LEVERAGING COMMONLY USED ADMISSION CRITERIA TO IDENTIFY TRADITIONALLY OVERLOOKED APPLICANTS

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF SCIENCE

IN THE GRADUATE SCHOOL OF THE

TEXAS WOMAN'S UNIVERSITY

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

COLLEGE OF ARTS AND SCIENCES

BY

GREGORY GENGO, B.S.

DENTON, TEXAS

MAY 2021

DEDICATION

This thesis work is dedicated to my loving wife, Kate, and precious children, Ben and Sophia, all of whom have helped me in more ways than they can possibly imagine. Thank you for your patience and love. I am incredibly fortunate to have them in my life. This work is also dedicated to all of my parents, who have loved and supported me unconditionally and played a role in nurturing me so that I could pursue this goal.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Mark S. Hamner, for his encouragement, support, and enduring patience in pushing me across the finish line. It was in large part because of his belief in my abilities that this thesis was completed. His enthusiasm was infectious and always helped keep me on track.

I also wish to thank my committee members, Dr. Don Edwards and Dr. Brandi Falley, for generously offering their support. Also of note is Dr. Edwards' support for the 13+ years that I have known him, and in helping me navigate some of the bureaucracy inherent in the process.

Finally, I would like to thank my wife and children for their patience, love, support, and for understanding why I was writing and researching so much. For my wife Kate, listening to me explain what I was working on was probably tiresome, but she supported me throughout the entire process and during the end stretch, helped by taking on so much responsibility. I am thankful for this support and love.

ABSTRACT

GREGORY GENGO

LEVERAGING COMMONLY USED ADMISSION CRITERIA TO IDENTIFY TRADITIONALLY OVERLOOKED APPLICANTS

MAY 2021

Universities have long had criteria that must be met in order to be admitted to the institution. The purpose of admissions criteria is to determine students who are well suited to begin their academic journey at the university level. The threshold level that these criteria must meet varies across the university landscape, but most, if not all, universities have some commonality with regard to the actual criteria that they require for consideration. In the course of this study, there was no significant literature that addressed holistically how the threshold levels, against which applicants are measured, were determined. Therefore, this study's goal was to determine a statistical framework for which any university can determine these levels to better identify applicants that are likely to be successful as students. A logistic regression model was built using an innovative dependent variable, to predict the probability of an applicant accruing a specific number of semester credit hours in a given time period, the specifics of which is discussed throughout this study. Because of the predictive accuracy, this model can serve as a framework for predicting the likely success of an applicant as a student.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Introduction

While the impact of the COVID-19 pandemic caused a significant decrease in the number of first-time postsecondary student enrollment in the fall of 2020, there are still over 2 million of these students estimated to have enrolled across the United States for that semester (National Student Clearinghouse Research Center, 2020). Each of these students had to submit an application to the institution of their choice. In some cases, a student will apply to multiple institutions. At each institution, these applications must be processed before a decision can be made about whether or not to accept or deny a student's application.

Admission criteria vary at different institutions. Each institution establishes the requirements to determine their own population's readiness. They set their criteria, for what is assumed some vetted internal process that gives them insights into what it takes to succeed at their respective institution. However, research for this study did not find found any descriptions of this internal process as it relates to the overall admission of applicants without regard for specific majors when searching the literature. In general, there are common elements used to determine admission across many institutions. For example, the prescribed admissions criteria for Texas Woman's University (TWU) relies fundamentally on three variables: an applicant's high school grade point average (GPA), an applicant's high school class rank, and entrance assessment scores (ACT or SAT)

submitted by the applicant (Texas Woman's University, 2020). However, not all institutions utilize the variables in the same way. For example, it has become more popular for institutions to omit the entrance assessment scores, and focus more on utilizing fewer variables, like high school GPA, in an effort to increase diversity of their incoming student population (Zwick, 2019).

The Carnegie Classification of institutions of higher education classifies doctoral granting universities based on their level of research activity. For these institutions, Carnegie has three classification: R1, which indicates very high research activity; R2, which indicates high research activity; and D/PU, which includes all other doctoral and professional universities. Given that there are three different classifications within this group, it stands to reason that the admissions criteria would vary across the different classifications as well. It may be the case, for example, that the admissions criteria at an R1 university are generally higher than those of other universities that have a lower research oriented Carnegie Classification.

A sample of different admissions criteria cutoffs, including one of TWU's local competitors and one of TWU's peer group institutions as defined by the Texas Higher Education Coordinating Board (THECB; Texas Woman's University, 2021) are shown in Table 1.1.

Table 1.1

A Comparison of Admission Requirements for Texas Woman's University and Select Peers

Institution	High School Class Rank	SAT 2016	ACT	Carnegie Classification
Texas Woman's University	Top 50%	1080	21	D/PU
Texas A&M Kingsville	Top 50%	910	17	R2
University of North Texas	Top 50%	1130	23	R1

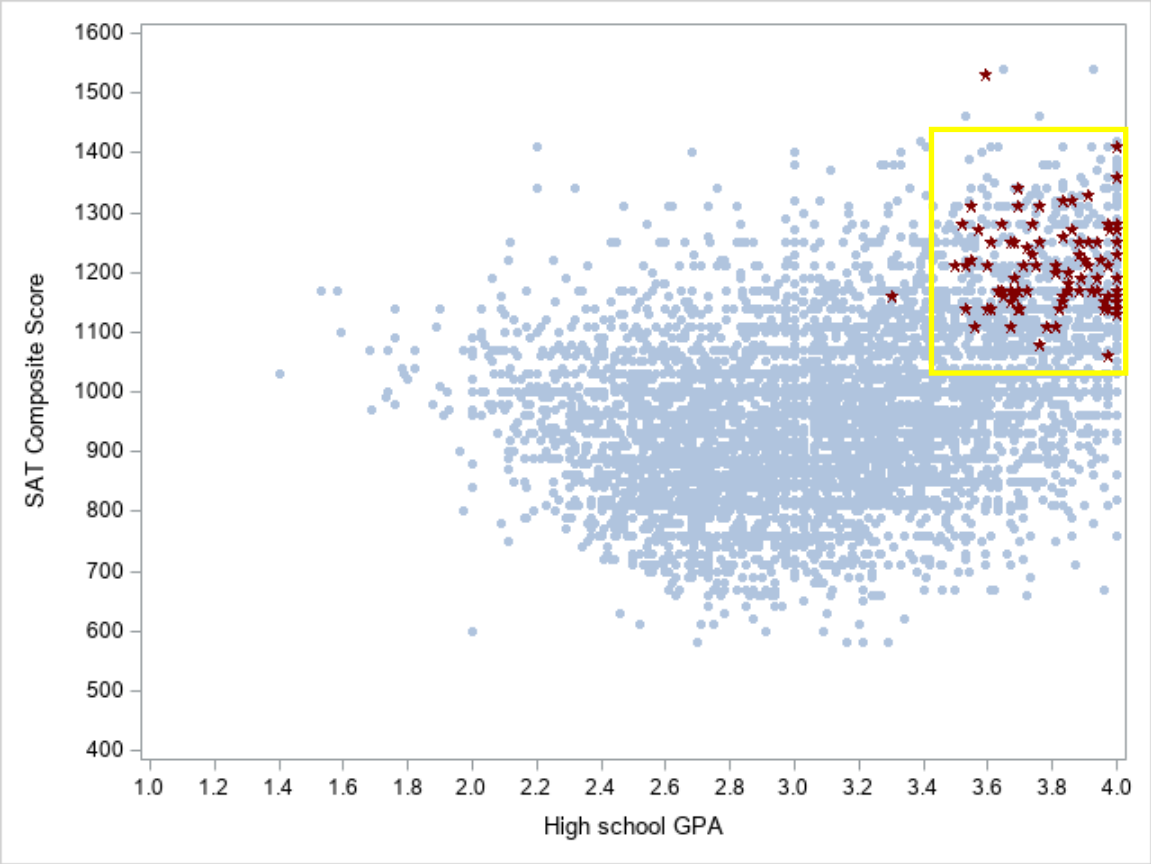
Examining the different criteria, while controlling for high school class rank, Table 1.1 shows that variability exists among the various institutions. For example, the University of North Texas (UNT), a school with an R1 Carnegie Classification, has the highest entrance assessment requirement as would be expected. TWU, with a D/PU Carnegie Classification, has the next highest entrance assessment requirement. Texas A&M – Kingsville, with an R2 Carnegie Classification, has the lowest entrance assessment requirement. Although there are clear increments of research activity that specify higher or lower levels of Carnegie Classification, this type of consistent rank order cutoffs with admission criteria is absent and in fact seems to indicate an arbitrary determination for how schools actually set their entrance requirements. Accordingly, this research establishes a more structured process of determining admissions criteria that connects to student "success" in a way to be described subsequently.

Exploratory Analysis

Exploratory data analysis of TWU First Time in College (FTIC) students that were in a group designated as Early Admit Nursing (EANUR) at the time of admission and went on to successfully transition into a Nursing program, consistently show they had higher high school GPAs and class rank, when compared to TWU's minimum requirements, at the time of admission. Comparing the relevant metrics of this population to the overall TWU admission criteria in Table 1.1, these applicants have an average rank of top 11% versus the required top 50%, an average high school GPA of 3.8 versus the required 2.0, and an average entrance assessment score of 1211 versus the required 1080. In these examples, notice that applicants who apply for TWU's most competitive second admit program have higher cutoff admission criteria than the minimum standards. Figure 1.1 illustrates these higher standards for EANUR applicants by plotting admission factors in a two dimensional graph, with an applicant's GPA as the x-axis and SAT score as the y-axis. The EANUR admitted applicants points are plotted (maroon stars) along with the same scores for all other TWU admitted applicants (blue circles). Applicants who were admitted as EANUR are clustered at the higher end of both axes.

Figure 1.1

A Comparison of All EANUR Admitted Applicants (maroon stars) vs All Other TWU FTIC Admitted Applicants (blue circles).



The yellow outlined area shown in Figure 1.1 above establishes the concept of a two dimensional zone, which presumes that an applicant's Euclidean positioning in such a zone indicates a higher expected performance that would meet the criteria for TWU's very competitive second admit nursing program. Accordingly, this concept extends to the hypothesis that there are subsets of zones for which the expected performance outcomes of the applicants, in their respective zones, would have significant variations of "success" across different zones. This speaks to the concept of determining clusters/zones that have

homogenized performance outcomes that from a set theoretic perspective create a partition of the Euclidean space defined by two commonly used factors or dimensions in admission processing.

To determine zones in the Euclidean space, a performance outcome or dependent variable was associated to zones created by the independent variables that define the Euclidean space. Since a university student's success in terms of graduation is inherently linked to accumulation of university semester credits hours, this study will be utilizing accumulated hours over a relevant time period to define the dependent variable. This dependent variable looks beyond persistence of a student. Instead, this dependent variable includes successful accumulated semester credit hours over a time period that is fundamentally influenced by persistence. This approach creates a more enhanced dependent variable that includes aspects of potential investment and financial sustainability in applicant recruitment. By considering semester credit hours in the development of the dependent variable, this research is now incorporating a significant revenue source for any university. Thus, a university can potentially determine sustainable investments by identifying the expected cumulative credit hours over time of applicant recruitment. To determine this type of dependent variable, historical data on applicants that were admitted and enrolled at TWU from an FTIC student route was used to calculate a cutoff score, C, of accumulated credit hours that discriminates between persisting and non-persisting over a specified time period. Utilizing the prescribed cutoff, C, "success" for each student will be defined by a dichotomous variable, y, which

indicates whether or not a student accumulates at least the prescribed cumulative cutoff hours or not. Specifically, for a specified time period,

$$y = \begin{cases} 1 \text{ if a student's cumulative semester credit hours is greater than or equal to } C \\ 0 \text{ if a student's cumulative semester credit hours is less than } C \end{cases}$$

By utilizing currently established admissions criteria, this study identifies multiple zones such that any group of students in a zone will have an expected rate or proportion of $y = 1$ that will be different from students in other identified zones.

Thus, this study sought to answer the questions: 1) By leveraging commonly used admission criteria, can "performance zones" be identified based on accumulated hours of TWU students after a prescribed time period? 2) Are there significant differences in proportions of success, $y = 1$, within each performance zone?

In order to answer these questions, the objectives of this research are as follows:

1) Utilize a statistical procedure, like Youden's index, to determine an overall cutoff point of the accumulated hours, within a specified timeframe, that distinguishes between $y = 1$ ("success") and $y = 0$ ("not-success").
2) Use the cutoff point as the dependent variable to create student performance zones by means of a classifying algorithm.
3) Using cross validation and chi-square analysis, determine if there are significant differences in the proportions of success within each zone as well as examine the descriptive statistics of the zones.

The data for this study was provided by the TWU Institutional Research & Data Management (IRDM) Department and consisted of a de-identified dataset of TWU FTIC

fall undergraduate applicants over a 5–7 year period. All data analysis were conducted using SAS statistical software.

Significance of Study

In a climate of student enrollment growth, universities have a competitive interest in helping recruit students that will be successful by virtue of accumulating university semester credit hours towards their degree. In addition, it is important for universities to recognize and possibly invest in potentially highly productive students that may have been overlooked by standard admission evaluation criteria. This could be particularly relevant and insightful for a growing demographic population, like is seen in Texas that are increasingly minority and are first-generation applicants. Given these pressures, finding applicants whose investment proposition may be overlooked at the time of admission but yet have potential to be productive students by virtue of their identified performance zone is vitally important. Currently, there is a lot of variation in admissions criteria and no documented process for establishing various performance zones of "success" utilizing basic admissions criteria. This study seeks to establish a structured process that can be utilized to help target and competitively recruit applicants that have desirable expected performance outcomes.

CHAPTER II

LITERATURE REVIEW

Introduction

The idea that universities have requirements that must be met in order to be granted admission is not unusual, or unexpected. In fact, there is a long history of universities having admission criteria that must be met before being admitted. The earliest study found regarding the subject of university admission criteria was the work of Edwin Broome, "A Historical and Critical Discussion of College Admission Requirements," published in 1903. In this work, Broome researched and described more than 150 years of college admissions requirements, starting with those of Harvard University in 1642. The requirements in place at all major American universities during this time period were an examination of a student's character, background, and proficiency in reading Latin and Greek. Toward the end of the 18th century, a working knowledge of arithmetic was added (Broome, 1903). After World War II, university entrance requirements started becoming more uniform (Beale, 1970). Based on a meta-analysis by Beale (1970), it was during this time that universities tended to admit students on the basis of six major factors: high school graduation, a minimum number of specific courses taken in high school, high school class rank, some form of entrance assessment test scores, a personal interview, and the recommendation of the principal. This looks much closer to the requirements that many universities require to be submitted with an application as of this study.

It is not difficult to see that there is a similarity of Harvard's early position of assessing a student's proficiency in reading Latin and Greek and a working knowledge of mathematics to a modern assessment of a student's proficiencies in high school, measured broadly as GPA. However, with the continued and steady increase of applications, universities are having to look beyond the question of "do you meet the minimum standards" and instead ask, "is this a good fit for both parties." One example of this increase can be found by comparing Harvard's first graduating class, which had nine students (The Crimson, 1890), to last year's graduating class, in which they conferred over 8,000 degrees and certificates (Harvard University Fact Book, 2020).

The ultimate goal, of both the student and the university that admits them, is to graduate. Many university administrators look at student persistence as an indicator or intermediate step towards this ultimate goal of student success (Levitz et al., 1999). Towards this end, the purpose of having admissions criteria is for each respective university's admission office to evaluate an applicant's likely preparedness to successfully navigate the course work required to earn a degree at that institution (Marsh et al., 2008). Thus, it can be said that student success truly begins from the point at which decisions are made, that is, at the time of admission.

However, in searching for literature related to this topic, it was difficult finding anything that approaches the establishment of admission criteria from a holistic point of view. Most studies that have addressed the relationship between admissions criteria and academic performance focus on specialized programs such as Nursing, Dental, and Medical programs. For instance, Yousafzai and Jamil (2019) found that among nursing

students, "…rigorous admissions criteria can predict academic performance…" (p. 1), as measured by cumulative GPA. Salvatori (2001) conducted a meta-analysis of health professions literature and found that pre-admissions GPA was the best overall predictor of academic performance among admissions criteria. Additionally, this relationship overwhelmingly extends across all health care professions. No literature was found that documented a methodology to establish similar admissions criteria holistically across all undergraduate applicants, regardless of program, for any university.

Predictive Analytics in Institutional Research

The first published statistical study in American higher education is attributed to Frederick Barnard, the president of Columbia University in 1866 (Rice et al., 2011). While this study employed descriptive statistics that analyzed the enrollments of 10 universities that Columbia identified as competitors at the time, statistical analysis in this field has only increased in usage and complexity. It was around the 1960s that predictive statistics seems to first been formally employed to address specific topics, such as predicting enrollment.

Since then, predictive analytics has only become more commonplace, but usually when looking at student success, the focus is on retention and persistence while modeling behaviors of the students *after* they have been admitted and enrolled. Most work in applying predictive analytics to student recruitment seem to focus more on identifying which students will apply, to prioritize contact and recruitment efforts, than which students will graduate (Eduventures, 2013).

In reviewing previous research and presentations by colleagues, a conference presentation was found that discussed the use of admission criteria, specifically application high school GPA (AP_GPA) and SAT scores, to create an "admission grid" that might lend insight into which students would be better performers once admitted (Langston et al., 2018). That research attempted to employ a *moneyball* approach to determine a "sweet spot" of student admission criteria that would allow the university to "attract, retain, and graduate undervalued students who persist with power." For those unfamiliar with the term, *Moneyball* was a book published in 2003 by Michael Lewis chronicling the Oakland Athletics baseball team's focus on using analytics to evaluate and identify players that would produce results on a competitively smaller budget when compared to the rest of their peers. To this end, this research is an extension of this *moneyball* approach that will be able to determine zones of performance, or production, from which recruitment can be focused while having a basic understanding of return on investment (ROI).

Additionally, through another research project, I became aware of a statistical approach primarily used in the medical field that is used to determine cutoff values: Youden's index. Youden first proposed this "index of performance" in 1950 (Hsieh & Turnbull, 1996), and it has been used in a variety of different medical research when attempting to determine an optimal cutoff point. For instance, it has been used in assessing cognitive screening instruments (Larner, 2015) as well as interpreting various diagnostic tests (Davidson, 2002). Youden's index was employed as the starting point to

determine the dependent variable mentioned in Chapter 1 and will discuss it in more detail in a subsequent chapter.

The goal of this study was to develop a holistic approach that will utilize what is known about applicants at the time of decision making. In the pursuit of this goal, this study used various statistical techniques, including but not limited to logistic regression, and these techniques were implemented by using SAS software.

CHAPTER III

THE LOGISTIC REGRESSION MODEL

Introduction

This chapter introduces the notation for the concept of determining performance zones in a two dimensional Euclidean space determined by two commonly used admission variables. Utilizing logistic regression, this research associated a dependent variable to admissions variables or independent variables that define the Euclidean space. On the basis of the dependent variable, this study was able to create rank ordered performance zones within the two dimensional space defined by the admission criteria. Since university student success in terms of graduation is inherently linked to accumulation of university semester credits hours, this research utilized accumulated hours over a relevant time period to define the dependent variable.

The dependent variable created goes beyond persistence of a student over a specified time period. Instead, this dependent variable includes successful accumulated semester credit hours over a time period that is fundamentally influenced by persistence. To illustrate this concept, let t represent a period of time after a student's initial enrollment for which the tracking of accumulated hours will be determined. For example, suppose the goal was to track a student over a time period, t, of one academic year. Suppose this admitted applicant enrolls at TWU and successfully completes 12 hours in the fall semester, 15 hours in the following spring semester, and 3 hours in the summer. This student during this time period t, one academic year, will have accumulated 30 total

semester credit hours. It is worth noting here that if a student makes an F in a course during this time period or takes a course that does not count toward earning a degree, like a developmental course, these hours will not contribute to their total accumulated semester credit hours over this time period. This aspect of the dependent variable is important because it creates a higher standard to track students towards the ultimate goal, which is graduation.

To graduate within any time period, students will need to accumulate hours that count toward their degree, and in Texas, the amount of accumulated semester credit hours needs to be at least 120 in total to obtain an undergraduate baccalaureate degree. By considering semester credit hours in the development of the dependent variable, this research is now incorporating a significant revenue source for any university. Accordingly, this approach creates a more enhanced dependent variable that includes aspects of potential investment and financial sustainability in applicant recruitment. Not only does this provide the university with a sustainable revenue source, but by targeting applicants who are likely to succeed, the goal is to hopefully reduce the potential for students to be burdened with debt without them graduating. This represents a potentially more holistic approach to the admission of applicants that is good for the university in terms of return on investment but also inherently beneficial to the student.

Using notation to facilitate this discussion, $\mathbf{P} = \{1,2,\ldots,N\}$ represents the finite population of size N of all TWU FTIC applicants that provide admission criteria values (i.e., the independent variable values). Using set theory notation, the magnitude of $\mathbf{P}$ is represented by $|\mathbf{P}| = N$, where $|.|$ is the magnitude (i.e., the number of elements) of a given

set. *C* represents a threshold or cutoff score of accumulated credit hours in time *t*. The specifics of how *C* is determined is discussed in more detail in the subsequent chapter, but for now assume *C* is some fixed integer value. Using *C*, the dependent variable is defined, for each applicant $i \in \mathbf{P}$, as:

$$y_i = \begin{cases} 1 \text{ if a student's cumulative semester credit hours is greater than or equal to C} \\ 0 \text{ if a student's cumulative semester credit hours is less than C} \end{cases} \quad (3.1)$$

Admissions criteria has minimal standards for which any student meeting or exceeding that standard is admitted. For the purposes of this research, the specific interest is in knowing the proportion of applicants that will have $y_i = 1$, which indicates success in this study. Specifically, the proportion of success for population **P** was defined as

$$\pi = \frac{\sum_{i=1}^{N} y_i}{N} \quad (3.2)$$

To introduce the concept of performance zones, the elements of **P** were partitioned into *k* zones, such that:

$$P = P_1 \cup P_2 \cup ... \cup P_k \quad (3.3)$$

where $\mathbf{P}_j$ represents a subset of indices from **P**, such that $\mathbf{P}_j \cap \mathbf{P}_l = \varnothing$, where $j \neq l$ and j and l are integer values between 1 and k. Let $|\mathbf{P}_j| = n_j$ represent the magnitude of $\mathbf{P}_j$ for

j=1,2,…,k. Given Equation 3.3 and since $\mathbf{P}_j \cap \mathbf{P}_l = \varnothing$, $j \neq l$, this implies N = $|\mathbf{P}_1|+|\mathbf{P}_2|+\ldots+|\mathbf{P}_k| = n_1 + n_2 + \ldots + n_k$.

Utilizing the partition P represented by Equation 3.3, and utilizing the concept of Equation 3.2, a relative proportion of success for each successive zone can be established as

$$\pi_j = \frac{\sum_{i \in \mathbf{P}_j} y_i}{n_j} = \frac{1}{n_j}\left(\sum_{i \in \mathbf{P}_j} y_i \right). \tag{3.4}$$

Equation 3.4 represents the relative proportion of applicants that would exceed the cutoff threshold determined by C.

To determine the zones that define Equation 3.3, a probability p_i, $0 \leq p_i \leq 1$, that represents the probability that $y_i = 1$, will be associated to each value y_i. The value $(1 - p_i)$ represents the probability that $y_i = 0$. Each response, y_i, follows a Bernoulli distribution with the probability mass function:

$$f(y_i \mid p_i) = (p_i)^{y_i}(1 - p_i)^{1-y_i}$$

By definition, the mean and variance of the Bernoulli distribution are

$$\mu_{y_i} = E(y_i) = p_i \tag{3.5}$$

and

$$\sigma_{y_i}^2 = p_i(1-p_i) \tag{3.6}$$

respectively. Equation 3.5 represents the expected value of y_i and Equation 3.6 is the associated variance for the random variable y_i. Utilizing Equation 3.5, the expected value or estimate of Equation 3.4 becomes

$$\hat{\pi}_j = E(\pi_j) = \frac{1}{n_j}\left(\sum_{i\in \mathbf{P}_j} E(y_i)\right) = \frac{1}{n_j}\left(\sum_{i\in \mathbf{P}_j} p_i\right). \tag{3.7}$$

Equation 3.7 was how the performance outcome of each zone represented by the partitions in Equation 3.3 was measured.

Logistic Regression

This study used logistic regression to associate the dependent variable defined in Equation 3.1 with two independent variables that are commonly used to make decisions at the time of admission. In particular, X_{AP_GPAj} represents the application GPA and X_{AP_Rankj} represents the application class rank corresponding to each applicant $i \in \mathbf{P}$. This study focused on these two independent variables because entrance assessment scores, such as SAT or ACT, are not well populated fields, meaning that obtaining this type of data from applicants is not always reliable. In the dataset of all completed applications,

that is, an application about which a decision can be made, 87.51% had both a high school GPA and a high school class rank. Compare this to the 45.85% of applications that had both a high school GPA and an entrance assessment score. For definitional purposes, it is important to note that AP_GPA is a TWU unweighted calculation on a 4.0 scale and includes an applicant's grades, from their official high school transcripts, in English, mathematics, science, social studies, and academic electives. Also consider that major universities, such as the University of Texas, are suspending their requirements for entrance assessment scores through at least the fall 2022 semester due to the ongoing impact of the COVID-19 pandemic (McGee, 2021), while some entire university systems, such as the University of California system, are phasing out the use of these tests entirely (Hubler, 2020).

This study used logistic regression to connect the independent variables (X_{AP_GPA}, X_{AP_Rank}) indirectly to y_i, Equation 3.1. Using the notation derived above, for each applicant $i \in \mathbf{P}$, associate the independent variables to p_i, which probability $y_i = 1$, with the following equation

$$logit(p_i) = \ln(odds_i) = \beta_0 + \beta_1 x_{AP\_GPA_i} + \beta_2 x_{AP\_Rank_i} \tag{3.8}$$

where β_0 is an intercept term, β_1 is the coefficient for the independent variable application GPA (X_{AP_GPA}), β_2 is the coefficient for the independent variable application rank (X_{AP_Rank}) and

$$odds_i = \frac{probability\ yi=1}{probability\ yi=0} = \frac{p_i}{1-p_i}.$$

The odds equation above can be rewritten with some algebra as

$$p_i = \frac{odds_i}{1+odds_i}. \tag{3.9}$$

Using Equation 3.9, it is worth noting that mathematically,

$$p_i = \frac{e^{\beta_0+\beta_1 x_{AP\_GPA_i}+\beta_2 x_{AP\_Rank_i}}}{1+e^{\beta_0+\beta_1 x_{AP\_GPA_i}+\beta_2 x_{AP\_Rank_i}}} = \frac{e^{\ln(odds_i)}}{1+e^{\ln(odds_i)}} = \frac{odds_i}{1+odds_i}. \tag{3.10}$$

Thus, Equation 3.10 shows how to formally utilize the admission criterion through logistic regression to estimate Equation 3.8, which allows us to measure the performance outcome of each zone.

Summary

We have now developed the mathematical framework for the analysis used in this study. The next chapter will introduce the logic associated with how to determine the cutoff value, *C*, for the dependent variable, Equation 3.1. Once the cutoff is determined, we will show the results of the logistic regression that estimate the parameters, coefficients and intercept, of Equation 3.8.

CHAPTER IV

THE DEPENDENT VARIABLE AND BUILDING THE MODEL

Introduction

This chapter introduces the logic associated with the creation of the cutoff value, C, for the dependent variable. In addition to meeting the minimum cutoff score, this research was also are interested in students who persist at TWU after 2 years, which we define in this study as P2. From research by Hamner et al. (2019), P2 has shown a strong correlation to graduation, which is, of course, the ultimate goal. Accordingly, we define time period, t, for this study, as 2 years. After development of this innovative dependent variable, we built a predictive model from historical data called the training dataset that identifies an applicant's likeliness to "succeed" or meet the desired cutoff score, C, by t = 2 academic years. In particular, this training dataset allows us to determine the parameter estimates of Equation 3.8 that we can associate to applicants at the time admission decisions are being made using commonly available admission variables: application high school GPA and application high school class rank. Accurately projecting these students will provide a formal structure for determining admissions criteria that will identify applicants as good candidates for not only admission but also potential investment. Thus, utilizing the outcome of Equation 3.8 will allow us to determine rank ordered performance zones as defined by Equation 3.7 that can be utilized to direct an institution's targeting initiatives, to increase enrollment yield on the basis of return on investment. Finally, to illustrate the predictive accuracy of the model developed from the

training dataset, a cross-validation on a testing dataset, which is a set of applicant data not used to determine the coefficients for Equation 3.8, was used.

The Dependent Variable

By introducing the concept of a cutoff score and tying it to a dichotomous outcome, such as P2 in this case, we automatically create a decision-making framework for determining an appropriate cutoff score. A common type of decision making framework in statistics, can be illustrated through hypothesis testing of a given null, H_0, and alternative hypothesis, H_a. In hypothesis testing, for example, a decision is made to either *reject the H_0* or *fail to reject the H_0*. For conceptual simplicity, the decision to fail to reject the H_0 is represented as "Accepting H_0" and the decision to reject H_0 as "Accepting H_a." The outcomes of a hypothesis decision-making process can generically be viewed through a classification table like Table 4.1 below.

Table 4.1

Hypothesis Testing Classification

		Assume	
		H_0 True	H_a True
Decide	Fail to Reject H_0 (Accepting H_0)	Correct Decision	Error
	Reject H_0 (Accepting H_a)	Error	Correct Decision

Similarly, this study utilized components of the classification table to illustrate how to determine the cutoff score, *C*. The components of the classification table we utilize will introduce concepts such as true positive, false positive, false negative, and true negative. More specifically, these terms within the study are defined as follows:

- True Positive – the number of students that met the minimum cutoff score while persisting for 2 years.

- False Positive – the number of students that met the minimum cutoff score but did not persist for 2 years.
- False Negative – the number of students that did not meet the minimum cutoff score but did persist after 2 years.
- True Negative – the number of students that did not meet the minimum cutoff score and did not persist for 2 years.

Additionally, these terms are displayed in the classification table, Table 4.2, below.

Table 4.2

Frequency Classification

		P2	
		Yes	No
Meet minimum Cutoff Score, *C*	Yes	True Positive	False Positive
	No	False Negative	True Negative

Table 4.2 represents the classification table where the rows are determined by the dichotomous outcome of a student either meeting the minimum cutoff score, 'Yes', or failure to meet the minimum cutoff score, 'No'. The associated columns of the Table 4.2 are determined by the dichotomous outcome of the student either persisting after 2 years, P2 ='Yes' or failing to persist after 2 years, P2 = 'No.' In general, the cutoff score, *C*, will be evaluated by the effect it has on the distribution of true positives and true negatives, which will also have an effect on the false negative and false positive components of classification Table 4.2.

Youden's Index

Through previous research, Hamner et al. (2019) considered Youden's index or Youden's *J* statistic for evaluating grade cutoff scores to determine gateway courses. Similarly, we now explore Youden's index to understand how its structure will affect selection of a cutoff score, *C*, for this study. Youden's index is calculated by

$$J = Sensitivity + Specificity - 1$$

where, using the definitions of the components in Table 4.2, *Sensitivity* and *Specificity* are defined by

$$Sensitivity = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

and

$$Specificity = \frac{\text{True Negatives}}{\text{True Negatives + False Positives}}.$$

To create a probability classification table we define the following definitions from the components of Table 4.2.

$$False\ Negative\ Error = \frac{\text{False Negatives}}{\text{False Negatives + True Positives}}$$

and

$$False\ Positive\ Error = \frac{\text{False Positives}}{\text{False Positives + True Negatives}}.$$

It is worth noting that mathematically Sensitivity + False Negative Error = 1 and Specificity + False Positive Error = 1. Utilizing the above probability definitions, we create Table 4.3 below.

Table 4.3

Probability Classification

		P2	
		Yes	No
Meet minimum Cutoff Score, *C*	Yes	Sensitivity	False Positive Error
	No	False Negative Error	Specificity
		1	1

Due to the mathematical constraints of the column's probabilities summing to 1 in Table 4.2, we note that Youden's index, *J*, can have a maximum value of 1 when the false negative error and false positive error are zero. In general, Youden's index, has a maximum value when both sensitivity and specificity, combined, produce the largest value of *J*. This suggest that Youden's index does not weight, in terms of importance, sensitivity over specificity, but rather simply considers whatever combination of both yields the largest value. That is to say, sensitivity and specificity are considered equally important in Youden's index. Thus, Youden's index will select a cutoff, *C*, that provides the maximum value of *J* regardless of a particular sensitivity or specificity value. The value *C* that maximizes *J* is typically referred to as the optimal cutoff score because it optimizes the differentiating ability of P2 in this research, for example, when equal weight is given to sensitivity and specificity (Ruopp et al., 2008).

Utilizing Youden's index, a baseline approach to determine a cutoff score, *C*, was operationalized using the training dataset. For this study, the training dataset contains all FTIC applicants that applied to TWU for any fall semester between 2014 and 2016. To explore a possible cutoff score from this training dataset, this study focused on applicants that matriculated to TWU. From this training data, this research observed all accumulated

semester credit hours for these students over time period $t = 2$ academic years, which is a continuous variable we denote as SCH_2, as well as a dichotomous outcome variable, P2, formally defined as

$$P2 = \begin{cases} 1 \text{ if a student persists after } t = 2 \\ 0 \text{ if a student does not persist after } t = 2 \end{cases}.$$

To automate this process this research notes the connection of Youden's index, a function of Sensitivity and Specificity, to the receiver operating characteristic (ROC) curves commonly utilized to evaluate model fit in logistic regression (Lawrence, 2020; Schisterman et al., 2005). The ROC curve is a two dimensional plot of Sensitivity (y-axis) versus 1 – Specificity (x-axis) at all possible cutoff points, C, for SCH_2. For a more extensive discussion on ROC curves, we refer the reader to Ahluwalia (2006). Accordingly, understanding the association of Youden's index with ROC curves from logistic regression allowed us to write SAS code that automates finding the cutoff score, C, that maximizes J by assigning the logistic regression model's dependent variable to P2 and the model's independent variable to SCH_2. See Appendix A for the SAS code that automates this process.

As a result of the SAS code and the training dataset, the initial baseline cutoff score considered that maximizes Youden's index is $C = 45$ SCH_2. Utilizing this cutoff score, we can fill in the components of Table 4.3 to generate Table 4.4 and Table 4.5, below.

Table 4.4

$SCH_2 \geq C = 45$, *Frequency Classification*

$SCH_2 \geq C = 45$		P2		
		Yes	No	**Total**
Meet minimum Cutoff Score, *C*	Yes	1,820	310	**2,130**
	No	172	882	**1,054**
Total		**1,992**	**1,192**	**3,184**

Table 4.5

$SCH_2 \geq C = 45$, *Probability Classification*

$SCH_2 \geq C = 45$		P2	
		Yes	No
Meet minimum Cutoff Score, *C*	Yes	.914	.260
	No	.086	.740
Total		**1**	**1**

From Table 4.5 and Table 4.3 we can evaluate the cutoff that was determined by Youden's index. With equal weight given to both Sensitivity and Specificity, the cutoff value, C = 45 SCH_2, that maximized *J* generated Sensitivity = .914 (91.4%) and Specificity = .740 (74.0%) for an index value of *J* = .654 (.914 + .740 – 1). As a consequence of this combination of Sensitivity and Specificity that maximized *J*, the resulting False Negative Error = .086 (8.6%) and False Positive Error = .260 (26.0%). Interestingly, this low False Negative Error in relation to the higher False Positive Error is likely the preferred relationship in the clinical or medical research community from which Youden's index is prominently discussed (Larner, 2015).

To conceptualize, consider utilizing Youden's index to calibrate or find a cutoff value for a medical test that will discriminate between, for example, "cancer" or "no cancer." Ideally, you want Sensitivity and Specificity to be 1 so that the errors are zero

and patients that test positive get appropriate medical attention. In the absence of a perfect test, consider the consequences of the respective errors. Suppose a patient tests positive for cancer and then the medical doctor immediately schedules for a follow-up procedure that indicates the patient does not, in fact, have cancer. This is an example of a false positive and the consequences of this error is inconvenience to the patient by requiring a follow-up medical procedure. However, suppose a patient tests negative for cancer but in fact has cancer, the potential consequences of this false negative is that the patient does not seek appropriate medical attention to deal with the cancer. With the potential health consequences associated with such high stakes medical testing, the calibration of such a test will focus on minimizing the false negative error and consequently boosting Sensitivity over Specificity. Keep in mind, however, that we are exploring an investment proposition in this study on all students that meet the minimum cutoff score, *C*, meaning we are considering investing in all students on Row = Yes. Subsequently, we have to evaluate the selection of the cutoff score for this research on the basis of the distribution of the classification table's components as well. From Table 4.3, if we were to accept Youden's index cutoff value, C = 45 SCH_2, then we are identifying 67% (2,130/3,184) of all matriculating students as a target for an investment commitment, which may not be realistic given finite monetary resources. This resource heavy commitment includes the 310 students that failed to persist after 2 years and that are no longer generating revenue for the university. Given the relative position of these students that are no longer persisting and the very large potential investment commitment, we are concerned with the false positive error of .260 (26.0%), because this

means that more than one quarter of all the students who do *not* persist after 2 years meet the Youden's cutoff score. With limited financial resources for targeted student recruitment and the concern of return on investment (ROI), it is reasonable to conclude that a higher standard, in the form of a larger cutoff value *C*, should be considered that favors increased specificity and a reduced false positive error.

Considering Youden's cutoff selection, $C = 45$ SCH_2, within the context of a merit-based investment towards the ultimate goal of graduation, it becomes evident that 45 accumulated semester credit hours within 2 years is not a high standard. In fact, this would fail to meet the minimum standard required by the federal government for Pell recipients, which requires full-time enrollment for undergraduates seeking a college degree. In two academic years, a typical full-time student would accumulate a minimum of 48 SCH_2 over the time period of interest. Considering that we are looking for a higher standard worthy of investment, we are going to require a higher standard beyond the 48 SCH_2 minimum of a full-time student, and use $C = 52$ SCH_2 as the cutoff value (see Tables 4.6 and 4.7). At TWU, all FTIC students are required to take a 1 hour First-Year Experience Course, so essentially the one step beyond full-time amounts to one additional 3 hour course at TWU within 2 years.

Table 4.6

$SCH_2 \geq C = 52$ *Frequency Classification*

$SCH_2 \geq C = 52$		P2		
		Yes	No	**Total**
Meet minimum Cutoff Score, *C*	Yes	1,375	182	**1,557**
	No	617	1,010	**1,627**
Total		**1,992**	**1,192**	**3,184**

Table 4.7

$SCH_2 \geq C = 52$ *Percentage Classification*

$SCH_2 \geq C = 52$		P2	
		Yes	No
Meet minimum Cutoff Score, *C*	Yes	.690	.153
	No	.310	.847
Total		**1**	**1**

We can see that the higher cutoff score $C = 52$ SCH_2 increased Specificity = .847 (84.7%) and reduced the False Positive Error = .153 (15.3%). Now, only 15.3% of all students in the training dataset do *not* persist after two years, which is over a 10% reduction to this error compared to the results utilizing Youden's cutoff score. Although Youden's index value of $J = .538$ (.690+.847– 1) has decreased with the higher standard cutoff score, the investment proposition is more selective and includes a nearly 20% reduced investment commitment of 48.9% (1,557/3,184) of all matriculating students. This more selective criteria also increases potential ROI. To illustrate this potential increase in ROI we calculate a simple ratio of true positive/false positive across the investment Row = Yes of Tables 4.4 and 4.6, which are respectively 5.87 and 7.55. We can think of this ratio as a simple win/loss ratio. For the higher standard cutoff, this ratio means that for every non-persisting student you mistakenly invest in, you have 7.55 students that persist and continue to generate revenue through accumulated SCH after time period $t = 2$. In this context, the higher standard cutoff, when compared to the Youden's cutoff, will compensate for each non-persisting student by having nearly two more persisting students (1.68) generate continued revenue. More formally, the higher standard cutoff increases the precision or positive predicted value (PPV), defined by

$$PPV = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}},$$

from Youden's PPV = .854 to PPV = .883. The PPV is also prominently discussed in medical research (Manrai et al., 2014). Using this more selective cutoff $C = 52$ SCH_2, we have now defined the dependent variable, Equation 3.1.

Variable Selection, Building the Model, and Assessing Fit

Utilizing Equation 3.8, we can now associate the dependent variable to the admission variables we are using for this study. These independent variables are expected to be widely available, well occupied, and are critical in the evaluation of whether or not an applicant to a university is admitted. The independent variables selected for this study are listed and described in Table 4.8. From the training dataset we obtain the logistic regression coefficients, Table 4.9, and fill in Equation 3.8 to obtain Equation 4.1

$$logit(p_i) = \ln(odds_i) = -3.4161 + 1.1499x_{AP\_GPA_i} - 0.00594x_{AP\_Rank_i} \quad . \qquad (4.1)$$

Notice that both independent variables are significant at the 0.05 level and application GPA, in particular, is significant at the 0.01 level. Still, and not unexpectedly, these variables are negatively and almost strongly correlated (−0.67255) with each other, which can cause multicollinearity issues and problems with model coefficient interpretation. However, we are more concerned with predictive accuracy than we are with model coefficient interpretation. In addition, the evaluation of the model's predictive accuracy will be through cross-validation. Although both variables are significant, it is not an uncommon practice to include variables that may not be significant, but for which

the literature suggests are variables that are generally significant. In a like way, because these variables are so commonly used across the admissions processing landscape and are well occupied in the data, their inclusion is relevant to almost all admission decision making processes at universities across the nation. Their face-validity, based on the findings in the literature, lends operational credibility utilizing them in the model (Lang, 2007). To this point, despite potential multicollinearity issues, it is worth noting through the odds ratio, 3.158, calculated from Table 4.9 as $\left(e^{1.1499}\right)$, how important the application GPA is to accumulating credit hours. For each .5 increase in application GPA, the odds of meeting the cutoff $C = 52$ SCH_2 increases by a factor of $1.9\left(e^{.5*1.1499}\right)$.

Table 4.8

Variables Selected for Model Building

Variable	Type	Coded Name	Description
High school GPA	Quantitative	AP_GPA	$0 \leq GPA \leq 4.00$
High school Class Rank	Quantitative	AP_Rank%	$0\% < Rank \leq 100\%$

The coefficients for these variables are shown in Table 4.8 below.

Table 4.9

Results from Fitting Logistic Regression Model to Quantitative Variables

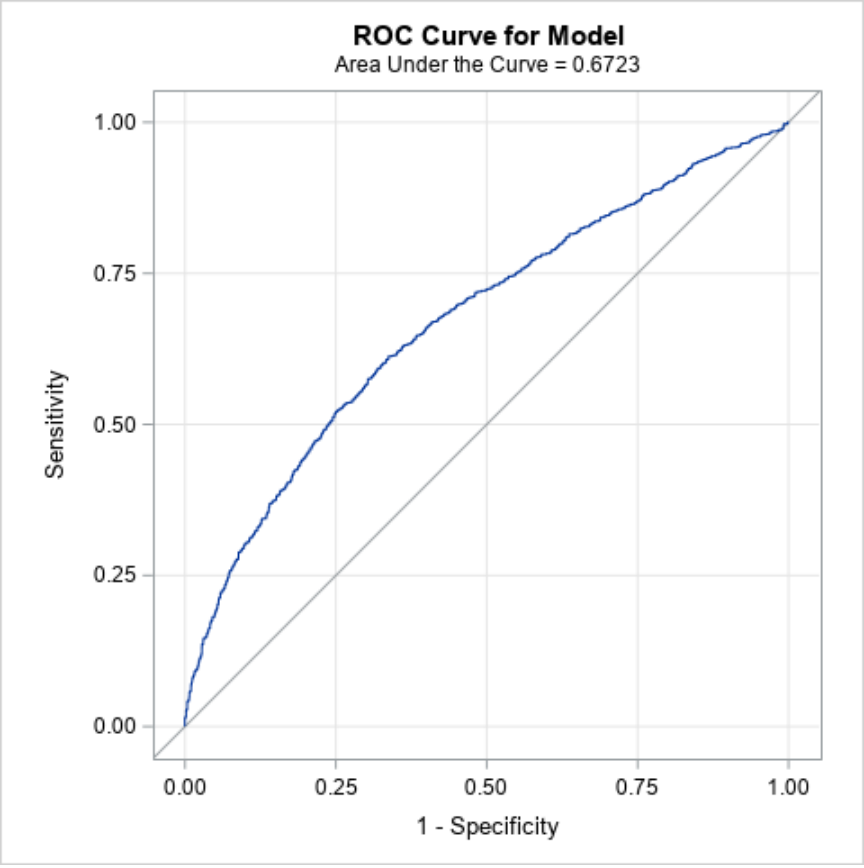
Variable	Coefficient Estimate	Standard Error	P-Value
Intercept	-3.4161	0.3863	<.0001
AP_GPA	1.1499	0.1072	<.0001
AP_Rank%	-0.00594	0.00244	0.0149

Although the ultimate judgment of the model is through its ability to predict future outcomes through cross-validation, we will now briefly discuss model fit assessment through interpretation of the resulting ROC curve, Figure 4.1.

A perfect fit of the model will have an ROC value of 1, which would mean the model perfectly discriminates between those that meet the cutoff value and those that do not. Whereas an ROC value of 0.5 indicates no discrimination. From Figure 4.1, the ROC value of 0.6723 indicates acceptable discrimination. Given that we are only utilizing two independent variables, the adequate model fit is fairly impressive.

Figure 4.1

ROC Curve



As discussed in Chapter 3, the objective is to create performance zones, Equation 3.3, that have relative proportions of success, Equation 3.4, indicating the expected outcomes of applicants meeting the derived cutoff score. We do this because we want to identify a rank ordered performance outcomes from which we can identify zones of potential investment. To create these zones, we will use Equation 4.1 and Equation 3.10 to initially partition the model outcomes (p) of the training dataset on the basis of decile intervals, as shown in Table 4.10. From Table 4.10, the "Average Probability" column of this table are the results of Equation 4.1 and Equation 3.7. The "SCH Cutoff Rate" column of Table 4.11 are the realized outcomes, Equation 3.4.

Table 4.10

Creating Partitions for Training Data on the Basis of Model Outcomes

Train Apply FTIC Grid						
		Apply Grid Model			Apply Grid Actual Results	
Model Group	Probability Range	Enroll Total	Probability Sum	Average Probability	SCH Cutoff Sum	SCH Cutoff Rate
03	$.7 \leq p < .8$	326	240	73.6%	258	79.1%
04	$.6 \leq p < .7$	509	330	64.8%	331	65.0%
05	$.5 \leq p < .6$	625	342	54.7%	342	54.7%
06	$.4 \leq p < .5$	737	332	45.0%	288	39.0%
07	$.3 \leq p < .4$	647	227	35.0%	237	36.6%
08	$.2 \leq p < .3$	307	80	26.1%	84	27.3%
09	$.1 \leq p < .2$	32	6	17.3%	17	53.1%
10	$0 \leq p < .1$	1	0	9.6%	0	0.0%
Total		3,184	1,557	48.9%	1,557	48.9%

Examining Table 4.10, due to the lower average probabilities in Groups 08–10 and the smaller enrollment totals, we decided to combine these groups with Group 07. Since we encountered no observations in Group 01 or Group 02 with combined

probability model scores between $.8 \leq p \leq 1$, we have decided to combine these two groups with Group 03. This new set of groups now totals five zones instead of ten, and are rewritten as follows, see Table 4.11: Group 01 = $.7 \leq p \leq 1$, Group 02 = $.6 \leq p < .7$, Group 03 = $.5 \leq p < .6$, Group 04 = $.4 \leq p < .5$, and Group 05 = $0 \leq p < .4$, see Table 4.10.

Table 4.11

Regrouped Partitions for Training Data on the Basis of Model Outcomes

Train Apply FTIC Grid						
		Apply Grid Model			Apply Grid Actual Results	
Model Group	Probability Range	Enroll Total	Probability Sum	Average Probability	SCH Cutoff Sum	SCH Cutoff Rate
01	$.7 \leq p \leq 1$	326	240	73.6%	258	79.1%
02	$.6 \leq p < .7$	509	330	64.8%	331	65.0%
03	$.5 \leq p < .6$	625	342	54.7%	342	54.7%
04	$.4 \leq p < .5$	737	332	45.0%	288	39.0%
05	$0 \leq p < .4$	987	313	31.6%	338	34.2%
Total		3,184	1,557	48.9%	1,557	48.9%

If you want to shape your class to have higher than average performance, the most logical approach would be to increase the yield from every strata that exceeds the overall results average (48.9%). These groups are worthy of investment at likely different rates that align to the expected performance outcomes. In this way, the training dataset identifies three groups, Groups 01–03, that can be used to shape the subsequent class in a positive way.

All universities are likely to compete for applicants in Group 01. These applicants are going to have both a high GPA and a high class rank percentage. However, in the training dataset, Group 02 has application GPA's as low as 3.36 and Group 03 has application GPA's as low as 3.0. Similarly, Group 02 has High School ranks that exceed

50% and Group 03 ranks that go down even further. Certainly, Group 02 and Group 03 contain potentially over looked and undervalued students that are worthy of financially incentivized targeting to increase yield and, consequently, increase the overall performance of the class. To illustrate this, we will now utilize Equation 4.1 and Equation 3.10 on the testing dataset, see Table 4.12.

Table 4.12

Partitions for Testing Data on the Basis of Model Outcome Zones

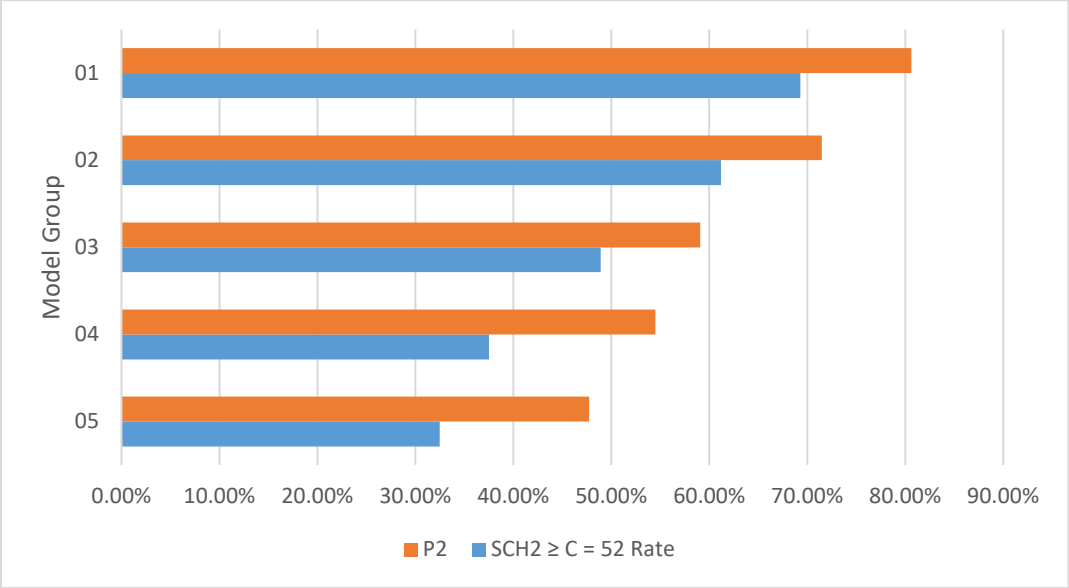
Test Apply FTIC Grid Projection							
		Apply Grid Model			Apply Grid Actual Results		
Model Group	Probability Range	All Applied	Probability Sum	Average Probability	Enroll Total	SCH Cutoff Sum	SCH Cutoff Rate
01	$.7 \leq p \leq 1$	545	398	73.0%	160	111	69.3%
02	$.6 \leq p < .7$	925	600	64.8%	214	131	61.2%
03	$.5 \leq p < .6$	974	536	55.0%	237	116	48.9%
04	$.4 \leq p < .5$	867	392	45.1%	200	75	37.5%
05	$0 \leq p < .4$	1,415	404	28.5%	243	79	32.5%
Total		4,726	2,330	49.2%	1,054	512	48.5%

Table 4.12 shows the results of the cross-validation on the testing dataset. Recall that the testing dataset, is a set of applicant data not used to determine the coefficients for Equation 4.1. The testing dataset contains all FTIC applicants that applied to TWU for the 2018 fall semester that provide admission criteria values and define $|\mathbf{P}| = 4{,}726$. Utilizing this performance zone modeling approach, during the application process, we would expect that students coming from Group 01 would be the top performers, followed by Group 02, Group 03, Group 04 and Group 05. In other words, we expect rank order performances on the basis of their model group and the results of the cross-validation show that exact relationship, see right hand side of Table 4.12. For example, 2 years after entry 69.3% of all the enrolled students from Group 01 met the expected cutoff whereas

only 32.5% of all the enrolled students from Group 05 met the expected cutoff. Comparing these results, we can generalize that, on average, a student from Group 02 will be almost twice as likely to meet the desired cutoff as a student from Group 05. In other words, you would likely need two students from Group 05 to meet the SCH production of one student from Group 02. Interestingly, the rank order performance of these groups does not limit itself to the SCH_2 cutoff score. In fact, Figure 4.2 shows Model Group rank order for both outcomes: 2-year persistence (P2) and SCH_2 Cutoff Rate.

Figure 4.2

Testing Dataset, Enrolled Rank order of P2 and SCH_2 Cutoff Rate by Model Group

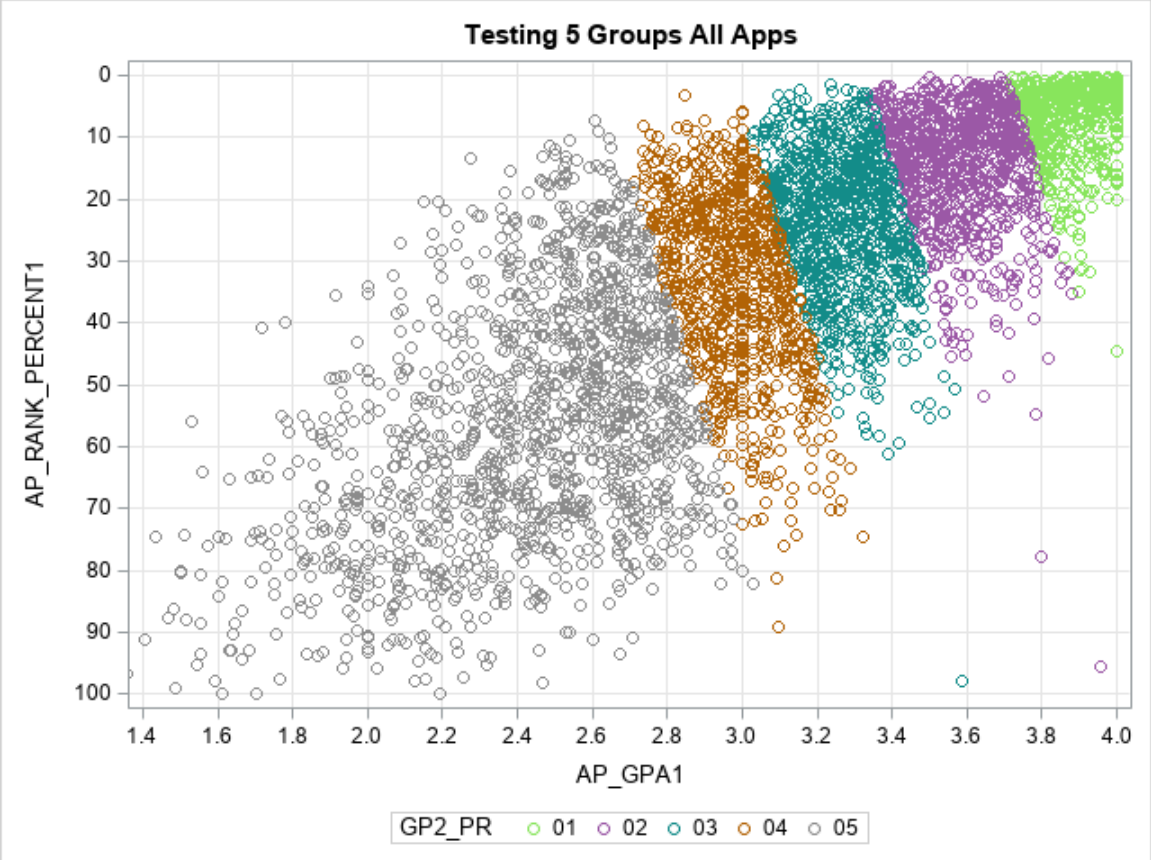


Now that we have illustrated the rank order reliability of the performance based zoning on the basis of model groups, we conclude with a brief discussion on shaping a class. Recall that the modeling technique we developed can be applied to all applicants

prior to their actual enrollment. From Table 4.12, we would be able to target and recruit any of the 4,726 applicants on the basis of their potential performance, see also performance grid visual Figure 4.3.

Figure 4.3

Testing Dataset, 5 Groups, All Applied Performance Grid



By focusing, for example, on the performance grid visual of Group 03, it is interesting to note the general compensatory pattern of application GPA and Rank. In particular, notice how that group creates a pattern that slopes down from left to right, which indicates that a student can compensate for a higher rank by an increased

application GPA (This pattern is present in the training data as well, see Appendix B). As a simple illustration on targeting yield by model group, suppose an enrollment manager increased the yield of Groups 01–03 by 10% respectively and decreased the yield of Groups 04 and 05 by 10% respectively, then the overall class size of the testing enrollment group would have increased by just 16 students to 1070 total, but the overall SCH cutoff rate would have increased by over 5% from 48.5% to 53.8% overall.

CHAPTER V

CONCLUSION

Introduction

In this chapter, this study used chi-square analysis to formally test the $SCH_2 \geq C = 52$ rates across the testing dataset groups modeled in Chapter 4. Further, this research utilized ANOVA and post hoc test to determine significant pairwise differences of the average accumulated SCH_2 between these groups. Additionally, these model groups will be briefly summarized and contextualize through the exploration of an investment proposition. To provide additional context, we will expand on the Chapter 4 discussion regarding additional outcome metrics associated with the testing model groups, by including some demographic breakdowns. Finally, we will discuss some alternative approaches to explore for future research.

Chi-Square and ANOVA Analysis

From the cross-validation example in Chapter 4, it was illustrated that the expected rank order outcome by model Groups 01–05 for the $SCH_2 \geq C = 52$ rates, in fact, held true as well as the persistence rank order, and even the rank order of both the first and second year TWU GPA. Utilizing chi-square analysis we tested the null hypothesis that there is no association between meeting $C = 52$ SCH_2 cutoff and the five model groups categories. The alternative hypothesis is that meeting the cutoff is dependent on the model group categories. To see the contingency table associated with this chi-square analysis see Appendix C. From Table 5.1, at the .01 significance level, we

reject the null hypothesis in favor of the alternative indicating that there is a significant (χ^2 (4) = 294.07, p < .0001) association (dependence) between meeting the cutoff C = 52 SCH_2 and the five model groups categories. This chi-square analysis implies a significant difference between the ranked ordered cutoff rates by groups, see Figure 5.1, but without identifying exactly which pairs of groups significantly differ.

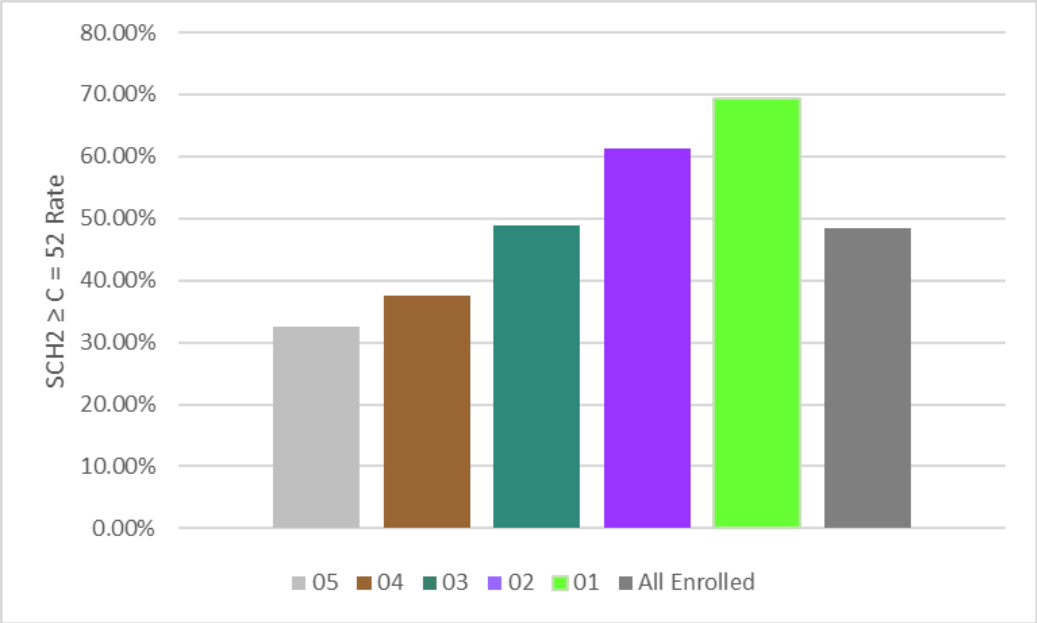
Table 5.1

Results from Chi-Square Analysis on the Training Data

Statistic	Degrees of Freedom	Value	Probability
Chi-Square	4	294.0651	<.0001

Figure 5.1

Testing Dataset, Enrolled SCH_2 Cutoff Rate by Model Group



To determine where the pairwise differences are occurring, we use post-hoc ANOVA analysis, see Appendix D. First, at the .01 significance level, the ANOVA rejects the null hypothesis that the mean SCH_2 across the five model groups are equal, which indicates that at least two mean SCH_2 model groups differ significantly ($F = 64.46$, $p < .0001$). From the Tukey's Honest Significant Difference test, a post hoc analysis, at the 0.05 significance level, we found that Group 01 mean SCH_2 is pairwise significantly different from all other group mean SCH_2, Group 02 mean SCH_2 is pairwise significantly different form all other group mean SCH_2 except Group 03 mean SCH_2, Group 03 mean SCH_2 is pairwise significantly different from all other group mean SCH_2 except Group 02 mean SCH_2, and Group 04 mean SCH_2 is pairwise significantly different from all other group mean SCH_2. This post hoc analysis highlights how Groups 01, 02, and 03, which have higher than average $SCH_2 \geq C = 52$ rates, are significantly different from Groups 04 and 05 and therefore are good candidates for targeted recruitment and investment efforts.

The Investment Proposition

For simplicity, without the nuance of formula funding and overhead cost, we present a simple, albeit contrived, way to conceptualize the potential revenue impact of the testing dataset model groups by their SCH_2 cutoff rate. For this exercise, assume 100 students enrolled in each model group, see Table 5.2, and utilize their respective realized SCH_2 cutoff rates from Table 4.12. Further, we assume the students that meet the cutoff rate generate exactly the minimum cutoff value $C = 52$ SCH_2. With these basic assumptions, we are controlling for extraneous factors to facilitate a simple model group revenue comparison. Thus, using Table 5.2, we considered the potential impact of an

investment in Group 02. According to the constraints set earlier in this example, we can expect that out of 100 students in this sub-cohort, 61 will generate 52 SCH_2 since their respective cutoff rate is 61.2%. Thus, if 61 students for Group 02 generate 52 SCH_2 then the total SCH_2 generated by those students is 3,172 SCH_2. TWU's current published undergraduate resident tuition rate is $230.70 per semester credit hour, which implies the 61 students from model Group 02 will generate $731,780.40 in SCH_2 revenue. Comparing the revenue from the students that meet the cutoff rate of Group 02 to Group 04, Group 02 generates $287,913.60 more revenue than Group 04. Similarly, comparing the revenue from the students that meet the cutoff rate of Group 02 to Group 05, Group 02 generates $347,895.60 more revenue than Group 05. Considering the potential revenue benefits of the model groups, we leave it up to the university's decision makers to determine how best to allocate resources in order to increase the yield rate of applicants-to-enrolled in the model groups that generate credit hours for the university.

Table 5.2

Test Data, Revenue of Model Groups by SCH Cutoff Rate

Model Group	Enroll Total	SCH Cutoff Rate	# Meet Cutoff	SCH_2 Cutoff	Total SCH_2	SCH_2 Revenue
01	100	69.3%	69	52	3,588	$ 827,751.60
02	100	61.2%	61	52	3,172	$ 731,780.40
03	100	48.9%	48	52	2,496	$ 575,827.20
04	100	37.5%	37	52	1,924	$ 443,866.80
05	100	32.5%	32	52	1,664	$ 383,884.80
Total	500	48.5%	247	52	12,844	$2,963,110.80

The obvious goal of an investment proposition should be to increase the yield from high performing zones. By doing so, the yield increase in these higher performing zones will lead to increased revenue generated for the university as well as increased P2 and graduation rates. However, an additional goal can be to address under performance in yield considering other factors as well. For example Table 5.3 presents the overall proportional distribution of applications by ethnicity for the training dataset, which is the data we would utilize to inform the tactical targeting for the testing model data. Notice that Hispanics make up 53% of all applications, which is more than all other ethnicities combined. If we narrow the focus on the high performance zone Group 02, we see that Hispanics make up almost half, 49%, of all applicants in this group.

Table 5.3

Various Ethnic Breakdowns for Training Data Applicants

Group	% White	% Black	% Hispanic	% Asian	% Other
01	36.2%	7.6%	36.3%	17.7%	2.2%
02	24.4%	12.3%	49.2%	11.6%	2.4%
03	16.3%	17.0%	55.2%	9.5%	2.0%
04	13.6%	20.6%	56.8%	7.2%	1.9%
05	9.7%	27.9%	54.5%	6.0%	1.8%
Total	**16.1%**	**20.3%**	**52.9%**	**8.7%**	**2.0%**

However, in Table 5.4, Hispanics make up only 38% of all actual enrolled students, a 15% disparity to their application percentage. Additionally, Hispanics account for only 32% of the Group 02 actual enrolled students, a 17% disparity to their application percentage specific to this group. On the other hand, all other ethnicities are over represented in their respective enrollment percentage relative to their application

percentage in both the overall and Group 02 specific totals. This over representation is especially true for white applicants.

Table 5.4

Various Ethnic Breakdowns for Training Data Enrolled Students

Group	% White	% Black	% Hispanic	% Asian	% Asian
01	52.5%	8.0%	25.5%	12.0%	2.1%
02	36.0%	13.8%	32.4%	15.5%	2.4%
03	27.0%	19.5%	39.5%	11.4%	2.6%
04	18.3%	27.1%	43.6%	8.3%	2.7%
05	16.6%	32.2%	39.4%	9.8%	1.9%
Total	**25.8%**	**23.1%**	**37.8%**	**10.9%**	**2.3%**

It is even more striking to see this phenomenon in terms of yield rate, or modeling group zones, see Table 5.5. We can clearly see that Hispanics have a much lower matriculation rate of 18% overall, which is 22% below the 40% yield rate of white applicants. Further, looking at the high performance zone Group 02 in the same table, we see nearly identical yield rates respective to each of these two ethnicities. To revisit Table 5.3 to make a more direct comparison, Hispanic applicants are more than twice the number of white applicants in the high performance zone Group 02 (49.2% to 24.4%), but white applicants enroll at more than twice the rate of Hispanic applicants (40.5% to 18.1%) for this same group. As an institution in the state of Texas with a growing demographic of Hispanics, it may be wise to address this yield problem through targeted initiatives utilizing performance zones from which appropriate investment and ROI are considered. For a more in depth look at the training data breakdowns, please see Appendix E.

Table 5.5

Percentage of Applicants (App) who Enroll (Enr) for Training Data

Group	White	Black	Hispanic	Asian	Other
	Enr/App	Enr/App	Enr/App	Enr/App	Enr/App
01	46.6%	33.8%	22.6%	21.8%	31.8%
02	40.5%	30.6%	18.1%	36.6%	26.7%
03	42.4%	29.3%	18.3%	30.7%	33.3%
04	38.2%	37.2%	21.7%	32.6%	40.8%
05	36.7%	24.9%	15.6%	35.1%	22.6%
Total	**40.7%**	**29.0%**	**18.2%**	**31.9%**	**29.8%**

Missing Data and Other Techniques

As mentioned in Chapter 3, in this research we defined the set **P** to contain only those applicants that submitted both application GPA and Rank. Before restricting the training dataset with this criteria, the number of applicants who submitted both application GPA and Rank accounted for 84.3% of all completed applications. However, there were 11.2% of students that had an application GPA but no Rank. Interestingly, some high schools are doing away with ranking their students (Balingit, 2015). To deal with these missing data, future research entails imputing high school rank, which will then allow assignment to performance zone modeling. For further research, we are also considering using alternative statistical approaches, such as k-means clustering, to determine the performance zones. Such an approach would substitute the dichotomous dependent variable in this study and replace it with a continuous variable, that is the sum of all SCH accumulated by a student over time period t = two academic years.

REFERENCES

Ahluwalia, Preet Mohan Singh. (2006). Enrollment prediction using Bayesian multiple logistic regression. ProQuest Dissertations Publishing.

Balingit, M. (2015, July 13). *High schools are doing away with class rank. What does that mean for college admissions?* The Washington Post. https://www.washingtonpost.com/news/grade-point/wp/2015/07/13/high-schools-are-doing-away-with-class-rank-what-does-that-mean-for-college-admissions/

Beale, A. V. (1970). the evolution of college admission requirements November 1970. *The National ACAC Journal*, *15*(3), 20–22.

Broome, E. C. (1903). *A Historical and Critical Discussion of College Admission Requirements*. Macmillan.

Davidson, M. (2002). The interpretation of diagnostic tests: A primer for physiotherapists. *Australian Journal of Physiotherapy*, *48*(3), 227–232. https://doi.org/10.1016/S0004-9514(14)60228-2

Eduventures. (2013). *Predictive Analytics in Higher Education: Data-Driven Decision-Making for the Student Life Cycle*. [White paper]. https://www.immagic.com/eLibrary/ARCHIVES/GENERAL/EDUVEN_US/E130131P.pdf

Hamner, M., Stankey, M., & Gengo, G. (2019). Architecting big data for Bayesian gateway course analysis. *AIR Forum 2019*.

Harvard University Fact Book. (2020). *Degrees Awarded Summary*. https://oir.harvard.edu/fact-book/degrees-awarded-summary

Hsieh, F., & Turnbull, B. W. (1996). Nonparametric methods for evaluating diagnostic tests. *Statistica Sinica*, *6*(1), 47–62.

Hubler, S. (2020). University of California will end us of SAT and ACT in admissions. *The New York Times*. https://www.nytimes.com/2020/05/21/us/university-california-sat-act.html

Lang, D. M. (2007). Class rank, GPA, and valedictorians: How high schools rank students. *American Secondary Education*, *35*(2), 36–48. http://www.jstor.org/stable/41406287

Langston, R., Hamner, M., & Stankey, M. (2018). Hidden in plain sight. *Strategic Enrollment Management Conference*, *79*(2), 58–61. https://doi.org/10.1215/00666637-8620384

Larner, A. J. (2015). Optimising the cutoffs of cognitive screening instruments in pragmatic diagnostic accuracy studies: Maximising accuracy or the youden index? *Dementia and Geriatric Cognitive Disorders*, *39*(3–4), 167–175. https://doi.org/10.1159/000369883

Lawrence, J. (2020). *Goodness of fit in logistic regression*. 1–17. http://www.medicine.mcgill.ca/epidemiology/joseph/courses/epib-621/logfit.pdf

Levitz, R. S., Noel, L., & Richter, B. J. (1999). Strategic moves for retention success. *New Directions for Higher Education*, *1999*(108), 31–49. https://doi.org/10.1002/he.10803

Manrai, A. K., Bhatia, G., Strymish, J., Kohane, I. S., & Jain, S. H. (2014). Medicine's uncomfortable relationship with math: Calculating positive predictive value. *JAMA Internal Medicine*, *174*(6), 991–993. https://doi.org/10.1001/jamainternmed.2014.1059

Marsh, C. M., Vandehey, M. A., & Diekhoff, G. M. (2008). A comparison of an introductory course to SAT/ACT scores in predicting student performance. *The Journal of General Education*, *57*(4), 244–255. https://doi.org/10.1353/jge.0.0024

McGee, K. (2021, February 25). *UT-Austin won't require SAT or ACT scores for 2022 applications due to COVID-19*. Texas Tribune. https://www.texastribune.org/2021/02/25/ut-austin-texas-sat-act-application/

National Student Clearinghouse Research Center. (2020). *Term Enrollment Estimates Fall 2020.* https://nscresearchcenter.org/wp-content/uploads/CTEE_Report_Fall_2020.pdf

Rice, G., Coughlin, M. A., & Howard, R. (2011). *The Association for Institutional Research : The First 50 Years*. Association for Institutional Research https://www.airweb.org/docs/default-source/documents-for-pages/reports-and-publications/air-first-50-years.pdf

Ruopp, M. D., Perkins, N. J., Whitcomb, B. W., & Schisterman, E. F. (2008). Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *National Institutes of Health*, *23*(1), 1–7.

Salvatori, P. (2001). Reliability and validity of admissions tools used to select students for the health professions. *Advances in Health Sciences Education*, *6*(2), 159–175. https://doi.org/10.1023/A:1011489618208

Schisterman, E. F., Perkins, N. J., Liu, A., & Bondell, H. (2005). Optimal cut-point and its corresponding Youden index to discriminate individuals using pooled blood samples. *Epidemiology*, *16*(1), 73–81. https://doi.org/10.1097/01.ede.0000147512.81966.ba

Texas Woman's University. (2020, November 20). *First Year Admissions Requirements*. https://twu.edu/admissions/first-year-admissions/first-year-admission-requirements/

Texas Woman's University. (2021, January 13). *Texas State Accountability System*. https://twu.edu/institutional-research/texas-state-accountability-system/

The Crimson. (1890, December 9). *The First Harvard Graduate*. https://www.thecrimson.com/article/1890/12/9/the-first-harvard-graduate-the-following/

Yousafzai, I. I., & Jamil, B. (2019). Relationship between admission criteria and academic performance: a correlational study in nursing students. *Pakistan Journal of Medical Sciences*, *35*(3), 858–861. https://doi.org/10.12669/pjms.35.3.217

Zwick, R. (2019). Assessment in American higher education: The role of admissions tests. *Annals of the American Academy of Political and Social Science*, *683*(1), 130–148. https://doi.org/10.1177/0002716219843469

APPENDIX A

SAS Code

```
%INCLUDE 'C:\Users\SAS_Credentials\mycreds.sas';

/*    Take de-identified data and only keep students
      who have both AP_GPA and AP_RANK_PERCENT  */

DATA THESIS.FTIC_APPLY_14_18_GPARANK_SCRUB;
      SET THESIS.FTIC_APPLY_14_18_SCRUB;

      IF AP_GPA1 = . THEN DELETE;
      IF AP_RANK_PERCENT1 = . THEN DELETE;
      IF AP_RANK_PERCENT1 = 0 THEN DELETE;
/*    We found one occurrence in the data where a
      Missing AP_RANK_PERCENT1 was changed to a 0.
      This was removed also */
RUN;


/*    Split this dataset into two datasets:
      * 14,15,16 students who enrolled (_SUBSET)
      * everyone else (_LEFT)
*/

DATA  THESIS.FTIC_APPLY_14_16_GPARANK_SUBSET
      THESIS.FTIC_APPLY_14_18_GPARANK_LEFT;
      SET THESIS.FTIC_APPLY_14_18_GPARANK_SCRUB;

      IF SUBSTR(TERM,1,2) IN ('14','15','16') AND SCH27 NE .
            THEN OUTPUT THESIS.FTIC_APPLY_14_16_GPARANK_SUBSET;
      ELSE OUTPUT THESIS.FTIC_APPLY_14_18_GPARANK_LEFT;
RUN;

/*    Move the main table to the work library   */
DATA CUTOFF_DATA;
      SET THESIS.FTIC_APPLY_14_16_GPARANK_SUBSET;

RUN;

/*    Set up the Youden Macro to run for multiple
      different parameters                                    */
%MACRO YOUDEN(P=,SCH=,);
%LET CONTINUOUS_VAR = &SCH.;

PROC LOGISTIC DATA=CUTOFF_DATA DESCENDING NOPRINT;
      MODEL &P. = &CONTINUOUS_VAR / OUTROC=ROC_STATS_CUTOFF;
/*    ROC_STATS_CUTOFF dataset contains the sensitivity and
      specificity values for different probability cutoffs.
      We will use Youden to select which probability cutoff (_Prob_)
      is optimal and key (link) it back to the CUTOFF_LOGIT_DATA.
*/
      OUTPUT OUT=CUTOFF_LOGIT_DATA
/*Probability scored dataset using the continuous variable as a
predictor.*/
      P=PROB
```

```
/*Probility field (Prob) in the CUTOFF_LOGIT_DATA is a key to link
optimal cutoff.*/
      XBETA=LOGIT;
RUN;

DATA CHECK;
      SET ROC_STATS_CUTOFF;
      _SPECIF_ = (1 - _1MSPEC_);
      J = _SENSIT_ + _SPECIF_ - 1;
/*    Youden's, Youden's J index is used to select the optimal
      predicted probability cut-off. It is the maximum vertical
      distance between ROC curve and diagonal line. The idea is to
      maximize the difference between True Positive and False Positive.
*/
RUN;

PROC SQL NOPRINT;
      CREATE TABLE CUTOFF AS
            SELECT _PROB_,
/* Use this variable to link back to dataset for the cutoff record */
                  J,
                  _SENSIT_,
                  _SPECIF_,
                  _1MSPEC_

            FROM  CHECK

            HAVING J = MAX(J);
/*criteria for determining cutoff record(s)*/
QUIT;

/* Attach the record from above back to CUTOFF_LOGIT_DATA dataset to
find the corresponding date that maximizes Youden */
PROC SQL;
      CREATE TABLE CUTOFF_LOGIT_DATA2 AS
            SELECT LD.* ,
                  CO._PROB_,
                  J,
                  _SENSIT_,
                  _SPECIF_,
                  _1MSPEC_

            FROM CUTOFF_LOGIT_DATA LD
                 LEFT JOIN CUTOFF  CO ON (LD.PROB = CO._PROB_)

            ORDER BY CO._PROB_ DESC;
QUIT;

DATA MACRO_DATE_DATA;
      SET CUTOFF_LOGIT_DATA2;
      IF _PROB_ NE . THEN CALL SYMPUT
("OPTIMAL_CUTOFF",STRIP(&CONTINUOUS_VAR.));
RUN;
```

```
%PUT &OPTIMAL_CUTOFF;
/*    Check the log if desired      */

DATA EXPLORE_&P.;
      SET CUTOFF_LOGIT_DATA2;

    OPTIMAL_CUTOFF_GTET=.;
    IF &CONTINUOUS_VAR NE . THEN DO;
            IF (&CONTINUOUS_VAR >= &OPTIMAL_CUTOFF.) THEN
OPTIMAL_CUTOFF_GTET=1;
        ELSE OPTIMAL_CUTOFF_GTET=0;
      END;
RUN;

PROC FREQ DATA=EXPLORE_&P.;
            TABLES OPTIMAL_CUTOFF_GTET*&P.;

RUN;

PROC MEANS N STD NMISS MEAN Q1 MEDIAN Q3 QRANGE
            DATA=EXPLORE_&P.;
            VAR &SCH.;
RUN;
%MEND YOUDEN;

/*
      Below is the macro to call the above code.
      We are passing P2 variables. This could be done
      With other persistence variables as well.
*/
%YOUDEN(P=P2,SCH=P2_SCH);


/*    Combine main table back with everything else    */
DATA COMBINE;
      SET
            WORK.CUTOFF_LOGIT_DATA2
             THESIS.FTIC_APPLY_14_18_GPARANK_LEFT;

RUN;


/*    Create flags for various cutoff thresholds      */
DATA THESIS.EXPLORE_P2_GT_AND_GTET_FULL_2;
      RETAIN SCRUB_STU_ID _LEVEL_ LOGIT PROB _PROB_
             GT_OPTIMAL_CUTOFF_45 GTET_OPTIMAL_CUTOFF_45
             GT_OPTIMAL_CUTOFF_51 GTET_OPTIMAL_CUTOFF_51;
      SET COMBINE;

GT_OPTIMAL_CUTOFF_45=.;
    IF P2_SCH NE . THEN DO;
        IF (P2_SCH > 45) THEN GT_OPTIMAL_CUTOFF_45=1;
        ELSE GT_OPTIMAL_CUTOFF_45=0;
    END;
```

```
GTET_OPTIMAL_CUTOFF_45=.;
    IF P2_SCH NE . THEN DO;
        IF (P2_SCH >= 45) THEN GTET_OPTIMAL_CUTOFF_45=1;
        ELSE GTET_OPTIMAL_CUTOFF_45=0;
    END;

      GTET_OPTIMAL_CUTOFF_48=.;
    IF P2_SCH NE . THEN DO;
        IF (P2_SCH >= 48) THEN GTET_OPTIMAL_CUTOFF_48=1;
        ELSE GTET_OPTIMAL_CUTOFF_48=0;
    END;

GT_OPTIMAL_CUTOFF_48=.;
    IF P2_SCH NE . THEN DO;
        IF (P2_SCH > 48) THEN GT_OPTIMAL_CUTOFF_48=1;
        ELSE GT_OPTIMAL_CUTOFF_48=0;
    END;

GT_OPTIMAL_CUTOFF_51=.;
    IF P2_SCH NE . THEN DO;
        IF (P2_SCH > 51) THEN GT_OPTIMAL_CUTOFF_51=1;
        ELSE GT_OPTIMAL_CUTOFF_51=0;
    END;

GTET_OPTIMAL_CUTOFF_51=.;
    IF P2_SCH NE . THEN DO;
        IF (P2_SCH >= 51) THEN GTET_OPTIMAL_CUTOFF_51=1;
        ELSE GTET_OPTIMAL_CUTOFF_51=0;
    END;
RUN;


/*
      Now that we have the various Cutoff Score options, we move onto
the implementing its usage. We knew from a %PUT statement earlier that
Youden 'picked' 45 - but we created other variables within a reasonable
range.
*/
```

```
/*MAKE DEPENDENT VARIABLE A MACRO VARIABLE*/
%LET DEPENDENT_VAR=GT_OPTIMAL_CUTOFF_51;
*%LET DEPENDENT_VAR=GTET_OPTIMAL_CUTOFF_51;
*%LET DEPENDENT_VAR=GT_OPTIMAL_CUTOFF_48;
*%LET DEPENDENT_VAR=GTET_OPTIMAL_CUTOFF_48;
*%LET DEPENDENT_VAR=GT_OPTIMAL_CUTOFF_45;
*%LET DEPENDENT_VAR=GTET_OPTIMAL_CUTOFF_45;

/*****************************************************************/
/*
Train the model by using 3-years of Fall data. Since two year
persistence is part of the model, the most recent year in the Train
dataset should be 2 years prior to Test Fall year of interest.  For
this study, I will use Fall Apply 14-16 data to predict Fall Apply 18
outcome.
Thus, using 3-years of data to Train the model and 1-year of data to
Test the model.
/*****************************************************************
*****************/

%LET PREDICT_COHORT=18; /*FALL APPLY COHORT*/

/*YEAR_BACK1 CREATES Primary data that will be used to make the model*/
%LET YEAR_BACK2=%SYSFUNC(PUTN(%EVAL(&PREDICT_COHORT-2),Z2.));/*NEED TWO
DIGIT VALUES*/
%LET YEAR_BACK3=%SYSFUNC(PUTN(%EVAL(&PREDICT_COHORT-3),Z2.));/*NEED TWO
DIGIT VALUES*/
%LET YEAR_BACK4=%SYSFUNC(PUTN(%EVAL(&PREDICT_COHORT-4),Z2.));/*NEED TWO
DIGIT VALUES*/

/**********************/
/*Bring in Thesis Data*/
/**********************/


DATA GRID_DATA;
      SET GG_DATA.Explore_p2_gt_and_gtet_full_2;

/* ONLY CONSIDERING COMPLETE APPLICATIONS, ONES ABOUT WHICH A DECISION
CAN BE MADE */
      IF AP_COMPLETE='YES';

/* POSSIBLE INDPENDENT VARIABLES: AP_GPA1,AP_RANK_PERCENT1,
      SAT_GRID<--CONVERTS act TO sat, AP_SAT_VM_OLD*/
/* POSSIBLE DEPENDENT VARIABLES: GT_OPTIMAL_CUTOFF_45,
GTET_OPTIMAL_CUTOFF_45, GT_OPTIMAL_CUTOFF_51, GTET_OPTIMAL_CUTOFF_51*/
      /*MINOR DATA CLEAN UP*/
      ARRAY VARIABLES {1:3} AP_GPA1 AP_RANK_PERCENT1 AP_SAT_VM_OLD;
      DO I=1 TO 3;
            /*CLEAN*/
            IF VARIABLES{I}=0 THEN VARIABLES{I}=. ;
                  ELSE VARIABLES{I}=VARIABLES{I};
      END;
RUN;
```

```
/*--CREATE TRAINING(MODEL BUILDING)& TESTING DATASETS--*/
DATA GRID_DATA_TRAIN GRID_DATA_TEST;
      SET GRID_DATA;

      /*TERMS TO BUILD MODEL & TEST*/
      IF SUBSTR(TERM,1,2) IN
("&YEAR_BACK2","&YEAR_BACK3","&YEAR_BACK4") THEN OUTPUT
GRID_DATA_TRAIN;
      IF SUBSTR(TERM,1,2) IN ("&PREDICT_COHORT") THEN OUTPUT
GRID_DATA_TEST;

RUN;


/*----------------1:LOGISTIC MODEL SELECTION--------------------------
---*/

/*macronize the indepependent variables*/

%let Explanatory=
/*------Application VARIABLES------*/
/*HAVE_EXAM REFERENCE GROUP=0 (DID NOT SUBMIT ENTRY TEST SCORES)*/
/*GENDER REFERENCE GROUP='M'*/
AP_GPA1 /**/
AP_RANK_PERCENT1 /**/
/*AP_SAT_VM_OLD */
;


PROC LOGISTIC DATA=GRID_DATA_TRAIN DESCENDING OUTEST=FIT_LOGISTIC
COVOUT OUTMODEL=TRY;
   TITLE "FTIC APPLY GRID ENROLLMENT MODEL";
/*DEPENDENT VARIABLE DEALS WITH YES/NO SCH ACCUMULATION AT CUTOFF OR
NOT*/
   MODEL &DEPENDENT_VAR = &Explanatory

                  /SELECTION=none
                  CTABLE PPROB = (0 TO 1 BY .05)
                  LACKFIT
                  RISKLIMITS
                  PPROB=.4    /*CUTOFF POINT*/
                  OUTROC=ROC;

   OUTPUT out=TRAIN_RESULTS p = prob xbeta =logit;
/*APPEND TO THE DATASET (FOR EACH OBS) LOGIT VALUE AND PROBABILITY*/
   ods output ParameterEstimates=LOGISTIC_PARMS;
/*THIS IS FOR POOLING MODEL RESULTS*/

/*----------2: CALCULATE PROBILITY ON TESTING DATASET----------------*/
      score data = GRID_DATA_TEST out = TEST_DATA_SCORED ;
RUN;
```

```
/*----3: REPORT RESULTS OF MODEL PREDICTION ON TRAINING DATASET----*/

/* THIS IS JUST APPLICANTS WHO ENROLLED FROM TRAINING DATASET*/
DATA TRAIN_MODEL_RESULTS;
      SET TRAIN_RESULTS;

      IF &DEPENDENT_VAR. NE .;

      PROB_SCH=PROB2;

/*DELETE THOSE THAT ENROLLED BUT DIDN'T HAVE BOTH RANK AND AP_GPA*/
      IF PROB_SCH=. THEN DELETE /*GP_PR='00'*/;

      IF  0<=PROB_SCH<.1 THEN GP_PR='10';
      ELSE IF .1<=PROB_SCH<.2 THEN GP_PR='09';
      ELSE IF .2<=PROB_SCH<.3 THEN GP_PR='08';
      ELSE IF .3<=PROB_SCH<.4 THEN GP_PR='07';
      ELSE IF .4<=PROB_SCH<.5 THEN GP_PR='06';
      ELSE IF .5<=PROB_SCH<.6 THEN GP_PR='05';
      ELSE IF .6<=PROB_SCH<.7 THEN GP_PR='04';
      ELSE IF .7<=PROB_SCH<.8 THEN GP_PR='03';
      ELSE IF .8<=PROB_SCH<.9 THEN GP_PR='02';
      ELSE IF .9<=PROB_SCH<=1  THEN GP_PR='01';
      ELSE GP_PR=' ';

/*AFTER VIEWING THE RESULTS OF THE ABOVE DECILES*/
      IF  0<=PROB_SCH<.4 THEN GP2_PR='05';/*7,8,9,10 above*/
      ELSE IF .4<=PROB_SCH<.5 THEN GP2_PR='04';/*6 above*/
      ELSE IF .5<=PROB_SCH<.6 THEN GP2_PR='03';/*5 above*/
      ELSE IF .6<=PROB_SCH<.7 THEN GP2_PR='02';/*4 above*/
      ELSE IF .7<=PROB_SCH<=1 THEN GP2_PR='01';/*1,2,3 above*/
      ELSE GP2_PR=' ';

      COUNT=1;
RUN;

/* THESE ARE ALL APPLICANTS FROM TRAINING COHORT */
DATA TRAIN_MODEL_RESULTS_2;
      SET TRAIN_RESULTS;

      /*ONLY KEEP DATA FOR THOSE THAT HAVE BOTH INDEPENDENT VARIABLES*/
      IF HAVE_GPA = 1 AND HAVE_RANK = 1;

      /*GIVE UNIQUE COUNT TO THOSE WHO HAVE ENROLLED*/
      IF &DEPENDENT_VAR. NE . THEN COUNT = 1;

      PROB_SCH=PROB2;
/*DELETE THOSE THAT ENROLLED BUT DIDN'T HAVE BOTH RANK AND AP_GPA*/
      IF PROB_SCH=. THEN DELETE /*GP_PR='00'*/;

      IF  0<=PROB_SCH<.1 THEN GP_PR='10';
      ELSE IF .1<=PROB_SCH<.2 THEN GP_PR='09';
      ELSE IF .2<=PROB_SCH<.3 THEN GP_PR='08';
      ELSE IF .3<=PROB_SCH<.4 THEN GP_PR='07';
```

```
      ELSE IF .4<=PROB_SCH<.5 THEN GP_PR='06';
      ELSE IF .5<=PROB_SCH<.6 THEN GP_PR='05';
      ELSE IF .6<=PROB_SCH<.7 THEN GP_PR='04';
      ELSE IF .7<=PROB_SCH<.8 THEN GP_PR='03';
      ELSE IF .8<=PROB_SCH<.9 THEN GP_PR='02';
      ELSE IF .9<=PROB_SCH<=1 THEN GP_PR='01';
      ELSE GP_PR=' ';
      /*after viewing results above*/
      IF  0<=PROB_SCH<.4 THEN GP2_PR='05';/*7,8,9,10 above*/
      ELSE IF .4<=PROB_SCH<.5 THEN GP2_PR='04';/*6 above*/
      ELSE IF .5<=PROB_SCH<.6 THEN GP2_PR='03';/*5 above*/
      ELSE IF .6<=PROB_SCH<.7 THEN GP2_PR='02';/*4 above*/
      ELSE IF .7<=PROB_SCH<=1 THEN GP2_PR='01';/*1,2,3 above*/
      ELSE GP2_PR=' ';

      ALL_COUNT=1;

RUN;


TITLE;
TITLE 'Training 10 Groups Enrolled';
%sort(dsn=train_model_results,var=GP_PR);
proc sgplot data=TRAIN_MODEL_RESULTS;
      Xaxis                VALUES =(1.4 TO 4   BY .2) GRID/**/;
      yaxis reverse  VALUES =(0 TO 100 BY 10)         GRID/**/;
    *styleattrs datacontrastcolors=(grey red orange yellow green blue
purple);
      scatter x=AP_GPA1 y=AP_RANK_PERCENT1 / group=GP_PR;
      *reg   y=AP_RANK_PERCENT1 x=AP_GPA1 / group=GP_PR ;
run;

TITLE;
TITLE 'Training 10 Groups All Apps';
%sort(dsn=train_model_results_2,var=GP_PR);
proc sgplot data=TRAIN_MODEL_RESULTS_2;
      Xaxis                VALUES =(1.4 TO 4   BY .2) GRID/**/;
      yaxis reverse  VALUES =(0 TO 100 BY 10)         GRID/**/;
    *styleattrs datacontrastcolors=(grey red orange yellow green blue
purple);
      scatter x=AP_GPA1 y=AP_RANK_PERCENT1 / group=GP_PR;
      *reg   y=AP_RANK_PERCENT1 x=AP_GPA1 / group=GP_PR ;
run;

TITLE;
TITLE 'Training 5 Groups Enrolled';
%sort(dsn=train_model_results,var=GP2_PR);
/*after viewing results above*/
proc sgplot data=TRAIN_MODEL_RESULTS;
      Xaxis                VALUES =(1.4 TO 4   BY .2) GRID/**/;
      yaxis reverse  VALUES =(0 TO 100 BY 10)         GRID/**/;
    styleattrs datacontrastcolors=(ggr vibg vio bilg lip);
      scatter x=AP_GPA1 y=AP_RANK_PERCENT1 / group=GP2_PR;
      *reg     y=AP_RANK_PERCENT1 x=AP_GPA1 / group=GP2_PR ;
```

```
run;

TITLE;
TITLE 'Training 5 Groups All Apps';
%sort(dsn=train_model_results_2,var=GP2_PR);
/*after viewing results above*/
proc sgplot data=TRAIN_MODEL_RESULTS_2;
      Xaxis                VALUES =(1.4 TO 4   BY .2) GRID/**/;
      yaxis reverse  VALUES =(0 TO 100 BY 10)         GRID/**/;
    styleattrs datacontrastcolors=(vibg ggr vio bilg lip);
      scatter x=AP_GPA1 y=AP_RANK_PERCENT1 / group=GP2_PR;
      *reg     y=AP_RANK_PERCENT1 x=AP_GPA1 / group=GP2_PR ;
run;


PROC REPORT DATA=TRAIN_MODEL_RESULTS PS=60 LS=110 MISSING SPLIT='\'
NOWINDOWS

      STYLE(REPORT)={BACKGROUND=WHITE /*CELLPADDING = 1.2PT CELLSPACING
= 0PT frame=box rules=groups*/}
      STYLE(HEADER)={FONT=("times new roman",12PT)
BACKGROUND=lightSTEELblue FOREGROUND=MAROON FONT_WEIGHT=BOLD}
      STYLE(COLUMN)={FONT=("times new roman",11PT) FOREGROUND=BLACK
/*CELLWIDTH=1.25IN*/};

      TITLE1 'Texas Woman''s Uninversity';
      TITLE2 'Train Model Data';
      TITLE3 "Dependent Variable:&DEPENDENT_VAR";


      COLUMN ("Train Apply FTIC Grid Prediction" GP2_PR COUNT ("Apply
Grid Model"  PROB_SCH conv1) ("Apply Grid Actual Results"
&DEPENDENT_VAR conv2));

      DEFINE GP2_PR      / GROUP /*FORMAT=$GP_FMT. */ORDER=INTERNAL
WIDTH=12    'Model\Level';
      DEFINE COUNT       / ANALYSIS SUM FORMAT=COMMA6.
      WIDTH=10          'Enroll\Total' CENTER;

      DEFINE PROB_SCH  / ANALYSIS sum FORMAT=COMMA6.        WIDTH=8
      'Probability\Sum' CENTER;
      DEFINE CONV1       / COMPUTED FORMAT=PRCNT_NEW. WIDTH=9
'AVG\Probability';

      DEFINE &DEPENDENT_VAR / ANALYSIS SUM FORMAT=COMMA6.   WIDTH=8
      'SCH_Cutoff\Sum' CENTER;
      DEFINE CONV2      / COMPUTED FORMAT=PRCNT_NEW. WIDTH=9
'AVG\SCH_Cutoff';

   COMPUTE CONV1;
        IF COUNT.SUM=0 THEN CONV1=.;ELSE
           CONV1=PROB_SCH.SUM/COUNT.SUM;
   ENDCOMP;
```

```
   COMPUTE CONV2;
        IF COUNT.SUM=0 THEN CONV2=.;ELSE
           CONV2=&DEPENDENT_VAR..SUM/COUNT.SUM;
   ENDCOMP;

      RBREAK AFTER /SUMMARIZE STYLE={FONT_WEIGHT=BOLD FOREGROUND=MAROON
BACKGROUND=lightSTEELblue};
      COMPUTE AFTER;
            CALL DEFINE ('GP_PR', "STYLE", "STYLE={PRETEXT='Total'
FONT_WEIGHT=BOLD JUST=LEFT}");
    ENDCOMP;

RUN;


PROC REPORT DATA=TRAIN_MODEL_RESULTS_2 PS=60 LS=110 MISSING SPLIT='\'
NOWINDOWS

      STYLE(REPORT)={BACKGROUND=WHITE /*CELLPADDING = 1.2PT CELLSPACING
= 0PT frame=box rules=groups*/}
      STYLE(HEADER)={FONT=("times new roman",12PT)
BACKGROUND=lightSTEELblue FOREGROUND=MAROON FONT_WEIGHT=BOLD}
      STYLE(COLUMN)={FONT=("times new roman",11PT) FOREGROUND=BLACK
/*CELLWIDTH=1.25IN*/};

      TITLE1 'Texas Woman''s Uninversity';
      TITLE2 'Train Model Data';
      TITLE3 "Dependent Variable:&DEPENDENT_VAR";

      COLUMN ("Train Apply FTIC Grid Prediction" GP2_PR ("Apply Grid
Model" PROB_SCH conv1) ("Apply Grid Actual Results" COUNT
&DEPENDENT_VAR conv2));

      DEFINE GP2_PR      / GROUP /*FORMAT=$GP2_FMT. */ORDER=INTERNAL
WIDTH=12    'Model\Level';

      DEFINE PROB_SCH  / ANALYSIS sum FORMAT=COMMA6.        WIDTH=8
      'Probability\Sum' CENTER;
      DEFINE CONV1       / COMPUTED FORMAT=PRCNT_NEW. WIDTH=9
'AVG\Probability';

      DEFINE COUNT       / ANALYSIS SUM FORMAT=COMMA6.
      WIDTH=10          'Enroll\Total' CENTER;
      DEFINE &DEPENDENT_VAR / ANALYSIS SUM FORMAT=COMMA6.   WIDTH=8
      'SCH_Cutoff\Sum' CENTER;
      DEFINE CONV2      / COMPUTED FORMAT=PRCNT_NEW. WIDTH=9
'AVG\SCH_Cutoff';

   COMPUTE CONV1;
        IF COUNT.SUM=0 THEN CONV1=.;ELSE
           CONV1=PROB_SCH.SUM/COUNT.SUM;
   ENDCOMP;
   COMPUTE CONV2;
        IF COUNT.SUM=0 THEN CONV2=.;ELSE
```

```
            CONV2=&DEPENDENT_VAR..SUM/COUNT.SUM;
   ENDCOMP;

      RBREAK AFTER /SUMMARIZE STYLE={FONT_WEIGHT=BOLD FOREGROUND=MAROON
BACKGROUND=lightSTEELblue};
      COMPUTE AFTER;
            CALL DEFINE ('GP2_PR', "STYLE", "STYLE={PRETEXT='Total'
FONT_WEIGHT=BOLD JUST=LEFT}");
    ENDCOMP;

RUN;



/*----4: REPORT RESULTS OF MODEL PREDICTION ON TESTING DATA SET----*/
DATA TEST_Model_RESULTS;
      SET TEST_DATA_SCORED;

      IF &DEPENDENT_VAR. NE .;
      PROB_SCH=P_1;

/*DELETE THOSE THAT ENROLLED BUT DIDN'T HAVE BOTH RANK AND AP_GPA*/
      IF PROB_SCH=. THEN DELETE /*GP_PR='00'*/;

      IF  0<=PROB_SCH<.4 THEN GP2_PR='05';/*7,8,9,10 above*/
      ELSE IF .4<=PROB_SCH<.5 THEN GP2_PR='04';/*6 above*/
      ELSE IF .5<=PROB_SCH<.6 THEN GP2_PR='03';/*5 above*/
      ELSE IF .6<=PROB_SCH<.7 THEN GP2_PR='02';/*4 above*/
      ELSE IF .7<=PROB_SCH<=1 THEN GP2_PR='01';/*1,2,3 above*/
      ELSE GP2_PR=' ';

      COUNT=1;
RUN;


DATA TEST_Model_RESULTS_2;
      SET TEST_DATA_SCORED;

      /*ONLY KEEP DATA FOR THOSE THAT HAVE BOTH INDEPENDENT VARIABLES*/
      IF HAVE_GPA = 1 AND HAVE_RANK = 1;

      /*GIVE UNIQUE COUNT TO THOSE WHO HAVE ENROLLED*/
      IF &DEPENDENT_VAR. NE . THEN COUNT = 1;
      PROB_SCH=P_1;
      *IF COUNT = 1 THEN PROB_SCH2 = PROB_SCH;

/*DELETE THOSE THAT ENROLLED BUT DIDN'T HAVE BOTH RANK AND AP_GPA*/
      IF PROB_SCH=. THEN DELETE /*GP_PR='00'*/;

      IF  0<=PROB_SCH<.4 THEN GP2_PR='05';/*7,8,9,10 above*/
      ELSE IF .4<=PROB_SCH<.5 THEN GP2_PR='04';/*6 above*/
      ELSE IF .5<=PROB_SCH<.6 THEN GP2_PR='03';/*5 above*/
      ELSE IF .6<=PROB_SCH<.7 THEN GP2_PR='02';/*4 above*/
```

```
      ELSE IF .7<=PROB_SCH<=1 THEN GP2_PR='01';/*1,2,3 above*/
      ELSE GP2_PR=' ';

      ALL_COUNT=1;
RUN;


PROC REPORT DATA=TEST_Model_RESULTS PS=60 LS=110 MISSING SPLIT='\'
NOWINDOWS

      STYLE(REPORT)={BACKGROUND=WHITE /*CELLPADDING = 1.2PT CELLSPACING
= 0PT frame=box rules=groups*/}
      STYLE(HEADER)={FONT=("times new roman",12PT)
BACKGROUND=lightSTEELblue FOREGROUND=MAROON FONT_WEIGHT=BOLD}
      STYLE(COLUMN)={FONT=("times new roman",11PT) FOREGROUND=BLACK
/*CELLWIDTH=1.25IN*/};

      TITLE1 'Texas Woman''s Uninversity';
      TITLE2 'Test Model Data';
      TITLE3 "Dependent Variable:&DEPENDENT_VAR";

      COLUMN ("Test Apply FTIC Grid Prediction" GP2_PR COUNT ("Apply
Grid Model"  PROB_SCH conv1) ("Apply Grid Actual Results"
&DEPENDENT_VAR conv2));

      DEFINE GP2_PR      / GROUP /*FORMAT=$GP2_FMT. */ORDER=INTERNAL
WIDTH=12    'Model\Level';
      DEFINE COUNT       / ANALYSIS SUM FORMAT=COMMA6.
      WIDTH=10          'Enroll\Total' CENTER;

      DEFINE PROB_SCH  / ANALYSIS sum FORMAT=COMMA6.        WIDTH=8
      'Probability\Sum' CENTER;
      DEFINE CONV1       / COMPUTED FORMAT=PRCNT_NEW. WIDTH=9
'AVG\Probability';

      DEFINE &DEPENDENT_VAR / ANALYSIS SUM FORMAT=COMMA6.   WIDTH=8
      'SCH_Cutoff\Sum' CENTER;
      DEFINE CONV2      / COMPUTED FORMAT=PRCNT_NEW. WIDTH=9
'AVG\SCH_Cutoff';

   COMPUTE CONV1;
        IF COUNT.SUM=0 THEN CONV1=.;ELSE
           CONV1=PROB_SCH.SUM/COUNT.SUM;
   ENDCOMP;
   COMPUTE CONV2;
        IF COUNT.SUM=0 THEN CONV2=.;ELSE
           CONV2=&DEPENDENT_VAR..SUM/COUNT.SUM;
   ENDCOMP;

      RBREAK AFTER /SUMMARIZE STYLE={FONT_WEIGHT=BOLD FOREGROUND=MAROON
BACKGROUND=lightSTEELblue};
      COMPUTE AFTER;
            CALL DEFINE ('GP2_PR', "STYLE", "STYLE={PRETEXT='Total'
FONT_WEIGHT=BOLD JUST=LEFT}");
```

```
    ENDCOMP;

RUN;


PROC REPORT DATA=TEST_Model_RESULTS_2 PS=60 LS=110 MISSING SPLIT='\'
NOWINDOWS

      STYLE(REPORT)={BACKGROUND=WHITE /*CELLPADDING = 1.2PT CELLSPACING
= 0PT frame=box rules=groups*/}
      STYLE(HEADER)={FONT=("times new roman",12PT)
BACKGROUND=lightSTEELblue FOREGROUND=MAROON FONT_WEIGHT=BOLD}
      STYLE(COLUMN)={FONT=("times new roman",11PT) FOREGROUND=BLACK
/*CELLWIDTH=1.25IN*/};

      TITLE1 'Texas Woman''s Uninversity';
      TITLE2 'Test Model Data 2';
      TITLE3 "Dependent Variable:&DEPENDENT_VAR";

      COLUMN ("Test Apply FTIC Grid Prediction" GP2_PR ALL_COUNT
("Apply Grid Model"  PROB_SCH conv1) ("Apply Grid Actual Results" COUNT
&DEPENDENT_VAR conv2));

      DEFINE GP2_PR      / GROUP /*FORMAT=$GP2_FMT. */ORDER=INTERNAL
WIDTH=12    'Model\Level';
      DEFINE ALL_COUNT / ANALYSIS SUM FORMAT=COMMA6. WIDTH=10 'All
Applicants' CENTER;

      DEFINE PROB_SCH  / ANALYSIS SUM FORMAT=COMMA6.        WIDTH=8
      'Probability\Sum' CENTER;
      DEFINE CONV1       / COMPUTED FORMAT=PRCNT_NEW. WIDTH=9
'AVG\Probability';

      DEFINE COUNT       / ANALYSIS SUM FORMAT=COMMA6.
      WIDTH=10          'Enroll\Total' CENTER;
      DEFINE &DEPENDENT_VAR / ANALYSIS SUM FORMAT=COMMA6.   WIDTH=8
      'SCH_Cutoff\Sum' CENTER;
      DEFINE CONV2      / COMPUTED FORMAT=PRCNT_NEW. WIDTH=9
'AVG\SCH_Cutoff';

      COMPUTE CONV1;
            IF COUNT.SUM=0 THEN CONV1=.;
            ELSE CONV1=PROB_SCH.SUM/ALL_COUNT.SUM;
      ENDCOMP;
      COMPUTE CONV2;
            IF COUNT.SUM=0 THEN CONV2=.;
            ELSE CONV2=&DEPENDENT_VAR..SUM/COUNT.SUM;
      ENDCOMP;

      RBREAK AFTER /SUMMARIZE STYLE={FONT_WEIGHT=BOLD FOREGROUND=MAROON
BACKGROUND=lightSTEELblue};
      COMPUTE AFTER;
            CALL DEFINE ('GP2_PR', "STYLE", "STYLE={PRETEXT='Total'
FONT_WEIGHT=BOLD JUST=LEFT}");
```

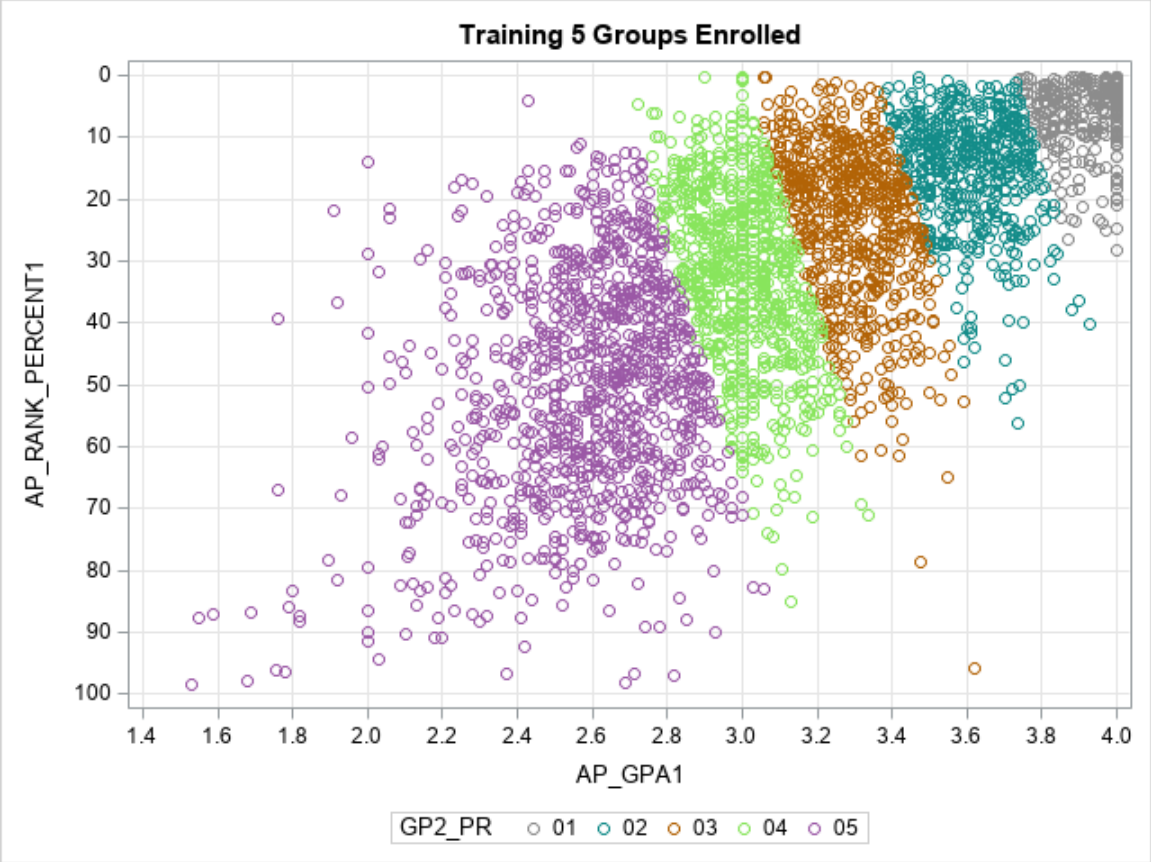
```
    ENDCOMP;

RUN;

TITLE;
TITLE 'Testing 5 Groups Enrolled';
%sort(dsn=test_model_results,var=GP2_PR);
proc sgplot data=TEST_Model_RESULTS;
      Xaxis                VALUES =(1.4 TO 4    BY .2) GRID/**/;
      yaxis reverse  VALUES =(0 TO 100 BY 10)          GRID/**/;
    styleattrs datacontrastcolors=(ggr vibg vio bilg lip);
      scatter x=AP_GPA1 y=AP_RANK_PERCENT1 / group=GP2_PR;
      *reg     y=AP_RANK_PERCENT1 x=AP_GPA1 / group=GP2_PR ;
run;

TITLE;
TITLE 'Testing 5 Groups All Apps';
%sort(dsn=test_model_results_2,var=GP2_PR);
proc sgplot data=TEST_Model_RESULTS_2;
      Xaxis                VALUES =(1.4 TO 4    BY .2) GRID/**/;
      yaxis reverse  VALUES =(0 TO 100 BY 10)          GRID/**/;
    styleattrs datacontrastcolors=(bilg lip vibg vio ggr);
      scatter x=AP_GPA1 y=AP_RANK_PERCENT1 / group=GP2_PR;
      *reg     y=AP_RANK_PERCENT1 x=AP_GPA1 / group=GP2_PR ;
run;



/* Code for Chi-Square Analysis */

title;
title 'Frequency/Expected/Chi-square - Train';
proc freq data=train_model_results;
 tables gt_optimal_cutoff_51*gp2_pr / expected chisq cellchi2 norow
nocol;
run;

title;
title 'Frequency/Expected/Chi-square - Test';
proc freq data=test_model_results;
 tables gt_optimal_cutoff_51*gp2_pr / expected chisq cellchi2 norow
nocol;
run;
```

```
/*    Code for ANOVA Analysis */

%sort(dsn=train_model_results,var=gp2_pr);
title;
title 'ANOVA Post Hoc - Train';
proc anova data = train_model_results;
      class gp2_pr;
      model cum_sch_after_s45 = gp2_pr;
      means gp2_pr / tukey scheffe;
run;

title;

title 'ANOVA Post Hoc - Test';
%sort(dsn=test_model_results,var=gp2_pr);
proc anova data = test_model_results;
      class gp2_pr;
      model cum_sch_after_s45 = gp2_pr;
      means gp2_pr / tukey scheffe;
run;



/*    Code for Correlation    */

title;
title 'AP GPA with AP Rank Correlation - Train';
proc corr data = train_model_results;
      var ap_gpa1 ap_rank_percent1;
run;

title;
title 'AP GPA with AP Rank Correlation - Test';
proc corr data = test_model_results;
      var ap_gpa1 ap_rank_percent1;
run;
```

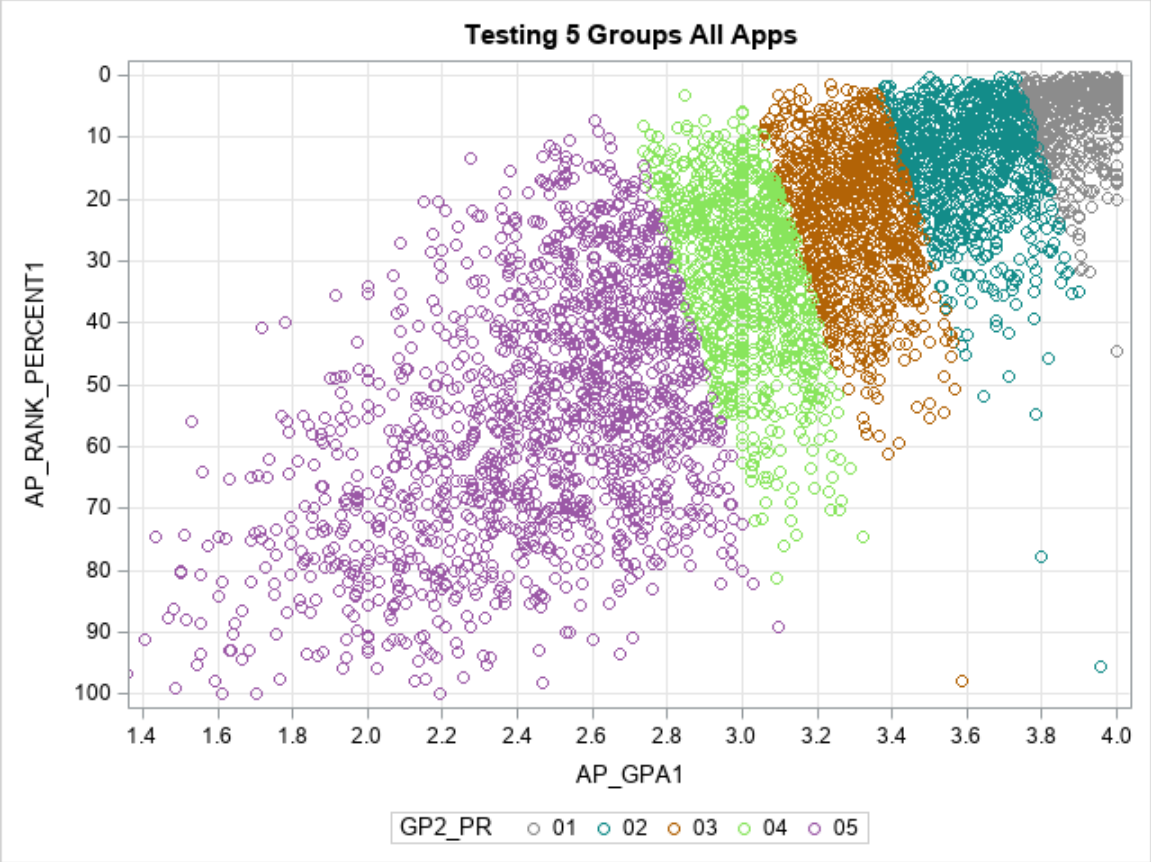
APPENDIX B

Scatterplots of Training and Testing Data

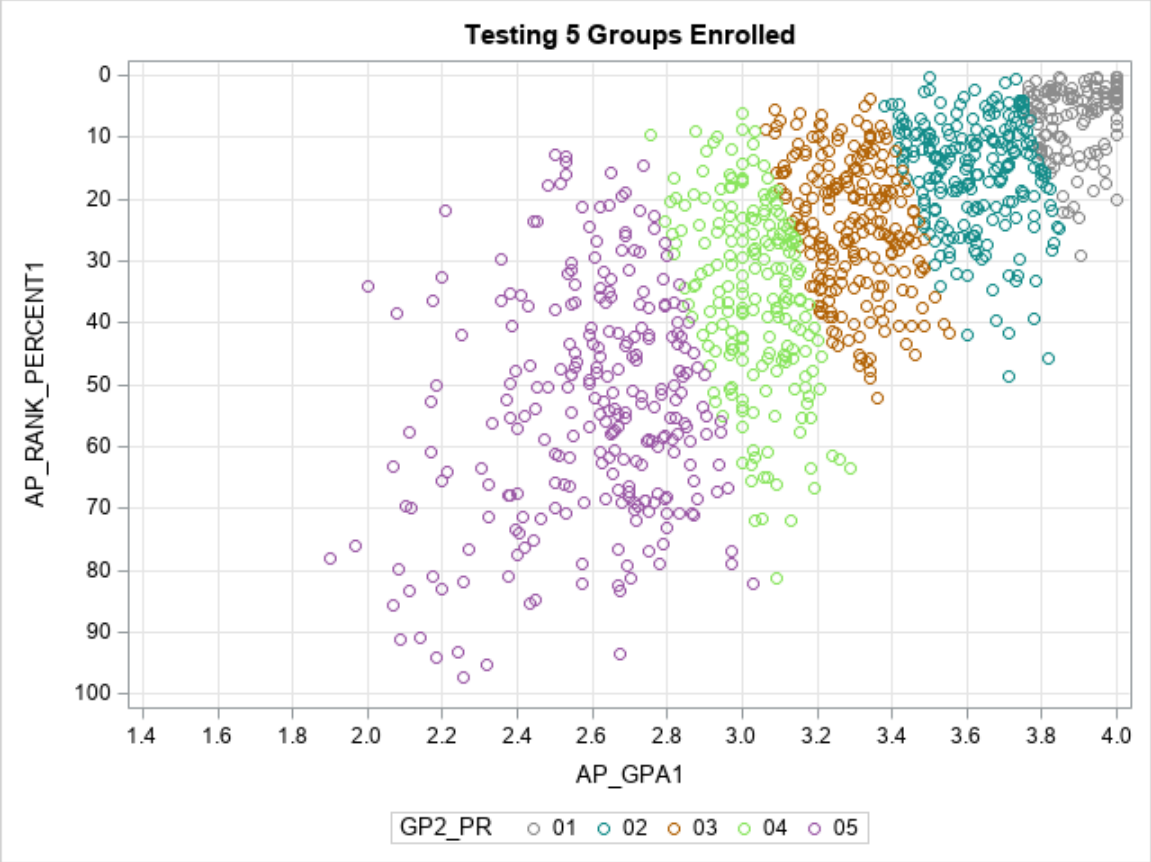
Training Dataset, 5 Groups, All Enrolled Performance Grid



Testing Dataset, 5 Groups, All Applied Performance Grid



Testing Dataset, 5 Groups, All Enrolled Performance Grid



APPENDIX C

Chi-Square Output

Chi-Square Statistics for Training Dataset

Table of GT_OPTIMAL_CUTOFF_51 by GP2_PR						
GT_OPTIMAL_CUTOFF_51	**GP2_PR**					
Frequency **Expected** **Cell Chi-Square** **Percent**	**01**	**02**	**03**	**04**	**05**	**Total**
0	68 166.58 58.341 2.14	178 260.1 25.912 5.59	283 319.37 4.1419 8.89	449 376.6 13.918 14.10	649 504.35 41.487 20.38	1627 51.10
1	258 159.42 60.964 8.10	331 248.9 27.077 10.40	342 305.63 4.3281 10.74	288 360.4 14.544 9.05	338 482.65 43.352 10.62	1557 48.90
Total	326 10.24	509 15.99	625 19.63	737 23.15	987 31.00	3184 100.00

Statistics for Table of GT_OPTIMAL_CUTOFF_51 by GP2_PR

Statistic	**DF**	**Value**	**Prob**
Chi-Square	4	294.0651	<.0001
Likelihood Ratio Chi-Square	4	303.9437	<.0001
Mantel-Haenszel Chi-Square	1	286.1707	<.0001
Phi Coefficient		0.3039	
Contingency Coefficient		0.2908	
Cramer's V		0.3039	

APPENDIX D

ANOVA Output

ANOVA – P2_GPA by GP2_PR (Groups)

Class Level Information		
Class	**Levels**	**Values**
GP2_PR	5	01 02 03 04 05

Number of Observations Read	3184
Number of Observations Used	3184

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	64269.7731	16067.4433	64.46	<.0001
Error	3179	792361.6628	249.2487		
Corrected Total	3183	856631.4359			

R-Square	Coeff Var	Root MSE	CUM_SCH_AFTER_S45 Mean
0.075026	34.39652	15.78761	45.89887

Source	DF	Anova SS	Mean Square	F Value	Pr > F
GP2_PR	4	64269.77313	16067.44328	64.46	<.0001

Tukey's Honest Significant (Post-Hoc) Test for Training Dataset

Alpha	0.05
Error Degrees of Freedom	3179
Error Mean Square	249.2487
Critical Value of Studentized Range	3.85987

Comparisons significant at the 0.05 level are indicated by *.**				
GP2_PR Comparison	**Difference Between Means**	**Simultaneous 95% Confidence Limits**		
01 - 02	4.7409	1.6842	7.7976	***
01 - 03	7.2569	4.3130	10.2007	***
01 - 04	10.3021	7.4360	13.1683	***
01 - 05	14.0826	11.3300	16.8351	***
02 - 01	-4.7409	-7.7976	-1.6842	***
02 - 03	2.5160	-0.0567	5.0886	
02 - 04	5.5612	3.0778	8.0446	***
02 - 05	9.3416	6.9903	11.6930	***
03 - 01	-7.2569	-10.2007	-4.3130	***
03 - 02	-2.5160	-5.0886	0.0567	
03 - 04	3.0452	0.7021	5.3883	***
03 - 05	6.8257	4.6230	9.0284	***
04 - 01	-10.3021	-13.1683	-7.4360	***
04 - 02	-5.5612	-8.0446	-3.0778	***
04 - 03	-3.0452	-5.3883	-0.7021	***
04 - 05	3.7804	1.6827	5.8782	***
05 - 01	-14.0826	-16.8351	-11.3300	***
05 - 02	-9.3416	-11.6930	-6.9903	***

Comparisons significant at the 0.05 level are indicated by ***.				
GP2_PR Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
05 - 03	-6.8257	-9.0284	-4.6230	***
05 - 04	-3.7804	-5.8782	-1.6827	***

Scheffe's Test for Training Dataset

Alpha	0.05
Error Degrees of Freedom	3179
Error Mean Square	249.2487
Critical Value of F	2.37473

Comparisons significant at the 0.05 level are indicated by ***.				
GP2_PR Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
01 - 02	4.7409	1.2892	8.1926	***
01 - 03	7.2569	3.9326	10.5811	***
01 - 04	10.3021	7.0656	13.5386	***
01 - 05	14.0826	10.9743	17.1908	***
02 - 01	-4.7409	-8.1926	-1.2892	***
02 - 03	2.5160	-0.3891	5.4211	
02 - 04	5.5612	2.7569	8.3655	***
02 - 05	9.3416	6.6864	11.9969	***
03 - 01	-7.2569	-10.5811	-3.9326	***
03 - 02	-2.5160	-5.4211	0.3891	
03 - 04	3.0452	0.3994	5.6911	***

Comparisons significant at the 0.05 level are indicated by ***.				
GP2_PR Comparison	**Difference Between Means**	**Simultaneous 95% Confidence Limits**		
03 - 05	6.8257	4.3383	9.3130	***
04 - 01	-10.3021	-13.5386	-7.0656	***
04 - 02	-5.5612	-8.3655	-2.7569	***
04 - 03	-3.0452	-5.6911	-0.3994	***
04 - 05	3.7804	1.4116	6.1492	***
05 - 01	-14.0826	-17.1908	-10.9743	***
05 - 02	-9.3416	-11.9969	-6.6864	***
05 - 03	-6.8257	-9.3130	-4.3383	***
05 - 04	-3.7804	-6.1492	-1.4116	***

APPENDIX E

Additional Academic and Demographic Metrics for Training Dataset

All Enrolled and P2 =1 – AP_GPA and AP_Rank Breakdown

Group	Average AP_GPA	High AP_GPA	Low AP_GPA	High AP_Rank	Low AP_Rank
01	3.904112727	4	3.72	0.16	40.25
02	3.590650538	3.84	3.36	0.37	52.25
03	3.264402948	3.62	3	0.6	95.91
04	2.968438202	3.32	2.71	0.4	83.05
05	2.555778905	2.97	1.68	4.15	97.97

All Enrolled and P2 = 1 – P2_GPA and P2_SCH Breakdown

Group	Average P2_GPA	High P2_GPA	Low P2_GPA	Average P2_SCH	High P2_SCH	Low P2_SCH
01	3.7058875	4	1.85	58.56	94	21
02	3.46507046	4	1.08571429	56.63709677	90	5
03	3.17316976	4	1.04545455	54.93611794	76	19
04	2.89184832	4	1	52.15505618	82	12
05	2.739878	4	0.97777778	52.86612576	88	19

All Enrolled (P2 =1 and P2 = 0) – AP_GPA and AP_Rank Breakdown

Group	Average AP_GPA	High AP_GPA	Low AP_GPA	High AP_Rank	Low AP_Rank
01	3.901104294	4	3.72	0.16	40.25
02	3.584713163	3.84	3.36	0.37	52.25
03	3.2652736	3.62	3	0.2	100
04	2.970339213	3.32	2.63	0.4	83.05
05	2.54851773	2.97	1.53	4.15	98.42

All Enrolled (P2 =1 and P2 = 0) – P2_GPA and P2_SCH Breakdown

Group	Average P2_GPA	High P2_GPA	Low P2_GPA	Average P2_SCH	High P2_SCH	Low P2_SCH
01	3.648258808	4	1.48	54.83128834	94	6
02	3.279851077	4	0	50.09037328	90	0
03	2.941817724	4	0	47.5744	76	0
04	2.587533397	4	0	44.52917232	82	0
05	2.219696204	4	0	40.74873354	88	0

All Applicants (Enrolled = 1 and Enrolled = 0) – AP_GPA and AP_Rank Breakdown

Group	Average AP_GPA	High AP_GPA	Low AP_GPA	High AP_Rank	Low AP_Rank
01	3.896528134	4	3.71	0.16	50.77
02	3.571774246	3.84	3.33	0.2	52.25
03	3.250027016	3.79	3	0.2	100
04	2.955791091	3.33	2.63	0.4	94.17
05	2.399612214	3.06	1	0.4	100

All Applicants – Ethnicity Frequency Breakdown

Group	Apply Total	White	Black	Hispanic	Asian	Other
01	1,013	367	77	368	179	22
02	1,856	452	229	914	216	45
03	2,443	399	416	1,349	231	48
04	2,604	353	537	1,478	187	49
05	4,585	447	1,277	2,501	276	84
Total	**12,501**	**2,018**	**2,536**	**6,610**	**1,089**	**248**

All Enrolled – Ethnicity Frequency Breakdown

Group	Apply Total	White	Black	Hispanic	Asian	Other
01	326	171	26	83	39	7
02	509	183	70	165	79	12
03	625	169	122	247	71	16
04	737	135	200	321	61	20
05	987	164	318	389	97	19
Total	**3,184**	**822**	**736**	**1,205**	**347**	**74**

Proportion Enrolled From Applicants for Each Group

Group	White	Black	Hispanic	Asian	Other
01	0.46594	0.33766	0.22554	0.21788	0.31818
02	0.40487	0.30568	0.18053	0.36574	0.26667
03	0.42356	0.29327	0.18310	0.30736	0.33333
04	0.38244	0.37244	0.21719	0.32620	0.40816
05	0.36689	0.24902	0.15554	0.35145	0.22619
Total	**0.40733**	**0.29022**	**0.18230**	**0.31864**	**0.29839**

All Applicants – Ethnicity Percentage Breakdown

Group	Apply Total	White	Black	Hispanic	Asian	Other
01	36.23%	7.60%	36.33%	17.67%	2.17%	36.23%
02	24.35%	12.34%	49.25%	11.64%	2.42%	24.35%
03	16.33%	17.03%	55.22%	9.46%	1.96%	16.33%
04	13.56%	20.62%	56.76%	7.18%	1.88%	13.56%
05	9.75%	27.85%	54.55%	6.02%	1.83%	9.75%
Overall	**16.14%**	**20.29%**	**52.88%**	**8.71%**	**1.98%**	**16.14%**

All Enrolled – Ethnicity Percentage Breakdown

Group	Apply Total	White	Black	Hispanic	Asian	Other
01	52.45%	7.98%	25.46%	11.96%	2.15%	52.45%
02	35.95%	13.75%	32.42%	15.52%	2.36%	35.95%
03	27.04%	19.52%	39.52%	11.36%	2.56%	27.04%
04	18.32%	27.14%	43.55%	8.28%	2.71%	18.32%
05	16.62%	32.22%	39.41%	9.83%	1.93%	16.62%
Overall	**25.82%**	**23.12%**	**37.85%**	**10.90%**	**2.32%**	**25.82%**