MODELING NURSE-ENTRY SUCCESS BY INTEGRATING STUDENT CHARACTERISTICS WITH GATEWAY COURSE PERFORMANCE

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF MATHEMATICS

IN THE GRADUATE SCHOOL OF THE

TEXAS WOMAN'S UNIVERSITY

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

COLLEGE OF ARTS AND SCIENCES

BY

EMILY APPICELLI, B.S.

DENTON, TEXAS

MAY 2020

DEDICATION

For my precious son and loving husband,

without whom any personal accomplishment would be vain.

ACKNOWLEDMENTS

I would like to extend my sincere gratitude to my thesis advisor, Dr. Mark S. Hamner, for teaching me more than any single other educator that I have encountered in my lifetime. As an advisor, a professor, a supervisor, and a mentor, he never failed to provide invaluable assistance and encouragement and made pursuing higher education at Texas Woman's University an unforgettable experience.

I also wish to thank my committee members, Dr. Don Edwards and Dr. Brandi Falley, for generously offering their time and guidance. To the TWU Office of Institutional Research and Data Management and the incredible Nora Sierra: thank you for your data resources, scheduling flexibility, and constant support throughout this project.

I am greatly indebted to my parents, who have always believed in me more than I deserve, and to my husband who enthusiastically championed me every step of the way. Finally, thanks to my son, Luke, who attended graduate courses with me in utero and has become my sweetest source of motivation.

ABSTRACT

EMILY APPICELLI
MODELING NURSE-ENTRY SUCCESS BY INTEGRATING STUDENT CHARACTERISTICS WITH GATEWAY COURSE PERFORMANCE
MAY 2020

Decades of undergraduate retention research has uncovered common predictors of student success and methods, such as early-alert intervention, that promote persistence among academically at-risk students. This study addresses low retention rates among nurse-entry majors enrolled at Texas Woman's University (TWU) by integrating a gateway course performance variable with other student characteristic variables and using predictive analytics to identify at-risk nurse-entry students after their first semester of coursework. A logistic regression model was built to predict the probability of being admitted into the upper division nursing program after persisting for two to three years. Significant interaction between gateway course performance and first-semester GPA was detected, conveying valuable insight into the odds of success for these students based on first semester behavior. Because this model exhibits exceptional predictive accuracy, it may realistically serve as a basis for early-alert intervention programs among TWU nurse-entry students in the future.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Introduction

The new fall semester will bring over 3 million first-time freshmen onto university campuses across the nation (Hussar & Bailey, 2011). Each one of these students represents a unique blend of social, cultural, and academic experience; yet, all have chosen to pay the high price of investing in postsecondary education. As they excitedly embark on this new stage of life, many may be unaware of the challenges that lie ahead. Based on the average undergraduate graduation rate in this country, only about half of them can expect to persist to graduation and earn the degree that they seek.

Undergraduate retention has been defined as "an institution of higher education's ability to retain a student from admission until graduation" (Demetriou & Schmitz-Sciborski, 2011, p. 1). Unwaveringly low undergraduate retention rates in this country have had researchers asking "why?" since as early as the 1930s. The U.S. Department of Interior and Office of Education commissioned a nation-wide study in 1937 to determine the extent to which students withdraw from college and attempt to pinpoint the reasons for their departure (McNeely, 1938). Since then, the study of student retention within higher education has continued to grow and evolve. Global events, social change, and political legislation brought a boom in university enrollment in the 1960s which, in turn, spurred a wealth of new research on the topic (Demetriou & Schmitz-Sciborski, 2011). It was during this period that systematic and theoretical studies in this field began to emerge

(Aljohani, 2016). In the 1970s, men such as William Spady and Vincent Tinto made lasting contributions to the field by developing a theoretical framework by which to understand the phenomenon of student retention and attrition as a whole. Much of today's present research is founded on Tinto's 1975 Student Integration Model and its subsequent revisions, in which he emphasizes the effect of academic and social integration along with institutional factors on student departure (Voigt & Hundrieser, 2008). Later theorists, such as John Bean (1982) identified the importance of student demographic and background characteristics in determining retention. More recent retention research has focused on increasing retention of underrepresented and/or disadvantaged student populations and calls upon campus-wide, cross-departmental support for student retention improvement (Demetriou & Schmitz-Sciborski, 2011).

After working for over 20 years across hundreds of institutions, three retention and enrollment management experts observed that "the success of an institution and the success of its students are inseparable" (Levitz, Noel, & Richter, 1999, p. 31). It is in a school's best interest to foster the success of its students and develop strategies to promote the best likelihood for their success. The 2004 ACT Policy Report concerning improving college retention found:

> The most successful retention strategies often use an early alert, assessment, and monitoring system based on academic factors such as high school and/or college GPA, test scores (ACT Assessment, tests in college courses), and other performance indicators. . . . This information alerts institutions to students who may have potential difficulties and enables them to direct these students into

> retention programs before their risk of dropping out increases. (Lotkowski, Robbins, & Noeth, 2004, p. 20)

Thus, predicting the outcome of student success early and with as much accuracy as possible is an extremely valuable topic of research that has the potential of an "early alert" system for appropriate intervention.

Tinto observes that the "key to successful student retention lies with the institution…in the ability of faculty and staff to apply what is known about student retention to the specific situation in which the institution finds itself" (Tinto, 1993, p. 4). Similarly, Voigt & Hundrieser (2008) encourage the identification of "specialty sub-populations" within the student body and the use of data to drive a deeper analysis of where "persistence and progression issues occur" (p. 4). With the advent of the internet and increasing computational power, the newly emerging field of predictive analytics within higher education is making identifying such target groups for academic intervention easier and more effective than ever before.

Purpose

We will use predictive analytics to address the low retention rates of a special sub-population enrolled at Texas Woman's University (TWU). TWU boasts of some unique characteristics, including the fact that it is the largest university primarily for women in the nation. It has an average enrollment number of 15,000 students, which is made up of about 90% females and 10% males. In 2018, non-white students made up over 60% of the student body, with 30% of those identifying as Hispanic and 18%

identifying as black. Nearly half of the university's students are the first in their families to go to college.

TWU offers a competitive nursing program with a large amount of applicants each year. In fact, students enrolled in the TWU College of Nursing make up about 20% of the student body. The nursing program is four years in length; however, a student must first complete 46 hours of nursing prerequisite courses before being admitted into the upper division program. Students in the process of completing these prerequisite courses are declared "nurse-entry" majors and make up 33-40% of the total number of incoming first-time-in-college (FTIC) students every fall.

Data obtained through the Office of Institutional Research and Data Management (IRDM), however, reveals that only a small percentage of these students are actually being admitted into the upper division nursing program two to three years later. Some will migrate and successfully complete an alternate major at TWU while the alarming majority are lost from the TWU system completely. We will use predictive analytics to address the low retention rates of FTIC nurse-entry students at TWU.

This study seeks to build a model for FTIC nurse-entry students at TWU that would, after their very first semester, accurately predict the probability of their being admitted into the upper division nursing program two to three years later. On average, for an incoming cohort of nurse-entry students, about 75% who are admitted into the upper division program are admitted after persisting or continuing to accumulate hours as a nurse-entry major, for two years. Another 15% percent are admitted after persisting an additional year. Thus, for the purpose of our model, we define "success" among nurse-

entry majors as admission into the upper division nursing program after persisting for either two or three years, as shown in Figure 1.

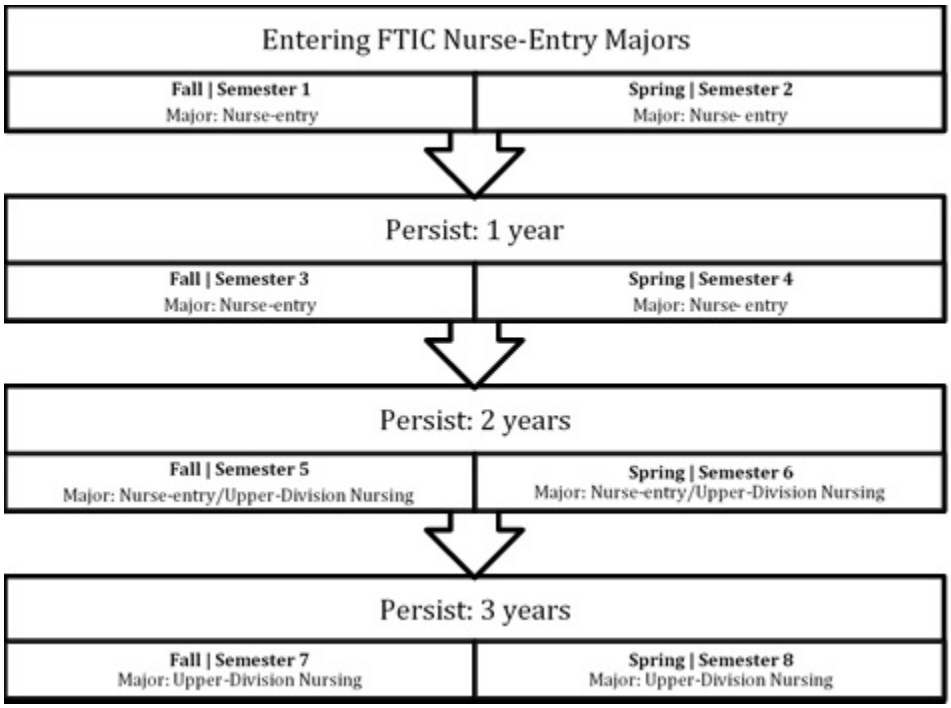


Figure 1. FTIC Nurse-entry "success" path

Historically, the main predictors for entering student academic success have been SAT scores and GPA. These two variables are “independent, positive predictors of undergraduate grades” (Harackiewicz, Barron, Tauer, & Elliot, 2002, p. 562). Another interesting variable to consider in predicting academic “success” is performance in a gateway course. A gateway course is a mandatory first or second year course considered foundational to a field of study (Flanders, 2015; Marsh, Vandehey, & Diekhoff, 2008).

We hypothesize that adding gateway course performance alongside traditionally used predictor variables (SAT scores and GPA) and other student characteristic variables in a prediction model will enhance its predictive accuracy. Through the integration of

gateway course performance with other student characteristic variables, the objectives of this research are as follows:

(i) Identify key independent variables and build a logistic regression model which predicts the probability of admission into the upper division nursing program for FTIC nurse-entry majors who have completed their first semester at TWU;

(ii) Perform analysis to explain the relationship between independent and dependent variables;

(iii) Assess overall model fit and predictive accuracy using cross-validation techniques on existing university data;

(iv) Interpret results and draw conclusions.

Significance of Study

Significant results of this study would have many positive implications. The ability to more accurately predict student success means the potential to make the most efficient use of university resources by targeting at-risk students. It could be the catalyst for early intervention by academic advisors to help facilitate student success through appropriate guidance (Flanders, 2015; Marsh et al., 2008; Stankus, Hamner, Stankey, & Mancuso, 2018). Ultimately, it should reduce the amount of time and money wasted by students and allow them to avoid unnecessary disappointment and frustration. If identified early, “academic rescue efforts could be initiated that would help the students avoid failure and simultaneously enhance the students' connection to the institution” (Marsh et al., 2008, p. 246). As a result, the university could see improved retention and

graduation rates and experience higher student satisfaction. Ultimately, our purpose in developing this statistical model is to foster the success of the student population at Texas Woman's University and see higher numbers persist to graduation.

CHAPTER II

LITERATURE REVIEW

Introduction

This chapter presents a review of the body of literature relevant to our study. We begin with a brief overview of the early theoretical framework that has shaped the last four decades of student retention research. We then turn to predictive analytics, a new and exciting field of study emerging in higher education, and discuss its common methodology and important empirical studies. We identify significant variables that have been shown to be predictive of student success and also explore a lesser known variable that may hold important predictive potential. Finally, we will discuss criticisms and pitfalls related to the use of predictive analytics within higher education.

Theoretical Framework

William Spady (1971) can be credited with producing the first theoretical model in student retention literature. He was also the first to associate student attrition to Durkheim's famous sociological work, "Suicide" (Durkheim, 1951) and draw similarities between suicidal behavior and student attrition behavior. Namely, both behaviors can be thought of as a "form of voluntary withdrawal from a particular society" due to a "lack of social and intellectual integration" into that society (Aljohani, 2016, p. 3). Vincent Tinto (1975, 1993) agreed with the Durkheim suicide theory analogy and built upon Spady's work to develop his famous and widely cited student integration model. This model

explains the college experience as being comprised of an academic and a social element, both of which a student must be integrated into in order to persist at their chosen institution (Metz, 2004). He also draws upon Arnold Van Gennep's work, The Rites of Passage (1960), to bring to light three stages through which a student must successfully pass: separation, transition, and incorporation. The identifies five critical components for understanding individual departure from institutions: pre-entry attributes, goals/commitments, institutional experiences, academic integration, and social integration. This work laid an important theoretical foundation upon which researchers have based empirical studies for decades (Metz, 2004). In fact, "virtually all models of persistence in higher education have evolved from Tinto's (1993) interactionist theory" (Barbera, Berkshire, Boronat, & Kennedy, 2017, p. 17).

Predictive Analytics

While retention research has a history dating back nearly one hundred years, predictive analytics is a relatively new field of study made possible by powerful computer algorithms. Predictive analytics is an area of statistical analysis that, by uncovering relationships and patterns within large volumes of data, can be used to predict behavior and events (Van Barneveld, Arnold, & John, 2012). Although widely employed in business settings, only recently have institutions of higher education begun taking advantage of its capabilities. Used in combination with student retention theory, predictive analytics offers exciting new possibilities to accurately predict student behaviors in the academic setting "– notably in the areas of learning outcomes, recruitment, and retention" (Eduventures, 2013, p. 4).

Although rooted in the theoretical framework that came before it, predictive analytics in higher education answers to some of the frustrations and criticisms made of Tinto's student integration model (1975). In particular, institutional researchers have criticized the model for its lack of generalizability (and thus, practicality) and the resource-intensive procedures it requires to implement (Caison, 2006; Metz, 2004). Often robust survey methodology must be employed in order to gather the necessary information required of the model. The field of predictive analytics, however, focuses less on understanding the phenomenon of student attrition as a general topic and more on developing a timely and efficient method to guide institution-specific decision-making. It "combines select institutional data, statistical analysis, and predictive modeling to create intelligence upon which students, instructors, or administrators can change academic behavior" (Baepler & Murdoch, 2010, p. 3).

Along this line, Caison (2006) compared traditional survey-based retention research methodology with an approach that relies on data commonly available in institutional student databases. She hypothesized that using data already routinely collected for each student, based on Tinto's theoretical model, can result in a model with predictive capability comparable to those using resource-intensive survey methodology. She chose sixteen independent variables that have bearing on the components of Tinto's retention theory, drawn from already existing institutional databases. When comparing the predictive capability of a survey-based model with the institutional research database model, findings suggest that the model based on institutional data actually outperformed the survey-based model.

Also using historical data from institutional databases, Delen (2011) developed models capable of predicting freshmen students who are mostly likely to drop out after their first year. The model was designed to be able to predict after first semester in order to properly develop intervention programs during the next semester in order to retain them. Using neural network, decision tree, and logistic regression they built three separate models that were between 74 and 81% accurate in classifying students into attrition or retention groups after their first semester of college.

In 2009, Purdue University launched *Signals*, an academic early-alert system based on a predictive model that draws upon 20 independent variables. It is integrated into Blackboard and within the first two weeks of the term can indicate to a student their likelihood of success in a number of core courses (Baepler & Murdoch, 2010). A green, yellow, or red light is displayed upon Blackboard login to indicate a particular student's current trajectory. Students are encouraged through faculty feedback to alter academic behavior in order to improve their results. During its pilot studies in 2006 and 2007, 78% of students who received a red light improved their grade by mid-term (Tally, 2009).

Predictive analytics seeks to arm institutional researchers with tools for developing programs to effect positive and timely change (Caison, 2006). The studies mentioned in this section demonstrate the viability of predictive analytics as a means to understand and predict student attrition at specific universities as well as help institutions take evidence-based action towards intervention (Delen, 2011; Baepler & Murdoch, 2010). This tool is helping to identify at-risk students in real-time, allowing institutions the ability to intervene with support faster than ever before.

Methodology

In order to accurately predict student outcomes, predictive analytics within higher education makes use of statistical analysis and predictive modeling. Some commonly used models are classification algorithms such as decision trees (Costa, Fonseca, Santana, de Araújo, & Rego, 2017; Delen, 2011; Raju & Schumacker, 2015;), neural networks (Costa et al., 2017; Delen, 2011; Raju & Schumacker, 2015), Naïve Bayes classifiers (Bydžovská, 2016; Costa et al., 2017), and logistic regression (Callahan & Belcheir, 2015; Delen, 2011; Glynn, Sauer, & Miller, 2011; Peng, So, Stage, & St. John, 2002; Stankus, Hamner, Stankey, & Mancuso, 2018; Yin & Burger, 2003).

These algorithms classify observations, based on input data, into two or more categories. Logistic regression, specifically, has been popular in retention research because it is so well suited towards explaining the relationship between a dichotomous dependent variable (i.e., persist or not persist) and a combination of quantitative and categorical independent variables. Peng et al. (2002) examined three leading higher education journals between the years of 1988 and 1999 and found 52 articles using logistic regression as the primary analytical tool. Of the 52 articles, over half were related to university enrollment and retention.

Logistic regression can explain quantitatively the factors that lead to classification as well as give a ratio of the odds of success on the dependent variable for two different values of an independent variable. It does not require the independent variables to behave normally or be linearly related to the dependent variable and can analyze all types of predictor variables (Mertler & Vannatta, 2010). Because of its established use in retention

literature, model simplicity, interpretability of parameters, and relatively small computational demands (Duh, Walker, Pagano, & Kronlund, 1998), we chose to use logistic regression as the analytical tool for this study. We develop a more thorough mathematical discussion of the logistic regression model in Chapter 3.

Significant Variables

What are the some of the important variables in predicting undergraduate retention? Decades of retention research has uncovered a number of consistently significant variables that wield influence over whether or not a university student will persist through their academic journey.

Raju and Schumacker (2015) attempted to use earliest available data to predict retention leading to graduation. They chose 11 variables that were known before the student began college and three variables that were gathered after the first semester. Their logistic regression model identified first-semester in college GPA, earned hours, student status, and high school GPA as significant variables. They found that first-semester GPA is especially important in predicting student graduation rate. Raju and Schumacker (2015) showed that "Universities do not have to wait until the end of year or even later to find students that need help….Early identification of potential leavers and successful intervention programs are the key for improving student graduation" (p. 587).

In 2017, Barbera et al. conducted a review of relevant retention research studies published after 2010 in order to identify patterns and critical variables useful to the field. They describe how variables including college preparatory assessments (SAT/ACT), academic readiness (GPA and high school curriculum), demographic characteristics,

socioeconomic status, personality traits, receipt of financial aid, first-generation enrollment, transfer history, and institutional characteristics have all shown significance in different studies aimed at improving student retention. Some important findings to note are "that low-income, first-generation, and underrepresented minority students are less likely to graduate" (Barbera et al., 2017, p. 18). Additionally, higher SAT scores and higher GPAs (both high school and first-semester college) were found to be almost invariably linked with persistence across different contexts.

With respect to nursing students specifically, Yin and Burger (2003) examined the relationship of pre-nursing admission variables to passing the National Council Licensure Examination-Registered Nurse (NCLEX-RN), a "standard benchmark for success of nursing graduates" (p. 232). Only data available at the time of nursing admissions was used to build their model, with the hope of providing a tool for admissions committees to select high potential applicants. The most important predictor of success was found to be college GPA prior to admission to the nursing program. Natural science GPAs, introductory psychology course grade, and high school rank were also found to be related to success.

In another attempt to help nursing admissions committees identify quality applicants, Cunningham, Manier, Anderson, and Sarnosky (2014) offer a regression-based selection method for upper division nursing admission. Like Yin and Burger (2003), they use only information available at the time of upper division admissions decision. In their review of the predictive value of admission criteria for undergraduate nursing programs, they found that academic predictors such as GPA, SAT/ACT scores,

and performance in college-level biology, social science, and chemistry have been linked to positive nursing educational outcomes. Overall GPA at entry, science GPA, number of completed prerequisites, and the ATI-TEXAS score (a standardized test for nursing education) were identified as significant predictors.

We can see that whether in reference to first-year retention, persistence to graduation, passing scores on the NCLEX-RN, or nursing school admission, GPA is perhaps the strongest predictor of student success across all student retention studies. Standardized test scores have also shown to be significant in predicting success for most of the literature cited in this section.

The Gateway Course

As we review the variables commonly studied in undergraduate retention, we note that historically the most influential predictors for entering student success in college have been academic indicators such as SAT scores and GPA. One less explored variable in student retention research is performance in a gateway course. A gateway course is typically defined a foundational course marked by high enrollment and high rates of D, F, or W grades,

Of the studies investigating the relationship of gateway course performance to student retention, most focus on its simple effect on the success outcome. For example, Callahan and Belcheir (2015) observed first-time, full-time freshmen who enrolled in both English and math courses during their first semester at Boise State University. They found that those students who earned an “A” in first-year English were three times more likely to persist after one year than their peers who did not pass the course. Flanders

(2015) studied first-to-second semester retention rates among freshmen who declared a major and attempted a gateway course in their field compared to those who declared a major but did not attempt a gateway course in their field and to those who simply did not declare a major at all. The students in his study who successfully completed the gateway course with a passing grade had the highest likelihood of returning for their second semester than any other group.

Marsh, Vandehay, and Diekhoff (2008) took a further step by using multiple regression analysis to combine a gateway course performance variable with SAT/ACT scores in an attempt to predict cumulative GPA of students three semesters later. They found that students' midterm grades in a gateway course were predictive of their future GPA. They assert that gateway course performance is just as powerful as SAT/ACT scores, if not more so, in predicting subsequent academic success.

Finally, Stankus, Hamner, Stankey, and Mancuso (2019) used a Bayesian statistical framework to 1) identify gateway courses that predict admission into upper division nursing and 2) determine a specific "threshold grade" for those courses that would make the biggest difference on eventual success. This research was conducted at our institution on interest, Texas Woman's University. In the context of nurse-entry students, Stankus et al. (2017) define a gateway course as "one in which a threshold grade highly discriminates between persisters and non-persisters and acts as a barrier to progress toward graduation" (p. 2). Their analysis uncovered that about 75% of students who made an A in Introductory Psychology were successfully admitted into the upper

division nursing program, while only around 18% of those who did not make an A were admitted.

It is clear that gateway course performance holds some predictive power with regards to student retention. We note, however, that a gap in the literature exists when it comes to integrating gateway course performance with other commonly used variables in a predictive model built to detect at-risk students. In fact, no studies have done this with nurse-entry students as the focus and upper division admission as the outcome measure. This particular gap in the literature forms the basis of our research study.

Criticism and Pitfalls

We now turn to some criticism of predictive analytics within higher education. In their study, Cunningham et al. (2014) claim that using a regression-based selection method to determine nursing admission can improve a school's ability to identify high-potential applicants and save institutional time and resources spent on rationally justified weighting of students. However, in *In The Promise and Peril of Predictive Analytics in Higher Education: A Landscape Analysis* (Palmer, Iris, Ekowo, Manuela, 2016), the authors warn of the possibility for predictive tools to inadvertently reinforce institutional bias that may have existed for years. They caution that "predictive models can discriminate against historically underserved groups because demographic data such as age, race, gender, and socioeconomic status are often central to their analyses" (Palmer et al., 2016, p. 13).

This is especially concerning in the field of nursing education, where efforts to promote a diverse student population have become a common topic in nursing literature.

Studies have shown that “minority health care providers are more likely to work in underserved areas and provide care for minority populations” (Wros & Noone, 2018, p. 211), yet as the number of ethnic minorities in the U.S. rises, the same trend is not being observed in the nursing workforce (Stankus et al., 2018). Scott & Zerwic (2015) state that “there is a critical need for more diversity in nursing, from the education arena to the workforce” (p. 488).

One approach that some nursing programs are using to increase numbers of diverse graduates is called a holistic admissions review process, in which an applicant’s experiences, attributes, and academic metrics are all considered in a balanced manner. The emphasis here is in identifying those applicants who would not only be successful students but also exhibit qualities that indicate they could make real contributions to their profession in the future (Scott & Zerwic, 2015). This approach often involves one-on-one or group interviews by faculty members to identify desirable abilities which may not be able to be determined through standardized tests or GPA alone. In fact, advocates of the holistic admissions process often wish to move away from an overreliance on academic metrics as they believe this could create a barrier for underrepresented students.

In their landscape analysis, Palmer et al. (2016) send out a call for institutional researchers to use predictive analytics to “mitigate rather than reinforce inequity” (p.14). While in 2016, Mount Saint Mary’s University was found using early-alert predictive analytics to encouraging high-risk students to leave the institution in order to artificially boost retention rates (Palmer et al., 2016; Patterson, 2016); Stankus et al. (2018) illustrate how predictive modeling can be used as a tool to help university programs identify a

more diverse student population for admission. Clearly, whether predictive modeling reinforces or mitigates inequality rests in the hands of the researcher. Armed with an awareness of possible dangers and pitfalls, institutional researchers using predictive analytics have the incredible opportunity to affect positive change within higher education and create enriched possibilities for new generations of students.

Summary

We have now seen a general sketch of the historical and theoretical background behind undergraduate retention research in this country as well as explored common methodology and significant variables associated with modern predictive analytics. In the following chapters we will go on to lay the mathematical framework for our study and to build a predictive model using logistic regression. We will then predict upper division nursing admission based on first-semester data in an effort to identify at-risk nurse-entry students as early as possible in their academic journey. As we thus make use of predictive analytics within the setting of higher education, we take care that our model truly supports student success without inadvertently discriminating against any groups.

CHAPTER III

THE LOGISTIC REGRESSION MODEL

Introduction

In this study, we build a model for FTIC nurse-entry students at TWU that will, after their very first semester, predict the probability of being admitted into the upper division nursing program two to three years later. Therefore, we are interested in estimating the total number of upper division admits that will result from a given term cohort. To facilitate this modeling, we now develop some helpful notation and lay the mathematical groundwork for our statistical analysis

Let $\boldsymbol{P} = \{1, 2,\ldots,N\}$ be the set of labels of the units of a finite population size N nurse-entry students for any cohort of interest at the beginning of their first semester. Associated with each unit $i \in \mathbf{P}$ is the value y_i, which indicates whether or not a student is admitted into the upper division program two to three years later. Specifically,

$$y_i = \begin{cases} 0 & \text{if a student is not admitted into upper division nursing} \\ 1 & \text{if a student is admitted into upper division nursing} \end{cases}$$

for $i = 1,2,\ldots,N$. Here, y is a dichotomous variable with the value 1 indicating "success" and the value 0 indicating "failure."

The total number of students in a given cohort who are admitted into upper division nursing will be represented as

$$T = \sum_{i=1}^{N} y_i \tag{3.1}$$

Our estimate of equation (3.1) is represented by

$$\hat{T} = E(T) = \sum_{i=1}^{N} E(y_i) \tag{3.2}$$

Since expectation is a linear operator, meaning we can distribute it over the summation, E(y_i) represents the expected value of the individual responses. To determine the expected value for each y_i, we associate with the random variable y_i a value π_i, $0 \leq \pi_i \leq 1$, which represents the probability that the i^{th} individual will be admitted into the upper division nursing program. The value $(1 - \pi_i)$ represents the probability that the i^{th} individual will not be admitted.

Each response, y_i, follows a Bernoulli distribution with the probability mass function:

$$f(y_i|\,\pi_i) = (\,\pi_i)^{y_i}(1 - \,\pi_i)^{1-y_i}$$

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

$$\mu_{y_i} = E(y_i) = \,\pi_i \tag{3.3}$$

and

$$\sigma^2 = \,\pi_i(1 - \,\pi_i) \tag{3.4}$$

respectively. Equation (3.3) represents the outcome of the expected value for each y_i and equation (3.4) is the associated variance for the random variable y_i. Substituting equation (3.3) into equation (3.2), equation (3.2) becomes

$$\hat{T} = E(T) = \sum_{i=1}^{N} E(y_i)$$

$$= \sum_{i=1}^{N} \pi_i \tag{3.5}$$

which is the estimate of the total number of admits resulting from a given cohort of interest.

Linear Regression

To understand how we would model equation (3.5), we now present a basic overview of simple linear regression to help guide us into the development of the multiple logistic regression model used in this study. Simple linear regression allows us to model mathematically the relationship between two continuous variables. It seeks to predict the values for a continuous dependent variable, y_i, based on given values for a single independent variable, x_i. Using the notation derived above, we associate the relationship between x_i and y_i with the following equation

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{3.6}$$

where β_0 is an intercept term, β_1 is the coefficient for the independent variable, and ε_i is an error term for the unexplained variation in y_i. The errors here are equal to the differences between the observed y values and predicted y values. They can be expressed mathematically by algebraically rearranging equation (3.6) to become

$$\varepsilon_i = y_i - (\beta_0 - \beta_1 x_i). \tag{3.7}$$

Using equation (3.7), we define the sum of the squared errors, SSE, as

$$SSE = \sum (y_i - \beta_0 - \beta_1 x_i)^2 . \tag{3.8}$$

Minimizing the SSE by taking the appropriate partial derivatives and setting them equal to zero will give us the estimators of β_0 and β_1 from equation (3.6)

$$\frac{\partial SSE}{\partial \beta_0} = -2\sum(y_i - \beta_0 - \beta_1 x_i) = 0 \tag{3.9}$$

$$\frac{\partial SSE}{\partial \beta_0} = -2\sum x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \tag{3.10}$$

Leisy (2019) shows the step-by-step calculations for solving equations (3.9) and (3.10). The results are the following expressions, which are most often denoted b_0 and b_1, and are the estimates of the parameters from equation (3.6)

$$b_0 = \bar{y} - b_1\bar{x} \tag{3.11}$$

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{3.12}$$

We have now demonstrated the ordinary least squares (OLS) method for finding a fitted regression line, or "best fitting line" for the sample data. This line fits the data "best" because it minimizes the SSE. The equation for the fitted regression line is

$$\hat{y}_i = b_0 + b_1 x_i \tag{3.13}$$

where $\hat{y}$ represents the predicted value of the dependent variable. Accordingly, the estimate of error from equation (3.7) is

$$\hat{\varepsilon}_i = y_i - \hat{y}_i \tag{3.14}$$

To justify the use of linear regression to model the association between *x* and *y*, several assumptions need to be met and checked (most of which are related to the error term). The assumptions of linear regression are listed in Table 1.

Table 1
Assumptions of Linear Regression Model

Assumption	Description
1. Linearity	The relationship between *x* and *y* is linear
2. Normality	Errors are normally distributed
3. Homoscedasticity	Errors have constant variance
4. Independent	Errors are independent

The multiple regression model extends the principles of simple linear regression and allows us to perform linear regression analyses on data that include more than one independent variable. Therefore, when there are *k* independent variables present we may write

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \tag{3.15}$$

where β_k, k = 1,2,…k, represents the associated regression coefficients for each independent variable. As demonstrated in equations (3.6) through (3.13), we again minimize the SSE and obtain parameter estimates using the sample data. Thus, the equation used to explain the relationship between one continuous dependent variable and two or more independent variables is

$$\hat{y_i} = b_0 + b_1 x_{i1} + b_2 x_{i1} + \cdots + b_k x_{ik} \tag{3.16}$$

where b_k, k = 1,2,…k, are the estimates of the parameters from equation (3.15).

Logistic Regression

Attempting to run linear regression on data with a dichotomous dependent variable becomes problematic. Allison & SAS Institute (2003) proves in detail that binary data does not meet the normality or homoscedasticity assumptions required of linear regression. In other words, the assumptions of linear regression require that the dependent variable, *y*, be continuous and unbounded and this, of course, is not the case when *y* can only take on the values of 0 or 1.

Logistic regression allows us to take advantage of the mathematical construct of the linear model described in the previous section for a binary dependent variable by means of the relationship between probability, that is, π_i associated with each y_i, and odds. By establishing this relationship and making use of some mathematical manipulation, we may achieve a continuous, unbounded function for the left-hand side of equation (3.16) when *y* is dichotomous. To demonstrate how this works, we first define the odds of an event as the ratio of the probability that an event will occur to the probability that it will not occur

$$odds_i = \frac{probability\ of\ event}{probability\ of\ no\ event} = \frac{\pi_i}{1-\pi_i}\ . \qquad (3.17)$$

Solving for π_i in equation (3.17) gives us the simple relationship between probability and odds

$$\frac{\pi_i}{1-\pi_i} = odds_i$$

$$\pi_i = odds_i(1-\pi_i)$$

$$\pi_i = odds_i - odds_i(\pi_i)$$

$$\pi_i + odds_i(\pi_i) = odds_i$$

$$\pi_i(1 + odds_i) = odds_i$$

$$\pi_i = \frac{odds_i}{1 + odds_i} \tag{3.18}$$

Odds have zero as a lower bound but, unlike probability, have no upper bound; in other words, $0 < odds < \infty$. If we then take the natural log of equation (3.17) we obtain an unbounded, continuous function such that $-\infty < odds < \infty$. Now, analogous to equation (3.15) where y_i was continuous, we have for each y_i of the dichotomous dependent variable the continuous function, $\ln(odds_i)$. Thus, the logit function, from which logistic regression derives its name, is defined as the natural log of equation (3.17) and is equivalent to the linear combination of independent variables from equation (3.15)

$$logit(\pi_i) = \ln(odds_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \tag{3.19}$$

Before moving on to explain how to estimate the coefficients of equation (3.19), we first introduce some matrix notation in order to simplify the subsequent discussion. Let $\mathbf{x_i} = (1, x_{1k}, x_{2k}, \ldots, x_{ik})$ be a row vector with length $k+1$, where k is the number of independent variables specified in the model. The first element of $\mathbf{x_i}$ takes the value of 1, holding the place for the intercept term, β_0.

We also define

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k+1} \end{bmatrix}$$

to be the column vector of regression coefficients with length $k + 1$. Now we can rewrite equation (3.19) in matrix notation:

$$logit(\pi_i) = \ln(odds_i) = \boldsymbol{\beta x_i} \tag{3.20}$$

It is worth noting that the natural logarithm function, $\ln(x)$, is the inverse function of the exponential function, e^x. In other words, these two functions are opposite functions that reverse one another. For example, we know from equation (3.20) that the logit function, $logit(\pi)$, is equal to $\ln(odds)$ and $\boldsymbol{\beta x_i}$. Thus, its inverse can be written

$$e^{\ln(odds_i)} = odds$$

or,

$$e^{\boldsymbol{\beta x_i}} = odds. \tag{3.21}$$

We may use equation (3.21) to rewrite equation (3.18) as

$$\pi_i = \frac{e^{\boldsymbol{\beta x_i}}}{1 + e^{\boldsymbol{\beta x_i}}} \tag{3.22}$$

which represents the predicted probability of success for y_i given a linear combination of independent variables. Summing equation (3.22) over all values of *i* will give us the estimated total from equation (3.5).

The assumptions of logistic regression differ from those of linear regression and, most notably, do not require linearity between the dependent and independent variables or normality and homoscedasticity among errors. The assumptions that are required of the logistic regression model are listed in Table 2.

Table 2
Assumptions of Logistic Regression Model

Assumption	Description
1. Outcome structure	The dependent variable is binary
2. Independence	Observations are independent of each other
3. Multicollinearity	Independent variables are not highly correlated with each other
4. Linearity with log odds	Independent variables are linearly related to the log odds
5. Sample size	Large sample size

Maximum Likelihood Estimation

In logistic regression, the regression coefficients, **β**, are not estimated using the OLS method described previously but by maximum likelihood estimation (MLE). While OLS is a distance-minimizing approximation method, MLE is a likelihood maximization method. The basic principle of MLE is to choose estimates of the parameter values for which the probability of the observed data is greatest. MLE achieves this through two steps: 1) constructing a likelihood equation, and 2) finding the values of the parameters that maximize that equation.

Using MLE for logistic regression, we must estimate k + 1 unknown parameters. Although this method has been derived and explained in many sources, we reference Czepiel (2002) for the derivation demonstrated in this paper.

The estimators are derived from the probability distribution of the dependent variable. Since we are interested in the total number of successes (i.e., admission into the upper division nursing program) in a sequence of Bernoulli trials, **Y** follows a binomial distribution with the probability mass function:

$$f(y_i) = \frac{n_i!}{y_i!\,(n_i - y_i)} {\pi_i}^{y_i}(1 - \pi_i)^{n_i - y_i}$$

The likelihood function, *L*, has the same form as the joint probability function and is thereby constructed by repeatedly multiplying probability functions for every i^{th} observation*:*

$$L = \prod_{i=1}^{N} \frac{n_i!}{y_i!\,(n_i - y_i)} {\pi_i}^{y_i}(1 - \pi_i)^{n_i - y_i} \tag{3.23}$$

We find the maximum of equation (3.23) by taking the first derivative with respect to **β** and setting it equal to zero. We note that the factorial term does not contain any π terms and thus acts as a constant and can be ignored. Leaving out the factorial term and rearranging the other terms algebraically, equation (3.23) becomes:

$$L = \prod_{i=1}^{N} \left(\frac{\pi_i}{1 - \pi_i}\right)^{yi} (1 - \pi_i)^{n_i} \tag{3.24}$$

To help simplify equation (3.24), we recall the relationship shown in equation (3.17) and take the exponential of both sides of equation (3.20)

$$\left(\frac{\pi_i}{1 - \pi_i}\right) = e^{\boldsymbol{\beta x_i}} \tag{3.25}$$

When we substitute equation (3.25) along with equation (3.22) into the likelihood equation we obtain

$$L = \prod_{i=1}^{N} \left(e^{\boldsymbol{\beta x_i}}\right)^{yi} \left[1 - \left(\frac{e^{\boldsymbol{\beta x_i}}}{1 + e^{\boldsymbol{\beta x_i}}}\right)\right]^{n_i} \tag{3.26}$$

Rewriting the 1 from equation (3.26) as $\frac{1+ e^{\boldsymbol{\beta x_i}}}{1+ e^{\boldsymbol{\beta x_i}}}$, we may simplify the equation further

$$L = \prod_{i=1}^{N} \left(e^{\boldsymbol{\beta x_i}}\right)^{yi} \left(\left(\frac{1 + e^{\boldsymbol{\beta x_i}}}{1 + e^{\boldsymbol{\beta x}}}\right) - \left(\frac{e^{\boldsymbol{\beta x_i}}}{1 + e^{\boldsymbol{\beta x_i}}}\right)\right)^{n_i}$$

$$L = \prod_{i=1}^{N} \left(e^{\boldsymbol{\beta x_i}}\right)^{yi} \left(\frac{1}{1 + e^{\boldsymbol{\beta x_i}}}\right)^{n_i}$$

$$L = \prod_{i=1}^{N} \left(e^{\boldsymbol{\beta x_i}}\right)^{yi} \left(1 + e^{\boldsymbol{\beta x_i}}\right)^{-n_i} \tag{3.27}$$

Equation (3.27) is the likelihood function to be maximized by taking partial derivatives with respect to **β**. These calculations can be quite complex so in practice it is common to take the derivative of the natural log of equation (3.14), or the "log-likelihood" function, instead. This is perfectly acceptable because the log function is a monotonically increasing function, meaning it will always increase as *x* increases. Thus, the parameters that maximize the likelihood function will also maximize its log with the added benefit that the calculations are greatly simplified. Taking the log of equation (3.27) gives us the log-likelihood function:

$$log\ L\ = \sum_{i=1}^{N} y_i\ (\boldsymbol{x_i \beta}) - n_i\ \cdot log(1 + \ e^{\boldsymbol{\beta x_i}})$$

We take the derivative of this function with respect to **β**

$$\frac{\partial logL}{\partial \beta} = \sum_{i=1}^{N} y_i \boldsymbol{x_i} - n_i\ \cdot \frac{1}{1 + e^{\boldsymbol{\beta x_i}}} \cdot \frac{\partial}{\partial \beta}(1 + e^{\boldsymbol{\beta x_i}})$$

$$= \sum_{i=1}^{N} y_i \boldsymbol{x_i} - n_i \cdot \frac{1}{1 + e^{\boldsymbol{\beta x_i}}} \cdot e^{\boldsymbol{\beta x_i}} \cdot \frac{\partial}{\partial \beta} \boldsymbol{\beta x_i}$$

$$= \sum_{i=1}^{N} y_i \boldsymbol{x_i} - n_i \cdot \frac{1}{1 + e^{\boldsymbol{\beta x_i}}} \cdot e^{\boldsymbol{\beta x_i}} \cdot \boldsymbol{x_i}$$

Using equation (3.22) to simplify, we obtain

$$\frac{\partial logL}{\partial \beta} = \sum_{i=1}^{N} y_i \boldsymbol{x_i} - n_i \pi_i \boldsymbol{x_i} \tag{3.28}$$

The maximum likelihood estimates for **β** can be found by setting each of the k + 1 equations in equation (3.28) to zero and solving for each β_k.

Equation (3.28) results in a system of k + 1 nonlinear equations that cannot be solved algebraically. The solutions must be successively approximated using an iterative method. In this study we rely on SAS software, specifically the LOGISITIC procedure, to compute the approximations. PROC LOGISITIC in SAS uses the Newton-Raphson algorithm to calculate the approximations. We refer the reader to Allison & SAS Institute (2003) and Czepiel (2002) for a more thorough discussion on the Newton-Raphson algorithm.

Summary

We have now developed the mathematical framework for the analysis used in this study. We identified equation (3.5) as the equation that will estimate the total number of upper division admits that we are interested in as well as derived the principles behind the logistic regression model. In the following chapters, we will focus on the actual building of the model with TWU nurse-entry data. We will detail the process of preparing the

data, selecting the most meaningful independent variables to be included, and assessing the degree to which the model fits the data. Additionally, we will provide interpretations for the MLE coefficients derived above.

CHAPTER IV

DATA PREPARATION AND MODEL BUILDING

Introduction

The intent of this study is to build a predictive model that, after one semester, estimates the total number of nurse-entry students who will be admitted into the upper division nursing program two to three years later. Accurately projecting this total after only one semester of academic progress could create an early-alert intervention opportunity for at-risk students and help TWU direct student support services more efficiently.

Before we begin to build this model, we must first understand the TWU nurse-entry population by examining our data and identifying key characteristics and general patterns within it. This practice is called exploratory analysis and allows us to understand and summarize our dataset in order to prepare for advanced modeling. In this study, extensive exploratory analysis was conducted to gain a deep familiarity with the dataset and aid in initial variable selection and creation. Only then can we follow a step-by-step process to systematically build our predictive model. Once we are satisfied with the final version of this model, we assess the degree to which it fits the data as well as test its predictive validity. When it comes to the model fit, the highest standard to achieve is predictive validity. A well-fitting model results in predicted values that are close to the

observed values even on data that was not used to estimate its regression coefficients. Such a model will thus output meaningful values for prediction purposes.

Data Description

All data in this study was provided through the TWU Office of Institutional Research and Data Management (IRDM) and is historical, secondary data that does not contain fields with identifiable information. Furthermore, only general patterns are analyzed and summary reports published in order to avoid any loss of confidentiality in the study. The data consists of 8 years of institutional data for all full-time, first time in college (FTIC) nurse-entry students who entered TWU between (and including) Fall 2008 and Fall 2015. TWU Honors students and "early-admit" students who were granted guaranteed admission into the upper division nursing program at the outset of their college career were not considered in the study. There were 2,975 total observations in the sample. About 40% of these students were first-generation in college and around 58% received a Pell grant (a federal grant awarded to students who exhibit exceptional financial need) during their first semester. The demographic characteristics of the sample are described in Table 3.

Table 3
Demographic Characteristics of TWU Nurse-entry Data

Variable	Description	Percentage
Gender	Female	95%
	Male	5%
Ethnicity	White	22%
	Black	30%
	Hispanic	34%
	Other	14%

Academic performance varied widely among the 2,975 FTIC nurse-entry students. Figure 2 displays a scatterplot of composite SAT scores by first-semester GPA for all records in the dataset.

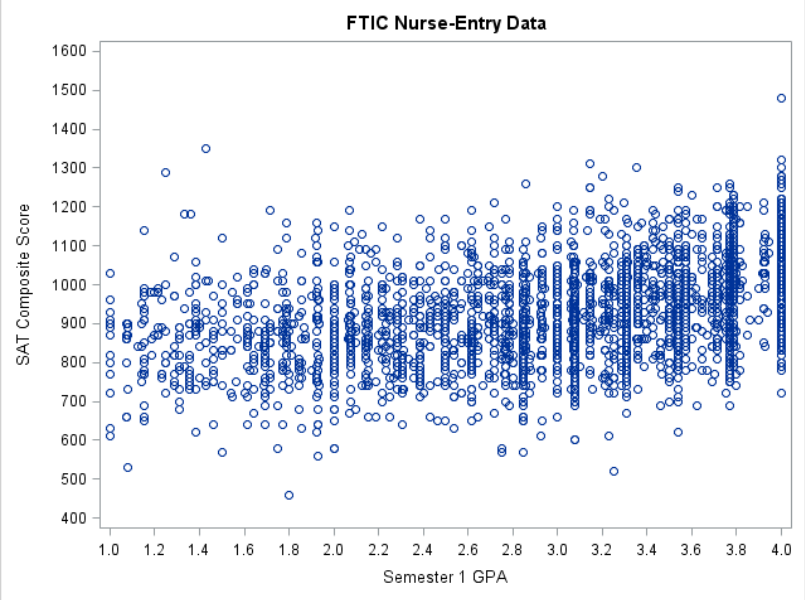


Figure 2. Scatterplot of SAT scores versus first-semester GPA for FTIC nurse-entry students

Data Preparation

We were provided two de-identified raw datasets for this study. In order to prepare the data for regression analysis, it was necessary to preprocess the data to create variables of interest and extract information pertinent to our research. The first dataset, which for simplicity we will refer to as FTIC_All, was a comprehensive list of all entering FTIC nurse-entry students between Fall 2008 and Fall 2015. This dataset consisted of one entry per student with a total of 3,180 records and 142 descriptive variables. The second dataset, which we will call Cohort_Courses, catalogued all the courses that each individual student had taken during their TWU career, with one record per course. This dataset had a total of 56,767 records and included 62 variables. Linking coursework completed at TWU from the Cohort_Courses dataset with important student characteristics found in the FTIC_All dataset is what will make this study unique to the literature. We will synthesize gateway course performance analysis into our predictive model in our attempt to gain fresh insight on the problem and improve predictive accuracy.

From the FTIC_All dataset, we were able to determine which nurse-entry students were admitted into the upper division nursing program and at what point in their academic journey they were admitted. A variable was created and named P_ADMIT to explain whether or not a student was admitted into the upper division program after persisting two or three years. Figure 3 shows the average distribution of students admitted into the upper division nursing program from Fall 2008 to Fall 2015 after persisting three full years. Figure 4 gives the composite SAT scores versus first-semester GPA of those

students who were admitted. Juxtaposing Figure 4 with Figure 2 gives us a good sense of how the upper division admits perform academically in comparison to all incoming FTIC nurse-entry students.

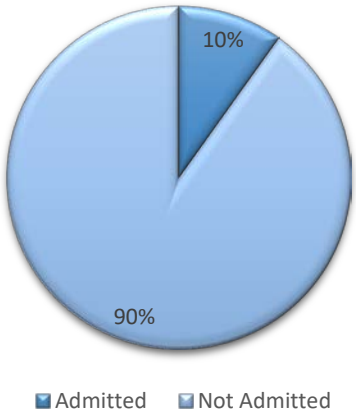


Figure 3. Average distribution of nurse-entry students from Fall 2008 to Fall 2015

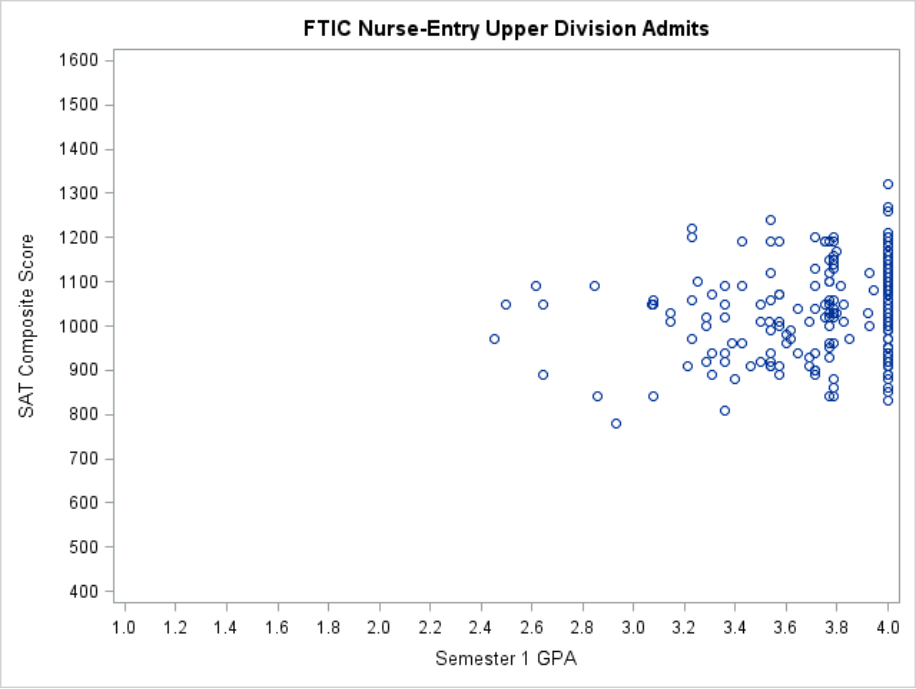


Figure 4. Scatterplot of composite SAT score versus first-semester GPA for upper division admits

We used the Cohort_Courses dataset to calculate the GPA associated with each student's first semester at TWU. A new variable, which we call GPA_S1, was created to report this value. All subsequent references to GPA in this paper refer to this variable.

The Cohort_Courses dataset also allowed us to determine which students took our specified gateway course during their first semester and identify observations that met the gateway threshold grade. Based on the findings of Stankus et al. (2018), the gateway course we evaluated for the TWU pre-nursing cohort is Introduction to General Psychology (PSY 1013). Not only has this course shown to be a discriminating factor in upper division nursing admission, but nearly 70% of the students in our sample took this class during their first semester, making it a strategic choice for early intervention efforts. The threshold grade associated with this gateway course and determined through a Bayesian statistical framework is an A (Stankus et al., 2018). Through exploratory analysis of our sample, we observed that of the students who made an A in the gateway course over 30% were admitted into the upper division nursing program two to three years later. In contrast, of the students who did not make an A in gateway only 3% were eventually admitted.

Next, we created a series of design variables, or dummy variables, that would allow us to include categorical variables such as ethnicity and gateway performance in the model. Notationally, if a categorical variable x_j has p_j possible values, then $p_j - 1$ dummy variables are needed. We denote the dummy variable as D_{jl} and the regression coefficient associated with each dummy variable as β_{jl} where l = 1,2,…, $p_j - 1$ (Hosmer &

Lemeshow, 2000) . Expanding equation (3.19) to include both continuous variables and dummy variables, we may write

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_{i1} + \cdots + \left(\sum_{l=1}^{p_j - 1} \beta_{jl} D_{jl} \right) + \beta_k x_k \,. \qquad (4.1)$$

Our ethnicity variable has four levels: “White,” “Black,” “Hispanic,” and “Other.” Thus, three dummy variables were created. Using the notation from equation (4.1), we may label these variables D_{E1}, D_{E2}, D_{E3}. If a student identifies as “White,” all the dummy variables are set equal to 0; this is called the “reference group.” If a student identifies as “Black,” $D_{E1} = 1$ and D_{E2} and D_{E3} are set to 0. This pattern continues until for “Other,” $D_{E3} = 1$ and all others are set to 0, as shown in Figure 5.

	D_{E1}	D_{E2}	D_{E3}
WH	0	0	0
BL	1	0	0
HS	0	1	0
OT	0	0	1

Figure 5. Dummy variable coding for four levels of ethnicity variable

The gateway course performance variable was coded using three levels: students who made an A in Psychology 1013 during their first semester (reference group), students who took the gateway course during their first semester but did not make an A, and students who simply did not take the gateway course during their first semester at all.

Variable Selection

Variables of interest for this study were extracted out from the 142 descriptive variables available in the FTIC_All dataset. Variables for our initial model were selected based on our literature review of commonly significant variables in the study of student retention. As Barbera et al. (2017) demonstrated, key demographic attributes, indicators of academic readiness, and financial challenges all appear to be influential predictors of student retention. Initial selection of variables included nine variables, coded as shown in Table 4.

Table 4
Initial Variables Selected for Model Building

Variable	Type	Coded Name	Description
Ethnicity	Categorical	WH	White
		BL	Black
		HS	Hispanic
		OT	Other
First-generation	Categorical	FIRST_GEN	1 = Yes; 0 = No
GPA	Quantitative	GPA_S1	$0 \leq GPA \leq 4.00$
Gateway	Categorical	A	A in gateway
		NOT A	Not A in gateway
		DID NOT TAKE	Did not take first semester
Gender	Categorical	GENDER	1 = Female; 0 = Male
First semester hours	Quantitative	HOURS_TAKEN_S1	$0 \leq Hours \leq 19$
Pell recipient	Categorical	PELL_S1	1 = Yes; 0 = No
SAT Math Score	Quantitative	SAT_M	$0 \leq Score \leq 800$
SAT Verbal Score	Quantitative	SAT_V	$0 \leq Score \leq 800$

Missing Data

As in most research, we encountered a proportion of missing data in our dataset. The impact of missing data on quantitative research can lead to biased estimates of parameters and decreased statistical power so it is an important issue to take into consideration. Through exploratory analysis of our datasets, we identified that by far the highest proportion of missing values on the variables SAT_M and SAT_V. This is because, under the Texas Ten Percent Plan (TTP), Texas high school students who graduate in the top 10% of their class are automatically admitted into any public university in the state (Daugherty, 2014). TWU extends this rule further, providing freshman applicants who are ranked in the top 25% of their class automatic admission. Consequently, many of these students do not submit standardized test scores with their application.

The default handling method for missing data in SAS is listwise deletion, in which every observation that contains a missing data field is removed from the analysis. Listwise deletion is most often criticized in the literature because it can remove a large proportion of the sample and thereby lead to decreased statistical power (Croninger & Douglas, 2005). However, Cheema (2014) writes that in cases where the sample size is large enough that adequate power is not a concern, then listwise deletion is one of the least risky and most efficient missing data handling methods. Because our study employed a sample size of $n = 2975$, we have chosen listwise deletion as our method of handling missing data. We believe the gains in accuracy offered by other methods would be trivial and would not justify the increased complexity in statistical analysis required to

implement them (Cheema, 2014). A total of 602 observations were deleted due to missing values, leaving 2,373 observations for analysis. Our model consistently captures 80% of the incoming FTIC nurse-entry students per term cohort. This percentage is high enough to draw meaningful conclusions and make recommendations for early-intervention efforts on campus.

Model Building

With a cleaned and processed dataset and nine initial variables selected, we may now begin the systematic process of model building. In building a final model for analysis, the principle of parsimony is to minimize the number of variables included while still reflecting the true outcomes of the data (Zhang, 2016). Including too many variables in a model will produce greater estimated standard errors and the model becomes increasingly dependent on the observed data (Hosmer & Lemeshow, 2000). Although SAS software offers several options of automated variable selection methods (forward, backward, and stepwise selection), we chose to use purposeful selection to build our model strategically, employing "part science, part statistical methods, and part experience and common sense" (Hosmer & Lemeshow, 2000, p. 91). To systematically choose the most meaningful variables for our model we followed the five-steps outlined below, as described by Hosmer and Lemeshow (2000) and demonstrated by Zhang (2016):

1) Univariable analysis of independent variables on outcome;
2) Multivariable analysis of covariates;
3) Model assumptions check;

4) Interactions check;

5) Model fit assessment.

Univariable Analysis

We first ran a univariable analysis on each variable to identify the association between the individual variables and the outcome. For quantitative variables, we fit a univariate logistic regression model to the variable. Significance of variables is determined using the Wald test statistic, W_k, which for each individual coefficient assigns a null hypothesis that the coefficient is equal to zero. The Wald statistic follows a standard normal distribution and is calculated

$$W_k = \frac{\hat{\beta}_k}{\widehat{SE}(\hat{\beta}_k)}$$

where $\widehat{SE}(\hat{\beta}_k)$ represents the standard error associated with each $\hat{\beta}_k$. Variables whose univariable test resulted in a p-value < 0.25 were used as candidates for the multivariable model, as advised by Hosmer and Lemeshow (2000). The resulting $\hat{\beta}_k$, $\widehat{SE}(\hat{\beta}_k)$, and p-values from the univariate logistic regression models for each quantitative variable are shown in Table 5. The p-value for the Hours_taken_S1 variable is not statistically significant, so we dropped this variable from further analysis.

Table 5
Results from Univariate Logistic Regression Models for Quantitative Variables

Variable	Coefficient Estimate	Standard Error	P-value
Hours_taken_S1	0.0620	0.0620	0.3694
GPA_S1	3.5578	0.4005	<.0001
SAT_V	0.0121	0.00175	<.0001
SAT_M	0.0118	0.00173	<.0001

For categorical variables, we created contingency tables of outcomes versus the k_j levels of the independent variable. The likelihood ratio chi-square test with $k_j - 1$ degrees of freedom is equal to the value of the likelihood ratio test for the significance of the coefficients for the $k_j - 1$ dummy variables in a univariable logistic regression model that contains that single independent variable (Hosmer & Lemeshow, 2000). Thus, we used the likelihood ratio chi-square test with $k_j - 1$ degrees of freedom to test for significance of categorical variables. The results of the likelihood ratio chi-square test for each separate categorical variable are shown in Table 6. Based on their large p-values, we dropped Gender and First_Gen from the model.

Table 6
Results from Likelihood Ratio Chi-Square Tests for Categorical Variables

Variable	Likelihood Ratio Chi-Square	P-value
Gender	0.7856	0.3754
Gateway	148.8320	<.0001
First_Gen	0.0010	0.9748
Pell_S1	19.3567	<.0001
Ethnicity	32.1708	<.0001

Multivariable Analysis

The variables that show statistical significance in the univariable analysis are GPA_S1, SAT_M, SAT_V, GATWAY, PELL_S1, and Ethnicity. We then ran a multivariable logistic regression model that includes only these variables. Again, we examined the Wald test statistic to test for significance of the coefficients. Literature suggests that using a traditional p-value of 0.05 here will often exclude important variables from the analysis (Hosmer & Lemeshow, 2000; Mickey & Greenland, 1989); thus, we held a more liberal p-value cutoff of 0.25. If any one of the $k_j - 1$ dummy variables for an independent variable, x_j, showed significance, we kept all the dummy variables in the model as they each represent one part of the whole x_j.

Using the same procedures outlined in the univariable analysis, we then fit an initial multivariate logistic regression model. The SAS output table for this step is shown in Appendix A. All three Ethnicity dummy variables show significance as do GPA_S1, SAT_V, and the Gateway variable. Pell_S1 has an associated p-value of 0.4276 and was thus dropped from further analysis.

Beyond simple statistical significance, however, there is also room for interpretation by the researcher in the variable selection decision. If we believe a variable is important, we can keep it in the model even if it does not show traditional statistical significance. Likewise, we can remove variables from the analysis that show statistical significance if we have justification from literature and experience.

For example, as mentioned above all three ethnicity dummy variables show statistical significance. However, from exploratory analysis of the data we know that

ethnicity is highly associated with the receipt of a Pell grant, which in turn is an indicator of socioeconomic status (SES). This indicated that including ethnicity as a variable in a model which predicts upper division nursing admission could entangle us in the pitfall warned against by Palmer et al. (2016). Namely, that by using race as a variable in the predictive model we might inadvertently "only pinpoint students who are traditionally "at-risk": underserved populations" (Palmer et al., 2016, p. 14). This could prompt misleading and discriminatory model interpretations as well as recommendations that reinforce institutional bias. Thus, we chose to remove the ethnicity variable from our model even though it showed statistical significance in the multivariable analysis.

On the other hand, the p-value for SAT_M did not show significance at 0.3119. However, SAT math scores were found repeatedly to be important in the literature. Conversing with subject matter experts at the university, we decided to keep both SAT verbal and SAT math variables in the model as the TWU nursing admissions criteria weights both equally.

The model building process is cyclical, and this method of adding and deleting variables and fitting and refitting the model continued until all the variables in the model are logically and statistically significant (Zhang, 2016). After dropping Pell_S1 and Ethnicity from the model, we ran the multivariate model a second time and obtain the results shown in Table 7.

Table 7
Results of Refitting Multivariate Logistic Regression Model for Select Variables

Parameter	Level	Coefficient Estimate	Standard Error	Wald Chi-Square	P-Value
Intercept		-12.5081	1.9542	40.9672	<.0001
GPA_S1		2.1952	0.4776	21.1240	<.0001
SAT_M		0.00237	0.00214	1.2285	0.2677
SAT_V		0.00461	0.00215	4.6143	0.0317
GATEWAY	Did not take	-0.9095	0.3200	8.0758	0.0045
GATEWAY	Not A	-1.7072	0.5515	9.5844	0.0020

Table 7 represents our preliminary final model in which all the variables are either statistically or practically significant. Filling in equation (4.1) with regression coefficients from Table 7, we may express the preliminary final model mathematically as

$$\begin{aligned}\text{logit}(\pi_i) = \ & -12.5081 + 2.1952\left(\text{GPA}_{\text{S1}_i}\right) \\ & +0.00237\left(\text{SAT}_{\text{M}_i}\right) + 0.00461\left(\text{SAT}_{\text{V}_i}\right) - 2.6167(\text{Gateway}_i). \end{aligned} \quad (4.2)$$

Checking Model Assumptions

Once we were reasonably sure that equation (4.2) was close to the final model, it was important to check that it met model assumptions. Logistic regression assumes that the independent variables have a linear relationship with the logit of the outcome and that none of the independent variables are highly correlated with one another. We can check the linearity assumption for all continuous variables by plotting variable values to the

logit of the outcome values on a scatter plot. Figure 6 shows that the linearity assumption was met because the variables GPA_S1, SAT_M, and SAT_V all have a linear association to the logit of the outcome.

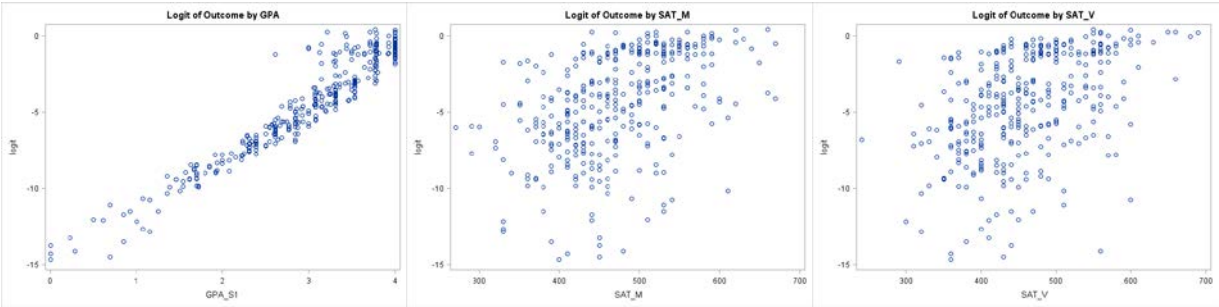


Figure 6. Scatterplots of GPA_S1, SAT_M, and SAT_V against logit of the outcome

We checked the multicollinearity assumption by estimating the equivalent linear model using PROC REG (Allison & Sas Institute, 2003). Collinearity output results are shown in Appendix B. Low tolerances and high variance inflation (VIF) are associated with high multicollinearity. Literature suggests that a common rule of thumb for multicollinearity concern occurs when tolerance < 0.10 and VIF > 10 (Williams, 2015). All tolerance values are above 0.30 and all VIF values are less than 3 for our model, we are satisfied that the multicollinearity assumption was met.

Interaction Terms

Finally, we checked our model for possible interactions. When interaction is present in a model it indicates that the effect of one independent variable on the outcome is different at different values of another independent variable. In such cases we must add a new variable to the model, an interaction term, to account for the effect. Mathematically, if an interaction exists between independent variables x_a and x_b, the

interaction term is represented as the product of the two variables $(x_a \times x_b)$ and will have a unique coefficient associated with it.

First, we examined the face validity of the independent variables to identify possible pairs that may have a basis for interaction with each other. We noted that the Gateway variable, including information on whether a student made an A in Psychology 1013, and GPA_S1, describing first-semester GPA, were logically related. When we added the interaction term to the preliminary final model and ran the multivariate analysis again, we obtained a p-value of 0.0017 for the significance of its coefficient.

To explore the effect of the interaction term further, we can conduct a partial likelihood ratio test (LRT) to compare the model that includes the interaction term to the preliminary final model from equation (4.2). An LRT assesses the goodness-of-fit between two models (one of which contains fewer parameters than the other) based on the ratio of their likelihoods, L_1 and L_2. SAS automatically outputs the value of $[-2\log(L)]$ for any model run in PROC LOGISTIC. With this information, an LRT can be quickly and easily calculated using the following formula:

$$\text{LRT} = |-2[\log(L_1) - \log(L_2)]|$$

The test follows a chi-square distribution with degrees of freedom *k*, the difference in number of independent variables between the two models. If the p-value was greater than 0.05, we assumed that there is no significant difference between the fit of the two models and we moved forward with the model containing the least number of independent variables.

When we conducted an LRT to compare the models with and without the interaction term, we obtained a p-value of 0.001. We have statistically significant evidence that there was a difference between the fit of the two models. Based on these results, we chose to add the interaction term between Gateway and GPA_S1 to the model. Table 8 represents the results of a multivariate analysis, which includes the interaction term.

Table 8
Results of Final Logistic Regression Model for TWU Nurse-entry Data

Parameter	Level	Coefficient Estimate	Standard Error	Wald Chi-Square	P-value
Intercept		-5.6009	2.5776	4.7217	0.0298
GPA_S1		0.3241	0.6759	0.2300	0.6315
SAT_M		0.00283	0.00214	1.7430	0.1868
SAT_V		0.00445	0.00212	4.4140	0.0356
GATEWAY	Did not take	-13.7271	4.3722	9.8572	0.0017
GATEWAY	Not A	-11.6961	5.5628	4.4208	0.0355
GPA_S1*GATEWAY	Did not take	3.4220	1.1607	8.6917	0.0032
GPA_S1*GATEWAY	Not A	2.7476	1.6060	2.9270	0.0871

Table 8 represents the final version of the TWU nurse-entry model. Filling in equation (4.1) with the final regression coefficients from Table 8, we may express the final model mathematically as

$$\text{logit}(\pi_i) = -5.6009 + 0.3241\left(\text{GPA}_{S1_i}\right) + 0.00283(\text{SAT\_M}_i) + 0.00445(\text{SAT\_V}_i) - 25.4232(\text{Gateway}_i) + 6.1696(\text{GPA\_S1}_i \times \text{Gateway}_i). \quad (4.3)$$

Assessing Model Fit

Finally, before using the model from equation (4.3) to draw conclusions or predict future outcomes, we must be confident in the degree to which the model fits the data. We employed two methods to check the fit of the final model. The first is the Hosmer-Lemeshow test, which assesses the calibration of the model. In other words, we tested if the probabilities generated by the model reflected the true outcome observed in the data.

The HL test statistics was calculated by predicting probabilities for all observations based on a specified model. The probabilities were then grouped into intervals and within each interval the predicted probabilities are summed to produce an expected frequency. Expected frequencies were compared to observed frequencies using a chi-square statistic with degrees of freedom equaling the number of intervals minus 2 (Allison). A p-value < 0.05 indicates a significant difference in the expected and observed frequencies. Table 9 shows the results of the HL test for the final model. The very large p-value indicates there is no evidence of a significant difference between what was expected and what was observed. That is, the model seems to fit the data very well.

Table 9
Hosmer and Lemeshow Goodness-of-Fit Test for Final Model

Chi-Square	DF	P-value
3.4326	8	0.8517

We also examined the area under the receiver operating characteristic (ROC) curve to understand the discrimination power of the model. The area under the ROC curve shows the model's ability to discriminate between those subjects who experienced

the outcome of interest (admission into upper division nursing) and those who did not. On the y-axis, we plotted sensitivity, the ratio of true positives to all favorable outcomes and on the x-axis we plotted (1-specificity), the ratio of false positives to all favorable outcomes. For a more extensive discussion on the specifics of ROC curves, we refer the reader to Ahluwalia (2006). Table 10 shows the general rules for interpreting the area under the ROC curve and Figure 7 shows the ROC curve produced by the final model. As we can see, the model exhibits outstanding discrimination.

Table 10
General Rules for Interpreting Area ROC Curve

Area under ROC	Interpretation
ROC = 0.5	No discrimination
$0.7 \leq ROC < 0.8$	Acceptable discrimination
$0.8 \leq ROC < 0.9$	Excellent discrimination
$ROC \geq 0.9$	Outstanding discrimination

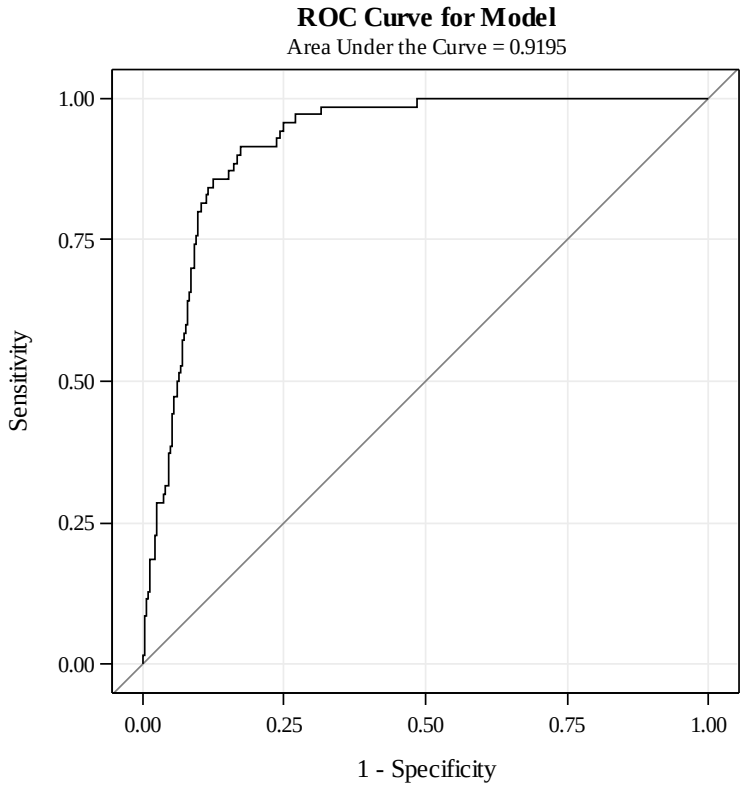


Figure 7. Receiver operating characteristic (ROC) curve for final model

Summary

Extensive exploratory analysis was conducted to gain familiarity with the raw data and understand the general patterns occurring within it. From there, we selected and coded nine initial variables, five categorical and four quantitative, to be used in the analysis. Listwise deletion was chosen as the method for dealing with missing data in our dataset.

In building equation (4.3), we first took into consideration the unadjusted association between each independent variable and the outcome. Next, we fit a multivariable model to the data and selected variables that were statistically and/or logically meaningful. We verified that the logistic regression model assumptions were

met and identified a significant interaction effect between two of the independent variables, Gateway and GPA_S1. After including the new interaction term, we assessed model fit for two different criteria, calibration and discrimination, and determined that our model fit the sample data exceedingly well on both accounts.

We were then ready to perform regression analysis and obtain the associated output in SAS. All the SAS code that was written and executed in this research can be found in Appendix C. In the following chapters we discuss the interpretation of the output, evaluate the predictive accuracy of the model, and conclude by addressing the answers to our initial research questions.

CHAPTER V

INTERPRETATION AND PREDICTIVE ACCURACY

Introduction

The interpretation of the regression coefficients from equation (4.3) provides us with general insights into how each independent variable affects the outcome. We also want to quantify the predictive accuracy of the model to give us an indication of its validity, particularly the extent of its predictive validity. Equation (4.3) was built systematically using well-documented practices found in literature; however, its ultimate purpose is to be used on real-world data so we must be able to trust that it can correctly identify at-risk students if we plan to implement early-intervention programs. We test the model's predictive accuracy by executing a specially designed cross-validation technique and iteratively running the model on data that was not used to estimate the coefficients. In this way, we simulate a real-time application of the model. We then compare predicted values to observed values to measure the achieved accuracy.

To facilitate the interpretation of our model, we give a brief overview of how to interpret regression coefficients for different types of independent variables and then move on to the actual interpretation for the variables we selected in Chapter Four. We end the chapter by measuring and reporting the predictive utility of the model based on a 3-fold cross-validation method specifically designed for real-world prediction application.

Odds Ratios

In the linear regression model described in Chapter Three, we may interpret the coefficients from equation (3.16) as the estimated change in the continuous dependent variable for every 1 unit increase in x_k, holding all other variables constant. In logistic regression, however, we used the logit transformation in equation (3.19) to set the linear combination of independent variables equal to the log odds of our outcome. This makes the interpretation of the regression coefficients slightly less intuitive than that of the linear model.

The mechanism used for interpreting the regression coefficients, $\boldsymbol{\beta}$, from equation (3.19) is called an odds ratio (OR). Odds ratios are, just as they sound, the ratio of two odds and are a simple way to present the strength of association between an independent variable and the outcome in logistic regression. ORs play an important role in the logistic regression model because they hold a special relationship with the logit parameters. To demonstrate this relationship, we show mathematically how the odds of success would change with a 1 unit increase in the independent variable x_j. Using equation (3.19) and (3.21), we let $odds_{x_j+1}$ represent the odds of success when x_j is increased by 1 unit

$$\frac{odds_{x_j+1}}{odds} = \frac{e^{\beta_0+\beta_1x_1+\cdots+\beta_j(x_j+1)+\cdots+\beta_kx_k}}{e^{\beta_0+\beta_1x_1+\cdots+\beta_jx_j+\cdots+\beta_kx_k}}.$$

Applying the exponent rule

$$\frac{e^a}{e^b} = e^{(a-b)}$$

and simplifying the expression we obtain

$$\frac{odds_{x_j+1}}{odds} = e^{(\beta_j(x_j+1)-\beta_j x_j)} = e^{\beta_j}$$

or,

$$OR_{x_j} = e^{\beta_j}. \tag{5.1}$$

Thus, for the continuous variable, x_j, equation (5.1) means that a change of 1 unit in x_j changes the estimated odds of success by a factor of e^{β_j}. We may also say that a 1 unit change in x_j changes the log odds ratio by the value of β_j.

Like odds, $0 < \text{ORs} < \infty$. For interpretation purposes, we are interested in an ORs relation to the number 1. Equation (5.1) shows that an OR = 1 corresponds to a regression coefficient of 0. This means that if an OR = 1, there is no change in the odds of success due to an increase in x_j and we may say that the change in the odds of success for that ratio is not significant. In general, if the OR < 1, odds of success are decreased for a given outcome, or in our case, for admission into upper division nursing. If OR > 1 the odds of success are increased for the outcome.

The 100 – α% confidence interval (CI) estimate for the odds ratio is calculated by exponentiating the two endpoints of the confidence interval for β_k such that

$$CI_{OR} = exp\left[\hat{\beta}_k \pm z_{\frac{\alpha}{2}} \times \widehat{SE}\left(\hat{\beta}_k\right)\right] \tag{5.2}$$

where $z_{\frac{\alpha}{2}}$ represents the z-score associated with 100-α% confidence and $\widehat{SE}\left(\hat{\beta}_k\right)$ is the standard error associated with each coefficient estimate.

Often, however, a 1 unit change in an independent variable is not conducive to a meaningful interpretation. For many continuous variables, we were usually more

interested in the change in odds for a more general *c* unit change. For a change in *c* units, the odds ratios can be calculated

$$OR = e^{c\beta_k} = \left(e^{\beta_k}\right)^c \tag{5.3}$$

and the endpoints of a 100 – α% (CI) estimate of OR(c) are found by

$$CI_{OR(c)} = exp\left[c\hat{\beta}_k \pm z_{\frac{\alpha}{2}} c \times \widehat{SE}\left(\hat{\beta}_k\right)\right]. \tag{5.4}$$

For a categorical independent variable, x_k, with two levels ($x_k = 1$; $x_k = 0$) an odds ratio represents the ratio of the odds of success when $x_k = 1$ to the odds of success when $x_k = 0$, as shown below

$$OR = \frac{odds_{x_k=1}}{odds_{x_k=0}}\ . \tag{5.5}$$

When a categorical variable has more than two levels, the numerator of equation (5.2) represents the odds of success for a given level of the independent variable and the denominator corresponds to the odds of success for the reference group specified in the design variable coding detailed in Chapter Four.

In multivariate models, such as the one used in this paper, the interpretations described above are based upon adjusted ORs. This means that we are comparing observations that differ only in the independent variable of interest and have the values of all other variables held constant. The adjustment is statistical, estimating what might be expected to be observed if the observations differed only on that independent variable (Hosmer & Lemeshow, 2000).

Interpretation of Individual Coefficients

The estimates of the ORs and associated confidence intervals for SAT_M and SAT_V obtained by using equation (5.1) and (5.2), are shown in Table 11.

Table 11
Odds Ratio Estimates and Confidence Intervals for SAT Variables

Effect	Odds Ratio Estimate	95% Confidence Limits	
SAT_M	1.003	0.999	1.007
SAT_V	1.004	1.000	1.009

Recall that since SAT_M and SAT_V are continuous variables, exponentiating the regression coefficients from Table 11 will give the odds change for a 1 unit increase in SAT score. As mentioned in the discussion above, a 1 unit change is not meaningful for every independent variable. In the context of SAT scores, we will need to choose a more appropriate value for c by which to measure the change in odds of success for these variables. Official SAT statistics indicate that the average improvement of SAT scores for test-takers is 60 to 70 points (Magana, 2019). Using 70 units as a reference point, then, we let $c = 70$ and use equations (5.4) and (5.5) to populate Table 12 with the OR estimates and associated confidence intervals for a 70 point increase in SAT scores.

Table 12
Odds Ratio Estimates and Confidence Intervals for a 70 Unit Change

Effect	Odds Ratio Estimate	95% Confidence Limits	
SAT_M	1.233	0.932	1.630
SAT_V	1.322	1.000	1.872

The estimated OR for SAT_M is 1.233 with a 95% confidence interval of (0.932, 1.630). Because 1.00 is include in the CI, the difference in odds of admission into upper division nursing for an increase of 70 points in SAT math is not significant. For SAT_V, however, the estimated OR is 1.322 with a 95% CI of (1.000, 1.872). This means that the odds of admission into upper division nursing is estimated to be 1.322 times larger with every increase of 70 points in SAT verbal score.

Interpretation of Interaction

We now turn to interpreting the results for GPA_S1 and Gateway, which interact in the model. As mentioned in Chapter Four, in the case of an interaction the OR of one variable involved in the interaction depends upon the value of the other. Interaction terms are more complicated to interpret in a model; but, they can greatly expand our understanding of the relationships among the variables. We now detail mathematically the method for determining estimated ORs in the presence of interaction, using similar notation derived by Hosmer & Lemeshow (2000).

Let C denote a student characteristic, X a covariate in the model, and C × X their interaction. Observing realized values from the sample data where $c \in C$ and $x \in X$, we model the logit as

$$g(c,x) = \beta_0 + \beta_1 c + \beta_2 x + \beta_3 c \times x .$$

Suppose we want to compare the ORs for two levels of C, C = c_0 and C = c_1 when X = x. To do this, we must write down the expressions for the logit at the two levels of C

$$g(c_0,x) = \beta_0 + \beta_1 c_0 + \beta_2 x + \beta_3 c_0 \times x$$

and

$$g(c_1,x) = \beta_0 + \beta_1 c_1 + \beta_2 x + \beta_3 c_1 \times x .$$

Next, we algebraically simplify the difference between the two logits

$$g(c_0,x) - g(c_1,x) =$$

$$(\beta_0 + \beta_1 c_0 + \beta_2 x + \beta_3 c_0 \times x) - (\beta_0 + \beta_1 c_1 + \beta_2 x + \beta_3 c_1 \times x) =$$

$$\beta_1(c_0 - c_1) + \beta_3 x(c_0 - c_1) . \quad (5.6)$$

Finally, we exponentiate equation (5.6) to obtain the estimator

$$OR = exp[\beta_1(c_0 - c_1) + \beta_3 x(c_0 - c_1)] . \quad (5.7)$$

When C is dichotomous variable ($c_0 = 1;\ c_1 = 0$), equation (5.7) can be simplified even further

$$OR = exp[\beta_1 + \beta_3 x] . \quad (5.8)$$

Through simple coding, SAS carries out the above calculations and allows us to automatically view the results of a detailed analysis of interaction effects. For explicit instructions of the SAS procedures and statements used to obtain our output, we reference "Analyzing and Visualizing Interactions in SAS" (UCLA Institute, 2019).

We may now understand how GPA and Gateway performance are working together in the model to affect the odds of admission into upper division nursing. Table 13 shows that the effect of GPA is significant for both those who did not take gateway and those who did not make an A in gateway. In both cases, increasing GPA increases the odds of being admitted into upper division nursing. Each additional grade point increases the odds of admission into upper division by a factor of 42.4 and 21.58 for those who did not take gateway and those who did not make an A in gateway, respectively. On the other hand, the effect of GPA is not significant for those who made an A in the gateway course as the confidence interval OR contains 1.00.

Table 13
Odds Ratio Estimates and Confidence Intervals for GPA_S1 at Every Level of Gateway

Odds Ratio	Estimate	95% Confidence Limits	
GPA_S1 at GATEWAY=Did not take	42.359	6.461	277.722
GPA_S1 at GATEWAY=Not A	21.580	1.228	379.309
GPA_S1 at GATEWAY=A	1.383	0.368	5.201

We can use equation (5.4) to see that even an increase of 0.25 in GPA will more than double the odds of admission into upper division nursing both for those who did not take gateway and those who did not make an A.

We can also compare levels of Gateway at different values of GPA. When comparing OR estimates between two dummy variables of x_j, what we are actually

comparing is a ratio of odds ratios (ROR). Using the notation for dummy variables described in Chapter Four, we may express this situation mathematically

$$ROR = \frac{OR(D_{j1})}{OR(D_{j2})} = \frac{\left(odds_{D_{j1}}/odds_{ref}\right)}{\left(odds_{D_{j2}}/odds_{ref}\right)} \quad (5.9)$$

where $OR(D_{j1})$ and $OR(D_{j2})$ each have odds of success for the reference group, $odds_{ref}$, in their denominator. Algebraically, $odds_{ref}$ will cancel out leaving

$$ROR = \frac{odds_{D_{j1}}}{odds_{D_{j2}}}$$

by which we may compare the odds of dummy variable D_{j1} with the odds of dummy variable D_{j2}.

Figure 8 shows RORs for all combinations of Gateway performance levels at different values of GPA. The graph spans from 2.5 to 3.5 on the x-axis but the data follows the same trend if we were to expand it out to include all values of GPA from 0 to 4.0. As GPA decreases from 3.5, the RORs grow exponentially for the A vs. not A and A vs. Did Not Take level comparisons. This means that making an A in gateway has an increasingly large effect on the success outcome as first semester GPA decreases. For students whose GPAs fall lower than 3, making an A in becomes more and more critical to their odds of being admitted into the upper division program.

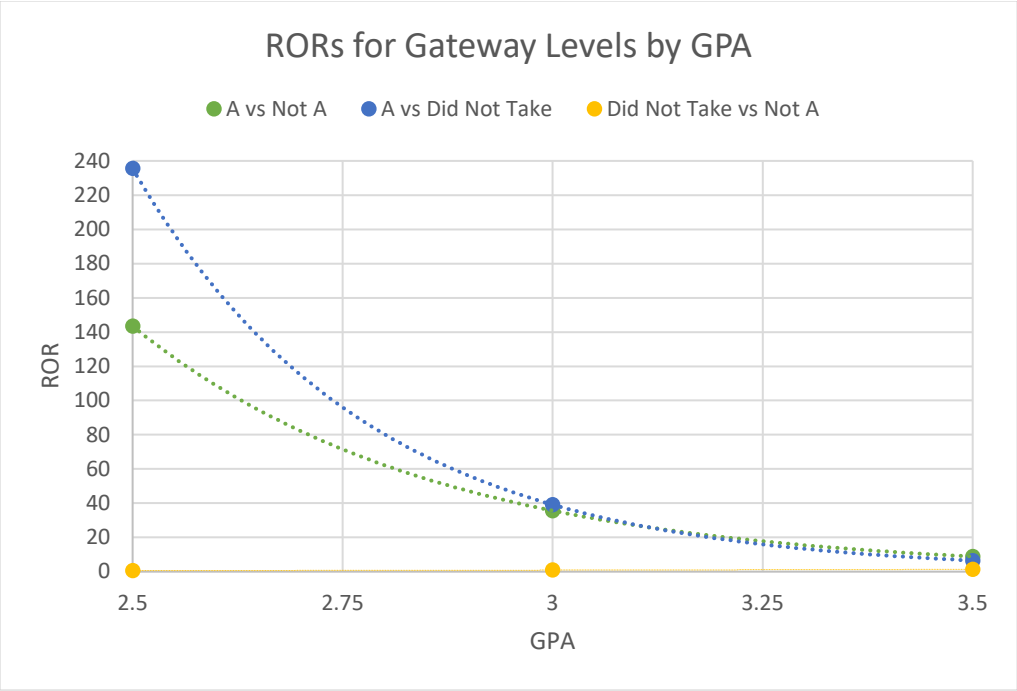


Figure 8. Ratio of odds ratios comparing gateway variable levels by GPA

Table 14 zooms in on the GPA values at 3.5 from Figure 8. Comparing the odds of success for these high GPA students, the odds of upper division admission are still 8 times greater for those who made an A vs those who did not make an A. They are nearly 6 times larger for those who made an A vs those who chose not to take gateway during their first semester. Between those who did not take gateway and those who did take it but did not meet the threshold grade of an A, the OR confidence intervals contains 1 and is thus not significant.

Table 14
ROR Estimates and Confidence Intervals for Gateway Levels at GPA_S1=3.5

Gateway Level Comparison	Estimate	95% Confidence Limits	
GATEWAY A vs NOT A	8.000	2.729	23.455
GATEWAY A vs DID NOT TAKE	5.754	2.457	13.477
GATEWAY DID NOT TAKE vs NOT A	1.39	0.425	4.544

Although ORs are informative, they can sometimes be misleading because they do not tell you the initial probability of success for an outcome. For example, we might interpret an OR of 100 by saying that the odds of success are 100 times larger for every 1 unit increase in a certain continuous independent variable. This seems to be a remarkable change, but, if the initial probability of success is very small, say only 0.00001, then using equations (3.17) and (3.18) we can calculate the new probability of success to be only 0.0009. Although the OR value was quite large, the probability of success remains extremely small. In order to understand if the increase in odds was truly meaningful, that is, if the likeliness of affecting the outcome was substantial, it is important to know if the probability of success is high to begin with. Thus, we estimated probabilities of the outcomes across various values of GPA to give the ORs discussed above more context.

The estimated probabilities shown in Figure 9 give the ORs from Table 14 more meaning. Students with a 3.0 GPA who do not make an A in gateway or did not take it first semester have less than 1% probability of being admitted in the upper division program. This probability gets smaller and smaller as GPA decreases. The probability of admission for these two groups begins to increase exponentially as GPA grows larger

than 3.0, with the probability of admission with a GPA of 3.75 being about 11% for those who did not take gateway and 6.8% for those who did not make an A. This trend highlights the fact that in a highly competitive major such as nursing, earning a minimum of a 3.0 GPA during the first semester is vital for upper division admission. Figure 4 illustrates this fact.

Yet, for every value of GPA, the probability of being admitted is greatest for those who make an A in gateway. Making an A in this course seems to be an important way to stand out to the admissions committee if a student fails to do so by earning a high GPA during their first semester. It may be that these students are able to raise their GPAs in subsequent semesters but based on the estimated probabilities, making an A in this specific course must give them an advantage over other students with low first-semester GPAs.

This demonstrates that, among nurse-entry students, not all GPAs are created equal. Not only is having a high GPA important for upper division admission but excelling in specific courses also matters. For this reason, making an A in Psychology 1013 can truly be described as a "gateway" into upper division nursing, as described by Stankus et al.(2019) and confirmed again in this analysis. This presents a powerful opportunity for academic intervention, but the timing is crucial. Students who make an A in gateway during their first semester but have a low overall GPA are great candidates for early-intervention outreach. These are the students who have the most to benefit from the help. Tutoring resources or academic counseling, if deployed at this critical first-semester

juncture, could redirect the trajectories of these at-risk students and as a result increase institutional retention rates.

Finally, we note that for every gateway level in Figure 9, the probability of admission into upper division nursing is always highest for those who achieve high GPAs in their first semester. Therefore, a positive first semester "launch" is an important factor for future success.

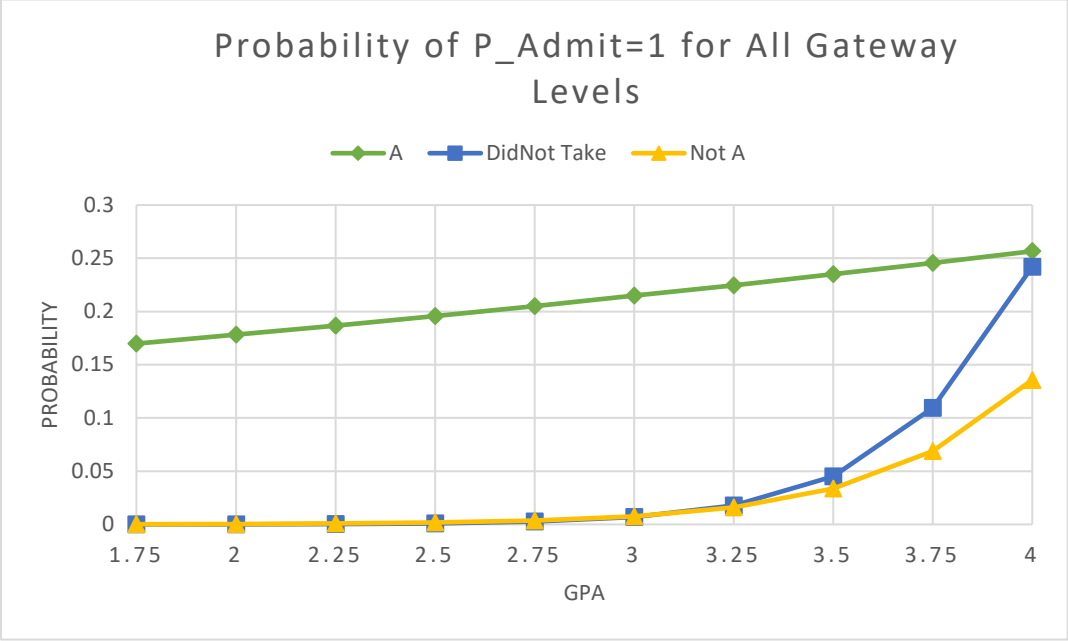


Figure 9. Probability of success for three levels of gateway variable by GPA value

Cross-Validation

Understanding the interpretation of the regression coefficients is central for explaining relationships among variables and making recommendations based on model results. Another key model component to explore is predictive accuracy. The goal of predictive modeling is ultimately to predict future outcomes and events by using a limited set of data to estimate the unknown parameters. Thus, it is important to be able to test the

model's predictive accuracy. To do this, we must run our model on data that was not used to build it; that is, on data that was not used for the estimation of the model parameters.

It is possible in the training of a model for the model to capture the noise of the sample dataset and fit the data too well. In these cases, called overfitting, the model results in good accuracy for the dataset used to create it but poor accuracy on new datasets and is thus of little practical use in the real world (Shaikh, 2018). This is one reason why we included some fundamental variables that may not necessarily show statistical significance in the model. Their face-validity, based on our findings in literature, lends depth to the model. Since being able to use this model on real-world data is of critical importance to our research goals, we implemented a cross validation (CV) technique to mitigate over-fitting.

For the TWU nurse-entry data, we used a unique cross-validation technique designed specifically for testing the model in the same conditions that it would be used for early intervention. In order to help identify at-risk students, we must structure our analysis such that we are testing the model just as if it was being applied in real-time. Because it takes three full years to know the success outcome for a nurse-entry term cohort it is not realistic for us to, for example, train on the Fall 2011 cohort and test on the Fall 2012. The upper division admission data for the Fall 2011 cohort would not be available simply because not enough time would have passed in order to know who was admitted. Thus, if you want to predict admission for the Fall 2012 cohort, the most recent usable data to train the model on would be the Fall 2009 cohort data. In general, the most

recent term cohort usable for training the model, where x is the term for prediction, would be

$$training\ term\ cohort = (x - 3)\ .$$

Furthermore, because there is such a small number of admits in each cohort (about 36 on average) we chose to train the model on three term cohorts instead of just one in order to enrich the training model and, again, avoid overfitting. Thus, if we want to predict admission into the upper division nursing program for the Fall 2012 cohort, we would train the model on Fall 2009, Fall 2008, and Fall 2007 data. Generally, the range of training years would be

$$[(x - 5), (x - 3)]\ .$$

The data in this study was split into training data sets and testing data sets, similar to a k-fold cross validation design described by Shaikh (2018). We ran the model in three iterations to test predictive accuracy. Figure 10 shows the custom cross-validation technique that we designed for this study which enables us to use the model for real-time prediction of nurse-entry cohorts.

Iteration 1	08	09	10	11	12	13	14	15
Iteration 2	08	09	10	11	12	13	14	15
Iteration 3	08	09	10	11	12	13	14	15

Train

Test

Figure 10. Cross-validation technique used for modeling real-time prediction of nurse-entry cohorts

The results detailed in this paper are based on the output obtained from the third iteration testing on the most recent term cohort data, Fall 2015. Results obtained for Fall 2014 and Fall 2013 tests can be found in Appendix D.

Predictive Accuracy

We now assess the predictive accuracy of the model to determine how well the model was able to discriminate between those who were truly admitted into the upper division program and those who were not. We use the results from the 3rd iteration of training/testing shown in Figure 10, which is the most recent data available.

Using the notation derived in Chapter Three, we let $\boldsymbol{P_T} = \{1, 2,\ldots,N_T\}$ be the set of labels of the units of a finite population size N_T nurse-entry students for the test data cohort at the beginning of their first semester. Associated with each unit $i \in \mathbf{P_T}$ is the value y_i which indicates for $i = 1,2,\ldots,N_T$ whether or not a student is admitted into the upper division program two to three years later. The total number of students in the testing data set who were admitted into upper division nursing was represented as

$$T_T = \sum_{i=1}^{N} y_i \tag{5.10}$$

and estimated by

$$\hat{T}_T = \sum_{i=1}^{N} \pi_i \,. \tag{5.11}$$

The model calculated π_i for every i^{th} observation and then classified them into groups based on these probabilities. Table 15 shows the numbers of predicted and actual admits from the 2015 cohort data.

Table 15
Nurse-Entry Upper Division Admission Prediction for Fall 2015

Enrolled 15/FA		Admitted 17/FA or 18/FA	
Probability Group	Cohort Total	Predict Admit	Actual Admit
01	235	3	7
02	23	4	7
03	35	9	8
04	34	12	12
05	12	6	5
06	7	4	2
07	1	1	0
Total	347	37	41

As shown, the model underestimated the total number of admits by only four observations. To quantify predictive accuracy, we subtracted the observed total number of students admitted from equation (5.10) to our estimated total calculated by equation (5.11; Perlow, 2014). We may express the prediction accuracy of our model as a percentage by

$$\frac{T_T - \hat{T}_T}{N_T} . \tag{5.12}$$

Using equation (5.12), prediction accuracy for the Fall 2015 data is calculated to be 0.0115. In other words, using only first-semester data, the model predicted within 1.15% the number of upper division admits from the Fall 2015 incoming students. All three model iterations predicted the number of upper division admits accurately within 10%. The Fall 2014 prediction was within 8.23% and Fall 2013 was less than 1% off from the true number of admits.

Identifying Sub-Groups

We divided the predicted probabilities from Table 16 into quartiles to obtain Table 15. Here we see that of the 174 students with lower half probability estimates, only two were admitted in to the upper division nursing program. On the other hand, among those with upper half probability estimates, 39 out of 173 were admitted. This means lower half observations had a 0.01% probability of being admitted while upper-half observations had a 22% probability of being admitted. Recall that in Chapter Four, exploratory analysis revealed that the average percent of nurse-entry students admitted into the upper division program per term cohort was 10%. We have now uncovered two sub-groups of that population who have a much lower, and much higher, probability of admittance.

Table 16
Fall 2015 Nurse-Entry Upper Division Admission Probability Prediction by Quartile

Predicted Probability Quartiles	Admitted		
	No	Yes	Total
0 - 25%	86	1	87
26 - 50%	86	1	87
51 - 75%	75	12	87
76 - 100%	59	27	86
Total	306	41	347

Comparing Models

In Chapter One, we hypothesized that adding gateway course performance alongside traditionally used predictor variables (SAT scores and GPA) would enhance its predictive accuracy. To explore this theory, we constructed a "GPA/SAT" model that contained only these two academic metrics as the independent variables.

A comparison of predictive accuracy between our full, final model and the GPA/SAT model on all three iterations of cross-validation is shown in Appendix E. Adding the gateway course performance variable neither increased nor diminished the predictive accuracy of the model due to the tremendous predictive power of first-semester GPA. The full model, however, while retaining excellent predictive utility, also gives us the added benefit of useful interpretation helping us to plan early-intervention strategies for at-risk students.

Summary

Based on the interpretation of the ORs in this chapter, the greatest impact on the odds of being admitted into the upper division nursing program seems to lie within the interaction between GPA and gateway course performance. That is, GPA and Gateway are working together in the model to affect the odds of admission. First, achieving a high GPA is essential in this competitive major and students who start their first semester by earning a 3.0 or higher are setting themselves up for a higher likelihood of success. However, as we discovered through this analysis, not all GPAs are created equal. Earning an A in Psychology 1013 raises a student's probability for admission for all values of GPA, even low first semester GPAs. In fact, these students who make an A in gateway

but have an overall low GPA are precisely the ones that we have pinpointed as strategic candidates for early academic intervention. For students who have a relatively high GPA, say 3.5, they are still almost 9 times more likely to be admitted into upper division nursing if they make an A in gateway than if they do not make an A in the course.

With respect to model's predictive accuracy, using only first-semester information we were able to predict the total number of upper division admits for the Fall 2015 term cohort 98.85% accurately. Using a custom designed cross-validation technique, we replicated the test for predictive accuracy three times as a form of reliability analysis and, on average, achieved a model accuracy of 96.57%. Additionally, by dividing predicted probabilities into quartiles, we were able to identify out of the 347 students in the Fall 2015 cohort 174 students who had less than 1% chance of being admitted into upper division nursing two to three years later and 173 students whose chances of admission are 22%.

In summary, this model gives us detailed insight into the effects of ethnicity, SAT scores, GPA, and gateway course performance on the odds of upper division admission. It has also proven to predict extremely accurately which students will eventually be admitted into the program two to three years later, even identifying a specific group of particularly at-risk students.

CHAPTER VI

CONCLUSION

Summary

The new fall semester will bring hundreds of FTIC nurse-entry students onto the TWU campus. Each one of these students represents a unique combination of cultural, academic, and financial background; yet, all have chosen to invest in nursing as an educational and career path. As they set out on this journey, many may be unaware of the challenges that lie ahead. Based on the analysis performed in this study, only about 10% of them can expect to be admitted into the upper division nursing program two to three years later.

The aim of this study has been to build a predictive model that can identify, after only one semester, which students have a high probability of being admitted into the upper division program and which are most likely not to be admitted. The final model used for analysis included first-semester GPA, math and verbal SAT scores, and a gateway course performance variable. The model predicted with an average accuracy of 96.56% and successfully classified TWU nurse-entry students into two groups: 1) those with over a 20% chance of being admitted, and 2) those with less than a 1% chance of admission.

We also posed the initial research question of how including gateway course information would affect the predictive accuracy of this model. Our findings indicated

that GPA alone is an exceptional predictor of admission into the upper division nursing program. When we compared the predictive accuracy of our study model with a model that only included first-semester GPA and SAT scores, there was no difference in their abilities to predict upper division admits. However, adding the gateway course performance variable into the model gave us incredible insight into the TWU nurse-entry plight that would have been otherwise unexplored in a GPA/SAT-only model.

We now understand that making an A in Psychology 1013, our pre-determined gateway course, bears considerable impact on whether a student is admitted into the upper division program. At every level of GPA, students who make an A in this class are more likely to be admitted than those who do not. Comparing high GPA students, those who make an A in gateway still have an odd of admission that is 8 times greater than their counterparts who do not make an A. Of the total upper division admits from the Fall 2015 term cohort, 63% made an A in the gateway course during their first semester. The other 37% had an average first semester GPA of 3.51.

Recommendations

The purpose of this model is to support student success at TWU. As literature clearly demonstrates, early intervention for students at-risk academically can be the deciding factor of whether a student is retained or lost from an institutional system. For this reason, predictive analytics focuses specifically on bringing about action. Because our model so accurately forecasts future behavior, it could realistically be used as a basis for early-alert intervention among TWU nurse-entry students. Model interpretation indicated that students who earn an A in gateway but have a low first semester GPA are

an expedient choice for early academic intervention due to their high probability of success in relation to other low GPA students.

Additionally, the model output a list of 174 students who, after their first semester, have less than a 1% chance of being admitted into the upper division program. This list could be used as a springboard for institutional outreach. Whether students are simply informed of their current probability of admission, assisted in their academic efforts, or redirected to pursue another path should be the discussion of faculty, staff, and administrators at TWU who wish to see improved retention rates among this population.

Limitations and Future Research

The analysis of this study was limited to 80% of the total nurse-entry students because of missing values for SAT scores. Although this majority is more than enough to motivate early-alert intervention, future research may include an imputation of SAT score values and concordance scores for ACT values so that all incoming students could be included in the model.

Additionally, our model predicts upper division admission after a student's first semester. An extension of this study could follow student progress further and iteratively include second semester data, third semester data, and so on. We expect that the predictive accuracy of the model would continue to increase with each additional semester of data.

In closing, with retention rates in this country stagnant at 50% for decades, there is more research still to be accomplished in this field. This study did a part in addressing retention issues at our university and demonstrated that research like this can be

performed on any campus and within any field of study, regardless of the ability to employ resource-intensive survey methodology. With access to institutional data, specific knowledge about their university, and a passion for supporting student success, researchers using predictive analytics can make a real impact on undergraduate retention rates and encourage a greater sense of student satisfaction on campuses across the nation.

REFERENCES

Ahluwalia, P. (2006). *Enrollment Prediction Using Bayesian Multiple Logistic Regression* (Master's thesis). Texas Woman's University. ProQuest Dissertations & Theses Global.

Aljohani, O. (2016). A comprehensive review of the major studies and theoretical models of student retention in higher education. *Higher Education Studies, 6*(2), 1. doi:10.5539/hes.v6n2p1

Allison, P. D., & Sas Institute. (2003). *Logistic regression using the SAS system : Theory and application*. Cary, N.C.: Sas Institute.

Baepler, P., & Murdoch, C. (2010). Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching and Learning, 4*(2). doi:10.20429/ijsotl.2010.040217

Barbera, S. A., Berkshire, S. D., Boronat, C. B., & Kennedy, M. H. (2017). Review of undergraduate student retention and graduation since 2010: Patterns, predictions, and recommendations for 2020. *Journal of College Student Retention: Research, Theory & Practice*. doi:10.1177/1521025117738233

Bean, J. (1982). Conceptual models of student attrition: How theory can help the institutional researcher. *New Directions for Institutional Research, 1982*(36), 17-33. doi:10.1002/ir.37019823604

Bydžovská, H. (2016). A comparative analysis of techniques for predicting student performance. *International Conference on Educational Data Mining (EDM)* (pp. 306-311). International Educational Data Mining Society. Retrieved from https://eric.ed.gov/?id=ED592637

Caison, A. L. (2006). Analysis of institutionally specific retention research: A comparison between survey and institutional database methods. *Research in Higher Education*, *48*(4), 435–451. doi:10.1007/s11162-006-9032-5

Callahan, J., & Belcheir, M. (2015). Testing our assumptions. *Journal of College Student Retention: Research, Theory & Practice*, *19*(2), 161–175. doi:10.1177/1521025115611620

Cheema, J. R. (2014). Some general guidelines for choosing missing data handling methods in educational research. *Journal of Modern Applied Statistical Methods*, *13*(2), 53–75. doi:10.22237/jmasm/1414814520

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, *73*, 247–256. doi:10.1016/j.chb.2017.01.047

Croninger, R. G., & Douglas, K. M. (2005). Missing data and institutional research. *New Directions for Institutional Research*, *2005*(127), 33–49. doi:10.1002/ir.154

Cunningham, C. J. L., Manier, A., Anderson, A., & Sarnosky, K. (2014). Rational versus empirical prediction of nursing student success. *Journal of Professional Nursing*, *30*(6), 486–492. doi:10.1016/j.profnurs.2014.03.006

Czepiel, S. (2002). *Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation.* Retrieved from http://czep.net

Daugherty, L. (2014, April 22). *The Texas Ten Percent Plan's Impact on College Enrollment*. Education Next. Retrieved October 29, 2019, from https://www.educationnext.org/texas-ten-percent-plans-impact-college-enrollment/

Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice*, *13*(1), 17–35. doi:10.2190/cs.13.1.b

Demetriou, C., & Schmitz-Sciborski, A. (2011). Integration, motivation, strengths and optimism: Retention theories past, present, and future. In R. Hayes (Ed.), *Proceedings of the 7th National Symposium on Student Retention* (pp. 300–312). Norman, OK: University of Oklahoma.

Dreiseitl, S., & Osl, M. (2012). Testing the calibration of classification models from first principles. *AMIA Annual Symposium Proceedings*, *2012*, 164–169.

Duh, M.-S., Walker, A. M., Pagano, M., & Kronlund, K. (1998). Prediction and cross-validation of neural networks versus logistic regression: Using hepatic disorders as an example. *American Journal of Epidemiology*, *147*(4), 407–413. doi:10.1093/oxfordjournals.aje.a009464

Durkheim, E. (1951). *Suicide* (J. A. Spaulding & G. Simpson, Trans.). Glencoe, IL: The Free Press.

Eduventures. (2013). *Predictive Analytics in Higher Education: Data-Driven Decision-Making for the Student Life Cycle* (pp. 1–12). Boston, MA: IBM.

Flanders, G. R. (2015). The effect of gateway course completion on freshman college student retention. *Journal of College Student Retention: Research, Theory & Practice*, *19*(1), 2–24. doi:10.1177/1521025115611396

Glynn, J. G., Sauer, P. L., & Miller, T. E. (2011). A logistic regression model for the enhancement of student retention: The identification of at-risk freshmen. *International Business & Economics Research Journal (IBER)*, *1*(8). doi:10.19030/iber.v1i8.3970

Harackiewicz, J. M., Barron, K. E., Tauer, J. M., & Elliot, A. J. (2002). Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. *Journal of Educational Psychology*, *94*(3), 562–575. doi:10.1037/0022-0663.94.3.562

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). New York, NY: Wiley-Interscience.

Hussar, W., & Bailey, T. (2011). *Projections of Education Statistics in 2020.* Washington, D.C.: U.S. Government Printing Office: U.S. Department of Education, National Center for Education Statistics.

Leisy, R. (2019, May 8). *Linear Regression Derivation*. Medium. Retrieved August 14, 2019, from https://towardsdatascience.com/linear-regression-derivation-d362ea3884c2

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

Lotkowski, V., Robbins, S., & Noeth, R. (2004). *The Role of Academic and Non-Academic Factors in Improving College Retention*. Iowa City, IA: ACT, Inc.

Magana, D. (2019). *Test Day: A Look into SATs and ACTs*. The Viking Press. Retrieved October 8, 2019, from https://thevikingpresstvs.com/948/focus/test-day-a-look-into-sats-and-acts/

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. doi:10.1353/jge.0.0024

McNeely, J. H. (1938). *Project in Research in Universities: College Student Mortality*. Bulletin, 1937, No. 11. Office of Education, United States Department of the Interior. Retrieved from https://eric.ed.gov/?id=ED542540

Mertler, C. A., & Vannatta, R. A. (2010). *Advanced and Multivariate Statistical Methods : Practical Application and Interpretation*. Glendale, CA: Pyrczak.

Metz, G. W. (2004). Challenge and changes to Tinto's persistence theory: A historical review. *Journal of College Student Retention: Research, Theory & Practice*, *6*(2), 191–207. doi:10.2190/m2cc-r7y1-wy2q-upk5

Mickey, R., & Greenland, S. (1989). The impact of confounder selection criteria on effect estimation. *American Journal of Epidemiology*, *129*(1), 125–137. doi:10.1093/oxfordjournals.aje.a115101

Palmer, Iris, Ekowo, Manuela. (2016). *The Promise and Peril of Predictive Analytics in Higher Education: A Landscape Analysis*. New America. Retrieved June 13, 2019, from https://eric.ed.gov/?id=ED570869

Patterson, M. (2016, March). *Colleges v. Corporations: The Sad Tale of Mount St. Mary's*. America Magazine. Retrieved October 3, 2019, from https://www.americamagazine.org/content/all-things/colleges-v-corporations-sad-tale-mount-st-marys

Peng, C. J., So, T. H., Stage, F. K., & St. John, E. P. (2002). The use and interpretation of logistic regression in higher education journals: 1988 - 1999. *Research in Higher Education*, *43*(3), 259–293. doi:10.1023/a:1014858517172

Perlow, E. (2014). *Hamner Enrollment Prediction Model: Transition to College for Individuals with Access Needs* (Doctoral dissertation). Texas Woman's University. ProQuest Dissertations & Theses Global.

Raju, D., & Schumacker, R. (2015). Exploring student characteristics of retention that lead to graduation in higher education using data mining models. *Journal of College Student Retention: Research, Theory & Practice*, *16*(4), 563–591. doi:10.2190/cs.16.4.e

Scott, L. D., & Zerwic, J. (2015). Holistic review in admissions: A strategy to diversify the nursing workforce. *Nursing Outlook*, *63*(4), 488–495. doi:10.1016/j.outlook.2015.01.001

Shaikh, R. (2018, November 26). *Cross Validation Explained: Evaluating Estimator Performance*. Medium. Retrieved August 6, 2019, from https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85

Spady, W. G. (1971). Dropouts from higher education: Toward an empirical model. *Interchange, 2*(3), 38–62. doi:10.1007/bf02282469

Stankus, J.-A., Hamner, M., Stankey, M., & Mancuso, P. (2018). Successful modeling of factors related to recruitment and retention of prenursing students. *Nurse Educator*, *44*(3), 147–150. doi:10.1097/nne.0000000000000579

Stankus, J.-A., Hamner, M., Stankey, M., & Mancuso, P. (2019). A Bayesian approach to gateway course identification with implications for the prenursing curriculum. *Nurse Educator.* Advance online publication. doi:10.1097/nne.0000000000000728

Tally, S. (2009). *Signals tells students how they're doing even before the test*. Purdue. Retrieved July 12, 2019, from https://news.uns.purdue.edu/x/2009b/090827ArnoldSignals.html

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45*(1), 89–125. doi:10.2307/1170024

Tinto, V. (1993). *Leaving college: Rethinking the Causes and Cures of Student Attrition* (2nd ed.). Chicago, IL: University Of Chicago Press.

UCLA Institute for Digital Research and Education Statistical Consulting. (2019). *Analyzing and Visualizing Interactions in SAS*. Retrieved September 15, 2019, from https://stats.idre.ucla.edu/sas/seminars/analyzing-and-visualizing-interactions/

Van Barneveld, A., Arnold, K., & John, C. (2012). Analytics in higher education: Establishing a common language. *Educause Learning Initiative*. 1-11. Retrieved from https://library.educause.edu/resources/2012/1/analytics-in-higher-education-establishing-a-common-language

Van Gennep, A. (1960). *The Rites of Passage* (M. Vizedine & G. Caffee, Trans.). Chicago, IL: University of Chicago Press.

Voigt, L., & Hundrieser, J. (2008). Student success, retention, and graduation: Definitions, theories, practices, patterns, and trends. *Noel-Levitz Rentention Codifications*. Cedar Rapids, IA: Ruffalo Noel Levitz.

Williams, R. (2015). *Multicollinearity Stata Example*. University of Notre Dame. Retrieved from https://www3.nd.edu/~rwilliam/stats2/l11.pdf

Wros, P., & Noone, J. (2018). Holistic admissions in undergraduate nursing: One school's journey and lessons learned. *Journal of Professional Nursing*, *34*(3), 211–216. doi:10.1016/j.profnurs.2017.08.005

Yin, T., & Burger, C. (2003). Predictors of NCLEX-RN success of associate degree nursing graduates. *Nurse Educator*, *28*(5), 232–236. doi:10.1097/00006223-200309000-00011

Zhang, Z. (2016). Model building strategy for logistic regression: Purposeful selection. *Annals of Translational Medicine, 4*(6), 111–111. doi:10.21037/atm.2016.02.15

APPENDIX A

SAS Output for Model Building Step

Intermediate Model Building Step

Parameter	Level	Coefficient Estimate	Standard Error	Wald Chi-Square	P-value
Intercept		-12.1496	2.0038	36.7645	<.0001
PELL_S1		-0.2352	0.2965	0.6292	0.4276
GPA_S1		2.1912	0.4777	21.0418	<.0001
SAT_M		0.00217	0.00215	1.0226	0.3119
SAT_V		0.00434	0.00217	4.0095	0.0452
GATEWAY	Did not take	-0.9060	0.3201	8.0119	0.0046
GATEWAY	Not A	-1.7004	0.5514	9.5093	0.0020

APPENDIX B

SAS Collinearity Output using PROC REG

Multicollinearity Check using Estimated Equivalent Linear Model

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	-0.04193	0.06291	-0.67	0.2704	.	0
GPA_S1	0.02742	0.00837	3.28	0.0022	0.71957	1.38972
SAT_M	0.00020548	0.00011723	1.75	0.1805	0.68625	1.45719
SAT_V	0.00032228	0.00011720	2.75	0.0015	0.71029	1.40788
GATEWAY NOT A	-0.24214	0.02537	-9.54	<.0001	0.33733	2.96443
GATEWAY DID NOT TAKE	-0.21303	0.02479	-8.59	<.0001	0.37595	2.65995

APPENDIX C

SAS Code

```
/*****************************************************************
            Modeling Nurse-Entry Success by Integrating Student
              Characteristics with Gateway Course Performance

                                  May 2020

                              Emily Appicelli
                    Major Professor: Dr. Mark S. Hamner
                Department of Mathematics and Computer Science
                           Texas Woman's University
*****************************************************************/

LIBNAME Share 'X:\My Shared Folders\KME_Space\THESIS\SAS_DATA';
LIBNAME Thesis 'C:\temp\Data';
LIBNAME Data 'X:\My Personal Folder\Classes\Thesis\SAS_Data';

/*-----------------------------------------------------------------------*
      Import SAS datafiles: cohort and course info
-----------------------------------------------------------------------*/
DATA Courses;
      SET share.cohort_courses_08_16_de_ident;
      If term_cohort='16/FA' THEN DELETE;
RUN;

PROC SORT DATA=courses;
      BY id_num;
RUN;

DATA FTIC_ALL;
      SET share.ftic_nurse_entry_08_16_de_ident;
      IF FULL_TIME_COHORT = '0' THEN DELETE; /*Include only full-time
students*/
      If term_cohort='16/FA' THEN DELETE;
RUN;

/*-----------------------------------------------------------------------
 *    Create Gateway variables:
 *
 *    1) 'Gateway' -> 0 = did not take course
 *    2) 'Gateway_S1' -> 0 = did not take
 *    3) 'Gateway_S2' -> 0 = did not take
 *    4) 'Threshold_met' -> 0 = did not take make A
 *    5) 'Threshold_S1' -> 0 = did not take make A in gateway semester 1
 *    6) 'Threshold_S2' -> 0 = did not take make A in gateway semester 2
-----------------------------------------------------------------------*/
DATA Gateway;
      SET courses;
      IF find(course_name, 'PSY*1013', 1) THEN Gateway=1;
      ELSE gateway=0;
      IF gateway=1 and candy_land='S1' THEN gateway_S1=1;
      ELSE gateway_S1=0;
      IF gateway_S1=0 then do; gate_var1=0; gate_var2=0; END;
```

```
      IF gateway_S1=1 and grade NE 'A' then do; gate_var1=1; gate_var2=0;
END;
      IF gateway_S1=1 and grade='A' then do; gate_var1=0; gate_var2=1; END;

RUN;

PROC MEANS DATA=Gateway MAX nway noprint; /*collapses gateway variables into
one record per ID_num */
      CLASS ID_num;
      VAR gateway_S1 gate_var1 gate_var2;
      OUTPUT OUT=gateway_for_merge MAX=;
RUN;

DATA Gateway_merge; /* prepares gateway data to be merged */
      SET Gateway_for_merge;
      KEEP ID_num gateway_S1 gate_var1 gate_var2;
RUN;

/*---------------------------------------------------------------------------*
            Calculate GPA by semester
----------------------------------------------------------------------------*/

DATA GPA_calc;
      SET courses;
      char_credit_hours=SUBSTR(COURSE_NAME, length(COURSE_NAME),1); /*takes
      last character of 'course_name' as credit hours for the course*/
      course_credit_hours=input(char_credit_hours, 1.); /* changes credit
      hours to numeric variable*/
      IF GRADE='A' THEN NUM_GRADE=4; /*assigns numbers associated with each
      letter grade*/
      IF GRADE='B' THEN NUM_GRADE=3;
      IF GRADE='C' THEN NUM_GRADE=2;
      IF GRADE='D' THEN NUM_GRADE=1;
      IF GRADE='F' THEN NUM_GRADE=0;
      IF GRADE='W' OR GRADE='9' THEN do; /*W and 9 do not count toward gpa*/
            NUM_GRADE=0;
            course_credit_hours=0;
            end;
      IF GRADE='WF' THEN NUM_GRADE=0; /*WF does count toward gpa*/
      IF GRADE='1' THEN NUM_GRADE=4; /*1,3,5 are used for grading
      developmental courses and correspond to A,B,C respectively */
      IF GRADE='3' THEN NUM_GRADE=3;
      IF GRADE='5' THEN NUM_GRADE=2;

      GRADE_POINT=course_credit_hours*NUM_GRADE; /*grade points earned per
      course*/
RUN;

PROC MEANS DATA=GPA_calc sum maxdec=2 nway noprint; /*sums every semester's
credit hours and grade points for each ID_num*/
      CLASS ID_num term_groups candy_land;
      VAR course_credit_hours Grade_point;
      OUTPUT OUT=GPA_calc2 sum=;
RUN;
```

```
DATA GPA_by_Term;
      SET GPA_calc2;
      FORMAT gpa_by_term 4.2;
      gpa_by_term=grade_point/course_credit_hours; /*calculates GPA by term*/

      IF candy_land='S1' then hours_taken_S1=course_credit_hours; /* creates
      variable for how many hours were taken during semester 1*/
RUN;

PROC SORT data=GPA_by_Term;
 BY candy_land;
RUN;

DATA GPA_variables; /* creates variables for each semester's GPA */
      SET GPA_by_Term;
      FORMAT GPA_S1 4.2 GPA_S2 4.2 GPA_S2_5 4.2 GPA_S3
      GPA_S4 4.2 GPA_S4_5 4.2;
      IF candy_land = 'S1' THEN GPA_S1=gpa_by_term;
      IF candy_land = 'S2' THEN GPA_S2=gpa_by_term;
      IF candy_land = 'S2.5' THEN GPA_S2_5=gpa_by_term; /* summer term*/
      IF candy_land = 'S3' THEN GPA_S3=gpa_by_term;
      IF candy_land = 'S4' THEN GPA_S4=gpa_by_term;
      IF candy_land = 'S4.5' THEN GPA_S4_5=gpa_by_term; /* summer term*/
RUN;

PROC SORT data=GPA_variables;
      BY ID_NUM;
RUN;

PROC MEANS DATA=GPA_variables sum maxdec=2 nway noprint; /* collapses
semester GPAs, total credit hours, total grade points, and hours taken
semester 1 into one record per ID_num */
      CLASS ID_num;
      VAR GPA_S1 GPA_S2 GPA_S2_5 GPA_S3 GPA_S4 GPA_S4_5 course_credit_hours
      grade_point hours_taken_S1;
      OUTPUT OUT=GPA_merge sum=;
RUN;

DATA GPA_for_merge; /* prepares GPA data to be merged */
      SET GPA_merge;
      DROP _type_ _freq_;
RUN;

/*---------------------------------------------------------------------------*
            Sort and merge variables into one dataset
----------------------------------------------------------------------------*/
PROC SORT data=Gateway_merge;
      BY ID_NUM;
RUN;

PROC SORT data=GPA_for_merge;
      BY ID_NUM;
RUN;
```

```
PROC SORT data=FTIC_ALL;
      BY ID_NUM;
RUN;

DATA Variables;
      merge ftic_all(in=A)gateway_merge gpa_for_merge;
      by ID_num;
      if A;
      SAT_COMP = SAT_M + SAT_V; /*creates composite SAT variable: obtained by
      adding math and verbal scores*/
RUN;

/*---------------------------------------------------------------------------*
      Create variable for admission into upper division nursing program
----------------------------------------------------------------------------*/
DATA Admits;
      SET Variables;
      /*Upper division admits after 1 year completed (P1)*/
      IF major_campus_2='Nursing--Dallas' or major_campus_2 = 'Nursing-RN' or
      major_campus_2 =: 'Nursing -Pediatric' or major_campus_2 =: 'Nursing
      Education'
      OR sp_major_campus_2='Nursing--Dallas' or sp_major_campus_2 = 'Nursing
      RN' or sp_major_campus_2 =: 'Nursing -Pediatric' or sp_major_campus_2
      =: 'Nursing-Education'
      THEN P1_admit=1;
      ELSE P1_admit=0;

      /*Upper division admits after 2 year completed (P2)*/
      IF (major_campus_3='Nursing--Dallas' or major_campus_3 = 'Nursing-RN'
      or major_campus_3 =: 'Nursing -Pediatric' or major_campus_3 =:
      'Nursing-Education'
      OR sp_major_campus_3='Nursing--Dallas' or sp_major_campus_3 = 'Nursing
      RN' or sp_major_campus_3 =: 'Nursing -Pediatric' or sp_major_campus_3
      =: 'Nursing-Education')
      and p1_Admit=0
      THEN P2_admit=1;
      ELSE P2_admit=0;

      /*Upper division admits after 3 year completed (P3)*/
      IF (major_campus_4='Nursing--Dallas' or major_campus_4 = 'Nursing-RN'
      or major_campus_4 =: 'Nursing -Pediatric' or major_campus_4 =:
      'Nursing-Education'
      OR sp_major_campus_4='Nursing--Dallas' or sp_major_campus_4 = 'Nursing
      RN' or sp_major_campus_4 =: 'Nursing -Pediatric' or sp_major_campus_4
      =: 'Nursing-Education')
      AND (p1_Admit=0 AND p2_Admit=0)
      THEN P3_admit=1;
      ELSE P3_admit=0;

      /*Upper division admits after 4 year completed (P4)*/
      IF (major_campus_5='Nursing--Dallas' or major_campus_5 = 'Nursing-RN'
      or major_campus_5 =: 'Nursing -Pediatric' or major_campus_5 =:
      'Nursing-Education'
```

```
      OR sp_major_campus_5='Nursing--Dallas' or sp_major_campus_5 = 'Nursing
      RN' or sp_major_campus_5 =: 'Nursing -Pediatric' or sp_major_campus_5
      =: 'Nursing-Education')
      and (p1_Admit=0 AND p2_Admit=0 and p3_admit=0)
      THEN P4_admit=1;
      ELSE P4_admit=0;

      /*Upper division admits after 5 year completed (P5)*/
      IF (major_campus_6='Nursing--Dallas' or major_campus_6 = 'Nursing-RN'
      or major_campus_6 =: 'Nursing -Pediatric' or major_campus_6 =:
      'Nursing-Education'
      OR sp_major_campus_6='Nursing--Dallas' or sp_major_campus_6 = 'Nursing
      RN' or sp_major_campus_6 =: 'Nursing -Pediatric' or sp_major_campus_6
      =: 'Nursing-Education')
      and (p1_Admit=0 AND p2_Admit=0 and p3_admit=0 and p4_Admit=0)
      THEN P5_admit=1;
      ELSE P5_admit=0;

      IF P1_ADMIT=1 THEN ADMIT_TIME=1; /*Create variable recording year of
      admission*/
      IF P2_ADMIT=1 THEN ADMIT_TIME=2;
      IF P3_ADMIT=1 THEN ADMIT_TIME=3;
      IF P4_ADMIT=1 THEN ADMIT_TIME=4;
      IF P5_ADMIT=1 THEN ADMIT_TIME=5;

      IF P1_ADMIT=1 OR P2_ADMIT=1 OR P3_ADMIT=1 OR P4_ADMIT=1 OR P5_ADMIT=1
      THEN upp_div_admit=1; /*for a total of all admitted*/
      ELSE upp_div_admit=0;

      IF P2_ADMIT=1 OR P3_ADMIT=1 THEN P_ADMIT=1; /* This will be our
      dependent variable showing the number of students admitted after their
      2nd or 3rd year*/
      ELSE P_ADMIT=0;
RUN;


PROC MEANS data=Admits nmiss n;
RUN;

/*-------------------------------------------------------------------------*
            Begin structuring dataset to pass through model
-------------------------------------------------------------------------*/
Data Set_up;
      SET Admits;
      Keep ID_num term_cohort ETHNIC_COHORT GENDER_COHORT FULL_TIME_COHORT
      DEGREE_SEEK_COHORT /*keeps only variables of interest in our model*/
      ASSURED_TYPE_0 PELL_COHORT FIRST_GEN_NOCOLL_0 FIRST_GEN_NODEG_0 SAT_M
      SAT_COMP SAT_V ACT_M ACT_E ACT_R ACT_S ACT_COMP GATEWAY_S1 gate_var1
      gate_var2 GPA_S1 GPA_S2 GPA_S2_5 GPA_S3 GPA_S4 GPA_S4_5 hours_taken_S1
      P_ADMIT upp_div_admit Persist_2 SP_PERSIST_2 PERSIST_3 SP_PERSIST_3
      DROP_2 SP_DROP_2 DROP_3 SP_DROP_3 GRADUATE_2 SP_GRADUATE_2 GRADUATE_3
      SP_GRADUATE_3 P2_admit;

      IF ACT_M NE '.' AND
```

```
      ACT_E NE '.' AND
      ACT_R NE '.' AND
      ACT_S NE '.' THEN
      ACT_COMP = (ACT_M + ACT_E + ACT_R + ACT_S)/4; /*creates composite ACT
      variable: obtained by averaging four sections*/

      IF Assured_Type_0 = 'HONS' OR Assured_Type_0 = 'EANUR' THEN DELETE;
      /*Exclude those guaranteed admission*/
RUN;

/*---------------------------------------------------------------------------
 *                Create grid zone variable by combining semester 1 GPA
 *                with composite SAT score
---------------------------------------------------------------------------*/

DATA Grid_var;
      SET Set_up;
      KEEP term_cohort P_admit ID_num GPA_S1 SAT_COMP;
      /*IF SAT_COMP = '.' THEN delete; /*keeps only records that have
      submited SAT score*/
RUN;

%MACRO yhat(x1, y1, x2, y2); /*calculates yhat equations for grid zones*/
%LET b1 = %SYSEVALF((&y2 - &y1)/(&x2 - &x1)); /*Input the two points that
define each equation: (x1, y1) and (x2, y2)*/
%LET b0 = %SYSEVALF(&y1 - (&b1 * &x1));
%PUT yhat = &b0 + (&b1*x); /* yhat equation will be printed in the log*/
%MEND;

/* We input the following points for each line:
 *    - Line 5: (2.9, 400); (1.1, 1200) *
 *    - Line 4: (3.4, 400); (1.1, 1400) *
 *    - Line 3: (3.8, 400); (1.2, 1550) *
 *    - Line 2: (3.9, 550); (1.7, 1550) *
 *    - Line 1: (3.9, 750); (2.4, 1550) */

%yhat(2.9, 400, 1.1, 1200)
%yhat(3.4, 400, 1.1, 1400)
%yhat(3.8, 400, 1.2, 1550)
%yhat(3.9, 550, 1.7, 1550)
%yhat(3.9, 750, 2.4, 1550)

/* Using the equations obtained above,
 * the data set 'Functions' gives all *
 * x and y values for the lines from  *
 * x = 0 to x = 4 in 0.1 increments.  *
 * We use it to plot the lines later. */

DATA Functions;
DO x = 0 TO 4 BY 0.1;
    yhat5 = 1688.88888888888 + (-444.444444444444*x);
      yhat4 = 1878.26086956521 + (-434.782608695652*x);
      yhat3 = 2080.76923076922 + (-442.307692307692*x);
      yhat2 = 2322.72727272727 + (-454.545454545454*x);
```

```
      yhat1 = 2829.99999999999 + (-533.333333333333*x);
    OUTPUT;
    END;
RUN;

 /* Pass data through equations and assign zone to each observation.
 *
 *    Let the x variable be semester 1 GPA and the y variable be SAT score
 *    - Pass the x variable of each observation through the yhat equations
 *      to obtain expected y values for each zone
 *    - Compare expected y values with the actual y value and assign a
 *      zone category (zones 1 - 6) based on the relationship
 *    - Print new list with zones assigned to each observation
 *     (observations with missing data listed first)*/

Data Points;
      SET Grid_var;
      x_var = GPA_S1;
      y_var = SAT_COMP;
      yhat5 = round((1688.88888888888 + (-444.444444444444*x_var)), .01);
      yhat4 = round((1878.26086956521 + (-434.782608695652*x_var)), .01);
      yhat3 = round((2080.76923076922 + (-442.307692307692*x_var)), .01);
      yhat2 = round((2322.72727272727 + (-454.545454545454*x_var)), .01);
      yhat1 = round((2829.99999999999 + (-533.333333333333*x_var)), .01);
      IF y_var = '.' THEN Zone = 0;
      ELSE IF y_var GE yhat1 THEN Zone = 1;
      ELSE IF y_var GE yhat2 and y_var LT yhat1 THEN Zone = 2;
      ELSE IF y_var GE yhat3 and y_var LT yhat2 THEN Zone = 3;
      ELSE IF y_var GE yhat4 and y_var LT yhat3 THEN Zone = 4;
      ELSE IF y_var GE yhat5 and y_var LT yhat4 THEN Zone = 5;
      ELSE IF y_var LT yhat5 THEN Zone = 6;
RUN;

DATA List;
      SET Points;
      KEEP term_cohort ID_num Zone;
RUN;

/* ------------------------------------------------------------------
 *          GRID ZONE GRAPH - comment out until you want to see visual
 * ------------------------------------------------------------------
 * Sort 'List' data set by zone and   *
 * concatenate with the 'Function'    *
 * data set in order to create plot   *
 * and visualize results              *
 *-----------------------------------*/

PROC SORT DATA = Points;
      BY zone;
RUN;

DATA Plot;
      SET Functions Points;
RUN;
```

```
/* ------------- Scatterplot by GPA and SAT -----------------*/

PROC SGPLOT DATA = Plot;
      title "FTIC Nurse-Entry Enrollment, 11/FA";
      where p_admit=1 and term_cohort='11/FA';
      SCATTER x = x_var y = y_var / group = zone;
      series x=x y=yhat1;
      series x=x y=yhat2;
      series x=x y=yhat3;
      series x=x y=yhat4;
      series x=x y=yhat5;
      xaxis label = 'Semester 1 GPA' values = (1 to 4 by 0.1);
      yaxis label = 'SAT Composite Score' values = (400 to 1600 by 100);
RUN;

PROC SGPLOT DATA = Plot;
      title "FTIC Nurse-Entry Data";
      SCATTER x = x_var y = y_var;
      xaxis label = 'Semester 1 GPA' values = (1 to 4 by 0.1);
      yaxis label = 'SAT Composite Score' values = (400 to 1600 by 100);
RUN;

%MACRO scatterplot(YEAR=);
PROC SGPLOT DATA = Plot;
      title "FTIC Nurse-Entry Enrollment, &YEAR./FA";
      where p_admit=1 and term_cohort="&YEAR./FA";
      SCATTER x = x_var y = y_var / group = zone;
      series x=x y=yhat1;
      series x=x y=yhat2;
      series x=x y=yhat3;
      series x=x y=yhat4;
      series x=x y=yhat5;
      xaxis label = 'Semester 1 GPA' values = (1 to 4 by 0.1);
      yaxis label = 'SAT Composite Score' values = (400 to 1600 by 100);
RUN;

%MEND scatterplot;
%scatterplot(11);

/*-------------------------------------------------------------------------
 *          Merge grid zone variable with other variables in Set_up
-------------------------------------------------------------------------*/
PROC SORT DATA=LIST;
      BY ID_num;
RUN;

PROC SORT DATA=Set_up;
      BY ID_num;
RUN;

Data test1;
      MERGE Set_up List;
RUN;
```

```
Data Var_1;
      MERGE Set_up List;
      BY ID_num;
      IF SAT_COMP = '.' THEN Zone=0; /*Grid zone '0' means that the
      observation did not contain a SAT_comp score*/
      IF SAT_COMP='.' and ACT_COMP='.' THEN have_exam=0;
      ELSE have_exam=1;

/*-------------------------------------------------------------------
 *                Dummy coding for remaining variables:
-------------------------------------------------------------------*/

/*Create dummy code for gender: GENDER=0 MALE */
IF GENDER_COHORT='F' THEN GENDER=1;
ELSE GENDER=0;

/*Create dummy code for Pell: PELL_S1=0 NO PELL*/
IF PELL_COHORT='1' THEN PELL_S1=1;
ELSE PELL_S1=0;

/*Create dummy code for first generation: FIRST_GEN=0 NO*/
IF FIRST_GEN_NOCOLL_0='Y' OR FIRST_GEN_NODEG_0='Y' THEN FIRST_GEN=1;
ELSE FIRST_GEN=0;


/*Create dummy variables for ethnicity: REFERENCE GROUP=WHITES*/
IF ETHNIC_COHORT='01'   THEN DO; ETHNIC_ALT='WH'; E1=0; E2=0; E3=0; END;
/*WHITES*/
IF ETHNIC_COHORT='02'   THEN DO; ETHNIC_ALT='BL'; E1=1; E2=0; E3=0; END;
/*BLACKS*/
IF ETHNIC_COHORT='03'   THEN DO; ETHNIC_ALT='HS'; E1=0; E2=1; E3=0; END;
/*HISPANICS*/
IF ETHNIC_COHORT IN ('04','05','06','07','08')  THEN DO; ETHNIC_ALT='OT';
E1=0; E2=0; E3=1; END; /* OTHER*/

/*Create dummy variables for grid zone: REFERENCE GROUP=ZONE 1*/
IF Zone = 0       THEN DO; Z1=.; Z2=.; Z3=.; Z4=.; Z5=.; END;
IF Zone = 1       THEN DO; Z1=0; Z2=0; Z3=0; Z4=0; Z5=0; END;
IF Zone = 2       THEN DO; Z1=1; Z2=0; Z3=0; Z4=0; Z5=0; END;
IF Zone = 3       THEN DO; Z1=0; Z2=1; Z3=0; Z4=0; Z5=0; END;
IF Zone = 4       THEN DO; Z1=0; Z2=0; Z3=1; Z4=0; Z5=0; END;
IF Zone = 5       THEN DO; Z1=0; Z2=0; Z3=0; Z4=1; Z5=0; END;
IF Zone = 6       THEN DO; Z1=0; Z2=0; Z3=0; Z4=0; Z5=1; END;

/*Create dummy variables for Gateway variable*/
IF GATE_VAR1 = 0 AND GATE_VAR2 = 0 THEN GATEWAY = 'Did not take';
IF GATE_VAR1 = 1 AND GATE_VAR2 = 0 THEN GATEWAY = 'Not A';
IF GATE_VAR1 = 0 AND GATE_VAR2 = 1 THEN GATEWAY = 'A';

IF GATE_VAR1 = 0 AND GATE_VAR2 = 0 THEN Gate_Not = 1;
ELSE gate_not = 0;
RUN;
```

```
/*-------------------------------------------------------------------
 *          Final dataset containing all variables of interest
-------------------------------------------------------------------*/
DATA Final_Var;
      SET VAR_1;
      KEEP ID_NUM term_cohort GENDER ETHNIC_ALT E1 E2 E3 E4 PELL_S1 GATEWAY
      GATEWAY_S1 GATE_VAR1 GATE_VAR2 GATE_NOT HOURS_TAKEN_S1 FIRST_GEN
      HAVE_EXAM P_ADMIT UPP_DIV_ADMIT GPA_S1 GPA_LEVEL SAT_M SAT_V SAT_COMP;
RUN;

PROC MEANS DATA=Final_var nmiss n;
RUN;
/*-------------------------------------------------------------------------
 *                Split data into training sets and test sets:
 *
 *                1) Training set: 5, 4, and 3 years before test sets
 *                2) Test set: 14/FA, 15/FA, 16/FA
-------------------------------------------------------------------------*/
DATA TRAIN13 ;
      SET FINAL_VAR;
      IF term_cohort NE '08/FA' AND term_cohort NE '09/FA' AND term_cohort NE
      '10/FA' THEN DELETE;
RUN;

DATA TEST13 ;
      SET FINAL_VAR;
      IF term_cohort NE '13/FA' THEN DELETE;
RUN;

DATA TRAIN14 ;
      SET FINAL_VAR;
      IF term_cohort NE '09/FA' AND term_cohort NE '10/FA' AND term_cohort NE
      '11/FA' THEN DELETE;
RUN;

DATA TEST14 ;
      SET FINAL_VAR;
      IF term_cohort NE '14/FA' THEN DELETE;
RUN;

DATA TRAIN15 ;
      SET FINAL_VAR;
      IF term_cohort NE '10/FA' AND term_cohort NE '11/FA' AND term_cohort NE
      '12/FA' THEN DELETE;
RUN;

DATA TEST15 ;
      SET FINAL_VAR;
      IF term_cohort NE '15/FA' THEN DELETE;
RUN;

/*--------------------LOGISTIC MODEL ANALYSIS-------------------------*/
```

```
/*Macronize the indepependent variables since they are used in multiple
places. This part of the modeling is dynamic in that you start with full
explanatory list and settle in a reduced model*/

%let testyear = 15;
%let admityear1 = %eval(&testyear+2);
%let admityear2 = %eval(&testyear+3);
%let Explanatory=

/*----------------------SEMESTER-1----------------*/
/*ETHNIC_ALT      /*ETHNICITY: REFERENCE GROUP=WHITES*/
/*PELL_S1         /*REFERENCE GROUP=0 (NOT PELL)*/
/*FIRST_GEN       /*REFERENCE GROUP=0 (NOT FIRST GENERATION)*/
/*GENDER          /*REFERENCE GROUP='M'*/
GPA_S1
SAT_M SAT_V       /*SAT MATH AND VERBAL SEPERATELY */
/*SAT_COMP
/*S1_HOURS
GATEWAY
/*------------------INTERACTION TERMS--------------*/
GPA_S1*GATEWAY
;
ODS RTF FILE="X:\My Personal
Folder\Classes\Thesis\SAS_Ouput\&testyear.LOGISTIC_RESULTS_REDUCEDMODEL.RTF";

/*------------------RUN MODEL--------------*/

PROC LOGISTIC DATA=Train&testyear DESCENDING OUTEST=FIT_LOGISTIC COVOUT
OUTMODEL=TRY;
 CLASS GATEWAY (ref='A') ETHNIC_ALT/ PARAM=GLM ORDER = INTERNAL;
   TITLE "&testyear./FA FTIC NURSE-ENTRY UPPER DIVISION ADMISSION MODEL -
REDUCED";
   MODEL P_ADMIT = &Explanatory

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

      OUTPUT out=train&testyear._results p = prob xbeta =logit; /*APPEND TO
THE DATA SET (FOR EACH OBS) LOGIT VALUE AND PROBABILITY*/
      ods output ParameterEstimates=LOGISTIC_PARMS;/*THIS IS FOR POOLING
MODEL RESULTS*/

/*---------- CALCULATE PROBILITY ON IMPUTED TESTING DATA SET-------------*/
     score data = TEST&testyear out = TEST&testyear._DATA_SCORED;

RUN;
/*----------- REPORT RESULTS OF MODEL PREDICTION ON TESTING DATA SET-------*/

DATA TEST&testyear._Model_RESULTS;
     SET TEST&testyear._DATA_SCORED;
```

```
      IF P_1='.' THEN DELETE;
      ELSE IF  0< P_1<.1 THEN GP_PR='01';
      ELSE IF .1<=P_1<.2 THEN GP_PR='02';
      ELSE IF .2<=P_1<.3 THEN GP_PR='03';
      ELSE IF .3<=P_1<.4 THEN GP_PR='04';
      ELSE IF .4<=P_1<.5 THEN GP_PR='05';
      ELSE IF .5<=P_1<.6 THEN GP_PR='06';
      ELSE IF .6<=P_1<.7 THEN GP_PR='07';
      ELSE IF .7<=P_1<.8 THEN GP_PR='08';
      ELSE IF .8<=P_1<.9 THEN GP_PR='09';
      ELSE IF .9<=P_1<1  THEN GP_PR='10';
      ELSE GP_PR='11';
RUN;

PROC MEANS DATA=TEST&testyear._Model_RESULTS SUM nway noprint;
      CLASS GP_PR;
      VAR P_1 P_ADMIT;
      OUTPUT OUT=Predictions&testyear SUM=;
RUN;

OPTIONS ORIENTATION=portrait  NODATE NONUMBER;
/*MISSING='?'FMTSEARCH=(IT_URR.FORMAT_LIBRARY)*/
ODS RTF FILE = "X:\My Personal
Folder\Classes\Thesis\SAS_Ouput\&testyear.FA_Prediction_Report";

PROC REPORT DATA=Predictions&testyear 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 University';
TITLE2 'Nurse-Entry Admission Model';
TITLE3 "Fall &testyear Nurse-Entry Cohort: Predict Upper Division
 Admission -- Reduced Model;";

COLUMN ("Nurse-Entry Upper division Admission Prediction for
 &testyear./FA" ("Enrolled &testyear./FA" GP_PR _FREQ_)
("Admitted &admityear1./FA or &admityear2./FA"  P_1 P_ADMIT));

DEFINE GP_PR       / /*GROUP FORMAT=$GP_PRFMT.*/ ORDER=INTERNAL WIDTH=12
      'Probability\Group';
DEFINE _FREQ_      / ANALYSIS SUM FORMAT=COMMA6.            WIDTH=10
      'Cohort\Total' CENTER;
DEFINE P_1         / ANALYSIS SUM FORMAT=COMMA6.            WIDTH=8
      'Predict\Admit' CENTER;
DEFINE P_ADMIT     / ANALYSIS SUM FORMAT=COMMA6.      WIDTH=8
      'Actual\Admit' CENTER;
```

```
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;
ods rtf close;

/*----------- SPLIT PROBABILITY PREDICTIONS INTO QUARTILES -----------*/

PROC MEANS DATA=Test&testyear._MODEL_RESULTS MIN MAX MEDIAN Q1 Q3 MAXDEC=5
NOPRINT;
      VAR P_1;
      OUTPUT OUT=QUARTS&testyear MIN=MIN MAX=MAX MEDIAN=MEDIAN Q1=Q1 Q3=Q3;
RUN;

DATA JOIN&testyear;
      MERGE TEST&testyear._MODEL_RESULTS QUARTS&testyear;
      DROP _TYPE_ _FREQ_;

      RETAIN _MIN;
      IF not missing(MIN) then _MIN=MIN;
      ELSE MIN=_MIN;
      DROP _MIN;

      RETAIN _MAX;
      IF not missing(MAX) then _MAX=MAX;
      ELSE MAX=_MAX;
      DROP _MAX;

      RETAIN _MEDIAN;
      IF not missing(MEDIAN) then _MEDIAN=MEDIAN;
      ELSE MEDIAN=_MEDIAN;
      DROP _MEDIAN;

      RETAIN _Q1;
      IF not missing(Q1) then _Q1=Q1;
      ELSE Q1=_Q1;
      DROP _Q1;

      RETAIN _Q3;
      IF not missing(Q3) then _Q3=Q3;
      ELSE Q3=_Q3;
      DROP _Q3;
RUN;

 DATA Quartiles&testyear;
      SET JOIN&testyear;
      IF p_1='.' THEN DELETE;
      IF p_1 LE Q1 THEN Q=1;
      IF p_1 GT Q1 AND P_1 LE MEDIAN THEN Q=2;
      IF p_1 GT MEDIAN AND P_1 LE Q3 THEN Q=3;
      IF P_1 GT Q3 THEN Q=4;
```

```
RUN;

PROC FREQ DATA=QUARTILES&testyear;
      TABLES Q*P_ADMIT / NOCOL NOROW NOPCT;
RUN;
ods rtf close;

/*-----------------INTERACTION TERM ANALYSIS-----------------------------*/

PROC LOGISTIC DATA = train15 DESCENDING;
      CLASS GATEWAY (ref='A')/ PARAM=GLM ORDER = INTERNAL;
      MODEL P_ADMIT = GPA_S1 SAT_M SAT_V GATEWAY
      GATEWAY*GPA_S1 / EXPB;
      ODDSRATIO GPA_S1 / AT(GATEWAY=ALL);
      STORE LOGIT;
RUN;

PROC LOGISTIC DATA = train15 DESCENDING;
      CLASS GATEWAY (ref='A') / PARAM=GLM ORDER = INTERNAL;
      MODEL P_ADMIT = GPA_S1 SAT_M SAT_V GATEWAY GATEWAY*GPA_S1;
      STORE LOGIT;
RUN;

PROC LOGISTIC DATA = TRAIN15 DESCENDING;
      CLASS GATEWAY (ref='A') / PARAM=GLM ORDER = INTERNAL;
      MODEL P_ADMIT = GPA_S1 SAT_M SAT_V GATEWAY GATEWAY*GPA_S1 /
      EXPB;
      ODDSRATIO GATEWAY / AT(GPA_S1 = 1.75 2 2.25 2.5 2.75 3 3.25 3.5 3.75
      4);
      STORE LOGIT;
RUN;

PROC PLM SOURCE = LOGIT;
      LSMEANS GATEWAY / AT GPA_S1=1.75 ILINK PLOTS=NONE;
      LSMEANS GATEWAY / AT GPA_S1=2.0 ILINK PLOTS=NONE;
      LSMEANS GATEWAY / AT GPA_S1=2.25 ILINK PLOTS=NONE;
      LSMEANS GATEWAY / AT GPA_S1=2.5 ILINK PLOTS=NONE;
      LSMEANS GATEWAY / AT GPA_S1=2.75 ILINK PLOTS=NONE;
      LSMEANS GATEWAY / AT GPA_S1=3 ILINK PLOTS=NONE;
      LSMEANS GATEWAY / AT GPA_S1=3.25 ILINK PLOTS=NONE;
      LSMEANS GATEWAY / AT GPA_S1=3.5 ILINK PLOTS=NONE;
      LSMEANS GATEWAY / AT GPA_S1=3.75 ILINK PLOTS=NONE;
      LSMEANS GATEWAY / AT GPA_S1=4 ILINK PLOTS=NONE;
RUN;

PROC PLM RESTORE=LOGIT;
EFFECTPLOT INTERACTION (X=GATEWAY) / AT (GPA_S1 = 1.75 2 2.25 2.5 2.75 3 3.25
3.5 3.75 4) CLM;
RUN;
/**END FILE**/
```

APPENDIX D

Fall 2013 and 2014 Results and Predictions

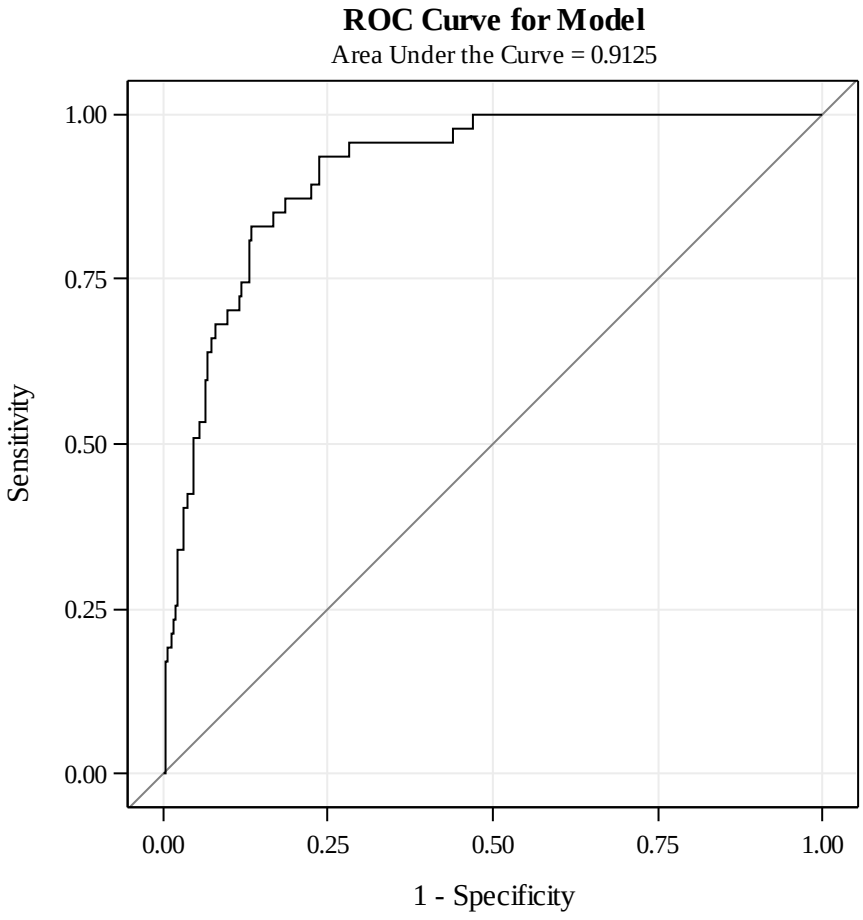
Fall 2014 Final Model Results and Predictions

Results of Final Logistic Regression Model for Fall 2014 TWU Nurse-entry Data

Parameter	Level	Coefficient Estimate	Standard Error	Wald Chi-Square	P-value
Intercept		-10.7610	5.0888	4.4718	0.0345
GPA_S1		1.7268	1.3028	1.7567	0.1850
SAT_M		0.00195	0.00259	0.5688	0.4507
SAT_V		0.00378	0.00255	2.1974	0.1382
GATEWAY	Did not take	-7.2949	5.9101	1.5235	0.2171
GATEWAY	Not A	-5.5852	7.0486	0.6279	0.4281
GPA_S1*GATEWAY	Did not take	1.9671	1.5502	1.6101	0.2045
GPA_S1*GATEWAY	Not A	1.3440	1.9544	0.4729	0.4917

Odds Ratio Estimates and Confidence Intervals for SAT Variables

Effect	Odds Ratio Estimate	95% Confidence Limits	
SAT_M	1.002	0.997	1.007
SAT_V	1.004	0.999	1.009



Hosmer and Lemeshow Goodness-of-Fit Test for Final Model

Chi-Square	DF	Pr > ChiSq
3.0641	8	0.9303

Nurse-Entry Upper Division Admission Prediction for Fall 2014

Enrolled 14/FA		Admitted 16/FA or 17/FA	
Probability Group	Cohort Total	Predict Admit	Actual Admit
01	254	4	19
02	34	5	9
03	23	6	9
04	12	4	8
05	5	2	4
Total	328	21	49

Fall 2014 Nurse-Entry Upper Division Admission Probability Prediction by Quartile

Predicted Probability Quartiles	Admitted		
	No	Yes	Total
0 – 25%	82	0	82
26 – 50%	77	5	82
51 – 75%	68	14	82
76 – 100%	52	30	82
Total	279	49	328

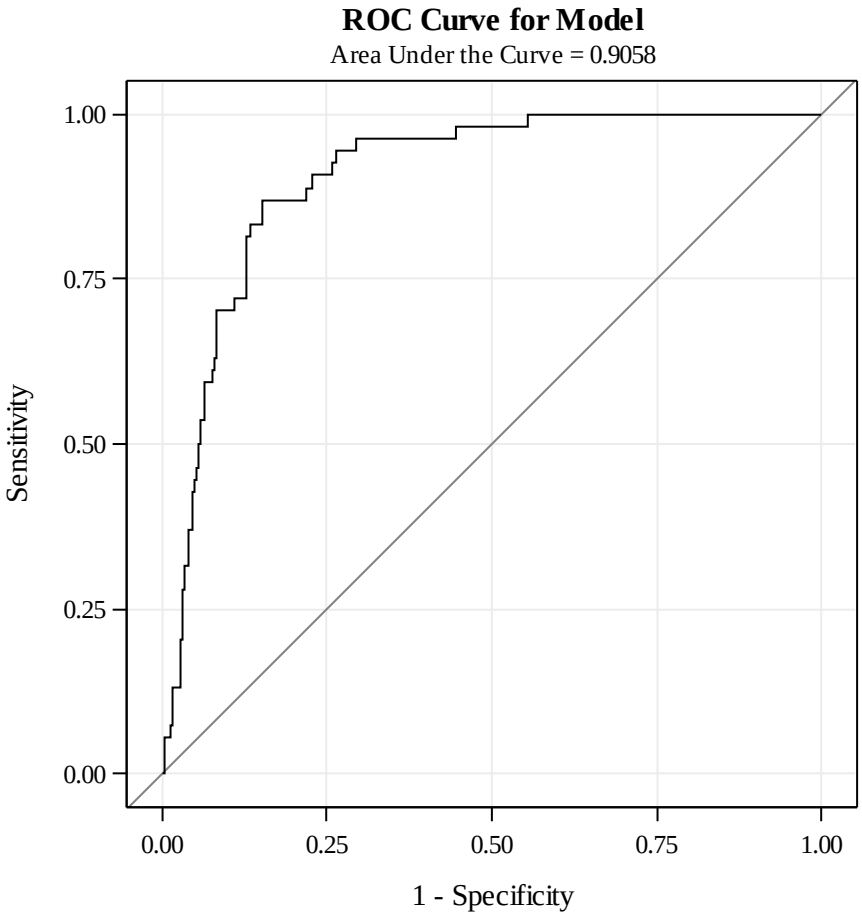
Fall 2013 Final Model Results and Predictions

Results of Final Logistic Regression Model for Fall 2013 TWU Nurse-entry Data

Parameter	Level	Coefficient Estimate	Standard Error	Wald Chi-Square	P-value
Intercept		-11.7090	4.0038	8.5524	0.0035
GPA_S1		1.8305	1.0170	3.2398	0.0719
SAT_M		0.00522	0.00250	4.3418	0.0372
SAT_V		0.00208	0.00242	0.7420	0.3890
GATEWAY	Did not take	-3.9405	4.5597	0.7469	0.3875
GATEWAY	Not A	-9.6743	7.8326	1.5256	0.2168
GPA_S1*GATEWAY	Did not take	1.0762	1.2102	0.7908	0.3739
GPA_S1*GATEWAY	Not A	2.6653	2.2380	1.4183	0.2337

Odds Ratio Estimates and Confidence Intervals for SAT Variables

Effect	Odds Ratio Estimate	95% Confidence Limits	
SAT_M	1.005	1.000	1.010
SAT_V	1.002	0.997	1.007



Hosmer and Lemeshow Goodness-of-Fit Test for Final Model

Chi-Square	DF	Pr > ChiSq
3.0667	8	0.9301

Nurse-Entry Upper Division Admission Prediction for Fall 2013

Enrolled 13/FA		Admitted 15/FA or 16/FA	
Probability Group	Cohort Total	Predict Admit	Actual Admit
01	205	3	0
02	51	7	8
03	35	9	12
04	10	3	2
05	23	10	15
06	6	3	3
07	1	1	0
Total	331	37	40

Fall 2013 Nurse-Entry Upper Division Admission Probability Prediction by Quartile

Predicted Probability Quartiles	Admitted		
	No	Yes	Total
0 – 25%	83	0	83
26 – 50%	83	0	83
51 – 75%	77	6	83
76 – 100%	48	34	82
Total	291	40	331

APPENDIX E

Predictive Accuracy Comparison

Predictive Accuracy Comparison between Final Model and GPA/SAT Model

Test Term	Final Model Accuracy	GPA/SAT Model Accuracy
Fall 2013	0.0091	0.0091
Fall 2014	0.0854	0.0853
Fall 2015	0.0086	0.0144