



SCHOOL of  
GRADUATE STUDIES  
EAST TENNESSEE STATE UNIVERSITY

East Tennessee State University  
**Digital Commons @ East  
Tennessee State University**

---

Electronic Theses and Dissertations

Student Works


---

5-2018

# Using Data Science and Predictive Analytics to Understand 4-Year University Student Churn

Joshua Lee Whitlock  
*East Tennessee State University*

Follow this and additional works at: <https://dc.etsu.edu/etd>

 Part of the [Educational Assessment, Evaluation, and Research Commons](#), [Educational Leadership Commons](#), and the [Higher Education Administration Commons](#)

---

## Recommended Citation

Whitlock, Joshua Lee, "Using Data Science and Predictive Analytics to Understand 4-Year University Student Churn" (2018). *Electronic Theses and Dissertations*. Paper 3356. <https://dc.etsu.edu/etd/3356>

This Dissertation - Open Access is brought to you for free and open access by the Student Works at Digital Commons @ East Tennessee State University. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Digital Commons @ East Tennessee State University. For more information, please contact [digilib@etsu.edu](mailto:digilib@etsu.edu).

# Using Data Science and Predictive Analytics to Understand 4-Year University Student Churn

---

A dissertation

presented to

the faculty of the Department of Educational Leadership and Policy Analysis

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Doctor of Education in Educational Leadership

---

by

Joshua Lee Whitlock

May 2018

---

Dr. Don Good, Chair

Dr. Bethany Flora

Dr. Virginia Foley

Dr. Sally Lee

---

Keywords: Artificial neural network, data mining, data science, decision analytics, decision tree, predictive analytics, random forest, student retention, support vector machine, transfer students

## ABSTRACT

Using Data Science and Predictive Analytics to Understand 4-Year University Student Churn

by

Joshua Lee Whitlock

The purpose of this study was to discover factors about first-time freshmen that began at one of the six 4-year universities in the former Tennessee Board of Regents (TBR) system, transferred to any other institution after their first year, and graduated with a degree or certificate. These factors would be used with predictive models to identify these students prior to their initial departure. Thirty-four variables about students and the institutions that they attended and graduated from were used to perform principal component analysis to examine the factors involved in their decisions. A subset of 18 variables about these students in their first semester were used to perform principal component analysis and produce a set of 4 factors that were used in 5 predictive models. The 4 factors of students who transferred and graduated elsewhere were “Institutional Characteristics,” “Institution’s Focus on Academics,” “Student Aptitude,” and “Student Community.” These 4 factors were combined with the additional demographic variables of gender, race, residency, and initial institution to form a final dataset used in predictive modeling. The predictive models used were a logistic regression, decision tree, random forest, artificial neural network, and support vector machine. All models had predictive power beyond that of random chance. The logistic regression and support vector machine models had the most predictive power, followed by the artificial neural network, random forest, and decision tree models respectively.

Copyright © 2018  
Joshua Lee Whitlock  
All Rights Reserved

## DEDICATION

To my loved ones and those who love me.

## ACKNOWLEDGEMENTS

The list of individuals to thank is extensive. I want to thank Dr. Ramona Williams, Ms. Sheryl Burnette, and Mr. Paul Hayes for their faith in and support of me all these many years. I aspire to have your wisdom, kindness, and grace with others. I am thankful for Dr. Evelyn Roach, Dr. Mike Hoff, and Ms. Mary Ellen Musick for their support and encouragement. I am grateful to Dr. Brian Noland of ETSU and Dr. Emily House at the Tennessee Higher Education Commission for helping me to obtain the data for my dissertation. I am indebted to many of the faculty of the ELPA program. Dr. Bethany Flora, Dr. Don Good, Dr. Jasmine Renner, Dr. Pamela Scott and Dr. Sally Lee were excellent professors and guides on my educational journey.

I am very grateful for my committee. Dr. Good and Dr. Flora were excellent instructors throughout the ELPA program. Dr. Foley was a role model of leadership as Faculty Senate President when I first met her. Dr. Lee was perhaps the first professor I interacted with at ETSU nearly two decades ago, and it was an honor to have her guide my doctoral capstone work.

I would like to thank the leadership of ETSU, of the Faculty Senate, and of the Staff Senate. My associations with those groups provided a unique opportunity to witness and participate in university governance. Those experiences were an indispensable part to of my education in the ELPA program

I want to thank my many friends. Robert Davidson, Gary Pleasant, and Ryan and Ashley Seale helped me retain my sanity in times that I felt overwhelmed throughout this work. I am grateful to fellow classmates in the ELPA program. There are many more friends who are part of the ETSU community and beyond that I am thankful for in helping me complete this work.

Finally, I am eternally grateful for the love and support of my family. My parents, Joy and Larry Whitlock, as well as my in-laws, Mike and Brenda McCalmont, have always

supported my education. My wife, Trisha, and my sons, Alexander and Elijah, were always patient with me as I completed my studies. My family has sustained me through this work, and I am very grateful for them.

## TABLE OF CONTENTS

	Page
ABSTRACT.....	2
DEDICATION.....	4
ACKNOWLEDGEMENTS.....	5
LIST OF TABLES.....	11
LIST OF FIGURES.....	13
Chapter	
1. INTRODUCTION.....	15
Statement of the Problem.....	19
Research Questions.....	19
Significance of the Study.....	20
Definitions of Terms.....	21
Limitations and Delimitations.....	25
Overview of the Study.....	27
2. LITERATURE REVIEW.....	29
The Rising Cost of Higher Education.....	29
Performance Funding.....	32
Additional Calls for Accountability.....	34
Available Data.....	35
Retention.....	35
Transfer Students.....	40
Data Mining in Higher Education.....	42
Data Mining Algorithms.....	44



Classification Algorithms .....	45
Clustering Algorithms.....	47
Data Mining and Classical Statistics.....	47
Factor Analysis .....	48
Data Mining Tools .....	49
Classifying Predictive Value.....	50
Predictive Model Comparison .....	52
Measuring Model Effectiveness .....	54
Pitfalls to Avoid .....	60
Data Mining Ethics .....	61
Chapter Summary .....	65
3. RESEARCH METHODOLOGY .....	69
Research Questions and Null Hypotheses .....	69
Population .....	72
Instrumentation .....	73
Data Collection .....	73
Data Analysis .....	78
Chapter Summary .....	83
4. DATA ANALYSIS AND RESULTS.....	84
Categorical Analysis .....	85
Research Question #1 .....	101
Research Question #2 .....	105
Research Question #3 .....	109
Research Question #4 .....	113
Research Question #5 .....	117

Research Question #6 .....	120
Research Question #7 .....	123
Research Question #8 .....	126
Research Question #9 .....	129
Research Question #10 .....	132
Chapter Summary .....	134
<b>5. SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS.....</b>	<b>135</b>
Summary of the Findings.....	136
Research Question #1 .....	137
Research Question #2 .....	137
Research Question #3 .....	138
Research Question #4 .....	139
Research Question #5 .....	140
Research Question #6 .....	142
Research Question #7 .....	143
Research Question #8 .....	144
Research Question #9 .....	145
Research Question #10 .....	146
Conclusions.....	147
Recommendations for Practice .....	148
Recommendations for Future Research .....	149
REFERENCES .....	160
APPENDICES .....	175
Appendix A – Data Sources from THEC .....	175
Appendix B – Predictive Models Python Code .....	178

Appendix C – Sample SPSS Factor Analysis Code.....	183
VITA.....	186

## LIST OF TABLES

Table	Page
1. Data Files .....	75
2. Merged Dataset .....	76
3. Factors for Graduation .....	103
4. Factors for Graduation from First Semester Information .....	105
5. Factors for Transfers .....	107
6. Factors for Transfers from First Semester Information .....	109
7. Factors for Nongraduates .....	111
8. Factors for Nongraduates from First Semester Information .....	113
9. Factors for Transfer Graduates .....	115
10. Factors for Transfer Graduates from First Semester Information .....	117
11. Logistic Regression Confusion Matrix .....	118
12. Logistic Regression Classification Report.....	118
13. C5 Decision Tree Confusion Matrix.....	121
14. C5 Decision Tree Classification Report .....	121
15. Random Forest Confusion Matrix .....	124
16. Random Forest Classification Report .....	124
17. Artificial Neural Network Confusion Matrix .....	127
18. Artificial Neural Network Classification Report .....	127
19. Support Vector Machine Confusion Matrix .....	130
20. Support Vector Machine Classification Report .....	130
21. Comparison of Model ROC AUC .....	134
22. Factors for Transfer Graduates, West Tennessee University.....	150
23. Factors for Graduation, Central Tennessee University.....	152

24. Factors for Transfers, North Central Tennessee University .....154

25. Factors for Nongraduates, East Central Tennessee University.....156

## LIST OF FIGURES

Figure	Page
1. Receiver Operating Characteristic (ROC) .....	56
2. Confusion Matrix .....	57
3. k Fold Validation .....	59
4. Multilayer Perception ANN .....	80
5. Support Vector Machine .....	81
6. Transfer Graduates Factor 1 Histogram.....	86
7. Transfer Graduates Factor 2 Histogram.....	87
8. Transfer Graduates Factor 3 Histogram.....	88
9. Transfer Graduates Factor 4 Histogram.....	89
10. Mann-Whitney U Test for Gender, Factor 1 .....	90
11. Mann-Whitney U Test for Gender, Factor 2.....	91
12. Mann-Whitney U Test for Gender, Factor 3.....	91
13. Mann-Whitney U Test for Gender, Factor 4.....	92
14. Kruskal-Wallis Test for Race, Factor 1 .....	93
15. Kruskal-Wallis Test for Race, Factor 2 .....	93
16. Kruskal-Wallis Test for Race, Factor 3 .....	94
17. Kruskal-Wallis Test for Race, Factor 4 .....	95
18. Kruskal-Wallis Test for Residency, Factor 1 .....	96
19. Kruskal-Wallis Test for Residency, Factor 2.....	96
20. Kruskal-Wallis Test for Residency, Factor 3.....	97
21. Kruskal-Wallis Test for Residency, Factor 4.....	98
22. Kruskal-Wallis Test for First-time Freshman Institution, Factor 1 .....	99
23. Kruskal-Wallis Test for First-time Freshman Institution, Factor 2 .....	99

24. Kruskal-Wallis Test for First-time Freshman Institution, Factor 3 .....	100
25. Kruskal-Wallis Test for First-time Freshman Institution, Factor 4 .....	101
26. Logistic Regression ROC Curve.....	119
27. Logistic Regression Sensitivity Analysis.....	120
28. C5 Decision Tree ROC Curve .....	122
29. C5 Decision Tree Sensitivity Analysis .....	123
30. Random Forest ROC Curve .....	125
31. Random Forest Sensitivity Analysis.....	126
32. Artificial Neural Network ROC Curve .....	128
33. Artificial Neural Network Sensitivity Analysis.....	129
34. Support Vector Machine ROC Curve.....	131
35. Support Vector Machine Sensitivity Analysis.....	132
36. Logistic Regression ROC Curve, West TN University .....	151
37. Logistic Regression ROC Curve, Central TN University.....	153
38. Logistic Regression ROC Curve, North Central TN University .....	155
39. Logistic Regression ROC Curve, East Central Tennessee University .....	157

## CHAPTER 1

### INTRODUCTION

According to Tennessee Higher Education Commission (THEC) fact books (2013, 2014, 2015, 2016) the Tennessee Board of Regents (TBR) reported approximately 11,500 first-time freshmen began at universities each fall between 2006 and 2009. Of that set an average of 5,800 students graduated within 6 years. Of those graduates an average of 950 students, or 17%, transferred to other institutions and graduated. The departure of such successful students is costly to institutions in two ways. First, institutions are losing the steady stream of tuition and fees from the students (Raisman, 2013). Second, institutions are losing the investment cost of retention efforts and advisement to these students prior to their departure (Johnson, 2012).

Public universities in Tennessee have experienced a consistent decrease in state funds per full-time student over the last several decades (THEC, 2013). This shift in funding from state appropriations to student tuition makes student departures more costly to institutions. The average cost of tuition for a full-time student taking 15 credit hours at a 4-year institution in Tennessee is approximately \$8,600 per year. A loss of 17%, or an average of 150 full-time students per university in the former TBR system, translates to roughly \$1,300,000 annually in foregone revenue per institution (College Tuition Compare, 2016).

As state appropriations per full-time student have decreased in Tennessee, the complexity of the process used to distribute funds has increased. The current funding-formula for Tennessee institutions consists of a mix of weighted outcomes and fixed cost calculations (THEC, n.d.-a). Institutions set their weights on student progression from 30, 60, and 90 hours, as well as bachelor, master, and doctorate outcomes. Institutions determine the weights for these items based on their institutional priorities and expectations that they can achieve high returns in each



category. Depending on the selected progression weights, even a 1% increase in a category due to increased retention can translate to an increase in state appropriations anywhere between a few thousand dollars and close to \$100,000.

Improving retention increases revenue from student tuition and state funding. Students who transfer and graduate elsewhere represent a substantial source of lost funding that could be retained if those students can be identified prior to their departure. This non-experimental quantitative study explores this population of students who transfer from their initial 4-year institution and graduate somewhere else. In addition this study determines if there is a predictive model for identifying such students prior to their transferring out.

Colleges and universities have been collecting massive amounts of student data for many years as part of “conducting business” (Soares, 2012, p. 1). Student information systems collect information including student addresses, emails, phone numbers, financial aid offers, grades in courses, ACT scores, housing information, meal plan information, social activities, and payments associated with the university (THEC, n.d.-b). Institutions in states such as Tennessee with performance funding initiatives must gather and store these data for state reporting requirements. Tennessee was the first state in the nation to implement performance based funding back in 1980. Connecticut, Missouri, and Kentucky implemented performance based funding systems in the next decade (McLendon & Hearn, 2013). Thirty-two states now have some form of performance funding (NCSL, 2015). This need to track and report student performance information means that institutions in the majority of states across the nation have an increasing record of historical student performance data that can be analyzed. In addition, just as in Tennessee, states across the nation have been investing fewer state dollars into higher education

(Leachman & Mai, 2014). Tennessee is not alone in the fiscal need to leverage data to improve student retention, progression, and outcomes.

Institutions across the United States that offer federal financial aid must also report institutional data points to the National Center for Educational Statistics (NPEC, 2009). This information is publicly available through the Integrated Postsecondary Education Data System (IPEDS). In addition, organizations like the National Student Clearinghouse (NSC) have enabled colleges to track students who attended their institution, left for another institution, and graduated. Institutions that are members of NSC have access to a small number of data points about former students such as the institution that the student transferred to, what program the student transferred into, and whether the student went on to graduate (NSC, n.d.).

Since the late 1980s these educational data have been stored electronically (Howard, McLaughlin, & Knight, 2012). Keeping electronic records has several unintended consequences. In addition to the primary data being stored, meta-data or data about the data can be tracked. This includes the time that the data were recorded, the fact that the data were not recorded, and who entered the information. In addition, the sheer amount of data that are collected over time enables researchers to find trends hidden in the data.

Data mining has been an emerging technique to analyze educational data and find those trends and make predictions from the data. Data mining combines the disciplines of computer science and statistics. Artificial intelligence is a subdiscipline within computer science that has been instrumental in data mining, as a key goal of artificial intelligence has been knowledge discovery. Machine learning techniques emerged from artificial intelligence. Supervised learning algorithms were developed in which the outcome categories for data are known beforehand. Such algorithms can be used to classify data records. For instance, a student could

be classified as a potential transfer student based on common characteristics of students who have previously transferred. Unsupervised learning algorithms were also developed. Unsupervised algorithms attempt to identify the outcome categories based on commonalities among data records (Provost & Fawcett, 2013). Given a set of students, an unsupervised algorithm could determine there are three distinct classifications of students: those who will stay at an institution and graduate, those who will drop out, and those who that will transfer elsewhere. In this manner unsupervised learning is similar to factor analysis.

Several studies have examined data mining techniques for use in identifying research variables for student retention as well as for predicting whether students will stay (Aguilar, 2015; Alpaydin, 2010; Delen, 2010, 2011; Herzog, 2006; Nandeshwar & Chaudhari, 2009). These studies have chosen several different data mining techniques as well as various means of testing the effectiveness of the selected algorithms. Several models are typically examined because each data set is different and one model may be more accurate than another for the particular data set in question. Chapter 2 of this study provides a more in-depth examination of data mining models and their use in higher education studies.

Higher education institutions may be able to leverage data they have been collecting for years rather than rely on costly annual surveys. The costs of conducting an in-house survey can include determining the population to survey, designing the survey, pretesting the instrument to ensure its validity and reliability, and hiring and training staff to administer the survey and collect results. Once the data are collected, issues such as response rates and what to do with nonresponses are introduced. Each of these costs and considerations incur the additional cost of time to deal with them. The time to properly conduct such a survey can range from “several months to a year” (Fairfax County, 2012, p. 1). The data that higher education institutions must

collect for state and federal reporting are already designed for valid and reliable instruments used at the state and federal level. In addition, response rates are not an issue because student participation is not voluntary. Businesses such as Amazon, Netflix, and Wal-Mart have been using data mining to effectively predict customer behavior for years (Amatriain, 2013, Harsoor & Patil, 2015). This study seeks to apply data mining techniques to the higher education issue of retaining students who are most likely to graduate.

### Statement of the Problem

The purpose of this quantitative study was to discover factors about first-time freshmen who began at a university in the former TBR system, transferred to any other institution after their first year, and graduated with a degree or certificate. In addition, this study determined if a predictive model can be generated to identify these particular students prior to their initial departure. An exploratory factor analysis was performed to identify a set of common student characteristics. These characteristics were used within five predictive models to identify such students prior to their departure: logistic regression, decision trees, random forest, support vector machines, and artificial neural networks.

### Research Questions

This study had 10 general research questions as listed below:

1. Which characteristics are most likely to predict graduation in 6 years after enrollment for first-time freshmen students under the age of 24 from a 4-year institution in Tennessee?
2. What characteristics identify first-time freshmen students under the age of 24 who transfer to another higher education institution?

3. What characteristics identify first-time freshmen students under the age of 24 who did not graduate from a 4-year institution in Tennessee?
4. What characteristics identify first-time freshmen students under the age of 24 who began at a 4-year institution in Tennessee, transferred to another higher education institution, and graduated?
5. Is the predictive power of logistic regression for determining which students will leave their home institution and graduate elsewhere greater than 50%?
6. Is the predictive power of decision trees for determining which students will leave their home institution and graduate elsewhere greater than 50%?
7. Is the predictive power of random forests for determining which students will leave their home institution and graduate elsewhere greater than 50%?
8. Is the predictive power of artificial neural networks for determining which students will leave their home institution and graduate elsewhere greater than 50%?
9. Is the predictive power of support vector machines for determining which students will leave their home institution and graduate elsewhere greater than 50%?
10. Which of the data mining techniques: logistic regression, decision tree, random forest, artificial neural network, or support vector machines provide the best classification result in predicting students who will leave their home institution and graduate somewhere else?

#### Significance of the Study

There has been an abundance of studies performed on retention and persistence. There has been a growing body of work on the use of data mining in higher education as well.

However, there have been no studies yet attempting to identify students prior to their departure that will transfer and graduate elsewhere using predictive analytics. This study thus contributed to the body of literature on retention and data mining. Tennessee institutions collect an abundance of student data, so this study tested the ability to use the collected data's utility for other big data research projects. A majority of states are similar to Tennessee in that they have implemented performance-based funding for the distribution of funds for higher education. The reporting necessary for such funding models, as well as reporting required for institutions accepting federal financial aid monies, means that many states also collect an abundance of data and may find this study to be informative. In addition to the utility of historical student data, this study will inform decision makers at the institutional level with information related to student characteristics that are likely to predict graduation, attrition and transfer. Predictive models may be useful in focusing resources effectively on such students to keep them from transferring out, thus improving the retention rate for the institution. This study furthermore contributed to the broader body of research concerning predictive models. The study design tested the limits of such models. The evaluation of several predictive models confirmed prior studies and provided new context for the applicability of different models.

#### Definitions of Terms

The following terms are used throughout this study. The definitions provided should be used during reading and interpretation of this work.

**Artificial Neural Network (ANN):** A more complex data mining algorithm that can be used for both classification (supervised learning) and clustering (unsupervised learning). For

classification tasks artificial neural networks take data inputs and assign weights through one or more “hidden” layers to generate outputs (Larose & Larose, 2015, p. 342).

**Bagging:** A technique to improve the predictive power of a model by combining multiple outputs into a single prediction. With bagging, each individual outcome is given equal weight in determining the combined prediction (Witten & Frank, 2011).

**Big Data:** An amorphous term used to describe the storage and use of large, complex data sets. Big data is typically described by four Vs: volume, velocity, variety, and value. The data typically come from a variety of sources and lack the structure of traditional data sources. New technologies such as machine learning are used to process the large volume of data collected. The processing velocity must increase as the volume increases. The goal of big data is to quickly analyze data to produce value via data-based decisions (Daniel, 2015; De Mauro, Greco, & Grimaldi, 2015; Ward & Barker, 2013).

**Boosting:** A technique to improve the predictive power of a model by combining multiple outputs into a single prediction. Successful outcomes are given more weight in determining the combined prediction (Witten & Frank, 2011).

**Data Mining:** The application of machine learning algorithms and statistics to identify trends and patterns within large data sets (Larose & Larose, 2015). Data mining combines several disciplines such as computer science and statistics to find these trends and make predictions from the data.

**Decision Tree:** A data classification algorithm that assigns probabilities to different outcomes. The tree is a structured sequence of probabilistic decisions that can be visually followed as the branches of a tree (Witten & Frank, 2011).

**First-time Freshman:** A degree-seeking freshman student starting in the fall or prior summer who has not attended college before. These students may enter with prior college credit from dual enrollment and advanced placement courses (IPEDS, 2016).

**Graduation Rate:** The rate at which degree-seeking first-time freshmen for an institution graduate. This is expressed as the number of students from a cohort that did not return in the fall because they graduated from the same institution since the prior fall term. This is typically calculated as a 6-year completion rate for 4-year institutions and is expressed as a percentage of the original entering cohort (IPEDS, 2016).

**Horizontal Transfer Student:** A student who transfers from one institution to another institution at the same level. For example, a lateral transfer student would transfer from one 4-year institution to another 4-year institution. Horizontal transfer students are also referred to as lateral transfer students (Tobolowsky & Cox, 2012).

**Machine Learning:** The collection of algorithms used to construct predictive models from large data sets. Originally machine learning was a subfield of artificial intelligence interested in knowledge discovery and creating predictions based on changes in data over time (Provost & Fawcett, 2013).

**Predictive Analytics:** The application of machine learning algorithms and statistics on large data sets to predict or estimate future outcomes (Larose & Larose, 2015). These algorithms use a training set of data for the predictive model to learn from and a test set of data for the predictive model to be evaluated against.

**Random Forest:** A series of decision trees evaluated together using bagging with the intent of improving the predictive power of the model (Witten & Frank, 2011).



**Retention Rate:** The rate at which degree-seeking, first-time freshmen for an institution return each fall term. This is expressed as the number of returning freshmen divided by the number of the original entering cohort. The rate is typically calculated for fall-to-fall enrollment and is expressed as a percentage (IPEDS, 2016).

**Reverse Transfer Student:** A student who transfers from a 4-year institution to a 2-year institution (Tobolowsky & Cox, 2012).

**Student Churn:** A concept derived from the business concept of customer churn. Students withdraw from a college in a similar fashion to customers who cease doing business with a business. New students must be brought in to make up for the lost revenue of the departed students (Ubi & Liiv, 2010).

**Supervised Learning:** Machine learning algorithms that have known target categories. These are used to classify data into the known categories. Examples of such algorithms include decision trees, random forest, support vector machines, and artificial neural networks (Provost & Fawcett, 2013).

**Support Vector Machine (SVM):** A data classification algorithm that uses multiple linear models to generate a maximally thick boundary between sets of data in order to accurately classify them (Provost & Fawcett, 2013).

**Test Set:** A sample set of data used by a predictive model to determine the effectiveness of the model. The model uses what it has learned from the data set that was used to train it in order to classify records in the test set. Because the actual outcomes of the test set are known, the outcomes from the predictive model can be compared to the actual outcomes to determine the effectiveness of the model (Tan, Steinbach, & Kumar, 2005).

**Training Set:** A sample set of data used by a predictive model to identify patterns. The predictive model algorithm learns how to classify outcomes based on patterns within the training set (Tan et al., 2005).

**Transfer Student:** Someone who leaves his or her current institution and enrolls at another institution (IPEDS, 2016). THEC further refines the definition of a transfer student to be someone who transfers to another institution after accumulating a minimum of 12 credit hours (THEC, 2016). For the purposes of this study, a transfer student is a student who transfers from his or her first-attended 4-year institution to any other higher education institution. The student must have attended the fall and spring semesters of their first year at their initial institution.

**Unsupervised Learning:** Machine learning algorithms that do not have known target categories. The algorithms attempt to cluster input data according to trends within the data. Examples of unsupervised algorithms include k-means clustering, k nearest neighbor, and hierarchical clustering (Provost & Fawcett, 2013).

**Vertical Transfer Student:** A student who transfers from a 2-year institution to a 4-year institution. For example, a student that transfers from a community college to a 4-year institution (Kirk-Kuwaye & Kirk-Kuwaye, 2007).

### Limitations and Delimitations

This study was limited by the data points available from THEC. While several data points related to retention, such as ACT score and high school GPA were available, other desirable data points such as the amount of financial aid needed, parental income, and first generation status were not available. Behavioral data points were not collected and thus could not be used, although certain behavioral data points could be inferred. For example, student

isolation is not a data point, but the number of students from particular population areas could be determined. This study was limited geographically by the choice of institutions selected for analysis. The institutions for the study were restricted to the state of Tennessee. The factors affecting students in Tennessee may be far different from those in other states. The predictive nature of this study required looking back to the cohorts of first-time freshmen students between 2006 and 2009. The conditions that existed in that timeframe may have been unique to those cohorts, further limiting the applicability of this study's results in practice. In addition, the enterprise data system being used by TBR institutions between 2006 and 2009 was upgraded from the Information Associates set of programs (Student Information System, Human Resource System, Financial Record System, and Alumni Development System) to the SCT/Sungard Banner system. TBR institutions began and completed their upgrades to the new system at different times between 2006 and 2009, meaning that data collection and entry in that timeframe may have been unstable. Finally, the TBR began a reverse transfer policy in 2014 that allowed students to graduate with an associate degree while enrolled in a bachelor program (TBR, 2014). This study used a differing definition for reverse transfer student that involved the student leaving a 4-year institution and enrolling at a 2-year institution. Students who take advantage of the Tennessee reverse transfer program remained enrolled at a 4-year institution but may appear to have departed and graduated elsewhere. Such cases could confound the results of this study. Identifying such cases was not possible.

This study was delimited to first-time freshmen under the age of 24 in the fall terms between 2006 and 2009. Furthermore, this study was delimited to students who began at a 4-year institution within the former Tennessee Board of Regents system. These institutions were Austin Peay State University, East Tennessee State University, Middle Tennessee State

University, Tennessee Technological University, Tennessee State University, and the University of Memphis. Students must have attended in the fall and spring semesters of their first year. Those students may have transferred to another institution within or outside of Tennessee. Transfer students may also have graduated with an associate, certificate, bachelor, or other type of undergraduate degree.

There were numerous data mining algorithms that could have been employed for the predictive analytics portion of this study. Five algorithms were selected for evaluation. A logistic regression model was used. Logistic regression is a general classifier model commonly used in data mining studies (Provost & Fawcett, 2013). A decision tree algorithm was used as it was commonly used by other researchers and intuitive to understand. A random forest algorithm was used as well for its intuitive nature and also because of its potential to provide more predictive power than a single execution of a decision tree. Support vector machines and artificial neural networks were included due to their use in other data mining studies. However, they are more complex algorithms that statisticians and those unfamiliar with data mining may find difficult to understand (Delen, 2010).

### Overview of the Study

Chapter 1 of this study provided an introduction to the topics included in this study. The rationale behind examining students who transfer out and graduate elsewhere was examined. The reasoning behind the use of secondary data and the application of data mining in this study was also examined. In addition, the introduction included a formal statement of the problem, research questions to be explored, definitions of terms used throughout this study, limitations on the applicability of the research, and delimitations for the sample used. Chapter 2 provides a

literature review that focuses on factors relating to first-time freshmen, retention, graduation rates, and transfer students. In addition, data mining was explored along with an in-depth examination of how data mining has been used with higher education data. Chapter 3 provides information about the methodology for this nonexperimental, quantitative study. Chapter 4 provides the results of the study. Each research question is addressed. Chapter 5 concludes with an analysis of the results, a summary of the study, and recommendations for future research.

## CHAPTER 2

### LITERATURE REVIEW

The purpose of this quantitative study is to discover factors about first-time freshmen that began at a university in the former Tennessee Board of Regents (TBR) system, transferred to any other institution after their first year, and graduated with a degree or certificate. In addition, this study determined if a predictive model can be generated to identify these particular students prior to their initial departure. Transfer students who graduate from another institution represent foregone revenue to their originating institution. In addition, due to performance funding, institutions can lose a portion of state funds as these students negatively impact progression and graduation rates.

State and federal reporting requirements provide institutions with a wealth of data that can be used for data mining. However, the intentions behind state and federal reporting may preclude the collection of data points that retention researchers have indicated as predictors of student persistence. Therefore, a review of the research for performance reporting, retention research, and transfer student retention was conducted. This was followed by a review of research in data mining, specifically higher education data mining. The types of student data points, data mining software, and techniques was examined. Common pitfalls of data mining research were reviewed to ensure that the research design of this study was robust. Finally, the ethics of data mining were reviewed to provide a humane context for this research.

#### The Rising Cost of Higher Education

An average of 17% of students who begin their higher education career at a Tennessee university transfer to another institution and go on to graduate there (THEC, 2013, 2014, 2015,

2016). Given that the average cost of tuition for a full-time student taking 15 credit hours at a university in Tennessee is \$8,600 annually, this is an average of \$1.3 million dollars in lost revenue per university in the former TBR system. In addition, state funds per full-time student for public institutions in Tennessee have been decreasing over the last several decades (THEC, 2013). Institutions have increased tuition and fees to make up for the lost funding source. This is a nationwide trend in higher education. Higher education institutions across America have seen decreases in state funding, increases in tuition and fees, and pressure to become more efficient (Noland, 2006, 2011). Between 1982 and 2007 the median income for families increased by 147% while college tuition and fees increased by an alarming 439% (Mendoza, Malcolm, & Parish, 2015). Gordon and Hedlund (2015) evaluated common reasons attributed to this rise in cost. Reasons included supply-side changes, demand-side changes, and macroeconomic forces. Supply-side changes were attributed to either Baumol and Bowen's (1966) notion of cost disease or to the decline of state funding with tuition and fees filling the gap in revenue. Cost disease is the concept that wages in one industry increase in response to wage increases in another industry as a means to retain top employees rather than due to productivity increases from those employees. Demand-side changes were expansions in grant aid and loans. Macroeconomic forces purportedly drove an increase in tuition as the demand for college degrees increased. Gordon and Hedlund (2015) concluded that the expansion in grant aid and loans actually drove the increase in tuition rather than declines in state funding. Gordon and Hedlund however used unadjusted costs of tuition between 1987 and 2010 for their study. Tuition and fees increased from \$6,600 in 1987 to \$14,500 in 2010. Using a consumer price index inflation calculator from the Bureau of Labor Statistics (CPI Inflation Calculator, n.d.), \$6,600 from 1987 would have \$12,700 purchasing power in 2010. At the same time, state

funding fell from \$8,200 per full-time equivalent in 1987 to \$7,300 in 2010. For perspective, \$8,200 in 1987 should have approximately \$15,700 of purchasing power in 2010. McCluskey (2017) also concluded that the rise in tuition was due to cost disease, increases in financial aid, and decreases in state appropriations per student. McCluskey further stated that tuition increases were larger than necessary to make up for the decline in appropriations. The net annual change per pupil for tuition and appropriations across the U.S. from 1990 to 2015 was an increase of \$57 (SHEF, 2016). McCluskey analyzed these increases by state and created four types of revenue changes per state: appropriations increased and tuition increased, appropriations decreased but tuition increased more, appropriations decreased but tuition increased less, and appropriations increased while tuition decreased. While McCluskey included the average net annual change per pupil for each category, the average total change in revenue was also included. The total change in revenue, computed as the number of full-time students multiplied by the total appropriations and tuition per student, was approximately \$47 million from 1990 to 2015 for the U.S.

McCluskey further emphasized this revenue increase with the use of graphs per state. A graph showing the increase in full-time enrollment per state, an average of approximately 3,400 students per year from 1990 to 2015 for the U.S., was omitted. There was a weak correlation ( $r=-0.26$ ) between the net annual change per pupil for tuition and appropriations and the change in full-time student enrollment per state from 1990 to 2015 (SHEF, 2016), indicating that the growth in enrollment over approximately 3 decades was not sensitive to the exchange between tuition and appropriations. There was a strong correlation ( $r=0.91$ ) between total change in revenue and change in full-time student enrollment per state from 1990 to 2015. Therefore, showing the increase in revenue to be a function of the increase in full-time student enrollment would have weakened McCluskey's argument that colleges and universities have been increasing



tuition and fees more than necessary. However, the sentiments expressed by Gordon, Hedlund, and McCluskey that higher education institutions have been increasing tuition and fees in order to collect more federal government dollars through loans and Pell grants, has led many states to adopt performance funding models (Dougherty, Natow, Bork, Jones, & Vega, 2013).

### Performance Funding

Tennessee was the first state in the nation to implement performance-based funding in 1980, and several states have adopted similar programs for funding higher education (McLendon & Hearn, 2013; National Conference of State Legislatures, 2015). Collecting data about students has been an ancillary outcome of performance modeling (Dougherty & Reddy, 2011). The intention of performance funding was to incentivize institutions to align their goals with the goals of state legislatures, namely to increase retention, graduation, and postgraduation employment (Kelly & Lautzenheiser, 2013). The success of performance funding has been inconsistent (Dougherty & Reddy, 2011; Hillman, Tandberg, & Fryar, 2015; Horn & Lee, 2017; Li, 2014; Tandberg & Hillman, 2014).

Rutherford and Rabovsky (2014) examined performance funding model outcomes at institutions across the United States between 1993 and 2010. They found that the models were unrelated to graduation rates, retention rates, and degree production. Sanford and Hunter (2011) analyzed retention and graduation rates at 4-year institutions in Tennessee between 1995 and 2009. In 2005 the state doubled the amount of money linked to retention and graduation rates, yet no improvement in rates was found. Dougherty et al. (2014) found that due to the focus on improving retention and outcomes performance funding may actually reduce access to higher education for disadvantaged students. Students with higher high school GPAs and ACT or SAT

scores are more likely to persist (Willingham, 1985). Therefore, an institution that wants to improve retention and graduation can simply raise admissions standards. This in turn reduces access to higher education for disadvantaged populations. This outcome was further supported by Umbricht, Fernandez, and Ortagus (2015) who found that performance funding in Indiana did not increase degree production rates. Instead, such funding models were associated with more rigorous admissions standards and lower enrollment of minority populations. Kelchen and Stedrak (2016) used IPEDS data to review Pell grant revenue in states with performance funding. Pell grant revenue was used to infer the number of low-income students in such states. Students from lower socioeconomic backgrounds have lower persistence rates than students from higher socioeconomic backgrounds (Astin, 1993). Kelchen and Stedrak found that colleges in states without performance funding had more Pell revenue than states with such funding models. Colleges in states with a performance funding model tended to have admissions standards that would bar low-income students from access.

The increase in the number of states adopting such models appears to lack good reason because performance funding has not resulted in obvious improvements of student persistence and graduation. Dougherty, Natow, Bork, Jones, and Vega (2013) suggested that one driving factor has been the increased ability for institutions and states to collect data related to outcomes. Another commonality among states that have adopted performance funding has been Republican state legislatures in search of ways to make state tax dollars more effective (Horn & Lee, 2017; McLendon, Hearn, & Deaton, 2006). Dougherty et al. (2013) claimed that performance funding models arose due to a skepticism about the mission of higher education and a reluctance to give state funds collected through taxes to institutions without accountability for outcomes desired by the state legislature.

The poor outcomes of performance funding models are due in part to the perception of the process (Dougherty & Reddy, 2011; Li, 2014). Li found that senior level administrators such as presidents and vice presidents and institutional research officers place special emphasis on the models. Below that level of administration the implications and goals of performance funding for an institution were not well known or understood. Dougherty and Reddy (2011) found that department chairs viewed reporting for performance funding to be a perfunctory task, as opposed to a valuable process. Thus, the data collected were not being used to actually work toward improvements in outcomes. Further detracting from internal application and analysis of data, administrators in Florida institutions had challenges with marshalling their collected data into the format required by the state reporting office. At a community college in Tennessee, the office of planning, research, and assessment had to expand to meet the data gathering and data massaging demands of the state (Shaw, 2000).

#### Additional Calls for Accountability

In addition to performance funding, the federal government has sought increasing accountability for higher education. In his 2009 address to Congress, President Obama challenged the nation to once more become the world leader in college graduates (Nichols, 2011). The College Scorecard was released in 2013 from the Department of Education, along with a financial aid shopping sheet as a method to implement performance reporting (Horn & Lee, 2017). Performance reporting does not directly impact funding but may cause students to choose another school that they view as a better investment. The Integrated Postsecondary Education Data System (IPEDS) has been used for performance reporting since 1985 for any higher education institutions that participate in federal financial aid programs (Fuller, 2011).

### Available Data

While reporting at the state and federal level may be a laborious process, requiring staff devoted to such work (Dougherty & Reddy, 2011), the data collection is not without opportunity. The data can be used in a number of longitudinal and data mining studies, having been collected electronically for decades. The IPEDS data are organized into 12 categories including institutional characteristics, enrollment, graduation outcomes, finances, and staffing information. There are over three thousand individual data fields, with an average of over 270 per category (IPEDS, n.d.). In 2012, the Tennessee Higher Education Commission (THEC) collected data files for enrollment and graduation outcomes that contained over 100 data fields. The THEC data fields include demographic information, precollege attributes such as ACT scores and high school GPA, and academic progress information.

Tennessee has six bordering states that also have some form of performance-based funding in place: Georgia, Mississippi, Arkansas, Missouri, Virginia, and North Carolina (National Conference of State Legislatures, 2015). Each state collects student information similar to Tennessee. Mississippi and North Carolina collect additional information about employees, scholarships, grades, housing, and admission practices (NCHED Forms, n.d.; Office of Strategic Research, n.d.). The average number of fields is over 100, and the fields deal primarily with demographic information as opposed to behavioral information such as student satisfaction or intention to remain enrolled.

### Retention

Retention has been studied in detail since the 1970s. The majority of research has been quantitative, focusing primarily on sociodemographic variables (Campbell & Mislevy, 2013;

Mendoza et al., 2015; Reason, 2009). Tinto (1975) initially compared dropping out of college to suicide as described by Durkheim (1961). Students who did not learn to fit into the society of the institution removed themselves from the institution. Tinto (1993) continued to write about three factors that influenced students' choice to withdraw: academic difficulties, social and intellectual integration, and issues between educational and occupational goals. Astin's 1993 research findings were an exception to the focus on sociodemographic factors. In an extensive longitudinal study from 1985 to 1989 involving 25,000 students from 200 institutions, Astin found that student peer group interaction had long-term effects on learning and development. Faculty and student interaction had the second largest impact on student development. Socioeconomic status had the most impact on baccalaureate degree completion. Peltier, Laden, and Matranga (1999) found gender, race, ethnicity, and socioeconomic status to be related to persistence. For example, more women than men persisted. This was also supported by work from Leppel (2002) who found that intervention efforts need to be customized to the needs of either gender.

Other research that examined covariates and the issue of imbalanced data sets has clarified the impact of sociodemographic factors though. Reason (2003) found that gender differences disappeared after controlling for interaction effects such as on-campus versus off-campus residence or institution type. In addition, where prior research indicated that white and Asian students experienced better persistence than other student groups, Reason found that such differences went away after controlling for socioeconomic status and precollege academic factors. Hu and St. John (2001) further supported Reason's research. Hu and St. John found that financial aid could mitigate the differences in persistence among ethnicities.

Other areas of student retention research outside of sociodemographic variables include academic preparation, student disposition, the student peer environment, individual student experiences, organizational factors, and external pressures (Bean 2005; Mendoza et al., 2015; Reason, 2009). Socioeconomic status and high school quality were found to be related to academic preparation of students (Cabrera, Burkum, & LaNasa, 2005; Pascarella & Terenzini, 2005). In addition, Adelman (2006) found that a greater number of higher level math courses in high school had a significant impact on student success and retention.

Research into student disposition has been predominantly in the field of psychology (Reason, 2009). Locus of control, self-efficacy, conscientiousness, and academic goals were factors found to influence student persistence (Braxton, Hirschy, & McClendon, 2004; Robbins et al., 2004; Tross, Harper, Osher, & Kneidinger, 2000). Bean (2005) found that a student's intention to stay or leave was the best predictor for student retention. White and Massiha (2016) found that self-confidence and a lack of barriers were key variables of persistence for women in science, technology, engineering, and math (STEM) programs.

The student peer environment influenced student outcomes due to its effects on social integration at college (Astin, 1993; Braxton, Jones, Hirschy, & Hartley, 2008). In confirmation of this finding, women's colleges and historically black colleges and universities were found to have higher retention and completion rates than their counterparts (Pascarella & Terenzini, 2005). Such heterogeneous student environments enabled social integration among students. Spruill, Hirt, and Mo (2014) examined persistence among males and found that peer views on what was important had a significant impact, further confirming that social integration is important.

Reason (2009) split individual student experiences into three kinds: curricular, classroom, and extra-curricular. In terms of curricular experiences, STEM program students were more likely to persist than students in other programs (Adelman, 1999; Leppel, 2002; Pascarella & Terenzini, 2005). However, this has likely been a result of a heterogeneous environment. STEM fields such as Computer Science have been male dominated (White & Massiha, 2016; Woodfield & O'Mahony, 2016). Education has been female dominated and has mostly nontraditional students, leading to lower retention rates. Business has been the most gender balanced area of study (Woodfield & O'Mahony, 2016). First year seminar courses were strongly related to persistence (Cuseo, 2007; Hunter & Linder, 2005; Strumpf & Hunt, 1993). Such courses assisted students with the transition to the college environment, a recommendation from Woodfield and Mahoney.

Active and engaged faculty had a positive impact on student social integration (Braxton et al., 2008; Pascarella, Seifert, & Whitt, 2008; Tinto, 1993; Wayne & Youngs, 2003; White & Massiha, 2016). Students felt a connection to the institution, and they felt that the instructor cared about their success when the faculty taught clearly and were organized with their instruction. Outside of the classroom, student involvement in academic activities such as studying indicated student engagement and translated to increased persistence (Astin, 1993; Baars & Arnold, 2014; Heller & Cassady, 2015). Pascarella and Terenzini (2005) found that involvement in student organizations had little to no direct impact on student persistence.

Berger and Milem (2000) explored two dimensions of organizations and their impact on students. Structural demographic dimensions such as community college versus university, public versus private, and Carnegie classification were typical institutional characteristics included in higher education studies. Organizational behavioral dimensions involved the culture

and environment produced by the institution. Berger (2001-2002) wrote about five types of organizational behaviors: bureaucratic, collegial, political, symbolic, and systematic. Collegial, symbolic, and systematic institutions enhanced student retention. Such organizational behaviors produced environments that showed care for students or for a higher ideal. Political and bureaucratic behaviors had a negative or no effect on retention. These organizational behaviors showed less care for students. A proxy for these types of behaviors was institutional expenditures. Expenditures for institutional/administrative support had a negative impact on student persistence, while expenditures on instruction and academic support had a positive impact on student persistence (Crawford, 2015; Gansemer-Topf & Schuh, 2006). However, expenditures on academic support only had a positive impact on persistence at selective institutions. Tinto (2010) identified four institutional aspects that impacted student success. Providing support academically, socially, and financially to students was one aspect that provided evidence that institutions who invest in student success will have more successful students. The other three institutional aspects included institutional expectations of students, good communication channels with students and seeking student involvement. Thus, institutions that focused on student outcomes rather than political or bureaucratic matters had greater student success.

The study of external pressures on student persistence has been a recent development (Heller & Cassady, 2015; Mendoza et al., 2015). In line with Reason (2009), Mendoza et al. viewed student retention as a multifaceted issue and used Bronfenbrenner's 1993 Ecological Systems theory for their phenomenological study of 45 undergraduate university students. The five ecological systems examined with the microsystem, the mesosystem, the exosystem, the macrosystem, and the chronosystem. The first three systems correlated with the previously



discussed dimensions of individual student experiences (microsystem), the student peer environment (mesosystem), and organizational factors (exosystem). However, Mendoza et al. provided more focus on external pressures such as employment while taking classes, family economic conditions, and the political and cultural norms of the time in which the individual student lived. The Great Recession caused a decrease in financial aid that made college less affordable and also made students have more anxiety about college completion. Students in the study reported participating in fewer social activities and working more hours, which interfered with studying. However, due to financial constraints, students were more committed to completing on time as a method to reduce the expense of education. More thought was applied to the choice of major, with choice based on job prospects after graduation. Heller and Cassidy (2015) found that goal setting behavior such as selecting a major for a desired career and graduating on time to begin a career were positively associated with persistence. Wilson et al. (2016) explored student connectedness to their home region. Retention was found to depend on social and regional tethering. Students from large families and students from distinct geographical areas such as Appalachia were less likely to persist if they were far from their cultural tethers.

### Transfer Students

Sixty percent of college students attend more than one institution over the course of their academic career (Adelman, 2006; Peter & Forrest Cataldi, 2005). Over a third of students who began college in 2008 transferred to another institution. (Shapiro, Dundar, Wakhungu, Yuan, & Harrell, 2015). Transfer students are thus a large, diverse group to study. This has led to inconsistent definitions for them. Students may be vertical transfers that moved from a 2-year

college to a 4-year institution, or they may be horizontal transfers that moved between the same level of institutions. In addition, students could be co-enrolled at community college and university programs, reverse transfers from a 4-year institution to 2-year school, or “swirling” back and forth among the various options (Ghusson, 2016, p. 28; Goldrick-Rab & Pfeifer, 2009, p.115; Tobolowsky & Cox, 2012, p. 390). Transfer students typically departed from their initial institution in the second year of college (Shapiro et al., 2015). These students have encountered difficulties due to their decision to transfer. Transfer shock, in which the student must adjust socially and academically to the new environment, has been a well-documented difficulty (Glass & Harrington, 2002; Hills, 1965; Ishitani, 2008; Laanan, 2001). In addition, transfer students often lose credits that will not transfer into the new institution (Monaghan & Attewell, 2014).

The first term GPA has consistently been identified as an indicator for transfer student success (McCormick, Sarraf, BreckaLorenz, & Haywood, 2009; McGuire & Belcheir, 2014; Pascarella & Terrenzini, 2005). The decline in GPA for transfer students during their first term at a new institution has been attributed to transfer shock. The new social and academic environment cause the student’s GPA to suffer. (Tobolowsky & Cox, 2012). Other factors have included student support structures of the transfer institution, the student’s perception about the institution, and nonacademic behaviors of transfer students. McCormick et al. (2009) found that transfer students were less likely to live on campus, thus self-selecting out of student support structures. Transfer students were also more likely to work off campus, be older than other students, and have more responsibilities outside of studies such as caring for children or aging parents.

Research into transfer students has focused predominately on vertical transfers from community colleges to 4-year institutions (Ghusson, 2016; Kirk-Kuwaye & Kirk-Kuwaye, 2007;

McGuire & Belcheir, 2014). Kirk-Kuwaye and Kirk-Kuwaye stated that vertical transfers from 2-year to 4-year institutions typically expected challenges and do better than other types of transfer students. However, McCormick et al. (2009) found that horizontal transfers were more likely to participate in research, study abroad opportunities, internships, and capstone projects than vertical transfers. Horizontal transfer students left their previous institution due to a number of reasons including academic, personal, and social dissatisfaction, financial difficulties, and pursuit of specific programs. Horizontal transfer students had a higher socioeconomic status than other types of transfers. Reverse transfer to a community college was more common among less affluent students and students whose parents had less education (Goldrick-Rab & Pfeifer, 2009). Reverse transfer students thus were more sensitive to academic and financial pressures than horizontal transfer students.

### Data Mining in Higher Education

Data mining has been used in a number of fields to perform pattern recognition, image processing, and outcome predictions (Ding, Shi, Tao, & An, 2016). For example, loan companies have been using data mining to make credit decisions, industrial companies have used data mining to diagnose mechanical devices, and oil companies have used data mining to improve the separation of gas from oil (Langley & Simon, 1995). More recently, Netflix has used data mining to predict user movie selections. Amazon and Wal-Mart have been using data mining to predict what products customers will purchase. Google has created a data collecting platform to allow companies to use data mining to identify web browsing and purchasing behavior. Financial institutions have been using data mining to detect fraud (Amatriain, 2013; Chen, Chiang, & Storey, 2012; Harsoor & Patil, 2015).

Predicting customer churn has been a common use for data mining outside of the higher education industry (Ballings & Van den Poel, 2012; Burez & Van den Poel, 2009; Coussement, Benoit, & Van den Poel, 2010; Luan, 2002). Pleskac, Keeney, Merritt, Schmitt, and Oswald (2011) noted the similarity between customer churn and student retention when offering an alternative to Bean's (1983) analogy of student withdrawal to employee turnover. Within higher education, data mining has been applied primarily to student retention and alumni donor issues (Durango-Cohen & Balasubramanian, 2015; Hashemi, Le Blanc, Bahrami, Bahar & Traywick, 2009; Le Blanc & Rucks, 2009; Luan, 2002; Luperchio, 2009; Skari, 2014).

Bogard, Helbig, Huff, and James (2011) made a distinction between classical stochastic and algorithmic research. Data mining was the application of these algorithms to conduct research. Luan (2002) gave a detailed description of data mining techniques and split them into four groups according to function. Classification techniques are used to assign binary values to output and can be useful for inferring missing values in a process called data imputing (Luan, 2002). Estimation techniques use data inputs representing past events to predict future outputs. Segmentation is used to cluster data into various groups. Description techniques are used to identify characteristics or rules of a general system. Alpaydin (2010) grouped data mining techniques into learning tasks. These tasks include learning associations, supervised learning, unsupervised learning, and reinforcement learning. Learning associations is analogous to Luan's description techniques as both seek rules to describe a system (Luan, 2002). Alpaydin listed regression and classification as types of supervised learning because the researcher is involved in selecting inputs and outputs for these techniques. Regression is analogous to Luan's estimation techniques as regression is used to predict an outcome given a particular set of inputs.

Unsupervised learning is analogous to Luan's segmentation as both techniques involve the clustering of similar data points together in the absence of specific guidance from the researcher.

Fayyad, Piatetsky-Shapiro, and Smyth (1996) condensed data mining into the two main tasks of prediction and description. Fayyad et al. went on to list specific techniques including Classification, Regression, Clustering, Summarization, Dependency Modeling, and Change and Deviation Detection. These techniques can be used for either prediction or description.

A consistent view of data mining emerges from the literature. Two main functions of data mining consist of predicting outcomes or describing data. Once a purpose is selected, the researcher selects a learning method consisting of either supervised or unsupervised learning. This selection is determined by the researcher's knowledge of or intention with the data set. If unknown patterns are sought, then unsupervised learning would be used. If evidence for suspected patterns is sought, then supervised learning would be used.

### Data Mining Algorithms

Several algorithms are used for supervised and unsupervised learning that can be used for predicting or describing data. Genetic algorithms, artificial neural networks, logistic regression, support vector machines, and decision trees are used for classification and estimation (Delen, 2010; Liao, Chu, & Hsiao, 2012; Luan, 2002). Market basket analysis, rule induction, and k-means are used for segmentation and description according to Luan (2002), although the full list of data mining algorithms is extensive.

## Classification Algorithms

Logistic regression is a common modeling technique used to perform regression analysis using a categorical dependent variable. The dependent variable is typically binary (e.g. yes or no). Logistic regression is similar to linear regression that has a continuous dependent variable. The output of a logistic regression is the odds, or probability, that a case belongs to a certain class (Provost & Fawcett, 2013). The use of logistic regression is widespread in higher education studies that use data mining algorithms as well as other types of higher education studies that focus on retention (Porter, 2002).

Decision trees are used to split a dataset into several homogeneous subsets distinguished by the state of a dependent variable at each level of the tree (Turban, Sharda, & Delen, 2010). The creation of the tree is an iterative process in which predictive variables are tested against a dependent variable. As the predictive power of a variable emerges through each case, the leaves of the tree are rearranged until the structure stabilizes. The leaves of the tree are the predictive variables, and they are arranged according to their influence on the dependent variable (Witten & Frank, 2011).

Genetic algorithms are meant to mimic the process of natural selection in evolution. Association rules are randomly generated for the input dataset. The rules are encoded so that crossover and mutation can easily occur. Crossover occurs when parts of rules are swapped. Mutation occurs when parts of rules are inverted. The fitness of the rules is evaluated against classification accuracy. Once accuracy is at an optimal level, the evolution of the dataset is complete (Han & Kamber, 2012; Langley & Simon, 1995).

Artificial Neural Networks (ANNs) are meant to mimic the neurons in the brain. ANNs consist of a series of layers that take inputs, compute weighted sums on the inputs, and generate

output probabilities (Langley & Simon, 1995). The main feature of ANNs is an S-shaped sigmoid function that returns values in the range of zero and one. The sigmoid function is applied to the weighted sums of the inputs to produce the output probabilities. Training the ANN is a crucial step as this is how the weights on inputs are learned. A feed-forward network model called multilayer perceptron (MLP) is the most commonly used ANN (Oztekin, 2016).

Support Vector Machines (SVMs) are similar to artificial neural networks. SVMs consist of an input, a layer of trained support vectors, and a classification output. SVMs use a training dataset to find a minimum, optimal distance between cases from two different classes or subsets of the dataset (Provost & Fawcett, 2013). Other data mining methods identify a separating hyperplane, or line when working with two-dimensional data, between different classes. An SVM identifies an optimal hyperplane by generating minimal margins that encompass all the valid, but suboptimal hyperplanes. The optimal hyperplane is then simply the middle of the margin between classes. These margins form the support vector. The support vector is used to minimize the number of incorrectly classified cases, enabling high generalization of cases (Cortes & Vapnik, 1995). SVMs can work with datasets that are linear and nonlinearly separable. A transformation function is used to map high dimensional datasets to a surface where the data are linearly separable. One such function that can be used is the sigmoid function, meaning an ANN can be created from an SVM. However, SVMs perform better than ANNs due to how SVMs generate an optimal hyperplane (Ding et al., 2016). In addition to the generalization advantage over ANNs, support vector machines work well with small datasets (Cortes & Vapnik, 1995; Cristianini & Shawe-Taylor, 2000).

In addition to these standard algorithms, ensemble methods are used. Ensemble methods simply combine data mining techniques in an attempt to improve model accuracy. Bagging and

boosting are the most common ensemble methods for decision trees (Provost & Fawcett, 2013; Witten & Frank, 2011). Bagging is used to combine multiple outputs into a single prediction. Each outcome is given equal weight in determining the combined prediction. Boosting also combines multiple outputs but more accurate outcomes are given more weight.

### Clustering Algorithms

Market basket analysis uses association rules to group items together. This type of clustering is primarily done in retail sales markets as a means of identifying subtle, but complementary product groupings, such as beer and potato chips (Witten & Frank, 2011). Rule induction subsumes decision trees, meaning rule induction and decision trees can be used for either classification or clustering. The ultimate goal of rule induction algorithms is to partition datasets into disjoint sets (Langley & Simon, 1995). K-means clustering is the most common clustering algorithm. A number of clusters to identify,  $k$ , is specified by the researcher. The algorithm then randomly selects  $k$  points as the cluster centers. In the first iteration, cases are assigned to the closest cluster center based on the mean distance to the  $k$  centers. A new mean center is then calculated for each of the  $k$  clusters, and all cases are reassigned based on the closest center. This process repeats until an iteration is redundant (Witten & Frank, 2011).

### Data Mining and Classical Statistics

Data mining and classical statistical methods are not mutually exclusive. Luan (2002) advocated the use of both when examining large data sets. Luan listed three strategies for data mining research. First, the results can be verified using classical statistical methods. Second, factor analysis and principal component analysis can be used to identify and remove



nonsignificant or highly correlated variables. Luan stated that data mining algorithms are more tolerant of correlated variables than classical methods. Thammasiri, Delen, Meesad, and Kasap (2014) made a similar point about the robustness of data mining methods. Data mining methods have fewer restrictions such as normality, independence, collinearity, etc. Finally, clustering and segmentation analysis can be used even though the target variables are known, as the analysis can reveal additional insights into the data.

### Factor Analysis

Although Luan (2002) lists factor analysis and principal component analysis as a means to identify and remove nonsignificant and correlated variables from datasets, relatively few studies perform this step. Instead, the predictive power of variables has typically been explored after the models have been executed. Techniques to test predictive power have included sensitivity analysis, Chi-Squared, and Pearson's Correlation (Aguiar, 2015; Delen, 2010, 2011; Herzog, 2006; Oztekin, 2016; Thammasiri et al., 2014). Baars and Arnold (2014) did use factor analysis in a manner similar to the one described by Luan. A survey at the University of Rotterdam was used to determine whether student motivation stemmed from aspects about the university, intrinsic student attributes, extrinsic student attributes, or extracurricular attributes. Maximum likelihood factor analysis was used to extract a limited number of motivational factors from the survey. Baars and Arnold's use of factor analysis was the typical application of the method, as opposed to using the reduced variable set in a data mining application. A closer approach to Luan came from Campbell and Mislevy (2013), who examined factors affecting student retention. While Campbell and Mislevy used a survey as well, they took variables from the three resulting factors (academic performance, institutional connectedness, and study skills)

as inputs for a multinomial logistic regression model. Campbell and Mislevy also used maximum likelihood as the factor extraction technique. Skari (2014) used logistic regression to predict alumni giving from a multistate sample of community college alumni. Skari used principal component factor analysis to reduce 14 student experience variables into a smaller set of three uncorrelated factors that were then used along with eight demographic variables for the logistic regression.

Factor analysis and clustering methods are similar in their intent. Factor analysis attempts to group variables together according to their power to explain variance between classes. Clustering attempts to identify groups of cases according to their similarities (Krebs, Berger, & Ferligoj, 2000). The selection of one method over the other depends upon the data. Factor analysis can also be used as a classification technique when a composite index is constructed. An index is used to assign weights to select variables (the variance ratios) in such a manner that each case within a dataset can be classified along a spectrum of factors. Index creation of this type is typically seen in the social sciences and in the field of finance (Brave & Butters, 2011; Kim & Rabjohn, 1980).

### Data Mining Tools

Several software tools exist to facilitate the use of data mining algorithms. Waikato Environment for Knowledge Analysis (WEKA) is a free data mining tool from the University of Waikato in New Zealand. The software comes with several data mining algorithms preloaded, allowing a researcher to focus on mining data rather than implementing mathematical models. Kabakchieva (2013) used WEKA to evaluate several data mining algorithms' abilities to predict student outcomes. Nandeshwar and Chaudhari (2009) used WEKA and Statistical Package for

Social Science (SPSS) to compare algorithms for predicting student enrollment. Pittman (2008) also used WEKA and SPSS to compare algorithms for predicting student retention. Bogard et al. (2011) and Raju and Schumacker (2015) used SAS Enterprise Miner to compare algorithms for predicting student retention. WEKA is often chosen for this type of research due to its free cost, low learning curve, and abundance of prepackaged algorithms. SPSS is a powerful statistical tool that has a lower learning curve than SAS (Liu, 2003).

In addition to algorithms and software tools, there are industry standards used for data mining. Luan (2002) discussed the Cross Industry Standard Process for Data Mining, CRISP-DM, that Daimler Chrysler developed in 1996 (Nandeshwar & Chaudhari, 2009). The standard consists of six steps. The first step is to understand the business domain. The second step is to identify data sources. The third step is extraction, transformation, and loading of data. The fourth step is to develop models to examine the data. The fifth step is evaluating each model against the data. The final step is to use the models in the decision-making process (Delen, 2010).

#### Classifying Predictive Value

Delen (2010) performed sensitivity analysis on neural networks, decision trees, support vector machines, and logistic regression. Credit hours, student age, residency, and retention time were found to have the greatest predictive weight. The sensitivity analysis from the neural networks was similar to beta coefficients from the regression model that Delen used. Credit hours, residency, and retention time had the most predictive value from the regression model. Herzog (2006) also used sensitivity analysis on the variables for a neural network model. Credit hours, student age, residency, and stop-out timing were the variables with the most predictive value. Oztekin (2016) performed a sensitivity analysis on decision trees, neural networks, and

support vector machines. Oztekin found that fall term GPA, housing status, and high school were the most predictive variables.

Raju and Schumacker (2015) used logistic regression, decision trees, and neural networks to identify attributes related to graduation outcomes. They found that first semester characteristics including end-of-term GPA, credit hours, and time status were important predictors. High school GPA was also found to be a graduation predictor.

Aguiar (2015) examined the use of an electronic portfolio program for an engineering department at Notre Dame as a means of improving early warnings for students at risk. Aguiar's approach of using a learning system to measure student engagement was unique. Other researchers examined only demographic and academic measures for predicting student retention. Aguiar used Information Gain, Gain Ratio, Chi-Squared, and Pearson's Correlation to rank variables in the study according to predictive power. Student use of the electronic portfolio was found to be the most important variable for retaining students, followed closely by whether the student had selected engineering as his or her major, and the student's SAT Math scores.

Mattern, Marini, and Shaw (2015) used cluster analysis to find patterns among students based on a broad range of variables. Mattern et al. used previous research to inform their selection of variables that would fall into eight general retention factors: intention to leave, attitudes, academic performance, social factors, bureaucratic factors, external environment, student's background, and financial factors. Hierarchical cluster analysis was then performed on approximately 19,000 nonreturning students. Three clusters emerged from the study: Affordability Issues, Unexpected Underperformers, and Underprepared and Facing Hurdles. Students in the Affordability Issues cluster had difficulty paying the high tuition at their institution. Students in the Unexpected Underperformers cluster were affluent, above-average

students prior to college who performed poorly in their first year and left. The Underprepared and Facing Hurdles cluster was the largest of the three clusters. Mattern et al. used the National Student Clearinghouse to track students who left. Students with affordability issues were most likely to enroll somewhere else that was more affordable. Approximately 35% of underprepared students and approximately 25% of the unexpected underperformers dropped out. This approach by Mattern et al. was one of the more original uses for data mining. The researchers effectively used demographics and financial measures to identify broader meta-data groups for the students in the study.

Tamhane, Ikbal, Sengupta, Duggirala, and Appleton (2014) used data mining techniques to predict which eighth graders would fail a state and national assessment test. Naïve Bayes, Decision Trees, and Logistic Regression were used. While Tamhane et al. used prior research to select the variables they used in the models, they also examined variable strength. Tamhane et al. were able to identify math test scores, ethnicity, and special education needs as variables that impacted prediction outcomes.

### Predictive Model Comparison

Delen (2010) examined four data mining techniques consisting of support vector machines, decision trees, artificial neural networks, and logistic regression. Delen also examined three ensemble methods: random forest, boosted trees, and information fusion. Support Vector Machines had the best predictive ability, followed by ensemble models and the information fusion model. The ensemble and information fusion models unsurprisingly had high predictive power due to their compounding effects. These types of combination models however did add complexity to the model. Delen favored decision trees because they were easier to understand

than the other algorithms. The ability to explain a model, especially when the prediction is wrong, may be more important to decision makers than accuracy as long as the accuracy is sufficiently high.

Strecht, Cruz, Soares, Mendes-Moreira, and Abreu (2015) examined seven data mining algorithms to predict whether students would pass or fail and what their final grade would be. Classification was used to determine if students would pass or fail, while regression was used to predict the final grade. The classification algorithms included Support Vector Machines, k-Nearest Neighbor, Random Forest, AdaBoost, Classification and Regression Trees, and Naïve Bayes, while the regression algorithms included Ordinary Least Squares, Support Vector Machines, Classification and Regression Trees, k-Nearest Neighbor, Random Forest, and AdaBoost.R2. The researchers found that for classification, Support Vector Machines had the most predictive power, followed by Classification & Regression Trees and Naïve Bayes. For regression, Support Vector Machines had the most predictive power while the Classification and Regression Trees had the least power. The researchers used small samples consisting of 700 courses with at least 100 students. This may have negatively impacted the predictive outcomes.

Herzog (2006) used logistic regression as a baseline for studying the effectiveness of data mining techniques for predicting student graduation times. Herzog examined incoming transfer students and freshmen with the intention of predicting who will graduate in 3 years and who will graduate in 6 years. Herzog found that neural network and decision tree algorithms were more effective than logistic regression for predicting students graduating in 3 years or less. The accuracy of these algorithms dropped when looking for students who would take 6 or more years to graduate. Pruned neural networks and decision trees were comparable to logistic regression in that case. The accuracy improved when looking only at freshmen and excluding the transfer

students. In this case, decision trees provided more accurate predictions. Herzog concluded that the accuracy of these data mining algorithms might improve if a larger set of input variables was examined. Tamhane et al. (2014) found this conclusion to be empirically true. Focusing on test scores for children between fourth and eighth grade, they found that their model predictions became more accurate as additional data was accumulated within that grade range.

Balakrishnan and Coetzee (2013) used Hidden Markov Models to predict student retention in a massively open online course (MOOC) using four student engagement indicators. These indicators were the number of times a student visited the course page, the accumulated percentage of videos watched, the number of discussion threads visited, and the number of discussion posts made. The researchers built Hidden Markov Models for each indicator and also used an ensemble approach that combined the indicators. Balakrishnan and Coetzee found that the Hidden Markov Models worked well at predicting positive outcomes, and the ensemble models were even better. However, the models were poor at predicting students who would drop out, which was explained as a balancing issue in which there are fewer engagement indicators for students who drop out.

### Measuring Model Effectiveness

Herzog (2006) used decision trees and neural networks to predict student retention and time to degree completion. Herzog explored predictive accuracy of these algorithms in the review of literature. Neural networks handle missing values and uncertainty better than other models. However, for neural networks to be effective, the sample size needs to be at least 500 due to the way that neural networks learn or are trained on the data input set. When analyzing

the effectiveness of output from neural networks, Herzog used the coefficients of determination ( $R^2$ ) and the number of accurate predictions.

Nandeshwar, Menzies, and Nelson (2011) used Bayesian networks because the model handles incomplete data well. Nandeshwar et al. also pointed out that the C4.5 algorithm available in the WEKA software uses decision trees in a way that deals well with missing data. In a thorough review of literature on data mining for student retention Nandeshwar et al. presented a set of equations to measure the predictive effectiveness of classifier models. Given  $TN$  = true negative,  $FN$  = false negative,  $FP$  = false positive, and  $TP$  = true positives, the following equations were used to determine various measures of models in the literature.

$$\text{true positive rate} = \text{recall} = \frac{TP}{FN + TP} \quad (1)$$

$$\text{false positive rate} = pf = \frac{FP}{TN + FP} \quad (2)$$

$$\text{precision} = \frac{TP}{FP + TP} \quad (3)$$

$$TP = \text{recall} * (FN + TP) \quad (4)$$

$$FP = pf * (TN + FP) \quad (5)$$

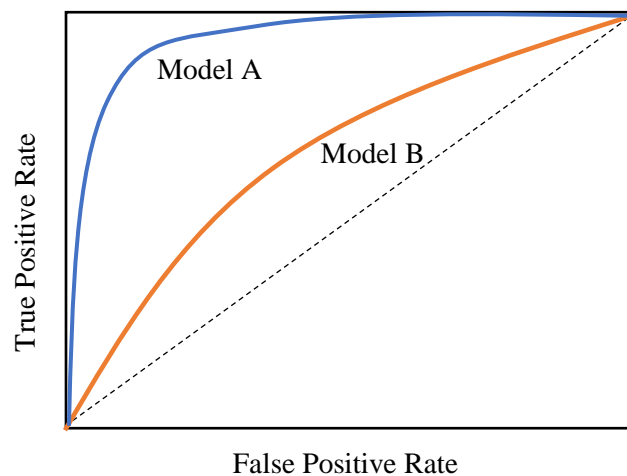
$$TN = FP * \frac{1}{pf - 1} \quad (6)$$

$$\text{accuracy} = \frac{TN + TP}{TN + FN + FP + TP} \quad (7)$$

These equations enable consistent evaluation for classifier models found in the research. They are an extension of the equations that Pittman (2008) described while examining the predictive ability of neural networks, logistic regression, Bayesian classifiers, and decision trees for student retention. Pittman further identified methods to measure predictive performance of classifier models. If the cost of false positives and false negatives is known, then a cost matrix can be used to further analyze a predictive model. Another method that is commonly used is



Receiver Operating Characteristic (ROC) analysis (Pittman, 2008, p. 81). ROC analysis consists of plotting the true positive rate, or recall, against the false positive rate. The resulting graph allows for simple intuitive interpretation of the accuracy of a classifier model (Hamel, 2008). Area under the ROC curve (AUC) can be calculated and used as a metric (Chawla, 2005). Figure 1 shows the characteristics of ROC curves. The dashed line that bisects the chart represents a 50% chance of accuracy. The line with label Model B is close enough to the dashed line that the model can be inferred to have poor predictive ability. The line with label Model A is much stronger in comparison.



*Figure 1.* Receiver Operating Characteristic (ROC)

Nandeshwar et al. (2011) used probability of detection, probability of false alarm, and variance between the two over cross-validation to test their models' predictive ability. Although classifiers were used, ROC was not specifically employed. The examination of detection rate versus false positive rate is essentially performing the same function as constructing an ROC graph, but it is less intuitive. Raju and Schumacker (2015) used only ROC to evaluate the effectiveness of the logistic regression, decision tree, and neural network models they used in

their study. Aguiar (2015) evaluated several methods for measuring model effectiveness in a study on improving early warning systems for students. Predictive accuracy was noted as a popular choice for evaluating classifiers. Aguiar pointed out that this method can be very misleading if used for imbalanced data sets.

Aguiar (2015) recommended the use of ROC and confusion matrices. Confusion matrices visually display true positives, false positives, true negatives, and false negatives. Figure 2 shows the layout of a confusion matrix. False positives are type I errors, while false negatives are type II errors (Witte & Witte, 2010).

		Prediction Outcome	
		Actual Positive	Actual Negative
Actual Outcome	Actual Positive	True Positive	False Negative (Type II Error)
	Actual Negative	False Positive (Type I Error)	True Negative

*Figure 2. Confusion Matrix*

Aguiar (2015) pointed out that ROC curves have a weakness in that they do not show the ratio of positive to negative associations in the dataset. ROC curves plot true positives to false positives. Precision-recall curves are similar to ROC curves but incorporate false negatives via the calculation for recall. Measuring true positives against both Type I and Type II enables a better interpretation of predictive model power with precision-recall curves (Davis & Goadrich, 2006).

Balakrishnan and Coetzee (2013) used a number of measures to gauge the effectiveness of their Hidden Markov Models. They used accuracy, precision, recall, the Matthews Correlation Coefficient, the ROC curve, and AUC. The F-score, or harmonic mean (Macari, 1985), was used to measure accuracy.

$$F1 = \frac{2(\textit{precision})(\textit{recall})}{\textit{precision} + \textit{recall}} \quad (8)$$

The Matthews Correlation Coefficient is similar to the Pearson Correlation Coefficient (Lund, Nielsen, Lundegaard, Kesmir, & Brunak, 2005).

$$MCC = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}} \quad (9)$$

The F1 score indicated that the Hidden Markov Models had weak predictive power. However, the AUC calculation showed that the predictive power of the models was actually quite strong.

Alpaydin (2010) provided a thorough treatment on evaluation methods for algorithms. Aside from ROC curve and confusion matrices for classifiers, several other means of evaluating data mining algorithms that are not classifiers were examined. McNemar's Test uses a structure similar to a confusion matrix to compare outcomes from two models. More common statistical tests were also listed including t test, Chi Square tests on cross-validated sets, analysis of variance (ANOVA), rank tests, Wilcoxon Signed Rank test, and the Kruskal-Wallis test. Alpaydin was a text-book introduction to machine learning as opposed to research into higher education issues. The majority of actual research in the area did not venture beyond measures of accuracy, recall, and ROC curves.

Cross-validation using k folds has been used to assist with model evaluations. The data set is divided into k mutually exclusive subsets. The model to be evaluated is then run against

the  $k$  folds. The model is trained on  $k - 1$  of the data sets. The model is then tested against the remaining data set. A mean and standard deviation can then be performed on the effectiveness outcomes of the model (Provost & Fawcett, 2013). This process is depicted in Figure 3.

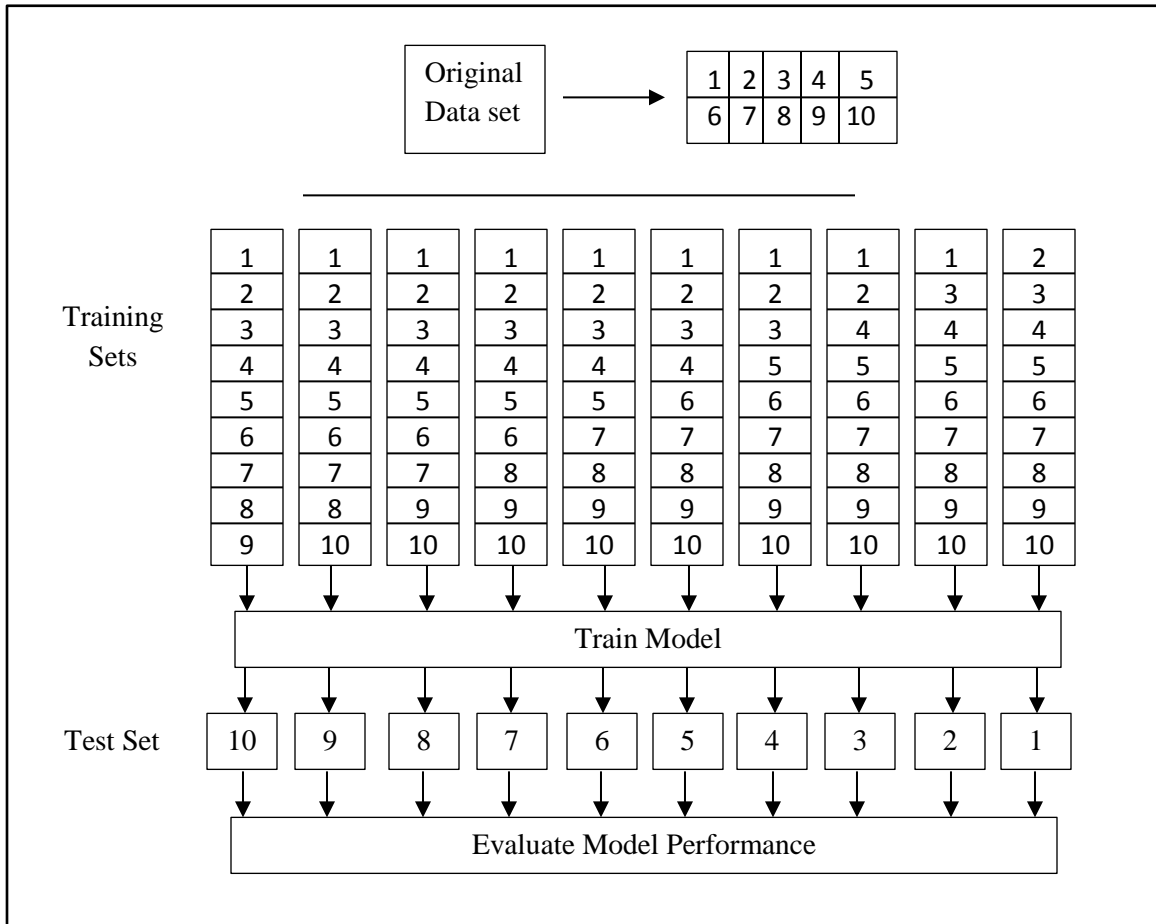


Figure 3. k Fold Validation

Kohavi (1995) stated that 10 was an optimal number of folds. Provost and Fawcett (2013) stated that 5 or 10 is an acceptable number of folds. Delen (2010) used 10-fold cross-validation to present an aggregated confusion matrix and to test accuracy of neural network, decision tree, support vector machine, and logistic regression models. Kabakchieva (2013) used 10-fold cross-validation to calculate an aggregated precision and true positive rate for decision tree, Bayesian

classifier, k-nearest neighbor, and rule learner models. Nandeshwar et al. (2011) used five-fold cross-validation to calculate probability of detection, probability of false alarm, and variance between the two for six different classifier models. Aguiar (2015) used 10-fold cross-validation to calculate for decision trees, naïve Bayes, random forest, and logistic regression models.

### Pitfalls to Avoid

Aguiar (2015) warned that imbalanced data sets can be very misleading. Classifying algorithms tend to have bias towards the majority class (Xu & Chow, 2006; Zhou & Liu, 2006). If a dataset contains a minority class that constitutes less than 35% of the total, then the dataset is imbalanced (Li & Sun, 2012). Imbalanced data sets are an important consideration in retention and completion studies when the number of nonreturners is greater than the number of students who persist to graduation, as is the case for universities formerly in the TBR system (THEC, 2016).

Kabakchieva (2013), in an introductory study, concluded that all of the models in the study had weak predictive power. Kabakchieva grouped students into five categories based on their total university score: excellent, very good, good, average, and bad. There were many more students in the very good and good categories than in the other categories. As a result, the predictive models barely registered output for the excellent, average, and bad categories. Strecht et al. (2015) found that none of the models they explored had impressive results. They used a sample size of 700 courses with at least 100 students per course. The intent of the study was to develop a model to predict student grades. It was not clear that the grades per course followed a normal distribution, so many more students may have passed the course than failed it. This would constitute an imbalanced data set scenario.

The issue of imbalanced data sets has been addressed primarily through over-sampling of the minority class and under-sampling of the majority class (Chawla, 2005; Thammisiri et al., 2014). Synthetic minority over-sampling (SMOTE) is another balancing technique that has been employed (Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Chawla, Lazarevic, Hall, & Bowyer, 2003; Han, Wang, & Mao, 2005; Thammisiri et al., 2014). SMOTE consists of generating a synthetic case based on the nearest neighbors of each minority case selected until the data sets are balanced. Thammisiri et al. used SMOTE in a study to predict freshmen attrition. Thammisiri et al. compared logistic regression, decision trees, artificial neural networks, and support vector machines using the original dataset, an over-sampled dataset, an under-sampled data, and a SMOTE dataset. The support vector machine using the SMOTE dataset performed best in terms of accurate classification. Delen (2010) used both imbalanced and balanced data sets in comparing data mining techniques. Delen used the under-sampling/over-sampling method to create the balanced set found that the balanced sets had better predictive value than the original, imbalanced set. Balancing the data set reduced bias. Burez and Van den Poel (2009) examined the impact of boosting and weighted random forests on imbalanced datasets. Weighted random forests were more accurate than random forests, but boosting did not outperform other techniques such as under-sampling or over-sampling.

### Data Mining Ethics

Ethical issues of data mining involve autonomy, transparency, privacy, and security (Beattie, Woodley, & Souter, 2014; Coglianese & Lehr, 2017; Daniel, 2015; Johnson, 2014, 2017; Jones, 2012; Richards & King, 2014). Johnson (2017) discussed data mining ethics from a structural justice perspective. Attempts to use predictive analytics in higher education violate

students' agency. Johnson explored the use of predictive analytics at Mount St. Mary's University where a survey was used to identify students with a high potential of being unsuccessful, with the ultimate goal being to dismiss those students. The president of the private institution resigned shortly after making remarks that the university needed to "drown the bunnies," meaning dismiss the students who had a low likelihood of being successful (O'Loughlin, 2016). Johnson (2017) also examined Austin Peay State University's Degree Compass system that recommended courses to students based on their prior academic history, and the EAB Student Success Collaborative that has been used to predict whether students are on track or need intervention based on their prior academic history. Johnson was critical of the black-box nature of these systems and the scientism, or strict adherence to the superiority of quantitative methods, employed to justify their use. The black-box nature of the systems made their results untrustworthy, while the impact on students was to guide them in certain directions instead of allowing the students to direct themselves. Johnson (2014) asserted that such violations of autonomy were paternalistic and unacceptable. Violations of autonomy should only be for exceptions such as when it may prevent waste of resources or when it is used to guide students lacking the knowledge or maturity to make optimal decisions. Beattie et al. (2014) expressed a similar deontological philosophy of how student data should be used. Beattie et al. described data mining analytics as creepy and intrusive, and they advocated for student data belonging to the student. Student data should be narrowly used only to improve learning outcomes. The analytic use of the data should be easily comprehensible to the student. Coglianese and Lehr (2017) stated that while machine learning algorithms are valuable for their accuracy, the mistrust presented by Johnson, Beattie et al., and others is due to the black-box nature of the technology. Coglianese and Lehr made the point that the black box nature of

machine learning techniques did not prevent the techniques from being examined and understood. Instead, machine learning was simply not as easily understood as more traditional analytical techniques. Given the complexity of the machine learning, transparency was identified as an important principle for the legitimacy of their use. Richards and King (2014) broadened the scope of transparency. Richards and King stated that organizations should be more open about how they collect, protect, use, and share data. This type of openness and transparency would garner more trust for data mining studies.

Privacy and security concerns for data mining and big data stem from the control of information (Johnson, 2014). Richards and King (2014) distinguished privacy as more than information that is kept secret. Privacy includes how information is used and shared. This nuanced definition led to the concept of confidentiality in which consumer information shared with providers is kept private. Richards and King made the point that individuals are willing to share personal information such as location tracking in order to use GPS for directions and cell phones for making phone calls. Dating sites use personal information to make matches, online bookstores use purchasing and browsing history to recommend new books to read, and social networking sites use personal information that is volunteered to find and connect friends. While organizations have been collecting consumer data for years, the technology of big data has enabled organizations to leverage the data collected to gain new insights into their customers. The customer may not want those insights being known according to Johnson (2014). Such personal insights can lead to manipulation of the customer. Johnson gave price discrimination and restrictive marketing as examples of such opportunistic uses of big data. Beattie et al. (2014) used the example of Facebook's naturalistic observation study in which the company measured how well it could manipulate its users' emotional states by presenting positive or negative posts



from friends (Kramer, Guillory, & Hancock, 2014). Richards and King (2014) stated that privacy and security positions are rapidly being added to organizations in an attempt to limit unethical uses and unintended consequences of data mining and big data. According to Richards and King all big data professionals should be concerned with protecting privacy and evaluating the ethics of their research. “Privacy by Design” (Richards & King, 2014, p. 430) was recommended as a way to eliminate the phenomenon of ad hoc data mining experiments and make privacy a central feature in experiments rather than an afterthought.

Chessell (2014) presented a pragmatic approach to the ethical use of big data and analytics that included legal considerations. Chessell stated that big data is “inherently ethics-agnostic.” Ethical use was presented as residing in the overlay of what is technologically possible, what is legally possible, and what an organization would like to do. Beattie et al. (2014) and Johnson (2014, 2017) discussed what organizations should do. Beattie et al. further discussed legal limits that institutions should place on themselves. A charter of student data rights was recommended as a means for institutions to proactively protect students’ data and to protect institutions from legal risks. Several sets of principles and codes such as the Belmont Report (NCPHS, 1978) were discussed, but actual laws affecting the use of big data were not discussed. Richards and King (2014) advocated for the establishment of legal rules to codify big data ethics. The Fair Credit Reporting Act was cited as a law enacted to protect financial consumer data. Richards and King discussed the establishment of data mining domain areas and domains where data mining should not be allowed. Voting was one such area, and the use of Twitter to sway a South Korean election was presented as an example of why the use of big data analytics in certain domains such as voting should be prohibited. Coglianese and Lehr (2017) examined the legal issues of machine learning algorithms in the context of use by federal

agencies. Anti-discrimination was a main principle that could be generalized to other organizations using data mining. Variables such as gender and ethnicity could lead to discriminatory results from machine learning models. However, the intent of such models is accuracy of prediction as opposed to discrimination. This difference led Coglianese and Lehr to conclude that machine learning models would not violate the equal protection requirements of the Constitution. Coglianese and Lehr warned against haphazard use of machine learning algorithms that could lead to distrust and subsequent legal issues over the implementation and use of the models.

### Chapter Summary

The rising cost of higher education has been attributed to numerous factors including cost disease, an increase in the demand for higher education, and decreases in real state appropriation dollars per full-time student over the past 3 decades. The decrease in state support has coincided with the rise of performance funding and reporting initiatives with the intent to hold public higher education accountable for outcomes. An unintended consequence of these initiatives has been an abundance of data about students and institutional operations that can be used for data mining. This data collection has been in response to both state and federal reporting requirements. The data collected has mainly been sociodemographic data, allowing the type of data collected to be uniform across states. Data on student experiences, peer environments, and external pressures have not been collected, although much research has been conducted in those areas. The intent of performance funding and reporting is to improve efficiency, yet the data collected does not fully empower institutions to analyze their operations and make improvements based on well-studied student retention research.

GPA and credit hour accumulation was a common predictive variable for studies in retention, transfer students, and data mining. Both data points could be interpreted as a proxy for the student attitudes examined by researchers such as Astin, Pascarella and Terenzini, and Reason. The data points are also commonly collected for state and federal reporting requirements. While retention research points to financial factors affecting students, such data points are not typically included in state reporting, but can be found in federal reporting. Institutional characteristics can also be found in federal reporting. Data mining can be used to take advantage of these different sources of data and find interesting patterns, despite the absence of other data points about student experiences.

Data mining's origins can be traced back to the late 1980s (Coenen, 2004). With the growth in computing power in the ensuing years, interest in data mining has grown tremendously. Despite over 20 years of research and countless publications, terminology, methodology and evaluation is still inconsistent. The CRISP-DM process is a positive step towards consistency in methodology. Measuring accuracy via precision, recall, ROC, and AUC appears to be more common among researchers. Many of the same machine learning algorithms are used in studies allowing researchers to see what works well in a certain domain.

The tools used for most research has remained consistent. WEKA is free and provides researchers with many of the data mining algorithms programmed into it. SPSS is a widely used academic software. SAS Enterprise Miner is a powerful tool used by statisticians and programmers. The statistical programming language R has not been seen in the research yet, but as data mining continues to merge the fields of computer science and statistics, R will likely be used as much as the other options.

Data mining and classical statistics are not mutually exclusive. The two can be used in tandem to better understand issues within the data. Logistic regression has typically been used as a baseline model of comparison for other data mining algorithms. Factor analysis can be used in a similar fashion to cluster analysis to reduce datasets down to heterogeneous groups.

Initial variable selection in data mining studies appears to be based on review of the literature for the problem domain. Sensitivity analysis is performed on variables after predictive models have been generated. Tests for variable appropriateness are not typically conducted prior to the creation of the models. Factor analysis and/or clustering of potential variables could further strengthen predictive models by ensuring that significant variables are included and insignificant variables are excluded.

The models used for prediction are predominately logistic regression, neural networks, variants of decision trees, ensemble methods, and support vector machines. Through the execution of these models, variables are shown to have certain predictive power within the model. Support vector machines and ensemble methods involving decision trees appear to be the most powerful predictive models. Neural Networks and logistic regression follow closely behind. Considerations for selecting data mining techniques for higher education research should include the ease of understanding the model and how well the model handles missing or incomplete data. Higher education data can have much incomplete data, especially if the data are self-reported by students through web interfaces.

There are numerous ways to measure the predictive power of data mining models. The simplest is to examine precision, recall, and accuracy based on true and false hits and misses. A better measure is the Receiver Operating Characteristic graph and the Area under the Curve calculation.

A noted pitfall to avoid with data mining is the use of imbalanced data sets. Imbalanced data sets can obscure the detection of minority datasets (Lu, Wang, Yang, & Zhao, 2011). This can be a problem for studies that use data mining to examine minority datasets. For example, the imbalanced data issue could be problematic in studying students who transfer from one 4-year institution to another and graduate.

Ethical issues of data mining include concerns about privacy, security, autonomy, and transparency. Privacy and security involve how collected information is stored, shared, and used. Autonomy stems from privacy concerns as advocates for autonomy fear paternalistic interference with individual agency based on the use of big data. Transparency is necessary to minimize ad hoc and unethical studies, build trust in predictive models, and limit legal risks.

## CHAPTER 3

### RESEARCH METHODOLOGY

This was a nonexperimental quantitative study involving data mining techniques. The first 10 research questions in this study focused on the development of predictive models for students who began college at a Tennessee university, transferred to another institution, and graduated. Five data mining techniques were selected based on their use in the review of the literature in other educational data mining studies. The five data mining techniques were logistic regression, decision trees, random forests, artificial neural networks, and support vector machines. The majority of the Cross Industry Standard Process for Data Mining (CRISP-DM) were followed for this research. The six steps of the CRISP-DM are 1) understanding the business domain, 2) identifying data sources, 3) extracting, transforming, and loading the data, 4) developing models to examine the data, 5) evaluating the models, and 6) using the models in the decision-making process (Delen, 2010). Step one was accomplished in Chapter 2 of this study. Step two was accomplished in this chapter. Steps three through five were accomplished in Chapters 4 and 5 of this study. Step six was dependent on the outcome of Chapters 4 and 5, and beyond the scope of this study.

#### Research Questions and Null Hypotheses

This study had 10 general research questions as listed below. Research Questions 1 – 4 were descriptive questions, while Research Questions 5 – 10 were associated with one or multiple research hypotheses.

Research Question 1.

Which characteristics are most likely to predict graduation in six years after enrollment for first-time freshmen students under the age of 24 from a 4-year institution in Tennessee?

Research Question 2.

What characteristics identify first-time freshmen students under the age of 24 who transfer to another higher education institution?

Research Question 3.

What characteristics identify first-time freshmen students under the age of 24 who did not graduate from a 4-year institution in Tennessee?

Research Question 4.

What characteristics identify first-time freshmen students under the age of 24 who began at a 4-year institution in Tennessee, transferred to another higher education institution, and graduated?

Research Question 5.

Is the predictive power of logistic regression for determining which students will leave their home institution and graduate elsewhere greater than 50%?

H<sub>0</sub>5: The predictive power of logistic regression for determining which students will leave their home institution and graduate elsewhere is less than or equal to 50%.

Research Question 6.

Is the predictive power of decision trees for determining which students will leave their home institutions and graduate elsewhere greater than 50%?

H<sub>0</sub>6: The predictive power of decision trees for determining which students will leave their home institution and graduate elsewhere is less than or equal to 50%.

Research Question 7.

Is the predictive power of random forests for determining which students will leave their home institution and graduate somewhere else greater than 50%?

H<sub>0</sub>7: The predictive power of random forests for determining which students will leave their home institution and graduate elsewhere is less than or equal to 50%.

Research Question 8.

Is the predictive power of artificial neural networks for determining which students will leave their home institution and graduate somewhere else greater than 50%?

H<sub>0</sub>8: The predictive power of artificial neural networks for determining which students will leave their home institution and graduate elsewhere is less than or equal to 50%.

Research Question 9.

Is the predictive power of support vector machines for determining which students will leave their home institution and graduate somewhere else greater than 50%?

H<sub>0</sub>9: The predictive power of support vector machines for determining which students will leave their home institution and graduate elsewhere is less than or equal to 50%.

Research Question 10.

Which of the data mining techniques: logistic regression, decision tree, random forest, artificial neural network, or support vector machines provide the best classification result in predicting students who will leave their home institution and graduate somewhere else?

H<sub>0</sub>10<sub>1</sub>: Logistic regression has no stronger predictive power than decision trees, random forests, artificial neural networks, or support vector machines.

H<sub>0</sub>10<sub>2</sub>: Decision trees have no stronger predictive power than logistic regression, random forests, artificial neural networks, or support vector machines.



H<sub>0</sub>10<sub>3</sub>: Random Forests have no stronger predictive power than logistic regression, decision trees, artificial neural networks, or support vector machines.

H<sub>0</sub>10<sub>4</sub>: Artificial neural networks have no stronger predictive power than logistic regression, decision trees, random forests, or support vector machines.

H<sub>0</sub>10<sub>5</sub>: Support vector machines have no stronger predictive power than logistic regression, decision trees, random forests, or artificial neural networks.

### Population

A purposeful sample procedure was used for this study. The sample for this study was all first-time freshmen under the age of 24 in the fall terms between 2006 and 2009 who attended a 4-year institution in the former TBR system. This population was selected because those students were expected to graduate within 6 years of their first fall semester, and 2009 is the earliest fall semester for the cohort graduating in May 2016.

The size of the population for this study was approximately 40,000 first-time freshmen. Large effects can be found in studies with small samples while small effects can be found in studies with large sample sizes (Witte & Witte, 2010). Finding these small effects within large sample sizes is the allure of big data research studies such as this one.

Two subgroups exist due to the nature of this study. The first subgroup was the set of students who transferred from their initial institution and graduated from some other institution. For this study, only students who attended in the fall and spring semesters of their first year, transferred, and graduated were used in this subgroup. The second subgroup was the set of remaining students from the sample. This subgroup was necessary for generating the predictive models. Probability is defined as the number of observed outcomes divided by the number of

possible outcomes. Thus while subgroup one was the focus of this research, both subgroups were necessary for training the predictive models to identify students with the highest probability of transferring and graduating.

### Instrumentation

The TBR required each member institution to submit a census extract file of student records on the 14<sup>th</sup> day each semester. In addition, each institution submitted a report of graduates at the end of the academic year. These files were then sent to the Tennessee Higher Education Commission (THEC). Along with these files, TBR institutions and THEC were and continue to be members of the National Student Clearinghouse (NSC).

The NSC enables member institutions to track students who transfer out. The NSC provides the name of the institutions that the students transferred to, the first semester of attendance, the last semester of attendance, the major pursued there, and whether the students graduated. IPEDS surveys collect information about institutions receiving federal funds. This information includes institution size, Carnegie classification, whether the institution is in a rural or metropolitan area, and how much the institution spends on academic and nonacademic support. The reporting files from the former TBR, data from IPEDS surveys, and data that THEC collects from the NSC comprised the instrument for this study.

### Data Collection

The data for this study came from the Tennessee Higher Education Commission (THEC). THEC keeps records of first-time freshmen students and whether the students graduate from their home institution, another Tennessee institution, from an out of state school, or not at all.

THEC uses the National Student Clearinghouse to identify students who transfer from their initial institution and graduate elsewhere. THEC has been collecting this data since the late 1990s, making the data gathering and storage highly standardized. The data were collected, kept, and used according to a THEC Data Sharing Agreement between THEC and the researcher. THEC stripped away any identifying information prior to sending any data to the researcher. No identifying information of students was available to the researcher. The data were kept secure using an encrypted, password protected drive that was kept with the researcher when in use and locked in an office when not in use.

The data collected by THEC has mainly been sociodemographic. However, certain data elements can be used to infer the peer environment at institutions. For example, high school and zip code data can be used to determine how many students from a particular high school or geographical area attended the same institution. The amount of Pell grant funds for a student can be used to infer external economic pressures on the student. These types of data points were used in this study. In addition, commonly used predictive data points such as GPA, ACT scores, and credit hour accumulation were used. Other demographic data points that were used included age, gender, ethnicity, and major. Information about the student's initial and graduation institutions were also examined. The IPEDS data system was used to determine institutional expenses on instruction, student services, and administration, institutional revenues from tuition and fees, state appropriations, and institutional size in terms of staff and students.

THEC provided six files with data about students including demographic information, enrollment information, and graduation information. Detailed information about these files is shown in Table 1.

Table 1

*Data Files*

<b>File</b>	<b>File Description</b>
Clearinghouse Enrollment	Contains records for each term a student was enrolled at a transfer institution
Clearinghouse Grads	Contains records of students who have graduated from a transfer institution
Demographics and FTF Enrollment	Contains demographic information for first-time freshmen from a 4-year TBR institution
FAFSA	Contains financial aid information for students
THEC Awards	Contains records of students who have graduated from a TN institution
THEC Enrollment	Contains records for each term a student was enrolled at a TN institution

These files contained 87 variables. Of those 87, sixteen variables were duplicated for the purpose of joining student records between files. In addition, there were 20 variables that duplicated information from other data points. These duplicated variables were either used once for joining the files or excluded from analysis. Another 24 variables were excluded from analysis because they could not be used as continuous or categorical variables in the factor analysis and predictive models. Factor analysis relies upon ordinal or continuous variables (Bartholomew, 1980; Yong & Pearce, 2013). Several categorical variables were thus used to generate continuous variables that could be used for factor analysis. For instance, student high school and majors were used to determine the number of other students from the same high school or pursuing the same major. Appendix I contains information about the THEC files and variables that were selected for the final merged data set.

In addition to the data files from THEC, several variables from IPEDS were collected on institutions that students first attended and graduated from. Staffing and enrollment variables included the number of full-time faculty, full-time nonfaculty staff, and full-time equivalent (FTE) students. Institution finance variables included tuition and fees, state appropriations, and expenses for instruction, research, academic support, and institutional support. The final data set used 30 variables from the original files and 20 IPEDS variables. The merged data set is shown in Table 2.

Table 2

*Merged Dataset*

<b>Variable Name</b>	<b>Description</b>
uniqueID	Student ID
ftf_age	Age when student was a first-time freshman
ftf_distance_from_home	Distance between student's FTF institution and their permanent address
transfer_distance_from_home	Distance between student's transfer institution and their permanent address
overallHSGPAGED	Student's high school GPA
ACTComposite	Student's composite ACT
FTF_Year	Year that the student was a first-time freshman
FTF_Semester_creditshours	The number of credit hours taken during the student's first semester
total_FTF_TSAA_Payment	Total TN state aid student received during their first semester
total_hours_at_ftf_inst	Total hours earned at the student's initial institution
total_semesters_at_ftf_inst	Total number of semesters student attended initial institution
total_semesters_after_ftf_inst	Total number of semesters student attended transfer institutions
hs_peers_cnt	Number of FTF that attended the same institution that were also from the student's high school
ftf_major_peers	Number of FTF that attended the same institution that were in the same major
ftf_major_changes	The number of major changes the student made
avg_term_creditshours	The average credit hours taken per semester by the student
graduation_indicator	An indicator of whether the student ever graduated from anywhere

<b>Variable Name</b>	<b>Description</b>
tn_4yr_grad_ind	An indicator of whether the student graduated from a 4-year TN school
transfer_grad_indicator	An indicator of whether the student transferred out and graduated
transferred_ind	An indicator of whether the student transferred out
Gender	The gender of the student
Racename	The IPEDS race/ethnicity of the student
residencyandcitizenshipstatus	The residency (in-state, out-of-state) or citizenship status (foreign) of the student
ftf_major	The major the student entered with
hs_name	The name of the student's high school
ftf_instname	The name of the first institution the student attended
thec_grad_inst	The name of the TN institution that the student graduated from
nsc_grad_inst	The name of the transfer institution that the student graduated from
ftf_full_time_faculty	The number of full-time faculty at the first institution the student attended
ftf_full_time_nonfaculty	The number of full-time non-faculty staff at the first institution the student attended
ftf_tuition_and_fees	The total tuition and fees collected at the first institution the student attended
ftf_state_approps	The amount of state appropriations for the first institution the student attended
ftf_instruction	Instruction expenses at the first institution the student attended
ftf_research	Research expenses at the first institution the student attended
ftf_acad_support	Academic support expenses at the first institution the student attended
ftf_stu_support	Student support expenses at the first institution the student attended
ftf_instit_support	Institutional support expenses at the first institution the student attended
ftf_total_fte	The number of full-time equivalent (FTE) students at the first institution the student attended
grad_full_time_faculty	The number of full-time faculty at the institution from which the student graduated
grad_full_time_nonfaculty	The number of full-time non-faculty at the institution from which the student graduated
grad_tuition_and_fees	The total tuition and fees collected at the institution from which the student graduated

<b>Variable Name</b>	<b>Description</b>
grad_state_approps	The amount of state appropriations at the institution from which the student graduated
grad_instruction	Instruction expenses at the institution from which the student graduated
grad_research	Expenses for research at the institution from which the student graduated
grad_acad_support	Academic support expenses at the institution from which the student graduated
grad_stu_support	Student support expenses at the institution from which the student graduated
grad_instit_support	Institutional support expenses at the institution from which the student graduated
grad_total_fte	The number of full-time equivalent (FTE) students at the institution from which the student graduated

Students without high school GPAs and ACT scores were excluded from the analysis. In addition, students had to appear in the THEC Enrollment file. This produced a final data set with 39,379 cases.

Data Analysis

A factor analysis was performed on a set of student characteristics relating to graduation for the first four research questions. The continuous variables used in the factor analysis were each standardized as scales for the variables could vary. A principal component analysis with varimax rotation was used. Varimax rotation produces factors that are easier to interpret (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Varimax is an orthogonal rotation. Orthogonal rotations are used when the researcher suspects that factors are not correlated (Pett, Lackey, & Sullivan). The factor analysis used only the subgroup of students who transferred and graduated. The factor analysis was used to retain or remove student characteristics that did not contribute to the variance between cases. The remaining set of characteristics were then converted to factor scores which were used to construct the predictive models for Research

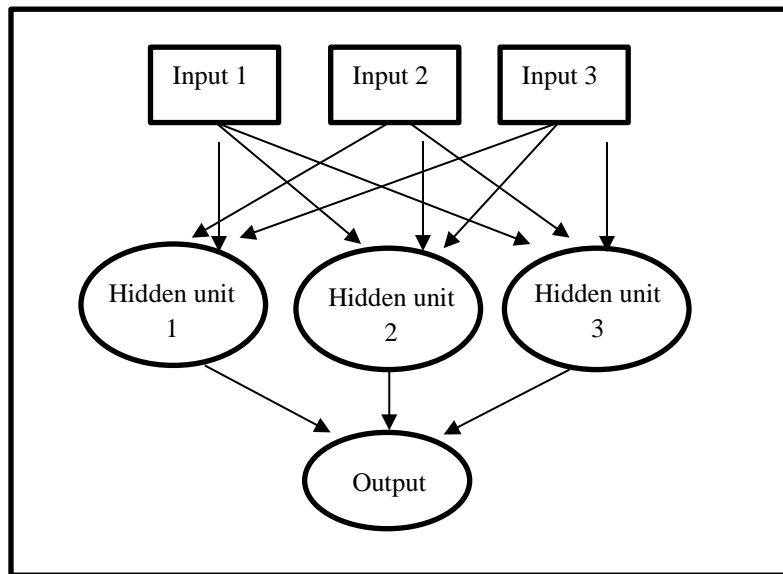
Questions 5 – 9, along with categorical variables that could not be included in the factor analysis. All variables for the predictive models were normalized to ensure the variables had the same unit scale (Pittman, 2008). Normalizing variables was essential for the artificial neural network and support vector machine models to produce accurate results due to the underlying algorithms of the models (Larose & Larose, 2015).

A typical logistic regression model was used to address Research Question 5. There are a number of decision tree algorithms that could have been used to address Research Question 6. Classification and regression tree (CART), Chi-squared automatic interaction detection (CHAID), Iterative Dichotomiser 3 (ID3), and C4.5, and C5 are commonly used algorithms (Thammasiri et al., 2014). C5 is an extension of C4.5, which in turn is an extension of ID3 (Quinlan, 1986, 1993). C4.5 and CART have been popular choices and listed in the top 10 data mining algorithms (Wu et al., 2007). C5 is more robust with missing values, and C5 is a faster algorithm than CART. CART and C4.5 were selected due to their popular use and subsequent availability in statistical software packages. A random forest algorithm was used to address Research Question 7. A random forest combines the results of multiple decision tree iterations into one aggregated result. The result of each iteration is weighted equally in a process called bagging.

As with decision trees, there were a number of different artificial neural networks (ANNs) that could have been used to address Research Question 8. Feed-forward ANNs have at least three layers that data moves forward through. A representation of a feed-forward ANN is presented in Figure 4. Recurrent ANNs allow data to move forward or backward through the model enabling a sort of memory. Multilayer perceptron (MLP) and radial basis function (RBF) networks are commonly used feed-forward ANNs (Witten & Frank, 2011). The multilayer

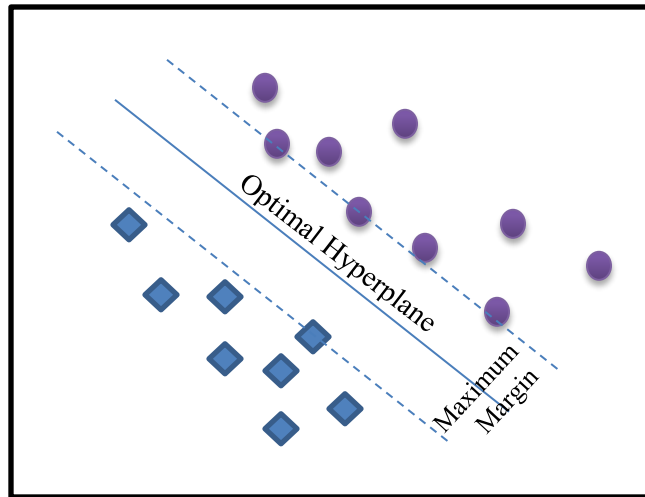


perceptron is the most commonly used feed-forward ANN (Oztekin, 2016) and was thus used to address Research Question 8.



*Figure 4.* Multilayer Perception ANN

Support vector machines (SVM) are similar to neural networks. SVMs that use a sigmoid transformation function produce a multilayer perceptron network. SVMs that use a Gaussian transformation function produce a radial basis function network. Standard support vector machines use a polynomial transformation function to construct maximum margin hyperplanes (Witten & Frank, 2011). A representation of a support vector machine with hyperplanes is shown in Figure 5. A support vector machine was used to address Research Question 9.



*Figure 5. Support Vector Machine*

For each predictive model, the data set was split into a training set and test set for each model per the literature. The strength of the models was examined using common predictive modeling techniques including accuracy, precision, recall, F1 score, receiver operator curve, area under the ROC curve (AUC), and confusion matrices. A 10-fold cross-validation method was used for splitting the data sets and evaluating the models. The area under the ROC curve was used as the evaluation metric for the 10-fold cross-validation.

The AUC measures the probability of correctly ranking a randomly selected case as a true positive. The AUC thus corresponds to the Mann-Whitney U test. The AUC values range from 0.5 (random chance) to 1 (perfect prediction) (Hanley & McNeil, 1982; Fawcett, 2006). The AUC was used to address Research Questions 5 - 10 pertaining to the predictive strength of each model. A simple comparison of model strength, rather than rigorous statistical testing, was used to answer Research Question 10.

The predictive models used both subgroups of the population. The training and test sets drew from the union of both subgroups. The two subgroups of students formed one sample from which the training and test sets were selected. Synthetic minority over-sampling (SMOTE) was

used to address the imbalance between subgroups. A Morris sensitivity analysis was used in addition to Research Questions 5 – 10 to determine which variables had the most predictive power in each model. A sensitivity analysis determines the predictive power of a variable by examining how the performance of the model changes when the variable is excluded from the model (Davis, 1989).

SPSS Modeler is an addition to the basic IBM statistical software package. Modeler provides a number of data mining models including decision trees, random forests, neural networks, logistic regression, and support vector machines (IBM, 2016). SAS provides similar functionality with their Enterprise Miner product, which is also a separate product from the basic statistical software option. The R statistical programming language is robust and free to download and use (The R Foundation, n.d.). The Python programming language offers numerous statistical packages such as pandas, NumPy, SciPy, and scikit-learn (Provost & Fawcett, 2013). WEKA is a data mining software package from the University of Waikato in New Zealand (Witten & Frank, 2011). Each of these software options comes with a graphical user interface and a command line interface. The advantage of a command line interface is that the code for the predictive models can be automated. Automation reduces time spent on running the model, reduces the chance of error due to human intervention, and maximizes time spent on applying the information from the output of the model (Hunt & Thomas, 2000). The ability to automate is important for the sixth step in the CRISP-DM process, which is to use the predictive model in the decision-making process (Delen, 2010). WEKA, R, and Python have the advantage of being freely available. R is specifically designed for statistical applications while Python can be used for web programming, desktop application programming, statistical programming, and

many other types of programming. Python is the more extensible option in terms of statistical software. WEKA offers a variety of stock algorithms per data model.

The SPSS statistical software package was used for the principal component analysis. Python was used to generate and evaluate the predictive models. Python was used to create scripts that can be automated.

### Chapter Summary

This was a nonexperimental quantitative study. The rationale of this study was to use the massive amount of data that TBR and THEC have collected as a means to better understand a retention issue and make predictions based on the data. A factor analysis was used on the initial dataset to explore which data elements should be used in the predictive models. Selecting elements that were associated with strong factors for the population increased the strength of the subsequent predictive models. Logistic regression, artificial neural networks, decision trees, random forests, and support vector machines were the predictive models used in this study. The predictive power of these models was examined using accuracy, precision, recall, F1 score, receiver operator curve, area under the ROC curve (AUC), and confusion matrices. The AUC was used as the main criteria for comparing each model. SPSS was used to conduct the factor analysis on the initial dataset. Python was used to generate and evaluate the predictive models. Python was used to ensure the models can be automated for compliance with the CRISP-DM process.

## CHAPTER 4

### DATA ANALYSIS AND RESULTS

The purpose of this quantitative study was to discover factors about first-time freshmen who began at a university in the former TBR system, transferred to any other institution after their first year, and graduated with a degree or certificate. In addition, this study sought to determine if a predictive model can be generated to identify these particular students prior to their initial departure. Factor analysis was conducted on a set of variables gathered from the Tennessee Higher Education Commission (THEC) and the Integrated Postsecondary Education Data System (IPEDS). A principal component analysis (PCA) was conducted twice for each research question concerned with characteristics of first-time freshmen retention decisions. An initial PCA was conducted to identify characteristics based on knowledge of where students transferred to and graduated from. A second PCA was conducted to identify characteristics based only on knowledge of the student in his or her first semester. This second PCA was instrumental to the research questions concerning predictive models. If there were no variables from that first semester that were useful for prediction, the predictive models would have limited to no use.

Factor analysis was performed for the first four research questions. The inputs for the factor analysis needed to be ordinal or continuous variables (Bartholomew, 1980; Yong & Pearce, 2013). The output of the factor analysis for the fourth research question dealing with students who transfer and graduate elsewhere was used along with a set of additional categorical variables about each student as the inputs to five predictive models: logistic regression, a decision tree, a random forest, an artificial neural network, and a support vector machine. Once

the factor analysis was completed for the fourth research question, an analysis was performed on the categorical variables to determine if the variables were appropriate for inclusion in the predictive models. The analysis of the categorical variables was included in this chapter prior to the results of the research questions for the sake of clarity.

### Categorical Analysis

Four categorical variables were used in the predictive models along with the resulting factor scores from the factor analysis of Research Question 4 concerning students who transfer and graduate somewhere other than their initial institution. The categorical variables were gender, race, residency, and first-time freshman institution. The distributions for the first two factor score variables were not normal (Figure 6, Figure 7), so nonparametric tests were performed on the categorical variables to determine if they were appropriate to include in the predictive models. The distributions for the other two factor score variables were normal (Figure 8, Figure 9), however nonparametric tests were still used for uniformity of analysis. The categorical variable would have been excluded if any of the categorical variables had no significant impact on the factor scores.

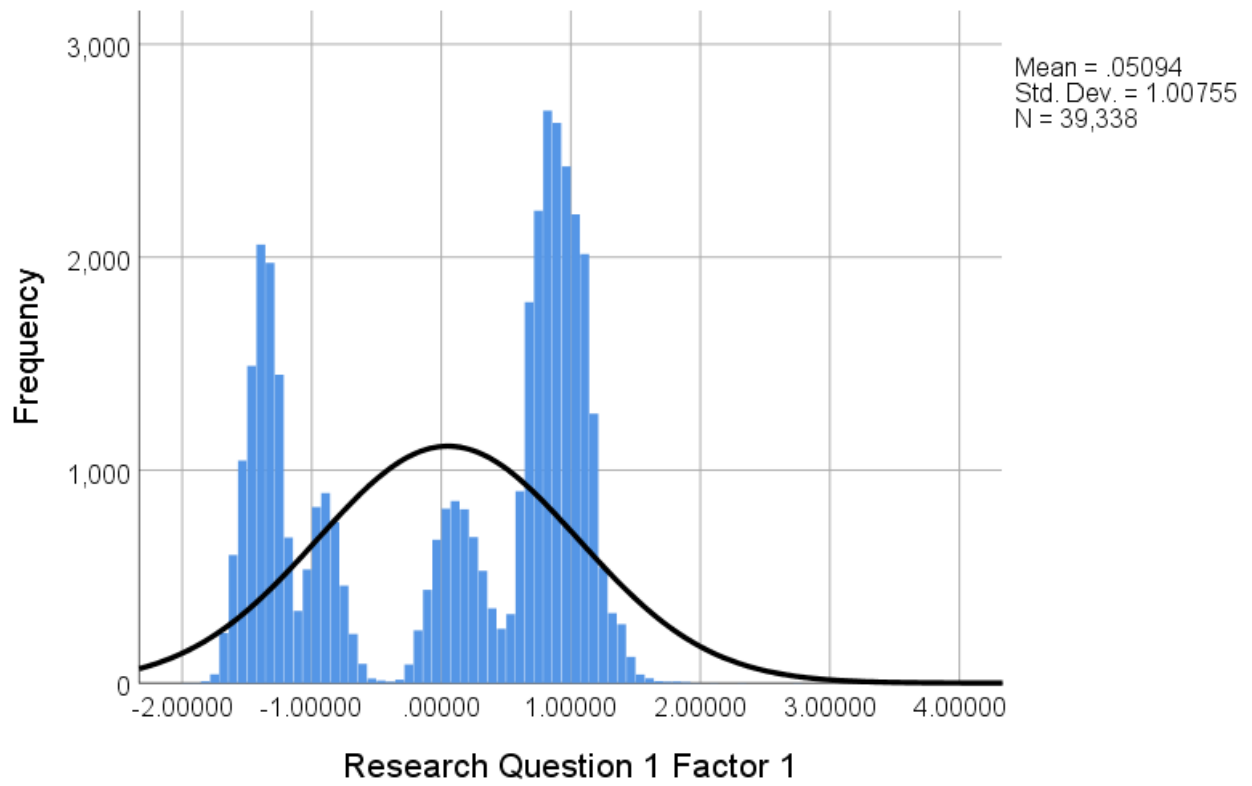
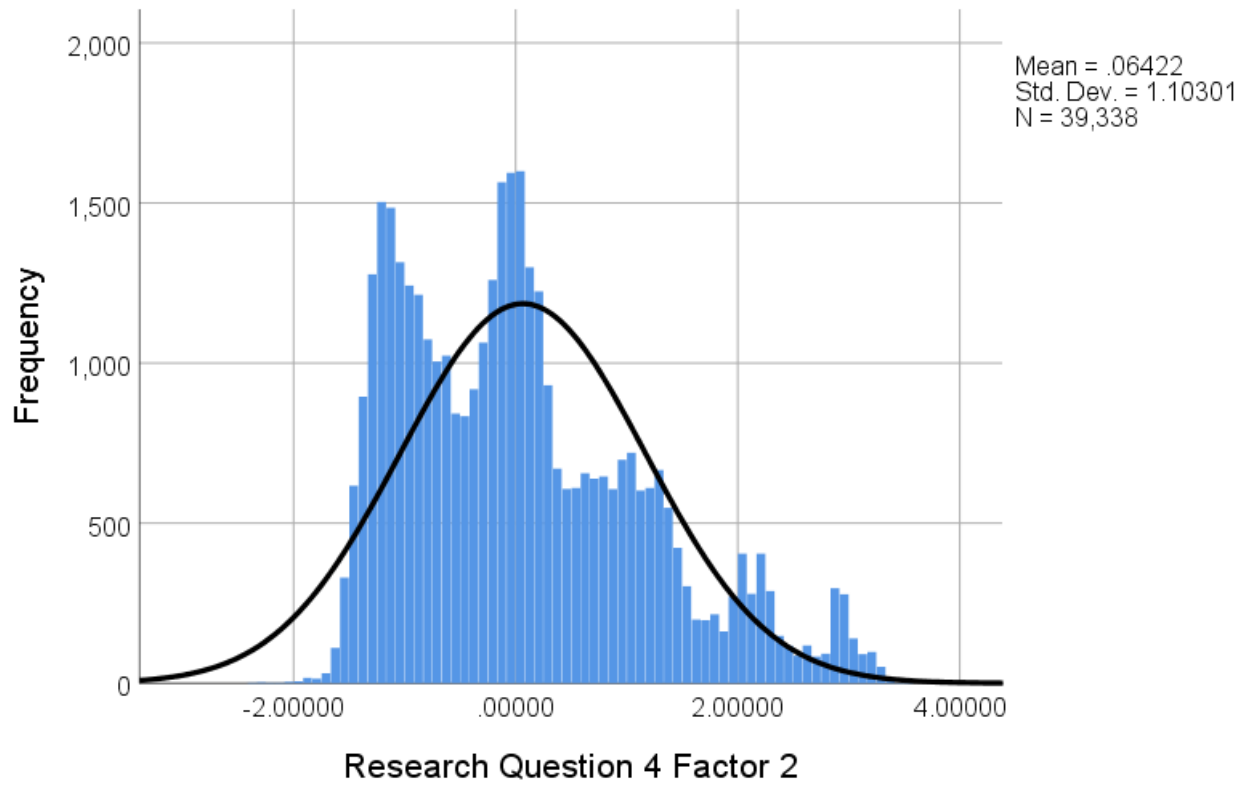
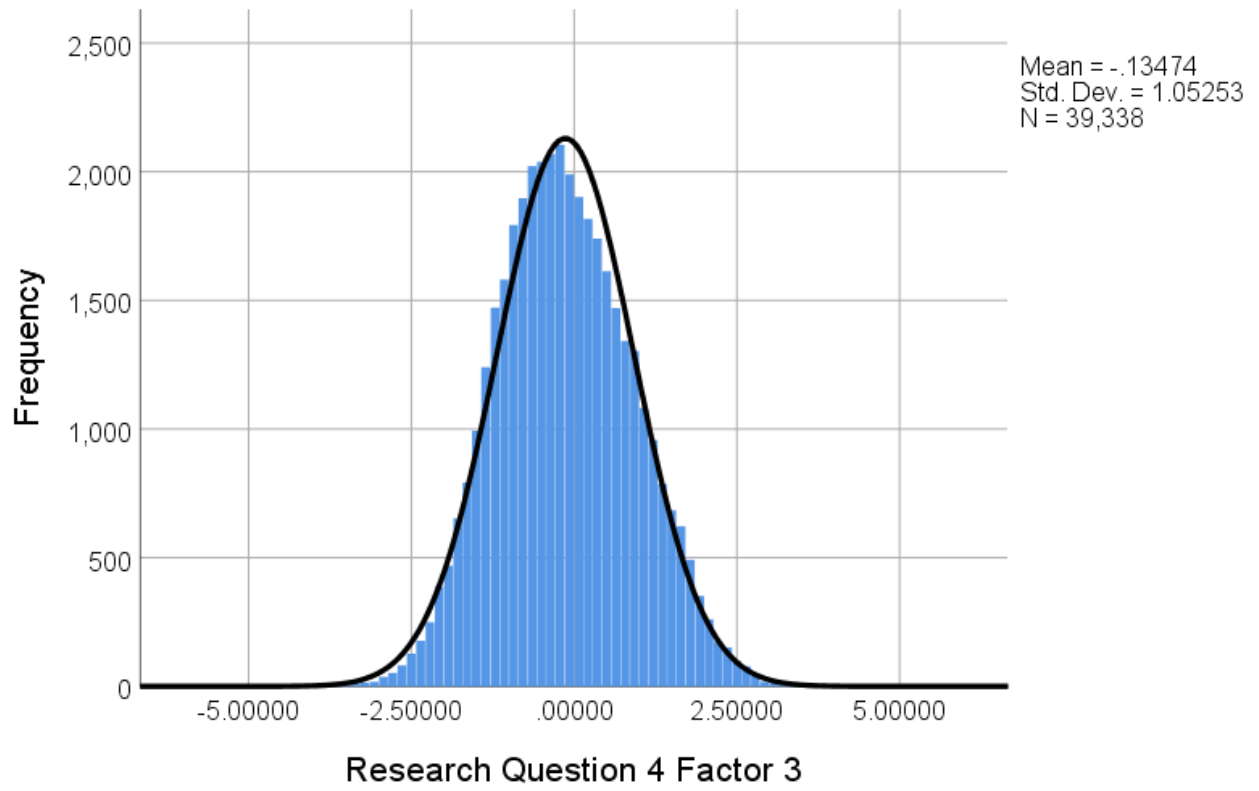


Figure 6. Transfer Graduates Factor 1 Histogram

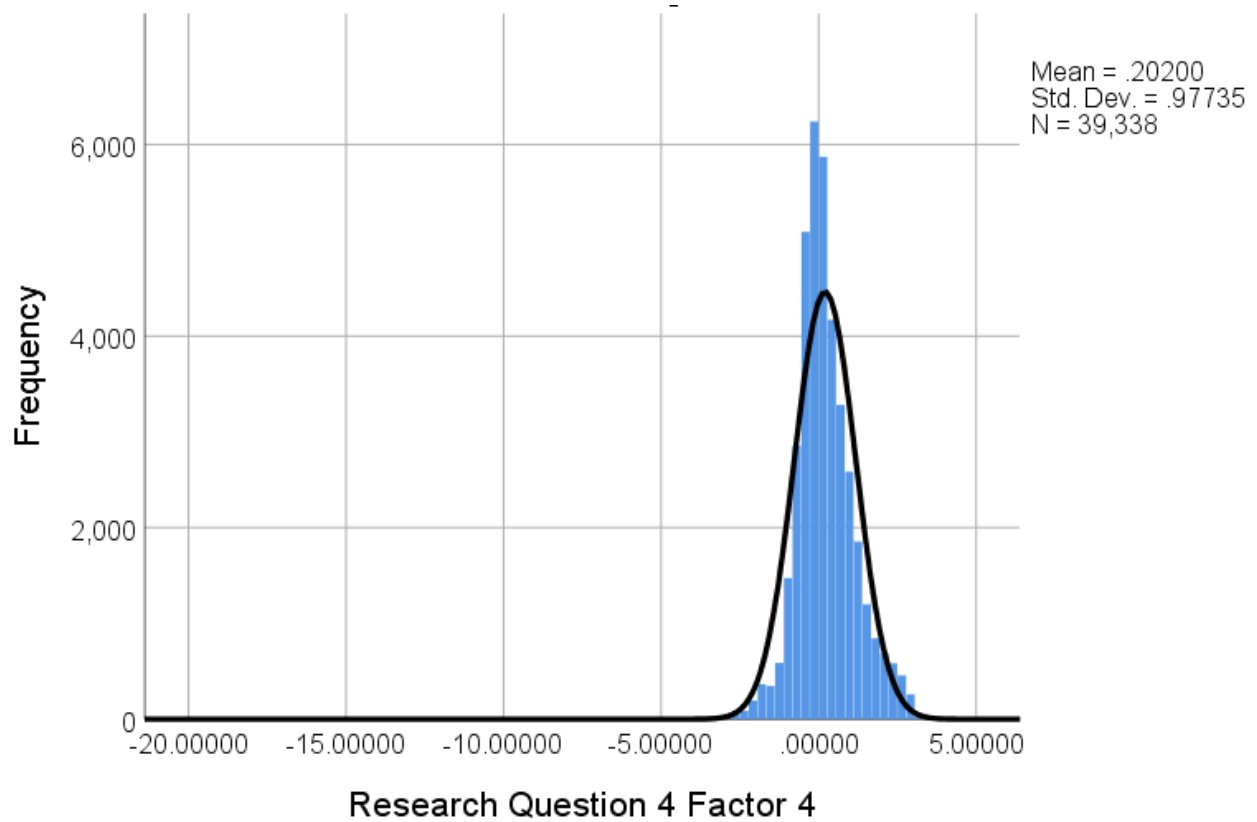


*Figure 7. Transfer Graduates Factor 2 Histogram*



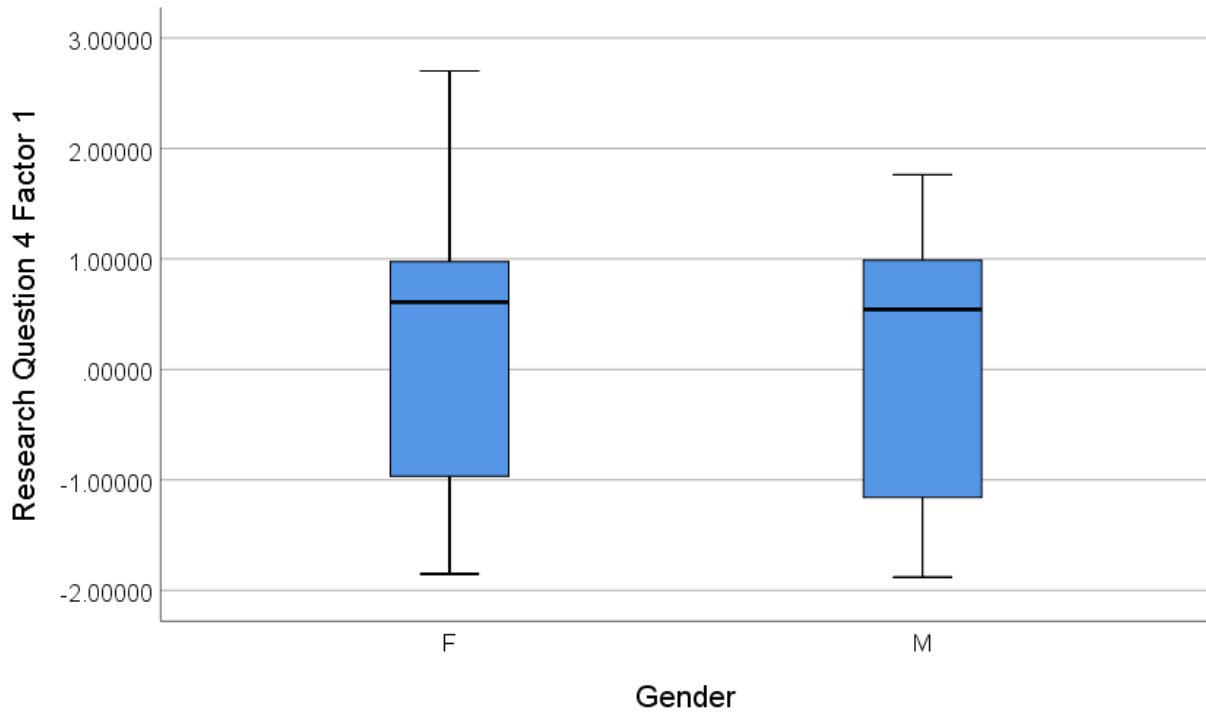


*Figure 8.* Transfer Graduates Factor 3 Histogram



*Figure 9.* Transfer Graduates Factor 4 Histogram

A Mann-Whitney U test was conducted to determine if the first factor score variable differed by gender. The results of the test were not significant,  $z=0.147$ ,  $p=0.883$ . Figure 10 shows the distributions of the scores on the first factor score for the two groups.



*Figure 10. Mann-Whitney U Test for Gender, Factor 1*

A Mann-Whitney U test was conducted to determine if the second factor score variable differed by gender. The results of the test were significant,  $z=10.007$ ,  $p<0.001$ . Men had a average rank of 19,030, whereas women had an average rank of 20,183. Figure 11 shows the distributions of the scores on the second factor score for the two groups.

A Mann-Whitney U test was conducted to determine if the third factor score variable differed by gender. The results of the test were not significant,  $z=1.517$ ,  $p=0.129$ . Figure 12 shows the distributions of the scores on the third factor score for the two groups.

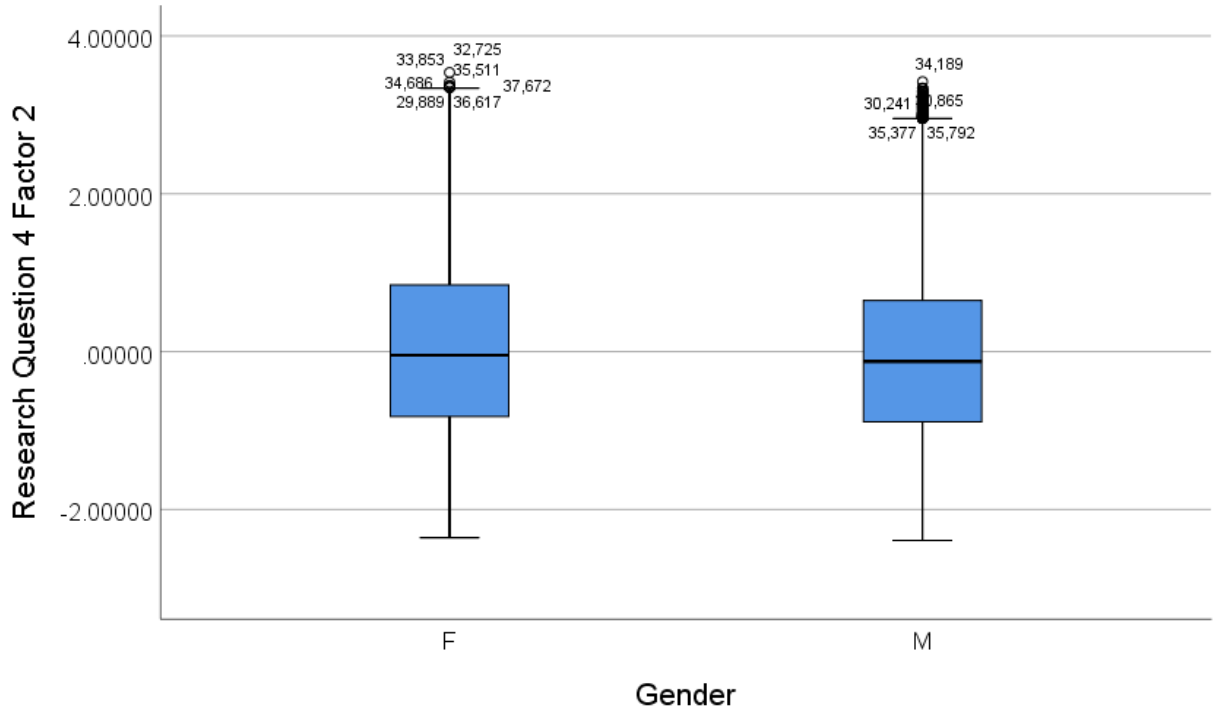


Figure 11. Mann-Whitney U Test for Gender, Factor 2

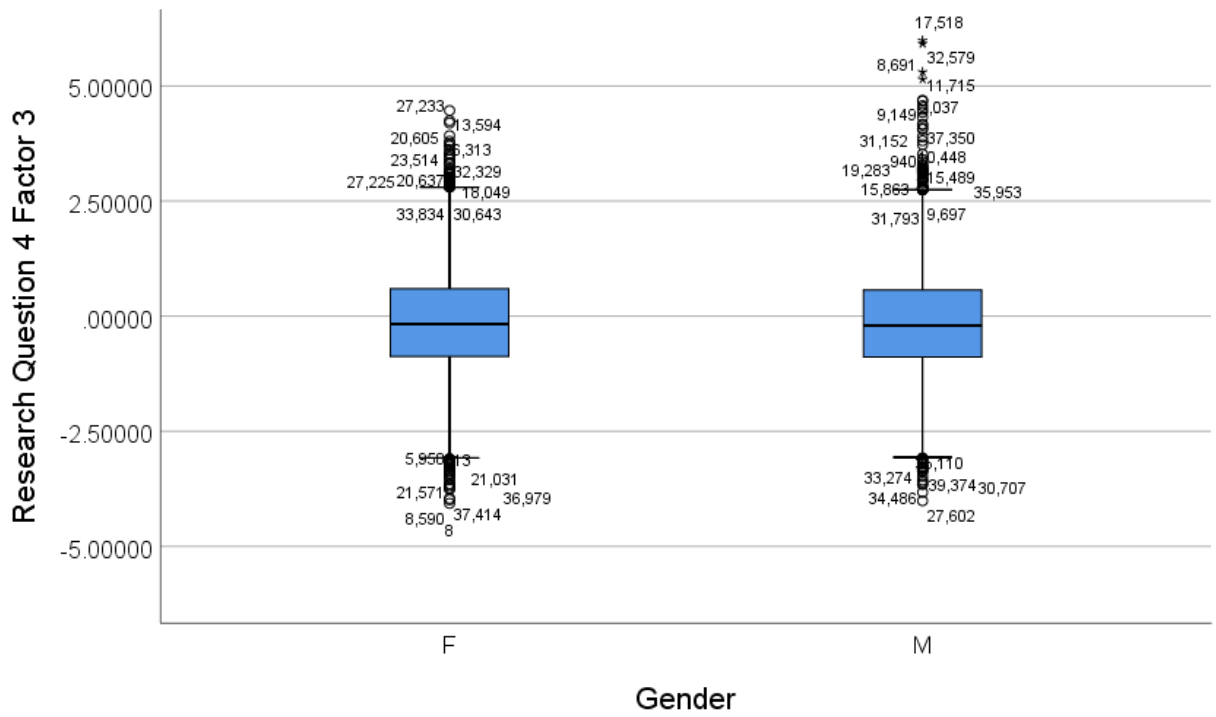


Figure 12. Mann-Whitney U Test for Gender, Factor 3

A Mann-Whitney U test was conducted to determine if the fourth factor score variable differed by gender. The results of the test were not significant,  $z=0.928$ ,  $p=0.353$ . Figure 13 shows the distributions of the scores on the fourth factor score for the two groups.

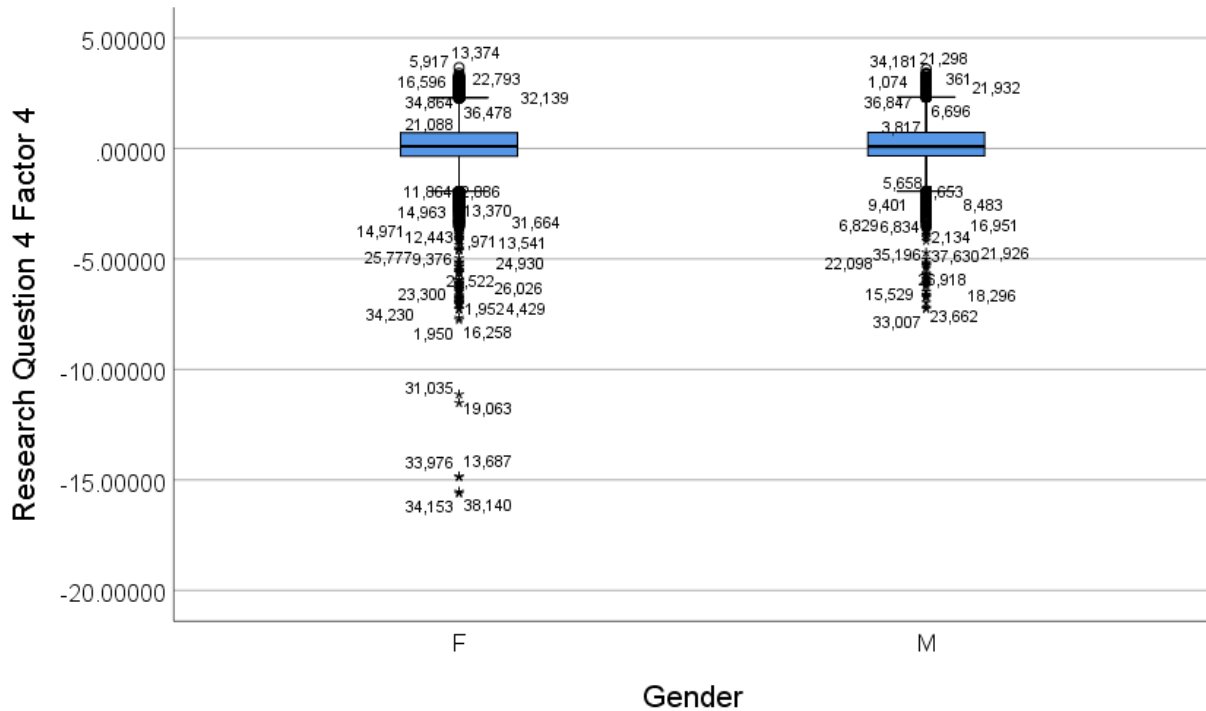


Figure 13. Mann-Whitney U Test for Gender, Factor 4

A Kruskal-Wallis test was conducted to determine if the first factor score variable differed by race/ethnicity. The test was significant,  $\chi^2(7, N=39,338) = 437.21$ ,  $p < 0.001$ . Figure 14 shows the distributions of the scores on the second factor score for each race/ethnicity.

A Kruskal-Wallis test was conducted to determine if the second factor score variable differed by race/ethnicity. The test was significant,  $\chi^2(7, N=39,338) = 828.23$ ,  $p < 0.001$ . Figure 15 shows the distributions of the scores on the first factor score for each race/ethnicity.

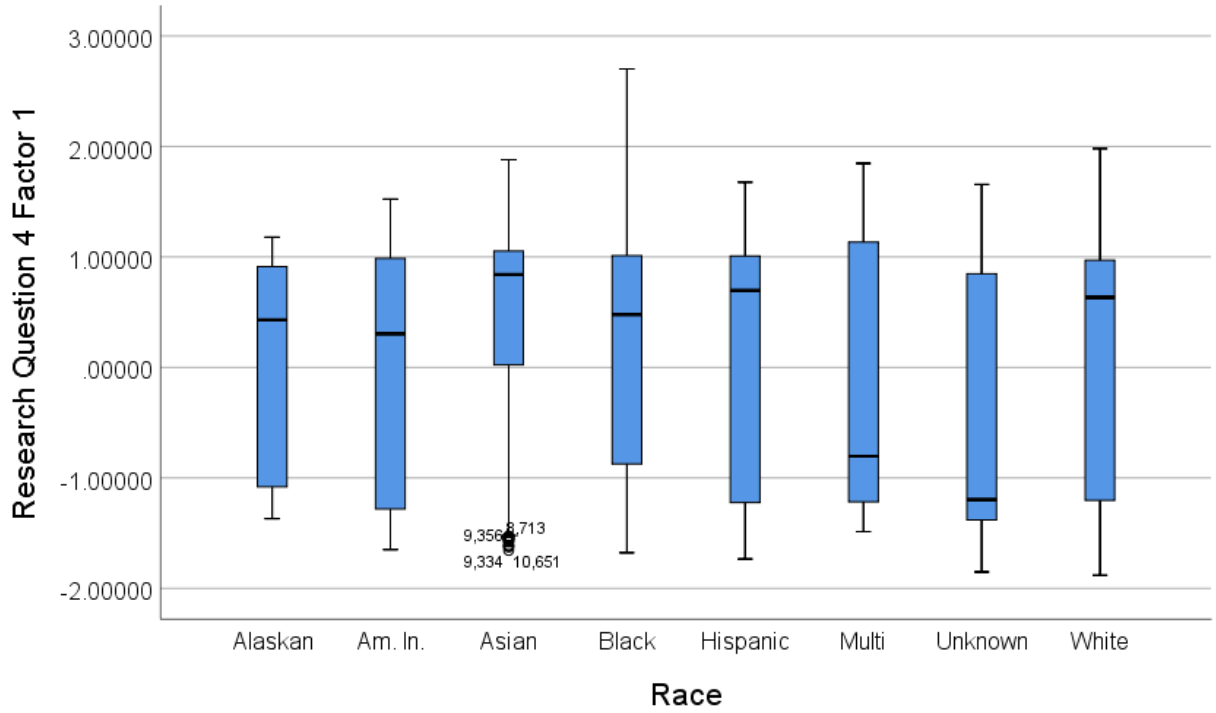


Figure 14. Kruskal-Wallis Test for Race, Factor 1

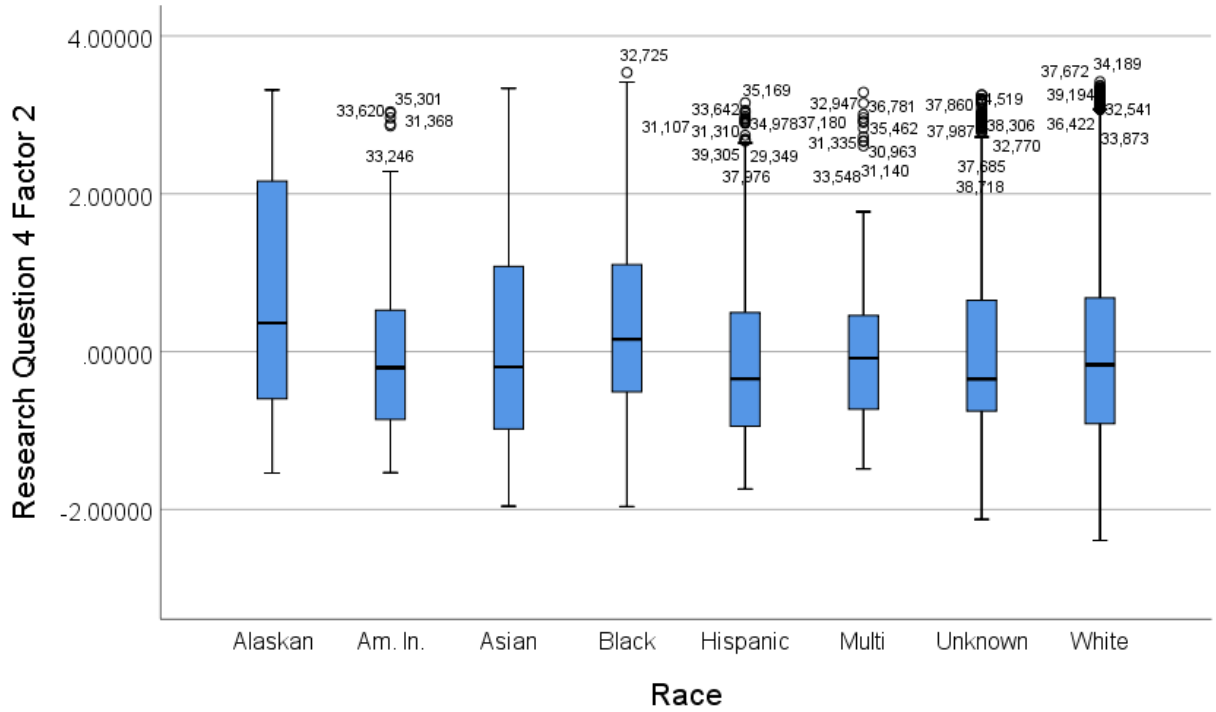


Figure 15. Kruskal-Wallis Test for Race, Factor 2

A Kruskal-Wallis test was conducted to determine if the third factor score variable differed by race/ethnicity. The test was significant,  $\chi^2(7, N=39,338) = 6,952.51, p < 0.001$ .

Figure 16 shows the distributions of the scores on the third factor score for each race/ethnicity.

A Kruskal-Wallis test was conducted to determine if the fourth factor score variable differed by race/ethnicity. The test was significant,  $\chi^2(7, N=39,338) = 3,507.31, p < 0.001$ .

Figure 17 shows the distributions of the scores on the fourth factor score for each race/ethnicity.

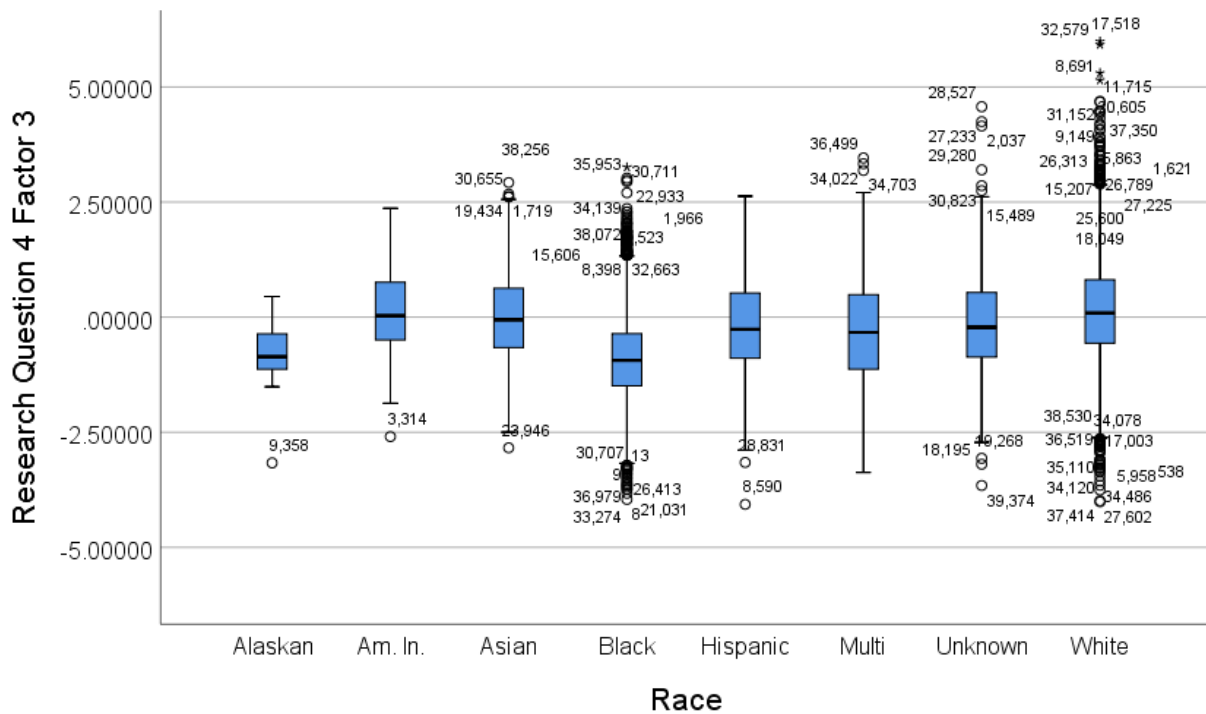


Figure 16. Kruskal-Wallis Test for Race, Factor 3





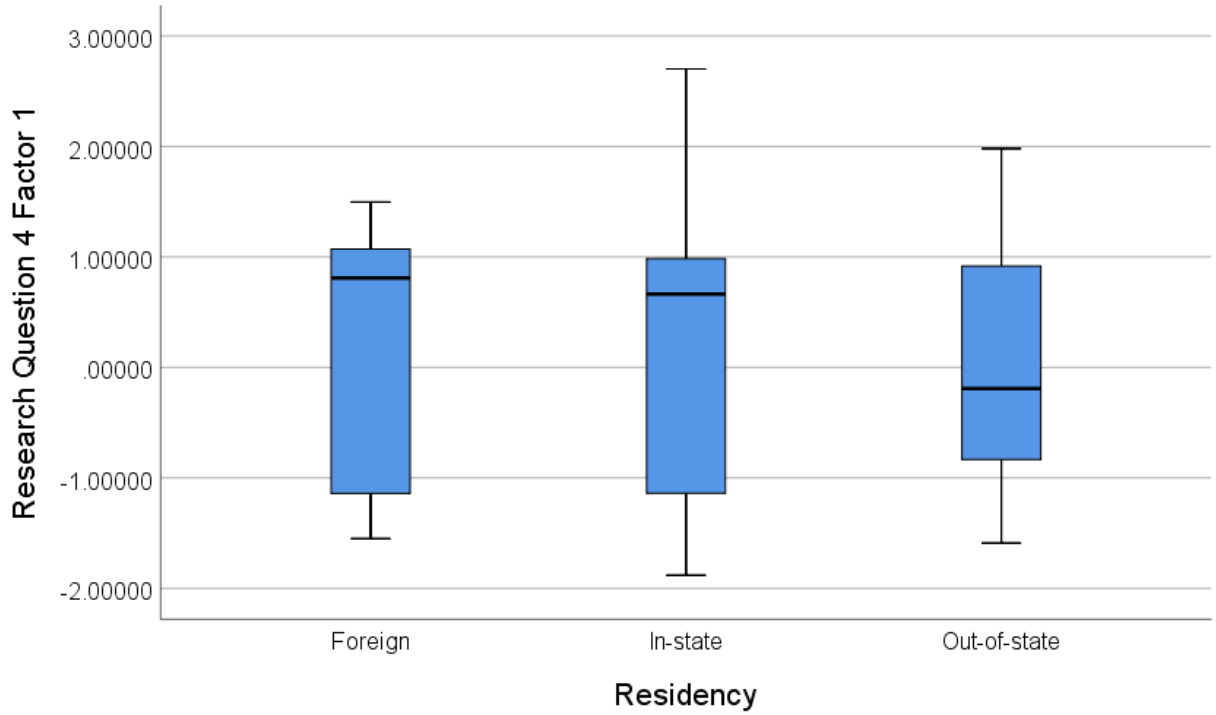


Figure 18. Kruskal-Wallis Test for Residency, Factor 1

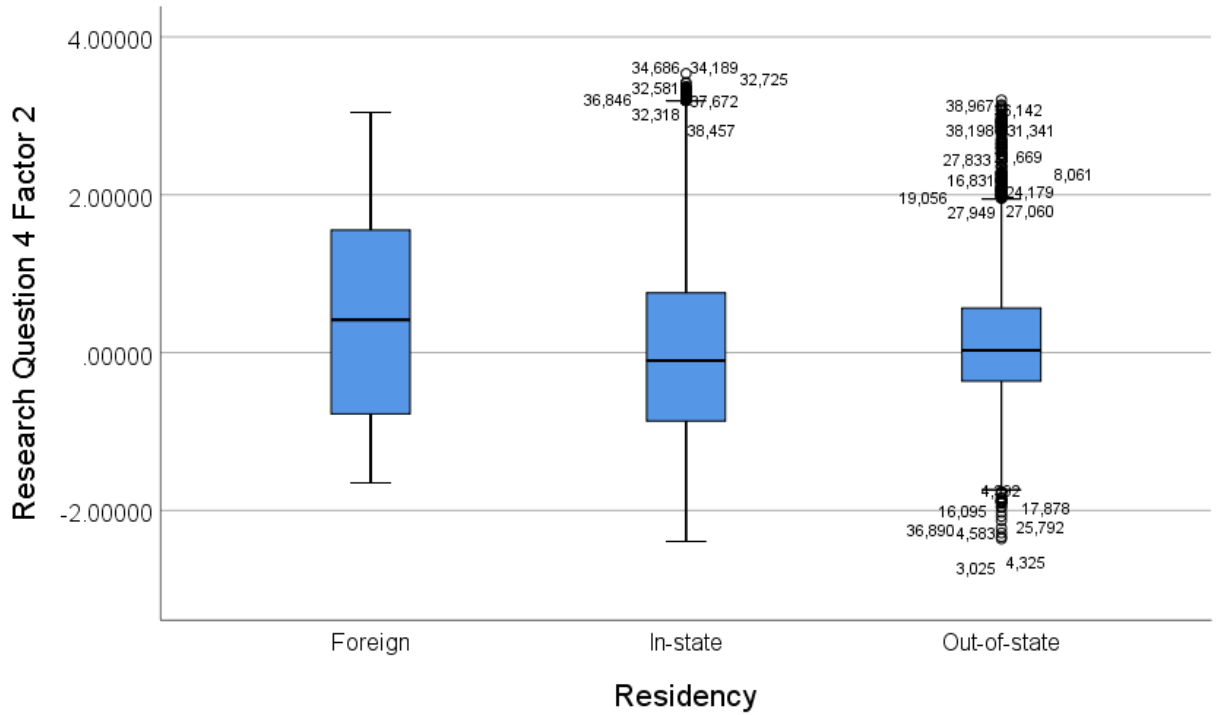


Figure 19. Kruskal-Wallis Test for Residency, Factor 2

A Kruskal-Wallis test was conducted to determine if the third factor score variable differed by race/ethnicity. The test was significant,  $\chi^2(7, N=39,338) = 44.01, p < 0.001$ . Figure 20 shows the distributions of the scores on the third factor score for each residency status.

A Kruskal-Wallis test was conducted to determine if the fourth factor score variable differed by race/ethnicity. The test was significant,  $\chi^2(7, N=39,338) = 4,786.65, p < 0.001$ . Figure 21 shows the distributions of the scores on the fourth factor score for each residency status.

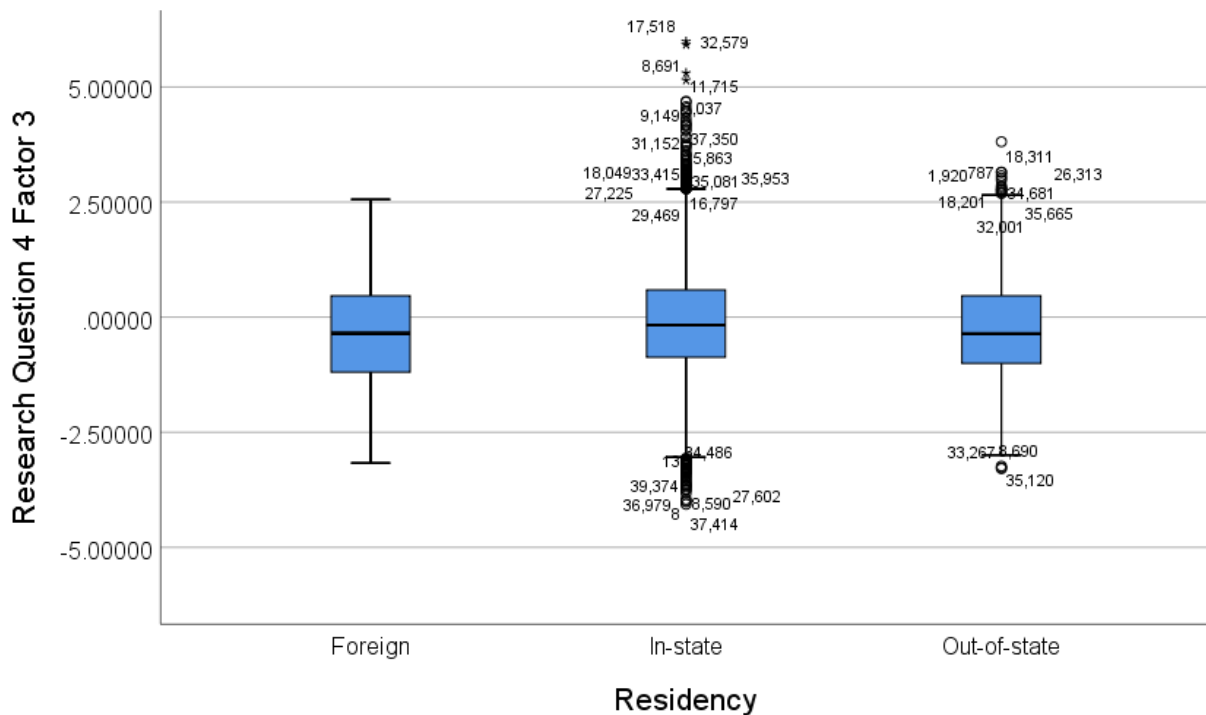


Figure 20. Kruskal-Wallis Test for Residency, Factor 3

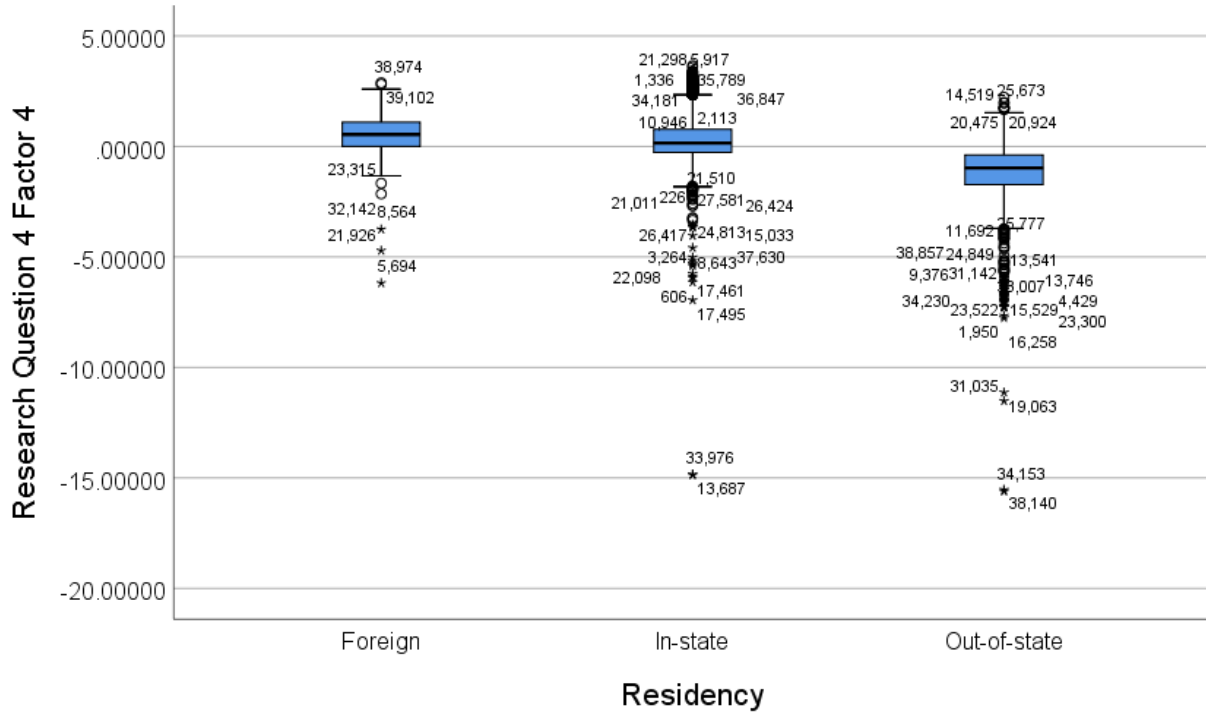


Figure 21. Kruskal-Wallis Test for Residency, Factor 4

A Kruskal-Wallis test was conducted to determine if the first factor score variable differed by institution. The test was significant,  $\chi^2(7, N=39,338) = 33,667.62, p < 0.001$ . Figure 22 shows the distributions of the scores on the first factor score for each institution.

A Kruskal-Wallis test was conducted to determine if the second factor score variable differed by institution. The test was significant,  $\chi^2(7, N=39,338) = 33,296.19, p < 0.001$ . Figure 23 shows the distributions of the scores on the second factor score for each institution.

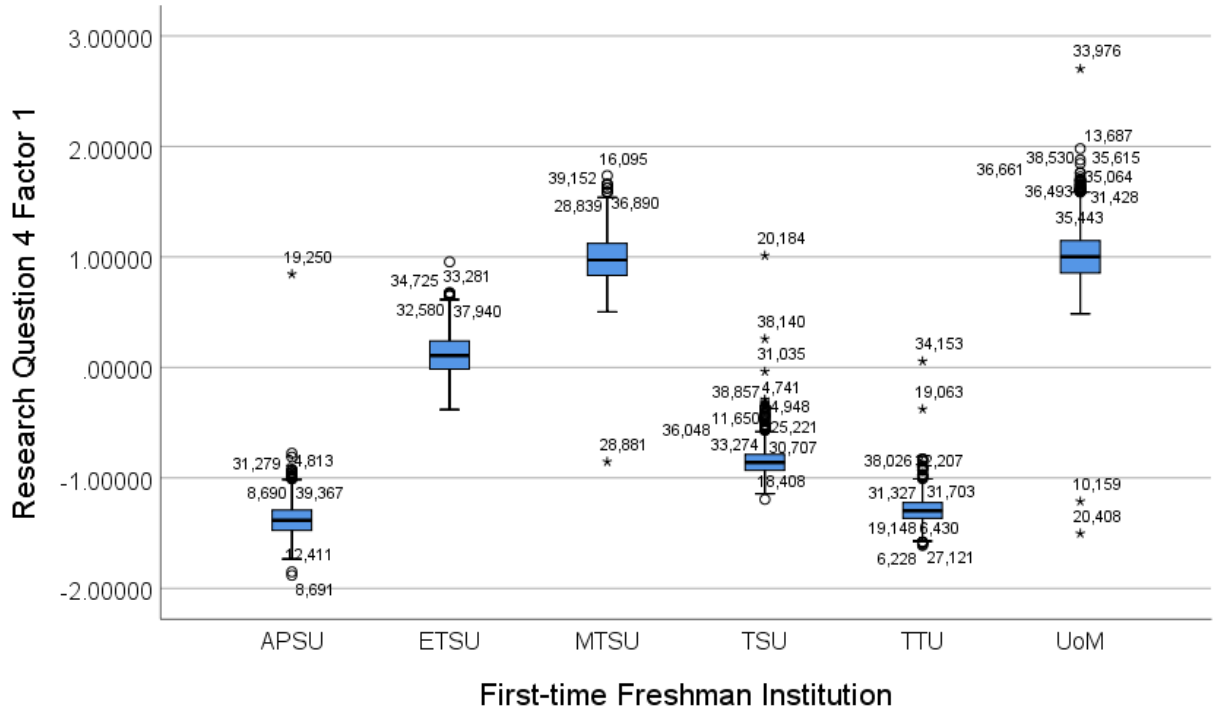


Figure 22. Kruskal-Wallis Test for First-time Freshman Institution, Factor 1

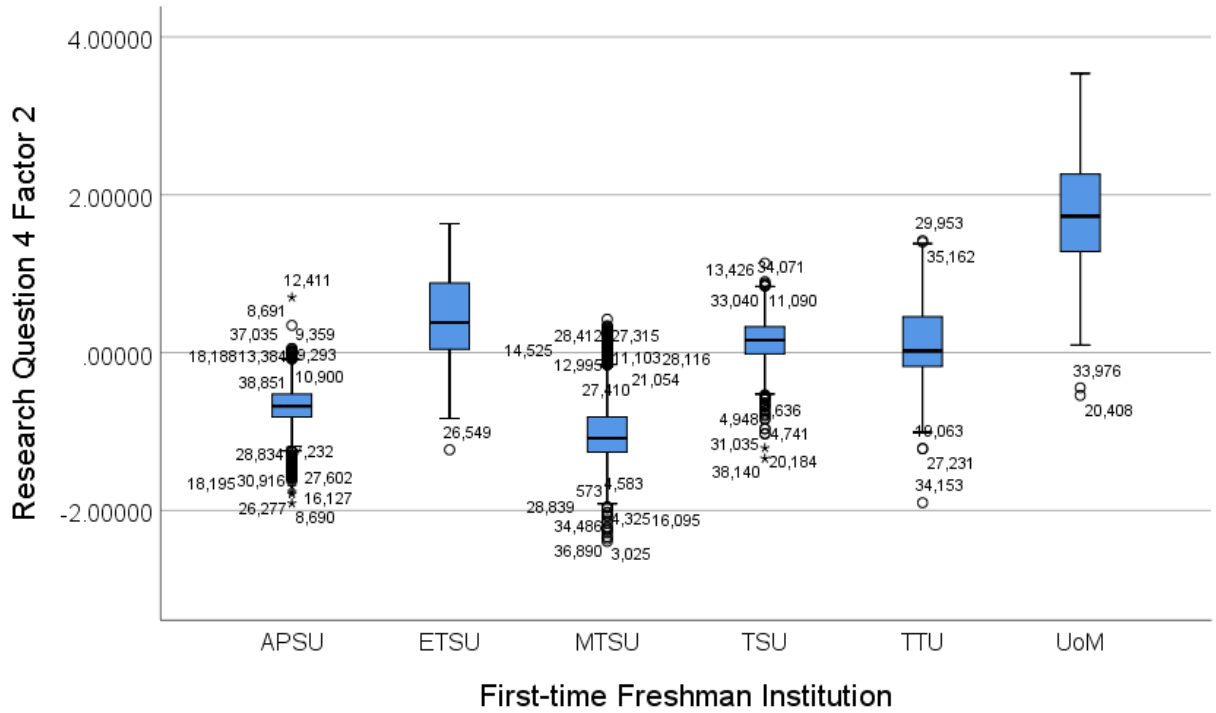


Figure 23. Kruskal-Wallis Test for First-time Freshman Institution, Factor 2

A Kruskal-Wallis test was conducted to determine if the third factor score variable differed by institution. The test was significant,  $\chi^2(7, N=39,338) = 4,458.60, p < 0.001$ . Figure 24 shows the distributions of the scores on the third factor score for each institution.

A Kruskal-Wallis test was conducted to determine if the fourth factor score variable differed by institution. The test was significant,  $\chi^2(7, N=39,338) = 4,691.53, p < 0.001$ . Figure 25 shows the distributions of the scores on the fourth factor score for each institution.

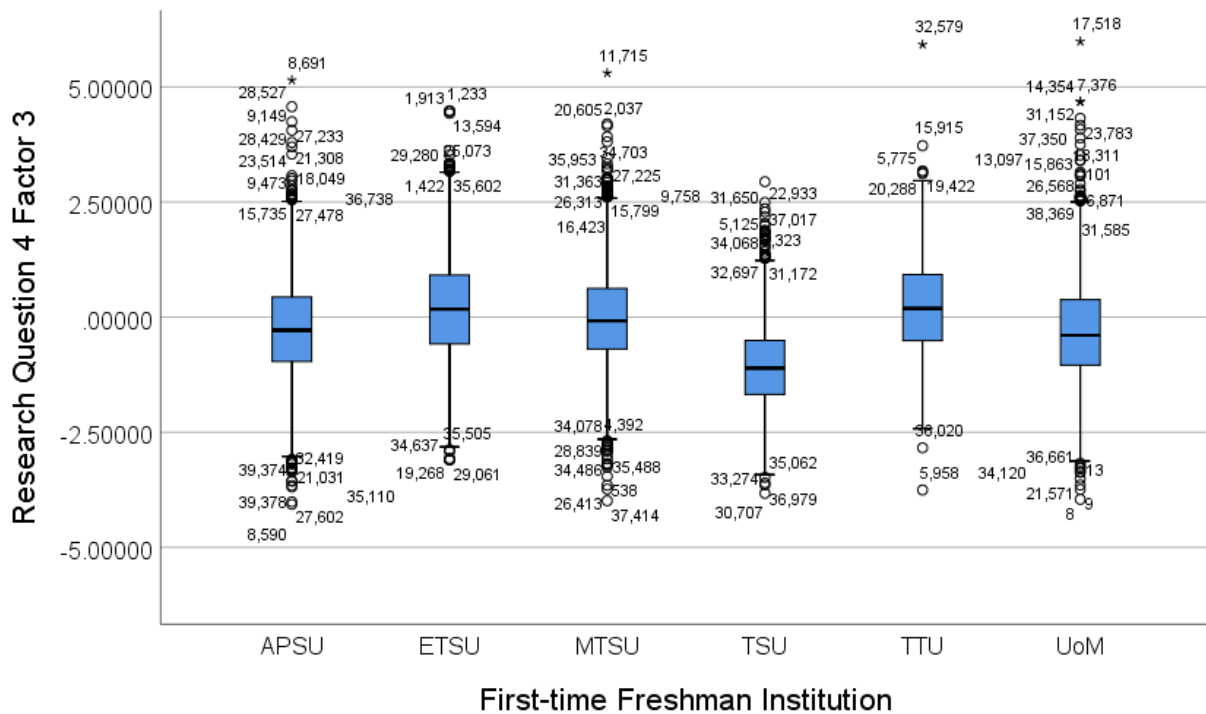


Figure 24. Kruskal-Wallis Test for First-time Freshman Institution, Factor 3

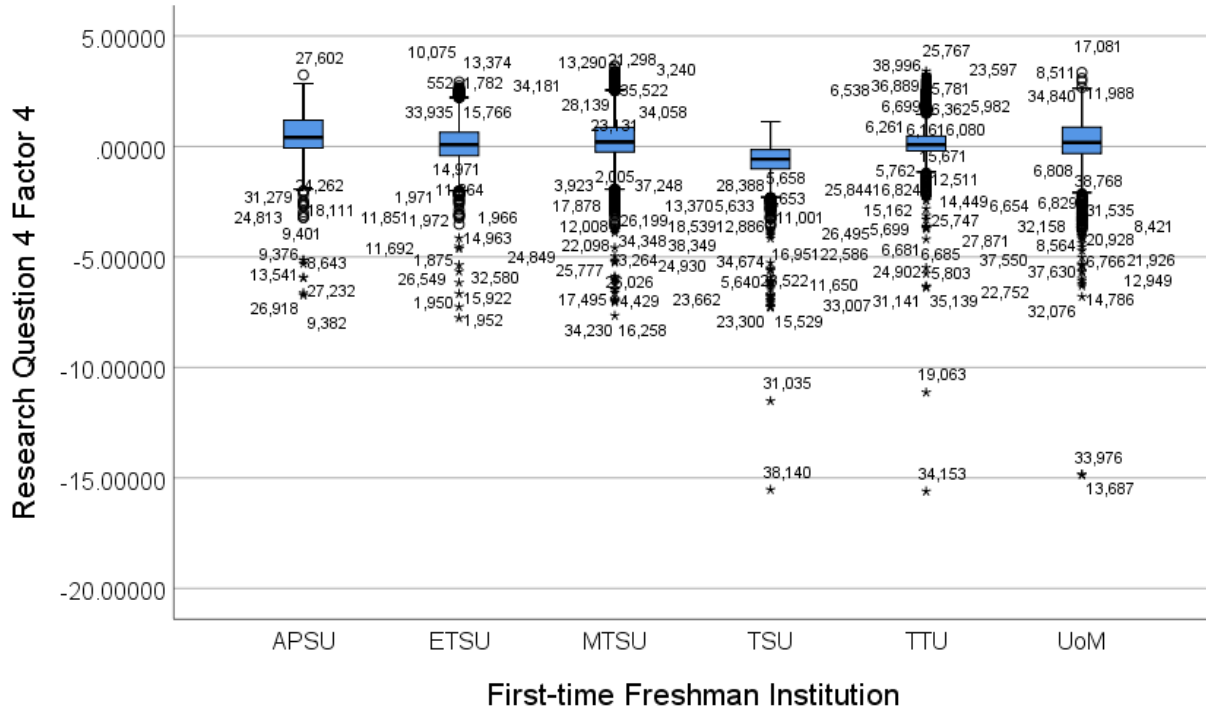


Figure 25. Kruskal-Wallis Test for First-time Freshman Institution, Factor 4

### Research Question #1

Which characteristics are most likely to predict graduation in 6 years after enrollment for first-time freshmen students under the age of 24 from a 4-year institution in Tennessee?

A principal component analysis was conducted to identify the characteristics most likely to predict graduation in 6 years after enrollment for first-time freshmen students under the age of 24 from a 4-year institution in Tennessee. An orthogonal, Varimax, rotation of initially 35 variables was conducted on 39,379 cases. Pairwise case deletion was used. Cases where columns with missing values would impact the results of the analysis were thus deleted. A total of 21,792 cases were analyzed. Thirty-three variables were used in the final analysis through an iterative process of eliminating component loadings of less than 0.5. The average credit hours taken per term and the first-time freshman year were the variables eliminated from the analysis

due to component loadings of less than 0.5. An examination of the Kaiser-Meyer Olkin measure of sampling adequacy suggested that the final sample was factorable ( $KMO=0.850$ ). The results of the orthogonal rotation are shown in Table 3. When loadings less than 0.5 were excluded, the factor analysis yielded nine factors accounting for 80.34% of variance in the data.

The variables of the first factor consisted of initial institution variables as well as the number of full-time equivalent students at the institution of graduation and the cost of student support expenditures at the institution of graduation. This factor accounted for approximately 27% of the variance in the sample. The second factor consisted of graduation institution characteristics. This factor accounted for approximately 22% of the variance in the sample. The third factor consisted of characteristics relating to the number of semesters spent in school prior to graduation. The fourth factor consisted of expenditures for research, academic support, and the initial institution. The fifth factor consisted of high school GPA, composite ACT, and the number of credit hours taken in the first semester. The sixth factor consisted of the distance from students' permanent address to their initial institution, as well as the number of students from the same high school attending the same institution. The seventh factor consisted of the number of students in the same major and the number of major changes made. The eighth factor consisted of the distance from students' permanent address to their transfer institution. The ninth factor consisted of age and the amount of state funds awarded.

Table 3

*Factors for Graduation*

Component	1	2	3	4	5	6	7	8	9
Age									0.77
Distance from home to initial institution						-0.76			
Distance from home to transfer institution								0.93	
High school GPA					0.78				
Composite ACT					0.76				
Number of credit hours taken in first semester					0.47				
Amount of state funds awarded									-0.64
Hours spent at initial institution			0.91						
Number of semesters at initial institution			0.90						
Number of semesters at transfer institutions			-0.77						
Number of FTF students from same high school						0.78			
Number of FTF students with same major							0.62		
Number of changes in major							0.84		
Number of full-time faculty at initial institution	0.96								
Number of full-time staff at initial institution	0.90								
Tuition and fees at initial institution	0.96								
State appropriations for initial institution	0.90								
Instruction expenditures at initial institution	0.91								
Research expenditures at initial institution	0.48			0.83					
Academic support expenditures at initial institution	0.84			0.44					
Student support expenditures at initial institution	0.89								
Institutional support expenditures at initial institution	0.90								
Student FTE at initial institution	0.85			-0.41					
Number of full-time faculty at graduation institution		0.90							
Number of full-time staff at graduation institution		0.93							
Tuition and fees at graduation institution		0.83							
State appropriations for graduation institution		0.87							
Instruction expenditures at graduation institution		0.93							
Research expenditures at graduation institution		0.92							
Academic support expenditures at graduation institution		0.94							
Student support expenditures at graduation institution	0.69	0.43							
Institutional support expenditures at graduation institution		0.91							
Student FTE at graduation institution	0.56	0.62							
Eigenvalues	8.88	7.41	2.30	1.60	1.57	1.36	1.24	1.12	1.03
Percentage of total variance	26.90	22.47	6.98	4.84	4.77	4.12	3.75	3.39	3.13
Number of test measures	12	10	3	3	3	2	2	1	2

\*Loadings  $\geq 0.5$ 

A second principal component analysis was conducted which excluded variables based on knowledge available after the first semester that the student attended. An orthogonal,



Varimax, rotation was used again on initially 19 variables and 39,379 cases. Pairwise case deletion was used. Cases where columns with missing values would impact the results of the analysis were thus deleted. A total of 21,792 cases were analyzed. Seventeen variables were used in the final analysis through an iterative process of eliminating component loadings of less than 0.5. Student age and the first-time freshman year were the variables eliminated from the analysis due to component loadings of less than 0.5. An examination of the Kaiser-Meyer Olkin measure of sampling adequacy suggested that the final sample was factorable ( $KMO=0.801$ ). The results of the orthogonal rotation are shown in Table 4. When loadings less than 0.5 were excluded, the factor analysis yielded four factors accounting for 72.26% of variance in the data. The first factor consisted of variables pertaining to the institution except for research expenditures. The first factor accounted for 43.12% of the variance in the data set. The second factor consisted of the number of students with the same major as well as institutional variables including research, academic support and institutional support expenditures. The second factor accounted for 12.15% of the variance. The third factor consisted of the student's high school GPA, composite ACT score, and the number of credit hours taken in the first semester. The third factor accounted for 9.02% of the variance. The fourth factor consisted of the distance of the institution from the student's home and the number of students from the same high school attending the same institution.

Table 4

*Factors for Graduation from First Semester Information*

Component	1	2	3	4
Distance from home to initial institution				0.77
High school GPA			0.77	
Composite ACT			0.81	
Number of credit hours taken in first semester			0.43	
Amount of state funds awarded				
Number of FTF students from same high school				-0.76
Number of FTF students with same major		0.56		
Number of full-time faculty at initial institution	0.98			
Number of full-time staff at initial institution	0.89			
Tuition and fees at initial institution	0.96			
State appropriations for initial institution	0.88			
Instruction expenditures at initial institution	0.93			
Research expenditures at initial institution		0.86		
Academic support expenditures at initial institution	0.77	0.59		
Student support expenditures at initial institution	0.85			
Institutional support expenditures at initial institution	0.84	0.48		
Student FTE at initial institution	0.92			
Eigenvalues	7.33	2.06	1.53	1.35
Percentage of total variance	43.12	12.15	9.02	7.97
Number of test measures	9	4	3	2

\*Loadings  $\geq 0.5$ Research Question #2

What characteristics identify first-time freshmen students under the age of 24 who transfer to another higher education institution?

A principal component analysis was conducted to identify the characteristics of first-time freshmen students under the age of 24 who transfer to another higher education institution. An orthogonal, Varimax, rotation of initially thirty-five variables was conducted on 39,379 cases. Pairwise case deletion was used. Cases where columns with missing values would impact the results of the analysis were thus deleted. A total of 10,625 cases were analyzed. Thirty variables

were used in the final analysis through an iterative process of eliminating component loadings of less than 0.5. The first-time freshman year, age, hours spent at the initial institution, and the number of changes in major were the variables eliminated from the analysis due to a component loading of less than 0.5. The amount of state funds awarded had a component loading greater than 0.5; however, it was excluded from the set of factors. An examination of the Kaiser-Meyer Olkin measure of sampling adequacy suggested that the final sample was factorable (KMO=0.845). The results of the orthogonal rotation are shown in Table 5. When loadings less than 0.5 were excluded, the factor analysis yielded seven factors accounting for 75.33 % of variance in the data.

The first factor consisted of graduation institution variables while the second factor consisted of initial institution variables. The first factor accounted for 27.78% of the variance in the sample, while the second factor accounted for 25.34% of the variance. The third factor consisted of characteristics relating to the number of first-time freshmen students with the same major and expenditures for research and academic support. The fourth factor consisted of the distance from the student's home, the number of credit hours taken in the first term, the number of first-time freshmen students from the same high school, and the average number of credits taken per semester. The fifth factor consisted of high school GPA and composite ACT. The sixth factor consisted of variables related to the number of semesters at the initial and transfer institutions. The seventh factor consisted of the distance between the student's permanent address and the transfer institution.

Table 5

*Factors for Transfers*

Component	1	2	3	4	5	6	7
Distance from home to initial institution				0.71			
Distance from home to transfer institution							0.88
High school GPA					0.72		
Composite ACT					0.73		
Number of credit hours taken in first semester				0.43			
Amount of state funds awarded							
Number of semesters at initial institution						0.76	
Number of semesters at transfer institutions						-0.72	
Number of FTF students from same high school				-0.68			
Number of FTF students with same major			0.62				
Average credit hours taken per term				0.46			
Number of full-time faculty at initial institution	0.99						
Number of full-time staff at initial institution	0.93						
Tuition and fees at initial institution	0.97						
State appropriations for initial institution	0.92						
Instruction expenditures at initial institution	0.94						
Research expenditures at initial institution	0.47	0.75					
Academic support expenditures at initial institution	0.85	0.45					
Student support expenditures at initial institution	0.88						
Institutional support expenditures at initial institution	0.89						
Student FTE at initial institution	0.87						
Number of full-time faculty at graduation institution	0.97						
Number of full-time staff at graduation institution	0.93						
Tuition and fees at graduation institution	0.90						
State appropriations for graduation institution	0.93						
Instruction expenditures at graduation institution	0.97						
Research expenditures at graduation institution	0.92						
Academic support expenditures at graduation institution	0.96						
Student support expenditures at graduation institution	0.79						
Institutional support expenditures at graduation institution	0.95						
Student FTE at graduation institution	0.88						
Eigenvalues	8.61	7.86	1.54	1.51	1.48	1.26	1.09
Percentage of total variance	27.78	25.34	4.97	4.87	4.78	4.06	3.52
Number of test measures	10	10	3	4	2	2	1

\*Loadings  $\geq 0.5$

A second principal component analysis was conducted which excluded variables based on knowledge available after the first semester that the student attended. An orthogonal, Varimax, rotation was used on initially 19 variables and 39,379 cases. Pairwise case deletion was used. Cases where columns with missing values would impact the results of the analysis were thus deleted. A total of 10,625 cases were analyzed. Seventeen variables were used in the final analysis through an iterative process of eliminating component loadings of less than 0.5. Student age and the first-time freshman year were the variables eliminated from the analysis due to component loadings of less than 0.5. An examination of the Kaiser-Meyer Olkin measure of sampling adequacy suggested that the final sample was factorable ( $KMO=0.803$ ). The results of the orthogonal rotation are shown in Table 6. When loadings less than 0.5 were excluded, the factor analysis yielded four factors accounting for 71.56% of variance in the data.

The first factor consisted of variables pertaining to the institution. The first factor accounted for 44.52% of the variance in the data set. The second factor consisted of the number of students with the same major as well as institutional variables including research, academic support, and institutional support expenditures. The second factor accounted for 10.77% of the variance. The third factor consisted of the distance of the institution from the student's home and the number of students from the same high school attending the same institution. The fourth factor consisted of the student's high school GPA, composite ACT score, and the number of credit hours taken in the first semester.

Table 6

*Factors for Transfers from First Semester Information*

Component	1	2	3	4
Distance from home to initial institution			-0.77	
High school GPA				0.74
Composite ACT				0.74
Number of credit hours taken in first semester				0.41
Amount of state funds awarded				
Number of FTF students from same high school			0.76	
Number of FTF students with same major		0.59		
Number of full-time faculty at initial institution	0.98			
Number of full-time staff at initial institution	0.91			
Tuition and fees at initial institution	0.96			
State appropriations for initial institution	0.90			
Instruction expenditures at initial institution	0.93			
Research expenditures at initial institution	0.40	0.80		
Academic support expenditures at initial institution	0.81	0.52		
Student support expenditures at initial institution	0.86			
Institutional support expenditures at initial institution	0.85	0.42		
Student FTE at initial institution	0.90			
Eigenvalues	7.57	1.83	1.39	1.38
Percentage of total variance	44.52	10.77	8.17	8.12
Number of test measures	10	4	2	3

\*Loadings  $\geq 0.5$ Research Question #3

What characteristics identify first-time freshmen students under the age of 24 who did not graduate from a 4-year institution in Tennessee?

A principal component analysis was conducted to identify the characteristics of first-time freshmen students under the age of 24 who do not graduate from a 4-year institution in Tennessee. An orthogonal, Varimax, rotation of initially thirty-five variables was conducted on 39,379 cases. Pairwise case deletion was used. Cases where columns with missing values would

impact the results of the analysis were thus deleted. A total of 17,588 cases were analyzed. Thirty-four variables were used in the final analysis through an iterative process of eliminating component loadings of less than 0.5. The first-time freshman year was the variable eliminated from the analysis due to a component loading of less than 0.5. Amount of state funds awarded had a component loading of greater than 0.5 but was not included in any factor. An examination of the Kaiser-Meyer Olkin measure of sampling adequacy suggested that the final sample was factorable (KMO=0.837). The results of the orthogonal rotation are shown in Table 7. When loadings less than 0.5 were excluded, the factor analysis yielded eight factors accounting for 74.06% of variance in the data.

The first factor consisted of graduation institution variables while the second factor consisted of initial institution variables. The first factor accounted for 24.24% of the variance in the sample, while the second factor accounted for 23.31% of the variance. The third factor consisted of variables related to the length of time the student spent at the initial or transfer institution and the number of changes in major. The fourth factor consisted of variables related to the number of students with the same major and expenditures for research and academic support at the student's initial institution. The fifth factor consisted of distance between the student's permanent address and the transfer institution as well as the number of major changes and average number of credit hours taken each semester. The sixth factor consisted of distance between students' permanent address and the initial institution as well as the number of students from the same high school. The seventh factor consisted of high school GPA and composite ACT score. The eighth factor consisted of the number of credit hours taken in the first semester and the student's age.

Table 7

*Factors for Nongraduates*

Component	1	2	3	4	5	6	7	8
Age								-0.74
Distance from home to initial institution						0.71		
Distance from home to transfer institution					0.51			
High school GPA							0.73	
Composite ACT							0.64	
Number of credit hours taken in first semester								0.59
Amount of state funds awarded								
Hours spent at initial institution			0.92					
Number of semesters at initial institution			0.93					
Number of semesters at transfer institutions					0.84			
Number of FTF students from same high school							-0.78	
Number of FTF students with same major				0.53				
Number of changes in major			0.50		0.56			
Average credit hours taken per term								
Number of full-time faculty at initial institution		0.99						
Number of full-time staff at initial institution		0.93						
Tuition and fees at initial institution		0.96						
State appropriations for initial institution		0.92						
Instruction expenditures at initial institution		0.94						
Research expenditures at initial institution		0.50		0.75				
Academic support expenditures at initial institution		0.86		0.43				
Student support expenditures at initial institution		0.88						
Institutional support expenditures at initial institution		0.89						
Student FTE at initial institution		0.87						
Number of full-time faculty at graduation institution	0.95							
Number of full-time staff at graduation institution	0.88							
Tuition and fees at graduation institution	0.90							
State appropriations for graduation institution	0.85							
Instruction expenditures at graduation institution	0.95							
Research expenditures at graduation institution	0.87							
Academic support expenditures at graduation institution	0.91							
Student support expenditures at graduation institution	0.82							
Institutional support expenditures at graduation institution	0.89							
Student FTE at graduation institution	0.88							
Eigenvalues	8.24	7.92	2.04	1.54	1.44	1.44	1.37	1.19
Percentage of total variance	24.24	23.31	6.01	4.53	4.23	4.22	4.04	3.49
Number of test measures	10	10	3	3	3	2	2	2

\*Loadings  $\geq 0.5$



A second principal component analysis was conducted which excluded variables based on knowledge available after the first semester that the student attended. An orthogonal, Varimax, rotation was used on initially 19 variables and 39,379 cases. Pairwise case deletion was used. Cases where columns with missing values would impact the results of the analysis were thus deleted. A total of 17,588 cases were analyzed. Seventeen variables were used in the final analysis through an iterative process of eliminating component loadings of less than 0.5. Student age and the first-time freshman year were the variables eliminated from the analysis due to component loadings of less than 0.5. An examination of the Kaiser-Meyer Olkin measure of sampling adequacy suggested that the final sample was factorable ( $KMO=0.806$ ). The results of the orthogonal rotation are shown in Table 8. When loadings less than 0.5 were excluded, the factor analysis yielded four factors accounting for 71.51% of variance in the data.

The first factor consisted of variables pertaining to the institution. The first factor accounted for 45.95% of the variance in the data set. The second factor consisted of the number of students with the same major as well as institutional variables including research and academic support expenditures. The second factor accounted for 9.77% of the variance. The third factor consisted of the distance of the institution from the student's home and the number of students from the same high school attending the same institution. The fourth factor consisted of the student's high school GPA, composite ACT score, and the number of credit hours taken in the first semester.

Table 8

*Factors for Nongraduates from First Semester Information*

Component	1	2	3	4
Distance from home to initial institution			-0.78	
High school GPA				0.74
Composite ACT				0.66
Number of credit hours taken in first semester				0.52
Amount of state funds awarded				
Number of FTF students from same high school			0.78	
Number of FTF students with same major		0.52		
Number of full-time faculty at initial institution	0.98			
Number of full-time staff at initial institution	0.93			
Tuition and fees at initial institution	0.96			
State appropriations for initial institution	0.92			
Instruction expenditures at initial institution	0.94			
Research expenditures at initial institution	0.47	0.76		
Academic support expenditures at initial institution	0.84	0.46		
Student support expenditures at initial institution	0.87			
Institutional support expenditures at initial institution	0.87			
Student FTE at initial institution	0.88			
Eigenvalues	7.81	1.66	1.40	1.28
Percentage of total variance	45.95	9.77	8.25	7.54
Number of test measures	10	3	2	3

\*Loadings  $\geq 0.5$ Research Question #4

What characteristics identify first-time freshmen students under the age of 24 who began at a 4-year institution in Tennessee, transferred to another higher education institution, and graduated?

A principal component analysis was conducted to identify the characteristics most likely to predict graduation in 6 years after enrollment for first-time freshmen student under the age of 24 from a 4-year institution in Tennessee. An orthogonal, Varimax, rotation of initially thirty-

five variables was conducted on 39,379 cases. Pairwise case deletion was used. Cases where columns with missing values would impact the results of the analysis were thus deleted. A total of 5,701 cases were analyzed. Thirty-three variables were used in the final analysis through an iterative process of eliminating component loadings of less than 0.5. The first-time freshman year was the variable eliminated from the analysis due to a component loading of less than 0.5. The amount of state funds awarded had a component loading of greater than 0.5, but the variable was not included in any factor. An examination of the Kaiser-Meyer Olkin measure of sampling adequacy suggested that the final sample was factorable ( $KMO=0.828$ ). When loadings less than 0.5 were excluded, the factor analysis yielded eight factors accounting for 74.37% of variance in the data. The results of the orthogonal rotation are shown in Table 9.

The first factor consisted of graduation institution variables, while the second factor consisted of initial institution variables. The first factor accounted for 24.05% of the variance in the sample, while the second factor accounted for 23.09% of the variance. The third factor consisted of variables related to the length of time the student spent at the initial or transfer institution. The fourth factor consisted of high school GPA, composite ACT score, and number of credit hours taken in the first semester. The fifth factor consisted of variables related to the number of students with the same major and expenditures for research and academic support at the student's initial institution. The sixth factor consisted of distance between students' permanent address and the transfer institution as well as the number of major changes and average number of credit hours taken each semester. The seventh factor consisted of distance between the student's permanent address and their initial institution as well as the number of students from the same high school. The eighth factor consisted of the student age and the amount of state funding the student received.

Table 9

*Factors for Transfer Graduates*

Component	1	2	3	4	5	6	7	8
Age								-0.69
Distance from home to initial institution						-0.73		
Distance from home to transfer institution							0.74	
High school GPA				0.75				
Composite ACT				0.75				
Number of credit hours taken in first semester				0.40				
Amount of state funds awarded								0.60
Hours spent at initial institution			0.91					
Number of semesters at initial institution			0.94					
Number of semesters at transfer institutions			-0.69					
Number of FTF students from same high school						0.78		
Number of FTF students with same major					0.58			
Number of changes in major							-0.65	
Average credit hours taken per term								
Number of full-time faculty at initial institution		0.99						
Number of full-time staff at initial institution		0.92						
Tuition and fees at initial institution		0.96						
State appropriations for initial institution		0.92						
Instruction expenditures at initial institution		0.93						
Research expenditures at initial institution		0.44			0.73			
Academic support expenditures at initial institution		0.85			0.45			
Student support expenditures at initial institution		0.90						
Institutional support expenditures at initial institution		0.91						
Student FTE at initial institution		0.88						
Number of full-time faculty at graduation institution	0.96							
Number of full-time staff at graduation institution	0.91							
Tuition and fees at graduation institution	0.87							
State appropriations for graduation institution	0.87							
Instruction expenditures at graduation institution	0.96							
Research expenditures at graduation institution	0.90							
Academic support expenditures at graduation institution	0.93							
Student support expenditures at graduation institution	0.79							
Institutional support expenditures at graduation institutio	0.91							
Student FTE at graduation institution	0.87							
Eigenvalues	8.18	7.85	2.30	1.63	1.46	1.41	1.36	1.09
Percentage of total variance	24.05	23.09	6.77	4.80	4.31	4.14	4.00	3.21
Number of test measures	10	10	3	3	3	2	2	2

\*Loadings  $\geq 0.5$

A second principal component analysis was conducted that excluded variables based on knowledge available after the first semester that the student attended. An orthogonal, Varimax, rotation was used on initially 19 variables and 39,379 cases. Pairwise case deletion was used. Cases where columns with missing values would impact the results of the analysis were thus deleted. A total of 5,701 cases were analyzed. Sixteen variables were used in the final analysis when components with a loading of less than 0.5 were excluded. Student age and the first-time freshman year were the variables eliminated from the analysis due to component loadings of less than 0.5. The amount of state funds awarded had a loading greater than 0.5, but the factor analysis did not include the data point in any of the four factors. An examination of the Kaiser-Meyer Olkin measure of sampling adequacy suggested that the final sample was factorable (KMO=0.793). The results of the orthogonal rotation are shown in Table 10. When loadings less than 0.5 were excluded, the factor analysis yielded four factors accounting for 71.46% of variance in the data.

The first factor consisted of variables pertaining to the institution except for research expenditures. The first factor accounted for 44% of the variance in the data set. The second factor consisted of the number of students with the same major as well as institutional variables including research, academic support, and institutional support expenditures. The second factor accounted for 10.68% of the variance. The third factor consisted of the student's high school GPA, composite ACT score, and the number of credit hours taken in the first semester. The fourth factor consisted of the distance of the institution from the student's home and the number of students from the same high school attending the same institution.

Table 10

*Factors for Transfer Graduates from First Semester Information*

Component	1	2	3	4
Distance from home to initial institution				-0.79
High school GPA			0.77	
Composite ACT			0.79	
Number of credit hours taken in first semester			0.45	
Amount of state funds awarded				
Number of FTF students from same high school				0.76
Number of FTF students with same major		0.56		
Number of full-time faculty at initial institution	0.98			
Number of full-time staff at initial institution	0.88			
Tuition and fees at initial institution	0.97			
State appropriations for initial institution	0.89			
Instruction expenditures at initial institution	0.92			
Research expenditures at initial institution		0.81		
Academic support expenditures at initial institution	0.79	0.55		
Student support expenditures at initial institution	0.89			
Institutional support expenditures at initial institution	0.87	0.40		
Student FTE at initial institution	0.91			
Eigenvalues	7.48	1.82	1.51	1.34
Percentage of total variance	44.00	10.68	8.90	7.87
Number of test measures	9	4	3	2

\*Loadings  $\geq 0.5$ Research Question #5

Is the predictive power of logistic regression for determining which students will leave their home institution and graduate elsewhere greater than 50%?

H<sub>0</sub>5: The predictive power of logistic regression for determining which students will leave their home institution and graduate elsewhere is less than or equal to 50%.

The logistic regression model was run against four factors from Research Question #4. Factor scores for each case were generated during the factor analysis. In addition to the four

factors, dummy variables for gender, race, residency, and first-time freshman institution were included in the model. The model was evaluated using 10-fold cross validation, resulting in an average of 3,936 cases per iteration. The resulting model had an accuracy score of 0.58. The confusion matrix is shown in Table 11, and the classification report is shown in Table 12. The ROC AUC score was 0.59. The plot of the ROC curve is shown in Figure 26. Each of the 10 cross validation iterations is shown in the plot of the ROC curve as well as a dashed line representing 50%, labeled “luck” in the graph. The null hypothesis was rejected. The predictive power of logistic regression for determining which students will leave their home institution and graduate elsewhere was greater than 50%.

Table 11

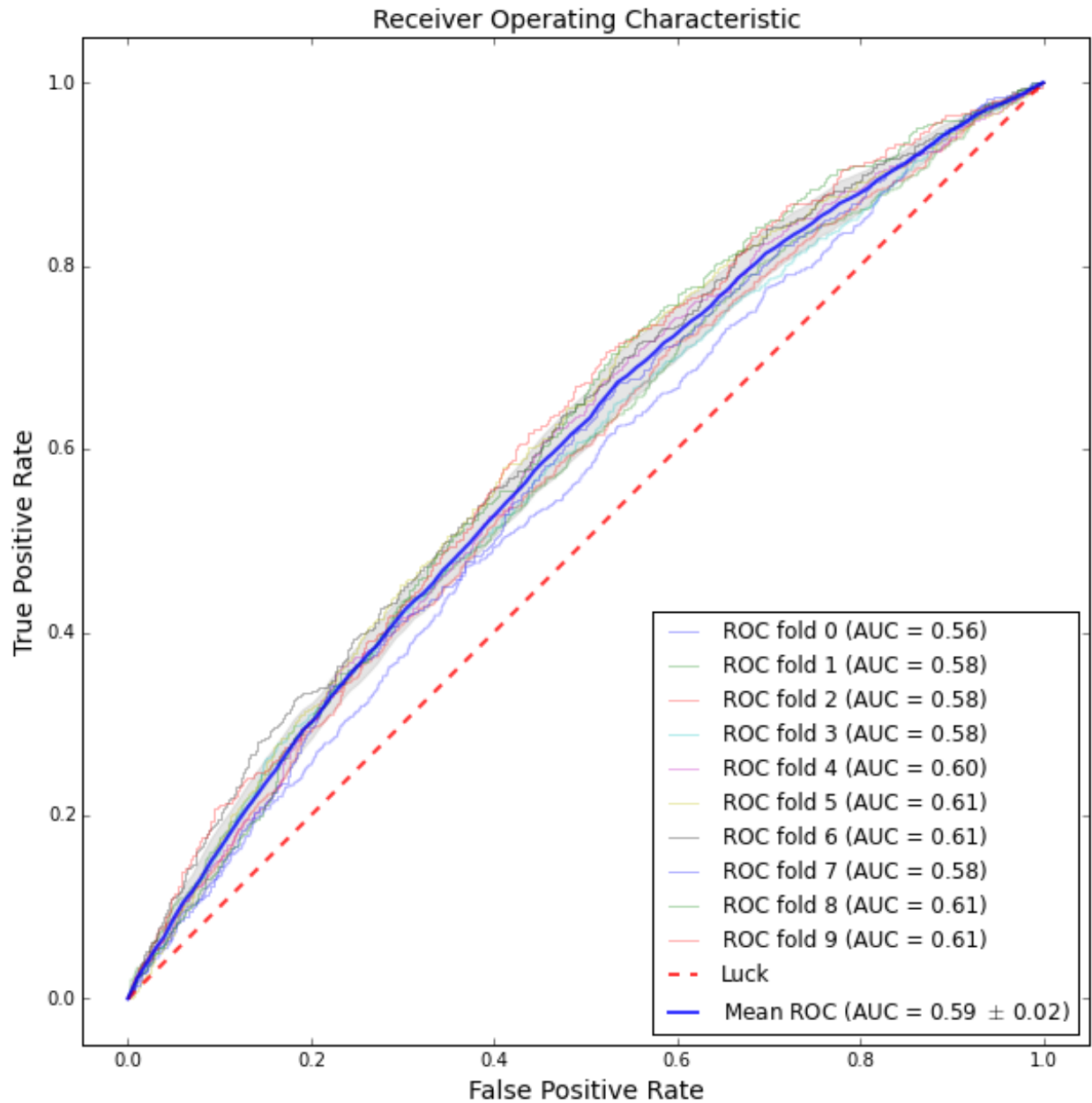
*Logistic Regression Confusion Matrix*

True	Predicted		All
	Did not transfer or graduate	Transferred and graduated	
Did not transfer or graduate	1963	1404	3367
Transferred and graduated	253	316	569
All	2216	1720	3936

Table 12

*Logistic Regression Classification Report*

accuracy	0.58
F1 score	0.27
precision	0.18
recall	0.55
ROC AUC	0.59



*Figure 26.* Logistic Regression ROC Curve

A Morris sensitivity analysis was performed on the four continuous and four categorical variables of the model. The variable effects from the sensitivity analysis indicated that gender, Student Community, being white, Student Aptitude, Institutional Characteristics, Focus on Academics, and being black had the most influence on the outcomes of the model. A bar chart and covariance chart of the sensitivity analysis are shown in Figure 27.



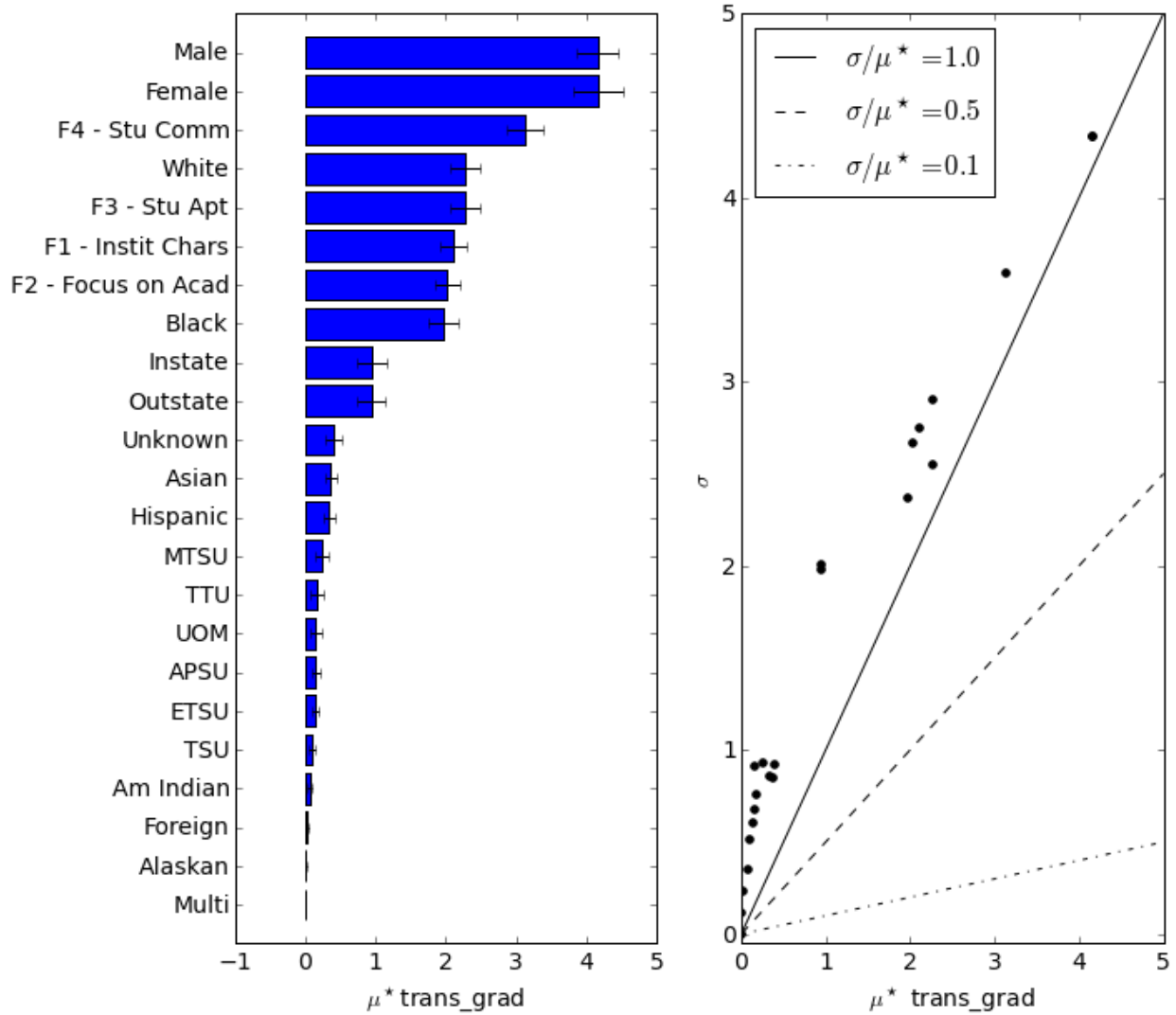


Figure 27. Logistic Regression Sensitivity Analysis

#### Research Question #6

Is the predictive power of decision trees for determining which students will leave their home institutions and graduate elsewhere greater than 50%?

H<sub>0</sub>6: The predictive power of decision trees for determining which students will leave their home institution and graduate elsewhere is less than or equal to 50%.

The decision tree model was run against four factors from Research Question #4. Factor scores for each case were generated during the factor analysis. In addition to the four factors, dummy variables for gender, race, residency, and first-time freshman institution were included in the model. The model was evaluated using 10-fold cross validation, resulting in an average of 3,936 cases per iteration. The resulting model had an accuracy score of 0.67. The confusion matrix is shown in Table 13, and the classification report is shown in Table 14. The ROC AUC score was 0.52. The plot of the ROC curve is shown in Figure 28. The null hypothesis was rejected. The predictive power of decision trees for determining which students will leave their home institution and graduate elsewhere was greater than 50%.

Table 13

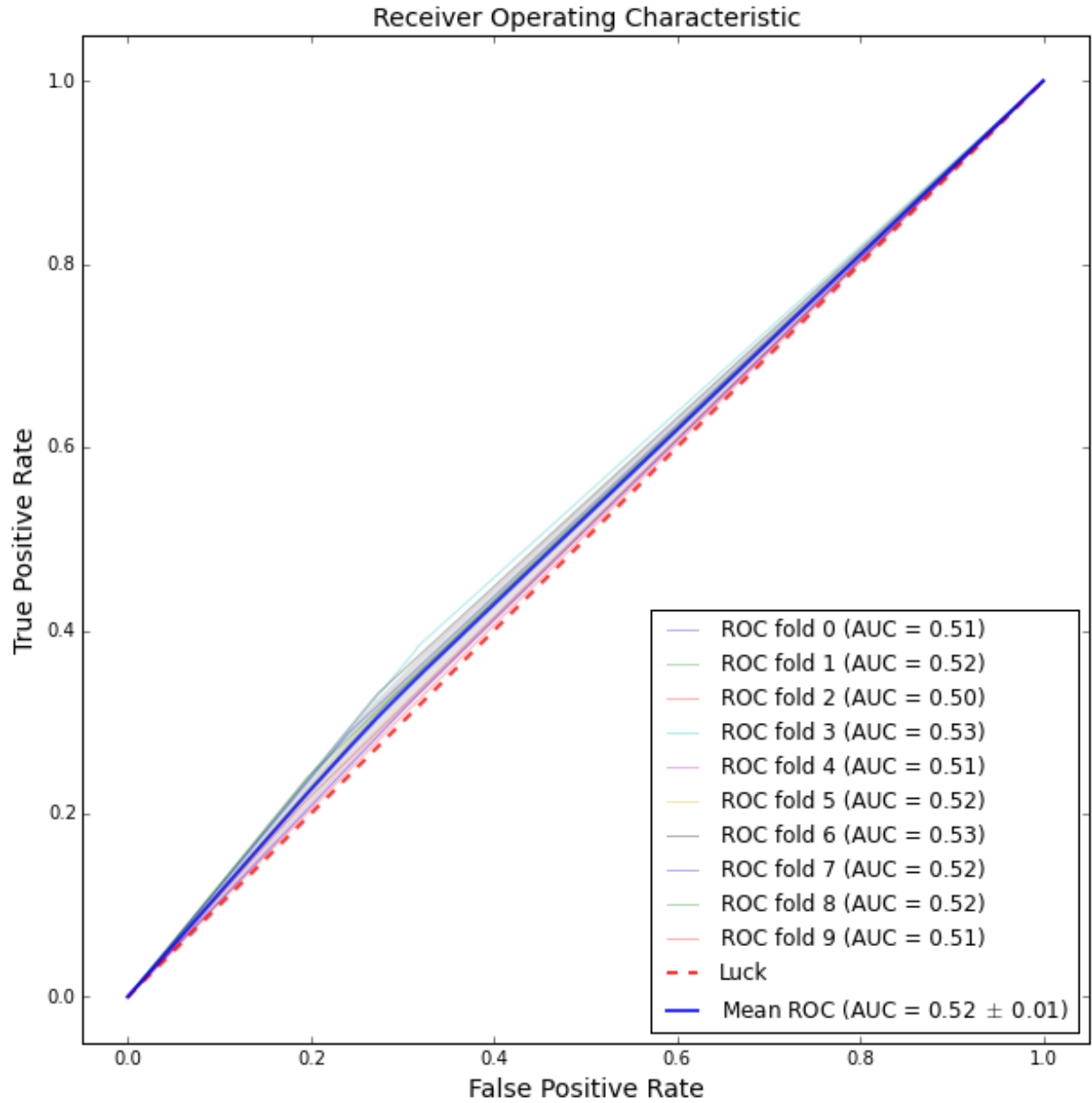
*C5 Decision Tree Confusion Matrix*

True	Predicted		All
	Did not transfer or graduate	Transferred and graduated	
Did not transfer or graduate	2464	903	3367
Transferred and graduated	393	176	569
All	2857	1079	3936

Table 14

*C5 Decision Tree Classification Report*

accuracy	0.67
F1 score	0.21
precision	0.16
recall	0.30
ROC AUC	0.52



*Figure 28. C5 Decision Tree ROC Curve*

A Morris sensitivity analysis was performed on the four continuous and four categorical variables of the model. The variable effects from the sensitivity analysis indicated that Institutional Characteristics, gender, Student Community Focus on Academics, Student Aptitude, and being white or black had the most influence on the outcomes of the model. A bar chart and covariance chart of the sensitivity analysis are shown in Figure 29.

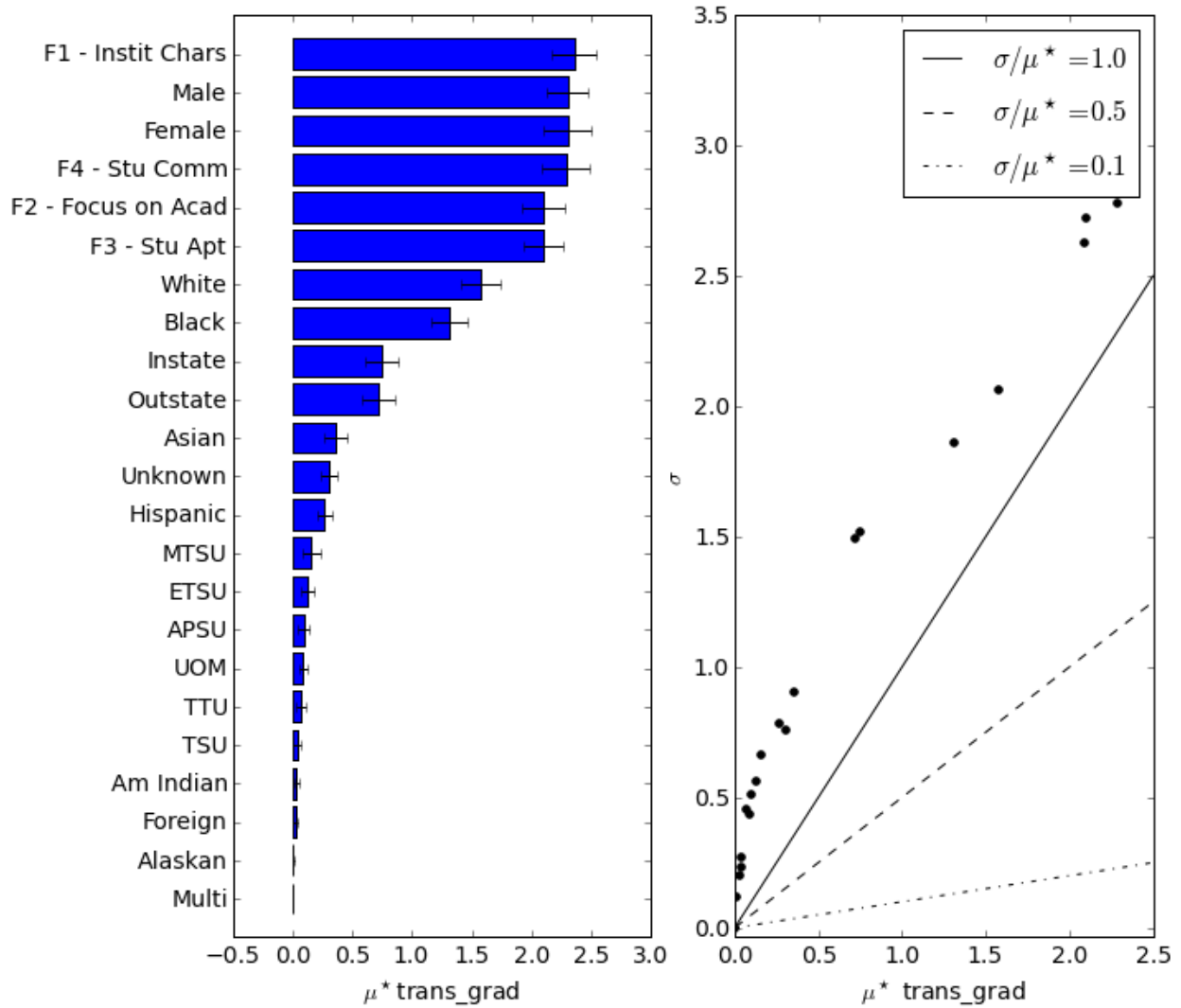


Figure 29. C5 Decision Tree Sensitivity Analysis

### Research Question #7

Is the predictive power of random forests for determining which students will leave their home institution and graduate somewhere else greater than 50%?

H<sub>0</sub>7: The predictive power of random forests for determining which students will leave their home institution and graduate elsewhere is less than or equal to 50%.

The random forest model was run against four factors from Research Question #4. Factor scores for each case were generated during the factor analysis. In addition to the four factors, dummy variables for gender, race, residency, and first-time freshman institution were included in the model. The model was evaluated using 10-fold cross validation, resulting in an average of 3,937 cases per iteration. The resulting model had an accuracy score of 0.76. The confusion matrix is shown in Table 15, and the classification report is shown in Table 16. The ROC AUC score was 0.54. The plot of the ROC curve is shown in Figure 30. The null hypothesis was rejected. The predictive power of random forests for determining which students will leave their home institution and graduate elsewhere was greater than 50%.

Table 15

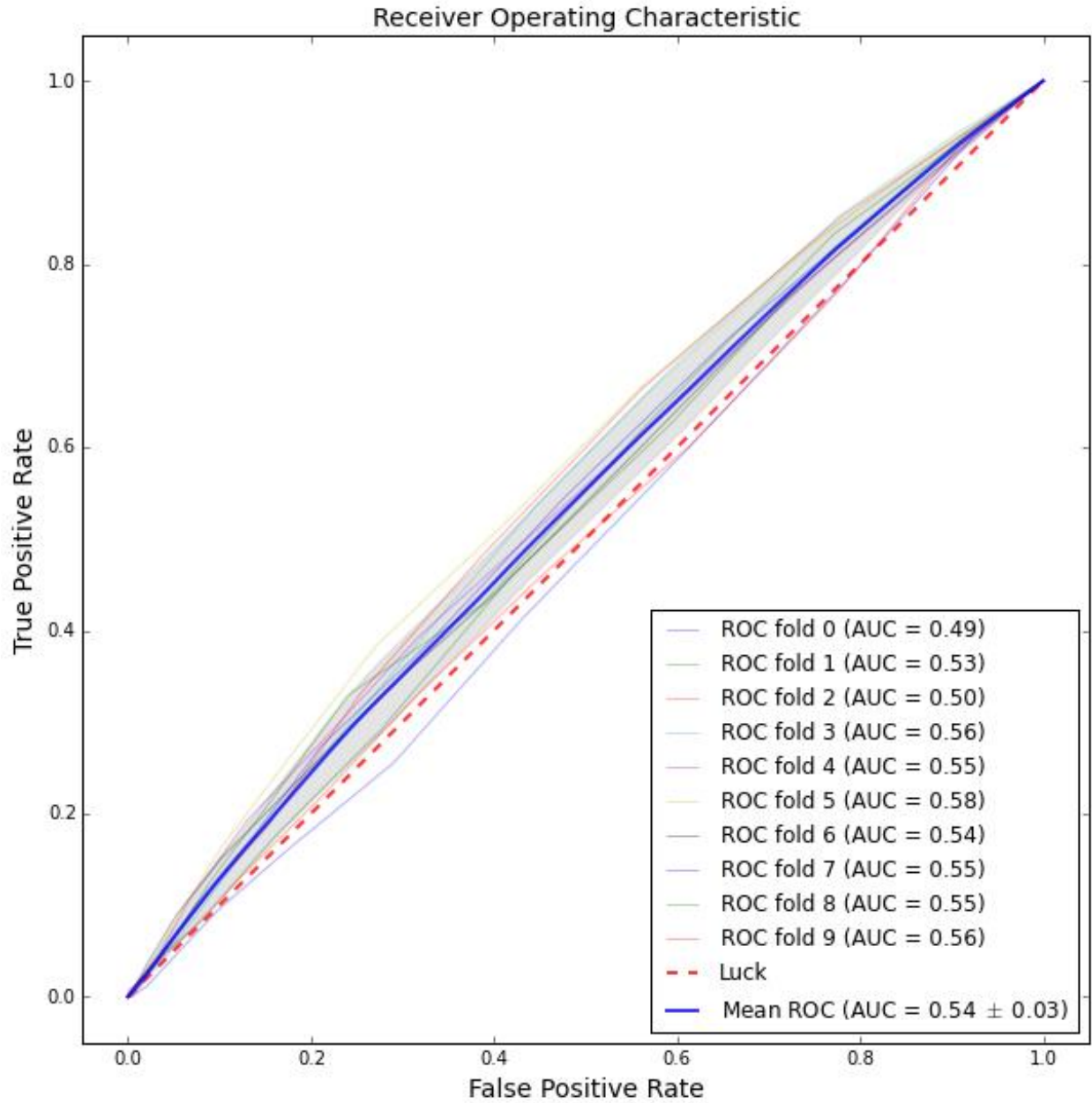
*Random Forest Confusion Matrix*

True	Predicted		All
	Did not transfer or graduate	Transferred and graduated	
Did not transfer or graduate	2820	547	3367
Transferred and graduated	454	115	569
All	3274	662	3936

Table 16

*Random Forest Classification Report*

accuracy	0.75
F1 score	0.18
precision	0.17
recall	0.20
ROC AUC	0.54



*Figure 30.* Random Forest ROC Curve

A Morris sensitivity analysis was performed on the four continuous and four categorical variables of the model. The variable effects from the sensitivity analysis indicated that Gender, Student Community, Institutional Characteristics, Student Aptitude, Focus on Academics, and race had the most influence on the outcomes of the model. A bar chart and covariance chart of the sensitivity analysis are shown in Figure 31.

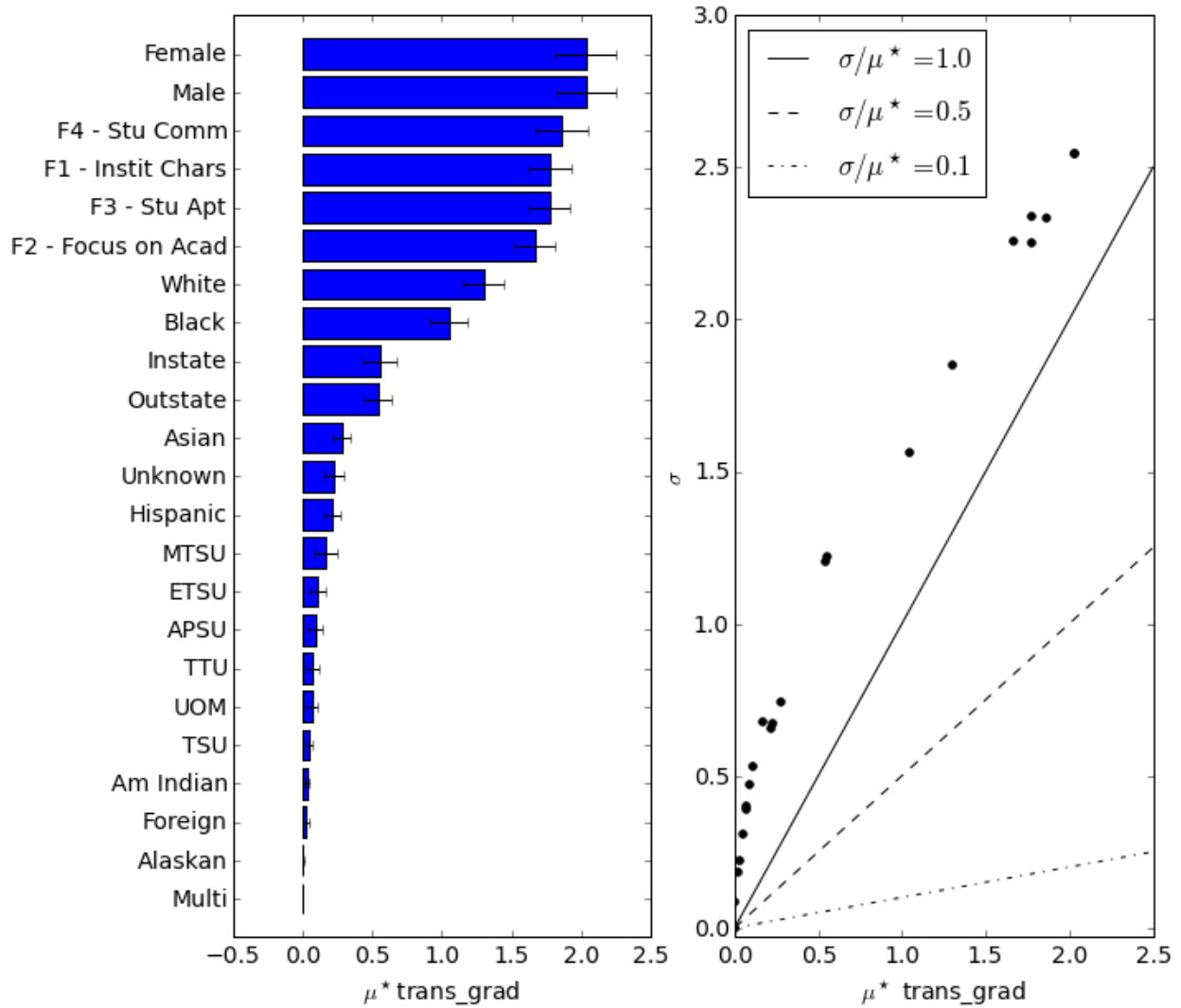


Figure 31. Random Forest Sensitivity Analysis

### Research Question #8

Is the predictive power of artificial neural networks for determining which students will leave their home institution and graduate somewhere else greater than 50%?

H<sub>0</sub>8: The predictive power of artificial neural networks for determining which students will leave their home institution and graduate elsewhere is less than or equal to 50%.

The artificial neural network model was run against four factors from Research Question #4. Factor scores for each case were generated during the factor analysis. In addition to the four factors, dummy variables for gender, race, residency, and first-time freshman institution were included in the model. The model was evaluated using 10-fold cross validation, resulting in an average of 3,937 cases per iteration. The resulting model had an accuracy score of 0.59. The confusion matrix is shown in Table 17, and the classification report is shown in Table 18. The ROC AUC score was 0.56. The plot of the ROC curve is shown in Figure 32. The null hypothesis was therefore rejected. The predictive power of artificial neural networks for determining which students will leave their home institution and graduate elsewhere was greater than 50%.

Table 17

*Artificial Neural Network Confusion Matrix*

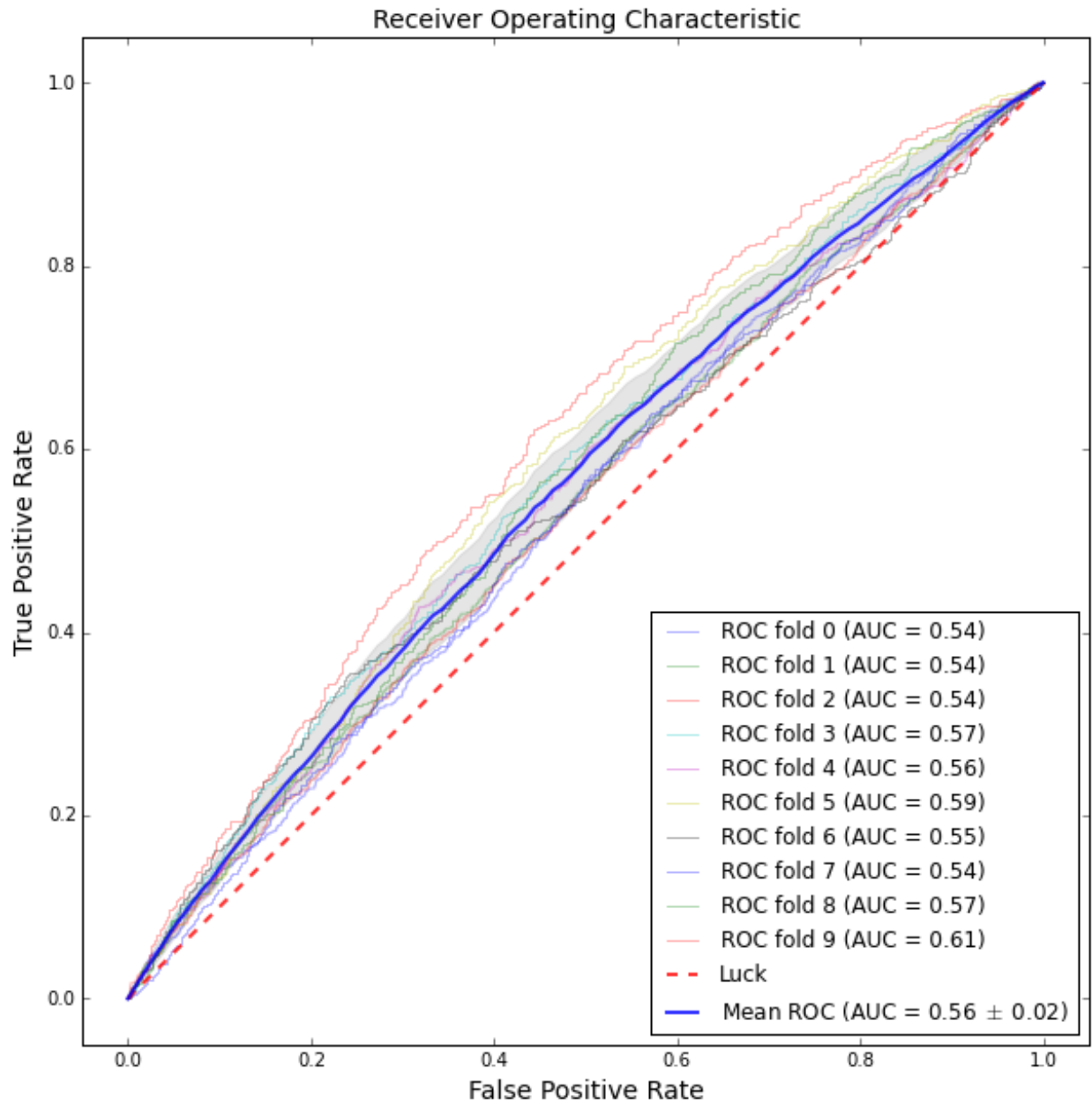
True	Predicted		All
	Did not transfer or graduate	Transferred and graduated	
Did not transfer or graduate	2063	1304	3367
Transferred and graduated	298	271	569
All	2361	1575	3936

Table 18

*Artificial Neural Network Classification Report*

accuracy	0.59
F1 score	0.25
precision	0.17
recall	0.46
ROC AUC	0.56





*Figure 32.* Artificial Neural Network ROC Curve

A Morris sensitivity analysis was performed on the four continuous and four categorical variables of the model. The variable effects from the sensitivity analysis indicated that gender, Student Community, Student Aptitude, being white, Institutional Characteristics, Focus on Academics, and being black had the most influence on the outcomes of the model. A bar chart and covariance chart of the sensitivity analysis are shown in Figure 33.

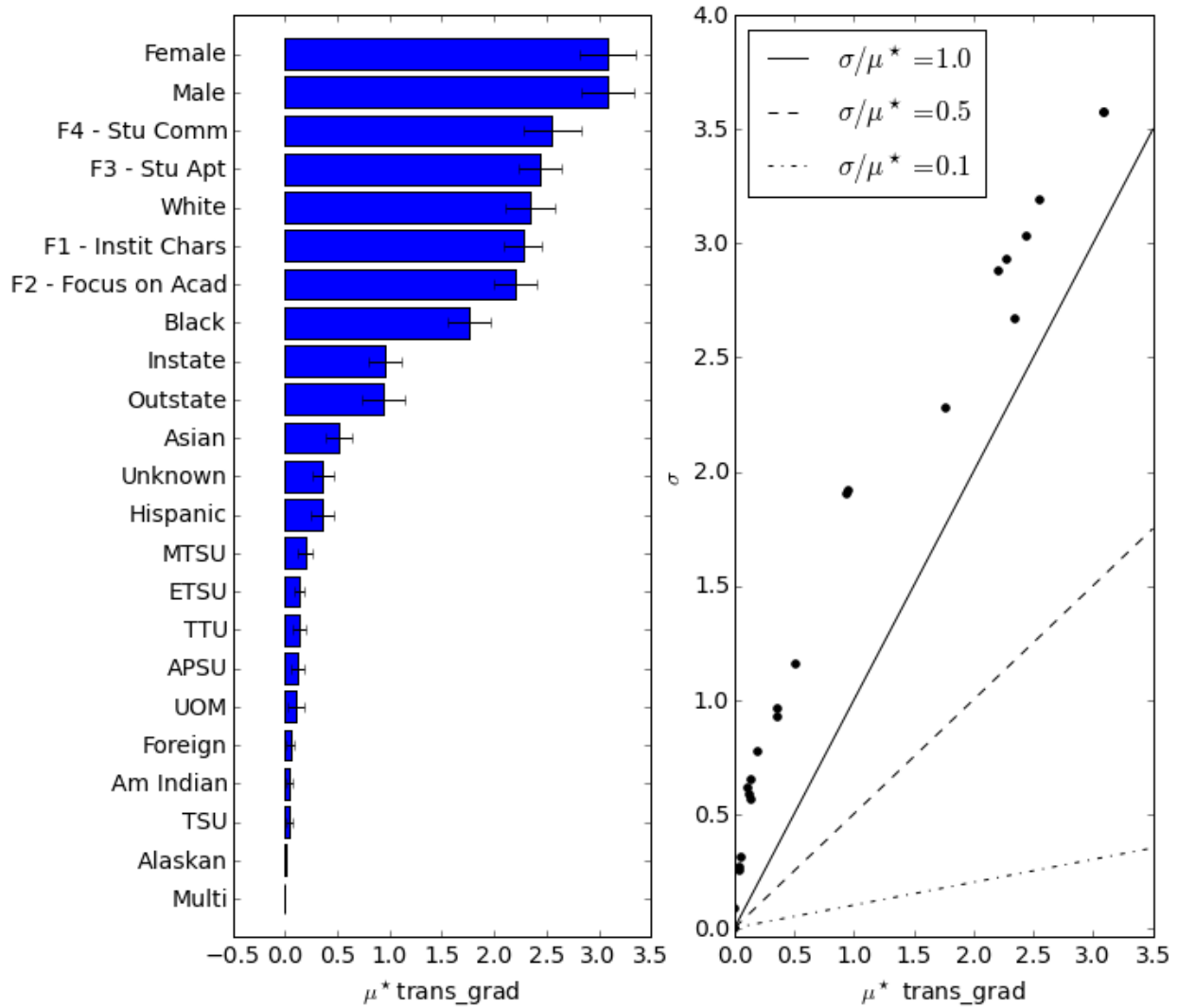


Figure 33. Artificial Neural Network Sensitivity Analysis

### Research Question #9

Is the predictive power of support vector machines for determining which students will leave their home institution and graduate somewhere else greater than 50%?

H<sub>0</sub>9: The predictive power of support vector machines for determining which students will leave their home institution and graduate elsewhere is less than or equal to 50%.

The support vector machine model was run against four factors from Research Question #4. Factor scores for each case were generated during the factor analysis. In addition to the four factors, dummy variables for gender, race, residency, and first-time freshman institution were included in the model. The model was evaluated using 10-fold cross validation, resulting in an average of 3,937 cases per iteration. The resulting model had an accuracy score of 0.58. The confusion matrix is shown in Table 19, and the classification report is shown in Table 20. The ROC AUC score was 0.59. The plot of the ROC curve is shown in Figure 34. The null hypothesis was therefore rejected. The predictive power of support vector machines for determining which students will leave their home institution and graduate elsewhere was greater than 50%.

Table 19

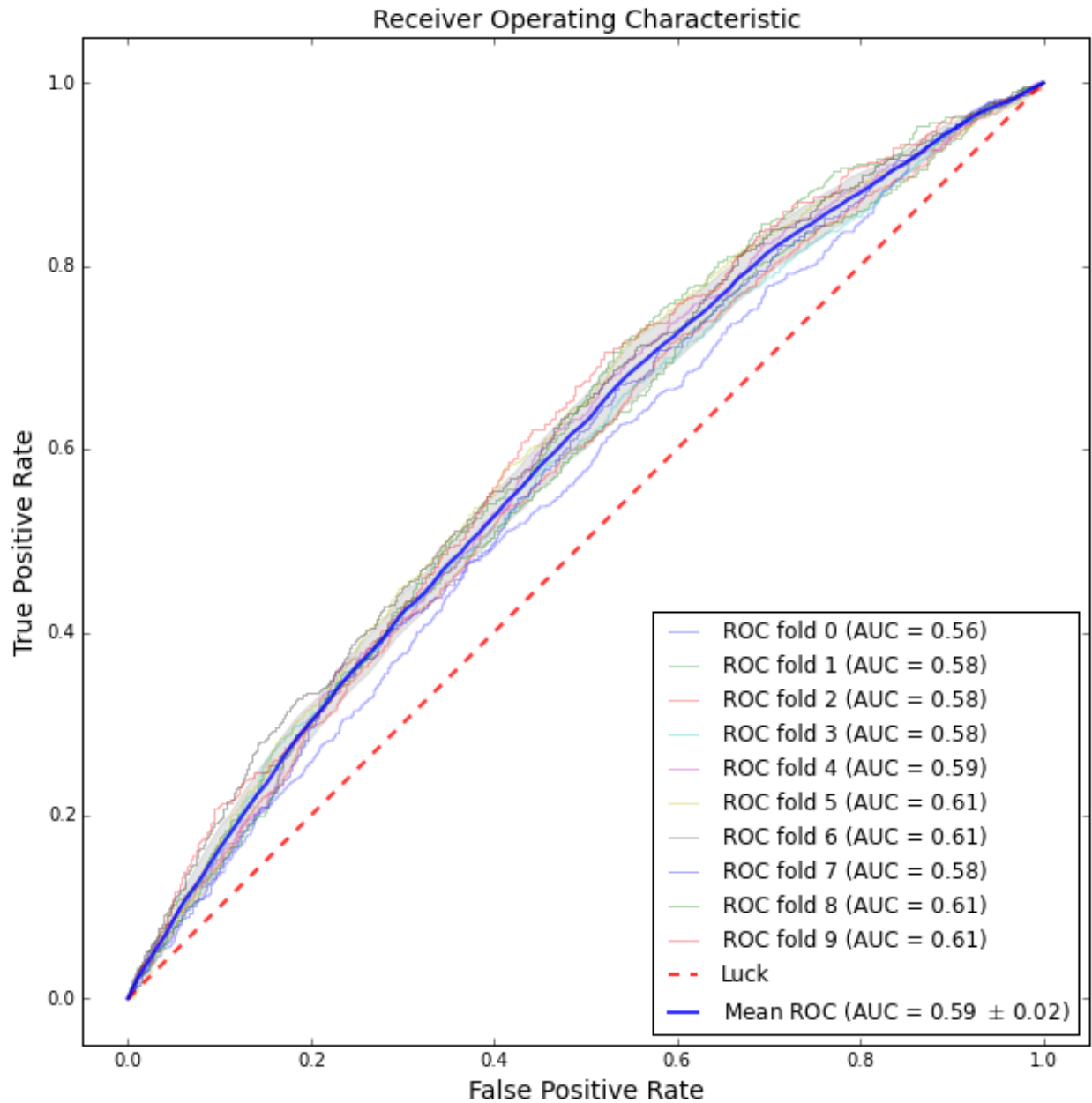
*Support Vector Machine Confusion Matrix*

True	Predicted		All
	Did not transfer or graduate	Transferred and graduated	
Did not transfer or graduate	1956	1411	3367
Transferred and graduated	253	317	570
All	2209	1728	3937

Table 20

*Support Vector Machine Classification Report*

accuracy	0.58
F1 score	0.27
precision	0.18
recall	0.55
ROC AUC	0.59



*Figure 34.* Support Vector Machine ROC Curve

A Morris sensitivity analysis was performed on the four continuous and four categorical variables of the model. The variable effects from the sensitivity analysis indicated that gender, Student Community, being white, Student Aptitude, Institutional Characteristics, Focus on Academics, and being black had the most influence on the outcomes of the model. A bar chart and covariance chart of the sensitivity analysis are shown in Figure 35.

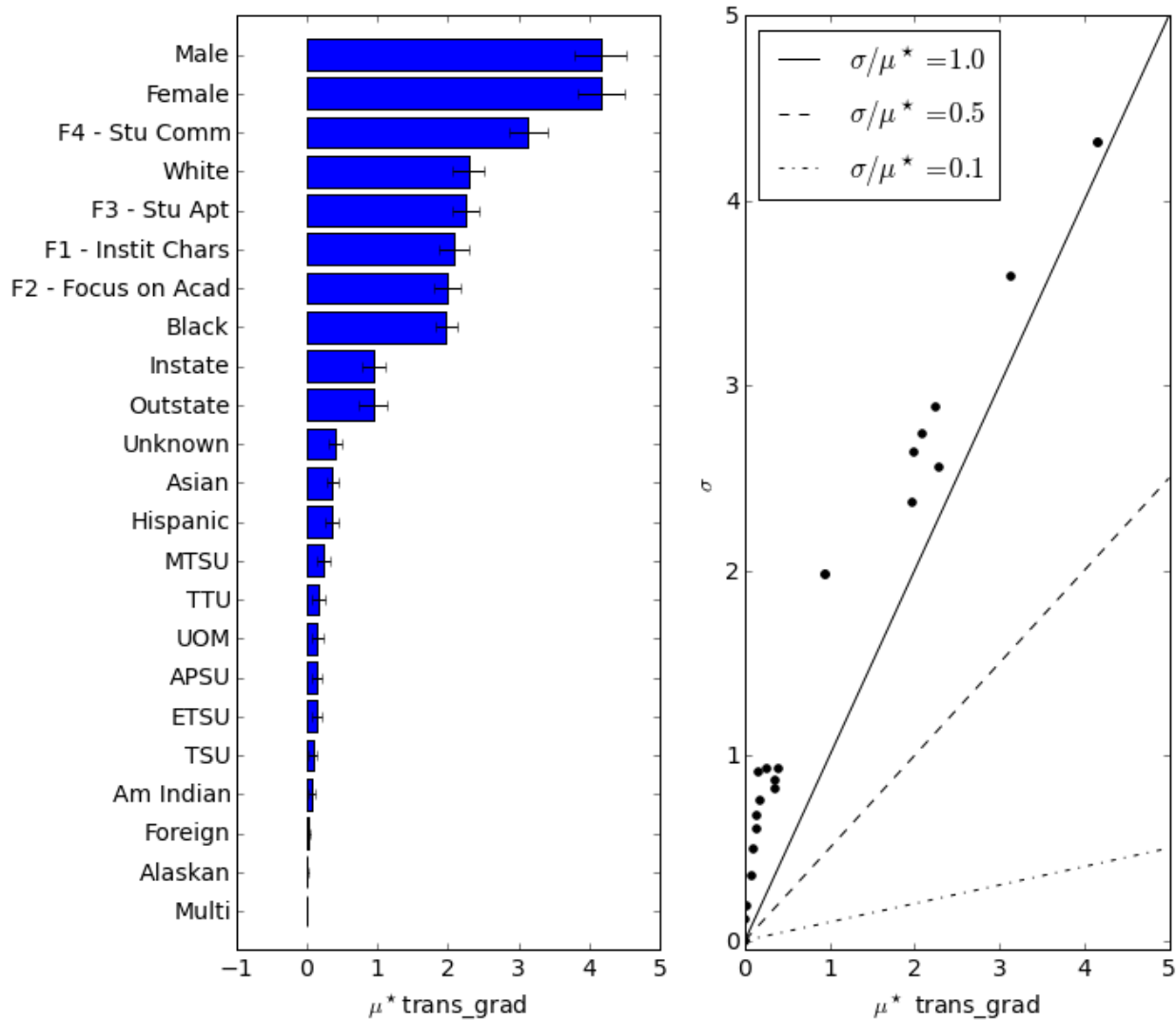


Figure 35. Support Vector Machine Sensitivity Analysis

### Research Question #10

Which of the data mining techniques: logistic regression, decision tree, random forest, artificial neural network, or support vector machines provide the best classification result in predicting students who will leave their home institution and graduate somewhere else?

$H_{010_1}$ : Logistic regression has no stronger predictive power than decision trees, random forests, artificial neural networks, or support vector machines.

H<sub>0</sub>10<sub>2</sub>: Decision trees have no stronger predictive power than logistic regression, random forests, artificial neural networks, or support vector machines.

H<sub>0</sub>10<sub>3</sub>: Random forests have no stronger predictive power than logistic regression, decision trees, artificial neural networks, or support vector machines.

H<sub>0</sub>10<sub>4</sub>: Artificial neural networks have no stronger predictive power than logistic regression, decision trees, random forests, or support vector machines.

H<sub>0</sub>10<sub>5</sub>: Support vector machines have no stronger predictive power than logistic regression, decision trees, random forests, or artificial neural networks.

Table 21 shows the ROC AUC scores for each model. Based on these values, the null hypothesis that logistic regression had no stronger predictive power than other models was rejected. The decision tree model had the weakest predictive power of all the models. Therefore, the null hypothesis that decision trees have no stronger predictive power than logistic regression, random forests, artificial neural networks, or support vector machines was maintained. The random forest model had stronger predictive power than the decision tree model. The null hypothesis that random forests have no stronger predictive power than logistic regression, decision trees, artificial neural networks, or support vector machines was rejected. The artificial neural network model had the third strongest predictive power next to the logistic regression and support vector machine models. The null hypothesis that artificial neural networks have no stronger predictive power than logistic regression, decision trees, random forests, or support vector machines was rejected. The support vector machine had the same predictive power as the logistic regression model. Therefore, the null hypothesis that support vector machines have no stronger predictive power than logistic regression, decision trees, random forests, or artificial neural networks was rejected.

Table 21

*Comparison of Model ROC AUC*

Model	ROC AUC
Logistic Regression	0.59
Decision Tree	0.52
Random Forest	0.54
Artificial Neural Network	0.56
Support Vector Machine	0.59

Chapter Summary

This chapter presented the factor analysis of variables related to students who began at a university in the former Tennessee Board of Regents system. Factor analysis was performed for students who graduated within 6 years, students who did not graduate, students who transferred out, and students who transferred and graduated. Factor analysis was performed for each group of students twice: once with complete information and once with data from just the student's first semester. The resulting factors for students who transferred and graduated were used in five predictive models to determine if such students could be identified within their first semester. All five predictive models had predictive power greater than 50%. The logistic regression and support vector machine models had the most predictive power at 59%, while the decision tree model had the least predictive power at 52%. A summary of these findings as well as conclusions, implications for practice, and recommendations for further study are presented in Chapter 5.

## CHAPTER 5

### SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

This chapter includes a summary of findings, conclusions, implications for practice, and recommendations for future research. The purpose of this quantitative study was to discover factors about first-time freshmen that began at a university in the former Tennessee Board of Regents (TBR) system, transferred to any other institution after their first year, and graduated with a degree or certificate. In addition, this study sought to determine if a predictive model can be generated to identify these particular students prior to their initial departure. Student data were provided by the Tennessee Higher Education Commission (THEC) for first-time freshmen cohorts between 2006 and 2009. Demographic, enrollment, financial aid, transfer, and graduation data were available for 39,379 students. Institutional data from the Integrated Postsecondary Education Data System (IPEDS) were also collected and merged with the student data. IPEDS data points included faculty counts, nonfaculty staff counts, student enrollment, tuition and fees, state appropriations, and expenditures on instruction, research, student support, academic support, and institutional support. These data were analyzed with factor analysis to determine which ones influenced student choices on transferring to another institution and graduating. The resulting factors were then used with a small set of demographic data points to generate predictive models for identifying first-time freshmen who will eventually transfer out from their initial institution and graduate somewhere else.



### Summary of the Findings

Chapter 1 of this dissertation presented 10 research questions used as the basis for statistical analyses. The 10 questions were again reported in Chapter 3 along with the corresponding hypotheses. Principal component analysis was used for Research Questions 1 through 4. Logistical regression was used for Research Question 5. A decision tree was used for Research Question 6. A random forest model was used for Research Question 7. A multilayer perceptron artificial neural network was used for Research Question 8, and a support vector machine was used for Research Question 9. A simple comparison of the predictive power of each model from Research Questions 5 through 9 was used for Research Question 10.

Characteristics of the initial and graduation institution had the greatest impact on students' decisions to remain enrolled at their initial institution, transfer to another institution, and graduate. Various student characteristics such as aptitude, in the form of high school GPA and ACT scores, and distance from home impacted student decisions to a lesser degree. The results of the factor analyses supported much of the research examined in the literature review.

Logistic regression and support vector machines had the most predictive power, followed by the artificial neural network and random forest models. The decision tree model had the least predictive power. The Receiver Operator Curve and confusion matrix were the most useful metrics for evaluating the models. Sensitivity analysis for each model indicated that gender had the greatest impact on the outcomes of the model, followed by the four factor scores. Race and residency impacted outcomes more than the initial institution.

### Research Question #1

Which characteristics are most likely to predict graduation in 6 years after enrollment for first-time freshmen students under the age of 24 from a 4-year institution in Tennessee?

A principal component analysis of all variables known about students in the selected population resulted in nine factors. The factors were described in order as “Initial Institution Characteristics,” “Graduation Institution Characteristics,” “Time to Graduation,” “Initial Institution’s Focus on Academics,” “Student Aptitude,” “Student Community,” “Student Major,” “Transfer Institution Distance,” and “Student Financial Aid.” The first two factors accounted for half of the variance in the data set, indicating that the institution is important to students’ enrollment and retention choices.

A principal component analysis of variables known about students only in their first semester resulted in four factors. The factors were described in order as “Institutional Characteristics,” “Institution’s Focus on Academics,” “Student Aptitude,” and “Student Community.” The Institutional Characteristics factor accounted for most of the variance at 43%, while the Institution’s Focus on Academics factor accounted for 12% of the variance.

The institutional characteristic factors from both PCAs supported the research of Crawford (2015) and Gansemer-Topf and Schuh (2006). Factors relating to a sense of student community supported the research on social and regional tethering of Wilson et al. (2016), as well as the research on social integration by Braxton et al. (2008). Factors relating to students’ high school GPA and ACT scores supported the research of Willingham (1985).

### Research Question #2

What characteristics identify first-time freshmen students under the age of 24 who transfer to another higher education institution?

A principal component analysis of all variables known about students in the selected population resulted in eight factors. The factors were described in order as “Graduation Institution Characteristics,” “Initial Institution Characteristics,” “Initial Institution’s Focus on Academics,” “Student Connectedness,” “Student Aptitude,” “Length of Time at Institution,” and “Transfer Institution Distance.” The first two factors accounted for half of the variance in the data set, indicating that the institution is important to students’ enrollment and retention choices.

A principal component analysis of variables known about students only in their first semester resulted in four factors. The factors were described in order as “Institutional Characteristics,” “Institution’s Focus on Academics,” “Student Community,” and “Student Aptitude.” The Institutional Characteristics factor accounted for most of the variance at 44%.

Institution characteristics were once again important factors. For the first PCA, factors relating to the number of semesters in school, the amount of credit hours taken each semester, and the number of major changes supported Astin’s (1993), Pascarella and Terenzini’s (2005), and Reason’s (2009) research into student attitudes contributing to retention. The factor for institution’s focus on academics supported the research of McCormick et al. (2009) who found that horizontal transfers were more likely to focus on research and other academic pursuits at their transfer institution. The Student Connectedness and Transfer Institution Distance also supported the research into students’ sense of community and social or regional tethering (Braxton et al., 2008; Wilson et al., 2016).

### Research Question #3

What characteristics identify first-time freshmen students under the age of 24 who did not graduate from a 4-year institution in Tennessee?

A principal component analysis of all variables known about students in the selected population resulted in eight factors. The factors were described in order as “Graduation Institution Characteristics,” “Initial Institution Characteristics,” “Time Spent at Institution,” “Initial Institution’s Focus on Academics,” “Transfer Community,” “Initial Institution Community,” “Student Aptitude,” and “Student Financial Aid.” The first two factors accounted for half of the variance in the data set, indicating that the institution is important to students’ enrollment and retention choices.

A principal component analysis of variables known about students only in their first semester resulted in four factors. The factors were described in order as “Institutional Characteristics,” “Institution’s Focus on Academics,” “Student Community,” and “Student Aptitude.” The Institutional Characteristics factor accounted for most of the variance at 46%.

The factors relating to this population of students were very similar to the factors for students who did graduate. Graduation institution characteristics were an important factor as many of these students transferred away and graduated at either a 2-year institution or one outside of Tennessee. Changes in major and the number of other students with the same major were not a factor for this group. Students who did not graduate were not as focused on their major as those students who did graduate. That behavior supported Heller and Cassady’s 2015 research on student goal setting behavior.

#### Research Question #4

What characteristics identify first-time freshmen students under the age of 24 who began at a 4-year institution in Tennessee, transferred to another higher education institution, and graduated?

A principal component analysis of all variables known about students in the selected population resulted in eight factors. The factors were described in order as “Graduation Institution Characteristics,” “Initial Institution Characteristics,” “Time to Graduation,” “Student Aptitude,” “Initial Institution’s Focus on Academics,” “Student Community,” “Transfer Community,” and “Student Financial Aid.” The first two factors accounted for half of the variance in the data set, indicating that the institution is important to students’ enrollment, and retention choices.

A principal component analysis of variables known about students only in their first semester resulted in five factors. The factors were described in order as “Institutional Characteristics,” “Institution’s Focus on Academics,” “Student Aptitude,” and “Student Community.” The Institutional Characteristics factor accounted for most of the variance at 41%, while the Institution’s Focus on Academics factor accounted for 10% of the variance.

The factors focused mainly on the institution of graduation and the ability of the student to graduate. As with the factors for Research Question #2, students appeared to focus on academics and taking more credit hours. Delen (2010) and Herzog (2006) found that credit hour accumulation contributed to the power and sensitivity of their predictive models. Student focus on high class loads supported Delen’s and Herzog’s findings. In addition, factors relating to students’ sense of community appeared have an impact, further supporting the research of Wilson et al. (2016) and Braxton et al. (2008).

#### Research Question #5

Is the predictive power of logistic regression for determining which students will leave their home institution and graduate elsewhere greater than 50%?

The logistic regression model tied with the support vector machine model for having the most predictive power of the models generated. The predictive power of the model is likely why it is commonly used as a baseline model in predictive model analysis (Porter, 2002; Provost & Fawcett, 2013). However, at 59% probability, the predictive power of the model in this study was weak. The confusion matrix showed a high number of false positives (Type I errors) and false negatives (Type II errors). The model correctly identified a little over half of the students who actually transferred and graduated elsewhere, leading to a high recall rate. However, the model also incorrectly labeled over four times as many students who did not transfer or graduate. The low precision and F1 score values reflected the model's poor ability to predict relevant cases.

Skari (2014) used a similar experimental design to the one in this study. Factor analysis on 14 variables produced a set of three factors that were used in a logistic regression model, along with eight demographic variables. The predictive outcomes had high precision and accuracy. Given Skari's success with factor analysis and logistic regression, the poor predictive power of the logistic regression in this study was likely due to the selection of variables. The factors used in the model accounted for approximately 71% of the variance in the dataset for which the factor analysis was performed on, indicating that nearly 30% of the variance was left unaccounted for. There are likely data points missing that would improve the predictive power of the logistic regression.

The sensitivity analysis indicated that gender was an important factor for the logistic regression model. The Mann-Whitney U test for the gender variable was significant for the second factor, Focus on Academics. Student Community was the third most sensitive variable. Based on Reason's 2003 research, this may have been a result of interaction effects. Further

analysis may be warranted. The sensitivity of the model to the races of white and black represented the imbalance of these two races in comparison to other races. Factor one, Institutional Characteristics, accounted for the majority of the variance in the data set of continuous variables. However, factor one had the third least impact of the factor variables on the model.

#### Research Question #6

Is the predictive power of decision trees for determining which students will leave their home institutions and graduate elsewhere greater than 50%?

The decision tree model had the worst predictive power of the models. The model barely outperformed mere chance with an ROC score of 0.52. The confusion matrix showed that few students who transferred out and graduated elsewhere were correctly predicted. Twice the number of true positives were predicted to not graduate or transfer. The model had thus had a higher accuracy due to the high number of students being correctly predicted as not transferring or graduating. The data set was imbalanced because the number of students in the data set who transferred and graduated was approximately 15% of the selected population. Increasing the number of predictions where the student did not transfer or graduate thus increased the accuracy of the model by virtue of the imbalanced data set.

The weakness of the decision tree model appeared to contradict the research of Delen (2010) and Herzog (2006). Delen favored the use of decision trees due to their ease of understanding. However, if the results have no predictive power, the ease of understanding them is irrelevant. Herzog's research found that decision trees outperformed logistic regression when students graduated within 3 years. The predictive power of decision trees dropped for students

who took 6 or more years to graduate. The majority of students in the data model took more than 3 years to graduate, which could explain the deviation from Herzog's findings.

The results of the sensitivity analysis for the decision tree were similar to that of the logistic regression. Institutional Characteristics had the greatest impact, followed by gender. Student Community, Focus on Academics, and Student Aptitude had the next largest impact. In-state and Out-of-state residence followed the races of white and black, once more reflecting imbalances in the dataset.

#### Research Question #7

Is the predictive power of random forests for determining which students will leave their home institution and graduate somewhere else greater than 50%?

The random forest model had more predictive power than the single decision tree. This was reflected by the accuracy score and area under the curve. However, the confusion matrix showed that this was another result of the imbalanced data set. Very few students were predicted to transfer and graduate, resulting in the higher accuracy. Many more students were correctly predicted to not transfer or graduate. The precision, recall, and F1 score showed how the higher area under the curve was misleading in terms of determining the model's ability to correctly identify students who would transfer out and graduate.

The power of the random forest over the single decision tree was expected per Witten and Frank (2011). The use of synthetic minority over-sampling (SMOTE) to account for the imbalanced dataset did not improve the accuracy of the model. The use of over-sampling should have improved the predictive power of the random forest according to Burez and Van den Poel (2009), but that did not occur.



The sensitivity analysis for the random forest showed that once more gender had the most influence on the output of the model. However the fourth, first, third, and second factors, Student Aptitude, Institutional Characteristics, Student Community, and Focus on Academics had the next most influence. Again, this may have been a result of interaction effects (Reason, 2003). The races white and black along with Out-of-state and In-state residence had the next most influence, similar to their impact on the other predictive models. The sensitivity of the variables for the random forest was similar to that of the decision tree model, which should be expected as the random forest model was a set of decision trees.

#### Research Question #8

Is the predictive power of artificial neural networks for determining which students will leave their home institution and graduate somewhere else greater than 50% ?

The artificial neural network (ANN) model had the third most predictive power at 56%. The confusion matrix showed that the model predicted cases only slightly worse than the logistic regression model. The accuracy, F1 score, precision, and recall reflected this performance of the model.

The artificial neural network (ANN) model outperformed the decision tree and random forest models. Herzog (2006) found that, similar to decision trees, artificial neural networks outperformed logistic regression when students graduated within 3 years. However, the predictive power of the ANN model dropped for students who took 6 or more years to graduate. The majority of students in the data model took more than 3 years to graduate, which could explain the difference from Herzog's findings. The impact of the dataset size on the ANN model, along with a review of the data variables used with the model may still be warranted.

Based on the sensitivity analysis, gender, Student Community, Student Aptitude, and being white had the most influence on the artificial neural network model. Institutional Characteristics, Focus on Academics, and being black had the next most influence on the model. As with the other models, In-state and Out-of-state residency followed in terms of influencing the model output. The results of the sensitivity analysis for the ANN model supported research about students' sense of community influencing retention and graduation decisions (Braxton et al., 2008; Wilson et al., 2016). In addition, the sensitivity of gender and race supported persistence research (Leppel, 2002; Peltier et al., 1999). As with the sensitivity analysis results of the other predictive models, these findings may have been the result of interaction effects (Reason, 2003). The ANN model had more of a black box nature than the previous models, making an accurate and thoughtful interpretation of these results important (Coglianese & Lehr, 2007). Further research into the results of the sensitivity analysis would be warranted if the ANN model were to be used in practice.

#### Research Question #9

Is the predictive power of support vector machines for determining which students will leave their home institution and graduate somewhere else greater than 50%?

The support vector machine (SVM) had the most predictive power, along with the logistic regression model at 59%. The results of the support vector machine were contradictory to other researchers' findings in that it did not outperform the logistic regression model. Delen (2010, 2011); Strecht et al. (2015); and Ding et al. (2016) found that support vector machines had more predictive power than decision trees, artificial neural networks, and logistic regressions. Delen's research found logistic regression to have the least predictive power of all

the predictive models. The weak predictive power of all the models indicated the need to review and revise the inputs to the models.

In addition to weak predictive power, the support vector machine model took hours to run with a nonlinear kernel, as opposed to a few minutes with the other predictive models. This was due to the nature of nonlinear kernels for SVMs that can have a quadratic order of complexity,  $O(n^2)$ , where  $n$  is the number of cases for the model to process (Joachims, 2016).

#### Research Question #10

Which of the data mining techniques: logistic regression, decision tree, random forest, artificial neural network, or support vector machines provide the best classification result in predicting students who will leave their home institution and graduate somewhere else?

Logistic regression and the support vector machine (SVM) provided the best classification results in predicting students who will leave their home institution and graduate somewhere else. Both models had the best Area under the Receiver Operating Curve (AUC) and confusion matrices, indicating better classification of cases. The outcome of the logistic regression model indicated why it was consistently used in other research as a baseline (Herzog, 2006; Porter, 2002). The artificial neural network (ANN) came close in performance to the logistic regression and support vector machine models, providing better predictive performance than the decision tree and random forest models. The performance of the ANN and SVM models was contradictory to research that would place their performance above the logistic regression mode (Delen, 2010; Ding et al., 2016; Strecht et al., 2015). The decision tree and random forest models had higher accuracy because they made more negative classifications. Because the majority of cases were for students who did not transfer or graduate, the decision tree and random forest models gained accuracy by classifying more students as such.

## Conclusions

Institutional expenditures, the number of faculty and staff, and the number of enrolled students mattered in terms of student decisions to transfer and graduate. Individual student factors such as distance from institution, number of students at the same institution in the same major or from the same high school, and ACT scores and high school GPA factored into student choices as well, but not to the degree of institutional characteristics. Student preferences toward institutions that the student attended, transferred to, and ultimately graduated from appeared to have a higher impact on student outcomes than what the student brought to that institution in terms of ability and social behaviors.

The data elements of the individual student factors came from the Tennessee Higher Education Commission (THEC) datasets. THEC data were a result of reporting for state performance funding. The institutional factors came from the Integrated Postsecondary Education Data System (IPEDS) in order to receive federal financial aid funds (Fuller, 2011; NPEC, 2009). The institutional factors had a greater impact on the predictive models than the individual student factors. This conclusion affirmed what Dougherty and Reddy (2011) found concerning attitudes about state performance funding. Reporting for performance funding was seen by institutional stakeholders as a perfunctory task with limited power to improve outcomes. In contrast, federal performance reporting through IPEDS appeared to provide more valuable institutional insights for why students chose to remain enrolled, transfer, or stop out.

Logistic regression and support vector machines outperformed other models including decision trees, random forests, and artificial neural networks. Support vector machines and artificial neural networks should have had more predictive power than the logistic regression model according to the research literature. Skari's 2014 study that used factor scores as inputs to

a logistic regression model indicated that the use of factor scores would be acceptable for any type of predictive model. However, the models may still have lacked variables that would have contributed to their predictive power. This, along with the small dataset used, could explain the results of the artificial neural network and support vector machine models that contradicted the research literature.

Evaluating each model by the graph of the Receiver Operating Curve and confusion matrix was instrumental in determining how well the models performed. While there were many metrics to consider, these two combined were the best metrics to examine. The Receiver Operating Curve and confusion matrix allowed for visual and intuitive interpretation of the predictive power of the various models (Aguilar, 2015; Chawla, 2005; Hamel, 2008).

### Recommendations for Practice

There are several recommendations for practice based on the results of this research. The first recommendation is to focus on logistic regression modeling. Aside from having the most predictive power of any model in this study except for support vector machines, it is a more common statistical method that other researchers will find easier to understand (Delen, 2010).

Support Vector Machines with nonlinear kernels should not be used on data sets with a large number of dimensions or variables. The order of complexity will cause the run time for the model to increase exponentially as additional variables are added (Joachims, 2016). Linear kernels relying on regression may be more appropriate for data sets with a large number of dimensions.

The choice of evaluation metrics for predictive models should be decided carefully. Although Area under the Receiver Operating Curve (AUC) was common in the research

literature, the AUC should not be the sole metric relied upon (Lobo, Jimenez-Valverde, & Real, 2007). A confusion matrix should be used along with the AUC, as these two metrics provide visual aid for the outcomes of the model. In addition, several other metrics such as the accuracy, precision, recall, and F1 score can be generated based on the values of the confusion matrix.

Even though the predictive power of the logistic regression model is low, it is not so low as to be negligible. Instead, the amount of resources invested in using the predictions should take the predictive power into account. Hiring new advisors to contact hundreds of students who have a 60% chance of leaving and being successful elsewhere would not be a good allocation of university resources. Tracking these students via an existing student information system and reaching out to them via email retention campaigns instead would be a low-cost intervention.

#### Recommendations for Future Research

Given the amount of data available, and the amount of data not yet collected, there are many recommendations for future research. The first recommendation is to run the predictive models against this data set again but focus on individual institutions rather than all the 4-year institutions in the state of Tennessee. Although the sensitivity analysis for each of the predictive models indicated that the individual institution had little influence on the model output the results of the Kruskal-Wallis test indicated that the continuous factors differed by institution. The inclusion of additional data points for the model, another recommendation for future research, could include student high school and major code. Such data points could be specific to the institution. For instance, Middle Tennessee State University has an aeronautical engineering program and East Tennessee State University has a Bluegrass program.

Factor analysis should be repeated per institution to focus on the individual institutions using the factor scores as in this research. Table 22 shows what the output for such a factor analysis would look like if done for a large public 4-year university in West Tennessee and for Research Question #4.

Table 22

*Factors for Transfer Graduates, West Tennessee University*

Component	1	2	3
Age			-0.79
Distance from home to initial institution			0.62
High school GPA		0.79	
Composite ACT		0.79	
Number of credit hours taken in first semester		0.47	
Number of FTF students with same major	0.58		
Tuition and fees at initial institution	0.97		
Institutional support expenditures at initial institution	0.91		
Student FTE at initial institution	0.96		
Eigenvalues	3.06	1.59	1.06
Percentage of total variance	34.01	17.62	11.77
Number of test measures	4	3	2

\*Loadings  $\geq 0.5$

Running a logistic regression based on the factor scores from that institution specific factor analysis and the categorical variables of gender, race, and residency would produce an ROC curve as shown in Figure 36.

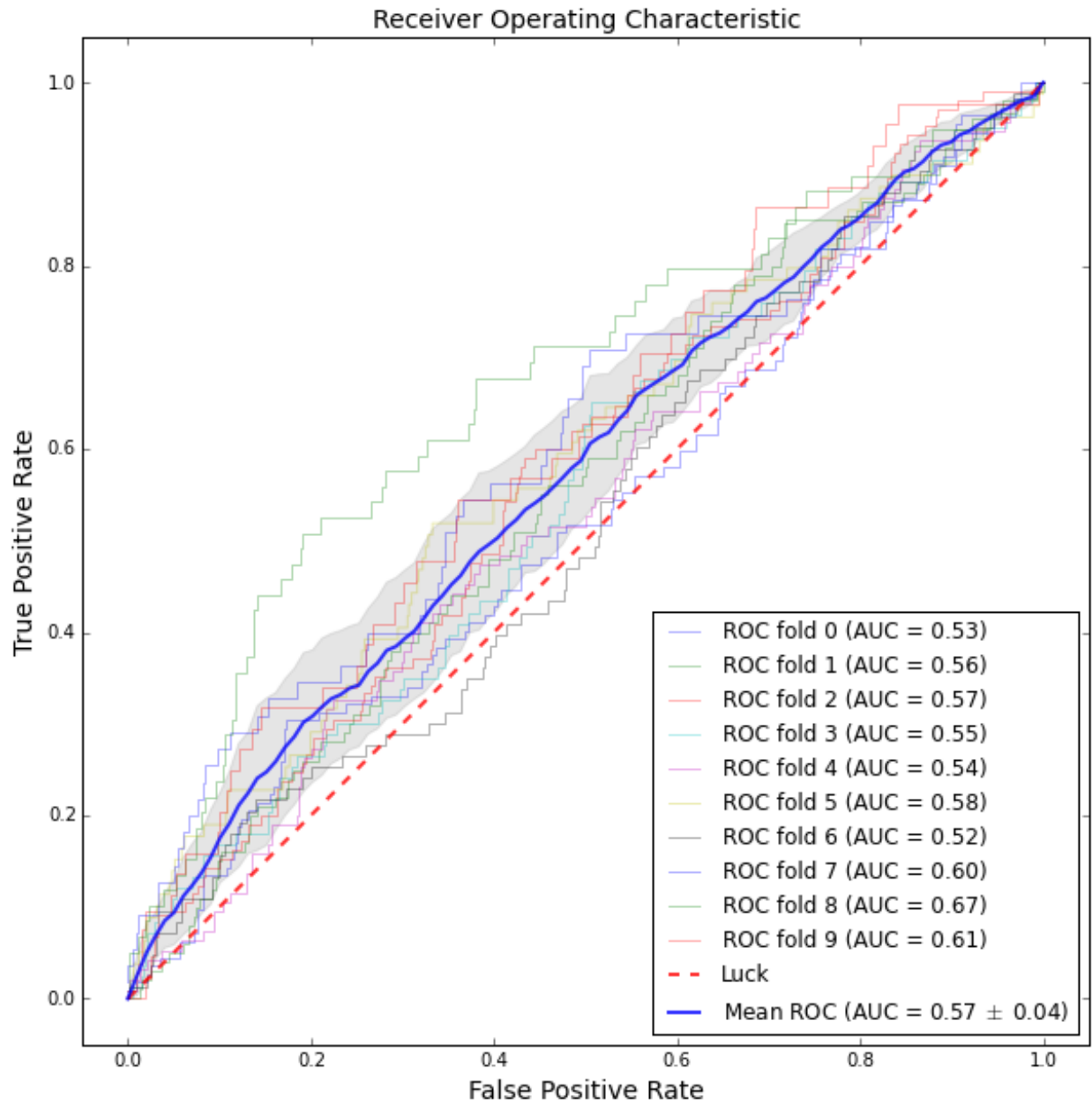


Figure 36. Logistic Regression ROC Curve, West TN University

The next recommendation is to explore the predictive models against the first research question relating to any type of graduate. This research focused on predicting students who would transfer out and graduate. The results did not produce strong predictive models. One cause of this was the imbalanced nature of the set of students who transfer out and graduate elsewhere. The entire set of students who graduated was more balanced at approximately 50%



(THEC, 2016). Applying the same form of institution specific factor analysis to the first research question for a large public 4-year university in Central Tennessee would result in output as shown in Table 23.

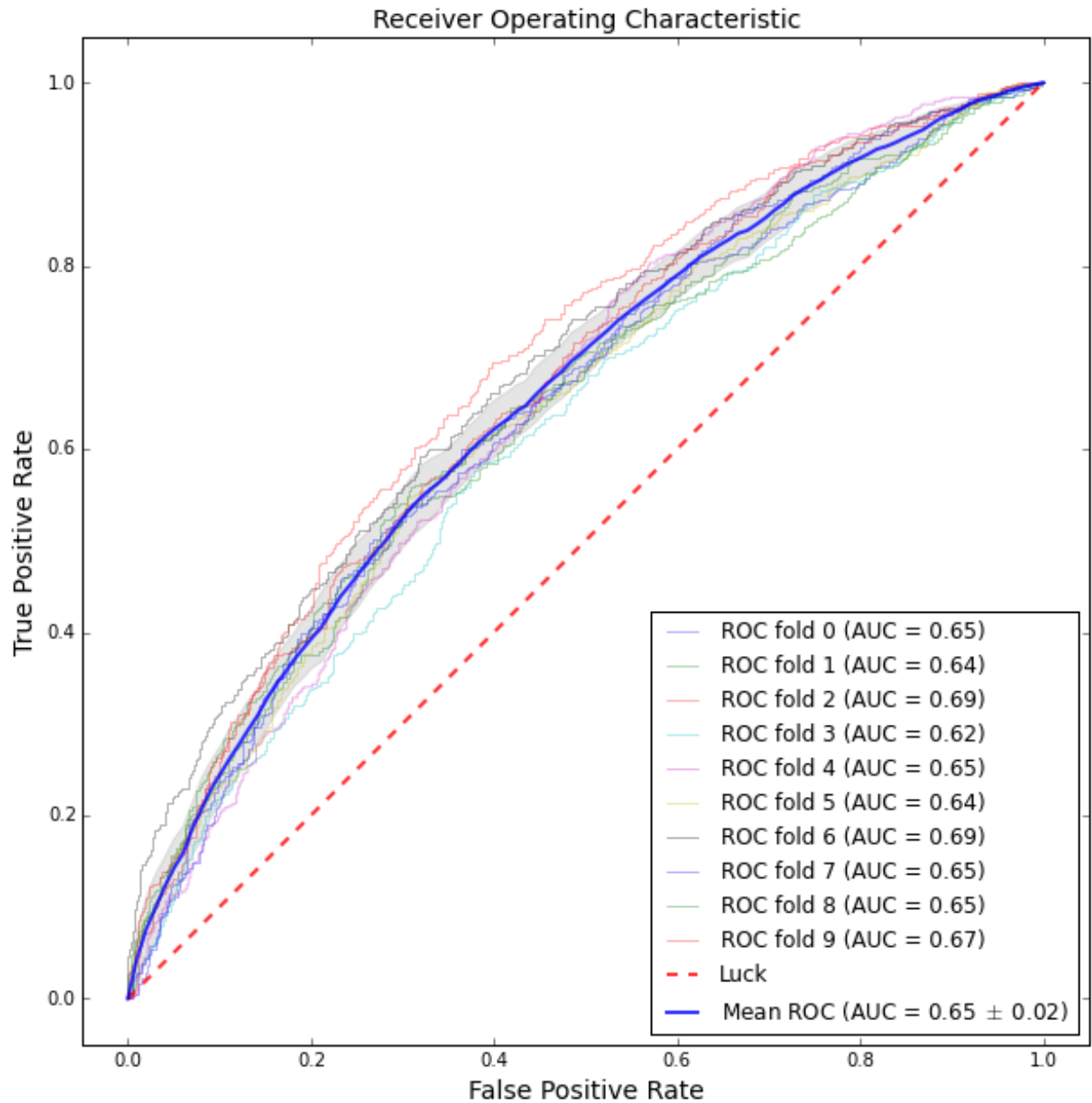
Table 23

*Factors for Graduation, Central Tennessee University*

Component	1	2	3
Distance from home to initial institution			0.78
High school GPA		0.75	
Composite ACT		0.78	
Number of credit hours taken in first semester		0.44	
Amount of state funds awarded			
Number of FTF students from same high school			-0.79
Number of full-time faculty at initial institution	0.96		
Tuition and fees at initial institution	0.98		
Student FTE at initial institution	0.99		
Eigenvalues	2.87	1.40	1.40
Percentage of total variance	31.94	15.55	15.55
Number of test measures	3	3	2

\*Loadings  $\geq 0.5$

Running a logistic regression based on the factor scores from that institution specific factor analysis and the categorical variables of gender, race, and residency would produce an ROC curve as shown in Figure 37. The more balanced data set caused the predictive power to increase by approximately 10%. For a single institution a logistic regression can correctly predict graduation between 62% and 70% of the time.



*Figure 37. Logistic Regression ROC Curve, Central TN University*

The next recommendation is to explore the predictive models against the second research question relating to transfer students, regardless of graduation. Applying the same form of institution specific factor analysis to the second research question for a medium-sized public 4-year university in North Central Tennessee would result in output as shown in Table 24.

Running a logistic regression based on the factor scores from that institution specific factor analysis and the categorical variables of gender, race, and residency would produce an ROC curve as shown in Figure 38. For a single institution a logistic regression can correctly predict transfers between 56% and 64% percent of the time.

Table 24

*Factors for Transfers, North Central Tennessee University*

Component	1	2	3	4
Age		-0.47		-0.43
Distance from home to initial institution			-0.72	
High school GPA		0.67		
Composite ACT		0.60		
Number of credit hours taken in first semester		0.60		
Amount of state funds awarded				0.76
Number of FTF students from same high school			0.80	
Number of FTF students with same major				0.42
Number of full-time faculty at initial institution	0.87			
Tuition and fees at initial institution	0.98			
Institutional support expenditures at initial institution	0.89			
Student FTE at initial institution	0.99			
Eigenvalues	3.52	1.41	1.39	1.09
Percentage of total variance	29.33	11.75	11.60	9.08
Number of test measures	4	4	2	3

\*Loadings  $\geq 0.5$

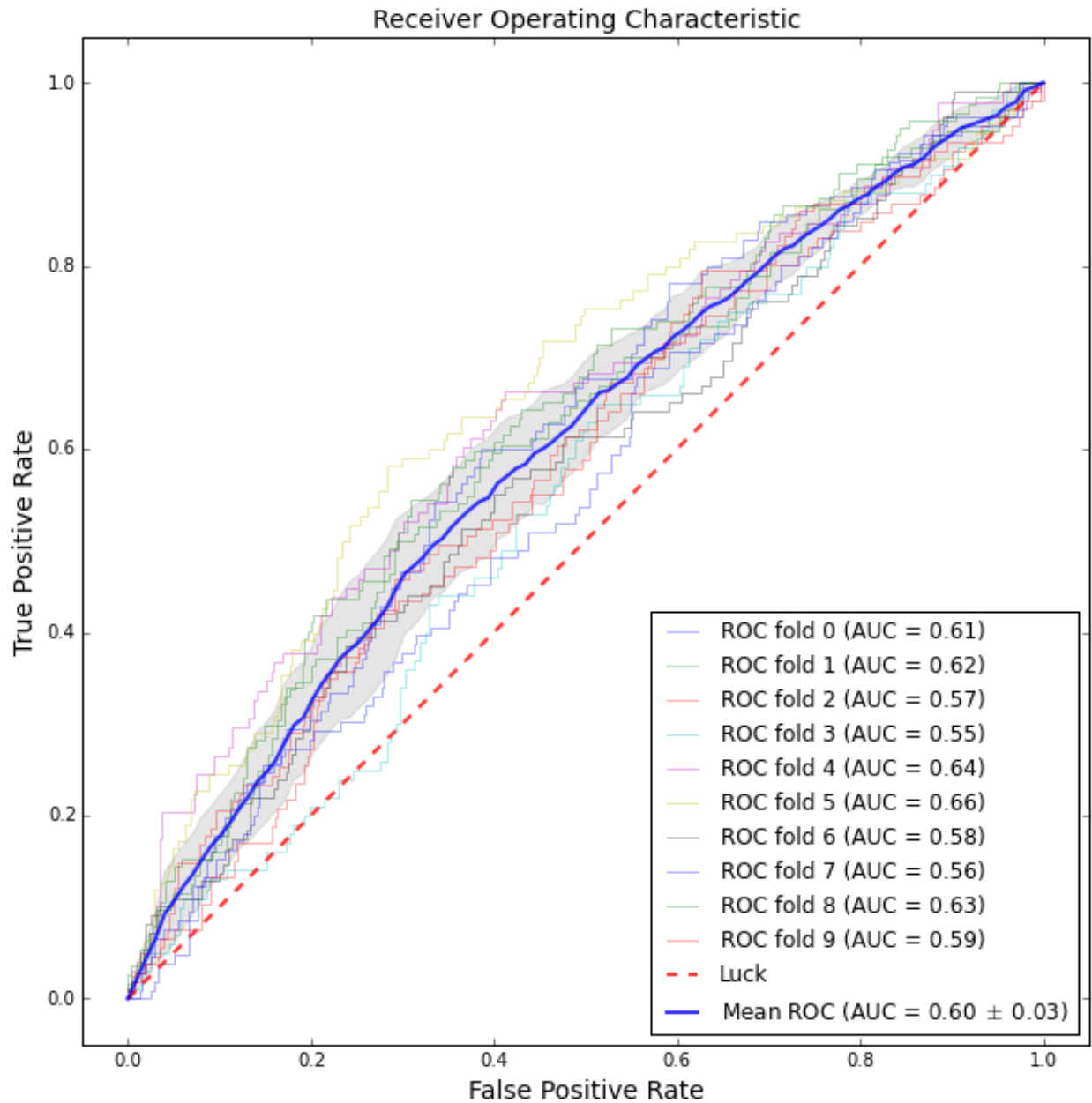


Figure 38. Logistic Regression ROC Curve, North Central TN University

The next recommendation is to explore the predictive models against the third research question relating to students who do not graduate from a 4-year institution in Tennessee.

Applying the same form of institution specific factor analysis to the second research question for a medium-sized 4-year public university in East Central Tennessee would result in output as shown in Table 25.

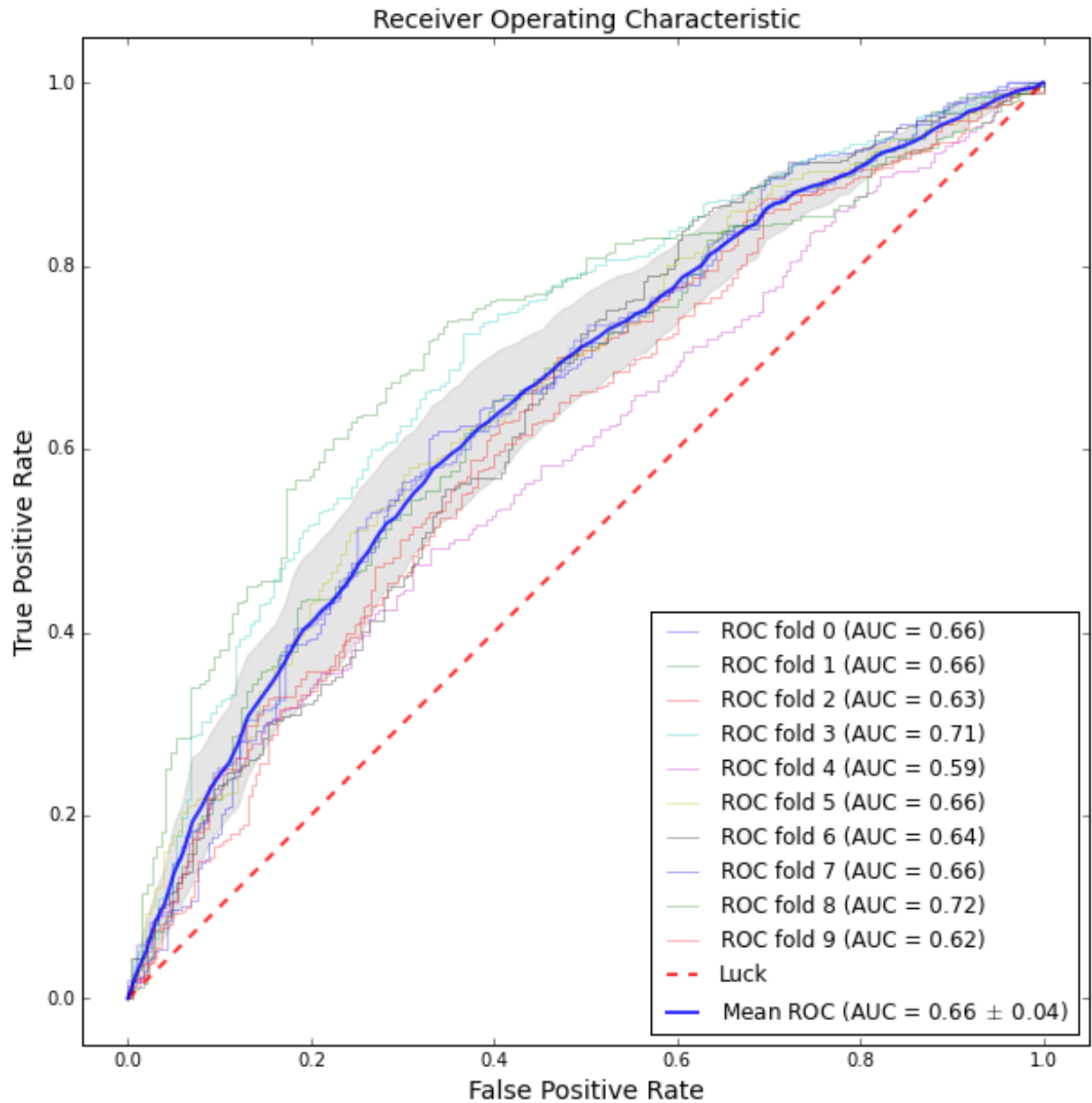
Table 25

*Factors for Nongraduates, East Central Tennessee University*

Component	1	2	3
Age			0.54
Distance from home to initial institution			0.54
High school GPA		0.47	
Composite ACT		0.70	
Number of credit hours taken in first semester		0.67	
Amount of state funds awarded			-0.67
Number of FTF students with same major	0.43	-0.49	
Number of full-time faculty at initial institution	-0.92		
Tuition and fees at initial institution	0.84		
Student FTE at initial institution	0.94		
Eigenvalues	2.70	1.47	1.09
Percentage of total variance	26.97	14.75	10.87
Number of test measures	4	4	3

\*Loadings  $\geq 0.5$ 

Running a logistic regression based on the factor scores from that institution specific factor analysis and the categorical variables of gender, race, and residency would produce an ROC curve as shown in Figure 39. For a single institution a logistic regression can correctly predict students who will not graduate from a 4-year institution in Tennessee between 62% and 70% of the time.



*Figure 39.* Logistic Regression ROC Curve, East Central Tennessee University

The next recommendation is to examine the research of others such as Delen (2010); Strecht et al. (2015); and Ding et al. (2016) to determine why decision trees, artificial neural networks, and support vector machines were not better predictive models for the data in this research. In particular, determining why logistic regression was better than decision trees and

artificial neural networks and comparable to support vector machines would be informative. A closer look or replication of their work could highlight deficiencies in the methods of this work.

Another recommendation is to include more categorical data points such as high schools and majors. These were not included in this research due to the complexity that creating dummy variables would have introduced. Including dummy variables for majors would have added 160 additional dimensions, while including dummy variables for high schools would have added 300 additional dimensions. The complexity could be reduced in future research by limiting the number of high schools and majors to examine. This reduction could be done through simple selection of the top 10 in each group, although this could lead to important latent information loss. Potdar, Pardawala, and Pai (2017) explored various ways of encoding categorical data that could be useful in reducing dimensionality. Potdar et al. used ordinal and binary encoding on nonordinal data and found that the resulting accuracy in a neural network model was similar to using one-hot encoding. One-hot encoding is an exchangeable term for dummy variable encoding for the purposes of this research. Binary encoding first converts data to an ordinal scale and then represents the ordinal values in a set of binary columns. Pasta (2009) argued that using ordinal data as though it were continuous is acceptable, and the results of Potdar et al. supported Pasta's argument. However, the use of ordinal encoding on nonordinal data imputes a distance between values that does not actually exist. This could lead to "spurious and meaningless findings" by the model (Larose & Larose, 2015, p. 341). Therefore, careful consideration for the encoding of majors and high schools should be explored in future research.

The next recommendation is to examine the sensitivity analysis conducted after each predictive model. The sensitivity analysis was meant to show which input variables had the most impact on the outcomes of the model. Intuitively, the four factor scores should have had the

most impact on the outcomes. The results of the Mann-Whitney U test on gender indicated that only the second factor varied significantly by gender. However, the sensitivity analysis indicated that gender had the most impact on the outcomes. In addition, race had a large impact on model outcomes according to the sensitivity analysis. The predictive models should be run again with data sets split out according to the sensitivity analysis.

The final recommendation is to supplement further research into predictive modeling of this data with qualitative research. This research sought to identify who would leave and graduate elsewhere, but it did not and could not identify why they left. Qualitative research in the form of surveys and case studies should be used to improve the understanding of the issue and improve the collection of behavioral data.



## REFERENCES

- Adelman, C. (1999). *Answers in the toolbox: Academic intensity, attendance patterns, and bachelor's degree attainment*. Washington, DC: U.S. Department of Education.
- Adelman, C. (2006). *The toolbox revisited: Paths to degree completion from high school through college*. Washington, DC: U.S. Department of Education.
- Aguiar, E. (2015). *Identifying students at risk and beyond: A machine learning approach* (Unpublished doctoral dissertation). University of Notre Dame, IN.
- Alpaydin, E. (2010). *Introduction to machine learning second edition adaptive computation and Machine Learning* (2nd ed.). Cambridge, MA: The MIT Press.
- Amatriain, X. (2013). Big and personal: Data and models behind Netflix recommendations. The 2nd international workshop on big data, streams and heterogeneous source mining: Algorithms, systems, programming models and applications. Chicago, IL: ACM.
- Astin, A. (1993). *What matters in college: Four critical years revisited*. San Francisco, CA: Jossey-Bass.
- Baars, G., & Arnold, I. (2014). Early identification and characterization of students who drop out in the first year at university. *Journal of College Student Retention: Research, Theory & Practice*, 16(1), 95-109.
- Balakrishnan, G., & Coetzee, D. (2013). *Predicting student retention in massive open online courses using hidden Markov models* (Technical Report No. UCB/EECS-2013-109). University of California at Berkeley, Electrical Engineering and Computer Sciences.
- Ballings, M., & Van den Poel, D. (2012). Customer event history for churn prediction: How long is long enough? *Expert Systems with Applications*, 39(18), 13517-13522. doi: <http://doi.org/10.1016/j.eswa.2012.07.006>
- Bartholomew, D. (1980). Factor analysis for categorical data. *Journal of the Royal Statistical Society Series B (Methodological)*, 42(3), 293-321.
- Baumol, W., & Bown, W. (1966). *Performing arts: The economic dilemma; a study of problems common to theater, opera, music, and dance*. New York, NY: Twentieth Century Fund.
- Bean, J. (1983). The application of a model of turnover in work organizations to the student attrition process. *The Review of Higher Education*, 6, 127-148.
- Bean, J. (2005). Nine themes of college student retention. In A. Seidman (Ed.), *College student retention: formula for success* (pp. 215-244). Lanham, MD: Rowman & Littlefield.

- Beattie, S., Woodley, C., & Souter, K. (2014). Creepy analytics and learner data rights. In B. Hegarty, J. McDonald & S.-K. Loke (Eds.), *Rhetoric and reality: Critical perspectives on educational technology - conference proceedings* (pp. 422–425). Dunedin, NZ: ASCILITE.
- Berger, J. (2001-2002). Understanding the organizational nature of student persistence: Empirically-based recommendations for practice. *Journal of College Student Retention*, 3, 3-21.
- Berger, J. & Milem, J. (2000). Organizational behavior in higher education and student outcomes. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. XV, pp. 268-338). New York: Agathon.
- Bogard, M., Helbig, T., Huff, G., & James, C. (2011). *A comparison of empirical models for predicting student retention*. [White paper]. Retrieved November 12, 2016, from [http://www.wku.edu/institutes/documents/comparison\\_of\\_empirical\\_models.pdf](http://www.wku.edu/institutes/documents/comparison_of_empirical_models.pdf)
- Brave, S., & Butters, R. (2011). Monitoring financial stability: A financial conditions index approach. Federal Reserve Bank of Chicago, *Economic Perspectives*, 35(1), 22-43.
- Braxton, J., Jones, W., Hirschy, A., & Hartley, H. (2008). The role of the classroom in college student persistence. *New Directions for Teaching and Learning*, 115, 71-83.
- Braxton, J., Hirschy, A., McClendon, S. (2004). Understanding and reducing college student departure. *ASHE-ERIC Higher Education Report*, 30(3). San Francisco, CA: Jossey-Bass.
- Bronfenbrenner, U. (1993). The ecology of cognitive development: Research models and fugitive findings. In R. H. Wozniak & K.W. Fischer (Eds.), *Development in context: Acting and thinking in specific environments* (pp. 3-44). Mahwah, NJ: Erlbaum.
- Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Application*, 36(3 Part 1), 4626-4636.
- Cabrera, A., Burkum, K., La Nasa, S. (2005). Pathways to a 4-year degree: Determinants of transfer and degree completion. In A. Seidman (Ed.), *College student retention: A formula for student success*. (pp. 155-214). Lanham, MD: Rowman & Littlefield.
- Campbell, C. & Mislavy, J. (2013). Students' perceptions matter: Early signs of undergraduate student retention/attrition. *Journal of College Student Retention: Research, Theory & Practice*, 14(4), 467-493. doi: <http://doi.org/10.2190/CS.14.4.c>
- Chawla, N. (2005). Data mining for imbalanced datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook* (pp. 853–867). doi: [http://doi.org/10.1007/0-387-25465-X\\_40](http://doi.org/10.1007/0-387-25465-X_40)

- Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(2002), 321-357.
- Chawla, N., Lazarevic, A., Hall, L., & Bowyer, K. (2003, September). SMOTE-boost: Improving prediction of the minority class in boosting. In *Seventh European Conference on Principles and Practice of Knowledge Discovery in Databases* (pp. 107-119). Dubrovnik, Croatia.
- Chen, H., Chiang, R., & Storey, V. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), 1165-1188.
- Chessell, M. (2014). *Ethics of big data and analytics*. IBM Corporation. Retrieved April 18, 2017 from [http://www.ibmbigdatahub.com/sites/default/files/whitepapers\\_reports\\_file/TCG%20Study%20Report%20-%20Ethics%20for%20BD&A.pdf](http://www.ibmbigdatahub.com/sites/default/files/whitepapers_reports_file/TCG%20Study%20Report%20-%20Ethics%20for%20BD&A.pdf)
- Coenen, F. (2004). Data mining: Past, present and future. *The Knowledge Engineering Review*, 00:0(1-24), 1–5. doi: <http://doi.org/10.1017/S0000000000000000>
- Coglianesse, C., & Lehr, D. (2017). Regulating by robot: Administrative decision making in the machine-learning era. *Faculty Scholarship*. 1734. Retrieved April 18, 2017 from [http://scholarship.law.upenn.edu/faculty\\_scholarship/1734](http://scholarship.law.upenn.edu/faculty_scholarship/1734)
- College Tuition Compare (2016). *2016 tuition, fees, and living costs comparison*. Retrieved November 13, 2016 from <http://www.collegetuitioncompare.com/compare/tables/?state=TN&degree=Undergraduate&type=Public&level=4-year%20or%20High>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. doi: <http://doi.org/10.1007/BF00994018>
- Coussement, K., Benoit, D., & Van den Poel, D. (2010). Improved marketing decision making in a customer churn prediction context using generalized additive models. *Expert Systems with Applications*, 37(3), 2132-2143. doi: <http://doi.org/10.1016/j.eswa.2009.07.029>
- CPI Inflation Calculator. (n.d.). Retrieved November 12, 2016, from <http://data.bls.gov/cgi-bin/cpicalc.pl>
- Crawford, G. (2015). The academic library and student retention and graduation: An exploratory study. *Portal: Libraries and the Academy*, 15(1), 41-57. <http://doi.org/10.1353/pla.2015.0003>
- Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods* (1<sup>st</sup> ed.). Cambridge, U.K.: Cambridge University Press.

- Cuseo, J. (2007). The empirical case against large class size: Adverse effects on the teaching, learning, and retention of first-year students. *Journal of Faculty Development*, 21(1), 5-21.
- Daniel, B. (2015). Big Data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology*, 46(5), 904–920.  
<http://doi.org/10.1111/bjet.12230>
- Davis, G. (1989). Sensitivity analysis in neural net solutions. *IEEE Transactions on Systems, Man, and Cybernetics*, 19, 1078-1082.
- Davis, J., & Goadrich. (2006, June). The relationship between precision-recall and ROC curves. In *Proceedings of the 23<sup>rd</sup> international conference on machine learning*. Pittsburgh, PA: ACM.
- Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4), 498–506. doi:10.1016/j.dss.2010.06.003
- Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention*, 13(1), 17-35. doi: 10.2190/CS.13.1.b
- De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and review of key research topics. *AIP Proceedings 1644*(1), 97-104. doi:  
<http://dx.doi.org/10.1063/1.4907823>
- Ding, S., Shi, Z., Tao, D., & An, B. (2016). Recent advances in support vector machines. *Neurocomputing*. doi: <http://doi.org/10.1016/j.neucom.2016.06.011>
- Dougherty, K., Jones, S., Lahr, H., Natow, R., Pheatt, L., & Reddy, V. (2014). Envisioning performance funding impacts: The espoused theories of action for state higher education performance funding in three states. (Working Paper No. 63). New York, NY: Community College Research Center.
- Dougherty, K., Natow, R., Bork, R., Jones, S., & Vega, B. (2013). Accounting for higher education accountability: Political origins of state performance funding for higher education. *Teachers College Record*, 115(1), 1-50.
- Dougherty, K. & Reddy, V. (2011). *The impacts of state performance funding systems on higher education institutions. Research literature review and policy recommendations*. New York, NY: Community College Research Center, Teachers College, Columbia University.
- Durango-Cohen, E., & Balasubramanian, S. (2015). Effective segmentation of university alumni: Mining contribution data with finite-mixture models. *Research in Higher Education*, 56(1), 78-104. doi: <http://doi.org/10.1007/s11162-014-9339-6>
- Durkheim, E. (1961). *Suicide*. Glencoe, IL: The Free Press.

- Fabrigar, L., Wegener, D., MacCallum, R., & Strahan, E. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272-299.
- Fairfax County Department of Neighborhood and Community Services (2012). *Common pitfalls in conducting a survey* [Brochure]. Fairfax, VA: Author.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <http://doi.org/10.1016/j.patrec.2005.10.010>
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 37–54. doi:10.1145/240455.240463
- Fuller, C. (2011). *The history and origins of survey items for the integrated postsecondary education data system*, (NPEC 2012-833). U.S. Department of Education. Washington, DC: National Postsecondary Education Cooperative. Retrieved November 12, 2016, from <http://nces.ed.gov/pubsearch>.
- Gansemer-Topf, A., & Schuh, J. (2006). Institutional selectivity and institutional expenditures: Examining organizational factors that contribute to retention and graduation. *Research in Higher Education*, 47, 613-641.
- Ghusson, M. (2016). *Understanding the engagement of transfer students in 4-year institutions: A national study*. (Unpublished doctoral dissertation). Seton Hall University, South Orange, NJ.
- Glass, J., & Harrington, A. (2002). Academic performance of community college transfer students and “native” students at a large state university. *Community College Journal of Research and Practice*, 26(5), 415-430.
- Goldrick-Rab, S., & Pfeifer, F. (2009). Beyond access: Explaining socioeconomic differences in college transfer. *Sociology and Education*, 82, 101-125.
- Gordon, G., & Hedlund, A. (2015). Accounting for the rise in college tuition. [White Paper]. Retrieved November 12, 2016, from National Bureau of Economic Research: <http://www.nber.org/chapters/c13711.pdf>
- Hamel, L. (2008). Model assessment with ROC curves. In *The Encyclopedia of Data Warehousing and Mining* (2nd ed.). Hershey, PA: Idea Group.
- Han, J., & Kamber, M. (2012). *Data mining: Concepts and techniques* (3<sup>rd</sup> ed.). Waltham, MA: Elsevier.
- Han, H., Wang, W., & Mao, B. (2005, March). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. *Proceedings of International Conference on Intelligent Computing*. doi: [http://dx.doi.org/10.1007/11538059\\_91](http://dx.doi.org/10.1007/11538059_91)

- Hanley, J., & McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29-36.
- Harsoor, A., & Patil, A. (2015). Forecast of sales of Walmart store using big data applications. *International Journal of Research in Engineering and Technology*, 4(6), 51-59.
- Hashemi, R., Le Blanc, L., Bahrami, A., Bahar, M., & Traywick, B. (2009). Association analysis of alumni giving: A formal concept analysis. *International Journal of Intelligent Information Technologies*, 5(2), 17-32.
- Heller, M., & Cassady, J. (2015). Predicting community college and university student success: A test of the triadic reciprocal model for two populations. *Journal of College Student Retention: Research, Theory & Practice*. doi: 10.1177/1521025115611130
- 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.
- Hillman, N., Tandberg, D., & Fryar, A. (2015). Evaluating the impacts of “new” performance funding in higher education. *Educational Evaluation and Policy Analysis*, 37(4), 501-519.
- Hills, J. (1965). Transfer shock: The academic performance of the junior college transfer. *Journal of Experimental Education*, 33, 201-216.
- Horn, A., & Lee, G. (2017). Evaluating the accuracy of productivity indicators in performance funding models. *Educational Policy*, 00(0), 1-32. doi: 10.1177/0895904817719521
- Howard, R., McLaughlin, G., & Knight, W. (2012). *The handbook of institutional research*. San Francisco, CA: Jossey-Bass.
- Hu, S., & St. John, E. (2001). Student persistence in a public higher education system: Understanding racial/ethnic differences. *The Journal of Higher Education*, 72(3), 265-286.
- Hunt, A., & Thomas, D. (2000). *The pragmatic programmer: From journeyman to master*. Reading, MA: Addison-Wesley.
- Hunter, M., & Linder, C. (2005). First-year seminars. In M. L. Upcraft, J. N. Gardner, B. O. Barefoot, & Associates, *Challenging and supporting the first-year student: A handbook for improving the first year of college* (pp. 275-291). San Francisco, CA: Jossey-Bass.
- IBM. (2016). *IBM SPSS Modeler*. Somers, NY: Author. “Integrated Postsecondary Education Data System: IPEDS Glossary”. (n.d.). National Center for Education Statistics. Revised

- August 2016. Retrieved November 12, 2016, from <https://surveys.nces.ed.gov/ipeds/VisGlossaryAll.aspx>
- IPEDS Data Center. (n.d.) Retrieved November 12, 2016 from <https://nces.ed.gov/ipeds/datacenter>
- Ishitani, T. (2008). How do transfers survive after “transfer shock”? A longitudinal study of transfer student departure at a 4-year institution. *Research in Higher Education*, 49, 403-419. doi:10.1007/s11162-008-9091-x
- Joachims, T. (2006). Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 217-226). Philadelphia, PA: Association for Computing Machinery.
- Johnson, N. (2012). The institutional cost of student attrition. *American Institutes for Research*. Retrieved November 13, 2016 from <http://www.deltacostproject.org/sites/default/files/products/Delta-Cost-Attrition-Research-Paper.pdf>
- Johnson, J. (2014). The ethics of big data in higher education. *IRIE International Review of Information Ethics*, 4(April 2013), 3–10.
- Johnson, J. (2017, February). *Structural justice in student analytics, or, the silence of the bunnies*. Paper presented at the annual meeting of the Eastern Sociological Society in Philadelphia, PA.
- Jones, S. (2012). Technology review: the possibilities of learning analytics to improve learner-centered decision-making. *The Community College Enterprise*, 18(1), 89–92.
- Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies*, 13(1), 61–72. doi:10.2478/cait-2013-0006
- Kelchen, R., & Stedrak, L. (2016). Does performance-based funding affect colleges’ financial priorities? *Journal of Education Finance*, 41(3), 302-321.
- Kelly, A., & Lautzenheiser, D. (2013). Taking charge: A state-level agenda for higher education reform. American Enterprise Institute. Retrieved November 12, 2016, from [https://www.aei.org/wp-content/uploads/2013/07/-taking-charge\\_103835525600.pdf](https://www.aei.org/wp-content/uploads/2013/07/-taking-charge_103835525600.pdf)
- Kim, J., & Rabjohn, J. (1980). Binary variables and index construction. In Karl F. Schuessler (Ed.), *Sociological Methodology* (pp. 120-159). San Francisco, CA: Jossey-Bass.
- Kirk-Kuwaye, C., & Kirk-Kuwaye, M. (2007). A study of engagement patterns of lateral and vertical transfer students during their first semester at a public research university. *Journal of the First-Year Experience & Students in Transition*, 19(2), 9-27.

- Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (pp. 1137–1143). San Francisco, CA: Morgan Kaufmann.
- Kramer, A., Guillory, J., & Hancock, J. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111(24): 8788-8790.
- Krebs, D., Berger, M., & Ferligoj, A. (2000). Approaching achievement motivation – comparing factor analysis and cluster analysis. *New Approaches in Applied Statistics*, 16, 147-171. Retrieved November 12, 2016, from <http://dk.fdv.uni-lj.si/metodoloskizvezki/mz16/default.htm>
- Laanan, F. (2001). Transfer student adjustment. *New directions for community colleges*, 114, 5-13. San Francisco, CA: Jossey-Bass.
- Larose, D., Larose, C. (2015). *Data mining and predictive analytics* (2nd ed.). Hoboken, NJ: John Wiley.
- Langley, P., & Simon, H. (1995). Applications of machine learning and rule induction. *Communications of the ACM*, 38, 55-64.
- Le Blanc, L., & Rucks, C. (2009). Data mining of university philanthropic giving: Cluster-discriminant analysis and Pareto effects. *International Journal of Educational Advancement*, 9(2), 64-82.
- Leachman, M., & Mai, C. (2014). *Most states still funding schools less than before the recession*. Washington, DC: Center on Budget and Policy Priorities.
- Leppel, K. (2002). Similarities and differences in the college persistence of men and women. *The Review of Higher Education*, 25(4), 433-450.
- Li, A. (2014). Performance funding in the states: An increasingly ubiquitous public policy for higher education. *Higher Education in Review*, 11(Online), 1-29. Retrieved November 12, 2016, from <http://sites.psu.edu/higheredinreview/wp-content/uploads/sites/36443/2016/02/Li-2014.pdf>
- Li, H., & Sun, J. (2012). Forecasting business failure: The use of nearest-neighbor support vectors and correcting imbalanced samples – evidence from the Chinese hotel industry. *Tourism Management*, 33(3), 622-634.
- Liao, S., Chu, P., & Hsiao, P. (2012). Data mining techniques and applications. A decade review from 2000 to 2011. *Expert Systems with Applications*, 39(12), 11303-11311. doi: <http://doi.org/10.1016/j.eswa.2012.02.063>



- Lobo, J., Jimenez-Valverde, A., & Real, R. (2007). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17, 145-151. doi: 1111/j.1466-8238.2007.00358.x
- Luan, J. (2002). Data mining and its applications in higher education. *New Directions for Institutional Research*, 2002(113), 17-36: doi: 10.1002/ir.35
- Liu, L. (2003, November). Using SAS in educational research. *Western Users of SAS Software 2003*. Retrieved November 13, 2016, from [http://www.lexjansen.com/wuss/2003/DataAnalysis/c-using\\_sas\\_in\\_educational\\_research.pdf](http://www.lexjansen.com/wuss/2003/DataAnalysis/c-using_sas_in_educational_research.pdf)
- Lu, B., Wang, X., Yang, Y., & Zhao, H. (2011). Learning from imbalanced data sets with a min-max modular support vector machine. *Frontiers of Electrical and Electronic Engineering*, 6(1), 56-71. doi:10.1007/s11460-011-0127-1
- Lund, O., Nielsen, M., Lundegaard, C., Kesmir, C., & Brunak, S. (2005). *Immunological bioinformatics*. Cambridge, MA: The MIT Press.
- Luperchio, D. (2009). *Data mining and predictive modeling in institutional advancement: How ten schools found success*. New York, NY: Council for the Advancement and Support of Education (CASE) and SPSS Inc.
- Macari, N. (1985). *Analysis of a machine learning algorithm*. (Unpublished master's thesis). Dresden University of Technology, Germany.
- Mattern, K., Marini, J., & Shaw, E. (2015). Identification of multiple non-returner profiles to inform the development of targeted college retention interventions. *Journal of College Student Retention: Research, Theory & Practice*, 17, 18-43.
- McCluskey, N. (2017). *Not just treading water: In higher education, tuition often does more than replace lost appropriations*. CATO Institute Policy Analysis February 15, 2017. No. 810. Retrieved February 26, 2017 from <https://www.cato.org/publications/policy-analysis/higher-education-tuition-lost-appropriations>.
- McCormick, A., & Sarraf, S., BrckaLorenz, A., & Haywood, A. (2009). *Examining the transfer student experience: Interactions with faculty, campus relationships and overall satisfaction*. Paper presented at the annual meeting of the Association for the Study of Higher Education in Vancouver, Canada.
- McGuire, S., & Belcheir, M. (2014). Transfer student characteristics matter. *Journal of College Student Retention*, 15(1), 37-48. doi: 10.2190/CS.15.1.c
- McLendon, M., & Hearn, J. (2013). The resurgent interest in performance-based funding for higher education. *Academe*, 99(6), 25-30.

- McLendon, M. Hearn, J. & Deaton, R. (2006). Called to account: Analyzing the origins and spread of state performance-accountability policies for higher education. *Educational Evaluation and Policy Analysis*, 28(1), 1-24.
- Mendoza, P., Malcolm, Z., & Parish, N. (2015). The ecology of student retention: Undergraduate students and the great recession. *Journal of College Student Retention: Research, Theory & Practice*, 16, 461-485.
- Monaghan, D., & Attewell, P. (2014, April). *Academic momentum at the gate: Does first-semester credit load affect postsecondary completion?* Paper presented at the meeting of the American Educational Research Association in Philadelphia, PA.
- Nandeshwar, A., & Chaudhari, S. (2009). *Enrollment prediction models using data mining*. Retrieved November 12, 2016, from [http://nandeshwar.info/wp-content/uploads/2008/11/DMWVU\\_Project.pdf](http://nandeshwar.info/wp-content/uploads/2008/11/DMWVU_Project.pdf)
- Nandeshwar, A., Menzies, T., & Nelson, A. (2011). Learning patterns of university student retention. *Expert Systems with Applications*, 38(12), 14984–14996. <http://doi.org/10.1016/j.eswa.2011.05.048>
- National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1978). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Bethesda, MD: The Commission.
- National Conference of State Legislatures. (2015, August). Performance-based funding for higher education. Retrieved November 12, 2016, from <http://www.ncsl.org/research/education/performance-funding.aspx>
- National Postsecondary Education Cooperative. (2009). Information required to be disclosed under the higher education act of 1965: Suggestions for dissemination (Updated) (NPEC 2010-831v2). Washington, DC: Coffey Consulting.
- National Student Clearinghouse (n.d.) Student tracker. Retrieved February 5, 2017 from <http://www.studentclearinghouse.org/colleges/studenttracker>
- Nichols, A. (2011). Developing 20/20 vision on the 2020 degree attainment goal: The threat of income-based inequality in education. Washington, DC: The Pell Institute. Retrieved November 12, 2016, from [http://www.pellinstitute.org/downloads/publicationsDeveloping\\_2020\\_Vision\\_May\\_2011.pdf](http://www.pellinstitute.org/downloads/publicationsDeveloping_2020_Vision_May_2011.pdf)
- NCHED Forms (n.d.) Retrieved November 12, 2016, from <http://analytics.northcarolina.edu/nched/html/download.html>
- Noland, B. (2011). *The West Virginia experience*. *National Cross Talk*, 19(12-13). Retrieved April 18, 2017 from <http://www.highereducation.org/crosstalk/ct0511/pf>

- Noland, B. (2006). Changing perceptions and outcomes: The accountability paradox in Tennessee. *New Directions for Higher Education*, 135(59-67). doi:10.1002/he.228
- Office of Strategic Research (n.d.) Retrieved November 12, 2016, from <http://www.mississippi.edu/research/admin.html>
- O'Loughlin, M. (2016, March 1). 'Drown the bunnies' president out at Mount St. Mary's. *CruX*. Retrieved April 18, 2017 from <https://cruxnow.com/church/2016/03/01/drown-the-bunnies-president-out-at-mount-st-marys/>
- Oztekin, A. (2016). A hybrid data analytic approach to predict college graduation status and its determinative factors. *Industrial Management & Data Systems*, 116(8). doi: <http://doi.org/10.1108/02635570710734262>
- Pascarella, E., Seifert, T., & Whitt, E. (2008). Effective instruction and college student persistence: Some new evidence. *New Directions for Teaching and Learning*, 115, 55-70.
- Pascarella, E., & Terenzini, P. (2005). *How college affects students. A third decade of research*. (2<sup>nd</sup> ed.). San Francisco, CA: Jossey-Bass.
- Pasta, D. (2009, March). Learning when to be discrete: Continuous vs. categorical predictors. *Proceedings of the SAS Global Forum*. Paper 248. Washington, DC:SAS.
- Peltier, G., Laden, R., & Matranga, M. (1999). Student persistence in college: A review of research. *Journal of College Student Retention*, 1(4), 357-375.
- Peter, K., & Forrest Cataldi, E. (2005). *The road less traveled? Students who enroll in multiple institutions* (NCES 2005-157). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.
- Pett, M., Lackey, N., & Sullivan, J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Thousand Oaks, CA: Sage.
- Pittman, K. (2008). *Comparison of data mining techniques used to predict student retention* (Unpublished doctoral dissertation). Nova Southeastern University, Fort Lauderdale, FL.
- Pleskac, T., Keeney, J., Merritt, S., Schmitt, N., & Oswald, F. (2011). A detection model of college withdrawal. *Organizational Behavior and Human Decision Processes*, 115(1), 85-98. doi: <http://doi.org/10.1016/j.obhdp.2010.12.001>
- Porter, S. (2002). Including transfer-out behavior in retention models: Using the NSC Enrollment Search data. *AIR Professional File*, 82, 1-16.

- Potdar, K., Pardawala, T., & Pai, C. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications* 175(4):7-9. doi: 10.5120/ijca2017915495
- Provost, F., & Fawcett, T. (2013). *Data science for business*. Sebastopol, CA: O'Reilly Media.
- Quinlan, J. (1986). Induction of decision trees. *Machine Learning*, 1, 81-106.
- Quinlan, J. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.
- Raisman, N. (2013). The cost of college attrition at 4-year colleges and universities. *Educational Policy Institute*. Retrieved November 13, 2016 from [http://www.educationalpolicy.org/publications/pubpdf/1302\\_PolicyPerspectives.pdf](http://www.educationalpolicy.org/publications/pubpdf/1302_PolicyPerspectives.pdf)
- Raju, D., & Schumacker, R. (2015). Exploring student characteristics of retention that lead to graduation in. *Journal of College Student Retention: Research, Theory & Practice*, 16(4), 563–591.
- Reason, R. (2003). Student variables that predict retention: Recent research and new developments. *NASPA Journal*, 40(4), 172-191.
- Reason, R. (2009). An examination of persistence research through the lens of a comprehensive conceptual framework. *Journal of College Student Development*, 50, 659-682. doi:10.1353/csd.0.0098
- Richards, N, & King, J. (2014). Big data ethics. *Wake Forest Law Review*, 2014. Retrieved April 18, 2017 from <http://ssrn.com/abstract=2384174>
- Robbins, S., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, 130, 261-288. doi: 10.1037/0033-2909.130.2.261
- Rutherford, A., & Rabovsky, T. (2014). Evaluating impacts of performance funding policies on student outcomes in higher education. *The ANNALS of the American Academy of Political and Social Science*, 655(1), 185-208. doi: 10.1177/0002716214541048
- Sanford, T. & Hunter, J. (2011). Impact of performance funding on retention and graduation rates. *Education Policy Analysis Archives*, 19(33), 1-27. doi: <http://dx.doi.org/10.14507/epaa.v19n33.2011>
- Shapiro, D., Wakhungu, P., Yuan, X., & Harrell, A. (2015). *Transfer and mobility: A national view of student movement in postsecondary institutions, Fall 2008 Cohort* (Signature Report No. 9). Herndon, VA: National Student Clearinghouse Research Center.
- Shaw, T. (2000). *An evaluation of Tennessee's performance funding policy at Walters State Community College* (Unpublished doctoral dissertation). Knoxville, TN.

- Skari, L. (2014). Community college alumni: Predicting who gives. *Community College Review*, 42(1), 23-40. doi: <http://doi.org/10.1177/0091552113510172>
- Soares, L. (2012). *The rise of big data in higher education*. Retrieved February 5, 2017, from <https://er.educause.edu/~media/files/article-downloads/erm1237.pdf>
- Spruill, N., Hirt, J., & Mo. Y. (2014). Predicting persistence to degree of male college students. *Journal of College Student Retention: Research, Theory and Practice*, 16(1), 25-48.
- State Higher Education Executive Officers (2016). *State-by-state wave charts. SHEF—State higher education finance FY15*. Retrieved April 15, 2017 from <http://www.sheeo.org/projects/shef-fy15>.
- Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J., & Abreu, R. (2015, June). A comparative study of classification and regression algorithms for modelling students' academic performance. *Proceedings of the 8th International Conference on Educational Data Mining* (pp. 392-395). Madrid, ES: Educational Data Mining Society.
- Strumpf, G., & Hunt, P. (1993). The effects of an orientation course on retention and academic standing of entering freshmen, controlling for the volunteer effect. *The Journal of the Freshman Year Experience*, 5(1), 7-14.
- Tamhane, A., Ikbal, S., Sengupta, B., Duggirala, M., & Appleton, J. (2014, August). Predicting student risks through longitudinal analysis. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1544-1552). New York, NY: Association for Computing Machinery.
- Tan, P., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Boston, MA: Pearson Addison-Wesley.
- Tandberg, D., & Hillman, N. (2014). State higher education performance funding: Data, outcomes, and policy implications. *Journal of Education Finance*, 39(3), 222-243.
- Tennessee Board of Regents. (2014). *Reverse transfer: Policies, procedures, and guidelines: 2:02:00:02*. Retrieved February 5, 2017, from <https://policies.tbr.edu/policies/reverse-transfer-policies-procedures-and-guidelines-0>
- Tennessee Higher Education Commission. (n.d.-a). *2015-2020 Outcomes based funding formula*. Retrieved November 12, 2016, from [https://www.tn.gov/content/dam/tn/thec/bureau/fiscal\\_admin/fiscal\\_pol/obff/2015-2020\\_Outcomes-based\\_Funding\\_Formula\\_Final\\_Website\\_-\\_101615.xlsx](https://www.tn.gov/content/dam/tn/thec/bureau/fiscal_admin/fiscal_pol/obff/2015-2020_Outcomes-based_Funding_Formula_Final_Website_-_101615.xlsx)
- Tennessee Higher Education Commission. (n.d.-b). *File Checklist*. Retrieved February 5, 2017 from <https://thec.ppr.tn.gov/THECSIS/FileCheck/FileCheckForm.aspx>

- Tennessee Higher Education Commission. (2012). *Public institution data dictionary*. Nashville, TN: Author.
- Tennessee Higher Education Commission. (2013). *2012-2013 Tennessee higher education fact book*. Nashville, TN: Author.
- Tennessee Higher Education Commission. (2014). *2013-2014 Tennessee higher education fact book*. Nashville, TN: Author.
- Tennessee Higher Education Commission. (2015). *2014-2015 Tennessee higher education fact book*. Nashville, TN: Author.
- Tennessee Higher Education Commission. (2016). *2015-2016 Tennessee higher education fact book*. Nashville, TN: Author.
- Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attribution. *Expert Systems with Applications*, *41*(2), 321-330. doi: <http://doi.org/10.1016/j.eswa.2013.07.046>
- The R Foundation. (n.d.) *The R project for statistical computing*. Retrieved November 12, 2016, from <https://www.r-project.org>
- Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, *45*, 89-125.
- Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition*. Chicago, IL: University of Chicago Press.
- Tinto, V. (2010). From theory to action: Exploring the institutional conditions for student retention. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research*, *25*, 51-89. New York, NY: Springer.
- Tobolowsky, B., & Cox, B. (2012). Rationalizing neglect: An institutional response to transfer students. *The Journal of Higher Education*, *83*(3), 389-410. doi: <http://doi.org/10.1353/jhe.2012.0021>
- Tross, S., Harper, J., Osher, L., & Kneidinger, L. (2000). Not just the usual cast of characteristics: Using personality to predict college performance and retention. *Journal of College Student Development*, *41*, 323-334.
- Turban, E., Sharda, R., & Delen, D. (2010). *Decision support and business intelligence systems* (9<sup>th</sup> ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

- Ubi, J., & Liiv, I. (2010, June). A review of student churn in the light of theories on business relationships. *Third International Conference on Educational Data Mining*, 329-330. Pittsburg, PA: International Educational Data Mining Society.
- Umbricht, M., Fernandex, F., & Ortagus, J. (2015). An examination of the (un)intended consequences of performance funding in higher education. *Educational Policy*. doi: 10.1177/0895904815614398
- Ward, J., Barker, A., (2013). Undefined by data: a survey of big data definitions. *arXiv preprint*. arXiv:1309.5821
- Wayne, A., & Young, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, 73(1), 89-122.
- White, J., & Massiha, G. (2016). The retention of women in science, technology, engineering, and mathematics: A framework for persistence. *International Journal of Evaluation and Research in Education*, 5(1), 1-8.
- Willingham, W. (1985). *Success in college: The role of personal qualities and academic ability*. New York, NY: College Entrance Examination Board.
- Wilson, S., Gore, J., Renfro, A., Blake, M., Muncie, E., & Treadway, J. (2016). The tether to home, university connectedness, and the Appalachian student. *Journal of College Student Retention: Research, Theory & Practice*. doi: 10.1177/1521025116652635
- Witte, R., & Witte, J. (2010). *Statistics (9th Ed.)*. Hoboken, NJ: John Wiley.
- Witten, I., & Frank, E. (2011). *Data mining: Practical machine learning tools and techniques*. Burlington, MA: Morgan Kaufmann.
- Woodfield, R., & O'Mahony, J. (2016). *Undergraduate student retention and attainment: Phase two overview report*. York, UK: The Higher Education Academy.
- Wu, X., Kumar, V., Quinlan, J., Ghosh, J., Yang, Q., Motoda, ... Steinberg, D. (2007). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14:1, pp. 1-37.
- Xu, L., & Chow, M. (2006). A classification approach for power distribution systems fault cause identification. *IEEE Transactions on Power Systems*, 21(1), 53-60.
- Yong, A., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, 9(2), 79-94. doi: 10.20982/tqmp.09.2.p079
- Zhou, Z., & Liu, X. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18, 63-77.

APPENDICES

APPENDIX A – Data Source from THEC

File Source	Variable Name	Notes
THEC Awards	UniqueID	Not used - Duplicate variable
THEC Awards	Award_System_Name	Not used
THEC Awards	Award_InstCode	
THEC Awards	Award_InstName	
THEC Awards	Award_Semester_Sequence	Not used
THEC Awards	Award_TermYear	
THEC Awards	Award_Degree	Not used
THEC Awards	Award_Level	Not used
THEC Awards	Award_CIP6	Not used
THEC Awards	Award_Major	Not used
Clearinghouse Enrollment	UniqueID	Not used - Duplicate variable
Clearinghouse Enrollment	NSC_CollegeCodeBranch	Not used - Duplicate variable
Clearinghouse Enrollment	NSC_CollegeName	Not used - Duplicate variable
Clearinghouse Enrollment	NSC_CollegeState	Not used - Duplicate variable
Clearinghouse Enrollment	NSC_2yr_4yr	Not used - Duplicate variable
Clearinghouse Enrollment	NSC_Public_Private	Not used - Duplicate variable
Clearinghouse Enrollment	NSC_SemesterSequence	Not used
Clearinghouse Enrollment	NSC_TermYear	Not used
Clearinghouse Enrollment	NCS_Semester_EnrollmentStatus	Not used
Clearinghouse Enrollment	NSC_ClassLevel	Not used
Clearinghouse Enrollment	NSC_6DigitCIP1	Not used
Clearinghouse Enrollment	NSC_EnrollmentMajor1	Not used - Similar to other variable
Clearinghouse Grads	UniqueID	Not used - Duplicate variable
Clearinghouse Grads	NSC_CollegeCodeBranch	Not used - Duplicate variable
Clearinghouse Grads	NSC_CollegeName	Not used - Duplicate variable
Clearinghouse Grads	NSC_CollegeState	Not used - Duplicate variable
Clearinghouse Grads	NSC_2yr_4yr	Not used - Duplicate variable
Clearinghouse Grads	NSC_Public_Private	Not used - Duplicate variable
Clearinghouse Grads	NSC_Grad_Semester_Sequence	Not used
Clearinghouse Grads	NSC_Grad_Term_Year	Not used
Clearinghouse Grads	NSC_DegreeTitle	Not used
Clearinghouse Grads	NSC_DegreeMajor1	Not used - Similar to other variable
Clearinghouse Grads	NSC_CIP1	Not used



<b>File Source</b>	<b>Variable Name</b>	<b>Notes</b>
Demographics and FTF Enrollment	UniqueID	Not used - Duplicate variable
Demographics and FTF Enrollment	Gender	
Demographics and FTF Enrollment	RaceName	
Demographics and FTF Enrollment	Birthyear	
Demographics and FTF Enrollment	PermZip	
Demographics and FTF Enrollment	PermStateCode	Not used - Similar to other variable
Demographics and FTF Enrollment	PermStateName	Not used - Similar to other variable
Demographics and FTF Enrollment	PermCountyCode	Not used - Similar to other variable
Demographics and FTF Enrollment	PermCountyName	Not used - Similar to other variable
Demographics and FTF Enrollment	ResidencyAndCitizenshipStatus	
Demographics and FTF Enrollment	OverallHSGPAGED	
Demographics and FTF Enrollment	ACT_HighSchoolCode	Not used - Similar to other variable
Demographics and FTF Enrollment	HighSchoolName	
Demographics and FTF Enrollment	HSCounty	Not used - Similar to other variable
Demographics and FTF Enrollment	ACTComposite	
Demographics and FTF Enrollment	ACTEnglish	Not used - Similar to other variable
Demographics and FTF Enrollment	ACTReading	Not used - Similar to other variable
Demographics and FTF Enrollment	ACTScience	Not used - Similar to other variable
Demographics and FTF Enrollment	ACTMath	Not used - Similar to other variable
Demographics and FTF Enrollment	ACTWriting	Not used - Similar to other variable
Demographics and FTF Enrollment	SATComposite	Not used
Demographics and FTF Enrollment	SATMath	Not used - Similar to other variable
Demographics and FTF Enrollment	SATVerbal	Not used - Similar to other variable
Demographics and FTF Enrollment	Ever_Pell_Eligible	Not used
Demographics and FTF Enrollment	FTF_System	Not used
Demographics and FTF Enrollment	FTF_InstCode	Not used - Similar to other variable
Demographics and FTF Enrollment	FTF_InstName	
Demographics and FTF Enrollment	FTF_Year	
Demographics and FTF Enrollment	FTF_SemesterSequence	Not used
Demographics and FTF Enrollment	FTF_Semester_CreditHours	
Demographics and FTF Enrollment	FTF_CIP6Digit	
Demographics and FTF Enrollment	FTF_MajorName	
Demographics and FTF Enrollment	In_THEC_Enrollment_File	Not used - Similar to other variable
Demographics and FTF Enrollment	In_THEC_Award_File	
Demographics and FTF Enrollment	In_NSC_Enrollment_File	Not used - Similar to other variable
Demographics and FTF Enrollment	In_NSC_Grad_File	
Demographics and FTF Enrollment	In_FAFSA_File	Not used - Similar to other variable

<b>File Source</b>	<b>Variable Name</b>	<b>Notes</b>
THEC Enrollment	UniqueID	Not used - Duplicate variable
THEC Enrollment	THEC_InstCode	Not used - Similar to other variable
THEC Enrollment	THEC_InstName	
THEC Enrollment	THEC_Semester_Sequence	
THEC Enrollment	Term_Year	Not used
THEC Enrollment	THEC_Term_CreditHours	
THEC Enrollment	THEC_CumulativeHoursEarned	
THEC Enrollment	THEC_MajorCIP6	
THEC Enrollment	THEC_MajorName	
THEC Enrollment	THEC_Semester_StudentLevel	Not used
FAFSA	Fall_Year	
FAFSA	UniqueID	Not used - Duplicate variable
FAFSA	PellGrantEligibility	Not used
FAFSA	LotteryEligibilityStatus	Not used
FAFSA	GrantEligibilityStatus	Not used
FAFSA	Total_TSAA_Pmt	
FAFSA	Total_Other_Pmts	

## APPENDIX B – Predictive Models Python Code

```
"""
FILE NAME: dissertation_model_eval.py
PURPOSE: Predictive Model Evaluation
AUTHOR: Joshua Whitlock
"""
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import metrics
from sklearn.metrics import roc_curve, auc
from sklearn.cross_validation import train_test_split
from sklearn.cross_validation import cross_val_score
from sklearn.model_selection import KFold
from imblearn.over_sampling import SMOTE
from sklearn.preprocessing import LabelEncoder
from SALib.analyze import sobol
from SALib.plotting.morris import horizontal_bar_plot, covariance_plot
from scipy import interp

def load_ALL_data ():
    # LOAD FILE, NAME COLUMNS
    input_file = "data.csv"
    df = pd.read_csv(input_file)
    df.columns = ['uniqueID', 'transfer_grad_indicator', 'gender',
                 'racename', 'residencyandcitizenshipstatus', 'ftf_instname',
                 'RQ4_2_1', 'RQ4_2_2',
                 'RQ4_2_3', 'RQ4_2_4']
    return df

def prepare_data (df):
    # PREPARE DATA
    for column in df.columns:
        if df[column].dtype == type(object):
            le = LabelEncoder()
            df[column] = le.fit_transform(df[column])

    # Create dummy variables - pd.get_dummies actually does
    # one-hot encoding. Need to add drop_first=True if you want
    # actual dummy variables
    gender = pd.get_dummies(df['gender'])
    race = pd.get_dummies(df['racename'])
    residency = pd.get_dummies(df['residencyandcitizenshipstatus'])
    ftf_instname = pd.get_dummies(df['ftf_instname'])
    df = pd.concat([df, gender, race, residency, ftf_instname], axis=1)
    X, y = df.iloc[:,5:], df.iloc[:, 1]
```

```

# File listing continued...

X['RQ4_2_1'] = X['RQ4_2_1'].astype(float)
X['RQ4_2_2'] = X['RQ4_2_2'].astype(float)
X['RQ4_2_3'] = X['RQ4_2_3'].astype(float)
X['RQ4_2_4'] = X['RQ4_2_4'].astype(float)
# Flatten y into a 1-D array
y = np.ravel(y)

X_scaled = preprocessing.scale(X)

return X_scaled, y, df

def evaluate_predictive_model(model, X, y, estimator):
# Evaluate the model by splitting into train and test sets
X = X.values
cv = KFold(n_splits=10)
tprs = []
aucs = []
accuracy_scores = []
f1_scores = []
precision_scores = []
recall_scores = []
confusion_matrices = []
prediction_probabilities = []
mean_fpr = (np.linspace(0,1,100))

plt.figure(figsize=(10,10))

i=0
for train, test in cv.split(X, y):
sm_set = SMOTE(random_state=0)
X_train, y_train = sm_set.fit_sample(X[train], y[train])
probas_ = model.fit(X_train, y_train).predict_proba(X[test])
preds_ = model.fit(X_train, y_train).predict(X[test])
# Compute ROC curve and area the curve
fpr, tpr, thresholds = roc_curve(y[test], probas_[ :, 1])
y_true=y[test]
y_pred=np.where(preds_ > 0.5, 1, 0)
prediction_probabilities.append(preds_[0:3936])
accuracy_scores.append(metrics.accuracy_score(y_true, y_pred))
f1_scores.append(metrics.f1_score(y_true, y_pred))
precision_scores.append(metrics.precision_score(y_true, y_pred))
recall_scores.append(metrics.recall_score(y_true, y_pred))
confusion_matrices.append(metrics.confusion_matrix(y_true, y_pred))
tprs.append(interp(mean_fpr, fpr, tpr))
tprs[-1][0] = 0.0

```

```

# File listing continued...
    roc_auc = auc(fpr, tpr)
    aucs.append(roc_auc)
    plt.plot(fpr, tpr, lw=1, alpha=0.3,
             label='ROC fold %d (AUC = %0.2f)' % (i, roc_auc))
    i += 1

plt.plot([0, 1], [0, 1], linestyle='--', lw=2, color='r',
         label='Luck', alpha=.8)

mean_tpr = np.mean(tprs, axis=0)
mean_tpr[-1] = 1.0
mean_auc = auc(mean_fpr, mean_tpr)
std_auc = np.std(aucs)

plt.plot(mean_fpr, mean_tpr, color='b',
         label=r'Mean ROC (AUC = %0.2f $\pm$ %0.2f)' %
             (mean_auc, std_auc),
         lw=2, alpha=.8)
std_tpr = np.std(tprs, axis=0)
tprs_upper = np.minimum(mean_tpr + std_tpr, 1)
tprs_lower = np.maximum(mean_tpr - std_tpr, 0)
plt.fill_between(mean_fpr, tprs_lower, tprs_upper, color='grey',
                 alpha=.2,
                 label=r'$\pm$ 1 std. dev.')

plt.xlim([-0.05, 1.05])
plt.ylim([-0.05, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic')
plt.legend(loc="lower right")
plt.show()

print "Mean Accuracy: " + str(np.mean(accuracy_scores))
print "Mean F1 score: " + str(np.mean(f1_scores))
print "Mean Precision: " + str(np.mean(precision_scores))
print "Mean Recall: " + str(np.mean(recall_scores))
print r'Mean ROC (AUC = %0.2f $\pm$ %0.2f)' % (mean_auc, std_auc)
print "Confusion Matrix: "
print np.sum(confusion_matrices, axis=0) / 10

print ""
print "Sensitivity analysis:"
y_sa = np.ravel(prediction_probabilities)
y_sa = y_sa[0:round_down(len(y_sa), 10)]
perform_sensitivity_analysis(X.values, y_sa)

def round_down(num, divisor):
    return num - (num%divisor)

```

```

# File listing continued...

def perform_sensitivity_analysis(X, Y):
    #bounds are the range of the factors
    problem = {
        'num_vars': 23,
        'names': ['F1 - Instit Chars', 'F2 - Focus on Acad',
                 'F3 - Stu Apt', 'F4 - Trans Comm',
                 'Female', 'Male', 'Alaskan',
                 'Am Indian', 'Asian', 'Black',
                 'Hispanic', 'Multi', 'Unknown',
                 'White', 'Foreign', 'Instate', 'Outstate',
                 'APSU', 'ETSU', 'MTSU', 'TSU', 'TTU', 'UOM'],
        'bounds': [[0.0, 1.0], [0.0, 1.0], [0.0, 1.0],
                   [0.0, 1.0], [0.0, 1.0], [0.0, 1.0], [0.0, 1.0],
                   [0.0, 1.0], [0.0, 1.0], [0.0, 1.0], [0.0, 1.0],
                   [0.0, 1.0], [0.0, 1.0], [0.0, 1.0], [0.0, 1.0],
                   [0.0, 1.0], [0.0, 1.0], [0.0, 1.0], [0.0, 1.0],
                   [0.0, 1.0], [0.0, 1.0], [0.0, 1.0], [0.0, 1.0]]
    }

    Si = morris.analyze(problem, X, Y, conf_level=0.95,
                        print_to_console=True,
                        num_levels=4, grid_jump=2, num_resamples=50)

    print 'Convergence index:', max(Si['mu_star_conf']/Si['mu_star'])

    fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(10,10))
    horizontal_bar_plot(ax1, Si, {}, sortby='mu_star', unit=r"trans_grad")
    covariance_plot(ax2, Si, {}, unit=r"trans_grad")

```

```

"""
File Name: predictive_model.py
Purpose: Run Predictive Models
Author: Joshua Whitlock
"""
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.neural_network import MLPClassifier
from sklearn import svm
import dissertation_model_eval as dme

# PREPARE DATA
df_input = dme.load_ALL_data()
X, y, df_output = dme.prepare_data(df_input)

# MODEL EVALUATION USING A VALIDATION SET
# Logistic Regression
model_LR = LogisticRegression()
dme.evaluate_predictive_model(model_LR, X, y, LogisticRegression())

# Decision Tree
model_DT = DecisionTreeClassifier()
dme.evaluate_predictive_model(model_DT, X, y, DecisionTreeClassifier())

# Random Forest
model_RF = RandomForestClassifier(n_jobs=10)
dme.evaluate_predictive_model(model_RF, X, y, RandomForestClassifier())

# Artificial Neural Network
# Instantiate a multilayer perceptron (neural network) model
model_ANN = MLPClassifier(hidden_layer_sizes=(30,30,30))
dme.evaluate_predictive_model(model_ANN, X, y, MLPClassifier())

# Support Vector Machine
model_SVM = svm.SVC(probability=True)
dme.evaluate_predictive_model(model_SVM, X, y, svm.SVC(probability=True))

```

## APPENDIX C – Sample SPSS Factor Analysis Code

```
* FILE NAME: RQ4_Prediction
* PURPOSE: PCA Code for Research Question 4, excluding
*           variables with values known.
*           only after the student's first semester.
* AUTHOR: Joshua Whitlock
FACTOR
  /VARIABLES
  Zftf_distance_from_home
  ZoverallHSGPAGED
  ZACTComposite
  ZFTF_Semester_credithours
  Zhs_peers_cnt
  Zftf_major_peers
  Zftf_full_time_faculty
  Zftf_instit_support
  Zftf_tuition_and_fees
  /MISSING PAIRWISE
  /ANALYSIS
  Zftf_distance_from_home
  ZoverallHSGPAGED
  ZACTComposite
  ZFTF_Semester_credithours
  Zhs_peers_cnt
  Zftf_major_peers
  Zftf_full_time_faculty
  Zftf_instit_support
  Zftf_tuition_and_fees
  /SELECT=transfer_grad_indicator(1)
  /PRINT UNIVARIATE INITIAL CORRELATION SIG DET KMO INV
        REPR AIC EXTRACTION ROTATION FSCORE
  /FORMAT BLANK(.4)
  /PLOT EIGEN ROTATION
  /CRITERIA MINEIGEN(1) ITERATE(100)
  /EXTRACTION PC
  /CRITERIA ITERATE(100)
  /ROTATION VARIMAX
  /METHOD=CORRELATION.
```



```
* FILE NAME: RQ_4
* PURPOSE: PCA Code for Research Question 4, including all data points
* AUTHOR: Joshua Whitlock
```

**FACTOR**

```
  /VARIABLES
Zftf_age
Zftf_distance_from_home
Ztransfer_distance_from_home
ZoverallHSGPAGED
ZACTComposite
ZFTF_Semester_credithours
Ztotal_FTF_TSAA_Payment
Ztotal_hours_at_ftf_inst
Ztotal_semesters_at_ftf_inst
Ztotal_semesters_after_ftf_inst
Zhs_peers_cnt
Zftf_major_peers
Zftf_major_changes
Zavg_term_credithours
Zftf_full_time_faculty
Zftf_full_time_nonfaculty
Zftf_tuition_and_fees
Zftf_state_approps
Zftf_instruction
Zftf_research
Zftf_acad_support
Zftf_stu_support
Zftf_instit_support
Zftf_total_fte
Zgrad_full_time_faculty
Zgrad_full_time_nonfaculty
Zgrad_tuition_and_fees
Zgrad_state_approps
Zgrad_instruction
Zgrad_research
Zgrad_acad_support
Zgrad_stu_support
Zgrad_instit_support
Zgrad_total_fte
  /MISSING PAIRWISE
  /ANALYSIS
```

```

* File listing continued...
Zftf_age
Zftf_distance_from_home
Ztransfer_distance_from_home
ZoverallHSGPAGED
ZACTComposite
ZFTF_Semester_creditshours
Ztotal_FTF_TSAA_Payment
Ztotal_hours_at_ftf_inst
Ztotal_semesters_at_ftf_inst
Ztotal_semesters_after_ftf_inst
Zhs_peers_cnt
Zftf_major_peers
Zftf_major_changes
Zavg_term_creditshours
Zftf_full_time_faculty
Zftf_full_time_nonfaculty
Zftf_tuition_and_fees
Zftf_state_approps
Zftf_instruction
Zftf_research
Zftf_acad_support
Zftf_stu_support
Zftf_instit_support
Zftf_total_fte
Zgrad_full_time_faculty
Zgrad_full_time_nonfaculty
Zgrad_tuition_and_fees
Zgrad_state_approps
Zgrad_instruction
Zgrad_research
Zgrad_acad_support
Zgrad_stu_support
Zgrad_instit_support
Zgrad_total_fte
/SELECT=transfer_grad_indicator(1)
/PRINT UNIVARIATE INITIAL CORRELATION SIG DET KMO INV
      REPR AIC EXTRACTION ROTATION FSCORE
/FORMAT BLANK(.4)
/PLOT EIGEN ROTATION
/CRITERIA MINEIGEN(1) ITERATE(100)
/EXTRACTION PC
/CRITERIA ITERATE(100)
/ROTATION VARIMAX
/METHOD=CORRELATION.

```

## VITA

### JOSHUA LEE WHITLOCK

Personal Data:      Date of Birth                      October 03, 1980  
                                 Place of Birth                      Martinsville, Virginia

Education:              Fieldale-Collinsville High School, Collinsville, Virginia  
                                 East Tennessee State University, Johnson City, Tennessee  
                                 B.S., Computer Science, Digital Media,  
                                 summa cum laude, May 2003  
                                 East Tennessee State University, Johnson City, Tennessee  
                                 M.S., Information Systems, May 2005  
                                 East Tennessee State University, Johnson City, Tennessee  
                                 M.B.A., Business Administration, May 2013  
                                 East Tennessee State University, Johnson City, Tennessee  
                                 Ed.D., Educational Leadership & Policy Analysis, May 2018

Professional  
Experience:              Applications Developer, MarketResearch.com, Rockville, Maryland, 2005-  
                                 2006  
                                 Programmer/Analyst 2, East Tennessee State University, Johnson City,  
                                 Tennessee, 2006-2007  
                                 Analyst 3, East Tennessee State University, Johnson City, Tennessee,  
                                 2007-2010  
                                 Database Administrator, East Tennessee State University, Johnson City,  
                                 Tennessee, 2010-2013  
                                 Director of Technical Systems for Enrollment Services, East Tennessee  
                                 State University, Johnson City, Tennessee, 2013-2014  
                                 Director of Institutional Research Applications and Data Systems, East  
                                 Tennessee State University, Johnson City, Tennessee, 2014 to  
                                 Present

Academic  
Experience:              Instructor – Math for Computer Science, East Tennessee State University,  
                                 College of Business and Technology, 2007-2008  
                                 Instructor – Unix Fundamentals, East Tennessee State University, College  
                                 of Business and Technology, 2013-2014

Professional  
Memberships:              Association of Institutional Researchers, 2015-Present

Certifications and  
Accomplishments:      Most Valuable Team Member of the Banner Student Team, 2009  
                                 Oracle Database SQL Certified Expert, 2010  
                                 Oracle Database 11g Administrator Certified Associate, 2010  
                                 President, Staff Senate, East Tennessee State University, 2014-2016  
                                 Staff Senator, East Tennessee State University, 2011-2018