Two Methodologies: How Well Can Universities Predict Retention

Tiffany Lynette Gregory

University of Mississippi

Follow this and additional works at: https://egrove.olemiss.edu/etd

Part of the Statistics and Probability Commons

Recommended Citation


https://egrove.olemiss.edu/etd/671

This Dissertation is brought to you for free and open access by the Graduate School at eGrove. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of eGrove. For more information, please contact egrove@olemiss.edu.
ABSTRACT

Student retention has been a long standing focus in higher education research with one of the earliest work dating back to 1937. Many researchers have proposed factors that affect a student’s decision to depart from the university without successfully completing a degree. It is important to not only research different attributes and characteristics that affect student departure but it is also important to study different statistical methodologies. With the advancement in technology, new methodologies such as the Classification and Regression Tree (CART) have proven to yield significant results in a variety of research fields. As these new statistical methodologies emerge, it is always worthwhile to compare the modern approaches with the longstanding classical statistical approaches. The present study utilized historical archived data in order to compare the performance of the Logistic Regression (LR) methodology with the CART methodology in predicting first-year retention for new freshmen at the University of Mississippi. It was found that the logistic regression method was more accurate than the CART methodology, with the overall accuracy of 83.3% and 82.6% respectively. However, the CART methodology was more specific than the logistic methodology, meaning that the CART model correctly predicted more students to not be retained. The logistic regression model failed to identify at-risk students. Note that 98% of the time the CART model and the logistic regression model yielded the same classification result. Among those 2% that the classification decision differed, the CART model was more accurate than the logistic model to predict non-retained
students. Thus using the prediction outcomes of the two methodologies in tandem of each other leads to more accurate results overall.
DEDICATION

This thesis is dedicated to everyone who aided me in the completion of this major life accomplishment either through inspiration or by service. I especially would like to thank, my parents, Rick and Judy for their continual commitment to the success of unlimited number of university students across the world, including myself. I want to thank them for being an amazing example of how to live life and for continuously providing me with the encouragement that I have needed throughout my life.
# LIST OF ABBREVIATIONS AND SYMBOLS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>American College Testing</td>
</tr>
<tr>
<td>CART</td>
<td>Classification and Regression Tree</td>
</tr>
<tr>
<td>cp</td>
<td>Complexity Parameter</td>
</tr>
<tr>
<td>CPC</td>
<td>College Preparatory Curriculum</td>
</tr>
<tr>
<td>FTFT</td>
<td>First-Time, Full-Time, Degree-Seeking Freshmen</td>
</tr>
<tr>
<td>GPA</td>
<td>Grade Point Average</td>
</tr>
<tr>
<td>HS</td>
<td>High School</td>
</tr>
<tr>
<td>IR&amp;A</td>
<td>The Office of Institutional Research and Assessment</td>
</tr>
<tr>
<td>LDA</td>
<td>Linear Discriminant Analysis</td>
</tr>
<tr>
<td>ln</td>
<td>Natural logarithm</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum Likelihood Estimation</td>
</tr>
<tr>
<td>MS</td>
<td>Mississippi</td>
</tr>
<tr>
<td>P</td>
<td>Probability</td>
</tr>
<tr>
<td>Res</td>
<td>Mississippi Resident</td>
</tr>
<tr>
<td>SAT</td>
<td>Scholastic Assessment Test</td>
</tr>
<tr>
<td>STEM</td>
<td>Science, Technology, Engineering, and Mathematics</td>
</tr>
<tr>
<td>UM</td>
<td>The University of Mississippi</td>
</tr>
<tr>
<td>α</td>
<td>Alpha</td>
</tr>
<tr>
<td>β</td>
<td>Beta</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

I want to express my sincere appreciation for my advisor and mentor, Dr. Xin Dang. I could not have completed this thesis without her continual support, guidance, and patience. I also want to especially thank my co-worker, mentor, and friend, Brenda Wimberly, for aiding me in the extraction of the data that were utilized in this research. Additionally, I want to thank all the other co-workers in the Office of Institutional Research and Assessment for their continual support throughout this process. Last, but not least, I would like to thank Coulter Ward for his persistent support and encouragement throughout this entire process.
# TABLE OF CONTENTS

ABSTRACT .................................................................................................................................... ii

DEDICATION ................................................................................................................................ iv

LIST OF ABBREVIATIONS AND SYMBOLS ........................................................................... v

ACKNOWLEDGEMENTS .......................................................................................................... vi

LIST OF TABLES ...................................................................................................................... viii

LIST OF FIGURES .................................................................................................................. x

INTRODUCTION .................................................................................................................. 1

BACKGROUND .................................................................................................................... 4

METHODOLOGY .................................................................................................................. 21

RESULTS .............................................................................................................................. 37

DISCUSSION ......................................................................................................................... 59

REFERENCES ....................................................................................................................... 62

APPENDIX A ......................................................................................................................... 71

APPENDIX B ......................................................................................................................... 81

VITA ........................................................................................................................................ 87
LIST OF TABLES

Table 1. Six Year Graduation Rates for 4 Year Public Institutions in the US, Gender and Ethnicity (White and Black Only) ................................................................. 13

Table 2. Retention and Graduation for New Freshmen Cohorts at UM, Gender Differences..... 13

Table 3. Retention and Graduation Rates for New Freshmen Cohorts at UM, Ethnicity Differences ........................................................................................................... 15

Table 4. Generic Contingency Table .................................................................................. 28

Table 5. High School GPA Threshold Comparison for Residency/Gender Combinations and School/Colleges ............................................................................. 41

Table 6. Overall and Sub-Model Variable Coefficient and Significance ............................... 42

Table 7. Cross-Validation and Accuracy Summary of Training and Testing Groups .......... 43

Table 8. Contingency Tables - "Before" Logistic Regression Models .................................. 43

Table 9. Specificity and Sensitivity of the Sub-models – “Before” Logistic Regression Models 43

Table 10. 1st Semester College GPA Threshold Comparison for Residency and School/Colleges ........................................................................................................ 47

Table 11. Overall and Sub-Model Variable Coefficient and Significance ............................... 47

Table 12. Cross-Validation and Accuracy Summary of Training and Testing Groups .......... 48

Table 13. Contingency Tables – “After” Logistic Regression Models ................................... 48

Table 14. Specificity and Sensitivity of the Sub-models – “After” Logistic Regression Models 49

Table 15. Overall and Sub-Model Variable Coefficient and Significance ............................... 52
Table 16. Cross Validation and Accuracy Summary of Training and Testing – “After” CART Models........................................................................................................................................... 55
Table 17. Contingency Tables – “After” CART Models............................................................. 55
Table 18. Specificity and Sensitivity of the Sub-models – “After” CART Models .................... 56
Table 19. Contingency Tables - Combined Outcomes Approach for the Matched Classification Conclusions................................................................................................................................... 57
Table 20. Contingency Tables – Combined Outcomes of the Logistic and CART Models – Mismatched Conclusions, using CART Model Only .......................................................... 57
Table 21. Contingency Tables – Combined Outcomes of the Logistic and CART Models – Mismatched Conclusions, using Logistic Regression Model Only .............................................. 58
LIST OF FIGURES

Figure 1. Relationship between Probabilities, Odds Ratios, and Log Odds ....................... 25
Figure 2. Generic Logistic Curve ....................................................................................... 26
Figure 3. Generic Classification Tree ................................................................................. 31
Figure 4. Generic Partitioning Visualization Example ...................................................... 33
Figure 5. Generic Visualization of Complexity Parameter .................................................. 34
Figure 6. For Every 0.05 Point Change in High School GPA, Percent Change in the Odds of
Being Retained .................................................................................................................... 38
Figure 7. For Every Unit Change in ACT, Percent Change in the Odds of Being Retained..... 39
Figure 8. Residency, Gender, and School/College Probability of Retention ....................... 40
Figure 9. For Every Unit Change in 1st Semester College GPA, Percent Change in the Odds of
Being Retained .................................................................................................................... 45
Figure 10. Residency and School/College Probability of Retention ..................................... 46
Figure 11. Overall Tree - Unpruned .................................................................................... 50
Figure 12. Overall Tree - Pruned ....................................................................................... 50
Figure 13. Complexity Parameter Evaluation ...................................................................... 51
Figure 14. Sub-Tree 1 - 1234 ............................................................................................. 53
Figure 15. Sub-Tree 2 - 1235 ............................................................................................. 53
Figure 16. Sub-Tree 3 - 1245 ............................................................................................. 54
Figure 17. Sub-Tree 4 - 1345 ............................................................................................. 54
Figure 18. Sub-Tree 5 - 2345 ............................................................................................. 54
CHAPTER 1

INTRODUCTION

An abundance of research has been conducted in the higher education realm on the success of university students, especially relating to retention and graduation. Researchers and higher education practitioners have tried to determine the reasons some students depart from the university while their peers not only stay but end up graduating from the university. Some have proposed theoretical models of student departure to illuminate when and why a student departs.

The implications of a good model for prediction of student retention and graduation would potentially affect universities’ admission policies, advising policies, retention rates, academic performance metrics, graduation rates, all of which impact the overall public reputation of the university. Tinto (2006) states that “What is needed and what is not available is a model of institutional action that provides guidelines for the development of effective policies and programs that institutions can reasonably employ to enhance the persistence of all their students” (p. 6). It is imperative for the success of universities to approach problems such as retention and graduation proactively.

It is extremely important for researchers and universities to continue their endeavors in trying to uncover the best practices for specifically identifying and ultimately providing support for their at-risk student populations. If a student departs from a university, there are losses on many levels (student level, university level, and societal level). Veenstra (2009), in an article
written about strategic efforts for improving freshmen college retention, claims that if a student is not retained in the current university or transfers to another university then it creates a loss to not only the university in terms of investment but also a loss to society (p. 21).

Some studies utilize questionnaires and surveys that evaluate aspects of a student such as personality traits (Moses et al., 2011; Pidcock, Fischer, & Munsch, 2001; McLaughlin, Moutray, & Muldoon, 2008), perceived academic stress (Daugherty & Lane, 1999; Perrine, 1998), emotional intelligence (Parker, Hogan, Eastabrook, Oke, & Wood, 2006), student involvement (Baker & Robnett, 2012), and social risk factors (Pidcock, Fischer, & Munsch, 2001). They utilize the responses from the various instruments in order to see how these attributes affect the success of college students. Most of these studies also consider pre-college performance metrics as well as demographic information.

Researchers not only need to consider the impact of new and different variables, but they also need to compare different types of statistical modeling techniques in order to achieve the best possible results. A byproduct of the rapid advancement in technology is that newer statistical methodologies have been developed that can be applied to many different scenarios and they may be more accurate than classical statistical methodologies. This study will hopefully aid the University in a proactive step to identify at-risk students by providing a comparison of multiple classification methodologies using student attributes that have been found to be significant.

Problem Statement

The purpose of this study is to compare the differences as well as the prediction accuracy of classical and modern statistical classification procedures by utilizing pre-college
characteristics, demographics, and first-semester college GPA to predict first-year retention at
The University of Mississippi. Some research questions that will be proposed:

(1) What is the impact of demographics and previous high school academic performance
indicators on first-year retention?

(2) What is the impact of demographics, previous high school academic performance
indicators, and first-semester college GPA on first-year retention?

(3) What is the strongest predictor of first-year retention?

Each of these research questions will be analyzed by both the classical (i.e. Logistic Regression)
and modern (i.e. Classification and Regression Tree) statistical methodologies in order to
compare the differences and the effectiveness of each approach.
CHAPTER 2

BACKGROUND

Generally academic success is measured by graduation, but in order to achieve graduation a student must be retained from one year to the next. Davis (2008) wrote in an article in the Chronicle of Higher Education that,

“Both senior administrators and governing boards should also learn to base decisions on indicators, as opposed to lagging indicators. Graduation rates are a lagging indicator. Retention rates are a leading indicator. Do we really want to wait six years to recognize and then eloquently explain low graduation rates, or do we want to focus on annual retention and nip problems in the bud? In short, college administrators must be taught to use leading indicators to alter future conditions. That is the essence of what leaders do – manage risks and alter futures” (p. A64).

If one follows Davis’s opinion, retention is an important issue that must be studied.

There are a vast number of ways researchers and practitioners can explore problems of interest such as retention. Using multiple methodologies on a single research interest and comparing them may yield some interesting findings, not only in the results about the significant variables themselves but also in the findings about the different modeling techniques. With development of modern statistical techniques, it is important to compare the effectiveness of new and longstanding classical statistical techniques.

In statistics there are several classification methods including some that are extremely popular such as linear discriminate analysis (LDA) and logistic regression (LR). LDA is method that attempts to classify data into two or more groups based on a linear combination of
independent variables, which is similar in idea to the logistic regression methodology. “Both of them are appropriate for the development of linear classification models, i.e. models associated with linear boundaries between the groups. Nevertheless, the two methods differ in their basic idea. While LR makes no assumptions on the distribution of the explanatory data, LDA has been developed for normally distributed explanatory variables” (Pohar, Blas, & Turk, 2004, p. 144). Logistic regression, which is explained in detail in the methodology section, was chosen over LDA as the classical approach for this research. This choice was made in part because of the underlying assumptions of the explanatory data of LDA which potentially makes this approach less general than the logistic regression method. Another reason why the logistic regression technique was chosen is that many fields, including the higher education realm, use this classification technique and practitioners are potentially more familiar with this approach.

Due to the influx of the advancement in technology there are many choices for modern classification approaches, including but not limited to Classification and Regression Trees (CART), Random Forests, Artificial Neural Networks (ANN), and Support Vector Machines (SVMs). CART, which will be explained in detail in the methodology section, is a method that recursively partitions data into two binary subgroups until the data cannot be partitioned anymore. After all possible partitions have been made, terminal nodes are established which is then classified according to majority voting in the node. “A classification tree is the result of asking an ordered sequence of questions, and the type of questions asked at each step in the sequence depends upon the answers to the previous questions of the sequence. The sequence terminates in a prediction of the class” (Izenman, 2008, p. 282). Random Forests is an “ensemble” methodology, where many classification trees are used to predict the final classification of the observation. “The classifier predicts the class of that observation by that
class that enjoys the largest number of total votes over all of the trees” (Izenman, 2008, p. 536). Artificial Neural Networks is a methodology that tries to emulate the human mind when it comes to processing data and to get a decision. This methodology uses “a network of highly interconnected nonlinear computing elements” in order to make decisions based on all the input data (Izenman, 2008, p. 316). Support Vector Machines methodology determines an optimal boundary (either linear or nonlinear) between the two classes of the data. This boundary is established in order to maximum the distance between the support vectors of either class (Izenman, 2008).

CART was chosen as the modern classification approach because the structure of a binary decision tree is intuitively interpretable and user friendly. It has also been found to provide very accurate results in a wide array of fields of research including but not limited to banking research (Emrouznejad & Anouze, 2010), psychology (Poulsen, Johnson, & Ziviani, 2011), ecological research (De’ath & Fabricius, 2000), environmental hazards research (Vega, Matias, Andrade, Reigosa, & Covelo, 2009), ocean research (Mahjoobi & Etemad-Shahidi, 2008), cardiac research (Quantin, et al., 2011), diabetes research (Goel, et al., 2009), cancer research (Barlin, et al., 2013), and epidemiology (Marshall, 2001; Porter, 2011).

In order to best predict retention of university students, a consideration of significant variables is mandatory. A multitude of theorists and researchers have proposed, tested and evaluated different theoretical models for student success and departure. As stated in the book *College Student Retention: Formula for Student Success*, “one of the earliest studies of student attrition was conducted by McNeely (1937). Specifically, McNeely was interested in determining the extent to which students withdrew from college and the factors responsible for such behavior” (Seidman, 2012, p. 63). Retention research has a long standing history but some
notable theorists on student retention include Alexander Astin in 1975 and again in 1985, Vincent Tinto in 1987, and Ernest Pascarella in 1980. All three of these theorists have proposed different models containing various aspects to explain student departure from higher education institutions. However, all three of these theorists proposed models that considered inputs such as demographics, student background, and pre-college attributes. Other areas of consideration include institutional characteristics, commitment to institutions and personal goals, and academic and social integration. They propose that some, if not all, of these areas of interest impact a student’s retention to the university (Seidman, 2012).

One aspect of the theoretical models on student departure that are focused on in this research, due in part to the availability of data, are students’ entry characteristics both in terms of demographics and academic abilities. The other reason for only utilizing these variable categories is a timing issue. The earlier a student can be identified as “at-risk for departure”, the more time the practitioners can have to provide an intervention strategy to hopefully prevent departure. One final note for limiting the research to these data metrics is that no additional monetary funds and resources are required since all of these variables are available during the admission process to the University and the first semester.

Colleges and universities have utilized previous academic performance within their admissions standards throughout history. Pre-college academic measurements included in admission policies range from high school GPA, high school rank, advanced placement credits, to standardized scores such as the SAT and ACT. These factors provide universities some insights to the future performance of their prospective students by past academic achievements. In a study of predicting academic performance, it was concluded that past performance academically was a significant predictor for future college performance (Elias & MacDonald,
2007, p. 2526). The University of Mississippi uses a combination of high school GPA on the
College Preparatory Curriculum (CPC) as well as standardized scores such as the ACT or the
SAT (The University of Mississippi Office of Admissions, 2013, p. 3).

Pre-College Characteristics

Predicting human behaviors is a convoluted science in all realms of life and entire fields
are devoted to studying human behaviors. When researching human behaviors it is a natural
inclination to investigate past behaviors. As part of past behaviors, previous academic
experience may shed light on future academic performance. It is then reasonable that there are
admission policies for universities that include submitting records of past academic performance
whether it is high school grade point average, results of standardized exams, or other metrics.
Most all research on student retention has some sort of consideration to past academic
performance.

High school academic metrics such as high school rank (Scott, Tolson, & Huang, 2009,
p. 23) and high school GPA (Bowen, Chingos, & McPherson, 2009, p. 113-114; Murtaugh,
Burns, & Schuster, 1999, p. 369; Rohr, 2013; Daughtrey & Lane, 1999, P. 359; Moses et al.,
2011, p. 240; D’Amico & Dika, 2013-2014, p. 181) have been found to be statistically
significant in predicting academic success of university students.

High School Grade Point Average

In research conducted by Adebayo (2008), there was a fairly positive correlation between
first-semester college GPA and high school GPA (p. 19). In the book, Crossing the Finish Line,
the researchers state that, “High school grades are a far better predictor of both four-year and six-
year graduation rates than are SAT/ACT test scores—a central finding that holds within each of
the six sets of public universities that we study” (Bowen, Chingos, & McPherson, 2009, p. 113-
In a study of predicting the retention of university students, the researchers stated that they were surprised by the results that high school GPA had a “superior predictive value” over SAT scores (Murtaugh, Burns, & Schuster, 1999, p. 369). In a retention study of STEM and business students, it was determined that “college preparatory GPA was found to be a significant predictor of retention of science, technology, engineering, mathematics, and business students” (Rohr, 2013, p. 204). The researchers determined that “for every point increase in GPA, the odds were more than twice as much that student would be retained” in the STEM or business fields (Rohr, 2013, p. 195). Daughtrey and Lane (1999) also found this relationship of lower academic metrics, specifically secondary school GPA and SAT scores, “were associated with increased vulnerability to attrition” (p. 359). In a research study on retention in Engineering in college, “the scores from the ALEKS and high school GPA did add significantly to the model” where “ALEKS is a measure specific to calculus readiness” (Moses et al., 2011, p. 240). Some schools use a transformed high school academic performance measure in their admission procedures. Researchers then utilized this score as a proxy measurement of previous academic achievement. The PGPA utilizes SAT scores and weighted grades in high school courses and was found to be a significant predictor of retention for both First Generation College Students (FGCSs) and non-FGCSs (D’Amico & Dika, 2013-2014, p. 181). Utilizing high school GPA is an important variable to investigate.

High School Rank

In a research study conducted by Scott, Tolson, & Huang (2009), that looked specifically at a subgroup of students that were enrolled in a Math, Engineering, or Science degree program (STEM related fields), the researcher found that by using three pre-college characteristics (high school rank, SAT verbal and SAT math scores) “could be used to correctly place 75.5 percent of
the students into the appropriate student group (students retained to math and science v. those who changed from math and science with a GPA less than 2.0 at the time of change)” (p. 23). In a graduation study, it was found that “students with higher high school rank (i.e., better students in high school) were significantly more likely (p=.001) to graduate or be retained compared to students with lower high school rank” (Whalen, Saunders, & Shelley, 2010, p. 420). These researchers utilized high school rank as a proxy for academic motivation. Including high school rank may be important.

*College Credit in High School and College Preparatory Curriculum*

Many high school students have started to enroll in college courses for both high school and college credit. Some high schools across the nation partner with colleges, universities, community colleges, and junior colleges to offer college level credit to high school students. Some high school students enroll in college or community college courses during summer sessions or even during the regular academic year in lieu of or in addition to their regular course load in high school. It was determined that completing college preparatory curriculum led to a “1.16 times increase in odds of persistence” in college (Johnson, 2008, p. 788). Results from a study conducted by Allen and Dadgar (2012), suggested “that completing one or more College Now duel enrollment courses is associated with positive and substantial gains including earning more credits during the first semester of college and a higher college GPA” (p. 15). They also found that even when controlling for demographics, high school GPA, test scores, and specific high school that they attended that these results still remained true as well as increasing the chances of retention in future semesters. They determined that demonstrates that “taking one or more College Now credit-bearing class is associated with almost one additional credit earned during the first semester, 0.16 points higher GPA in the first semester, and 5 percentage points
greater likelihood of reenrolling in the third semester” (p. 15). Research has suggested that college credit attainment prior to college has an effect on future college performance.

**Testing Scores from Standardized Exams**

Traditional variables used to predict success in college are standardized exams such as the SAT or ACT as well as high school grade performance. Some of these traditional variables are used in admission standards across the nation in higher education. “As the nation’s most widely used college admission test, the SAT is the first step toward higher education for students of all backgrounds. It’s taken by more than two million students every year and is accepted by virtually all colleges and universities” (The College Board, 2013). The Scholastic Assessment Test (SAT) was first introduced in 1901. The SAT has been adjusted throughout the years. As this test evolves, research continues to be completed to test the validity and reliability of the SAT as it relates to future academic success in college.

Likewise, the American College Testing (ACT) is also a standardized test that sheds light on academic and mental abilities. The ACT was first administered in November of 1959 to offer students an alternative standardized college admissions test to the SAT (ACT, Inc., 2013). In most research as well as most institutions that utilize standard test scores as part of the admissions policies, ACT and SAT scores are recalibrated using standard concordance tables so they are used interchangeably. These concordance tables have been constructed by the collaboration of the College Board and ACT through research. The most current concordance tables were established in 2006 (The College Board, 2013).

Research completed in 2007, after the most recent adjustment of the SAT exam, showed that using both high school GPA and SAT scores provides the best combination to predict first-year cumulative college GPA (Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008, p. 1). Other
research has found that “lower SAT scores and secondary school GPA’s were associated with increased vulnerability to attrition” (Daugherty & Lane, 1999, p. 359). In a retention research study on STEM and Business students, the researcher stated that “the SAT was found to be a significant predictor of retention” (Rohr, 2013, p. 204). In a research study on social network related to retention it was determined that along with attrition scores and retention scores, SAT scores were significant in predicting retention. The researchers did say that the impact was “so small as to be negligible” (Eckles & Stradley, 2012, p. 177).

Demographics

Demographics are another aspect of a student that is widely considered within much of the research. Studies that have included demographics such as gender (Johnson, 2008; Pidcock, Fischer, & Munsch, 2001, p. 812), ethnicity (Murtaugh, Burns & Schuster, 1999, p. 368; D’Amico & Dika, 2013-2014, p. 181), socio-economic background (Johnson, 2008; Bowen, Chingos, & McPherson, 2009), residency status (Johnson, 2008; Whalen, Saunders, Shelley, 2010) and parental educational achievements (D’Amico & Dika, 2013-2014, p. 186; Johnson 2008; Nandeshwar, Menzies, & Nelson, 2011, p. 14994) have found statistical difference between each category.

Gender

Gender differences in persistence and in graduation have been found to exist. Johnson (2008) highlighted an interesting trend that females are more likely than males to persist to the second year. They also determined that if the females do not return to the university, it is not usually due to factors that are related to academic performance (p. 788). In another research study that compared Hispanics with their Anglo counterparts it was concluded that “Hispanic females left school at the highest rate whereas Hispanic males stayed in school at the second
highest rate. It appeared that Hispanic females represented a group that is at particular risk to leave college” (Pidcock, Fischer, & Munsch, 2001, p. 812).

Nationally, as shown in Table 1, there is a difference nationally in six-year graduation rates between males and females. Females have higher graduation rates than males.

Table 1. Six Year Graduation Rates for 4 Year Public Institutions in the US, Gender and Ethnicity (White and Black Only)

<table>
<thead>
<tr>
<th>Year</th>
<th>Total</th>
<th>White</th>
<th>Black</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>54.8%</td>
<td>57.1%</td>
<td>40.8%</td>
<td>48.1%</td>
<td>57.7%</td>
</tr>
<tr>
<td>2002</td>
<td>54.9%</td>
<td>57.4%</td>
<td>39.4%</td>
<td>51.3%</td>
<td>57.5%</td>
</tr>
<tr>
<td>2003</td>
<td>55.7%</td>
<td>58.6%</td>
<td>38.6%</td>
<td>51.7%</td>
<td>58.1%</td>
</tr>
<tr>
<td>2004</td>
<td>56.0%</td>
<td>58.9%</td>
<td>38.3%</td>
<td>52.9%</td>
<td>58.5%</td>
</tr>
</tbody>
</table>


At the University of Mississippi, over the past 10 years, females in the New Freshmen cohort have a higher retention and graduation rate than their male counterparts (The University of Mississippi Office of Institutional Research and Assessment, 2013). Table 2 illustrates these trends between males and females.

Table 2. Retention and Graduation for New Freshmen Cohorts at UM, Gender Differences

<table>
<thead>
<tr>
<th>Cohort Year</th>
<th>Total Cohort Cont to 2nd Yr</th>
<th>% Cont in 6 Yrs</th>
<th>Male Cont to 2nd Yr</th>
<th>% Grad in 6 Yrs</th>
<th>Female Cont to 2nd Yr</th>
<th>% Grad in 6 Yrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>76.0%</td>
<td>55.6%</td>
<td>72.9%</td>
<td>52.6%</td>
<td>78.9%</td>
<td>58.3%</td>
</tr>
<tr>
<td>2003</td>
<td>81.0%</td>
<td>60.5%</td>
<td>78.3%</td>
<td>55.8%</td>
<td>83.3%</td>
<td>64.3%</td>
</tr>
<tr>
<td>2004</td>
<td>79.1%</td>
<td>58.7%</td>
<td>78.1%</td>
<td>57.1%</td>
<td>79.9%</td>
<td>60.2%</td>
</tr>
<tr>
<td>2005</td>
<td>80.3%</td>
<td>60.4%</td>
<td>76.4%</td>
<td>57.5%</td>
<td>83.8%</td>
<td>62.9%</td>
</tr>
<tr>
<td>2006</td>
<td>80.5%</td>
<td>58.4%</td>
<td>78.3%</td>
<td>56.0%</td>
<td>82.4%</td>
<td>60.5%</td>
</tr>
<tr>
<td>2007</td>
<td>78.3%</td>
<td></td>
<td>74.8%</td>
<td></td>
<td>81.4%</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>81.0%</td>
<td></td>
<td>78.1%</td>
<td></td>
<td>83.3%</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>83.1%</td>
<td></td>
<td>79.7%</td>
<td></td>
<td>86.2%</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>81.2%</td>
<td></td>
<td>79.4%</td>
<td></td>
<td>82.8%</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>80.8%</td>
<td></td>
<td>77.6%</td>
<td></td>
<td>83.5%</td>
<td></td>
</tr>
</tbody>
</table>

Source: UM IR&A Official Retention and Graduation Rates
Both from past research results and the trends nationally as well as at The University of Mississippi, it is important that gender differences are investigated when researching retention. 

*Ethnicity*

The US Department of Education published on their website the ethnic breakout of the total undergraduate enrollment in degree-granting postsecondary institutions in 2010. They stated that 60.3% were white, 14.8% were black, 14.1% were Hispanic, 6.0% were Asian or Pacific Islander, 1% were American Indian or Alaskan Native, and 1.6% were two or more races (The US Department of Education, 2011). Not only is there a difference in enrollment when it comes to ethnicity, there is a difference in graduation rates as well. As shown in Table 1, nationally Black students have a much lower graduation rate than their white counterparts.

The University of Mississippi closely represents these US trends as well. Over the past five years, the minority enrollment for the total undergraduate population at The University of Mississippi ranges from 19.8 percent in Fall 2009 to 23.5 percent in Fall 2013 (The University of Mississippi Office of Institutional Research and Assessment, 2013). Mirroring the national trends, there is also a ethnicity differences in retention and graduation rates at The University of Mississippi. Table 3 showcases these differences.
Table 3. Retention and Graduation Rates for New Freshmen Cohorts at UM, Ethnicity Differences

<table>
<thead>
<tr>
<th>Year</th>
<th>Total Cohort %Cont to_2nd_Yr</th>
<th>Total Cohort %Grad in_6_Yrs</th>
<th>Black %Cont to_2nd_Yr</th>
<th>Black %Grad in_6_Yrs</th>
<th>White %Cont to_2nd_Yr</th>
<th>White %Grad in_6_Yrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>76.0%</td>
<td>55.6%</td>
<td>71.7%</td>
<td>42.9%</td>
<td>76.9%</td>
<td>57.9%</td>
</tr>
<tr>
<td>2003</td>
<td>81.0%</td>
<td>60.5%</td>
<td>83.3%</td>
<td>51.4%</td>
<td>81.1%</td>
<td>62.1%</td>
</tr>
<tr>
<td>2004</td>
<td>79.1%</td>
<td>58.7%</td>
<td>73.6%</td>
<td>41.0%</td>
<td>79.5%</td>
<td>60.9%</td>
</tr>
<tr>
<td>2005</td>
<td>80.3%</td>
<td>60.4%</td>
<td>75.2%</td>
<td>43.6%</td>
<td>81.2%</td>
<td>62.6%</td>
</tr>
<tr>
<td>2006</td>
<td>80.5%</td>
<td>58.4%</td>
<td>77.9%</td>
<td>47.9%</td>
<td>81.3%</td>
<td>60.8%</td>
</tr>
<tr>
<td>2007</td>
<td>78.3%</td>
<td></td>
<td>77.8%</td>
<td></td>
<td></td>
<td>78.6%</td>
</tr>
<tr>
<td>2008</td>
<td>81.0%</td>
<td></td>
<td>81.1%</td>
<td></td>
<td></td>
<td>81.7%</td>
</tr>
<tr>
<td>2009</td>
<td>83.1%</td>
<td></td>
<td>81.8%</td>
<td></td>
<td></td>
<td>83.6%</td>
</tr>
<tr>
<td>2010</td>
<td>81.2%</td>
<td></td>
<td>76.9%</td>
<td></td>
<td></td>
<td>82.2%</td>
</tr>
<tr>
<td>2011</td>
<td>80.8%</td>
<td></td>
<td>79.2%</td>
<td></td>
<td></td>
<td>81.6%</td>
</tr>
</tbody>
</table>

Source: UM IR&A Official Retention and Graduation Rates

However, in a retention research study at Oregon State University, it was determined that even though African American students expressed concerns about academic success, if they enter with a similar academic preparation as white counterparts, African American students graduate at a higher rate than any other ethnic group (Murtaugh, Burns & Schuster, 1999, p. 368). Additionally, in one study on First Generation College Students, it was determined that “being White versus African American or Asian lowered likelihood of retention” (D’Amico & Dika, 2013-2014, p. 181). Like gender, considering ethnicity is imperative when researching retention rates, since clearly both nationally and at The University of Mississippi there are differences in rates for these different groups as well as those researchers showcasing conflicting trends to the National and UM trends.

Socio-Economic Background

In an article, released by the United States Census Bureau in September 2012 regarding the results of the 2011 American Community Survey, Mississippi was ranked with the highest poverty rate (22.6 percent) out of all the states in the United States (United States Census
It has been found by Johnson (2008) that “the odds of persistence increase with the increase with the increase of parents’ income.” (p. 788). In another study it was determined that, “student’s and parent’s income capacity and levels affected student retention” (Nandeshwar, Menzies, & Nelson, 2011, p. 14993). Socio-economic status may have an impact on student retention at The University of Mississippi, especially considering the poverty rate for the state of Mississippi.

Residency Status

Several studies have determined that there is in fact a difference in retention trends of in state and out of state students. The research findings conclude that in-state students are more likely to be retained than out-of-state students (Murtaugh, Burns & Schuster, 1999, p. 368; D’Amico & Dika, 2013-2014, p. 181; Whalen, Saunders, & Shelley, 2010, p. 418). Including the residency status may yield significant results.

Parental Educational Achievements

First Generation College Students (FGCSs) are defined by those students whose parents did not obtain a bachelor’s degree or higher. These students are potentially navigating the university system in a different way than their non-FGCSs who may have assistance from parents that are familiar with university life. Some research on FGCSs has determined that “one’s status as a FGCS may present a barrier to academic performance in college” (D’Amico & Dika, 2013-2014, p. 186). Johnson (2008) determined that “the odds of persistence for first-generation students are 0.83 times the odds of persistence for those students whose parent(s) completed a Bachelor’s degree” (p. 788). Likewise in another study they concluded that “Parent’s education level had a positive effect on student retention. Students whose parents did
not attend college had a lower retention compared to students whose parents did attend college” (Nandeshwar, Menzies, & Nelson, 2011, p. 14994).

College Characteristics

**College GPA**

Many researchers have found that cumulative college GPA is related to retention. Gifford, D.D., Briceño-Perriott, J., & Mianzo, F. (2006) found that the students that were retained into their sophomore year had statistically significantly higher first-year cumulative GPAs than those that did not continue to the sophomore year (p. 23). Voelkle & Sander (2008) determined, through their research on dropouts within university students, that an important predictor of potential future dropout was average university grade. Johnson (2008) in his empirical study on student persistence found that “college GPA has the most substantial effect on persistence. One-point increase in first-semester GPA is associated with 3.01 times increase in odds of persistence” (p. 788). Bowen, Chingsos, and McPherson (2009) found that “first-year grades do in fact have a powerful “independent effect” on graduation rates.”(p. 55). Additional research by Baker and Robnett (2012) confirmed this finding as well, specifically for Latino students. They found that “first-year cumulative GPA was a significant predictor for staying enrolled for Latino students: the odds of staying enrolled increased more than 16 times for every 1-point increase in GPA” (p. 331).

**Intended School or College of Degree Program**

There has been several retention studies on different subgroups related to academic areas of interest of the entire freshmen populations. For instance, the retention research of Engineering students (Moses et al., 2011), STEM students (Scott, Tolson & Huang, 2009) and STEM with Business (Rohr, 2013) students mentioned previously indicates that some researchers
hypothesize that there are different retention trends that occur with these subgroups. Some researchers have included each of the different areas of academic schools or colleges to see if they were significant such as Murtaugh, Burns, & Schuster (1999). Integrating intended school or college of the student’s degree program may be a significant factor in predicting retention of the student.

Other Areas of Significance and Limitations

Researchers have tried to find other means to gage students’ success in college other than pre-college academic history and demographics. Even though previous academic behaviors give insight into potential future academic success, as well as demographics, this does not encompass all aspects of success of a student and their future success or failure at a university, suggested by theorist and researchers. As previously stated many researchers use surveys and questionnaires that measure many aspects of a student and their experiences. Even though these surveys and questionnaires yield significant results in explaining student retention the cost, there are definite limitations in utilizing these measurements. These limitations include time to administer and obtain results from the surveys. If the results of a survey take a long period of time to administer, collect, and analyze then important time is lost to identify and help potential at-risk students. Additionally, surveys that have been developed and validated tend to cost money to administer the survey as well as having staff to assist. With the increase of costs and the decreased of funds available this creates problems in utilizing these resources for universities. Required and non-required surveys and questionnaires have a drawback in and of themselves. It is difficult to obtain a good response rate and may result in a biased sample. At UM, an example of such kind of survey is the National Survey of Student Engagement (NSSE). Participation in the survey annually costs approximately $7,800 for a university the size of UM. In 2012, UM
had a 23% response rate and a national response rate of 25% (The University of Mississippi Office of Institutional Research and Assessment, 2013). So therefore, universities must consider greatly the cost benefit ratio of tools of such nature.

Limitations of the Research Study

In terms of analyzing first-year retention, the data used in this study are only related to demographics, pre-college characteristics, admission data known prior to Freshmen Orientation, and first-semester college GPA. There are neither data included about other aspects of the student such as personality traits, study habits, expectations, support structure, student involvement, student commitment, etc. nor any aspects of institutional characteristics that may have an effect on success in college as theorized by many researchers.

Institutional procedural changes may limit the results of this study. For instance, in March of 2011, the Mississippi Intuitions of Higher Learning Board allowed The University of Mississippi to adjust the admission procedure for non-resident applicants, which may limit the results of this study to predict college success on future cohorts. The retention to Fall 2013 rate for the Fall 2012 New Freshmen Cohort was 85.6% compared to previous retention rates of 80.8% and 81.2% for the Fall 2011 and Fall 2010 cohorts respectively. In terms of residency, the retention rate for non-residents for the Fall 2012 New Freshmen Cohort was 85.5% and for the residents the retention rate was 85.7%. For the Fall 2011 New Freshmen Cohort these rates were 79.1% for non-residents and 82.8% for residents. The non-resident retention rate increased by 6.4 percentage points, while the resident retention rate only increased by 2.9 percentage points (UM Institutional Research and Assessment, 2013). Therefore the procedural change may have impacted the retention rate for the non-resident students as well as the overall retention rate for the University.
There may also be limits of generalization to other universities and schools other than The University of Mississippi due to the possible inherent differences of the student bodies. Additionally due to the transferring of the University’s data management platform in 2003 some of The University of Mississippi’s historical records of student metrics were not saved and therefore not available to utilize. Also, in the early years of this new data warehousing platform, there were issues related to the quality of data on some metrics, therefore limiting the utilization of these data.

Summary of the Literature Review

Extensive research has been completed throughout the years on how best to gauge a student’s success in college. Unfortunately, there is not perfect model that can predict student behavior regarding retention because humans are by no means completely predictable. Many faucets of a student have to be considered in order to understand the success or failure of the student in college. It is important though to keep trying to find the best way student success can be predicting by the use of many tools of investigation, such as the different variables to predict student success as well as different statistical modeling procedures. Universities must also weigh the costs and benefits of using a model that will be efficient and precise enough to produce reliable results in a timely fashion.
CHAPTER 3

METHODOLOGY

Participants

The study included first-time, full-time, degree-seeking freshmen (FTFT) from The University of Mississippi that began in the Fall 2003, Fall 2004, or Fall 2005. Seventeen students were excluded from the analysis due to death. Eighty-four students were excluded due to missing high school GPA or an ACT score. In this group that contained missing information, 43 were female (51%) and 41 were male (49%). Thirty-eight were missing a test score, 46 were missing a high school GPA, and 3 were missing both metrics. Twenty students were partitioned off from the main cohort to analyze separately in future research because these students withdrew from the University within the first-semester of enrollment and did not earn any college GPA. Within these 20 students, 6 were female (30%) and 14 were male (70%), 19 were US citizens of which 17 were white (85%), 2 were black (10%), and 1 was a non-US citizen (5%). This group is considered extremely high risk for not being retained and therefore may inherently be different from the main cohort.

The main cohort consisted of 6,652 students. There were 3,572 females (54%) and 3,080 males (46%). The ethnic breakdown of this cohort consisted of 85% White, 12% African American, and 3% unknown or other ethnicity. Mississippi residents consisted of 53% of the cohort. The range in an ACT or converted SAT score was from 12 to 35, with the mean score of
23.2 (SD=4.07). The overall high school GPA ranged from 1.07 to 4.0, with a mean of 3.22 (SD=0.56). The overall first-semester college GPA ranged from 0.0 to 4.0, with a mean of 2.57 (SD=1.00).

**Cross Validation**

The main cohort was split into five randomly selected disjoint subgroups to perform a five-fold cross validation in subsequent procedures. These subgroups were used to build, test, and compare the model performance in the used methodologies. “Much like exploratory and confirmatory analysis should not be done on the same sample of data, fitting a model and then assessing how well that model performs on the same data should be avoided” (Starkweather, 2011). For example, in the first iteration, subgroups one through four were used to construct the model and subgroup five was used to validate the model. In a five-fold cross validation, this process will be completed five times for different subgroups to construct and validate the models for comparison.

**Data Collection and Variable Definitions**

Demographics, previous academic performance, and other historical information were provided by the Office of Institutional Research and Assessment at The University of Mississippi. The data were extracted from official University data that were frozen on the University’s official census dates and deadlines for semester grades. The statistical package, R, was used to perform all statistical methodologies. Refer to Appendix A for the Logistic Regression code for R and Appendix B for the CART code for R.

Demographic data included gender (0=Male; 1=Female), black (0=White/Caucasian and Other/Unknown; 1=Black/African-American), other (0=White/Caucasian and Black; 1=Other/Unknown), Mississippi residency status (0=Non-Resident; 1=MS Resident), from a
contiguous US state to Mississippi (0=No; 1=Yes). Previous academic performance metrics of interest were the overall high school grade point average on a scale from 0 to 4 points and the highest ACT composite score. SAT composite scores were converted to an ACT scale and if both exam scores were submitted, the highest of all submitted scores was used. The range of ACT scores was integers from 1 to 36. High school rank, college credit prior to enrolling at UM, first generation and financial status was not available for this dataset.

Other historical data included initial school or college within UM, initial degree program (0=Undecided; 1=Declared), first-year, fall-to-fall retention to the University (0=Not Retained; 1=Retained), and first-semester college grade point average. First-year retention is defined as a student returning to the consecutive fall term at UM. By the University’s definition, a student does not have to stay continuously enrolled in each semester post their first initial fall. The student does have to be enrolled as of the University’s official census date of the next fall to be counted as a retained student. First-semester college GPA was configured using SAP, the campus data management system for the University. The variable consisted of the total grade points earned divided by the total credit hours earned as of the end of the initial fall semester. This first-semester college GPA does not contain any transfer hours or hours earned while in high school. If a student withdrew from the University in the initial fall and received a “W”, meaning withdrew while passing under the University’s guidelines, this student was excluded from the main cohort due to potential inherent differences within this group.

Classical Statistical Methodologies

Binary logistic regression is a statistical method used when the dependent variable is a dichotomous variable Y, and there are potential combinations of one or more categorical and/or
continuous independent variables \(X_1, \ldots, X_n\). The goal of logistic regression is to use information 
\((X_1, \ldots, X_n)\) to predict \(Y\). In our case, \(Y\) is an ordinary Bernoulli random variable.

“The basic logistic regression analysis begins with logit transformation of the dependent 
variable through utilization of maximum likelihood estimation” (Healy, 2006, p. 4). The logit 
equation, or log odds, of the simple logistics regression model for the case of a single 
independent variable is

\[
\text{logit}(\pi) = \ln \left( \frac{p}{1-p} \right) = \alpha + \beta X
\]

(1)

where \(p\) is the probability that \(Y = 1\) given \(X\). The unknown parameters \(\alpha\) and \(\beta\) are the 
intercept and regression coefficient, respectively. “The logit is the logarithm of the odds of 
success, the ratio of the probability of success to the probability of failure” (Weisberg, 1985, p. 
268). This ratio

\[
\frac{p}{1-p}
\]

(2)

is called the odds ratio. The range of the odds ratio is from zero to infinity because probabilities 
range from zero to one. Note that the logit equation implies

\[
\frac{p}{1-p} = e^{(\alpha + \beta X)}.
\]

(3)

With some algebraic manipulations of Formula 3, we obtain

\[
p = \frac{1}{1+e^{-(\alpha + \beta X)}}.
\]

(4)

This equation is known as the logistic regression equation. “The transformation from probability 
to odds is a monotonic transformation, meaning the odds increase as the probability increases or 
vice versa” (UCLA Statistical Consulting Group, 2013). Therefore, by use of probabilities, odds, 
and log odds one can determine the probability that an event will occur, the odds that an event 
will occur versus the odds that an event will not occur, and the odds an event will occur given a
specific scenario. See Figure 1 for an example of the relationship between probabilities, odds, and log odds.

Figure 1. Relationship between Probabilities, Odds Ratios, and Log Odds

In determining the estimate of the parameters, $\alpha$ and $\beta$, the maximum likelihood estimation (MLE) is calculated. This estimate maximizes the conditional probability or likelihood of the data. Estimates are obtained through an iterative numerical process because there is no closed form of the solution. The log-likelihood is given by

$$
\log L(\alpha, \beta) = \sum_{i=1}^{n} y_i (\alpha + \beta X_i) - \sum_{i=1}^{n} \ln(1 + e^{\alpha + \beta X_i}),
$$

and the estimates of $\alpha$ and $\beta$, $\hat{\alpha}$ and $\hat{\beta}$, are solutions to

$$
\arg\max_{\alpha, \beta} \log L(\alpha, \beta).
$$

“Based on the assumption that the relationship between the dichotomous dependent variable and an independent variable can be represented by a logistic distribution, the probability of the dependent variable [to be 1 (or 0)] is estimated for each group (in the case of grouped data) or for each subject (in case of individual data)” using MLE method (Cabrera, 1994, p. 229). See Figure 2 for the case of one independent variable.
The regression coefficient $\beta$ is the log odds of success for a unit change in $X$ as shown below

$$\ln \left[ \frac{P(Y=1|X+1)/P(Y=0|X+1)}{P(Y=1|X)/P(Y=0|X)} \right] = (\alpha + \beta(X+1)) - (\alpha + \beta X) = \beta. \quad (7)$$

In the case of multiple independent variables, Equation 1 can be extended with all properties of the simple example transferred as follows

$$\text{logit } P(Y=1|X_1, X_2, ..., X_n) = \ln \left[ \frac{P(Y=1|X_1, X_2, ..., X_n)}{1-P(Y=1|X_1, X_2, ..., X_n)} \right]$$

$$\text{logit } P(Y=1|X_1, X_2, ..., X_n) = \ln \left[ \frac{P(Y=1|X_1, X_2, ..., X_n)}{P(Y=0|X_1, X_2, ..., X_n)} \right] \quad (8)$$

$$\text{logit } P(Y=1|X_1, X_2, ..., X_n) = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n.$$ 

Specifically for this research a backward selection procedure is used for the model selection process. The backward selection method consists of including all variables within the model and dropping one variable at a time. At each removal of the variable, the nested model is compared to the previous model and evaluated to determine if the model was worsened by this removal. "Such trimming not only yields a more parsimonious model but also increases the statistical
power of the analysis” (Jaccard, 2001, p. 66). Since the major impetus of this research is to compare the classical logistic models with the more modern CART approach, there was no model selection process completed on the sub-models. All variables selected in the overall model were included on all the sub-models. This is the method utilized to control the complexity of the logistic models. No transformation technique was considered for the independent variables within this research. Also no inclusions of interactions between independent variables were utilized in this process. This could be a question for future research. The assumptions of logistic regression include that the dependent variables must be Bernoulli random variables, that no over or under-fitting of the model should occur, and finally that observations need to be independent.

Specifically in this research, a binary logistic regression method was used to determine the impact of demographics, previous high school academic performance indicators, whether a student has decided to declare a specific degree program or not at time of application submission, and specific school or college in which their intended degree program is associated with at the time of application submission on first-year retention. All these factors are known prior to a student beginning their first-semester at the University. Secondly, another binary logistic regression was performed using demographics, previous high school academic performances indicators, whether a student has decided to declare a specific degree program or not, specific school or college in which their degree program is associated with, and first-semester college GPA at the end of the initial fall semester to determine first-year retention. In the logistic models for retention, the value of the dependent variable was Y=0 for not retained and Y=1 for retained. Notice that within this research, any individual student whose estimated probability of success was greater than the threshold of 0.5 was categorized as retained and a contingency table using
the ground truth was made to assess the overall accuracy, sensitivity, and specificity of each model.

Also in terms of this research, sensitivity is defined as the proportion of retained students that were predicted to be retained. Specificity is defined as the proportion of not retained students that were predicted to not be retained. In a research study conducted by Juana-Maria Vivo and Manuel Franco (2008), it was stated that both sensitivity and specificity must be high for the classification model to be both useful for the classification of success and classification of failure respectively (p. 330). “An ideal diagnostic test has a high sensitivity combined with a high specificity” (Lütkenhöner & Basel, 2013, p. 1). A generic example shown below in Table 4, the sensitivity would be 99.8% and specificity would be 28.2%.

Table 4. Generic Contingency Table

<table>
<thead>
<tr>
<th>Figure 2 Results</th>
<th>Predicted to Not be Retained</th>
<th>Predicted to Retain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actually Not Retained</td>
<td>100</td>
<td>255</td>
</tr>
<tr>
<td>Actually Retained</td>
<td>8</td>
<td>3,206</td>
</tr>
</tbody>
</table>

One of the benefits of the logistic regression approach is that this methodology is rooted in sound statistical theory, which is accepted and used by many researchers across many fields of study. Logistic regression has the ability to measure the relative strengths of the independent variables as well as give a scale of probabilities that an event will occur. This may be helpful, especially in determining those cases that are close to the decision threshold.

Limitations also exist. Many non-statisticians or end users of the research may find it difficult to interpret the results. Even if the research is sound, if the end users are not able to understand the results, then the research is somewhat pointless. Another limitation is the fact
that it may be difficult and often very time consuming to investigate interactions between variables. It is also labor intensive to investigate other variable optimization methods such as transformation of the independent variables. This limitation may lead to a failure in uncovering important interactions in the structure of the data which would lead to a less accurate model.

Modern Statistical Methodologies

The modern statistical methodology that will be utilized in this research is the Classification and Regression Tree (CART). The CART methodology was first developed by Jerome Friedman, Richard Olshen, Leo Breiman, and Charles Stone during the 1970s and 1980s (Breiman, L., Friedman, J. H., Olshen, R. A., & Stones, C. J., 1984, p. ix). Many statisticians dismissed this method in the early years because of the lack of the theoretical constructs. This methodology is more in line with data mining than the classical approach. Instead of finding means and testing hypotheses, this method looks more on the individual level. As stated in the book, *Classification and Regression Trees*, “…the basic purpose of a classification study can be either to produce an accurate classifier or to uncover the predictive structure of the problem” (Breiman, Friedman, Olshen, & Stone, 1984, p. 6). “CART is often able to uncover complex interactions between predictors which may be difficult or impossible to uncover using traditional multivariate techniques” (Lewis, R.J., p. 2). The CART methodology has been used heavily in the pharmaceutical and health care realm due to the ease of interpretation of the results. Over the years, many other industries and fields have adopted this methodology.

There are two types of trees within the CART methodology: the Classification Tree and the Regression Tree. The Classification Tree, is used when the outcome variable is categorical. In this research the outcome variables will be Retained at UM and Not Retained at UM, thus it is
the setup of a Classification Tree. The Regression Tree is used when the outcome variable is a continuous variable.

As mentioned before, the overarching concept of the Classification Tree is to classify data based on a series of hierarchical questions with binary answers forming a decision tree. In essence, there are three main objectives to consider when constructing a Classification Tree:

- What variables and where within the variable should the split occur?
- What stopping rule to use?
- Which class to assign the terminal node?

The process of creating a Classification Tree begins with what is termed as growing the initial tree on the training dataset. This initial tree is usually very complex and overfits the training data. This tree usually lacks the ability to accurately predict classification on new datasets because of this overfitting issue. In the first step of growing a Classification Tree, the software, such as R, takes the data and determines the variable to split so that the data is partitioned into two different independent groups. This process splits the root node and establishes two children nodes (See Figure 2’s Root Node and the two child nodes: Terminal Node 2 and Node 3).
The decision threshold is determined in order to partition the data so that it maximizes correctly predicted outcomes, or minimize the expected error rate or impurity for subsequent nodes. As noted in the book, *Classification and Regression Trees*, “…use the rule that assigns an object selected at random from the node \((t)\) to class \(i\) with the probability \(p(i|t)\). The estimated probability that the item is actually in class \(j\) is \(p(j|t)\). Therefore, the estimated probability of misclassification under this rule is the Gini index” which is defined as follows

\[
GL(t) = \sum_{i \neq j} p(i|t)p(j|t) \tag{8}
\]

In the binary classification problem, the Gini index is as follows

\[
GL(t) = p(1|t)(1 - p(1|t)) + p(0|t)(1 - p(0|t)) \tag{9}
\]

\[
GL(t) = 2p_t(1 - p_t) \tag{10}
\]

where,

\[
p_t = \frac{1}{N_t} \sum_{x_i \in \text{node}(t)} I(y_i = 1) = p(i|t). \tag{11}
\]
Here \( N_t \) is the number of observations in the node \( t \). In other words, \( p_t \) is the proportion of retention in node \( t \). We classify the observation in node \( t \) to be retained if \( p_t > 0.5 \). The overall goal in deciding what variable to split and where the split should occur is to maximize the reduction in the Gini index. The smaller the Gini index the smaller the impurity for subsequent nodes. There are other methods of splitting criteria but it has been determined that “within a wide range of splitting criteria the properties of the final tree selected are surprisingly insensitive to the choice of splitting rule” (Breiman, Friedman, Olshen, & Stone, 1984).

If the splitting decision utilizes a categorical variable the partition is made between the categorical groups. For example, if Gender is used, one branch would include all the females and one branch would include the males (See Second Partition in Figure 4 for an example). If a continuous variable is used the variable is partitioned into two groups (above and below the decision threshold) such that there is a maximum decrease in impurity for the subsequent nodes. For instance, if Age is used and <18.5 is established as the best decision threshold this would partition the data into two separate subgroups: the individuals that were younger than 19 and those that were 19 and older (see First Partition in Figure 4 for an example). This binary decision nodes process continues on each subset, creating parent and child nodes until the process reaches a stopping rule, creating a terminal node.
There are several stopping rules. One stopping rule, which is not recommended and was not used, is that the tree is continually split until the terminal nodes only contain one case. This stopping rule overfits the model and there is very little generalization power to other datasets. Another stopping rule allows the recursive partitioning to occur until the terminal node contains a minimum number of cases. The stopping rule utilized in this research was controlling the complexity parameter. In this way, the tree is allowed to grow until it reaches a certain complexity. Then, after the tree is grown, the complexity parameter is updated to fit, but not overfit, the metrics in the learning dataset. This optimal complexity parameter (cp) is the cp that minimizes the relative error. For example in Figure 5, one would choose cp=0.059. If there are multiple cps that are “tied” or very close to being tied, then choosing the cp associated with the smallest tree size is generally suggested.
This optimal complexity parameter was then applied to the sub-trees to compare the accuracy with the logistic regression sub-models. Each sub-tree model was grown using the defined complexity parameter and not pruned such as the overall model was pruned. This process is paralleled to not performing the model selection on the sub-models for the logistic regression.

Notice that, at each terminal node, there was an assigned class. This process of assigning the class takes into account the greatest accuracy. Referring to Figure 3’s Terminal Node 4, the class assignment was “Not Retained”. This is because out of those 108 females under the age of 19, 100 were not retained and eight were retained. Therefore, using the class “Not Retained” classified 93 percent of the cases accurately. If this node is assigned the classification of “Retained”, then there is a 93 percent misclassification rate for this node. An estimated conditional probability was also determined at each terminal node.

To obtain the overall accuracy of the Classification Tree, two by two contingency tables were constructed which summarized the results of the terminal nodes. Contingency tables were constructed for both the training and testing datasets. Sensitivity and specificity of the models were also assessed.
Once this process has been applied using the training dataset, the model was introduced and assessed using a testing set of data. Thus, each model has a training accuracy and a testing accuracy. It is normal for the training accuracy to be slightly better than the testing accuracy. However, if the testing accuracy is significantly worse than the training accuracy, this may indicate that the model has overfit the training dataset.

Some of the strengths of this type of methodology include: easily understood diagnostic tool, automatic discovery of the useful patterns that are present in the original dataset, all variables are allowed to interact, no underlying distribution assumptions, can handle all types of data including missing data, and can identify important variables that are those mostly used for splitting nodes. Some of the weaknesses include:

- Creating too complex of trees may restrict the generalization power when new data is introduced.
- The tree structures are unstable.
- There is a lack of strong theoretical construct.
- Some researchers and practitioners cannot question the results.

“Because CART analysis is unlike other analysis methods it has been accepted relatively slowly” (Lewis, R. J., 2000, p 2).

Both methodologies have their own advantages and disadvantages. Thus it will be worthwhile to use both methodologies to accurately predict retention for these students. A comparison of the students’ predicted retention classification using both methodologies was completed. For those students whose predicted retention classification matched on both methodologies, a contingency table was constructed to assess accuracy, specificity, and sensitivity. For those students whose predicted retention classification differed between the two
methodologies, a contingency table for each methodology was also constructed on this subset of students in order to determine which methodology was the best in terms of accuracy, specificity, and sensitivity.
CHAPTER 4

RESULTS

Comparison of Classical and Modern Models with Respect to Research Questions

(1) What is the impact of demographics, previous high school academic performance indicators, and first-year retention?

The logistic regression model for data known prior to the initial fall was used to estimate factors which may influence first-year retention behavior for first-time, full-time, degree-seeking freshmen. The overall logistics regression model highlights are as follows. The coefficient associated with gender was statistically significant with p-value less than and was positive for females. It was found that the odds for females are 18% higher than the odds for males to be retained. Associations to Liberal Arts, Applied Sciences, Engineering and Pharmacy were also found to be significant. The coefficients were all negative with p-values ranging from .000 to .01.

- The odds for Liberal Arts students are about 26% lower than the rest of the other schools/college to be retained.
- The odds for Applied Science students are about 33% lower than the rest of the other schools/college to be retained.
- The odds for Engineering students are about 46% lower than the rest of the other schools/college to be retained.
- The odds for Pharmacy students are about 38% lower than the rest of the other schools/college to be retained.

Mississippi residency was significant with p-value of order $10^{-4}$. Unlike the association with the different schools, the coefficient is positive. The odds for MS Residents are about 28% higher than the odds for Non MS Residents to be retained. Both Overall high school GPA and ACT (converted SAT) scores were significant with p-value of order $10^{-4}$. For one point increase in high school GPA, we expect to see about 95% increase in the odds of being retained. Figure 6 shows the incremental unit change in high school GPA for less than a full point increase. For one point increase in ACT score, we expect to see about a 4% increase in the odds of being retained. Figure 7 showcases the incremental unit change in ACT score for more than one unit.

Figure 6. For Every 0.05 Point Change in High School GPA, Percent Change in the Odds of Being Retained
It was found that there were differences among students within different schools and colleges, gender, and residency status with regards to their probability of retention. Figure 8 shows an interesting way to visualize these differences. These graphs use the estimated probability of retention on the y-axis and the overall high school GPA on the x-axis for each of the gender and residency combinations by their association to the different schools and colleges, while accounting for ACT scores. For each of residency and gender combinations, students in Other Schools/Colleges have the lowest GPA thresholds. This indicates that even with extremely low high school GPA's (especially for the MS Resident Females), students are expected to be retained. Liberal Arts ranks second for having the lowest GPA thresholds. Non MS Resident Males have the greatest GPA thresholds, ranging from 0.92 for Other to 1.84 for Engineering. This means that in order to be expected to be retained, a student from this group needs to have better high school GPAs than those in other residency and gender combinations. Overall, females tend to have a lower GPA threshold than their male counterparts (see Table 5 for all of the GPA thresholds by residency, gender, school/college, and GPA combinations).
Figure 8. Residency, Gender, and School/College Probability of Retention

**Applied Science Students**

- MS Res Females
- MS Res Males
- Non MS Res Females
- Non MS Res Males
- Threshold

<table>
<thead>
<tr>
<th>GPA Threshold</th>
<th>MS Res Females</th>
<th>MS Res Males</th>
<th>Non MS Res Females</th>
<th>Non MS Res Males</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.92</td>
<td>1.18</td>
<td>1.29</td>
<td>1.53</td>
</tr>
</tbody>
</table>

**Engineering Students**

- MS Res Females
- MS Res Males
- Non MS Res Females
- Non MS Res Males
- Threshold

<table>
<thead>
<tr>
<th>GPA Threshold</th>
<th>MS Res Females</th>
<th>MS Res Males</th>
<th>Non MS Res Females</th>
<th>Non MS Res Males</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.20</td>
<td>1.48</td>
<td>1.84</td>
<td></td>
</tr>
</tbody>
</table>

**Pharmacy Students**

- MS Res Females
- MS Res Males
- Non MS Res Females
- Non MS Res Males
- Threshold

<table>
<thead>
<tr>
<th>GPA Threshold</th>
<th>MS Res Females</th>
<th>MS Res Males</th>
<th>Non MS Res Females</th>
<th>Non MS Res Males</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.02</td>
<td>1.30</td>
<td>1.89</td>
<td>1.65</td>
</tr>
</tbody>
</table>
Additionally five sub-models were constructed using the significant variables of the overall regression model. When examining all the sub-models for the logistic regression that were constructed, it was found that gender was not significant on all sub-models. In Model 2 and Model 3, the p-value was greater than .05. In Model 3, the School of Pharmacy was not significant at .05. All other variables were significant with all p-values greater than .05 for all
sub-models. See Table 6 for the level of significance of the parameters of all the logistic regression models.

Table 6. Overall and Sub-Model Variable Coefficient and Significance

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-1.86599 ***</td>
<td>-1.61697 ***</td>
<td>-1.61615 ***</td>
<td>-1.66237 ***</td>
<td>-1.25113 ***</td>
<td>-1.600329 ***</td>
</tr>
<tr>
<td>FEMALE</td>
<td>0.17754 *</td>
<td>0.14283</td>
<td>0.11228</td>
<td>0.16867 *</td>
<td>0.20716 **</td>
<td>0.161565 *</td>
</tr>
<tr>
<td>HSGPA</td>
<td>0.67371 ***</td>
<td>0.70109 ***</td>
<td>0.69053 ***</td>
<td>0.69874 ***</td>
<td>0.57099 ***</td>
<td>0.666794 ***</td>
</tr>
<tr>
<td>LIBARTS</td>
<td>-0.24892 **</td>
<td>-0.32082 ***</td>
<td>-0.26074 **</td>
<td>-0.34161 ***</td>
<td>-0.3384 **</td>
<td>-0.302131 ***</td>
</tr>
<tr>
<td>APPSCI</td>
<td>-0.3152 *</td>
<td>-0.35636 *</td>
<td>-0.42872 **</td>
<td>-0.4387 **</td>
<td>-0.49415 **</td>
<td>-0.405414 **</td>
</tr>
<tr>
<td>ENGR</td>
<td>-0.60894 ***</td>
<td>-0.59608 ***</td>
<td>-0.62714 ***</td>
<td>-0.63082 ***</td>
<td>-0.61086 ***</td>
<td>-0.615198 ***</td>
</tr>
<tr>
<td>PHARM</td>
<td>-0.52693 **</td>
<td>-0.48904 *</td>
<td>-0.38224</td>
<td>-0.4691 *</td>
<td>-0.48713 *</td>
<td>-0.470855 **</td>
</tr>
<tr>
<td>MSRES</td>
<td>0.2958 ***</td>
<td>0.20545 **</td>
<td>0.257 ***</td>
<td>0.18348 *</td>
<td>0.28874 ***</td>
<td>0.245844 ***</td>
</tr>
<tr>
<td>HIGHSCORE</td>
<td>0.05089 ***</td>
<td>0.03995 ***</td>
<td>0.04014 ***</td>
<td>0.04313 ***</td>
<td>0.04127 ***</td>
<td>0.042998 ***</td>
</tr>
</tbody>
</table>

*** p<0.001, ** p<0.01, * p<0.05, . p<0.1

As shown in Table 7, the average training accuracy was 80.72%, compared to the testing accuracy of 80.68% for all the sub-models. Even though the average training accuracy is slightly higher than the average testing accuracy, in all of the sub-models, except Model 5, the testing accuracy was slightly better than the training accuracy. Model 5 brought the overall average down quite a bit in the testing accuracy. Model 5 may have overfit the training data and therefore the generalization to the testing data failed to produce results similar the other sub-models. However, the fact that Models 1 through 4 are slightly above the training accuracy is not generally what occurs. Usually the training accuracy is better than the testing accuracy. This does indicate that the models can be generalized fairly well to new sets of data. The models would be considered relatively stable. Note that within the contingency tables (both in the training and testing) there are very few students predicted to not be retained (see Table 8). This indicates that these models are not very specific. This is a major concern of the usefulness of this model. These models are considered highly sensitive, meaning that the proportion of retained students that were predicted to be retained is very high. See Table 9 for specificity and sensitivity of each of the sub-models.
Even though the logistic regression methodology produced results at this time, though poor as it might be, the CART methodology could not produce a viable tree past the root node. The fact that no splits could be made past the root node means that the data could not be split so that the impurity function or misclassification could not be decreased if a split was made. Therefore, none of these variables at this time yielded any significant tree.
(2) What is the impact of demographics, previous high school academic performance indicators, and first-semester college GPA on first-year retention?

Another logistic regression model was constructed utilizing the same variables as before but also including first-semester college GPA. It was found that high school performance metrics, such as Overall HS GPA and ACT scores were no longer significant as previously used in the model. The only variables found to be significant where the p-value is less than .05 were the association to the School of Engineering, the association to the College of Liberal Arts, residency status, and first-semester college GPA. It was found that the odds of being retained differed by association to school or college.

- Liberal Arts students are about 55% lower than other schools/college to be retained.
- Engineering students are about 31% lower than other schools/college to be retained.

The coefficients of these school or college associations were significant with p-value less than .01. It was also shown that the odds for MS Residents are about 29% higher than the odds for Non MS Residents to be retained. Finally, for one point increase in 1st Semester College GPA, we expect to see about a 163% increase in the odds of being retained. Figure 9 illustrates the odds of being retained for incremental units less than one point in first-semester college GPA. Generally, these findings are similar to the previous logistic regression models: significant negative coefficients on the association with the different schools or colleges, positive coefficients with MS residency, and positive coefficients on academic metrics.
Figure 9. For Every Unit Change in 1st Semester College GPA, Percent Change in the Odds of Being Retained

As clearly shown in Figure 10, there were differences among residency and the different schools and colleges in terms of the probability of retention. These graphs show the estimated probability of retention on the y-axis and the first-semester college GPA on the x-axis for each of the residency statuses by their association to the different schools and colleges. Since only one continuous variable was found to be significant, no other variables were accounted for within these graphs. Engineering Students had to have better first-semester college GPAs to be expected to retain. Non MS Resident had larger GPA thresholds than their MS Resident counterparts, ranging from 0.83 for Other to 1.21 for Engineering. This means that in order to be expected to be retained, the Non MS Residents had to have higher 1st Semester College GPAs than MS Residents.
Figure 10. Residency and School/College Probability of Retention

**Liberal Arts Students**

<table>
<thead>
<tr>
<th>1st Sem GPA Thresholds</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MS Res</td>
<td>0.77</td>
</tr>
<tr>
<td>Non MS Res</td>
<td>1.03</td>
</tr>
</tbody>
</table>

**Engineering Students**

<table>
<thead>
<tr>
<th>1st Sem GPA Thresholds</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MS Res</td>
<td>0.55</td>
</tr>
<tr>
<td>Non MS Res</td>
<td>1.21</td>
</tr>
</tbody>
</table>

**All Other School/Colleges**

<table>
<thead>
<tr>
<th>1st Sem GPA Thresholds</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MS Res</td>
<td>0.57</td>
</tr>
<tr>
<td>Non MS Res</td>
<td>0.83</td>
</tr>
</tbody>
</table>
As mentioned earlier, five logistic regression sub-models were built to compare to the five CART sub-models, each using the same training and testing datasets. These sub-models included the significant variables of the overall model. Table 11 showcases the level of significance of the parameters within each of the models. Liberal Arts was the only variable to be not significant to the p-value level less than .05 in the sub-models. All the other variables were significant in every sub-model. First Semester College GPA with p-value of order $10^{-4}$ in all the models.

Table 11. Overall and Sub-Model Variable Coefficient and Significance

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.90114 ***</td>
<td>-0.76533 ***</td>
<td>-0.84879 ***</td>
<td>-0.78923 ***</td>
<td>-0.66464 ***</td>
<td>-0.79343 ***</td>
</tr>
<tr>
<td>LIBARTS</td>
<td>-0.13553</td>
<td>-0.21988 **</td>
<td>-0.15617</td>
<td>-0.25156 **</td>
<td>-0.21164 **</td>
<td>-0.19501 **</td>
</tr>
<tr>
<td>ENGR</td>
<td>-0.32545 *</td>
<td>-0.35856 *</td>
<td>-0.38442 *</td>
<td>-0.38685 *</td>
<td>-0.37836 *</td>
<td>-0.36725 **</td>
</tr>
<tr>
<td>MSRES</td>
<td>0.29846 ***</td>
<td>0.22373 **</td>
<td>0.2738 ***</td>
<td>0.20312 **</td>
<td>0.26302 ***</td>
<td>0.252 ***</td>
</tr>
<tr>
<td>FIRSTSEMGPA</td>
<td>0.98995 ***</td>
<td>0.95465 ***</td>
<td>0.97635 ***</td>
<td>0.98731 ***</td>
<td>0.92081 ***</td>
<td>0.96559 ***</td>
</tr>
</tbody>
</table>

"." p<0.1, "*" p<0.05, "**" p<0.01, "***" p<0.000

The average accuracy of the training dataset was 83.37%, which is just slightly better than the average accuracy of the testing dataset of 83.31%. Table 12 showcases the training and testing accuracy for each of the built models. For two of the sub-models (i.e. Models 2 and 3) the testing accuracy was slightly better than the training accuracy. Model 5 had the lowest testing accuracy at 81.74%. Overall though, the models produced decent accuracy results for the new dataset, confirming the stability of the sub-models. One major difference in the logistic models not utilizing first-semester college GPA and those that do incorporate first-semester college GPA is clearly shown in the contingency tables below (see Table 13). Even though these models are a little lower in their sensitivity than the previously built models they are still highly sensitive, with values ranging from 97.5% to 98.1% on the training dataset and 97.4% to 98.3%
in the testing dataset. The overall model sensitivity was 97.6%. The most significant difference, in terms of model sensitivity and specificity, is that in these models we were able to predict more not retained students than in the other models not utilizing first-semester college GPA. This equates to a change in specificity of the models. Now instead of having a specificity close to zero like before, these specificities range from 21.9% to 25.5% on the training dataset and 20.3% to 27.2% on the testing dataset. The overall specificity was 24.2%. See Table 14 for all the specificity and sensitivity percentages for each of the sub-models.

Table 12. Cross-Validatıon and Accuracy Summary of Training and Testing Groups

| Retention Predicted "AFTER" | P(Y=Retention (1)|IVs Known Prior to Orientation AND 1st Semester Cumulative College GPA) | % Accurate of Training | % Accurate of Testing |
|-----------------------------|-------------------------------------------------------------------------------------------------|------------------------|----------------------|
| Model 1 TRAIN TRAIN TRAIN TRAIN TEST | 83.33% 83.23% | | |
| Model 2 TRAIN TRAIN TRAIN TEST | 83.03% 84.81% | | |
| Model 3 TRAIN TRAIN TEST TRAIN TRAIN | 83.15% 83.83% | | |
| Model 4 TRAIN TEST TRAIN TRAIN TRAIN | 83.54% 82.99% | | |
| Model 5 TEST TRAIN TRAIN TRAIN TRAIN | 83.80% 81.74% | | |
| Average | | | 83.38% 83.32% |

Table 13. Contingency Tables – “After” Logistic Regression Models

<table>
<thead>
<tr>
<th>&quot;AFTER&quot; MODELS</th>
<th>Model 1: Training</th>
<th>Predicted to Not be Retained</th>
<th>Predicted to be Retain</th>
<th>Model 3: Training</th>
<th>Predicted to Not be Retained</th>
<th>Predicted to be Retain</th>
<th>Model 5: Training</th>
<th>Predicted to Not be Retained</th>
<th>Predicted to be Retain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Retained</td>
<td>250</td>
<td>381</td>
<td>240</td>
<td>Not Retained</td>
<td>219</td>
<td>382</td>
<td>Not Retained</td>
<td>311</td>
<td>976</td>
</tr>
<tr>
<td>Retained</td>
<td>106</td>
<td>4,185</td>
<td>108</td>
<td>Retained</td>
<td>104</td>
<td>4,181</td>
<td>Retained</td>
<td>130</td>
<td>2,238</td>
</tr>
<tr>
<td>Model 2: Training</td>
<td>Predicted to Not be Retained</td>
<td>Predicted to be Retain</td>
<td>Model 4: Training</td>
<td>Predicted to Not be Retained</td>
<td>Predicted to be Retain</td>
<td>Overall</td>
<td>Predicted to Not be Retained</td>
<td>Predicted to be Retain</td>
<td></td>
</tr>
<tr>
<td>Not Retained</td>
<td>246</td>
<td>802</td>
<td>264</td>
<td>Not Retained</td>
<td>240</td>
<td>372</td>
<td>Not Retained</td>
<td>311</td>
<td>976</td>
</tr>
<tr>
<td>Retained</td>
<td>101</td>
<td>4,173</td>
<td>104</td>
<td>Retained</td>
<td>104</td>
<td>4,181</td>
<td>Retained</td>
<td>130</td>
<td>2,238</td>
</tr>
<tr>
<td>Model 1: Test</td>
<td>Predicted to Not be Retained</td>
<td>Predicted to be Retain</td>
<td>Model 3: Test</td>
<td>Predicted to Not be Retained</td>
<td>Predicted to be Retain</td>
<td>Model 5: Test</td>
<td>Predicted to Not be Retained</td>
<td>Predicted to be Retain</td>
<td></td>
</tr>
<tr>
<td>Not Retained</td>
<td>59</td>
<td>197</td>
<td>Not Retained</td>
<td>65</td>
<td>190</td>
<td>Not Retained</td>
<td>61</td>
<td>225</td>
<td></td>
</tr>
<tr>
<td>Retained</td>
<td>26</td>
<td>1,048</td>
<td>Retained</td>
<td>25</td>
<td>1,050</td>
<td>Retained</td>
<td>18</td>
<td>1,027</td>
<td></td>
</tr>
<tr>
<td>Model 2: Test</td>
<td>Predicted to Not be Retained</td>
<td>Predicted to be Retain</td>
<td>Model 4: Test</td>
<td>Predicted to Not be Retained</td>
<td>Predicted to be Retain</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Retained</td>
<td>65</td>
<td>174</td>
<td>Not Retained</td>
<td>51</td>
<td>200</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retained</td>
<td>28</td>
<td>1,063</td>
<td>Retained</td>
<td>27</td>
<td>1,053</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

48
Table 14. Specificity and Sensitivity of the Sub-models – “After” Logistic Regression Models

<table>
<thead>
<tr>
<th></th>
<th>Training Specificity</th>
<th>Training Sensitivity</th>
<th>Testing Specificity</th>
<th>Testing Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>24.2%</td>
<td>97.5%</td>
<td>23.0%</td>
<td>97.6%</td>
</tr>
<tr>
<td>Model 2</td>
<td>23.5%</td>
<td>97.6%</td>
<td>27.2%</td>
<td>97.4%</td>
</tr>
<tr>
<td>Model 3</td>
<td>23.5%</td>
<td>97.5%</td>
<td>25.5%</td>
<td>97.7%</td>
</tr>
<tr>
<td>Model 4</td>
<td>25.5%</td>
<td>97.6%</td>
<td>20.3%</td>
<td>97.5%</td>
</tr>
<tr>
<td>Model 5</td>
<td>21.9%</td>
<td>98.1%</td>
<td>21.3%</td>
<td>98.3%</td>
</tr>
</tbody>
</table>

With the inclusion of first-semester college GPA into the logistic model, the overall accuracy was improved. This inclusion equates to an increase in predicting the not retained students and therefore improving the specificity of this approach. It was also determined that with the use of residency status, association to Liberal Arts and Engineering, and first-semester college GPA, the models were fairly stable when applied to new datasets.

In the first step of the model building process for the CART, a tree was built using the entire database of new freshmen and a complexity parameter of cp=.001, meaning that for every continuous splitting of nodes, there is at least 0.1% probability gain. The intent was to construct a tree that was very complex as shown on Figure 11. The training accuracy of this unpruned tree was 84.5%. This is the most accurate of all the trees grown, which is not surprising due to its complexity (see Table 6).
Figure 11. Overall Tree - Unpruned

This tree is very complex and results in overfitting of the data. There are a total of 25 decision nodes, including 27 terminal nodes. The only variables not used in this tree are “Black” and “Engineering”. The root node, like all the trees to follow, utilized the first-semester college GPA variable. In the case of the unpruned and the pruned tree (see Figure 12), the root node splits the first-semester college GPA at 1.20.

Figure 12. Overall Tree - Pruned

After the initial very complex tree was grown, the complexity parameter was reevaluated and the best complexity parameter was chosen such that the tree could be pruned. The
A complexity parameter of 0.003108003 was determined to be the most optimal for the pruned tree, which minimizes the relative error and takes into account the size of the tree. This cp was then applied to all subsequent sub-trees.

Figure 13. Complexity Parameter Evaluation

The pruned tree was found to be 84.0% accurate as compared to the unpruned tree accuracy of 84.5%. However, there were only eight decision nodes compared to the 25 in the unpruned tree. This loss of 0.5% accuracy is definitely justified by the fact that the pruned tree is a much simpler tree. The variables used in this pruned tree included first-semester college GPA, business, gender, and ACT score. First-semester college GPA was used a total of five of the eight decision nodes and as the first split variable in the root node. This indicates that first-
semester college GPA is the most important predictor of retention within the CART methodology.

All CART trees and sub-trees used most frequently first-semester college GPA in the decision nodes. All trees’ root nodes used first-semester GPA with splits at varying thresholds. The unpruned and pruned overall trees first split the database using a GPA of 1.20, the sub-trees at 1.2, 1.4, 1.5, 1.2, and 1.2. As shown in Table 15, the unpruned tree used first-semester college GPA at 6 of 25 nodes, the pruned tree at 5 of the 8 nodes, the Sub-Tree 1 at 4 of the 8 nodes, the Sub-Tree 2 at 6 of the 15 nodes, the Sub-Tree 3 at 4 of the 9 nodes, the Sub-Tree 4 at 3 of the 4 nodes, and the Sub-Tree 5 at 4 of the 6 nodes. Therefore showcasing that by far, first-semester is clearly the strongest predictor of first-year retention in the CART methodology. Table 15 shows how many times the variable is used as a decision node throughout each model and sub-model.

Table 15. Overall and Sub-Model Variable Coefficient and Significance

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unpruned Tree</td>
<td>Pruned Tree</td>
<td>Tree1234</td>
<td>Tree1235</td>
<td>Tree1245</td>
</tr>
<tr>
<td>FIRSTSEMGP A</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>HIGHSCORE</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>HSGPA</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CONTIGSTATE</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIBARTS</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FEMALE</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>BUSINESS</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DECIDEDPRGM</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSRES</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLACK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENGR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total Times Used</strong></td>
<td><strong>25</strong></td>
<td><strong>8</strong></td>
<td><strong>8</strong></td>
<td><strong>15</strong></td>
<td><strong>9</strong></td>
</tr>
</tbody>
</table>

Shown below are the different trees produced in the sub-models. These images show one of the limitations of this methodology: the unstable nature of the tree. Notice that some of the trees are quite simple (Model 4) while others are seemingly complex (Model 2). Both of these models have equal complexity parameters. Model 4 has 4 decision nodes, while Model 2 has 15
decision nodes. Notice also that within each of the sub-models, different variables are utilized and within the variables, different decision thresholds are utilized. However, even with the vast differences in the tree structures, the accuracies are relatively similar.

Figure 14. Sub-Tree 1 - 1234

Figure 15. Sub-Tree 2 - 1235
Figure 16. Sub-Tree 3 - 1245

Figure 17. Sub-Tree 4 - 1345

Figure 18. Sub-Tree 5 - 2345
The average training accuracy for all sub-models is 84.15% and the 82.62% for the testing accuracy. This is very common to have the testing accuracy lower than the training accuracy. Both the training and testing accuracy are relatively close, indicating that the models are fairly accurate in terms of assessing new sets of data. As shown in Table 16, the most accurate in terms of training accuracy is Model 2, which is also the most accurate in terms of testing accuracy. Model 5, just as the logistic regression Model 5 is the lowest in terms of testing accuracy.

Table 16. Cross Validation and Accuracy Summary of Training and Testing – “After” CART Models

<table>
<thead>
<tr>
<th>Retention</th>
<th>SG1 (n=1331)</th>
<th>SG2 (n=1331)</th>
<th>SG3 (n=1330)</th>
<th>SG4 (n=1330)</th>
<th>SG5 (n=1330)</th>
<th>% Accurate of Training</th>
<th>% Accurate of Testing</th>
<th>Complexity Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>TRAIN TRAIN TRAIN TRAIN TEST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>84.18%</td>
<td>82.18%</td>
<td></td>
</tr>
<tr>
<td>Model 2</td>
<td>TRAIN TRAIN TRAIN TEST TRAIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>84.42%</td>
<td>83.23%</td>
<td></td>
</tr>
<tr>
<td>Model 3</td>
<td>TRAIN TRAIN TEST TRAIN TRAIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>84.05%</td>
<td>83.16%</td>
<td></td>
</tr>
<tr>
<td>Model 4</td>
<td>TRAIN TEST TRAIN TRAIN TRAIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>83.93%</td>
<td>82.57%</td>
<td></td>
</tr>
</tbody>
</table>
| Model 5   | TEST TRAIN TRAIN TRAIN TRAIN | | | | | 84.16% | 81.97% | CP=0.003108003
| Average   | | | | | | 84.15% | 82.62% | |
| Pruned Overall | | | | | | | 84.0% | |
| Unpruned Overall | | | | | | | 84.5% | CP=.001 |

Table 17. Contingency Tables – “After” CART Models

The CART models are still more sensitive than specific. Most of the error occurs in predicting more to retain than actually do retain. See Table 17 for the contingency tables showcasing how many freshmen were predicted to retain and not retain and who in actuality...
were retained and were not retained, respectively. The overall specificity of the pruned tree is 29.2% and the sensitivity is 97.1%. The overall specificity of the unpruned tree is 29.7% and the sensitivity is 97.7%. As presented in Table 18, the specificities of the sub-models range from 27.2% to 32.5% on the training dataset and 22.7% to 30.5% on the testing dataset. The sensitivity ranges from 96.4% to 96.9% on the training dataset and on the testing dataset from 94.8% to 97.5%.

Table 18. Specificity and Sensitivity of the Sub-models – “After” CART Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Training Specificity</th>
<th>Training Sensitivity</th>
<th>Testing Specificity</th>
<th>Testing Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>31.9%</td>
<td>96.7%</td>
<td>24.6%</td>
<td>95.9%</td>
</tr>
<tr>
<td>Model 2</td>
<td>32.4%</td>
<td>97.2%</td>
<td>30.5%</td>
<td>94.8%</td>
</tr>
<tr>
<td>Model 3</td>
<td>27.2%</td>
<td>97.7%</td>
<td>22.7%</td>
<td>97.5%</td>
</tr>
<tr>
<td>Model 4</td>
<td>32.5%</td>
<td>96.4%</td>
<td>29.1%</td>
<td>95.0%</td>
</tr>
<tr>
<td>Model 5</td>
<td>29.3%</td>
<td>96.9%</td>
<td>28.3%</td>
<td>96.7%</td>
</tr>
</tbody>
</table>

Combined Outcomes of the Logistic and CART Models

When comparing the two methodologies (using the Overall Logistic Model and the Pruned CART tree), the models produced the same conclusion of predicting the students’ retention classification on 6,516 students of the total 6,652 students. In other words, the two models produced the same classification result for 98% of the students. Table 19 shows the contingency tables for those 6,516 students. The accuracy for this combined outcomes approach was 84.4%. The specificity was 25.2% and the sensitivity was 97.9%. Compared to the CART model alone, the specificity declined about 4.5 percentage points and increased in the sensitivity about 0.2 percentage points. Compared to the overall logistic model alone, the specificity increased approximately 1 percentage point and increased the sensitivity 0.3 percentage points. Therefore, using the combined outcomes approach is better than using the logistic regression approach alone but may not be the case for the CART methodology.
Table 19. Contingency Tables - Combined Outcomes Approach for the Matched Classification Conclusions

<table>
<thead>
<tr>
<th>Combined Logistic and CART Models</th>
<th>Predicted to Not be Retained</th>
<th>Predicted to be Retain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Retained</td>
<td>306</td>
<td>906</td>
</tr>
<tr>
<td>Retained</td>
<td>111</td>
<td>5,193</td>
</tr>
</tbody>
</table>

The models differed in the classification conclusion for 136 students. The CART model yielded significantly more accurate results than the logistic model. The accuracy rates for the CART model with this group was 65.4% while the logistic regression was 34.6%. For the logistic regression, the estimated predicted probability of those 85 students that were misclassified ranged from 0.46 to 0.66, with 76.4% in the range from 0.46 to 0.56. With this low accuracy of the logistic model and the high percentage of those that were misclassified within the 0.46 to 0.56 range, there may be an issue for those that have an estimated predicted probability of being retained that hover around the 0.5 decision threshold for classification.

For the CART model on these 136 students, the specificity was extremely high at 93.3%, correctly predicting 70 of the 75 that were actually not retained to UM. However, in terms of the sensitivity, the model was not good of 31.1%. The opposite was found for the logistic model, 6.7% specificity and 68.9% sensitivity.

Table 20. Contingency Tables – Combined Outcomes of the Logistic and CART Models – Mismatched Conclusions, using CART Model Only

<table>
<thead>
<tr>
<th>CART Used on Mismatched Conclusions</th>
<th>Predicted to Not be Retained</th>
<th>Predicted to be Retain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Retained</td>
<td>70</td>
<td>5</td>
</tr>
<tr>
<td>Retained</td>
<td>42</td>
<td>19</td>
</tr>
</tbody>
</table>
Table 21. Contingency Tables – Combined Outcomes of the Logistic and CART Models – Mismatched Conclusions, using Logistic Regression Model Only

<table>
<thead>
<tr>
<th>Logistic Regression Used on Mismatched Conclusions</th>
<th>Predicted to Not be Retained</th>
<th>Predicted to be Retain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Retained</td>
<td>5</td>
<td>70</td>
</tr>
<tr>
<td>Retained</td>
<td>19</td>
<td>42</td>
</tr>
</tbody>
</table>
CHAPTER 5

DISCUSSION

Overall, the logistic regression model was more accurate than the CART model. The average testing accuracy for the five logistic sub-models was 83.3%, while the average testing accuracy for the five CART sub-trees was 82.6%. Only one of the CART sub-models (Model 5) outperformed the corresponding logistic model. The testing accuracy for the logistic regression sub-model 5 was 81.7%, while the testing accuracy for the CART sub-model 5 was 82.0%.

It was found that the logistic models tended to predict more people to be retained than the CART models. All five sub-models for the logistic regression accuracy for predicting retention for those that were actually retained was better than the CART approach. It would be concluded that the logistic regression methodology was overall more sensitive than the CART methodology. By the logistic regression being more sensitive, this approach produced more false positives than the CART methodology. In terms of retention, the logistic regression models were failing to identify at-risk students. They were in essence, “slipping through the cracks” and by not being able to identify these students there could be a loss of student because potential intervention strategies would not be applied to these students. However, if the resources of the intervention strategies were constricted, this approach would produce a smaller at-risk population than the CART modeling technique.
Even though the logistic regression model was slightly more accurate and more sensitive, the CART models tended to be more specific than the logistic models. Being more specific indicates that the CART models tended to predict more students to not be retained than the logistic models. Four of the five CART sub-models outperformed the logistic regression in terms of better predicting the proportion of students that were not retained who were actually not retained. More false negatives occurred with the CART models, meaning that in terms of retention, some students that were marked at-risk were actually not at-risk in reality. The CART methodology however predicts a larger at-risk population, with more false negatives, and therefore the allocation of intervention resources would have to be stretched.

Some practitioners may determine that false negatives are more detrimental to occur than false positives (or vice versa) and thus place a weight on those error totals to truly determine the best model or methodology. This method is subjective but may yield better results when used in practice. This process was not done for this research but may be a question to consider in future research.

As shown in the result section, both the CART model and the logistic models yielded the same results 98% of the time. However, for the 2% where the classification results differed, the CART model was overall more accurate than the logistic model. However, to get the best results possible to predict those that would not be retained, the CART model is the best methodology to utilize and to predict those that would be retained, the logistic regression model would be the best methodology. Thus using the models in tandem of each other, may lead to a more accurate approach as shown in the results section.

As for variable importance and significance, in the logistic models as well as in the CART models, first semester college GPA was the most significant variable which is consistent with the
previous research. In the logistic models, first semester college GPA has a p-value less than $10^{-4}$ and in the CART models it was used not only as the initial splitting variable but the most frequently used variable split within the tree structure.

Future Research

For improvements of the logistic methodology, there may be some benefit into looking at different transformation and interactions of the independent variables. Adding other variables about the students such as financial data, student involvement data, social risk factors, and more high school performance metrics that were not available on this older dataset such as high school rank and credit earned prior to starting at UM may help with improving both the logistic model as well as the CART models. Also including institutional characteristics about previous high schools such as high school size, public or private institution, and average ACT or SAT scores for graduating classes may yield improvements to the models. Some other areas of future research may include other modeling techniques such as the LDA, Random Forests, ANN, and Support Vector Machines that were mentioned earlier or even a combination of all of these methods as an ensemble approach.
LIST OF REFERENCES


LIST OF APPENDICES
APPENDIX A: LOGISTIC REGRESSION R CODE
## Logistic Regression R Code ##

### Load Final Data File ###
```r
definal <- read.csv("finaldata.csv", header=TRUE)
```

### Set up subsets from the dataset ###
```r
subset1234 <- subset(final, SUBGROUP != 5)
subset1235 <- subset(final, SUBGROUP != 4)
subset1245 <- subset(final, SUBGROUP != 3)
subset1345 <- subset(final, SUBGROUP != 2)
subset2345 <- subset(final, SUBGROUP != 1)
subset1 <- subset(final, SUBGROUP == 1)
subset2 <- subset(final, SUBGROUP == 2)
subset3 <- subset(final, SUBGROUP == 3)
subset4 <- subset(final, SUBGROUP == 4)
subset5 <- subset(final, SUBGROUP == 5)
```

### BEFORE MODELS -- TRAINING/TESTING ###

#### All Model on the training and testing subsets to get coefficients from the training subset to apply to testing subset -- forcing all variables into the model

#### Building the Overall BEFORE Model
```r
glmALL.TESTBEFORE <-
glm(RETURNEDYEARTWO ~ FEMALE + HGPA + BLACK + OTHER + LIBARTS + BUSINESS + APPSCI + ACCY + ENGR + PHARM + DECIDEDPRGM + MSRES + CONTIGSTATE + HIGHSCORE, data = final, family = binomial)
```
```r
summary(glmALL.TESTBEFORE)
```

#### Backward Direction Test

#### Testing the model selection using the AIC (lower AIC = better model)
```r
step(glmALL.TESTBEFORE, direction="backward")
```

#### Final Model from STEP
```r
stepglm.BEFORE <-
glm(RETURNEDYEARTWO ~ FEMALE + HGPA + OTHER + LIBARTS + APPSCI + ENGR + PHARM + DECIDEDPRGM + MSRES + CONTIGSTATE + HIGHSCORE, family = binomial, data = final)
```
```r
summary(stepglm.BEFORE)
```

#### FINAL MODELS FOR BEFORE
```r
glmALL.FINALBEFORE <-
glm(RETURNEDYEARTWO ~ FEMALE + HGPA + LIBARTS + APPSCI + ENGR + PHARM + MSRES + HIGHSCORE, data = final, family = binomial)
```
```r
summary(glmALL.FINALBEFORE)
```
## Model Comparisons between STEP Model and Reduced STEP Model

```r
anova(glmALL.FINALBEFORE, stepglm.BEFORE, test="Chisq")
```

### FINAL MODELS FOR BEFORE
### FINAL binary logistic models for each training set - BEFORE models##

```r
glmALL.FINAL1234BEFORE <-
glm(RETURNEDYEARTWO~FEMALE+HSGPA+LIBARTS+APPSCI+ENGR+PHARM+MS
   RES+HIGHSCORE, data=subset1234, family=binomial)

glmALL.FINAL1235BEFORE <-
glm(RETURNEDYEARTWO~FEMALE+HSGPA+LIBARTS+APPSCI+ENGR+PHARM+MS
   RES+HIGHSCORE, data=subset1235, family=binomial)

glmALL.FINAL1245BEFORE <-
glm(RETURNEDYEARTWO~FEMALE+HSGPA+LIBARTS+APPSCI+ENGR+PHARM+MS
   RES+HIGHSCORE, data=subset1245, family=binomial)

glmALL.FINAL1345BEFORE <-
glm(RETURNEDYEARTWO~FEMALE+HSGPA+LIBARTS+APPSCI+ENGR+PHARM+MS
   RES+HIGHSCORE, data=subset1345, family=binomial)

glmALL.FINAL2345BEFORE <-
glm(RETURNEDYEARTWO~FEMALE+HSGPA+LIBARTS+APPSCI+ENGR+PHARM+MS
   RES+HIGHSCORE, data=subset2345, family=binomial)

glmALL.FINALBEFORE <-
glm(RETURNEDYEARTWO~FEMALE+HSGPA+LIBARTS+APPSCI+ENGR+PHARM+MS
   RES+HIGHSCORE, data=dataset, family=binomial)
```

### Summary of binary logistic models for BEFORE models##

```r
summary(glmALL.FINAL1234BEFORE)
summary(glmALL.FINAL1235BEFORE)
summary(glmALL.FINAL1245BEFORE)
summary(glmALL.FINAL1345BEFORE)
summary(glmALL.FINAL2345BEFORE)

summary(glmALL.FINALBEFORE)
```

### Looking at each model and the contingency tables associated with each model - uses overall
### model independent variables

### MODEL 1 BEFORE##

```r
glmALL.FINAL1234BEFORE <-
glm(RETURNEDYEARTWO~FEMALE+HSGPA+LIBARTS+APPSCI+ENGR+PHARM+MS
   RES+HIGHSCORE, data=subset1234, family=binomial)

summary(glmALL.FINAL1234BEFORE)
```

### Contingency Table for training group=1234##

```r
predict1234B <- predict(glmALL.FINAL1234BEFORE, newdata=subset1234, type="response")
thresh <- 0.5

pred1234B <- cut(predict1234B, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))
```
cTab1234B<--table(subset1234$RETURNEDYEARTWO, pred1234B, dnn=c("actual","predicted"))
addmargins(cTab1234B)

sort1234B<-sort(predict1234B)
plot(sort1234B)

##Contingency Table for testing group=5##
predict5B<-predict(glmALL.FINAL1234BEFORE,newdata=subset5, type="response")
thresh<-0.5
pred5B<-cut(predict5B, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))
cTab5B<--table(subset5$RETURNEDYEARTWO, pred5B, dnn=c("actual","predicted"))
addmargins(cTab5B)

##MODEL2 BEFORE##
glmALL.FINAL1235BEFORE<-
glm(RETURNEDYEARTWO~FEMALE+HSGPA+LIBARTS+APPSCI+ENGR+PHARM+MSRES+HIGHSCORE, data=subset1235, family=binomial)
summary(glmALL.FINAL1235BEFORE)

##Contingency Table for training group=1235##
predict1235B<-predict(glmALL.FINAL1235BEFORE,newdata=subset1235, type="response")
thresh<-0.5
pred1235B<-cut(predict1235B, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))
cTab1235B<--table(subset1235$RETURNEDYEARTWO, pred1235B, dnn=c("actual","predicted"))
addmargins(cTab1235B)

##Contingency Table for testing group=4##
predict4B<-predict(glmALL.FINAL1235BEFORE,newdata=subset4, type="response")
thresh<-0.5
pred4B<-cut(predict4B, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))
cTab4B<--table(subset4$RETURNEDYEARTWO, pred4B, dnn=c("actual","predicted"))
addmargins(cTab4B)

##MODEL3 BEFORE##
glmALL.FINAL1245BEFORE<-
glm(RETURNEDYEARTWO~FEMALE+HSGPA+LIBARTS+APPSCI+ENGR+PHARM+MSRES+HIGHSCORE, data=subset1245, family=binomial)
summary(glmALL.FINAL1245BEFORE)

##Contingency Table for training group=1245##
predict1245B<-predict(glmALL.FINAL1245BEFORE,newdata=subset1245, type="response")
thresh<-0.5
pred1245B<-cut(predict1245B, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))

cTab1245B<-table(subset1245$RETURNEDYEARTWO, pred1245B, dnn=c("actual","predicted"))
addmargins(cTab1245B)

##Contingency Table for testing group3##
predict3B<-predict(glmALL.FINAL1245BEFORE,newdata=subset3, type="response")
thresh<-0.5
pred3B<-cut(predict3B, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))
cTab3B<-table(subset3$RETURNEDYEARTWO, pred3B, dnn=c("actual","predicted"))
addmargins(cTab3B)

##MODEL4 BEFORE##
glmALL.FINAL1345BEFORE<-glm(RETURNEDYEARTWO~FEMALE+HSGPA+LIBARTS+APPSCI+ENGR+PHARM+MSRES+HIGHSCORE, data=subset1345, family=binomial)
summary(glmALL.FINAL1345BEFORE)

##Contingency Table for training group=1345##
predict1345B<-predict(glmALL.FINAL1345BEFORE,newdata=subset1345, type="response")
thresh<-0.5
pred1345B<-cut(predict1345B, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))
cTab1345B<-table(subset1345$RETURNEDYEARTWO, pred1345B, dnn=c("actual","predicted"))
addmargins(cTab1345B)

##Contingency Table for testing group=2##
predict2B<-predict(glmALL.FINAL1345BEFORE,newdata=subset2, type="response")
thresh<-0.5
pred2B<-cut(predict2B, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))
cTab2B<-table(subset2$RETURNEDYEARTWO, pred2B, dnn=c("actual","predicted"))
addmargins(cTab2B)

##MODEL5 BEFORE##
glmALL.FINAL2345BEFORE<-glm(RETURNEDYEARTWO~FEMALE+HSGPA+LIBARTS+APPSCI+ENGR+PHARM+MSRES+HIGHSCORE, data=subset2345, family=binomial)
summary(glmALL.FINAL2345BEFORE)

##Contingency Table for training group=2345##
predict2345B<-predict(glmALL.FINAL2345BEFORE,newdata=subset2345, type="response")
thresh<-0.5
pred2345B<-cut(predict2345B, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))
cTab2345B<-table(subset2345$RETURNEDYEARTWO, pred2345B, dnn=c("actual","predicted"))
addmargins(cTab2345B)
## Contingency Table for testing group=1 ##

\[
predict1B <- \text{predict(glmALL.FINAL2345BEFORE, newdata=subset1, type="response")}
\]

\[
\text{thresh} <- 0.5
\]

\[
pred1B <- \text{cut(predict1B, breaks=c(-Inf, \text{thresh}, Inf), labels=c("DNR","R"))}
\]

\[
cTab1B <- \text{table(subset1$RETURNEDYEARTWO, pred1B, dnn=c("actual","predicted"))}
\]

\[
\text{addmargins(cTab1B)}
\]

## OVERALL BEFORE MODEL ##

\[
\text{glmALL.FINALBEFORE} <- \text{glm(RETURNEDYEARTWO~FEMALE+HSGPA+LIBARTS+APPSCI+ENGR+PHARM+MSRES+HIGHSCORE, data=dataset, family=binomial)}
\]

\[
\text{summary(glmALL.FINALBEFORE)}
\]

## Contingency Table for ALL ##

\[
predictALLB <- \text{predict(glmALL.FINALBEFORE, newdata=dataset, type="response")}
\]

\[
\text{thresh} <- 0.5
\]

\[
predALLB <- \text{cut(predictALLB, breaks=c(-Inf, \text{thresh}, Inf), labels=c("DNR","R"))}
\]

\[
cTabALLB <- \text{table(dataset$RETURNEDYEARTWO, predALLB, dnn=c("actual","predicted"))}
\]

\[
\text{addmargins(cTabALLB)}
\]

\[
\text{sortALLB} <- \text{sort(predictALLB)}
\]

\[
\text{plot(sortALLB)}
\]

## Contingency Table for ALL Before with decision threshold of 0.1 increments ##

\[
predictALLB <- \text{predict(glmALL.FINALBEFORE, newdata=dataset, type="response")}
\]

\[
\text{thresh3} <- 0.3
\]

\[
\text{thresh4} <- 0.4
\]

\[
\text{thresh5} <- 0.5
\]

\[
\text{thresh6} <- 0.6
\]

\[
\text{thresh7} <- 0.7
\]

\[
\text{thresh8} <- 0.8
\]

\[
\text{thresh9} <- 0.9
\]

\[
predALLB <- \text{cut(predictALLB, breaks=c(-Inf,thresh5,thresh6,thresh7,thresh8,thresh9,Inf), labels=c("<.50",".50-.59",".60-.69",".70-.79",".80-.89",".90-1.0"))}
\]

\[
cTabALLB <- \text{table(dataset$RETURNEDYEARTWO, predALLB, dnn=c("actual","predicted"))}
\]

\[
\text{addmargins(cTabALLB)}
\]

\[
\text{barplot(cTabALLB, beside=T, main="Misclassification Counts")}
\]
## AFTER MODELS--TRAINING/TESTING ##

## All Model on the training and testing subsets to get coefficients from the training subset to apply to testing subset--forcing all variables into the model 

## Building the Overall AFTER Model 
```
glmALL.TESTAFTER<-
glm(RETURNEDYEARTWO~FEMALE+HSGPA+BLACK+OTHER+LIBARTS+BUSINESS +APPSCI+ACCY+ENGR+PHARM+DECIDEDPRGM+MSRES+CONTIGSTATE+HIGHSOC RE+FIRSTSEMGPA, data=dataset, family=binomial)
summary(glmALL.TESTAFTER)
```

## Backwards Direction Test 
## Testing the model selection using the AIC (lower AIC = better model) 
```
step(glmALL.TESTAFTER, direction="backward")
```

## Final Model from STEP 
```
stepglm.AFTER<-glm(formula = RETURNEDYEARTWO ~ LIBARTS + APPSCI + ENGR + PHARM + DECIDEDPRGM + MSRES + FIRSTSEMGPA, family = binomial, data = dataset)
summary(stepglm.AFTER)
```

## Final Model from Reduced STEP 
```
glmALL.FINALAFTER<-
glm(RETURNEDYEARTWO~LIBARTS+ENGR+MSRES+FIRSTSEMGPA, data=dataset, family=binomial)
summary(glmALL.FINALAFTER)
```

## Model Comparisons between STEP Model and Reduced STEP Model 
```
anova(glmALL.FINALAFTER, stepglm.AFTER, test="Chisq")
```

## FINAL MODELS FOR AFTER 
## FINAL binary logistic models for each training set - AFTER models##
```
glmALL.FINAL1234AFTER<-
glm(RETURNEDYEARTWO~LIBARTS+ENGR+MSRES+FIRSTSEMGPA, data=subset1234, family=binomial)
glmALL.FINAL1235AFTER<-
glm(RETURNEDYEARTWO~LIBARTS+ENGR+MSRES+FIRSTSEMGPA, data=subset1235, family=binomial)
glmALL.FINAL1245AFTER<-
glm(RETURNEDYEARTWO~LIBARTS+ENGR+MSRES+FIRSTSEMGPA, data=subset1245, family=binomial)
glmALL.FINAL1345AFTER<-
glm(RETURNEDYEARTWO~LIBARTS+ENGR+MSRES+FIRSTSEMGPA, data=subset1345, family=binomial)
glmALL.FINAL2345AFTER<-
glm(RETURNEDYEARTWO~LIBARTS+ENGR+MSRES+FIRSTSEMGPA, data=subset2345, family=binomial)
```
glmALL.FINALAFTER<- 
glm(RETURNEDYEARTWO~LIBARTS+ENGR+MSRES+FIRSTSEMGPA, data=dataset, 
family=binomial)

###summary of binary logistic models for AFTER models##
summary(glmALL.FINAL1234AFTER)
summary(glmALL.FINAL1235AFTER)
summary(glmALL.FINAL1245AFTER)
summary(glmALL.FINAL1345AFTER)
summary(glmALL.FINAL2345AFTER)
summary(glmALL.FINALAFTER)

###Looking at each model and the contingency tables associated with each model - uses overall 
model independent variables
###MODEL1 AFTER##
glmALL.FINAL1234AFTER<- 
glm(RETURNEDYEARTWO~LIBARTS+ENGR+MSRES+FIRSTSEMGPA, data=subset1234, 
family=binomial)
summary(glmALL.FINAL1234AFTER)

##Contingency Table for training group=1234##
predict1234<-predict(glmALL.FINAL1234AFTER,newdata=subset1234, type="response")
thresh<-0.5
pred1234<-cut(predict1234, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))
cTab1234<-table(subset1234$RETURNEDYEARTWO, pred1234, dnn=c("actual","predicted"))
addmargins(cTab1234)

##Contingency Table for testing group=5##
predict5<-predict(glmALL.FINAL1234AFTER,newdata=subset5, type="response")
thresh<-0.5
pred5<-cut(predict5, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))
cTab5<-table(subset5$RETURNEDYEARTWO, pred5, dnn=c("actual","predicted"))
addmargins(cTab5)

###MODEL2 AFTER##
glmALL.FINAL1235AFTER<- 
glm(RETURNEDYEARTWO~LIBARTS+ENGR+MSRES+FIRSTSEMGPA, data=subset1235, 
family=binomial)
summary(glmALL.FINAL1235AFTER)

##Contingency Table for training group=1235##
predict1235<-predict(glmALL.FINAL1235AFTER,newdata=subset1235, type="response")
thresh<-0.5
pred1235<-cut(predict1235, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))
cTab1235<-table(subset1235$RETURNEDYEARTWO, pred1235, dnn=c("actual","predicted"))
addmargins(cTab1235)

##Contingency Table for testing group=4##
predict4<-predict(glmALL.FINAL1235AFTER,newdata=subset4, type="response")
thresh<-0.5
pred4<-cut(predict4, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))
cTab4<-table(subset4$RETURNEDYEARTWO, pred4, dnn=c("actual","predicted"))
addmargins(cTab4)

##MODEL3 AFTER##
glmALL.FINAL1245AFTER<-
glm(RETURNEDYEARTWO~LIBARTS+ENGR+MSRES+FIRSTSEMGPA, data=subset1245,
family=binomial)
summary(glmALL.FINAL1245AFTER)

##Contingency Table for training group=1245##
predict1245<-predict(glmALL.FINAL1245AFTER,newdata=subset1245, type="response")
thresh<-0.5
pred1245<-cut(predict1245, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))
cTab1245<-table(subset1245$RETURNEDYEARTWO, pred1245, dnn=c("actual","predicted"))
addmargins(cTab1245)

##Contingency Table for testing group3##
predict3<-predict(glmALL.FINAL1245AFTER,newdata=subset3, type="response")
thresh<-0.5
pred3<-cut(predict3, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))
cTab3<-table(subset3$RETURNEDYEARTWO, pred3, dnn=c("actual","predicted"))
addmargins(cTab3)

##MODEL4 AFTER##
glmALL.FINAL1345AFTER<-
glm(RETURNEDYEARTWO~LIBARTS+ENGR+MSRES+FIRSTSEMGPA, data=subset1345,
family=binomial)
summary(glmALL.FINAL1345AFTER)

##Contingency Table for training group=1345##
predict1345<-predict(glmALL.FINAL1345AFTER,newdata=subset1345, type="response")
thresh<-0.5
pred1345<-cut(predict1345, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))
cTab1345<-table(subset1345$RETURNEDYEARTWO, pred1345, dnn=c("actual","predicted"))
addmargins(cTab1345)

##Contingency Table for testing group=2##
predict2<-predict(glmALL.FINAL1345AFTER,newdata=subset2, type="response")
thresh<-0.5
pred2<-cut(predict2, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))
cTab2<-table(subset2$RETURNEDYEARTWO, pred2, dnn=c("actual","predicted"))
addmargins(cTab2)

##MODEL5 AFTER##
glmALL.FINAL2345AFTER<-
glm(RETURNEDYEARTWO~LIBARTS+ENGR+MSRES+FIRSTSEMGPA, data=subset2345,
family=binomial)
summary(glmALL.FINAL2345AFTER)

##Contingency Table for training group=2345##
predict2345<-predict(glmALL.FINAL2345AFTER,newdata=subset2345, type="response")
thresh<-0.5
pred2345<-cut(predict2345, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))
cTab2345<-table(subset2345$RETURNEDYEARTWO, pred2345, dnn=c("actual","predicted"))
addmargins(cTab2345)

##Contingency Table for testing group=1##
predict1<-predict(glmALL.FINAL2345AFTER,newdata=subset1, type="response")
thresh<-0.5
pred1<-cut(predict1, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))
cTab1<-table(subset1$RETURNEDYEARTWO, pred1, dnn=c("actual","predicted"))
addmargins(cTab1)

##ALL AFTER##
glmALL.FINALAFTER<-
glm(RETURNEDYEARTWO~LIBARTS+ENGR+MSRES+FIRSTSEMGPA, data=dataset,
family=binomial)
summary(glmALL.FINALAFTER)

##Contingency Table for training group=2345##
predictALLA<-predict(glmALL.FINALAFTER,newdata=dataset, type="response")
thresh<-0.5
predALLA<-cut(predictALLA, breaks=c(-Inf, thresh, Inf), labels=c("DNR","R"))
cTabALLA<-table(dataset$RETURNEDYEARTWO, predALLA, dnn=c("actual","predicted"))
addmargins(cTabALLA)
APPENDIX B: CART CODE FOR R
## Coding for CART Tree with First Semester GPA ##

## Load Dataset ##
```r
getwd()
dataset<-read.csv("finaldata.csv", header=TRUE)
```

## Set up training and testing datasets ##
```r
datasettrain1234<-subset(dataset, SUBGROUP!=5)
datasettest5<-subset(dataset, SUBGROUP==5)

datasettrain1235<-subset(dataset, SUBGROUP!=4)
datasettest4<-subset(dataset, dataset$SUBGROUP==4)

datasettrain1245<-subset(dataset, SUBGROUP!=3)
datasettest3<-subset(dataset, dataset$SUBGROUP==3)

datasettrain1345<-subset(dataset, SUBGROUP!=2)
datasettest2<-subset(dataset, dataset$SUBGROUP==2)

datasettrain2345<-subset(dataset, SUBGROUP!=1)
datasettest1<-subset(dataset, dataset$SUBGROUP==1)
```

require(rpart)
library(partykit)
library(rpart.plot)
library(mvpart)

set.seed(200)

## CP Selection on entire dataset ##
## Set up initial tree with cp=0.001 ##
```r
tree1<-
rpart(RETURNEDYEARTWO~HSGPA+HIGHSCORE+FEMALE+BLACK+OTHER+LIBARTS+BUSINESS+APPSCI+ACCY+ENGR+PHARM+DECIDEDPRGM+MSRES+CONTIGSTATE+FIRSTSEMGPA, data=dataset, method="class", control=rpart.control(cp=.001))
```

# Obtain the value of the cost complexity parameter for trees of different sizes--look to find that xerror has achieved an interior minimum.
```r
printcp(tree1)
```

# Graphical Display of cp table -- Look where the minimum cross-validation error occurred for trees listed in the cp table
```r
plotcp(tree1)
```

# minimum cross-validation error
```r
min(tree1$cptable[,"xerror"])
```
# location of minimum in cp table
which.min.results<-which.min(tree1$cptable[,"xerror"])

# the tree with the minimum cross-validation error
tree1$cptable[which.min.results,]

# Setting cp
cp.choice<-tree1$cptable[which.min.results,"CP"]
cp.choice

# Pruning tree1, using defined CP
pruned.tree<-prune(tree1, cp=cp.choice)

# See Results of pruned tree
pruned.tree
plot(pruned.tree, margin=0.1)
text(pruned.tree, cex=.9, use.n=T)

# Contingency Table for tree1
predicttree1<-predict(tree1, newdata=dataset, type='class')
table(dataset$RETURNEDYEARTWO, predicttree1)

# Contingency Table for pruned.tree
predictpruned.tree<-predict(pruned.tree, newdata=dataset, type='class')
table(dataset$RETURNEDYEARTWO, predictpruned.tree)

## Train Tree 1234-where 1234 is the training set and 5 is the testing set using a set CP
tree1234<-
  rpart(RETURNEDYEARTWO~HSGPA+HIGHSCORE+FEMALE+BLACK+OTHER+LIBAR
TS+BUSINESS+APPSCI+ACCY+ENGR+PHARM+DECIDEDPRGM+MSRES+CONTIGST
ATE+FIRSTSEMGPA, data=datasettrain1234, method="class",
  control=rpart.control(cp=cp.choice))

## Predict Training Data
predictiontrain1234 <- predict(tree1234, newdata=datasettrain1234, type='class')

## Contingency Table for Training Data
table(predictiontrain1234, datasettrain1234$RETURNEDYEARTWO)

## Predict Testing Data
predictiontest5 <- predict(tree1234, newdata=datasettest5, type='class')

## Contingency Table for Testing Data
table(predictiontest5, datasettest5$RETURNEDYEARTWO)
## Show tree1234 using set cp
prp(tree1234, extra=1)
title(main=paste("Tree1234"))

## Train Tree 1235—where 1235 is the training set and 4 is the testing set using a set CP
rpart(RETURNEDYEARTWO~HSGPA+HIGHSCORE+FEMALE+BLACK+OTHER+LIBARTS+BUSINESS+APPSCI+ACCY+ENGR+PHARM+DECIDEDPRGM+MSRES+CONTIGSTATE+FIRSTSEMGPA, data=datasettrain1235, method="class", control=rpart.control(cp=cp.choice))

## Predict Training Data
predictiontrain1235 <- predict(tree1235, newdata=datasettrain1235, type='class')

## Contingency Table for Training Data
table(predictiontrain1235, datasettrain1235$RETURNEDYEARTWO)

## Predict Testing Data
predictiontest4 <- predict(tree1235, newdata=datasettest4, type='class')

## Contingency Table for Testing Data
table(predictiontest4, datasettest4$RETURNEDYEARTWO)

## Show tree1235 using set cp
prp(tree1235, extra=1)
title(main=paste("Tree1235"))

## Train Tree 1245—where 1245 is the training set and 3 is the testing set using a set CP
rpart(RETURNEDYEARTWO~HSGPA+HIGHSCORE+FEMALE+BLACK+OTHER+LIBARTS+BUSINESS+APPSCI+ACCY+ENGR+PHARM+DECIDEDPRGM+MSRES+CONTIGSTATE+FIRSTSEMGPA, data=datasettrain1245, method="class", control=rpart.control(cp=cp.choice))

## Predict Training Data
predictiontrain1245 <- predict(tree1245, newdata=datasettrain1245, type='class')

## Contingency Table for Training Data
table(predictiontrain1245, datasettrain1245$RETURNEDYEARTWO)

## Predict Testing Data
predictiontest3 <- predict(tree1245, newdata=datasettest3, type='class')

## Contingency Table for Testing Data
table(predictiontest3, datasettest3$RETURNEDYEARTWO)
## Show tree1245 using set cp
prp(tree1245, extra=1)
title(main=paste("Tree1245"))

## Train Tree 1345-where 1345 is the training set and 2 is the testing set using a set CP
tree1345<-
  rpart(RETURNEDYEARTWO~HSGPA+HIGHSCORE+FEMALE+BLACK+OTHER+LIBARTS+BUSINESS+APPSCI+ACCY+ENGR+PHARM+DECIDEDPRGM+MSRES+CONTIGSTATE+FIRSTSEMGPA, data=datasettrain1345, method="class",
  control=rpart.control(cp=cp.choice))

## Predict Training Data
predictiontrain1345 <- predict(tree1345, newdata=datasettrain1345, type='class')

## Contingency Table for Training Data
table(predictiontrain1345, datasettrain1345$RETURNEDYEARTWO)

## Predict Testing Data
predictiontest2 <- predict(tree1345, newdata=datasettest2, type='class')

## Contingency Table for Testing Data
table(predictiontest2, datasettest2$RETURNEDYEARTWO)

## Show tree1345 using set cp
prp(tree1345, extra=1)
title(main=paste("Tree1345"))

## Train Tree 2345-where 2345 is the training set and 1 is the testing set using a set CP
tree2345<-
  rpart(RETURNEDYEARTWO~HSGPA+HIGHSCORE+FEMALE+BLACK+OTHER+LIBARTS+BUSINESS+APPSCI+ACCY+ENGR+PHARM+DECIDEDPRGM+MSRES+CONTIGSTATE+FIRSTSEMGPA, data=datasettrain2345, method="class",
  control=rpart.control(cp=cp.choice))

## Predict Training Data
predictiontrain2345 <- predict(tree2345, newdata=datasettrain2345, type='class')

## Contingency Table for Training Data
table(predictiontrain2345, datasettrain2345$RETURNEDYEARTWO)

## Predict Testing Data
predictiontest1 <- predict(tree2345, newdata=datasettest1, type='class')

## Contingency Table for Testing Data
table(predictiontest1, datasettest1$RETURNEDYEARTWO)
## Show tree2345 using set cp
prp(tree2345, extra=1)
title(main=paste("Tree2345"))

## Additional Code
pruned.tree<-prune(tree1, cp=cp.choice)
plot(pruned.tree, margin=.1)
text(pruned.tree, use.n=T, cex=.9)
par(mfrow=c(2,1))
par(mar=c(0.5,2.1,0.5,0))
plot(pruned.tree, margin=.3)
text(pruned.tree, use.n = T, cex = .9)
par(xpd=T)
par(mar=c(1.1,3.1,0.1,1))
mosaicplot(table(pruned.tree$where, dataset$RETURNEDYEARTWO), main="", xlab="", las=1)
par(xpd=F)
par(mar=c(4.1,5.1,2.1,1.1))
par(mfrow=c(1,1))
dotchart(dataset$FIRSTSEMGPA, pch = dataset$RETURNEDYEARTWO, xlab = "Range",
ylab = "Sample", main = "FIRST SEM GPA")
abline(v=1.24,lty=2,col=2)

summary(pruned.tree)
VITA

Tiffany L. Gregory
307 Falkner Hall · University, MS 38677 · (662) 915-5482 · tlgregor@olemiss.edu

Education
Baylor University
Bachelor of Science, Psychology
Aug. 2002-May 2005
Waco, TX

Relevant Work Experience
Data Analyst
June 2010-Present
The University of Mississippi, Institutional Research
Oxford, MS

- Conducted statistical analysis using statistical software including SPSS, R, Rapid Insights Analytics, and Microsoft Excel
- Built predictive models using Regression Analysis, Logistics Regression, and Classification and Regression Trees for Strategic Enrollment, Student Success, and Retention Efforts
- Analyzed data and illustrated trends in order to highlight movements within the data for internal and external constituents.
- Prepared newsletters and reports for internal and external distribution
- Built self-generating databases to streamline reports
- Constructed scenario builders in order to showcase many what-if situations that utilizes a vast array of data sources and trends
- Updated and maintained the departmental website using both Dreamweaver and Word Press formats
- Established a method for auditing campus data related to Enrollment and Applications
- Created, administered, and evaluated qualitative and quantitative surveys ranging from student transitions, student college choice, student expectations, campus climate survey on sexual harassment, and alumni surveys
- Worked with the campus marketing team on providing data and trends to aid in the strategic marketing efforts
- Committee member of the following committees:
  - Retention Task Force (March 2012-Present)
  - Buildings, Grounds and Renovations (Sept 2012-Sept 2013)
  - First-Year Experience Advisory Committee (Oct 2013-Present)
Program Coordinator March 2008-June 2010
The University of Mississippi, Institutional Research Oxford, MS
 Conducted statistical analysis using statistical software, SPSS and SAS
 Built predictive models
 Analyzed data and illustrated trends in order to highlight trends
 Prepared newsletters and reports for internal and external distribution
 Built self-generating databases to streamline reports

Frontier Strategies/Jackson-Alvarez Group Jackson, MS
 Compiled extensive informational databases using twelve years of legislative journals
 Researched records using newspapers, legal documents, and court records
 Summarized large amounts of information in reports
 Dealt with sensitive information in a professional and confidential manner

SMIC International Private School-Beijing Beijing, China
 Organized, coordinated and implemented school policies and procedures
 Edited and maintains the school website
 Coordinated meetings with parents
 Integrated and created supplemental curriculum
 Designed and regulated school disciplinary system
 Instructed English, Math, Science and Physical Education classes for 1st and 2nd grades

English Teacher July 2004-Aug. 2004
SMIC International Private School-Shanghai Shanghai, China
 Certified in English as a Second Language
 Instructed an English class of fifteen second grade children
 Adjusted and created curriculum to accommodate students’ fluency levels
 Worked closely with administrators and assistants in the school

Technical Skills
Microsoft Office: Access, Excel, PowerPoint, Publisher, and Word
Statistical Packages: SPSS, R, and Rapid Insights Analytics
Other: SAP, Hobsins, Dreamweaver (including writing html/css code), WordPress, MapLand (visual mapping tool), Rapid Insights Veera, and Adobe
Conference Presentations and Workshops

MAIR 2012: Models for Predicting Student Success
Abstract: This discussion will look into two models of prediction for student success defined in different ways. The discussion will showcase some implications and cautions of both of these studies and their effect on the University of Mississippi New Freshmen population now and in the future. There will also be a discussion of some potential future studies that will be occurring to gauge and predict student success on campus, not limited to New Freshmen.

SAIR 2013: Predicting Retention of New Freshmen: Two Methodologies
Abstract: Often in the world of data driven decisions, utilizing multiple methodologies to evaluate a problem at hand is very helpful. In this presentation I will showcase the different results, including the strengths and weaknesses, of the Classification and Regression Tree (CART) methodology and the Logistics Regression methodology. I will also talk about different variables that have been shown to be significant in predicting freshmen retention.

SAIR 2013 Workshop: A Gentle Introduction to R for IR Practitioners
Abstract: The goal of the workshop is to familiarize participants with the free software system R. Written and supported by academics and programmers across disciplines, R provides users with a flexible and powerful array of packages and options for describing, displaying, and analyzing data. Participants will get experience with data management, manipulation, and analysis using postsecondary datasets. At the conclusion of the workshop, participants will get access to software code and resources to further explore the capabilities of R.

Internal Research Projects (Not Published)
Spring 2008: Gender Pay Equity Research Study
Spring 2010: Predicting Student Success in Terms of 1st Semester Cumulative College GPA for New Freshmen
Spring 2012: Predicted the Probability of Success of New Freshmen in Math 261 (Calculus I). Where “Success” was defined by a student receiving an A through a C
Summer 2013: Predicting 1st Semester College GPA for MS Resident New Freshmen
Summer 2013: Predicting 1st Year Retention for MS Resident New Freshmen
Fall 2013: Predicting Enrollment of New Freshmen Applicants to the University of Mississippi

Professional Affiliations
Mississippi Association of Institutional Researchers (MAIR)
Southern Association of Institutional Researchers (SAIR)