Can Characteristics of Students Predict Completion of the College-Now Program?

Written by: Nick Paternostro

Abstract

Administrators have been posed with a difficult decision when it comes to admitting stents to the College Now program. While they hope to gain a potential graduate candidate after completing the program, it is difficult to predict if first-year students, who are technically not admitted as full-time students, will complete the program without knowing their characteristics. Since the University of Massachusetts Dartmouth spends large amounts of funds and resources in developing the students in this program, administrators are interested in predicting whether a student will complete the program based on the following categories: (1)personal characteristics, (2)psychological characteristics, (3)student behavior, or (4)academic performance. Based on the results of a psychological study conducted by an outside group, a model has been created to classify the probability of a student completing the course based on different traits. First, a study was conducted to determine the proportion of students who have completed the course for each predictor variable. These results were plotted and an S-shaped curve was created. Logistic regression was then used to model the probability of the student completing the program based on the predictor variables. The goal of creating a logistic model is to estimate the probability a student completes the course based on a linear combination of the predictor variables. Using single-variable logistic regression, logistic models were generated for each predictor variable independently. Then using multi-variable logistic regression, models for different combinations of predictor variables were created. First, all predictor variables were divided into the four subcategories listed above and a model was created for each category. The logistic model using student behavior variables as predictors generates the most accurate model. Then, a systematic approach, based on intuition, was used to build logistic models with a high percentage of accurate predictions but a small number of predictor variables. Excluding all psychological variables, thus ridding the need for the psychological study, a logistic model was created with roughly 90% accuracy. Administrators can use this model to classify the probability a student will complete the program based on the predic-

tor variables utilized. Lastly, a brief gender study was conducted to determine which gender has a high proportion of students completing the program.

Introduction

The college now program offers at a wide range of universities across the United States allows prospective students the opportunity to attend college even if they are not academically inclined due to social or personal difficulties. The University of Massachusetts Dartmouth offers an alternative admissions program for students who may be questioning whether they are “college material”. College Now gives students with academic disadvantages the opportunity to be a full-time students with the rights, privileges, and responsibilities of a normal student at the University of Massachusetts Dartmouth. The program allows students to enter the university after completing a semester in a preliminary program. Due to the amount of time spent with these students preparing them for their future, the university exhausts a handful of resources for the program to succeed. Whether it be academic counseling, instruction, or individual attention, the university spends serious funds on the program and all its members. Since accepting a student into the College Now program is a serious investment, administrators are interested in predicting whether students will complete the program based on the students’ answers to some questions. Using physiological and personal characteristics, previous academic performance, and student behavior (more accurately their future behavior) as predictor variables, a study can be conducted to determine the effect each predictor has on the desired outcome, competing for the College Now program. Logistic regression is used to model probabilities of binary outcomes given values of predictor variables. Using logistic regression will allow administrators to classify whether students will complete the course based on predictor variables.

Methods

By generating a logistic regression, a model will be generated to classify the probability of a student completing the program based on the values of independent variables. Each predictor variable lies under one of the four categories: personal characteristics, psychological characteristics, student behavior, or academic performance. Each predictor was evaluated independently using a single variable logistic regression. Then four categories were created, listed above, and each variable was placed in the category corresponding to the type of the variable shown in Table 1 below. NOTE federal ethnic group and gender were not used for any of the following studies, but are listed in Table 1.

Table 1: Four Sub-Categories for Independent (Predictor) Variables

Out[•]=

Personal Characteristic	Psychological Characteristic	Student Behavior	Academic Performacne
Gender	Dropout Proneness	Attended Orientation	High School GPA
Ethnic Group	Predicted Academic Difficulty	Completed Summer Bridge	SAT Score
Athlete	Educationla Stress	Attended Experience Day	Cumulative GPA
Resident/Commuter	Receptivity to Institutional Help	Completed Campus Event Requirement	Number of Credits Earned
	Receptivity to Academic Assistance	Completed the Community Service Requirement	
	Receptivity to Personal Counseling	Number of Faculty Advisor Meetings	
	Receptivity to Social Engagement	Number of Peer Mentor Meetings	
	Receptivity to Career Guidance	Number of Workshops Attended	
	Receptivity to Finanical Guidance		

Regression methods are utilized in scenarios where it is desirable to describe the relationship between a response variable and one or more independent variables. All forms of regression are interested in finding the best-fitting model that has practical use in the real world. Logistic regression is used to model the probability of binary outcomes given the value of predictor variables, similar to linear regression finds the relationships for a discrete outcome and the explanatory variables. Additionally, logistic regression can (1) estimate the probability that an event occurs for a randomly selected independent variable, (2) predict the effect of a sequence of variables on a binary response variable, (3) and classify observations by estimating the probability that a predictor variable is in a particular category. The logistic model generates a model that will try to fit a linear model between the predictor variable and the logarithm of the “odds”: log (p/(1-p)) where p is the probability of the binary outcome given a particular sequence of values of the predictor variables.

First, to generate an understanding, the summation of all values of the independent variables in the student behavior(except attended orientation/experience day) was taken as the predictor variable. The predictor variable ranged from 2-19, and each predictor variable was treated as one interval. For each predictor variable, the number of students who completed the program and proportion of student who completed the program was found and plotted in Figure 1. When generating logistic models, the S-shaped curve created in Figure 1 should be replicated. The quantity p/(1-p) is known as the odds of success given a predictor variable (or sequence of predictor variables). Using a logistic model, one can

determine if it is a good fit for the data based on whether it accurately predicts success or failure in the program. If the model is accurate, it should follow the S-shaped pattern presented in Figure 1.

The data provided by the university was first categorized into the groups of Table 1. Since some students failed to complete the psychological survey, the data had some variables without any values. To correct this, all students who had no numerical value for an independent variable were ignored and removed from the sample set for the given study. However, removing multiple individuals from the study will cause the sample size to be smaller, leading to a large variation in the results when outliers are presented. An alternative option instead of removing the individuals from consideration is to have a replace the empty variable with 0; however, before proceeding, one should understand the limitations and assumptions made when altering the data set. For each study, the students with blank independent variables were removed from consideration and new sample size was taken. With the new sample set, the independent variables with the corresponding outcome variable for each student were loaded into Mathematica. Using Mathematica's LogitModelFit, a logistic model can be generated for the given independent variables. For a single-variable logistic regression, the predictor variable is imported with the binary outcome variable. The model is generated using a single independent variable and evaluated for the single predictor variable. This will generate an estimated probability that a student completed the course and that will be compared with the actual outcome for that student. For the multivariable logistic regression, multiple predictor variables will be utilized to generate a logistic model. Instead of using LogtiModelFit for one variable, multiple independent variables will be used as the predictor variables to generate a logistic model. This logistic model will contain multiple independent variables in the equation, each corresponding to a predictor variable. Then, for each student, the model is used to get an estimated probability that the student competed in the course for the given predictor variables of the student. Following that process, the accuracy of the models can be found based on how many accurate predictions the model made relative to the actual binary outcome. The model produced an estimated probability of completion and this value was compared with the actual outcome whether the student completed the course or not. For the predicated value generated by the model to be counted as, either the estimated probability of successful completion is greater than 0.5 and the student did complete the program or the estimated probability of successful completion is less than 0.5 and the student did not complete the program. Using the results from the accuracy, a percentage of accurate predictions can be found. Keep in mind, the probabilities of successful completion are close to 0.5, therefore there is no surprise if the probability is near 0.5, the model fails.

Once the accuracy of the logistic model is quantified for one set of predictor variables, different combinations of predictor variables were tested to determine if the accuracy of the logistic model could be maximized based on the sequence of predictor variables. First, all variables listed in Table 1 were taken into consideration except the variables listed under academic performance. This creates a baseline, multivariable logistic model. However, the goal is to minimize the number of variables needed to generate an accurate model. Therefore a systematic approach is used to slowly lower the number of variables but still maintain a high level of accuracy. Iterations tried to shrink the number of predictor

variables, thus each successive iteration eliminated a predictor variable from the baseline model. By limiting the predictor variables, one can find that there may a model where there is no need for the psychological variables and can eliminate the expense the university spends on obtaining these variables. Each iteration was created based on intuition for which parameters would be beneficial in classifying if a student would complete a course.

Results

Summation of Predictor Variables

The easiest way of creating a logistic regression involves summing the independent variables and computing the proportion of success for intervals. By creating intervals for the independent variable and computing the average value of the success variable for a value of the predictor variable, some variation in the data can be removed. Figure 1 represents the proportion of success for a given predictor variable. As can be seen, by the shape and the visualizations of the model, the goal of a logistic model is to generate an S-curve that fits a shape like this. This figure demonstrates what the average value of the success variable is for a given value of the predictor. The S-shaped curved generated by plotting this data is a baseline for how the logistic model should be generated. By plotting the model against the real values, it becomes easier to visualize how a more accurate model follows a sharper version of the S-curve built in Figure 5. Using these curves, a conclusion can be made whether the logistic model for a given predictor variable is an accurate fit for classifying whether the student has completed the course.

Table 2: Proportion of Students Completed the College Now program based on the Summated Multi-Predictor Variables

Out[•]=

Predictor	N	Number of Failure	Number of Success	Proportion of Success
2	2	2	0	0
3	1	2	0	0
4	2	2	0	0
5	3	3	0	0
6	6	5	1	0.166
7	5	3	2	0.4
8	7	4	3	0.429
9	5	2	3	0.6
10	11	3	8	0.727
11	7	2	5	0.714
12	15	2	13	0.866
13	10	2	8	0.8
14	9	0	9	1.0
15	15	1	14	0.933
16	3	0	3	1.0
17	3	0	3	1.0
19	1	0	1	1.0

Figure 1: Student Behavior Predictor Variables, except attendances vs. Proportion of Success

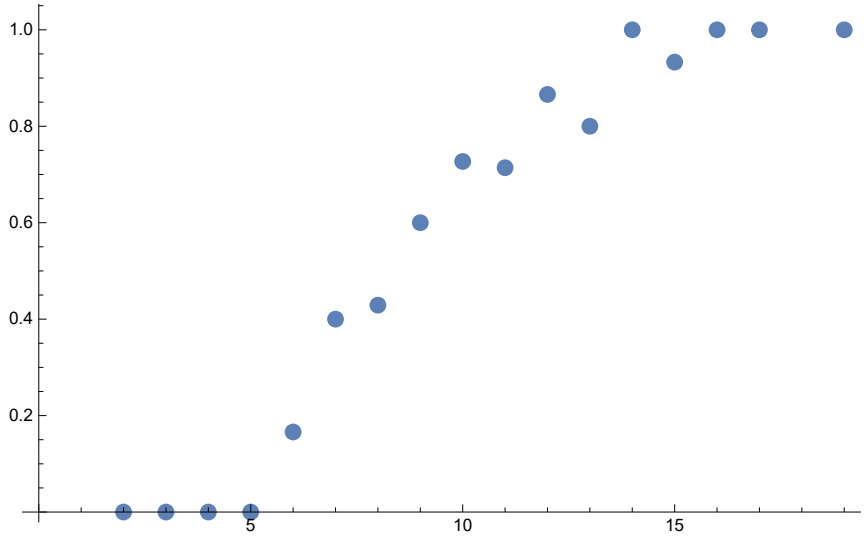
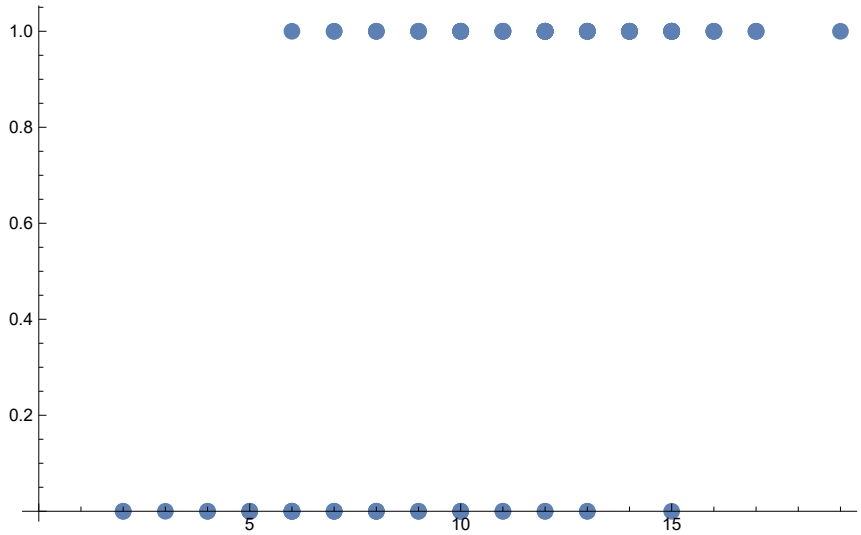


Figure 2: Student Behavior Predictor Variables, except attendances, and Students completing the course



There is some tendency for the individuals who have not completed the College Now Program to have low predictor values whereas individuals who have completed the program typically have higher predictor values. There is some difference in this data set, although, for some high predictor values, such as above 10, there are still some individuals who have failed in completing the course. The overlap of students will make predicting the success/ failure based on the predictor more difficult. Since there is no clear cut-off between the range of predictor values from 6-13, the odds of a student in the predictor range finding success in completing the course is roughly 1:1, that is the odds are somewhat even. Since the odds are even, it is difficult to predict whether any given student in the range of predictor values of 6-13 has succeeded/failed in completing the course.

Single Variable Logistic Regression

Single-Variable Logistic Regression was applied to all independent variables. The results from this study were used to guide the optimization study, however, it was found that using intuition is better. Listed below are some of the best predictor variables for classifying the probability of a student completed the program. Single-variable logistic regression is not the best approach for generating an accurate model. A stronger emphasis is placed on multi-variable logistic regression as it will produce a more accurate model. However, single-variable logistic models may be useful if administrators are interested in finding one predictor variable that can accurately predict whether a student completed the program.

Figure 3: Logistic Model (dashed) Based on Cumulative GPA

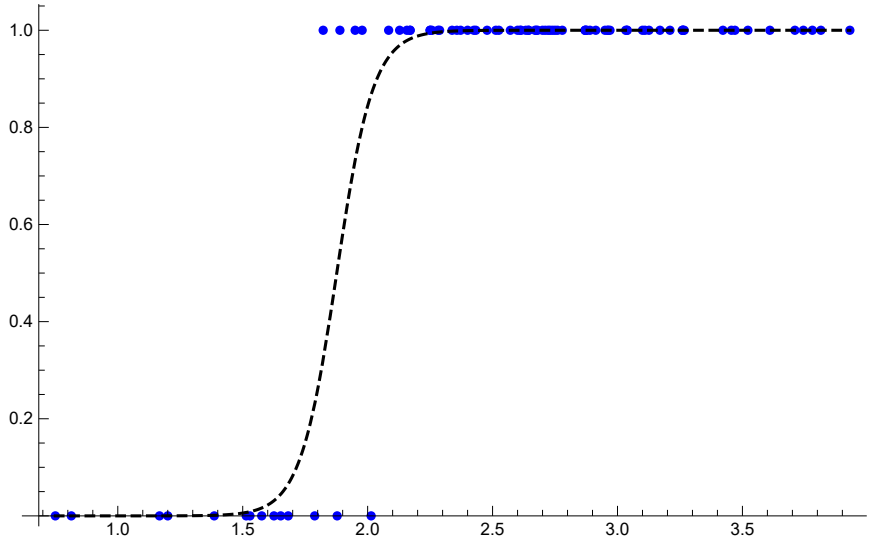


Figure 4: Logistic Model (dashed) Based on Number of Peer Mentor Meetings Attended

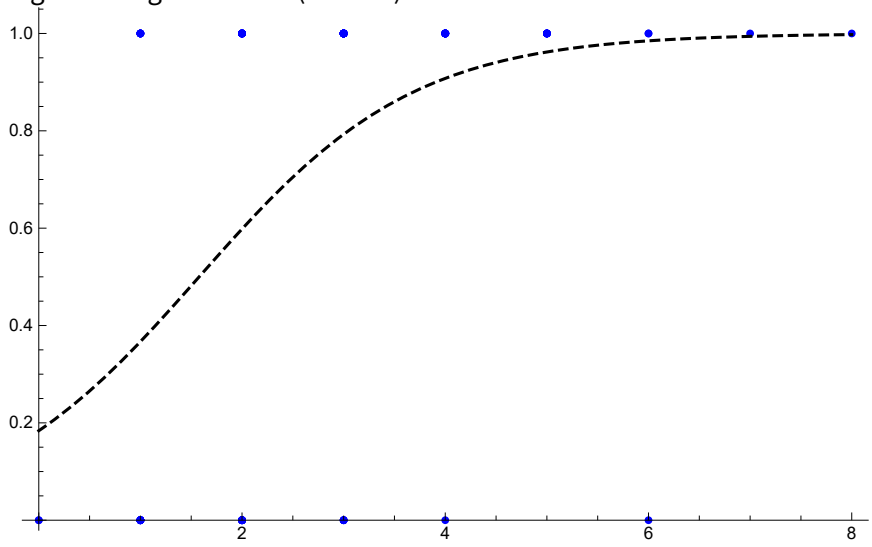


Figure 5: Logistic Model (dashed) Based on Number of Workshops Attended

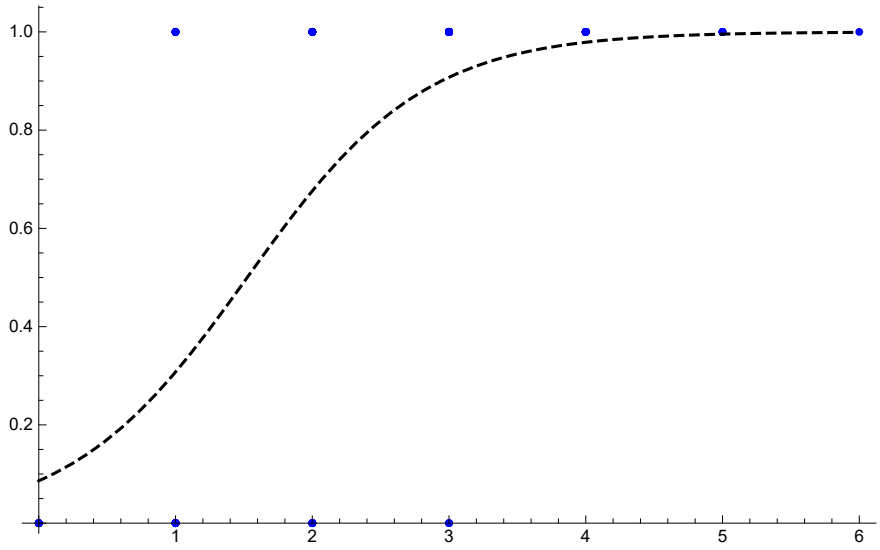


Figure 6: Logistic Model (dashed) Based on Completing Community Service requirement

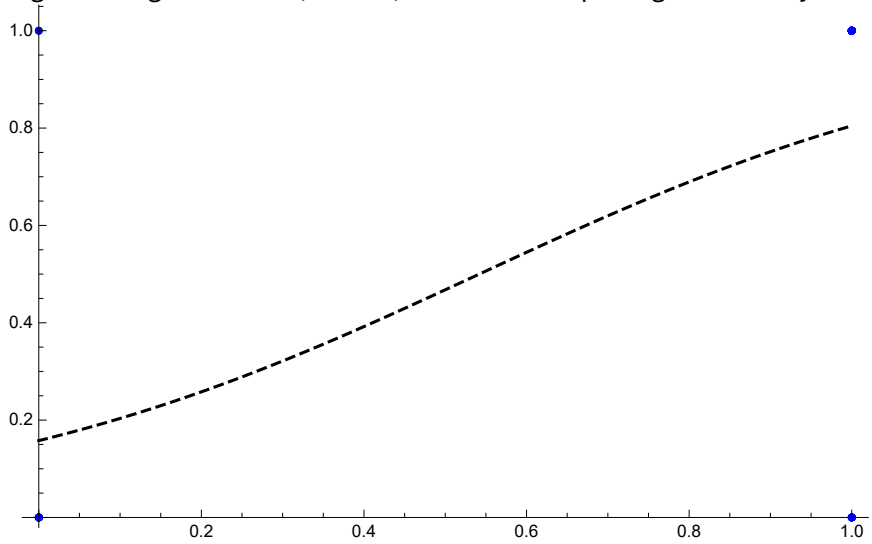


Table 3: Percentage of Accurate Predictions for Given Logistic Models

Out[•]=

Independent Variable	Percentage of Accurate Predictions
Cumulative GPA	96.5%
Number of Peer Mentor Meeting Attended	77.0%
Number of Workshops Attended	81.0%
Completed Community Service Requirement	81.1%

Multi-Variable Logistic Regression

The goal of multi-variable logistic regression is to estimate the probability that a student will complete the College Now program for a linear combination of the predictor variables. While single-variable logistic regression did provide understanding for how to utilize a logistic regression model, multi-

variable logistic regression will rely on more independent variables to generate an estimated probability, thus being more accurate in most cases. Using multivariable logistic models creates a model with multiple independent variables and then the probability that a student will succeed can be classified based on the predictor variables. Reported below is the total number of accurate predictions and percentage of accurate predictors from each logistic regression model (Table 4). To minimize the number of predictor variables but still create an accurate model was the purpose of Table 5.

Table 4: Model Accuracy for 4 sub-categories

Out[•]=

Category	Total Number of Accurate Predictions	Percentage of Accurate Predictions
Personal Characteristics	67	72.04%
Physcological Characteristics	69	74.19%
Academic Performance	70	68.63 %
Student Behavior	92	88.46%

Table 5: Table of Iterations to Limit Number of Predictor Variables

Out[•]=

Iteration	Variables Evaluated	Total Number of Accurate Predictions	Percentage of Accurate Predictions
1	All	89	97.8%
2	All – (Resident/Commuter and Athlete)	89	97.8%
3	All – (Receptivity Variables)	78	85.7%
4	All – (Res/Com, Athlete,Attend)	88	96.7%
5	Student Behavior – (Attend)	82	90.1%
6	Without Physc and Acad	94	90.4%

Gender Study

The last study evaluated the role gender has when predicting if a student completes the course. Instead of using a logistic model, the probability of a student completing the course based on their gender was computed. From Table 5 below, given a student is female, they are more likely to complete the College Now program as opposed to if the student were male. This can be demonstrated in Table 5 since the proportion of success, or the average value of success given the gender is larger if a student was female as opposed to male.

Table 6: Gender Effects on Portion of Students Completing the Course

Out[•]=

Number of Female	41
Number of Male	65
Number of Female Completed	30
Number of Male Completed	44
Probability of Female Completing the Course	0.732
Probability of Male Completing the Course	0.676

Discussion and Conclusion

When administrators admit a student to the college-now program, they are investing. First, the student will complete a year of school attending courses and participating in the college now program, proving they are suitable for attending the university of interest. Then, once the course has been completed, the student becomes a full-time student and member of the campus community. However, if the students fail to complete the course, the university has wasted a serious amount of time and money. Therefore, administrators are interested in predicting whether a student will complete the course based on the results of a psychological study. To model the probability of a student completing the course (binary outcome variable) based on categorial/numerical independent variables, logistic regression was used. The results provided above will indicate which predictor variables are useful in classifying whether a student will complete the College Now program.

Before using single or multi-variable logistic regression, the summation of multiple predictor variables was used to get an idea of what a typical model curve should resemble. The summation of the predictor variables ranged from 2-19, creating a new set of independent variables. Intervals were created to treat each independent variable as a predictor variable and then find the correlating proportion of success for a given predictor variable where the proportion of success is the number of students who completed the College Now program relative to the total cohort of students in the program. When the plot of “Proportion of Success” and “Predictor” was built, the corresponding curve was shaped like an S. This curve demonstrates the average value of the success variable for a given value of the predictor variable. Later, when logistic modeling is utilized, the model plots will resemble an S-shaped curve; the more accurate the model, the sharper the curve and it almost resembles a unit step function.

When conducting a single variable logistic regression and evaluating each predictor variable independently, the cumulative GPA and number of credits earned is a very strong predictor of success in the College Now program(Figure 3). However, this is expected as the cumulative GPA and number of credits earned can be used to decide success in the course. Using predictor variables, the cumulative GPA can be predicted from other predictor variables. Therefore, the cumulative GPA dictates whether a student will complete the College Now program or not and cannot be used as a predictor variable. The predicator variables under the category of student behavior were the most accurate predictors. If the student had completed the community service requirement, roughly 81% of the predictions made by

the model are accurate predictions. For students who are truly interested in attending the university as a full-time student and want the opportunity to be a graduate candidate, they will surely complete the community service requirement while others who are not as committed may neglect this; thus, this predictor variable is an accurate predictor (Figure 6)Another accurate predictor variable is the number of workshops attended. When students attend more workshops, they are more likely to complete the program. Attending more workshops demonstrates engagement in the program by the students and correlates to a student's willingness to succeed(Figure 5). The number of peer mentor meetings attended also is a decent predictor variable which leads to accurate predictions for about 77% of the time (Figure 4). Interestingly, when the psychological variables were used as the predictor variable (i.e. educational stress, dropout proneness, receptivity to guidance and assistance, etc.) for the single variable logistic regression the accuracy is any of the predictor variables does not exceed 71%. That is, when any of the psychological variables are used as the predictor variable, the model is not highly accurate, thus whether a student will complete the College Now program cannot be accurately predicted based on these variables. Based on intuition, I would believe athletes are more likely to complete the program compared to non-athletes just due to motivation to stay in the program. However, this is not true; whether a student is an athlete is not an accurate prediction for completing the course. Similarly, based on intuition, I would expect students coming in with higher SAT scores to complete the program with ease. On the other hand, since College Now is for students with academic disadvantages, the students in the program may be motivated, but the grades do not coincide with the effort they may put in. The results seem to follow the latter since the SAT scores are not an accurate model for predicting whether students will complete the course. Instead of using a single variable to generate a logistic model and estimate the probability that a student will compete in the College Now program, multiple variable logistic regression will produce a more accurate model.

To gain a more accurate logistic model, multi-variable logistic regression is utilized. Several studies were conducted using multivariable logistic regression to find a combination of predictor variables that would lead to the most accurate model. Before using intuition or previous results from single-variable logistic regression, the independent variables were broken into four categories: personal characteristics, psychological characteristics, student behavior, or academic performance. As shown in Table 4, personal characteristics, such as whether the student was an athlete, a commuter or resident, or even the desire to transfer had roughly 72% accuracy when classifying the students' probability for completing the course. The psychological variables (Table 1) produce a slighter more accurate model. With about 74% accuracy, this model should be used with caution. Depending on how the administration wants to continue, this prediction could help determine if the psychology group hired to obtain these answers from the students is worth it. Using previous academic performances, such as high school GPA and SAT scores, another logistic model was created to determine if a student completing the course can be classified based on his/her high school GPA and SAT scores. Similar to the single-variable logistic regression case, the cumulative GPA and number of credits earned are not considered because these variables directly rely on whether the student completed the course. The last categorical group examined, before moving to intuition, was based on student behavior (Table 1). This cate-

gory consists of independent variables which the student has had an opportunity to complete. The logistic model generated for this scenario was the most accurate of all the models. Notably, whether a student completes the course can be classified with the student behavior by about 88% accuracy. By far, this model produces the largest amount of accurate predictions. However, depending on when the events take place, some variables cannot be predicted before the student's study year. Instead, some predictor variables (number of advisor meetings, number of peer mentor meetings, and a few other categories) are not taken until the end of the study year and cannot help predict the probability of completing the College Now program. Anyway, student behavior offered the most accurate model, but the administration could still utilize one of the other three categorical groups.

Once the most accurate logistic model of the four groups was found, a study was conducted to determine if the accuracy of the logistic model could be optimized. At first, all independent (predictor) variables, except the academic performance variables, listed in Table 1 were implemented into a multivariable logistic regression model. The results of successive iterations are shown in table 3. TO start, all categories listed in Table 1 were considered, except the variables based on academic performance since these variables generated a low accuracy logistic model when the multivariable logistic regression was conducted. As shown by Table 3, the accuracy of the multi-variable logistic model created when utilizing most of the predictor variables produced a model with 97% accuracy. However, since this model relies on all variables it is essentially "more expensive". That is, more predictor variables are required to generate such an accurate model. It should be of interest to lower the number of predictor variables and still obtain a model on the same order of accuracy. After the baseline model (iteration 1) was found, the predictor variables were re-examined to find which predictor variables were disposable. The next iteration eliminated the resident vs. commuter variable and the athlete variable. As expected, when a logistic model was generated for the second iteration, the accuracy did not vary. This may be because the variables removed have little effect on predicting the outcome compared to the other predictor variables, but this is just speculation. The next iteration dismissed some psychological variables, mainly the ones that dealt with receptivity to guidance and assistance. By eliminating some psychological variables and analyzing the accuracy for prediction, the need for an outside group to conduct a psychological investigation can be questioned. The model accuracy decrease by roughly 12%, a quite significant jump; however, with 85% accuracy, the independent variables still produce an accurate logistic model. The next iteration eliminated the attendance from the predictor variables but replaced the psychological variables removed in the previous step. The reasoning for removing the two variables including attendance is because these events took place in the summer. Not being able to attend an event during the summer, when school is in recess, should not be a serious factor when predicting whether the students completed the course. As expected, the accuracy of the model did not drop significantly, only going from 97.8% to 96.7%. Therefore, since eliminating the predicting variables that required attendance had little impact on the accuracy of the logistic model, these variables were removed from the categorical study on student behavior. This iteration only focused on the student behavior category, and it happens that when the predictor variables that required attendance were removed, the model became more accurate, going from

88.5% to 90.1%. Lastly, since the goal of this study is to generate an accurate model with the least amount of predictor variables, the entirety of the psychosocial variable category was removed from iteration 1. By removing all psychological variables, there is no longer a need for an outside group to come in and conduct a study to evaluate the psychological variables. When removing the psychological variables, the accuracy of the logistic model is 90.4%. While this accuracy s not as high as when all variables are examined in a multivariable logistic regression, this is accuracy level may raise questions as to whether the psychological variables are needed. Deciding whether to continue with the psychological studies comes down to the administration's decisions and a risk assessment. However, the results do demonstrate that an accurate model can be generated by classifying whether a study will complete the program based on variables other than psychological variables.

In future work, it would be appropriate to find use bootstrapping or a p-value to find if the model can or cannot be trusted for a given set of predictor variables. While all predictor elements were utilized independently in a single-variable logistic regression, not all permutations of predictor variables were taken when using multivariable logistic regression. Using intuition, different predictor variables were used to make a logistic model and then this model was tested for accuracy. However, in future studies, it may be beneficial to test all possible permutations of predictor variables to find the minimum number of independent variables needed to keep the accuracy above 90%. When conducting future studies, it is important to consider changing the range of estimated probabilities correlating to a success. Since a probability close to 0.5 has close to an equally likely chance of being a failure or a success, it is difficult to predict the outcome variable in this range. An additional error may arise from the fact that a smaller sample size is used since all students who had empty variables were rejected. Smaller sample size would allow the model to be easily persuaded by extremities.

References

[1]Hosmer, D.W, Lemeshow, S. (2000). Applied Logistic Regression Second Edition. John Wiley & Sons Inc.

[2] Nolan, D., & Speed, T. P. (2000). Stat labs: Mathematical statistics through applications. Springer Science & Business Media.

Scientific writing made easy: A step-by-step guide to undergraduate writing in the biological sciences. (2016, October 3). The Ecological Society of America.

[3] Rodríguez, G. (2007). Lecture Notes on Generalized Linear Models. Chapter 3: Logit Models for Binary Data.

[4] Fultz, B Video on Logisitic Regression

[5] Smith, T. J., & McKenna, C. M. (2013). A comparison of logistic regression pseudo R2 indices. Multiple Linear Regression Viewpoints, 39(2), 17-26.

[6] Goodness of Fit in Logistic Regression. (n.d.). Faculty of Medicine, McGill University.