USING A STATISTICAL MODEL TO PREDICT STUDENT SUCCESS IN TEXAS WOMAN'S UNIVERSITY MATHEMATICS PROGRAM

A THESIS

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE IN THE GRADUATE SCHOOL OF THE

TEXAS WOMAN'S UNIVERSITY

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

COLLEGE OF ARTS AND SCIENCES

BY

ESTHER A. IBRAHIM

DENTON, TEXAS

AUGUST 2018

Copyright © 2018 by Esther A. Ibrahim

DEDICATION

I would like to dedicate this study to my parents, Mr. and Mrs. Ibrahim.

To my Father – Thank you sir for giving me quality education.

To my mother – Thank you ma'am for your prayers.

I would also like to thank my siblings - Seun, Funmi, Shile, and Femi.

To my uncle, Festus Oyewole - Thank you sir for your financial support.

To my best friend, Babatunde Aribikolasi – Thank you for believing in me.

ACKNOWLEDGEMENTS

I would like to thank my committee chairman, Dr. David Marshall, for the encouragement and support to complete this study. I sincerely appreciate Dr. Don Edwards, for his advice towards the completion of my program and Dr. Winifred Mallam, for her support and guidance on making my research a reality. I would also like to thank my parents, siblings, uncles, aunties, brothers, sisters, cousins, nephew and nieces for their support all through the journey of my master's degree because without them I will not have been able to achieve my educational goal. Lastly, I would like to give glory to God Almighty because He is the one that has given me the gift of life to finish this program.

ABSTRACT

ESTHER A. IBRAHIM

USING A STATISTICAL MODEL TO PREDICT STUDENT SUCCESS IN
TEXAS WOMAN'S UNIVERSITY MATHEMATICS PROGRAM

AUGUST 2018

Admission applications from high school students are received yearly by colleges but majority of the students that are being admitted eventually do not graduate. Colleges need to base their admission requirements on some specific factors to determine the students who will succeed. The purpose of this research was to utilize the data from first time incoming students who were enrolled full time and graduated from the mathematics department within five years to predict the success of future students seeking admission into the department. Data from Fall 2003 to Fall 2012 were used to build the predictive model. Success in this research is defined as the students that graduated from the mathematics program within five years of admission.

Two models were developed, one from doing a forward stepwise logistic regression on all the datasets and the second was using cross validation to build a model with the training datasets and checking the effectiveness on the testing datasets. The findings were that the SAT Mathematics score is the best predictor of success. A student with an

SAT Math score ≥ 590 has a probability of 0.5 of graduating within five years, while those < 590 are at risk of not graduating.

TABLE OF CONTENT

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Statement of the Problem

Graduates of high school enroll in colleges to earn a college degree, but some students do not graduate. Institutions need to rely on several factors that predict college success to determine which applicants to admit. The aim of these factors is to ensure that students have the necessary background, competencies, and determination to successfully complete their university degree. Significant effort has been made by researchers to predict student graduation in higher education and to understand the process of dropping out of college by developing theoretical models of student graduation and retention using associated factors. Despite these intensive efforts to improve student graduation, dropout rates are high across the United States (Yu, DiGangi, Jannasch-Pennell, & Kaprolet, 2010). The U.S. Department of Education's Center for Educational Statistics reported that only 50% of those who enroll in college earn a degree (Siedman, 2005). The goal of most colleges is not only to admit students, but also to retain students who would be successful at their institution. Therefore, with the increase in students seeking admission into colleges and universities, it is crucial that school

administrators and faculty use some measures or factors to predict the success of students in the program.

Every year, the Department of Mathematics and Computer Science at Texas Woman's University admits high school graduates into its undergraduate program. However, the number of students switching out of the mathematics major continues to be a prominent issue. To assist the mathematics students with timely advising early in their career, an effective prediction model of matriculation and graduation in mathematics that use available student data are highly desirable.

Purpose of the Study

The purpose of this research study is to use pre-college variables (i.e., ACT/SAT scores, high school GPA, high school rank, gender, and ethnicity) and college variables (GPA after first year and credit hours after first year) to devise a significant method that would help predict the success of students in the undergraduate mathematics program. College success in this research is defined as first year undergraduate students who graduated from the program within five years of graduation.

Research Questions

The examined research questions are:

1. Which variables are the best predictor of success?

2. What are the significant characteristics of students who graduated from the program?

3. What are the characteristics of students who did not graduate from the program?

Definition of Terms

Graduation: is defined as a first-time entering freshman student who graduates within five years of enrollment.

Retention: is defined as a first-time freshman student who gradually progresses and graduates within five years of enrollment.

Success: success is defined based on student graduation. A successful student gradually progresses through his/her degree and eventually graduates within five years of enrollment.

Variables: is defined as the characteristics or attributes of a student.

Grade Point Average: is the average obtained by dividing the total number of grade points by the total number of credits attempted.

American College Testing: (ACT) test is a curriculum and standards-based educational and career planning tool that assesses students' academic readiness for college.

Scholastic Aptitude Test: (SAT) is a standardized test widely used for college admissions in the United States.

Limitation and Delimitation

The population sample used in this study is comprised of students admitted and graduated from TWU. Therefore, the result of this research can only be applied to Texas Woman's University and cannot be generalized to other institutions in the United States. The data for this study were delimited to first time TWU undergraduate mathematics majors admitted into the program from Fall 2003 through Fall 2012. This research study also was delimited to student graduation within five years from their initial student enrollment.

In addition, standardized tests changed in 2016 from a scale of 1600 to 2400. Therefore, the present study only included students who had taken the SAT prior to admissions offices using the new format to make admissions decisions. The models developed in this study are only useful for predicting graduation within five years for students that took the SAT prior to the change in format. To account for the change, new models should be developed once substantial data is available for students that took the SAT after the change.

CHAPTER II

LITERATURE REVIEW

Theories on student graduation are examined since graduation is considered an important theme throughout this study. Also, the literature review will develop a clear understanding of the relationship of the variables which were used in this study with student graduation.

Theoretical Perspective on Student Graduation

Over the past decades, studies have been conducted on student graduation from an institution. Literature defines student graduation in terms of retention. Walleri (1981) defined retention as a timely graduation within four to five years. According to Thomas (2009), retention in an institutional setting refers to "the percentage of students in a particular year who neither graduate nor continue studying in an award course at the same institution in the following year" (p.9). Seidman and Hagedorn (2005) defined retention as the act of staying in college until the degree is completed. Wild and Ebbers (2002) defined retention as a measure and state that "student retention is significant for measuring institutional effectiveness in the prevailing environment of accountability and budgetary constraints" (p.503). Despite the numerous definitions of retention by

researchers, there are many common ideas, most of which focus on degree attainment from a college. While other views center on different grade levels of students.

Student Graduation: ACT/SAT and High School GPA

Studies have examined the empirical relationship between academic achievement and scores on standardized tests as a predictor of how well a student will perform in college. High school achievement factors such as high school GPA (HSGPA) and ACT /SAT scores are among the most consistent predictors of graduation and retention. Astin, Korn, and Green (1987), in their study of 8,000 students, found that students entering college with an "A" average from high school were seven times more likely to graduate within four years than students with a "C" average from high school. Furthermore, students with the highest SAT scores were six times more likely to graduate within four years than students with the lowest SAT scores.

The study conducted by Levitz, Noel, and Richter (1999) found that schools with the highest averages of test scores report a first- and second-year retention rate of 91%. Meanwhile, this rate was only 56% among students with the lowest test score averages. Sawyer (2010) admited that when predicting college success, the best choice is using both high school grade point average and ACT scores instead of using either by itself.

Tross, Harper, Osher, and Kneidinger (2000) found a stronger relationship between high school achievement variables and retention. Tross et al. (2000) research shows that high school grade point average and SAT/ACT score accounted for 29% of the variance in retention. Wolfe and Johnson (1995) cited the high multicollinearity between HSGPA and SAT/ACT. Multicollinearity is the correlation between two or more predictor variables. Studies indicated that high school GPA is the stronger predictor of the two variables, but multicollinearity issue could cause some of the predictive power of either variable to go undetected.

Student Graduation: First Year GPA and Credit Hours

Studies have shown that collegiate achievement variables like first year GPA, credit hours/course loads are also predictor of student success. First year GPA is very important as this gives the university early warning of students who are at risk of dropping out or not graduating within five years.

Murtaugh, Burns, and Schuster (1999) used first-quarter GPA to predict retention between the first and second years of college. In their findings, the likelihood of returning for a second year of college increased dramatically with higher GPA. Students with the lowest GPA (0.0–2.0) had a 57% probability of being retained, while students with the highest GPAs (3.3–4.0) had a 91% probability of being retained. Furthermore, Murtaugh

et al. (1999), in their multivariate model, reported that the value of the hazard ratio for GPA was 0.49. Therefore, for each point increase in GPA the probability of withdrawal from the university decreases by 49%. Also, Allen (1999) found that first-year college GPA was a statistically significant predictor of student's retention for both minority and nonminority students in the study. For both minority and nonminority students, first-year college GPA exerted the largest direct effect on whether a student was retained or graduated.

Researchers also found that the credit loads a student takes is a good predictor of success. 15 to Finish, a presentation by David Mongold and Joanne Itano (2017) is a campaign which originated at the University of Hawaii. It encourages students to take more credits, graduate on time and start earning faster. Smith (2016) reported in her publication that the Community College Research Center at the Teachers College at Columbia University released a report that says students who enroll in 15 credits' worth of classes in their first semester are more likely to graduate than those students who enroll with only 12 credits. Ahmed Shami, Ahmed Abo-Laban, and Ahmed Shami (1980) found in their studies that in every department student who registered for less than 12 credits had the lowest semester GPAs while students who registered for more than 17 credits had the highest GPAs. In a study conducted by Duby and Schartman (1997), they noted that first-semester credit load patterns tend to persist over time. Students who

begin with light credit loads tend to continue with light credit loads and students who begin with heavy credit loads tend to continue with heavy credit loads.

Student Graduation: Ethnicity and Gender

Ethnicity and gender are known as demographic factors that are also predictor of retention and graduation. Studies have identified gender and race/ethnicity as two of the four most consistent predictors of retention and that male students are less likely to persist than female students. Astin (1975), Astin et al. (1987), and Tinto (1987) found that gender was related to whether a student retained. Peltier, Laden, and Matranga (1999) reported relatively consistent findings that gender was predictive of persistence, with women more likely to persist than men.

Asian American and/or White students were most likely to retain in college, while other racial groups were less likely to retain (Astin, 1997; Murtaugh et al., 1999; Peltier et al., 1999). Murtaugh et al. (1999) and Leppel (2002) found relationships between gender and race that influenced retention. Leppel (2002), in a national study of 5,384 undergraduate students, also explicated the effects of marital status and age on the persistence of men and women.

CHAPTER III

METHODOLOGY

Data Collection

The data for this study were collected as de-identified data from the Institutional Research and Improvement Office at Texas Woman's University. Data were collected from all students that graduated from Texas Woman's University mathematics program between Fall 2003 and Fall 2012 with a bachelor's degree. Fall 2012 was chosen as the last year for analyses to allow sufficient time for students to have graduated. The dataset included only first-time entering undergraduates who began the first fall term of their academic career as full time (12 or more credits) students. Those who started as part-time students were not included in the dataset.

Data Analysis

Exploratory data analysis was used to gain insight on the data and to answer the research questions, statistical analysis of the data was conducted using a software called SPSS. Success is the dependent variable of this study. It is a dichotomous variable with a value of 0 for students who have not graduated within five years and 1 for students who graduate within five years. The independent variables that are the potential predictors are ACT/SAT scores, high school GPA, high school rank, gender, ethnicity, GPA after first

year and credit hours earned at end of first year. Table 1.1 gives a description of the variables. The categorical variables were converted into numerical variables. The ACT math score was converted into the SAT mathematics score using a concordance table in Appendix A. Cross validation was used to split the data randomly into 70% to build the model and 30% for model evaluation. Logistic regression was used to model the relationship between dependent variable and independent variable.

Table 1.1

List of Variables

Variable	**Type**	**Description**
HSGPA	Continuous	Cumulative high school G.P.A
HSRANK	Continuous	Percentile high school rank
SAT/ACT	Continuous	SAT score or converted ACT score
Gender	Categorical	Sex of the student
First GPA	Continuous	Average GPA after first year
Total Hours	Continuous	Hours earned after first year
Ethnicity	Categorical	Social group that student identifies with.

Variable Selection

Identifying the best subset among many variables to include in a model is arguably the hardest part of model building. Too many variables in a model might lead to overfitting and too few variables might lead to underfitting; therefore, it is crucial to choose the most optimal variables for a high-quality model. Many variable selection methods exist. Some of the most common variable selection methods include forward selection, backward elimination, stepwise, best subsets and all possible subsets. In this research the forward selection method will be used. The forward selection technique begins with just the intercept and then sequentially adds the variables that most improves the fit. The significance of any variable is evaluated in terms of a likelihood ratio chi-square test. The process terminates when no significant improvement can be obtained by adding any variable.

Logistic Regression

When the dependent variable is binary (dichotomous), such as the students' retention status, conventional linear regression is not appropriate as the modeling tool. Instead, logistic regression becomes a better suited statistical method for such models. For a linear regression model see Equation 1.1

Assuming:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n + e \quad 1.1$$

Where:

$\hat{y}$ is the dependent variable

x_i is the i^{th} independent variable

β_0 is the regression constant

β_i is the i^{th} regression coefficient

n is the number of independent variables

e is the error

Some of the assumptions associated with a multiple regression error term are:

1. The variance of the error term is constant regardless of the value for x_1, x_2, x_3 …. .

2. The error values are independent

3. The error term is a normally distributed random variable μ = 0 and variance = σ^2.

Logistic regression has been used in educational studies to predict a student's retention or graduation status because it does not require any of the stated assumptions (Fadlalla, 2005). It is a statistical method for analyzing datasets in which there are one or more independent variables that determine a categorical dependent variable. It provides an association between the independent variable and the logarithm of the odds of a

categorical dependent variable. The categorical dependent variable is measured with a dichotomous variable in which there are only two possible outcomes (success or failure). In this study, the target variable is graduation within five years which is defined as success or failure of the student. For every $k \in R$, $Y_k = 1$ represents graduating within five years of admission (success) and $Y_k = 0$ represents not graduating within five years of admission (failure). See Table 1.2.

Table 1.2

Defining Success and Failure

Success $(\boldsymbol{Y_k = 1})$	**Failure** $(\boldsymbol{Y_K = 0})$
Graduate within five years	Does not graduate within five years.

For each $k \in R$, there are only two outcomes for the dependent variable which mean Y_k can be modeled as a Bernoulli random variable. Defined the probability that $Y_k = 1$ as $P(Y_k = 1) = \left(\frac{n_{y=1}}{N}\right) =$, that is divide the cases of for whom $Y = 1$, by the total N, the predicted value can fall outside the range of 0 and 1. This problem can be solved by

first replacing the probability $y_k = 1$ with the odds of $y_k = 1$.The odd function, which is shown by equation (1.2), has a range of $(0, \infty)$.

$$odds(Y_k = 1) = \frac{P(Y_k=1)}{1-P(Y_k=1)} \quad 1.2$$

The second step is to take the natural log of the odds, called logit of Y_k, which produces a range of $(-\infty, \infty)$.

$$logit(Y_k) = \frac{\ln[P(Y_k=1)]}{1-P(Y_k=1)} \quad 1.3$$

By using the logit of $Y_k = 1$, the dependent variable will then be in the range of 0 to 1.

Let each independent variable be associated as a linear combination with an unknown coefficient as shown in Equation 1.4.

$$P(Y_k = 1) = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r = \beta_0 + \beta_k x_k \quad 1.4$$

The conditional mean or the expected value $E(Y_k|x_k)$ for logistic regression can then be written as shown in Equation 1.5.

$$P(Y_k = 1) = \frac{e^{\beta_0+\beta_k x_k}}{1+e^{\beta_0+\beta_k x_k}} \ ,$$

or

$$P(Y_k = 1) = \frac{1}{1+e^{-(\beta_0+\beta_k x_k)}}$$

then

$$1 - P(Y_k = 1) = 1 + \frac{e^{\beta_o + \beta_k x_k}}{1+e^{\beta_0 + \beta_k x_k}} \quad - \quad \frac{e^{\beta_0+\beta_k x_k}}{1+e^{\beta_0+\beta_k x_k}} = \frac{1}{1+e^{\beta_0+\beta_k x_k}}$$

and

$$\frac{P(Y_k=1)}{1-P(Y_k=1)} \; = \; e^{\beta_0+\beta_k x_k} \qquad 1.5$$

Combining Equation 1.3 and 1.5, results in $logit(Y_k)$ becoming Equation 1.6.

$$logit(Y_k) = \ln[e^{\beta_0+\beta_k x_k}]$$

$$= \beta_0 \; + \; \beta_k x_k \qquad 1.6$$

The parameter β of the model are estimated by the maximizing the log-likelihood function (see Equation 1.7).

$$-2LL = \; -2 \sum_{\{i=1\}} \{P(Y_k = 1) \ln[Y_k] + \big(1 - P(Y_k = 1)\big) \ln(1 - Y_k)\} \qquad 1.7$$

Success in prediction will be a probability over 0.50 that the event will happen, concomitant with the fact that the event did happen. While a prediction of failure will be a computed probability below 0.50 for a subject when in fact the event had happened.

Multicollinearity

Chapter 2 discussed some independent variables that are likely predictors of our success, that is, graduation within five years of admission. Some selected variables in a logistics regression model could result in multicollinearity. The generated model in this study was checked for multicollinearity.

Multicollinearity refers to a situation where several independent variables in a regression model are closely correlated to one another. Multicollinearity can lead to skewed or misleading results when a researcher or analyst is attempting to determine how well each one of several individual independent variables can most effectively be utilized to predict or understand the dependent variable in a statistical model. In general, multicollinearity can lead to wider confidence intervals and less reliable probability values (p values) for the independent variables.

One method used to measure multicollinearity is the variance inflation factor (VIF), which assesses how much the variance of an estimated regression coefficient increases if the predictors are correlated. If no factors are correlated, the VIFs will all be 1. If a collinearity issue arises, that is, the VIFs for a factor is near or above five then the highly correlated predictors or independent variable will be removed from the model. The data used in this research will be checked to see if there is a collinearity issue.

Cross Validation

Cross validation has been used to build and test models to avoid overfitting or under fitting. Cross validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments as shown in Figure 1. One is used to learn or train a model and the other is used to validate the model. In typical cross-validation, the training and validation sets must cross-over in successive rounds such that each data point has a chance of being validated against. In this research the holdout cross validation method would be used. In the holdout method, data points are randomly assigned to two sets; say x_0 and x_1, usually called the training set and the test set, respectively. The size of each of the sets is arbitrary, although typically the test set is smaller than the training set. x_0 will then be the training set and x_1 the testing set. Hold-out validation avoids the overlap between training data and test data, yielding a more accurate estimate for the generalization performance of the algorithm.

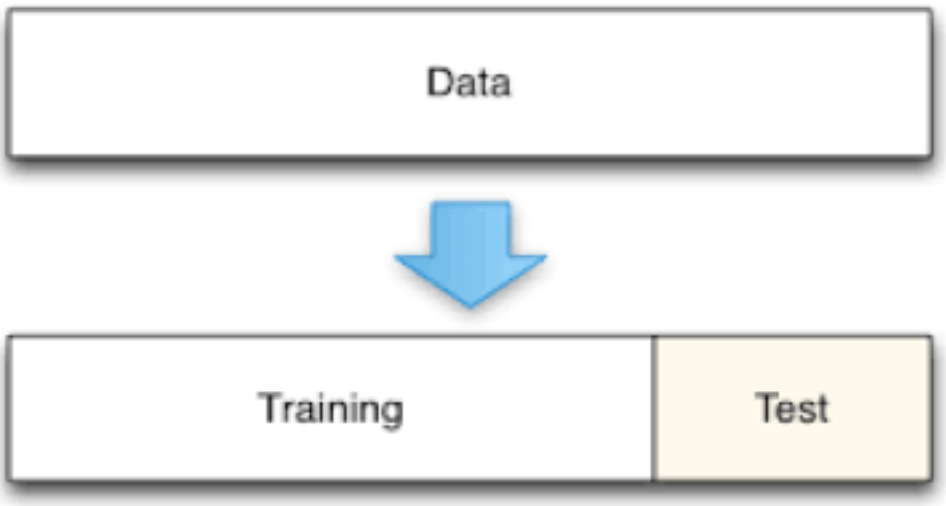


Figure 1. Cross validation of data into training and testing set.

CHAPTER IV

RESULTS

The purpose of this research is to predict the success of students in the Department of Mathematics and Computer Science at Texas Woman's University, Denton Texas. Data analysis for the research was performed using 97 sample datasets of first time incoming students from fall 2003 through fall 2012 who were enrolled as full-time.

Exploratory Data Analysis

Exploratory data analysis was used to gain insight about the data and incorporate domain knowledge about the data before performing statistical analysis. The data exploration actions included visual techniques that examined the dataset in terms of summary statistics with respect to student graduation. Every variable in the data set was explored to find patterns in student graduation for entering full-time, first-time freshmen from 2003 until 2012.

Table 1.3 shows the graduation and dropout rate for 10 cohorts of full time incoming students. The table summarizes the number of students admitted for each year and the percentage that graduated or dropout of the mathematics program within five years of admission. The year 2009 has the highest percentage of graduates with no

dropouts, while year 2010 has the lowest percentage of graduates. The year 2006 has the highest full-time incoming student enrollments.

Table 1.3

Graduation and Drop Rate by Cohort

Year	n	Graduate ($\leq 5 years$)	G%	Drop	D%
2003	7	4	57	0	0
2004	7	3	43	2	29
2005	8	4	50	3	38
2006	16	9	56	5	31
2007	8	3	38	2	25
2008	10	4	40	5	50
2009	7	5	71	0	0
2010	13	4	31	5	39
2011	10	5	50	4	40
2012	11	5	45	5	45
Total	97	46	47	31	32

Table 1.4 shows the graduation and dropout rate of students by gender for full time freshmen students enrolled from 2003 to 2012 at Texas Woman's University. Texas Woman's University is a college that comprises of 85% female and 15% male, therefore

the enrollment for male is lower compared to female. Table 1.4 shows that female students have higher percentage of persisting and graduating than the male students.

Table 1.4

Graduation and Drop Rate by Gender

		Male					Female				
Year	n	n	G	G (%)	D	D (%)	n	G	G (%)	D	D (%)
2003	7	2	1	50	0	0	5	3	60	0	0
2004	7	2	0	0	1	33	5	3	75	1	25
2005	8	0	0	0	0	0	8	4	50	3	37
2006	16	3	1	33	0	0	13	9	69	4	38
2007	8	1	1	100	0	0	7	3	43	2	29
2008	10	0	0	0	0	0	10	4	40	5	50
2009	7	0	0	0	0	0	7	5	71	0	0
2010	13	1	1	100	0	0	12	3	25	5	42
2011	10	2	1	50	1	50	8	4	50	2	25
2012	11	0	0	0	0	0	11	5	45	3	27
Total	97	11	5	45	2	18	86	43	50	25	29

The total number of female enrollment in the mathematics program constitute about 88.7%. See Table 1.4. The highest graduation rate for females was 2004 and the lowest was 2010. Also, in 2008 female students have the highest drop out of 50% while

no female students dropped from the program in 2003 and 2009. The total number of male enrollment into the mathematics program constitute about 11.3% (see Table 1.4). The highest graduation rate for males were in 2007 and 2010, while the lowest were in 2004, 2005, 2008, 2009 and 2012. Also, in 2011 male students have the highest drop of 50%, while no male students dropped from the program in 2003, 2005, 2006, 2007, 2008, 2009, 2010, and 2012.

Table 1.5 and Figure 2 show the graduation and dropout rate by ethnicity for full-time incoming students enrolled from Fall 2003 to Fall 2012 in the Department of Mathematics and Computer Science at Texas Woman's University. The column n shows the number of students in the ethnic categories for the sample of students admitted into the program from 2003 to 2012.

White students constituted 37.1% of the total full time incoming students. The highest graduation rates were 2009 and 2012 while the lowest was 2008 and 2010. The highest dropout rates were 2008 and 2010 while the lowest were 2003, 2007, 2009, and 2012. The overall graduation and dropout rates are 61% and 25%, respectively.

African American students constituted 20.6% of the total full time incoming students in the datasets. The highest graduation rates were in 2003 and 2004, while the lowest were in 2005, 2008, 2009 and 2012. The highest dropout rate was in 2008 while

the lowest drop rates were in 2003, 2004, 2005, 2007, 2009, 2010, and 2012. The overall graduation and dropout rates are 40% and 20%, respectively.

Table 1.5

Graduation and Drop Rate by Ethnicity

Year	White					African American					Hispanic					Asian					American Indian				
	n	G	D	G %	D %	n	G	D	G %	D %	n	G	D	G %	D %	n	G	D	G %	D %	n	G	D	G %	D %
2003	4	2	0	50	0	1	1	0	100	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2004	3	2	1	67	33	1	1	0	100	0	3	0	1	0	33	0	0	0	0	0	0	0	0	0	0
2005	5	3	1	60	20	0	0	0	0	0	2	1	1	50	50	1	0	1	0	100	0	0	0	0	0
2006	8	5	1	63	25	4	2	1	50	25	3	1	2	33	67	1	0	1	0	100	0	0	0	0	0
2007	2	1	0	50	0	2	1	0	50	0	4	1	2	25	50	0	0	0	0	0	0	0	0	0	0
2008	3	1	2	33	67	3	0	2	0	67	3	2	1	67	33	1	1	0	100	0	0	0	0	0	0
2009	4	4	0	100	0	0	0	0	0	0	2	1	0	50	0	1	0	0	0	0	0	0	0	0	0
2010	3	1	2	33	67	3	1	0	33	0	6	2	3	33	50	0	0	0	0	0	2	0	0	0	0
2011	3	2	1	67	33	4	2	1	50	50	3	1	0	33	0	0	0	0	0	0	0	0	0	0	0
2012	1	1	0	100	0	2	0	0	0	0	7	4	4	57	29	0	0	0	0	0	1	0	1	0	100
Total	36	22	9	61	25	20	8	4	40	20	34	13	14	38	41	4	1	2	25	50	3	0	1	0	33

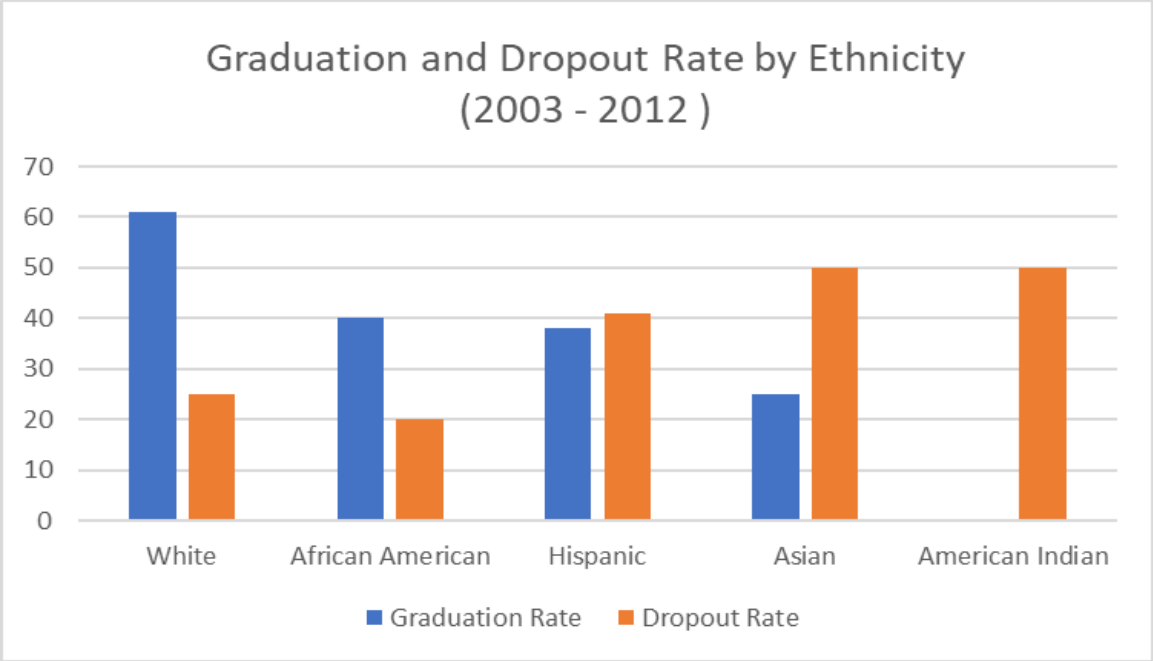


Figure 2. Bar chart of overall percentage of graduation and dropout rate by ethnicity.

Hispanic students constituted 35.1% of the total full incoming students. The highest graduation rate was in 2008, while the lowest were in 2003 and 2004. The highest drop rate was in 2006, while the lowest were in 2003, 2009, and 2011. The overall graduation and dropout rates are 38% and 41%, respectively. Asian students constituted 4.1% of the total full time incoming students. The highest graduation rates were experienced in 2008 while the dropout rates were high in 2005 and 2006. The overall graduation and dropout rate are 25% and 50%, respectively.

American Indian students constituted 2.1% of the total full time incoming students in the datasets. None of the students graduated from the program and there was a 100% drop rate in 2012. The overall graduation and dropout rate are 0% and 33%, respectively. In summary, in terms of enrollment and graduation white students have the highest percentage of graduates, while in terms of drop rate Hispanic students have the highest percentage.

Multicollinearity

Chapter 3 discussed multicollinearity and its effect on data. The dataset was analyzed to detect any outliers or collinearity before doing prediction. Appendix B shows the SPSS output of Tolerances and variance proportions. High school GPA and the high school rank are correlated with each other, an expected result. The ethnics indictors were also predictably redundant. However, since the prediction models were to be done in a step-wise manner, collinear variables were not expected to enter any equation. Multicollinearity is a more crucial concern when predictors are force entered.

Research Question One

To answer the first research question, a forward stepwise logistic regression was done to determine which of the variable was the best predictor of success. A forward stepwise regression was performed on all the data and after which a cross validation

was used to divide the data into training and testing sets. The training set comprises 70% of the data, while 30% of the data was used as testing sets. All independent variables were included in the analysis including the ones that were correlated. The variable that was considered as a significant predictor of success in the final model is SAT mathematics score. Appendix C shows the SPSS output analysis of all cases. SAT mathematics score is reported to be the only predictor of success. The fitted model is

$$\hat{y} = 0.012\,x - 6.946 \qquad 1.8$$

The model in Equation 1.8 is generated from the analysis of all datasets shown in Appendix C. In the model, -6.946 is the y-intercept, while 0.012 is the slope and x represent the SAT mathematics score of each student.

According to this model, it is predicted that students that have a high SAT math score have a higher probability of graduating within five years. This model will help the Department of Mathematics and Computer Science at Texas Woman's University predict the students who are at elevated risk of failing 84.6% of the time, likewise, it will identify students who will graduate within five years 45.5% of the time. The overall predictive power of this model is 70.5%.

To estimate how accurately this predictive model will perform, a cross validation of the datasets was done. A random Z score was generated, and the positive Z score was used

as the training data while the negative Z score was used as the testing data. Forward stepwise logistic regression was used to build the model on the training data and SAT mathematics score was found to be the only best predictor. The model that fit the training data is

$$\hat{y} = -6.939 + 0.012\,x \qquad 1.9$$

The model in Equation 1.9 is generated from the analysis of the training data Appendix C. In the model, -6.939 is the y-intercept, while 0.012 is the slop and X represents the SAT mathematics score of each student.

The model built with the training datasets correctly predicted 84.0% of the time that students will succeed (graduate within five years from the mathematics program) and 50% of the time students who will fail (not graduate within five years from the mathematics program). To confirm how effective, the model was, it was used for prediction on the testing datasets. It predicted failure correctly 71.4% of the time and predicted success correctly 75.0% of the time. The overall predictive power of this model is 72.2%.

Research Questions Two and Three

To ensure that the Department of Mathematics and Computer Science at Texas Woman's University identifies at an early stage student who have tendency of failing

from the program, it is essential to know the characteristics of these students, so they can provide academic help such as the Mathematics and Technology Success Center. This research found that students who have SAT mathematics scores less than 590 are at elevated risk of failing from the program. Appendix E shows a graph of SAT mathematics and the predicted probability. If the SAT mathematics score of a student is put into Equation 1.8 or 1.9 and the predicted probability falls below 0.5 or in the third quadrant of the graph in Appendix D, then the students are at risk in the program. This study also found out that a student with an SAT mathematics score of at least 590 has a probability of 0.5 graduating from the program.

CHAPTER V

CONCLUSION

The purpose of this study was to use pre-college variables and college variables to build a model that would help the Department of Mathematics and Computer Science at Texas Woman's University predict the success of full-time incoming freshmen students. The exploratory analysis of the data was done to gain insight about the data. Over 60% of the students admitted into the department from 2003 to 2012 were white. Female students comprise of 86.7%, while male students comprise of 11.3%. The highest graduation rate occurs in year 2009, while the lowest graduation rate was year 2010. In year 2008, 50% of the students dropped out of the program within five years, while there was no drop out within five years in year 2003 and 2009. The overall graduation rate within five years for these 10 cohorts is 48.1%, while the overall drop rates within five years was 29.7%.

SPSS was used for the statistical analysis. The dependent variable was success and the independent variables were high school GPA, high school rank, SAT mathematics score, ethnicity, gender, first year GPA and number of hours earned after first year. The categorical variables were coded into dummy variables and ACT mathematics scores of some students was converted to SAT mathematics score using

the concordance table in Appendix A. When the datasets were checked for collinearity, high school GPA and high school rank were found to have a strong correlation, of 0.758. Forward stepwise logistic regression was used to build the model by analyzing the data in two diverse ways. Analysis was done on all the dataset and to see how effective the model was, a cross validation was also done.

The variable found to be the best predictor of success was the SAT mathematics score. The findings were that a student who has an SAT mathematics score of at least 590 has a 0.5 probability of graduating within five years, while a student who has an SAT mathematics score of less than 590 is at risk of not graduating within five years. This model has an overall prediction power of about 72. 5%. With this model, the Department of Mathematics and Computer Science Department at Texas Woman's University can know the potential of students admitted into the department and can also provide the necessary support to students who have been detected to not graduate within five years.

Recommendations for Further Study

1. For future research, additional predictors can be examined. The age and grade level of the student when he or she took the standardized exam might influence

success. The scores of each section of the SAT score might also be included for improved prediction accuracy.

2. The amount of funding and the health status of students may also be considered, since financial aid and the health status of students may affect timely graduation.
3. In this study, success is defined as graduation within five years. Future study might measure success as whether a student gets a job immediately after graduation or went to graduate school. Given the high predictive accuracy of this study's regression model, the primary recommendation for future research is to replicate the study at other departments or institutions to predict enrollments.

REFERENCES

Ahmed Shami, M. A., Ahmed Abo-Laban, M., & Ahmed Shami, M. B. (1980). *Relationship of students'course loads with their grade point average scores*. Makkah, Saudi Arabia: King Adulaziz University (ERIC ED202441), 2-8.

Allen, D. (1999). Desire to finish college: An empirical link between motivation and persistence. *Research in Higher Education, 40*(4), 461–485.

Astin, A. W. (1975). *Preventing students from dropping out*. San Francisco: Jossey-Bass.

Astin, A. W., Korn, W., & Green, K. (1987). Retaining and satisfying students. *Educational Record, 68(*1*),* 36–42.

Astin, A. W. (1997). How ''good'' is your institution's retention rate? *Research in Higher Education, 38(6),* 647–658.

Duby, P., & Schartman, L. (1997). Credit hour loads at college onset and subsequent college performance: A multi-institution pilot project. AIR 1997 Annual Forum Paper.

Fadlalla, A. (2005). An experimental investigation of the impact of aggregation on the performance of data mining with logistic regression. *Information and Management,* 42(5), 695-707.

Leppel, K. (2002). Similarities and differences in the college persistence of men and women. *The Review of Higher Education* 25(4), 433–450.

Levitz, R., Noel, L., & Richter, B. J. (1999). Strategic Moves for Retention Success. *New directions for higher education, 1999*(108), 31-49.

Mongold, D., & Itano, J. (2017). *15 to Finish* [PowerPoint Presentation]. Retrieved from http:// hawaii.edu/hawaiigradinitiative/15-to-finish/.

Murtaugh, P. A., Burns, L. D., & Schuster, J. (1999). Predicting the retention of university students. *Research in Higher Education*, *40*(3), 355–371.

Peltier, G. L., Laden, R., Matranga, M. (1999). Student persistence in college: A review of research*. Journal of College Student Retention, Theory & Practice,* 1(4), 357–376.

Sawyer, R. L. (2010). Usefulness of high school average and ACT scores in making college admission decisions. ACT Research Report Series 2010-2. *ACT, Inc.*

Seidman, A. (2005). Where we go from here. College student retention: *Formula for student success,* 295.

Seidman, A., & Hagedorn, L. (2005). How to Define Retention. *College student retention: formula for student success.* Westport, CT: Praeger Publishers, 89-105.

Smith, A. (2016, June 30). Study finds first-year students who take 15 credits succeed. Retrieved from https://www.insidehighered.com/quicktakes/2016/06/30/study-finds-first-year-students-who-take-15-credits-succeed.

Thomas, Liz. (2009). Improving student retention in higher education. *Australian Universities Review 51(2),* 9-18.

Tinto, V. (1987). *Leaving college: Rethinking the causes and cures of student attrition*. Chicago: University of Chicago Press, 5801 S. Ellis Avenue, Chicago, IL 60637.

Tross, S. A., Harper, J. P., Osher, L. W., & Kneidinger, L. M. (2000). Not just the usual cast of characteristics: Using personality to predict college performance and retention. *Journal of College Student Development*, 41, 323–334.

Walleri, D. (1981) Student Retention and Attrition in the Community College. A Review and Research Design ERIC Number: ED210064, p. 37.

Wolfe, R. N., & Johnson, S. D. (1995). Personality as a predictor of college performance. *Educational and Psychological Measurement, 55*(2), 177–185.

Yu, C. H., DiGangi, S., Jannasch-Pennell, A., & Kaprolet, C. (2010). A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science, 8(2),* 307-32.

Wild, L., & Ebbers, L. (2002). Rethinking Student Retention in Community Colleges. *Community College Journal of Research and Practice* 26(6), 1-17.

APPENDIX A

SAT MATH TO ACT MATH CONCORDANCE TABLE

ACT MATH SCORE	SAT MATH SCORE (Before March 2016)	SAT MATH SCORE (After March 2016)
36	800	800
35	790	800
34	780	790
33	760	780
32	730	760
31	700	730
30	680	710
29	660	690
28	640	660
27	620	640
26	600	620
25	580	600
24	560	580
23	540	570
22	520	550
21	500	530
20	480	510
19	460	500
18	440	480
17	410	450
16	390	430
15	360	400
14	330	370
13	300	350
12	280	330
11	260	300

http://catalog.usu.edu/content.php?catoid=12&navoid=7347

APPENDIX B

Multicollinearity and Correlation Check among the Independent Variables

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	RANK_PERCENT_0	.393	2.544
	AP_GPA_0	.322	3.106
	SP_TOTAL_HOURS_1	.505	1.982
	SAT_M	.550	1.820
	white	.059	16.933
	afamer	.077	13.023
	hispan	.059	16.815
	asiapac	.201	4.972
	sex	.461	2.167

a. Dependent Variable: Success or failure

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions									
				(Constant)	RANK_PERCENT_0	AP_GPA_0	SP_TOTAL_HOURS_1	SAT_M	white	afamer	hispan	asiapac	sex
1	1	6.292	1.000	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	2	1.141	2.349	.00	.02	.00	.00	.00	.02	.02	.00	.01	.00
	3	1.007	2.499	.00	.00	.00	.00	.00	.00	.03	.01	.05	.00
	4	1.001	2.507	.00	.00	.00	.00	.00	.00	.00	.01	.12	.00
	5	.356	4.202	.00	.39	.00	.01	.00	.01	.01	.00	.01	.00
	6	.165	6.173	.00	.00	.00	.55	.00	.01	.00	.01	.01	.00
	7	.015	20.163	.00	.17	.03	.01	.26	.24	.14	.17	.19	.31
	8	.011	23.609	.00	.20	.03	.36	.20	.36	.33	.42	.24	.50
	9	.008	27.419	.04	.00	.24	.00	.50	.35	.40	.34	.34	.05
	10	.002	51.958	.96	.22	.70	.06	.04	.01	.07	.03	.03	.14

a. Dependent Variable: Success or failure

APPENDIX C

ANALYSIS OF ALL CASES

Dependent Variable Encoding

Original Value	Internal Value
failure	0
success	1

Classification Table[a,b]

			Predicted		
			Success or failure		
	Observed		failure	success	Percentage Correct
Step 0	Success or failure	failure	39	0	100.0
		success	22	0	.0
	Overall Percentage				63.9

a. Constant is included in the model.

b. The cut value is .500

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	9.908	1	.002
	Block	9.908	1	.002
	Model	9.908	1	.002

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	69.855[a]	.150	.206

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			Success or failure		
	Observed		failure	success	Percentage Correct
Step 1	Success or failure	failure	33	6	84.6
		success	12	10	45.5
	Overall Percentage				70.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SAT_M	.012	.004	7.477	1	.006	1.012
	Constant	-6.946	2.382	8.506	1	.004	.001

a. Variable(s) entered on step 1: SAT_M.

Correlation Matrix

		Constant	SAT_M
Step 1	Constant	1.000	-.993
	SAT_M	-.993	1.000

Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	SAT_M	-39.881	9.908	1	.002

Variables not in the Equation

			Score	df	Sig.
Step 1	Variables	white	.965	1	.326
		afamer	.806	1	.369
		hispan	1.498	1	.221
		asiapac	.103	1	.749
		AP_GPA_0	2.629	1	.105
		RANK_PERCENT_0	.043	1	.835
		SP_TOTAL_HOURS_1	1.095	1	.295
		sex	.017	1	.897
	Overall Statistics		7.929	8	.440

APPENDIX D

CROSS VALIDATION

Classification Table[a,b]

Observed			Predicted: Selected Cases[c] Success or failure		Selected Cases[c] Percentage Correct	Unselected Cases[d,e] Success or failure		Unselected Cases[d,e] Percentage Correct
			failure	success		failure	success	
Step 0	Success or failure	failure	25	0	100.0	14	0	100.0
		success	18	0	.0	4	0	.0
	Overall Percentage				58.1			77.8

a. Constant is included in the model.

b. The cut value is .500

c. Selected cases selection EQ 1

d. Unselected cases selection NE 1

e. Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.329	.309	1.129	1	.288	.720

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	white	2.965	1	.085
		afamer	.184	1	.668
		hispan	.688	1	.407
		asiapac	.096	1	.756
		AP_GPA_0	4.399	1	.036
		RANK_PERCENT_0	1.649	1	.199
		SAT_M	6.928	1	.008
		SP_TOTAL_HOURS_1	.714	1	.398
		sex	.057	1	.811
	Overall Statistics		11.163	9	.265

CROSS VALIDATION (CONT.)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	7.710	1	.005
	Block	7.710	1	.005
	Model	7.710	1	.005

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	50.756[a]	.164	.221

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c,d]		
			Success or failure		Percentage Correct	Success or failure		Percentage Correct
	Observed		failure	success		failure	success	
Step 1	Success or failure	failure	21	4	84.0	10	4	71.4
		success	9	9	50.0	1	3	75.0
	Overall Percentage				69.8			72.2

a. The cut value is .500

b. Selected cases selection EQ 1

c. Unselected cases selection NE 1

d. Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SAT_M	.012	.005	5.632	1	.018	1.012
	Constant	-6.939	2.830	6.012	1	.014	.001

a. Variable(s) entered on step 1: SAT_M.

Correlation Matrix

		Constant	SAT_M
Step 1	Constant	1.000	-.993
	SAT_M	-.993	1.000

Model if Term Removed

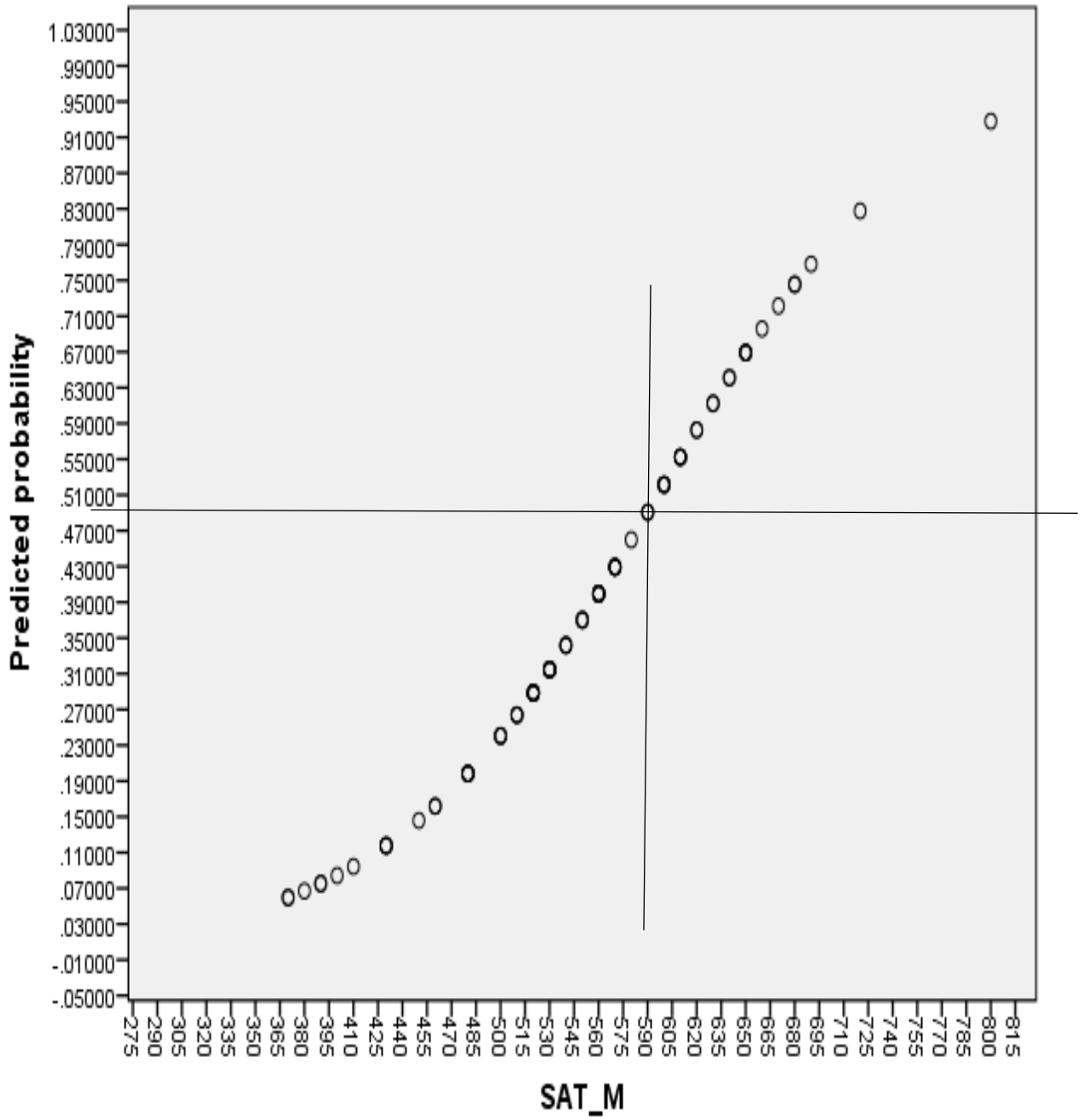
Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	SAT_M	-29.233	7.710	1	.005

Variables not in the Equation

			Score	df	Sig.
Step 1	Variables	white	.758	1	.384
		afamer	2.404	1	.121
		hispan	1.796	1	.180
		asiapac	.318	1	.573
		AP_GPA_0	.806	1	.369
		RANK_PERCENT_0	.182	1	.670
		SP_TOTAL_HOURS_1	.040	1	.841
		sex	.013	1	.909
	Overall Statistics		5.221	8	.734

APPENDIX E

GRAPH OF PREDICTED PROBABILITY


Predicted probability
1.03000
.99000
.95000
.91000
.87000
.83000
.79000
.75000
.71000
.67000
.63000
.59000
.55000
.51000
.47000
.43000
.39000
.35000
.31000
.27000
.23000
.19000
.15000
.11000
.07000
.03000
-.01000
-.05000
275 290 305 320 335 350 365 380 395 410 425 440 455 470 485 500 515 530 545 560 575 590 605 620 635 650 665 680 695 710 725 740 755 770 785 800 815
SAT_M