpa)
describe(TCC18to36_2018$age)
describe(TCC18to36_2018$gpa)
jpeg("rplot1.jpg", width = 350, height = 350)
hist(TCC18to36_2017$age)
dev.off()
jpeg("rplot2.jpg", width = 350, height = 350)
hist(TCC18to36_2017$gpa)
dev.off()
jpeg("rplot3.jpg", width = 350, height = 350)
hist(TCC18to36_2018$age)
dev.off()
jpeg("rplot4.jpg", width = 350, height = 350)
hist(TCC18to36_2018$gpa)
dev.off()
col_order <- c("age", "gpa", "racecode", "hsdipcode", "hopepos", "gencode", "gradcode", "sparcode", "dhomcode", "pellcode", "retained")
TCC18to36_2017reordered <- TCC18to36_2017[, col_order]
TCC18to36_2018reordered <- TCC18to36_2018[, col_order]
View(TCC18to36_2017reordered)
View(TCC18to36_2018reordered)
mixedCor(TCC18to36_2017reordered, c=1:2, p=3:5, d=6:11, correct = FALSE, smooth = TRUE, global = FALSE)
mixedCor(TCC18to36_2018reordered, c=1:2, p=3:5, d=6:11, correct = FALSE, smooth = TRUE, global = FALSE)
rcorr(as.matrix(TCC18to36_2017reordered))
rcorr(as.matrix(TCC18to36_2018reordered))
usdm::vif(as.data.frame(TCC18to36_2017reordered))
usdm::vif(as.data.frame(TCC18to36_2018reordered))
#####################################################################
myglm1 <- glm(retained ~ age + gpa, data = TCC18to36_2017, family = "binomial")
crPlots(myglm1)
#####################################################################
jpeg("rplot5.jpg", width = 350, height = 350)
```

```
boxplot(TCC18to36_2017$age)
dev.off()
boxplot(TCC18to36_2017$age, plot=FALSE)$out
jpeg("rplot6.jpg", width = 350, height = 350)
boxplot(TCC18to36_2017$gpa)
dev.off()
boxplot(TCC18to36_2017$gpa, plot=FALSE)$out
leveneTest(age ~ retained, TCC18to36_2017)
leveneTest(gpa ~ retained, TCC18to36_2017)
TCC18to36_2017$racecode<-as.factor(TCC18to36_2017$racecode)
TCC18to36_2017$gencode<-as.factor(TCC18to36_2017$gencode)
TCC18to36_2017$hsdipcode<-as.factor(TCC18to36_2017$hsdipcode)
TCC18to36_2017$gradcode<-as.factor(TCC18to36_2017$gradcode)
TCC18to36_2017$sparcode<-as.factor(TCC18to36_2017$sparcode)
TCC18to36_2017$dhomcode<-as.factor(TCC18to36_2017$dhomcode)
TCC18to36_2017$pellcode<-as.factor(TCC18to36_2017$pellcode)
TCC18to36_2017$retained<-as.factor(TCC18to36_2017$retained)
TCC18to36_2017$hopepos<-as.factor(TCC18to36_2017$hopepos)
TCC18to36_2018$racecode<-as.factor(TCC18to36_2018$racecode)
TCC18to36_2018$gencode<-as.factor(TCC18to36_2018$gencode)
TCC18to36_2018$hsdipcode<-as.factor(TCC18to36_2018$hsdipcode)
TCC18to36_2018$gradcode<-as.factor(TCC18to36_2018$gradcode)
TCC18to36_2018$sparcode<-as.factor(TCC18to36_2018$sparcode)
TCC18to36_2018$dhomcode<-as.factor(TCC18to36_2018$dhomcode)
TCC18to36_2018$pellcode<-as.factor(TCC18to36_2018$pellcode)
TCC18to36_2018$retained<-as.factor(TCC18to36_2018$retained)
TCC18to36_2018$hopepos<-as.factor(TCC18to36_2018$hopepos)

#create recipe
set.seed(100)
rec_obj_test <- recipe(retained ~ ., data = TCC18to36_2017) %>%
  step_knnimpute(all_predictors())
rec_obj_test
prep_obj_test <- prep(rec_obj_test, retain = TRUE)
prep_obj_test
juiced_test <- juice(prep_obj_test)

summary(TCC18to36_2017$hsdipcode)
summary(TCC18to36_2017$gradcode)
summary(juiced_test$hsdipcode)
summary(juiced_test$gradcode)

set.seed(100)
rec_obj <- recipe(retained ~ ., data = TCC18to36_2017) %>%
  step_knnimpute(all_predictors()) %>%
  step_YeoJohnson(all_numeric()) %>%
  step_dummy(all_predictors(), -all_numeric()) %>%
  step_normalize(all_numeric()) %>%
```

```
  step_corr(all_numeric(), threshold = .9) %>%
  step_zv(all_numeric()) %>%
  step_downsample(retained, skip = TRUE)
rec_obj

#all_nominal(), -all_numeric(), hsdipcode, one_hot = TRUE
#step_downsample(retained, skip = TRUE)
prep_obj <- prep(rec_obj, retain = TRUE)
prep_obj

juiced <- juice(prep_obj)
################################################################################
describe(juiced$age)
describe(juiced$gpa)

myglm2 <- glm(retained ~ age + gpa, data = juiced, family = "binomial")
crPlots(myglm2)
################################################################################
#baked <- bake(prep_obj, new_data = TCC18to36_2018)

#resampling/tuning
set.seed(300)
folds <- vfold_cv(TCC18to36_2017, strata = retained, v = 10, repeats = 5)

#glm
################################################################################
myglm <- glm(retained ~ ., data = juiced, family = "binomial")
performance_hosmer(myglm, n_bins = 16)
################################################################################

summary(myglm)
exp(cbind(OR = coef(myglm), confint(myglm)))
summ(myglm)
summ(myglm, scale = TRUE)
crPlots(myglm)

glm_spec <- logistic_reg() %>%
  set_engine("glm") %>%
  set_mode("classification")

#####################################################################
glm_wf <- workflow() %>%
  add_recipe(rec_obj) %>%
  add_model(glm_spec)

set.seed(400)
glm_rs2 <-
  glm_wf %>%
```

```
  tune_grid(resamples = folds,
            grid = 25,
            control = control_grid(save_pred = TRUE),
            metrics = metric_set(roc_auc, sens, spec))
glm_rs2

glm_rs2 %>%
  show_best(metric = "roc_auc")

best_glm <-
  glm_rs2 %>%
  select_best(metric = "roc_auc")
best_glm

glm_auc2 <-
  glm_rs2 %>%
  collect_predictions(parameters = best_glm) %>%
  roc_curve(retained, .pred_1, event_level = "second") %>%
  mutate(model = "Logistic Regression")
autoplot(glm_auc2)

set.seed(345)
last_glm_fit <-
  glm_wf %>%
  fit(TCC18to36_2017)
last_glm_fit
digits5 <- glance(last_glm_fit$fit$fit)
as.data.frame(digits5)
digits1 <- tidy(last_glm_fit)
as.data.frame(digits1)
digits2 <- tidy(last_glm_fit, exponentiate = TRUE, conf.int = TRUE)
as.data.frame(digits2)

set.seed(546)
last_glm_fit_test <-
  glm_wf %>%
  fit(TCC18to36_2018)
last_glm_fit_test
digits6 <- glance(last_glm_fit_test$fit$fit)
as.data.frame(digits6)
digits3 <- tidy(last_glm_fit_test)
as.data.frame(digits3)
digits4 <- tidy(last_glm_fit_test, exponentiate = TRUE, conf.int = TRUE)
as.data.frame(digits4)

jpeg("rplot7a.jpg", width = 350, height = 350)
last_glm_fit_test %>%
  pull_workflow_fit() %>%
```

```
  vip(geom = "col", num_features = 20) + ggtitle("Logistic Regression Importance")
dev.off()
######################################################################

glm_fit <- glm_spec %>%
  fit(retained ~ ., data = juiced)
glm_fit

jpeg("rplot7.jpg", width = 350, height = 350)
vip(glm_fit, num_features = 20) + ggtitle("Logistic Regression Importance")
dev.off()

tidy(glm_fit)
glance(glm_fit$fit)

set.seed(400)
glm_rs <- glm_spec %>%
  fit_resamples(
    rec_obj,
    folds,
    metrics = metric_set(roc_auc, sens, spec),
    control = control_resamples(save_pred = TRUE)
  )
glm_rs %>%
  collect_metrics()

glm_rs %>%
  show_best(metric = "roc_auc")

glm_auc <-
  glm_rs %>%
  collect_predictions() %>%
  roc_curve(retained, .pred_1, event_level = "second") %>%
  mutate(model = "Logistic Regression")
autoplot(glm_auc)

#lda
lda_spec <- discrim_linear(penalty = .1) %>%
  set_engine("MASS") %>%
  set_mode("classification")

######################################################################
lda_wf <- workflow() %>%
  add_recipe(rec_obj) %>%
  add_model(lda_spec)

set.seed(400)
lda_rs2 <-
```

```
 lda_wf %>%
 tune_grid(resamples = folds,
        grid = 25,
        control = control_grid(save_pred = TRUE),
        metrics = metric_set(roc_auc, sens, spec))
lda_rs2

lda_rs2 %>%
 show_best(metric = "roc_auc")

best_lda <-
 lda_rs2 %>%
 select_best(metric = "roc_auc")
best_lda

lda_auc2 <-
 lda_rs2 %>%
 collect_predictions(parameters = best_lda) %>%
 roc_curve(retained, .pred_1, event_level = "second") %>%
 mutate(model = "Linear Discriminant Analysis")
autoplot(lda_auc2)

set.seed(345)
last_lda_fit <-
 lda_wf %>%
 fit(TCC18to36_2017)
last_lda_fit

set.seed(501)
last_lda_fit_test <-
 lda_wf %>%
 fit(TCC18to36_2018)
last_lda_fit_test

absvalue <- abs(last_lda_fit_test$fit$fit$fit$scaling)
jpeg("rplot8a.jpg", width = 350, height = 350)
mgraph(absvalue,graph="IMP",col="gray",main = "LDA Importance")
dev.off()
#####################################################################

lda_fit <- lda_spec %>%
 fit(retained ~ ., data = juiced)
lda_fit

absvalue <- abs(lda_fit$fit$scaling)
jpeg("rplot8.jpg", width = 350, height = 350)
mgraph(absvalue,graph="IMP",leg=names(juiced),col="gray",Grid=10, main = "LDA
Importance")
```

```
dev.off()

set.seed(400)
lda_rs <- lda_spec %>%
  fit_resamples(
    rec_obj,
    folds,
    metrics = metric_set(roc_auc, sens, spec),
    control = control_resamples(save_pred = TRUE)
  )
lda_rs %>%
  collect_metrics()

lda_rs %>%
  show_best(metric = "roc_auc")

lda_auc <-
  lda_rs %>%
  collect_predictions() %>%
  roc_curve(retained, .pred_1, event_level = "second") %>%
  mutate(model = "LDA")
autoplot(lda_auc)

#tree
#tried to tune min_n
tree_spec <- decision_tree(cost_complexity = tune(), tree_depth = tune()) %>%
  set_engine("rpart") %>%
  set_mode("classification")
tree_spec

tree_wf <- workflow() %>%
  add_recipe(rec_obj) %>%
  add_model(tree_spec)

tree_grid <- grid_regular(cost_complexity(),
                          tree_depth(),
                          levels = 5)

set.seed(400)
tree_rs <-
  tree_wf %>%
  tune_grid(resamples = folds,
            grid = 25,
            control = control_grid(save_pred = TRUE),
            metrics = metric_set(roc_auc, sens, spec))
tree_rs

tree_rs %>%
```

```
  show_best(metric = "roc_auc")

best_tree <-
  tree_rs %>%
  select_best(metric = "roc_auc")
best_tree

tree_auc <-
  tree_rs %>%
  collect_predictions(parameters = best_tree) %>%
  roc_curve(retained, .pred_1, event_level = "second") %>%
  mutate(model = "Decision Tree")
autoplot(tree_auc)

last_tree_mod <-
  decision_tree(cost_complexity = 0.0000793, tree_depth = 4) %>%
  set_engine("rpart") %>%
  set_mode("classification")

last_tree_workflow <-
  tree_wf %>%
  update_model(last_tree_mod)

set.seed(345)
last_tree_fit <-
  last_tree_workflow %>%
  fit(TCC18to36_2017)
last_tree_fit

jpeg("rplot9.jpg", width = 350, height = 350)
last_tree_fit %>%
  pull_workflow_fit() %>%
  vip(geom = "col", num_features = 20) + ggtitle("Classification Tree Importance")
dev.off()

rpart.plot(last_tree_fit$fit$fit$fit)
summary(last_tree_fit$fit$fit$fit)
last_tree_fit$fit$fit$fit$splits

#####################################################################
set.seed(345)
last_tree_fit <-
  last_tree_workflow %>%
  fit(TCC18to36_2017)
last_tree_fit

set.seed(359)
last_tree_fit_test <-
```

```
  last_tree_workflow %>%
  fit(TCC18to36_2018)
last_tree_fit_test

jpeg("rplot9.jpg", width = 350, height = 350)
last_tree_fit %>%
  pull_workflow_fit() %>%
  vip(geom = "col", num_features = 20) + ggtitle("Classification Tree Importance")
dev.off()

rpart.plot(last_tree_fit$fit$fit$fit)
summary(last_tree_fit$fit$fit$fit)
last_tree_fit$fit$fit$fit$splits
last_tree_fit$fit$fit$fit$variable.importance

jpeg("rplot9a.jpg", width = 350, height = 350)
last_tree_fit_test %>%
  pull_workflow_fit() %>%
  vip(geom = "col", num_features = 20, scale = FALSE) + ggtitle("Classification Tree Importance")
dev.off()

rpart.plot(last_tree_fit_test$fit$fit$fit, roundint = FALSE)
summary(last_tree_fit_test$fit$fit$fit)
last_tree_fit_test$fit$fit$fit$splits
last_tree_fit_test$fit$fit$fit$variable.importance
##############################################################################

#rand
rand_spec <- rand_forest(mtry = tune(), min_n = tune()) %>%
  set_engine("randomForest") %>%
  set_mode("classification")
rand_spec

rand_wf <- workflow() %>%
  add_recipe(rec_obj) %>%
  add_model(rand_spec)

set.seed(345)
rf_res <-
  rand_wf %>%
  tune_grid(resamples = folds,
            grid = 25,
            control = control_grid(save_pred = TRUE),
            metrics = metric_set(roc_auc, sens, spec))

rf_res %>%
  show_best(metric = "roc_auc")
```

```
rf_best <-
  rf_res %>%
  select_best(metric = "roc_auc")
rf_best

rf_auc <-
  rf_res %>%
  collect_predictions(parameters = rf_best) %>%
  roc_curve(retained, .pred_1, event_level = "second") %>%
  mutate(model = "Random Forest")
autoplot(rf_auc)

last_rf_mod <-
  rand_forest(mtry = 6, min_n = 39) %>%
  set_engine("randomForest") %>%
  set_mode("classification")

last_rf_workflow <-
  rand_wf %>%
  update_model(last_rf_mod)

########################################################################
set.seed(345)
last_rf_fit <-
  last_rf_workflow %>%
  fit(TCC18to36_2017)
last_rf_fit

jpeg("rplot10.jpg", width = 350, height = 350)
last_rf_fit %>%
  pull_workflow_fit() %>%
  vip(geom = "col", num_features = 20) + ggtitle("Random Forest Importance")
dev.off()

lastrf_obj <- pull_workflow_fit(last_rf_fit)$fit
lastrf_obj$importance

set.seed(500)
last_rf_fit_test <-
  last_rf_workflow %>%
  fit(TCC18to36_2018)
last_rf_fit_test

jpeg("rplot10a.jpg", width = 350, height = 350)
last_rf_fit_test %>%
  pull_workflow_fit() %>%
  vip(geom = "col", num_features = 20) + ggtitle("Random Forest Importance")
dev.off()
```

```
lastrf_obj <- pull_workflow_fit(last_rf_fit_test)$fit
lastrf_obj$importance
#############################################################################

set.seed(345)
last_rf_fit <-
  last_rf_workflow %>%
  fit(TCC18to36_2017)
last_rf_fit

jpeg("rplot10.jpg", width = 350, height = 350)
last_rf_fit %>%
  pull_workflow_fit() %>%
  vip(geom = "col", num_features = 20) + ggtitle("Random Forest Importance")
dev.off()

lastrf_obj <- pull_workflow_fit(last_rf_fit)$fit
lastrf_obj$importance

#svm
svm_spec <- svm_rbf(rbf_sigma = tune(), cost = tune()) %>%
  set_engine("kernlab") %>%
  set_mode("classification")
svm_spec

svm_wf <- workflow() %>%
  add_recipe(rec_obj) %>%
  add_model(svm_spec)

svm_grid <- expand.grid(rbf_sigma = c(0.1,0.5,1,2), cost = c(0.1,1,10,100))
doParallel::registerDoParallel()

set.seed(500)
svm_rs <-
  svm_spec %>%
  tune_grid(
    rec_obj,
    resamples = folds,
    grid = svm_grid,
    control = control_grid(save_pred = TRUE, verbose = FALSE),
    metrics = metric_set(roc_auc, sens, spec)
  )
svm_rs

svm_rs %>%
  show_best(metric = "roc_auc")
```

```
svm_best <-
  svm_rs %>%
  select_best(metric = "roc_auc")
svm_best

svm_auc <-
  svm_rs %>%
  collect_predictions(parameters = svm_best) %>%
  roc_curve(retained, .pred_1, event_level = "second") %>%
  mutate(model = "SVM")
autoplot(svm_auc)

last_svm_mod <-
  svm_rbf(rbf_sigma = 0.1, cost = 0.1) %>%
  set_engine("kernlab") %>%
  set_mode("classification")

last_svm_workflow <-
  svm_wf %>%
  update_model(last_svm_mod)

########################################################################
set.seed(345)
last_svm_fit <-
  last_svm_workflow %>%
  fit(TCC18to36_2017)
last_svm_fit

jpeg("rplot11a.jpg", width = 350, height = 350)
last_svm_fit %>%
  pull_workflow_fit() %>%
  vip(method = "permute",
      target = "retained", metric = "auc", reference_class = "1",
      pred_wrapper = kernlab::predict, train = juiced, num_features = 20) + ggtitle("Support Vector
Machine Importance")
dev.off()

last_svm_fit %>%
  pull_workflow_fit() %>%
  vi(method = "permute",
     target = "retained", metric = "auc", reference_class = "1",
     pred_wrapper = kernlab::predict, train = juiced)

set.seed(501)
last_svm_fit_test <-
  last_svm_workflow %>%
  fit(TCC18to36_2018)
last_svm_fit_test
```

```
vi_scores <- last_svm_fit_test %>%
  pull_workflow_fit() %>%
  vi(method = "permute",
     target = "retained", metric = "auc", reference_class = "1",
     pred_wrapper = kernlab::predict, train = juiced)
vi_scores

jpeg("rplot11b.jpg", width = 350, height = 350)
vip(vi_scores, num_features = 20) + ggtitle("Support Vector Machine Importance")
dev.off()
##############################################################################

set.seed(345)
last_svm_fit <-
  last_svm_workflow %>%
  fit(TCC18to36_2017)
last_svm_fit

jpeg("rplot11.jpg", width = 350, height = 350)
last_svm_fit %>%
  pull_workflow_fit() %>%
  vip(method = "permute",
      target = "retained", metric = "auc", reference_class = "1",
      pred_wrapper = kernlab::predict, train = juiced, num_features = 20) + ggtitle("Support Vector
Machine Importance")
dev.off()

last_svm_fit %>%
  pull_workflow_fit() %>%
  vi(method = "permute",
     target = "retained", metric = "auc", reference_class = "1",
     pred_wrapper = kernlab::predict, train = juiced)

#roccurve for each model
jpeg("rplot12.jpg", width = 350, height = 350)
glm_rs %>%
  unnest(.predictions) %>%
  mutate(model = "glm") %>%
  bind_rows(lda_rs %>%
              unnest(.predictions) %>%
              mutate(model = "MASS")) %>%
  bind_rows(tree_rs %>%
              unnest(.predictions) %>%
              mutate(model = "rpart")) %>%
  bind_rows(rf_res %>%
              unnest(.predictions) %>%
              mutate(model = "randomForest")) %>%
```

```
  bind_rows(svm_rs %>%
              unnest(.predictions) %>%
              mutate(model = "kernlab")) %>%
  group_by(model) %>%
  roc_curve(retained, .pred_1, event_level = "second") %>%
  ggplot(aes(x = 1 - specificity, y = sensitivity, color = model)) +
  geom_line(size = 1.5) +
  geom_abline(
    lty = 2, alpha = 0.5,
    color = "gray50",
    size = 1.2)
dev.off()

jpeg("rplot13.jpg", width = 350, height = 350)
glm_auc %>%
  mutate(model = "glm") %>%
  bind_rows(lda_auc %>%
            mutate(model = "MASS")) %>%
  bind_rows(tree_auc %>%
              mutate(model = "rpart")) %>%
  bind_rows(rf_auc %>%
              mutate(model = "randomForest")) %>%
  bind_rows(svm_auc %>%
              mutate(model = "kernlab")) %>%
  group_by(model) %>%
  ggplot(aes(x = 1 - specificity, y = sensitivity, color = model)) +
  geom_line(size = 1.5) +
  geom_abline(
    lty = 2, alpha = 0.5,
    color = "gray50",
    size = 1.2)
dev.off()

jpeg("rplot14.jpg", width = 350, height = 350)
hist(juiced$age)
dev.off()
jpeg("rplot15.jpg", width = 350, height = 350)
hist(juiced$gpa)
dev.off()
jpeg("rplot16.jpg", width = 350, height = 350)
boxplot(juiced$age)
dev.off()
boxplot(juiced$age, plot=FALSE)$out
jpeg("rplot17.jpg", width = 350, height = 350)
boxplot(juiced$gpa)
dev.off()
boxplot(juiced$gpa, plot=FALSE)$out
```

```
#glm fit & predict
glm_fit %>%
  tidy() %>%
  arrange(-estimate)

glm_pred <- glm_fit %>%
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2017),
          type = "prob") %>%
  mutate(truth = TCC18to36_2017$retained) %>%
  roc_auc(truth, .pred_1)
glm_pred

glm_pred <- glm_fit %>%
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2017),
          type = "prob")
glm_pred

jpeg("rplot20.jpg", width = 350, height = 350)
roc_values <-
  roc_curve(glm_pred, TCC18to36_2017$retained, .pred_1, event_level = "second")
autoplot(roc_values)
dev.off()

glm_pred <- glm_fit %>%
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2018),
          type = "prob") %>%
  mutate(truth = TCC18to36_2018$retained) %>%
  roc_auc(truth, .pred_1)
glm_pred

glm_pred <- glm_fit %>%
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2018),
          type = "prob")
glm_pred

jpeg("rplot21.jpg", width = 350, height = 350)
roc_values <-
  roc_curve(glm_pred, TCC18to36_2018$retained, .pred_1, event_level = "second")
autoplot(roc_values)
dev.off()

glm_pred <- glm_fit %>%
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2017),
          type = "class") %>%
  mutate(truth = TCC18to36_2017$retained) %>%
  spec(truth, .pred_class)
glm_pred
```

```
glm_pred <- glm_fit %>%
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2017),
        type = "class")
confusionMatrix(glm_pred$.pred_class, TCC18to36_2017$retained, positive = "1",
mode="everything")

glm_pred <- glm_fit %>%
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2018),
        type = "class") %>%
  mutate(truth = TCC18to36_2018$retained) %>%
  spec(truth, .pred_class)
glm_pred

glm_pred <- glm_fit %>%
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2018),
        type = "class")
confusionMatrix(glm_pred$.pred_class, TCC18to36_2018$retained, positive = "1",
mode="everything")

#lda fit & predict
lda_fit

lda_pred <- lda_fit %>%
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2017),
        type = "prob") %>%
  mutate(truth = TCC18to36_2017$retained) %>%
  roc_auc(truth, .pred_1)
lda_pred

lda_pred <- lda_fit %>%
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2017),
        type = "prob")
lda_pred

jpeg("rplot18.jpg", width = 350, height = 350)
roc_values <-
  roc_curve(lda_pred, TCC18to36_2017$retained, .pred_1, event_level = "second")
autoplot(roc_values)
dev.off()

lda_pred <- lda_fit %>%
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2018),
        type = "prob") %>%
  mutate(truth = TCC18to36_2018$retained) %>%
  roc_auc(truth, .pred_1)
lda_pred

lda_pred <- lda_fit %>%
```

```
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2018),
        type = "prob")
lda_pred

jpeg("rplot19.jpg", width = 350, height = 350)
roc_values <-
  roc_curve(lda_pred, TCC18to36_2018$retained, .pred_1, event_level = "second")
autoplot(roc_values)
dev.off()

lda_pred <- lda_fit %>%
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2017),
        type = "class") %>%
  mutate(truth = TCC18to36_2017$retained) %>%
  spec(truth, .pred_class)
lda_pred

lda_pred <- lda_fit %>%
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2017),
        type = "class")
confusionMatrix(lda_pred$.pred_class, TCC18to36_2017$retained, positive = "1",
mode="everything")

lda_pred <- lda_fit %>%
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2018),
        type = "class") %>%
  mutate(truth = TCC18to36_2018$retained) %>%
  spec(truth, .pred_class)
lda_pred

lda_pred <- lda_fit %>%
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2018),
        type = "class")
confusionMatrix(lda_pred$.pred_class, TCC18to36_2018$retained, positive = "1",
mode="everything")

#classification tree predict
tree_pred <- predict(last_tree_fit, TCC18to36_2017, type = "prob") %>%
  mutate(truth = TCC18to36_2017$retained) %>%
  roc_auc(truth, .pred_1)
tree_pred

tree_pred <- predict(last_tree_fit, TCC18to36_2018, type = "prob") %>%
  mutate(truth = TCC18to36_2018$retained) %>%
  roc_auc(truth, .pred_1)
tree_pred

tree_pred <- predict(last_tree_fit, TCC18to36_2017, type = "prob")
```

```
tree_pred

jpeg("rplot22.jpg", width = 350, height = 350)
roc_values <-
  roc_curve(tree_pred, TCC18to36_2017$retained, .pred_1, event_level = "second")
autoplot(roc_values)
dev.off()

tree_pred <- predict(last_tree_fit, TCC18to36_2018, type = "prob")
tree_pred

jpeg("rplot23.jpg", width = 350, height = 350)
roc_values <-
  roc_curve(tree_pred, TCC18to36_2018$retained, .pred_1, event_level = "second")
autoplot(roc_values)
dev.off()

tree_pred <- predict(last_tree_fit, TCC18to36_2017, type = "class") %>%
  mutate(truth = TCC18to36_2017$retained) %>%
  spec(truth, .pred_class)
tree_pred

tree_pred <- predict(last_tree_fit, TCC18to36_2017, type = "class")
confusionMatrix(tree_pred$.pred_class, TCC18to36_2017$retained, positive = "1",
mode="everything")

tree_pred <- predict(last_tree_fit, TCC18to36_2018, type = "class") %>%
  mutate(truth = TCC18to36_2018$retained) %>%
  spec(truth, .pred_class)
tree_pred

tree_pred <- predict(last_tree_fit, TCC18to36_2018, type = "class")
confusionMatrix(tree_pred$.pred_class, TCC18to36_2018$retained, positive = "1",
mode="everything")

#random forest predict
rand_pred <- predict(last_rf_fit, TCC18to36_2017, type = "prob") %>%
  mutate(truth = TCC18to36_2017$retained) %>%
  roc_auc(truth, .pred_1)
rand_pred

rand_pred <- predict(last_rf_fit, TCC18to36_2018, type = "prob") %>%
  mutate(truth = TCC18to36_2018$retained) %>%
  roc_auc(truth, .pred_1)
rand_pred

rand_pred <- predict(last_rf_fit, TCC18to36_2017, type = "prob")
rand_pred
```

```
jpeg("rplot24.jpg", width = 350, height = 350)
roc_values <-
  roc_curve(rand_pred, TCC18to36_2017$retained, .pred_1, event_level = "second")
autoplot(roc_values)
dev.off()

rand_pred <- predict(last_rf_fit, TCC18to36_2018, type = "prob")
rand_pred

jpeg("rplot25.jpg", width = 350, height = 350)
roc_values <-
  roc_curve(rand_pred, TCC18to36_2018$retained, .pred_1, event_level = "second")
autoplot(roc_values)
dev.off()

rand_pred <- predict(last_rf_fit, TCC18to36_2017, type = "class") %>%
  mutate(truth = TCC18to36_2017$retained) %>%
  spec(truth, .pred_class)
rand_pred

rand_pred <- predict(last_rf_fit, TCC18to36_2017, type = "class")
confusionMatrix(rand_pred$.pred_class, TCC18to36_2017$retained, positive = "1",
mode="everything")

rand_pred <- predict(last_rf_fit, TCC18to36_2018, type = "class") %>%
  mutate(truth = TCC18to36_2018$retained) %>%
  spec(truth, .pred_class)
rand_pred

rand_pred <- predict(last_rf_fit, TCC18to36_2018, type = "class")
confusionMatrix(rand_pred$.pred_class, TCC18to36_2018$retained, positive = "1",
mode="everything")

#svm predict
svm_pred <- predict(last_svm_fit, TCC18to36_2017, type = "prob") %>%
  mutate(truth = TCC18to36_2017$retained) %>%
  roc_auc(truth, .pred_1)
svm_pred

svm_pred <- predict(last_svm_fit, TCC18to36_2018, type = "prob") %>%
  mutate(truth = TCC18to36_2018$retained) %>%
  roc_auc(truth, .pred_1)
svm_pred

svm_pred <- predict(last_svm_fit, TCC18to36_2017, type = "prob")
svm_pred
```

```
jpeg("rplot26.jpg", width = 350, height = 350)
roc_values <-
  roc_curve(svm_pred, TCC18to36_2017$retained, .pred_1, event_level = "second")
autoplot(roc_values)
dev.off()

svm_pred <- predict(last_svm_fit, TCC18to36_2018, type = "prob")
svm_pred

jpeg("rplot27.jpg", width = 350, height = 350)
roc_values <-
  roc_curve(svm_pred, TCC18to36_2018$retained, .pred_1, event_level = "second")
autoplot(roc_values)
dev.off()

svm_pred <- predict(last_svm_fit, TCC18to36_2017, type = "class") %>%
  mutate(truth = TCC18to36_2017$retained) %>%
  spec(truth, .pred_class)
svm_pred

svm_pred <- predict(last_svm_fit, TCC18to36_2017, type = "class")
confusionMatrix(svm_pred$.pred_class, TCC18to36_2017$retained, positive = "1",
mode="everything")

svm_pred <- predict(last_svm_fit, TCC18to36_2018, type = "class") %>%
  mutate(truth = TCC18to36_2018$retained) %>%
  spec(truth, .pred_class)
svm_pred

svm_pred <- predict(last_svm_fit, TCC18to36_2018, type = "class")
confusionMatrix(svm_pred$.pred_class, TCC18to36_2018$retained, positive = "1",
mode="everything")

glm_pred <- glm_fit %>%
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2018),
        type = "prob")
glm_pred

lda_pred <- lda_fit %>%
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2018),
        type = "prob")
lda_pred

tree_pred <- predict(last_tree_fit, TCC18to36_2018, type = "prob")
tree_pred

rand_pred <- predict(last_rf_fit, TCC18to36_2018, type = "prob")
rand_pred
```

```
svm_pred <- predict(last_svm_fit, TCC18to36_2018, type = "prob")
svm_pred

glm_pred %>%
  mutate(model = "glm (auc=0.672)") %>%
  bind_rows(lda_pred %>%
            mutate(model = "lda (auc=0.674)")) %>%
  bind_rows(tree_pred %>%
            mutate(model = "tree (auc=0.654)")) %>%
  bind_rows(rand_pred %>%
            mutate(model = "rand (auc=0.662)")) %>%
  bind_rows(svm_pred %>%
            mutate(model = "svm (auc=0.654)")) %>%
  group_by(model) %>%
  roc_curve(TCC18to36_2018$retained, .pred_1, event_level = "second") %>%
  ggplot(aes(x = 1 - specificity, y = sensitivity, color = model)) +
  geom_line(size = 1.5) +
  geom_abline(
    lty = 2, alpha = 0.5,
    color = "gray50",
    size = 1.2)

glm_pred <- glm_fit %>%
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2018),
          type = "prob") %>%
  mutate(truth = TCC18to36_2018$retained) %>%
  roc_auc(truth, .pred_1, event_level = "second")
glm_pred

lda_pred <- lda_fit %>%
  predict(new_data = bake(prep_obj, new_data = TCC18to36_2018),
          type = "prob") %>%
  mutate(truth = TCC18to36_2018$retained) %>%
  roc_auc(truth, .pred_1, event_level = "second")
lda_pred

tree_pred <- predict(last_tree_fit, TCC18to36_2018, type = "prob") %>%
  mutate(truth = TCC18to36_2018$retained) %>%
  roc_auc(truth, .pred_1, event_level = "second")
tree_pred

rand_pred <- predict(last_rf_fit, TCC18to36_2018, type = "prob") %>%
  mutate(truth = TCC18to36_2018$retained) %>%
  roc_auc(truth, .pred_1, event_level = "second")
rand_pred

svm_pred <- predict(last_svm_fit, TCC18to36_2018, type = "prob") %>%
```

```
  mutate(truth = TCC18to36_2018$retained) %>%
  roc_auc(truth, .pred_1, event_level = "second")
svm_pred
```