Fall 9-21-2017

Investigating the Student Enrollment Decision at WKU

Alec Brown
Western Kentucky University, alec.m.brown@gmail.com

INVESTIGATING THE STUDENT ENROLLMENT DECISION AT WKU

A Capstone Project Presented in Partial Fulfillment

of the Requirements for the Degree Bachelor of Arts

with Honors College Graduate Distinction at

Western Kentucky University

By

Alec M. Brown

May 2017

*****

CE/T Committee:

Dr. Melanie Autin, Chair

Dr. Ngoc Nguyen

Dr. Dennis Wilson

Dedication statement

This thesis is dedicated to my parents, Steve and Diane, to my sister, Dana, and to my second home, Western Kentucky University.

ACKNOWLEDGEMENTS

I would not be anywhere near having a finished product without the continued patience, guidance, and thoughtfulness of my thesis advisor, Dr. Melanie Autin. Without her, this project never would have gotten off the ground. I am also incredibly thankful to Dr. Brian Meredith, Ms. Sharon Hunter, Dr. Tuesdi Helbig, and Ms. Cindy Burnette for providing me with a wonderful treasure trove of data with which to work. I would not have been able to find this passion for analytics without the personal growth I have achieved as a member of the Chi Eta Chapter of Phi Gamma Delta. I am proud to have shared three years of my collegiate experience with these men. Finally, a special thanks to Ms. Allison Smith, Ms. Aimee Bettersworth, Ms. Rebekah Russell, Ms. Shelia Houchins, Ms. Torie Cockriel, Ms. Julia McDonald, Ms. Lauren Osello, Ms. Deborah Wilkins, Ms. Andrea Anderson, Ms. Freida Eggleton, Mr. Brian Campbell, Dr. Craig Cobane, and President Gary Ransdell for providing me with the opportunities I have had to make this collegiate experience the best four years of my life so far.

ABSTRACT

The purpose of this research is to investigate the relationships between the enrollment decision of first-time, first-year students admitted to Western Kentucky University and the amount of financial aid awarded, as well as demographic information. The Division of Enrollment Management provided a SAS dataset containing various information about all WKU students admitted in 2013, 2014, and 2015. Additionally, information about the 2016 class of admitted students was provided. The data has been analyzed in SAS Enterprise Miner. We performed analysis using decision tree modeling and logistic regression modeling. Results of these two procedures indicated the importance of credit hours earned by students before attending WKU, the student's academic performance in high school, and the financial aid package offered to the student by the University.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

Enrolling in college is one of the most beneficial decisions a high school student can make relative to his or her long-term prospects in life. Numerous studies have shown that enrolling in college leads to a better quality of life. As a college degree has become more and more of a necessity in order to succeed in the modern world, the cost of attending has risen dramatically. Tuition to attend Western Kentucky University (WKU) in the early 1970s was $200 a semester. Now it is upwards of $8000 for in-state students and even more than that for students from outside the state of Kentucky (Western Kentucky University, 2016).

As costs have increased, the importance of financial aid for students has also increased. WKU provides academic merit-based scholarships, need-based scholarships, and need-based merit scholarships to incoming students as incentives to enroll here as opposed to choosing to attend some other university. This aid can be the deciding factor between a student choosing to enroll at WKU versus at some other university or sometimes even in choosing to not enroll in college at all.

WKU is a four-year degree-granting institution with its largest campus situated in Bowling Green, KY. The institution was established in 1906 and has grown into a school with an enrollment of over 20,000 graduate and undergraduate students. The University's current President, Dr. Gary A. Ransdell, has made it his goal during his twenty-year term to create a "Leading American University with International Reach." His term has brought WKU into prominence, with the athletic programs continuing their long-standing traditions of success, the rise of the Honors College, the creation of the Gatton Academy of

Mathematics and Science, and the rebuilding of a campus in need of repair. In recent years, the University has struggled to make up for decreasing amounts of higher-education funding from the State of Kentucky in its yearly budget. One way to address this continuing deficit would be to increase enrollment at WKU.

The research for this thesis seeks to find a statistical model that WKU can use to predict the enrollment decision of admitted students. Colleges and universities across the globe choose to admit students knowing that some will choose not to actually enroll at that particular institution. In order to make this decision, students typically must consider a host of factors, some potentially weighing higher than others on the final decision. The ultimate goal of this research is to provide information to WKU that can then be used to increase its yield of enrolling students. Predicting enrollment at WKU will benefit the University in three key areas. Within budgeting, WKU's Student Financial Assistance Office can produce a prediction of scholarship dollars spent by the University. This prediction will not be exactly what the University will spend, but it can help the Office to stay within the bounds of their budgeted amount of merit-based, need-based, and need-based-merit scholarships. Within housing, a prediction of enrollment can help Housing and Residence Life to determine the approximate number of beds they will need for first-time, first-year freshmen heading to campus. This can allow for more clarity and precision in their messaging to current residents of freshman residence halls about moving out, and it will also allow them to have an idea of whether or not they can let current on-campus residents move off campus in greater numbers than before. Finally, in programming, the University can create an estimate of the number of instructors, classrooms, and staff members to bring in for M.A.S.T.E.R. Plan orientation week and for the rest of the semester. If the University

has a valid prediction of the first-time, first-year class, they can provide a better-informed number of courses for those students so that they can take the correct courses at the very beginning of their time in college.

The data used in this research was provided by the WKU Division of Enrollment Management, with contributions from the Department of Student Financial Assistance. It contains information about every first-time, first-year accepted student to WKU from the 2013-14, 2014-15, and 2015-16 academic years.

Chapter two reviews the existing literature on modeling the collegiate enrollment decision. In chapter three, the dataset and WKU's specific financial aid components are discussed. In chapter four, decision trees are introduced and the WKU model is revealed. A logistic regression model is presented in chapter five, and chapter six compares the predictive performance of both models. Chapter seven summarizes this research, discusses some concerns and applications, and indicates areas where future research could be completed.

CHAPTER TWO: LITERATURE REVIEW

Modeling the student enrollment decision has consistently been an interesting topic in higher education. As time has passed, several authors have done multiple studies to update or change their model to better represent changing student desires, changing financial conditions, and the changing job market requirements. These studies typically fall into one of two categories: (1) a long-term study of high school students' decision to enroll in any college versus choosing to go straight to the workforce or (2) a study done by a specific college to model what factors are causing students to attend that institution.

Within the first category, several key points are echoed across different studies. The first is that financial aid is important to students who are looking to enroll in college. In fact, McPherson and Shapiro (McPherson & Schapiro, 1991), Perna (Perna, 2000), and (Doyle, 2010) confirm in their separate works that financial aid is a significant factor in students' choosing to enroll in an institution of higher learning.

McPherson and Shapiro (1991) cite evidence of students responding positively to either price cuts or aid increases. They also suggest that students' decision to enroll changes based on changes in relative pricing. Finally, they confirm a study by Manski in 1983 that affirmed the positive effects of Pell Grants on enrollment. They find that students consistently respond to price cuts and aid increases, but their main claim is that students from low-income families are most likely to change their decision based on an aid increase or a price cut. Students from high-income families will find a college to attend regardless of federal aid provided, but students from low-income families often need that federal funding in order to affordably attend college.

Perna (2000) elaborated on these findings by examining different demographic groups. He found that white students, on average, received more in financial aid but that the cost of attending college was relatively higher. Hispanic and African-American students were more likely to receive financial aid than Caucasian students, and African-American students were most likely to also receive federal loans to help pay for school than other demographic groups.

Doyle (2010) summarized this category of research succinctly in 2010 by studying the changes in institutional aid over the 10-year period between 1992 and 2003. He came to the conclusion that institutions of higher learning have responded to known positive effects of increasing financial aid by creating systems that respond to a student's academic characteristics rather than a student's financial need. Essentially, schools use their financial aid capabilities to attract students who are likely to enroll and graduate rather than students who need funding in order to attend college.

In the second category, there have been several different modeling strategies that institutions have taken to approach the student enrollment decision. Williams College economists Nurnberg, Schapiro, and Zimmerman (2012) published a paper about their model of the student matriculation decision at Williams. They found that applicant quality (measured by tests and grade point average), the net price, geographic origin, race, and artistic, athletic, or academic interests significantly affect the matriculation decision of a student. This private, liberal arts college is an incredibly small and selective institution, and as such it is important for them to manage enrollment. Both over and under yields can be devastating to the College. In order to use this information, the college admissions department was able to attach a probability of matriculation to each applicant so that it

could forecast its financial aid budget, potential housing situations, and other college-specific aspects of the first-time, first-year freshman class. An important caveat to this research is that they were not able to provide evidence that Williams College could actively manipulate any of the key variables considered in the student's enrollment decision to actually change that decision. Essentially, they were unable to prove that changing the amount of financial aid offered to a student would cause an increase in the likelihood of that student enrolling there.

In 1993, Leppel created what she called a gravity model for enrollment at a small private college in Pennsylvania. She hypothesized that the closer a student is geographically to this college, the more likely that student is to enroll. In creating this model, she used Newton's Law of Gravitation to create her model for matriculation at this college. This model creates a probability based on two utility functions, the utility of attending College A and the utility of not attending College A. She used data that included academic characteristics of students, and she also collected geographical location data and assigned the students into four groups to measure this distance effect. Within her analysis, she found that only geography and academic ability were statistically significant predictors of enrollment. She concluded that at her specific college, students did not consider academic programs or other characteristics when it came to making the final decision to enroll. She hypothesized that this distance effect is partly caused by a lack of information about the college that is received by the students as distance from the college increases.

WKU Institutional Research conceived of a strategy to increase enrollment and retention in 2013 (Bogard, 2013). Bogard and his team suggested that WKU could improve its enrollment and retention by simply accepting students that are more likely to graduate.

He hypothesized that creating more strenuous admissions standards by which to naturally create this process was impractical based on the current conditions of the University, so he and his team created a model instead based on the data available to them. They categorized students in separate enrollment and retention models based on four levels of likelihood. The goal for the university was to enroll more students who are likely to persist and graduate from the institution. This model was created to quantify the enrollment situation so that university employees could use the information in admissions and enrollment decisions.

Another type of model that has been used in the literature is a regression discontinuity (RD) model, proposed by Van der Klauww (2002). This study is the most similar to the situation at WKU, in that it was performed at a large public four-year institution, using admitted-student data. Van der Klauww wisely remembered to consider omitted variable bias within the model, because admitted-student data lacks key pieces of information that students will use to make their final matriculation decision (namely other financial aid offers from competing schools and interest in a particular academic major or curriculum). This analysis is centered on a regression discontinuity within the institution's financial aid program. The school creates an ability index via a weighting of high school GPA and the student's standardized test score. This ability index has a direct effect on financial aid received from the university. Because a university has no control over these two variables, the author hypothesizes that there will be significant differences in enrollment at the boundaries of the ability scores when students are not all that different from one another on one side or the other. This RD approach eliminates some of the bias in ordinary least-squares regression analysis in this situation and creates accurate

estimators to predict enrollment at the college. The author expects that the results would be slightly different given another college's unique system to assign institutional aid, but the effects should be fairly similar once the model is adjusted appropriately. Financial aid offered by the institution remains a significant predictor of matriculation.

The research conducted both by WKU and outside of WKU clearly demonstrates the importance of financial aid to students. Other models similarly demonstrate the importance of geographical, academic program (choice of major), and other academic ability-related factors are statistically significant predictors based on the specific institution. In order to accurately characterize the matriculation decision at WKU, I will extend the decision tree modeling approach, create a logistic regression model, compare the predictive performance of both of these models, and focus on how WKU can utilize this information to increase enrollment in first-time, first-year freshmen.

CHAPTER THREE: A PICTURE OF WKU FIRST-TIME, FIRST-YEAR ENROLLMENT

The data used for creating the statistical models in this research is from a dataset containing information about all students who were accepted to WKU as first-time, first-year freshmen in the 2013-14, 2014-15, and 2015-16 academic years. This amounts to 21,442 individuals, with key information such as gender, age, ethnicity, high school GPA, state of residence, ACT scores, SAT scores, prior credit hours earned, financial aid offered and awarded, loans offered and awarded, Pell grant eligibility, and the student's enrollment decision, along with other several other variables that are not used in this analysis. All data was provided by the Division of Enrollment Management, the Division of Student Financial Assistance, and WKU Institutional Research.

WKU is a comprehensive, four-year degree-granting institution located in Bowling Green, Kentucky. The institution is not highly selective in nature, but instead attracts diverse students from nearly all fifty states, students from abroad, students in the Honors College, students who barely meet the admissions requirements, students from many different ethnic backgrounds, and students of all socioeconomic classes. WKU prides itself upon being a "Leading American University with International Reach." This slogan describes a university that sends its students around the world, but it also applies to the students who attend classes on the campuses in Bowling Green, Elizabethtown, Glasgow, and Owensboro as well. Although WKU chooses to admit on a non-selective basis, the yield rate of first-time, first-year freshmen in each incoming class is less than 50% (Bogard, 2013).

The results of this research are meant to be used by University employees to inform decisions related to the admissions process, so the information used to build the model must be available to the employees before the student chooses to enroll. The dataset contains a total of 49 variables about each admitted student. For this research, we focus on twelve of these variables to predict the enrollment decision of each student. These variables contain information about demographic information, geographic information, high school performance, credit hours earned at WKU, academic interests, and the financial aid package offered to the student. With this information gathered by WKU's online application, we are able to analyze students' decision to enroll based on changing characteristics.

The student demographic information is all provided on a volunteer basis by the student. A student can choose to answer all, some, or none of the demographic questions. Some of the variables that are collected include an ethnicity description, a first-generation college student identifier, a gender variable, and an age variable. Gender consists of three possible values: male, female, and prefer not to answer. Nearly 60% of the admitted students are female. Figure 3.1 displays the gender of all students within the dataset. Nearly 40% of the admitted students in the dataset are first-generation college students. Figure 3.2 displays the first-generation status of all students within the dataset. Finally, the average age of the admitted first-time, first-year students is 18.33. All figures in this section are separated based on enrollment decision (where "ENROLL = N" indicates that the student did not ultimately enroll at WKU and "ENROLL = Y" indicates that the student did choose to enroll at WKU).

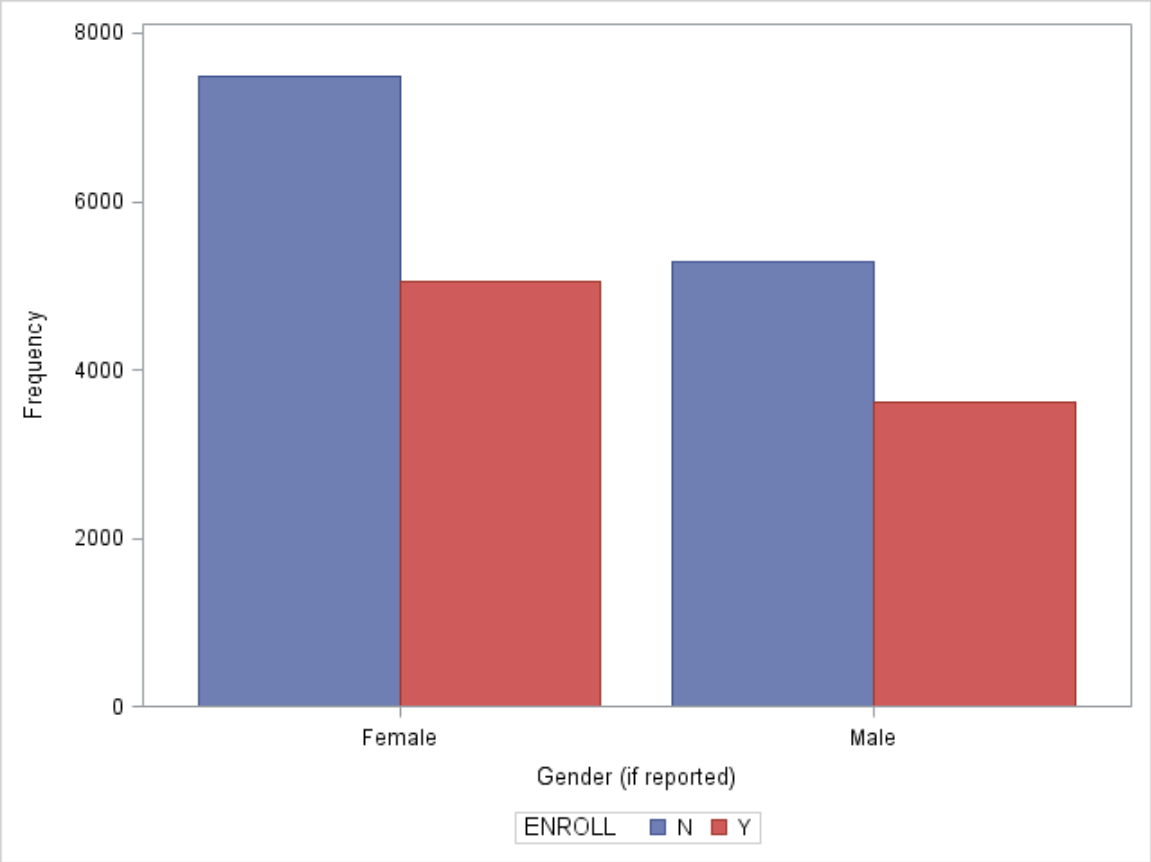

Figure 3.1: Gender of Admitted Students

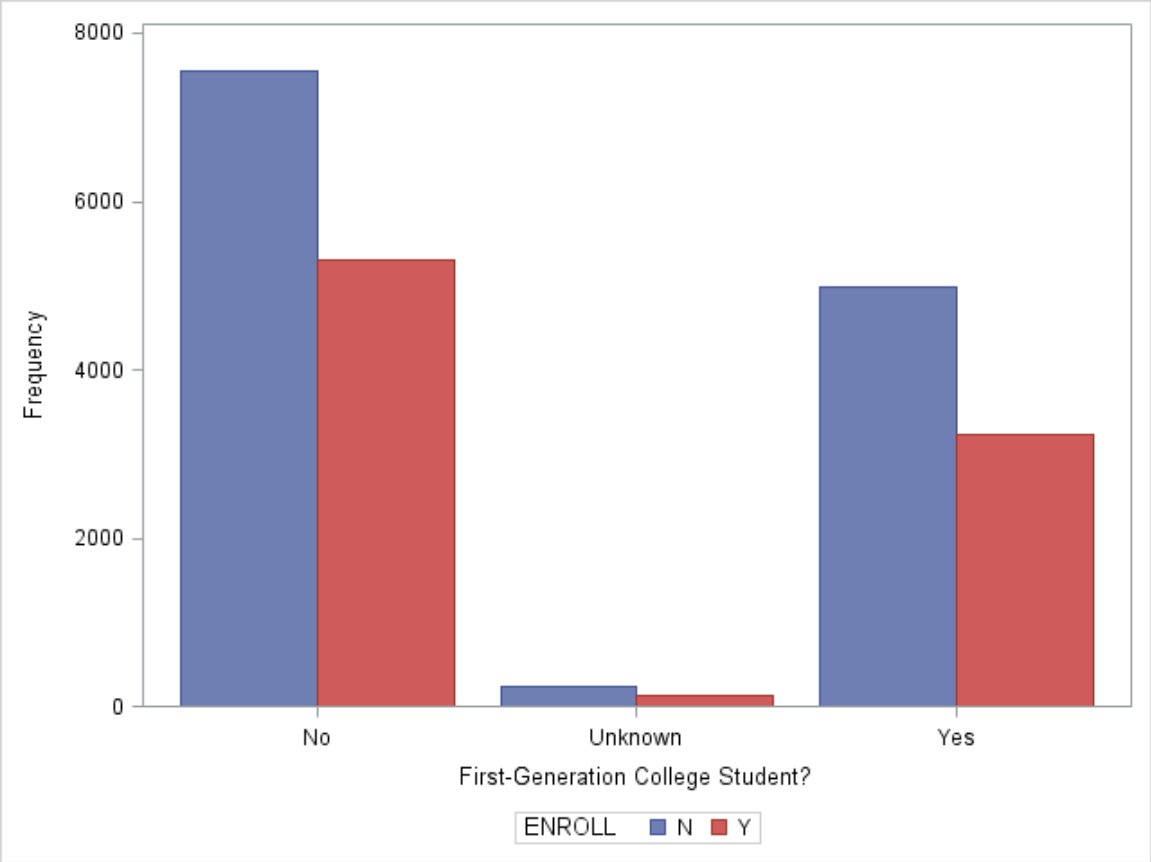

Figure 3.2: First-Generation Status of Admitted Students

Another key piece of the admitted-student data is the geographic component stored in each application. The student's high school, residential location, and county are collected. For this analysis, students were characterized as either in-state students, border-state students (Tennessee, Virginia, West Virginia, Ohio, Indiana, Illinois, and Missouri), or from somewhere else, based on the location of their high school. Of the admitted students, 15,068 were from Kentucky, and the majority of the remaining students are from bordering states. Only 1449 of accepted students were from other locations. Figure 3.3 displays the amount of admitted students from Kentucky, states that share a border with Kentucky, and states that do not share a border with Kentucky.

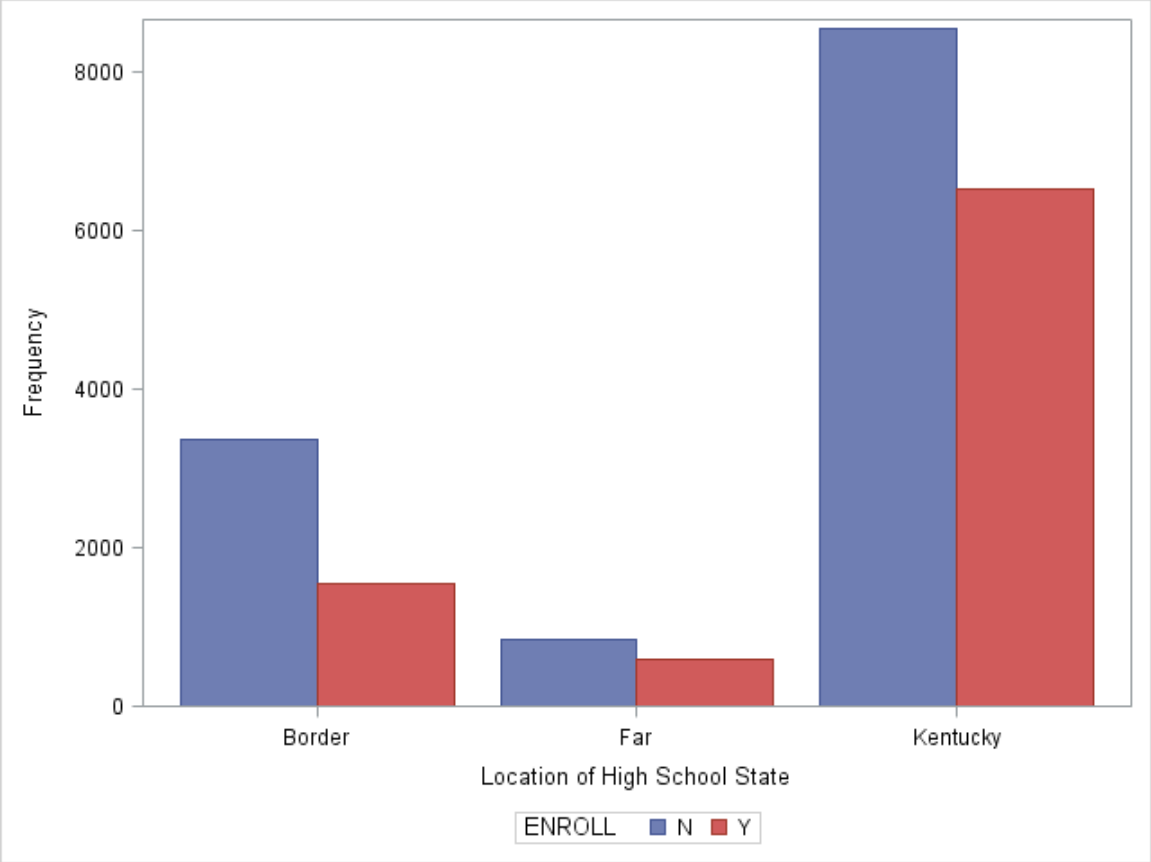

Figure 3.3: Location of High School State of Admitted Students

The ethnicity description present in the WKU application allows the student to select one of nine ethnicity identifiers (white, black or African-American, Hispanic, Asian, pacific islander/native Hawaiian, native Alaskan/American Indian, two or more races, non-resident alien, or unknown). Although the large majority of students who apply are white, there are instances of each of these ethnicities choosing to enroll and choosing not to enroll at WKU. In Figure 3.4, the ethnicities of the admitted students in our dataset are displayed grouped by enrollment decision.

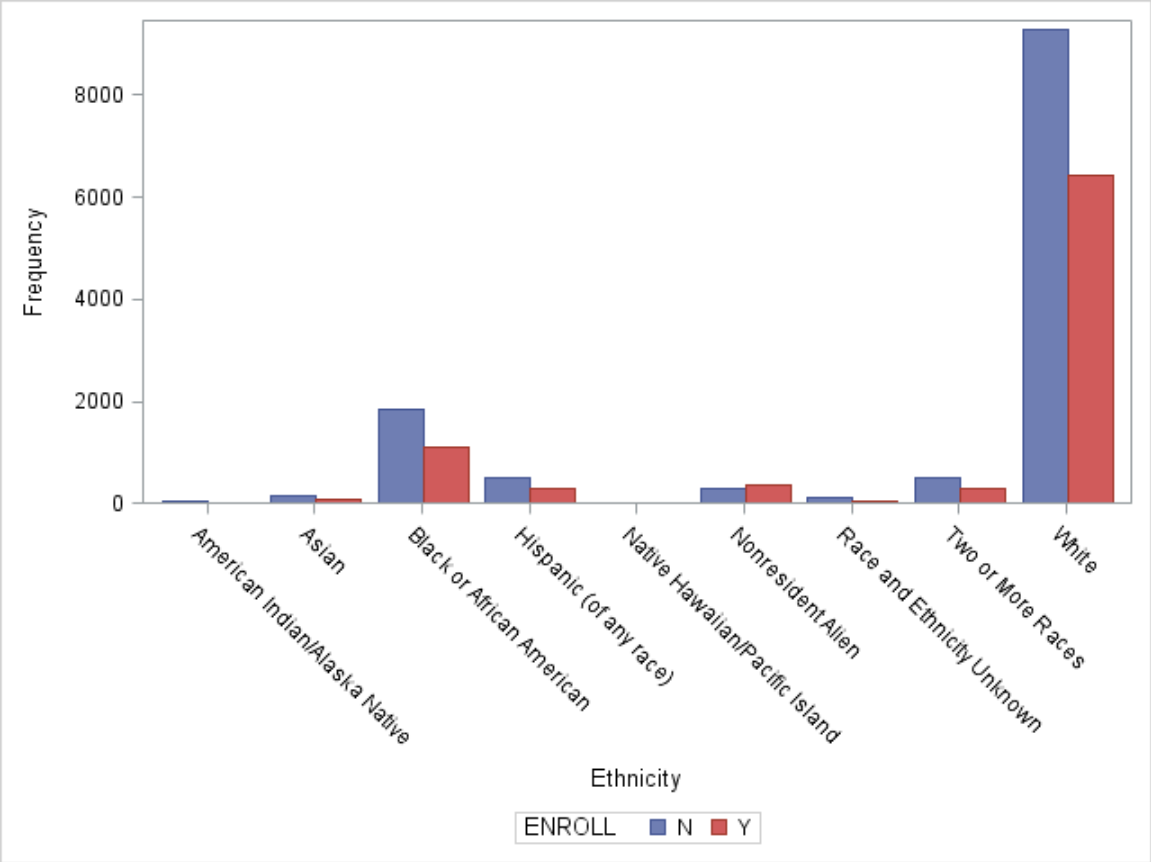

Figure 3.4: Ethnicity of Admitted Students

Within high school performance, there are two key quantitative components of interest. The first is the student's standardized test performance, which is captured by either the SAT or the ACT. WKU accepts scores from both tests with no preference. In order to preserve consistency, we used the application data to create a new variable that measures a student's "best" score out of the reported values. To do this, we converted SAT scores into a comparable ACT score (http://www.studypoint.com/ed/sat-to-act-conversion/, 2017). Then, for students who took both tests, we used the maximum value between their ACT score and their converted SAT score. The other quantitative area of interest within high school data is a student's unweighted GPA. This is entered directly into the

application. The average value of the admitted students' standardized test score is 22.35, and the average high school GPA is 3.24. The distribution of the maximum standardized test scores is shown in Figure 3.5, and the distribution of high school GPA is shown in Figure 3.6.

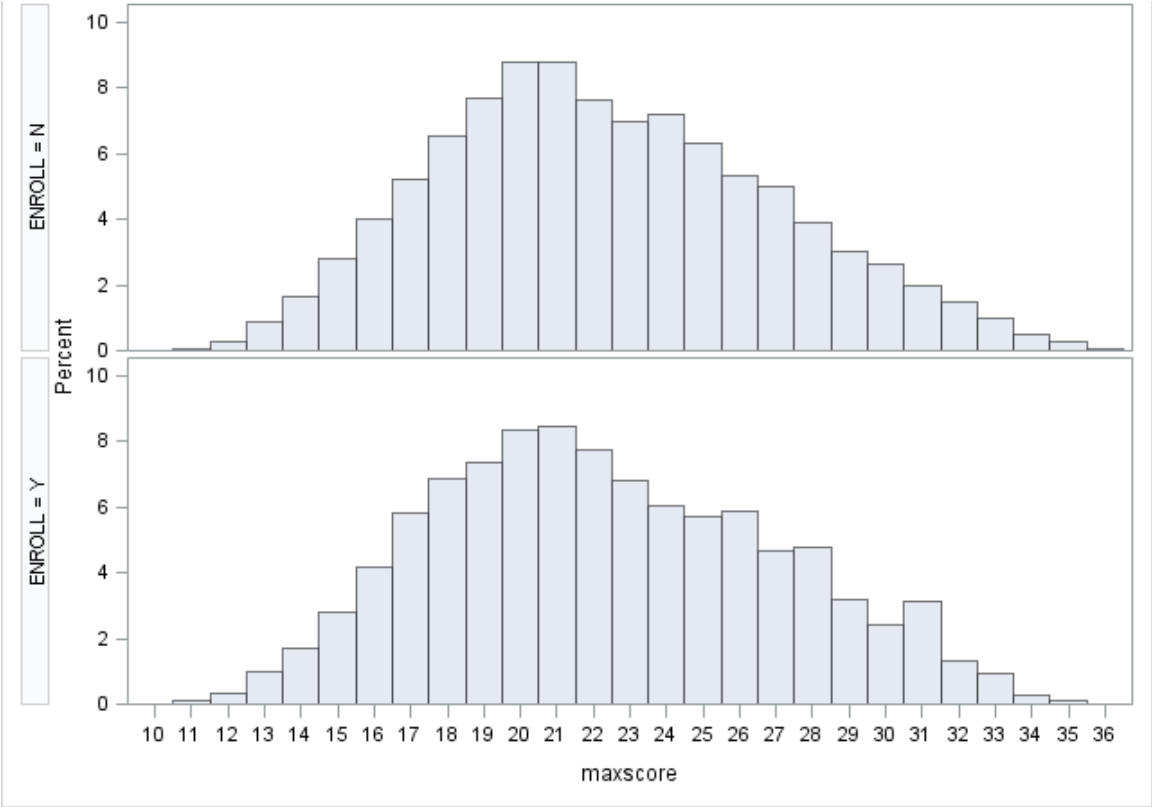

Figure 3.5: Maximum Standardized Test Score of Admitted Students

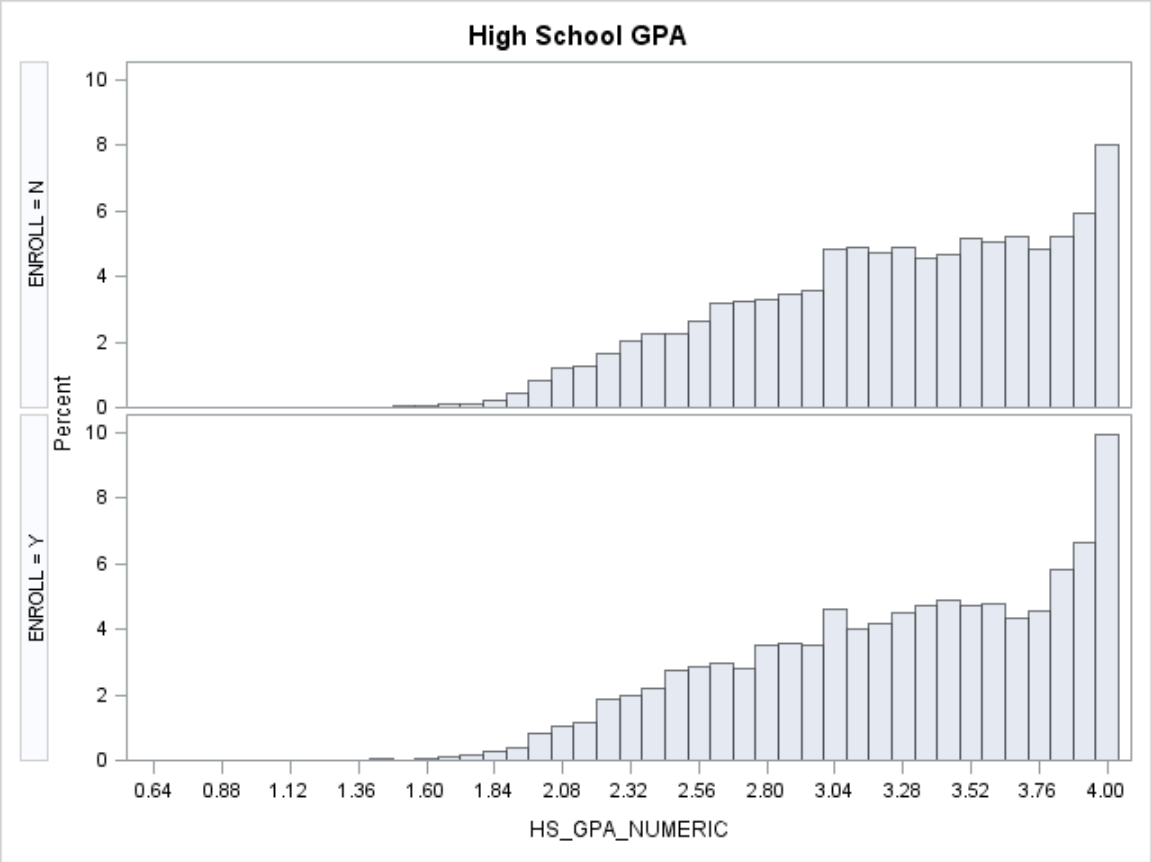

Figure 3.6: High School GPA of Admitted Students

The dataset also contains information about the number of credit hours a student has earned through work prior to college. Some students take dual-credit, Advanced Placement, or International Baccalaureate courses that WKU accepts for introductory classes. Others earn specific credit for component test scores (e.g., English 100 credit is given to all students with a 26 or above on the ACT English), departmental exams, or various other opportunities (Western Kentucky University, 2016). The average admitted student has earned 2.88 credit hours at WKU, which is just short of the credit for one typical class. In context, this means that more students are admitted with 0 hours than non-zero hours. Although students are not required to earn prior credit in order to meet the 120

credit-hour obligation for graduation in four years, many who do earn this credit are able to take advantage of the head start by graduating early or pursuing more complex course schedules. These credits are specific to WKU and are not necessarily reciprocated by the other institutions a student may be interested in attending. A summary of the number of prior hours earned by admitted students is displayed in Figure 3.7.

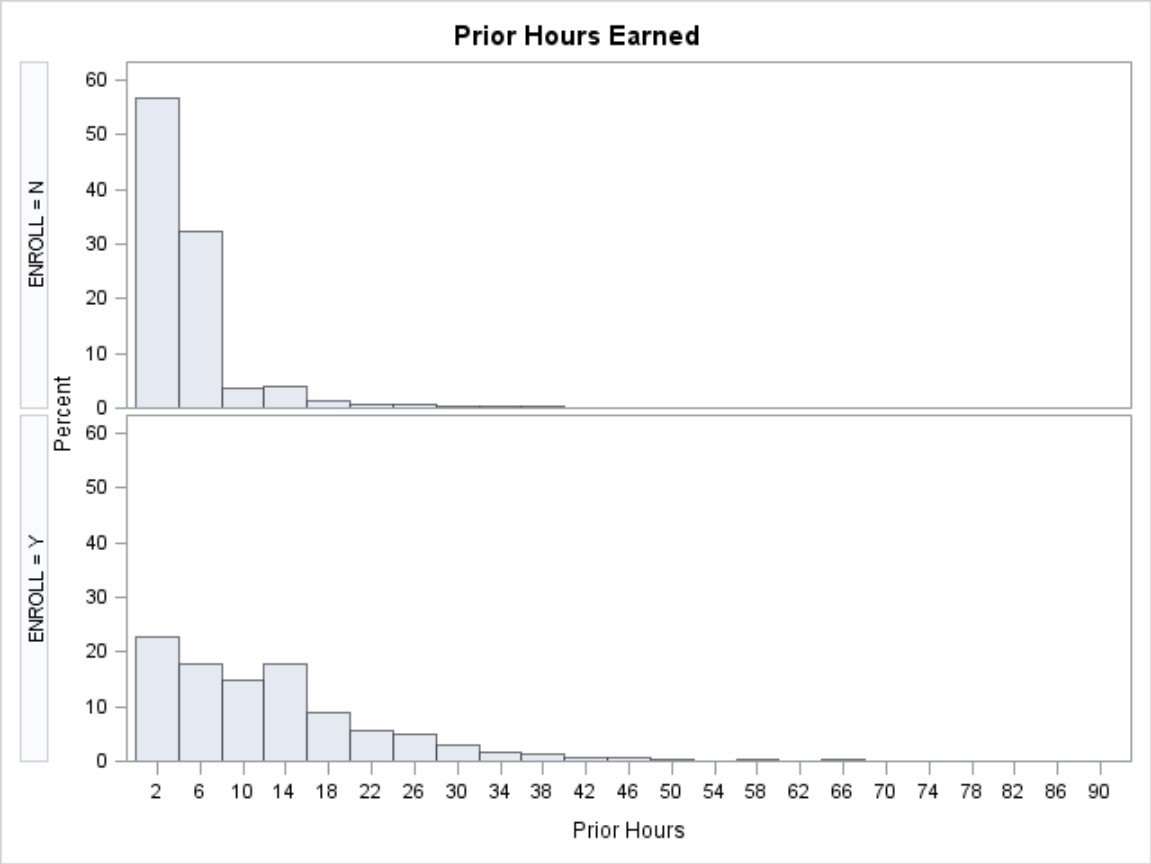

Figure 3.7: Number of WKU Prior Credit Hours Earned by Admitted Students

WKU also collects information about an applicant's potential academic interests, including a potential major and the academic college that major is housed in. The most popular incoming potential major within the dataset is the intended Bachelor of Science in

Nursing, and the second most popular is a Bachelor of Science in Biology. Accordingly, the most popular academic colleges for a first-time, first-year student to enter into are Ogden College of Science and Engineering (OCSE) and the College of Health and Human Services (CHHS). Figure 3.8 displays the amounts of admitted students entering in to each of the academic colleges at WKU.

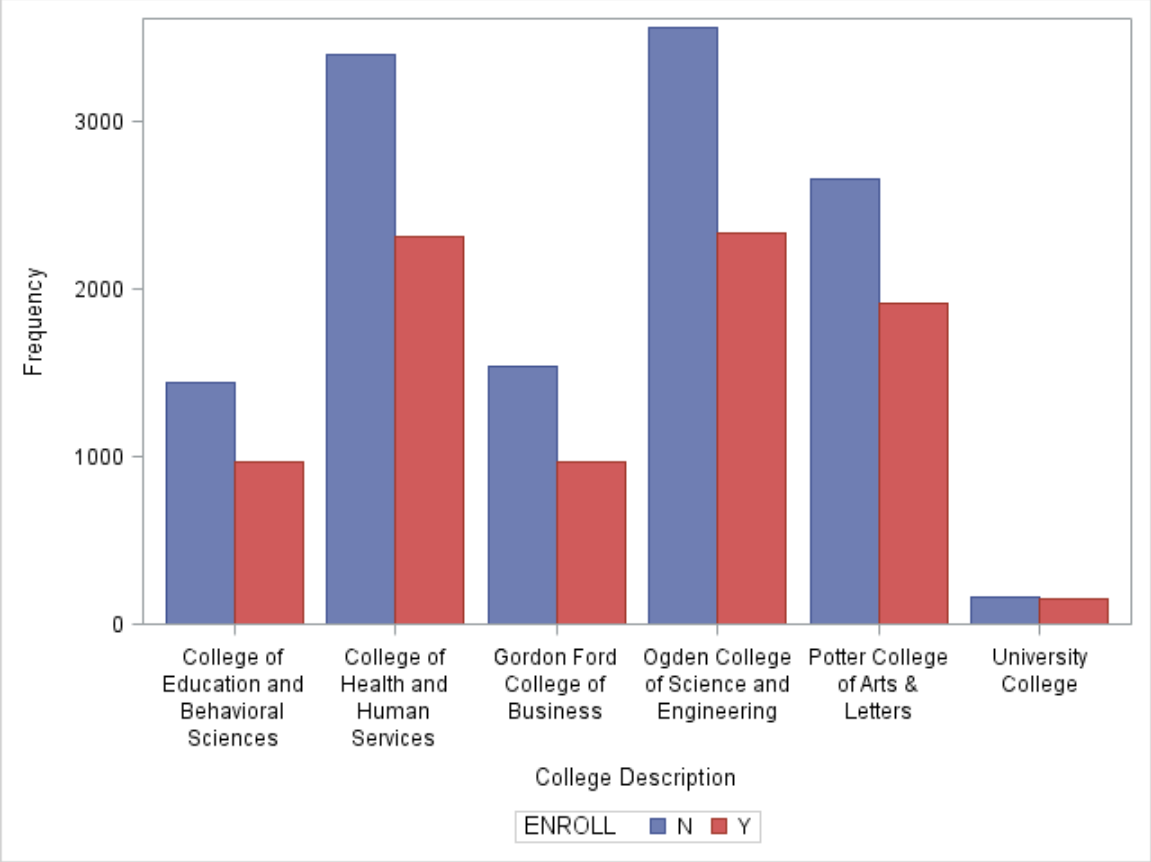

Figure 3.8: Intended College of Admitted Students

The final area of importance contained within the dataset is the financial aid information. The dataset contains the amount of financial aid offered and accepted in merit awards, need-based awards, and need-based merit awards. Merit awards are determined entirely by prior academic performance in high school through test scores and high school

GPA (Western Kentucky University, 2016). Need-based awards are selected through the state and federal government, specifically through the FAFSA application. Need-based merit awards are given through a similar process as need-based awards; however, there is a qualification process based on prior academic performance. The mean amount of funding offered for all admitted students on a merit basis is $1854.86. The mean amount of need-based funding for all admitted students is $12.72, and the mean amount of need-based merit funding offered is $44.04.

In order to qualify for merit-based funding at WKU, one needs at least a 25 composite score on the ACT and a 3.3 unweighted high school GPA (Western Kentucky University, 2016). In 2016, WKU changed the way this merit-based funding was allocated. Now, this funding is split into four block grant award amounts: a $1500 award, a $2500 award, a $4000 award, and an $8000 award; these are all renewable for four years of attendance with a maintained 3.0 cumulative GPA. There are two merit-based scholarships above these grants: a $12000 award and a $16000 award. These are competitively awarded to WKU's top applicants and are renewable with a maintained 3.4 cumulative GPA. Prior to 2016, any student who had above a 31 on the ACT and above a 3.8 unweighted GPA received a scholarship that is approximately equivalent to the $12000 award, with the most successful 20 students receiving an award approximately equivalent to the $16000 award.

There are also separate awards sponsored by private donors, academic units, and the University at large. These awards are made available to students through the wku.edu/topdollar application. When a student fills out that application, it will provide him or her with the opportunity to enter scholarship competitions that he or she qualifies for based on differing characteristics (Western Kentucky University, 2016).

The State of Kentucky provides need-based and merit-based scholarship funding through the Kentucky Lottery. However, its need-based funding scholarships have recently seen a decrease in the total amount of dollars provided by the Kentucky State Legislature. Although the lottery funding is structured to provide 55% of its revenue to need-based grants and 45% to the merit-based Kentucky Educational Excellence Scholarship (KEES), the Legislature has not provided funding in that ratio over the past five years. In 2015, for example, $221.1 million was handed over to the state by the Lottery Corporation. Need-based funding should have received $120 million of this money, but only $91.9 million was given to those programs (Kentucky Center for Economic Policy, 2016). From 2011-2015, 15,000 students were not able to access need-based funding specifically due to the need-based scholarships not being fully funded after the actions of the Kentucky State Legislature (Kentucky Center for Economic Policy, 2016).

In order to qualify for need-based or need-based merit funding, a student has to show financial need for his or her family. In order to qualify for the automatic need-based state awards, students must demonstrate financial need through the FAFSA application. This application also qualifies students for need-based merit scholarships provided by the state, the campus, and other groups. The state awarded "need-based" scholarships are automatically given to students and are renewable as long as the family continues to file the FAFSA every year. Need-based merit scholarships are awarded competitively and are renewable based on a student's college grade point average.

The independent variables that cause variation within the student enrollment decision at WKU each demonstrate key differences in how students behave when making their collegiate decision. On the whole, the group of first-time, first-year students at WKU

looks similar to the group of students that choose not to enroll at WKU in nearly each of the variables (with exceptions seen in some variables, such as the number of prior credit hours earned). As such, further analysis is needed to truly determine how variation in these variables affects the student enrollment decision at WKU. To complete this analysis, we use both a decision tree model and a logistic regression model, and we compare the results. The created models were then tested using a second dataset consisting of the same information about all students accepted for admission at WKU as a first-time, first-year freshman in the 2016-17 academic year; this consists of data for 8,475 students.

CHAPTER FOUR: DECISION TREES

4.1 Decision Tree Algorithms

In order to perform an analysis of WKU enrollment and inform predictive analytics, decision tree modeling was first utilized. A decision tree is a decision support tool that uses a tree-like graph to illustrate a model of decisions and their possible consequences. A decision tree is created through an algorithm that identifies different ways to split the data based on the response of a target variable (in this research, the enrollment decision) relative to input variables. These are called decision rules.

The discovery of the decision rule to form the branches or segments underneath the root node is based on a method that extracts the relationship between the object of analysis (that serves as the target field in the data) and one or more fields that serve as input fields to create the branches or segments. The values in the input field are used to estimate the likely value in the target field (de Ville & Neville, 2013).

De Ville and Neville go on to explain, "Rules can be selected and used to display the decision tree, which provides a means to visually examine and describe the tree-like network of relationships that characterize the input and target values. Decision rules can predict the values of new or unseen observations that contain values for the inputs, but might not contain values for the targets" (de Ville & Neville, 2013). An important process in identifying decision rules exists when deciding how to "bin" the data together. Each decision rule is mutually exclusive, guaranteeing that a single observation cannot qualify for more than one bin.

Decision trees attempt to find a strong relationship between input values and target values in a group of observations that form a data set. When a set of input values is identified as having a strong relationship to a target value, then all of these values are grouped in a bin that becomes a branch on the decision tree. (de Ville & Neville, 2013).

This test utilized is a Chi-Square Test of Independence. If the input variable being considered (with a set of bins being tested) is not statistically significant, then there should be no difference in the distribution of the value of the target variable (i.e., successes and failures) for the bins. If a contingency table of the dataset is created with *k* rows for the *k* bins being considered and two columns for the two values of the binary target variable, then the test statistic is

$$Q = \sum_{j=1}^{k} \sum_{i=1}^{2} \frac{(y_{ij} - n_j \hat{p}_i)^2}{n_j \hat{p}_i}, \tag{4.1}$$

where y_{ij} is the number of observations falling into row j and column i of the table, and $\boldsymbol{n_j \hat{p}_i}$ is the expected number that would fall into that cell if there is no relationship between the target variable and the potential input variable (binned in the way that is being considered). Since the distribution of Q is approximately $\boldsymbol{\chi^2}$ with $k - 1$ degrees of freedom, the p-value is found by finding $P(\boldsymbol{\chi^2_{k-1} \geq Q})$ (Hogg & Tanis, 2015). When the p-value is small, a statistically significant relationship has been detected by the algorithm generating the decision tree. Once this relationship is found, the algorithm tests all of the binning possibilities to determine which has the strongest relationship with the target variable outcome. To do this, the p-value generated by the test statistic is transformed to a measure called logworth, where logworth = -log(p-value). The binning possibility with the largest logworth is the one that is added to the decision tree (SAS, 2017). Continuing the process,

the algorithm will run tests of significance on the decision rule created to determine whether the input variable is a significant descriptor of the target value and to assess its relative strength compared to other input variable rules (de Ville & Neville, 2013). This process ensures that only statistically significant relationships, relationships demonstrably different from random effects, are included in the model. It also addresses biases in selection of input variables in branch partitioning. Tree growth is terminated when there is no significant relationship created when splitting the data in further branches (de Ville & Neville, 2013).

SAS Enterprise Miner was used to create models and to perform all analyses (SAS). This program was created as an offshoot of base SAS, specifically for business analytics and predictive modeling. It runs SAS code based on a point-and-click diagram created by the user and creates output that varies based on the path, process, and analysis that the user wants to perform. Although many different types of analysis are possible with SAS Enterprise Miner, this portion of the research relied solely on the program's decision tree modeling. SAS Enterprise Miner's Decision Tree Algorithm relies on traditional significance testing as mentioned above. This test essentially answers the question: "Are there differences in magnitude among the groups so great that the null hypothesis of no differences can be rejected as not tenable?" for each individual branch of the tree (de Ville & Neville, 2013). The outcome of the tree can change based on the input selection, but the process for growing and pruning the tree remains the same.

SAS also uses a validation method to verify the integrity of a statistical model. To do this, it sets aside a selected percentage of the original dataset to test the performance of the model. This partitioning of the data into training, validation, and test data creates

safeguards within the dataset to check the validity of the predictions. SAS Enterprise Miner has a command called data partition, which separates the data into these three sub-sets, with the percentages of the overall dataset allotted to each group chosen by the user. The statistical testing to determine the decision rules is run with the training data set. These rules are then used in the validation data set. Both sets of results are displayed in the decision tree in order to highlight the similar outcomes exhibited in training and validation data. The test data is used for later modeling not present in this research.

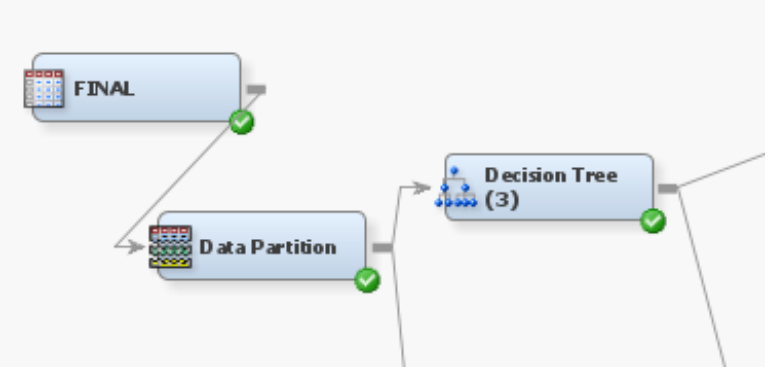

Figure 4.1: An Example SAS Enterprise Miner Diagram

A path for creating decision trees in SAS Enterprise Miner looks similar to what is shown in Figure 4.1.

The first node (FINAL) is the SAS Data Table containing the dataset. The next step is the Data Partition node, which separates the data into the training, validation, and testing sets. After this node, there is the decision tree node, which initiates the modeling. The user can decide several key settings within the decision tree node before the analysis is run by the program. In determining these settings, de Ville and Neville suggest considering the

following: how will input categories be combined to form branches, how will branches be sorted, how many nodes are allowed on a branch, what is the predictive power between branches, how are branches evaluated, when will the decision tree processor stop identifying potential branches and nodes, and will branch growth be based on hypothesis tests or empirical tests of accuracy (de Ville & Neville, 2013). All of these factors can be decided by the user and incorporated into the decision tree algorithm. These decisions will affect the output of the process, which is why it is important for the user to carefully consider the effects that changing settings can have on the model.

The settings of the decision tree algorithm in SAS Enterprise Miner can create differences in the final tree outcome. One can make upwards of thirty different adjustments to the decision tree algorithm. Three key settings in the decision tree model that produce noticeable changes are "maximum branch," "use input once," and "maximum depth." The maximum branch setting changes the number of branches into which an input variable can split. Its default setting is two, so changing this will create the opportunity for branches of more than two bins. Another key decision is whether or not to use an input variable more than once in a decision tree. When this option is chosen to be "no," a variable is used for at most one splitting rule and then cannot be used again at a point later in the algorithm. A "yes" setting for this option would allow for further decision rules to be created with an input variable. For example, if a tree's first splitting rule was in high school GPA, a "yes" setting would allow the algorithm to continue testing for new significant relationships with that same variable as the data splits. High school GPA could re-appear at any point further down the tree with a new decision rule. The final key setting that a user can change is the maximum depth of the tree. This setting controls the number of generations a tree algorithm

can produce before being forced to stop. The default setting is six, so changing this can create larger trees or smaller trees depending on the user's preference.

4.2 The WKU Decision Tree

To create the decision tree for the student enrollment decision at WKU, we had to decide which variables to include and which not to include. The goal of our list of variables was to avoid anything that would automatically indicate that a student chose to enroll at WKU; i.e., any amount of financial aid actually accepted and Math Placement Exam score. We also had to ensure that our variables used in the decision tree model would be able to also be used in a logistic regression modeling approach (chapter 5) for comparison purposes. Finally, we had to make sure that our variables were easily gleaned from looking at application data, so that the model could be used in admissions, financial aid, and enrollment decisions in the future.

The potential input variables for the decision tree include: high school GPA, the maximum standardized test score, amount of merit scholarships awarded, amount of need-based scholarships awarded, amount of need-based merit scholarships awarded, the first-generation identifier, gender, the number of prior hours earned, the geographical categorical variable, the ethnicity categorical variable, the student's age, and the student's choice of academic college to enter. In this analysis, we separate the data into 50% training and 50% validation because we have the data from the 2016-17 admitted students to use as test data. To create this decision tree, we selected a maximum branch of five, to only use input variables one time, and to select a maximum depth of six. The results of the tree are displayed in Figure 4.2.

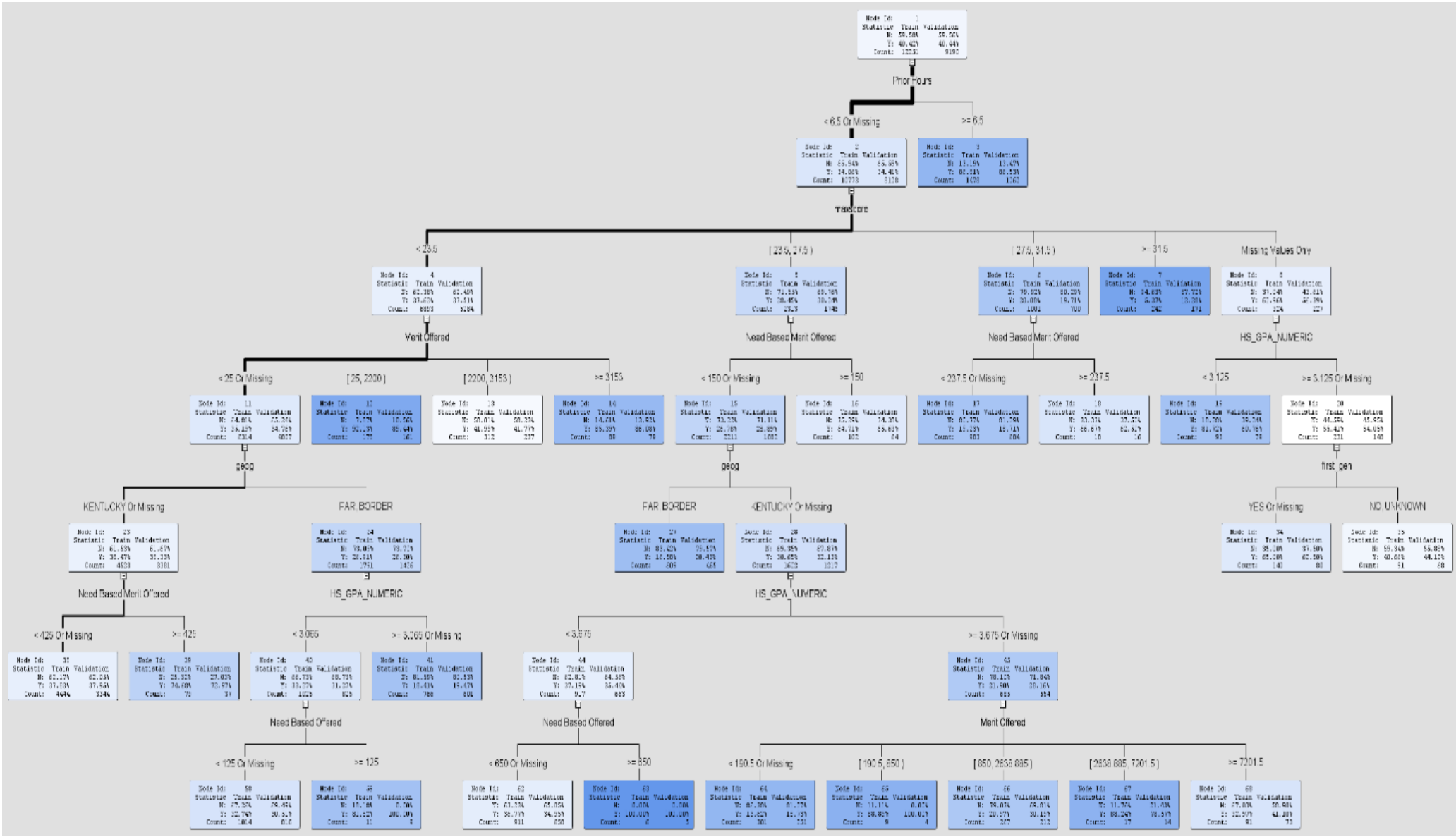

Figure 4.2: The WKU Student Enrollment Decision Tree Model

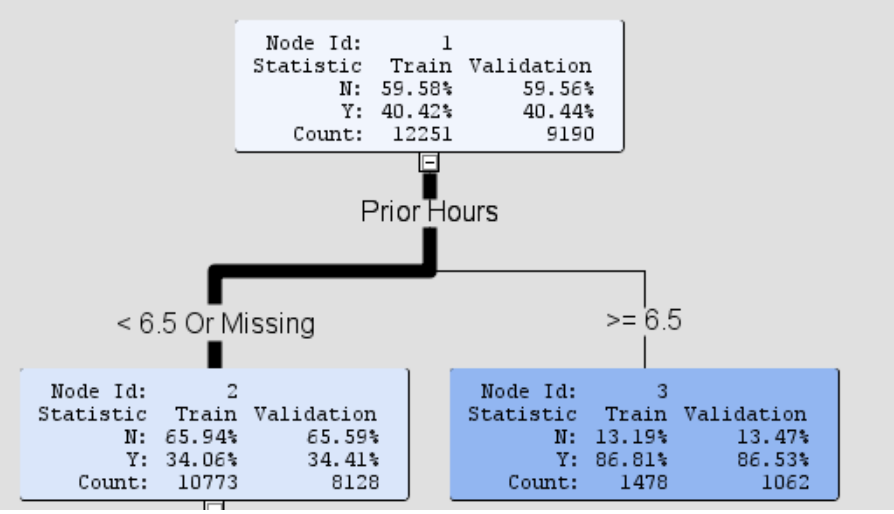

Figure 4.3: Decision Tree Model, Generation 1

Figure 4.2 is displayed in more detail in Figures 4.3 to 4.9. Figure 4.3 displays the root node of our decision tree model and also its first decision rule. The root node, which is the first box, summarizes the response variable in our data separated into two sets, the training data and the validation data. These two sets of data display the basic equilibrium of the WKU enrollment situation. About 40% of admitted students actually enroll at WKU, and nearly 60% choose not to attend this institution. As shown in Figure 4.3, prior hours (the number of credit hours the student earned prior to attending WKU) is the first split, meaning it is the most valuable predictor out of the potential input variables considered. Students with more than 6.5 prior hours chose to enroll at WKU just over 86% of the time. This statistically significant relationship between prior hours and the student enrollment decision does not split any further; no other predictor variables were found to be significant for this group. The students with fewer than 6.5 hours of prior credit chose to enroll at only

a 34% rate, which is smaller than the overall percentage of students that chose to enroll at WKU. This type of split is exactly why decision trees were created; that is to observe if groups make decisions differently and then predict future decisions based on the prior data.

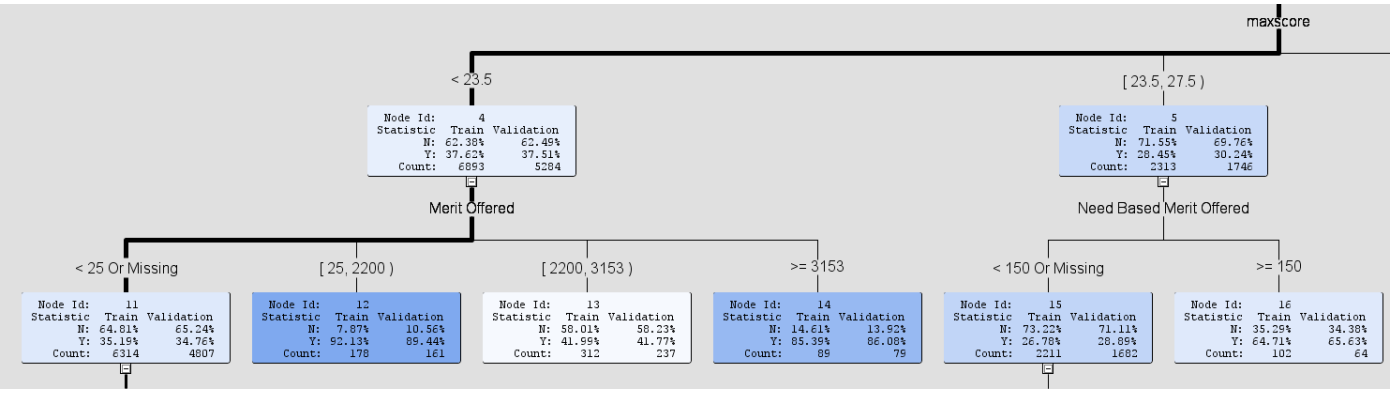

Figure 4.4: Decision Tree Model, Generations 2 and 3A

Figure 4.4 displays the left side of the second and third generations of the final decision tree model. This split in the dataset is for the students who have earned fewer than 6.5 hours of prior credit. The decision rule that is the next most significant predictor is the maximum standardized test score (denoted as “maxscore”). This means that the maximum test score has a statistically significant interaction effect with the number of prior credit hours earned for students that have fewer than 6.5 hours of prior credit earned. Students accepted to WKU that score lower than a 23.5 (or equivalent) on their standardized test, while also earning fewer than 6.5 hours of prior credit, chose to enroll at WKU at a similar rate to our overall training population. As the test score increases, the percentage of students accepted to WKU that chose to enroll at WKU decreases. When examining the interaction, this makes logical sense. If a student is receiving fewer than 6.5 hours of prior college credit but has a high standardized test score, that student is most likely going to

enroll at a different university. In generation three, the students with the lowest test scores have a splitting rule based on the amount of merit scholarships offered. If these students, who have earned fewer than 6.5 hours of prior credit and lower than a 23.5 maximum test score, get merit scholarship opportunities, they are more likely to enroll at WKU than any random student from the training population. Students who have a maximum test score between 23.5 and 27.5 split based on need-based merit scholarships being offered, and they are more likely to enroll at WKU if they receive an offer from the need-based merit pool of funding. These findings are also intuitive; students that are not able to secure prior credits and have mid-range test scores but find a way to secure merit or need-based merit funding at WKU will likely enroll here because this may be the only offer of financial aid on the table. Students respond to changes in net price, and this aspect certainly changes the net price of attendance for these specific students.

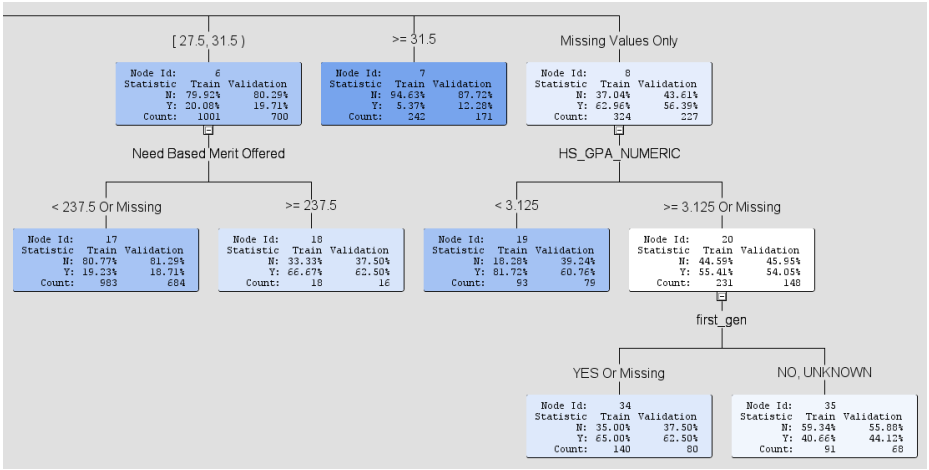

Figure 4.5: Decision Tree Model, Generation 2, 3, and 4B

The significant splits in Figure 4.5 are in two separate areas. The node displaying the statistically significant split in need-based merit scholarship amounts offered comes from the subset of students that earned fewer than 6.5 hours of prior credit and a maximum standardized test score between 27.5 and 31.5. Although not many students received need-based merit scholarships above $250, the ones that did enrolled at WKU around 66% of the time, which is a massive increase relative to the rest of this subset of the student population.

The other area where a split occurs is the subset of students that have not taken a standardized test (the top right node in Figure 4.5). Those particular students already enroll at WKU in a higher proportion than the general accepted-student population. The split, which occurs at the 3.125 mark in high school GPA, displays about a 26% gap between the two groups of students in the training datasets. This split displays the fact that even within this area of students, those who achieve highly in the high school classroom are more likely to choose not to enroll at WKU. A further split for the students who are on the higher end of this spectrum is also displayed. It represents students with fewer than 6.5 hours of prior credit earned, no standardized test score, a high school GPA above 3.125 or a missing value, and then separates based on the first-generation college student identifier. Students that are first-generation college students with these characteristics enroll around 60% of the time, while students that are not enroll at about the same percentage as the overall admitted student population does at WKU.

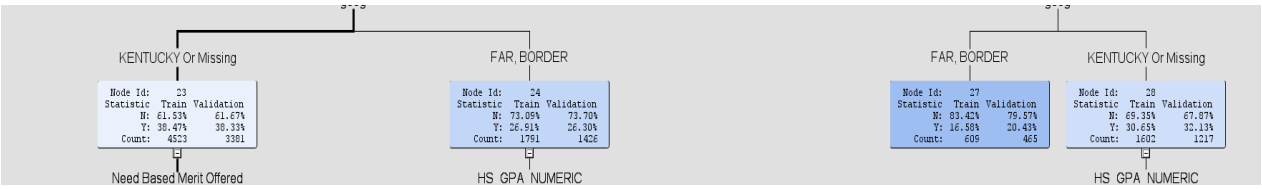

Figure 4.6: Decision Tree Model, Generation 4A

Figure 4.6 spawns from previous nodes that include students with fewer than 6.5 prior credit hours, lower than a 23.5 (or equivalent) on the standardized test, and less than $25 in merit scholarships offered (on the left of Figure 4.6), or they have fewer than 6.5 prior hours, between a 23.5 and a 27.5 on the standardized test, and they have received less than $150 in need-based merit scholarships (on the right of Figure 4.6). The students are further subdivided based on their geographic areas, and unsurprisingly, students with these characteristics who are from Kentucky continue to enroll at WKU at roughly the same rate as the overall admitted-student population if they fall below a 23.5 on the standardized test and are grouped by merit scholarship dollars offered. If these students are instead slightly higher achieving and grouped by need-based merit scholarship dollars offered, they enroll only around 30% of the time. In both cases, students not from Kentucky experience a drop in the likelihood of enrolling at WKU. In the case of the students who are below a 23.5 on their standardized test, the drop is around 10%, whereas the students in the higher areas of academic achievement experience a 15% decrease in likelihood of enrolling at WKU.

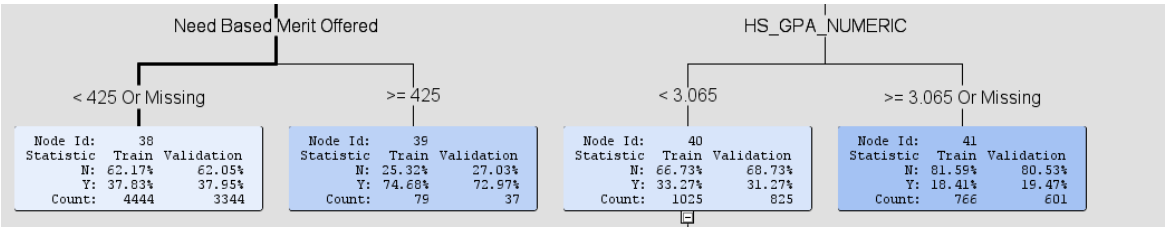

Figure 4.7: Decision Tree Model, Generation 5A

Continuing into the fifth generation of the tree, the splits in Figure 4.7 come from the nodes which grouped students with fewer than 6.5 hours earned, lower than a 23.5 on the standardized test, less than $50 of merit scholarships earned, and then either being from Kentucky or not from the state. Those students who are from the state of Kentucky (on the left side of Figure 4.7) are further split by the amount of need-based merit scholarship dollars offered. In this case, students who are offered $425 or more are much more likely to enroll than the students who do not receive this much funding. The out-of-state students (on the right side of Figure 4.7) are instead split by high school GPA. Continuing the trend, the more successful a student is in the high school classroom, the less likely s/he is to enroll at WKU.

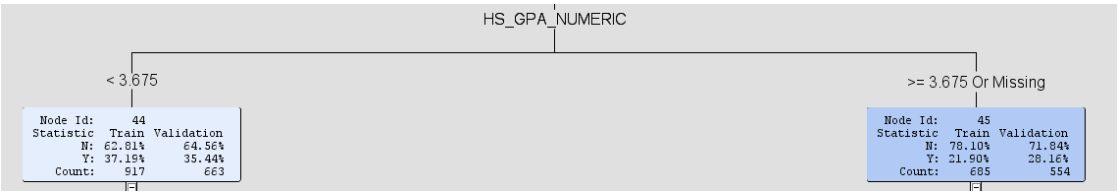

Figure 4.8: Decision Tree Model, Generation 5B

Figure 4.8 comes directly from the node describing Kentucky students with less than $425of need-based merit scholarships offered, between a 23.5 and 27.5 on the standardized test, and fewer than 6.5 hours of prior credit earned (from Figure 4.6). This split again shows the trend that students who demonstrate more success in the high school classroom are less likely to choose to attend WKU.

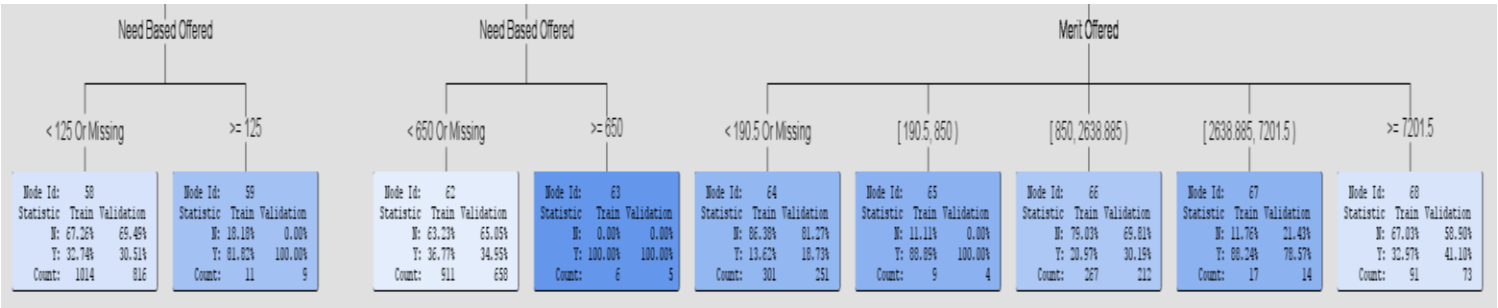

Figure 4.9: Decision Tree Model, Generation 6

Figure 4.9 is the last figure depicting a section of our decision tree model. As the modeling process gets to this point, one will observe that we do not have many students left available in our subset populations. The largest nodes at this point contain only around 1900 students total (training and validation combined) after beginning with nearly 22000, and the smallest nodes in this section contain fewer than 15 total observations. As such, it is difficult to rely on this section to learn anything necessarily new about the dataset. However, we can continue to observe patterns. Each of these nodes has to do with the amount of funding a student receives, either through need-based, need-based merit, or merit-based scholarships. As the amount a student is offered increases, we see an increased likelihood in that student enrolling at WKU.

To summarize the decision tree findings, the most important statistical indicator of student enrollment decision is the student's number of prior credit hours earned. If a student who is admitted as a first-time, first-year freshman has earned more than 6.5 hours of credit, that student is much more likely to enroll (over 86%) than the average admitted student in his or her class (approximately 40%). The second key conclusion we can take from this model is that increased academic success, whether it is through high school GPA or the standardized test score, decreases a student's likelihood of enrolling at WKU if that student has earned fewer than 6.5 hours of prior credit. In section six, we will further discuss the predictive performance of this model when used to model the enrollment decisions of the 2016-17 first-time, first-year freshman class, and also compare it to the logistic regression model presented in chapter five.

CHAPTER FIVE: LOGISTIC REGRESSION

5.1: Logistic Regression Analysis

Logistic regression was the second modeling approach utilized to analyze this dataset. This type of regression modeling is typically used to model the relationship of a binary response variable and several predictor variables (Mendenhall & Sincich, 2012). The response variable is coded as 1 if the response is a success and 0 if the response is a failure. Since the response variable in least-squares linear regression is not restricted to be between 0 and 1, it is not appropriate to use when the response is binary.

The logistic regression model ensures that the estimated response variable lies between 0 and 1, so it provides an appropriate model for the student enrollment decision at WKU. It does this by utilizing the odds and log-odds of an event occurring. The odds of an event occurring are defined as

$$\mathbf{odds} = \frac{\boldsymbol{\pi}}{\mathbf{1}-\boldsymbol{\pi}} = \frac{\boldsymbol{P(y=1)}}{\boldsymbol{P(y=0)}}, \tag{5.1}$$

and the log-odds are defined as

$$\text{log-odds} = \mathbf{ln}\left(\frac{\boldsymbol{\pi}}{\mathbf{1}-\boldsymbol{\pi}}\right) = \mathbf{ln}\left(\frac{\boldsymbol{P(y=1)}}{\boldsymbol{P(y=0)}}\right), \tag{5.2}$$

where π is the true probability of a successful response. When there are k predictor variables, we fit the model

$$\mathbf{ln}\left(\frac{\boldsymbol{\pi}}{\mathbf{1}-\boldsymbol{\pi}}\right) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \boldsymbol{\beta_2 x_2} + \cdots + \boldsymbol{\beta_k x_k}, \tag{5.3}$$

which looks very similar to a multiple linear regression model. As $\pi \to 0$, the log-odds$\to -\infty$; as $\pi \to 1$, the log-odds$\to \infty$. Thus, the response is no longer restricted to be between 0 and 1. Using algebraic operations, (5.3) can be rewritten in the general logistic regression model form,

$$\boldsymbol{E(y) = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}}, \quad (5.4)$$

where the expected value of y, E(y), is equal to the probability of a successful response occurring, π, which is restricted to a value between 0 and 1 (Mendenhall & Sincich, 2012).

As shown in (5.4), the general logistic regression model is not a linear function of the β_i parameters. Instead, estimates of the parameters are obtained by a method of maximum likelihood estimation. Since this is not a linear estimation, one has to be careful interpreting the coefficients in the regression equation. In the case of least-squares linear regression, the estimate of β_i, denoted $\boldsymbol{\hat{\beta}_i}$, quantifies the estimated change in the expected value of the response variable for each one-unit increase in x_i, assuming all other predictor variables are held constant. However, this is not the way that $\boldsymbol{\hat{\beta}_i}$ can be interpreted in logistic regression. The β_i parameter in a general logistic regression model would represent the additive change in log-odds for every one-unit increase in x_i, assuming all other variables are held constant (Mendenhall & Sincich, 2012). To make this more interpretable, consider $\boldsymbol{e^{\beta_i}}$, the odds ratio. This represents the multiplicative change in the odds of an event for every one-unit increase in x_i, assuming all other variables are held constant. If x_i is binary, then $\boldsymbol{e^{\beta_i}}$ represents the ratio of the odds when $x_i = 1$ to the odds when $x_i = 0$; i.e., the odds ratio between these two groups.

Since our goal is to estimate the probability of a successful response (π), it can be helpful to consider how the value of β_i is related to π. Assuming all other variables are held constant, if $\boldsymbol{e^{\beta_i} > 1}$, then π increases as x_i increases. If $\boldsymbol{e^{\beta_i} < 1}$, then π decreases as x_i increases. If $\boldsymbol{e^{\beta_i} = 1}$, then π does not change as x_i increases.

5.2: WKU Logistic Regression Model

We performed logistic regression modeling in SAS Enterprise Miner with a very similar process to the Decision Tree Model. A pre-built logistic regression process is included with the software, so we use that node to perform our logistic regression analysis. In order to accurately characterize the WKU enrollment decision and to make valid comparisons between models, we chose to use the same predictor variables in the logistic regression model that were outlined previously in the decision tree modeling process. However, some changes needed to be made. For each of the qualitative predictors, "dummy" or indicator variables had to be created. For a qualitative predictor with j categories, $j - 1$ dummy variables were created; each of these indicates whether or not an individual falls into a particular category. For example, consider the geographical variable, which has three categories: Kentucky, border state, and other (called "far"). Since there are three categories, two dummy variables were created. For the first dummy variable, a student is assigned a value of 1 if s/he is from Kentucky and a 0 otherwise. For the second dummy variable, the student is assigned a value of 1 if s/he is from a bordering state and a 0 otherwise. A third dummy variable for students from elsewhere is not needed; values of 0 for both of the already created dummy variables indicate that a student falls into the third category. All variables utilized in the logistic regression model are defined in Table 5.1.

i	x_i
1	1 if College of Education and Behavioral Sciences, 0 otherwise
2	1 if College of Health and Human Services, 0 otherwise
3	1 if Gordon Ford College of Business, 0 otherwise
4	1 if Ogden College of Science and Engineering, 0 otherwise
5	1 if Potter College of Arts and Letters, 0 otherwise
6	number of prior hours earned at WKU
7	amount of merit scholarship dollars offered
8	amount of need-based merit scholarships dollars offered
9	amount of need-based scholarship dollars offered
10	student age
11	high school GPA
12	maximum standardized test score
13	1 if student is from Kentucky, 0 otherwise
14	1 if student is from a bordering state, 0 otherwise
15	1 if Asian, 0 otherwise
16	1 if black, 0 otherwise
17	1 if Hispanic, 0 otherwise
18	1 if Hawaiian or Pacific Islander, 0 otherwise
19	1 if multiple races, 0 otherwise
20	1 if Native American, 0 otherwise
21	1 if unknown race, 0 otherwise
22	1 if white, 0 otherwise
23	1 if first-generation college student, 0 otherwise
24	1 if male, 0 otherwise

Table 5.1: Predictor Variables in the Logistic Regression Model

The resulting logistic regression equation is

$$\ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \mathbf{0.4346} - \mathbf{0.0374}x_1 - \mathbf{0.1099}x_2 - \mathbf{0.0247}x_3 - \mathbf{0.1717}x_4 +$$

$$\mathbf{0.0296}x_5 + \mathbf{0.1866}x_6 + \mathbf{0.000058}x_7 + \mathbf{0.000943}x_8 + \mathbf{0.00161}x_9 +$$

$$\mathbf{0.1069}x_{10} - \mathbf{0.3672}x_{11} - \mathbf{0.1243}x_{12} + \mathbf{0.8010}x_{13} + \mathbf{0.2326}x_{14} + \mathbf{0.1336}x_{15} -$$

$$\mathbf{0.1824}x_{16} - \mathbf{0.2887}x_{17} + \mathbf{0.5773}x_{18} - \mathbf{0.0675}x_{19} - \mathbf{0.4102}x_{20} - \mathbf{0.1919}x_{21} +$$

$$\mathbf{0.1127}x_{22} - \mathbf{0.2262}x_{23} + \mathbf{0.0545}x_{24}, \qquad (5.5)$$

where x_1 through x_{24} are as defined in Table 5.1.

Of these variables, only age, high school GPA, the in-state dummy variable, all three levels of financial aid, the first-generation college student variable, the maximum standardized test score, and number of prior hours earned are statistically significant at a 0.01 significance level.

As such, the effects of a one-unit increase in these statistically significant independent variables (holding all else constant) on the binary student enrollment decision are as summarized in Tables 5.2 and 5.3.

Variable	Effect on Odds of Enrolling
student age	11.28% higher
high school GPA	30.73% lower
amount of merit scholarship dollars offered	0.000058% higher
amount of need-based merit scholarships dollars offered	0.000943% higher
amount of need-based scholarship dollars offered	0.00161% higher
maximum standardized test score	11.69% lower
number of prior hours earned at WKU	20.51% higher

Table 5.2: Effects of a One-Unit Increase in Statistically Significant Variables

Variable	Effect on Odds of Enrolling
student from Kentucky	122.78% higher
first-generation college student	20.24% lower

Table 5.3: Effects of a Categorical Change in Statistically Significant Variables

As an example, consider the student age (x_{10}). The estimated logistic regression coefficient for this variable is 0.1069, as shown in (5.5). Since $e^{0.1069} - 1 = 0.1128$, each one-year increase in student age increases the odds of the student enrolling at WKU by 11.28% (assuming all other variables are held constant). As a second example, consider the maximum standardized test score (x_{12}), which has an estimated coefficient of -0.1243. Since $e^{-0.1243} - 1 = 0.1169$, each one-point increase in maximum standardized test score

decreases the odds of the student enrolling at WKU by 11.69% (assuming all other variables are held constant). Now consider the categorical dummy variable x_{13}, which has an estimated coefficient of 0.8010. Since $e^{0.8010} - 1 = 1.2278$. The odds of a student from Kentucky enrolling at WKU are 122.78% higher than the odds of a student not from Kentucky enrolling at WKU (assuming all other variables are held constant).

The logistic regression model demonstrates some general trends within the enrollment situation at WKU with a look to specific discrete changes in independent variables. One way to easily characterize these trends is to examine the odds ratio for each statistically significant parameter. For instance, an increase of 1 point in high school GPA decreases the probability of a student actually enrolling at WKU if every other independent variable is held constant because $e^{-0.3672} < 1$. A similar negative effect is seen with a 1-point increase in the maximum standardized test score. This demonstrates a general struggle to recruit academically successful students (when they are equivalent to other students in all other categories).

Since the standard errors vary among the coefficient estimates, they cannot be directly compared to investigate the relative strength of the predictors within the model. In order to accurately compare the magnitude of the expected change in a student's probability of enrolling, one must examine the standardized coefficients of the logistic regression model. Standardized coefficients eliminate the units of the variables and are all on the same scale, allowing us to observe which predictor variables are associated with the largest magnitude change in our response variable. The higher the absolute value of the standardized coefficient, the greater the relative strength of the predictor. The standardized

coefficients for the significant predictors in the logistic regression model are displayed in Table 5.4.

Variable	Standardized Coefficient
student age	0.0488
high school GPA	-0.1125
amount of merit scholarship dollars offered	0.1330
amount of need-based merit scholarships dollars offered	0.2098
amount of need-based scholarship dollars offered	0.1598
maximum standardized test score	-0.3209
number of prior hours earned at WKU	0.6866
student from Kentucky	0.1980
first-generation college student	-0.0606

Table 5.4: Standardized Coefficients of Statistically Significant Variables

Another general trend that we can deduce from the logistic regression results is the overall increase in the odds of enrolling when a student's financial aid package is increased, regardless of whether that increase comes from need-based, merit-based, or need-based merit scholarship funds. As shown in Table 5.4, the magnitude of the change in odds is larger for need-based scholarships and need-based merit scholarships than it is for exclusively merit-based scholarships, which demonstrates WKU's ability to recruit students with demonstrated financial need better than students who are offered merit scholarships.

Table 5.4 highlights a key distinction in students who are successful in the classroom in high school. The standardized coefficient on the maximum test score variable (-0.3209) is nearly triple the standardized coefficient on high school GPA, (-0.1125). This means that the decrease in the odds of enrolling that is associated with an increase in the maximum test score is much larger than the decrease associated with an increase in GPA.

WKU struggles generally with converting highly successful admitted students, but the students with higher standardized test scores are the demographic WKU struggles with the most to convert into enrolled students.

The standardized coefficient for the number of prior credit hours earned is much larger (in absolute value) than any other value in Table 5.4. This finding demonstrates WKU's ability to convert admitted students who have earned credit from the University before pursuing admission as a first-time, first-year student on campus. WKU also effectively converts admitted students from the state of Kentucky into enrolled students quite well, but on the flip side this result displays a lack of ability to effectively recruit admitted students from outside the state to actually enroll at WKU. The University also struggles with converting first-generation college students from admitted to enrolled students. WKU typically does a good job of converting admitted students of higher ages to first-time, first-year college enrollees, but as displayed in Table 5.4, this is the smallest standardized coefficient of any statistically significant variable in the model.

The standardized coefficients and the odds ratios are helpful estimators for determining important characteristics to look for when it comes to increasing enrollment. First and foremost, examining Tables 5.2-5.4 reveals that WKU does an excellent job recruiting students from the state of Kentucky, students with more hours of prior credit earned at WKU, and students with demonstrated financial need. Some areas where the institution could put a priority on improving with the goal of lifting enrollment numbers include out-of-state students, students who are high achieving in the high school classroom, and first-generation college students.

It is important also to recognize that this model does not predict large changes in the odds of enrolling for students who are exactly alike in every way except for the amount of merit scholarships they were offered. This suggests that a change in the merit scholarship plan to give more students money, while decreasing the amount of funding given to students at the top of the academic performance metrics, may actually increase enrollments.

CHAPTER SIX: PREDICTIVE PERFORMANCE

Within each model, there are different indicators of model performance in a predictive setting. The average squared errors are given as part both models' outputs in SAS Enterprise Miner. For the decision tree, this value was 0.19, and for the logistic regression model it was 0.20. This would suggest that the decision tree modeling approach is perhaps slightly more accurate in predictive measures. We put this claim to the test by creating a prediction from each model based on the 2016-17 admitted student dataset and then comparing that prediction to the actual student's enrollment decision. This prediction is created by assigning a probability of enrolling to each student in the dataset using the characteristics defined in the modeling procedure. In decision trees, the program finds the correct node from Figure 4.2 for each student and then assigns the probability of enrollment from this bin to that student. In logistic regression modeling, the program inserts the actual values of the variables for each student into (5.5), and solves the resulting value for $\widehat{\boldsymbol{\pi}}$ to create a predicted probability of enrollment. If the predicted value is greater than 0.5, the student is assigned as a "yes," and if the predicted value is less than 0.5, the student is assigned as a "no." Figure 6.1 displays the results of the predictions of the decision tree model, and Figure 6.2 displays the results of the predictions of the logistic regression equation.

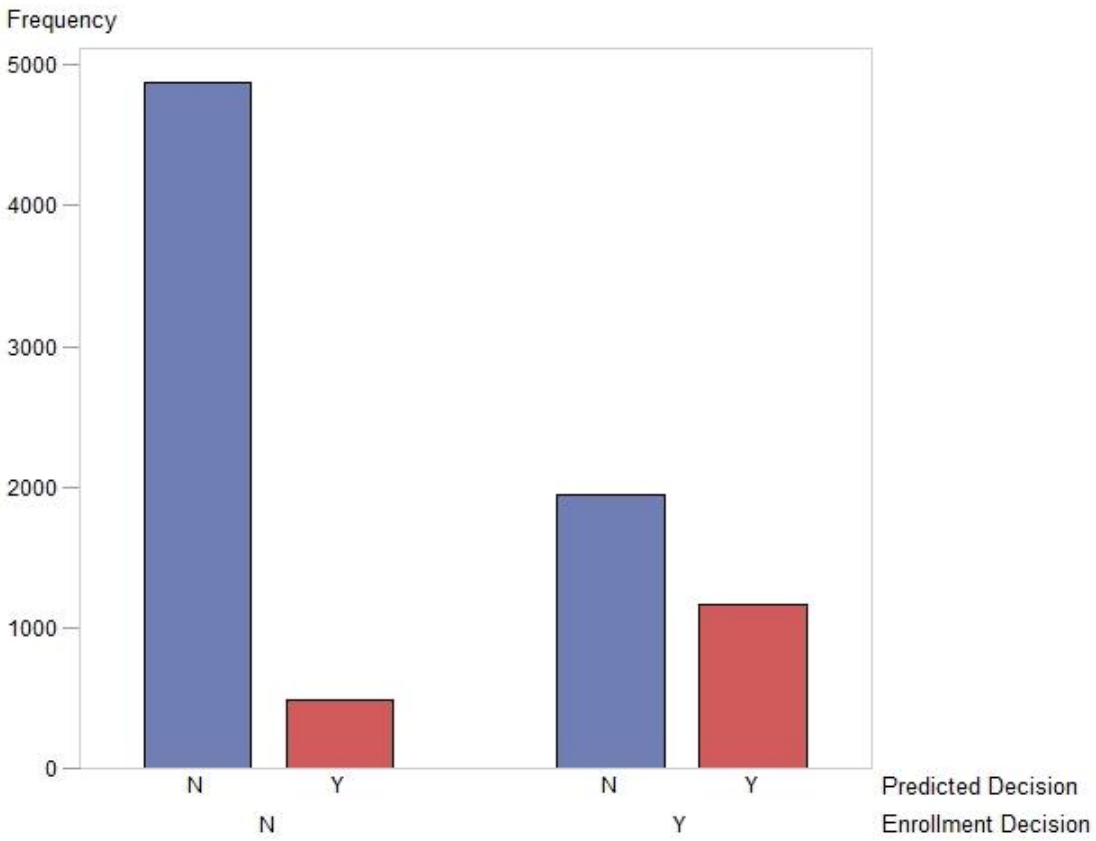

Figure 6.1: Decision Tree Model Predictions

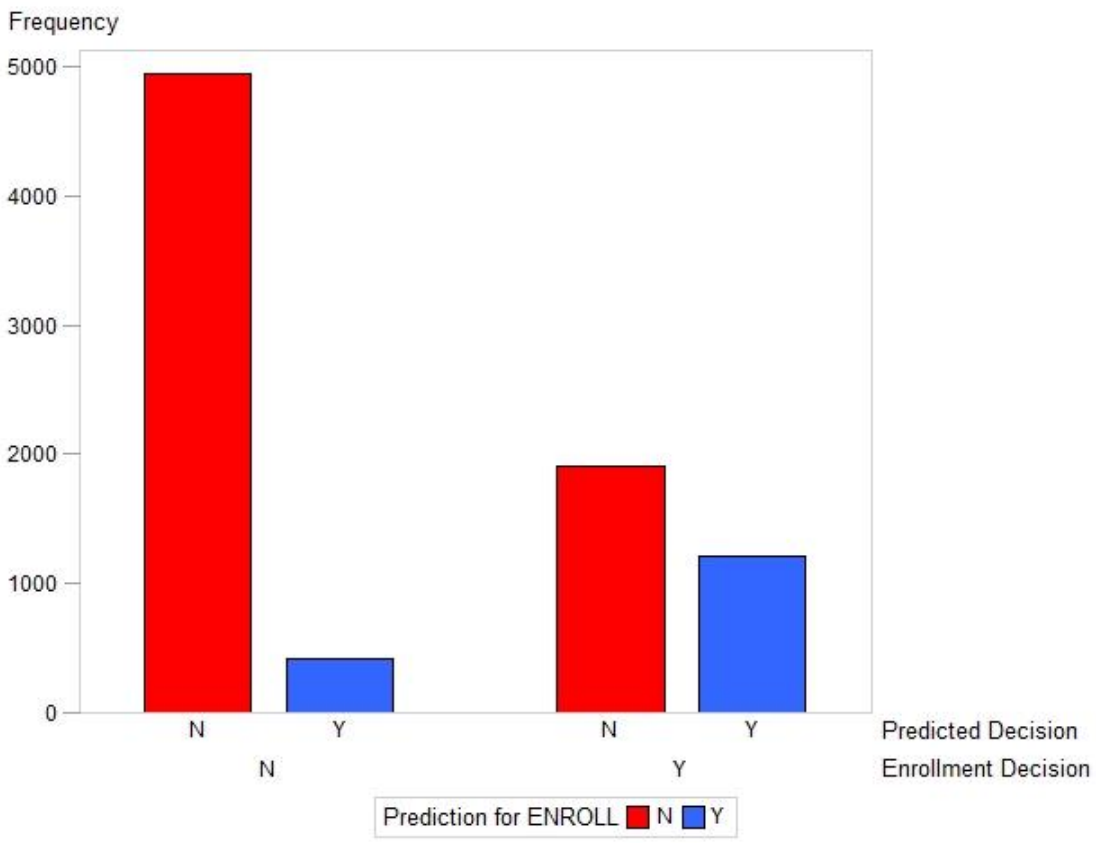

Figure 6.2: Logistic Regression Predictions

In both figures displayed above, the students who chose not to enroll at WKU are in the two left-most bars in the graph. The students in the right half of the graph are those who enrolled at WKU as first-time, first-year freshman. Within the groups, the students who are not predicted to enroll at WKU are on the left, and the students who are predicted to enroll are on the right. To truly compare the two models, we need to split the data into four categories: students who are predicted to enroll at WKU and do matriculate (a true positive), students who are predicted to enroll and do not matriculate (a false positive), students who are predicted not to enroll at WKU that do matriculate (a false negative), and students who are predicted not to enroll at WKU that do not matriculate (a true negative).

The false positive rate is around 5%. This can be seen through examining the students who did not choose to enroll at WKU that were predicted to enroll. However, because of the relative strength of correctly predicting students to not enroll when they actually do not enroll at WKU, both models perform poorly when predicting the number of admitted students who will actually enroll at WKU. The rate of a false negative is around 40%, as displayed in Figures 6.1 and 6.2. Both models classified over half of the class that actually enrolled at WKU as not likely to enroll. This suggests that perhaps many of the students that chose to enroll at WKU for the academic year of 2016-2017 did not make similar decisions to students with similar characteristics from the previous classes of first-time, first-year students that were used to create these models. This could be due to year–to-year variation, but we also must consider that there was a potential change in student behavior based on the change in merit-based financial aid strategies that occurred just before these students were recruited. This must be quantified in the future in order to fully understand the effects of this change in the long-term modeling process.

In all of the predictive areas, the logistic regression model performs slightly better than the decision tree model. However, there are benefits to using both of these models alongside one another. The logistic regression model is easily interpretable, allowing one to estimate the change in odds of enrolling associated with changes in the values of the predictor variables. This information can be useful, for example, to admissions counselors when communicating with students about the opportunity to earn collegiate credit at WKU before finishing their high school curriculum. The decision tree model allows WKU Admissions staff to quickly categorize a student into a group and assign a rough probability of enrollment. It can be used by the staff members to quickly decide which groups of

prospective students need to receive important information about WKU, and this information could be tailored so that all students with similar probability receive information designed to increase their likelihood of matriculation.

CHAPTER SEVEN: DISCUSSION

In summary, decision trees and logistic regression are both tools used to inform predictive analytics. Using either model presented in this paper, one can determine that the key variables that are useful in predicting enrollment at WKU include the number of prior credit hours earned, the maximum standardized test score, the amount of merit scholarship dollars offered, the amount of need-based merit scholarship dollars offered, the amount of need-based scholarship dollars offered, state residence, age, and high school GPA. When these models are completed, measures of fit such as the average squared error and misclassification rate demonstrate that both models produce similar predictions, with the logistic regression model performing slightly better. Both models lean towards predicting that students will not choose to enroll at WKU. This follows the pattern we have seen across the data which has shown that only around 40% of admitted students choose to enroll at WKU.

There are several key applications to take from this research with an eye toward increasing enrollment at WKU. This analysis demonstrates that students are more likely to attend WKU if they have earned more than 6.5 hours of credit prior to attending. There are also increased odds of attendance if a student is from the state of Kentucky and if the student is receiving some type of financial aid. However, WKU can clearly improve its yield rate with out-of-state students and with students who are more successful in their high school performance.

To begin in immediate applications for this data, WKU can recognize the limited number of students who are predicted to enroll. Since this number is much smaller than the

total number of applicants, the University can specifically design a communication structure that delivers these select students through the acceptance and enrollment stages of their college decision so that they ultimately end up choosing to become a WKU student. A student that is predicted to enroll that actually enrolls is one that the University can count on in budgeting strategies, housing decisions, and classroom offerings, all of which have a great deal of variables in play. Consistently delivering on projections can help WKU better deal with the rest of its programming in a more efficient way. This strategy can also help to identify why students who are predicted to enroll choose not to do so. WKU is not the right institution for everyone, so the more information the institution can gather from all of its admitted students, the better the predictive modeling will be in the future.

Another immediately applicable aspect of this data lies in analyzing all of the students who are not predicted to enroll at WKU. One limitation of these models is that many of these students who are not predicted to enroll at WKU actually do enroll at the institution. A little under 28% of admitted students are misclassified in this way according to our predictive models. We can be fairly confident that when the predicted probability of enrolling is close to zero, WKU should spend less resources on communicating with these students than others because these students are more likely to not enroll at WKU. However, our model currently leaves some gray area where WKU could learn more about this particular group of students' decisions while also reaping the benefits of a false negative prediction by actually enrolling the student on campus.

Overall, the impact of merit-based scholarship aid is small, but it does exist. Especially if a student is not on the highest end of the high school academic performance spectrum (a standardized test score between 23 and 27 and a GPA less than 3.6 for

instance), providing this student with a small merit-based scholarship to decrease the sticker price of their education could be beneficial to WKU in the long run. If a student performs more successfully in high school, WKU may need to highlight the increased amount of credit those students can earn with a few additional classes or tests before officially enrolling at the University.

The increased odds of enrolling from an increase in need-based aid or need-based merit aid are magnitudes larger than the merit-based consideration, and could also provide WKU with the marginal ability to improve its enrollment situation if the university is able to recoup some of its losses from the state of Kentucky.

If WKU pursues these changes, the long-run proportion of students who are accepted that choose to enroll at WKU may increase from 40%. This increase in conversion rate of accepted students into enrolled students will help stabilize the institution's budget in the long-run, especially considering the continually decreasing assistance from the state and increasing costs of higher education across the nation.

There are some caveats to the application of this research. Throughout this process, we did not do anything to prove that WKU has the ability to cause a student to change from a "no" to a "yes" in their enrollment decision. Although relationships were determined, we cannot classify those relationships as being causal since observational (rather than experimental) data was used. As such, we have to be very careful when it comes to the applications. In reality, WKU could do everything "correctly" and still have students decide not to enroll at this institution. The research we have done allows anyone with access to the application data to develop a predicted probability of enrollment that could update as a student enters more information into the WKU application. This prediction will never

be perfect, but it can be used to better inform decisions made by WKU in spending issues, admissions team contact, departmental involvement, and a whole host of other institutional factors entirely in WKU's control.

To advance upon this research, and continue to build on the study completed by Bogard in 2013, WKU needs to combine this application data with some information about admissions contact, collect more data on the effects of changing the merit scholarship award process, and connect it all to retention and persistence.

I suggest combining information from the Division of Enrollment Management, the Office of Student Financial Assistance, and the Office of Institutional Research together with data from the Office of Admissions so that we have a more holistic picture of each individual student. The models presented in this paper still predicted false negatives for over half of the 2016-17 first-time, first-year freshman class (as seen in the right-hand sides of Figures 6.1 and 6.2). Clearly, there is more variation within the student enrollment decision that our dataset could not accurately measure. Adding admissions data, such as contact with a counselor, a high school visit, a tour completed on campus, a view book sent, etc., will most likely add more valuable information to improve the predictive accuracy of this data. Another key piece of additional information that the University could pursue in order to better predict student behavior is to follow up with students that chose not to enroll at WKU. If we had information about why a student chose to attend a different college, it would help to determine comparative strengths and weaknesses of WKU relative to its competitors in higher education. Both pieces of additional information can help increase the predictive accuracy of the model.

A second way to advance this research is to continue to study the effect of changing the merit scholarship program. Drawing from my own personal experience as a tour guide for the Honors College, I know that there are students who have chosen not to enroll because of the changes in the merit scholarship program. I have no doubts that there are students who have decided to enroll at WKU because of this change as well. This topic needs to be fully investigated to determine if the change had the desired outcome on enrollment. If it did not have that effect, then learning more about this particular aspect of the enrollment decision will continue to provide valuable information to WKU so that the next change to the scholarship program will be a successful one.

Finally, WKU needs to connect this data into more information about persistence to graduation. Much of WKU's budgetary struggles the past few years have stemmed from a small enrollment decline and a much larger problem with retention once a student is actually enrolled here. WKU currently has the data necessary to make predictions on both enrollment and persistence of a student, classify each student into a specific category, and then make decisions with regard to specific classes of students to attempt to head-off the retention problem. Much of the difficulty within this task is combining all of the data into one central location that provides usable information to any staff or faculty member with access to it.

If WKU can successfully further this research by better predicting student enrollment through use of admissions and application data, continued understanding of financial aid's impact on admitted student behavior, and connect all of this information to a model that continuously tracks the probability of persistence to graduation, the benefits are far reaching across several key areas of WKU's campus. First, the serious budget

shortfall problem caused by struggles in enrollment and retention would finally be heading in a beneficial direction for the University. Enrolling more students and keeping more of them present until graduation has positive effects for the students and for the financial aspect of campus as well. Finally, WKU's Admissions and Enrollment teams will be much more efficient in their communication with prospective students and their contact with current students. Modeling behavior is not a perfect solution to all of our recruitment and retention questions, but the results of the models can certainly inform decision makers so that the outcomes are more cost efficient and beneficial for students.

In conclusion, modeling the student enrollment decision at WKU with a decision tree and a logistic regression model led to three key takeaways. First, WKU can improve at recruiting high-achieving high school students, students from outside the state of Kentucky, and first-generation college students. Second, WKU does very well at recruiting students with a high number of credits earned at the institution, students from the state of Kentucky, and students who receive financial aid through either merit-based, need-based, or need-based merit sources. Finally, WKU can improve upon these models and become more data-driven in their strategies implemented across the campus departments by integrating data from applications, admissions, and retention to make decisions focused on solving the budgetary shortfalls and student persistence to graduation.

REFERENCES

Bogard, M. (2013). A Data Driven Analytic Strategy for Increasing Yield and Retention at Western Kentucky. Bowling Green.

de Ville, B., & Neville, P. (2013). Decision Trees for Analytics Using SAS Enterprise Miner. SAS.

Doyle, W. R. (2010). Changes in Institutional Aid, 1992-2003: The Evolving Role of Merit Aid. *Research in Higher Education*, 789-810.

Hogg, R. V., & Tanis, E. A. (2015). *Probability and Statistical Inference.* Pearson Education.

http://www.studypoint.com/ed/sat-to-act-conversion/. (2017). Retrieved from studypoint.com.

Kentucky Center for Economic Policy. (2016). *Kentucky Lottery Funding.* Analysis.

Leppel, K. (1993). Logit Estimation of a Gravity Model of the College Enrollment. *Research in Higher Education*, 387-398.

McPherson, M. S., & Schapiro, M. O. (1991). Does Student Aid Affect College Enrollment? New Evidence on a Persistent Controversy. *The American Economic Review*, 309-318.

Mendenhall, W., & Sincich, T. (2012). Regression Analysis. Boston: Prentice Hall.

Nurnberg, P., Schapiro, M., & Zimmerman, D. (2012). Students Choosing Colleges: Understanding the matriculation decision at a highly selective private institution. *Economics of Education Review*, 1-8.

Perna, L. W. (2000). Differences in the Decision to Attend College among African Americans, Hispanics, and. *The Journal of Higher Education*, 117-141.

SAS. (2017). SAS Support Usage Note 24329: In the SAS® Enterprise Miner Tree node, what is the logworth statistic?

SAS. (n.d.). SAS Enterprise Miner Version 13.2.

van der Klauww, W. (2002). Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-. *International Economic Review*, 1249-1287.

Western Kentucky University. (2016). *Academic Merit Awards.* Retrieved December 6, 2016, from wku.edu.

Western Kentucky University. (2017). *Prior Credit.* Retrieved from wku.edu.