

Determining Academic, Background, and Financial Predictors of Community College  
First Year Retention using Data Mining Techniques

A Proposal 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

June 2021

Camille Gasaway Pace

M.S.A.S., Kennesaw State University, 2010  
B.S. Biology, Kennesaw State University, 1994

© Copyright 2021 Camille Gasaway Pace

All Rights Reserved

This dissertation, "Determining Academic, Background, and Financial Predictors of Community College First Year Retention using Data Mining Techniques," by Camille G. Pace, is approved by:

**Dissertation  
Committee  
Chair**

Lantry L. Brockmeier

Lantry L. Brockmeier, Ph.D.

Professor of Leadership, Technology, and Workforce Development

**Committee  
Member**

Michael J. Bochenko

Michael J. Bochenko, Ed.D.

Assistant Professor of Leadership, Technology, and Workforce  
Development

**Committee  
Member**

Daesang Kim

Daesang Kim, Ph.D.

Associate Professor of Leadership, Technology, and Workforce  
Development

**Associate Provost  
for Graduate  
Studies and  
Research**

Becky K. de Cruz

Becky K. de Cruz, Ph.D., J.D.

Professor of Criminal Justice

**Defense Date**

6/15/2021

FAIR USE

This dissertation is protected by the Copyright Laws of the United States (Public Law 94-553, revised in 1976). Consistent with fair use as defined in the Copyright Laws, brief quotations from this material are allowed with proper acknowledgment. Use of the material for financial gain without the author's expressed written permission is not allowed.

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

*Camille Gasaway Pace*

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

Signature \_\_\_\_\_

## ABSTRACT

Even with extensive retention research dating from the 1960s, community colleges still struggle to identify the reasons why students do not return to college. Data mining has allowed these retention models to evolve to identify new patterns among student populations and variables. The purpose of this study was to create a predictive model for student retention using background, academic, and financial factors serving as a guide for other community colleges to use when investigating institutional retention. Four different data mining models (neural networks, random forest trees, support vector machines, and logistic regression) identified significant factors for retention. The models were compared to identify if one outperformed the others on five different evaluation metrics.

The number of credit hours was consistently the most important variable in retention. In addition, the interactions between the number of credit hours, GPA, and financial aid variables were significant in student retention in their first year. The interaction between GPA, financial aid variables, and the number of remedial hours was also crucial for the first-year retention. There were no consistent variables among the retention models that can predict students' nonretention in the first year of their college career. Many background predictors (age, gender, race, or ethnicity) were not significant in predicting retained or nonretained students. The comparison of the retention models found the random forest model had the best performance for accurately classifying the nonretained and retained students overall and the retained students individually.

## TABLE OF CONTENTS

Chapter I: INTRODUCTION .....	1
Statement of the Problem.....	6
Purpose of the Study .....	6
Research Questions.....	7
Research Methodology .....	7
Significance of the Study.....	10
Theoretical Basis of the Study.....	11
Limitations of the Study.....	13
Definition of Terms.....	14
Organization of the Study .....	17
Chapter II: LITERATURE REVIEW.....	19
Community College Populations .....	20
Community College Enrollment Trends.....	22
Community College Funding.....	23
Community College Retention .....	24
Bean and Metzner’s Retention Model .....	25
Importance of Individualized Retention Models .....	27
Retention Variables.....	28
Background Variables.....	28

Age.....	29
Gender.....	31
Race or Ethnicity.....	33
High School GPA .....	36
Academic Factors.....	37
College GPA .....	38
Online Courses.....	40
Remedial Courses .....	44
Number of Courses Completed.....	47
Financial Aid Factors.....	50
Amount of Financial Aid Awarded.....	52
Amount of Financial Aid Paid .....	53
FASFA Completion .....	55
Introduction of Data Science and Big Data .....	59
Data Mining .....	61
Educational Data Mining .....	62
Classifiers.....	63
Cross Validation Methods.....	64
Decision Trees and Random Forest Trees .....	65
Support Vector Machines (SVM) .....	68

Neural Network.....	72
Logistic Regression.....	74
Interpretation of Binary Classifier Models .....	76
Evaluation Metrics for Comparing Classifier Models .....	77
Accuracy, Sensitivity, and Specificity .....	78
F1-Scores. ....	78
Receiver Operating Characteristic (ROC) Curves .....	78
Validation of Evaluation Metrics.....	80
Summary .....	80
Chapter III: METHODOLOGY.....	83
Research Design.....	83
Participants.....	85
Instrumentation .....	87
Data Collection .....	88
Data Analysis.....	88
Inferential Statistics.....	91
Random Forest .....	92
Supported Vector Machine (SVM).....	93
Neural Network.....	93
Logistic Regression.....	93



Summary .....	95
Chapter IV: RESULTS .....	97
Demographic Characteristics for Individual Cohorts .....	98
Descriptive Statistics for Students .....	100
Correlation Coefficients for Students .....	102
Categorical Variable Analysis of Combined Cohorts.....	104
Missing Data Analysis of Combined Cohorts .....	105
Cross Validation Method .....	106
Outliers and Normality of Combined Cohorts.....	106
Outlier Capping, Transformation, and Normalization.....	108
Research Question 1 .....	111
Random Forest .....	112
Support Vector Machine with Polynomial Kernel.....	118
Support Vector Machine with Radial Kernel.....	125
Neural Network.....	132
Logistic Regression.....	139
Comparison of Variable Importance.....	150
Research Question 2 .....	151
Random Forest .....	153
Support Vector Machine with Polynomial Kernel.....	155

Support Vector Machine with Radial Kernel .....	157
Neural Network.....	159
Logistic Regression.....	161
Overall Model Comparison with ROC Curves.....	163
Inferential Tests for Model Comparison.....	165
Summary .....	173
Chapter V: SUMMARY, DISCUSSION, and CONCLUSIONS.....	176
Overview of the Study .....	177
Related Literature.....	177
Classification Models.....	178
Individual and Sector-based Models.....	178
Predictive Factors.....	178
Methodology .....	180
Participants.....	181
Variables Studied .....	181
Background Factors .....	181
Academic Factors.....	182
Financial Factors.....	183
Procedures.....	183
Summary of Findings.....	184

Research Question 1 .....	184
Research Question 2 .....	189
Discussion of Findings.....	191
Research Question 1.....	191
Research Question 2 .....	193
Limitations of the Study.....	194
Implications for Future Research.....	197
Conclusions.....	198
REFERENCES .....	201
APPENDIX A: R Code for Modeling Building and Variable Importance.....	230
APPENDIX B: R Code for Inferential Statistics Tests.....	259
APPENDIX C: Institutional Review Board Protocol Exemption Report.....	264
APPENDIX D: Data Sharing Agreement.....	266

## LIST OF TABLES

Table 1: <i>Confusion Matrix Retention Example</i> .....	76
Table 2: <i>Demographic Characteristics for Students in Both Cohorts</i> .....	99
Table 3: <i>Descriptive Statistics for Fall 2017 Cohort before Data Transformations</i> .....	100
Table 4: <i>Descriptive Statistics for Fall 2018 Cohort before Data Transformations</i> .....	101
Table 5: <i>Pearson Correlation Coefficients for Fall 2017 Cohort</i> .....	102
Table 6: <i>Pearson Correlation Coefficients for Fall 2018 Cohort</i> .....	103
Table 7: <i>Descriptive Statistics for Both Cohort before Outlier Capping, Transformation, and Normalization</i> .....	108
Table 8: <i>Univariate Normality Test for Both Cohort before Outlier Capping, Transformation, and Normalization</i> .....	109
Table 9: <i>Descriptive Statistics for Both Cohort after Outlier Capping, Transformation, and Normalization</i> .....	110
Table 10: <i>Univariate Normality Test for Both Cohort after Outlier Capping, Transformation, and Normalization</i> .....	111
Table 11: <i>Variable Importance for Random Forest Final Model with Training Data</i> ...	114
Table 12: <i>Variable Importance for Random Forest Final Model with Test Data</i> .....	116
Table 13: <i>Variable Importance for SVM with Polynomial kernel Final Model with Training Data</i> .....	120
Table 14: <i>Variable Importance for SVM with Polynomial kernel Final Model with Test Data</i> .....	122
Table 15: <i>Variable Importance for SVM Radial Final Model with Training Data</i> .....	127
Table 16: <i>Variable Importance for SVM Radial Final Model with Test Data</i> .....	130
Table 17: <i>Variable Importance for Neural Network Final Model with Training Data</i> ..	134
Table 18: <i>Variable Importance for Neural Network Final Model with Test Data</i> .....	137

Table 19: <i>Variables Used to Predict Retention Utilizing Logistic Regression (Training Data)</i> .....	140
Table 20: <i>Variable Importance for Logistic Regression Final Model with Training Data</i> .....	142
Table 21: <i>Variables Used to Predict Retention Utilizing Logistic Regression (Test Data)</i> .....	146
Table 22: <i>Variable Importance for Logistic Regression Final Model with Test Data</i> ...	148
Table 23: <i>Confusion Matrix Results for the Test Data Set using the Final Random Forest Model</i> .....	155
Table 24: <i>Confusion Matrix Results for the Test Data Set using the Final SVM with Polynomial Kernel Model</i> .....	157
Table 25: <i>Confusion Matrix Results for the Test Data Set using the Final SVM with Radial Kernel Model</i> .....	159
Table 26: <i>Confusion Matrix Results for the Test Data Set using the Final Neural Network Model</i> .....	161
Table 27: <i>Confusion Matrix Results for the Test Data Set using the Final SVM with Polynomial Kernel Model</i> .....	163
Table 28: <i>Train and Test Data Set Evaluation Metrics for Classification Models</i> .....	175

## LIST OF FIGURES

Figure 1: <i>Missing data heatmap</i> .....	106
Figure 2: <i>Missing data co-occurrence plot</i> .....	107
Figure 3: <i>Retention variable importance plot for random forest model using the training data</i> .....	115
Figure 4: <i>Nonretention variable importance plot for random forest model using the training data</i> .....	115
Figure 5: <i>Retention variable importance plot for random forest model using the test data</i> .....	117
Figure 6: <i>Retention variable importance plot for random forest model using the test data</i> .....	118
Figure 7: <i>Retention variable importance plot for svm with the polynomial kernel using the training data</i> .....	121
Figure 8. <i>Nonretention variable importance plot for svm with the polynomial kernel using the training data</i> .....	122
Figure 9: <i>Retention variable importance plot for svm with the polynomial kernel using the test data</i> .....	124
Figure 10: <i>Nonretention variable importance plot for svm with the polynomial kernel using the test data</i> .....	125
Figure 11: <i>Retention variable importance plot for svm with the radial kernel using the training data</i> .....	128
Figure 12: <i>Nonretention variable importance plot for svm with the radial kernel using the training data</i> .....	129
Figure 13: <i>Retention variable importance plot for svm with the radial kernel using the test data</i> .....	131
Figure 14: <i>Nonretention variable importance plot for svm with the radial kernel using the test data</i> .....	131
Figure 15: <i>Retention variable importance plot for neural networks using the training data</i> .....	135

Figure 16: <i>Nonretention Variable Importance Plot for Neural Network using the Training Data</i> .....	135
Figure 17: <i>Retention variable importance plot for neural networks using the test data</i>	138
Figure 18: <i>Retention variable importance plot for neural networks using the test data</i>	138
Figure 19: <i>Retention variable importance plot for logistic regression model using the training data</i> .....	143
Figure 20: <i>Nonretention variable importance plot for logistic regression model using the training data</i> .....	144
Figure 21: <i>Retention variable importance plot for logistic regression model using the test data</i> .....	149
Figure 22: <i>Nonretention variable importance plot for logistic regression model using the test data</i> .....	150
Figure 23: <i>ROC curve results for the training data set using the final random forest model</i> .....	154
Figure 24: <i>ROC curve results for the test data set using the final random forest model</i>	154
Figure 25: <i>ROC curve results for the training data set using the final svm with polynomial kernel model</i> .....	156
Figure 26: <i>ROC curve results for the test data set using the final svm with polynomial kernel model</i> .....	156
Figure 27: <i>ROC curve results for the training data set using the final svm with radial kernel model</i> .....	158
Figure 28: <i>ROC curve results for the test data set using the final svm with radial kernel model</i> .....	158
Figure 29: <i>ROC curve results for the training data set using the final neural network model</i> .....	160
Figure 30: <i>ROC curve results for the test data set using the final neural network model</i> .....	160
Figure 31: <i>ROC curve results for the training data set using the final logistic regression model</i> .....	162

Figure 32: <i>ROC curve results for the test data set using the final logistic regression model</i> .....	162
Figure 33: <i>ROC curve results for the training data set with all the final models</i> .....	164
Figure 34: <i>ROC curve results for the test data set with all the final models</i> .....	164
Figure 35: <i>Boxplots of four models using the training data and the accuracy evaluation metrics</i> .....	166
Figure 36: <i>Boxplots of four models using the training data and the f1-value evaluation metrics</i> .....	168
Figure 37: <i>Boxplots of four models using the training data and the roc_auc evaluation metrics</i> .....	169
Figure 38: <i>Boxplots of four models using the training data and the sensitivity evaluation metrics</i> .....	170
Figure 39: <i>Boxplots of four models using the training data and the specificity evaluation metrics</i> .....	172



## ACKNOWLEDGMENTS

A huge thank you to Dr. Lantry Brockmeier for his support and help throughout this process. It was terrific to work with someone who pushed me to think bigger and keep learning. His guidance in my coursework and through my dissertation was a gift that I can never repay. Thank you to each of my other committee members, Dr. Michael Bochenko and Dr. Daesang Kim, for reading my technical, complex work and helping me translate it to others.

To my best friend, Elaine Westbrook, for the support, calls, zoom meetings, and just being there for me during this process and the past 35 years. I cannot think of anyone I would ever want to go with this process through. To GHC President Dr. Don Green, you pushed me throughout this journey and ignited my passion for community college students and their stories. Finally, to my cohort, Mandy, Matt, Amber, Jennifer, Melanie, Tony, Justin, Christian, John, Regina, and especially Josie, I am overwhelmed with the support, advice, and humor you all gave during this process. Remember never to compare yourself to the right or left but keep looking ahead.

To my stepmom, Diane Gasaway, thank you for listening to me and being proud of me during this process. To Neil Pace, thank you for your patience, humor, and support. To my sons, Chandler and Tony, a huge thank you for everything, including the support and love you have always given. Keep being amazing! And finally, even though they are no longer here on Earth, this dissertation would have never been a goal of mine if it had not been for my dad, Tony Gasaway, and his dad, Lawson Gasaway. They always pushed me to keep learning and never quit.

## **Chapter I**

### **INTRODUCTION**

Community colleges play a crucial role in the educational landscape of the United States of America. With decreased funding, increased competition from the for-profit sector, and a competitive job market, community colleges face lower student retention rates. Students leaving community colleges reduce these institutions' revenue and funding since their budgets often depend on students' enrollment and graduation rates. For the academic year 2015-2016, community colleges collected almost 17 billion dollars from tuition, not including additional funding from the various government entities that the colleges rely on (American Association of Community Colleges, 2018). Students may depart for numerous reasons which might have consequences such as incurred debt or untransferable credits for work completed. These students have a tougher time transferring or completing a bachelor's degree than students who begin their educational careers in a traditional four-year institution (Pascarella & Terenzini, 2005). These students' needs have sparked initiatives and entire research centers to identify ways to increase graduation and retention rates (Crisp, Carales, & Núñez, 2016; Pascarella & Terenzini, 2005). Community colleges need to be aware of these crucial factors of student retention on an institutional and student level since they serve up 35% to 46% of higher education students (Cohen, Brawer, & Kisker, 2014; NCES, 2019b). The entire retention process should actively be studied at the institutional level to understand why students leave a specific institution (Aljohani, 2016; Berger, Ramirez, & Lyon, 2012).

Retention and graduation rates are two significant challenges community colleges have dealt with during the past decades (Aljohani, 2016). State and federal level funding reduced the amount of money institutions received, which previously helped relieve financial burdens while expecting public institutions to improve retention and graduation rates (Kerkvliet & Nowell, 2014). Graduation rates for the United States are measured with a three-year window for degree completion, with the current graduation rate for the 980 public two-year institutions being 25.4%, with slightly higher rates for females than males (Juszkiewicz, 2017). The retention rates for public United States two-year institutions were 62% and represented students who enrolled in the same institution the following fall (NCES, 2019c). Both rates need to be monitored and increased to help students continue and finish their academic careers. Historically, most community colleges were primarily concerned with recruitment and did not focus on the importance of retention (Astin & Higher Education Research Inst., 1975; Tinto, 1999). Community colleges' strategic plans have shifted to include retention initiatives based on specific student population's needs (Fike & Fike, 2008). Community colleges need to identify and track factors helping students remain retained at their institutions since these students may have different goals than traditional four-year college students (Wild & Ebbers, 2002).

Individual and sector-based retention models can help community colleges with retention. Since the 1960s, retention models help identify specific factors important to students' retention and focused on preventing dropouts (Aljohani, 2016; Berger et al., 2012). During the 1960s, higher education retention models expanded to include students' social and background characteristics in addition to academic success and integration

(Berger et al., 2012). This realization shifted the research of retention into individual student characteristics and greater emphasis on diversity (Berger et al., 2012).

Additionally, large-scale studies of withdrawal and persistence focused on demographic and psychological factors (Berger et al., 2012).

The 1970s was the era of building theories on various perspectives ranging from "psychological, sociological, organizational, environmental, interactional, and economic" (Aljohani, 2016, p. 3). Models created by Tinto, Spady, and Astin laid the theoretical groundwork for a more comprehensive examination of the factors surrounding retention (Astin & Higher Education Research Inst., 1975; Spady, 1970; Tinto, 1975).

Additionally, the studies built on these models served as a knowledge base and set the standard for a systematic approach to investigating retention.

Models in the 1980s focused on managing enrollment prompted by the decline in enrollment and a need to enroll and retain students (Berger et al., 2012). The idea of "enrollment management," developed by Jack Maguire, encouraged multiple parts (admissions, financial aid, registration, and research) to work together to address retention at the institution (Berger et al., 2012; Hossler & Bontrager, 2014). The focus of retention at the various levels of institutions included student life, disseminated across professional associations, regional and national conferences, and campuses across the country (Berger et al., 2012). New models were created based on existing frameworks and expanded to address the retention of previously excluded populations (Bean & Metzner, 1985). Students over the age of 24 years old, first-generation students, and students of various racial and ethnic backgrounds became the focus of studying to determine their persistence factors (Berger et al., 2012).

The 1990s saw retention research shift to validate Tinto's model and focus on the addition of influences to help identify retention factors (Berger et al., 2012; Braxton, 2000). Financial factor research became more prevalent, with an emphasis on identifying financial barriers that can affect students' retention (Berger et al., 2012). Studies on student learning initiatives emphasized learning communities and the impact of social and academic factors on students (Berger et al., 2012). Student diversity was given more attention during this decade, with newer models focusing on how students of different races and ethnicity interact and persist on predominantly Caucasian campuses (Berger et al., 2012). The importance of persistence became more critical with researchers recognizing that students may attend numerous higher education institutions to earn an undergraduate degree (Berger et al., 2012).

Building on the past decades' research, retention has become a significant initiative in higher education, with numerous studies addressing the topic. Even with all of the emphasis on retention at the institutional, state, and national level, retention rates remain low overall and for specific demographic populations (Berger et al., 2012). Current trends in retention models continue to study the role of race and ethnicity on student retention as well as students from lower socioeconomic backgrounds (Berger et al., 2012). Distance learning has exploded over the past two decades and has become a focus on retention studies. While this course delivery format can help disseminate education to different populations, the retention rates of these students are historically lower than traditional face-to-face students (Aragon & Johnson, 2008).

With the focus of retention models shifting, retention models themselves have started to evolve in methodology with new data analytics. Data science or data analytics

introduced new data processes to discover new patterns or relationships among populations and variables (Roiger, 2017). Data science has been mainly used in engineering and business but shifted into social sciences and educational fields to include higher education (Provost & Fawcett, 2013). Higher education can harness data analytics to target specific populations and institutional issues (Drigas & Leliopoulos, 2014). Within the field of data science, data mining (DM) is the term describing a group of computer-driven methods to discover the structure and identify patterns in data sets (Attewell & Monaghan, 2015; Bharati & Ramageri, 2010). Educational data mining research has focused on retention and identifying new patterns (Chacon, Spicer & Valbuena, 2012; Herzog, 2006; Huebner, 2013; Luan, 2002, Lin 2012; Yu, DiGangi, Jannasch-Pennell & Kaprolet, 2010).

One type of data mining technique is classifiers which are different models that predict which classes the dependent variables individual cases belong to (Attewell & Monaghan, 2015; Bharati & Ramageri, 2010; Breiman, 1999; Breiman et al., 1984; Han, Pei & Kamber, 2011). Classifiers look at the past behavior of the variables to predict where cases belong based on specific categories (Breiman et al., 1984). Wolpert's No Free Lunch Theorem indicated that no one classifier could handle all data sets and suggested using multiple data mining techniques to discover the most accurate model (Wolpert, 1996). Evaluation metrics can assess the model's accuracy using the confusion matrix. Individual evaluation metrics can compare different classifiers using nonparametric tests to determine if one model performs better than the other models (Demšar, 2006).

## **Statement of the Problem**

Community colleges rely on retention numbers for funding and may not understand the reasons students are leaving. Students depend on community colleges to provide them with academic content and the resources needed to be successful. With only 48.9% of students retained after the first year at community colleges, a need exists to determine why students do not return (National Student Clearinghouse Research Center, 2019). As the costs of college increase and more jobs require some college education level, community colleges need to meet these students' needs to finish their degrees. If students leave before completing their degrees, they can be left without the requirements to earn more income and may struggle to pay acquired student debt. Retention models can provide a framework for institutions to understand specific populations and identify certain factors critical in student persistence. Higher education can use data mining techniques to determine certain factors causing students to drop out and help institutions develop initiatives to retain them (Drigas & Leliopoulos, 2014). However, institutions may not have the software or personnel to harness these data mining techniques and rely on past initiatives or assumptions about students' retention habits.

## **Purpose of the Study**

The purpose of this study is to create a predictive model for student retention using background, academic, and financial factors serving as a guide for other community colleges to use when investigating institutional retention. This study will examine four different data mining models (neural networks, random forest trees, support vector machines, and logistic regression) with each model having a different algorithm to identify the significant factors of retention. Each model will have evaluation metrics

(accuracy, specificity, sensitivity, F1-scores, and AUC) derived from its confusion matrixes and providing measurements of each model's accuracy. Inferential statistical tests on these evaluation metrics will determine which of the four models is most accurate and will serve as the study's predictive model. These data mining models and significant factors can serve as a starting point in investigating individual retention analysis in the community college sector.

### **Research Questions**

The research questions for this study are:

1: Are background factors (age, gender, race or ethnicity, and high school GPA), academic factors (college GPA, percentage of courses taken in an online format, number of remedial courses taken, and the number of credits earned during the first academic year), and financial factors (FAFSA completion, amount of financial aid awarded, and amount of financial aid paid to the student during the first academic year) significant in predicting first-year student retention for community college students?

2: Does one of the data mining models (random forests, support vector machines, neural networks, or logistic regression) generate a more accurate classifier performance overall based on the evaluation metrics of accuracy, sensitivity, specificity, area under the curve (ROC\_AUC) and f-measure (F<sub>1</sub>) scores?

### **Research Methodology**

This study is a nonexperimental, correlational classification research design created to predict students' retention who completed three consecutive semesters at seven community colleges in Georgia. Classification studies are optimal for research design since the goal is to determine which academic, background and financial variables are



significant in predicting students' retention status (Mills & Gay, 2019). The use of a correlational methodology is appropriate for this study since statistical methods are needed to predict significant factors in the retention of first-year students (Creswell, 2014). Correlational quantitative research aims to measure the association between two or more variables and expand to larger models showing more complex relationships (Creswell, 2012; Creswell, 2014). This study is considered nonexperimental since the population is not randomly assigned to groups and seeks to understand the relationship between two distinct groups; retained and nonretained students (Belli, 2008). Four data mining models (random forest trees, support vector machines, neural networks, and logistic regression) will identify possible significant academic, background, and financial factors critical to student retention. The models will be compared to each other using evaluation metrics (accuracy, sensitivity, specificity, ROC\_AUC, and F<sub>1</sub> scores) to determine which model(s) produces the most accurate results.

The archival data used in the study will come from the University System of Georgia (USG) in a Microsoft Excel format. Data analysis occurs in two separate parts based on each research question and uses the current version of R, statistical software, and the various packages within the Tidyverse and Caret collections (Korkmaz, Goksuluk & Zararsiz, 2014; Kuhn et al., 2019).

For the first research question, the study design will use the dataset in three phases: data preparation, descriptive statistics, and inferential statistics. Data preparation will identify multivariate and singular variable outliers and multivariate normality, with all of these factors affecting the classification models' performances (Kuhn & Johnson, 2019). Additional preparation will focus on identifying missing data which can occur as

singular events or as a subset of the predictors with appropriate techniques to handle these missing values (Kuhn & Johnson, 2013). The second part of the research question will focus on descriptive statistics of each predictor variable. The third part of research question one will focus on building the models using a 10-fold cross validation process creating training and test data sets for the data. The four classifiers (random forest trees, support vector machines, neural networks, and logistic regression) will use the training data sets to build models and test data sets to test the models. Each classifier will create an optimal model and produce a summary output containing a confusion matrix, accuracy, sensitivity, and specificity using the test data set. The identification of the significant predictors of each model will occur using the VIP function.

The second research question will compare the data mining models (random forests, support vector machines, neural networks, or logistic regression) to identify if one of the models generates a more accurate classifier performance based on the confusion matrix. The second research question will display accuracy, sensitivity, specificity, the  $F_1$  scores, and the ROC\_AUC value for each model. When comparing these performance metrics for all the models, the differences will be measured using statistical methods (Hothorn, Leisch, Zeileis, & Hornik, 2005; Kuhn & Johnson; 2013; Kuhn & Johnson; 2019). The analytical approach will be the Mann-Whitney U test to allow for comparison of the models from training and test models, a Friedman's test identifying if there are differences between the models for each evaluation metric, and the Wilcoxon signed-rank test serving as the ad hoc test for pairwise comparison (Demšar, 2006, Fernández-Delgado et al., 2014).

The study's overall design purposely minimizes threats to the validity and

reliability of the results by using archival data of a population. Threats to external validity are reduced since the population is very specialized and the results may only be applied to other populations with these similar factors (Creswell, 2014). Threats to internal validity included sample selection, size of the sample, and data misrepresentation. (Creswell, 2014). The choice of the students will not be random but will consist of the entire population of students. The inclusion of data from two or three academic years of data and seven institutions should allow for a large sample size. If the dataset is too small, additional data can be requested and added to increase the number of students to the correct size for the classification models.

Since the study did not use an instrument for collecting the data, the reliability of the instrument was not applicable. Data accuracy is a critical component of the reliability of the study. The entire archival data is retrieved directly from the USG to maintain consistency of the data among the seven institutions. The Research and Policy Analysis office at the Board of Regents for the University System of Georgia (USG) handles all USG data. This office collects each institution's data six times a year and pulls the data elements in the USG data warehouse. The collected data is formatted into variables that allow for accuracy and consistency for the seven institutions' data. The use of numerous data mining techniques, additional binary classifier comparison using evaluation metrics, and nonparametric statistical tests increase the reliability of the study's results.

### **Significance of the Study**

The significance of this study is to identify factors predicting retention for community college students and examining data mining classifiers to detect these factors and possible relationships that might not be apparent. Many student retention studies

focus on the social phenomenon with data collected from student surveys (Delen, 2010). Critics of survey-based studies cite the relevancy to other institutions and the issues associated with survey instruments (Delen, 2010). While survey-based studies capture the students' interactions at the institutions, they do not always provide the most accurate predictive factors (Caison, 2007).

This study will not produce an instrument for predicting factors associated with student retention in a specific population. Instead, this study hopes to explain trends at the sector level to help gain a new understanding using historical institutional data. This technique is similar to "churn analysis" found in marketing to identify customers who may leave a company and create initiatives to persuade them to stay (Delen, 2010). With higher education funding tied to student retention numbers, these models can help institutions identify and intervene with students at risk for not returning, which can help stabilize their budget. In addition, different audiences can gain value from the results of these models, including but not limited to college administrations, admissions offices, financial offices, institutional research offices, faculty, and anyone dealing with student retention at community colleges.

### **Theoretical Basis of the Study**

Bean and Metzner's nontraditional undergraduate student attrition model will serve as a framework for this study. John Bean created the student attrition model, which compared students dropping out from college to workers leaving the workplace with students' satisfaction tied to various student's beliefs and institutional factors (Aljohani, 2016; Bean, 1982; Morrison & Silverman, 2012). He revised this model with Barbara Metzner, the nontraditional undergraduate student attrition model, to emphasize essential

retention factors for the nontraditional student population (Aljohani, 2016; Bean & Metzner, 1985; Johnson, Wasserman, Yildirim, & Yonai, 2014). Within their new model, five different categories of variables correlated with student retention: high school and college performance, psychological and environmental outcomes, and background variables (Aljohani, 2016; Bean & Metzner, 1985). In addition, the model indicated that nontraditional students drop out for academic reasons unrelated to social interaction (Metzner & Bean, 1987).

While many community college students who begin their educational journey classified as traditional students, they often have factors similar to nontraditional students such as attendance, age, demographics, socioeconomic status, residence, financial concerns, and college preparedness (Provasnik & Planty, 2008; Schuetz, 2008; Travers, 2016). The background factors for this study: age, gender, race or ethnicity, and high school GPA, are in their model (Bean & Metzner, 1985). The academic factors (college GPA, percentage of course taking in an online format, number of remedial courses taken, and the number of credits earned during the first academic year) differ in this study from Bean and Metzner's model but are factors relevant to current community college students (Bean & Metzner, 1985). With the average cost of college increasing 260% from 1980 to 2014, and 58% of community college students receiving some financial aid to attend college, finances play a role in whether students can afford to attend college (Jackson, 2015; Radwin et al., 2018). Bean and Metzner's model included finances under environmental variables and will be expanded in this study to include FASFA completion, amount of financial aid awarded, and amount of financial aid paid to the student during the first academic year (Bean & Metzner, 1985).

## **Limitations of the Study**

This study will focus on past students who attended any of the seven colleges from the academic years of Fall 2017 through Fall 2019. The population for the study will include students who attended the institution as first-year students and exclude dual enrollment or transfer students. In Georgia, community colleges are classified in the state college sector with a total of nine schools. The seven colleges selected are community colleges located in various parts of Georgia and are limited to colleges awarding associate degrees. The two schools not included in this sector are specialized institutions not meeting the criteria for the study. The specificity of this study may limit the applicability of the results to other higher education institutions.

Another limitation is the exclusion of social and educational integration in this study since these integrations are not as crucial to the nontraditional student population and require survey data. Survey-based studies can capture the students' interactions at the institutions but may not provide the most accurate predictive factors (Caison, 2007). Therefore, this study aims to develop predictive models explaining trends at the sector and institutional level using existing variables without student interventions.

The definition of retention may limit the findings since retained status is determined by enrollment after the drop/add period for three consecutive semesters after initial attendance. The exact period allows students to complete 30 credit hours and move from first-year students to sophomore status. This method may cause a lack of randomness since it may exclude students who left for environmental factors beyond their control, like natural disasters, and may reenter after the study's defined time.

Within the field of classification models, there are many different families of

classifiers with numerous models in each of these families (Fernández-Delgado et al., 2014). Fernández-Delgado, Cernadas, Barro & Amorim (2014) measured the accuracy rates on 179 different classifiers on 121 data sets to determine classifier behavior and accuracy regardless of the data sets. Their results found that random forests, support vector machines (SVM), and neural networks had the most accurate results among the 121 different data sets. This study used 3 of their top models and added the logistic regression model since it is one most common classification methods used in higher education and dates back to the 1960s (Cabrera, 1994). By only using four models, the study might not identify predictors or patterns that other classifiers could identify.

The data for this study accurately reflects the data institutions are required to report to the state and the federal government. The exclusion of qualitative data in this study may have identified additional and hidden factors pertinent to students' retention. The data collection was limited to two academic years and may not include students who started in later semesters of this period. These limited-time results may not apply to a more extended period due to population or variable changes (Lau, 2017; Levin, 2006). To maximize replication of the techniques, the software and software packages accessed for this study are free and accessible to anyone.

### **Definition of Terms**

For this study, the following terms were defined:

- *Accuracy* - a numeric value of how well the model identified the true positives and true negatives from a model's confusion matrix.
- *Area under the ROC curve (ROC\_AUC)* - a numeric value for the accuracy of the ROC model.

- *Classifiers* - different models that predict which classes the dependent variables individual cases belong to (Attewell & Monaghan, 2015; Bharati & Ramageri, 2010; Breiman, Friedman, Olshen, & Stone, 1984; Han, Pei & Kamber, 2011).
- *Cross validation* - a method which divides the data set performance by random assignment in different groups; a training set used to build the models, and a testing set is used only once to assess the models' (Attewell & Monaghan, 2015; Bost, Popa, Tu & Goldwasser, 2015; Efron, 1979; Han et al., 2011; Kuhn & Johnson, 2013).
- *Data mining (DM)* - the name for a group of computer-driven methods that discover the structure and identify patterns in data sets (Attewell & Monaghan, 2015; Bharati & Ramageri, 2010).
- *Data science* - the process of using data of all types to determine new patterns or ideas (Roiger, 2017). Also called data analytics.
- *F1-scores* - a measurement of the "harmonic mean of precision and recall and gives a better measure of the incorrectly classified cases than the accuracy metric" (Huilgol, 2019). F1-scores use precision and recall values (also called sensitivity) derived from the confusion matrix (Huilgol, 2019)
- *Federal student aid application (FAFSA)* - the government required form that identifies critical financial information to help determine the expected family financial contributions (EFC) (Choitz & Reimherr, 2013; Denning, 2019).



- *Friedman test* - a nonparametric alternative used to determine any statistically significant differences between the distributions of three or more related groups (Laerd Statistics, 2015b).
- *Full-time student* - student who attends under 12 hours at higher education institutions.
- *Graduation* - The award of an academic degree.
- *Logistic regression* - a model similar to a linear regression that has two distinct events using the binomial distribution.
- *Mann-Whitney U test* - A nonparametric alternative used to determine any differences between two groups of continuous or ordinal variables (Laerd Statistics, 2015c).
- *Neural network* - a classification model that mimics the operations of biological neurons with neurons passing information to other neurons with the ability to learn based on previous errors (Attewell & Monaghan, 2015).
- *Nontraditional student* - a student who fits one or more of the following traits: older than 24 years, does not live in a campus residence, or a part-time student. This student is affected by different factors than traditional students (Bean & Metzner, 1985).
- *Part-time student* - a student who attends under 12 hours at higher education institutions.
- *Persistence* - The act of enrolling and remaining enrolled until completion of a degree (Hagedorn, 2012).

- *Random forest trees* - an ensemble method using a group of decision trees to form the forest, and the tree with the most votes becomes the model used (Han et al., 2011).
- *Receiver operating characteristic (ROC) curve* - a graph that plots the specificity on the Y-axis and 1- specificity on the X-axis (Attewell & Monaghan, 2015; Knowles, 2015; Kuhn & Johnson, 2013).
- *Retention* - The state of retaining students at institutions (Hagedorn, 2012).
- *Sensitivity* - the number of true positives divided by all of the true and false positives in a confusion matrix (Attewell & Monaghan, 2015; Knowles, 2015).
- *Specificity* - the number of true negatives divided by all of the true and false negatives in a confusion matrix (Attewell & Monaghan, 2015; Knowles, 2015).
- *Support Vector Machines (SVM)* - a type of classifier or regression function that helps determine inputs in high dimensional feature space (Delen, 2010)
- *Traditional student* - a student who likely resides on campus and is affected by different factors than nontraditional students (Bean & Metzner, 1985).
- *Wilcoxon signed-rank test* - A nonparametric alternative used to determine any statistically significant median differences between the distributions of paired or matched observations (Laerd Statistics, 2015d).

### **Organization of the Study**

Chapter 1 of this proposal described the introduction and significance of understanding the factors of first-year community college student retention. Additionally, this chapter contained the theoretical basis, methodology, research questions, and hypotheses of the study. Chapter 2 will present the literature review on community

college student retention, including Bean and Metzner's (1985) nontraditional undergraduate student attrition model. The chapter will explain the different types of variables that are critical in the retention of these students. This chapter will also explore the research on data analytics, including the four data mining models and evaluation metrics. Chapter 3 will describe the research design, the participants, instrumentation, data collection, and data analysis that will help answer the research questions.

## **Chapter II**

### **LITERATURE REVIEW**

The creation of community colleges developed from several different needs: serve students who were unable to attend selective schools due to academic challenges, help improve the skills of the workforce, and remove the pressure off universities to educate freshmen and sophomores (Berger et al., 2012; Cohen et al., 2014; Jurgens, 2010; Milliron, de Los Santos, & Browning, 2003). Community colleges can go by many different names like "junior college" or "technical colleges," but each could be defined differently (Cohen et al., 2014; Milliron et al., 2003). Junior colleges' roles shifted throughout the decades and eventually morphed into community colleges. Community colleges are defined "as any not-for-profit institution regionally accredited to award the associate in arts or the associate in science as its highest degree." (Cohen et al., 2014, p. 5). Technical colleges focus on providing professional, vocational, and career training. The Truman Commission Report in 1947 called for the creation of public community colleges that were free or low cost, attract a diverse group of students, and blend the content of the junior and technical colleges (Jurgens, 2010; Milliron et al., 2003). These colleges continued to grow after the World Wars, with the G.I. Bill allowing for 1.1 million soldiers returning to the classroom (Berger et al., 2012; Crisp & Mina, 2012; Milliron et al., 2003). The Civil Rights movement in the 1950s helped educate minority students who may not have been able to attend school earlier and students who faced socio-economic constraints (Berger et al., 2012). Currently, 1,476 two-year institutions in

the United States offer associate degrees and certificates, with 876 institutions classified as public (NCES, 2019a).

Current colleges reside in two different categories; one group of institutions prepares students to transfer to four-year colleges or universities, and the other group provides occupational skills and experience for immediate employment after graduation (Sanburn & Watertown, 2017). The majority of community college students (81%) begin in a two-year institution to pursue a bachelor's degree or higher but have lower graduation rates than students who start in a four-year institution (Horn & Skomsvold, 2011; Monaghan & Attewell, 2015; Stephan, Rosenbaum & Person, 2009). The completion rates for a Bachelor's in Arts for students are 23% to 30% lower if they start at a community college but have higher graduation rates if they obtain an associate degree from those institutions (Alfonso, 2006; National Student Clearinghouse Research Center, 2012; Stephan et al., 2009). Community colleges have recently begun to focus on short and long-term certification programs outside of the degree programs that provide specialized training in a short time frame (Jurgens, 2010; Milliron et al., 2003). These institutions adjust to the changing climate of societal needs and trends, including developing partnerships to increase students' education and experiences (Milliron et al., 2003). In some environments like rural areas, community colleges may be the only traditional public higher education institutions available and must serve the needs of their communities (Hicks & Jones, 2011).

### **Community College Populations**

The population for community college students differs from traditional four-year colleges by attendance, age, race or ethnicity, socioeconomic status, and college

preparedness (Provasnik & Planty, 2008; Schuetz, 2008; Travers, 2016). Historically, community colleges serve students who have characteristics associated with lower retention rates, such as financial difficulties and attendance status (Burns, 2010). Most community college students (59%) enrolled in credit-seeking programs do not attend full-time (AACC, 2018). Women attend community colleges in higher numbers (56%) than men (44%) (NCES, 2019d). The average age (28) and median age (24) of community college students is higher than four-year colleges, with roughly half of the students under 21 years old (NCES, 2013). Full-time students under 25 attend public two-year institutions at a higher rate (79%) than full-time students 25 years and older (NCES, 2019a). Part-time enrollment has increased in higher educational institutions over the past 50 years, with 32% of students classified as part-time in 1970 and rising to 38% of students considered part-time in 2017 (NCES, 2019a; O'Toole, Stratton, & Wetzel, 2003). Part-time students under 25 attend public two-year institutions at a higher rate (61%) than part-time students 25 years and older (NCES, 2019a). Full-time students, regardless of age, attend public institutions in more significant numbers, while part-time students attend public, private, and for-profit institutions (NCES, 2019a).

Community colleges provide educational access to roughly half of all minority undergraduate students with the breakdown of credit-seeking students being primarily White (47%); Hispanic, (24%); Black, (13%); Asian/Pacific Islander, (6%) with 36% of the overall student population identifying as first-generation students (Horn, Nevill & Griffith, 2006; Mullin, 2012; NCES, 2013; NCES, 2018b). Minority students attend community colleges in higher numbers than any other higher education institution (Martin, Galentino, & Townsend, 2014). More than half of two-year community college

students are employed while attending classes (Goldrick-Rab, 2010). Roughly 40% of community college students live at or below the poverty line and require financial assistance to continue their education (Mullin, 2012). Many students receive financial aid (58%), with the majority of the funding coming from federal grants (34%) and state aid (23%) (NCES, 2018b).

### **Community College Enrollment Trends**

In 2017, 67% of high school students completed some course work at either two or four-year college level and continued this trend in 2000 and 2010 (NCES, 2019c). Of these high school students, 23% enrolled in two-year schools, with a slightly higher percentage of males than females entering these institutions (NCES, 2019c). The population of high school graduates was at one of the highest levels in 2007-2008 and will exceed those levels by 2021-2022 with an increase in graduation rates for the southeastern part of the United States (Prescott, 2008). From 2021-2022, the increase in high school graduates will continue until the mid-2020s, followed by a sharp decline (Bransberger, Michelau & Western Interstate Commission for Higher Education, 2017). The southern part of the United States will have almost 47% of all high school graduates in the country, with an overall shift to increased minority graduation rates from previous years (Bransberger et al., 2016). While high school graduates will increase, the trends are different in higher education. Student retention in community colleges in the southeastern United States of America will continue to be an essential issue since projections show a decline in student attendance of 2.5% to 7.5% between 2012 to 2029 (Grawe, 2018). These fluctuations in future enrollment will impact community colleges' funding and student population.

## **Community College Funding**

Community college funding has been unpredictable due to government funding changes, increased operating costs, and variations in student enrollments (Phelan, 2014). As a result, these institutions have begun to minimize programs that cater to academically underprepared students and move to self-sustaining programs, undermining their original mission of serving all students who enroll (Phelan, 2014). Rural community colleges are often hit the hardest and may cut entire departments or essential programs that fit their student population (Phelan, 2014). In addition, with community colleges' funding tied to state budgets, they may make up the deficits in funding by raising tuition that can place additional financial burdens on the students (Kennamer, Katsinas, & Schumacker, 2010). Since 1980, tuition and fees have doubled for community colleges, with the types of aids decreasing with their overall effect to cover these costs (Scott-Clayton, 2012). In the academic year 2017-2018, full-time undergraduate students received an average of \$14,790 in financial aid, with the majority being from grants (Baum, 2018). With these increases, finances could play a more significant role in students' concerns than before. Even with college completers earning more income in employment than non-college completers, many students cannot finish their degrees due to tuition costs (Denning, 2019).

Even with increased costs and financial burdens, students continue to enroll in higher education, with college completion seen as an economic and social success (Schudde & Goldrick-Rab, 2015). Traditionally underserved populations have benefited from the social mobility a college education provides, including insulation from unemployment in economic recessions and have financial success outside of their



demographic backgrounds (Hout, 2012; Schudde & Goldrick-Rab, 2015; Torche, 2011). College graduates tend to have more desirable jobs and earn more money over their careers than non-college graduates, with men making 1.1 million dollars more and women earning \$636 thousand more (Hout, 2012). Even with the increase in fees and tuition, students have benefited from obtaining a degree than non-completers or non-attenders of college (Hout, 2012). While community college degrees do not provide the same level of monetary gains as a traditional four-year institution, students who complete two-year degrees have better financial success than students with no degrees (Marcotte, Bailey, Borkoski, & Kienzl, 2005). If half of the students entering community colleges nationwide in 2006 completed their degrees, their earnings would have been an additional 1.4 billion dollars in 2010 with additional federal tax revenue of 200 million dollars and state revenue of 60 million dollars (Schneider & Yin, 2012).

### **Community College Retention**

Historically, community colleges have been unable to retain half of their student population (48.9%) (National Student Clearinghouse Research Center, 2019). Students leave higher education institutions for a myriad of reasons, including environmental and financial obstacles. Students from lower socioeconomic status or families that struggle with educational expenses may have to leave before finishing a degree (Terriquez & Gurantz, 2015). The retention levels are higher for full-time two-year public institution students (60.1%) than part-time two-year public institution students (44.9%), with 48.9% of all students enrolling in the same institution the following fall (National Student Clearinghouse Research Center, 2019). Over the last nine years, the percentage of full-time students has remained the same as the percentage of part-time students gradually

increased (National Student Clearinghouse Research Center, 2019). White students' retention is a slightly higher percentage (49.6 %) than the overall average but at a lower rate than Hispanic students (52.8%) and Asian students (55.7%) (National Student Clearinghouse Research Center, 2019). Black students have the lowest retention rate of the groups, with 42.0% of these students retained (National Student Clearinghouse Research Center, 2019). While the National Student Clearinghouse reporting system may underreport data on the for-profit institutional sector, they can help accurately capture 96% of all two-year public institutions (Dynarski, Hemelt, & Hyman, 2015). The National Center for Educational Statistics has similar results to the NSC reports on public two-year retention rates (NCES, 2019b).

### **Bean and Metzner's Retention Model**

Retention frameworks identified students' social and performance factors, including their collegiate relationships in higher education institutions (Aljohani, 2016; Berger et al., 2012). The increase of opportunities for more students to attend college, regardless of ethnicity, race, or socioeconomic status, was driven by the end of the world wars and the Civil Rights movement (Berger et al., 2012).

John Bean created a model, the student attrition model, which compared college retention and workplace turnover factors (Bean, 1982). His model emphasized student satisfaction and used four main variables: student's background, environmental and personal beliefs, institutional factors, and outcomes (Aljohani, 2016; Morrison & Silverman, 2012). He revised his model with Barbara Metzner, and they developed the nontraditional undergraduate student attrition model, which built onto Tinto's and Bean's earlier models but emphasized the nontraditional student population (Aljohani, 2016;

Bean & Metzner, 1985; Johnson et al., 2014). Their model hypothesized five different categories of variables relevant to student retention and concluded the following results: high school and college performance are linked, psychological and environmental outcomes play a more significant role than academic variables, and background variables influence student persistence (Aljohani, 2016; Bean & Metzner, 1985). They validated their model in their 1987 study, which found that nontraditional students drop out for academic reasons unrelated to social interaction (Metzner & Bean, 1987). Bean's later work with Eaton (2001) identified psychological attributes for student success in social and academic integration, including self-efficacy, approaches to the social and educational challenges that arise, and positive attitudes to attending their institution. Higher education institutions addressing these student issues could help increase retention through learning communities, first-year students' interest groups, and tutoring (Bean & Eaton, 2001).

Nontraditional students were affected by environmental factors more than traditional students, and those factors can shape their entire college experience socially and academically (Bean & Metzner, 1985). Their definition of environmental factors included finances, hours of employment, outside encouragement, family responsibilities, and opportunity to transfer (Bean & Metzner, 1985). Social integration variables had a direct effect but were not as important as the other groups of variables for nontraditional students (Bean & Metzner, 1985). Nontraditional students are less affected by social interactions on campus since environmental factors can lessen the chances for students to build great social connections on campus (Bean & Metzner, 1985). Without the same access to academic services and support due to environmental factors, nontraditional

students may not perform as well as traditional students. The latter has higher chances of using educational services and support (Bean & Metzner, 1985).

### **Importance of Individualized Retention Models**

While traditional retention models like Tinto's could be the base of models in various higher education institutions, specialized institutional-based retention models may provide a deeper understanding of the factors of these specific populations. Metzner and Bean (1987) believed that "samples of nontraditional students tend to be heterogeneous and probably differ substantially from university to university so that the combination of several schools might not produce additive effects" (p. 34), contributing to the overall need for individualized retention models. Specialized institution-based models help identify specific factors that can allow for planning, assessments, and policymaking that are essential to that student population retention and help develop early intervention initiatives (Aguilar et al., 2014; Herzog, 2006; Nevarez & Wood, 2010). Another type of modeling is sector-based retention models that identify relationships not detected in institutions with smaller populations. Sector-based modeling uses similar schools within a sector or area and completes analyzes on these schools. Identifying at-risk students could lead to specialized interventions that could retain them at the institution or sector level (Herzog, 2006). With roughly 31% of community college students transferring to four-year institutions, specialized models could help the community college students begin at and the institutions they move to (Shapiro et al., 2017). The institutions in this study are all labeled state colleges and primarily serve students in their first two years of higher education.

## **Retention Variables**

Over the past sixty years, various retention models identified distinct variables that helped predict retention. The previous retention models use similar variables in their models but group them into different categories (Bean, 1982; Bean & Metzner, 1985; Cabrera et al., 1993; Tinto, 1975; Tinto, 1993). The study will focus on community college students' academic, background, and financial characteristics to predict retention using Bean and Metzner's (1985) nontraditional undergraduate student attrition model. The majority of this population is commuter students attending community colleges with limited residential housing options. Many of the academic, background, and financial factors identified by Bean and Metzner (1985) will be the variables used in answering the research questions. Social and educational integration will not be included in this model since these integrations are not as crucial to the nontraditional student population and require survey data. Most student retention studies focus on social phenomena, relying on surveys to develop and validate theories (Delen, 2010). Critics of survey-based studies cite the relevancy to other institutions and the issues associated with survey instruments (Delen, 2010). Survey-based studies capture the students' interactions at the institutions but do not always provide the most accurate predictive factors (Caison, 2007). There are often negative perceptions of survey-based studies' results, especially with low response rates (Fosnacht, Sarraf, Howe & Peck, 2017). This study aims to develop predictive models that can explain trends at the sector and institutional levels, which can help gain a new understanding at these levels.

## **Background Factors**

The student background characteristics gathered at enrollment reflected

demographic information, such as high school performance and other demographic factors (Johnson et al., 2014). Tinto (1975) referred to these variables as pre-entry attributes: highlighting family, background, skills, abilities, and prior schooling. In Bean's first model, the definition of background variables included student's past performance, socioeconomic status, state residency, distance from home, hometown size, and how students interact with the college environment (Bean, 1980, Kerby, 2015). His later model with Barbara Metzner expanded on background variables, including age, enrollment status, residence, educational goals, high school performance, ethnicity, and gender (Bean & Metzner, 1985). For this study, the background variables examined are age, gender, race or ethnicity, and high school GPA.

**Age.** Many researchers divide students into two distinct categories; traditional, referring to students under 24 years old, and nontraditional, which include ages 24 and older. Additional definitions of nontraditional students include other categories outside of the age range, such as attending part time, working full time, delayed entry, financial independence, or being a parent (Iloh, 2018; Layne, Boston, & Ice, 2013). Nontraditional students account for half of higher education enrollment, with community colleges enrolling a large number of these types of students, and they tend to be older than students at traditional four-year institutions (Howell, Williams, & Lindsay, 2003; Parsad, Lewis & National Center for Education Statistics, 2008; Wood, 2013).

The research on age produced mixed results of its importance as a factor for retention rates at community colleges (Bean & Metzner, 1985; Metzner & Bean, 1987; Pascarella & Chapman, 1983). Metzner and Bean (1987) expanded on their previous model (1985) to build a newer conceptual model focused on older, part time students at a

Midwestern university in the United States. Their regression models focused on retention factors of nontraditional students with data collection derived from a questionnaire given to all students in English composition courses and some academic variables from the registrar's office (Metzner & Bean, 1987). The 26 variables that focused on academic factors, background factors, social integration, environmental, and physiological variables explained 29% of the variance for dropping out (Metzner & Bean, 1987). Age was not significant as a factor in students dropping out but ranked 7<sup>th</sup> out of the variables (Metzner & Bean, 1987). Their model showed that older students in the study had higher GPAs and more invested than younger students but drop out at the same rates (Metzner & Bean, 1987). Feldman (1993) used a forward, stepwise logistic regression to determine background variables (gender, ethnicity, age, enrollment status, goals, basic skill needs, and high school GPA) A forward, stepwise logistic regression determined that high school GPA ( $B = .459$ ,  $W = 34.029$ ,  $p < .01$ ), age ( $B = 1.770$ ,  $W = 26.127$ ,  $p < .01$ ), and enrollment status ( $B = 2.227$ ,  $W = 14.480$ ,  $p < .01$ ) are significant in predicting first year retention (Feldman, 1993). Ethnicity, gender, goals, and basic skills were not statistically significant at the 0.05 level in the logistic regression model (Feldman, 1993). Mertes and Hoover (2014) created a Chi-square analysis to determine what variables (age, gender, ethnicity, credit hour load, educational goal, remedial need (English and math), grade in an introductory technology course, and receipt of financial aid) were significant in retention at midwestern community college. Age (fall 2007,  $\chi^2(3, n = 587) = 20.682$ ,  $p < .001$  and fall 2010,  $\chi^2(3, n = 872) = 16.877$ ,  $p < .001$ ), gender (fall 2007,  $\chi^2(1, n = 587) = 5.265$ ,  $p < .05$  and fall 2010,  $\chi^2(1, n = 872) = 8.179$ ,  $p < .01$ ), program of study fall 2007,  $\chi^2(3, n = 587) = 17.634$ ,  $p < .001$  and fall 2010,  $\chi^2(2, n = 872) = 17.637$ ,  $p <$

.001), and grade in the introductory course (fall 2007,  $\chi^2(3, n = 587) = 140.976, p < .001$  and fall 2010,  $\chi^2(3, n = 872) = 20.356, p < .001$ ) were significant for retention in the two semesters of data (Mertes & Hoover, 2014). Windham, Rehfuss, Williams, Pugh, and Tincher-Ladner (2014) used a logistic regression model to find the effects of a study skills course on retention at a Southeastern community college and found gender ( $\beta = 0.663, p < .001$ ), entry reading score significantly predict student retention ( $\beta = 0.012, p < .001$ ), and age were significant in predicting student retention. Out of the four categories for age, under 18, 19-24 ( $\beta = -0.296, p < .05$ ), and over 40 years old ( $\beta = 0.535, p < .05$ ) were significant in predicting retention using a study skills course (Windham et al., 2014). Younger students (age 19-24) were retained (25.7%) at lower rates than students under 18-year old for a community college with multiple locations (Windham et al., 2014). They also discovered that students over 40 years old were retained at higher rates (70.7%) than students under 18 years old (Windham et al., 2014). Ethnicity or race, and socioeconomic status were not significant in predicting retention when paired with a study skills course (Windham et al., 2014).

**Gender.** The research on gender indicates that students who identify as female account for more than half of high education enrollment and have higher persistence rates than students who identify as male (Bean and Metzner, 1985; Chee, Pino, & Smith, 2005; Howell et al., 2003; Jaggars & Xu, 2010; Wladis, Conway, & Hachey, 2017; Xu & Jaggars, 2011). Bean and Metzner (1985) included gender in their model since males and females can have "indirect effects on attrition through family responsibilities (positive effects for women) and opportunity to transfer (negative effect for women)" (p. 498).

For two-year public community college students, females had a slightly higher



graduation rate than males, 26% versus 24% and higher GPAs but overall, there are no significant differences with student persistence based on gender (NCES, 2019e; Peter & Horn, 2005; Stewart, Lim & Kim, 2015). Corbett, St. Rose, and Hill (2008) studied descriptive longitudinal data for the nation and found women earn the majority of associate degrees with a percentage increase between the 1970s (47%) to the 2000s (62%). Mertes and Hoover (2014) used a Chi-square analysis to determine if age, gender, ethnicity, credit hour load, educational goal, remedial need (English and math), and receipt of financial aid were significant for retention at a midwestern community college using two different fall semester for their data. Gender (fall 2007,  $\chi^2(1, n = 587) = 5.265, p < .05$  and fall 2010,  $\chi^2(1, n = 872) = 8.179, p < .01$ ), program of study fall 2007,  $\chi^2(3, n = 587) = 17.634, p < .001$  and fall 2010,  $\chi^2(2, n = 872) = 17.637, p < .001$ ), grade in the introductory course (fall 2007,  $\chi^2(3, n = 587) = 140.976, p < .001$  and fall 2010,  $\chi^2(3, n = 872) = 20.356, p < .001$ ), and age (fall 2007,  $\chi^2(3, n = 587) = 20.682, p < .001$  and fall 2010,  $\chi^2(3, n = 872) = 16.877, p < .001$ ) were significant in retention (Mertes & Hoover, 2014). Windham et al. (2014) used a logistic regression model to find the effects of a study skills course on retention at a Southeastern community college and found gender was a significant predictor ( $\beta = 0.663, p < .01$ ) with female retention higher than male retention. They also found entry reading scores significantly predict student retention ( $\beta = 0.012, p < .01$ ) and age ( $\beta = -0.296, p < .05$ ) (Windham et al., 2014). Yu (2017) created a hierarchical generalized linear model to identify college completion variables at community college using Integrated Postsecondary Education Data (IPEDS) and Beginning Postsecondary Students Longitudinal Study (BPS) data. The study found gender ( $\beta = 0.185, p < .05$ ), ethnicity or race ( $\beta = -0.212, p < .05$ ), high school GPA ( $\beta =$

0.185,  $p < .05$ ), and attending status ( $\beta = 0.457, p < .01$ ) are significant to college completion in six years (Yu, 2017). Yu (2017) also identified a negative correlation between institutions with a large percentage of female students ( $\beta = -0.024, p < .05$ ) and completion at 2-year community colleges. Wang (2012) used an OLS regression and the National Education Longitudinal Study of 1988 dataset and found being female ( $\beta = 0.146, p < .01$ ), taken remedial courses in math ( $\beta = -0.079, p < .01$ ), having a higher college GPA before transfer ( $\beta = 0.581, p < .01$ ), and continuous enrollment ( $\beta = 0.168, p < .01$ ) had higher GPAs at the next institution they attended (Wang, 2012)

While most of the studies show gender is significant to retention, Stewart, Lim, and Kim (2015) found gender was not significant. Stewart et al. (2015) investigated the effects of demographic variables, family characteristics, pre-college and college academic performance factors, and remedial courses on retention at a large, residential university using a factorial analysis of variance (ANOVA). Ethnicity ( $F(1, 3212) = 8.386, p < .01$ ), financial aid ( $F(1, 3212) = 30.862, p < .01$ ), and remedial status ( $F(1, 3212) = 9.582, p < .05$ ) were found to have an effect on retention (Stewart et. al, 2015). They found no statistically significant effect of gender on retention,  $F(1, 3212) = .399, p = .528$ , for first-time first-year students during their first two academic years at a large, residential university (Stewart et al., 2015).

**Race or Ethnicity.** Over the last forty years, the workforce and educational population in the United States have become more diverse (Bransberger et al., 2016; Burke, 2019; Zumeta, Breneman, Callan, & Finney, 2012). By 2024-2025, the South will graduate 40% of all Black high school graduates and 60% of all Hispanic high school graduates (Bransberger et al., 2016). With this shift, community colleges need to be

aware of the diverse backgrounds and cultures of their current and potential students to help increase retention and graduation (Astin, Keup, & Lindholm, 2002; Burke, 2019). Bean and Metzner's model theorized that "the primary indirect effects of ethnicity for nontraditional students are through a strong negative influence on GPA due to the comparatively poorer education provided for minority students at the secondary level" (Bean & Metzner, 1985, p. 498). Their stepwise regression model supported this claim, with minority students ( $\beta = -0.19, p \leq .001$ ) having lower academic grades, resulting in higher dropout rates (Bean & Metzner, 1985).

Stewart et al. (2015) investigated the effects between demographic variables, family characteristics, precollege and college academic performance factors, and remedial courses and retention at a large, residential university using a factorial analysis of variance (ANOVA). Ethnicity ( $F(1, 3212) = 8.386, p < .01$ ), financial aid ( $F(1, 3212) = 30.862, p < .01$ ), and remedial status ( $F(1, 3212) = 9.582, p = .047$ ) were found to have an effect on retention (Stewart et. al, 2015). Yu (2017) created a hierarchical generalized linear model to identify college completion variables at community college using Integrated Postsecondary Education Data (IPEDS) and Beginning Postsecondary Students Longitudinal Study (BPS) data. The study found hours worked, high school GPA, attending status, institution size, and percentage of minority students are significant to college completion in three years (Yu, 2017). The study found gender ( $\beta = 0.185, p < .05$ ), ethnicity or race ( $\beta = -0.212, p < .05$ ), high school GPA ( $\beta = 0.185, p < .05$ ), attending status ( $\beta = 0.457, p < .01$ ), tuition and fees ( $\beta = 0.000, p < .10$ ), institution size ( $\beta = -0.233, p < .10$ ) are significant to college completion in six years (Yu, 2017). Yu

(2017) also found having a significant minority population ( $\beta = -0.004, p < .10$ ) decreases the odds of degree completion during a three year period.

Within the different race or ethnicity groups, retention rates differ with Asian students having higher retention rates than Hispanic/Latino and African American students (Community College FAQs, n.d.). Corbett, Rose, and Hill (2008) studied descriptive longitudinal data for the nation and found African American men and Hispanic women and men earn bachelor's degrees at lower rates than other demographic populations and have declined since the 1970s. Wang (2012) discovered African American community college students' GPAs ( $\beta = -0.282, p < .01$ ) are significantly lower than their White counterparts using an OLS regression for the National Education Longitudinal Study of 1988 dataset. The study also found being female ( $\beta = 0.146, p < .01$ ), taken remedial courses in math ( $\beta = -0.079, p < .01$ ), have a higher college GPA before transfer ( $\beta = 0.581, p < .01$ ), and remained continuous enrolled ( $\beta = 0.168, p < .01$ ) had higher GPAs at the next institution they attended (Wang, 2012). Mertes and Hoover (2014) used a Chi-square analysis to determine if age, gender, ethnicity, credit hour load, educational goal, remedial need (English and math), and receipt of financial aid were significant for retention at a midwestern community college using two different fall semester of data. Gender (fall 2007,  $\chi^2 (1, n = 587) = 5.265, p < .05$  and fall 2010,  $\chi^2 (1, n = 872) = 8.179, p < .01$ ), program of study fall 2007,  $\chi^2 (3, n = 587) = 17.634, p < .001$  and fall 2010,  $\chi^2 (2, n = 872) = 17.637, p < .001$ ), grade in the introductory course (fall 2007,  $\chi^2 (3, n = 587) = 140.976, p < .001$  and fall 2010,  $\chi^2 (3, n = 872) = 20.356, p < .001$ ), and age (fall 2007,  $\chi^2 (3, n = 587) = 20.682, p < .001$  and fall 2010,  $\chi^2 (3, n = 872) = 16.877, p < .001$ ) were significant in retention (Mertes & Hoover, 2014). Race or

ethnicity was statistically significant in retention for fall 2010 only,  $\chi^2(6, n = 821) = 13.853, p = .031$  with African-American and Hispanic students having lower retention rates than their White counterparts (Mertes & Hoover, 2014).

**High School GPA.** Bean and Metzner's research indicated that commuter students have a lower high school GPA than traditional residential students, and older commuter students have a lower high school GPA than younger commuter students (Bean & Metzner, 1985). One issue with a high school GPA is the lack of a standard formula because the GPA formula can vary from school to school, leading to inconsistency in the measurement (Porter & Polikoff, 2012).

High school GPA is a predictor for student persistence in higher education for the first year, where lower high school GPAs are an indicator of dropping out (Feldman, 1993; Huerta & Watt, 2015; Yu, 2017). Feldman (1993) used a forward, stepwise logistic regression analysis to determine pre enrollment variables (gender, ethnicity, age, enrollment status, goals, basic skill needs, and high school GPA significant for first year students' retention. High school GPA ( $B = .459, W = 34.029, p < .01$ ), age ( $B = 1.770, W = 26.127, p < .01$ ), and enrollment status ( $B = 2.227, W = 14.480, p < .01$ ) are significant in predicting first year retention (Feldman, 1993). The model also found that a one-point increase in high school GPA was correlated with a decrease in the dropout rate by 0.46 (Feldman, 1993). Ethnicity, gender, goals, and basic skills were not statistically significant at the 0.05 level in the logistic regression model (Feldman, 1993). Yu (2017) created a hierarchical generalized linear model with Integrated Postsecondary Education Data (IPEDS) and Beginning Postsecondary Students Longitudinal Study (BPS) data and found gender ( $\beta = 0.185, p < .05$ ), ethnicity or race ( $\beta = -0.212, p < .05$ ), high school

GPA ( $\beta = 0.185, p < .05$ ), and attending status ( $\beta = 0.457, p < .01$ ) are significant to college completion in six years (Yu, 2017). Yu (2107) also found that a higher high school GPA ( $\beta = 0.081, p < .05$ ) increased the odds of degree completion in three years. Huerta and Watt (2015) tracked a group of 329 high school students through college and found high school GPA was a significant predictor of first year retention ( $\beta = 1.455, W = 8.348, p < .01$ ) and second-year retention ( $\beta = 1.615, W = 23.052, p < .001$ ) using a logistic regression model. They also found that completion of college credits in high school was a significant predictor of first-year retention ( $\beta = 1.695, W = 4.747, p < .05$ ) and second-year retention ( $\beta = 1.885, W = 30.206, p < .001$ ) (Huerta & Watt, 2015). Other high school variables like the number of Advanced Placement (AP) courses, number of years in Advancement Via Individual Determination (AVID) system, and taking the SAT were not significant in first-year retention (Huerta & Watt, 2015). Belfield, Crosta, and Columbia University (2012) examined transcript data from statewide community college system students. They found that high school GPA ( $\beta = 0.845, p < .001$ ) was statistically significant in predicting the first-year cumulative GPA and accounted for 21% of the variation using a regression model (Belfield et al., 2012). They discovered that student's college GPA tends to be one grade notch below their high school GPA, indicating that high school grades can help predict future college academic success (Belfield et al., 2012).

### **Academic Factors**

Academic performance is an indicator of future performance in retention and graduation of community college students at their current and future institutions (Pascarella & Terenzini, 2005). Bean and Metzner (1985) felt the main reason

community college students enroll in their colleges is purely academic and differs from students who attend traditional four-year institutions that may seek out more social and educational integration. Additionally, Metzner and Bean (1987) found that social inclusion was not an important reason for nontraditional students to drop out, with part-time students leaving primarily due to academic performance and their lack of commitment to the institution attended. With two-year students having less time for social integration through campus activities than traditional four-year students, the classroom becomes their primary source of academic and social inclusion (Townsend & Wilson, 2009). In addition, students in community colleges feel more connected to their faculty, and other students since the number of students in a course may be smaller than traditional four-year institutions and allow for more personal interaction (Townsend & Wilson, 2006). With the classroom being the primary source of integration for these students, performance in courses could help identify potential student retention factors. The academic variables for this study are college GPA, percentage of courses taking in an online format, number of remedial classes taken, and the number of credits earned during the first academic year.

**College GPA.** The research on college GPA suggests it is a strong predictor of students' persistence (Tinto, 1975). Metzner and Bean (1987) created a stepwise regression model that significantly predicted that GPA ( $\beta = -0.36$ ,  $p < .001$ ), intent to leave ( $\beta = 0.28$ ,  $p < .001$ ), hours enrolled ( $\beta = -0.16$ ,  $p < .001$ ), and study skills ( $\beta = 0.09$ ,  $p < .05$ ) were significant in predicting students' not returning to college. College GPA is the best predictor of students' not returning to college than any other social integration, environmental, or physiological variables (Metzner & Bean, 1987). Stewart et al. (2015)

ran a multiple regression analysis and determined first-semester college GPA ( $\beta = 0.859$ ,  $R^2 = .241$ ,  $p < .01$ ) was the most significant predictor of persistence. This study also examined the relationship between ACT composite scores, high school GPA, and persistence (Stewart et al. 2015). The first-semester college cumulative GPA variable accounted for slightly over 24% (.241) of variance on the model and had a strong correlation (.491) on persistence (Stewart et al., 2015). This study also examined the relationship between gender, race or ethnicity, ACT composite score, high school GPA, family income, financial aid status, college cumulative GPA, and remedial (Stewart et al. 2015). DeNicco, Harrington, and Fogg (2015) developed a logistic regression model for 1,800 students in a public state college system. They identified college GPA ( $\beta = 0.097$ ,  $p < .01$ ) as a significant factor in the first-year retention among other demographics, high school characteristics, placement test scores, freshman year performance, and remedial course work (DeNicco et al., 2015). For each point above the mean GPA, there is an increase (9.7%) in the likelihood of their retention (DeNicco et al., 2015). Wang (2012) used an OLS regression and the National Education Longitudinal Study of 1988 dataset and found being female ( $\beta = 0.146$ ,  $p < .01$ ), taken remedial courses in math ( $\beta = -0.079$ ,  $p < .01$ ), have a higher college GPA before transfer ( $\beta = 0.581$ ,  $p < .01$ ), and remained continuous enrolled ( $\beta = 0.168$ ,  $p < .01$ ) had higher GPAs at the next institution they attended (Wang, 2012). Nakajima, Dembo, and Mossler (2012) examined multiple factors, including demographic, financial, academic, academic integration, and psychosocial, to determine their relationship to student retention for 427 students at a community college in California. A 63-item survey was completed in the fall semester with additional course data collected for that fall and the following spring semester



(Nakajima, Dembo & Mossler, 2012). Student retention was negatively associated with students' age ( $r = -.104, p < .05$ ); off-campus employment hours ( $r = -.161, p < .01$ ); total employment hours ( $r = -.130, p < .01$ ); and English proficiency ( $r = -.099, p < .05$ ) (Nakajima, Dembo & Mossler, 2012). Nakajima, Dembo, and Mossler (2012) found positive correlations to retention for credit hours enrolled ( $r = .179, p < .01$ ); receipt of financial aid ( $r = .122, p < .05$ ); and cumulative college GPA ( $r = .125, p < .05$ ). There was no significant correlation between any of the psychosocial variables or the academic integration variable to retention (Nakajima, Dembo & Mossler, 2012). The second part of their study used community college students' data and t-tests and determined the following variables were significant in their association to student retention: age ( $t = 2.127, p < .05$ ); receipt of financial aid ( $t = 2.814, p < .01$ ); credit hours attempted ( $t = 2.246, p < .05$ ); number of credit hours completed ( $t = 2.218, p < .05$ ); number of current credit hours enrolled ( $t = 5.442, p < .001$ ); high school graduation year ( $t = 3.307, p < .001$ ); cumulative college GPA ( $t = 2.559, p < .01$ ); English proficiency ( $t = 3.307, p < .001$ ); off-campus employment hours ( $t = 3.363, p < .001$ ); and total employment hours ( $t = 3.097, p < .01$ ) (Nakajima, Dembo & Mossler, 2012). They reinforced these findings with logistic regression with cumulative college GPA ( $\beta = 2.014, p < .01$ ) being the most reliable predictor in their model followed by Fall 2007 enrollment ( $\beta = 1.012, p < .05$ ), and English proficiency ( $\beta = 0.622, p < .05$ ) (Nakajima, Dembo, & Mossler, 2012). Students who had higher cumulative GPAs were twice as likely to stay in college (Nakajima, Dembo & Mossler, 2012).

**Online Courses.** Community colleges were early adopters of online learning, with 97% of institutions offering courses in this format and serving more students online

than other higher education institutions (Allen & Seaman, 2010; Parsad et al., 2008; Travers, 2016). The definition of online learning varies based on institutions and reporting agencies (Cejda, 2010). The growth in online offerings has played a role in meeting enrollment increases without an increased expense of physical infrastructure with 2/3 of chief academic officers including this course delivery format in future planning (Allen & Seaman, 2013; Gregory & Lampley, 2016; Jaggars, Edgecombe, & Stacey, 2013). Online course growth is ten times higher than traditional course delivery, with community college students taking many of these courses (Shea & Bidjerano, 2014). Older female students account for more than half of the students enrolled in online classes (Howell et al., 2003).

Online learning can provide a convenient solution for students who cannot attend traditional face-to-face courses, with almost 7 million U.S. students taking at least one online course in 2014 (Allen, Seaman, Poulin, & Straut, 2016). In addition, community colleges have large numbers of nontraditional students attending online classes, especially with students who cited issues with attending face to face courses due to outside obligations (Gregory & Lampley, 2016; Pontes & Pontes, 2012). However, one problem facing online learning is that dropout rates 20% higher than traditional face to face courses (Aragon & Johnson, 2008).

The research on the success of these students in online courses is mixed. Xu and Jaggars (2011) used percentage comparisons to examine Washington community and technical college students and found that students were more likely to complete face to face courses (90%) than online courses (82%). They also found that students who took online courses in the first term (32%) in their academic careers had a slight but

significant increase in dropping out of school than students who took online courses during their first year (19%) (Xu & Jaggars, 2011). Their research for withdrawal and failure rates in online courses was repeated in Virginia with two different cohorts and found similar results in Washington (Jaggars & Xu, 2010). They continued to use percentage comparisons to examine the 2004 cohort of community college students in Virginia and found that students were more likely to complete face to face courses (81%) than online courses (68%) (Jaggars & Xu, 2010). The 2008 cohort had similar results, with face to face course completion rates of 79% and online course completion at 67% (Jaggars & Xu, 2010). Shea and Bidjerano (2014) created a logistic regression model using a nationwide sample of 16,100 students to examine the relationship of personal, family, and institutional variables with degree obtainment and online course delivery and found a different outcome. The results of the model indicate that students who were female ( $\beta = 0.342, p < .001$ ), older ( $\beta = 0.024, p < .001$ ), from larger families ( $\beta = 0.077, p < .05$ ), have a higher amount of financial aid ( $\beta = 0.000, p < .01$ ), location from institution ( $\beta = -0.255, p < .05$ ) and have loans ( $\beta = 0.000, p < .05$ ) were more likely to take distance education courses with the other variables not being significant (Shea & Bidjerano (2014). Students who take an online course early in their academic career complete community college credentials significantly higher ( $R^2 = .004, \text{Wald } F(1, 240) = 5.76, p < .05$ ) than students who only take face-to-face courses (Shea & Bidjerano (2014). Johnson and Mejia (2014) also created descriptive statistics logistic regression models using student demographics, course enrollment, and student outcomes to investigate California community colleges. Their analysis indicated that only 11% of online courses were highly successful, and students who took online courses were only successful 60%

of the time for the academic year 2013-2014 (Johnson & Mejia, 2014). Students have lower odds of passing an online course ( $\beta = -0.147, p < .01$ ) than a traditional course compared to face-to-face students (Johnson & Mejia, 2014). They also found a negative correlation between the success rates of minority students and the completion of online courses with African American ( $\beta = -0.043, p < .01$ ) and Latino ( $\beta = -0.023, p < .01$ ) students performing lower than White students (Johnson & Mejia, 2014). Hart, Friedmann, and Hill (2015) found similar results about online courses using an OLS regression and percentage comparison focusing on course enrollment, course outcomes, student characteristics, and instructor characteristic variables. Their study focused on first-time students in California community colleges for four academic years (2008 to 2011) (Hart, Friedmann, & Hill, 2018). Students were less likely to complete online courses ( $\beta = -0.056, p < .01$ ), less likely to pass online courses ( $\beta = -0.065, p < .01$ ), with the overall completion rate of online courses (78.99%) lower than face to face courses (84.58%) (Hart, Friedmann & Hill, 2015). James, Swan, and Daston (2016) used predictive analytics for five different community colleges located in various geographical regions of the United States and found that students who take only face-to-face courses were retained at a higher percentage (51%) than students who only take online courses (30%). Students who took a combination of both types of courses were retained at a higher level (58%) than each type of course (James, Swan, & Daston, 2016). Aragon and Johnson (2008) investigated student background, enrollment, academic, and self-directed learning factors of online students in one community college. Their Chi-Square and t-tests found no significant differences in most factors for students' completion of online courses (Aragon & Johnson, 2008). The significant differences were found in gender ( $\chi^2$

(1, n = 305) = 5.64,  $p < .05$ ) with female students (66%) had a higher rate of completion than male students (52%) and a higher GPA average ( $t(303) = 4.45$ ,  $p < .001$ ) between completers and non-completers (Aragon & Johnson, 2008).

**Remedial Courses.** Many students arrive underprepared for the rigor of their community college courses and enroll in one or more remedial courses (Xu & Dadgar, 2018). Across the country, at least 60% of community college students need remedial support in at least one area, with less than 25% of these students completing a degree in eight years (Bailey, Cho & Columbia University, 2010). For students who leave after one semester, 80% of students had a learning deficiency and were required to take some form of remediation (Burley, Butner & Cejda, 2001). Remedial or co-requisite courses assignment happens based on students' standardized test or an institutional-based assessment exam score. These courses represent 10% of all courses earned at community colleges but often do not count towards degree completion (Scott-Clayton & Rodriguez, 2015). These additional courses add time to a degree which could decrease the retention of the student as well as re-emphasis the material learned in high school and can frustrate students for having to pay for these "redundant" courses (Adelman, 2006; Barbatis, 2010; Boylan & Saxon, 1999).

While Bean and Metzner's model (1985) identified areas of student academic unpreparedness, they never addressed the relationship of remedial courses with retention. Scott-Clayton and Rodriguez (2015) developed a regression discontinuity design for six different urban community colleges and explored student demographic, academic, and financial factors concerning remediation (Scott-Clayton & Rodriguez, 2015). Students being assigned to remedial courses didn't have an impact on degree completion ( $\beta = -$

0.002,  $p > .05$ ), student persistence ( $\beta = -0.008$ ,  $p > .05$ ), dropout ( $\beta = 0.010$ ,  $p > .05$ ), and semesters enrolled ( $\beta = -0.004$ ,  $p > .05$ ) (Scott-Clayton & Rodriguez, 2015). Belfield and Crosta (2012) examined transcript data from statewide community college system students to examine if eight different academic placement tests or high school performance could predict college performance, including course grades. The first part of their study used pairwise Pearson correlations and found that placement exams had lower correlations ( $r = .08$  to  $.18$ ) to developmental course grades and could not predict students' grades (Belfield & Crosta, 2012). High school GPA had higher correlations ( $r = .34$  to  $.36$ ) and was a better predictor of students' grades in the remedial courses (Belfield & Crosta, 2012). The second part of the research examined the differences of students who placed in the highest and lowest quartile of the placement exams and found that students whose scores were in the highest quartile had an average of nine credits more than a student with a placement test score in the lowest quartile (Belfield & Crosta, 2012). Students who placed in the lowest quartile also earned more developmental credits (5.4) than students whose scores were in the highest quartile (Belfield & Crosta, 2012).

Aragon and Johnson (2008) investigated demographic, enrollment, academic, and self-directed learning factors of online students in one community college. Their Chi-Square and t-tests found no significant differences in most of the factors, including course completion based on placement in developmental reading, mathematics, or writing in online courses for student's completion of online courses (Aragon & Johnson, 2008). The significant differences were found in gender ( $\chi^2(1, n = 305) = 5.64, p < .05$ ) and GPA ( $t(303) = 4.45, p < .001$ ) between completers and non-completers (Aragon & Johnson, 2008). Crisp and Delgado (2014) conducted a study on the persistence of students in

remedial courses. Their nationwide study included students under 24 years old who began at two-year colleges with the intent to transfer to four-year institutions for the study (Crisp & Delgado, 2014). They used propensity scores to explore the effect of remediation on students and discovered that remedial (79%) and non-remedial (77%) students persisted in their second academic year at similar rates (Crisp & Delgado, 2014). There was also no significant relationship was found between the type of remedial courses (mathematics, reading, or English) and students' persistence decisions (Crisp & Delgado, 2014). They did find remedial students were significantly different than non-remedial students by gender ( $MD = 1.000, p < .001$ ), ethnicity ( $MD = 0.075, p < .001$ ), first-generation status ( $MD = 0.076, p < .001$ ), high school grade point average ( $MD = -0.057, p < .001$ ), highest mathematics class taken in high school ( $MD = 0.088, p < .001$ ), earning college credit during high school ( $MD = -0.065, p < .001$ ), and delaying entry into college ( $MD = 0.074, p < .001$ ) (Crisp & Delgado, 2014). Stewart et al. (2015) investigated the effects between demographic variables, family characteristics, pre-college and college academic performance factors, and remedial courses and retention at a large, residential university using a factorial analysis of variance (ANOVA). Remedial status  $F(1, 3212) = 9.582, p = .047$ , had a significant effect on retention along with financial aid and ethnicity (Stewart et al., 2015). Xu and Dadgar (2018) developed a regression discontinuity design with degree goal, background factors, including dual enrollment status, academic factors for Virginia community college students who took the Prealgebra COMPASS exam. Students enrolled in the lowest-level course (15%) were less likely to earn a credential in four years than students enrolled in the middle-level courses (9%) (Xu & Dadgar, 2018). The other variables were not significant in the

models, with the authors acknowledging that the differences in the community colleges could have impacted the results (Xu & Dadgar, 2018).

**Number of Courses Completed.** Community college students lag behind traditional four-year students in the number of courses completed due to noncontinuous enrollment throughout their academic careers (Monaghan & Attewell, 2015). Within the USG system, students are encouraged to take at least 15 credit hours each semester and graduate in four years (What is a Momentum Year, 2019). Their research has found that students who complete the 30 or more credits during that year are more likely to graduate than just taking 15 hours in the first semester and decreasing the number in the second semester (What is a Momentum Year, 2019). Students may not begin their educational journey in the fall semester but enroll in the following spring or summer semester (Leinbach & Jenkins, 2008). Many retention models subdivide students into part-time and full-time status based on the number of hours attempted every semester to identify differences in these two groups.

Mertes and Hoover (2014) used a Chi-square analysis to determine if age, gender, ethnicity, credit hour load, educational goal, remedial need (English and math), and receipt of financial aid were significant for retention at a midwestern community college using two different fall semesters of data. Gender (fall 2007,  $\chi^2(1, n = 587) = 5.265, p < .05$  and fall 2010,  $\chi^2(1, n = 872) = 8.179, p < .01$ ), program of study fall 2007,  $\chi^2(3, n = 587) = 17.634, p < .001$  and fall 2010,  $\chi^2(2, n = 872) = 17.637, p < .001$ ), grade in the introductory course (fall 2007,  $\chi^2(3, n = 587) = 140.976, p < .001$  and fall 2010,  $\chi^2(3, n = 872) = 20.356, p < .001$ , and age (fall 2007,  $\chi^2(3, n = 587) = 20.682, p < .001$  and fall 2010,  $\chi^2(3, n = 872) = 16.877, p < .001$  were significant in retention (Mertes & Hoover,



2014). Race or ethnicity was statistically significant in retention for fall 2010 only,  $\chi^2$  (6, n = 821) = 13.853,  $p = .031$  with African-American and Hispanic students have lower retention rates than their White counterparts (Mertes & Hoover, 2014). Mertes and Hoover (2014) found retention rates were higher for students who took 12 or more credit hours in a midwestern community college in two different year's data using a Chi-square analysis.

Nakajima, Dembo, and Mossler (2012) discovered the number of courses completed was significant in student retention using community college students' data and t-tests. They examined multiple factors including demographic, financial, academic, academic integration, and psychosocial to determine their relationship to student retention for 427 students at a community college in California and found positive correlations to retention for credit hours enrolled ( $r = .179, p < .01$ ); receipt of financial aid ( $r = .122, p < .05$ ); and cumulative college GPA ( $r = .125, p < .05$ ). There was no significant correlation between any of the psychosocial variables or the academic integration variable to retention (Nakajima, Dembo & Mossler, 2012). The second part of their study used community college students' data and t-tests and determined the following variables were significant in their association to student retention: age ( $t = 2.127, p < .05$ ); receipt of financial aid ( $t = 2.814, p < .01$ ); credit hours attempted ( $t = 2.246, p < .05$ ); number of credit hours completed ( $t = 2.218, p < .05$ ); number of current credit hours enrolled ( $t = 5.442, p < .001$ ); high school graduation year ( $t = 3.307, p < .001$ ); cumulative college GPA ( $t = 2.559, p < .01$ ); English proficiency ( $t = 3.307, p < .001$ ); off-campus employment hours ( $t = 3.363, p < .001$ ); and total employment hours ( $t = 3.097, p < .01$ ) (Nakajima, Dembo & Mossler, 2012).

Fike and Fike (2008) also created a logistic model based on four years of public urban community college academic and background students' data to determine from one semester to the next semester retention status and one complete academic year retention status. Gender, ethnicity, enrollment in a remedial writing course, and completion of a remedial writing course were not statistically significant in the regression model for one semester retention (Fike & Fike, 2008). Positive predictors of one semester retention are passing a remedial reading course ( $\beta = 1.197, p < .001$ ), taking online courses ( $\beta = 0.947, p < .001$ ), participating in a support services program ( $\beta = 0.803, p < .001$ ), not taking a developmental reading course ( $\beta = 0.787, p < .001$ ), passing a developmental mathematics course ( $\beta = 0.762, p < .001$ ), receiving financial aid ( $\beta = 0.473, p < .001$ ), father having some college education ( $\beta = 0.247, p < .001$ ), semester hours enrolled in the first fall semester ( $\beta = 0.153, p < .001$ ), and student age ( $\beta = 0.011, p < .001$ ) (Fike & Fike, 2008). Variables that reduce the odds of fall-to-spring retention in the regression model included not taking a developmental mathematics course ( $\beta = -0.245, p < .001$ ), mother having some college education ( $\beta = -0.157, p < .001$ ), and semester hours dropped in the first fall semester ( $\beta = -0.156, p < .001$ ) (Fike & Fike, 2008). For one academic year retention, age, gender, ethnicity, and not taking a developmental writing course were not statistically significant in the regression model (Fike & Fike, 2008). Positive predictors of one academic year retention are passing a remedial reading course ( $\beta = 1.184, p < .001$ ), taking online courses ( $\beta = 1.151, p < .001$ ), not taking a developmental reading course ( $\beta = 0.978, p < .001$ ), participating in a support services program ( $\beta = 0.756, p < .001$ ), passing a developmental writing course ( $\beta = 0.704, p < .001$ ), passing a developmental mathematics course ( $\beta = 0.698, p < .001$ ), receiving

financial aid ( $\beta = 0.342, p < .001$ ), father having some college education ( $\beta = 0.184, p = .005$ ), mother having some college education ( $\beta = 0.137, p = .029$ ), and semester hours enrolled in the first fall semester ( $\beta = 0.067, p < .001$ ) (Fike & Fike, 2008). Variables that reduce the odds of fall-to-spring retention in the regression model included not taking a developmental mathematics course ( $\beta = -0.412, p < .001$ ) and semester hours dropped in the first fall semester ( $\beta = -0.111, p < .001$ ) (Fike & Fike, 2008). Both models identified a reduction in credit hours during the fall semester and not taking a developmental mathematics course decreased the odds of student retention (Fike & Fike, 2008).

### **Financial Aid Factors**

An additional concern for community colleges is the ability of their students to afford their education. Bean and Metzner's model defines finances as a component in students' ability to be retained in higher education (Bean & Metzner, 1985). Students' financial attitudes slightly impact retention for students in the nontraditional student attrition model and student retention integrated model (Bean & Metzner, 1985; Cabrera et al., 1993). Astin studied the financial aspects and found that students who used parental or personal savings to fund their education had higher degree completion rates (Astin & Oseguera, 2005). The average adjusted public college tuition has increased by over 270% since 1973, while median household earnings increased by 5% (Mitchell & Leachman, 2015). Wohlgemuth et al. (2007) used regression analysis and found that first-year retention increases with grants, scholarships, and work-study, where proceeding years show increases with all types of aid.

Overall yearly U. S. education borrowing decreased in the 2017-2018 academic

year, with full-time undergraduate students borrowing an average of \$4,510 during that period (Baum, 2018). Community colleges in rural areas serve low-income students who may rely heavily on financial aid for persistence in college (Hurford, Ivy, Winters, & Eckstein, 2017). Walker (2016) found a significant correlation between retention rates and the amount of funding available, and the net price of students' educational costs. The government requires students to complete the federal student aid application (FAFSA), which identifies assets, income, demographics, and family structure, that helps determine the expected family financial contributions (EFC) (Choitz & Reimherr, 2013; Denning, 2019). However, EFC can be misleading since many families' economic circumstances do not allow for additional funds to be given to students to support their educational pursuits (Bound, Lovenheim, & Turner, 2012).

Additionally, students are classified as financially independent if they turn 24 years old by January 1st of the following academic year can affect their EFC (Denning, 2019). The EFC, with the subtraction of scholarships and grants, is referred to as the net price (Choitz & Reimherr, 2013). The Pell Grant, a grant awarded based on financial needs, is adjusted based on the EFC and varies yearly due to financial aid rules (Park & Scott-Clayton, 2018). Park and Scott-Clayton (2018) found that full time students who receive the Pell Grant have higher enrollment rates for the spring semester of their first year and fall semester of their second year than non Pell Grant recipients using a regression discontinuity model.

When financial aid and additional resources are unable to pay for their education costs, students at community colleges work longer hours to pay for their courses, which can harm academic performance and completion times for degrees (Bound et al., 2010;

Scott-Clayton, 2012; Johnson & Rochkind, 2009). The shift in student employment hours has increased for all students over the last 40 years but is more pronounced in the community college population (Bound et al., 2010). Astin and Oseguera (2005) examined pre-college characteristics and found that students who worked off-campus and who planned to work full-time had statistically significantly lower completion rates. An increase in employment hours has been associated with longer college completion times for community college students (Bound et al., 2010; O'Toole et al., 2003). The financial variables selected for this study are FAFSA application completion, the amount of financial aid awarded, and the amount of financial aid paid to the student during the first academic year.

**Amount of Financial Aid Awarded.** With 58% of community college students receiving some financial aid to attend college, finances play a significant role in whether students can afford college (Radwin et al., 2018). For the 2015-2016 academic year, 72% of higher education students receive aid, with 39% of the funding coming from the Pell Grant, which awarded an average amount of \$3,700 per student (NCES, 2018a). The Pell Grant is one of the primary sources of financial aid, with awarded amounts based on financial need (Federal Pell Grants, 2019). The Pell Grant provided over 9 million students with 30.3 billion dollars of assistance during the 2014-2015 academic year (Federal Pell Grants, 2019; Park & Scott-Clayton, 2018). The need for aid is more significant for students in lower socio-economic groups, with 73% of Pell Grant recipients came from families with incomes below \$40,000 for 2017-2018 (Baum, 2018; Breier, 2010). DesJardins, Ahlburg, and McCall (2002) developed a hazard model with new first-year students at a large, urban university and found grants (RR = 1.03) had no

significant effect on retention. They discovered that scholarships (RR = 0.28) and work-study (RR = 0.50) had the most significant impact on increasing retention in the first two years (DesJardins, Ahlburg, & McCall, 2002).

**Amount of Financial Aid Paid.** Unsubsidized and subsidized loans often cover the amount of tuition and fees that are not covered by scholarships and grants can shape a student's financial future. Students using federal loans are required to pay them back once they have been out of school for six months, even if they do not complete a degree. The average loan amount for non-completers in 2009 was \$5,700 (Wei & Horn, 2013). Kofoed (2017) used the national postsecondary student aid survey to create propensity scores to find the amounts of aid students miss out on if they don't apply for FAFSA. Students whose family or individual incomes are less than \$10,000 would have a total grant aid of around \$5,464.45, and students whose family or individual incomes making more than \$100,000 would have \$1,784.43 in aid (Kofoed, 2017). The average overall aid across years and income levels is \$3,254.87 (Kofoed, 2017). Herzog (2018) also implemented propensity scores for two cohorts of first-year students at a public research university. He determined that students who take out the maximum amount of subsidized loans have a slightly higher risk of non-persistence than students who do not receive the maximum amount (Herzog, 2018). The study also found Pell Grant-eligible students who took out loans (63%) and students who receive higher amounts of loan aid to pay for their college costs (74%) had an elevated risk of departure after the first year of college compared to students who took out no loans (79%) or less than \$10,000 an academic year (81%) (Herzog, 2018). Gross, Hossler, Ziskin, and Berry (2015) used first time freshmen longitudinal data from all public colleges and universities in Indiana to build a hazard

model to predict the time to departure based on academic preparation, student background characteristics, collegiate academic domain, collegiate social domain, and finances. The model found that age ( $\beta = -0.009$ ,  $p < .001$ ), African American students ( $\beta = 0.188$ ,  $p < .001$ ), Asian and Asian American students ( $\beta = 0.095$ ,  $p < .01$ ), students who didn't indicate race ( $\beta = 0.441$ ,  $p < .001$ ), combined SAT scores ( $\beta = 0.0$ ,  $p < .05$ ), lived off-campus ( $\beta = 0.414$ ,  $p < .001$ ), lived with parents or guardians ( $\beta = 0.531$ ,  $p < .001$ ), college GPA ( $\beta = -0.120$ ,  $p < .001$ ), cumulative credits ( $\beta = -0.027$ ,  $p < .001$ ), and declared major ( $\beta = 0.254$ ,  $p < .001$ ) had a significant impact on the time to departure for college (Gross et al., 2015). All of the financial factors had a significant effect on the time of departure; received any aid ( $\beta = 0.179$ ,  $p < .001$ ), received need-based aid ( $\beta = 0.191$ ,  $p < .001$ ), had cumulative loans ( $\beta = 0.012$ ,  $p < .001$ ), applied for aid ( $\beta = -0.100$ ,  $p < .01$ ), net price ( $\beta = 0.028$ ,  $p < .001$ ), ratio of loans to total aid ( $\beta = -0.745$ ,  $p < .001$ ), institutional merit aid ( $\beta = -0.063$ ,  $p < .01$ ), and institutional need aid ( $\beta = -0.058$ ,  $p < .001$ ) (Gross et al., 2015). Merit-based aid recipients were less likely to leave their institutions with an increase of \$1000 merit-based aid resulting in a 5% decline in the odds of departing (Gross et al., 2015). They also found that with the addition of \$1000 to the net price of the students' education, the odds of departure (2.5%) increased (Gross et al., 2015). Jones-White, Radcliffe, Lorenz, and Soria (2014) created a multinomial regression model for first-year students at a large, mid-western research university to estimate the relationship between 6-year retention and financial student background, academic, and social factors. Significant factors for continuous retention at the current institution were identified as female (RRR = 0.8239,  $p < .01$ ), being an underrepresented minority (RRR = 1.4669,  $p < .01$ ), being a first generation student (RRR = 1.4007,  $p <$

.001), being older than 19 years old (RRR = 3.4570,  $p < .001$ ), having a composite ACT score (RRR = 1.0252,  $p < .05$ ), number of advanced placement (AP) credits (RRR = 0.9612,  $p < .001$ ), remedial coursework (RRR = 2.9302,  $p < .001$ ), first semester course completion percentage (RRR = 0.9598,  $p < .001$ ), number of C grades awarded during the first semester (RRR = 1.4969,  $p < .001$ ), , number of D grades awarded during the first semester (RRR = 1.9247,  $p < .001$ ), living on campus in a non-living learning community (RRR = 0.7729,  $p < .01$ ), living on campus in a living learning community (RRR = 0.7322,  $p < .05$ ), and having athlete status (RRR = 0.5342,  $p < .01$ ) (Jones-White et al., 2014). The significant financial factors were the amount of loan aid received (RRR = 1.0751,  $p < .001$ ) and the amount of merit aid received (RRR = 0.5763,  $p < .001$ ), which revealed that the larger a first-year student's financial needs are unmet, the higher the risk they are for not completing their degrees (Jones-White et al., 2014). Additionally, they found that as students' monetary awards increase, their risk for non-completion also increases, suggesting that students facing larger loan debt may find continuing their education as cost-prohibitive (Jones-White et al., 2014).

**FASFA Completion.** In 2007-2008, 69.7% of students failed to fill out a FAFSA could have received additional financial aid which could have to minimize their employment hours (Kantrowitz, 2009, McKinney & Novak, 2012). McKinney and Novak (2015), using the beginning postsecondary student survey, found that students attending part-time have lower FAFSA filing behavior than full-time students, which could have helped with affording college. Using a regression model, they found students with the following factors; males ( $\beta = 1.607$ ,  $p < .001$ ), African Americans ( $\beta = 0.556$ ,  $p < .001$ ), parents with less than an associate's degree ( $\beta = 0.597$ ,  $p < .001$ ), part time status ( $\beta =$



1.745,  $p < .001$ ), delayed enrollment a year or more ( $\beta = 1.752$ ,  $p < .001$ ), undeclared major ( $\beta = 1.425$ ,  $p < .01$ ), and expected family contribution ( $\beta = 1.045$ ,  $p < .001$ ) were associated with not filing a FAFSA (McKinney & Novak, 2015). Kofoed (2017) used the national postsecondary student aid survey to build a multinomial logit model to identify factors of FAFSA completion. Populations who are Black ( $\beta = 0.144$ ,  $p < .01$ ), Hispanic ( $\beta = 0.087$ ,  $p < .01$ ), female ( $\beta = 0.038$ ,  $p < .01$ ), in a higher socioeconomic status ( $\beta = -0.086$ ,  $p < .01$ ), and dependent financially on their parents ( $\beta = 0.129$ ,  $p < .01$ ) are more likely to complete the FAFSA (Kofoed, 2017).

McKinney and Novak (2012) investigated FAFSA filing status and first-year retention for community colleges using the beginning postsecondary student survey. They found that gender, ethnicity, English as a primary language, parents' education, high school mathematics level, high school GPA, remedial coursework, major, Pell-eligible, have dependents on their taxes, and hours worked were not significant factors for first-year retention (McKinney & Novak; 2012). The factors that were significant for retention were delayed enrollment ( $\beta = 0.58$ ,  $p < .01$ ), college GPA ( $\beta = 1.67$ ,  $p < .001$ ), part-time status ( $\beta = 0.33$ ,  $p < .001$ ), meeting with an advisor ( $\beta = 1.43$ ,  $p < .05$ ), and had filed a FAFSA application ( $\beta = 1.79$ ,  $p < .01$ ) (McKinney & Novak, 2012). Community college students who filed a FAFSA resulted in 79% higher odds of persisting when controlling for the other predictors in the model with a greater impact on part-time students (McKinney & Novak, 2012). Part-time students who filed a FAFSA application had 100% higher odds of being retained at community colleges than part-time students who did not file a FAFSA (McKinney & Novak, 2012). In Georgia, students who complete the FAFSA are eligible for the HOPE scholarship, a merit-based

scholarship program that pays for tuition for in-state colleges and technical schools. The HOPE scholarship has no income restriction on the recipients, which could help students in every financial bracket

pay their tuition (Cornwell, Mustard, & Sridhar, 2006). Cornwell, Mustard, and Sridhar (2006) did not find a rise in two-year level students based on the HOPE scholarship, but White and Black enrollment increased as a result.

Students who do not fill out the FAFSA have a higher net cost (43.9%) for college than students who do (Choitz & Reimherr, 2013). Many students fail to fill out the FAFSA, believing they will not qualify for financial aid or can't afford to attend (McKinney & Novak, 2015; Oreopoulos & Dunn, 2013). Oreopoulos and Dunn (2013) discovered that high school students who did not believe they could afford college due to cost changed their mindset after watching a video tutorial on higher education costs. Other noncompletion reasons are privacy concerns, funding from employment or other sources, country residence status, confusion about the form, and missing the application deadline (Kantrowitz, 2009; McKinney & Novak, 2015). Bettinger, Long, Oreopoulos, and Sanbonmatsu (2012) created an experiment in Ohio and Charlotte, North Carolina, where students and their families could receive FAFSA assistance after filing their taxes. Their regression model found that this assistance leads to an increase in different groups of students filling out the FAFSA: students who were classified as dependents ( $\beta = 0.035$ ,  $p < .01$ ), students who were not dependent on their parents and had not attended college ( $\beta = 0.009$ ,  $p < .01$ ), and students who were not dependent on their parents and had attended college previously ( $\beta = 0.012$ ,  $p < .01$ ) (Bettinger et al., 2012). They also found that enrollment rates for dependent students whose families received help filling out the

FAFSA application ( $\beta = 0.035, p < .05$ ) were 8% higher than families who did not receive assistance (Bettinger et al., 2012). Adult students with no prior college experience enrollment rates were 16% higher ( $\beta = 0.007, p < .05$ ) than other adult students who did not participate in the study (Bettinger et al., 2012). Owen and Westlund (2016) investigated the role of school counseling financial aid on FAFSA completion and college attendance in a large urban school district with 21 high schools in the southwestern region of the United States of America. They ran a linear probability model and found students who received financial counseling had higher FAFSA completion rates ( $\beta = 0.103, p < .001$ ) and attended college in higher rates ( $\beta = 0.117, p < .001$ ) than non-counseled students (Owen & Westlund, 2016). The investigation of financial factors in the student's ability to continue at a community college needs to be localized to the institution level to help identify factors for that specific population (Gross, Hossler, Ziskin & Berry, 2015; Herzog, 2018).

Students who file out the FAFSA early have higher retention rates than students who delay with early filers receiving more financial aid money than late registrants with early FAFSA filers received \$700 more in aid than late filers in the 2003-2004 academic years (LaManque, 2009; McKinney & Novak, 2012; McKinney & Novak, 2015). McKinney and Novak (2015) used academic, background, and financial data from the beginning postsecondary student study to investigate FAFSA filing behavior for students attending community colleges and four-year colleges. Using a regression model, they found students with the following factors; males ( $\beta = 1.273, p < .05$ ), no high school degree or GED ( $\beta = 2.300, p < .01$ ), took mathematic courses lower than Algebra II in high school ( $\beta = 1.430, p < .01$ ), and delayed enrollment a year or more ( $\beta = 2.811, p <$

.001) were associated to file a FAFSA late (McKinney & Novak, 2015). They found that community college students were more likely to file their FAFSA later than students attending other institutions with financial aid amounts 60% less than their peers who filed earlier. They also found that community college students who complete the FAFSA early were awarded approximately \$700 more in funds than late filers (McKinney & Novak, 2015). Feeney and Heroff (2013) created a logistic regression model from the financial records of first-generation students from Illinois and found that first generation students filed FAFSA applications later ( $\beta = 0.708, p < .001$ ) than other students. Additionally, they found that students who had a weaker academic performance in high school ( $\beta = 0.180, p < .001$ ) were significantly more likely not to complete the FAFSA early as well as female students ( $\beta = 1.220, p < .05$ ) compared to male students (Feeney & Heroff, 2013). Students who had no expected family contribution were significantly less likely to complete the FAFSA by the priority date ( $\beta = 0.784, p < .01$ ) or late cutoff ( $\beta = 0.555, p < .001$ ) (Feeney & Heroff, 2013). Even when a student completes the FAFSA early, their application can be flagged for income verification, which can stall the process for completion. Page, Castleman, and Meyer (2016) found that economically disadvantaged students (42%) and non-White non-economically disadvantaged students (39%) applications were flagged in higher numbers than White non-economically disadvantaged students (26%).

### **Introduction of Data Science and Big Data**

The introduction of digital mediums in people's lives has increased the amount of data produced worldwide and ushered in the era of "big data." Big data is more than just the data itself; but the amount of data collected, the data collection rate, and the types of

data collected (McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012). During 2016-2017 alone, 2.5 quintillion bytes of data were created daily (Marr, 2018). In addition, the rise of social media and people's digital daily routines have introduced new forms of data that needed to be analyzed. Statistical methods are limited to numeric data and tend to capture the behavior of smaller, clean data sets (Hand, 1998).

Data science or data analytics is the process of using data of all types to determine new patterns or ideas (Roiger, 2017). Data science's overall goal is to produce models to solve problems through data acquisition, data processing, modeling determination and use, cross validation, reporting results, and repetition of results (Roiger, 2017). Data science has recently transitioned into social sciences and educational fields, including higher education (Provost & Fawcett, 2013). Data-driven decisions help companies become more productive and profitable than their competitors, which could implement higher education (McAfee et al., 2012). Higher education can use data analytics to target specific issues that affect their students and help redefine strategies for the institution's future (Drigas & Leliopoulos, 2014). IBM software group (2018) highlighted four different ways of data analytics, specifically predictive analytics, that could help educators better serve their students:

1. Improving student performance with predictive analytics;
2. Developing effective student retention strategies;
3. Building a 360-degree profile of students; and
4. Enhancing enrollment management.

With available tools to harness available big data and analytics, higher education could help increase student retention using the data they have available and investigating

new directions such as learning management systems analytics (Krüger, Merceron & Wolf, 2010; Picciano, 2012). Proponents of big data call for ethical processes for handling the data, understanding the limitations of using big data, and a moral responsibility to address inequality issues that arise from the results of these analytics (Eynon, 2013). Larger companies have the technical components to handle “big data” with these requirements limiting the overall feasibility of individuals and smaller institutions to run similar methods (Attewell & Monaghan, 2015; Provost & Fawcett, 2013). Newer software has allowed for smaller personal computers to handle data sets that were once too large for them to handle (Attewell & Monaghan, 2015). Packages like SPSS, R, RapidMiner, Weka, and TraMiner can be run on smaller computers and have a lower price point than business enterprise software, allowing for a greater audience of users (Attewell & Monaghan, 2015). Additionally, researchers can run more extensive models with numerous iterations and types to optimize their results.

Outside of the uses of data analytics within higher education, companies will need employees trained in data analytics, which means higher education must be able to train and educate on these topics (Drigas & Leliopoulos, 2014). Free and reduced cost software allows for higher education to educate these students without a significant financial commitment in software and hardware costs. Students can replicate their classroom experiences at minimal expense once they leave the institutions.

### **Data Mining**

Within data science, data mining (DM) is the name for a group of computer-driven methods that discover the structure and identify patterns in data sets (Attewell & Monaghan, 2015, Bharati & Ramageri, 2010). DM can analyze text, sound, and visual

data, as well as numerical data in large datasets with many variables (Attewell & Monaghan, 2015). In addition, DM can handle large, unclean data sets and identify significant predictors among numerous variables (Attewell & Monaghan, 2015). The idea of DM is not new; statisticians have used data dredging to identify new patterns in the data but faced limitations on the validity of the patterns being random or consistent (Hand, 1998). DM algorithms and techniques are categorized by their functions: classification, clustering, prediction, and association (Bharati & Ramageri, 2010).

### **Educational Data Mining**

According to the website [educationaldatamining.org](http://educationaldatamining.org). (n.d.) “educational data mining is an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings and using those methods to understand students better, and the settings in which they learn.” In educational data mining, relationship mining was predominant in 1995-2005, with a shift to predictive methods by 2008-2009 (Baker & Yacef, 2009). Educational data mining research has focused on student retention and attrition, personal learning environments, and recommender systems (Huebner, 2013). Data mining can allow new patterns to emerge from commonly used data and allow for precise interventions for specific students (Chacon, Spicer & Valbuena, 2012; Herzog, 2006; Lin 2012; Luan, 2002; Yu, DiGangi, Jannasch-Pennell & Kaprolet, 2010). Luan (2002) used classification trees, neural networks, and cluster analysis to identify at-risk students by demographics and course enrollment patterns. Chacon, Spicer, and Valbuena (2012) took the data mining results, which identified at-risk students and implemented specific notifications to staff and faculty to address these students’ needs. Lin (2012) used data mining methods to

predict which new students were at risk of dropping out. Herzog (2006) found that data mining methods, decision trees, and neural networks performed comparably to regression models for freshmen retention. Yu et al. (2010) found that transfer hours, residency, and ethnicity were essential to student retention using data mining techniques of classification trees, neural networks, and multivariate adaptive regression splines (MARS).

### **Classifiers**

One of the most common types of data mining is classifiers, which are different models that predict which classes the dependent variables individual cases belong to (Attewell & Monaghan, 2015; Bharati & Ramageri, 2010; Breiman, 1999; Breiman et al., 1984; Han et al., 2011). The past behavior of the variables is measured to predict where the cases belong and which cases belong together (Breiman et al., 1984). Classification studies may identify classifiers or define the overall predictive nature of a group of variables to describe a phenomenon (Breiman et al., 1984). Patterns of behaviors can identify the significance of the variables within the classification process (Han et al., 2011). Classifiers come from numerous families, including decision trees, neural networks, support vector machines, random forests, regression models, and other methods. Researchers often use classification methods they are familiar with and may not be the most accurate classifier for the problem (Fernández-Delgado, Cernadas, Barro & Amorim, 2014). One issue is that classifiers come from different disciplines such as statistics, mathematics, or computer science, and researchers may not be aware of all of the possibilities available (Fernández-Delgado et al., 2014). An additional consideration is the numerous software packages and languages that run the classifiers and the knowledge needed to compute the classifiers within each one successfully.



Wolpert's No Free Lunch Theorem indicated no one classifier could handle all data sets (Wolpert, 1996). Fernández-Delgado, Cernadas, Barro & Amorim (2014) measured the accuracy rates on 179 different classifiers on 121 data sets to determine classifier behavior and accuracy regardless of the data sets. Their results found that random forests, support vector machine (SVM), neural networks, and boosting ensembles had the most accurate results among the 121 different data sets (Fernández-Delgado et al., 2014). Kuhn and Johnson (2013) recommend that researchers start with complex models with the most flexibility and less interpretability to give the most accurate results. Simpler models like logistic regression can be used with sophisticated models to compare the accuracy and interpretation of the equations (Attewell & Monaghan, 2015; Kuhn & Johnson, 2013). Additionally, researchers can combine the predictions from multiple classifiers models in a technique called ensemble learning that can yield more accurate results than singular classifier models (Attewell & Monaghan, 2015).

### **Cross Validation Methods**

Data mining models may overfit the data, leading to the model's poor predictive ability with decreased accuracy (Attewell & Monaghan, 2015; Kuhn & Johnson, 2013). Overfitting is when a predictive model is created for one data set and will not perform as adequate on other datasets (Attewell & Monaghan, 2015). One way to minimize overfitting is using a technique called cross validation which divides the data set performance by random assignment in different groups; a training set used to build the models, and a testing set is used to assess the models' performance (Attewell & Monaghan, 2015; Bost et al., 2015; Efron, 1983; Han et al., 2011; Kuhn & Johnson, 2013). There are various cross validation methods, with one of the simplest being the

holdout method, where one training set and one testing set are created and used (Attewell & Monaghan, 2015). This method is ideal for large datasets with many variables.

Methods for smaller datasets, including bootstrapping, develop numerous iterations of the data set into training and testing sets. These various random samples provided a range of results when averaged together, obtained an overall result, and validated if patterns or trends occur (Attewell & Monaghan, 2015; Kuhn & Johnson, 2013; Kuhn & Johnson; 2019). Bootstrapping takes random samples of the sample with replacement and creates numerous training and testing sets that help produce the more accurate model (Attewell & Monaghan, 2015; Kuhn & Johnson, 2013; Kuhn & Johnson; 2019). The benefits of bootstrapping include a smoother estimate of the error rate, standard error of the prediction, standard error of the error rate, and a nonparametric standard error that is not dependent on the data distribution (Attewell & Monaghan, 2015). Cross validation can help with model comparison for each model using the same training and test data set. The output results can be compared to the models to determine how robust each model is individually and to each other.

### **Decision Trees and Random Forest Trees**

With the introduction of data science, newer techniques have allowed different methods to explore the data, including decision trees. Decision trees structurally mimic trees found in nature and provide a flowchart structure that is easy to interpret (Han et al., 2011). Data is classified starting at the root and moving throughout the tree down to the leaves. The root node is the beginning point of the classification and extends throughout the internal node to test a specific characteristic; then, there are branches that show the outcomes of the test and leaves that define the particular class label (Han et al., 2011).

Decision trees can be used for exploratory analysis and are intuitive (Han et al., 2011). This type of classification modeling can identify factors and relationships that other statistical models may not find (Mendez, Buskirk, Lohr & Haag, 2008).

There are multiple types of decision trees. Quinlan developed one of the earliest examples of decision trees called iterative dichotomiser (ID3), which takes one part of the training set, called a window, into a tree and uses this tree to classify the remaining portion of the training set until everything is accounted for (Quinlan, 1986). Non-classified data is added to the window to develop a new tree (Quinlan, 1986). Quinlan went on to form another type called C4.5, which can handle continuous and discrete data, handles missing data, and prunes the tree where each non-leaf subtree becomes a leaf (Quinlan, 2014). Another type of trees called CART (classification and regression trees) was developed in 1983 by a group of statisticians. CART's two analytic methods are based on the dependent variable's classification - classification trees for nominal or categorical data and regression trees for interval or continuous data (Breiman et al., 1984; Ma, 2018). CARTs use the predictors to create a model that divides the data into categories based on different interests (Attewell & Monaghan, 2015; Ma, 2018). Ma (2018) describes this type of classifier "as a data-mining technique, is a new tool for exploratory data analysis often performed in inductive research or data-driven research." (p. 2). Classification trees can predict the outcome variable while developing a complex set of classification choices that span the entire data set (Attewell & Monaghan, 2015). This classifier can handle any outcome variable, making it more advantageous than traditional predictive classifiers like logistic regression, and is easily interpreted (Attewell & Monaghan, 2015; Kuhn & Johnson, 2013). All three of these decision trees use a

greedy method that does not allow for backtracking but a top-down approach starting with the training phase (Han et al., 2011). These trees can overfit the data and may not handle large datasets effectively (Ali, Khan, Ahmad, & Maqsood, 2012). The process of pruning the decision trees can help decrease the overfitting and increase the accuracy of the models. Decision trees are an ideal model to help predict the retention behavior of students in higher education (Yadav, Bharadwaj, & Pal, 2012).

Random forest trees are an ensemble method using a group of decision trees to form the forest, and the tree with the most votes becomes the model used (Han et al., 2011). This method allows for a group of weak classifiers to form one robust classifier model using bootstrapping methods (Attewell & Monaghan, 2015; Mao & Wang, 2012). Each model has a different set of predictors that allows for nonreplicated structure and content among the models (Attewell & Monaghan, 2015). These differences in the predictors for each tree provide identification on low correlation and helps increase the reliability of the overall model (James, Witten, Hastie & Tibshirani, 2013). In addition, the randomness of the trees allows for better prediction (Brieman 1999). This model is better equipped to handle errors and outliers than individual decision trees, and it addresses the issues of overfitting that occur with an increase in generalization to other independent samples (Attewell & Monaghan, 2015; Han et al., 2011). Random forests are applicable for large data sets and estimate variable importance (Han et al., 2011). This classifier is non-parametric in behavior and can handle binary, continuous, and categorical data (Ali, Khan, Ahmad & Maqsood, 2012).

Among classifiers, random forest trees have some of the highest accuracy rates among different data sets and disciplines (Caruana & Niculescu-Mizil, 2006; Dissanayake,

Robinson & Al-Azzam, 2016; Fernández-Delgado et al., 2014; He, Levine, Fan, Beemer & Stronach, 2018). Within higher education, random forest trees have predicted student progress, student performance, completion and graduation rates, and licensing rates, but they are less used than decision trees (Goga et al., 2015; Hardman, Paucar-Caceres & Fielding, 2013; He et al., 2018; Hutt, Gardener, Kamentz, Duckworth & D'Mello, 2018; Langan, Harris, Barrett, Hamshire & Wibberley, 2018). Goga et al. (2015) examined admissions data from a specific university and identified random forest trees as the most accurate method (99.908%) for predicting background variables significant for student performance compared to other classifiers. Hardman, Paucar-Caceres, and Fielding (2013) used virtual learning environments and management information systems data to create random forest trees that identified significant predictors for a university in the United Kingdom. Hutt et al. (2018) developed a random forest tree model that correctly predicted four-year graduation rates of students using the Common Application data for the United States. Random forest trees can handle data that regression methods could not: a huge number of predictors compared to sample size, multicollinearity, and nonlinearity and provide reliable estimates of variable importance (He et al., 2018; Liaw & Wiener, 2002).

### **Support Vector Machines (SVM)**

Another type of classifier is SVM, which can be used as a classifier or regression function and help determine inputs in high dimensional feature space (Delen, 2010).

Among classifiers, SVM has a high accuracy rate among numerous data sets and disciplines (Caruana & Niculescu-Mizil, 2006; Delen, 2010; Fernández-Delgado et al., 2014). Delen (2010) found SVM models produced the most accurate model for predicting

students who are more likely to drop out after the first years using a balanced dataset (81.18%) and an unbalanced dataset (87.23%). For higher education data, SVMs have been used to predict student retention with mixed results to other classifier methods in accuracy (Delen, 2010; Lauría, Baron, Devireddy, Sundararaju & Jayaprakash, 2012; Zhang, Oussena, Clark & Kim, 2010). Lauria et al. (2012) identified SVMs as performing comparable to logistic regression and outperforming decision trees in identifying academic risk factors. Zhang, Oussena, Clark, and Kim (2010) also found that SVMs outperformed decision trees but were not as accurate as Naïve Bayes when predicting academic risk factors.

The SVM algorithm has been around since the 1960s, with the first paper on them was developed by Boser, Guyon, and Vapnik in 1992 and revised to its current form in 1995 by Vapnik and Cortes (Boser, Guyon & Vapnik, 1992; Cortes, & Vapnik, 1995; Han et al., 2011). SVMs were developed for binary classification but can handle multiclassification, regression, clustering, anomaly detection, and feature selection (Attewell & Monaghan, 2015). SVM behaves similarly to other classifiers within classification methods by separating data into groups while minimizing error (Attewell & Monaghan, 2015). Two-dimensional data has linear separation, three-dimensional data has separation by a plane, and n-dimensional data separation by a hyperplane (Attewell & Monaghan, 2015; Han et al., 2011). There are infinite linear boundaries for binary classification that could classify the data, with many accuracy measurements being equivalent (Kuhn & Johnson, 2013).

Margin creations on both sides of the hyperplane help to standardize the multiple boundaries into one optimal solution. These margins measure the distance from the

boundary between the classes and their closest data points. The boundary with the largest margin is called the maximum marginal hyperplane (Han et al., 2011; James et al., 2013; Kuhn & Johnson, 2013). Hyperplanes with the largest margins have the most generalized accuracy since they give the largest separation between the classes using SVM (Han et al., 2011). The formula for the hyperplane

$$w_0 + w_1x_1 + w_2x_2 = 0$$

where  $w_1$  and  $w_2$  are the weights,  $x_1$  and  $x_2$  are the values for the attributes for the variables, and  $w_0$  is the bias of the model (Han et al., 2011). The formula for the area above the hyperplane, H1, is

$$w_0 + w_1x_1 + w_2x_2 > 0$$

where any tuples above or on H1 are greater than 0 (Han et al., 2011). The formula for the area below the hyperplane, H2, is

$$w_0 + w_1x_1 + w_2x_2 < 0$$

where any tuples below or on H2 are less than 0 (Han et al., 2011). The maximal margin hyperplane formula makes sure each tuple is on the right side of the hyperplane and maximizes the hyperplane margin (James et al., 2013). Most of the tuples in the training set will fall above H1 or below H2 and are easily classified. Any tuples that fall on H1 or H2 are called support vectors (Han et al., 2011). While they are difficult for classification purposes, they can provide the most information for classification (Han et al., 2011).

One limitation of the maximal margin hyperplane is noisy data, which can skew the robustness of the hyperplane (James et al., 2013). In these cases, the hyperplane might not separate the data into two classes, with the model providing greater robustness and

better classification for the majority of the training data set (James et al., 2013). SVMs, which can also be called soft margin classifiers, allow for data to be on the wrong side of the margin or hyperplane. The data points that lie on the margin or the wrong side of the margin are called support vectors and affect the support vector classifier (SVC) more than correctly classified data (James et al., 2013). SVC works well with linear data but handles non-linear data poorly (James et al., 2013). Using the quadratic, cubic, higher-order polynomial, and other functions for the predictors may correct this issue (James et al., 2013). These functions could lead to complex computations that could be difficult to handle (James et al., 2013).

SVMs help with this issue using kernel functions that transform the data based on the kernel function used (Han et al., 2011; James et al., 2013; Ngemu, Elisha, William & Bernard, 2015). Numerous kernel functions transform the data based on the specific function (Fernández-Delgado et al., 2014; James et al., 2013). Examples of kernels are the linear kernel, the polynomial kernel, the Gaussian kernel, and the radial kernel. The optimal kernel discovery is a process of trial and error (Attewell & Monaghan, 2015). SVMs can adjust their margins using the cost parameter where a small cost value allows for wide margins where many support vectors may occur on the margin or violate the margin (James et al., 2013). Conversely, larger cost values create smaller margins and decrease the number of support vectors on the margin or violate margin (James et al., 2013). One drawback to SVM is its sensitivity to predictors with skewed distributions and outliers (Kuhn & Johnson, 2019). Another disadvantage is that SVM often overfits the data, and cross validation is essential to correct this issue (Attewell & Monaghan, 2015).



## Neural Network

Neural networks, also called artificial neural networks, are classification models that mimic the operations of biological neurons, with neurons passing information to other neurons with the ability to learn based on previous errors (Attewell & Monaghan, 2015). The neural network description is a “set of connected input/output units in which each connection has a weight associated with it (page 398)” (Han et al., 2011). Neural networks have predicted student course selection, institutional application, retention, and graduation times (Delen 2010; González & DesJardins, 2002; Herzog, 2006; Kardan, Sadeghi, Ghidary & Sani, 2013; Luan, 2002). Delen (2010) discovered neural networks (86.45%) and logistic regression (86.12%) performed similarly for predicting students who are more likely to drop out after the first years. González and DesJardins (2002) showed that neural networks (78%) outperformed logistic regression (72%) when predicting which students would apply to a large research institution based on correct classification rates. Herzog (2006) found that neural networks performed similarly to logistic regression and decision trees to predict student retention. Kardan, Sadeghi, Ghidary, and Sani (2013) created two individual neural network models that both outperformed SVMs, K-nearest neighbors, and decision trees when predicting student course selection in online higher education institutions. Luan (2002) found neural networks outperformed classification trees in predicting the retention of community college students.

Neural networks start with a dataset with various independent variables called “inputs,” with each input variable assigned a random weight comparable to a regression coefficient (Attewell & Monaghan, 2015). The input information is gathered through

summation and transformed into a nonlinear function to an output (Attewell & Monaghan, 2015). The process of assigning weights, sums, and transformations occur in the hidden layer and are called hidden nodes (Han et al., 2011). This hidden layer primarily performs a nonlinear regression with the input variables (Attewell & Monaghan, 2015; Han et al., 2011). A simple neural network with numerous inputs, one hidden node, and one output is similar to logistic regression (Attewell & Monaghan, 2015). The addition of more hidden nodes can improve accuracy compared to traditional logistic regression since the weights of the inputs are randomly assigned and recalculated for each hidden node (Attewell & Monaghan, 2015). Throughout the process, the neural formula starts at a random point and goes through each observation until it learns from its predictive error and fine-tunes the parameters (Attewell & Monaghan, 2015).

Neural networks are self-sufficient and do not require transformations or interactive terms found in other classifier methods since the model is mapping itself (Attewell & Monaghan, 2015). This type of model has a tuning requirement that can help maximize the accuracy in predicting (Attewell & Monaghan, 2015). While neural networks can map out the results, the same dataset with the same variables can produce two different results since the hidden layer assigns random weights (Attewell & Monaghan, 2015). To avoid this issue, a researcher should use a fixed seed to help with weights in the hidden layer (Attewell & Monaghan, 2015). Neural networks can handle nonlinear and missing data among the variables but produce better results with sample sizes greater than 500 (Herzog, 2006). This model can handle various types of data: binary, continuous, and categorical and is considered superior in prediction accuracy over regressions and classification tree models (Attewell & Monaghan, 2015).

Even with higher classification accuracy, neural networks can be challenging to interpret the relationship between the inputs and outputs (Attewell & Monaghan, 2015; González & DesJardins, 2002). They also require longer training times but can often recognize patterns that are undetected in the other types of classifiers (Etheridge, Sriram & Hsu, 2000; Han et al., 2011). Neural networks can overfit the data and may need correction with cross validation methods or decreasing the number of hidden nodes (Attewell & Monaghan, 2015).

### **Logistic Regression**

One of the most common classification methods used in higher education is logistic regression and dates back to the 1960s (Cabrera, 1994). Higher education research using logistic regression range from predictors for student retention, student graduation, and numerous subjects dealing with students and faculty (Astin & Oseguera, 2005; Chatterjee, Marachi, Natekar, Rai & Yeung, 2018; Delen, 2010; Herzog, 2006; Lauría et al., 2012; Pyke & Sheridan, 1993). Delen (2010) discovered that logistic regression's (86.12%) accuracy was lower than SVM (87.23%), decision trees (87.16%), and neural networks (86.45%) for predicting students who are more likely to drop out after the first years. Herzog (2006) found that logistic regression performed comparably on neural networks and decision trees to predict student retention. Lauria et al. (2012) identified logistic regression as outperforming SVMs and decision trees in identifying academic risk factors.

With the logistic regression's similarity to linear regression and a more straightforward interpretation of the results versus other data mining techniques, many educational researchers choose logistic regression as their statistical method (Gunu, Lee,

Gyasi & Roe, 2017; Peng, So, Stage & John, 2002). Between 1988 and 1999, 52 articles in the three higher education journals used logistic regression as their statistical method (Peng et al., 2002). While many educational researchers use logistic regression, they often violate the parameters for using this classifier method correctly, including small sample sizes, incorrect transform interpretation, and sampling bias (Peng et al., 2002). With the introduction of data mining, logistic regression may not be the best classifier for current and future educational research.

Logistic regression has its roots in statistics and is considered a supervised-learning binary classification system where the outcome is binary (Homer, Lemeshow, & Sturdivant, 2013). The overall outcome of this retention model has two distinct events using the binomial distribution ideal for this outcome, with  $p$  representing the probability of an event occurring and  $1-p$  representing the probability of an event not occurring. With probabilities ranging from 0 to 1 for the outcome, logistic regression uses the log odds of the event as the liner function with the following equation with  $P$  representing the number of predictors in the model (Kuhn & Johnson, 2013):

$$\log (p/(1-p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_P x_P$$

Logistic regression requires no turning parameters and allows for a simplistic prediction equation (Kuhn & Johnson, 2013). This type of model behaves like linear regression with slope parameters for each predictor and an overall intercept (Kuhn & Johnson, 2013). The outputs of logistic regression are similar to linear regression, with both models producing coefficients, standard errors, z-statistics, and p-values (James et al., 2013). The difference between linear and logistic regression is the classification of the dependent variable, with linear regression being continuous and logistic regression being

dichotomous. Logistic regression's coefficients allow for calculating the probabilities and predict the behavior based on the dichotomous classifier (James et al., 2013).

Logistic regression's independent variables are on the continuous or nominal scale with nominal scale variables converted to dummy variables. All observations are independent of each other, including the dependent and nominal independent variables being mutually exclusive (Laerd Statistics, 2015a).

### Interpretation of Binary Classifier Models

For binary classification, a confusion matrix can show the model's accuracy and predict the probabilities of events occurring based on the model (Attewell & Monaghan, 2015; James et al., 2013). The four groups are categorized based on the outcomes: the actual event values based on the positive and negative outcomes and the predicted event values based on the positive and negative outcomes (Table 1) (Knowles, 2015).

Table 1

#### *Confusion Matrix Retention Example*

		Actual Event	
		Non-Retained (+)	Retained (-)
Predicted Event	Non-Retained (+)	True Positive n=75	False Negative n=50
	Retained (-)	False Positive n=125	True Negative n=275

The precision value of the model is a percentage of the true positives (n=75) divided by true positives and false negatives (n = 125) (Attewell & Monaghan, 2015; Knowles, 2015). The 60% represents the percent of predicted non-retained students not retained. The model's error rate is the false negatives (n = 50), and false positives (n = 125) divided by all of the outcomes (n = 525) and represents the percentage of misidentified counts (Attewell & Monaghan, 2015). The 33.3% represents the percentage

of incorrectly identified students as non-retained or retained in this model. The sensitivity of the model is the true positives ( $n = 75$ ) divided by all of the true and false positives ( $n = 200$ ) (Attewell & Monaghan, 2015; Knowles, 2015). This percentage, 37.5%, indicates the actual non-retained students identified correctly. The specificity of the model is the true negatives ( $n = 275$ ) divided by all of the true and false negatives ( $n = 325$ ) (Attewell & Monaghan, 2015; Knowles, 2015). This percentage, 84.6%, represents the actual retained student identified correctly. The false alarm rate of the model is the percentage of students predicted as non-retained but remained at the institution divided by all of the negatives (Attewell & Monaghan, 2015; Knowles, 2015). In this example, the false alarm is 15.4 % or  $1 - \text{specificity}$ . The accuracy of the model is a measurement of how well the model identified the true positives ( $n = 75$ ) and true negatives ( $n = 275$ ) divided by all of the cases ( $n = 525$ ). The model could predict who was retained and non-retained 66.7% of the time with slightly better odds than chance.

### **Evaluation Metrics for Comparing Classifier Models**

There is no set standard for comparing classification models in data mining (Demšar, 2006). Comparisons of classifiers for their accuracy allow researchers to understand their differences and similarities (Attewell & Monaghan, 2015). For binary classifier models, a comparison of the probabilities from the confusion matrixes (accuracy, sensitivity, and specificity) identifies which models performed better on these metrics (Attewell & Monaghan, 2015). The evaluation metrics should favor the minority class more than the majority class, especially in imbalanced class scenarios (Hossin & Sulaiman, 2015).

**Accuracy, Sensitivity, and Specificity.** The accuracy measurement allows for overall model comparison with other methods and can include the accuracy of the training and test data sets to interpret model predictive ability (Attewell & Monaghan, 2015). Accuracy may not be the exact measurement of a specific model since false negatives can have higher consequences than correctly predictive data. The comparison of sensitivity and specificity helps identify performance-based true positives and negatives for the predictability of the models. Medical diagnosis tests use specificity, sensitivity, and accuracy to quantify the results and reliability of the tests (Zhu, Zeng, & Wang, 2010).

**F1-Scores.** F1-scores are a measurement of the “harmonic mean of precision and recall and give a better measure of the incorrectly classified cases than the accuracy metric” (Huigol, 2019). The formula for F1-scores use precision and recall values (also called sensitivity) derived from the confusion matrix (Huigol, 2019):

$$F1\text{-score} = \left( \frac{Recall^{-1} + Precision^{-1}}{2} \right)^{-1} \text{ or } 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

F1-scores are a better measurement than accuracy when false positive and negative events are crucial and imbalanced classes exist. (Huigol, 2019; Zhang, Wang, and Zhao, 2015).

**Receiver Operating Characteristic (ROC) Curves.** Using the confusion matrix information, receiver operating characteristic (ROC) curves provide a visual way to compare said models (Attewell & Monaghan, 2015). Their design is to “determine an effective threshold such that values above the threshold are indicative of a specific event (pg. 262) (Kuhn & Johnson, 2013). The ROC curve evaluates class probabilities for the model across multiple thresholds and plots the specificity (the rate of the true positives)

on the Y-axis and 1 - specificity (the rate of false alarms) on the X-axis (Attewell & Monaghan, 2015; Knowles, 2015; Kuhn & Johnson, 2013). This visual representation shows the benefit and cost of the model's classification by the percentage of correctly classified observations and false alarm rates (Attewell & Monaghan, 2015; Knowles, 2015). Using the previous table example, the ROC curve of that data would show the correctly identified non-retained students along the y-axis compared to the incorrectly identified students as non-retained when they were retained (Knowles, 2015). With higher education retention, the false negatives (1-sensitivity) are also significant since these students misidentified as being retained when they were not retained (Knowles, 2015). Students who were misidentified as non-retained but retained do not carry the same consequences as the false negative students since they remain at the institution (Knowles, 2015).

ROC curves that closely follow the Y-axis and curve parallel to the X-axis are ideal. The area under the ROC curve (ROC\_AUC) helps provide a numeric value for the model's accuracy. Models with ROC\_AUC values closer to 1 have a higher level of accuracy for predicting the correct outcomes, while models with ROC\_AUC values near .5 are not accurate in their predictive abilities (Attewell & Monaghan, 2015). The .5 value is seen as a 45-degree line on the ROC curve and is no better than chance (Attewell & Monaghan, 2015). One benefit of ROC curves is their ability to handle data with low and high rates of occurrence for the positive or negative predicted or actual events (Hastie, Tibshirani & Friedman, 2009). The visual representation can help researchers understand their relationship to one another. For comparison of different classification models, multiple ROC curves can be imprinted on one ROC model and show the overall



relationship to one another (Bowers, Sprott & Taff, 2012). The ROC\_AUC values allow for the comparison of different classifiers.

### **Validation of Evaluation Metrics**

Inferential statistical tests on these evaluation metrics provide reliable measurements to determine which models have fit the data best. Demšar (2006) recommended using a nonparametric test such as the Wilcoxon signed-rank test for two classifiers and the Friedman test with post ad hoc analysis for more than two classifiers. These nonparametric tests help reassure the validity of the results and handle data that does not follow a normal distribution or homogeneity of the variances by comparing the means of the groups to find a statistically significant difference between them.

### **Summary**

Student retention has remained a significant problem in higher education, including at the community college level. Many different models have been created to understand higher education problems but often don't focus on community colleges (Astin et al., 1975; Bean, 1982; Tinto, 1975; Tinto, 1999). Community college student populations often differ from other higher education institutions and need specific research to understand the needs of these students. Bean and Metzner (1985) developed the nontraditional undergraduate student attrition model built on earlier models but focused on the nontraditional student population whose characteristics aligned closer to community colleges. They also recommended specialized institution-based retention models since nontraditional student populations can differ between institutions (Bean & Metzner, 1985). Sector-based retention models, which combine student population data

from similar schools within a sector or area, may also identify relationships not detected in individual institutions and better understand sector trends.

Within the retention models, variable selection is critical for understanding the populations being studied. Using Bean and Metzner's (1985) model, this study will focus on community college students' academic, background, and financial factors to predict retention. The student background characteristics are gathered when enrollment occurred and include high school GPA, age, race or ethnicity, and gender (Johnson et al., 2014). Academic performance can serve as an indicator of future performance in retention, with Bean and Metzner (1985) believing that community college students enroll in college for purely academic reasons (Pascarella & Terenzini, 2005). The academic variables for this study are college GPA, percentage of courses taken in an online format, number of remedial classes taken, and the number of credits earned during the first academic year. Bean and Metzner's (1985) model also included finances as a component in students' ability to be retained in higher education. With the average adjusted public college tuition increasing by 270% since 1973, many students struggle to afford their education and may drop out before completing their degrees (Mitchell & Leachman, 2015). The financial variables selected for this study are FAFSA application completion, the amount of financial aid awarded, and the amount of financial assistance paid to the student during the first academic year.

The models created in this study rely on educational data mining techniques that focus on computer-driven methods to identify patterns in data sets (Attewell & Monaghan, 2015; Bharati & Ramageri, 2010). Data mining techniques have identified new models in retention that allow for precise interventions for specific student

populations (Chacon, Spicer & Valbuena, 2012; Herzog, 2006; Luan, 2002, Lin 2012; Yu, DiGangi, Jannasch-Pennell & Kaprolet, 2010). One type of data mining modeling, classifiers, predicts which classes the dependent variables' cases belong to and is derived from different algorithms (Attewell & Monaghan, 2015; Bharati & Ramageri, 2010; Breiman, 1999; Breiman et al., 1984; Han et al., 2011). The goal of the models is to predict retention through the classification of students who are retained or not retained after the first academic year. Wolpert's No Free Lunch Theorem indicated that no one classifier could handle all data sets and established the need for numerous models to validate the data in the study (Wolpert, 1996). Three of the classification models chosen for the study (random forests, support vector machine (SVM), and neural networks) have deemed the most accurate results among the 121 different data sets (Fernández-Delgado et al., 2014). The final classification model, logistic regression, is easier to interpret and is included in the study to compare all the equations' accuracy and interpretation (Attewell & Monaghan, 2015; Kuhn & Johnson, 2013).

The overall models' validation relies on cross validation techniques which divide the data set into training sets used to build the models, and a testing set is used only once to assess the models' performance (Attewell & Monaghan, 2015; Bost et al., 2015; Efron, 1983; Han et al., 2011; Kuhn & Johnson, 2013). Creating a confusion matrix for each of the models can show the accuracy of the individual model and predict the probabilities of events occurring (Attewell & Monaghan, 2015; James et al., 2013). Comparing the probabilities through inferential statistical tests from the individual confusion matrixes (accuracy, sensitivity, specificity, F1-scores, and ROC curves) identify which models performed better on these metrics (Attewell & Monaghan, 2015).

## **Chapter III**

### **METHODOLOGY**

The purpose of this chapter was to present the research methods used in this study. Using Bean and Metzner's model as a framework, this study identified significant variables that may have influenced if a student stays or drops out of community college and which classification model was most accurate in identifying these variables. This chapter is divided into five sections, starting with the study's research design, including the independent and dependent variables. The second section discussed the participants for this study. The third section focused on the instrumentation of the research and focuses on the accuracy of the data. The fourth section dealt with the collection of the data. The fifth section was the descriptions of the data analysis procedures broken down by the research question. The final section provided a summary of this chapter.

#### **Research Design**

This nonexperimental, correlational study used archival data from the University System of Georgia. The correlation aspect of this study focused on the relationship of the variables to the retention of first year students. This study was also considered a classification study since the goal was to determine which variables are significant in predicting which students remained or left school after their first year (Mills & Gay, 2019). The research design was divided into two different components, with the first part building four classification models, which identified if any of these models identified factors that predict retention. The second part of the study compared the classification

models using evaluation metrics to determine which model(s) produce the most accurate results.

This study's 12 independent variables were defined into three distinct groups, with the introduction of multiple independent variables allowed for a higher degree of accuracy than single variables. The background variables were age, gender, race or ethnicity, and high school GPA. The variable of age was a ratio measurement for length of time (in years) from their birthdate to the first day of their first semester of attendance at their institution. Gender, a nominal measurement, was dummy coded with 0 for males and 1 for females. The race or ethnicity variable, a nominal measurement, was dummy-coded using six exhaustive and mutually exclusive dichotomous variables for White, American Indian or Alaska Native, Asian, Black, Other/Unknown/Multiple, and Hispanic. The method used to make the dummy variables were the one-hot encoding function in R. The high school GPA was measured as an interval variable with two decimal places. The academic variables were college GPA, percentage of courses taking in an online format, remedial classes, and the number of credits earned and are calculated using three consecutive semesters. College GPA's final value was a weighted mean comprised of each semester's course hours and final letter grades, which produced an interval variable. The percentage of online courses' calculations, a ratio variable, was a percentage consisting of all three semesters total course hours and total online course hours. The remedial classes were ratio measurements for the number of mathematics, reading, and writing required remedial courses taken. The number of credit hours was also a ratio measurement scale that represented the sum of the hours completed. The financial variables were FAFSA completion, the date of FAFSA completion, the amount

of financial aid awarded, and the amount of financial assistance paid to the student during the first academic year and are calculated using three consecutive semesters. FAFSA completion was a dichotomous, nominal value dummy coded with 0 for completion and 1 for non-completion. If no FAFSA were submitted, the variable was left blank. The amount of financial aid awarded to the students was a ratio variable rounded to the nearest dollar amount. The amount of financial assistance paid to the students was a ratio variable rounded to the nearest dollar amount.

Retention status, the dependent variable, was determined by enrollment after the drop/add period for three consecutive semesters after initial attendance. This variable was defined using the first four consecutive semesters for each student. Students enrolled after the initial three following three-semester periods were labeled retained and labeled as 1. Students not registered after the initial three consecutive three-semester periods were considered nonretained and marked with a 0.

### **Participants**

The target population for this study was community college students attending public institutions in Georgia beginning with their freshman year. In Fall 2017, 5.8 million students attended two-year public institutions in the United States, with women (56%) attend these schools in higher numbers than men (44%) (AACC, 2018; Ginder, Kelly-Reid, & Mann, 2018). The average of these students was 28 years old, and the median age was 24 (AACC, 2018). Students' enrollment status identified 37% of the students attended full time (12 hours or more) and 63% as part-time (under 12 hours) (AACC, 2018). The demographics for the community college student population was 47% identified as white, 24% as Hispanic, 13% as Black, 6% as Asian or Pacific

Islander, and 10% as Native American, two or more races, other, and nonresident alien (AACC, 2018).

The accessible population for this study was students who attend seven community colleges in Georgia. The participants of this research were past students who attended their respected colleges from the academic years of Fall 2017 through Fall 2019 without dual enrollment or transfer status. Historically, the freshmen year had the most significant decrease in retention for community college students (Wyman, 1997). These seven institutions awarded associate and bachelor's degrees and resided in the state college sector for Georgia. The combined undergraduate student enrollment of the seven institutions for the 2017-2018 academic year was 26,122 students (U.S. Department of Education, n.d.). The average percentage of female attendance (64%) was higher than male attendance (36%) (U.S. Department of Education, n.d.). IPEDS reporting divided age into two distinct categories, 24 and under and 25 and older (U.S. Department of Education, n.d.). These seven institutions' students' age were predominantly 24 and under (81%), with the minority of students labeled as nontraditional (19%) (Ginder, Kelly-Reid, & Mann, 2018). More than half of these students (58%) classified as full-time students (12 credit hours or more) (U.S. Department of Education, n.d.). The ethnicity and racial classification were diverse for these seven schools (U.S. Department of Education, n.d.). The number of students classified as White ranged from 2% to 66%, 4% to 90% as Black, and 4% to 28% as Hispanic, 1% to 2% as Asian, 1% and less for Native American or Alaska Native, and 2% to 7% as Other/Unknown/Multiple (U.S. Department of Education, n.d.). The percentage of undergraduates who received the Pell grant ranged from 42% to 68% at these institutions, with the percentage of grant and scholarship aid

ranging from 73% to 86% (U.S. Department of Education, n.d.). The percentage of students who received federal aid assistance ranges from 21% to 54% (U.S. Department of Education, n.d.).

The subjects in this study were the total population of first time, first year students for a consecutive four-semester period in seven community colleges. The theoretical sample for this study needed at least 473 students. The determination of the minimum sample size was determined using the priori method in the G\*Power calculator with a power level of .80, an effect size of .50, and a significance level of .05. The minimum population size exceeded the ideal sample size since the estimated population was 6297 students for one academic year, excluding one college whose first-time freshmen rates were not published. By increasing the period to two years, the population size was over 10,000 students and allowed for enough data for all four modeling techniques. The population size was large enough for all four models with students who attended schools in the state college sector and shared similar characteristics. The different schools in this sector were located through Georgia and helped capture these students' diverse demographics.

### **Instrumentation**

The study focused on student background, academic, and financial factors for student retention. The entire archival data was retrieved directly from the USG to maintain the consistency of the data among the seven institutions. Data accuracy was a critical component of the study. The Research and Policy Analysis office at the Board of Regents for the University System of Georgia (USG) handled all USG data. This office collected each institution's data six times a year and pulled the data elements in the USG



data warehouse. The collected data was formatted into variables that allowed for accuracy and consistency for the seven institutions' data. The requested data set included all first time first year students with the exclusion of other populations that attended the college during those years. The data only contained the required data without identifiable components to allow for students' anonymity. Misrepresented data were avoided through discussions with the Research and Policy Analysis at the Board of Regents for the University System of Georgia to ensure the selected variables were appropriate. The overall process of the study purposely minimized threats to the accuracy of the data.

### **Data Collection**

Once the Institutional Review Board (IRB) granted permission, a data request was made to the Research and Policy Analysis at the Board of Regents for the University System of Georgia (Appendices C and D). This request focused on all seven institutions' data and included two academic years for all first-time freshmen. The data had identifiable information removed, and informed consent was not required. Each variable's required data was identified and requested in the necessary documentation. The data request asked for unmanipulated data for these students in a single Microsoft Excel file with student information were linked together by a newly created student identification number. The data was encrypted and stored in a password protected file with multiple backups produced for replication. The two research questions used the same data set, which contained the information for first year students and contained no unique identification, which allowed for students' anonymity.

### **Data Analysis**

Data analysis occurred in two separate parts based on each research question and

used the current version of R, statistical software, and the various packages within the tidyverse collection (Appendices A and B, Korkmaz, Goksuluk & Zararsiz, 2014; Kuhn et al., 2019). These packages were an evolving collection of different techniques for modeling functions (Kuhn, 2008). The study design for research question 1 used the dataset in two phases: data preparation and inferential statistics.

Data preparation focused on transforming the data to reduce the impact of outliers and skewness, improving the classification models' performances (Kuhn & Johnson, 2019). The first step in this process was to create new variables from existing data, as previously described. New quantitative variables were made for college GPA, percentage of courses taking in an online format, number of remedial classes, number of credits earned, the date of FAFSA completion, amount of financial aid awarded, and amount of financial assistance paid to the student. Dummy variables for gender, race or ethnicity, FAFSA completion, and retention status were created after descriptive statistics using the recipes package. Descriptive statistics (sample size, mean, standard deviation, median, skewness, and kurtosis) were calculated for each predictor variable using the skimr package that showed the overall summarization.

The next step was the identification of missing data using the dplyr, ggplot2, and complex heat map packages, which can occur as singular events or as a subset of the predictors (Kuhn & Johnson, 2013). Students who drop out after one semester had missing data for academic and financial factors. These missing values were coded as 0 to allow for their inclusion in the descriptive statistics and modeling and represented that the students didn't return. Students who don't attend for one semester but come back had these 0 values as their values and had additional values that showed their academic and

financial career in the four semesters.

Row and percentage plots displayed the missing data by predictors and compared the entire data set (Kuhn & Johnson, 2019). In addition, a co-occurrence plot showed the frequency of missing predictor combinations using the complex heat map package (Kuhn & Johnson, 2019). Background variables that have missing data were treated as genuinely missing data and excluded from the sample since assumptions were not made on these variables. Other variables used the `bagImpute` function from the `caret` package to impute new similar values in place of missing values (Kuhn, 2008).

Individual histograms and Q-Q plots helped identify characteristics of the discrete and continuous variables using the `MVN` package (Korkmaz, Goksuluk & Zararsiz, 2014; Kuhn & Johnson, 2019). Both types of graphs identified skewed data for predictors that could affect models like logistic regression (Korkmaz, Goksuluk & Zararsiz, 2014; Kuhn & Johnson, 2019). A simple transformation of the predictors, such as Box-Cox or logarithm function, changed the skewed data into symmetric distribution and possibly removed the appearance of outliers in individual variables (Kuhn & Johnson, 2019). Outlier determination for these variables occurred by univariate analysis within the `MVN` package that used the following tests: Cramer-von Mises test, Lilliefors test, and Anderson-Darling test (Korkmaz, Goksuluk & Zararsiz, 2014). A value flagged as an outlier was investigated for data entry errors and validity as a value for that variable. The categorical variables were plotted with bar charts of each individual variable and their retention outcome (Kuhn & Johnson, 2019).

The `MVN` package also determined multivariate normality through the following tests: Mardia's test, Henze-Zirkler's test, Royston's test, and Doornik-Hansen's test

(Korkmaz, Goksuluk & Zararsiz, 2014). Multivariate outliers and influence points were displayed and evaluated using the Cook's distance (Korkmaz, Goksuluk & Zararsiz, 2014). Outliers were handled based on an individual level since they are actual occurrences for our sample and may show an unknown pattern that exists. Transformations were implemented using the recipes package to handle the identified outliers.

The second part of research question 1 was the cross validation step using the institutional combined dataset, which created the initial training and test data sets (Kuhn & Johnson, 2019). The group assignment was done randomly with the R-software with a 70% split of data in the training data sets and the remaining 30% in the test data sets, which used the `initial_split` function. The initial training data set was split into ten different sample sets for the data mining models using the 10-fold cross validation method. (Attewell & Monaghan, 2015; Kuhn & Johnson, 2013; Kuhn & Johnson, 2019). The `set.seed` function allowed similar results in the datasets using R's random number generator (James et al., 2013). Within the tidyverse package of "R," the `tidy_kfolds` function used the number of "10" to signify ten iterations of this method and the strata of "retention" for the dependent variable (Kuhn et al., 2019). The use of numerous sample data sets helped measure the variability and differences in the models (Kuhn & Johnson, 2019).

### **Inferential Statistics**

The four types of models (random forest, supported vector machines, neural networks, and logistic regression) in this study allowed for different methods to explore the data. A random forest model is a form of decision trees that classified the data, which

started with an individual tree at the root and moved throughout the tree down to the leaves. A group of decision trees formed the forest based on classifiers to build one robust classifier model (Attewell & Monaghan, 2015; Han et al., 2011; Mao & Wang, 2012). The SVM behaved similarly to other classifiers by separating data into groups and created margin creations on both sides of the hyperplane to help standardized the multiple boundaries into one optimal solution. (Attewell & Monaghan, 2015). Neural networks were classification models that mimic the operations of biological neurons where the neural formula started at a random point and going through each observation until it learned from its predictive error and tuned the parameters (Attewell & Monaghan, 2015). Logistic regression was similar to linear regression, where the overall outcome of the model had two distinct events using the binomial distribution and produced slope parameters for each predictor and an overall intercept (Kuhn & Johnson, 2013).

**Random Forest.** The random forest analysis used the method, `rand_forest`, in the `parSNIP` package (James et al., 2013; Kuhn et al., 2019; Kuhn & Johnson, 2013). The model constructed using the default setting in the package. One parameter that needed to be adjusted in the random forest was the `mtry` function, which determined the number of predictors used at each split in the model. The recommended value for this parameter was one-third of the predictors, which was roughly four for this study (James et al., 2013; Kuhn et al., 2019; Kuhn & Johnson, 2013). The model was rerun with the `mtry` set to blank to test for the optimal number of splits and to confirm if four was appropriate for this model. At the same time, the model was trying to find the optimal min nodes, which was the minimum number of terminal nodes in the forest. The highest `ROC_AUC` score determined the ideal min nodes. The optimal number of trees was found using the tree

function. The random forest model created a confusion matrix, accuracy value, sensitivity value, specificity value, ROC\_AUC value, F1 scores, and ROC curves.

**Supported Vector Machine (SVM).** For support vector machines (SVM), the models used two different kernels: `svm_rbf`, for the nonlinear kernel, and `svm_poly` for the polynomial kernel since Attewell and Monaghan (2015) recommended trial and error of several kernels to discover the optimal kernel. The radial and poly functions were in the `parSNIP` package. The tuning parameter for this model was the cost, which was adjusted to increase the model's accuracy (James et al., 2013). Multiple cost values were run to find the optimal value for the two models (James et al., 2013). The models' output contained the optimal tuning parameter, `sigma`, by discovering the optimal cost value. The SVM produced a confusion matrix, accuracy value, sensitivity value, specificity value, ROC\_AUC value, F1 scores, and ROC curves.

**Neural Network.** The neural networks model used the method, `mlp`, in the `nnet` package, which allowed for the creation of hidden units and penalties (Kuhn et al., 2019). The other parameters in this method remained as the "R" defaults (Ripley & Venables, 2016). The optimal model was determined by the ROC\_AUC measurement for the model's hidden units, penalty, and epochs. The neural network model produced a confusion matrix, accuracy value, sensitivity value, specificity value, ROC\_AUC value, F1 scores, and ROC curves.

**Logistic Regression.** The logistic regression model was built using the `logistic_reg` function in the `parSNIP` package, allowing the model to be created with a binary variable for student retention (James et al., 2013). This logistic regression modeled the probability of a student not being retained:

Pr (default = NonRetained |balanced)

And allowed for the probability to fall between zero and one with a nonretained assigned value of 0 (James et al., 2013). The investigation of multicollinearity and singularity measured variable relationships, and linear relationships review occurred through the tolerance measurements and variance inflation factors (VIF). The relationship between the continuous independent variables and the dependent variable needed to be linear. The last assumption was checking for outliers, high leverage points, and influence points identified and were handled on a case-by-case basis. The logistic regression model produced a confusion matrix, beta values, standard errors, odds ratios,  $\chi^2$  values, degrees of freedom, accuracy value, sensitivity value, specificity value, ROC\_AUC value, F1 scores, and ROC curves. The logistic regression model results showed the statistically significant variables by p-values and variable importance plots. The comparison of the five models identified the statistically significant factors in predicting first-year student retention for community college students using the results of the varImpPlot function from each model.

The second research question compared the data mining models (random forests, support vector machines, neural networks, or logistic regression) and identified if one of the models generated a more accurate classifier performance based on the confusion matrix. Each model's evaluation metrics were created to show how many test observations were classified correctly or incorrectly (Attewell & Monaghan, 2015; James et al., 2013). Additionally, the models were analyzed using overlaid ROC curves that allowed for a visual comparison of the models.

When comparing the accuracy, sensitivity, specificity, ROC\_AUC, and F1 scores for all the models, their differences were measured using statistical methods (Hothorn, Leisch, Zeileis, & Hornik, 2005; Kuhn & Johnson; 2013; Kuhn & Johnson; 2019). The first analytical method was a Mann-Whitney U test to determine if there was any difference in the models from the training and test data sets and reassured the validity of the results more than visual comparison. The next analytical method was Friedman's test that compared the models to each other using the evaluation metrics from all ten testing data sets and determined if the classifiers are significantly different. Finally, the Wilcoxon signed-rank test served as the ad hoc test for the pairwise comparison of the evaluation metrics from all ten testing data sets and ranked the individual classifiers. (Demšar, 2006, Fernández-Delgado et al., 2014). The results from the Friedman's and Wilcoxon signed-rank tests answered the second question and allowed for the identification of the most accurate classifier for predicting first year student retention for the state college students.

### **Summary**

This nonexperimental, correlational study used two academic years of data for all first-time freshmen students from seven community colleges. The data for this population of students contained the 12 independent variables (background, academic, and financial) and one dependent variable (retention outcome) used for the model creation. The software for the study's analysis was "R" with the caret and tidyverse packages.

The first research question focused on developing the five classifier models and determining if any of the 12 predictors are significant in predicting retention. Before the models' creation, data preparation transformed the data to lessen the impact of outliers,



skewness, and missing data at the individual and multivariate levels. The 10-fold cross validation method created ten different training and test datasets with the training sets used to develop the four types of models (random forest, supported vector machines, neural networks, and logistic regression). The testing sets validated the new models and provided the evaluation metrics and significant predictors in first year student retention for community college students.

The second research question used the evaluation metrics (accuracy, sensitivity, specificity, ROC\_AUC, and F1 scores) from the models' results, and additional statistical tests were run to determine which classifier models produced the most accurate results on various evaluation metrics. A stacked ROC curve allowed for a visual comparison of the models. The significant predictors of each model were compared to see if they were similar and strengthened the final model accuracy.

## **Chapter IV**

### **RESULTS**

The purpose of this chapter is to identify which background, academic, and financial factors were significant for student retention and identify if a specific data mining model produced more accurate results based on certain evaluation metrics. Two different cohorts of freshmen, fall 2017 and 2018, from seven community colleges were used to create the different data mining models. These models will help identify what factors are important for retaining community college freshmen and identify the optimal data mining model for this type of analysis. The analysis of this project focused on answering the following questions:

- 1: Are background factors (age, gender, race or ethnicity, and high school GPA), academic factors (college GPA, percentage of courses taken in an online format, number of remedial courses taken, and the number of credits earned during the first academic year), and financial factors (FAFSA completion, amount of financial aid awarded, and amount of financial aid paid to the student during the first academic year) significant in predicting first-year student retention for community college students?
- 2: Does one of the data mining models (random forests, support vector machines, neural networks, or logistic regression) generate a more accurate classifier performance overall based on the evaluation metrics of accuracy, sensitivity, specificity, area under the curve (ROC\_AUC) and f-measure ( $F_1$ ) scores?

This chapter shows the data analysis process with results for both research questions. The first part of this chapter will focus on the demographic characteristics, descriptive statistics, and Pearson correlation coefficients to compare both cohorts. The following section combines both cohorts' data and analyzes the categorical variables and missing data. The cross validation methods used to create the models are explored before identifying outliers and the overall normality of the combined cohorts. The following section describes the outlier capping, transformation, and normalization of combined cohort data before model building. The final sections address the research questions to identify which academic, background, and financial predictors are significant in predicting first-year student retention for community college students. The section will also determine if one of the data mining models (random forests, support vector machines, neural networks, or logistic regression) generates a more accurate classifier performance overall based on the evaluation metrics.

### **Demographic Characteristics for Individual Cohorts**

The University System of Georgia (USG) provided the community college student data with the Data Governance Committee's approval. The student data focused on two different cohorts of first-time freshmen who first attended Fall 2017 and Fall 2018, including their first four consecutive semesters of data. The demographic characteristics for both cohorts are displayed in Table 2. First-time freshmen were identified for each cohort with 6,834 (51.44%) students in the Fall 2017 cohort with 6,452 (48.56%) students in the Fall 2018 cohort to create a total of 13,286 students used in the data analysis. Females outnumbered male students in the Fall 2017 cohort (59.4% vs. 40.6) and Fall 2018 cohort (60.1% vs. 39.9%). Both cohorts had similar results for

Table 2

*Demographic Characteristics for Students in Both Cohorts*

Demographic Characteristics	Fall 2018		Fall 2019	
	N	%	N	%
<b>Gender</b>				
Female	4,062	59.4	3,878	60.1
Male	2,772	40.6	2,574	39.9
<b>Race or Ethnicity</b>				
American Indian or Alaska Native	15	0.2	19	0.3
Asian	78	1.1	76	1.2
Black or African American	2,296	33.6	2,258	35.0
Hispanic or Latino	827	12.1	804	12.5
Native Hawaiian or Other Pacific Islander	13	0.2	5	0.1
Unknown	72	1.1	53	0.8
Two or More Races	195	2.9	205	3.2
White	3,338	48.8	3,032	47.0
<b>Retention Status</b>				
Retained	3,558	52.1	3,446	53.4
Non retained	3,276	47.9	3,006	46.6
<b>FASFA Status</b>				
FASFA completed	6,308	92.3	5,901	91.5
FASFA not completed	526	7.7	551	8.5

race or ethnicity; White, with 48.8% for the Fall 2017 cohort and 47.0% for the Fall 2018 cohort. Students who identified as Black represented 33.6% of the Fall 2017 cohort and 35.0% of the Fall 2018 cohort. Hispanic or Latino students accounted for 12.1% of the Fall 2017 cohort and 12.5% of the Fall 2018 cohort. Students who identified as two or more races represented 2.9% of the Fall 2017 cohort and 3.2% of the Fall 2018 cohort. American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander and unknown race or ethnicity students represented 2.6% of the Fall 2017 cohort and 2.4% of the Fall 2018 cohort. The retained (52.1% and 53.4%) and nonretained (47.9% and

46.6%) students' rates were consistent for Fall 2017 and 2018 cohorts. The FASFA completion (92.3% and 91.5%) and noncompletion (7.7% and 8.5%) students' rates were similar in both cohorts.

### Descriptive Statistics for Students

Table 3 illustrates descriptive statistics for the continuous predictors for the Fall 2017 cohort. The average age of freshmen in this cohort was 18.71 ( $SD = 2.89$ ) and had an average high school GPA of 2.97 ( $SD = 0.52$ ). The average number of hours freshmen took was 18.45 ( $SD = 10.60$ ), with an average first-year GPA of 2.21 ( $SD = 1.15$ ). The average number of remedial courses taken by freshmen in this cohort was 0.73 ( $SD = 1.09$ ), and the average percentage of online courses taken was 10.53 ( $SD = 18.65$ ). The average amount of financial aid paid to the Fall 2017 cohort's freshmen was \$6,534.10

Table 3

#### *Descriptive Statistics for Fall 2017 Cohort before Data Transformations*

Variable	Min	Max	<i>Mdn</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
age	15.00	70.00	18.00	18.71	2.89	7.95	87.09
hsgpa	1.00	4.00	2.94	2.97	0.52	0.16	-0.78
credit	0.00	52.00	20.00	18.45	10.60	-0.21	-0.97
remed	0.00	6.00	0.00	0.73	1.09	1.35	0.85
online	0.00	100.00	0.00	10.53	18.65	2.54	7.41
gpa	0.00	4.00	2.36	2.21	1.15	-0.36	-0.90
paid	0.00	26210.00	5682.00	6534.10	4461.00	0.70	-0.10
award	0.00	27190.00	8887.50	9276.80	4363.80	0.35	-0.35
percaid	0.00	300.23	68.27	74.42	30.33	-0.56	-0.20

*Note.*  $n = 6,834$ . age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = college GPA for the first three semesters. percaid = percentage of financial aid paid divided by the amount of financial aid awarded paid to the student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters.

( $SD = \$4,461.00$ ), and the average amount of financial aid awarded was  $\$9,276.80$  ( $SD = \$4,363.80$ ). The variable, *percaid*, is the overall percentage of financial aid used and is created by the amount of financial aid paid divided by the amount of financial aid awarded. For the Fall 2017 cohort, the overall percentage of financial aid used was 74.42 ( $SD = 30.33$ ). Table 4 presents descriptive statistics for the continuous predictors for the Fall 2018 cohort. The average age of freshmen in this cohort was 18.79 ( $SD = 3.28$ ), and they had an average high school GPA of 2.96 ( $SD = 0.53$ ), which was higher than the Fall 2017 cohort. The average number of hours freshmen took was 18.05 ( $SD = 10.74$ ) with an average first-year GPA of 2.18 ( $SD = 1.17$ ), which was lower than the previous cohort. The average number of remedial courses taken by freshmen in this cohort was 0.92 ( $SD = 0.99$ ), and the average percentage of online courses taken was 14.27 ( $SD = 21.54$ ), with both measurements increasing from the Fall 2017 cohort. The average amount of

Table 4

*Descriptive Statistics for Fall 2018 Cohort before Data Transformations*

Variable	Min	Max	<i>Mdn</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
age	16.00	67.00	18.00	18.79	3.28	7.28	65.49
hsgpa	1.42	4.00	2.94	2.96	0.53	0.16	-0.86
credit	0.00	53.00	19.00	18.05	10.74	-0.13	-0.91
remed	0.00	5.00	1.00	0.92	0.99	0.82	0.01
online	0.00	100.00	0.00	14.27	21.54	2.11	4.79
gpa	0.00	4.00	2.36	2.18	1.17	-0.37	-0.95
paid	0.00	26969.00	6095.00	6788.40	4670.10	0.73	0.00
award	0.00	28124.00	9192.50	9690.70	4642.90	0.40	-0.23
percaid	0.00	100.01	68.41	76.27	30.07	-0.65	-0.71

*Note.*  $n = 6,452$ . age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = College GPA for the first three semesters. percaid = percentage of financial aid used by the student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters.

financial aid paid was \$6,788.40 ( $SD = \$4,670.10$ ). The average amount of financial aid awarded to the Fall 2018 cohort was \$9,690.70 ( $SD = \$4,642.90$ ), which increased from the previous cohort. For the Fall 2018 cohort, the overall percentage of financial aid used was 76.27 ( $SD = 30.07$ ) and was slightly higher than the Fall 2017 cohort.

### Correlation Coefficients for Students

An analysis was run to determine correlations between the quantitative variables using Pearson correlation coefficients, which range from -1, which indicates a strong negative correlation to +1, which means a strong positive correlation. A correlation matrix was run for each cohort's data. The correlation matrix for the Fall 2017 cohort variables is shown in Table 5. For the Fall 2017 cohort, there was a strong, positive correlation between financial aid awarded and paid,  $r(6832) = .81$ ,  $p < .001$ , which is not surprising since the amount paid is dependent on the amount awarded. GPA had a strong,

Table 5

#### *Pearson Correlation Coefficients for Fall 2017 Cohort*

Variable	1	2	3	4	5	6	7	8	9
1. age	1.00								
2. hsgpa	-.11**	1.00							
3. credit	-.08**	.40**	1.00						
4. remed	.08**	-.55**	-.28**	1.00					
5. online	.17**	.09**	.07**	-.13**	1.00				
6. gpa	.02	.50**	.78**	-.28**	.09**	1.00			
7. paid	-.02	-.07**	.28**	.17**	-.03*	.05**	1.00		
8. award	.00	.01	.22**	.10**	-.01	.06**	.81**	1.00	
9. percaid	-.01	-.14**	.17**	.14**	-.01	.00	.58**	.11**	1.00

*Note.*  $p < .001$  ‘\*\*\*’,  $p < .05$  ‘\*’. age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = college GPA for the first three semesters. percaid = percentage of financial aid used by the student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters.

positive correlation with credit hours,  $r(6832) = .78, p < .001$  and high school GPA,  $r(6832) = .50, p < .001$ . High school GPA had a moderately negative correlation with number of remedial courses taken,  $r(6832) = -.55, p < .001$ , and a moderate positive correlation with credit hours,  $r(6832) = .40, p < .001$ . The percentage of financial aid used had a moderately positive correlation with the amount of financial paid,  $r(6832) = .58, p < .001$ , and was not surprising since the amount of financial paid was in the overall calculation of the percentage of financial aid used.

The Fall 2018 cohort had similar variable correlations as the Fall 2017 cohort.

The correlation matrix for the Fall 2018 cohort variables is shown in Table 6. There was a strong, positive correlation between financial aid awarded and paid,  $r(6450) = .80, p < .001$ . GPA had a strong, positive correlation with credit hours,  $r(6450) = .80, p < .001$  and

Table 6

*Pearson Correlation Coefficients for Fall 2018 Cohort*

Variable	1	2	3	4	5	6	7	8	9
1. age	1.00								
2. hsgpa	-.14***	1.00							
3. credit	-.04***	.40***	1.00						
4. remed	.04**	-.48***	-.25***	1.00					
5. online	.20***	.05***	.05***	-.08***	1.00				
6. gpa	.04**	.52***	.80***	-.30***	.05***	1.00			
7. paid	.01	-.07***	.30***	.19***	.02	.08***	1.00		
8. award	.03*	.01	.23***	.10***	.02	.09***	.80***	1.00	
9. percaid	-.01	-.13***	.18***	.17***	.01	.00	.55***	.07***	1.00

*Note.*  $p < .001$  ‘\*\*\*’,  $p < .01$  ‘\*\*’,  $p < .05$  ‘\*’. age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = college GPA for the first three semesters. percaid = percentage of financial aid used by the student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters. Percaid = the overall calculation of the percentage of financial aid used.



high school GPA,  $r(6450) = .52, p < .001$ . High school GPA had a moderately negative correlation with number of remedial courses taken,  $r(6832) = -.48, p < .001$ , and a moderate positive correlation with credit hours,  $r(6450) = .40, p < .001$ . The percentage of financial aid used had a moderately positive correlation with the amount of financial paid,  $r(6832) = .55, p < .001$ , and again, was not surprising since the amount of financial. Both cohort datasets have similar demographic characteristics, descriptive statistics, and Pearson correlation coefficients. The data from both cohorts will be combined into one dataset for further analysis.

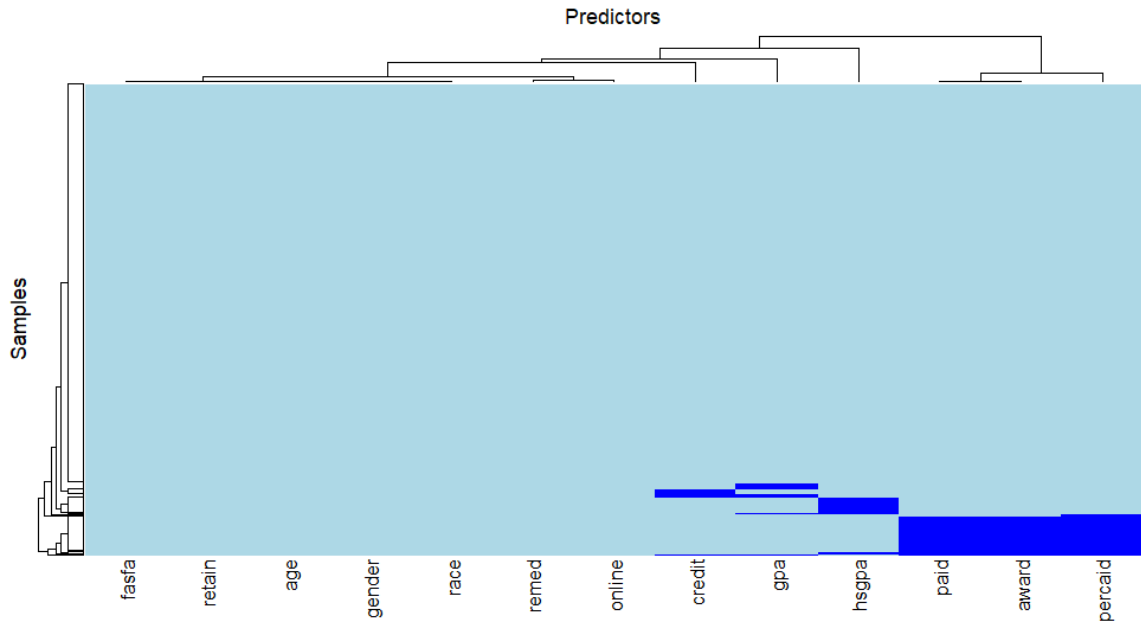
### **Categorical Variable Analysis of Combined Cohorts**

The categorical variables were plotted with stacked bar charts of each variable with retention status and confidence intervals plots by the retained students' proportion to show the differences within each categorical variable (Kuhn & Johnson, 2019). For the gender variable, the overall ratio of retention status was the same for male students. Female students (54% and 51%) were retained at a higher proportion than male students. Black or African American students were retained at lower rates (42%), while Hispanic or Latino students who had the third-highest populated category were retained in greater proportions, 67%, than all the races or ethnicities. The students in the Other category had retained status with a retention rate of around 55%, slightly less than White students. White students, who accounted for most of the students, were retained at a rate of approximately 57%. A comparison of retention status and FASFA completion showed that students who filed the FASFA (53% and 52%) had a slightly higher proportion of being retained. Comparing gender and FASFA completion showed that both genders filed the FASFA with females at 93% and males at 90%. For the race or ethnicity variable

compared to the FASFA completion, the majority of Black or African American students (98%), Hispanic or Latino students (88%), and students in the category of Other and White (90%) filed a FASFA. Dummy variables were created for gender, race or ethnicity, and FASFA status. The retention variable was coded as a factor with 0, indicating students who were not retained after the first academic year and 1, indicating students who were retained after the first academic year.

### **Missing Data Analysis of Combined Cohorts**

Before inferential statistics can be run, missing data values must be identified. A heatmap was created to display the missing data by predictors and compare the entire data set (Kuhn & Johnson, 2019). The heatmap (Figure 1) of missing data showed that the percaid variable had the highest amount of missing data, 8.85%. The second highest missing data variable was the financial amount of aid awarded and the financial amount of aid paid variables with the missing data of 8.40% of the missing data. GPA variables had the next highest amount of missing data with 3.99% for high school GPA and 2.80% for college GPA. The remaining variables were total credit hours with 2.24%, the number of remedial courses with 0.09%, and the percentage of online courses with 0.10% of its data missing. None of the demographic characteristics had missing data. A co-occurrence plot (Figure 2) identified the individual variables. The financial variables of percaid, the amount awarded, and the amount paid had the highest combination of missing data, 986. The combination of total credit hours and college GPA had 93 missing data values. The next set of combinations occurred with financial aid awarded and paid and other variables. The different varieties were smaller in number ranging from 35 to 1 pair. All the missing data in the dataset were replaced through the bagImpute function



*Figure 1.* Missing data heatmap. This heatmap displays the missing data by the individual predictors with the dark blue areas indicating the amount of missing data and the relation to the entire dataset.

from the caret package to impute new similar values in the recipe step (Kuhn, 2008).

### **Cross Validation Method**

Two distinct datasets were randomly created from the initial dataset with a 70% split of data in the training data sets ( $n = 9,301$ ) and the remaining 30% in the test data set ( $n = 3,985$ ) using the `initial_split` function in R. The initial training data set will be split into ten different sample sets for the data mining models using the 10-fold cross validation method (Attewell & Monaghan, 2015; Kuhn & Johnson, 2013; Kuhn & Johnson, 2019).

### **Outliers and Normality of Combined Cohorts**

Individual histograms and Q-Q plots were created to inspect the shape and visually identify outliers for each predictor in the training set except for the factors (gender, race or ethnicity, FASFA filing status, and retention status). All the numeric

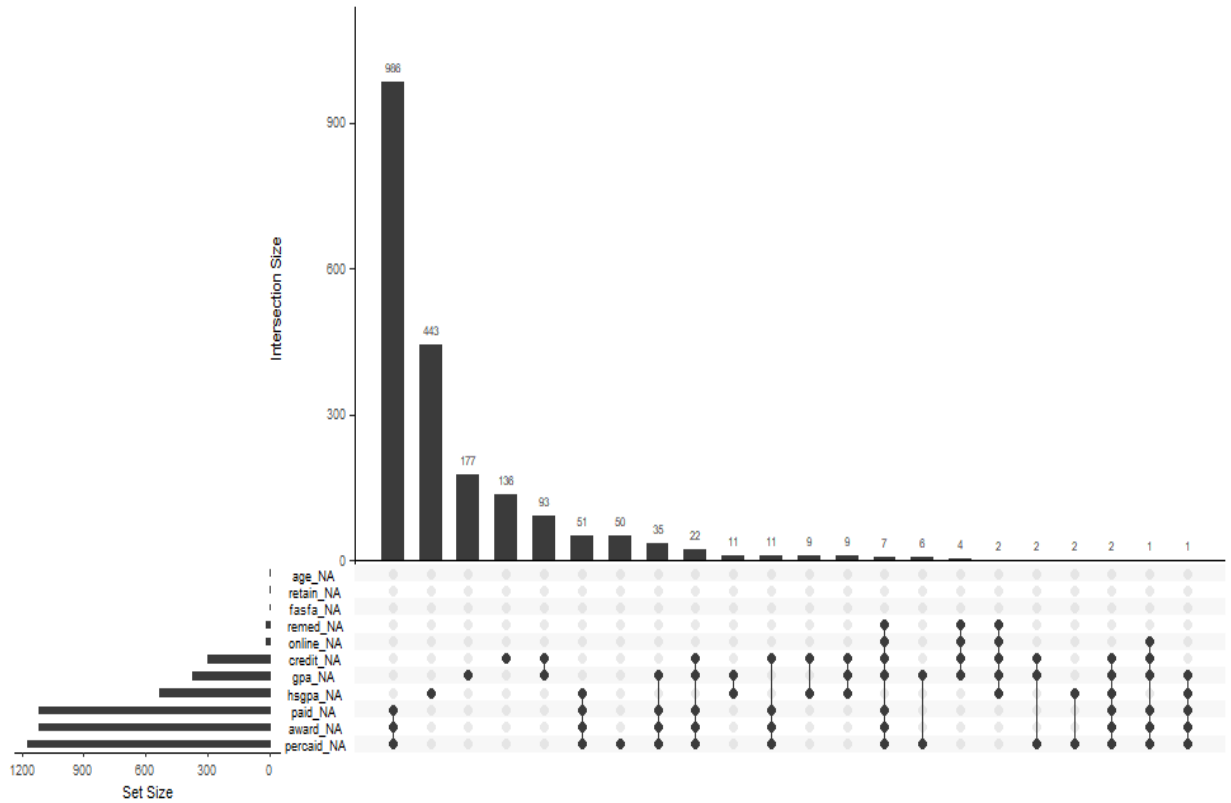


Figure 2. Missing data co-occurrence plot. This plot displays the missing data by the individual predictors with the black bars on the left and the number of predictor combinations of missing data with the black bars along the top. The dots on the bottom of the plot show the variety of predictors.

predictors contained outliers. Descriptive statistics were also calculated for the training set to inspect skewness and kurtosis (Table 7).

Three univariate normality tests, Anderson-Darling, Lilliefors, and Cramer-von Mises, were calculated on each numeric predictor variable before data transformations were performed. The training dataset was too large for the individual tests, and a random sample of 5000 rows was created to run the three tests. All predictor variables were found not to have a normal distribution for each of the three tests (Table 8).

Multivariate normality and outliers were assessed using the following tests, Mardia's, Henze-Zirkler's, and Doornik-Hansen's, and the Cook's distance, before any

Table 7

*Descriptive Statistics for Both Cohort before Outlier Capping, Transformation, and Normalization*

Variable	Min	Max	<i>Mdn</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
age	15.00	70.00	18.00	18.71	2.89	7.95	87.09
hsgpa	1.00	4.00	2.94	2.97	0.52	0.16	-0.78
credit	0.00	52.00	20.00	18.45	10.60	-0.21	-0.97
remed	0.00	6.00	0.00	0.73	1.09	1.35	0.85
online	0.00	100.00	0.00	10.53	18.65	2.54	7.41
gpa	0.00	4.00	2.36	2.21	1.15	-0.36	-0.90
percaid	0.00	300.23	68.27	74.42	30.33	-0.56	-0.20
paid	0.00	26210.00	5682.00	6534.10	4461.00	0.70	-0.10
award	0.00	27190.00	8887.50	9276.80	4363.80	0.35	-0.35

*Note.* n = 13,286. age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = college GPA for the first three semesters. percaid = percentage of financial aid used by the student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters.

data transformations using the same random sample of 5000 from the training dataset.

The Mardia test indicate multivariate nonnormality by measuring skewness;  $M(4999) = 96981.38$ ,  $p < .001$  and kurtosis;  $M(4999) = 429.06$ ,  $p < .001$ , The Henze-Zirkler's,  $HZ(4999) = 9.36$ ,  $p < .001$  and the Doornik-Hansen,  $D(18) = 93155.58$ ,  $p < .001$ , tests also support multivariate nonnormality before outlier capping. The Cook's distance identified multiple values that could be influencing the data.

**Outlier Capping, Transformation, and Normalization**

Since the data contained numerous outliers and was not normally distributed, different methods were used to improve both scenarios. Outlier capping, which adjusts the extreme outliers to four standard deviations from the mean, was applied to the training data set. The Yeo-Johnson transformation was applied to these variables in the

Table 8

*Univariate Normality Test for Both Cohort before Outlier Capping, Transformation, and Normalization*

Variable	Anderson-Darling		Lilliefors		Cramer-von Mises	
	Value	p	Value	p	Value	p
age	927.834	< .001	0.372	< .001	191.548	< .001
hsgpa	23.594	< .001	0.057	< .001	3.162	< .001
credit	55.516	< .001	0.096	< .001	8.967	< .001
remed	412.441	< .001	0.321	< .001	71.571	< .001
online	457.238	< .001	0.281	< .001	84.644	< .001
gpa	50.831	< .001	0.069	< .001	7.172	< .001
paid	58.114	< .001	0.101	< .001	9.369	< .001
award	22.338	< .001	0.075	< .001	3.577	< .001
percaid	155.269	< .001	0.149	< .001	23.949	< .001

*Note.* n = 5,000. age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = college GPA for the first three semesters. percaid = percentage of financial aid used by the student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters.

recipe stage. The numeric predictors were normalized to change the standard deviation to one and the mean to zero. These techniques did help improve the normality of the training set slightly (Table 9). Histograms and Q-Q plots were re-rerun and showed a slight improvement in the data distribution, but outliers still existed. The three univariate normality tests, Anderson-Darling, Lilliefors, and Cramer-von Mises, were calculated on each of the numeric predictor variables after the data transformations were performed with a random sample of 5000. Again, all predictor variables were found not to have a normal distribution for each of the three tests (Table 10). The multivariate normality tests, Mardia's, Henze-Zirkler's, and Doornik-Hansen, were calculated again and showed the sample's nonnormality after the transformations. The Mardia test indicate multivariate

Table 9

*Descriptive Statistics for Both Cohort after Outlier Capping, Transformation, and Normalization*

Variable	Min	Max	<i>Mdn</i>	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
age	-1.02	15.77	-0.24	0.00	1.00	7.79	78.74
hsgpa	-4.05	1.87	-0.01	0.00	1.00	-0.01	-0.75
credit	-1.85	2.85	0.22	0.00	1.00	-0.36	-0.86
remed	-0.91	1.83	-0.91	0.00	1.00	0.30	-1.72
online	-0.89	1.59	-0.89	0.00	1.00	0.30	-1.79
gpa	-1.76	1.65	0.11	0.00	1.00	-0.23	-1.06
paid	-2.43	2.92	0.00	0.00	1.00	-0.27	0.00
award	-3.04	3.15	0.02	0.00	1.00	-0.10	-0.43
percaid	-2.17	5.63	0.31	0.00	1.00	-0.59	-0.79

*Note.* n = 13,286. age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = college GPA for the first three semesters. percaid = percentage of financial aid used by the student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters.

nonnormality by measuring skewness;  $M(4999) = 87561.82$ ,  $p < .001$  and kurtosis;

$M(4999) = 403.05$ ,  $p < .001$ . The Henze-Zirkler's,  $HZ(4999) = 9.07$ ,  $p < .001$  and the

Doornik-Hansen,  $D(18) = 35900.58$ ,  $p < .001$ . The Cook's distance was also recalculated and continued to show numerous points that could be influencing the data.

Predictor interactions were created to identify potential relationships that could exist in the dataset. The focus of these new predictors was to examine the relationships between credit with the other academic and financial variables that could affect a student during their first year in college: credit by gpa, credit by online, credit by award, credit by paid, credit by percaid, credit by remed, credit by gpa by online, credit by gpa by remed, credit by online by remed. An additional set of predictors were created to study the relationship between gpa and the other academic and financial variables that could affect

Table 10

*Univariate Normality Test for Both Cohort after Outlier Capping, Transformation, and Normalization*

Variable	Anderson-Darling		Lilliefors		Cramer-von Mises	
	Value	p	Value	p	Value	p
age	930.695	< .001	0.369	< .001	191.666	< .001
hsgpa	22.539	< .001	0.056	< .001	3.121	< .001
credit	51.711	< .001	0.093	< .001	8.206	< .001
remed	418.981	< .001	0.324	< .001	72.776	< .001
online	439.839	< .001	0.279	< .001	80.879	< .001
gpa	52.862	< .001	0.071	< .001	7.557	< .001
paid	58.419	< .001	0.101	< .001	9.369	< .001
award	20.937	< .001	0.072	< .001	3.285	< .001
percaid	147.484	< .001	0.146	< .001	22.452	< .001

*Note.* n = 5,000. age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = college GPA for the first three semesters. percaid = percentage of financial aid used by the student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters.

a student during their first year in college: gpa by paid, gpa by online, gpa by award, gpa by percaid, and gpa by remed.

**Research Question 1**

1: Are background factors (age, gender, race or ethnicity, and high school GPA), academic factors (college GPA, percentage of courses taken in an online format, number of remedial courses taken, and the number of credits earned during the first academic year), and financial factors (FAFSA completion, amount of financial aid awarded, and amount of financial aid paid to the student during the first academic year) significant in predicting first-year student retention for community college students?



Five different types of models: logistic regression, random forest, support vector machine with the radial kernel, support vector machine with the poly kernel, and neural network were created with the training dataset. A seed value was used throughout the entire modeling process to allow for the reproduction of the results. The modeling process included a grid function that optimized the different parameters in the different models to find the best combination of parameters. The grid value for this study is 20, which creates 20 different models for each model type except for logistic regression, which only produced one model. The cross validation process first divided the entire data set into two different sets: the training and test data sets. The training data set is subdivided into ten individual files that are used to train the models. The results are from these 20 different models for each model type provide the evaluation metrics for comparison. The final model for each model type is selected using the highest ROC\_AUC value from the training models and identifies the critical variables for the training data sets. Finally, the final five models use the testing dataset to collect the final evaluation metrics and important variables.

**Random Forest.** The random forest models were created using the randomForest engine to determine the optimal mtry, trees, and min\_n for each evaluation metric. Random forest trees are an ensemble method that creates different decision trees to form the forest with the tree with the most votes becoming the model used (Han et al., 2011). The final random forest model chosen had a mtry value of 10, 1781 trees, and a min\_n value of 36. The mtry refers to the number of predictors to use at each split in the model, with the recommended value being one-third of the predictors (James et al., 2013; Kuhn et al., 2019; Kuhn & Johnson, 2013). The mtry value of 10 is roughly one-third of the

predictors and interactions used in the final model. The last random forest model had 1781 trees which refer to the number of decision trees in the mode. The `min_n` function finds the minimum number of terminal nodes in the forest, with the final model having a value of 36 for the `min_n`.

Variable importance was determined for the final random forest model using the training data set (Table 11). For the training dataset, the variable with the highest importance for retained students was the number of credit hours with a value of .0726 (SD = .0026) and referred to the number of credit hours taken (Figure 3). The next highest variable with slight importance to retention was the interaction between the number of credit hours and GPA for the first three semesters with a value of .0143 (SD = .0023). The third highest variable was high school GPA, with a value of .0117 (SD = .0010). The following four variables were similar in variable importance; the interaction between the number of credit hours and the percentage of financial aid used (.0104, (SD = .0007), GPA with a value of .0099 (SD = .0013), Black or African American students (.0097 (SD = .0008)), and the number of credit hours and the number of remedial classes (.0096 (SD = .0008)). Two variables that were comparable in their importance were the percentage of online courses with a value of 0.0082 (SD = .0005) and the percentage of financial aid used (0.0082 (SD = .0009)). The interaction between the number of credit hours, GPA, and the number of remedial courses was also important, with a value of .0075 (SD = .0010). There were no significant variables for predicting retention in first year students in the training dataset for the random forest model (Figure 4).

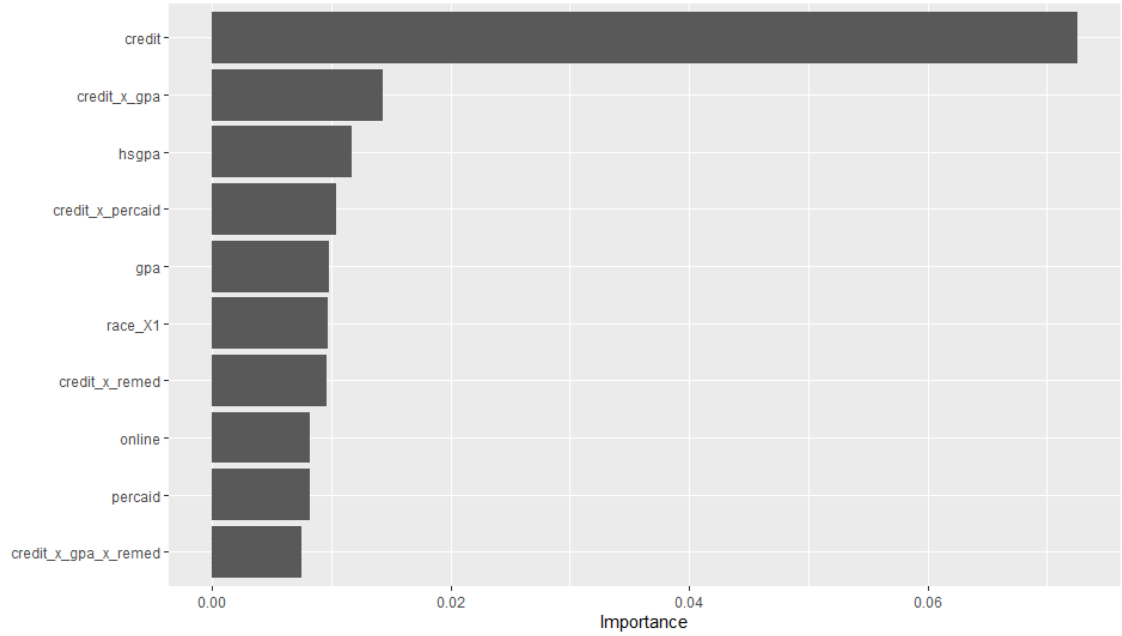
Variable importance was determined for the final random forest model using the test data set (Table 12). The variable importance was the same variable that had the

Table 11

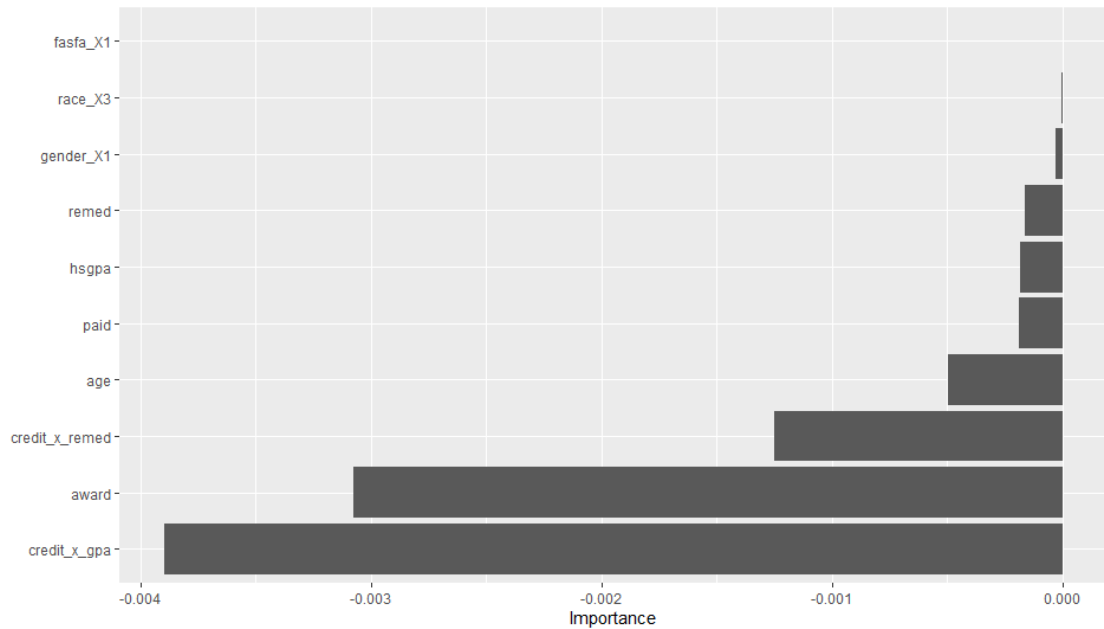
*Variable Importance for Random Forest Final Model with Training Data*

Variable	Importance	SD
credit	0.0726	0.0026
credit_x_gpa	0.0143	0.0023
hsgpa	0.0117	0.0010
credit_x_percaid	0.0104	0.0007
gpa	0.0099	0.0013
race_x1	0.0097	0.0008
credit_x_remed	0.0096	0.0008
online	0.0082	0.0005
percaid	0.0082	0.0009
credit_x_gpa_x_remed	0.0075	0.0010
gpa_x_percaid	0.0060	0.0006
gpa_x_award	0.0057	0.0006
gpa_x_online	0.0055	0.0004
credit_x_online	0.0055	0.0005
gpa_x_paid	0.0052	0.0005
gpa_x_remed	0.0047	0.0005
credit_x_paid	0.0047	0.0008
award	0.0047	0.0008
race_x2	0.0046	0.0005
credit_x_award	0.0041	0.0008
credit_x_gpa_x_online	0.0039	0.0003
age	0.0036	0.0005
remed	0.0023	0.0005
paid	0.0023	0.0005
credit_x_online_x_remed	0.0016	0.0001
gender_x1	0.0014	0.0002
fasfa_x1	0.0013	0.0002
race_x3	0.0002	0.0001

age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = college GPA for the first three semesters. percaid = percentage of financial aid used by the student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters. race\_x1 = Black or African American students. race\_x2 = Hispanic or Latino students. race\_x3 = other students. fasfa\_x1 = no FASFA completion. gender\_x1 = female students. the x indicates an interaction between two or three different variables.



*Figure 3.* Retention variable importance plot for random forest model using the training data. This plot displays variable importance from highest to lowest order with the variable names located on the right side and the length of the bar showing the level of importance from left to right.



*Figure 4.* Nonretention variable importance plot for random forest model using the training data. This plot displays variable importance from highest to lowest order with the variable names located on the right side and the length of the bar showing the level of importance from left to right at the 0.000 value. No variable was identified as important to nonretention in this model.

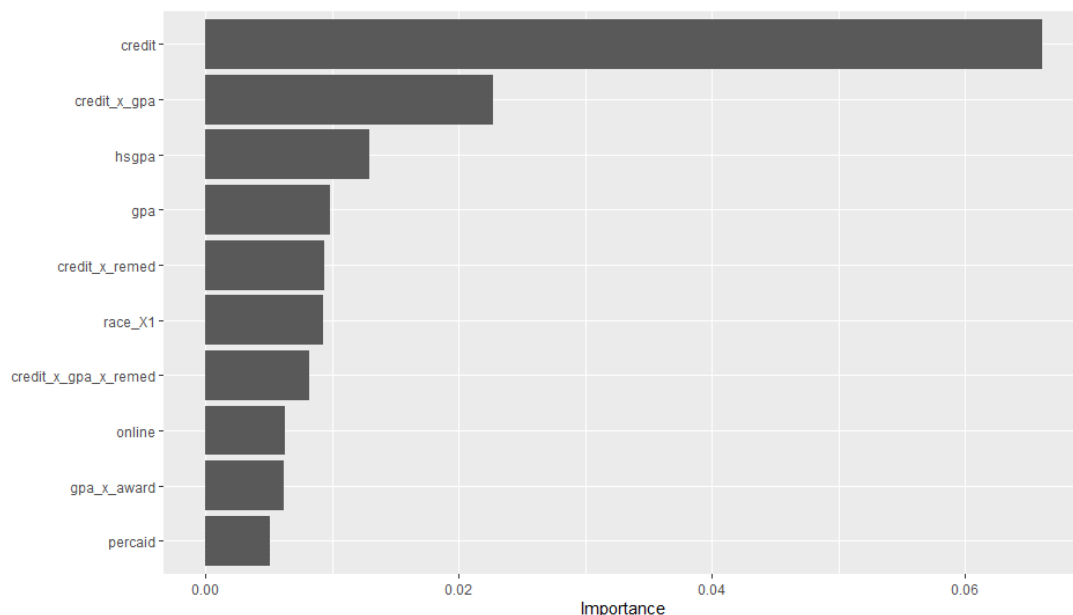
Table 12

*Variable Importance for Random Forest Final Model with Test Data*

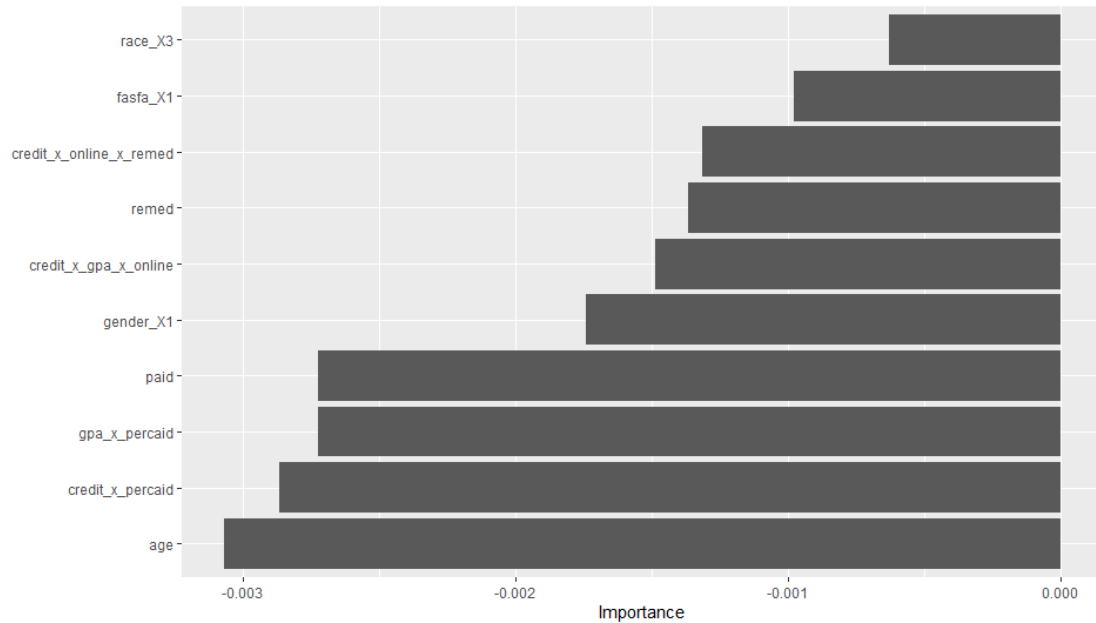
Variable	Importance	SD
credit	0.0661	0.0039
credit_x_gpa	0.0227	0.0022
hsgpa	0.0129	0.0011
gpa	0.0098	0.0014
credit_x_remed	0.0094	0.0007
race_x1	0.0093	0.0009
credit_x_gpa_x_remed	0.0082	0.0011
online	0.0063	0.0007
gpa_x_award	0.0062	0.0006
percaid	0.0050	0.0007
credit_x_paid	0.0049	0.0009
credit_x_award	0.0049	0.0008
gpa_x_remed	0.0047	0.0009
gpa_x_paid	0.0046	0.0008
gpa_x_online	0.0045	0.0005
credit_x_online	0.0042	0.0006
race_x2	0.0041	0.0008
award	0.0037	0.0003
age	0.0031	0.0004
credit_x_percaid	0.0029	0.0010
gpa_x_percaid	0.0027	0.0006
paid	0.0027	0.0009
gender_x1	0.0017	0.0004
credit_x_gpa_x_online	0.0015	0.0006
remed	0.0014	0.0005
credit_x_online_x_remed	0.0013	0.0003
fasfa_x1	0.0010	0.0001
race_x3	0.0006	0.0002

age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = college GPA for the first three semesters. percaid = percentage of financial aid used by the student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters. race\_x1 = Black or African American students. race\_x2 = Hispanic or Latino students. race\_x3 = other students. fasfa\_x1 = no FASFA completion. gender\_x1 = female students. the x indicates an interaction between two or three different variables.

highest importance for retained students as in the training set, which was the number of credit hours taken but decreased slightly in significance with a value of .0661 (SD = .00239) (Figure 5). The following two highest variables increased in their importance to retention with the interaction between the number of credit hours and GPA for the first three semesters, which had a value of .0227 (SD = .0022), and high school GPA had a value of .0129 (SD = .0011) The three variables that had similar importance values were GPA with a value of .0098 (SD = .0098), the interaction between the number of credit hours and the number of remedial classes (.0094 (SD = .0007)), and Black or African American students (.0093 (SD = .0009). The interaction between the number of credit hours, GPA, and the number of remedial courses was also important, with a value of .0082 (SD = .0011). There were no significant variables for predicting nonretention in first year students in the test data set (Figure 6). The random forest model was able to



*Figure 5.* Retention variable importance plot for random forest model using the test data. This plot displays variable importance from highest to lowest order with the variable names located on the right side and the length of the bar showing the level of importance from left to right.



*Figure 6.* Retention variable importance plot for random forest model using the test data. This plot displays variable importance from highest to lowest order with the variable names located on the right side and the length of the bar showing the level of importance from left to right. No variable was identified as important to nonretention in this model.

predict factors that were important for the retention of first year students, focusing on academic and background variables.

**Support Vector Machine with Polynomial Kernel.** SVM models create a margin hyperplane that separates the data into two classes and allows for correctly and incorrectly classified data on the two sides of the margin or hyperplane. The data points that lie along on the margin or the wrong side of the margin are called support vectors and may influence the model more than correctly classified data (James et al., 2013). SVM models are sensitive to predictors with skewed distributions and outliers and often overfit the data (Attewell & Monaghan, 2015; Kuhn & Johnson, 2019).

The SVM model with the polynomial kernel was created using the kernlab engine. The polynomial kernel fits the support vector classifier in a higher-dimensional

space and allows for more flexibility in the decision boundary (James et al., 2013). The tuning parameter for this SVM model was a cost value of 1.66 and was found by using the highest ROC\_AUC value from the tuning of the training data set (Attewell & Monaghan, 2015). Since the cost is a low value and will create larger margins, this model may underfit the data, but cross validation and tuning methods should help control these issues (Kuhn & Johnson, 2019).

Variable importance was determined for the final SVM model with the polynomial kernel model using the training data set (Table 13). For the training dataset, the variable with the highest importance for retained students was the number of credit hours taken with a value of .2977 (SD = .0049) (Figure 7). Another variable of importance to retention was the interaction between GPA for the first three semesters and the amount of financial aid awarded with a value of .1186 (SD = .0048). Two variables that were similar in importance are the interaction between GPA and the percentage of financial assistance used (.0706 (SD = .0030)) and the interaction between GPA and the number of remedial courses (.0624 (SD = .0025)). The number of credit hours had two interactions that were comparable in importance; the interaction between credit hours and the percentage of financial aid paid (.0367 (SD = .0031)) and the interaction between credit hours and the amount of financial aid awarded (.0344 (SD = .0036)).

There were four significant variables for predicting nonretention in first year students in the training dataset (Figure 8). The first variable was the percentage of financial aid used (the amount of financial aid paid divided by the amount of financial aid awarded) with a value of -.0006 (SD = .0007). The next two variables were slightly less important; the amount of financial aid awarded (-.0005, (SD = .0007)) and the interaction

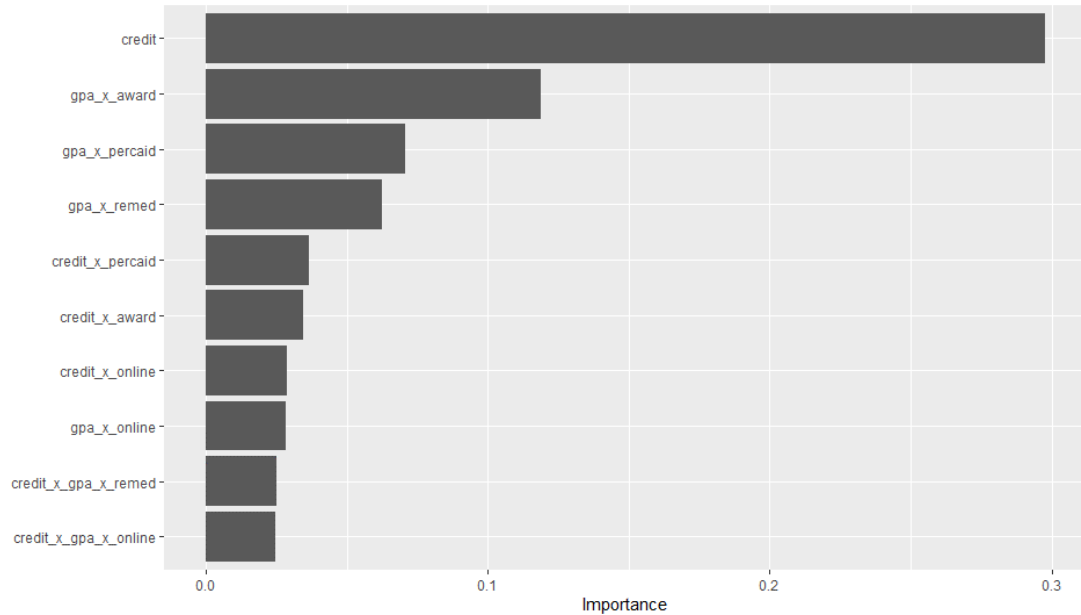


Table 13

*Variable Importance for SVM with Polynomial kernel Final Model with Training Data*

Variable	Importance	SD
credit	0.2977	0.0049
gpa_x_award	0.1186	0.0048
gpa_x_percaid	0.0706	0.0030
gpa_x_remed	0.0624	0.0025
credit_x_percaid	0.0367	0.0031
credit_x_award	0.0344	0.0036
credit_x_online	0.0286	0.0012
gpa_x_online	0.0285	0.0017
credit_x_gpa_x_remed	0.0250	0.0026
credit_x_gpa_x_online	0.0248	0.0034
credit_x_paid	0.0237	0.0025
gpa	0.0087	0.0015
online	0.0075	0.0018
paid	0.0030	0.0010
credit_x_gpa	0.0024	0.0015
gpa_x_paid	0.0016	0.0020
race_x2	0.0012	0.0009
remed	0.0006	0.0005
hsgpa	0.0002	0.0002
credit_x_online_x_remed	0.0001	0.0006
race_x3	0.0000	0.0000
race_x1	0.0000	0.0007
age	-0.0001	0.0002
fasfa_x1	-0.0001	0.0001
gender_x1	-0.0003	0.0004
credit_x_remed	-0.0005	0.0014
award	-0.0005	0.0007
percaid	-0.0006	0.0007

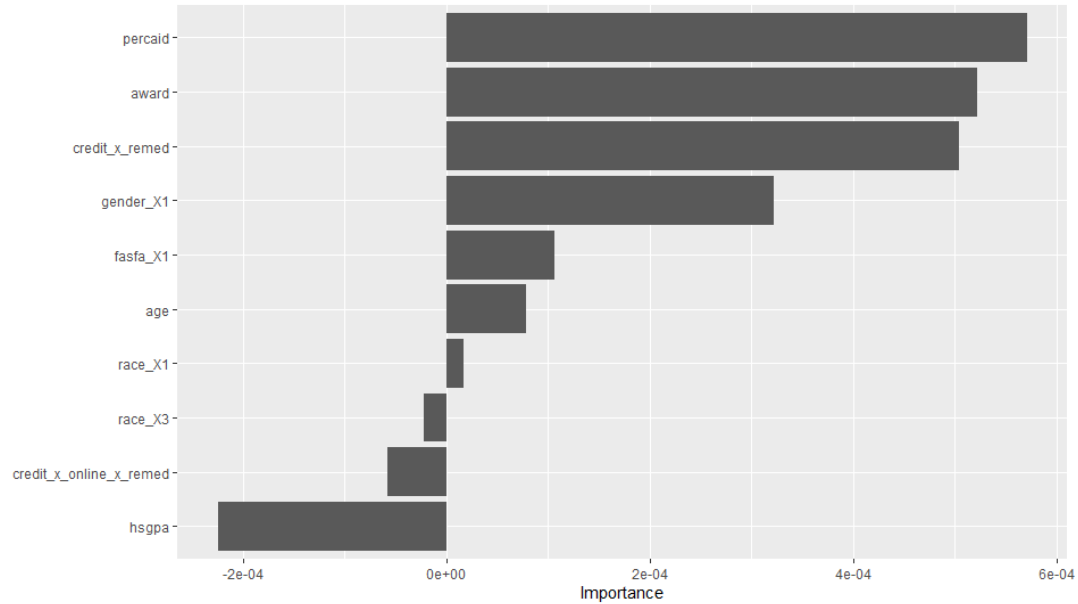
age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = college GPA for the first three semesters. percaid = percentage of financial aid used by the student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters. race\_x1 = Black or African American students. race\_x2 = Hispanic or Latino students. race\_x3 = other students. fasfa\_x1 = no FASFA completion. gender\_x1 = female students. the x indicates an interaction between two or three different variables.



*Figure 7.* Retention variable importance plot for svm with the polynomial kernel using the training data. This plot displays variable importance from highest to lowest order with the variable names located on the right side and the length of the bar showing the level of importance from left to right.

between the number of credit hours and the number of remedial courses taken (-.0005, (SD = .0014)). The fourth important variable for retention using this model was female students with a value of (-.0003, (SD = .0004)). The other variables had a minimal or no impact on retention status for first-year students.

Variable importance was determined for the final SVM with the polynomial kernel using the test data set (Table 14). The variable importance was the same for retained students in the training set: the number of credit hours taken with a value of 0.2974 (SD = .0041) and decreased slightly in importance in the test data set (Figure 9). The interaction between the GPA for the first three semesters and the amount of financial aid awarded (0.1239 (SD = .0033)) was important in the test model. The GPA interactions between the percentage of financial aid used (0.0861 (SD = .0041)), the number of remedial courses (0.0638 (SD = .0048)), and the percentage of



*Figure 8.* Nonretention variable importance plot for svm with the polynomial kernel using the training data. This plot displays variable importance from highest to lowest order with the variable names located on the right side and the length of the bar showing the level of importance from left to right.

online courses were also important in retention. There were six different interaction with the number of credit hours that were significant to retention: with the percentage of financial aid used (.0429 (SD = .0037)), the amount of financial aid awarded (.0362 (SD = .0041)), the percentage of online courses (.0356 (SD = .0027)), GPA and the percentage of online courses (.0279 (SD = .0021)), the amount of financial aid paid (.0261 (SD = .0031)), and ), GPA and the number of remedial courses (.0251 (SD = .0016)).

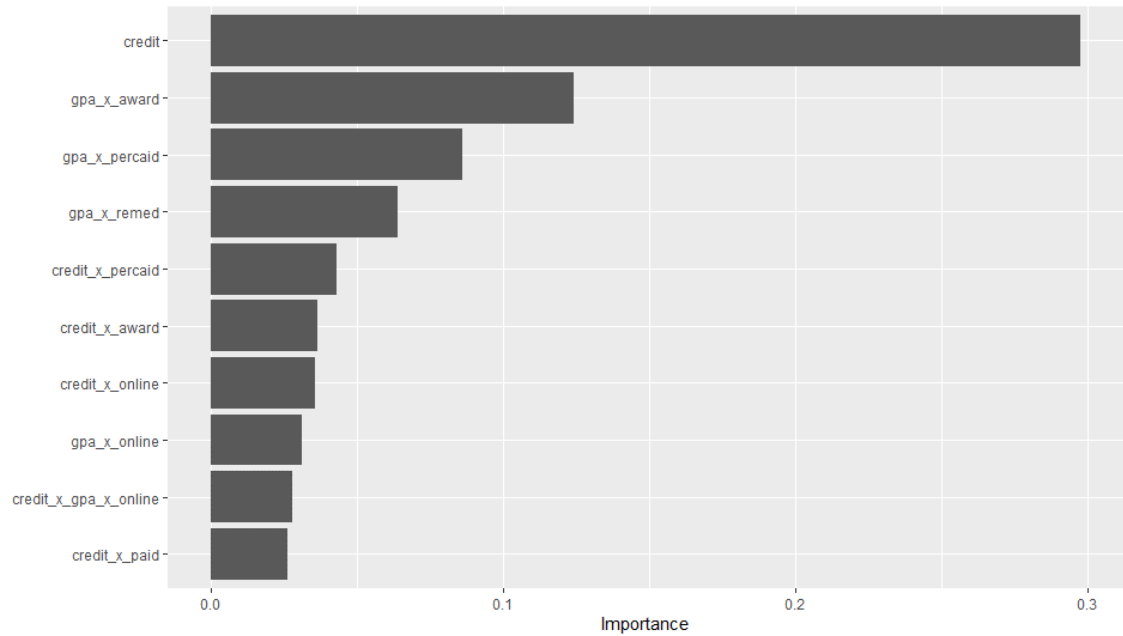
There was a significant change in the variables for predicting nonretention in first year students in the test data set from the earlier results seen in the training data set (Figure 10). The variable with the highest importance was the interaction between credit hours and the number of remedial courses (-.0013 (SD = .0009), which increased from the value in the training set. Another variable of importance for nonretention identified

Table 14

*Variable Importance for SVM with Polynomial kernel Final Model with Test Data*

Variable	Importance	SD
credit	0.2974	0.0041
gpa_x_award	0.1239	0.0033
gpa_x_percaid	0.0861	0.0041
gpa_x_remed	0.0638	0.0048
credit_x_percaid	0.0429	0.0037
credit_x_award	0.0362	0.0041
credit_x_online	0.0356	0.0027
gpa_x_online	0.0312	0.0020
credit_x_gpa_x_online	0.0279	0.0021
credit_x_paid	0.0261	0.0031
credit_x_gpa_x_remed	0.0251	0.0016
gpa	0.0102	0.0026
online	0.0092	0.0015
paid	0.0055	0.0011
credit_x_gpa	0.0043	0.0033
gpa_x_paid	0.0037	0.0026
race_x2	0.0006	0.0008
remed	0.0003	0.0008
race_x3	0.0000	0.0000
gender_x1	0.0000	0.0004
hsgpa	-0.0001	0.0002
age	-0.0002	0.0002
fasfa_x1	-0.0002	0.0001
percaid	-0.0005	0.0008
award	-0.0006	0.0009
credit_x_online_x_remed	-0.0008	0.0007
race_x1	-0.0009	0.0006
credit_x_remed	-0.0013	0.0009

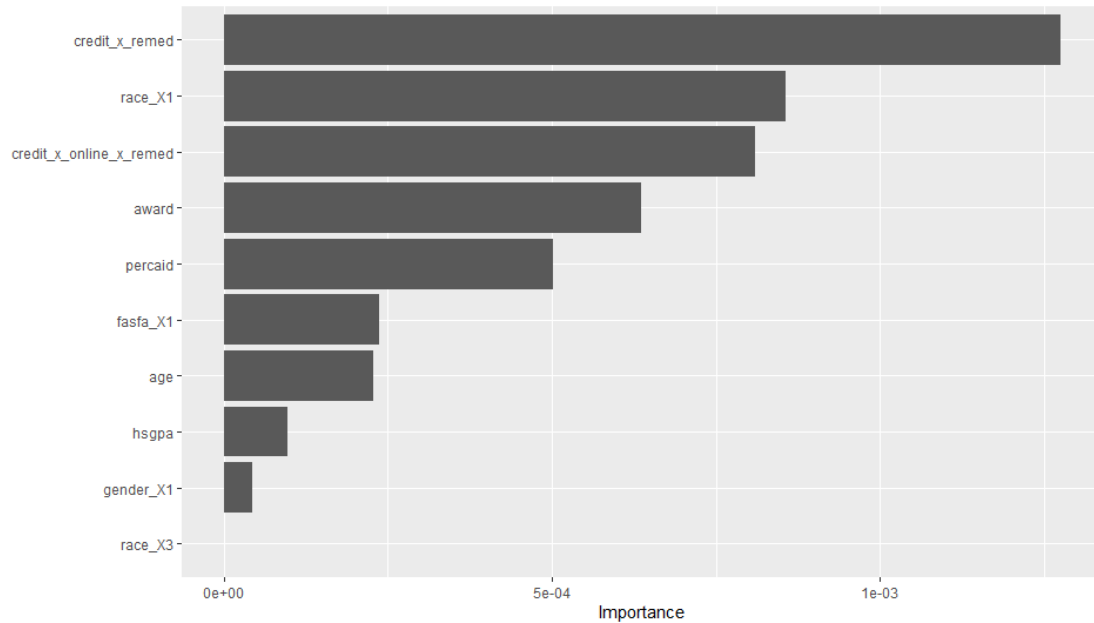
age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = college GPA for the first three semesters. percaid = percentage of financial aid used by the student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters. race\_x1 = Black or African American students. race\_x2 = Hispanic or Latino students. race\_x3 = other students. fasfa\_x1 = no FASFA completion. gender\_x1 = female students. the x indicates an interaction between two or three different variables.



*Figure 9.* Retention variable importance plot for svm with the polynomial kernel using the test data. This plot displays variable importance from highest to lowest order with the variable names located on the right side and the length of bar showing the level of importance from left to right.

Black or African American students with a value of  $-.0009$  (SD =  $.0004$ ). This variable increased in importance from the training set. The interaction between the number of credit hours, the percentage of online courses, and the number of remedial classes with a value of  $-0.0008$  (SD =  $0.0007$ ) were not identified as necessary in the training set. The amount of financial aid awarded with a value of  $-.0006$  (SD =  $.0009$ ) and the percentage of financial assistance used ( $-.0005$  (SD =  $.0008$ )) remained of the same importance in the test data set.

The final SVM with polynomial kernel model was able to predict factors that were important for the retention of first year students, which focused on credit hours and the interactions between the academic and financial factors. These interactions centered around GPA or credit hours and the financial factors or types of courses taken (remedial



*Figure 10.* Nonretention variable importance plot for svm with the polynomial kernel using the test data. This plot displays variable importance from highest to lowest order with the variable names located on the right side and the length of bar showing the level of importance from left to right.

or online). The model identifies the significance of the relationship between the academic and financial factors on retention. The variables for nonretention were the interaction between credit hours and types of courses, Black or African American students, and financial aid variables.

**Support Vector Machine with Radial Kernel.** The SVM model with the radial kernel was created using the kernlab engine. The radial kernel differs from the polynomial kernel by comparing the Euclidean distance between two points and classifying the data based on the distance. This SVM model's tuning parameter is the cost value and the sigma value, which was found by using the highest ROC\_AUC value from the training data set's tuning (Attewell & Monaghan, 2015). The final SVM model with a radial kernel has a cost of 0.024 and a sigma of 0.030. Since both tuning values are low,

this model may underfit the data due to larger, inflexible margins. The use of cross validation and grid search methods should help control these issues (Kuhn & Johnson, 2019).

Variable importance was determined for the final SVM model with the radial kernel using the training data set (Table 15). For the training dataset, the variable with the highest importance for retained students was the number of credit hours with a value = .0155 (SD = .0014) (Figure 11). Different variable interaction with the number of credit hours was identified as of importance to retention; the interaction with GPA for the first three semesters (0.0106 (SD = 0.0017)), the interaction with the amount of financial aid awarded (0.0061 (SD = 0.0011)), the interaction with the amount of financial aid paid (0.0045 (SD = 0.0008)), and the interaction with the percentage of financial aid used (0.0045 (SD = 0.0014)). The interaction between GPA and the percentage of financial aid used (0.0026 (SD = 0.0012)) and the interaction between GPA and the amount of financial aid paid (0.0026 (SD = 0.0012)) had the same variable importance value in this model. The interaction between the number of credit hours, GPA, and percentage of online hours (0.0023 (SD = 0.0011)) was also significant to retention.

Four variables were identified as having slight importance for predicting nonretention in first-year students in the training dataset (Figure 12). The first variable was the percentage of finances used with a value of -.0007 (SD = .0005). The following two variables were slightly less significant: the interaction between GPA and the number of remedial courses taken (-.0006, (SD = .0006)) and the number of remedial classes taken (-.0005, (SD = .0014)). The fourth important variable for retention using this model was the percentage of online credit hours taken (-.0004 (SD = .0003)).

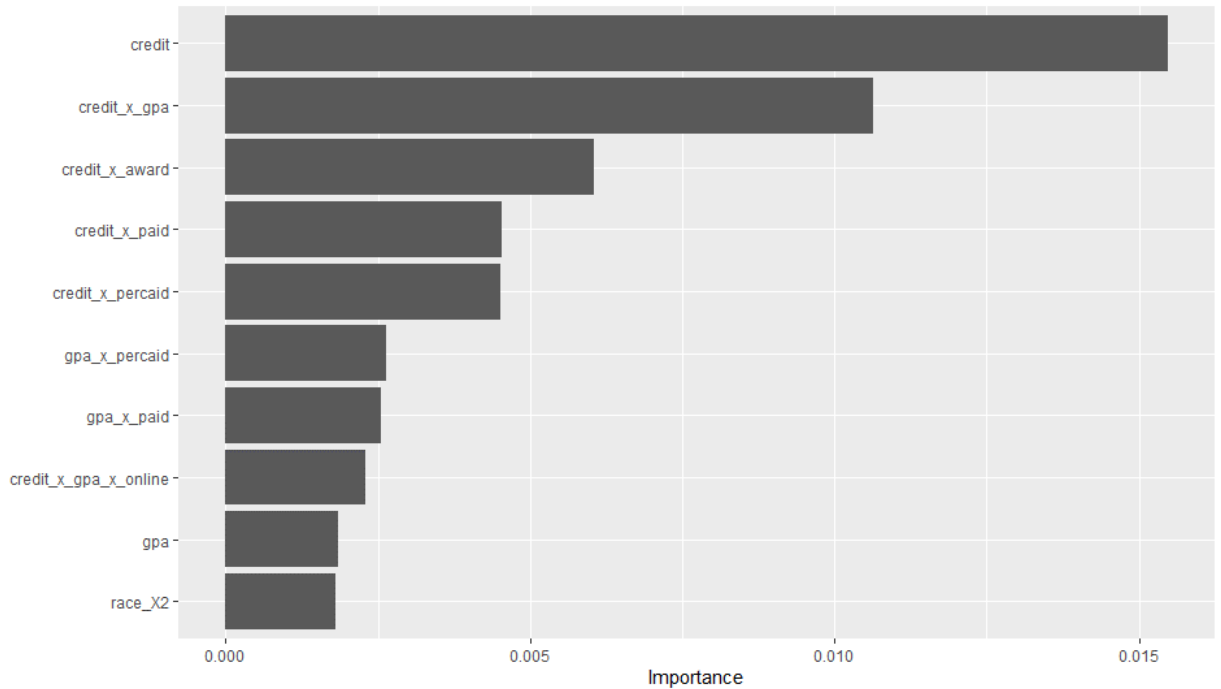
Table 15

*Variable Importance for SVM Radial Final Model with Training Data*

Variable	Importance	SD
credit	0.0155	0.0014
credit_x_gpa	0.0106	0.0017
credit_x_award	0.0061	0.0011
credit_x_paid	0.0045	0.0008
credit_x_percaid	0.0045	0.0014
gpa_x_percaid	0.0026	0.0012
gpa_x_paid	0.0026	0.0012
credit_x_gpa_x_online	0.0023	0.0011
gpa	0.0019	0.0013
race_x2	0.0018	0.0009
gpa_x_award	0.0018	0.0014
credit_x_online	0.0015	0.0014
credit_x_gpa_x_remed	0.0012	0.0006
credit_x_online_x_remed	0.0009	0.0006
race_x1	0.0009	0.0008
gpa_x_online	0.0007	0.0005
gender_x1	0.0006	0.0004
credit_x_remed	0.0006	0.0010
award	0.0006	0.0006
hsgpa	0.0003	0.0004
fasfa_x1	0.0002	0.0004
race_x3	0.0001	0.0003
paid	-0.0001	0.0006
age	-0.0001	0.0005
online	-0.0004	0.0003
remed	-0.0005	0.0005
gpa_x_remed	-0.0006	0.0006
percaid	-0.0007	0.0005

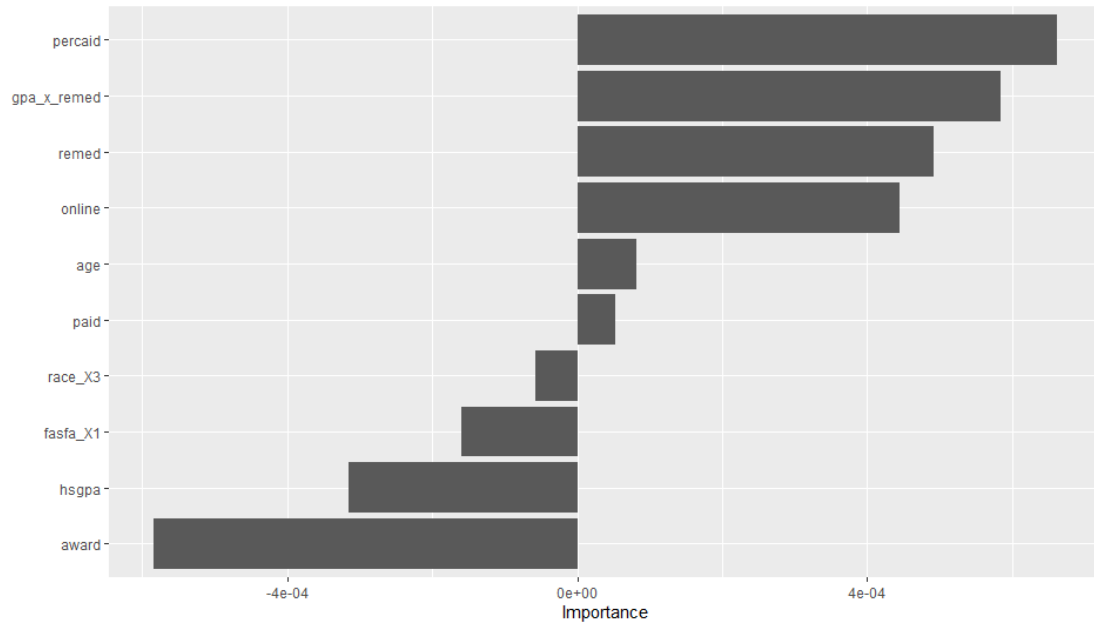
age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = College GPA for the first three semesters. percaid = percentage of financial aid used by the student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters. race\_x1 = Black or African American students. RACE\_X2 = Hispanic or Latino students. race\_x3 = other students. fasfa\_x1 = no FASFA completion. gender\_x1 = female students. The x indicates an interaction between two or three different variables.





*Figure 11.* Retention variable importance plot for svm with the radial kernel using the training data. This plot displays variable importance from highest to lowest order with the variable names located on the right side and the length of bar showing the level of importance from left to right.

The final SVM model with the radial kernel using the test data set calculated the variable importance (Table 16). The variable with the highest importance continued to be the number of credit hours (.0151 (SD = .0023)), which slightly decreased from the value in the training set (Figure 13). The second highest variable of importance was the interaction between the number of credit hours taken and GPA during the first three semesters, with a value of .0115 (SD = .0016). This variable had a slight increase in importance from the training set. Three interactions between the number of credit hours and the financial factors were significant: credit hours and percentage of financial aid used (.0068 (SD = .0009)), credit hours and amount of financial aid awarded (.0066 (SD = .0014)), and credit hours and amount of financial aid paid (.0044 (SD = .0011)). The



*Figure 12.* Nonretention variable importance plot for svm with the radial kernel using the training data. This plot displays variable importance from highest to lowest order with the variable names located on the right side and the length of bar showing the level of importance from left to right.

three interactions between GPA and financial variables were also significant: GPA and the amount of financial aid paid (.0044 (SD = .0011)), GPA and the percentage of financial aid used (.0026 (SD = .0014)), and GPA and the amount of financial aid awarded (.0025 (SD = .0011)). The interaction between credit hours and other academic factors; credit hours, GPA, and percentage of online courses (.0027 (SD = .0005)) and credit hours and percentage of online courses together (.0022 (SD = .0005)) were also identified as critical to retention.

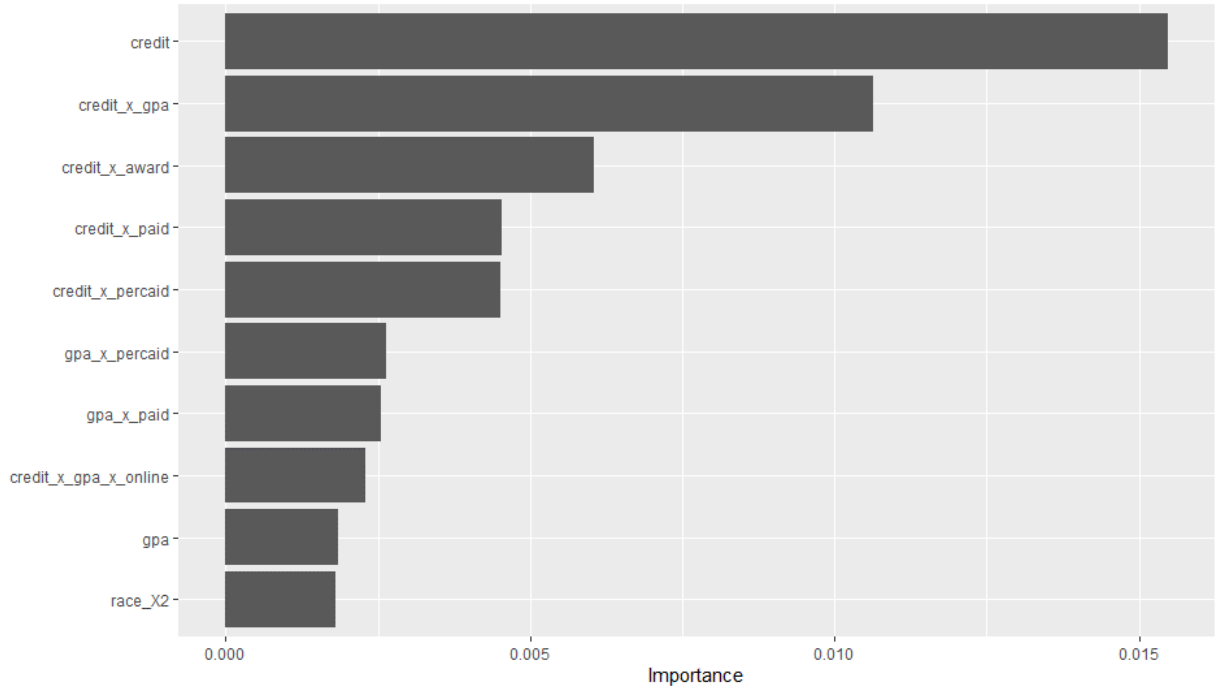
Only two variables were identified as having slight importance for predicting retention in first-year students in the test dataset (Figure 14). The first variable was the percentage of online credit hours taken with a value of -.0005 (SD = .0007) and had a similar value for importance in the training set. The variable, the interaction between

Table 16

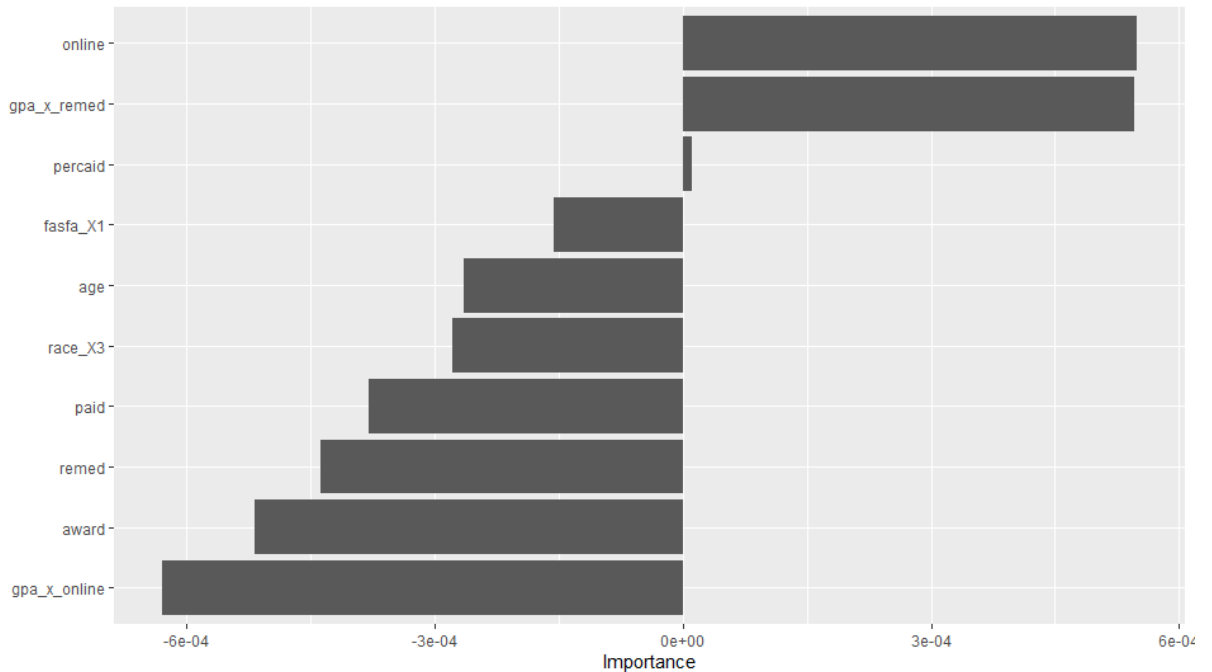
*Variable Importance for SVM Radial Final Model with Test Data*

Variable	Importance	SD
credit	0.015086	0.002185
credit_x_gpa	0.011536	0.001585
credit_x_percaid	0.006822	0.000933
credit_x_award	0.006589	0.001434
credit_x_paid	0.004384	0.001127
gpa_x_paid	0.003158	0.001213
credit_x_gpa_x_online	0.002784	0.000529
gpa_x_percaid	0.002640	0.001423
gpa_x_award	0.002489	0.001116
credit_x_online	0.002162	0.000563
gpa	0.001991	0.001353
credit_x_gpa_x_remed	0.001538	0.001036
race_x2	0.001371	0.000977
race_x1	0.001103	0.000429
gender_x1	0.000870	0.000286
hsgpa	0.000726	0.000404
credit_x_online_x_remed	0.000666	0.000407
credit_x_remed	0.000642	0.000983
gpa_x_online	0.000629	0.000421
award	0.000517	0.000589
remed	0.000438	0.000674
paid	0.000380	0.000559
race_x3	0.000279	0.000316
age	0.000265	0.000485
fasfa_x1	0.000156	0.000280
percaid	-0.000011	0.000723
gpa_x_remed	-0.000547	0.000868
online	-0.000549	0.000682

age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = College GPA for the first three semesters. percaid = percentage of financial aid used by the student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters. race\_x1 = Black or African American students. race\_x2 = Hispanic or Latino students. race\_x3 = other students. fasfa\_x1 = no FASFA completion. gender\_x1 = female students. The x indicates an interaction between two or three different variables.



*Figure 13.* Retention variable importance plot for svm with the radial kernel using the test data. This plot displays variable importance from highest to lowest order with the variable names located on the right side and the length of bar showing the level of importance from left to right.



*Figure 14.* Nonretention variable importance plot for svm with the radial kernel using the test data. This plot displays variable importance from highest to lowest order with the variable names located on the right side and the length of bar showing the level of importance from left to right.

GPA and the number of remedial courses taken (-.0005, (SD = .0009)), also had a similar value as seen in the training set. None of the other variables had importance in predicting retention.

The final SVM with the radial kernel model also predicted factors that were important for the retention of first-year students were credit hours and the interaction between academic and financial factors. The variables for nonretention were academic factors but had low importance levels. Both SVM models (polynomial and radial) identified similar factors for retaining students, with the radial kernel model variables having lower levels of importance overall.

**Neural Network.** Neural networks are a type of classification model that behave like biological neurons that pass information to other neurons to learn based on previous errors (Attewell & Monaghan, 2015). The neural network models were created using the keras engine with three different parameters: hidden units, epochs, and the penalty. The hidden unit represents the input combinations (predictor variables) transformed (Attewell & Monaghan, 2015; Kuhn & Johnson; 2019). The epoch describes the number of times the data is feed through the neural network model while the penalty controls the model weights (Attewell & Monaghan, 2015; Kuhn & Johnson; 2019).

The final neural network model: the hidden unit of 4, the penalty was 0.540, and the epochs were 811, were identified using the highest ROC\_AUC value from the 20 training models (Attewell & Monaghan, 2015). A hidden unit is a low number that may decrease the model's complexity and cause the overfitting of the data (Attewell & Monaghan, 2015).

Variable importance was determined for the final neural network model using the

training data set (Table 17). For the training dataset, the variable with the highest importance for retained students was the number of credit hours with a value = .3012 (SD = .0077) (Figure 15). The next variable regarding retention was the interaction between credit hours and the amount of financial aid paid for the first three semesters (.2824 (SD = .0017)). The interaction between the number of credit hours and the amount of financial aid awarded (.0837 (SD = .0037)) and the number of credit hours and the percentage of financial aid used interaction (.0768 (SD = .0022)) were also identified as important. The interactions between credit hours and other academic factors were significant as well; the interaction between credit hours, GPA, and the number of remedial courses (.0687 (SD = .0029)), the interaction between credit hours and percentage of online hours (.0600 (SD = .0025)), and the interaction between credit hours, GPA, and percentage of online hours (.0598 (SD = .0032)).

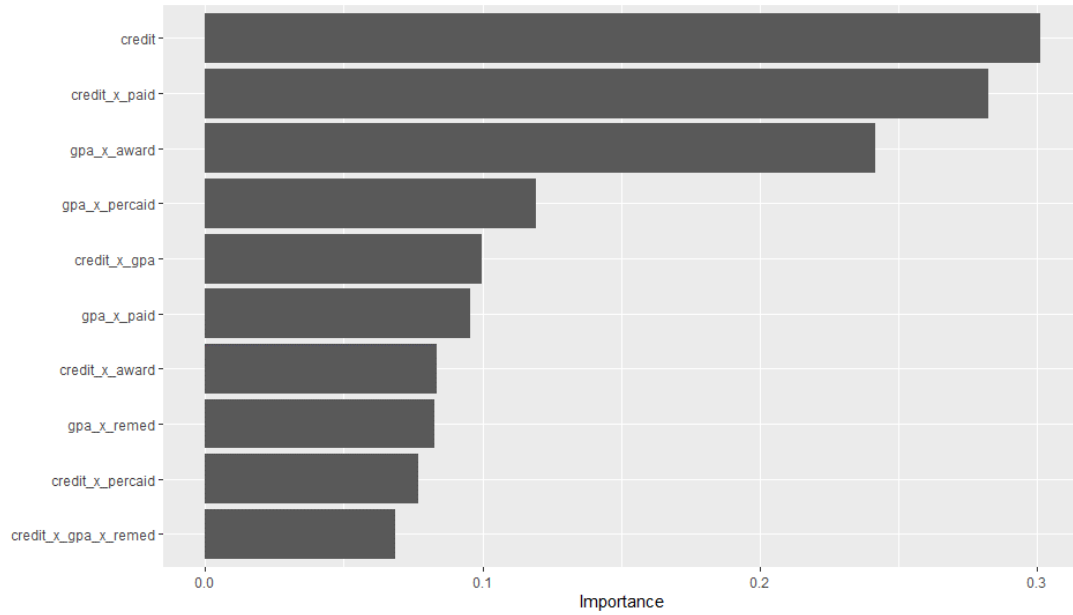
The interaction between GPA and financial variables were also identified as important to the retention; GPA and the amount of financial aid awarded (.2419 (SD = .0053)), GPA and the percentage of financial aid used (.1192 (SD = .0038)), and GPA and the amount of financial aid paid (.0958 (SD = .0043)). Another interaction had a lesser impact but was similar; interaction between GPA and the percentage of financial aid used (.1192 (SD = .0038)). Two other interactions between GPA and academic factors were significant: GPA and number of remedial courses (.0828 (SD = .0023)) and GPA and the percentage of online courses (.0600 (SD = .0025)). No variables were identified as important to nonretention in the neural network final model using the training data set.

Table 17

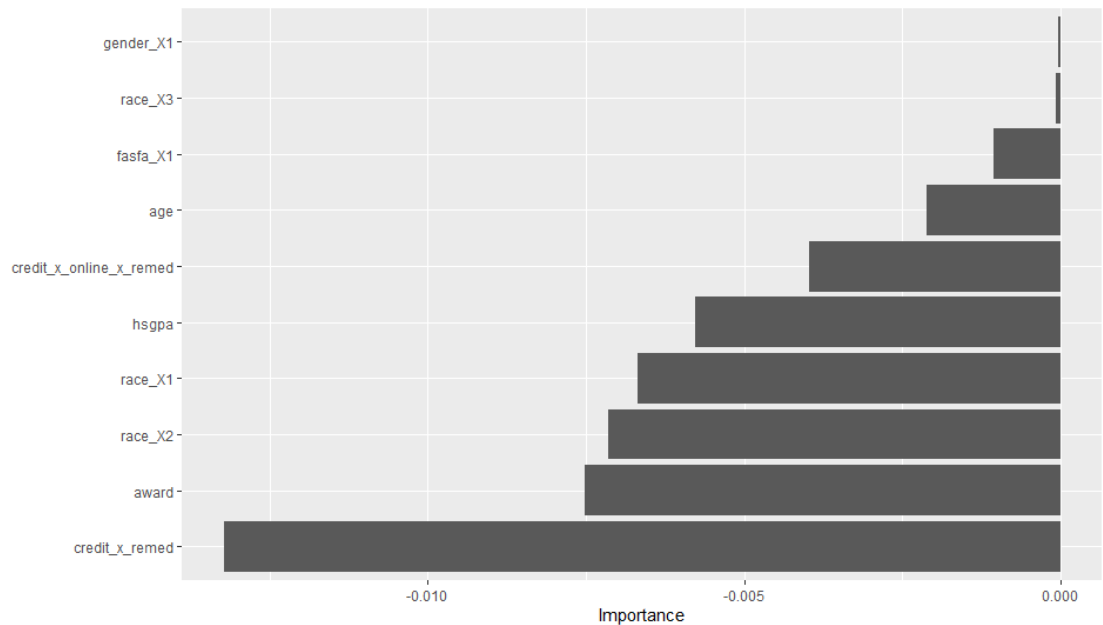
*Variable Importance for Neural Network Final Model with Training Data*

Variable	Importance	SD
credit	0.3012	0.0077
credit_x_paid	0.2824	0.0057
gpa_x_award	0.2419	0.0053
gpa_x_percaid	0.1192	0.0038
credit_x_gpa	0.1000	0.0047
gpa_x_paid	0.0958	0.0043
credit_x_award	0.0837	0.0037
gpa_x_remed	0.0828	0.0023
credit_x_percaid	0.0768	0.0022
credit_x_gpa_x_remed	0.0687	0.0029
gpa_x_online	0.0664	0.0022
credit_x_online	0.0600	0.0025
credit_x_gpa_x_online	0.0598	0.0032
gpa	0.0445	0.0024
online	0.0197	0.0015
percaid	0.0181	0.0014
remed	0.0167	0.0017
paid	0.0166	0.0012
credit_x_remed	0.0132	0.0010
award	0.0075	0.0008
race_x2	0.0071	0.0006
race_x1	0.0067	0.0011
hsgpa	0.0058	0.0008
credit_x_online_x_remed	0.0040	0.0007
age	0.0021	0.0005
fasfa_x1	0.0011	0.0004
race_x3	0.0001	0.0001
gender_x1	0.0000	0.0001

age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = College GPA for the first three semesters. percaid = percentage of financial aid used by the student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters. race\_x1 = Black or African American students. race\_x2 = Hispanic or Latino students. race\_x3 = other students. fasfa\_x1 = no FASFA completion. gender\_x1 = female students. the x indicates an interaction between two or three different variables.



*Figure 15.* Retention variable importance plot for neural networks using the training data. This plot displays variable importance from highest to lowest order with the variable names located on the right side and the length of bar showing the level of importance from left to right.



*Figure 16.* Nonretention variable importance plot for the neural network using the training data. This plot displays variable importance from highest to lowest order with the variable names located on the right side and the length of bar showing the level of importance from left to right. No variable was identified as important to nonretention in this model.



The final neural network model used the test data set to calculate the variable importance (Table 18). The variable with the highest importance was the interaction of the number of credit hours and GPA during the first three semesters with a value of .2615 (SD = .0056), which increased the value in the training set (Figure 17). The second highest variable of importance was the number of credit hours taken with a value of .2221 (SD = .0041) and decreased importance from the training set. The three interactions between the number of credit hours and financial factors were significant: credit hours and amount of financial aid paid (.1460 (SD = .0040)), credit hours and amount of financial aid award (.0541 (SD = .0033)), and credit hours and percentage of financial aid used (.0442 (SD = .0026)). The interaction between credit hours, GPA, and percentage of online was also identified as critical to retention (.0331 (SD = .0022)). The GPA for the first three semesters and its interactions with one financial variable and two academic variables were also identified as important to retention. The interaction between GPA and the amount of financial aid award (.0410 (SD = .0023) and GPA (.0235 (SD = .0019) by itself were significant. The interaction between GPA and the percentage of online courses (.0238 (SD = .0016) and the interaction between GPA and the number of remedial courses (0.0163 (SD = .0024) were also identified.

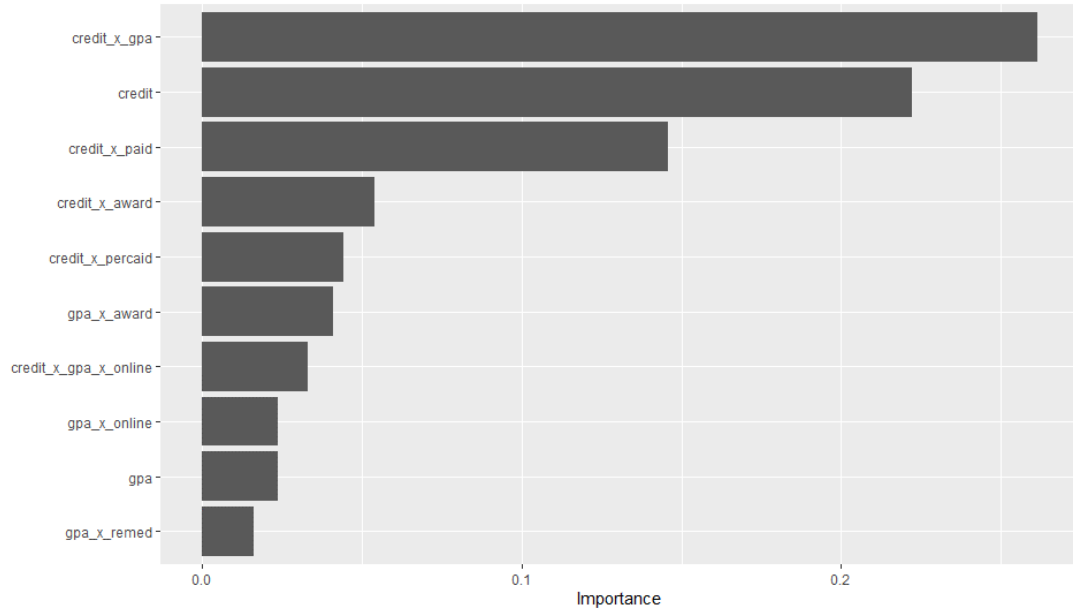
Only one variable was identified as important for predicting nonretention in first-year students in the test dataset for the final neural network (Figure 18). The variable was high school GPA (-.0002 (SD = .0003) and was not identified in the training set. The overall variable importance for neural networks shifted from training and test data sets and could indicate the model's overfitting. The interactions between academic and financial factors, as well as credit hours, were identified as important for retention.

Table 18

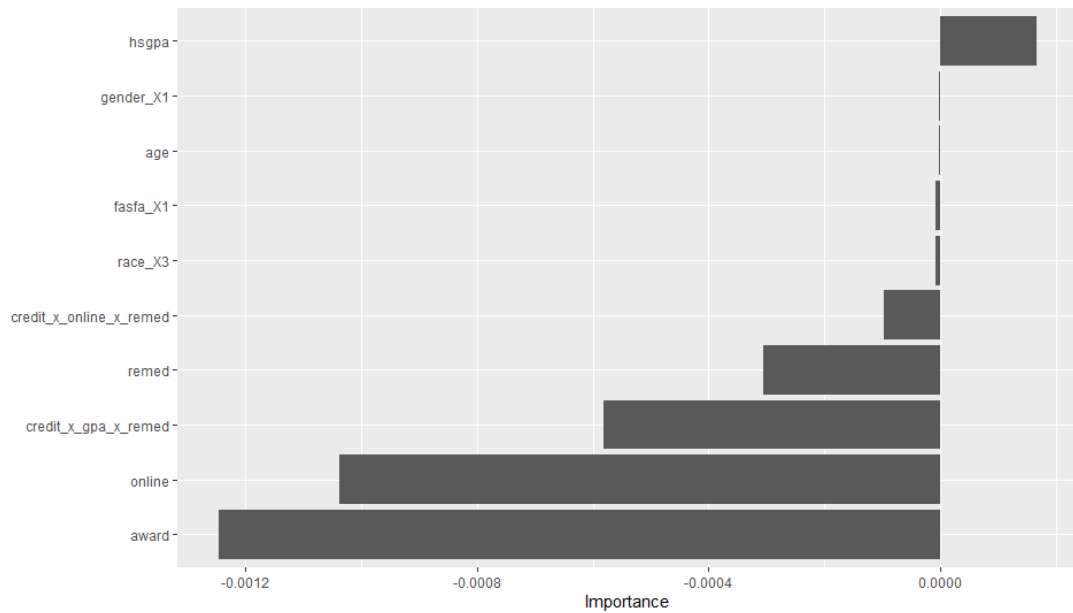
*Variable Importance for Neural Network Final Model with Test Data*

Variable	Importance	SD
credit_x_gpa	0.2615	0.0056
credit	0.2221	0.0041
credit_x_paid	0.1460	0.0040
credit_x_award	0.0541	0.0033
credit_x_percaid	0.0442	0.0026
gpa_x_award	0.0410	0.0023
credit_x_gpa_x_online	0.0331	0.0022
gpa_x_online	0.0238	0.0016
gpa	0.0235	0.0019
gpa_x_remed	0.0163	0.0024
credit_x_online	0.0078	0.0012
paid	0.0072	0.0008
percaid	0.0065	0.0012
race_x2	0.0043	0.0004
gpa_x_paid	0.0042	0.0012
credit_x_remed	0.0041	0.0009
gpa_x_percaid	0.0040	0.0004
race_x1	0.0019	0.0003
award	0.0012	0.0004
online	0.0010	0.0003
credit_x_gpa_x_remed	0.0006	0.0004
remed	0.0003	0.0001
credit_x_online_x_remed	0.0001	0.0000
race_x3	0.0000	0.0000
fasfa_x1	0.0000	0.0000
age	0.0000	0.0001
gender_x1	0.0000	0.0001
hsgpa	-0.0002	0.0003

age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = College GPA for the first three semesters. percaid = percentage of financial aid used by the student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters. race\_x1 = Black or African American students. race\_x2 = Hispanic or Latino students. race\_x3 = other students. fasfa\_x1 = no FASFA completion. gender\_x1 = female students. The x indicates an interaction between two or three different variables.



*Figure 17.* Retention variable importance plot for neural networks using the test data. This plot displays variable importance from highest to lowest order with the variable names located on the right side and the length of bar showing the level of importance from left to right.



*Figure 18.* Retention variable importance plot for neural networks using the test data. This plot displays variable importance from highest to lowest order with the variable names located on the right side and the length of bar showing the level of importance from left to right.

**Logistic Regression.** The logistic regression models were created using the glm engine and did not require any tuning like the other models. Logistic regression is like a linear regression model but differs since the linear regression has a continuous dependent variable while the logistic regression has a dichotomous dependent variable. The logistic regression model was consistent throughout the entire modeling process using the training and test data sets.

The results of the logistic regression model for the training set are listed in Table 19. Logistic regression has additional assumptions that need to be met before the models are interpreted and were calculated using the blorr package. To determine the overall performance of the logistic regression model, the Hosmer and Lemeshow goodness of fit test is ideal for binary classification models and compares the observed and expected frequencies of retention status. For the logistic regression model using the training data set, the goodness of fit test ( $\chi^2(8) = 34.23, p < 0.001$ ) found that the observed nonretained students differ significantly from the expected value of nonretained students. The small p-value of this test indicates the model did not fit the data well. Another measurement of model performance is the pseudo R squared value which is specifically for logistic regression and is similar to R squared in ordinary least-squares (OLS) regression (Smith & McKenna, 2013). Nagelkerke's pseudo R squared value, 0.346, shows that the academic, background and financial factors could account for 34.6% of the retention status of the students in this model. A different pseudo R squared value, McFadden's, had a value of 0.230, indicating 23% of the model could account for the variance in retention status. Eleven of the 28 predictors were statistically significant in predicting retention using the training data set: number of credit hours, GPA, Black or African American

Table 19

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

Variable	Log Odds	SE	Z	Pr(> z )		95% Confidence Interval		
						OR	Lower	Upper
(Intercept)	1.645	0.532	3.092	0.002	**			
age	0.041	0.026	1.571	0.117		1.042	0.991	1.097
hsgpa	0.021	0.035	0.567	0.571		1.022	0.952	1.093
credit	3.067	0.303	10.131	$p < .001$	***	21.478	11.862	38.964
paid	-0.027	0.318	-0.085	0.932		0.973	0.532	1.848
award	0.142	0.247	0.573	0.567		1.152	0.702	1.851
remed	0.027	0.109	0.245	0.807		1.027	0.831	1.274
online	-0.172	0.112	-1.537	0.124		0.842	0.674	1.045
gpa	-0.677	0.265	-2.557	0.011	*	0.508	0.302	0.854
percaid	0.291	0.201	1.452	0.146		1.337	0.895	1.961
gender_X1	-0.023	0.052	-0.441	0.661		0.977	0.883	1.082
race_X1	-0.344	0.067	-5.112	$p < .001$	***	0.709	0.622	0.809
race_X2	0.599	0.084	7.150	$p < .001$	***	1.821	1.547	2.149
race_X3	-0.039	0.111	-0.351	0.727		0.962	0.774	1.197
fasfa_X1	0.024	0.099	0.242	0.812		1.024	0.843	1.243
credit_x_gpa	-0.006	0.006	-1.038	0.299		0.994	0.982	1.005
credit_x_online	0.018	0.011	1.618	0.106		1.018	0.996	1.041
credit_x_award	0.006	0.002	-3.287	0.001	**	0.999	0.998	1.001
credit_x_paid	0.008	0.005	1.717	0.086		1.001	0.999	1.002
credit_x_percaid	-0.001	0.004	-3.746	$p < .001$	***	0.999	0.998	1.001
credit_x_remed	-0.016	0.034	-0.482	0.631		0.984	0.921	1.052
credit_x_gpa_x_online	-0.008	0.003	-2.177	0.029	*	0.993	0.986	0.999
credit_x_gpa_x_remed	-0.055	0.011	-4.875	$p < .001$	***	0.947	0.926	0.968
credit_x_online_x_remed	0.023	0.006	3.903	$p < .001$	***	1.023	1.012	1.035
gpa_x_paid	-0.002	0.002	-0.844	0.399		0.998	0.994	1.002
gpa_x_online	0.047	0.047	1.001	0.317		1.048	0.957	1.148
gpa_x_award	0.002	0.007	2.301	0.021	**	1.002	1.001	1.003
gpa_x_percaid	0.002	0.001	1.814	0.072		1.002	0.999	1.004
gpa_x_remed	0.886	0.154	5.753	$p < .001$	***	2.425	1.794	3.282

Note. AIC: 9968.40.  $p < 0.001$  '\*\*\*',  $p < 0.01$  '\*\*',  $p < 0.05$  '\*'

age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = college GPA for the first three semesters. percaid = percentage of financial aid paid divided by the amount of financial aid awarded to the student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters. race\_x1 = Black or African American students. race\_x2 = Hispanic or Latino students. race\_x3 = Other students. fasfa\_x1 = no FASFA completion. gender\_x1 = female students. The x indicates an interaction between two or three different variables.

students, Hispanic or Latino students, the interaction between the number of credit hours and amount of financial aid awarded to the student, the interaction between the number of credit hours and the percentage of financial aid used by the student, the interaction between the number of credit hours, GPA, and the percentage of online courses taken, the interaction between the number of credit hours, GPA, and the number of remedial classes taken, the interaction between the number of credit hours, percentage of online courses taken and the number of remedial classes taken, the interaction between GPA and amount of financial aid awarded to the student, and the interaction between GPA and the number of remedial classes taken. The other variables were not significant to the model for predicting retention.

Students who take more credit hours in their first three semesters are 21.478 times greater for being retained opposed to students who do not ( $z = 10.131$ ,  $p < .001$ , odds ratio = 21.478, 95% CI = 11.862 to 38.964) when all other variables remain unchanged. Another significant variable was the interaction between GPA and number of remedial courses ( $z = 5.753$ ,  $p < .001$ , odds ratio = 2.425, 95% CI = 1.794 to 3.282) had 2.425 times odds of students being retained given all other variables are constant. Hispanic or Latino students ( $z = 7.150$ ,  $p < .001$ , odds ratio = 1.821, 95% CI = 1.547 to 2.149) had 1.821 times odds of being retained given the other variables remain unchanged.

Variable importance was determined using the training dataset for the logistic regression model (Table 20). The variable importance for logistic regression is based on the absolute values of the z-statistic and will show both the most and least influential predictors. The variables with the highest importance for retained students was the number of credit hours with a value of .4218 (SD = .0058), the interaction between

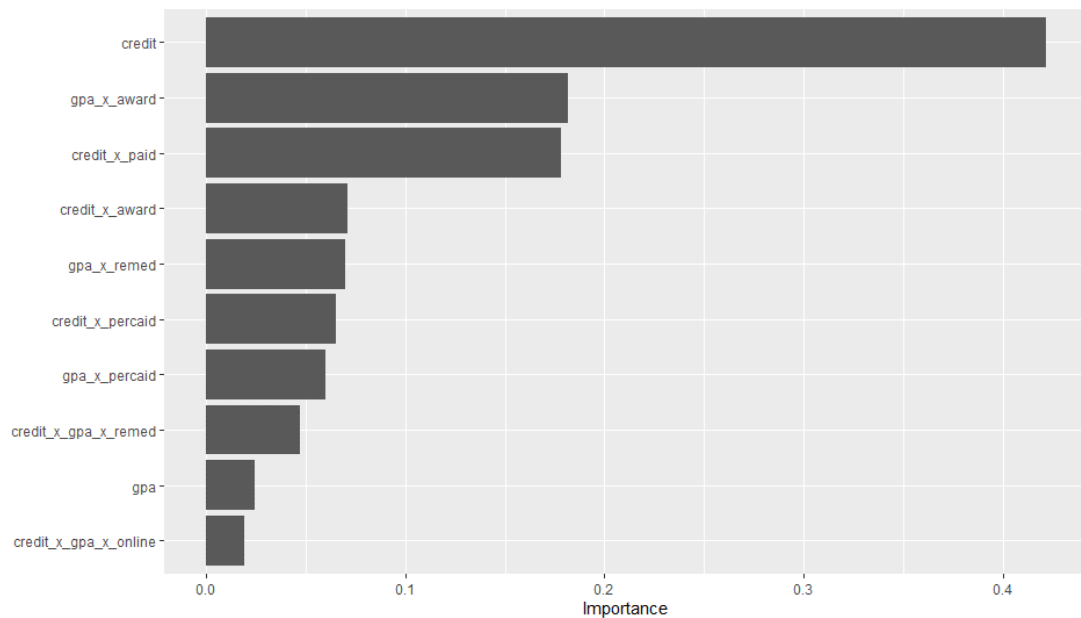
Table 20

*Variable Importance for Logistic Regression Final Model with Training Data*

Variable	Importance	SD
credit	0.421836618	0.005844103
gpa_x_award	0.18148342	0.006049901
credit_x_paid	0.178060272	0.004353836
credit_x_award	0.070840407	0.003821329
gpa_x_remed	0.069574219	0.002637632
credit_x_percaid	0.065257325	0.002118494
gpa_x_percaid	0.059773039	0.00262367
credit_x_gpa_x_remed	0.047014496	0.001582332
gpa	0.024296152	0.001662356
credit_x_gpa_x_online	0.019392847	0.001894772
gpa_x_paid	0.017801919	0.001590386
credit_x_online	0.016573625	0.001518067
percaid	0.012811559	0.001439412
race_x2	0.006699701	0.000679063
race_x1	0.005654569	0.000825256
online	0.005183641	0.000687316
credit_x_online_x_remed	0.004874785	0.000543576
gpa_x_online	0.004809781	0.000923209
credit_x_gpa	0.003895379	0.000775032
award	0.003077368	0.000650027
credit_x_remed	0.00124825	0.000325769
age	0.000499295	0.000352438
paid	0.000189265	0.000131387
hsgpa	0.000186119	0.000090171
remed	0.000163583	0.000110422
gender_x1	0.000029077	0.000056682
race_x3	0.000003218	0.000052595
fasfa_x1	-0.000000821	0.000028050

age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = college GPA for the first three semesters. percaid = percentage of financial aid used by a student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters. race\_x1 = Black or African American students. race\_x2 = Hispanic or Latino students. race\_x3 = other students. fasfa\_x1 = no FASFA completion. gender\_x1 = female students. The x indicates an interaction between two or three different variables.

the GPA and the amount of financial aid awarded (.1815 (SD = .0060)), and the interaction between credit hours and the amount of financial assistance paid (.1781 (SD = .0044)) (Figure 19). Another group of factors important to retention is the interaction between credit hours and the amount of financial assistance awarded (.0708 (SD = .0038)), the interaction between the GPA and the number of remedial courses (.0696 (SD = .0026)) the interaction between credit hours and the percentage of financial assistance used (.0653 (SD = .0021)), and the interaction between GPA and the percentage of financial aid used (.0598 (SD = .0026)). Additionally, the interaction between credit hours, GPA, and the number of remedial courses (.0470 (SD = .0016)) and the GPA for the first three semesters (.02430 (SD = .0016)) were also identified as significant to the retention of students during their first year. There were no critical variables for

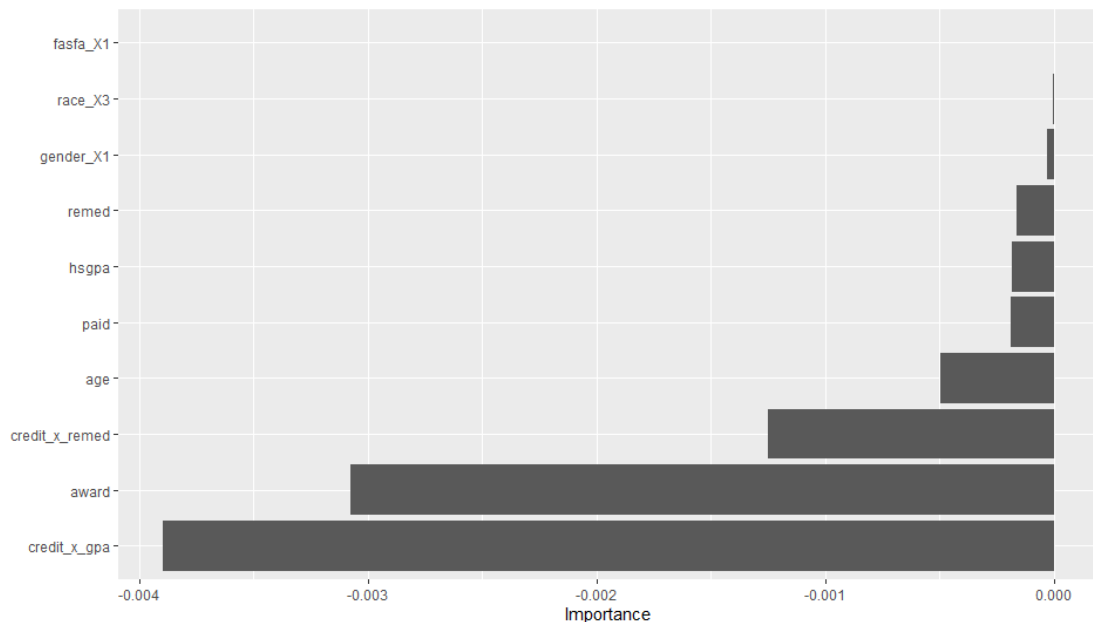


*Figure 19.* Retention variable importance plot for logistic regression model using the training data. This plot displays variable importance from highest to lowest order with the variable names located on the right side and the length of the bar showing the level of importance from left to right.



nonretention of first-year students using the logistic regression (Figure 20).

The variables identified as significant in both the logistic regression model and the variable importance plots were credit hours and the interaction between GPA and the number of remedial courses. The model also identified variables that were not as significant and were also identified in the variable importance plots: the interaction between GPA and amount of financial aid awarded, the interaction between credit hours and amount of financial aid awarded, the interaction between credit hours and percentage of financial assistance used, the interaction between credit hours, GPA, and the number of remedial courses, and GPA. The variable importance plot did not identify any background variables as significant, where the model identified the variable for Hispanic or Latino students as being influential to retention. The variable importance plot identified the interaction of credit hours and amount of financial aid paid and the



*Figure 20.* Nonretention variable importance plot for logistic regression model using the training data. This plot displays variable importance from highest to lowest order with the variable names located on the right side and the length of the bar showing the level of importance from left to right. No variable was identified as important to nonretention in this model.

interaction of GPA and the percentage of financial assistance used as important variables, but these interactions were not significant in the model.

The logistic regression model results for the test data set are listed in Table 21. The Hosmer and Lemeshow goodness of fit test results ( $\chi^2(8) = 30.48, p < 0.001$ ) were significant, indicating that the overall model does not fit the data well. The Nagelkerke's pseudo R squared value for the test data set was 0.375 and was 2.9% higher than the training data set. McFadden's pseudo R squared value had a value of 0.239, which was less than a 1% increase in the overall model performance than the training data set.

Nine of the 28 predictors were statistically significant in predicting retention using the test data set: number of credit hours, female students, Hispanic or Latino students, the interaction between the number of credit hours and amount of financial aid awarded to the student, the interaction between the number of credit hours and the amount of financial assistance paid to the student, the interaction between the number of credit hours and the percentage of financial aid used by the student, the interaction between the number of credit hours, GPA, and the number of remedial courses taken, the interaction between the number of credit hours, percentage of online courses taken and the number of remedial classes taken, and the interaction between GPA and the number of remedial courses taken. The other nineteen variables were not significant to the model for predicting retention. Seven of the variables were previously identified as important in the training data model. Students who take more credit hours in their first three semesters are 31.887 times greater for being retained opposed to students who do not ( $z = 3.462, p < .001, \text{odds ratio} = 31.887, 95\% \text{ CI} = 12.059 \text{ to } 86.182$ ). The interaction with GPA and number of remedial courses ( $z = 6.032, p < .001, \text{odds ratio} = 4.455, 95\% \text{ CI} = 2.748 \text{ to}$

Table 21

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

Variable	Log Odds	SE	Z	Pr(> z )		95% Confidence Interval		
						OR	Lower	Upper
(Intercept)	2.678	0.874	3.064	0.002	**			
age	0.017	0.041	0.419	0.675		1.017	0.939	1.104
hsgpa	-0.016	0.054	-0.298	0.766		0.984	0.885	1.093
credit	3.462	0.501	6.906	p < .001	***	31.887	12.059	86.182
paid	-0.904	0.511	-1.771	0.077		0.405	0.149	1.107
award	0.674	0.393	1.715	0.086		1.961	0.909	4.241
remed	-0.103	0.172	-0.597	0.551		0.902	0.644	1.268
online	-0.187	0.179	-1.046	0.296		0.829	0.579	1.167
gpa	-0.539	0.422	-1.281	0.201		0.583	0.255	1.331
percaid	0.537	0.334	1.609	0.108		1.711	0.892	3.301
gender_X1	-0.182	0.079	-2.284	0.022	*	0.834	0.713	0.974
race_X1	-0.181	0.103	-1.751	0.082		0.835	0.682	1.022
race_X2	0.425	0.129	3.282	0.001	**	1.529	1.189	1.976
race_X3	0.023	0.172	0.131	0.896		1.029	0.731	1.437
fasfa_X1	-0.234	0.158	-1.483	0.138		0.792	0.581	1.077
credit_x_gpa	-0.008	0.009	-0.842	0.399		0.992	0.974	1.011
credit_x_online	0.009	0.019	0.509	0.611		1.009	0.973	1.048
credit_x_award	-0.001	0.001	-2.959	0.003	**	0.999	0.998	1.001
credit_x_paid	0.002	0.001	2.111	0.035	*	1.001	0.999	1.002
credit_x_percaid	-0.002	0.004	-2.896	0.004	***	0.999	0.998	1.001
credit_x_remed	-0.033	0.056	-0.587	0.557		0.968	0.867	1.079
credit_x_gpa_x_online	-0.004	0.006	-0.763	0.445		0.996	0.985	1.007
credit_x_gpa_x_remed	-0.085	0.018	-4.703	p < .001	***	0.919	0.887	0.952
credit_x_online_x_remed	0.029	0.009	3.059	0.002	**	1.031	1.012	1.049
gpa_x_paid	0.003	0.003	1.058	0.291		1.003	0.997	1.009
gpa_x_online	0.049	0.076	0.655	0.513		1.051	0.906	1.122
gpa_x_award	0.004	0.001	0.104	0.917		1.001	0.997	1.003
gpa_x_percaid	0.001	0.002	-0.036	0.972		0.999	0.996	1.003
gpa_x_remed	1.494	0.248	6.032	p < .001	***	4.455	2.748	7.261

Note. AIC: 4255.225. p < 0.001 ‘\*\*\*’, p < 0.01 ‘\*\*’, p < 0.05 ‘\*’

age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = College GPA for the first three semesters. percaid = percentage of financial aid paid divided by the amount of financial aid awarded to the student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters. race\_x1 = Black or African American students. race\_x2 = Hispanic or Latino students. race\_x3 = Other students. fasfa\_x1 = no FASFA completion. gender\_x1 = female students. The\_x indicates an interaction between two or three different variables.

7.261) 4.425 times odds of students being retained. Hispanic or Latino students had ( $z = 3.282$ ,  $p = .001$ , odds ratio = 1.529, 95% CI = 1.189 to 1.976) had 1.529 odds of being retained.

The logistic regression model used the test data set to calculate the importance of the variables (Table 22). The variable with the highest importance for retained students was the number of credit hours with a value of .4240 (SD = .0039) (Figure 21). The next two variables with similar significant values to retention were the interaction between the GPA and the amount of financial aid awarded (.1980 (SD = .0041)) and the interaction between credit hours and the amount of financial aid paid (.1701 (SD = .0036)). Another group of factors important to retention is the interaction between GPA and the percentage of financial assistance used (.0837 (SD = .0033)), the interaction between GPA and the number of remedial courses (.07235 (SD = .0041)), the interaction between credit hours and the percentage of financial assistance used (.0710 (SD = .00382)), and the interaction between credit hours and the amount of financial aid awarded (.0698 (SD = .0042)). Additionally, the interaction between credit hours, GPA, and the number of remedial courses (.0467 (SD = .0018)), the GPA for the first three semesters (.0277 (SD = .0018)), the interaction between GPA and the amount of financial aid paid (.0225 (SD = .0024)), and the interaction between credit hours, GPA, and percentage of online courses (.0206 (SD = .0016)) were also identified as significant to the retention of students during their first year. There were no significant variables for first year students' nonretention using the logistic regression (Figure 22).

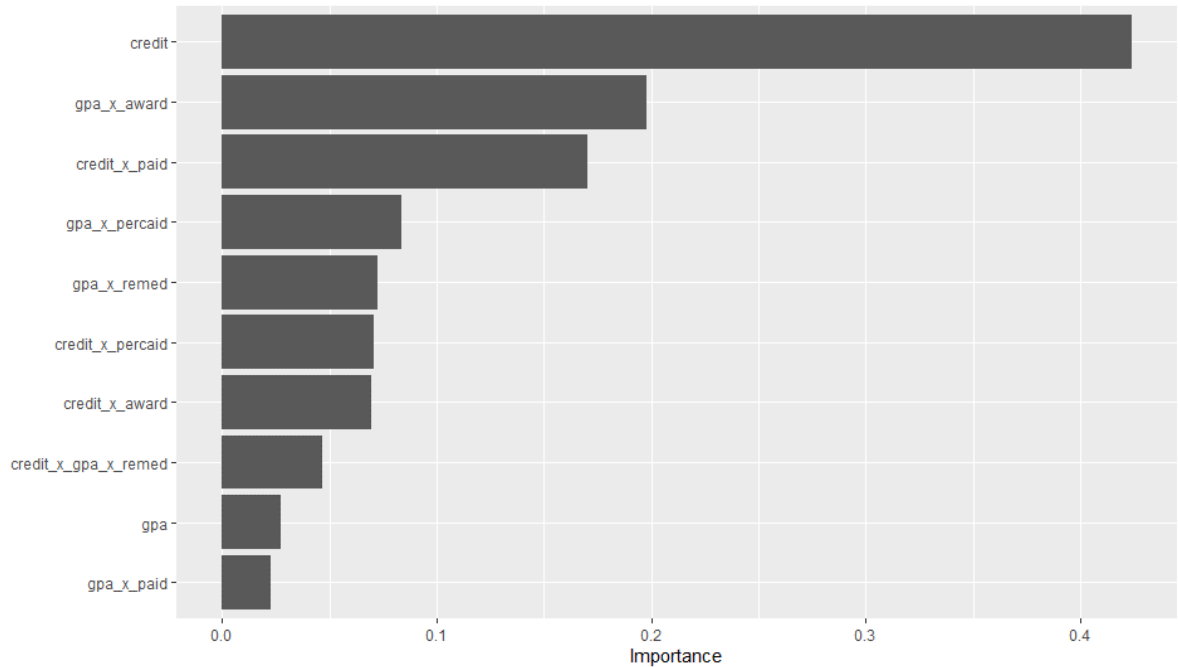
The variables identified as significant in both the logistic regression model and the variable importance plots for the testing data were credit hours and the interaction

Table 22

*Variable Importance for Logistic Regression Final Model with Test Data*

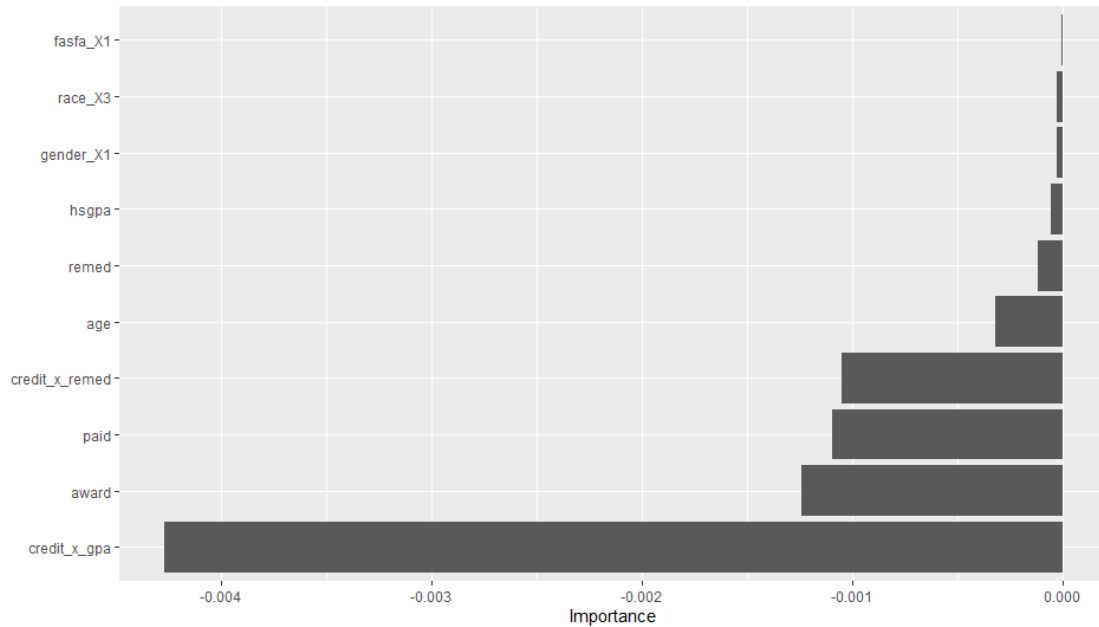
Variable	Importance	SD
credit	0.424030104	0.003984104
gpa_x_award	0.198033134	0.00413566
credit_x_paid	0.170145344	0.003618471
gpa_x_percaid	0.083690842	0.003314372
gpa_x_remed	0.07253872	0.004124679
credit_x_percaid	0.070994274	0.003193877
credit_x_award	0.069849065	0.004274963
credit_x_gpa_x_remed	0.046780507	0.001787635
gpa	0.027689384	0.001786602
gpa_x_paid	0.022528851	0.002409847
credit_x_gpa_x_online	0.020585131	0.001587576
credit_x_online	0.018719984	0.001609185
percaid	0.011199443	0.001210464
race_x2	0.006236796	0.000958282
gpa_x_online	0.005532292	0.000850328
online	0.005273075	0.000638466
race_x1	0.004836372	0.001107719
credit_x_online_x_remed	0.004478562	0.000531005
credit_x_gpa	0.004268071	0.001070734
award	0.001241196	0.000293454
paid	0.001092671	0.000308525
credit_x_remed	0.001048114	0.000439013
age	0.000318359	0.000204609
remed	0.0001187	0.0001109
hsgpa	0.0000539	0.0001244
gender_x1	0.0000257	0.0000620
race_x3	0.0000248	0.0000491
fasfa_x1	-0.0000030	0.0000348

age = age in years. hsgpa = high school GPA. credit = amount of credit hours taken in the first three semesters. remed = number of remedial courses taken in the first three semesters. online = percentage of online courses taken in the first three semesters. gpa = college GPA for the first three semesters. percaid = percentage of financial aid used by the student in their first three semesters. paid = amount of financial aid paid to the student in their first three semesters. award = amount of financial aid awarded to the student in their first three semesters. race\_x1 = Black or African American students. race\_x2 = Hispanic or Latino students. race\_x3 = other students. fasfa\_x1 = no FASFA completion. gender\_x1 = female students. The x indicates an interaction between two or three different variables.



*Figure 21.* Retention variable importance plot for logistic regression model using the test data. This plot displays variable importance from highest to lowest, with the variable names located on the right side and the length of the bar showing the level of importance from left to right.

between GPA and the number of remedial courses. The model also identified variables that were not as significant but were also identified in the variable importance plots: the interaction between credit hours and amount of financial aid paid, the interaction between GPA and the number of remedial courses taken, the interaction between credit hours and percentage of financial assistance used the interaction between credit hours and the amount of financial aid awarded, and the interaction between credit hours, GPA, and the number of remedial courses. The variable importance plot did not identify any background variables, including race or ethnicity, as significant, where the model identified the variable for Hispanic or Latino students being critical to retention. The variable importance model identified the interaction between GPA and amount of



*Figure 22.* Nonretention variable importance plot for logistic regression model using the test data. This plot displays variable importance from highest to lowest order with the variable names located on the right side and the length of the bar showing the level of importance from left to right. No variable was identified as important to nonretention in this model.

financial aid awarded, and the interaction of GPA and percentage of financial assistance used, GPA, the interaction between GPA and amount of financial assistance paid, and the interaction between the number of credit hours, GPA, and percentage of online courses but these interactions were not significant in the model. The significant variables through the model and the variable importance plots are primarily the academic and financial variables except for Hispanic or Latino students.

### **Comparison of Variable Importance**

Several patterns emerged among the variable importance plots for the retention of students in the first year. The number of credit hours by itself and its interactions were consistent in the models as an important predictor of retention. The number of credit hours was consistently the most critical variable in retention for four of the models. These

interactions between credit hours and financial variables (amount of financial aid awarded, amount of financial aid paid, and percentage of financial aid used) were found in all the models except for random forest. The number of credit hours and financial assistance plays a role in student retention in their first year. Three of the models (random forest, SVM with radial kernel, and neural network) identified the importance of the interaction between the number of credit hours and the first three semesters' GPA. Other interactions between credit hours, GPA, and the percentage of online courses or the number of remedial classes were identified as important to retention.

Four of the models identified the interactions between GPA and financial variables (amount of financial aid awarded, financial aid amount paid, and percentage of financial assistance used) as critical to retention. Additionally, GPA and the number of remedial courses were significant as well. The random forest model identified high school GPA and Black or African American students as crucial to retention and was the only model to identify a background factor as significant to retention. All the academic and financial variables play a role in the retention of first year students.

There were no consistent variables that can predict students' nonretention in the first year of their college career. The random forest and logistic regression models did not identify any variables for nonretention. Many background predictors (age, gender, race or ethnicity) were not significant in predicting retained or nonretained students.

## **Research Question 2**

2: Does one of the data mining models (random forests, support vector machines, neural networks, or logistic regression) generate a more accurate classifier performance overall based on the evaluation metrics of accuracy, sensitivity,



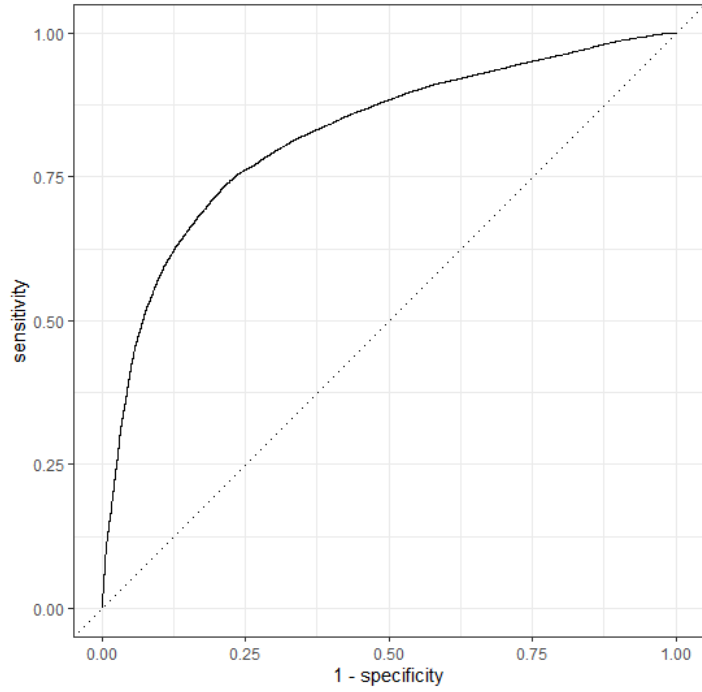
specificity, area under the curve (ROC\_AUC), and F1-values scores?

All five models had ROC curves, confusion matrices, and evaluation metrics (accuracy, ROC\_AUC, specificity, sensitivity, and F1-measures) for the training and test data sets. The ROC curve evaluates class probabilities for the model across multiple thresholds and plots the specificity (the rate of the true positives) on the Y-axis and 1-specificity (the rate of false alarms) on the X-axis (Attewell & Monaghan, 2015; Knowles, 2015; Kuhn & Johnson, 2013). This visual representation, a curve, shows the benefit and cost of the model's classification by the percentage of correctly classified observations and false alarm rates (Attewell & Monaghan, 2015; Knowles, 2015). The ROC curves for this study show the correctly identified nonretained students along the y-axis compared to the students who were incorrectly identified as nonretained when they were retained (Knowles, 2015). The area under the curve is calculated and reported as the ROC\_AUC score. Models with ROC\_AUC values near to 1 have a higher level of accuracy for predicting the correct outcomes, while models with ROC\_AUC values near .5 are not accurate in their predictive abilities since it is considered no better than chance (Attewell & Monaghan, 2015).

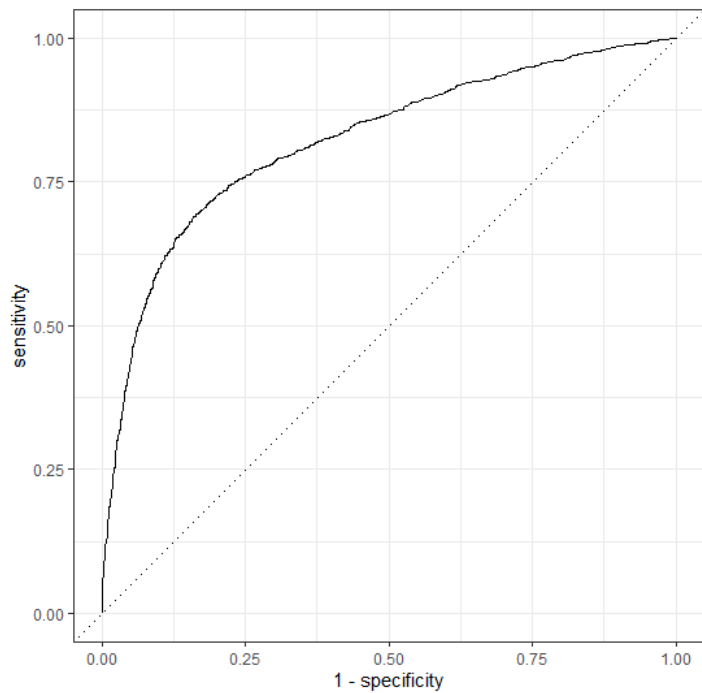
The confusion matrix predicts the probabilities of events occurring based on the specific model using four groups categorized based on the outcomes: the actual event values based on the positive and negative outcomes and the predicted event values based on the positive and negative outcomes (Attewell & Monaghan, 2015; James et al., 2013; Knowles, 2015). The confusion matrix also provides the Type I (false positives) and Type II (false negatives) errors given by each classification model.

Other types of evaluation metrics are calculated from the confusion matrix for the training and test data set. The accuracy measurement of the model represents the percentage of correctly identified true positives and true negatives divided by all the positives and negatives. In this study, accuracy signifies the correctly identified students as retained and nonretained over all the students. Specificity is the measurement of the true negatives divided by the true negative and false positives. This measurement calculates the correctly classified students as nonretained divided by the students classified as nonretained (correct and incorrect). The measure for sensitivity explores the percentage of true positives divided by the true positives and false negatives. Sensitivity calculates the students who are correctly classified as retained divided by students classified as retained (correct and incorrect). Both sensitivity and specificity can help identify performance-based true positives and negatives among the five models. F1-scores measure the "harmonic mean of precision and recall and maybe a better measure of false positive and negative events (Huilgol, 2019; Zhang, Wang & Zhao, 2015).

**Random Forest.** The ROC curves show both data set has a similar shape with little difference (Figure 23 and 24). The random forest model had a ROC\_AUC value of .823 for the training and .821 for the test data set. These values are higher than the other models in either training or test stages. The confusion matrix and ROC curve for the training data set using the final random forest (Table 23) set shows a true-positive rate (sensitivity) of .697 and a false-positive rate of .182. The confusion matrix and ROC curve for the final random forest model using the test data set shows a true-positive rate (sensitivity) of .707 and a false-positive rate of .182. The random forest model did not have a high rate of predicting students who would be nonretained but had a low rate for



*Figure 23.* ROC curve results for the training data set using the final Random Forest model. The area under the curve: 0.823.



*Figure 24.* ROC curve results for the test data set using the final Random Forest model. The area under the curve: 0.821.

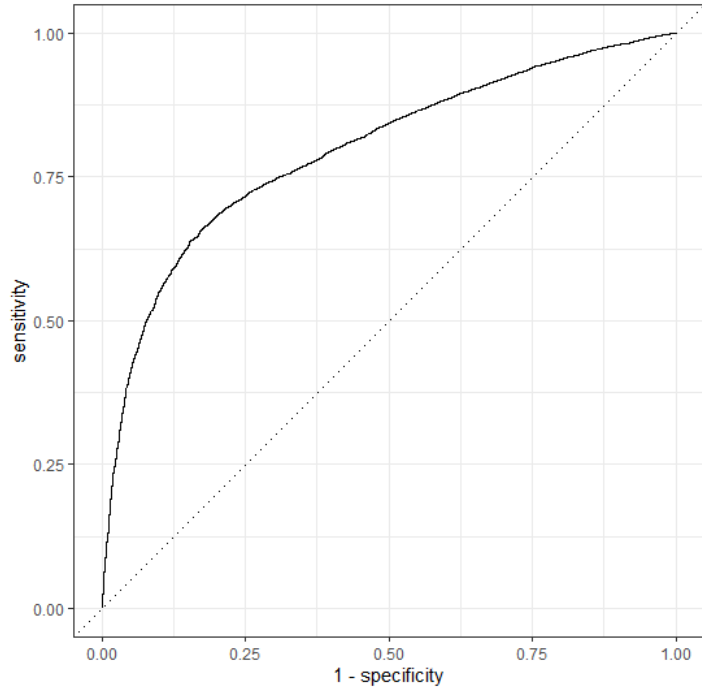
Table 23

*Confusion Matrix Results for the Test Data Set using the Final Random Forest Model*

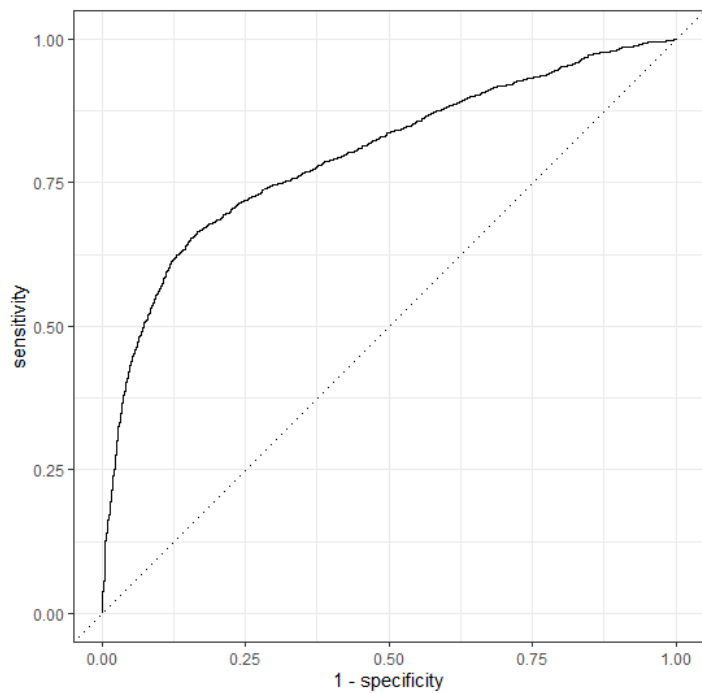
	Actual Nonretained	Actual Retained
Predicted Nonretained	1332	381
Predicted Retained	552	1720

misidentifying students as nonretained when they were retained. This model had an accuracy rate of .766, which was higher than any other final model. The F1 value of .741 was higher than the other models and could mean the random forest model is better for identifying false positives and negatives.

**Support Vector Machine with Polynomial Kernel.** The SVM with the polynomial kernel had a ROC\_AUC value of .796 for the training and .797 for the test data set. Again, these values are lower than the random forest model but aligned with the SVM values with the radial kernel for both data sets. With the ROC\_AUC values being similar, it is not surprising that the ROC curve has a similar shape for both data sets (Figure 25 and 26). The confusion matrix and ROC curve for the training data set using the final random forest shows a true-positive rate (sensitivity) of .591 and a false-positive rate of .124. The confusion matrix for the test data set using the final SVM with polynomial kernel model (Table 24) set shows a true-positive rate (sensitivity) of .602 and a false-positive rate of .118 lowest of all the models. This model had the highest specificity value of any model in the training and test data set. The model has a low rate of predicting students who would be nonretained. This model had an accuracy rate of .750, which was higher than SVM with the radial kernel model and would not be the best model for predicting retention. The F1 value of .695 was higher than the other



*Figure 25.* ROC curve results for the training data set using the final SVM with polynomial kernel model. The area under the curve: 0.796.



*Figure 26.* ROC curve results for the test data set using the final SVM with polynomial kernel model. The area under the curve: 0.797.

Table 24

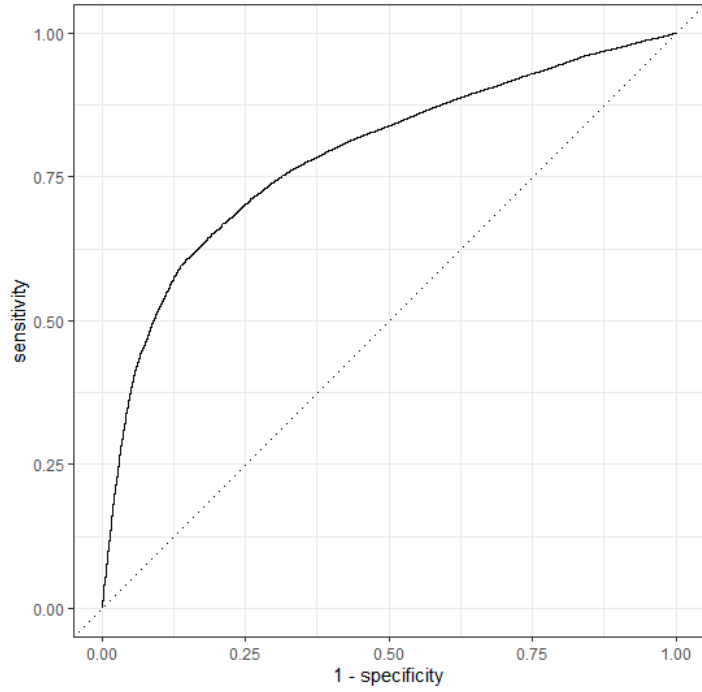
*Confusion Matrix Results for the Test Data Set using the Final SVM with Polynomial Kernel Model*

	Actual Nonretained	Actual Retained
Predicted Nonretained	1135	248
Predicted Retained	749	1853

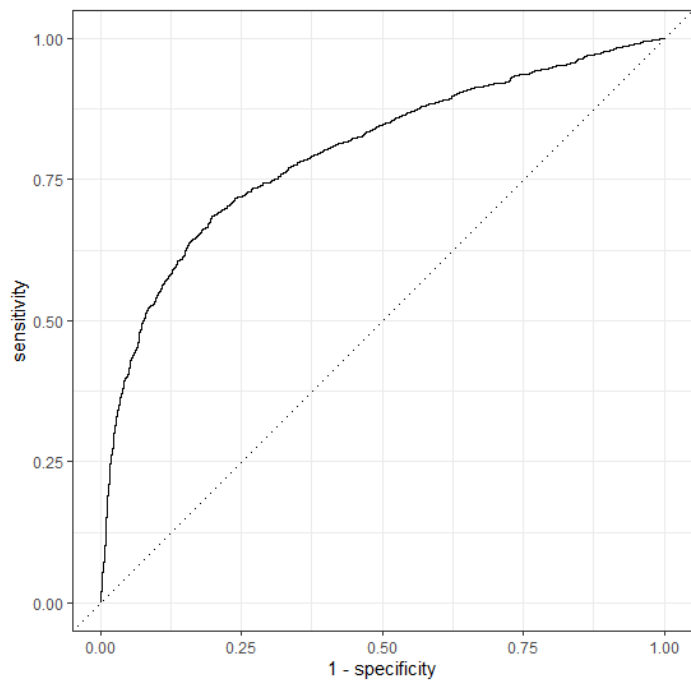
SVM model but lower than the other three models.

**Support Vector Machine with Radial Kernel.** The SVM with the radial kernel had a ROC\_AUC value of .795 for the training and .797 for the test data set. These values are lower than the random forest model but were close to the SVM with the polynomial kernel for both data sets. With the ROC\_AUC values being similar, it is not surprising that the ROC curve has a similar shape for both data sets (Figure 27 and 28). This model's ROC curve for the training data set shows a true-positive rate (sensitivity) of .585 and a false-positive rate of .132.

The confusion matrix for the test data set using the final SVM with radial kernel model (Table 25) shows a true-positive rate (sensitivity) of .592 and a false-positive rate of .132. The model has a low rate of predicting students who would be nonretained and has the lowest rate for all the models. This model had an accuracy rate of .738, which was the lowest of any of the other final models and would not be the best model for predicting retention. The F1 value of .681 was the lowest of the other models and could mean this model is not ideal for identifying false positives and negatives.



*Figure 27.* ROC curve results for the training data set using the final SVM with radial kernel model. The area under the curve: 0.795.



*Figure 28.* ROC curve results for the test data set using the final SVM with radial kernel model. The area under the curve: 0.797.

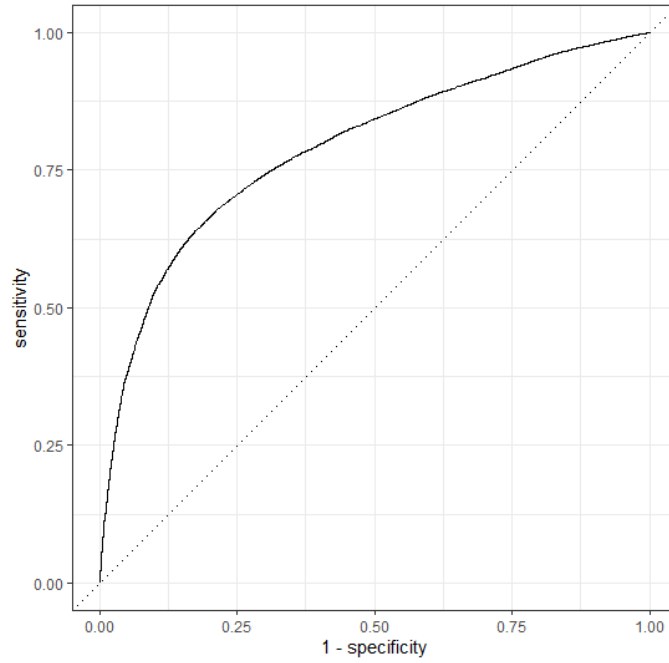
Table 25

*Confusion Matrix Results for the Test Data Set using the Final SVM with Radial Kernel Model*

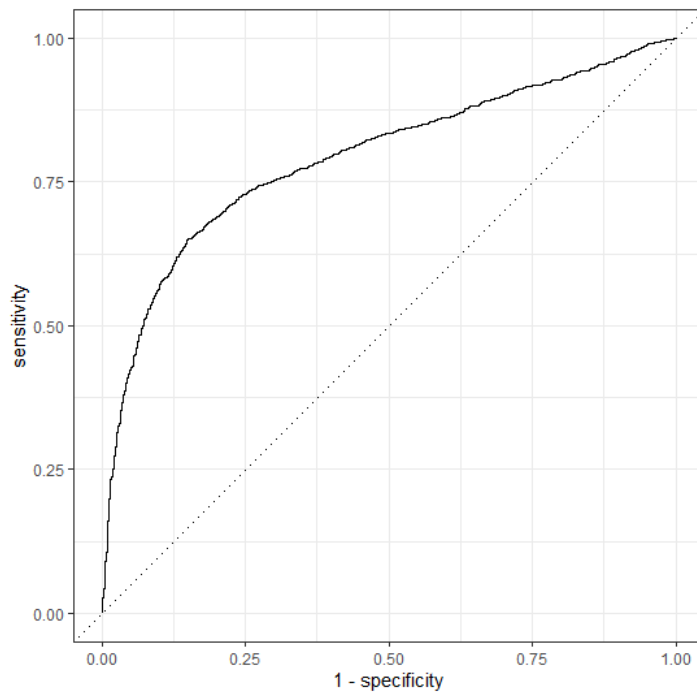
	Actual Nonretained	Actual Retained
Predicted Nonretained	1116	277
Predicted Retained	768	1824

**Neural Network.** The neural network model had a ROC\_AUC value of .807 for the training and .791 for the test data set. While these values are lower than the random forest model, the test data set's value was lower than any other model. There was a difference in the training and test modeling phase scores and is seen in the smaller area for the test data set in the ROC curve (Figure 29 and 30). The ROC curve of this model for the training data set shows a true-positive rate (sensitivity) of .670 and a false-positive rate of .178. The confusion matrix for the test data set using the final neural network model (Table 26) set shows a true-positive rate (sensitivity) of .650 and a false-positive rate of .151. The neural network model's sensitivity decreased from the training to the test set while the false positive rate decreased. This model had an accuracy rate of .755 and an F1 value of .715, which was higher than all the other models but the random forest model.





*Figure 29.* ROC curve results for the training data set using the final neural network model. Area under the curve: 0.807.



*Figure 30.* ROC curve results for the test data set using the final neural network model. Area under the curve: 0.791.

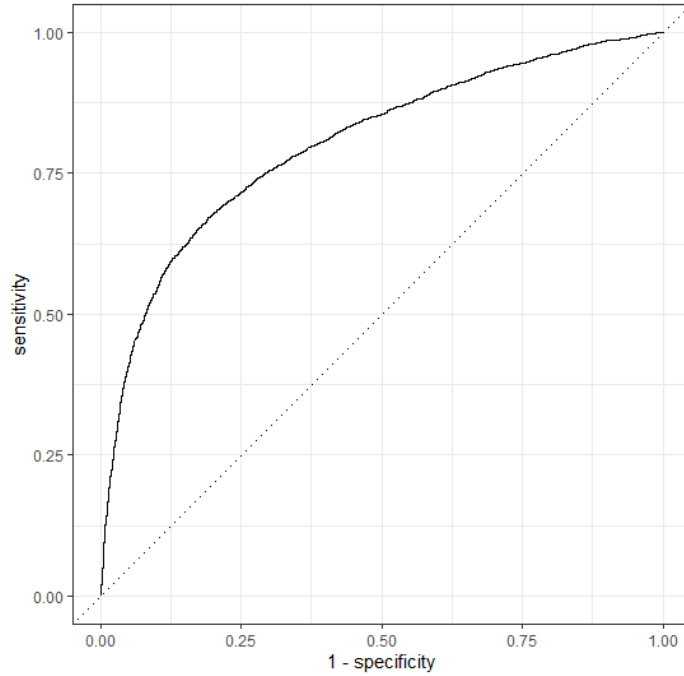
Table 26

*Confusion Matrix Results for the Test Data Set using the Final Neural Network Model*

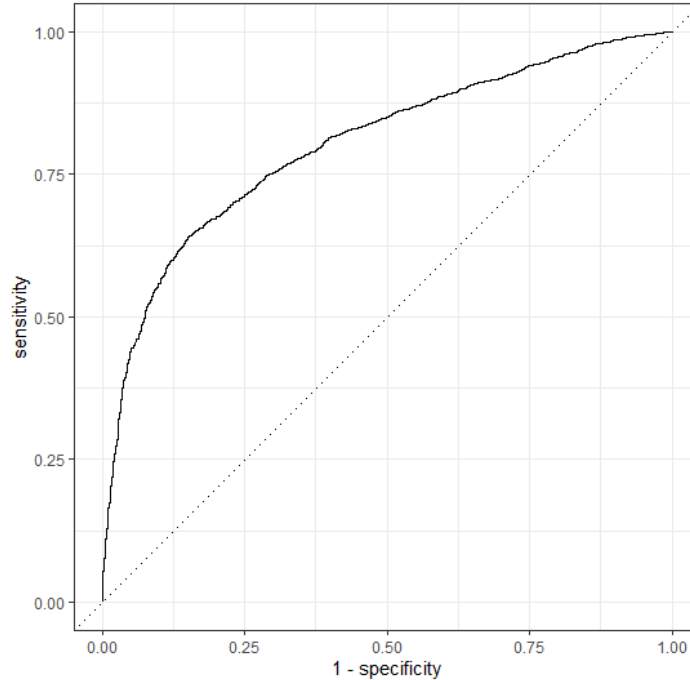
	Actual Nonretained	Actual Retained
Predicted Nonretained	1224	317
Predicted Retained	660	1784

**Logistic Regression.** The logistic regression model had a ROC\_AUC value of .802 for the training and had the same value for the test data set. There was no tuning with the logistic regression model, so the evaluation metrics' changes were due to the data. These values are lower than the random forest model but higher than the other models. The logistic regression model ROC\_AUC values were the same, and the ROC curves had the same shape from both the training and test data set (Figures 31 and 32).

The confusion matrix and ROC curve for the training data set using the final logistic regression model (Table 27) shows a true-positive rate (sensitivity) of .638 and a false-positive rate of .163. Using the test data set, the final logistic regression model (Table 20) shows a true-positive rate (sensitivity) of .647 and a false-positive rate of .161. This model had the second-lowest specificity value, .839 of the other models in the test data set, and stayed consistent through both modeling phases. This model had an accuracy rate of .748, which was the second-lowest of all the models and, again, might not be the best model for predicting retention. The F1 value of .708 and increased slightly from the training to the test set.



*Figure 31.* ROC curve results for the training data set using the final logistic regression model. The area under the curve: 0.802.



*Figure 32.* ROC curve results for the test data set using the final logistic regression model. The area under the curve: 0.802.

Table 27

*Confusion Matrix Results for the Test Data Set using the Final SVM with Polynomial Kernel Model*

	Actual Nonretained	Actual Retained
Predicted Nonretained	1218	338
Predicted Retained	666	1763

### **Overall Model Comparison with ROC Curves**

The individual ROC curves for the training data were combined to compare the models (Figure 33). The random forest model has a slightly higher curve than the other models. The other models were grouped and are consistent with the values described before. The individual ROC curves for the test data were combined to compare the models (Figure 34) and showed the random forest model having a higher curve than the other models. The graph shows more separation of the curves with a slight increase from the training set with logistic regression.

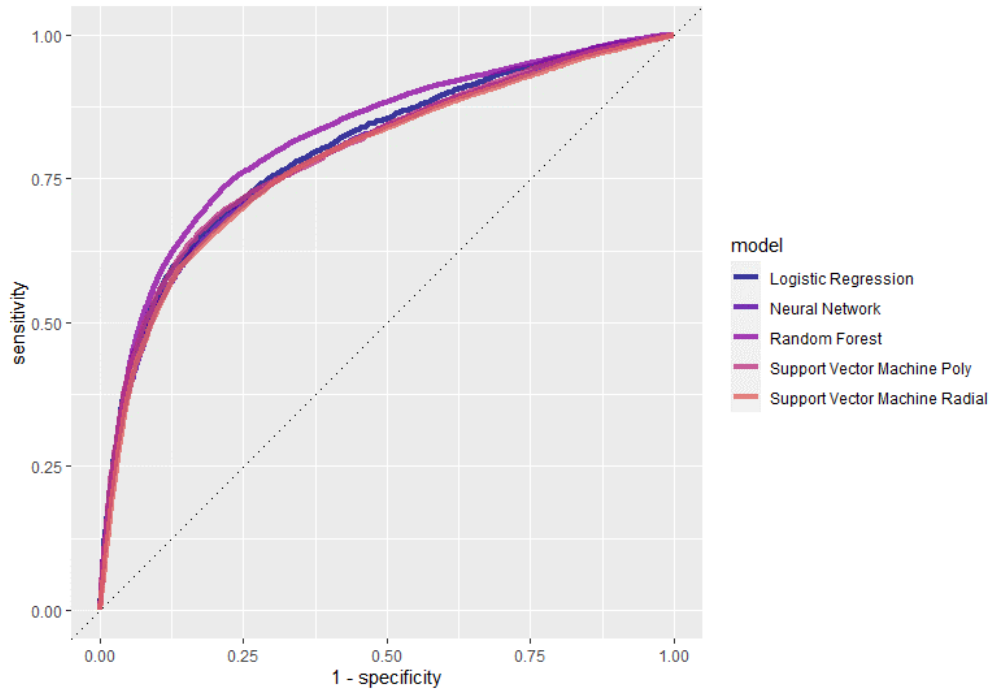


Figure 33. ROC curve results for the training data set with all the final models.

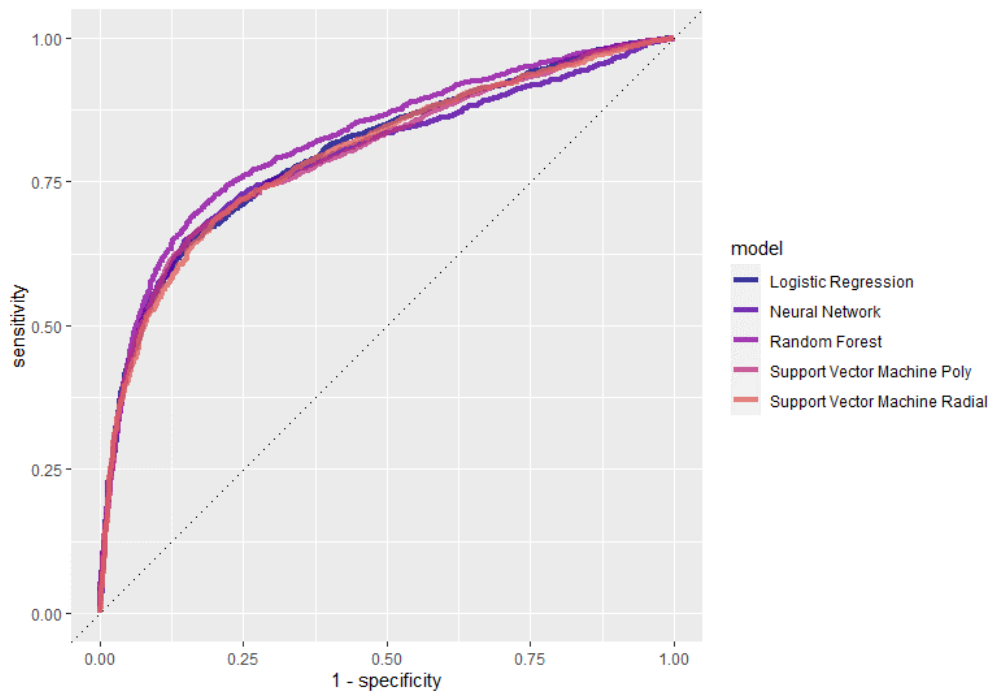


Figure 34. ROC curve results for the test data set with all the final models.

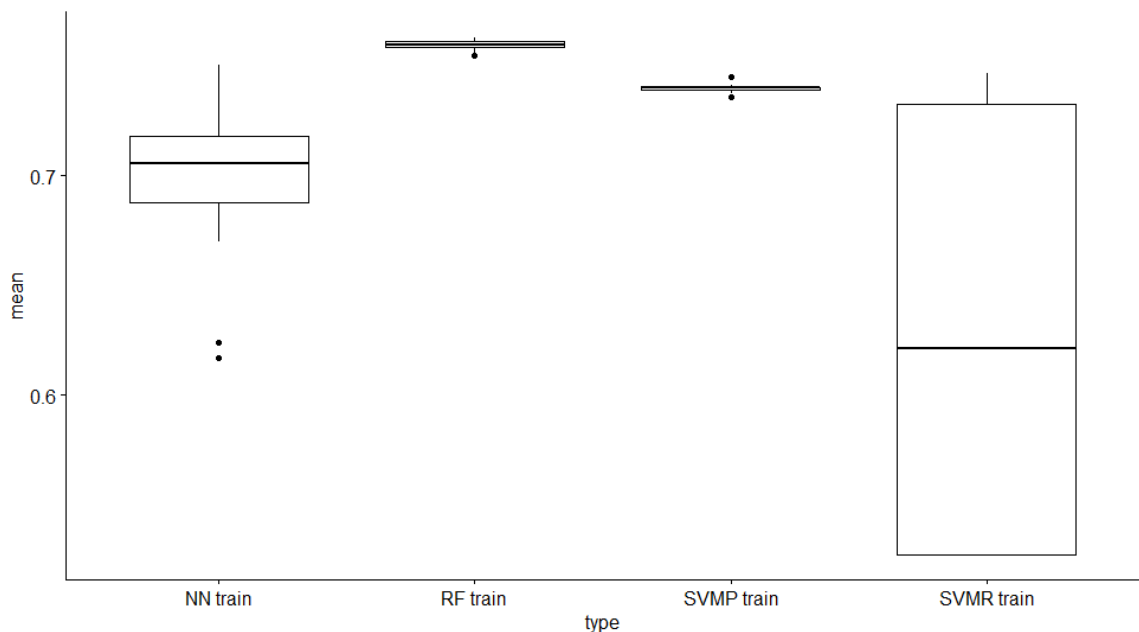
## Inferential Tests for Model Comparison

For the evaluation metrics, a Mann-Whitney U test was used to determine if there was any difference in the models from the training and test data sets to reassure the validity of the results more than visual comparison. For the accuracy metric, there was no difference among the models between the training and test data set,  $U(N_{training} = .743, N_{test} = .750,) = 16.00, p = .531$ . For the F1-values metric, there was no difference among the models between the training and test data set,  $U(N_{training} = .701, N_{test} = .708,) = 14.00, p = .835$ , with all the models underfitting the training data. There was no difference among the models between the training and test data set,  $U(N_{training} = .802, N_{test} = .797,) = 10.00, p = .676$  for the ROC\_AUC scores. For the sensitivity metric, there was no difference among the models between the training and test data set,  $U(N_{training} = .638, N_{test} = .647,) = 15.00, p = .676$ . There was no difference among the models between the training and test data set for the specificity metric,  $U(N_{training} = .837, N_{test} = .849,) = 16.00, p = .531$ .

The second type of inferential test was completed to see differences between the models for each evaluation metric using a Friedman's test (Demšar, 2006, Fernández-Delgado et al., 2014). Except for logistic regression, each model had different runs to evaluate the final models using the training data set. The models were displayed in box plots to show the models' overall shape for each evaluation metric. The effect size for Friedman's test was determined using the Kendall's W coefficient and post hoc pairwise comparisons using paired Wilcoxon signed-rank test with a Bonferroni multiple testing correction method. The analysis was only available for the training set since the models

created with the test data set used the final model and provided one set of evaluation metrics per model.

The evaluation metric, accuracy, was statistically significantly different among the various run of the models by type,  $\chi^2(3) = 48.10$ ,  $p < .001$  with large effect size,  $W = .801$  using a Kendall's  $W$  coefficient with the box plots showing the different distribution between the models (Figure 35). The random forest and support vector machine with polynomial kernel models had a small range but higher mean values than the other models. The neural networks had two outliers with a more extensive range than random forest and the support vector machine with polynomial kernel models. The support vector machine with the radial kernel model had the largest range of the mean values. Post hoc analysis using the Wilcoxon pairwise ranked sign test revealed statistically significant differences in the accuracy values from the various models for random forest ( $Mdn =$



*Figure 35.* Boxplots of four models using the training data and the accuracy evaluation metrics. This plot displays the distribution of each model's accuracy values for the different variations of models created with the training data set. The model type is on the x-axis, and the values for the mean on the y-axis.

0.76) to the support vector machine with polynomial kernel ( $Mdn = 0.74$ ) ( $p < .001$ ) and the support vector machine with radial kernel ( $Mdn = 0.62$ ) ( $p < .001$ ) (Demšar, 2006, Fernández-Delgado et al., 2014). The random forest model was more accurate in predicting retention than both support vector machine models. There were similar results with neural networks ( $Mdn = 0.71$ ) which was statistically significant from the other models; random forest ( $Mdn = 0.76$ ) ( $p < .001$ ) and support vector machine with polynomial kernel ( $Mdn = 0.74$ ) ( $p < .001$ ). The neural network model was less accurate in predicting retention than the random forest and support vector machine with the polynomial kernel. There was a difference between the support vector machine with the polynomial kernel ( $Mdn = 0.74$ ) and the support vector machine with the radial kernel ( $Mdn = 0.62$ ) ( $p < .001$ ). The polynomial kernel outperformed the radial kernel producing the best support vector machine model for accuracy. There was not a significant difference between neural networks and support vector machines with the radial kernel. The ranking of the models from highest to lowest for the accuracy metric is random forest, support vector machine with the polynomial kernel, neural networks, and support vector machine with the radial kernel.

The evaluation metric, F1-Value, was also statistically significantly different among the various run of the models by type,  $\chi^2(3) = 21.2$ ,  $p < .001$  with a large effect size,  $W = .708$  using a Kendall's W coefficient with box plots to show the models' distribution (Figure 36). The neural network model had the largest range of all the models. Both support vector machine models have small ranges with outliers. Again, the random forest model had the highest mean values of all the models. The post hoc analysis revealed statistically significant differences between the random forest models and the



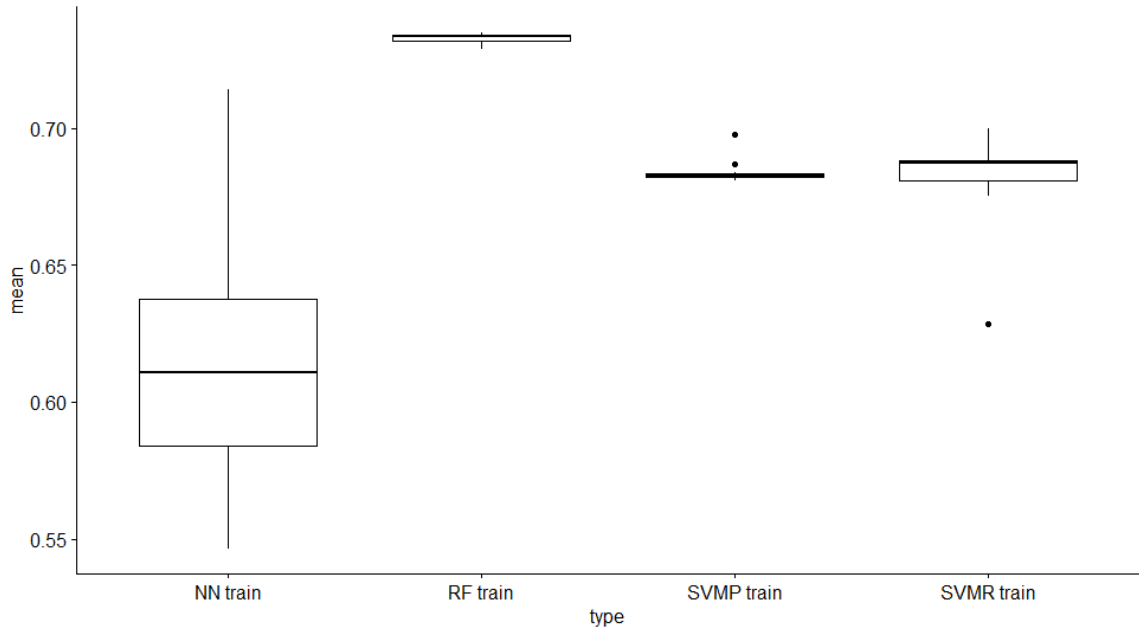
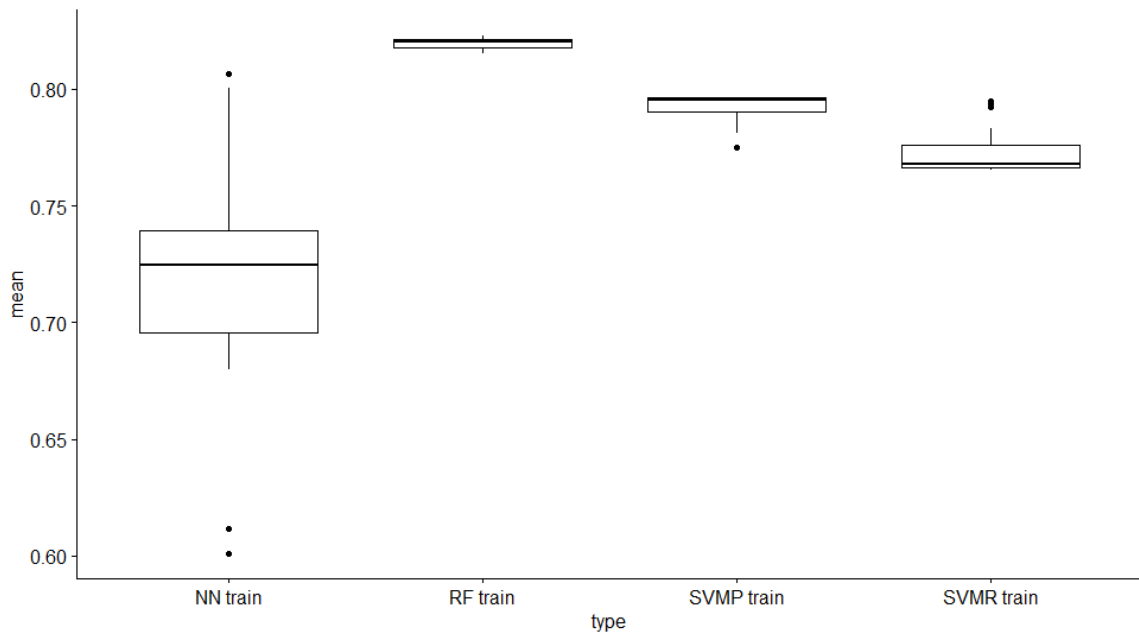


Figure 36. Boxplots of four models using the training data and the f1-value evaluation metrics. This plot displays the distribution of each model's F1-Values for the different variations of models created with the training data set. The model type is on the x-axis, and the values for the mean on the y-axis.

models. The random forest model had the highest F1- Value ( $Mdn = 0.73$ ) than the support vector machine with polynomial kernel ( $Mdn = 0.68$ ) ( $p < .05$ ), the support vector machine with radial kernel ( $Mdn = 0.69$ ) ( $p < .05$ ), and neural network ( $Mdn = 0.61$ ) ( $p < .05$ ). There was not a significant difference between neural networks and the support vector machine models. The ranking of the models from highest to lowest for the F1- Value metric is random forest, support vector machine with the radial kernel, support vector machine with the polynomial kernel, and neural networks. The evaluation metric, ROC\_AUC, was also statistically significantly different among the various run of the models by type,  $\chi^2(3) = 51.40$ ,  $p < .001$  with a large effect size,  $W = .856$  using a Kendall's W coefficient with box plots to show the overall shape within and among the models for accuracy (Figure 37). The neural network model had the largest ranges with

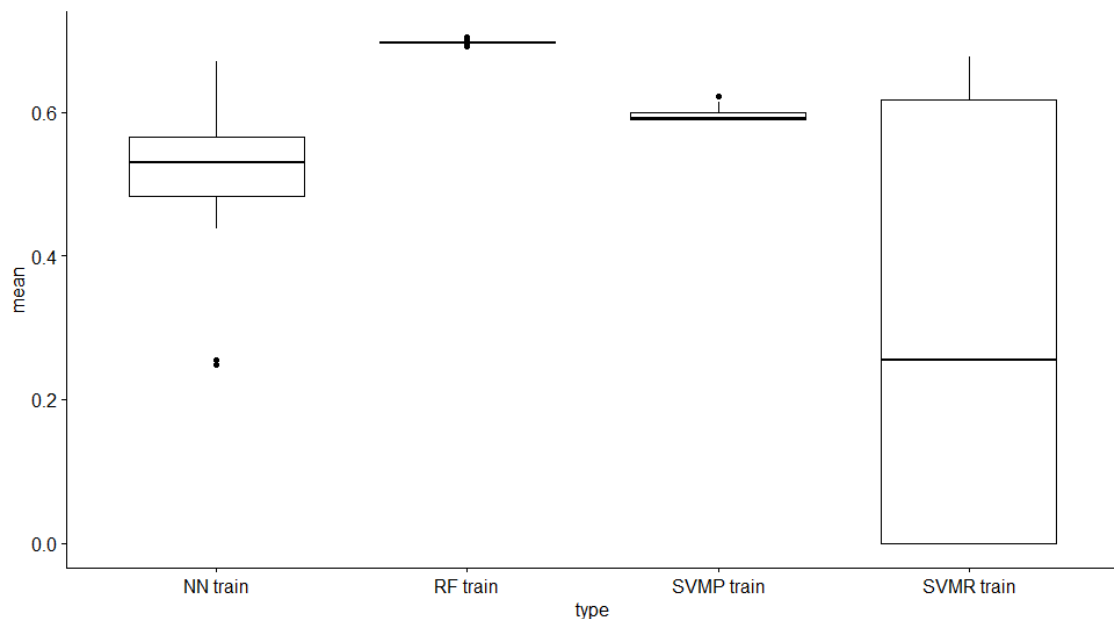
outliers among the models. The random forest model had the highest mean values of all the models. The post hoc analysis revealed statistically significant differences in all the ROC\_AUC values from the various models. The neural network model had the lowest ROC\_AUC median of all the models ( $Mdn = 0.73$ ) compared to the support vector machine with polynomial kernel ( $Mdn = 0.80$ ) ( $p < .001$ ), the support vector machine with radial kernel ( $Mdn = 0.77$ ) ( $p < .001$ ), and random forests ( $Mdn = 0.82$ ) ( $p < .001$ ). The random forest models had the highest ROC\_AUC values ( $Mdn = 0.82$ ) compared to the support vector machine with polynomial kernel ( $Mdn = 0.80$ ) ( $p < .001$ ) the support vector machine with radial kernel ( $Mdn = 0.77$ ) ( $p < .001$ ), and neural network ( $Mdn = 0.73$ ) ( $p < .001$ ). The support vector machine models with the polynomial kernel ( $Mdn = 0.80$ ) had higher ROC\_AUC values than the support vector machine with ROC\_AUC values for all the models, followed by the support vector machine model with the radial



*Figure 37.* Boxplots of four models using the training data and the roc\_auc evaluation metrics. This plot displays the distribution of each model's accuracy values for the different variations of models created with the training data set. The model type is on the x-axis, and the values for the mean on the y-axis.

kernel ( $Mdn = 0.77$ ) ( $p < .001$ ). The random forest model gave the highest with the polynomial kernel, the support vector machine with the radial kernel, and the neural network model.

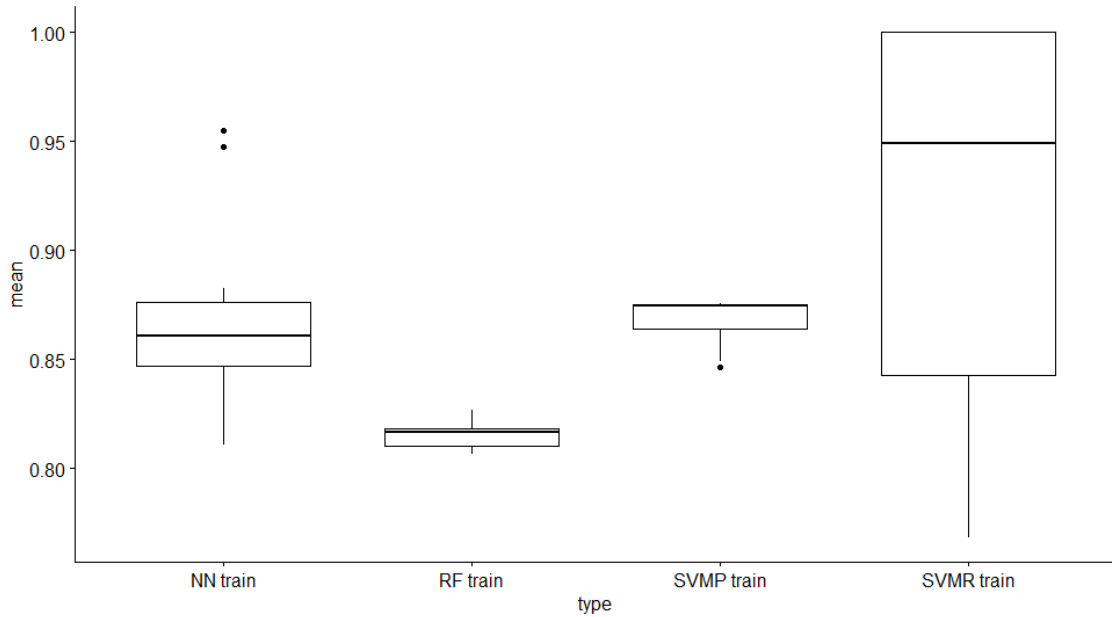
The evaluation metric, sensitivity, was also statistically significantly different among the various run of the models by type,  $\chi^2(3) = 41.50$ ,  $p < .001$  with a large effect size,  $W = .691$  using a Kendall's  $W$  coefficient with box plots to show the different models' distribution (Figure 38). The random forest and support vector machine with polynomial kernel models had a small range, with the random forest model having the highest mean value. The neural networks had two outliers and a larger range than the support vector machine with polynomial kernel and random forest models. The support vector machine with the radial kernel model had the largest range of the mean values among the four models. The post hoc analysis revealed statistically significant differences



*Figure 38.* Boxplots of four models using the training data and the sensitivity evaluation metrics. This plot displays the distribution of each model's sensitivity values for the different variations of models created with the training data set. The model type is on the x-axis, and the values for the mean on the y-axis.

in the highest sensitivity values ( $Mdn = 0.70$ ) compared to the support vector machine with the polynomial kernel ( $Mdn = 0.59$ ) ( $p < .001$ ), the support vector machine with radial kernel between the random forest models and the models again. The random forest model had ( $Mdn = 0.26$ ) ( $p = < .001$ ), and neural network ( $Mdn = 0.53$ ) ( $p < .001$ ). There was a difference between the neural network ( $Mdn = 0.53$ ) and the support vector machine with the polynomial kernel ( $Mdn = 0.59$ ) ( $p < .05$ ). There was no significant difference between neural networks, the support vector machine with the radial kernel, and the support vector machine models. The ranking of the models from highest to lowest for the sensitivity metric is random forest, support vector machine with the polynomial kernel, support vector machine with the radial kernel, and neural networks.

Specificity was also statistically significantly different among the various run of the models by type,  $\chi^2(3) = 31.70$ ,  $p < .001$  with a moderate to large effect size,  $W = .529$  using a Kendall's  $W$  coefficient with box plots to show the distribution among the types of models (Figure 39). The random forest and support vector machine with polynomial kernel models had the smallest ranges. The neural networks had two outliers and had a larger range than the support vector machine with polynomial kernel and random forest models. The support vector machine with the radial kernel model had the largest range of the mean values among the four models and the lowest mean value for specificity. The post hoc analysis revealed statistically significant differences between the random forest models and the models. The random forest model had the lowest specificity metrics ( $Mdn = 0.82$ ) compared to the support vector machine with the polynomial kernel ( $Mdn = 0.88$ ) ( $p < .001$ ), the support vector machine with radial kernel ( $Mdn = 0.95$ ) ( $p < .001$ ), and neural network ( $Mdn = 0.86$ ) ( $p < .001$ ). There was not a significant difference



*Figure 39.* Boxplots of four models using the training data and the specificity evaluation metrics. This plot displays the distribution of each model's specificity values for the different variations of models created with the training data set. The model type is on the x-axis, and the values for the mean on the y-axis.

between neural networks and the support vector machines models. The ranking of the models from highest to lowest for the specificity metric is different from the other evaluation metrics with support vector machine with radial kernel having the highest value. The three remaining models were support vector machine with the polynomial kernel, neural networks, and random forest having the lowest specificity value.

Random forest models produced the highest evaluation metrics for accuracy, F1-Value, ROC\_AUC, and sensitivity. These higher values meant that random forests have the best performance for accurately classifying the students who were nonretained and retained (accuracy and F1-Value), the largest area under the ROC curve (ROC\_AUC), and correctly classifying the students who were retained (sensitivity). The support vector machine with the polynomial kernel had similar results, generating the second highest

evaluation metrics for accuracy, ROC\_AUC, and sensitivity. The support vector machine with the polynomial kernel had the highest value for the specificity metric, indicating the optimal model for predicting true negatives in the confusion matrix. In the retention classification for the models, the true negatives refer to the model accurately classifying the students who were nonretained. All the other evaluation metrics for the support vector machine with the radial kernel were the third or fourth lowest values among the models. The neural network evaluation metrics never ranked higher than the third or fourth place in the models. The actual performance of the models was consistent from the training to the test data set, as seen with the results of the Mann-Whitney U tests.

### **Summary**

Five data mining models (random forests, support vector machines, neural networks, or logistic regression) were created on a training data set, and the optimal model for each type was generated using the highest ROC\_AUC value. This test data set was run through these final models to see if one of them had the highest accurate classifier performance overall based on the evaluation metrics of accuracy, sensitivity, specificity, area under the curve (ROC\_AUC), and f-measure ( $F_1$ ) scores. Along with creating the evaluation metrics, variable importance was determined for the academic, background, and financial factors.

When deciding what academic, background and financial predictors are essential for student retention for first-year students at community colleges, the number of credit hours was consistently the most critical variable in retention. There was also an important relationship between the number of credit hours and financial aid in student retention in

their first year. Another meaningful relationship is a student's GPA for their first three semesters and credit hours. The interaction between GPA and financial aid and the interaction between GPA and the number of remedial hours was also crucial for first year retention. Based on the five models in this analysis, the significant factors for retention would be the number of credit hours and its interactions with GPA during the first year and the student's financial aid. The interaction of GPA with financial assistance and how many remedial courses the student took that first year are also important.

There were no consistent variables that can predict students' nonretention in the first year of their college career. Many background predictors (age, gender, race or ethnicity) were not significant in predicting retained or nonretained students. In addition, the FAFSA's completion had no impact and could be explained by the high number of students completing it. Another variable, the percentage of online courses taken, was not identified as critical to the retention of first-year students.

The five different classification models were compared to see if one would have the highest accurate classifier performance based on the different evaluation metrics. The actual performance of the models was consistent from the training to the test data set, as seen with the results of the Mann-Whitney U tests. Visual inspection of the evaluation metrics shows that the random forest model performed better than the other models in accuracy, F1-values, ROC\_AUC, and sensitivity (Table 28). The Friedman's and Wilcoxon signed-rank tests confirmed that the random forest model did have the best performance for accurately classifying the nonretained and retained students (accuracy and F1-Value), highest ROC\_AUC score, and classifying the students who were retained (sensitivity) compared to the other models. The SVM with polynomial kernel had the

Table 28

*Train and Test Data Set Evaluation Metrics for Classification Models*

Classification Model	Accuracy	F1-Values	ROC AUC	Sensitivity	Specificity
Training Data Set					
Random Forest	.761	.734	.823	.697	.818
SVM Polynomial	.741	.684	.796	.591	.876
SVM Radial	.734	.675	.795	.585	.868
Neural Networks	.750	.717	.807	.670	.822
Logistic Regression	.743	.701	.802	.638	.837
Test Data Set					
Random Forest	.766	.741	.821	.707	.818
SVM Polynomial	.750	.695	.797	.602	.882
SVM Radial	.738	.681	.797	.592	.868
Neural Networks	.755	.715	.791	.650	.849
Logistic Regression	.748	.708	.802	.647	.839

*Note.* ROC\_AUC is ROC Area Under the Curve.

highest value for the specificity (nonretained students) and was supported by the Friedman's and Wilcoxon signed-rank tests. All the evaluation metrics for the support vector machine with the radial kernel and the neural network never ranked higher than second place. Logistic regression was not included in the inference tests but never had the highest rank for any evaluation metrics. Visual inspection of the grouped ROC curves also showed the random forest as the optimal model for the highest accurate classifier for the first-year retention.



## **Chapter V**

### **SUMMARY, DISCUSSION, and CONCLUSIONS**

Community colleges play a vital part in the educational landscape serving the needs of nontraditional students and more than half of the minority students in the country. However, the retention of community college students continues to be a concern as the overall enrollment of students continues to decrease, and half of freshmen students do not return to continue their education (Juszkiewicz, 2020). In addition, community college students may face environmental factors such as employment, family obligations, and financial insecurity affecting their ability to remain enrolled. Specialized retention models can help colleges identify important variables for student retention serving as the basis for new programs and initiatives. Additionally, these models can provide a framework for institutions to understand the needs of specific populations.

Data mining methods in retention models have increased over the last decade with numerous techniques such as neural networks, decision trees, SVMs, logistic regression, and random forest models (Cardona, Cudney, Hoerl & Snyder, 2020). The introduction of these techniques allows higher education institutions to expand from models whose performance is affected by skewed data or outliers. Instead, institutions can quickly create models to understand the factors currently affecting students and compare them to previous models.

## **Overview of the Study**

Individual community colleges may create retention models to understand student populations but may miss patterns or significant variables due to sample size. This study expands the impact of a school-specific model to include seven different community colleges to identify important predictors and relationships among the state college sector schools. The models created in this research represented freshmen students during two years and can provide scalability to individual schools and whole sectors of community colleges. The background, academic, and financial predictors in this study aligned with Bean and Metzner's nontraditional undergraduate student attrition model. Additionally, the study aims to identify which of the four types of models produces the most accurate results for classifying student retention. The model selection of random forests, SVMs, and neural networks were based on the recommendations of Fernández-Delgado, Cernadas, Barro & Amorim (2014).

## **Related Literature**

Retention frameworks identified students' social and performance factors, including their collegiate relationships in higher education institutions (Aljohani, 2016; Berger et al., 2012). Bean and Metzner (1985) developed the nontraditional undergraduate student attrition model built onto previous retention models and focused primarily on the nontraditional student population (Aljohani, 2016; Bean & Metzner, 1985; Johnson et al., 2014). Their model theorized that student retention depended on the link between high school and college performance, psychological and environmental outcomes playing a more significant role than academic variables, and background variables influencing student persistence (Aljohani, 2016; Bean & Metzner, 1985). The

model was tested and found that nontraditional students drop out for academic reasons unrelated to social interaction (Metzner & Bean, 1987).

**Individual and Sector-based Models** Metzner and Bean (1987) theorized that "samples of nontraditional students tend to be heterogeneous and probably differ substantially from university to university so that the combination of several schools might not produce additive effects" (p. 34), indicating the need for individualized retention models. Another type of modeling is sector-based retention models using similar schools within a sector or area to identify relationships not detected in institutions with smaller populations (Herzog, 2006). With roughly 31% of community college students transferring to four-year institutions, specialized sector models could help the community colleges where students begin and the institutions to where they move (Shapiro et al., 2017).

**Predictive Factors** Many of the academic, background, and financial factors identified by Bean and Metzner (1985) will be the variables used in answering the research questions. The student background characteristics gathered at enrollment reflected demographic information such as high school performance and other demographic factors (Johnson et al., 2014). Academic performance is an indicator of future performance in retention and graduation of community college students at their current and future institutions (Pascarella & Terenzini, 2005). Students' financial attitudes slightly impact retention for students in the nontraditional student attrition model and student retention integrated model (Bean & Metzner, 1985; Cabrera et al., 1993).

**Classification Models** Classification models predict the classes to which the dependent variables individual cases belong and are ideal for retention models (Attewell

& Monaghan, 2015; Bharati & Ramageri, 2010; Breiman, 1999; Breiman et al., 1984; Han et al., 2011). Wolpert's No Free Lunch Theorem indicated that no one classifier could handle all data sets (Wolpert, 1996). Kuhn and Johnson (2013) recommend that researchers start with complex models with the most flexibility and less interpretability to give the most accurate results. Fernández-Delgado et al. (2014) found researchers often use familiar classification methods and are not the most accurate classifier for the problem. They measured the accuracy rates on 179 different classifiers on 121 data sets to determine classifier behavior and found random forests, support vector machine (SVM), neural networks, and boosting ensembles models had the most accurate results among the 121 different data sets (Fernández-Delgado et al., 2014).

Among classification models, random forest trees have some of the highest accuracy rates among different data sets and disciplines (Caruana & Niculescu-Mizil, 2006; Dissanayake, Robinson & Al-Azzam, 2016; Fernández-Delgado et al., 2014; He, Levine, Fan, Beemer & Stronach, 2018). Random forest models have predicted student progress, student performance, completion and graduation rates, and licensing rates, but these are less used than decision trees (Goga et al., 2015; Hardman, Paucar-Caceres & Fielding, 2013; He et al., 2018; Hutt, Gardener, Kamentz, Duckworth & D'Mello, 2018; Langan, Harris, Barrett, Hamshire & Wibberley, 2018). For higher education data, SVMs have been used to predict student retention with mixed results compared to other classifier methods in accuracy (Delen, 2010; Lauría, Baron, Devireddy, Sundararaju & Jayaprakash, 2012; Zhang, Oussena, Clark & Kim, 2010). SVMs use different kernel functions transforming the data based on the specific function with the optimal kernel determination occurring through trial and error (Attewell & Monaghan, 2015, Fernández-

Delgado et al., 2014; James et al., 2013). Neural networks have predicted student course selection, institutional application, retention, and graduation times (Delen 2010; González & DesJardins, 2002; Herzog, 2006; Kardan, Sadeghi, Ghidary & Sani, 2013; Luan, 2002). Even with higher classification accuracy, neural networks can be challenging to interpret the relationship between the inputs and outputs (Attewell & Monaghan, 2015; González & DesJardins, 2002). A standard classification method used in higher education is logistic regression and dates back to the 1960s (Cabrera, 1994). Higher education research using logistic regression range in topics from student retention, student graduation, and the interactions between students and faculty (Astin & Oseguera, 2005; Chatterjee, Marachi, Natekar, Rai & Yeung, 2018; Delen, 2010; Herzog, 2006; Lauría et al., 2012; Pyke & Sheridan, 1993). With the logistic regression's similarity to linear regression and a more straightforward interpretation of the results versus other data mining techniques, many educational researchers choose logistic regression as their statistical method (Gunu, Lee, Gyasi & Roe, 2017; Peng, So, Stage & John, 2002).

## **Methodology**

This study using archival data is a nonexperimental, correlational classification research design created to predict students' retention who completed three consecutive semesters at seven community colleges in Georgia. Five classification models (random forest trees, support vector machine with the radial kernel, support vector machine with the polynomial kernel, neural networks, and logistic regression) were created to determine significant background, academic, and financial factors of retention. The models were compared to each other using respective evaluation metrics and inferential statistics to see if one model outperformed the other models.

## **Participants**

The target population for the study is community college students attending public institutions in Georgia, beginning with their freshman year. The participants of this research are past students who attended their respected colleges from the academic years of Fall 2017 through Fall 2019 without dual enrollment or transfer status. First-time freshmen were identified for two different cohorts with 6,834 (51.44%) students in the Fall 2017 cohort and 6,452 (48.56%) students in the Fall 2018 cohort to create a total of 13,286 students used in the data analysis.

## **Variables Studied**

The first area of research in this study wanted to identify if there were background factors (age, gender, race or ethnicity, and high school GPA), academic factors (college GPA, percentage of courses taken in an online format, number of remedial classes taken, and the number of credits earned during the first academic year), or financial factors (FAFSA completion, amount of financial aid awarded, and amount of financial assistance paid to the student during the first academic year) that were significant in predicting first-year student retention for community college students.

**Background Factors.** The background predictors (age, gender, race or ethnicity, and high school GPA) were chosen based on Bean and Metzner's model. The variable of age had mixed results in the research of its importance as a factor for retention rates at community colleges (Bean & Metzner, 1985; Metzner & Bean, 1987; Pascarella & Chapman, 1983). The research on gender indicated that female students accounted for more than half of high education enrollment and had higher persistence rates than male students (Bean and Metzner, 1985; Chee, Pino, & Smith, 2005; Howell et al., 2003;

Jaggars & Xu, 2010; Wladis, Conway, & Hachey, 2017; Xu & Jaggars, 2011). Bean and Metzner's model believed race or ethnicity hurt college GPAs during secondary education (Bean & Metzner, 1985). Research on high school GPA shows it is an accurate predictor for student persistence in higher education for the first year, especially for the first-year retention in community colleges where students with lower high school GPAs have a higher risk of dropping out (Feldman, 1993; Huerta & Watt, 2015; Yu, 2017).

**Academic Factors.** The academic predictors (college GPA, percentage of courses taken in an online format, number of remedial classes taken, and the number of credits earned during the first academic year) were chosen based on Bean and Metzner's theory. They theorized community college students enroll in their colleges is purely academic, which differs from traditional four-year institutions that may seek out more social and educational integration (Bean & Metzner, 1985). Community college students may have environmental factors limiting their time and resources to explore campus activities, decreasing their social and educational integration. The research on college GPA suggests it is a strong predictor of students' persistence (DeNicco et al., 2015; Metzner & Bean, 1987; Nakajima, Dembo & Mossler, 2012; Stewart et al., 2015; Tinto, 1975). Another predictor is the percentage of online courses taken during the first year. The overall success of these students in online classes is mixed but should be included since many community college students take online courses (Shea & Bidjerano, 2014). The number of remedial courses was included in the study since these courses represent 10% of all courses earned at community colleges (Scott-Clayton & Rodriguez, 2015). The graduation rate for students who take at least one remedial support class area is less than 25% (Bailey et al., 2010). Additionally, the number of credits earned during the first

academic year significantly impacts community college student retention (Mertes & Hoover, 2014; Nakajima, Dembo & Mossler, 2012).

**Financial Factors.** The financial factors (FAFSA completion, amount of financial aid paid to the student during the first academic year) were included in the study. Bean and Metzner's model defined finances as a component in students' ability to be retained in higher education (Bean & Metzner, 1985). With 58% of community college students receiving some financial aid to attend college, this additional financial support plays a significant role in whether students can afford college (Radwin et al., 2018). Students who do not fill out the FAFSA have higher education costs and may wrongly believe they will not qualify for financial aid or cannot afford to attend college (Choitz & Reimherr, 2013; LaManque, 2009; McKinney & Novak, 2012; McKinney & Novak, 2015; Oreopoulos & Dunn, 2013). Students who have to borrow more money to remain in school may find continuing their education as cost-prohibitive with even the addition of \$1000 to the net price of the students' education and the odds of departure (2.5%) increase (Gross et al., 2015; Jones-White et al., 2014).

## **Procedures**

Each of the two cohorts was analyzed separately and had similar student demographic characteristics and predictor values. The cohorts were combined into one final dataset used for answering the research questions. The dataset was divided with a 70% split of data in the training data set ( $n = 9,301$ ) and the remaining 30% in the test data set ( $n = 3,985$ ). Numeric transformation of outlier capping, Yeo-Johnson, normalization, and bagImputation was applied to the training data set before model creation. Additional interactions were created to identify possible relationships between



academic and financial predictors. Each model was created using the training data set and produced model-specific evaluation metrics and variable importance. The highest ROC\_AUC value from each model identified the optimal settings to create the final models using the test data set. The final models yielded the evaluation metrics and significant predictors used to answer the two research questions.

### **Summary of Findings**

This study focused on two research questions to identify predictors of retention and the overall performance of the models. The findings of both questions can provide a starting point for community colleges wanting to create or modify retention models. The significant predictors were chosen based on the nontraditional student population of community colleges for the first-year retention. The models allowed for different methods in determining significant predictors and the overall classification rates of the students.

**Research Question 1** Are background factors (age, gender, race or ethnicity, and high school GPA), academic factors (college GPA, percentage of courses taken in an online format, number of remedial courses taken, and the number of credits earned during the first academic year), and financial factors (FAFSA completion, amount of financial aid awarded, and amount of financial aid paid to the student during the first academic year) significant in predicting first-year student retention for community college students?

The average age of the students in the study was approximately 18.75 years, with the median age being 18. There was an extensive range for ages 15 to 70, with the data being positively skewed. The community college population is diverse, with different

types of people enrolling at different stages in their lives, explaining the considerable variation in this variable. Several categorizations of age were tried. None of these improved the normality of the data, including dividing the dataset into traditional and nontraditional students. Age was not significant to retention or nonretention in the findings. The SVM training models did identify age as a very weak predictor of nonretention, but it was only identified in the testing phase for SVM with the polynomial kernel. The value was very close to 0 and would indicate very slight importance to nonretention.

The dataset supported the finding of the higher population, with female students accounting for 60% of the people, but there was no significant finding that gender influenced retention in any of the models. Among the five models, race or ethnicity were not consistent in the retention or nonretention of community college students. The random forest model found that being a Black or African American student was significant to retention, whereas the SVM with polynomial kernel found being a Black or African American student was important to nonretention. The logistic regression model identified the variable for Hispanic or Latino students critical to retention, but the variable importance plot did not identify any background variables, including race or ethnicity, as significant.

Both SVM models and the logistic regression model found no significance in the high school GPA for retention or non-retention. The neural network model showed a very weak significance for nonretention. The random forest model identified high school GPA as critical to retention after the number of credit hours and the interaction between credit

hours and GPA. While the findings were not consistent throughout the models, high school GPAs should still be considered in future retention models.

The results showed that GPA had a slight significance on retention in all the models in the first three semesters. The interaction of GPA with other variables had higher importance in retention than just GPA alone. In three models, random forest, SVM with the radial kernel, and neural network, the interaction between credit hours and GPA was significant to retention ranking in the first or second position. The SVM with the polynomial kernel, neural network, and the logistic regression model indicates the interaction between GPA and the amount of financial aid awarded was significant to retention in modeling. The interaction between GPA and the percentage of financial assistance used was substantial in the SVM with polynomial kernel model and logistic regression model. The SVM with polynomial kernel model and logistic regression also identified that the interaction between GPA and the number of remedial courses was significant. These two models placed a greater significance on the interactions of GPA than the other three models.

The percentage of courses taken in an online format had little impact on retention in this study. The SVM with radial kernel ranks the percentage of online courses as the most critical variable in non-retention on the test data set, but the variable importance was very low. Online courses may not be the best indicator of retention since the median percentage of online courses taken was 0 courses. The average percentage for online courses taken for the population was 12.42%, equivalent to one 3-hour course the entire three semesters, indicating students are not taking many online courses. While online

courses may be necessary for some students, they had minimal significance in the models overall.

Remedial courses represent roughly 10% of all courses earned at community colleges, but students may not get college credit for these courses (Scott-Clayton & Rodriguez, 2015). These additional courses can delay the overall time to complete a degree and negatively affect the retention of students required to take them (Stewart et al., 2015; Xu & Dadgar, 2018). The number of remedial courses does not appear to be a good indicator of retention since the median number was 0 and the mean number was 0.83. Most students are not taking remedial courses, and if required, most of those students are taking one remedial course. Therefore, the impact of one class would not impact their ability to graduate on time.

The number of remedial courses concerning retention in the study was not very consistent among the models. The variable by itself (with no interaction) was not found significant in any of the models. The interaction for the number of remedial courses with other variables was identified in four models necessary to retention. Three of the models found the interaction between GPA and the number of remedial courses critical to the retention of students. The random forest model identified the interaction between credit hours and the number of remedial classes as significant, while three models found the interaction between credit hours, GPA, and the number of remedial courses as critical to retention. The interaction between remedial classes and other academic variables (GPA and the number of credit hours) is significant to retention.

The seven community colleges in the study have been included in the Momentum Year approach, where students are encouraged to take at least 15 credit hours each

semester and graduate in four years (What is a Momentum Year, 2019). Every model except for the neural network had the number of credit hours as the top predictor of student retention and supported the theory that the number of credit hours in the first year of attendance is significant. The interaction between credit hours and GPA was significant to retention, ranking in the first or second position for three models. The interaction between the number of credit hours and the different variables for financial aid was significant to retention in every model except for the random forest model.

The financial factors can significantly impact students since they may rely on financial aid to persist in college (Hurford et al., 2017). If students' financial assistance and additional resources are unable to pay for their education costs, students may have to work longer hours to pay for their education or drop out (Bound et al., 2010; Scott-Clayton, 2012; Johnson & Rochkind, 2009). For students to qualify for financial aid, students must complete the FASFA application. The FAFSA completion rate for the students in this study was around 92% indicating most students fill out the form. The only model to have FASFA completion as a significant variable was the logistic model and the overall impact on nonretention was very weak. With most students completing the FASFA, the variable may have a minimal effect on the overall model.

With 72% of students receiving financial aid during the 2015-2016 academic year, financial assistance is vital to their retention (NCES, 2018a). Three different financial variables were used to measure the impact of financial aid on student retention; the amount of financial assistance awarded, the amount of financial aid paid, and the amount of financial aid used during the first academic year. The different variables for financial aid were not significant to retention or nonretention in any of the models.

However, the financial variables had a more significant impact on retention when applied as part of the interactions with the number of credit hours or GPA. All the models, except for random forest, had these interactions ranked as significant to the retention of students.

**Research Question 2** Does one of the data mining models (random forests, support vector machines, neural networks, or logistic regression) generate a more accurate classifier performance overall based on the evaluation metrics of accuracy, sensitivity, specificity, area under the curve (ROC\_AUC) and f-measure ( $F_1$ ) scores?

Five models (random forest, support vector machine with the polynomial kernel, support vector machine with the radial kernel, neural network, and logistic regression) were evaluated using ROC curves, confusion matrices, and evaluation metrics (accuracy, ROC\_AUC, specificity, sensitivity, and F1-value) for the training and test data sets. In addition, the models were compared to each other visually and through inferential tests. The Mann-Whitney tests confirmed no significant differences between any evaluation metrics from the training to the test data sets. This finding indicates that the final models were consistent in the training and test phase.

The random forest model was created using the randomForest engine and determined the optimal values for the final random forest model as a mtry value of 10, 1781 trees, and a min\_n value of 36 using the highest ROC\_AUC value in the training of the model. This final model had an accuracy rate of .766, ROC\_AUC value of .821, specificity value of .818, sensitivity value of .707, and F1-value of .741. The combined ROC curves for the test data showed the random forest model having a higher curve than the other models. The Friedman's and Wilcoxon signed-rank tests confirmed that the

random forest model produced the highest ROC\_AUC value, sensitivity value, accuracy values, and F1-value compared to the other models.

The SVM model with the polynomial kernel was created using the kernlab engine with the polynomial kernel fitting the support vector classifier in a higher-dimensional space, allowing for more flexibility in the decision boundary (James et al., 2013). The tuning parameter for this SVM model was a cost value of 1.66 and was found by using the highest ROC\_AUC value (Attewell & Monaghan, 2015). This final model had an accuracy rate of .750, ROC\_AUC value of .797, a specificity value of .882, a sensitivity value of .602, and an F1 value of .695. The SVM with polynomial kernel had the highest value for the specificity among the models and was supported by the Friedman's and Wilcoxon signed-rank tests.

The SVM model with the radial kernel was created using the kernlab engine with the tuning parameters of cost and sigma (Attewell & Monaghan, 2015). The final SVM model with a radial kernel has a cost of 0.024 and a sigma of 0.030 from the highest ROC\_AUC value in the training phase. Since both tuning values are low, the model may have under fitted the data due to its larger, inflexible margins. This final model had an accuracy rate of .738, a ROC\_AUC value of .797, a specificity value of .868, a sensitivity value of .592, and an F1 value of .681. While the SVM with radial kernel had the most extensive range for accuracy, sensitivity, and specificity, it did not have any significant evaluation metrics in the inferential tests.

The neural network models were created using the keras engine with three different parameters: the hidden unit of 4, the penalty was 0.540, and the epochs were 811, which were identified using the highest ROC\_AUC value (Attewell & Monaghan,

2015). This final model had an accuracy rate of .755, ROC\_AUC value of .791, a specificity value of .849, a sensitivity value of .650, and an F1 value of .715. The neural network evaluation metrics constantly ranked in third or fourth place in the models.

The logistic regression model was created using the glm engine and did not require tuning throughout the modeling process. The Hosmer and Lemeshow goodness of fit test significance ( $\chi^2(8) = 30.48, p < 0.001$ ) indicated that the logistic regression model does not fit the data well. Nagelkerke's pseudo R squared value of 0.375 and McFadden's pseudo R squared value of 0.239 describe the variation (37.5% and 23.9%) of the academic, background, and financial factors contribution to the retention status of the students in this model. This final model had an accuracy rate of .748, ROC\_AUC value of .802, a specificity value of .839, a sensitivity value of .647, and an F1 value of .708.

### **Discussion of Findings**

The study focused on identifying significant factors of first-year student retention in community colleges and if one model could outperform the other models based on evaluation metrics. The results did find significant academic and financial predictors for student retention, and one model produced the highest evaluation metrics.

**Research Question 1** Academic and financial factors play an essential role in retaining community college students during their first year. Bean and Metzner (1985) theorized that community college students enroll in their colleges for academic reasons, with the academic interaction serving as their primary integration method. During the first year, the number of credit hours was the most significant variable in any of the models in predicting first-year retention. The workload associated with more credit hours may influence the mindset of students to devote more time and resources to complete the



work. Conversely, students who take few credit hours may have external factors limiting the number of classes they can take and the time they can commit to completing coursework.

Three of the models (random forest, SVM with radial kernel, and neural network) identified the importance of the interaction between the number of credit hours and the first three semesters' GPA. Students who pass their courses have higher GPAs than students who fail courses and accumulate more credit hours. Interactions between credit hours, GPA, and the percentage of online courses or the number of remedial classes were identified as slightly significant to retention.

Four models (SVM with the polynomial kernel, SVM with the radial kernel, neural network, and logistic regression) identified the interactions between GPA and financial variables (amount of financial aid awarded, financial aid amount paid, and percentage of financial assistance used) as significant to retention. The logistic model found that GPA and the number of remedial courses were important to retention.

There were no consistent variables among the five models predicting first-year students' nonretention. Two models, random forest and logistic regression, did not identify any significant variables for nonretention. The SVM with polynomial kernel model identified the following predictors as significant to nonretention: the interaction between the number of credit hours and remedial courses, being a Black or African American student, the interaction between the number of credit hours, percentage of online courses, and remedial courses, the amount of financial aid awarded, and the percentage of financial assistance used. The percentage of online courses and interaction between GPA and the number of remedial classes were significant to nonretention in the

SVM with the radial kernel. The neural network model indicated that high school GPA was significant to nonretention. The absence of significant factors for nonretention could suggest important variables were not included in these models, such as environmental factors like employment and family obligations.

All the academic and financial variables play a role in the retention of first-year community college students. The interactions between credit hours and other academic and financial variables were also important to retention, signifying that academic and financial factors can impact students' number of credit hours.

**Research Question 2** The different classification models were compared to see if any models would have a higher classifier performance based on the different evaluation metrics and inferential tests. The random forest model performed better than the other models in accuracy, F1-values, ROC\_AUC, and sensitivity. Visual inspection of the grouped ROC curves showed the random forest could be the optimal model as the highest accurate classifier for the first-year retention. The SVM with polynomial kernel had the highest value for the specificity of all the models identifying the students who were correctly classified as nonretained.

The research by Fernández-Delgado et al. (2014) measured the accuracy rates on 179 different classifiers on 121 data sets to determine classifier behavior and accuracy regardless of the data sets and found that random forests, support vector machine (SVM), and neural networks had the most accurate results among the 121 different data sets. The random forest model was the optimal classification model in the study based on the highest accuracy and inferential tests. This finding is aligned with the five classifier models' results for retention, with the random forest model having the highest accuracy

value and being significantly higher than the other models. Additionally, Fernández-Delgado et al. (2014) ranked the logistic regression model lower than the SVM and neural network models in overall accuracy, yet this was different from the findings in this study. While the neural network model had a higher accuracy metric for predicting students' retention than both the logistic regression and SVM model, the logistic regression model had a higher accuracy metric than both SVM models. Therefore, the ideal model for predicting community college first-year retention is the random forest model.

### **Limitations of the Study**

One purpose of this study is to add to the existing body of research around the academic, background, and financial factors important for the first-year retention of community college students. There are some limitations of this study that could impact the generalization of the results. The seven colleges used in this study were from one state in the southeastern part of the United States of America. The results of this study may deviate from the different geographical locations such as state, region, and country. The students in the study were from Fall 2017 and 2018, with the last date of the collection being Fall 2019. The data was not influenced by the global COVID-19 pandemic and may not reflect the current factors critical to retention during this period.

The variables in the study were chosen using the results of other retention models but may not account for all variables critical to student retention. While the models captured the significant variables for seven colleges in the state college sector, the overall findings may not be specific enough for each school. Within each of these schools, requirements and regulations may cause differences in the remedial course requirements.

One of the original variables being investigated, the date of FASFA completion, was not included in the study due to the data not being collected for the first cohort of students. Environmental factors such as employment hours and family obligations could play a more significant role in the ability of community college students to be retained.

Another limitation is the omission of social and educational integration since these integrations are not as crucial to the nontraditional student population and require survey data. While survey-based research may capture the students' interactions at the institutions, the results may not provide the most accurate predictive factors (Caison, 2007). This study aims to develop predictive models explaining trends at the sector level using existing variables from archival data without student interventions.

The five models created represent different types of algorithms to provide a greater range of models. Even with the different models, significant predictors or patterns may not be identified if other classifiers were used. The models had different assumptions needing to be met before the training step creation. Random forest models are nonparametric in nature, can handle heavily skewed data, and only require larger sample sizes to properly run (Ali, Khan, Ahmad & Maqsood, 2012). The SVM models are sensitive to predictors with skewed distributions and outliers and may overfit the data (Attewell & Monaghan, 2015; Kuhn & Johnson, 2019). Neural network models can handle various data types but may overfit the data (Attewell & Monaghan, 2015). The logistic regression model had the largest number of assumptions of all models: independence of observations among predictors, linear relationship between the predictors and retention variables, no multicollinearity, and no significant outliers or influential points (Laerd Statistics, 2015a).

Before the models were built, several steps were taken to address missing values, skewed data, and outliers. The dataset was checked for missing data and found three of the financial variables (the percentage of financial aid used, the amount of financial aid awarded, and the amount of financial assistance paid) had the largest amount of missing data. The missing data were replaced through the bagImpute function to impute new similar values. The new values were similar but may have influenced the results. Next, individual histograms and Q-Q plots were created to inspect the shape and visually identify outliers with three univariate normality tests, Anderson-Darling, Lilliefors, and Cramer-von Mises, to verify normality. All predictor variables were found not to have a normal distribution for each of the three tests and contain outliers. Multivariate normality and outliers were assessed using the following tests: Mardia's, Henze-Zirkler's, Doornik-Hansen's, and the Cook's distance multivariate nonnormality and multiple values that could be influencing the data. Outlier capping, Yeo-Johnson transformation, and normalization of the numeric predictors did help improve the normality of the training set slightly.

Logistic regression's assumptions were still violated after the data transformations, including the linear relationship between the predictors and retention variable, significant outliers, and influential points. Multicollinearity and the independence of observations among predictors were also violated since the amount of financial aid paid was depended on the amount of financial aid awarded. Additionally, the percentage of financial assistance used variable was created from two of the other financial variables. These violations could impact the performance of the logistic regression model in its evaluation metrics and variable importance.

The cross validation method divided the data with a 70% split of data in the training data set (n = 9,301) and the remaining 30% in the test data set (n = 3,985) to help with the overfitting of the data. This procedure was only done with one seed and was not repeated with other seeds. Model creation occurred with the tidyverse and tidymodels packages to allow consistency among the models with the preprocessing steps. Only one run of the cross validation method and particular libraries could have limited the results. The final models made from the training results were based on the highest ROC\_AUC value for each model. By choosing one evaluation metric to establish the final models, the results could have been limited and may not have shown each evaluation metric's most accurate model.

### **Implications for Future Research**

The research indicated several academic and financial significant factors for the retention of first-year community college students. The critical academic factors could be expanded to create predictive models finding the ideal number of credit hours and optimal GPA for the retention of these students. While the background predictors were not identified as significant, models with specific demographics, like gender, race, or ethnicity as the study sample, may allow for a greater understanding of these student populations. While financial aid predictors were not significant individually, future models could expand to include new variables, including interactions with academic variables.

The study identified the most accurate data mining model, random forest, for identifying significant factors. As data mining methods keep improving and expanding, other types of models, including a stacked ensemble method, could be used for

comparison. The models' overall performance was based on the highest ROC\_AUC values using the training data. Specific models build on individual evaluation metrics (specificity or sensitivity) may provide more accurate classification results for retention. The models were built using one training and test data set and run once using a specific seed. The expansion of multiple creations of the training and test data sets and different runs of model creations using different seeds can reinforce the results.

The impact of COVID-19 on higher education is continuing and will need to be assessed on first-year students' retention. The models described in this research can be run for different time periods, such as before, during, and after COVID, with comparisons of the significant predictors. This analysis will allow for a greater understanding of what affects freshmen students during the first year and serve as a reference for future events that can arise.

## **Conclusions**

With community colleges providing educational access to roughly half of all undergraduate students in the United States of America, retention models need to be created to serve their populations better (Horn, Nevill & Griffith, 2006; Mullin, 2012; NCES 2013; NCES 2018b). Among the seven community colleges, the number of credit hours was consistently the most critical variable in retention. The interactions between the number of credit hours, GPA, and financial aid variables were significant in student retention in their first year. Additionally, the interaction between GPA, financial aid variables, and the number of remedial hours was crucial for the first-year retention. Combining these variables shows that academic and financial variables are interconnected and may need multiple messages to reach students. Specific marketing

campaigns about the benefits of reaching credit hour milestones, GPA requirements, and FASFA deadlines for financial aid may help impact these variables. Schools with similar populations could work together to share ideas and resources to help gain a broader reach.

No consistent variables among the retention models predicted students' nonretention in the first year of their college career. Many background predictors (age, gender, race, or ethnicity) were not significant in predicting retained or nonretained students. Even though these variables were not significant in these sector-based models, individual community colleges may want to include these variables to understand students' backgrounds better. While FAFSA's completion had no impact on the model, students must complete it for financial aid and should be included in any retention message to understand its importance.

The comparison of the retention models found the random forest model had the best overall performance for accurately classifying the nonretained and retained students together and the retained students individually. The SVM with polynomial kernel had the highest value for the specificity, which identifies the nonretained students. Logistic regression, a commonly used model for retention analysis, did not perform as well as the other models because of the skewed data and correlated variables. The support vector machine with the radial kernel and neural network evaluation metrics never performed better than the random forest and SVM with the polynomial kernel. While the random forest model may not be commonly used in retention models, it may be an ideal model for identifying the overall retention of community college students since it can handle different data (binary, categorical, ordinal) and not be affected by outliers. The study



shows that the use of more than one model allows for validating the variable's importance and potential patterns. Free statistical software makes model creation affordable with no cost training and information to learn software such as R or Python. With higher education funding tied to student retention numbers, specialized retention models can help institutions and systems identify and intervene with students at risk for not returning and help stabilize their funding.

## REFERENCES

- Adelman, C. (2006). *The toolbox revisited: Paths to degree completion from high school through college*. Washington, D.C.: US Department of Education.
- Aguiar, E., Ambrose, G. A. A., Chawla, N. V., Goodrich, V., & Brockman, J. (2014). Engagement vs performance: Using electronic portfolios to predict first semester engineering student persistence. *Journal of Learning Analytics, 1*(3), 7-33.
- Alfonso, M. (2006). The impact of community college attendance on baccalaureate attainment. *Research in Higher Education, 47*(8), 873-903.
- Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues, 9*(5), 272-278.
- Aljohani, O. (2016). A comprehensive review of the major studies and theoretical models of student retention in higher education. *Higher Education Studies, 6*(2), 1-18.
- Allen, I. E., & Seaman, J. (2010). *Learning on Demand: Online Education in the United States, 2009*. Sloan Consortium.
- Allen, I. E., & Seaman, J. (2013). *Changing course: Ten years of tracking online education in the United States*. Sloan Consortium.
- Allen, I. E., Seaman, J., Poulin, R., & Straut, T. T. (2016, February). Online report card: Tracking online education in the United States. Babson Survey Research Group.
- American Association of Community Colleges [AACC]. (2018). *Fast Facts 2018*
- Aragon, S. R., & Johnson, E. S. (2008). Factors influencing completion and noncompletion of community college online courses. *The American Journal of Distance Education, 22*(3), 146-158.

- Astin, A. W., & Higher Education Research Inst., Inc (1975). Financial Aid and Student Persistence. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=ED112804&site=eds-live&scope=site>.
- Astin, A. W., Keup, J. R., & Lindholm, J. A. (2002). A Decade of Changes in Undergraduate Education: A National Study of System" Transformation". *The Review of Higher Education*, 25(2), 141-162.
- Astin, A. W., & Oseguera, L. (2005). Pre-college and institutional influences on degree attainment. *College student retention: Formula for student success*, (pp. 245-276). Plymouth: Rowman & Littlefield Publishers.
- Attewell, P., & Monaghan, D., (2015). *Data mining for the social sciences: An introduction*. Oakland: University of California Press.
- Bailey, T., Cho, S.-W., & Columbia University, C. C. R. C. (2010). *Issue Brief: Developmental Education in Community Colleges*. Community College Research Center, Columbia University. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=ED512399&site=eds-live&scope=site>
- Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3-17.
- Barbatis, P. (2010). Underprepared, Ethnically Diverse Community College Students: Factors Contributing to Persistence. *Journal of Developmental Education*, 33(3), 16.
- Baum, S. (2018). Trends in Student Aid, 2018. New York: 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*(2), 155-187.
- Bean, J. P. (1982). Conceptual models of student attrition: How theory can help the institutional researcher. *New Directions for Institutional Research, 1982*(36), 17-33.
- Bean, J., & Eaton, S. B. (2001). The psychology underlying successful retention practices. *Journal of College Student Retention: Research, Theory & Practice, 3*(1), 73-89.
- Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research, 55*(4), 485-540.
- Belfield, C. R., Crosta, P. M., & Columbia University, C. C. R. C. (2012). *Predicting Success in College: The Importance of Placement Tests and High School Transcripts*. CCRC Working Paper No. 42. Community College Research Center, Columbia University. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=ED529827&site=eds-live&scope=site>
- Belli, G. (2008). Quantitative Research. Retrieved from <http://www.pdfslibforme.com/gabriella-belliquantitative-research.pdf>
- Berger, J., Ramirez, G. B., & Lyon, S. (2012). Past to present: A historical look at retention. *College student retention: Formula for student success*, (pp. 7-34). Plymouth: Rowman & Littlefield Publishers.
- Bettinger, E. P., Long, B. T., Oreopoulos, P., & Sanbonmatsu, L. (2012). The role of application assistance and information in college decisions: Results from the

- H&R Block FAFSA experiment. *The Quarterly Journal of Economics*, 127(3), 1205-1242.
- Bharati, M., & Ramageri, M. (2010). Data mining techniques and applications. *Indian Journal of Computer Science and Engineering*, 1(4), 301-305.
- Bost, R., Popa, R. A., Tu, S., & Goldwasser, S. (2015, February). Machine learning classification over encrypted data. In NDSS (Vol. 4324, p. 4325). Retrieved from <http://iot.stanford.edu/pubs/bost-learning-ndss15.pdf>
- Bound, J., Lovenheim, M. F., & Turner, S. (2012). Increasing time to baccalaureate degree in the United States. *Education Finance and Policy*, 7(4), 375-424.
- Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, 144-152. Retrieved from <https://dl.acm.org/doi/10.1145/130385.130401>
- Bowers, A. J., Sprott, R., & Taff, S. A. (2012). Do we know who will drop out? A review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *The High School Journal*, 96(2), 77-100.
- Boylan, H. R., & Saxon, D. P. (1999). What works in remediation: Lessons from 30 years of research. Retrieved from <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.734.1771&rep=rep1&type=pdf>
- Bransberger, P., Michelau, D. K., & Western Interstate Commission for Higher Education. (2017). *Knocking at the College Door: Projections of High School Graduates*, 9th Edition. Revised. Western Interstate Commission for Higher

- Education. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=ED573115&site=eds-live&scope=site>
- Braxton, J. M. (2000). *Reworking the student departure puzzle*. Nashville: Vanderbilt University Press.
- Breier, M. (2010). From 'financial considerations' to 'poverty': Towards a reconceptualisation of the role of finances in higher education student drop out. *Higher Education, 60*(6), 657-670.
- Breiman L. (1999). Random Forests—random features. Statistics Department, University of California, Berkeley, Technical Report 567, September 1999.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Boca Raton: Chapman & Hall/CRC.
- Burke, A. (2019). Student retention models in higher education: A literature review. *College and University, 94*(2), 12-21.
- Burley, B., Butner, B., & Cejda, H. (2001). Dropout and stopout patterns among developmental education students in Texas community colleges. *Community College Journal of Research and Practice, 25*(10), 767-782.
- Burns, K. (2010). At issue: Community college student success variables: A review of the literature. *The Community College Enterprise, 16*(2), 33.
- Cabrera, A. F. (1994). Logistic regression analysis in higher education: An applied perspective. *Higher education: Handbook of theory and research, 10*, 225-256.
- Cabrera, A. F., Nora, A., & Castaneda, M. B. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. *The Journal of Higher Education, 64*(2), 123-139.

- Caison, A. L. (2007). Analysis of institutionally specific retention research: A comparison between survey and institutional database methods. *Research in Higher Education, 48*(4), 435-451.
- Cardona, T., Cudney, E. A., Hoerl, R., & Snyder, J. (2020). Data Mining and Machine Learning Retention Models in Higher Education. *Journal of College Student Retention: Research, Theory & Practice, 13*(1), 17–35
- Caruana, R. & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning*, 161-168. Retrieved from <http://lacam.di.uniba.it:8000/people/courses/IA/IA0809/sistemi/caruana.icml06.pdf>
- Cejda, B. (2010). Online education in community colleges. *New Directions for Community Colleges, (150)*, 7-16.
- Chacon, F., Spicer, D., & Valbuena, A. (2012). Analytics in support of student retention and success. *Research Bulletin, 3*, 1-9.
- Chatterjee, A., Marachi, C., Natekar, S., Rai, C., & Yeung, F. (2018). Using logistic regression model to identify student characteristics to tailor graduation initiatives. *College Student Journal, 52*(3), 352-360.
- Chee, K. H., Pino, N. W., & Smith, W. L. (2005). Gender differences in the academic ethic and academic achievement. *College student journal, 39*(3), 604-619.
- Choitz, V., & Reimherr, P. (2013). Mind the Gap: High Unmet Financial Need Threatens Persistence and Completion for Low-Income Community College Students. *Center for Law and Social Policy, Inc.*

- Cohen, A. M., Brawer, F. B., & Kisker, C. B. (2014). *The American community college*. San Francisco: Jossey-Bass.
- Community College FAQs. (n.d.). Retrieved from <https://ccrc.tc.columbia.edu/Community-College-FAQs.html>.
- Corbett, C., Hill, C., & St. Rose, A. (2008). Where the Girls Are: The Facts about Gender Equity in Education. American Association of University Women Educational Foundation. 1111 Sixteenth Street NW, Washington, DC 20036.
- Cornwell, C., Mustard, D. B., & Sridhar, D. J. (2006). The Enrollment Effects of Merit-Based Financial Aid: Evidence from Georgia's HOPE Program. *Journal of Labor Economics*, 24(4), 761–786.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.
- Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating*. Upper Saddle River: Merrill.
- Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oakes: Sage.
- Crisp, G., Carales, V. D., & Núñez, A. M. (2016). Where is the research on community college students?. *Community College Journal of Research and Practice*, 40(9), 767-778.
- Crisp, G., & Delgado, C. (2014). The impact of developmental education on community college persistence and vertical transfer. *Community College Review*, 42(2), 99-117.



- Crisp, G., & Mina, L. (2012). The community college. *College student retention: Formula for student success*, (pp. 147-165). Plymouth: Rowman & Littlefield Publishers.
- Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4), 498-506.
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan), 1-30.
- DeNicco, J., Harrington, P., & Fogg, N. (2015). Factors of one-year college retention in a public state college system. *Research in Higher Education Journal*, 27.
- Denning, J. T. (2019). Born under a lucky star financial aid, college completion, labor supply, and credit constraints. *Journal of Human Resources*, 54(3), 760-784.
- DesJardins, S. L., Ahlburg, D. A., & McCall, B. P. (2002). Simulating the longitudinal effects of changes in financial aid on student departure from college. *Journal of Human Resources*, 653-679.
- Dissanayake, H., Robinson, D., & Al-Azzam, O. (2016). Predictive modeling for student retention at St. Cloud state university. *Proceedings of the International Conference on Data Mining*, 215-221. Retrieved from <https://search.proquest.com/docview/1806428521?pq-origsite=gscholar>
- Drigas, A. S., & Leliopoulos, P. (2014). The use of big data in education. *International Journal of Computer Science Issues*, 11(5), 58.
- Dynarski, S. M., Hemelt, S. W., & Hyman, J. M. (2015). The missing manual: Using national student clearinghouse data to track postsecondary outcomes. *Educational Evaluation and Policy Analysis*, 37, 53S-79S.

- Educationaldatamining.org. (n.d.). Retrieved from <http://educationaldatamining.org/>.
- Efron, B. (1979). Computers and the theory of statistics: Thinking the unthinkable. *SIAM review*, 21(4), 460-480.
- Etheridge, H. L., Sriram, R. S., & Hsu, H. K. (2000). A comparison of selected artificial neural networks that help auditors evaluate client financial viability. *Decision Sciences*, 31(2), 531-550.
- Eynon, R. (2013). The rise of Big Data: What does it mean for education, technology, and media research?. *Learning, Media, and Technology*, 38(3), 237-240.
- Federal Pell Grants. (2019, July). Retrieved from <https://studentaid.ed.gov/sa/types/grants-scholarships/pell>
- Feeney, M., & Heroff, J. (2013). Barriers to Need-Based Financial Aid: Predictors of Timely FAFSA Completion Among Low-Income Students. *Journal of Student Financial Aid*, 43(2), 2.
- Feldman, M. J. (1993). Factors associated with one-year retention in a community college. *Research in Higher Education*, 34(4), 503-512.
- Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. *The Journal of Machine Learning Research*, 15(1), 3133-3181.
- Fike, D. S., & Fike, R. (2008). Predictors of first-year student retention in the community college. *Community college review*, 36(2), 68-88.
- Fosnacht, K., Sarraf, S., Howe, E. & Peck, L. K. (2017). How important are high response rates for college surveys? *The Review of Higher Education* 40(2), 245-265.

- Ginder, S. A., Kelly-Reid, J. E., & Mann, F. B. (2018). Postsecondary Institutions and Cost of Attendance in 2017-18; Degrees and Other Awards Conferred, 2016-17; and 12-Month Enrollment, 2016-17: First Look (Provisional Data). NCES 2018-060rev. National Center for Education Statistics.
- Goga, M., Kuyoro, S., & Goga, N. (2015). A recommender for improving the student academic performance. *Procedia-Social and Behavioral Sciences*, *180*, 1481-1488.
- Goldrick-Rab, S. (2010). Challenges and opportunities for improving community college student success. *Review of Educational Research*, *80*(3), 437-469.
- González, J. M. B., & DesJardins, S. L. (2002). Artificial neural networks: A new approach to predicting application behavior. *Research in Higher Education*, *43*(2), 235-258.
- Grawe, N. D. (2018). *Demographics and the demand for higher education*. Baltimore: JHU Press
- Gregory, C. B., & Lampley, J. H. (2016). Community college student success in online versus equivalent face-to-face courses. *Journal of Learning in Higher Education*, *12*(2), 63-72.
- Gross, J. P., Hossler, D., Ziskin, M., & Berry, M. S. (2015). Institutional merit-based aid and student departure: A longitudinal analysis. *The Review of Higher Education*, *38*(2), 221-250.
- Gunu, E. A., Lee, C., Gyasi, W. K., & Roe, R. M. (2017). Modern predictive models for modeling the college graduation rates. *Proceedings of the 15th International*

- Conference on Software Engineering Research, Management and Applications*, 39-45. Retrieved from <https://ieeexplore.ieee.org/abstract/document/7965705>
- Hagedorn, L. S. (2012). How to define student retention: A new look at an old problem. *College student retention: Formula for student success*, (pp. 81-100). Plymouth: Rowman & Littlefield Publishers.
- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Waltham: Elsevier.
- Hand, D. J. (1998). Data mining: Statistics and more?. *The American Statistician*, 52(2), 112-118.
- Hardman, J., Paucar-Caceres, A., and Fielding, A. (2013). Predicting students' progression in higher education by using the random forest algorithm. *Systems Research and Behavioral Science*, 30(2), 194-203.
- Hart, C., Friedmann, E., & Hill, M. (2018). Online course-taking and student outcomes in California community colleges. *Education Finance and Policy*, 13(1), 42-71.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.
- He, L., Levine, R. A., Fan, J., Beemer, J., & Stronach, J. (2018). Random forest as a predictive analytics alternative to regression in institutional research. *Practical assessment, research & evaluation*, 23(1), 1-16.
- Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. *New Directions for Institutional Research*, 131, 17-33.

- Herzog, S. (2018). Financial aid and college persistence: Do student loans help or hurt?. *Research in Higher Education, 59*(3), 273-301.
- Hicks, C., & Jones, S. (2011). At Issue: survival tactics for small, rural-serving community colleges. *Community College Enterprise, 17*(2), 28-45.
- Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. (Vol. 398). John Wiley & Sons.
- Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process, 5*(2), 1-11.
- Hossler, D., & Bontrager, B. (2014). *Handbook of strategic enrollment management*. San Francisco: John Wiley & Sons.
- Horn, L., Nevill, S., & Griffith, J. (2006). Profile of undergraduates in US postsecondary education institutions, 2003-04: With a special analysis of community college students. *Statistical Analysis Report*. NCES 2006-184. National Center for Education Statistics.
- Horn, L., & Skomsvold, P. (2011). Community college student outcomes: 1994-2009. Washington, DC: National Center for Education Statistics.
- Hothorn, T., Leisch, F., Zeileis, A., & Hornik, K. (2005). The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics, 14*(3), 675-699.
- Hout, M. (2012). Social and economic returns to college education in the United States. *Annual Review of Sociology, 38*, 379-400.

- Howell, S. L., Williams, P. B., & Lindsay, N. K. (2003). Thirty-two trends affecting distance education: An informed foundation for strategic planning. *Online Journal of Distance Learning Administration*, 6(3), 1-18.
- Huebner, R. A. (2013). A survey of educational data-mining research. *Research in Higher Education Journal*, 19(1-13).
- Huerta, J., & Watt, K. M. (2015). Examining the college preparation and intermediate outcomes of college success of AVID graduates enrolled in universities and community colleges. *American Secondary Education*, 43(3), 20.
- Huilgol, P. (2019, August 24). Accuracy vs. F1-Score. Retrieved from <https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2>
- Hurford, D. P., Ivy, W. A., Winters, B., & Eckstein, H. (2017). Examination of the variables that predict freshman retention. *The Midwest Quarterly*, 3, 302.
- Hutt, S., Gardener, M., Kamentz, D., Duckworth, A. L., & D'Mello, S. K. (2018). Prospectively predicting 4-year college graduation from student applications. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 280-289. Retrieved from <https://dl.acm.org/doi/10.1145/3170358.3170395>
- IBM Software Group. (2018). IBM predictive analytics for education. Armonk, NY., Retrieved from <https://www.ibm.com/downloads/cas/DG75RMX0>
- Iloh, C. (2018). Toward a new model of college “choice” for a twenty-first-century context. *Harvard Educational Review*, 88(2), 227-244.

- Jackson, A. (2015, July 20). This chart shows how quickly college tuition has skyrocketed since 1980. Retrieved from <https://www.businessinsider.com/this-chart-shows-how-quickly-college-tuition-has-skyrocketed-since-1980-2015-7>
- Jaggars, S. S., Edgecombe, N., & Stacey, G. W. (2013). *What we know about online course outcomes*. New York, NY: Columbia University, Teachers College, Community College Research Center.
- Jaggars, S. S., & Xu, D. (2010). *Online learning in the Virginia Community College System*. New York, NY: Columbia University, Teachers College, Community College Research Center.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.
- James, S., Swan, K., & Daston, C. (2016). Retention, progression and the taking of online courses. *Online Learning, 20*(2), 75-96.
- Johnson, H. P., & Mejia, M. C. (2014). *Online learning and student outcomes in California's community colleges*. Public Policy Institute.
- Johnson, J., & Rochkind, J. (2009). With their whole lives ahead of them: Myths and realities about why so many students fail to finish college. *Public Agenda*.
- Johnson, D. R., Wasserman, T. H., Yildirim, N., & Yonai, B. A. (2014). Examining the effects of stress and campus climate on the persistence of students of color and white students: An application of Bean and Eaton's psychological model of retention. *Research in Higher Education, 55*(1), 75-100.
- Jones-White, D. R., Radcliffe, P. M., Lorenz, L. M., & Soria, K. M. (2014). Priced out?. *Research in Higher Education, 55*(4), 329-350.

- Jurgens, J. C. (2010). The evolution of community colleges. *College Student Affairs Journal*, 28(2), 251–261.
- Juszkiewicz, J. (2017). Trends in Community College Enrollment and Completion Data, 2017. *American Association of Community Colleges*.
- Juszkiewicz, J. (2020). Trends in community college enrollment and completion data, 2020. *American Association of Community Colleges*.
- Kantrowitz, M. (2009). Analysis of why some students do not apply for financial aid. Student Aid Policy Analysis. Retrieved from <http://www.finaid.org/educators/studentaidpolicy.phtml>.
- Kardan, A. A., Sadeghi, H., Ghidary, S. S., & Sani, M. R. F. (2013). Prediction of student course selection in online higher education institutes using neural network. *Computers & Education*, 65, 1-11.
- Kenamer, M. A., Katsinas, S. G., & Schumacker, R. E. (2010). The moving target: Student financial aid and community college student retention. *Journal of College Student Retention: Research, Theory & Practice*, 12(1), 87-103.
- Kerby, M. B. (2015). Toward a new predictive model of student retention in higher education: An application of classical sociological theory. *Journal of College Student Retention: Research, Theory & Practice*, 17(2), 138-161.
- Kerkvliet, J., & Nowell, C. (2014). Public subsidies, tuition, and public universities' choices of undergraduate acceptance and retention rates in the USA. *Education Economics*, 22, 652-666.



- Knowles, J. E. (2015). Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *Journal of Educational Data Mining*, 7(3), 18-67.
- Kofoed, M. S. (2017). To apply or not to apply: FAFSA completion and financial aid gaps. *Research in Higher Education*, 58(1), 1-39.
- Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An R package for assessing multivariate normality. *The R Journal*, 6(2), 151-162.
- Krüger, A., Merceron, A., & Wolf, B. (2010). A Data Model to Ease Analysis and Mining of Educational Data. *Educational Data Mining 2010*, 131-140
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.
- Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5), 1-26.
- Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. Boca Raton; CRC Press.
- Kuhn, M, Williams, C.K., Engelhardt, A., Cooper, T., Mayer, Z., Ziem, A., Scrucca, L., Tang, Y., Candan, C., & Hunt, T. (2019). Package ‘caret’. Retrieved from <https://topepo.github.io/caret/>
- Laerd Statistics (2015a). Binomial logistic regression using SPSS statistics. Statistical tutorials and software guides. Retrieved from <https://statistics.laerd.com/>
- Laerd Statistics (2015b). Friedman test using SPSS Statistics. Statistical tutorials and software guides. Retrieved from <https://statistics.laerd.com/>
- Laerd Statistics (2015c). Mann-Whitney U test using SPSS Statistics. Statistical tutorials and software guides. Retrieved from <https://statistics.laerd.com/>

- Laerd Statistics (2015d). Wilcoxon signed-rank test using SPSS Statistics. Statistical tutorials and software guides. Retrieved from <https://statistics.laerd.com/>
- LaManque, A. (2009). Factors associated with delayed submission of the free Application for federal financial aid. *Journal of Applied Research in the Community College*, 17(1), 6.
- Langan, A. M., Harris, W. E., Barrett, N., Hamshire, C., & Wibberley, C. (2018). Benchmarking factor selection and sensitivity: a case study with nursing courses. *Studies in Higher Education*, 43(9), 1586-1596.
- Lau F. (2017). *Handbook of ehealth evaluation: An evidence-based approach [internet]*. Victoria (BC): University of Victoria.
- Lauría, E. J., Baron, J. D., Devireddy, M., Sundararaju, V., & Jayaprakash, S. M. (2012). Mining academic data to improve college student retention: An open source perspective. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* pp. 139-142. Retrieved from <http://cs.colby.edu/courses/S16/cs251-labs/final-lauria-studentRetention-LAK2012.pdf>
- Layne, M., Boston, W. E., & Ice, P. (2013). A longitudinal study of online learners: Shoppers, swirlers, stoppers, and succeeders as a function of demographic characteristics. *Online Journal of Distance Learning Administration*, 16(2), 1-12.
- Leinbach, D. T., & Jenkins, D. (2008). Using longitudinal data to increase community college student success: A guide to measuring milestone and momentum point attainment. CCRC Research Tools No. 2. *Community College Research Center, Columbia University*.

- Levin, K. A. (2006). Study design III: Cross-sectional studies. *Evidence-based Dentistry*, 7(1), 24.
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3), 18-22.
- Lin, S. H. (2012). Data mining for student retention management. *Journal of Computing Sciences in Colleges*, 27(4), 92-99.
- Luan, J. (2002, June). *Data Mining and Knowledge Management in Higher Education Applications*. Paper presented at the Annual Forum for the Association for Institutional Research, Toronto, Ontario, Canada. Retrieved from <http://eric.ed.gov/ERICWebPortal/detail?accno=ED474143>
- Ma, X. (2018). *Using classification and regression trees: A practical primer*. Charlotte: IAP.
- Mao, W., & Wang, F. (2012). *New advances in intelligence and security informatics*. Waltham: Academic Press.
- Marcotte, D. E., Bailey, T., Borkoski, C., & Kienzl, G. S. (2005). The returns of a community college education: Evidence from the National Education Longitudinal Survey. *Educational Evaluation and Policy Analysis*, 27(2), 157-175.
- Marr, B. (2018, May). How much data do we create every day? The mind-blowing stats everyone should read. *Forbes*, Retrieved from <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#4772f32360ba>

- Martin, K., Galentino, R., & Townsend, L. (2014). Community college student success: The role of motivation and self-empowerment. *Community College Review, 42*(3), 221-241.
- McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: The management revolution. *Harvard Business Review, 90*(10), 60-68.
- McKinney, L., & Novak, H. (2012). The relationship between FAFSA filing and persistence among first-year community college students. *Community College Review, 41*(1), 63-85.
- McKinney, L., & Novak, H. (2015). FAFSA filing among first-year college students: Who files on time, who doesn't, and why does it matter?. *Research in Higher Education, 56*(1), 1-28.
- Mendez, G., Buskirk, T. D., Lohr, S., & Haag, S. (2008). Factors associated with persistence in science and engineering majors: An exploratory study using classification trees and random forests. *Journal of Engineering Education, 97*(1), 57-70.
- Mertes, S. J., & Hoover, R. E. (2014). Predictors of first-year retention in a community college. *Community College Journal of Research and Practice, 38*(7), 651-660.
- 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.
- Milliron, M. D., de los Santos, G. E., & Browning, B. (2003). Feels like the third wave: The rise of fundraising in the community college. *New Directions for Community Colleges, 2003*(124), 81-93.

- Mills, G. E., & Gay, L. R. (2019). *Educational research: Competencies for analysis and applications*. Pearson. One Lake Street, Upper Saddle River, New Jersey 07458.
- Mitchell, M., & Leachman, M. (2015). Years of cuts threaten to put college out of reach for more students. *Center on Budget and Policy Priorities*, 13.(2015), 16.
- Monaghan, D. B., & Attewell, P. (2015). The community college route to the bachelor's degree. *Educational Evaluation and Policy Analysis*, 37(1), 70-91.
- Morrison, L., & Silverman, L. (2012). Retention theories, models, and concepts. *College student retention: Formula for student success*, (pp. 61-80). Plymouth: Rowman & Littlefield Publishers.
- Mullin, C. M. (2012). Why Access Matters: The Community College Student Body. AACC Policy Brief 2012-01PBL. *American Association of Community Colleges (NJ)*.
- 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.
- National Center for Educational Statistics [NCES] (2013). 2011–12 National Postsecondary Student Aid Study Retrieved from <https://nces.ed.gov/pubs2013/2013165.pdf>
- National Center for Educational Statistics [NCES] (2018a). 2015–16 National Postsecondary Student Aid Study Retrieved from <https://nces.ed.gov/pubs2018/2018466.pdf>
- National Center for Educational Statistics [NCES] (2018b). Enrollment and Employees in

Postsecondary Institutions Fall 2016; and Financial Statistics and Academic Libraries, Fiscal Year 2016. Retrieved from <https://nces.ed.gov/pubs2018/2018002.pdf>

National Center for Educational Statistics [NCES] (2019a). National Center for Educational Statistics (2019a). Characteristics of Degree-Granting Postsecondary Institutions. Retrieved from [https://nces.ed.gov/programs/coe/indicator\\_csa.asp](https://nces.ed.gov/programs/coe/indicator_csa.asp)

National Center for Educational Statistics [NCES] (2019b). Characteristics of Postsecondary Students. Retrieved from [https://nces.ed.gov/programs/coe/indicator\\_csb.asp](https://nces.ed.gov/programs/coe/indicator_csb.asp)

National Center for Education Statistics. [NCES] (2019c). Immediate College Enrollment Rate. Retrieved from [https://nces.ed.gov/programs/coe/indicator\\_cpa.asp](https://nces.ed.gov/programs/coe/indicator_cpa.asp)

National Center for Educational Statistics [NCES] (2019d). Undergraduate Enrollment. Retrieved from [https://nces.ed.gov/programs/coe/indicator\\_ctr.asp](https://nces.ed.gov/programs/coe/indicator_ctr.asp)

National Center for Educational Statistics [NCES] (2019e). Undergraduate Retention and Graduation Rates. Retrieved from [https://nces.ed.gov/programs/coe/indicator\\_cha.asp](https://nces.ed.gov/programs/coe/indicator_cha.asp)

National Student Clearinghouse Research Center. (2012). Snapshot report on degree attainment. Retrieved from <https://nscresearchcenter.org/wp-content/uploads/SnapshotReport8-GradRates2-4Transfers.pdf>

National Student Clearinghouse Research Center. (2019). Persistence & Retention - 2019. Retrieved from <https://nscresearchcenter.org/snapshotreport35-first-year-persistence-and-retention/>

- Nevarez, C., & Wood, J. L. (2010). *Community college leadership and administration: Theory, practice, and change (Vol. 3)*. New York: Peter Lang.
- Ngemu, J. M., Elisha, O. O., William, O. O., & Bernard, M. (2015). Student retention prediction in higher learning institutions: The Machakos university college case. *International Journal of Computer and Information Technology*, 4(2), 489-497.
- O'Toole, D. M., Stratton, L. S., & Wetzel, J. N. (2003). A longitudinal analysis of the frequency of part-time enrollment and the persistence of students who enroll part time. *Research in Higher Education*, 44(5), 519-537.
- Oreopoulos, P., & Dunn, R. (2013). Information and college access: Evidence from a randomized field experiment. *The Scandinavian Journal of Economics*, 115(1), 3-26.
- Owen, L., & Westlund, E. (2016). Increasing college opportunity: School counselors and FAFSA completion. *Journal of College Access*, 2(1), 3.
- Page, L. C., Castleman, B. L., & Meyer, K. (2016). Customized nudging to improve FAFSA completion and income verification. *Educational Evaluation and Policy Analysis*, 0162373719876916.
- Park, R. S. E., & Scott-Clayton, J. (2018). The impact of Pell Grant eligibility on community college students' financial aid packages, labor supply, and academic outcomes. *Educational Evaluation and Policy Analysis*, 40(4), 557-585.
- Parsad, B., Lewis, L., & National Center for Education Statistics (ED). (2008). Distance Education at Postsecondary Institutions: 2006-07. First Look. NCES 2009-044. *National Center for Education Statistics*.

- Pascarella, E. T., & Chapman, D. W. (1983). A multiinstitutional, path analytic validation of Tinto's model of college withdrawal. *American Educational Research Journal*, 20(1), 87-102.
- Pascarella, E. T., & Terenzini, P. T. (2005). *How College Affects Students: A Third Decade of Research. Volume 2*. Indianapolis: Jossey-Bass.
- Peng, C. Y. J., So, T. S. H., Stage, F. K., & John, E. P. S. (2002). The use and interpretation of logistic regression in higher education journals: 1988–1999. *Research in Higher Education*, 43(3), 259-293.
- Peter, K., & Horn, L. (2005). Gender Differences in Participation and Completion of Undergraduate Education and How They Have Changed Over Time. Postsecondary Education Descriptive Analysis Reports. NCES 2005-169. *US Department of Education*.
- Phelan, D. J. (2014). The clear and present funding crisis in community colleges. *New Directions for Community Colleges*, 2014(168), 5-16.
- Picciano, A. G. (2012). The evolution of big data and learning analytics in American higher education. *Journal of Asynchronous Learning Networks*, 16(3), 9-20.
- Pontes, M. C. F., & Pontes, N. M. H. (2012). Distance education enrollment is associated with greater academic progress among first generation low-income undergraduate students in the US in 2008. *Online Journal of Distance Learning Administration*, 15(1). 1-5.
- Porter, A. C., & Polikoff, M. S. (2012). Measuring academic readiness for college. *Educational Policy*, 26(3), 394-417.



- Prescott, B. (2008). Knocking at the College Door: Projections of High School Graduates by State and Race/Ethnicity, 1992-2022. *Western Interstate Commission for Higher Education*.
- Provasnik, S., & Planty, M. (2008). Community colleges: Special supplement to the Condition of Education 2008 (NCES 2008-033). *National Center for Education Statistics*.
- Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big data*, 1(1), 51-59.
- Pyke, S. W., & Sheridan, P. M. (1993). Logistic regression analysis of graduate student retention. *Canadian Journal of Higher Education*, 23(2), 44-64.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81-106.
- Quinlan, J. R. (2014). *C4. 5: Programs for machine learning*. San Mateo: Morgan Kaufmann Publishers, Inc.
- Radwin, D., Conzelmann, J.G., Nunnery, A., Lacy, T.A., Wu, J., Lew, S., Wine, J. and Siegel, P., (2018). 2015-16 National Postsecondary Student Aid Study (NPSAS: 16): Student Financial Aid Estimates for 2015-16. First Look. NCES 2018-466. *National Center for Education Statistics*.
- Ripley, B., & Venables, W. (2016). Package ‘nnet’. R package version, 7, 3-12.
- Roiger, R. J. (2017). *Data mining: a tutorial-based primer*. Boca Raton: CRC Press.
- Sanburn, J., & Watertown, S. (2017). The case for community college. *Time*, 189(22), 44-47.

- Schudde, L., & Goldrick-Rab, S. (2015). On second chances and stratification: How sociologists think about community colleges. *Community College Review, 43*(1), 27-45.
- Schuetz, P. (2008). Developing a theory-driven model of community college student engagement. *New Directions for Community Colleges, 2008*(144), 17–28.
- Schneider, M., & Yin, L. M. (2012). Completion matters: The high cost of low community college graduation rates. *AEI Education Outlook, 2*, 1-10.
- Scott-Clayton, J. (2012). What explains trends in labor supply among US undergraduates?. *National Tax Journal, 65*(1), 181.
- Scott-Clayton, J., & Rodriguez, O. (2015). Development, discouragement, or diversion? New evidence on the effects of college remediation policy. *Education Finance and Policy, 10*(1), 4-45.
- Shapiro, D., Dundar, A., Huie, F., Wakhungu, P. K., Yuan, X., Nathan, A., & Hwang, Y. (2017). Tracking transfer: Measures of effectiveness in helping community college students to complete bachelor's degrees. *Signature Report, (13)*.
- Shea, P., & Bidjerano, T. (2014). Does online learning impede degree completion? A national study of community college students. *Computers & Education, 75*, 103-111.
- Smith, T. J., & McKenna, C. M. (2013). A comparison of logistic regression pseudo R2 indices. *Multiple Linear Regression Viewpoints, 39*(2), 17-26
- Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange, 1*(1), 64-85.

- Stephan, J. L., Rosenbaum, J. E., & Person, A. E. (2009). Stratification in college entry and completion. *Social Science Research, 38*(3), 572-593.
- Stewart, S., Lim, D. H., & Kim, J. (2015). Factors influencing college persistence for first-time students. *Journal of Developmental Education, 38*(3), 12-20
- Terriquez, V., & Gurantz, O. (2015). Financial challenges in emerging adulthood and students' decisions to stop out of college. *Emerging Adulthood, 3*(3), 204-214.
- Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45*(1), 89-125.
- Tinto, V. (1993). Building community. *Liberal Education, 79*(4), 16-21.
- Tinto, V. (1999). Taking retention seriously: Rethinking the first year of college. *NACADA Journal, 19*(2), 5-9.
- Torche, F (2011). Is a college degree still the great equalizer? Intergenerational mobility across levels of schooling in the United States. *American Journal of Sociology, 117*(3), 763.
- Townsend, B. K., & Wilson, K. B. (2006). "A hand hold for a little bit": Factors facilitating the success of community college transfer students to a large research university. *Journal of College Student Development, 47*(4), 439.
- Townsend, B. K., & Wilson, K. B. (2009). The academic and social integration of persisting community college transfer students. *Journal of College Student Retention: Research, Theory & Practice, 10*(4), 405-423.
- Travers, S. (2016). Supporting online student retention in community colleges. *Quarterly Review of Distance Education, 17*(4), 49.

- U.S. Department of Education, National Center for Education Statistics, Integrated Postsecondary Education Data System (IPEDS), (n.d.) 2017-2018, College Navigator. Retrieved from <https://nces.ed.gov/collegenavigator/>
- Walker, E. G. (2016). Predicting higher education outcomes and implications for a postsecondary institution ratings system. *Journal of Higher Education Policy & Management, 38*, 422-433.
- Wang, X. (2012). Academic performance of community college transfers: Psychological, sociodemographic, and educational correlates. *Community College Journal of Research and Practice, 36*(11), 872-883.
- Wei, C. C., & Horn, L. (2013). Federal Student Loan Debt Burden of Noncompleters. Stats in Brief. NCES 2013-155. *National Center for Education Statistics*.
- What is a Momentum Year? (2019). Retrieved from <https://completega.org/what-momentum-year>
- Wild, L., & Ebbers, L. (2002). Rethinking student retention in community colleges. *Community College Journal of Research and Practice, 26*(6), 503-519.
- Windham, M. H., Rehfuss, M. C., Williams, C. R., Pugh, J. V., & Tincher-Ladner, L. (2014). Retention of first-year community college students. *Community College Journal of Research and Practice, 38*(5), 466-477.
- Wladis, C., Conway, K., & Hachey, A. C. (2017). Using course-level factors as predictors of online course outcomes: A multi-level analysis at a US urban community college. *Studies in Higher Education, 42*(1), 184-200.

- Wohlgemuth, D., Whalen, D., Sullivan, J., Nading, C., Shelley, M., & Wang, Y. (2007). Financial, academic, and environmental influences on the retention and graduation of students. *Journal of College Student Retention: Research, Theory & Practice*, 8(4), 457-475.
- Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7), 1341-1390.
- Wood, J. L. (2013). The same... but different: Examining background characteristics among Black males in public two-year colleges. *The Journal of Negro Education*, 82(1), 47-61.
- Wyman, F. J. (1997). A predictive model of retention rate at regional two-year colleges. *Community College Review*, 25(1), 29-58.
- Xu, D., & Dadgar, M. (2018). How effective are community college remedial math courses for students with the lowest math skills?. *Community College Review*, 46(1), 62-81.
- Xu, D., & Jaggars, S. S. (2011). *Online and hybrid course enrollment and performance in Washington State community and technical colleges*. New York, NY: Columbia University, Teachers College, Community College Research Center.
- Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Mining Education data to predict student's retention: a comparative study. *International Journal of Computer Science and Information Security*, 10(2), 113-117
- Yu, H. (2017). Factors associated with student academic achievement at community colleges. *Journal of College Student Retention: Research, Theory & Practice*, 19(2), 224-239.

- 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(2), 307-325.
- Zhang, Y., Oussena, S., Clark, T., & Kim, H. (2010). Use Data Mining to Improve Student Retention in Higher Education-A Case Study. *ICEIS*, 1, 190-197.
- Zhang, D., Wang, J., & Zhao, X. (2015, September). *Estimating the uncertainty of average F1 scores*. In Proceedings of the 2015 International Conference on The Theory of Information Retrieval (pp. 317-320).
- Zhu, W., Zeng, N., & Wang, N. (2010). Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations. *NESUG proceedings: Health care and life sciences, Baltimore, Maryland*. 19(2010), 67.
- Zumeta, W., Breneman, D. W., Callan, P. M., & Finney, J. E. (2012). *Financing American higher education in the era of globalization*. Cambridge: Education Press.

## **APPENDIX A:**

R Code for Modeling Building and Variable Importance

### ***#Load libraries into R environment***

```
library(readxl)
library(tidymodels)
library(tidyverse)
library(workflows)
library(tune)
library(mlbench)
library(stacks)
library(caret)
library(vip)
library(psych)
library(corrplot)
library(ggcorrplot)
library(RColorBrewer)
install.packages("Rcpp", dependencies = TRUE)
library(usethis)
library(devtools)
library(gplots)
library(car)
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("ComplexHeatmap")
library(naniar)
library(visdat)
library(ComplexHeatmap)
library(MVN)
library(blorr)
```

### ***#Import dataset into R environment***

```
FinalData <- read_excel("C:/Users/camil/Desktop/Chapter 4/Data and Code for Chapter
4/data/FinalData.xlsx")
```

### ***#Recode Gender and Race or Ethnicity***

```
FinalData <- FinalData %>%
  mutate(gender = ifelse(gender == "Male",0,1))

FinalData<- FinalData %>% mutate(race=recode(race,
  `White`= 0,
  `Black or African American`=1,
  `Hispanic or Latino`= 2,
  `Other`=3))
```

### ***#Creation of the Percaid variable***

```
FinalData <- FinalData %>% mutate(percaid = paid/award)
```



```
FinalData$percaid<- FinalData$percaid * 100
FinalData$percaid
```

### ***#Descriptive statistics by individual cohort***

```
describeBy(FinalData, group="term")
describe(FinalData)
```

### ***#Tallys for categorical variables by individual cohort and overall***

```
table(FinalData$race, FinalData$term)
table(FinalData$gender, FinalData$term)
table(FinalData$retain, FinalData$term)
table(FinalData$fasfa, FinalData$term)
table(FinalData$race)
table(FinalData$gender)
table(FinalData$retain)
table(FinalData$fasfa)
```

### ***#Correlations by individual cohorts and bagimpute on the dataset***

```
Correlationdata <- FinalData
FinalDataCorrecipe<-
recipe(retain ~ term + gender + race + hsgpa + age + credit + paid + award + remed +
       online + gpa + fasfa + percaid,
data = Correlationdata) %>%
step_bagimpute(all_numeric())
```

```
Correlationdata <- FinalDataCorrecipe %>%
prep(Correlationdata) %>%
juice()
```

### ***#Cohort one correlation matrix, p-values, and correlogram***

```
cohortonecorr <- Correlationdata %>%
  filter(term == 20182)
cohortonecorr <- cohortonecorr [c(4:11, 13)]
corr1 <- round(cor(cohortonecorr), 2)
corr1
p.mat1 <- cor_pmat(cohortonecorr)
p.mat1
```

```
corrplot(corr1, type = "upper", order = "hclust", p.mat = p.mat1, sig.level = 0.05, col =
  gray.colors(100), tl.col = 'black')
```

### ***#Cohort two correlation matrix, p-values, and correlogram***

```
cohorttwocorr <- Correlationdata %>%
  filter(term == 20192)
cohorttwocorr <- cohorttwocorr [c(4:11, 13)]
corr2 <- round(cor(cohorttwocorr), 2)
```

```

corr2
p.mat2 <- cor_pmat(cohorttwocorr)
p.mat2

corrplot(corr2, type = "upper", order = "hclust", p.mat = p.mat2, sig.level = 0.05, col =
  gray.colors(100), tl.col = 'black')

#Cohort together correlation matrix, p-values, and correlogram
cohortcorr <- Correlationdata
cohortcorr <- cohortcorr [c(4:11, 13)]
allcorr <- round(cor(cohortcorr), 2)
allcorr
allp.mat <- cor_pmat(cohortcorr)
allp.mat

corrplot(allcorr, type = "upper", order = "hclust", p.mat = allp.mat, sig.level = 0.05, col =
  gray.colors(100), tl.col = 'black')

#Categorical variable analysis
plotmeans(RETAIN ~ RACE, data = FinalDataRace, frame = FALSE)
plotmeans(RETAIN ~ GENDER, data = FinalDataRace, frame = FALSE)
plotmeans(RETAIN ~ FASFA, data = FinalDataRace, frame = FALSE)
plotmeans(FASFA ~ GENDER, data = FinalDataRace, frame = FALSE)
plotmeans(FASFA ~ RACE, data = FinalDataRace, frame = FALSE)

#Creation of training and test set before missing data, normality, and outlier checks
FinalData <- FinalData[c(2:14)]
set.seed(42)
#Split the data into training (70%) and testing (30%)
FinalData_split <- initial_split(FinalData,
  prop = 7/10, strata = retain)
FinalData_split
FinalData_train <- training(FinalData_split)
FinalData_test <- testing(FinalData_split)
FinalData_cv <- vfold_cv(FinalData_train, strata = retain)

#Histograms and QQplots for outlier detection
hist(FinalData_train$AGE, main="Histogram for Age", xlab="Age in Year")
qqPlot(FinalData_train$AGE, main="Qqplot for Age", ylab="Age in Years")
hist(FinalData_train$HSGPA, main="Histogram for High School GPA ", xlab="High
  School GPA")
qqPlot(FinalData_train$HSGPA, main="Qqplot for High School GPA", ylab=" High
  School GPA")
hist(FinalData_train$CREDIT, main="Histogram for Total Credit Hours Earned",
  xlab="Total Credit Hours Earned")

```

```

qqPlot(FinalData_train$CREDIT, main="Qqplot for Total Credit Hours Earned",
       ylab="Total Credit Hours Earned")
hist(FinalData_train$PAID, main="Histogram for Total Financial Amount Paid",
     xlab="Total Financial Amount Paid")
qqPlot(FinalData_train$PAID, main="Qqplot for Total Financial Amount Paid ",
       ylab="Total Financial Amount Paid")
hist(FinalData_train$AWARD, main="Histogram for Total Financial Amount Awarded",
     xlab="Total Financial Amount Awarded")
qqPlot(FinalData_train$AWARD, main="Qqplot for Total Financial Amount Awarded ",
       ylab="Total Financial Amount Awarded")
hist(FinalData_train$REMED, main="Histogram for Number of Remedial Courses",
     xlab=" Number of Remedial Courses ")
qqPlot(FinalData_train$REMED, main="Qqplot for Number of Remedial Courses ",
       ylab="Number of Remedial Courses")
hist(FinalData_train$ONLINE, main="Histogram for Percentage of Online Courses",
     xlab=" Percentage of Online Courses ")
qqPlot(FinalData_train$ONLINE, main="Qqplot for Percentage of Online Courses",
       ylab=" Percentage of Online Courses")
hist(FinalData_train$GPA, main="Histogram for GPA ", xlab="GPA")
qqPlot(FinalData_train$GPA, main="Qqplot for GPA", ylab=" GPA")

```

***#Missing data with heatmaps and missing data co-occurrence plots***

***#Training data***

```

convert_missing <- function(x) ifelse(is.na(x), 0, 1)
Train_missing <- FinalData_train [c(1:13)]
Final_missing <- apply(Train_missing, 2, convert_missing)

```

***#Test data***

```

convert_missing <- function(x) ifelse(is.na(x), 0, 1)
all_missing <- FinalData [c(1:13)]
Final_missing <- apply(all_missing, 2, convert_missing)

```

***#Heatmap***

```

Heatmap(
Final_missing,
name = "Missing", #title of legend
column_title = "Predictors", row_title = "Samples",
col = c("blue", "light blue"),
show_heatmap_legend = FALSE,
row_names_gp = gpar(fontsize = 0) # Text size for row names
)

```

***#Missing data co-occurrence Plot***

```

gg_miss_upset(FinalData, nsets = 12)

```

***#Univariate and multivariate tests of normality before outlier capping***

```

Trainnumeric <- FinalData_train [c(3:4, 6:11, 13)]

#Anderson-Darling test
resultad <- mvn(data = Trainnumeric, univariateTest = "AD", desc = TRUE)
resultad$univariateNormality
#Lillie test
resultL <- mvn(data = Trainnumeric, univariateTest = "Lillie", desc = TRUE)
resultL$univariateNormality
#Cramer test
resultCVM <- mvn(data = Trainnumeric, univariateTest = "CVM", desc = TRUE)
resultCVM$univariateNormality

#Random sample of 5000 to allow multivariate tests to work
Trainnumericssample <- sample_n(Trainnumeric, 5000)
#Mardia test
resultmar<- mvn(data = Trainnumericssample, mvnTest = "mardia")
resultmar$multivariateNormality
#Henze-Zirkler's MVN test
resulthz <- mvn(data = Trainnumericssample, mvnTest = "hz")
resulthz$multivariateNormality
#Doornik-Hansen's MVN test
resultdh <- mvn(data = Trainnumericssample, mvnTest = "dh")
resultdh$multivariateNormality

#Cook's d graph
mod <- lm(retain ~ ., data=FinalData_train)
cooks_d <- cooks.distance(mod)
plot(cooks_d, pch="*", cex=2, main="Influential Obs by Cooks distance")
abline(h = 4*mean(cooks_d, na.rm=T), col="red") # add cutoff line
text(x=1:length(cooks_d)+1, y=cooks_d, labels=ifelse(cooks_d>4*mean(cooks_d,
  na.rm=T),names(cooks_d,""), col="red") # add labels
influential <- as.numeric(names(cooks_d)[(cooks_d > 4*mean(cooks_d, na.rm=T))]) #
  influential row numbers

#Outlier capping and transformation for training set
Normalitytrain<- FinalData_train

qnHSGPA = quantile(Normalitytrain$HSGPA, c(0.00006, 0.99994), na.rm = TRUE)
Normalitytrain = within(Normalitytrain, {HSGPA = ifelse(HSGPA < qnHSGPA[1],
  qnHSGPA[1], HSGPA)
  HSGPA = ifelse(HSGPA > qnHSGPA[2], qnHSGPA[2], HSGPA)})
qnAGE = quantile(Normalitytrain$AGE, c(0.00006, 0.99994), na.rm = TRUE)
Normalitytrain = within(Normalitytrain, {AGE = ifelse(AGE < qnAGE[1], qnAGE[1],
  AGE)
  AGE = ifelse(AGE > qnAGE[2], qnAGE[2], AGE)})
qnCREDIT = quantile(Normalitytrain$CREDIT, c(0.00006, 0.99994), na.rm = TRUE)

```

```

Normalitytrain = within(Normalitytrain, {CREDIT = ifelse(CREDIT < qnCREDIT[1],
  qnCREDIT[1], CREDIT)
  CREDIT = ifelse(CREDIT > qnCREDIT[2], qnCREDIT[2], CREDIT)})
qnAWARD = quantile(Normalitytrain$AWARD, c(0.00006, 0.99994), na.rm = TRUE)
Normalitytrain = within(Normalitytrain, {AWARD = ifelse(AWARD < qnAWARD[1],
  qnAWARD[1], AWARD)
  AWARD = ifelse(AWARD > qnAWARD[2], qnAWARD[2], AWARD)})
qnPAID = quantile(Normalitytrain$PAID, c(0.00006, 0.99994), na.rm = TRUE)
Normalitytrain = within(Normalitytrain, {PAID = ifelse(PAID < qnPAID[1], qnPAID[1],
  PAID)
  PAID = ifelse(PAID > qnPAID[2], qnPAID[2], PAID)})
qnREMED = quantile(Normalitytrain$REMED, c(0.00006, 0.99994), na.rm = TRUE)
Normalitytrain = within(Normalitytrain, {REMED = ifelse(REMED < qnREMED[1],
  qnREMED[1], REMED)
  REMED = ifelse(REMED > qnREMED[2], qnREMED[2], REMED)})
qnONLINE = quantile(Normalitytrain$ONLINE, c(0.00006, 0.99994), na.rm = TRUE)
Normalitytrain = within(Normalitytrain, {ONLINE = ifelse(ONLINE < qnONLINE[1],
  qnONLINE[1], ONLINE)
  ONLINE = ifelse(ONLINE > qnONLINE[2], qnONLINE[2], ONLINE)})
qnGPA = quantile(Normalitytrain$GPA, c(0.00006, 0.99994), na.rm = TRUE)
Normalitytrain = within(Normalitytrain, {GPA = ifelse(GPA < qnGPA[1], qnGPA[1],
  GPA)
  GPA = ifelse(GPA > qnGPA[2], qnGPA[2], GPA)})

```

```

FinalDataNormrecipe<-
recipe(RETAIN ~ academic_term.x + GENDER + RACE + HSGPA + AGE + CREDIT
  + PAID + AWARD + REMED + ONLINE + GPA + FASFA,
data =Normalitytrain) %>%
step_bagimpute(all_numeric()) %>%
step_YeoJohnson(all_numeric())

```

```

Normalitytrain <- FinalDataNormrecipe %>%
prep(Normalitytrain) %>%
juice()

```

### ***#Histograms and QQplots after outlier capping and transformation***

```

hist(Normalitytrain$AGE, main="Histogram for Age", xlab="Age in Year")
qqPlot(Normalitytrain$AGE, main="Qqplot for Age", ylab="Age in Years")
hist(Normalitytrain$HSGPA, main="Histogram for High School GPA ", xlab="High
  School GPA")
qqPlot(Normalitytrain$HSGPA, main="Qqplot for High School GPA", ylab=" High
  School GPA")
hist(Normalitytrain$CREDIT, main="Histogram for Total Credit Hours Earned",
  xlab="Total Credit Hours Earned")
qqPlot(Normalitytrain$CREDIT, main="Qqplot for Total Credit Hours Earned",
  ylab="Total Credit Hours Earned")

```

```

hist(Normalitytrain$PAID, main="Histogram for Total Financial Amount Paid",
      xlab="Total Financial Amount Paid")
qqPlot(Normalitytrain$PAID, main="Qqplot for Total Financial Amount Paid ",
        ylab="Total Financial Amount Paid")
hist(Normalitytrain$AWARD, main="Histogram for Total Financial Amount Awarded",
      xlab="Total Financial Amount Awarded")
qqPlot(Normalitytrain$AWARD, main="Qqplot for Total Financial Amount Awarded ",
        ylab="Total Financial Amount Awarded")
hist(Normalitytrain$REMED, main="Histogram for Number of Remedial Courses",
      xlab=" Number of Remedial Courses ")
qqPlot(Normalitytrain$REMED, main="Qqplot for Number of Remedial Courses ",
        ylab="Number of Remedial Courses")
hist(Normalitytrain$ONLINE, main="Histogram for Percentage of Online Courses",
      xlab=" Percentage of Online Courses ")
qqPlot(Normalitytrain$ONLINE, main="Qqplot for Percentage of Online Courses",
        ylab=" Percentage of Online Courses")
hist(Normalitytrain$GPA, main="Histogram for GPA ", xlab="GPA")
qqPlot(Normalitytrain$GPA, main="Qqplot for GPA", ylab=" GPA")

```

```

#Univariate & multivariate tests of normality after outlier capping and transformation
Normalitynumeric <- Normalitytrain [c(4:5, 7:12)]

```

***#Anderson-Darling test***

```

resultad <- mvn(data = Normalitynumeric, univariateTest = "AD", desc = TRUE)
resultad$univariateNormality

```

***#Lillie test***

```

resultL <- mvn(data = Normalitynumeric, univariateTest = "Lillie", desc = TRUE)
resultL$univariateNormality

```

***#Cramer test***

```

resultCVM <- mvn(data = Normalitynumeric, univariateTest = "CVM", desc = TRUE)
resultCVM$univariateNormality

```

***#Random sample of 5000 to allow tests to work***

```

Normalitynumericssample <- sample_n(Normalitynumeric, 5000)

```

***#Mardia test***

```

resultmar<- mvn(data = Normalitynumericssample, mvnTest = "mardia")
resultmar$multivariateNormality

```

***#Henze-Zirkler's MVN test***

```

resulthz <- mvn(data = Normalitynumericssample, mvnTest = "hz")
resulthz$multivariateNormality

```

***#Doornik-Hansen's MVN test***

```

resultdh <- mvn(data = Normalitynumericssample, mvnTest = "dh")
resultdh$multivariateNormality

```

***#Cook's d graph***

```

mod <- lm(RETAIN ~ ., data=Normalitytrain)

```

```

cooks_d <- cooks.distance(mod) plot(cooks_d, pch="*", cex=2, main="Influential Obs by
  Cooks distance")
abline(h = 4*mean(cooks_d, na.rm=T), col="red") # add cutoff line
text(x=1:length(cooks_d)+1, y=cooks_d, labels=ifelse(cooks_d>4*mean(cooks_d,
  na.rm=T),names(cooks_d),""), col="red") # add labels
influential <- as.numeric(names(cooks_d)[(cooks_d > 4*mean(cooks_d, na.rm=T))]) #
  influential row numbers

```

***#Data mining models creation code with original data. Some steps were previously  
#used but the modeling software reruns them.***

***#Recode and add factors***

```

finaldata$gender = as.factor(finaldata$gender)
finaldata$race = as.factor(finaldata$race)
finaldata$fasfa = as.factor(finaldata$fasfa)
finaldata$retain = as.factor(finaldata$retain)

```

***#Percaid variable creation***

```

finaldata <- finaldata %>% mutate(percaid = paid/award)
finaldata$percaid<- finaldata$percaid * 100

```

***#Train and test sets with new variable***

```

set.seed(42)
# Split the data into training (70%) and testing (30%)
finaldata_split <- initial_split(finaldata,
  prop =7/10, strata = retain)

finaldata_split
finaldata_train <- training(finaldata_split)
finaldata_test <- testing(finaldata_split)
summary(finaldata_train)

```

***#Outlier capping and creation of cross validation file***

```

qnhs_gpa = quantile(finaldata_train$hsgpa, c(0.00006, 0.99994), na.rm = TRUE)
finaldata_train = within(finaldata_train, {hsgpa = ifelse(hsgpa < qnhs_gpa[1], qnhs_gpa[1],
  hsgpa)
  hsgpa = ifelse(hsgpa > qnhs_gpa[2], qnhs_gpa[2], hsgpa)})
qnage = quantile(finaldata_train$age, c(0.00006, 0.99994), na.rm = TRUE)
finaldata_train = within(finaldata_train, {age = ifelse(age < qnage[1], qnage[1], age)
  age = ifelse(age > qnage[2], qnage[2], age)})
qncredit = quantile(finaldata_train$credit, c(0.00006, 0.99994), na.rm = TRUE)
finaldata_train = within(finaldata_train, {credit = ifelse(credit < qncredit[1], qncredit[1],
  credit)
  credit = ifelse(credit > qncredit[2], qncredit[2], credit)})
qnaward = quantile(finaldata_train$award, c(0.00006, 0.99994), na.rm = TRUE)
finaldata_train = within(finaldata_train, {award = ifelse(award < qnaward[1],
  qnaward[1], award)

```

```

award = ifelse(award > qnaward[2], qnaward[2], award)})
qnpaid = quantile(finaldata_train$paid, c(0.00006, 0.99994), na.rm = TRUE)
finaldata_train = within(finaldata_train, {paid = ifelse(paid < qnpaid[1], qnpaid[1], paid)
paid = ifelse(paid > qnpaid[2], qnpaid[2], paid)})
qnremed = quantile(finaldata_train$remed, c(0.00006, 0.99994), na.rm = TRUE)
finaldata_train = within(finaldata_train, {remed = ifelse(remed < qnremed[1],
qnremed[1], remed)
remed = ifelse(remed > qnremed[2], qnremed[2], remed)})
qnonline = quantile(finaldata_train$online, c(0.00006, 0.99994), na.rm = TRUE)
finaldata_train = within(finaldata_train, {online = ifelse(online < qnonline[1],
qnonline[1], online)
online = ifelse(online > qnonline[2], qnonline[2], online)})
qngpa = quantile(finaldata_train$gpa, c(0.00006, 0.99994), na.rm = TRUE)
finaldata_train = within(finaldata_train, {gpa = ifelse(gpa < qngpa[1], qngpa[1], gpa)
gpa = ifelse(gpa > qngpa[2], qngpa[2], gpa)})
qnpercaid = quantile(finaldata_train$percaid, c(0.00006, 0.99994), na.rm = TRUE)
finaldata_train = within(finaldata_train, {percaid = ifelse(percaid < qnpercaid[1],
qnpercaid[1], percaid)
percaid = ifelse(percaid > qnpercaid[2], qnpercaid[2], percaid)})
summary(finaldata_train)

```

### ***#Cross validation***

```
finaldata_cv <- vfold_cv(finaldata_train, strata = retain)
```

### ***#Recipe step and universal workflow created***

```
finaldatarecipe<-
```

### ***#which consists of the formula (outcome ~ predictors)***

```
recipe(retain ~ gender + race + age + hsgpa + credit + paid + award + remed + online +
gpa + fasfa + percaid,
data = finaldata_train) %>%
```

### ***#and some pre-processing steps***

```
step_bagimpute(all_numeric()) %>%
step_YeoJohnson(all_numeric()) %>%
step_dummy(all_nominal(), -all_outcomes())%>%
step_interact(terms = ~ credit:gpa) %>%
step_interact(terms = ~ credit:online) %>%
step_interact(terms = ~ credit:award) %>%
step_interact(terms = ~ credit:paid) %>%
step_interact(terms = ~ credit:percaid) %>%
step_interact(terms = ~ credit:remed) %>%
step_interact(terms = ~ credit:gpa:online) %>%
step_interact(terms = ~ credit:gpa:remed) %>%
step_interact(terms = ~ credit:online:remed) %>%
step_interact(terms = ~ gpa:paid) %>%
step_interact(terms = ~ gpa:online) %>%
step_interact(terms = ~ gpa:award) %>%

```



```

step_interact(terms = ~ gpa:percaid) %>%
step_interact(terms = ~ gpa:remed) %>%
step_normalize(age, hsgpa, credit, paid, award, remed, online, gpa, percaid)

```

```

finaldatarecipe %>% prep() %>% juice() %>% summary()

```

```

finaldataworkflow <- workflow() %>%
  add_recipe(finaldatarecipe)
finaldataprep <- prep(finaldatarecipe)

```

### ***#Defining tuning control***

```

ctrl_grid <- control_grid(save_pred = TRUE, save_workflow = TRUE)
model_metrics <- metric_set(roc_auc, accuracy, spec, f_meas, sens)

```

### **#Random Forest Training Model with VIP**

```

set.seed(42)
library(randomForest)
random_forest_model <- rand_forest(
  mtry = tune(), trees = tune(), min_n = tune()) %>%
  set_mode("classification") %>%
  set_engine("randomForest")

```

```

random_forest_workflow <- finaldataworkflow %>%
  add_model(random_forest_model)

```

```

random_forest_res <-
  tune_grid(
    random_forest_workflow,
    resamples = finaldata_cv,
    metrics = model_metrics,
    control = ctrl_grid,
    grid = 20
  )

```

```

rf <- random_forest_res
rfvip <- vip(rf, method = "permute", target = "retain", metric = "roc_auc",
  pred_wrapper = predict) + ggtitle("RF")

```

```

random_forest_pred <- random_forest_res %>%
  collect_predictions()

```

### ***#Confusion Matrix for training set***

```

random_forest_pred %>%
  conf_mat(truth = retain, estimate = .pred_class)

```

```

random_forest_train_resultsallmetrics <- random_forest_res %>%

```

```

collect_metrics()
random_forest_train_resultsallmetrics

write_csv(random_forest_train_resultsallmetrics, path =
  "random_forest_train_resultsallmetrics.csv")
best_auc_rf <- select_best(random_forest_res, metric = "roc_auc")
best_auc_rf

```

### ***#ROC curve for training***

```

rf_roc_curve <- random_forest_res %>%
  show_best(metric = "roc_auc")

```

```

random_forest_auc <-
  random_forest_res %>%
  collect_predictions(parameters = rf_roc_curve) %>%
  roc_curve(retain, .pred_0) %>%
  mutate(model = "Random Forest")

```

```

autoplot(random_forest_auc)

```

### ***#Variable Importance for Random Forest Training Model***

```

set.seed(42)
train_random_forest_impvip <- train_random_forest_impvip <- finaldataworkflow %>%
  add_model(final_random_forest_mod) %>%
  fit(finaldata_train) %>%
  pull_workflow_fit() %>%
  vi(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
train_random_forest_impvip

```

### ***#Variable Importance Plots for Random Forest Training Model***

```

set.seed(42)
train_random_forest_impvipnonplot <- train_random_forest_impvipnonplot <-
finaldataworkflow %>%
  add_model(final_random_forest_mod) %>%
  fit(finaldata_train) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )

```

```

train_random_forest_impvipnonplot

set.seed(42)
train_random_forest_impvipretplot <- train_random_forest_impvipretplot <-
finaldataworkflow %>%
  add_model(final_random_forest_mod) %>%
  fit(finaldata_train) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "1",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
train_random_forest_impvipretplot
write_csv(train_random_forest_impvip, path = "train_random_forest_impvip.csv")

#Final Random Forest Model using the highest ROC_AUC value
final_random_forest_mod <- finalize_model(
  random_forest_model,
  best_auc_rf)
final_random_forest_mod

test_random_forest_workflow <-
  random_forest_workflow %>%
  update_model(final_random_forest_mod)

set.seed(42)
test_random_forest_fit <-
  test_random_forest_workflow %>%
  last_fit(finaldata_split)

test_random_forest_fit %>%
  collect_metrics()

test_random_forest_predict <- test_random_forest_fit %>%
  collect_predictions()

#ROC curve for testing
rf_test_roc_curve <- test_random_forest_fit %>%
  show_best(metric = "roc_auc")

test_random_forest_auc <-
  test_random_forest_fit %>%
  collect_predictions(parameters = rf_test_roc_curve) %>%
  roc_curve(retain, .pred_0) %>%
  mutate(model = "Random Forest")

```

```

autoplot(test_random_forest_auc)

#Confusion matrix for the testing
test_random_forest_predict %>%
  conf_mat(truth = retain, estimate = .pred_class)

#Variable Importance for Random Forest Testing Model
set.seed(42)
test_random_forest_impvip <- test_random_forest_impvip <- test_random_forest_fit
  %>%
  pluck(".workflow", 1) %>%
  pull_workflow_fit() %>%
  vi(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
test_random_forest_impvip

#Variable Importance Plots for Random Forest Testing Model
set.seed(42)
test_random_forest_impvipnonplot <- test_random_forest_impvipnonplot <-
  test_random_forest_fit %>%
  pluck(".workflow", 1) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
test_random_forest_impvipnonplot

set.seed(42)
test_random_forest_impvipretplot <- test_random_forest_impvipretplot <-
  test_random_forest_fit %>%
  pluck(".workflow", 1) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "1",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
test_random_forest_impvipretplot

write_csv(test_random_forest_impvip, path = "test_random_forest_impvip.csv")

```

```

write_csv(test_random_forest_auc, path = "test_random_forest_auc.csv")

#Logistic Regression Training Model with VIP
set.seed(42)
lr_model <-
logistic_reg() %>%
set_engine("glm") %>%
set_mode("classification")

lr_workflow <- finaldataworkflow %>%
  add_model(lr_model)

lr_res <-
  tune_grid(
    lr_workflow,
    resamples = finaldata_cv,
    metrics = model_metrics,
    control = ctrl_grid,
    grid = 20
  )

lr_pred <- lr_res %>%
  collect_predictions()
lr_pred

#Logistic regression coefficients
prepped_recipe <- prep(finaldatarecipe, training = finaldata_train)
train_preprocessed <- bake(prepped_recipe, finaldata_train)
lr_reg <- glm(retain ~., data = train_preprocessed, family = binomial(link = 'logit'))

lr_reg %>%
  blr_step_aic_both() %>%
  plot()

blr_model_fit_stats(lr_reg)
blr_test_hosmer_lemeshow(lr_reg)

#Confusion matrix for training set
lr_pred %>%
  conf_mat(truth = retain, estimate = .pred_class)

lr_train_resultsallmetrics <- lr_res %>%
  collect_metrics()
lr_train_resultsallmetrics

write_csv(lr_train_resultsallmetrics, path = "lr_train_resultsallmetrics.csv")

```

```

best_auc_lr <- select_best(lr_res, metric = "roc_auc")
best_auc_lr

#ROC curve for training set
lr_roc_curve <- lr_res %>%
  show_best(metric = "roc_auc")

lr_auc <-
  lr_res %>%
  collect_predictions(parameters = lr_roc_curve) %>%
  roc_curve(retain, .pred_0) %>%
  mutate(model = "Logistic Regression")
autoplot(lr_auc)

#Final Logistic Regression Model using the highest ROC_AUC value
final_lr_mod <- finalize_model(
  lr_model,
  best_auc_lr)
final_lr_mod

#Variable Importance for Logistic Regression Training Model
set.seed(42)
train_lr_impvip <- train_lr_impvip <- finaldataworkflow %>%
  add_model(final_lr_mod) %>%
  fit(finaldata_train) %>%
  pull_workflow_fit() %>%
  vi(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
train_lr_impvip

#Variable Importance Plots for Logistic Regression Training Model
set.seed(42)
train_lr_impvipnonplot <- train_lr_impvipnonplot <- finaldataworkflow %>%
  add_model(final_lr_mod) %>%
  fit(finaldata_train) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
train_lr_impvipnonplot

```

```

set.seed(42)
train_lr_impvipretplot <- train_lr_impvipretplot <- finaldataworkflow %>%
  add_model(final_lr_mod) %>%
  fit(finaldata_train) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "1",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
train_lr_impvipretplot
write_csv(train_lr_impvip, path = "train_lr_impvip.csv")

```

### ***#Logistic Regression Testing Model with VIP***

```

test_lr_workflow <-
  lr_workflow %>%
  update_model(final_lr_mod)

```

```

set.seed(42)
test_lr_fit <-
  test_lr_workflow %>%
  last_fit(finaldata_split)

```

```

test_lr_fit %>%
  collect_metrics()

```

```

lr_test_roc_curve <- test_lr_fit %>%
  show_best(metric = "roc_auc")

```

### ***#ROC curve for testing set***

```

test_lr_auc <-
  test_lr_fit %>%
  collect_predictions(parameters = lr_test_roc_curve) %>%
  roc_curve(retain, .pred_0) %>%
  mutate(model = "Logistic Regression")
autoplot(test_lr_auc)

```

```

test_lr_predict <- test_lr_fit %>%
  collect_predictions()

```

### ***#Confusion matrix for testing set***

```

test_lr_predict %>%
  conf_mat(truth = retain, estimate = .pred_class)

```

### ***#Odds ratios***

```

lr_odds <- lr_odds <- test_lr_fit$.workflow[[1]] %>%
  tidy(exponentiate = TRUE)
#Logistic regression coefficients
testprepped_recipe <- prep(finaldatarecipedata, training = finaldata_test)
test_preprocessed <- bake(testprepped_recipe, finaldata_test)
testlr_reg <- glm(retain ~., data = test_preprocessed, family = binomial(link = 'logit'))

testlr_reg %>%
 blr_step_aic_both() %>%
  plot()
blr_model_fit_stats(testlr_reg)
blr_test_hosmer_lemeshow(testlr_reg)

#Variable Importance for Logistic Regression Testing Model
set.seed(42)
test_lr_impvip <- test_lr_impvip <- test_lr_fit %>%
  pluck(".workflow", 1) %>%
  pull_workflow_fit() %>%
  vi(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
test_lr_impvip

#Variable Importance Plots for Logistic Regression Testing Model
set.seed(42)
test_lr_impvipnonplot <- test_lr_impvipnonplot <- test_lr_fit %>%
  pluck(".workflow", 1) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
test_lr_impvipnonplot

set.seed(42)
test_lr_impvipretplot <- test_lr_impvipretplot <- test_lr_fit %>%
  pluck(".workflow", 1) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "1",
    pred_wrapper = predict, train = juice(finaldataprep)
  )

```



```

test_lr_impvipretplot
write_csv(test_lr_impvip, path = " test_lr_impvip.csv")
write_csv(test_lr_auc, path = " test_lr_auc.csv")

#Support Vector Machines Radial Training Model with VIP
set.seed(42)
svmr_model <-
  svm_rbf(
    cost = tune(),
    rbf_sigma = tune()
  ) %>%
  set_engine("kernlab") %>%
  set_mode("classification")

svmr_workflow <- finaldataworkflow %>%
  add_model(svmr_model)

svmr_res <-
  tune_grid(
    svmr_workflow,
    resamples = finaldata_cv,
    metrics = model_metrics,
    control = ctrl_grid,
    grid = 20
  )

svmr_pred <- svmr_res %>%
  collect_predictions()
svmr_pred

#Confusion matrix for training set
svmr_pred %>%
  conf_mat(truth = retain, estimate = .pred_class)

svmr_train_resultsallmetrics <- svmr_res %>%
  collect_metrics()
svmr_train_resultsallmetrics
write_csv(svmr_train_resultsallmetrics, path = "svmr_train_resultsallmetrics.csv")

best_auc_svmr <- select_best(svmr_res, metric = "roc_auc")
best_auc_svmr

#ROC curve for training
svmr_roc_curve <- svmr_res %>%
  show_best(metric = "roc_auc")

```

```

svmr_auc <-
  svmr_res %>%
  collect_predictions(parameters = svmr_roc_curve) %>%
  roc_curve(retain, .pred_0) %>%
  mutate(model = "Support Vector Machine Radial")
autoplot(svmr_auc)

```

### ***#Final SVMR Model using the highest ROC\_AUC value***

```

final_svmr_mod <- finalize_model(
  svmr_model,
  best_auc_svmr)
final_svmr_mod

```

### ***#Variable Importance for SVMR Training Model***

```

set.seed(42)
train_svmr_impvip <- train_svmr_impvip <- finaldataworkflow %>%
  add_model(final_svmr_mod) %>%
  fit(finaldata_train) %>%
  pull_workflow_fit() %>%
  vi(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
train_svmr_impvip

```

### ***#Variable Importance Plots for SVMR Training Model***

```

train_svmr_impvipnonplot <- train_svmr_impvipnonplot <- finaldataworkflow %>%
  add_model(final_svmr_mod) %>%
  fit(finaldata_train) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
train_svmr_impvipnonplot

```

```

train_svmr_impvipretplot <- train_svmr_impvipretplot <- finaldataworkflow %>%
  add_model(final_svmr_mod) %>%
  fit(finaldata_train) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "1",

```

```

    pred_wrapper = predict, train = juice(finaldataprep)
  )
train_svmr_impvipretplot
write_csv(train_svmr_impvip, path = "train_svmr_impvip.csv")
write_csv(train_svmr_impvip, path = "train_svmr_impvip.csv")

```

### ***#SVMR Testing Model with VIP***

```

test_svmr_workflow <-
  svmr_workflow %>%
  update_model(final_svmr_mod)

set.seed(42)
test_svmr_fit <-
  test_svmr_workflow %>%
  last_fit(finaldata_split)

test_svmr_fit %>%
  collect_metrics()

test_svmr_pred <- test_svmr_fit %>%
  collect_predictions()

```

### ***#ROC curve for testing***

```

test_svmr_auc <-
  test_svmr_fit %>%
  collect_predictions() %>%
  roc_curve(retain, .pred_0) %>%
  mutate(model = "Support Vector Machine Radial")
autoplot(test_svmr_auc)

```

### ***#Confusion Matrix for testing***

```

test_svmr_pred %>%
  conf_mat(truth = retain, estimate = .pred_class)

```

### ***#Variable Importance for SVMR Testing Model***

```

set.seed(42)
test_svmr_impvip <- test_svmr_impvip <- test_svmr_fit %>%
  pluck(".workflow", 1) %>%
  pull_workflow_fit() %>%
  vi(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
test_svmr_impvip

```

### ***#Variable Importance plots for SVMR Testing Model***

```
set.seed(42)
test_svmr_impvipnonplot <- test_svmr_impvipnonplot <- test_svmr_fit %>%
  pluck(".workflow", 1) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
test_svmr_impvipnonplot
```

```
set.seed(42)
test_svmr_impvipretplot <- test_svmr_impvipretplot <- test_svmr_fit %>%
  pluck(".workflow", 1) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "1",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
test_svmr_impvipretplot
write_csv(test_svmr_impvip, path = "test_svmr_impvip.csv")
write_csv(test_svmr_auc, path = "test_svmr_auc.csv")
```

### ***#Support Vector Machines Polynomial Training Model with VIP***

```
set.seed(42)
library(kernlab)
svmp_model <-
  svm_poly(cost = tune()) %>%
  set_engine("kernlab") %>%
  set_mode("classification")

svmp_workflow <- finaldataworkflow %>%
  add_model(svmp_model)

svmp_res <-
  tune_grid(
    svmp_workflow,
    resamples = finaldata_cv,
    metrics = model_metrics,
    control = ctrl_grid,
    grid = 20
  )
```

```

svmp_pred <- svmp_res %>%
  collect_predictions()
svmp_pred

#Confusion Matrix for training set
svmp_pred %>%
  conf_mat(truth = retain, estimate = .pred_class)

svmp_train_resultsallmetrics <- svmp_res %>%
  collect_metrics()
svmp_train_resultsallmetrics
write_csv(svmp_train_resultsallmetrics, path = "svmp_train_resultsallmetrics.csv")

best_auc_svmp <- select_best(svmp_res, metric = "roc_auc")
best_auc_svmp

#ROC curve for training set
svmp_roc_curve <- svmp_res %>%
  show_best(metric = "roc_auc")

svmp_auc <-
  svmp_res %>%
  collect_predictions(parameters = svmp_roc_curve) %>%
  roc_curve(retain, .pred_0) %>%
  mutate(model = "Support Vector Machine Poly")
autoplot(svmp_auc)

#Final SVM Model using the highest ROC_AUC value
final_svmp_mod <- finalize_model(
  svmp_model,
  best_auc_svmp)
final_svmp_mod

#Variable Importance for SVM Training Model
set.seed(42)
train_svmp_impvip <- train_svmp_impvip <- finaldataworkflow %>%
  add_model(final_svmp_mod) %>%
  fit(finaldata_train) %>%
  pull_workflow_fit() %>%
  vi(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
train_svmp_impvip

```

### ***#Variable Importance Plot for SVM Training Model***

```
set.seed(42)
train_svm_impvipnonplot <- train_svm_impvipnonplot <- finaldataworkflow %>%
  add_model(final_svm_mod) %>%
  fit(finaldata_train) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
train_svm_impvipnonplot
```

```
set.seed(42)
train_svm_impvipretplot <- train_svm_impvipretplot <- finaldataworkflow %>%
  add_model(final_svm_mod) %>%
  fit(finaldata_train) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "1",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
train_svm_impvipretplot
write_csv(train_svm_impvip, path = "train_svm_impvip.csv")
```

### ***#SVM Testing Model with VIP***

```
test_svm_workflow <-
  svm_workflow %>%
  update_model(final_svm_mod)
```

```
set.seed(42)
test_svm_fit <-
  test_svm_workflow %>%
  last_fit(finaldata_split)
```

```
test_svm_fit %>%
  collect_metrics()
```

```
test_svm_predict <- test_svm_fit %>%
  collect_predictions()
```

### ***#ROC curve for testing***

```
test_svm_auc <-
  test_svm_fit %>%
  collect_predictions() %>%
```

```

roc_curve(retain, .pred_0) %>%
mutate(model = "Support Vector Machine Poly")

autoplot(test_svm_auc)

#Confusion Matrix for testing
test_svm_predict %>%
  conf_mat(truth = retain, estimate = .pred_class)

#Variable Importance for SVM Testing Model
set.seed(42)
test_svm_impvip <- test_svm_impvip <- test_svm_fit %>%
  pluck(".workflow", 1) %>%
  pull_workflow_fit() %>%
  vi(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
test_svm_impvip

#Variable Importance Plot for SVM Testing Model
set.seed(42)
test_svm_impvipnonplot <- test_svm_impvipnonplot <- test_svm_fit %>%
  pluck(".workflow", 1) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
test_svm_impvipnonplot

set.seed(42)
test_svm_impvipretplot <- test_svm_impvipretplot <- test_svm_fit %>%
  pluck(".workflow", 1) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "1",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
test_svm_impvipretplot
write_csv(test_svm_impvip, path = "test_svm_impvip.csv")
write_csv(test_svm_auc, path = "test_svm_auc.csv")

```

### ***#Neural Networks Training Model with VIP***

```
set.seed(42)
nn_nnet_model <-
  mlp(hidden_units = tune(), penalty = tune(), epochs = tune()) %>%
  set_engine("nnet", trace = 0) %>%
  set_mode("classification")

nn_nnet_workflow <- finaldataworkflow %>%
  add_model(nn_nnet_model)

nn_nnet_res <-
  tune_grid(
    nn_nnet_workflow,
    resamples = finaldata_cv,
    metrics = model_metrics,
    control = ctrl_grid,
    grid = 20
  )

nn_nnet_pred <- nn_nnet_res %>%
  collect_predictions()
nn_nnet_pred

#Confusion Matrix for training set
nn_nnet_pred %>%
  conf_mat(truth = retain, estimate = .pred_class)

nn_nnet_train_resultsallmetrics <- nn_nnet_res %>%
  collect_metrics()
nn_nnet_train_resultsallmetrics

write_csv(nn_nnet_train_resultsallmetrics, path = "nn_nnet_train_resultsallmetrics.csv")

#Final NN Model using the highest ROC_AUC value
best_auc_nn <- select_best(nn_nnet_res, metric = "roc_auc")
best_auc_nn

#ROC curve for training
nn_nnet_roc_curve <- nn_nnet_res %>%
  show_best(metric = "roc_auc")

nn_nnet_auc <-
  nn_nnet_res %>%
  collect_predictions(parameters = nn_nnet_roc_curve) %>%
```



```

roc_curve(retain, .pred_0) %>%
mutate(model = "Neural Network")

autoplot(nn_nnet_auc)

final_nn_nnet_mod <- finalize_model(
  nn_nnet_model,
  best_auc_nn)
final_nn_nnet_mod

#Variable Importance for NN Training Model
set.seed(42)
train_nn_nnet_impvip <- train_nn_nnet_impvip <- finaldataworkflow %>%
  add_model(final_nn_nnet_mod) %>%
  fit(finaldata_train) %>%
  pull_workflow_fit() %>%
  vi(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
train_nn_nnet_impvip

#Variable Importance Plot for NN Training Model
set.seed(42)
train_nn_nnet_impvipnonplot <- train_nn_nnet_impvipnonplot <- finaldataworkflow
  %>%
  add_model(final_nn_nnet_mod) %>%
  fit(finaldata_train) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
train_nn_nnet_impvipnonplot

set.seed(42)
train_nn_nnet_impvipretplot <- train_nn_nnet_impvipretplot <- finaldataworkflow %>%
  add_model(final_nn_nnet_mod) %>%
  fit(finaldata_train) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "1",
    pred_wrapper = predict, train = juice(finaldataprep)
  )

```

```

)
train_nn_nnet_impvipretplot
write_csv(train_nn_nnet_impvip, path = "train_nn_nnet_impvip.csv")

#NN Testing Model with VIP

test_nn_nnet_workflow <-
  nn_nnet_workflow %>%
  update_model(final_nn_nnet_mod)

set.seed(42)
test_nn_nnet_fit <-
  test_nn_nnet_workflow %>%
  last_fit(finaldata_split)

test_nn_nnet_fit %>%
  collect_metrics()

test_nn_nnet_predict <- test_nn_nnet_fit %>%
  collect_predictions()

#ROC curve for testing
test_nn_nnet_roc_curve <- test_nn_nnet_fit %>%
  show_best(metric = "roc_auc")

test_nn_nnet_auc <-
  test_nn_nnet_fit %>%
  collect_predictions(parameters = test_nn_nnet_roc_curve) %>%
  roc_curve(retain, .pred_0) %>%
  mutate(model = "Neural Network")
autoplot(test_nn_nnet_auc)

#Confusion Matrix for testing
test_nn_nnet_predict %>%
  conf_mat(truth = retain, estimate = .pred_class)

#Variable Importance for NN Testing Model
set.seed(42)
test_nn_nnet_impvip <- test_nn_nnet_impvip <- test_nn_nnet_fit %>%
  pluck(".workflow", 1) %>%
  pull_workflow_fit() %>%
  vi(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )

```

```
test_nn_nnet_impvip
```

### ***#Variable Importance Plot for NN Testing Model***

```
set.seed(42)
test_nn_nnet_impvipnonplot <- test_nn_nnet_impvipnonplot <- test_nn_nnet_fit %>%
  pluck(".workflow", 1) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "0",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
test_nn_nnet_impvipnonplot
```

```
set.seed(42)
test_nn_nnet_impvipretplot <- test_nn_nnet_impvipretplot <- test_nn_nnet_fit %>%
  pluck(".workflow", 1) %>%
  pull_workflow_fit() %>%
  vip(
    method = "permute", nsim = 10,
    target = "retain", metric = "auc", reference_class = "1",
    pred_wrapper = predict, train = juice(finaldataprep)
  )
test_nn_nnet_impvipretplot
write_csv(test_nn_nnet_impvip, path = "test_nn_nnet_impvip.csv")
write_csv(test_nn_nnet_auc, path = "test_nn_nnet_auc.csv")
```

### ***#Create ROC curves for all model training***

```
bind_rows(random_forest_auc, lr_auc, svmr_auc, svmp_auc, nn_nnet_auc) %>%
  ggplot(aes(x = 1 - specificity, y = sensitivity, col = model)) +
  geom_path(lwd = 1.5, alpha = 0.8) +
  geom_abline(lty = 3) +
  coord_equal() +
  scale_color_viridis_d(option = "plasma", end = .6)
```

### ***#Create ROC curves for all model test***

```
bind_rows(test_random_forest_auc, test_lr_auc, test_svmr_auc, test_svmp_auc,
  test_nn_nnet_auc) %>%
  ggplot(aes(x = 1 - specificity, y = sensitivity, col = model)) +
  geom_path(lwd = 1.5, alpha = 0.8) +
  geom_abline(lty = 3) +
  coord_equal() +
  scale_color_viridis_d(option = "plasma", end = .6)
```

## **APPENDIX B:**

R Code for Inferential Statistics Tests

***#Load libraries into R environment***

```
library(tidyverse)
library(ggpubr)
library(rstatix)
```

***#Mann-whitney for evaluation metrics from training to test set comparison***

***#ACCURACY EVALUATION METRIC***

```
acmw<-wilcox.test(phaseac ~ typeac, data=FinalTrainingandTestEV, na.rm=TRUE,
paired=FALSE, exact=FALSE, conf.int=TRUE)
print(acmw)
```

***#F1-VALUE EVALUATION METRIC***

```
f1mw<-wilcox.test(phasef1 ~ typef1, data=FinalTrainingandTestEV, na.rm=TRUE,
paired=FALSE, exact=FALSE, conf.int=TRUE)
print(f1mw)
```

***#ROC\_AUC EVALUATION METRIC***

```
rocmw<-wilcox.test(phaseroc ~ typeroc, data=FinalTrainingandTestEV, na.rm=TRUE,
paired=FALSE, exact=FALSE, conf.int=TRUE)
print(rocmw)
```

***#SENSITIVITY EVALUATION METRIC***

```
senmw<-wilcox.test(phasesen ~ typesen, data=FinalTrainingandTestEV, na.rm=TRUE,
paired=FALSE, exact=FALSE, conf.int=TRUE)
print(senmw)
```

***#SPECIFICITY EVALUATION METRIC***

```
spemw<-wilcox.test(phasespe ~ typespe, data=FinalTrainingandTestEV, na.rm=TRUE,
paired=FALSE, exact=FALSE, conf.int=TRUE)
print(spemw)
```

***#Friedman's and Wilcoxon's tests for evaluation metric comparison of training set***

***#ACCURACY EVALUATION METRIC***

***#Import accuracy evaluation metrics dataset into R environment***

```
library(readxl)
Accuracy <- read_excel("C:/Users/camil/Desktop/Chapter 4/March28
Code/Accuracy.xlsx")
View(Accuracy)
```

***#Descriptive statistics and boxplot of accuracy evaluation metrics***

```
Accuracy %>%
  group_by(type) %>%
  get_summary_stats(mean, type = "common")
ggboxplot(Accuracy, x = "type", y = "mean", add = "jitter")
```

***#Friedman's test of accuracy evaluation metrics***

```
accres.fried <- Accuracy %>% friedman_test(mean ~ type |.config)
accres.fried
Accuracy %>% friedman_effsize(mean ~ type |.config)
```

***#Wilcoxon's test of accuracy evaluation metrics***

```
pwc <- Accuracy %>%
  wilcox_test(mean ~ type, paired = TRUE, p.adjust.method = "bonferroni")
pwc
```

***#F1-VALUE EVALUATION METRIC***

***#Import F1-Value evaluation metrics dataset into R environment***

```
library(readxl)
FMeasure1 <- read_excel("C:/Users/camil/Desktop/Chapter 4/March28
Code/FMeasure1.xlsx")
View(FMeasure1)
```

***#Descriptive statistics and boxplot of F1-Value evaluation metrics***

```
FMeasure1 %>%
  group_by(type) %>%
  get_summary_stats(mean, type = "common")
ggboxplot(FMeasure1, x = "type", y = "mean")
```

***#Friedman's test of F1-Value evaluation metrics***

```
rocres.fried <- FMeasure1 %>% friedman_test(mean ~ type |.config)
rocres.fried
FMeasure1%>% friedman_effsize(mean ~ type |.config)
```

***#Wilcoxon's test of F1-Value evaluation metrics***

```
pwc <- FMeasure1 %>%
  wilcox_test(mean ~ type, paired = TRUE, p.adjust.method = "bonferroni")
pwc
```

***#ROC\_AUC EVALUATION METRIC***

***#Import ROC\_AUC evaluation metrics dataset into R environment***

```
library(readxl)
ROC <- read_excel("C:/Users/camil/Desktop/Chapter 4/March28 Code/ROC.xlsx")
View(ROC)
```

***#Descriptive statistics and boxplot of ROC\_AUC evaluation metrics***

```
ROC %>%
  group_by(type) %>%
  get_summary_stats(mean, type = "common")
ggboxplot(ROC, x = "type", y = "mean", add = "jitter")
```

***#Friedman's test of ROC\_AUC evaluation metrics***

```
rocres.fried <- ROC %>% friedman_test(mean ~ type |.config)
rocres.fried
ROC %>% friedman_effsize(mean ~ type |.config)
```

***#Wilcoxon's test of ROC\_AUC evaluation metrics***

```
pwc <- ROC %>%
  wilcox_test(mean ~ type, paired = TRUE, p.adjust.method = "bonferroni")
pwc
```

***#SENSITIVITY EVALUATION METRIC***

***#Import sensitivity evaluation metrics dataset into R environment***

```
library(readxl)
Sens <- read_excel("C:/Users/camil/Desktop/Chapter 4/March28 Code/Sens.xlsx")
View(Sens)
```

***#Descriptive statistics and boxplot of sensitivity evaluation metrics***

```
Sens %>%
  group_by(type) %>%
  get_summary_stats(mean, type = "common")
ggboxplot(Sens, x = "type", y = "mean")
```

***#Friedman's test of sensitivity evaluation metrics***

```
accres.fried <- Sens %>% friedman_test(mean ~ type |.config)
accres.fried
Sens %>% friedman_effsize(mean ~ type |.config)
```

***#Wilcoxon's test of sensitivity evaluation metrics***

```
pwc <- Sens %>%
  wilcox_test(mean ~ type, paired = TRUE, p.adjust.method = "bonferroni")
pwc
```

***#SPECIFICITY EVALUATION METRIC***

***#Import specificity evaluation metrics dataset into R environment***

```
library(readxl)
Spec <- read_excel("C:/Users/camil/Desktop/Chapter 4/March28 Code/Spec.xlsx")
View(Spec)
```

***#Descriptive statistics and boxplot of specificity evaluation metrics***

```
Spec %>%
  group_by(type) %>%
  get_summary_stats(mean, type = "common")
ggboxplot(Spec, x = "type", y = "mean")
```

***#Friedman's test of specificity evaluation metrics***

```
accres.fried <- Spec %>% friedman_test(mean ~ type |.config)
```

```
accres.fried  
Spec %>% friedman_effsize(mean ~ type |.config)
```

```
#Wilcoxon's test of specificity evaluation metrics
```

```
pwc <- Spec %>%  
  wilcox_test(mean ~ type, paired = TRUE, p.adjust.method = "bonferroni")  
pwc
```



**APPENDIX C:**

Institutional Review Board Protocol Exemption Report



**Institutional Review Board (IRB)  
For the Protection of Human Research Participants**

**PROTOCOL EXEMPTION REPORT**

---

**Protocol Number:** 04077-2020

**Responsible Researcher:** Camille Pace

**Supervising Faculty:** Dr. Lantry Brockmeier

**Project Title:** *Determining Academic, Background, and Financial Predictors of Community College First Year Retention using Data Mining Techniques.*

---

**INSTITUTIONAL REVIEW BOARD DETERMINATION:**

This research protocol is **Exempt** from Institutional Review Board (IRB) oversight under Exemption **Category 4**. 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](mailto:irb@valdosta.edu)) before continuing your research.

---

**ADDITIONAL COMMENTS:**

- *Upon completion of this research study all data must be securely maintained (locked file cabinet, password protected computer, etc.) and accessible only by the researcher 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](mailto:irb@valdosta.edu) to ensure an updated record of your exemption.*

---

*Elizabeth Ann Olphie*      09.28.2020  
Elizabeth Ann Olphie, IRB Administrator

Thank you for submitting an IRB application.  
Please direct questions to [irb@valdosta.edu](mailto:irb@valdosta.edu) or 229-253-2947.

---

Revised: 06.02.10

**APPENDIX D:**  
Data Sharing Agreement

**Data Sharing Agreement  
Between the  
Board of Regents of the University System of Georgia  
and Camille Pace**

This data sharing agreement (“Agreement”) is entered into by the Board of Regents of the University System of Georgia (“BOR”) and Camille Pace

Purpose of Agreement

The purpose of this Agreement is to:

Obtain data to create a predictive model for student retention using background, academic, and financial factors, which can serve as a guide for other community colleges to use when they are investigating their institution’s retention.

Data

The BOR will provide Camille Pace with the following data:

- The request is for data of two different cohorts of first-time freshmen who first attended Fall 2017 and Fall 2018, including their first four consecutive semesters of data.
- Schools in Study: Atlanta Metropolitan College, Coastal College of Georgia, Dalton State College, East Georgia State College, Georgia Highlands College, Gordon State College, South Georgia State College
- Academic Terms Requested: For Fall 17 cohort, data from Fall 17, Spring 18, Summer 2018, and Fall 2018. For Fall 18 cohort, data from Fall 18, Spring 19, Summer 2019, and Fall 2019
- See DED Variable Selection form for specific data elements requested.

BOR agrees to share data with Camille Pace in a manner that safeguards the confidentiality of student data as defined by the Federal Family Educational Rights and Privacy Act (FERPA) and other applicable laws and regulations. FERPA establishes a right of privacy for student data based on a rule of non-release of individually identifiable data to anyone outside the student's institution or to persons inside the institution who have no legitimate need for the information without the express written permission of the student. However, FERPA contains a limited exception to the general rule when information is used by educational organizations for the purposes of conducting research to improve instruction. This Agreement fits under this limited exception to FERPA. See 20 U.S.C. § 1232 g (b)(1)(F).

Specifically, BOR agrees to share data with [party name] under the following stipulations.

- The data will be used only for purposes outlined in this agreement:

To create a predictive model for student retention using background, academic, and financial factors, which can serve as a guide for other community colleges to use when they are investigating their institution's retention.

- The parties agree that the transmittal of data shall be done in a secure manner.
- Camille Pace will limit access to the data to staff who require the data to develop, exchange, maintain, analyze and evaluate information for the purposes outlined in this agreement. Camille Pace shall maintain records of those individuals who are allowed access to the data and shall assure that each person is fully cognizant of the restrictions placed upon use of the data and the restrictions upon its disclosure.
- The data will be maintained in a secure environment and shall not be shared with other parties except as authorized by federal and/or state law.
- Camille Pace will utilize their best efforts to maintain the confidentiality of the data.
- The linked data will be destroyed after use or two years after the BOR shares data with Camille Pace.
- Camille Pace will indemnify and hold the BOR harmless against any claim, loss, expense, or demand incurred by the BOR as a result of [party name]'s access and use of the data.
- Camille Pace will provide any findings to be presented/published from the data to BOR at least two weeks prior to presentation/publication.
- Small cell sizes ( $N < 10$ ) cannot be published.

#### Termination

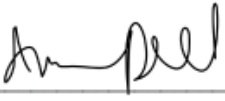
This Agreement shall take effect upon completion of signatures and remain in effect for one year or until terminated. This Agreement may be terminated by either BOR or [party name] upon notice to the other party. BOR may terminate this Agreement with or without cause at any time by providing written notice to Camille Pace thirty (30) calendar days prior to the termination date. Upon termination, all projects using the linked data must be immediately discontinued.

Board of Regents of the University System of Georgia

Name: Angela Bell

Title: Vice Chancellor for Research and Policy Analysis

Organization: Board of Regents of the University System of Georgia

Signature:  \_\_\_\_\_

Date: 12/15/20 \_\_\_\_\_

Name: Camille Pace

Title: Chair and Associate Professor, Doctoral Candidate

Organization: Georgia Highlands College, Valdosta State University

A handwritten signature in black ink that reads "Camille Pace". The signature is written in a cursive style with a large, prominent initial 'C'.

Signature: