Factors Affecting the Graduation Rate for Transfer Students at New Mexico Tech

by

Daulton Gray

Submitted in Partial Fulfillment
of the Requirements for the Degree of
Master of Science in Mathematics
with Specialization in Operations Research and Statistics

New Mexico Institute of Mining and Technology
Socorro, New Mexico
August, 2018

ABSTRACT

This study examines several variables to see if a transfer student's success at New Mexico Tech can be predicted. The model used to predict the success is logistic regression. ROC curves are used to evaluate the model. The model was built on data from transfer students that entered New Mexico Tech from 2006-2010 and the model was then used to see if students from 2011 were predicted to graduate and if they actually graduated. The best predictors after running the model are last transfer institution GPA, highest math transfer credit, and campus residence. The models were only able to correctly predict about 65-75% of the 2011 students correctly.

Keywords: Student Success; Logistic Regression; ROC Curve;

CONTENTS

LIST OF TABLES

LIST OF FIGURES

This thesis is accepted on behalf of the faculty of the Institute by the following committee:

Brian Borchers, Advisor

I release this document to the New Mexico Institute of Mining and Technology.

Daulton Gray Date

CHAPTER 1

INTRODUCTION & BACKGROUND

The aim of this study is to look at the success of transfer students at New Mexico Institute of Mining and Technology. In the past at New Mexico Tech, several studies have looked at student success for freshman entering New Mexico Tech but a study has never been done that looks at transfer student success at New Mexico Tech. The big question of this study is how well can we predict whether or not a transfer student will graduate and what factors affect graduation rate?

Transfer students are a huge group that often get forgotten in todays statistics, and there is currently a trend for students to transfer between schools and be at multiple schools before finishing their degree. Most transfer students are counted as drop outs from their first school and usually not included in graduation rates for schools. One third of all students will transfer at least once in five years according to the National Student Clearinghouse.[4]

Graduation rate is the percentage of students that finish their undergraduate degree in x years from their entry. This information can help students decide if New Mexico Tech is the right place for them or help New Mexico Tech administration with their decision on if they should admit a student or not. For freshman, 4, 6 and 8 year graduation rates are commonly used. This study will look at what variables affect transfer students graduation rates at 3, 4, and 6 years.

There have been three studies done at New Mexico Tech that have looked at student success and they all of been based on freshman. In August, 2000, Julie Luna wrote a thesis titled *Predicting Student Retention and Academic Success at New Mexico Tech*. Luna looked at incoming freshman and examined after three semesters what variables would predict retention and academic success at New Mexico Tech. Retention rate is the percentage of first year students who return to New Mexico Tech for their second year and are at least enrolled full time. Luna took a data set that consisted of first time enrolled students that were new students and admitted as freshman from 1993-1997. She used Logistic Regression, Discriminant Analysis, and Classification and Regression Trees in her study to predict the fall to fall retention rate for all students and students in good standing and academic success. [6]

In Luna's study, she did not fulfill all of the goals she tried to accomplish. Her models did end up doing a good job predicting GPA. Luna found that high school GPA and ACT math scores were the best variables to predict academic success at New Mexico Tech. However, the models could not predict retention.

Her reasoning on why the model did not work as well for retention was that the variables she used may have not been important to why a student leaves New Mexico Tech.[6]

Norelle Irene Shlanta wrote a thesis titled *Statistical Models of Student Retention at New Mexico Tech* in May, 2008. Shlanta asked the question “Can we determine if a student will be successful in college before they even set foot on campus?” Shlanta, like Luna, also looked at different statistical models to predict student retention at New Mexico Tech. The data set she used was a set of first-time, full-time freshman students who entered in the fall semesters of 2004-2006. The methods used were descriptive statistics, testing of variable means, Pearson correlation coefficients, principal component analysis, multiple linear regression, and logistic regression.

Shlanta was unable to answer her question "Can we determine if a student will be successful in college before they set foot on campus?" Shlanta was able to find that first semester college GPA and first semester college credit hours earned were the best predictors of third semester student retention at New Mexico Tech. She also concluded that students drop out of school based off of academic performance. 82% of the dropouts had less than a 2.0 GPA. [8]

Another thesis was completed in July of 2009 titled *An Examination of the Impact of Selected Pre-College and College Variables on Student Retention at New Mexico Tech* by Guohui Wu. Wu looked at 3rd semester student retention at New Mexico Tech. Using pre-college variables, college variables, campus residence, family socio-economic status, and financial aid to answer if any factors predicted 3rd semester retention, if financial aid was a factor of retention and what changes need to be made. Wu used a data set that included first-time full-time freshman enrolling in the fall semesters of 2003-2007. Wu used data imputation, descriptive statistics, Pearson correlation, regression analysis, and path analysis. The study discovered that 1st semester GPA, High School GPA, ACT score, campus residence and whether or not a student received financial aid were good indicators of student retention. [9]

This study will look at transfer students compared to the last three theses. Can success be predicted for transfer students at New Mexico Tech? What variables affect transfer students more than other variables? There is more data to look at today then there were when the other 3 students did their studies.

CHAPTER 2

DATA

The data that this study will look at comes from the registrar in a report generated by a software package called Argos. Argos is a software package that generates all kinds of reports. The report used here consists of every incoming student since 2006. The data contains basic information like gender, ethnicity, whether they are a incoming freshman or transfer student and other basic information. It also includes information from high school or past university, including GPA, ACT score, and transfer credits. [10]

This study will predominantly look at transfer students that entered New Mexico Tech during the fall semester between the years of 2006-2010. The total number of students in the data is 424. In the data, the problems that were encountered when it came to transfer students were a lot of information was missing. Some transfer students did not have GPAs from their past university or they had an incorrect GPA, for example a GPA that was greater than 4. Students without this information or with incorrect information will not be included in the study. After eliminating the data that is incorrect, the data set includes 305 students. Table 2.1 shows all the available data fields split up by data that is available before they enter, and data that is available after they enter.

Available Data	
Before Entering	After Entering
Entry Term	First Semester Math Grade
In-state or out-of-state Student	First Degree Major
Major on Entry	When Degree Completed
Race	Retained
Ethnicity	Attempted and Earned Credits
Gender	Semester GPAs
Low Income	Academic Warning
First Generation Student	Academic Probabtion
Lived on Campus	Academic Suspension
Last Transfer Institution	Grades in Math Classes
Last Transfer Institution GPA	Cumulative GPA for each semester
Total Transfer Credits	
Highest Math Transfer	
Highest Chemistry Transfer	
Highest Physics Transfer	

Table 2.1: Data Fields Available

Looking at the variables, the entry term is the term in which the student entered New Mexico Tech. The format for the entry term is 200710 where the first 4 digits refer to the school year and the last two refer to the semester. The year is encoded in terms of the spring semester and the end of the academic year. A 10 is summer, 20 is fall, and 30 is spring. For example, if a student is labeled 200730, they entered in the spring of 2007 and if a student is labeled 200720, they entered in the fall of 2006. This variable is used to determine if the student is a part of the training data, testing data or even in the data at all.

An in-state student is a transfer student that transfered from a school in New Mexico and an out-of-state student came from a school outside of New Mexico. Major on entry is the major the student picked when entering New Mexico Tech.

Race is the student's race which includes Native American, Asian, Black, Caucasian, etc. This variable is not very useful because the population of New Mexico Tech is predominantly Caucasian. There are 348 transfer students that are Caucasian which is about 82.1% of the data. The other races are all lower than 5%. American Indian is 4.5% of the data, Asian is 4.5%, Black is 5.0% and, Non-Resident is 3.7%.

Ethnicity is whether a student identifies themselves as "Hispanic or Latino" or "not Hispanic or Latino". 21.22% of New Mexico Tech transfer students are Hispanic and 75.47% are not. 3.30% of the transfer students are not listed.

Gender is whether the student identifies themselves as male or female.

Low income is whether or not the student was eligible for a Pell Grant.

First generation student is whether or not a student is the first from their family to go to college or not as reported by the student

Lived on campus is whether or not the student lived on campus in their first semester. 49.67% of the transfer students lived on campus their first semester.

Last transfer institution is the last school the transfer student attended. The problem with this variable is there are not enough transfer students from one particular school for it to be useful. The highest transfer school is Central New Mexico Community College with 42 students which is about 9.9% of the data. Another problem is this variable only indicates the last transfer school the student attended meaning if they transfered multiple times, the earlier institutions are not shown.

Last transfer institution GPA is the GPA from the previous school the student attended. This is where the problem was were a GPA not on a 4.0 scale was found and the GPA could have been entered incorrectly. In this study, the GPAs that were not on a 4.0 scale were not used.

Total transfer credits is how many total credits a transfer student brought to New Mexico Tech.

Highest math transfer credit is the highest math course the transfer student is transferring over from their previous school. The math transfer courses New Mexico Tech accepts are College Algebra (MATH 101), PreCalculus (MATH 103), Trigonometry (MATH 104), Calculus I (MATH 131), and Calculus II (MATH 132). New Mexico Tech does accept more advanced courses, but they are not in the data.

Highest chemistry and physics are identical to highest math transfer but instead of math it is chemistry and physics. Chemistry I (CHEM 121), Chemistry II (CHEM 122), Physics I (PHYS 121), and Physics II (PHYS 122) are the courses that New Mexico accepts transfer credits for.

First semester math grade is the grade the student got in their first math class at New Mexico Tech.

First degree major is the major the student completed their first degree under. This also indicates whether or not the student graduated.

Retained means the student returned the following year from when they entered. One problem with this variable is if a student took a break from school and came back, they are not considered retained.

Attempted and earned credits are the amount of credits the student took at New Mexico Tech and the amount of credit the student earned, meaning they passed the class.

Semester GPA is the GPA in each semester the student was at New Mexico Tech.

Academic warning, probation, and suspension mean whether or not the student was given a warning, put on probation, or was ever suspended. This is broken up by semester.

Grades in math classes are the grades the students got whenever they took a math class at New Mexico Tech.

Cumulative GPA is the GPA of all semesters combined. It is updated each semester.

CHAPTER 3

METHODS

Logistic Regression will be used in this study. Logistic Regression is a regression model where the dependent variable takes on a limited number of possible values and representing different groups. In this study, binomial logistic regression will be used. Binomial logistic regression only has two possible outcomes for the dependent variable, ether a "0" or a "1".[3] The reason for using binomial logistic regression is because it is binomial and can model whether or not a student will graduate. Using the data set and building a model to see if a student will graduate can be determined with gender, ethnicity, income, and other things. The logistic function can be defined as:

$$p(t) = \frac{1}{1+e^{-t}}$$

where $t = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ and x_i is one of the variables found in the data.

The output p is bounded by zero and one, which is what is needed to model probabilities. Logistic regression's distribution is "S" shaped meaning differences in the center influence the outcome more than the differences at the end. The upper half of the "S" curve show high probabilities and the lower half show low probabilities. [6]

A simple example of logistic regression would be trying to predict if a transfer student would graduate based on GPA from the institution they transfered from. The reason why it is a good example is because the dependent variable has only two outcomes, the student ether graduated or did not graduate which would be denoted as a "1" or a "0" respectively. Table 3.1 shows a small random sample of 20 transfer students from the data, with their GPA and whether they graduated or not.

GPA	2.25	2.26	2.43	2.55	2.72	2.94	3.07	3.10	3.14
Graduated	0	0	0	0	1	0	0	1	0
GPA	3.26	3.28	3.29	3.34	3.37	3.38	3.69	3.76	3.83
Graduated	0	1	1	0	1	1	1	1	0
GPA	4.00	4.00							
Graduated	1	1							

Table 3.1: Logistic Regression Example Data

Figure 3.1 shows the probability of graduating versus the last transfer GPA.

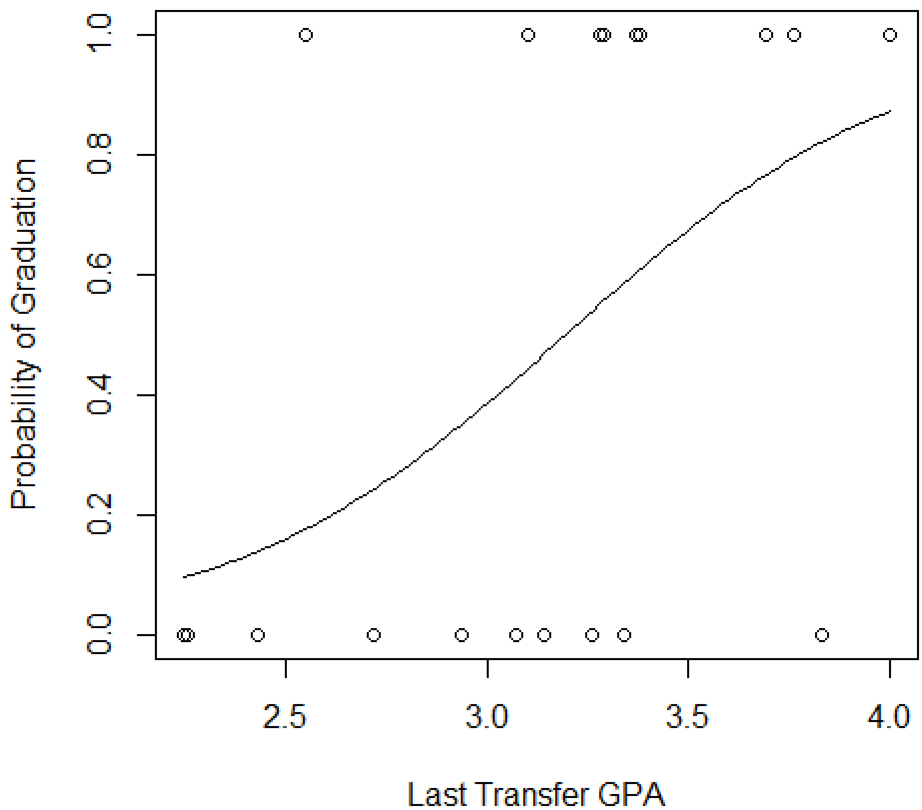

Figure 3.1: Graph of Logistic Example

The logistic regression output is in table 3.2.

	Coefficient	Std.Error	z-value	p-value
Intercept	-7.66	3.97	-1.93	0.05
GPA	2.40	1.22	1.96	0.05

Table 3.2: Logistic Regression Analysis

In this example, the last transfer GPA is significant to the probability of graduating because the p-value is at 5%. The output also gives the coefficients for the intercept and the GPA giving a logistic regression equation to estimate the odds of graduating.

$$t = 2.40x - 7.66$$

where t is the log-odds of graduating and x is the last transfer GPA. The probability then can be defined as

$$p(t) = \frac{1}{1 + e^{-(2.40x - 7.66)}}$$

The data will be split into two groups. Training and validating datasets. The training set is used to fit parameters. In this model, the training dataset is chosen randomly and it consists of 60% of the initial data set. The validation dataset consists of the other 40% of the data. After the logistic regression model, a cut-off probability will be created to make yes/no predictions. The students above the cut-off are predicted to graduate and the students below the cut-off are predicted to not graduate. This is where the validation data set is used. The cut-off usually starts at .5 and will move up or down in based on the accuracy of the model.[5]

Along with logistic regression, receiver operating characteristic curves or ROC curves will be used. An ROC curve is a graphical plot that is used to show how well a classification model works. A classification model is used to predict the value of outcomes given the inputs. [2]

To create an ROC curve, the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) must be found. To find these, outcomes are labeled as either positive or negative. The true positive is when an outcome was predicted positive and the actual outcome is positive. The false positive is when the outcome is predicted positive, but the actual outcome is negative. The true negative is when both the prediction and actual outcome is negative. The false negative is when the outcome is predicted negative, but the actual outcome is positive.[1]

prediction outcome

		p	**n**	**total**
actual value	**p′**	True positive	False negative	P′
	n′	False positive	True negative	N′
	total	P	N	

Figure 3.2: Confusion Matrix

Figure 3.2 is a confusion table. The ROC curve is created using the true positive rate as a function of the false positive rate. The true positive rate is the measure of how many positives that correctly predicted over the total number of positives.

$$TPR = \frac{TP}{TP + FN}$$

The true positive rate is also called hit rate, recall, or sensitivity. In this study, the true positive rate measures the number of students we correctly predict will graduate over the number that did graduate. [1]

The false positive rate is the measure of how many positives are incorrectly predicted over the number of total negatives.

$$FPR = \frac{FP}{FP + TN}$$

The false positive rate is also called false alarm rate. In this study, the false positive rate would be the number of students we predict will graduate but end up not graduating over the total number of students that do not graduate. [1]

An ROC curve will be created from the validation data. When graphing an ROC curve, the x-axis is the false positive rate and the y-axis is the true positive rate. The origin or the lower left corner of an ROC curve is where no outcomes are predicted as positive. The upper right corner or the point $(1, 1)$ is where there is all the outcomes are predicted as positive. The lower right corner or the point $(1, 0)$ is where the model is 100% incorrect and no outcome was predicted correctly. The lower left corner or the point $(0, 0)$ is where every outcome is predicted as negative. The upper left corner or the point $(0, 1)$ is where the model is perfect and every outcome was correctly predicted. The diagonal line $y = x$ is what the model would look like if it was completely random. The TPR and the FPR have to equal for it to be on the line $y = x$. For example, if the model randomly predicts half to be positive and the other half to be negative, the TPR and the FPR should both be 0.5. The best model will end up being the model closest to the upper left corner. [1] The model will go through a loop to find the best cutoff. The cutoff starts at 0.5 and will move up or down based on the the accuracy of the model. Once the ROC curve gives the best model, that cutoff will be used to predict the testing data. The ROC curve is created by the different TPR and FPR from the different cutoffs created during the loop.

To show an example of an ROC curve, the model from the logistic regression example and a new group of random 20 students that will serve as a validation data set will be used. Table 3.3 shows the validation data along with the probability that the student will graduate and the prediction if they will graduate.

GPA	Graduate	Probability	Prediction
2.22	0	0.09	False
2.26	0	0.10	False
2.43	0	0.14	False
2.44	1	0.14	False
2.50	0	0.16	False
2.60	1	0.19	False
2.60	0	0.19	False
2.66	0	0.22	False
2.98	1	0.37	False
3.20	0	0.50	False
3.35	1	0.59	False
3.59	1	0.72	True
3.73	1	0.78	True
3.75	0	0.79	True
3.75	1	0.79	True
3.88	1	0.84	True
3.96	1	0.86	True
4.00	1	0.88	True
4.00	1	0.88	True
4.00	1	0.88	True

Table 3.3: Data for ROC Example

The cutoff probability that was calculated was 0.591. The true positive rate is 0.750 and the false positive rate is 0.125. Figure 3.3 is the ROC curve that was generated from the data above.

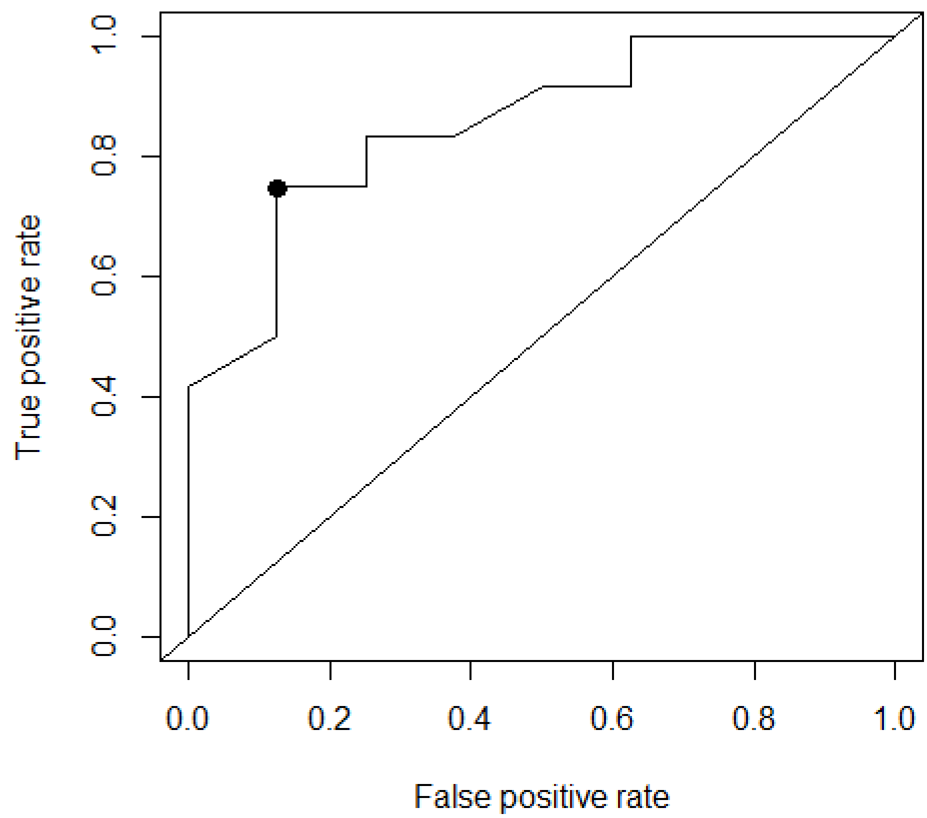

Figure 3.3: ROC example

The area under the curve is 0.854. The point marked in figure 3.3 is the point were the cutoff was chosen for the model because it is the closest point to the upper left corner. The point is $(0.125, 0.750)$.

After the ROC curve, a final testing dataset is used to test the model. The testing data will be the students that entered New Mexico Tech in 2011. The students in the testing data will be put into the model to see if they will graduate or not. The results from the model will then compare to what actually happened to the students who entered in 2011. [5]

The model is written in the programming language R. R is mainly used for statistical computation. Many different models are build using R, for example, linear models, nonlinear regression models and and other statistical procedures. [5]

CHAPTER 4

RESULTS

4.1 Breakdown of Data

Before the model was created, a breakdown of the data set was done. The breakdown looked at basic graduation rates for 3, 4, and 6 years. The breakdown also looked at the percentage of each variable in the data set along with the 3, 4, and 6 year graduation rate using these variables. The variables that were used were gender, ethnicity, low income, first generation status, student type, highest math transfer, highest chemistry transfer, highest physics transfer, transfer GPA and major.

The 3, 4, and 6 year graduation rate for transfer students are 20.99%, 34.90%, and 46.46% respectively. Table 4.1 shows the breakdown:

Variable	Percent	3 year grad	4 year grad	6 year grad
Student Type				
Out of State	45.52%	24.87%	39.38%	48.70%
In State	54.48%	17.75%	31.17%	44.59%
Gender				
Male	68.40%	20.34%	34.14%	46.90%
Female	31.60%	22.39%	36.57%	45.52%
Ethnicity				
Hispanic	21.95%	18.89%	32.22%	43.33%
Non-Hispanic	78.05%	20.94%	35.63%	47.81%
Income				
Low Income	38.44%	19.02%	31.29%	47.51%
Not Low Income	61.56%	22.22%	37.16%	44.79%
First Generation Status				
First Generation	23.58%	12.00%	23.00%	37.00%
Non First Generation	76.42%	23.77%	38.58%	49.38%

Table 4.1: Breakdown of Data

Variable	Percent	3 year grad	4 year grad	6 year grad
Highest Math Transfer				
None	19.40%	8.54%	14.63%	21.95%
Math 101	8.97%	0.00%	7.89%	28.95%
Math 103	7.78%	3.03%	18.18%	27.27%
Math 104	9.91%	9.52%	30.95%	50.00%
Math 131	15.33%	13.85%	30.77%	47.69%
Math 132	38.70%	41.46%	57.32%	65.24%
Highest Chemistry Transfer				
None	43.87%	6.45%	16.67%	31.72%
Chem 121	24.76%	19.05%	40.95%	48.57%
Chem 122	31.37%	42.86%	55.64%	65.41%
Highest Physics Transfer				
None	64.15%	9.93%	23.16%	35.66%
Phys 121	8.02%	23.53%	44.12%	55.88%
Phys 122	27.83%	45.76%	59.32%	68.64%
Transfer GPA				
[0.0, 2.0]	2.32%	14.29%	14.29%	14.29%
(2.0, 2.5]	11.59%	11.43%	17.14%	20.00%
(2.5, 3.0]	19.87%	18.33%	31.67%	41.67%
(3.0, 3.5]	29.47%	16.85%	33.71%	47.19%
(3.5, 4.0]	36.75%	27.93%	44.14%	55.86%
Race				
American Indian	4.47%	17.65%	47.06%	58.82%
Asian	4.47%	45.45%	63.64%	81.82%
Black	4.98%	22.22%	27.78%	44.44%
Caucasian	82.1%	18.41%	32.64%	42.68%
Non-Resident Alien	3.73%	25.00%	25.00%	25.00%
Campus Residence				
On Campus	49.67%	25.17%	40.40%	54.30%
Off Campus	50.33%	16.34%	30.07%	37.25%

Table 4.2: Breakdown of Data 2

Variable	Percent	3 year grad	4 year grad	6 year grad
First Major				
None	4.01%	0.00%	0.00%	0.00%
Biology	4.72%	25.00%	40.00%	50.00%
Basic Science	1.18%	0.00%	20.00%	20.00%
Civil Engineering	6.37%	33.33%	48.15%	62.96%
Chemical Engineering	7.55%	21.88%	37.50%	56.25%
Chemistry	9.43%	50.00%	50.00%	75.00%
Computer Science	6.37%	14.81%	22.22%	37.04%
Electrical Engineering	6.84%	24.14%	37.93%	68.97%
Environmental Engineering	6.84%	20.00%	33.33%	40.00%
Environmental Sciences	1.89%	12.50%	25.00%	25.00%
Earth Sciences	4.48%	26.32%	36.84%	47.37%
Engineering Undecided	0.47%	0.00%	50.00%	100.00%
General Studies	9.43%	25.00%	25.00%	25.00%
Information Technology	3.01%	0.00%	7.69%	23.08%
Materials Engineering	2.83%	25.00%	50.00%	58.33%
Mathematics	2.36%	20.00%	20.00%	40.00%
Mineral Engineering	2.36%	10.00%	60.00%	60.00%
Mechanical Engineering	14.15%	28.33%	50.00%	60.00%
Management	2.36%	20.00%	30.00%	40.00%
Petroleum Engineering	13.92%	23.20%	38.98%	47.46%
Physics	5.67%	0.00%	16.67%	25.00%
Psychology	1.18%	0.00%	0.00%	0.00%
Technical Communications	1.89%	0.00%	12.50%	12.50%
Last Transfer Institution (Only Institution Greater than 1% of the data)				
CNM	13.82%	9.52%	26.20%	33.33%
ENM	1.64%	20.00%	40.00%	40.00%
New Mexico Highlands	2.96%	33.33%	44.44%	44.44%
NMSU	2.96%	0.00%	11.11%	11.11%
San Juan	7.57%	47.83%	56.52%	78.26%
Santa Fe CC	2.96%	0.00%	0.00%	33.33%
UNM	11.51%	20.00%	31.43%	37.14%
First Semester GPA				
[0.0, 2.0]	30.10%	4.44%	5.56%	12.22%
(2.0, 2.5]	13.71%	12.20%	26.83%	43.90%
(2.5, 3.0]	15.05%	15.56%	31.11%	44.44%
(3.0, 3.5]	21.40%	32.81%	56.25%	70.31%
(3.5, 4.0]	19.73%	44.07%	69.49%	76.27%

Table 4.3: Breakdown of Data 3

Looking at the breakdown, the variables that seemed to show the biggest impact were first generation status, math transfer, chemistry transfer, physics transfer, transfer GPA, campus residence, and first semester GPA. So the model should show the same information.

4.2 Logistic Regression with All Variables on Transfer Students

After the breakdown a model is built using all the variables except major on entry and last transfer institution. Those variable were left out because there are so many different majors and institutes corresponding to such a small data set that they probably will not affect the model. A model with those variables will be created later to show that they do not affect graduation rate. The data is all the transfer students from 2006-2010 and is not split up into training and validation sets. When running the model, the variables with a p-value of 0.05 or lower show the variable being significant, meaning that variable is more of a factor to whether or not a student will graduate or not. Table 4.4 is the p-values for the 3, 4, and 6 year graduation model using all the variables. Only the variables that are significant will be shown.

Variable	3 Year	4 Year	6 Year
In State	0.040	0.104	0.381
Ethnicity	0.071	0.025	0.359
First Generation	0.242	0.003	0.825
Math 104	0.512	0.440	0.027
Math 132	0.286	0.019	0.003
Chem 121	0.766	0.8946	0.091
Chem 122	0.030	0.1588	0.513
Phys 121	0.077	0.1280	0.106
Transfer GPA	0.807	0.184	0.062
Campus Resident	0.221	0.075	0.039
First Semester GPA	0.0001	9.3×10^{-8}	1.98×10^{-8}

Table 4.4: Logistic Regression with All Variables on Transfer Students

This model shows that in-state or out-of-state, ethnicity, first generation, highest math transfer, highest chemistry transfer, highest physics transfer, last transfer GPA, campus residence and first semester GPA are good variables to start with. The problem with first semester GPA is that it will not help with deciding if an incoming transfer student will be successful or not until they are finished with their first semester.

4.3 Logistic Regression with Major on Entry

Major on entry has two many inputs to really have an impact on the model. To prove this point, a regression model with just major on entry was created for 3, 4, and 6 year graduation rate. The p-values will be in table 4.5

Variable	3 Year	4 Year	6 Year
Biology	0.997	0.995	0.993
Basic Science	1.000	1.000	0.993
Civil Engineering	1.000%	0.995	1.000
Chemical Engineering	0.997	0.995	0.993
Chemistry	1.000	1.000	0.993
Computer Science	0.997	0.995	0.993
Electrical Engineering	0.997	0.995	0.993
Environmental Engineering	0.997	0.995	0.993
Environmental Sciences	0.997	0.995	1.000
Earth Sciences	0.997	0.995	0.993
Engineering Undecided	0.997	0.995	0.992
General Studies	1.000	1.000	0.993
Information Technology	1.000	1.000	0.993
Materials Engineering	1.000	0.995	0.993
Mathematics	0.997	0.995	0.993
Mineral Engineering	0.997	0.995	0.993
Mechanical Engineering	1.000	0.995	0.993
Management	1.000	1.000	0.993
Petroleum Engineering	0.997	0.995	0.993
Physics	1.000	1.000	0.994
Psychology	1.000	1.000	1.000
Technical Communications	1.000	0.995	0.993

Table 4.5: Logistic Regression with Major on Entry

This model shows that this is not a good factor to predict graduation because the p-values are 1.000 or almost 1.000. The p-values less than 0.05 are significant and there is not a predictor close to 0.05.

4.4 Logistic Regression Model with Last Transfer Institution

Variable	3 Year	4 Year	6 Year
CNM	0.083	0.359	0.096
ENM	0.914	0.779	0.717
New Mexico Highlands	0.441%	0.524	0.823
NMSU	0.996	0.191	0.062
San Juan	0.013	0.045	0.012
Santa Fe CC	0.999	0.996	0.392
UNM	0.797	0.786	0.246

Table 4.6: Logistic Regression with Last Transfer Institution

This model is inconsistent. Only the Institutions that have more than 1% are shown in Table 4.6. All the other Institutions had a about a 0.999 p-value.This is because there are not enough students coming from each Institution.

4.5 3 Year Logistic Regression Models

4.5.1 Model 1

Using the breakdown and the models above, a model is built using the variables in state or out of state, ethnicity, first generation, highest math transfer, highest chemistry transfer, highest physics transfer, last transfer GPA, and campus residence to predict 3 year graduation. The transfer students are randomly split into training and validation sets. Table 4.7 gives the following information on the variables:

Variable	Coefficient	Std. Error	z-value	p-value
Intercept	-3.921	1.526	-2.569	0.010
In State Student	-0.389	0.487	-0.799	0.424
Not Hispanic or Latino	-0.549	0.571	-0.961	0.337
First Generation	0.0003	0.572	0.001	0.999
Math 101	-16.360	1466.000	-0.011	0.991
Math 103	-16.530	1817.000	-0.009	0.993
Math 104	-0.219	1.148	-0.190	0.8490
Math 131	-0.395	0.976	-0.405	0.686
Math 132	0.625	0.854	0.732	0.464
Chem 121	0.261	0.731	0.357	0.721
Chem 122	1.254	0.719	1.744	0.081
Phys 121	0.678	0.737	0.920	0.358
Phys 122	0.662	0.599	1.106	0.269
Transfer GPA	0.550	0.397	1.384	0.166
Campus Resident	1.057	0.481	2.196	0.028

Table 4.7: Model 1, Test 1 for 3 year graduation

This model shows that only campus residence is a significant variable. This can be seen because they have small p-values. Having transfer credit for Chem 122 is almost significant but still has a p-value greater than 0.05. The variables with the a high p-value are not statistically significant, especially first generation status with a p-value of 0.999. The coefficient is also almost zero, showing that this variable does not affect the model much. The ROC curve is shown below.

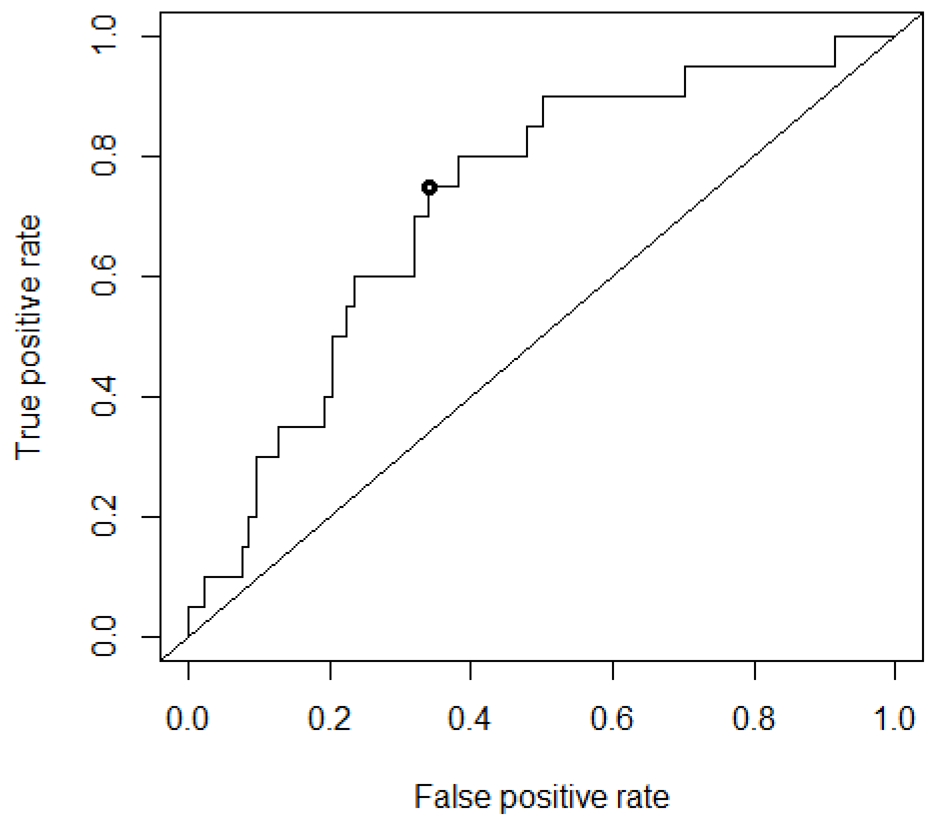

Figure 4.1: ROC Curve for Model 1, Test 1 for 3 year graduation

The area under the curve is 72.44% which is a decent model. The closer to 100% the better the model is. The point is where the cutoff was decided. The TPR was 0.750 and the FPR was 0.340. The cutoff was 0.258. The final test consisted of 63 students that entered in 2011. 52 students were correctly predicted. 12 of those were true positive and the other 40 were true negatives. The TPR was 0.700 and the FPR was 0.340 for the final test. Since the training and validation data is random, running the model a second and third time. The information is a little bit different.

Test 2				
Variable	Coefficient	Std. Error	z-value	p-value
Intercept	-3.338	1.5052	-2.218	0.027
In State Student	-0.852	0.503	-1.694	0.090
Not Hispanic or Latino	-0.886	0.565	-1.567	0.117
First Generation	-0.563	0.586	-0.962	0.336
Math 101	-16.957	1440.602	-0.012	0.991
Math 103	-16.907	1639.982	-0.010	0.992
Math 104	-0.289	0.884	-0.327	0.744
Math 131	-0.985	0.963	-1.023	0.306
Math 132	0.498	0.854	0.579	0.562
Chem 121	-0.166	0.789	-0.210	0.834
Chem 122	0.817	0.689	1.186	0.236
Phys 121	1.389	0.850	1.63	0.102
Phys 122	0.733	0.575	1.275	0.203
Transfer GPA	0.805	0.438	1.836	0.066
Campus Resident	0.840	0.444	1.892	0.058
Test 3				
Variable	Coefficient	Std. Error	z-value	p-value
Intercept	-2.569	1.333	-1.927	0.054
In State Student	0.061	0.460	0.132	0.895
Not Hispanic or Latino	-0.292	0.568	-0.514	0.608
First Generation	-0.262	0.541	-0.484	0.628
Math 101	-17.030	1346.969	-0.013	0.990
Math 103	-1.009	1.183	-0.853	0.394
Math 104	-0.461	0.861	-0.536	0.592
Math 131	-1.269	0.967	-1.312	0.189
Math 132	0.385	0.703	0.548	0.584
Chem 121	0.298	0.696	0.428	0.668
Chem 122	0.709	0.666	1.064	0.288
Phys 121	0.556	0.791	0.704	0.482
Phys 122	0.291	0.577	0.504	0.614
Transfer GPA	0.268	0.345	0.779	0.436
Campus Resident	0.632	0.440	1.437	0.151

Table 4.8: Model 1, Test 2 and 3 for 3 year graduation

The models are not as similar. In the second test, no variable is statistically significant. In state, transfer GPA, and campus residence have p-values less than 0.1, so those variables are almost significant. Variables that have high p-values and coefficients far from zero are having Math 101 and 103 credit. This variables show that having this math class as the highest math transfer the student is very unlikely to graduate in 3 years.

The third model also did not have any variables statistically significant. Math 101 was once again had a high p-value. Figure 4.2 and 4.3 are the ROc curves for test 2 and 3.

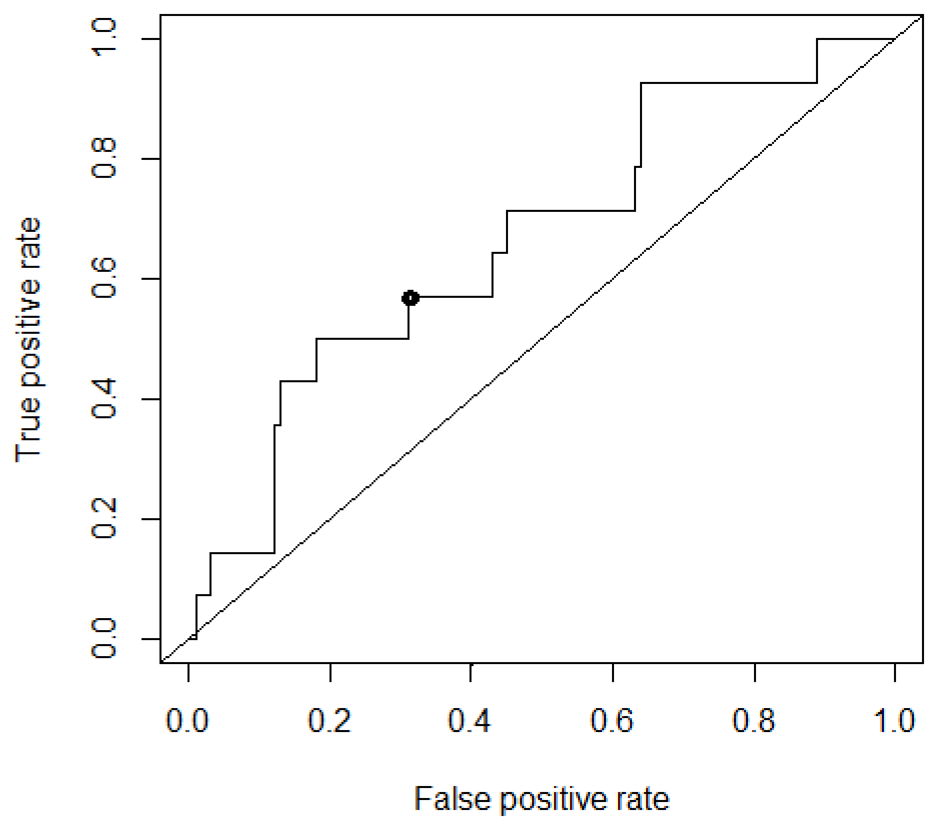

Figure 4.2: ROC Curve for Model 1, Test 2 for 3 year graduation

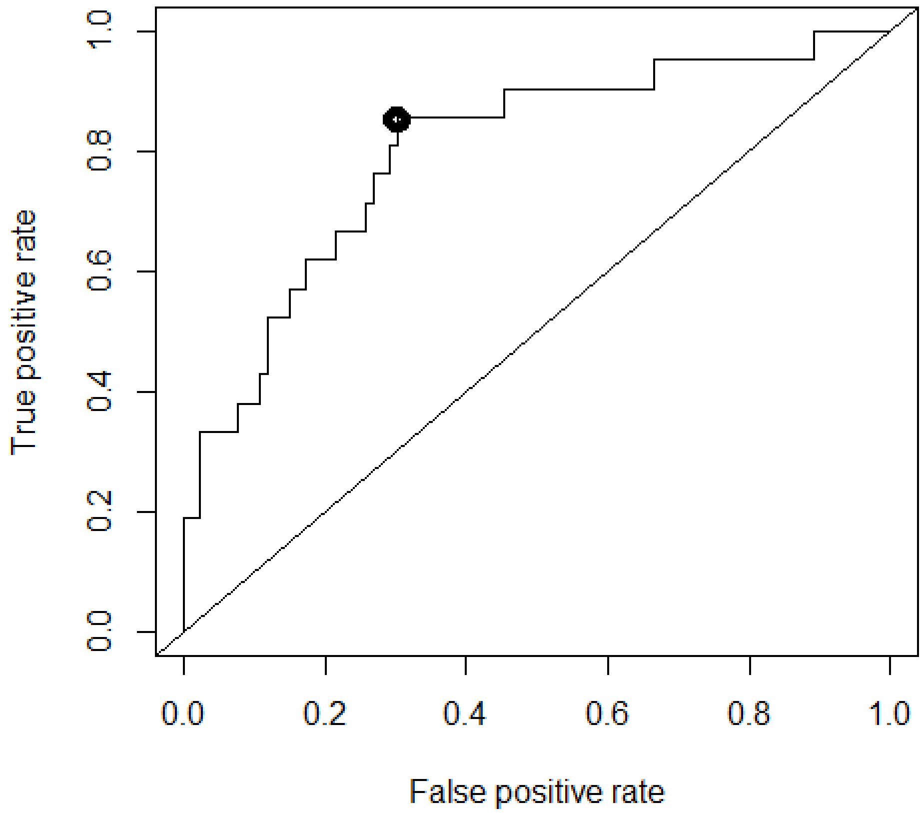

Figure 4.3: ROC Curve for Model 1, Test 3 for 3 year graduation

The area under the curve is 66.43% and 80.24% for test 2 and 3 respectively. With the TPR 0.571, FPR 0.310 and a cutoff of 0.399 for test 2. The TPR was 0.857 and the FPR was 0.301 for test 3. The cutoff for test 3 was 0.241. Test 2 had 52 correctly predicted with 9 being true positives and test 3 had 51 with 13 being true positive. For test 2, the TPR was 0.600 and the FPR was 0.104, and for test 3, the TPR was 0.867 and the FPR was 0.208.

4.5.2 Model 2

Using the variables that were significant in model 1, a new model is created using in state or out of state, highest chemistry transfer, transfer GPA, and campus residence.

Variable	Coefficient	Std. Error	z-value	p-value
Intercept	-4.722	1.481	-3.187	0.001
In State Student	-0.576	0.455	-1.266	0.206
Chem 121	0.624	0.683	0.914	0.361
Chem 122	1.823	0.556	3.281	0.001
Transfer GPA	0.690	0.4413	1.564	0.118
Campus Resident	0.359	0.451	0.796	0.426

Table 4.9: Model 2, Test 1 for 3 year graduation

In this model, Chem 122 is the only significant variable. Figure 4.4 is the ROC curve.

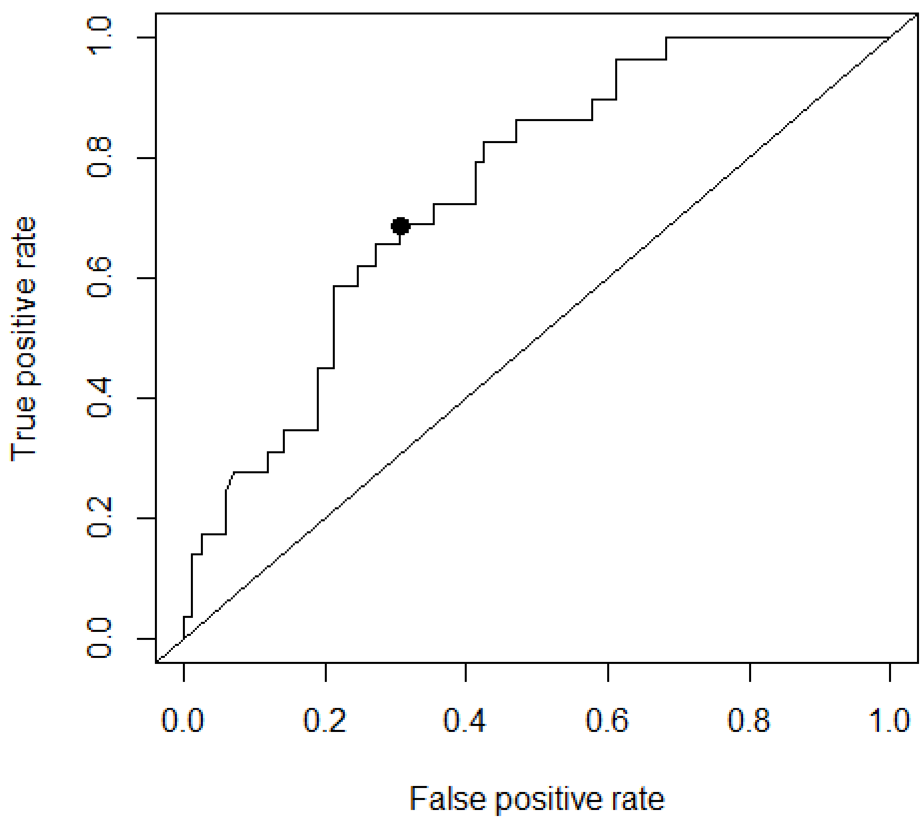

Figure 4.4: ROC Curve for Model 2, Test 1 for 3 year graduation

The area under the curve is 74.87%. The TPR is 0.690 and the FPR is 0.306. The cutoff is 0.143. The final test only predicted 46 correctly with 12 true positives and a TPR of 0.800 and 34 true negatives and a FPR of 0.292. Two more tests were ran for this model.

Test 2				
Variable	Coefficient	Std. Error	z-value	p-value
Intercept	-4.967	1.653	-3.005	0.003
In State Student	-0.824	0.480	-1.718	0.086
Chem 121	1.633	0.7382	2.12	0.027
Chem 122	2.187	0.678	3.227	0.001
Transfer GPA	0.670	0.481	1.394	0.163
Campus Resident	0.133	0.485	0.274	0.784
Test 3				
Variable	Coefficient	Std. Error	z-value	p-value
Intercept	-3.485	1.384	-2.517	0.012
In State Student	-0.760	0.462	-1.645	0.010
Chem 121	0.988	0.698	1.416	0.157
Chem 122	2.126	0.600	3.543	0.0004
Transfer GPA	0.167	0.406	0.412	0.681
Campus Resident	0.853	0.468	1.822	0.069

Table 4.10: Model 2, Test 2 and 3 for 3 year graduation

In test 2, Chem 121 and Chem 122 are significant variables and in test 3 just Chem 122. The ROC curves are in figure 4.5 and 4.6.

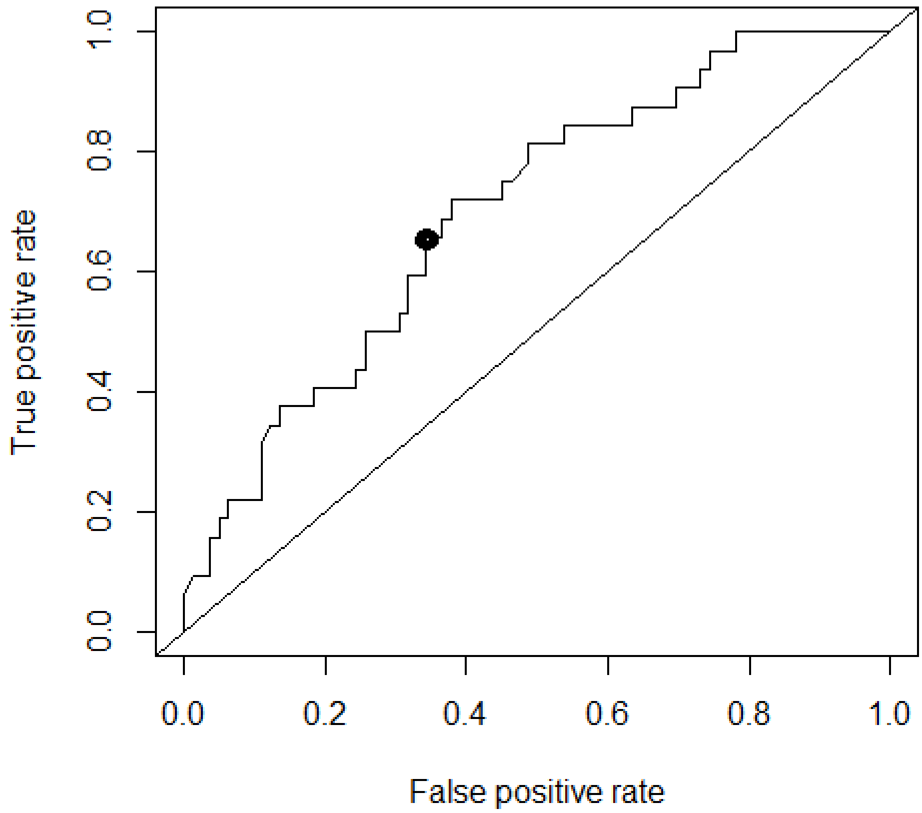

Figure 4.5: ROC Curve for Model 2, Test 2 for 3 year graduation

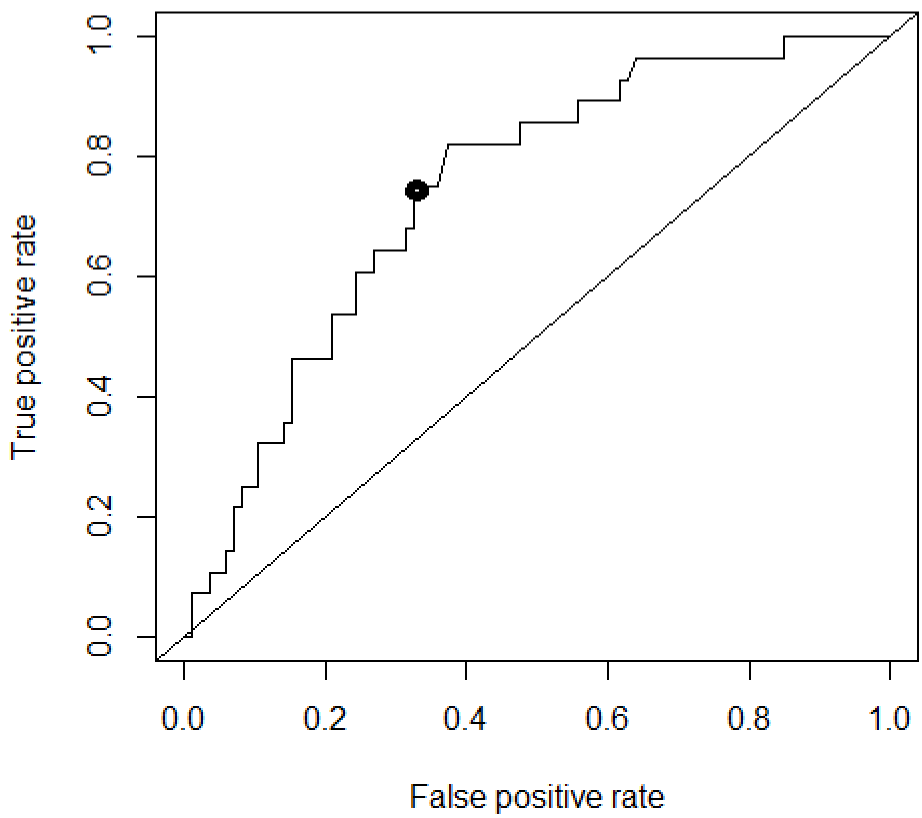

Figure 4.6: ROC Curve for Model 2, Test 3 for 3 year graduation

The area under the curve is 69.97% for test 2 and 74.48% for test 3. The TPR is 0.719 and the FPR is 0.378 for test 2. The cutoff is 0.158. For test 3, The TPR is 0.750 and the FPR is 0.326. The cutoff is 0.119. Test 2 correctly predicted 45 with a TPR of 0.800 and a FPR of 0.3125. Test 3 predicted 43 correctly with a TPR of 0.867 and a FPR of 0.375. This model is actually worse than model 1. It can be concluded that it is hard to predict 3 year graduation. This is because a transfer student will probably need a lot of specific transfer courses in order to graduate in 3 years. The first model did not have very many significant variables and when we created a model with the variables that were almost significant or significant the model did worse.

4.6 4 Year Logistic Regression Models

4.6.1 Model 1

A model for 4 year graduation rate was then created using the same variables as model 1.

Variable	Coefficient	Std. Error	z-value	p-value
Intercept	-6.333	1.647	-3.845	0.0001
In State Student	-0.776	0.470	-1.653	0.098
Not Hispanic or Latino	-1.008	0.570	-1.769	0.077
First Generation	-1.053	0.534	-1.973	0.049
Math 101	-1.090	1.185	-0.920	0.358
Math 103	-0.129	1.0015	-0.129	0.898
Math 104	-0.227	0.827	0.274	0.784
Math 131	0.265	0.772	0.343	0.732
Math 132	1.326	0.763	1.737	0.082
Chem 121	0.8205	0.583	1.408	0.159
Chem 122	0.946	0.572	1.655	0.098
Phys 121	1.209	0.773	1.564	0.118
Phys 122	0.372	0.592	0.628	0.530
Transfer GPA	1.661	0.480	3.463	0.0005
Campus Resident	0.858	0.431	1.992	0.046

Table 4.11: Model 1, Test 1 for 4 year graduation

This model has multiple significant variables. Those variables are first generation, transfer GPA, and campus residence. Transfer GPA has a p-value less than 0.01 showing that in this model, it is a great factor to predict graduation in 4 years. Note that in state, ethnicity, Math 132, and Chem 122 have small p-values, but not smaller than 0.05 so it is possible that in another model these variables may become statistically significant. In figure 4.7 the ROC curve is shown.

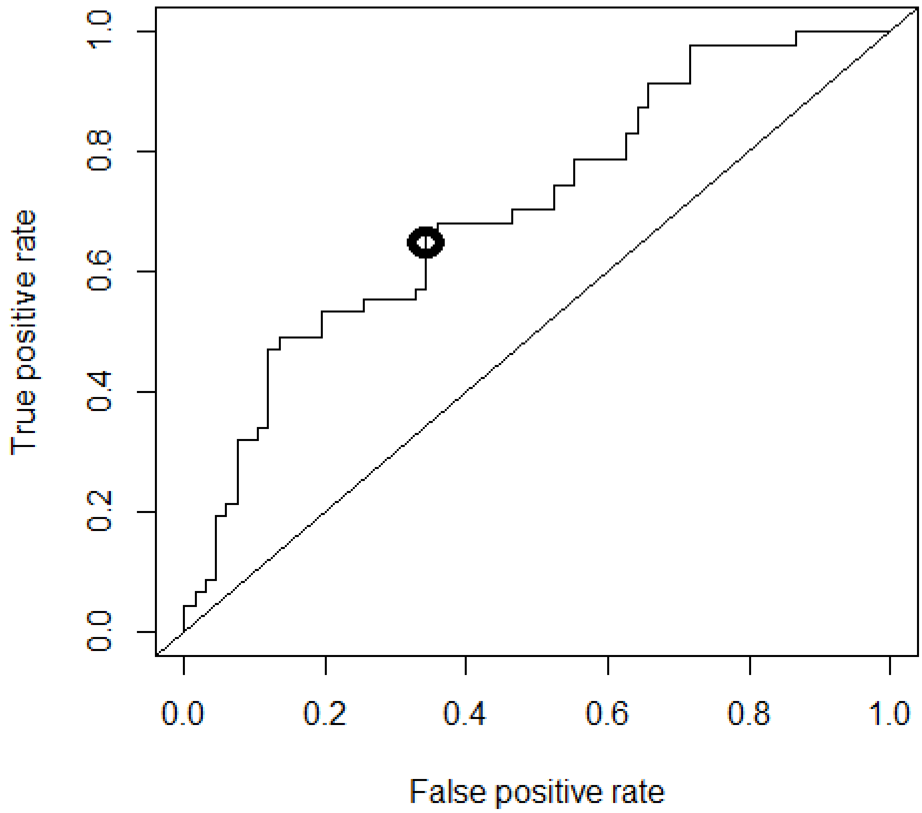

Figure 4.7: ROC Curve for Model 1, Test 1 for 4 year graduation

The area under the curve is 70.56%. The TPR for the ROC curve is a 0.681 and the FPR is 0.358. The cutoff is 0.404. In the final test, 49 students were correctly predicted. 14 of those were true positive and the other 35 were true negative. The TPR was 0.700 and the FPR was 0.186 for the final test.

Two more tests were done with this model.

Test 2				
Variable	Coefficient	Std. Error	z-value	p-value
Intercept	-3.727	1.4353	-2.597	0.009
In State Student	-0.593	0.421	-1.409	0.159
Not Hispanic or Latino	-0.740	0.538	-1.378	0.168
First Generation	-0.531	0.491	-1.082	0.279
Math 101	-1.413	1.126	-1.255	0.210
Math 103	-0.175	0.901	-0.194	0.846
Math 104	-0.259	0.765	-0.338	0.735
Math 131	0.358	0.651	0.550	0.582
Math 132	1.188	0.656	1.810	0.070
Chem 121	0.071	0.579	0.123	0.902
Chem 122	0.415	0.580	0.714	0.475
Phys 121	1.276	0.730	1.747	0.081
Phys 122	0.607	0.569	1.067	0.286
Transfer GPA	0.972	0.399	2.435	0.015
Campus Resident	0.459	0.391	1.174	0.240
Test 3				
Variable	Coefficient	Std. Error	z-value	p-value
Intercept	-3.535	1.328	-2.662	0.008
In State Student	-0.531	0.428	-1.239	0.215
Not Hispanic or Latino	-0.438	0.500	-0.875	0.382
First Generation	-1.727	0.548	-3.152	0.002
Math 101	-0.714	0.918	-0.777	0.437
Math 103	-0.633	0.948	-0.668	0.504
Math 104	-0.389	0.794	-0.489	0.625
Math 131	-0.165	0.679	-0.242	0.808
Math 132	0.826	0.675	1.224	0.221
Chem 121	0.509	0.545	0.933	0.351
Chem 122	0.338	0.532	0.635	0.525
Phys 121	0.125	0.776	0.161	0.873
Phys 122	0.838	0.555	1.509	0.131
Transfer GPA	0.983	0.369	2.664	0.008
Campus Resident	0.093	0.391	0.239	0.812

Table 4.12: Model 1, Test 2 and 3 for 4 year graduation

Only Transfer GPA is significant in both tests and it is the only statistically significant variable in test 2. In test 3, first generation is also statistically significant. The ROC curves are in figure 4.8 and 4.9.

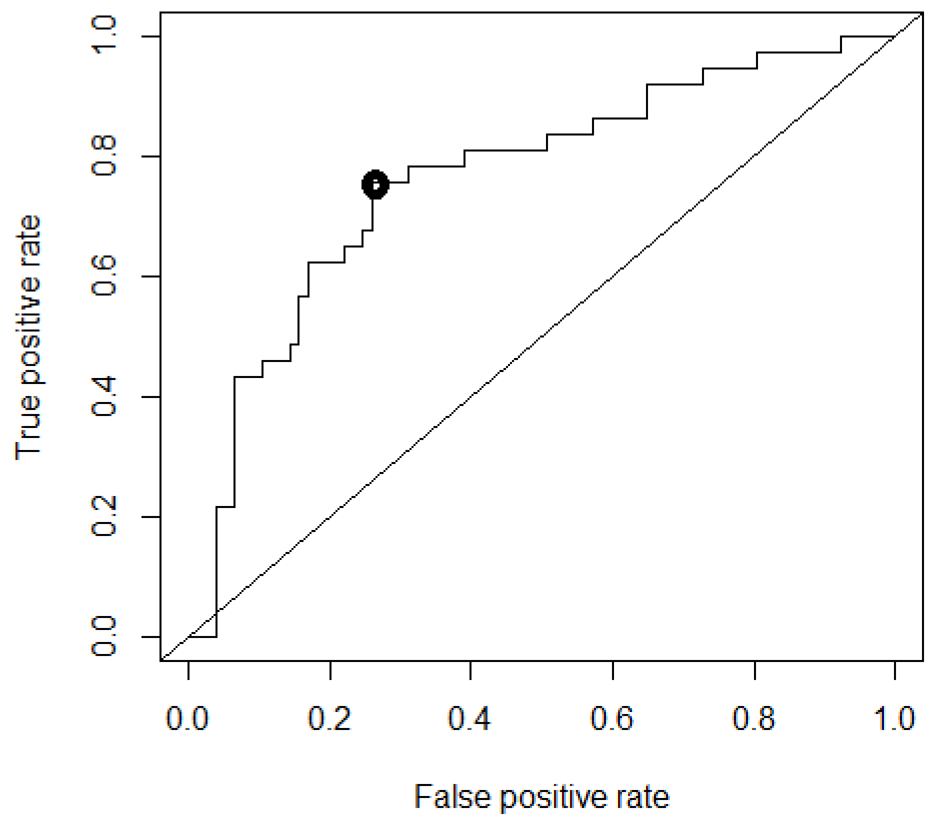

Figure 4.8: ROC Curve for Model 1, Test 2 for 4 year graduation

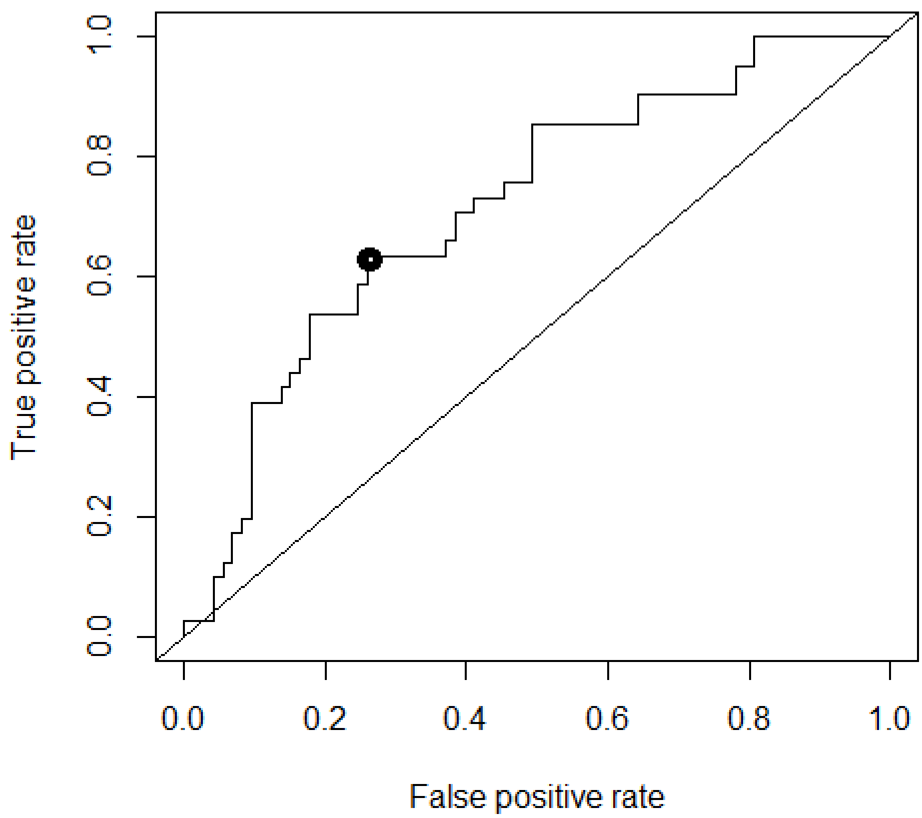

Figure 4.9: ROC Curve for Model 1, Test 3 for 4 year graduation

The area under the curve is 76.59% and 71.70% for test 2 and 3 respectively. With the TPR of 0.757 and FPR of 0.260 for test 2. The cutoff was 0.492. The TPR was 0.634 and the FPR was 0.260 for test 3. The cutoff was 0.394. Test 2 correctly predicted 54 and test 3 had 44 correctly predicted. Test two had 15 TP and 39 TN, and test three had 10 TP and 34 TN. For test 2, the TPR was 0.750 and the FPR was 0.093, and for test 3, the TPR was 0.500 and the FPR was 0.209.

4.6.2 Model 2

Based off the model above, a model using Transfer GPA, campus residence, and first generation because they were significant, and highest math transfer and highest chemistry transfer because they had small p-values.

Variable	Coefficient	Std. Error	z-value	p-value
Intercept	-5.989	1.247	-4.072	4.66×10^{-5}
First Generation	-1.072	0.528	-2.029	0.042
Math 101	0.404	0.871	0.463	0.643
Math 103	0.371	0.979	0.397	0.705
Math 104	0.772	0.773	0.999	0.318
Math 131	1.176	0.730	1.610	0.107
Math 132	2.118	0.718	2.951	0.003
Chem 121	0.194	0.588	0.329	0.742
Chem 122	0.674	0.576	1.171	0.242
Transfer GPA	1.206	0.386	3.125	0.002
Campus Resident	0.282	0.391	0.721	0.471

Table 4.13: Model 2, Test 1 for 4 year graduation

In this model, first generation, Math 132, and transfer GPA are significant variables. Transfer GPA and having math 132 are good indicators of graduating in 4 years. First Generation has a negative coefficient meaning being a first generation student is an indicator of the student has a good chance of not graduating. The ROC curve is in figure 4.10.

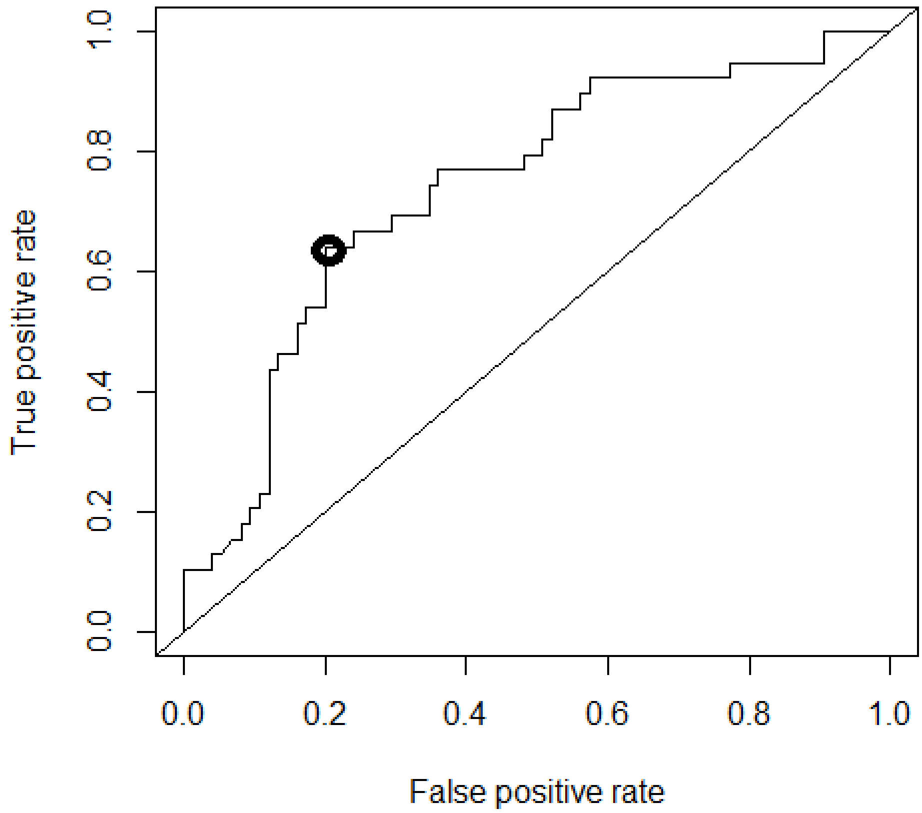

Figure 4.10: ROC Curve for Model 2, Test 1 for 4 year graduation

The area under the curve is 74.10%. The TPR is 0.667 and the FPR is 0.240. The cutoff is 0.396. The final test predicted 48 correct. 14 true positive and 34 true negatives with a TPR of 0.700 and FPR of 0.209. Two more tests were done.

Test 2				
Variable	Coefficient	Std. Error	z-value	p-value
Intercept	-6.849	1.518	-4.512	6.43×10^{-6}
First Generation	-1.344	0.525	-2.558	0.011
Math 101	-0.457	0.954	-0.479	0.632
Math 103	-0.245	0.955	-0.257	0.797
Math 104	0.096	0.784	0.122	0.903
Math 131	0.187	0.719	0.260	0.795
Math 132	1.684	0.682	2.470	0.014
Chem 121	0.595	0.0.567	1.049	0.294
Chem 122	0.902	0.551	1.638	0.101
Transfer GPA	1.574	0.418	3.767	0.0002
Campus Resident	0.389	0.395	0.988	0.323
Test 3				
Variable	Coefficient	Std. Error	z-value	p-value
Intercept	-6.632	1.550	-4.279	1.88×10^{-5}
First Generation	-0.937	0.491	-1.908	0.056
Math 101	0.179	1.002	0.179	0.858
Math 103	-0.471	1.265	-0.372	0.710
Math 104	0.433	0.861	0.503	0.615
Math 131	0.851	0.819	1.038	0.299
Math 132	1.657	0.800	2.070	0.038
Chem 121	1.073	0.603	1.178	0.075
Chem 122	0.687	0.583	1.178	0.239
Transfer GPA	1.395	0.403	3.460	0.0005
Campus Resident	0.153	0.392	0.389	0.697

Table 4.14: Model 2, Test 2 and 3 for 4 year graduation

Transfer GPA and Math 132 are significant in both tests. In test 2, first generation is also significant. The ROC curves are in figure 4.11 and 4.12.

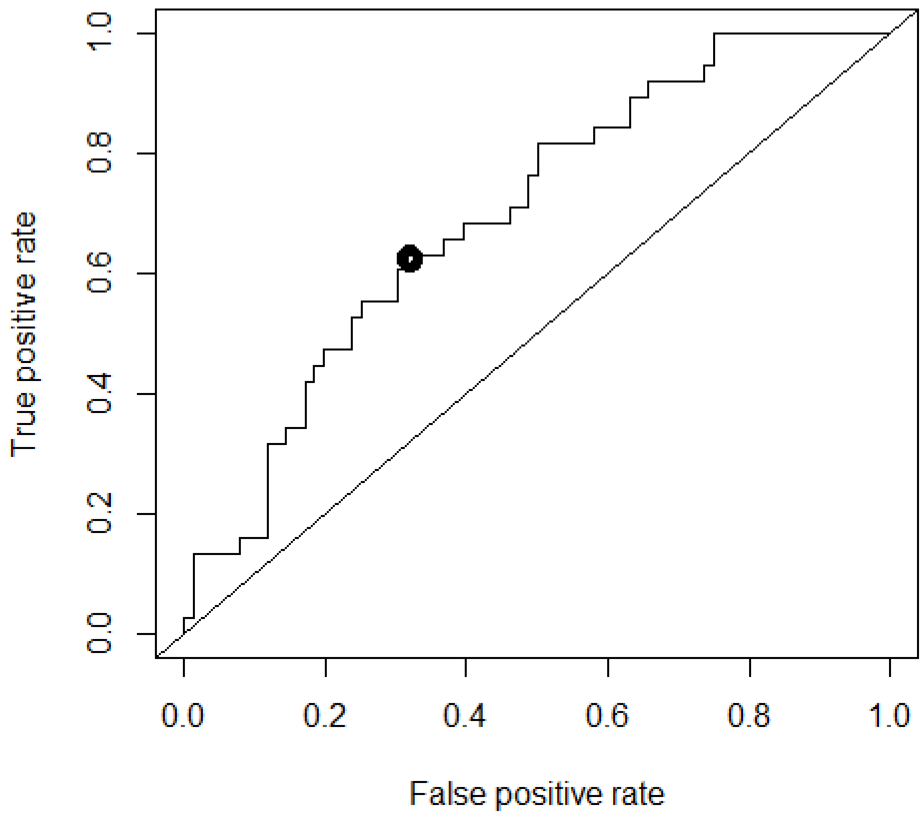

Figure 4.11: ROC Curve for Model 2, Test 2 for 4 year graduation

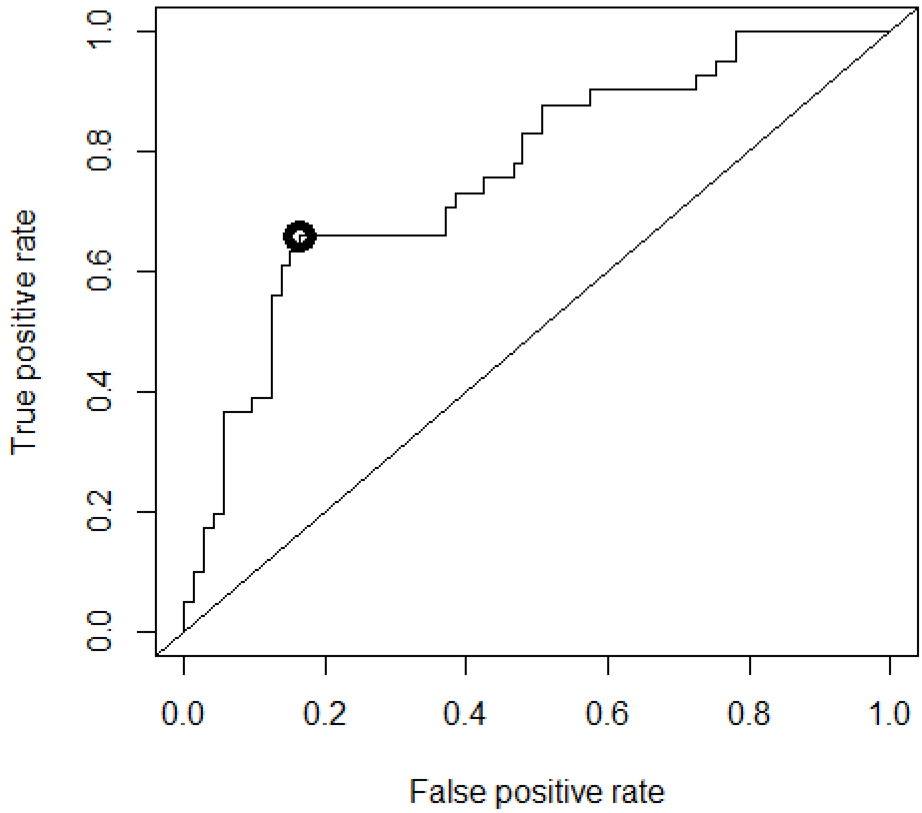

Figure 4.12: ROC Curve for Model 2, Test 3 for 4 year graduation

The area under the curve is 69.84% and 76.38% for test 2 and 3 respectively. The TPR was 0.632 and the FPR was 0.316 for test 2. The cutoff was 0.351.The TPR was 0.659 and the FPR was 0.164 for test 3. The cutoff was 0.378. Test 2 had 47 correctly predicted and test 3 had 44. For test 2, the TPR was 0.700 and the FPR was 0.233, and for test 3, the TPR was 0.550 and the FPR was 0.233. Model 2 did not have as great of a final test as model 1. The consistent thing from both models that we can gather for 4 year graduation is how much of an impact transfer GPA makes. In the second model, Math 132 was significant in every test and almost significant in every test for model 1. For 4 year graduation, transfer GPA and highest math transfer are the factors that affect graduation rate.

4.7 6 Year Logistic Regression Models

4.7.1 Model 1

With the same variables, a model for 6 year graduation was created.

Variable	Coefficient	Std. Error	z-value	p-value
Intercept	-4.268	1.285	-3.321	0.0009
In State Student	-0.246	0.392	-0.628	0.529
Not Hispanic or Latino	0.115	0.477	0.240	0.810
First Generation	-0.281	0.444	-0.632	0.527
Math 101	0.161	0.671	0.241	0.810
Math 103	-16.439	1112.954	-0.015	0.988
Math 104	0.912	0.731	1.247	0.212
Math 131	1.075	0.636	1.690	0.091
Math 132	1.331	0.668	1.991	0.046
Chem 121	-0.377	0.567	-0.665	0.506
Chem 122	-0.311	0.564	-0.551	0.582
Phys 121	0.015	0.662	0.022	0.982
Phys 122	1.102	0.537	2.052	0.040
Transfer GPA	0.945	0.328	2.882	0.004
Campus Resident	0.706	0.381	1.854	0.064

Table 4.15: Model 1, Test 1 for 6 year graduation

Math 132, Phys 122, and transfer GPA are significant variables for this model. Once again transfer GPA has a p-value less than 0.01, so it is a great factor that predicts 6 year graduation rate. Math 131 and campus residence are almost significant. The ROC curve is in figure 4.13.

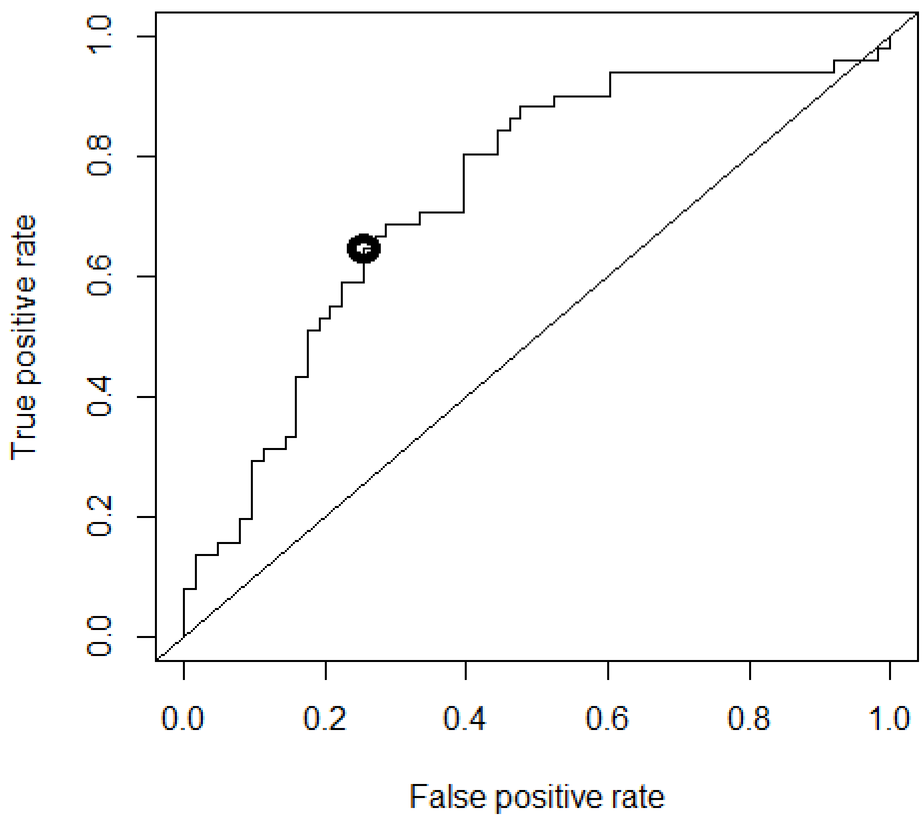

Figure 4.13: ROC Curve for Model 1, Test 1 for 6 year graduation

The area under the curve is 73.70%. The TPR is 0.686 and the FPR is 0.286. The cutoff was 0.470. For the final test, 40 cases were correctly predicted with 16 being true positive and 24 being true negative. The TPR and FPR are 0.571 and 0.314 respectively.

Two more tests were done with this model.

Test 2				
Variable	Coefficient	Std. Error	z-value	p-value
Intercept	-4.794	1.326	-3.614	0.0003
In State Student	-0.548	0.402	-1.362	0.173
Not Hispanic or Latino	0.313	0.467	0.671	0.502
First Generation	0.097	0.475	0.204	0.838
Math 101	-0.262	0.793	-0.330	0.741
Math 103	-1.315	0.947	-1.388	0.165
Math 104	0.959	0.662	1.448	0.148
Math 131	0.888	0.680	1.307	0.191
Math 132	1.525	0.666	2.288	0.022
Chem 121	0.413	0.525	0.786	0.432
Chem 122	0.081	0.522	0.156	0.876
Phys 121	0.143	0.702	0.203	0.839
Phys 122	0.101	0.560	0.181	0.857
Transfer GPA	1.114	0.369	3.023	0.002
Campus Resident	0.632	0.377	1.678	0.093
Test 3				
Variable	Coefficient	Std. Error	z-value	p-value
Intercept	-6.717	1.446	-4.644	3.41×10^{-6}
In State Student	0.105	0.390	0.270	0.787
Not Hispanic or Latino	0.494	0.458	1.078	0.281
First Generation	-0.455	0.434	-1.050	0.294
Math 101	0.344	0.703	0.490	0.624
Math 103	-0.601	0.807	-0.745	0.456
Math 104	-0.088	0.714	-0.124	0.901
Math 131	0.445	0.644	0.691	0.489
Math 132	1.223	0.700	1.748	0.080
Chem 121	0.369	0.572	0.646	0.519
Chem 122	0.011	0.547	0.020	0.984
Phys 121	0.555	0.686	0.809	0.419
Phys 122	0.431	0.601	0.717	0.473
Transfer GPA	1.625	0.393	4.135	3.55×10^{-5}
Campus Resident	1.024	0.380	2.695	0.007

Table 4.16: Model 1, Test 2 and 3 for 6 year graduation

In both tests, transfer GPA is a significant variable. In test 2, Math 132 is significant and almost significant in test 3. In test 3, campus residence is significant and almost significant in test 2. The ROC curves are in figure 4.14 and 4.15.

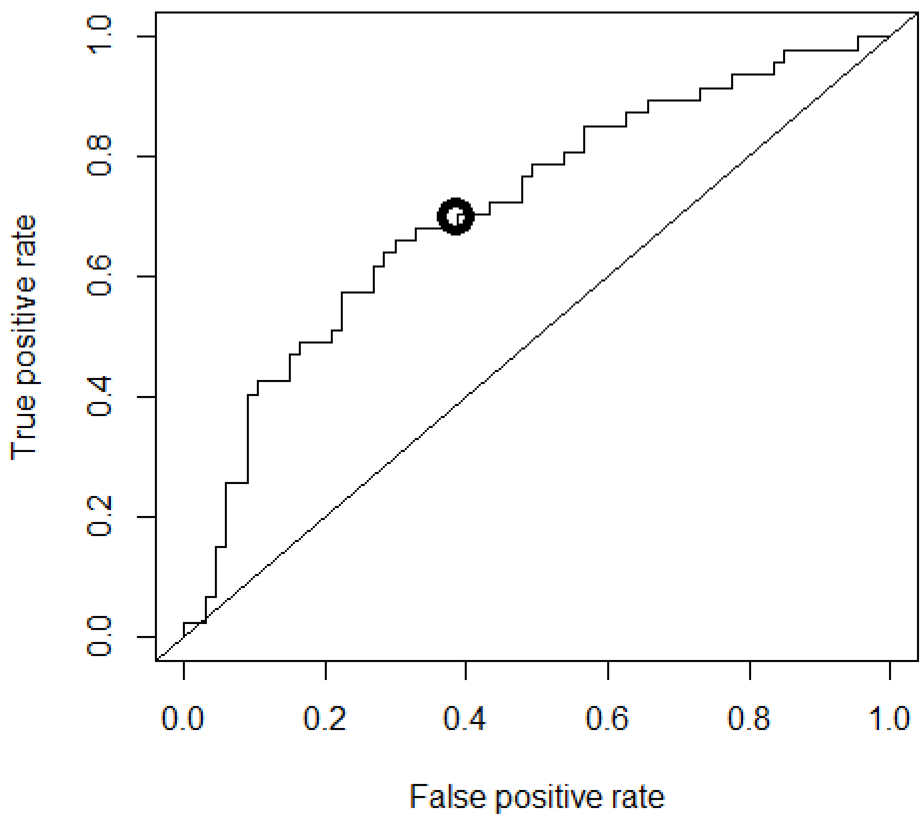

Figure 4.14: ROC Curve for Model 1, Test 2 for 6 year graduation

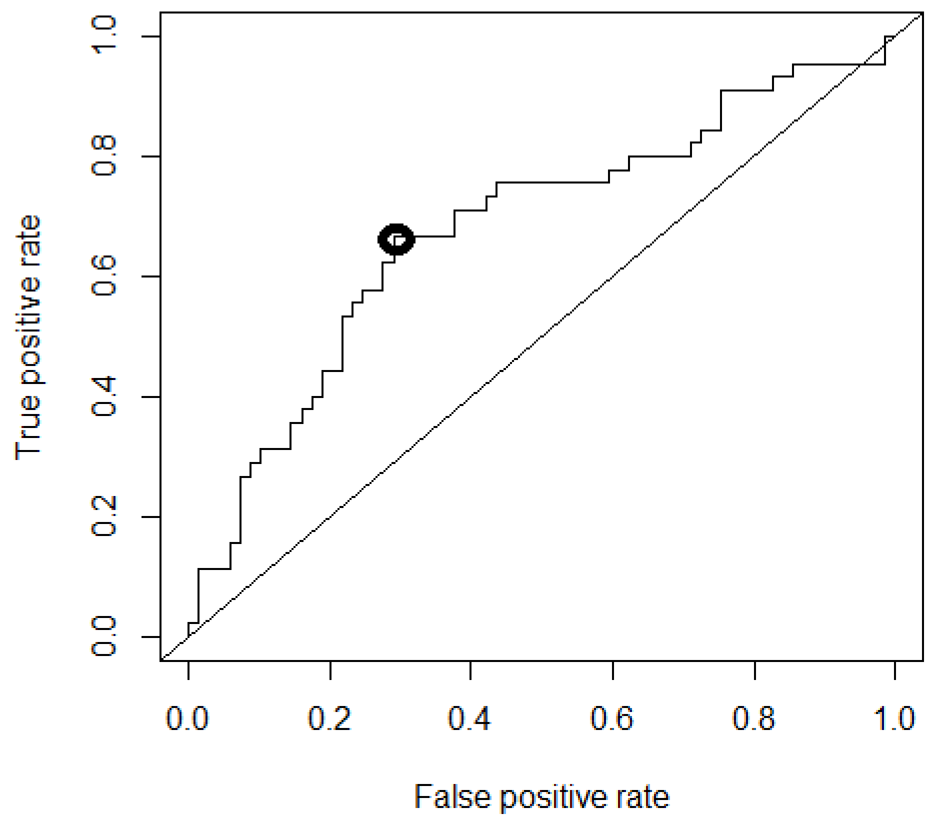

Figure 4.15: ROC Curve for Model 1, Test 3 for 6 year graduation

The area under the curve for test 2 was 71.42% with a TPR of 0.660 and a FPR of 0.299. The cutoff was 0.550. The area under the curve for test 3 was 68.05% with a TPR of 0.667 and a FPR of 0.290. The cutoff was 0.597. Test 2 predicted 37 correctly with a TPR of 0.536 and a FPR of 0.371. Test 3 predicted 39 with a TPR of 0.429 and a FPR of 0.229.

4.7.2 Model 2

Using model 1, a new model for 6 year graduation was created using transfer GPA, highest math transfer, highest physics transfer, and campus residence.

Variable	Coefficient	Std. Error	z-value	p-value
Intercept	-5.4218	1.276	-4.248	2.16×10^{-5}
Math 101	-0.167	0.735	-0.227	0.820
Math 103	-1.383	1.141	-1.211	0.226
Math 104	0.971	0.640	1.518	0.129
Math 131	0.445	0.638	0.698	0.485
Math 132	0.921	0.638	1.442	0.149
Phys 121	1.140	0.737	1.546	0.122
Phys 122	1.293	0.579	2.232	0.026
Transfer GPA	1.255	0.360	3.487	0.0004
Campus Resident	0.520	0.366	1.422	0.155

Table 4.17: Model 2, Test 1 for 6 year graduation

Phys 122, and transfer GPA are significant variables for this model. Transfer GPA has a p-value less than 0.01, so it is a great factor that predicts 6 year graduation rate. The ROC curve is in figure 4.16.

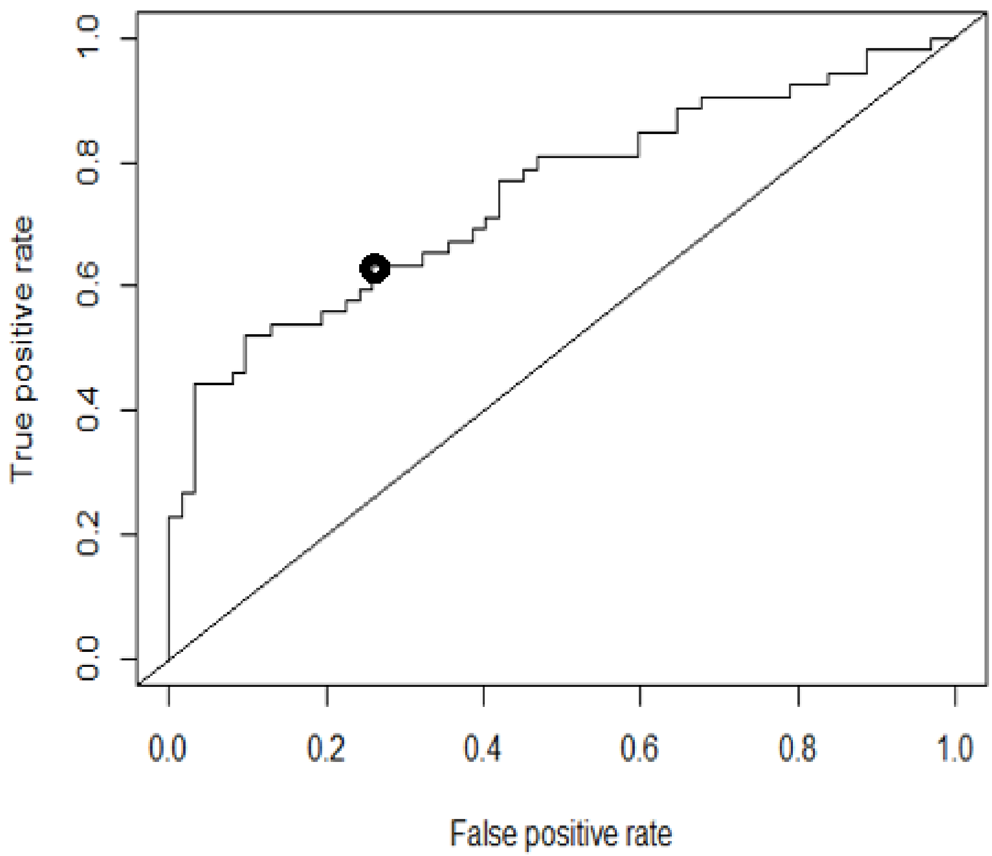

Figure 4.16: ROC Curve for Model 2, Test 1 for 6 year graduation

The area under the curve is 74.66%. The TPR is 0.635 and the FPR is 0.258. The cutoff is 0.473. For the final test, 44 cases were correctly predicted with 17 being true positive and 27 being true negative. The TPR and FPR are 0.607 and 0.229 respectively.

Two more tests were done with this model.

Test 2				
Variable	Coefficient	Std. Error	z-value	p-value
Intercept	-5.058	1.297	-3.900	9.63×10^{-5}
Math 101	0.351	0.682	0.516	0.606
Math 103	-0.831	0.879	-0.945	0.345
Math 104	1.345	0.638	2.108	0.035
Math 131	1.512	0.659	2.296	0.022
Math 132	1.516	0.572	2.650	0.008
Phys 121	0.400	0.681	0.585	0.558
Phys 122	0.641	0.506	1.267	0.205
Transfer GPA	1.039	0.354	2.933	0.003
Campus Resident	0.914	0.362	2.527	0.012
Test 3				
Variable	Coefficient	Std. Error	z-value	p-value
Intercept	-5.432	1.281	-4.242	2.22×10^{-5}
Math 101	-0.180	0.711	-0.253	0.801
Math 103	-0.367	0.818	-0.399	0.690
Math 104	0.591	0.682	0.866	0.386
Math 131	0.090	0.641	0.140	0.888
Math 132	1.020	0.610	1.673	0.094
Phys 121	0.663	0.691	0.960	0.337
Phys 122	0.678	0.529	1.280	0.200
Transfer GPA	1.315	0.364	3.609	0.0003
Campus Resident	0.739	0.356	2.077	0.038

Table 4.18: Model 2, Test 2 and 3 for 6 year graduation

Test 2 had multiple significant variables. Math 104, Math 131, Math 132, transfer GPA, and campus residence were statistical significant. In test 3, only transfer GPA and campus residence. The ROC curves are in figure 4.17 and 4.18.

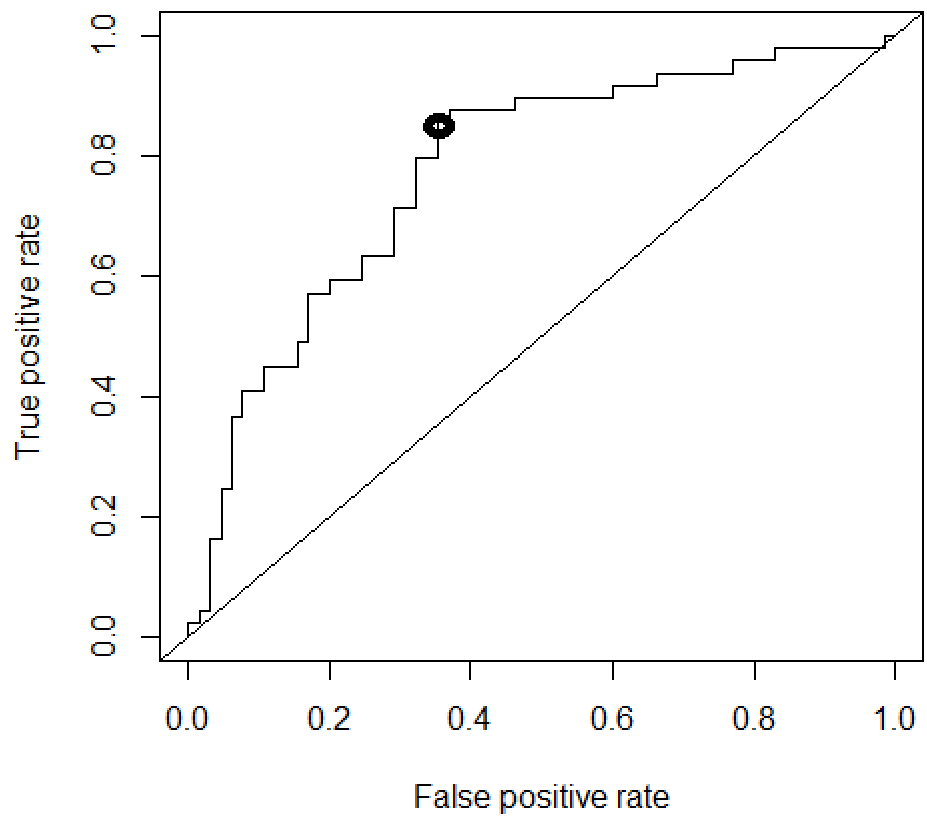

Figure 4.17: ROC Curve for Model 2, Test 2 for 6 year graduation

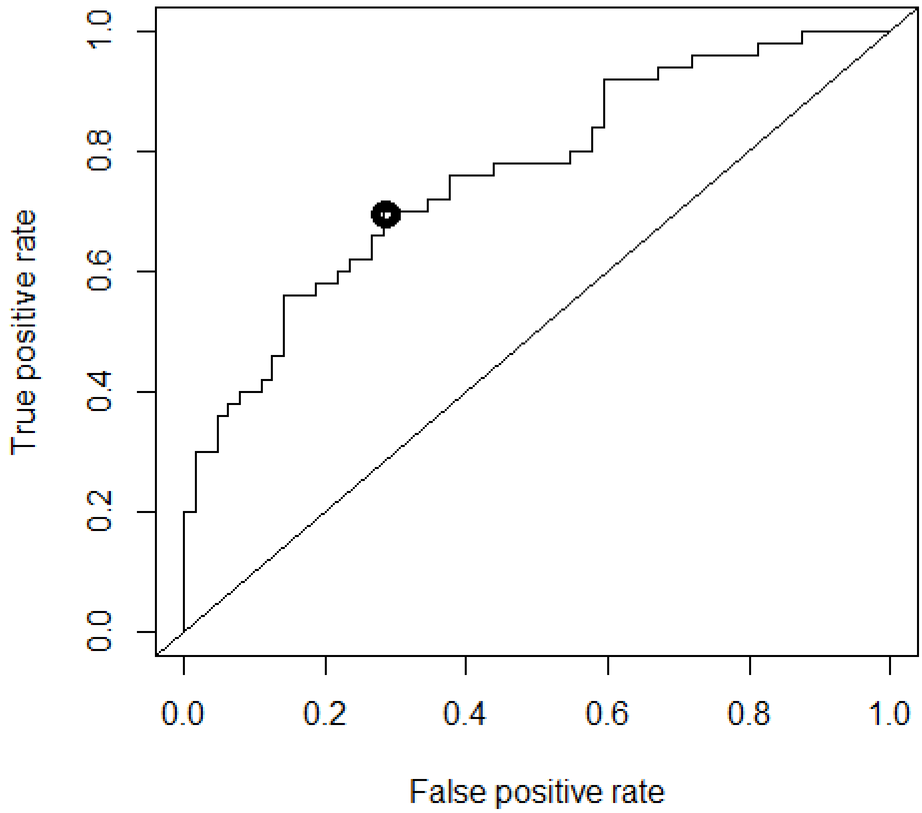

Figure 4.18: ROC Curve for Model 2, Test 3 for 6 year graduation

The area under the curve for test 2 was 77.55% with a TPR of 0.857 and a FPR of 0.354. The cutoff was 0.402. The area under the curve for test 3 was 76.31% with a TPR of 0.700 and a FPR of 0.281. The cutoff was 0.461. Test 2 predicted 38 correctly with a TPR of 0.750 and a FPR of 0.514. Test 3 predicted 36 with a TPR of 0.500 and a FPR of 0.343. This model shows that transfer GPA is a big factor for graduation in 6 years. Highest math transfer and campus residence is also a factor for 6 year graduation.

CHAPTER 5

CONCLUSION

The goal of the study was to answer the question how well can we predict whether or not a transfer student will graduate at New Mexico Tech and what factors affect graduation rate? The reason this question was asked was because in the past only freshman were ever studied at New Mexico Tech and factors that predict transfer student success have never been looked into. The data that this study used came from the registrar in a report generated by a software package called Argos. The report used consisted of every incoming student since 2006. This study specifically looked at transfer students that entered New Mexico Tech in the fall of 2006 to the fall of 2011. Logistic regression was used to test different data fields and to find which fields affect graduation rate the most. The model was build with 60% of the transfer students from 2006 to 2010. Once a model was created, a cutoff was found using the remaining 40% of the 2006 to 2010 students. The model was then tested on the 2011 incoming transfer students. ROC curves were used to determine how good the model was.

So can we predict whether or not a transfer student will graduate at New Mexico Tech? Yes and no. There are variables that are helpful, but also there are some things that cannot be predicted. The models to predict 3 year graduation rate could not give a good predictor. The models for 4 year and 6 year were more consistent and showed that last transfer GPA, highest math transfer and campus residence affect graduation rate.

New Mexico Tech is a science and tech school and tends to be harder than the average university, so last transfer GPA makes sense to why it can help predict the success of transfer students. Students that tend to do better at a previous university would have the study habits that could carry over once they transfer over.

With New Mexico Tech being a science and tech school, highest math transfer would make sense. All students have to take up to Math 132 at New Mexico Tech. At New Mexico Tech, the course programs are set up for students to take Math 131 and Math 132 their freshman year and are also prerequisites to many classes, so transfer students that do not have Math 131 or Math 132 credit are at the same point as freshman. New Mexico's Tech graduation rate for freshman is 48%, so it can be concluded that transfer students graduation rate if they do not have Math 131 or 132 could be about as freshman graduation rate. Also a transfer student that is at the same time frame as a freshman could make the transfer student discouraged because they have already spent time at another school and probably did not plan to transfer to New Mexico Tech and be there for another 4 to 6 years.

Campus residence as a factor is surprising. The models show that transfer students that live on campus in their first semester are predicted to do better than those that do not. This could be because it might be easier to develop study habits and finish work when a student is on campus.

Highest transfer chemistry and physics were sometimes factors in some models but not a lot of them. This could have the same affect as math transfer but unless the student is a chemistry or physics major, they do not need necessary need to get general chemistry and general physics done right away and are not as far back as if they did not have the math credit.

Most models were only able to correctly predict about 2/3 to 3/4 of the testing data. Most of the correct predictions were correctly predicting which students did not graduate rather than the students that did graduate. That means there are students that have the all the correct pieces to graduate but end up not graduating. This can also be seen in the breakdown where in the years of 2006-2010 only 55% of the transfer students with a GPA of 3.5 or higher graduated. This means that there are unknowns to why a student does not graduate. The unknowns could be personal things that vary from student to student for example motivation.

A problem with the study that can be fixed in the future is the size of the data set. The model was being built with a small number of data and because of that the models were inconsistent. The models were going to be different no matter what because of the random split between training and validation sets, but because the data was small, there is a higher chance for the models to be different.

This method could also be tested on the whole cohort of students or maybe just the freshman to see how similar the model compares to the past studies.

In summary, the factors that affect transfer graduation rate are a student's transfer GPA from the previous university, the highest math transfer credit, and campus residence. There is not enough data to build a great model but a somewhat useful model can be built to predict graduation.

REFERENCES

[1] Evisions. https://evisions.com/solutions/argos/

[2] Fawcett, Tom. *An introduction to ROC analysis*. Pattern Recognition Letters. Volume 27. Issue 8. December 19, 2005. Pages 861-874.

[3] Fuchs, Kirill. *Machine Learning: Classification Models*. March 28, 2017. https://medium.com/fuzz/machine-learning-classification-models-3040f71e2529

[4] Hegre, Havard. *Logistic Regression: Binomial, Multinomial and Ordinal*. September 23, 2011. https://havardhegre.files.wordpress.com/2014/03/logisticregression2011.pdf

[5] Hossler et al. (2012). *Transfer & Mobility: A National View of Pre-Degree Student Movement in Postsecondary Institutions*. National Student Clearinghouse Research Center. February 2012.

[6] James, Gareth (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

[7] Luna, J. (2000). *Predicting student retention and academic success at New Mexico Tech*. Master's Thesis. New Mexico Tech.

[8] Ramsey/Schafer. *The Statistical Sleuth*. 3rd Edition. Brooks/Cole Cengage Learning. 2013.

[9] Shlanta, N. I. (2008). *Statistical models of student retention at New Mexico Tech*. Master's Thesis. New Mexico Tech.

[10] Wu, G. (2009). *An examination of the impact of selected pre-college and college variables on student retention at New Mexico Tech*. Master's Thesis. New Mexico Tech.

Factors Affecting the Graduation Rate for Transfer Students at New Mexico Tech

by

Daulton Gray