University of Texas Rio Grande Valley
ScholarWorks @ UTRGV

Theses and Dissertations - UTB/UTPA

8-2014

Prediction of Time-to-Graduation for STEM Hispanic Undergraduate Students

Gejun Zhu
University of Texas-Pan American

Follow this and additional works at: https://scholarworks.utrgv.edu/leg_etd

 Part of the Applied Mathematics Commons, and the Higher Education Commons

Recommended Citation

Zhu, Gejun, "Prediction of Time-to-Graduation for STEM Hispanic Undergraduate Students" (2014). *Theses and Dissertations - UTB/UTPA*. 948.
https://scholarworks.utrgv.edu/leg_etd/948

This Thesis is brought to you for free and open access by ScholarWorks @ UTRGV. It has been accepted for inclusion in Theses and Dissertations - UTB/UTPA by an authorized administrator of ScholarWorks @ UTRGV. For more information, please contact justin.white@utrgv.edu, william.flores01@utrgv.edu.

PREDICTION OF TIME-TO-GRADUATION FOR STEM HISPANIC UNDERGRADUATE STUDENTS

A Thesis

by

GEJUN ZHU

Submitted to the Graduate School of
The University of Texas-Pan American
In partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

August 2014

Major Subject: Mathematics

PREDICTION OF TIME-TO-GRADUATION FOR STEM HISPANIC UNDERGRADUATE STUDENTS

A Thesis
by
GEJUN ZHU

COMMITTEE MEMBERS

Dr. Xiaohui Wang
Chair of Committee

Dr. George Yanev
Committee Member

Dr. Maria Cristina Villalobos
Committee Member

August 2014

Copyright 2014 Gejun Zhu
All Rights Reserved

ABSTRACT

Zhu, Gejun, Prediction of Time-to-graduation for STEM Hispanic Undergraduate Students. Master of Science (MS), August, 2014, 42 pages, 30 tables, 8 figures, 41 references, 41 titles.

In this thesis, we study the time-to-graduation problem for STEM Hispanic undergraduate students. The response, time-to-graduation, was treated in two different ways: as a binary variable with graduated (by the 6th year) and not-graduated values, and as an ordinal variable with values year-4, year-5, year-6, and not-graduate. Mathematics education plays critical role in students' timely graduation, especially for STEM students. We used students records data obtained from The University of Texas-Pan American to illustrate how mathematics background factors (including SAT math score, ACT math score, TASP math score) and mathematics performance variables (including mathematics GPA, number of dropped mathematics courses, number of repeated math courses), and some demographic factors (including gender, full-/part-time study status) are related to time-to-graduation. Logistic regression and backwards subset selection were employed to determine the significant variables and then form the model for the prediction of timely graduation. In addition, random forests method was used to predict the ordinal outcome, the year when the students graduated.

Keywords: Mathematics education; Logistic regression; Timely graduation; Backwards subset selection; Decision tree methods

DEDICATION

The completion of my master's studies would not have been possible without the love and support of my family. I dedicate this thesis to my parents who have been supporting me behind. They are a source of encouragement and inspiration to me throughout my life. It is their unconditional love that motivates me to set higher targets.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Xiaohui Wang for her patient guidance and tremendous help in making this work possible. I would like to thank my thesis committee members: Dr. Gorge Yanev and Dr. Maria Cristina Villalobos for their kind service and constructive comments on this thesis. I am also grateful to the Department of Mathematics of The University of Texas-Pan American for providing me with a generous teaching assistantship, which greatly supported my past two-year studies and research work.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

The disparity in achievement between Hispanic students and their Anglo classmates, especially in mathematics, continues to be prevalent in post-secondary education. One possible reason is that a large portion of Hispanic students enter universities unprepared, especially in mathematics. Seels (1980) likened mathematics to an "invisible filter" and concluded that students who do not acquire a sufficient background in mathematics prior to university virtually are eliminated from an inordinate number of careers ranging from those in the physical sciences and engineering to those in the social sciences and psychology. Beverly Anderson, the former Director of Minority Programs for the Mathematical Sciences Education Board of National Research Council stated: "Mathematics is an enabling force and a critical filter. Minorities-especially Blacks, Hispanics, and Native Americans-must be helped to see, and thereby be convinced, that mathematics education offers essential access to careers in a technologically, powerful society. They must be encouraged to view themselves positively as learners and, especially as learners of mathematics" (Anderson 1990). University graduation, especially timely graduation, is not only one of the institutional effectiveness measures but also an increasingly important policy issue (Hebel 1999). The timely university graduation rates in Hispanic population have drawn attention from legislatures, administration as well as researchers.

Moreover, one of the significant facts in the higher education institutes is the explosive growth of educational data. These data are increasing rapidly without any benefit to the improvement of education if they are just left out there. The main objective of higher education institutes is to provide high quality education to its students and to improve the quality of managerial deci-

sions. We believe that to manage this difficult task, it is required to apply new techniques and tools to process the large amount of generated data and extract some useful knowledge and information. One way to achieve highest level in higher education system is to study the main attributes that may affect the students' performance. The discovered knowledge can be used to offer helpful and constructive recommendations to the academic planners in higher education institutes to enhance their decision making process, to improve students' performance and trim down failure rate, to better understand students' behavior, to assist instructors to improve teaching and many other benefits.

It is of interest to find out what factors determined or contributed to timely graduation - the success in university education. In the literature of studying what possible determinants of success could be, researchers considered factors such as student demographic characteristics (Thomas, 1981), pre-preparation of students (DeBrock, Hendricks, and Koenker, 1996), financial aid (Bean and Vesper, 1990, Cabrera, Stampen, and Hanse, 1990, McPherson and Schapiro, 1998), different colleges (Des Jardins, Ahlburg and McCall, 1999, 2002, Velez, 1985), and attitudinal variables (DesJardins, Ahlburg and McCall 2002). Velez (1985) grouped factors affecting college completion into five categories: personal background, academic process, psychosocial process, institutional factors, and institutional integration. Personal background variables include parents' socioeconomic status and the student's gender, race and religion. Academic processes include high school curriculum enrollment, number of mathematics courses taken in high school, and high school and college grades. Psychological processes include the student's educational aspirations and the student's perception of mother's educational aspirations for her/him. Institutional factors include type of control and college type. Institutional integration is measured by two variables, living quarters and participation in a work-study program. Among those five categories, academic process is the only one on which university education can offer assistance through advisory, instructional, or remediation interventions. Our society seems to understand that mathematics education is the critical and irreplaceable component of academic processes in university education. However, we have not seen much study on relating mathematics academic performance in university to timely graduation.

By discovering the successful patterns of students in various categories, the university can predict the graduation status of each student and which, in turn, helps to identify students at risk early and allow the instructor to provide appropriate advice in a timely manner. Moreover, the undergraduate advisor can recommend some courses based on the recommendation system if some students want to graduate in a certain year successfully. One report from *The Chronicle of Higher Education* stated, "graduation rates can be predicted more precisely by examining students characteristics." University of California-Los Angeles (UCLA) have developed an online application, named Expected Graduation Rate Calculator, which can be used to predict expected degree completion figures for a single student or an entire cohort of students at a college or university. The prediction results (figures) are based on sex, SAT/ACT score, Race/Ethnicity, and High School Grade. The methodology applied in this calculator is detailed in the HERI monograph Completing College: Assessing Graduation Rate at Four-Year Institutions.

Yingkuachat et al.(2007) used Bayesian Belief Networks (BNN) model to determine the important variables for prediction of the education accomplishment and the relationship between them. A prediction model was constructed based on k-fold cross validation by WEKA software. They discovered the important variables affecting the graduation status were: previous GPA, mother and/or fathers career, the total income of the family, and first-year GPA.

Zwich et al.(2005) applied linear regression models to predict first year GPA (FGPA) using high shool grades and SAT scores, and assessed predictive effectiveness by measuring the degree of correspondence between predicted and observed FGPAs. Also, they used survival models to determine how well SAT scores, high school grades, and language minority status predicted degree attainment and whether the effects of SAT scores and high school grades on degree attainment differed among the four language/ethnicity groups assessed. They found that high score GPA and SAT score were found to jointly explain 22% of the variance in FGPA, and higher high school grades and SAT scores were associated with a higher probability of graduation, a finding that supports the use of these criteria in college admissions.

Moucary et al.(2011) presented a hybrid procedure based on Neural Network (NN) and data

clustering that enables academicians to predict their foreign language performance at a first stage, then classify the student in a well-defined cluster for further advising and follow up by forming a new system entry.

Sembiring et al.(2011) applied the kernel method to analyze the relationships between students' behavior and their success. They predicted the students' final grade by using Smooth Support Vector Machine (SSVM) classification and grouped the students according to their similar characteristics by kernel k-Means algorithms technique. The results of this study indicated that there existed strong correlation between mental condition of student and their final academic performance.

Tair et al.(2012) studied a case of graduate students from the college of Science and Technology - Khanyounis from 1993 to 2007. The variables they selected in the models were gender, speciality, city, Matriculation GPA, secondary school type. Association rules and Naïve Bayesian classifier were applied to predict the grades of the graduate students. Also, they clustered the students into groups using K-Means clustering algorithm. Finally, they used two outlier detection methods - Distance-based and Density-based - to detect all the outliers in the data and showed how we can benefit from the discovered knowledge to improve the performance of student.

Imam Tahyudin (2013) compared several data mining classification algorithms, especially the Decision Tree (DT), Naive Bayes (NB), Artificial Neural Network (ANN), Support Vector Machine (SVM) and Logistic Regression (LR) algorithms with cross validation evaluation and T -Test to predict the graduation student on time. They found SVM algorithm is the appropriate algorithm that used to predict the student graduation on time based on the comparison of performance score and t-test.

George et al.(1994) employed logistic regression to predict success in an engineering program in 1994. Logistic regression is employed due to the binary nature of the response variable (whether or not graduation occurs after 5 years). Multiple logistic regression on the grades obtained in various courses provides little additional predictive power over a simple logistic regression on the early term's average grade. The results appear to support the faculty's choice of minimum

average term grade for promotion to the next term in the programme.

Brijesh Kumar Bhardwaj et al.(2011) employed Bayes Naïve Classification algorithm to predict the performances of 300 students in the course BCA (Bachelor of Computer Application). This study will help the students and the teachers to improve the division of the student. It will also work to identify those students which needed special attention to reduce failing ration and taking appropriate action at the right time.

Dmitri Rogulkin (2011) used decision tree model to find the most influential factors in predicting graduation status for students in the fall 2004 cohort at California State University, Fresno. The results showed that staying on track, which was defined as reaching sophomore level no later than the end of the third semester, and cumulative GPA after the first year were the most influential factors in predicting six year graduation. The data mining clustering algorithm was also used to identify profiles of students who were at-risk of not graduating, those who had high odds of graduating, and those who might potentially benefit from some sort of support to improve the likelihood of graduating. Brijesh Kumar Baradwaj, Saurabh Pal (2012) also employed IID3, which is a simple decision tree, to predict the students' performance in the end of semester based on the previous database.

Table 1.1: Graduation Rates of UTPA from 2000 to 2006

Year	4th Year Graduation Rate	6th Year Graduation Rate
2000	5.8%	23.5%
2001	6.2%	22.9%
2002	7.7%	24.5%
2003	8.4%	26.2%
2004	10.2%	26.7%
2005	9.6%	30.0%
2006	13.2%	32.4%

We obtained the following graduation rates reports (see Table 1.1) from the Office of In-

stitutional Research and Effectiveness of UTPA. The reports showed that the graduation rates at 4th year and 6th year is not high at UTPA. It would be helpful if we can extract some useful information from the graduation data.

In this thesis, we used time-to-graduation as the response variable to measure students' success in university education. The population of interest is Hispanic undergraduate students. We investigated factors including gender, math scores from SAT, ACT, and TASP tests, full-/part-time study status, mathematics performance (measured by cumulative mathematics grade points average). We focused on two main research questions in this thesis. One is to predict whether the students can graduate within 6 years. The other is to predict which year the student would graduate based on the selected factors.

The rest of the present thesis is organized as follows: Chapter 2 presents theories and models we used in the model building; Chapter 3 describes the data set and the preparation methods performed; Chapter 4 reports the results of statistical learning techniques on the educational data; finally we conclude this paper with discussion and an outlook for future in Chapter 5.

CHAPTER II

THEORIES AND MODELS

In this thesis, we are going to predict whether the students will graduate successfully or not, and at which year they will graduate. Logistic regression is a good choice to predict whether they can graduate since the response variable is a binary variable. However, logistic regression can only be applied to binary response variable. To predict the students graduation years, we would like to employ the decision tree method. Here comes the basic idea behind the decision tree method. Let's say the possible values for graduation status are $\mathscr{Y} \in \{4,5,6,0\}$, where 4, 5, 6 denote that they can graduate at 4th-year, 5th-year, 6th-year respectively, and 0 denotes the student cannot graduate. One approach is to build several layers of predictive models to graduation status of different years. The idea would be that we would predict the graduation status at their 4th year. If the student is predicted to be graduated successfully, we would output graduate at 4th year as the final predicted result. Otherwise, we would run the observations through another algorithm to predict the graduation status for 5th year. We would repeat this process for other layers.

2.1 Logistic Regression

Logistic regression is a statistical analysis procedure commonly used for the analysis and prediction of research problems where there is a binary outcome. If we let Y be a binary response, say $y \in \{0,\ 1\}$, then logistic regression is to choose the Bernoulli family to model the conditional distribution of Y given X, $Pr(Y = 1|X = x)$ as a function of x. In other words, logistic regression is an approach to learning functions of the form $f:\ X \to Y$, or $P(Y|X)$ in the case where Y is discrete-valued, and $X = (X_1, ..., X_n)$ is any vector containing discrete or continuous variables. The value of $Pr(Y = 1|X)$, which we abbreviate as $p(X)$, will range from 0 and 1. To satisfy this condition,

we choose the *logistic function*,

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

Then, we have

$$e^{\beta_0+\beta_1 X} = \frac{p(X)}{1-P(X)}.$$

By taking logarithm of both sides, we arrive at:

$$logit(p) = log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0+\beta_1 X.$$

The quantity $p(X)/[1-p(X)]$ is called the *odds*, and can take on any value between 0 and ∞. This is the case for one variable problem. The left-hand side is called *logit* or *log-odds*. We can generalize the simple logistic regression to multiple regression as follows:

$$log\left(\frac{p(\mathbf{X})}{1-p(\mathbf{X})}\right) = \beta_0+\beta_1 X_1+\cdots+\beta_p X_p,$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ are p predictors. It can be rewritten as

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

Logistic regression applies maximum likelihood estimation after transformation of the dependent variable (graduation status) into logit variable (the natural log of the odds of the dependent response occurring or not); therefore, logistic regression will estimate the odds that an existing student graduated or not graduated. We will predict that the student graduate successfully if $p(x) \geq threshold$, not graduate otherwise.

To fit the logistics model, we minimize the Residual Sum of Squares (RSS)

$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2,$$

where y_i is the actual value, and $f(x_i)$ is the estimated value for predictor.

Logistic regression is used to predict the odds of being a case based on the values of the independent variables (predictors). The odds are defined as the probability that a particular outcome is a case divided by the probability that it is a non-cases.

2.2 Backwards Stepwise Selection

Adjusting the variables included in a regression model is a form of model tuning. Principle of simplicity puts that a model with fewer variables is preferable from the original model. In other words, simpler models are more interpretable and generalizable. Too many variables in a model might lead to over-fitting, while too few variables might lead to under-fitting. Therefore, it is necessary and utmost important to choose the most optimal variables for high quality model. With a large set of variables, we often would like to determine a smaller subset that exhibit the strongest effect. In order to get a "big picture", we are willing to sacrifice some of the small details.

The method we used to select variables is backwards stepwise selection. The backwards stepwise selection begins with full least squares model containing all p predictors, and then iteratively removes the variable with the largest p-value – that is, the variable that is the least statistically significant. The new $(p-1)$-variable model is fit, and the variable with the largest p-value is removed. This procedure will continue until a stopped rule is reached. For instance, we may stop when all remaining variables have a p-value below some threshold.

2.3 Decision Tree Method

Decision tree method is one of the most widely used and practical methods for inductive inference. It is a method for approximating discrete-valued functions, in which the learned function is represented by a decision tree. Learned trees can also be re-represented as sets of if-then rules to improve human readability. A decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. The topmost node in a tree is the root node. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of instance, and each branch descending from that node corresponding to one of possible values for this attribute. An instance is classified by starting at the root of the tree, testing the attribute specified by this node, then moving down the branch corresponding to the value of the attribute in the given example.This process is then repeated for the sub-tree rooted at the new

node. Classification and Regression Trees (CART) is a very popular non-parametric decision tree learning technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively.

During the tree construction, attribute selection measures are used to select the attribute that best partitions the data into distinct classes. The Gini index in used in CART. The Gini index measures the impurity of D, a data partition or set of training tuples, as

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2,$$

where p_i is the probability that a tuple in D belongs to class C_i and is estimated by $|C_{i,D}|/|D|$. The sum is computed over m classes. When considering a binary split, we compute a weighted sum of impurity of each resulting partition. For example, if a binary split on A partition D into D_1 and D_2, the Gini index of D is given that partitioning is

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2).$$

For each attribute, each of the possible binary splits is considered. For a discrete-valued attribute, the subset that gives minimum Gini index for that attribute is selected as its splitting subset.

After building the decision tree, many of the branches may reflect noise or outlier in the training data. Tree pruning attempts to identify and remove such branches, with the goal of improving classification accuracy on unseen data. The advanced decision tree method is called random forests. Random forests are an ensemble learning method for classification that operate by constructing a multitude of decision trees at training time and output the class that is the mode of the classes output by individual trees.

2.4 Model Evaluation Techniques

In practice, a binary classifier such as logistic regression can make two types of errors: it can incorrectly assign an student who graduated to the category not graduated, or it can incorrectly

Table 2.1: Receiver Operating Characteristics (ROC) Table

	Actual Graduation	
Prediction Graduation	No	Yes
No	True Negative (TN)	False Negative (FN)
Yes	False Positive (FP)	True Positive (TP)

assign an student who did not graduate into graduate category. We are interested in determining which of these two types of errors are being made. A confusion matrix, shown as table 2.1, is a convenient way to display this information.

Class-specific perform is important in medicine and biology, where the terms *sensitivity* and *specificity* characterize the performance of a classifier or screen testing. The overall accuracy rate measure gives the proportion of overall classification accuracy, including true positive and true negative. While overall error rate measures the proportion of the inaccurate ones including false negative and false positive. Sensitivity measures the proportion of actual positives which are correctly identified as such. In other words, the sensitivity measure gives the evaluation of the probability that a given statistic correctly predicts the correct existing with respect to the threshold. While specificity measures the proportion of negatives which are correctly identified as such. N is the number of observations.

$$Overall\ accuracy = \frac{TN+TP}{N},\ Overall\ error\ rate = \frac{FN+FP}{N},$$

$$Sensitivity = \frac{TP}{TP+FN},\ Specificity = \frac{TN}{TN+FP}.$$

Many binary classification algorithms compute a sort of classification score (sometimes but not always this is a probability of being in the target state), and they classify based upon whether or not the score is above a certain threshold. Different classification objectives might make one point on the curve more suitable for one task and another more suitable for a different

task. But how can we decide which threshold is the best? The *ROC curve* , where ROC stands for *receiver operating characteristics* is a popular graphic for simultaneously displaying the two types of errors for all possible thresholds. Viewing the ROC curve lets you see the tradeoff between sensitivity and specificity for all possible thresholds rather than just the one that was chosen by the modelling technique. The overall performance of a classifier, summarized over all possible thresholds, is given by the *area under the ROC curve* (AUC). If AUC equals 0.5, that means we obtain the prediction results by random guessing. The ideal ROC curve will hug the top left corner, so the larger the AUC the better the classifier. We employed these four indicators to evaluate our logistic regression model. As for decision tree model, we only calculated overall accuracy rate to evaluate it.

CHAPTER III

DATA COLLECTION AND PROCESSING

3.1 Data Collection

To conduct this study, we obtained students records data from The University of Texas-Pan American (UTPA), a comprehensive public coeducational institution located in Edinburg, Texas. More than 85% of the undergraduate students at this university are Hispanic students. Students records were obtained for first-time Freshmen students entering UTPA from the Fall semester of 2000 until the Fall semester of 2006. For each of those students, we collected operational data for the following variables: (1) gender; (2) ethnicity (we are only interested in Hispanic students, so we only select those whose ethnicity is Hispanic); (3) claimed college when entering UTPA; (4) full-/part-time status of the students when entering; (5) cumulative mathematics grade points average at the end of 4th year, 5th year, 6th year calculated based on all up-to-date mathematics courses (i.e. courses offered by the Department of Mathematics) taken by the students (if the student has taken the same courses for several times , then we collected the data for all of the courses taken by the student); (6) cumulative grade points average that is calculated by the university, either at the time of graduation or at the time of collecting data; (7) time-to-graduation is split into four groups, denoted by "graduated at 4th year ", "graduated at 5th year", "graduated at 6th year", and "not graduate" respectively. Table 3.1 shows details of all variables.

For math scores, only 98 out of 414 students had SAT, ACT, and TASP scores at the same time. Our idea was to find the percentile of these three courses respectively for each student and combine these three variables into one variable, named Entering Math Score. If the student only had the ACT percentile, then we used his/her ACT percentile as the result of Entering Math Score.

Table 3.1: List of Variables

Name	Type	Description
Gender	Binary	Student Gender
Math scores of SAT, ACT, THEA	Numerical	Scores of math testes
Mathematics Grade Point Average (MGPA)	Numerical	GPA for mathematics courses
Full-/Part-time status	Binary	Student status
Time-to-graduation	Categorical	Graduation status

If the student had more than one score, then we used the average percentile instead.

3.2 Data Processing for Math Courses

For math courses, some students took them once and they obtained a valid and satisfied grade and they didn't need to retake it; some took it once but didn't obtain a valid grade and they don't care; some took it once but didn't obtain a valid grade and decided to retake it. Therefore, there are four patterns for students taking math courses:

- took once, and obtained a valid grade;
- took once, but did not obtain any valid grades;
- repeated at least one time, and obtained a valid grade;
- repeated at least one time, but finally did not obtain any valid grades;

Table 3.2 shows the rules for us to calculate MGPA. For example, if the student obtained an A for that math course, then he/she can get 4 points for each credit hour. Grades such as DP, WP, DR will not be included in the calculation. Generally, to calculate students' GPA, multiply course hours by the points earned for the grades, and then divide by total course hours. However, there are two things which are different from usual GPA calculation: 1) we still count repeated courses, but UTPA just counts the last attempt; 2) we count Math1300, Math1334, while UTPA will not count those two developmental courses.

Table 3.2: Rules for Calculating MGPA

Grade	Points per Credit Hour
A	4
B	3
C	2
D	1
F	0
DF, WF	0

Moreover, we calculated the repeated times and drop times after each academic year because we believe that there must be some relationship between these two factors and graduation status. The more the student repeated some courses, the lower probability the students would graduate in time.

3.3 Data Processing for Graduation Data

We combined the math GPA data set with graduation data set by student ID (faked) after selecting all the Hispanic students. Since we are interested in graduation rate at 4th year and 6th year, we created two response variables. One is for graduate or not graduate during 6 years, and the other is for graduation year. In summary, the independent variables we have for prediction are: ACT math scores, SAT math scores, TASP math scores, gender, full-/part-time, math GPA for 6 years, repeat times for math courses after each academic year, drop times for math courses after each academic year. Finally, we split the data set into two sets, one of the data sets would be used for training the model, the other for testing the model. Generally, the split ratio is 6:4, which means 60% for training model and 40% for testing model. The model that fits the training data set perfectly doesn't necessarily mean that it would work well on the testing data set. This is why we separate the data set into two parts.

CHAPTER IV

RESULTS

In this thesis, we chose to consider (1) whether the students would graduate within 6 years; (2) if they can graduate, which year they will graduate. Therefore, the result would be divided into two parts: graduate or not, and which year they will graduate and the corresponding probabilities. We first used exploratory data analysis technique to take a look at the overall graduate rate for STEM undergraduate students.

4.1 Exploratory Data Analysis

Exploratory data analysis was used to inspect the student records using graphical charts and descriptive statistics. The data exploration actions includes visual techniques that examined the data set in terms of summary statistics with respect to student graduation.

Table 4.1 showed the graduation rates for the STEM undergraduate students enrolled from Fall semester 2000 at The University of Texas-Pan American. The dataset included 414 STEM Hispanic Undergraduate students, 44 (10.6%) students graduated at 4th year, 49 (11.8%) students

Table 4.1: Graduation Rates for STEM Undergraduate Students Enrolled Fall 2000

Year	Number	Graduation Rate
2004	44	10.6%
2005	49	11.8%
2006	49	11.8%
Total	142	34.2%

at 5th year, and 49 (11.8%) students at 6th year, 272 (65.7%) students did not graduate.

4.2 Results on Math Courses

For this part, we considered the math courses taken by students at College of Science from Fall 2000 to Fall 2006. First, we obtained the overall frequency of all math courses regardless of the repeat times. Here, if one student took one course for several times, we only count it as one time. There are 42 math courses offered by Department of Mathematics at UTPA. Math1340 (College Algebra) was taken by 333 students in total, Math1334 (Intermediate Algebra) by 269 students, Math1356 (Trigonometry) by 227 students, Math1401 (Calculus I) by 201 students etc. Table 4.2 showed the overall frequency of all math courses students have taken.

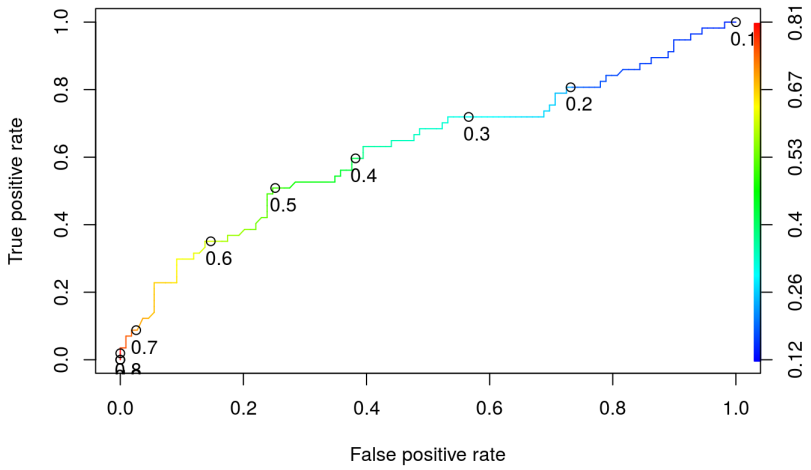
After obtaining this, we calculated the frequency of students who repeated math courses for once, twice, and more than twice. For the courses students repeated once, Math1334 (39 times), Math1356 (38 times), Math1340 (35 times), Math1401 (32 times), Math1357 (Pre-calculus, 30 times) were the top courses. For the courses students repeated twice, Math1334 (17 times), Math1356 (13 times), Math1401 (9 times), Math1357 (8 times) have the highest frequency. Math1356 (11 times), Math1334 (7 times), Math1340 (5 times), Math1402 (5 times) were top courses repeated more than twice by students.

4.3 Graduation Prediction When Entering

In this section, we employed logistic regression to construct the model because the predictor variable is binary. In the logistic regression algorithm, a threshold needed to be chosen to classify the outcomes. If the probability is larger than the threshold, then it would be classified as 1, otherwise it would be classified as 0. Typically, the threshold is 0.5. But for our model, we applied ROC curve to select an appropriate threshold.

In the beginning of their undergraduate studies, our goal was to classify whether the students could graduate within 6 academic years or not. Hence, the students who graduated, whatever the year was, will be classified as graduate successfully. The ones left were classified as not graduated. All the demographic factors (including gender, entering math score, full-/part-time student)

Figure 4.1: ROC Curve for Graduation Prediction When Entering UTPA



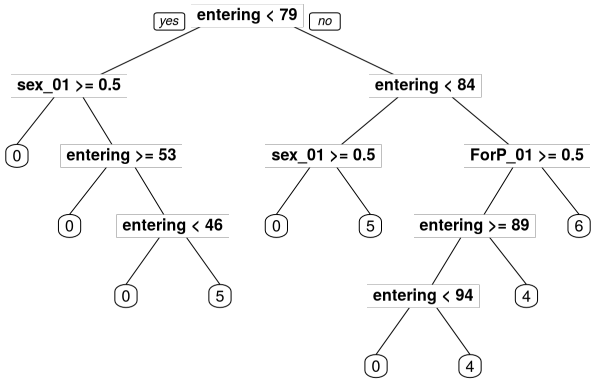
were used to construct the model. See the results from Table 4.3.

Backwards stepwise selection method was used to select significant variables. We removed one most insignificant variable which has the highest p-value each time. For example, we removed Full-/Part-time variable for the first time, and then we rebuilt the model based on the same data set. We repeated this procedure for several times until we obtained all the significant variables. Finally, we obtained the following significant variables: Entering Math Score and Gender. Then we used the selected variables to construct the model with the training data set.

From Figure 4.1 we see that the best threshold should locate around 0.6 because our goal is to predict those who cannot graduate precisely and at this point the true positive rate and the false positive rate are not very high. The false positive means that our algorithm predicted that the student can graduate while she/he cannot graduate actually. If the value of false positive rate is very high, the model would not be a good one since our goal is to warn those who might not graduate successfully within 6 years based on the model we built. After obtaining the significant variables and threshold, we used the model to make graduation predictions on the testing data set to evaluate it.

Table 4.4 is the confusion matrix/classification table with predicted graduation as the row

Figure 4.2: Tree Methods for Graduation Prediction When Entering UTPA



variable and actual graduation as the column variable. 64% students were predicted as not graduate while they actually didn't graduated and 4% students were predicted as graduate while they graduated successfully. The misclassification rate was calculated to evaluate the overall model performance with respect to the exact number of classification in the entire testing data. The diagonal numbers indicate the accurate classifications and the off diagonal elements show the inaccurate classifications. The overall accuracy of the test result is 0.67. The sensitivity of the test result is 0.11, and the specificity is 0.97. From Figure 4.1, we found the Area Under Curve (AUC) is 0.65.

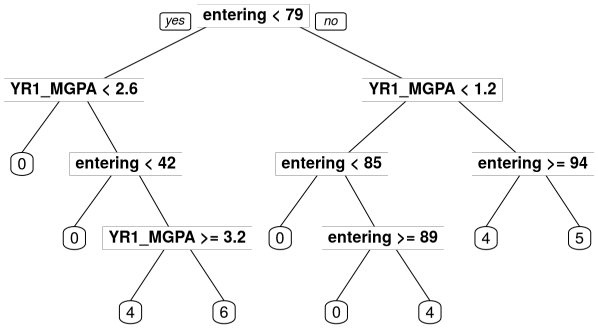
Next, we applied decision tree method to predict which year the student would graduate. Figure 4.2 showed the visualized decision tree in graduation prediction. Only the significant variable were selected and they were selected automatically.

Finally, we used random forests to predict the year student would graduate. Table 4.5 showed the graduation year prediction results. To calculate the accuracy rate of classification, we use the diagonal numbers of confusion matrix to divide by the size of testing data set. The accuracy rate for this prediction is 0.66.

4.4 Graduation Prediction at Freshmen

At this time, some students have taken some math courses. We considered the GPA of all math courses students took during the first year studies as an independent variable. If the

Figure 4.3: Tree Methods for Graduation Prediction at Freshmen

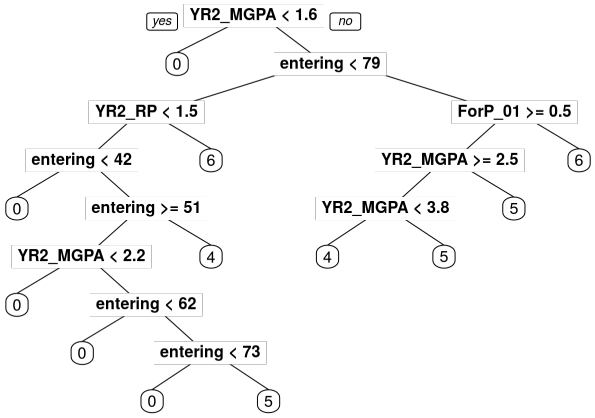


students didn't take any math course, the math GPA for him/her would be 0. The times students dropped the math courses and repeated were considered as independent variables as well. Hence, the independent variables are: gender, entering math score, full-/part-time, math GPA at Freshmen, dropped times of math courses, and repeated times of math courses. To build the prediction model, we used the same training and testing data set so that we can compare the models with those we built before. Table 4.6 showed that Gender, Entering math score, and Math GPA were significant. Based on the selected variables, we rebuilt our model and make predictions on testing data set.

At this point, 62% of the students were predicted not graduate while they actually didn't graduate. 8% of the students were predicted graduate successfully and they actually graduated. The accuracy rate for this prediction is 0.70, which is better than the prediction when entering UTPA. The area under curve (AUC) of this model is 0.73, which is also better than the prediction when entering UTPA. The sensitivity of the test result is 0.33, and the specificity is 0.94.

Next, we used decision tree method to predict the year they will graduate. This tree is much more complex than the previous (see Figure 4.3 for detail). Finally, we used random forests to improve the accuracy of prediction. Table 4.8 showed the confusion matrix when predicting years the students would graduate. The overall accuracy rate of the prediction is 0.65.

Figure 4.4: Tree Methods for Graduation Prediction at Sophomore



4.5 Graduation Prediction at Sophomore

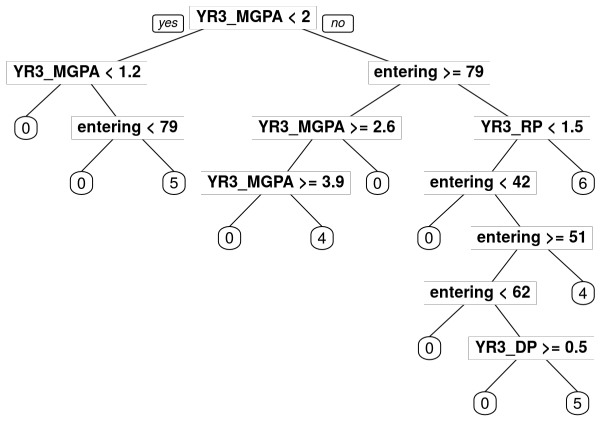
At this point, we included cumulated Math GPA, dropped time and repeated times for the past two years. Also, the times students dropped the math courses and repeated will be considered as independent variables. We repeated the steps as previous ones. First, we predicted whether the students can graduate or not.

Table 4.9 showed the significant variables in the fitting model. The significant variables were Entering math score, Gender, MGPA at Sophomore, Dropped times for math courses, Repeated times for math courses.

After finding the coefficients, we used the models to make predictions. Table 4.10 showed the confusion matrix predicted by the logistic regression model. The best threshold for the binary classification is 0.7 according to the ROC curve. The overall accuracy rate of this prediction is 0.72. The overall error rate is 0.27. The sensitivity of the test result is 0.23, and the specificity is 0.97. The Area Under Curve (AUC) is 0.77. Finally, we applied random forests model to predict the year students will graduate.

Table 4.11 showed confusion matrix for graduation prediction at Sophomore year. The overall accuracy rate for this prediction is 0.66.

Figure 4.5: Tree Methods for Graduation Prediction at Junior



4.6 Graduation Prediction at Junior

We repeated the steps as previous ones. First, we predicted whether the students can graduate or not. Table 4.12 showed the coefficients for logistic regression model, and the significant variables was only MGPA at Junior. We used the significant variables to make graduation prediction and we predicted 61% students as not graduate while they didn't graduated, 5% students as graduate while they graduate in fact.

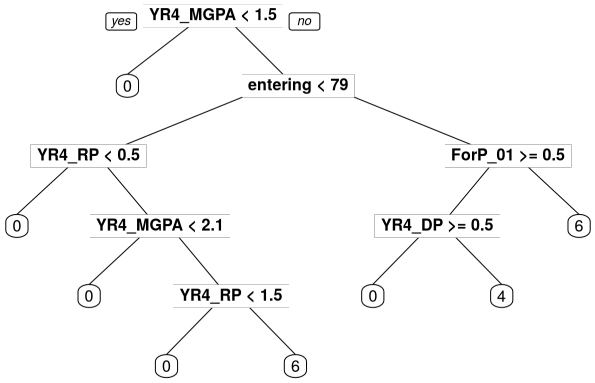
Table 4.13 showed the confusion matrix results for the graduation prediction. The best threshold for the binary classification is 0.7 according to the ROC curve. The overall accuracy rate of this prediction is 0.70. The overall error rate is 0.32. The sensitivity of the test result is 0.17, and the specificity is 0.94. The Area Under Curve (AUC) is 0.79. Finally, we used random forests to give graduation year prediction.

Table 4.14 showed confusion matrix for graduation prediction at Junior year. The overall accuracy rate for this prediction is 0.61.

4.7 Graduation Prediction at Senior

We repeated the steps as previous ones. First, we predicted whether the students can graduate or not. Table 4.15 showed the coefficients for logistic regression model, and the significant variables were Repeated times at Senior and MGPA at Senior. We used the significant variables to

Figure 4.6: Tree Methods for Graduation Prediction at Senior



make graduation prediction. We predicted 61% of the students didn't graduate while they didn't graduate in fact, 6% students would graduate while they actually graduated.

Table 4.16 showed the confusion matrix results for the graduation prediction. The best threshold for the binary classification is 0.7 according to the ROC curve. The area under curve (AUC) is 0.79. The overall accuracy rate of this prediction is 0.67. The overall error rate is 0.33. The sensitivity of the test result is 0.18, and the specificity is 0.93. Finally, we used random forests to give graduation year prediction.

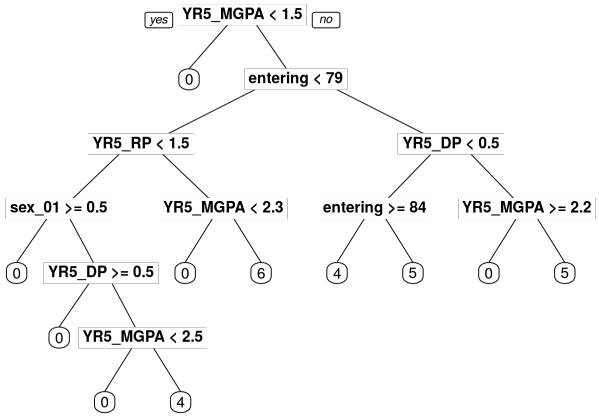
Table 4.17 showed confusion matrix for graduation prediction at Senior year. The overall accuracy rate for this prediction is 0.66.

4.8 Graduation Prediction at Fifth Year

We repeated the steps as previous ones. First, we predicted whether the students can graduate or not. Table 4.18 showed the coefficients for logistic regression model, and the significant variables were MGPA at 5th year and Entering math score. We used the significant variables to make graduation prediction and we predicted 60% of the students as not graduate while they actually didn't graduate, 8% students as graduate while they graduate in fact.

Table 4.19 showed the confusion matrix results for the graduation prediction. The best threshold for the binary classification is 0.7 according to the ROC curve. The AUC is 0.78. The

Figure 4.7: Tree Methods for Graduation Prediction at 5th year



overall accuracy rate of this prediction is 0.69. The overall error rate is 0.31. The sensitivity of the test result is 0.25, and the specificity is 0.92. Finally, we used random forests to give graduation year prediction.

Figure 4.7 showed decision tree fitted by the training data set.

Table 4.20 showed confusion matrix for graduation prediction at 5th year. The overall accuracy rate for this prediction is 0.65.

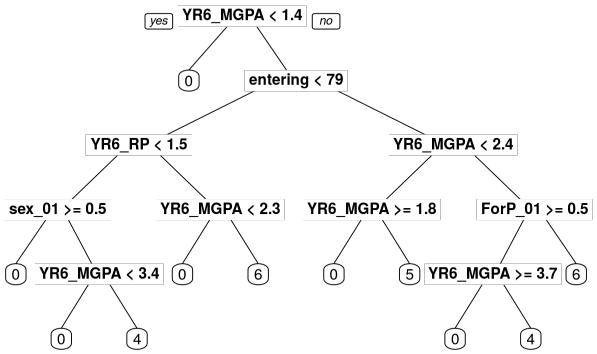
4.9 Graduation Prediction at Sixth Year

We repeated the steps as previous ones again. First, we predicted whether the students can graduate or not. Table 4.21 showed the coefficients for logistic regression model, and the significant variables were Entering math score and MGPA at 6th year. We used the significant variables to make graduation prediction and we predicted 103 students as not graduate while they actually didn't graduate, 14 students as graduate while they graduate.

Figure 4.8 showed decision tree fitted by the training data set.

Table 4.22 showed the confusion matrix results for the graduation prediction. The best threshold for the binary classification is 0.7 according to the ROC curve. The AUC is 0.78. The overall accuracy rate of this prediction is 0.68. The overall error rate is 0.32. The sensitivity of the test result is 0.19, and the specificity is 0.94. Finally, we used random forests to give graduation

Figure 4.8: Tree Methods for Graduation Prediction at 6th year



year prediction. Table 4.23 showed confusion matrix for graduation prediction at 5th year. The overall accuracy rate is 0.67.

In summary, we obtained the significant variables as shown in Table 4.24

Table 4.2: Overall Frequency of All Math Courses

	Course	Frequency		Course	Frequency
1	MATH1340	333	22	MATH1342	9
2	MATH1334	269	23	MATH2387	9
3	MATH1356	227	24	EMAT2307	8
4	MATH1401	201	25	MATH1341	7
5	MATH1357	195	26	MATH2306	6
6	STAT2330	163	27	MATH3337	6
7	MATH1402	136	28	MATH1460	5
8	MATH3349	108	29	MATH3303	5
9	MATH2401	100	30	MATH4319	4
10	MATH2345	48	31	MATH2307	3
11	MATH1300	47	32	MATH3368	3
12	MATH1321	40	33	MATH4302	2
13	MATH3373	33	34	MATH4318	2
14	MATH4339	30	35	MATH4360	2
15	MATH2346	24	36	MATH4379	2
16	MATH3311	19	37	MATH1470	1
17	MATH1322	14	38	MATH1487	1
18	MATH3304	14	39	MATH1488	1
19	MATH4351	14	40	MATH3355	1
20	MATH4357	14	41	MATH4364	1
21	EMAT2306	13			

Table 4.3: Coefficients for Logistic Regression Model When Entering

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	-1.7236	0.5996	-2.87	0.0040 **
Entering math score	0.0255	0.0065	3.92	0.0001 ***
Gender	-0.7293	0.2847	-2.56	0.0104 *
Full-/Part-time	-0.1323	0.4582	-0.29	0.7727

Signif. codes: '***' - 0.001, '**' - 0.01, '*' - 0.05, '.' - 0.1, ' ' - 1

Table 4.4: Confusion Matrix for Graduation Prediction When Entering UTPA

	Actual Graduation		
Predicted Graduation	No	Yes	Total
No	106	51	143
Yes	3	6	23

Table 4.5: Confusion Matrix for Graduation Year Prediction When Entering UTPA

			Actual Graduation Year		
Predicted Graduation Year	Not Graduate	4th year	5th year	6th year	Total
Not Graduate	109	15	18	24	166
4th year	0	0	0	0	0
5th year	0	0	0	0	0
6th year	0	0	0	0	0

Table 4.6: Coefficients for Logistic Regression Model at Freshmen

	Coefficient	Std. Error	z value	Pr($>|z|$)
(Intercept)	-1.8704	0.6267	-2.98	0.0028 **
Entering math score	0.0243	0.0070	3.46	0.0005 ***
Full-/Part-time	-0.2816	0.4785	-0.59	0.5561
Gender	-0.7127	0.3042	-2.34	0.0191 *
MGPA at Freshmen	0.4808	0.1096	4.39	0.0000 ***
Dropped times at Freshmen	-0.6379	0.4871	-1.31	0.1903
Repeated times at Freshmen	-0.7919	0.5543	-1.43	0.1531

Signif. codes: '***' - 0.001, '**' - 0.01, '*' - 0.05, '.' - 0.1, ' ' - 1

Table 4.7: Confusion Matrix for Graduation Prediction at Freshmen

	Actual Graduation		
Predicted Graduation	No	Yes	Total
No	103	43	146
Yes	6	14	20

Table 4.8: Confusion Matrix for Graduation Year Prediction at Freshmen

			Actual Graduation Year		
Predicted Graduation Year	Not Graduate	4th year	5th year	6th year	Total
Not Graduate	97	8	12	19	136
4th year	4	5	4	2	15
5th year	7	2	2	3	14
6th year	1	0	0	0	1

Table 4.9: Coefficients for Logistic Regression Model at Sophomore

	Coefficient	Std. Error	z value	Pr($>|z|$)
(Intercept)	-2.3955	0.5347	-4.48	0.0000 ***
Entering math score	0.0154	0.0078	1.99	0.0469 *
Gender	-0.7215	0.3187	-2.26	0.0236 *
MGPA at Sophomore	0.6655	0.1307	5.09	0.0000 ***
Repeated times for math courses	0.4387	0.1813	2.42	0.0155 *
Dropped times for math courses	-0.6866	0.3004	-2.29	0.0222 *

Signif. codes: '***' - 0.001, '**' - 0.01, '*' - 0.05, '.' - 0.1, ' ' - 1

Table 4.10: Confusion Matrix for Graduation Prediction at Sophomore

	Actual Graduation		
Predicted Graduation	No	Yes	Total
No	106	44	150
Yes	3	13	16

Table 4.11: Confusion Matrix for Graduation Year Prediction at Sophomore

			Actual Graduation Year		
Predicted Graduation Year	Not Graduate	4th year	5th year	6th year	Total
Not Graduate	98	4	11	18	131
4th year	3	8	5	5	19
5th year	7	4	1	0	12
6th year	1	0	1	1	3

Table 4.12: Coefficients for Logistic Regression Model at Junior

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	-2.7901	0.6830	-4.08	0.0000 ***
Entering math score	0.0142	0.0073	1.95	0.0517 .
Gender	-0.5105	0.3127	-1.63	0.1026
Full-/Part-time	0.0329	0.4983	0.07	0.9473
MGPA at Junior	0.7480	0.1360	5.50	0.0000 ***
Repeated times at Junior	0.2739	0.1520	1.80	0.0716 .
Dropped times at Junior	-0.2260	0.2219	-1.02	0.3084

Signif. codes: '***' - 0.001, '**' - 0.01, '*' - 0.05, '.' - 0.1, ' ' - 1

Table 4.13: Confusion Matrix for Graduation Prediction at Junior

	Actual Graduation		
Predicted Graduation	No	Yes	Total
No	102	43	145
Yes	7	14	21

Table 4.14: Confusion Matrix for Graduation Year Prediction at Junior

			Actual Graduation Year		
Predicted Graduation Year	Not Graduate	4th year	5th year	6th year	Total
Not Graduate	97	5	13	18	133
4th year	4	8	3	2	17
5th year	4	2	2	2	10
6th year	4	0	0	2	6

Table 4.15: Coefficients for Logistic Regression Model at Senior

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	-3.5011	0.7188	-4.87	0.0000 ***
Entering math score	0.0159	0.0073	2.19	0.0286 *
Gender	-0.5395	0.3239	-1.67	0.0958 .
Full-/Part-time	0.1553	0.5092	0.31	0.7603
MGPA at Senior	0.9215	0.1538	5.99	0.0000 ***
Repeated times at Senior	0.3777	0.1451	2.60	0.0092 **
Dropped times at Senior	-0.2801	0.2041	-1.37	0.1699

Signif. codes: '***' - 0.001, '**' - 0.01, '*' - 0.05, '.' - 0.1, ' ' - 1

Table 4.16: Confusion Matrix for Graduation Prediction at Senior

	Actual Graduation		
Predicted Graduation	No	Yes	Total
No	101	47	148
Yes	8	10	18

Table 4.17: Confusion Matrix for Graduation Year Prediction at Senior

			Actual Graduation Year		
Predicted Graduation Year	Not Graduate	4th year	5th year	6th year	Total
Not Graduate	98	5	13	20	136
4th year	6	9	3	2	20
5th year	2	1	2	1	6
6th year	3	0	0	1	4

Table 4.18: Coefficients for Logistic Regression Model at 5th Year

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	-3.6795	0.7375	-4.99	0.0000 ***
Entering math score	0.0144	0.0072	1.99	0.0466 *
Full-/Part-time	0.4408	0.5128	0.86	0.3900
Gender	-0.4883	0.3238	-1.51	0.1315
MGPA at 5th year	0.9519	0.1590	5.99	0.0000 ***
Dropped times at 5th year	-0.3467	0.2017	-1.72	0.0856 .
Repeated times at 5th year	0.3119	0.1266	2.46	0.0137 *

Signif. codes: '***' - 0.001, '**' - 0.01, '*' - 0.05, '.' - 0.1, ' ' - 1

Table 4.19: Confusion Matrix for Graduation Prediction at 5th Year

	Actual Graduation		
Predicted Graduation	No	Yes	Total
No	100	43	143
Yes	9	14	23

Table 4.20: Confusion Matrix for Graduation Year Prediction at 5th year

			Actual Graduation Year		
Predicted Graduation Year	Not Graduate	4th year	5th year	6th year	Total
Not Graduate	95	5	12	19	131
4th year	8	8	3	2	21
5th year	3	2	2	0	7
6th year	3	0	1	3	7

Table 4.21: Coefficients for Logistic Regression Model at 6th Year

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	-3.5116	0.7235	-4.85	0.0000 ***
Entering math score	0.0151	0.0070	2.16	0.0308 *
Full-/Part-time	0.3747	0.5006	0.75	0.4541
Gender	-0.5268	0.3164	-1.67	0.0959 .
MGPA at 6th year	0.8745	0.1556	5.62	0.0000 ***
Dropped times at 6th year	-0.1727	0.1694	-1.02	0.3078
Repeated times at 6th year	0.1574	0.0979	1.61	0.1080

Signif. codes: '***' - 0.001, '**' - 0.01, '*' - 0.05, '.' - 0.1, ' ' - 1

Table 4.22: Confusion Matrix for Graduation Prediction at 6th Year

	Actual Graduation		
Predicted Graduation	No	Yes	Total
No	102	46	148
Yes	7	11	18

Table 4.23: Confusion Matrix for Graduation Year Prediction at 6th Year

			Actual Graduation Year		
Predicted Graduation Year	Not Graduate	4th year	5th year	6th year	Total
Not Graduate	102	88	13	19	142
4th year	3	7	4	2	16
5th year	1	0	0	0	1
6th year	3	0	1	3	7

Table 4.24: Significant Variables for Logistic Regression Models

Year	Significant Variables Selected by Backwards Subset Selection
2000	gender, entering math score
2001	gender, entering math score, math GPA
2002	gender, entering math score, math GPA, dropped times and repeated times for math courses
2003	repeated times for math courses, math GPA
2004	repeated times for math courses, math GPA
2005	entering math score, math GPA
2006	entering math score, math GPA

Table 4.25: Evaluations for Logistic Regression Models

Year	Overall Accuracy Rate	Overall Error Rate	Specificity	Sensitivity	AUC
2000	0.67	0.33	0.97	0.11	0.65
2001	0.70	0.30	0.94	0.33	0.73
2002	0.72	0.28	0.97	0.23	0.77
2003	0.70	0.30	0.94	0.17	0.79
2004	0.67	0.33	0.93	0.18	0.79
2005	0.69	0.31	0.92	0.25	0.78
2006	0.68	0.32	0.94	0.19	0.78

Table 4.26: Evaluation for Decision Tree models

Year	Overall Accuracy Rate
2000	0.66
2001	0.65
2002	0.66
2003	0.61
2004	0.66
2005	0.65
2006	0.67

CHAPTER V

CONCLUSION AND DISCUSSIONS

In this thesis, we explored the patterns behind math courses taken by STEM undergraduate students. The frequency of some courses low such as Math4364 (Advanced Discrete Structures) and Math3355 (Linear Programming). It is because that 1) we only take Hispanic students into consideration; and 2) we only consider STEM undergraduate students. The top 9 math courses with have high frequency are: Math1340 (College Algebra), Math 1334 (Intermdiate Algebra), Math1356 (Trigonometry), Math1401 (Calculus I), Math 1357 (Pre-Calculus), Math1402 (Calculus II), Math3349 (Differential Equations), Math2401 (Calculus III), STAT2330 and Math2330 (Elementary Statistics and Probability). These courses are core courses in the degree plan for different colleges and students are required to take them to graduate. Also, these courses are the bases for the future research for STEM students.

We also studied time-to-graduate prediction of STEM undergraduate students. Logistic regression model and random forests model were employed to predict whether the students can graduate or not and which year they will graduate respectively. As we see from the results of prediction when students entered UTPA, there is no big difference between the area under curve (AUC) of prediction models except the model built when entering. But this is reasonable since we didn't have other information other than some demographic factors when entering. The stable AUC shows that logistic regression works very good on our data set. Our goal of graduation prediction is to warn those who might not graduate timely. Since all the values of specificity is high, that means we are good at predicting students who would not graduate. In the later prediction results, math GPA joined in our predictor variables and it was playing an important role in graduation

prediction since it was significant in every year prediction. Moreover, the coefficient is positive and approaches to 1, which emphasized its importance again. Therefore, students who want to graduate timely should pay attention to all the math courses he/she takes. Other variables were significant sometimes, but not always. For example, gender was significant when the student entered UTPA, first-year, and second-year. But after the end of second year, gender was no longer significant. In addition, student status (full-time, part-time) was never significant. Hence, we can concluded that gender and student status do not matter in timely graduation.

In the future work, we will split the data set for several times and then find the average accuracy rate, sensitivity, and specificity. In addition, we will include more variables such as reading scores, high school percentile because more realistic models should include more variables. Since the overall accuracy rates of graduation year prediction by random forest model were all around 60%, which is not very high. We will use multinomial logistic regression, support vector machines (SVM) models to make graduation year prediction and compare the results generated by random forest model. Moreover, we will include more data sets from other colleges so that we can generalize our model to all colleges in UTPA.

BIBLIOGRAPHY

[1] B. J. Anderson. Minorities and mathematics: The new frontier and challenge of the nineties. *Journal of Negro Education*, pages 260–272, 1990.

[2] S. L. Autry. Predicting student-athlete success: An analysis of graduation using precollege and college experience variables. 2010.

[3] R. Baker et al. Data mining for education. *International encyclopedia of education*, 7:112–118, 2010.

[4] B. K. Baradwaj and S. Pal. Mining educational data to analyze students' performance. *arXiv preprint arXiv:1201.3417*, 2012.

[5] J. Bean and N. Vesper. Quantitative approaches to grounding theory in data: Using lisrel to develop a local model and theory of student attrition. In *Annual Meeting AERA*, 1990.

[6] M. J. Berry and G. Linoff. *Data mining techniques: for marketing, sales, and customer support*. John Wiley & Sons, Inc., 1997.

[7] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[8] A. F. Cabrera et al. Exploring the effects of ability to pay on persistence in college. *Review of Higher Education*, 13(3):303–36, 1990.

[9] L. DeAngelo, R. Franke, S. Hurtado, J. H. Pryor, and S. Tran. *Completing college: Assessing graduation rates at four-year institutions*. Higher Education Research Institute, Graduation School of Education & Information Studies, University of California, Los Angeles, 2011.

[10] L. DeBrock, W. Hendricks, and R. Koenker. The economics of persistence: Graduation rates of athletes as labor market choice. *Journal of Human Resources*, pages 513–539, 1996.

[11] S. L. DesJardins, D. A. Ahlburg, and B. P. McCall. An event history model of student departure. *Economics of Education Review*, 18(3):375–390, 1999.

[12] S. L. DesJardins, D. A. Ahlburg, and B. P. McCall. A temporal investigation of factors related to timely degree completion. *Journal of Higher Education*, pages 555–581, 2002.

[13] C. El Moucary, M. Khair, and W. Zakhem. Improving student's performance using data clustering and neural networks in foreign-language based higher education. 2011.

[14] E. García, C. Romero, S. Ventura, and C. de Castro. A collaborative educational association rule mining tool. *The Internet and Higher Education*, 14(2):77–88, 2011.

[15] J. N. Gardner and G. L. Kramer. *Fostering student success in the campus community*, volume 138. John Wiley & Sons, 2009.

[16] G. George, E. Moore, and M. Patey. A simple model for predicting success in an engineering programme. *International Journal of Engineering Education*, 10:268–268, 1994.

[17] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[18] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009.

[19] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*. Springer, 2013.

[20] Z. Kovacic. Early prediction of student success: Mining students' enrolment data. 2010.

[21] J. Luan. Data mining applications in higher education. *SPSS Executive*, 7, 2004.

[22] R. P. Mendoza, N. Rubens, and T. Okamoto. Hierarchical aggregation prediction method. In *KDD Cup 2010 Workshop, ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2010.

[23] S. B. Merriam. *Qualitative Research and Case Study Applications in Education. Revised and Expanded from" Case Study Research in Education.".* ERIC, 1998.

[24] S. K. Mohamad and Z. Tasir. Educational data mining: A review. *Procedia-Social and Behavioral Sciences*, 97:320–324, 2013.

[25] C. E. Moucary, M. Khair, and W. Zakhem. Improving student's performance using data clustering and neural networks in foreign-language based higher education. ... *Research Bulletin of Jordan ACM, II*, II(Iii), 2011.

[26] R. Nisbet, J. Elder IV, and G. Miner. *Handbook of statistical analysis and data mining applications*. Academic Press, 2009.

[27] A. Nora, L. Barlow, and G. Crisp. Student persistence and degree attainment beyond the first year in college. *College student retention: Formula for success*, pages 129–153, 2005.

[28] O. Oyelade, O. Oladipupo, and I. Obagbuwa. Application of k means clustering algorithm for prediction of students academic performance. *arXiv preprint arXiv:1002.2425*, 2010.

[29] L. Pinkus. Using early-warning data to improve graduation rates: Closing cracks in the education system. *Washington, DC: Alliance for Excellent Education*, 2008.

[30] K. Pittman. *Comparison of data mining techniques used to predict student retention*. ProQuest, 2008.

[31] R. D. Reason. Student variables that predict retention: Recent research and new developments. *Naspa Journal*, 46(3), 2009.

[32] D. Rogulkin. Predicting 6-year graduation and high-achieving and at-risk students. *Online Submission*, 2011.

[33] C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1):135–146, 2007.

[34] J. M. Rothstein. College performance predictions and the sat. *Journal of Econometrics*, 121(1):297–317, 2004.

[35] S. Sembiring, M. Zarlis, D. Hartama, S. Ramliana, and E. Wani. Prediction of student academic performance by an application of data mining techniques. *International Proceedings of Economics Development & Research*, 6, 2011.

[36] I. Tahyudin, E. Utami, A. Amborowati, I. Tahyudin, E. Utami, and A. Amborowati. Comparison of Data Mining Classification Algorithms to Predict the Graduation Students on Time. In *Information Systems International Conference (ISICO) 2013*, 2013.

[37] M. Tair and A. M. El-Halees. Mining educational data to improve student's performance: A case study. *International Journal of Information and Communication Technology Research*, 2(2), 2012.

[38] C.-F. Tsai, C.-T. Tsai, C.-S. Hung, and P.-S. Hwang. Data mining techniques for identifying students at risk of failing a computer proficiency test required for graduation. *Australasian Journal of Educational Technology*, 27(3):481–498, 2011.

[39] J. W. Tukey. Exploratory data analysis. 1977.

[40] J. Yingkuachat, P. Praneetpolgrang, and B. Kijsirikul. An application of the probabilistic model to the prediction of student graduation using bayesian belief networks. *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology Association of Thailand (ECTI Thailand)*, pages 63–71, 2007.

[41] R. Zwick and J. C. Sklar. Predicting college grades and degree completion using high school grades and sat scores: The role of student ethnicity and first language. *American Educational Research Journal*, 42(3):439–464, 2005.

BIOGRAPHICAL SKETCH

Gejun Zhu was born in January 1986 in China. He obtained his bachelor degree in Mathematics from Tianjin University of Technology and Education, Tianjin, China in July of 2008. His main research interests focused on data mining applications and multiple attribute decision making. He joined the Mathematics department of The University of Texas-Pan American to pursue his Master's degree in August 2012. Before joining UTPA, he worked for KONE Corporation, an elevator company, as a customer support engineer. He can be reached at gejun.anderson@gmail.com in the future.