Shawn Koohy
24 April 2022
Dr. Davis
MTH332

Predicting Course Completion with Logistic Regression

Abstract

Logistic regression is an important aspect of creating models for data and finding probabilities of outcomes. Logistic regression is especially useful when the outcome of data is presented in binary form, 0’s and 1’s. This form of regression has many applications, such as statistical mechanics in physics and machine learning in computer science and computational mathematics. Here we will use logistic regression to predict whether a student will pass or fail a college introductory course, which would allow them to be admitted into a university with a major and degree path. We will be using a combination of variables mainly specified into three sets, academic performance, student behavior, and psychological characteristics. It will be shown that both student behavior and academic performance play a major role in a student’s ability to complete the course. Bettering the aspects in these two sets, student behavior and academic performance will allow the student to greatly increase their probability of completing the course.

Introduction & Background

The College Now/START program at the University of Massachusetts Dartmouth is an alternate admission program to support academically disadvantaged students. This program plans to prepare students to perform the best they can if they complete the course. They will be prepared by working with counselors to help aid them in “goal setting, course selection, academic achievement, and short/long-term program planning”. If they are admitted then the students can pursue a degree and major at the University of Massachusetts Dartmouth. The specific data used in this report comes from a total of 106 students. These students were tracked with specific variables including high school GPA, SAT score, fall grant eligible, attended orientation, attended experience day, resident/commuter, athlete, completed summer bridge, dropout proneness, predicted academic difficulty, academic stress, receptivity to academic assistance, receptivity to personal counseling, receptivity to social engagement, receptivity to career guidance, receptivity to financial guidance, desire to transfer, completed campus event requirement, completed community service requirement, number of faculty advisor meetings attended, number of per mentor meetings attended, number of workshops attended, fall semester GPA, spring semester GPA, cumulative GPA, number of credits earned and finally completed course. This research aims to discover which variables give the best prediction of whether a student will pass/fail the program. This will be done by using single variable and multivariable logistical regression. Logistic regression is a commonly used model when predicting the outcomes of a single binary outcome variable (ours is the variable: completed course).

Methods

We will first be looking at 3 sets of “randomly” chosen variables. These sets are just personal observations on certain variables that would seem to make a decent prediction. We will examine these 3 sets using singular variable logistic regression.

The first set of variables:

1. Completed summer bridge
2. Completed campus event requirement
3. Completed community service requirement
4. Number of faculty advisor meetings attended
5. Number of peer mentor meetings attended
6. Number of workshops attended

The second set of variables:

1. High school GPA
2. Attended orientation
3. Completed Summer bridge
4. Number of workshops attended
5. Fall Semester GPA
6. Spring semester GPA

The third set of variables:

1. Fall grant eligible
2. Completed campus event requirement
3. Number of faculty advisor meetings attended
4. Cumulative GPA
5. Number of workshops attended

Each of the variables is either determined by a numerical label such as 1=yes and 0=no, or a scale for a specific variable, we will list the description of them below.

Variables that exist in the range of 0-4:

1. High school GPA
2. Spring semester GPA
3. Fall semester GPA
4. Cumulative GPA

Variables determine by a numerical label, variables are associated with 1=yes, 0=no unless stated otherwise

1. Fall grant eligible
2. Attended orientation
3. Attended experience day
4. Resident/Commuter (1=resident, 0=commuter)
5. Athlete
6. Completed summer bridge (2=completed, 1=completed at least half, 0=incomplete)
7. Completed campus event requirement
8. Completed community service requirement
9. Completed course

Variables in the range of 0-100

1. Dropout proneness
2. Predicted academic difficulty
3. Academic stress

4. Receptivity to academic assistance
5. Receptivity to personal counseling
6. Receptivity to social engagement
7. Receptivity to career guidance
8. Receptivity to financial guidance
9. Desire to transfer

Other variable ranges:

10. Number of faculty advisor meetings attended (no limit)
11. Number of peer mentor meetings attended (no limit)
12. Number of workshops attended (no limit)
13. Number of credits (0-35)
14. SAT score (0-1600)
15. Desire to transfer

Before we continue there are a few areas that need to be covered. The first is the issues of missing data points. Any missing data points will be replaced with the data set's average. Example:

$$Data = (1,\ ,2),\ notice\ the\ missing\ data\ point$$

We then turn this data set into the following:

$$Data = (1, 1.5, 2)$$

The second issue or discrepancy that we have is the fact that some variables are much larger than others and so they may cause instability or will overweigh other variables. We will solve this by normalizing the larger variables to the same range in which the small variables exist. To do this we apply the following formula:

$$x_{norm} = \frac{x - r_{min}}{r_{max} - r_{min}}(t_{max} - t_{min}) + t_{min}$$

$$r_{min} = Minimum\ in\ the\ range\ of\ the\ measurement$$

$$r_{max} = Maximum\ in\ the\ range\ of\ the\ measurement$$

$$t_{min} = Minimum\ in\ the\ range\ of\ desired\ scaling$$

$$t_{min} = Maximum\ in\ the\ range\ of\ desired\ scaling$$

$$x = Value\ in\ data\ set\ which\ to\ normalize$$

For example, we will soon group SAT scores and GPA into one similar group. GPA is on a scale from 0-4 and SAT scores from 0-1600. We will scale the SAT scores into a range of 0-4 while still keeping the score's meaning/value. If someone had an SAT score of 1200 then the following would be the normalized value:

$$1200_{norm} = \frac{1200 - 0}{1600 - 0}(4 - 0) + 0 = 3$$

To perform logistical regression with a single variable for our first 3 sets we will be using Mathematica. To get the first set of variables into just one variable we will sum them up and pair it to whether that person passed or failed the course. Once we have our sum for the predictor variables we will use Mathematica's LogitModelFit to find the equation of our model. We can now plot our model on top of our completed course data. Once we have done that we will apply our model to the data points and calculated the probability of successfully completing the course. Since we already know if that student had completed the course we can compare them to see how true/accurate our model is. This will be done by seeing how many times our model's prediction was right. If our model predicted success with a probability of 70% and the student passed that would be a success. However, if a student failed the course and our model said that the student would pass with 85% then our model has failed for that case. This process and idea will be applied to all sets of variables we use, single and multivariable.

It might be obvious that many of these variables are related or that they belong to a specific category. We will consider 3 main categories of variables that are separate from the previous 3 we encountered. The variables can be split into these 3 groups: psychological characteristics, academic performance, and student behavior. We define these sets with the following variables:

Psychological Characteristics:

1. Dropout proneness

2. Predicted academic difficulty
3. Educational stress
4. Receptivity to academic assistance
5. Receptivity to personal counseling
6. Receptivity to social engagement
7. Receptivity to career guidance
8. Receptivity to financial guidance
9. Desire to transfer

Academic Performance:

1. High school GPA
2. SAT score
3. Fall semester GPA
4. Spring semester GPA
5. Cumulative GPA
6. Number of credits earned

Student Behavior:

1. Attended orientation
2. Attended experience day
3. Completed summer bridge
4. Completed campus Event Requirements
5. Completed community service requirement
6. Number of faculty advisor meetings attended
7. Number of peer mentor meetings attended
8. Number of workshops attended

For these three groups, we will repeat the same process for accuracy as described before, but we will perform both single and multivariable logistical regression. In the single variable case, we will have one variable which would be the sum of all variables whereas for the multivariable case we will let each variable be a prediction rather than sum them all up. The last important aspect specifically in the case of the psychological characteristic group is that of many missing data points. There were about 13 students who had no information on any of the variables. When performing the logistical regression for this group we will completely discard these students from our data, as I do not believe, with the amount of missing information, to simply just put the average amounts in as it may highly skew the model in various ways. It is better to examine the information we truly know much more about.

Results

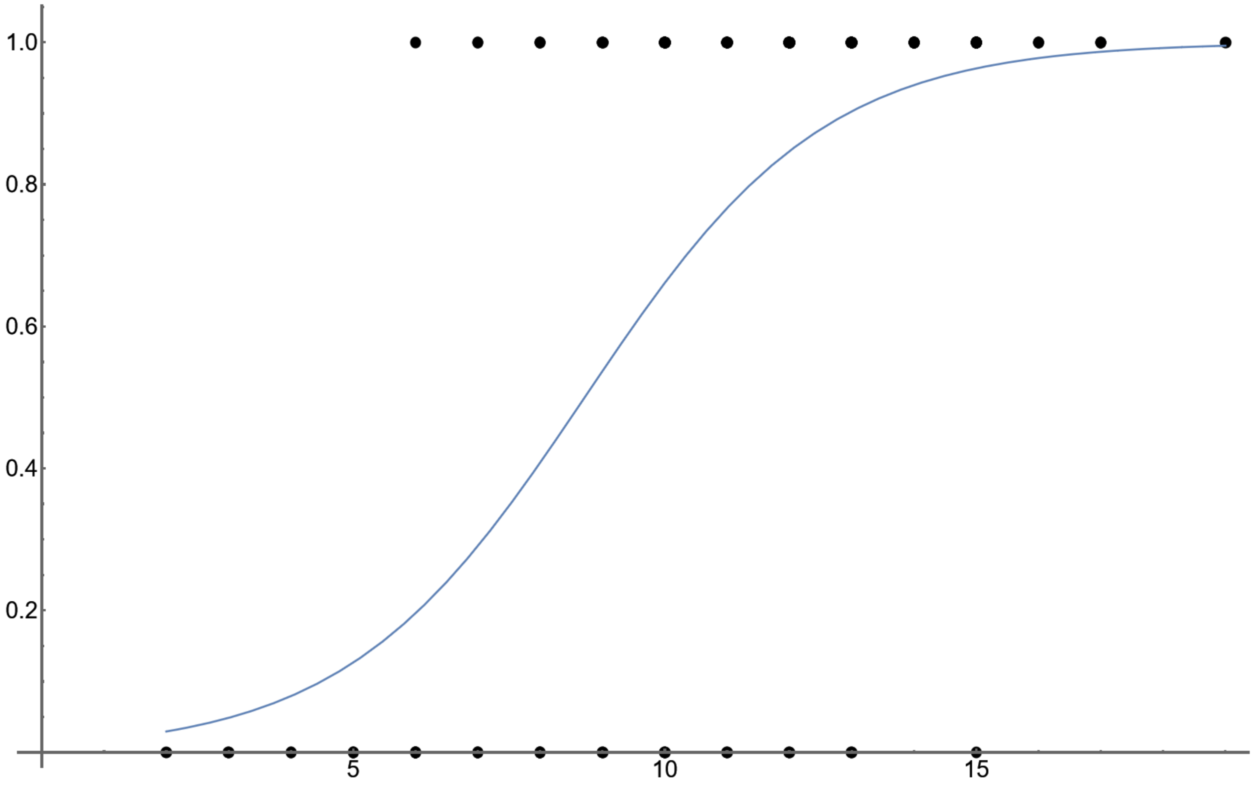


Figure 1. Plot of Logistic Model for the First Set

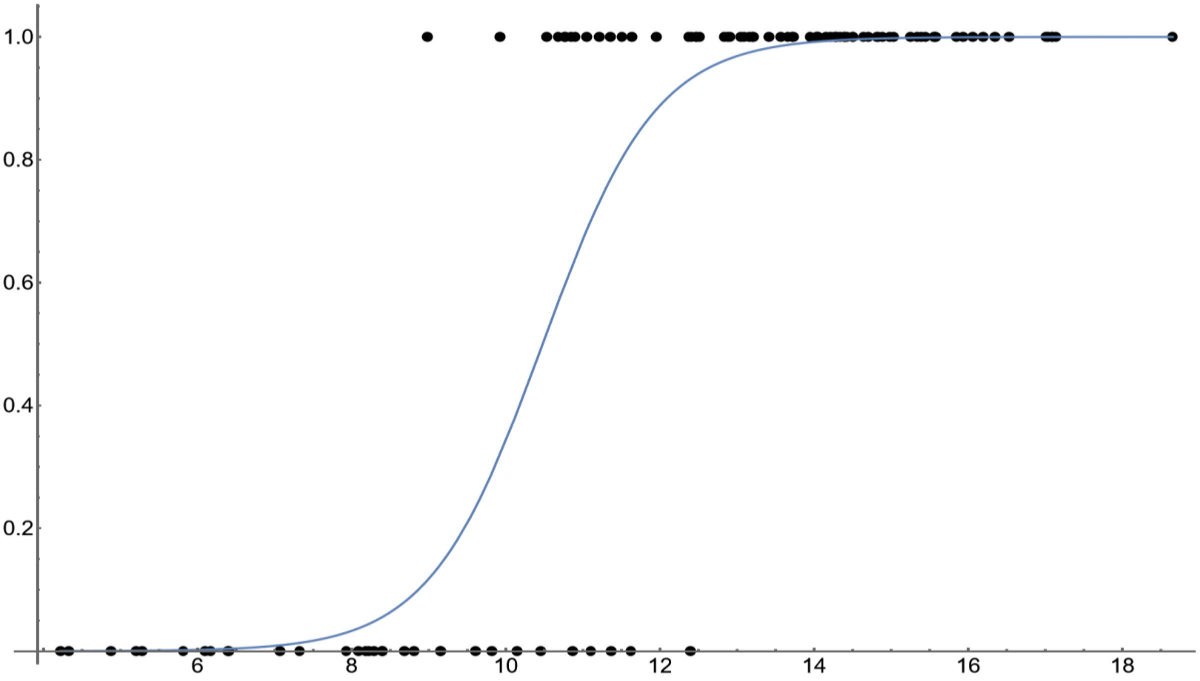


Figure 2. Plot of Logistic Model for the Second Set

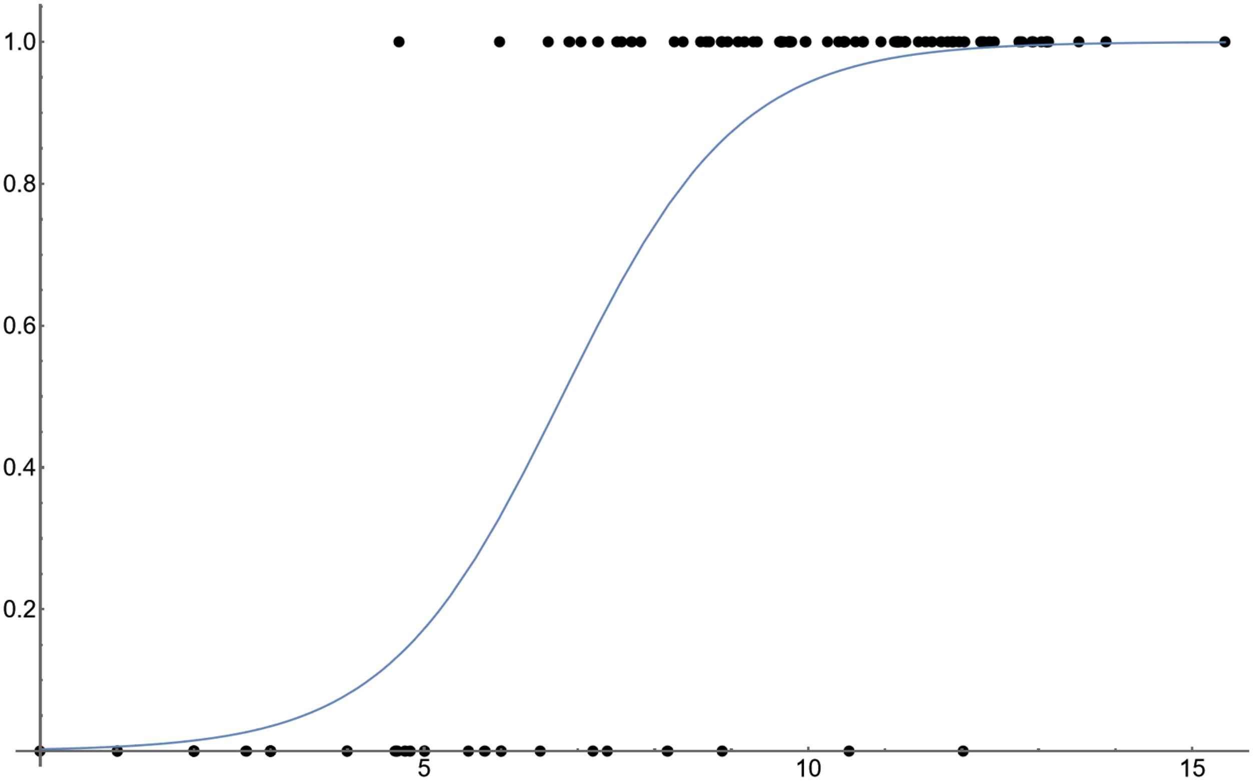


Figure 3. Plot of Logistic Model for the Third Set

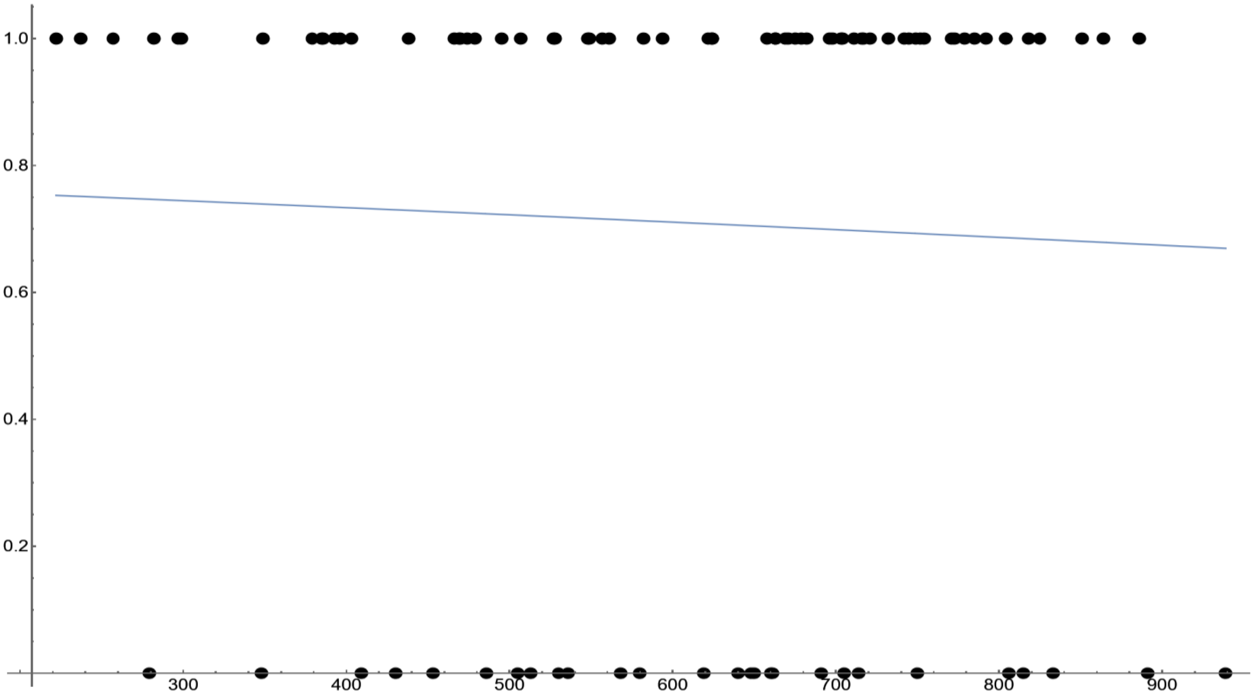


Figure 4. Plot of Logistic Model for the Psychological Characteristics (Single Variable)

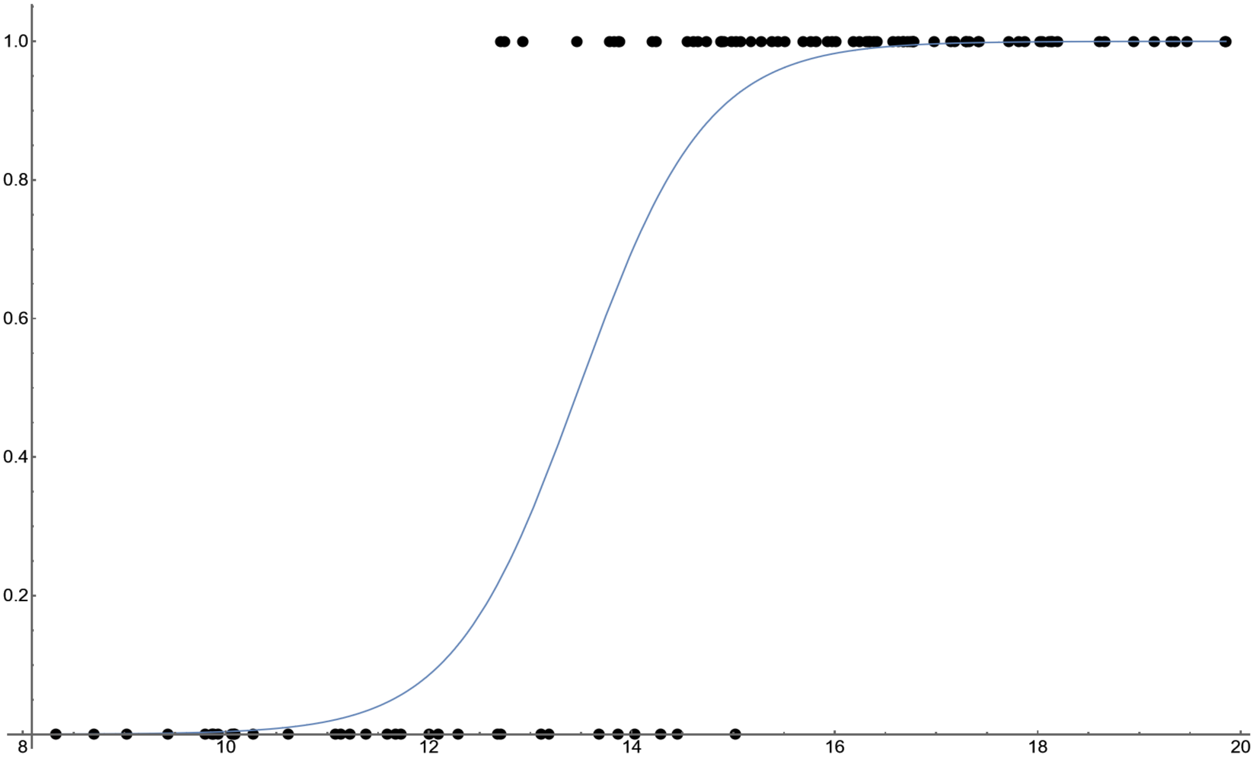


Figure 5. Plot of Logistic Model for the Academic Performance (Single Variable)

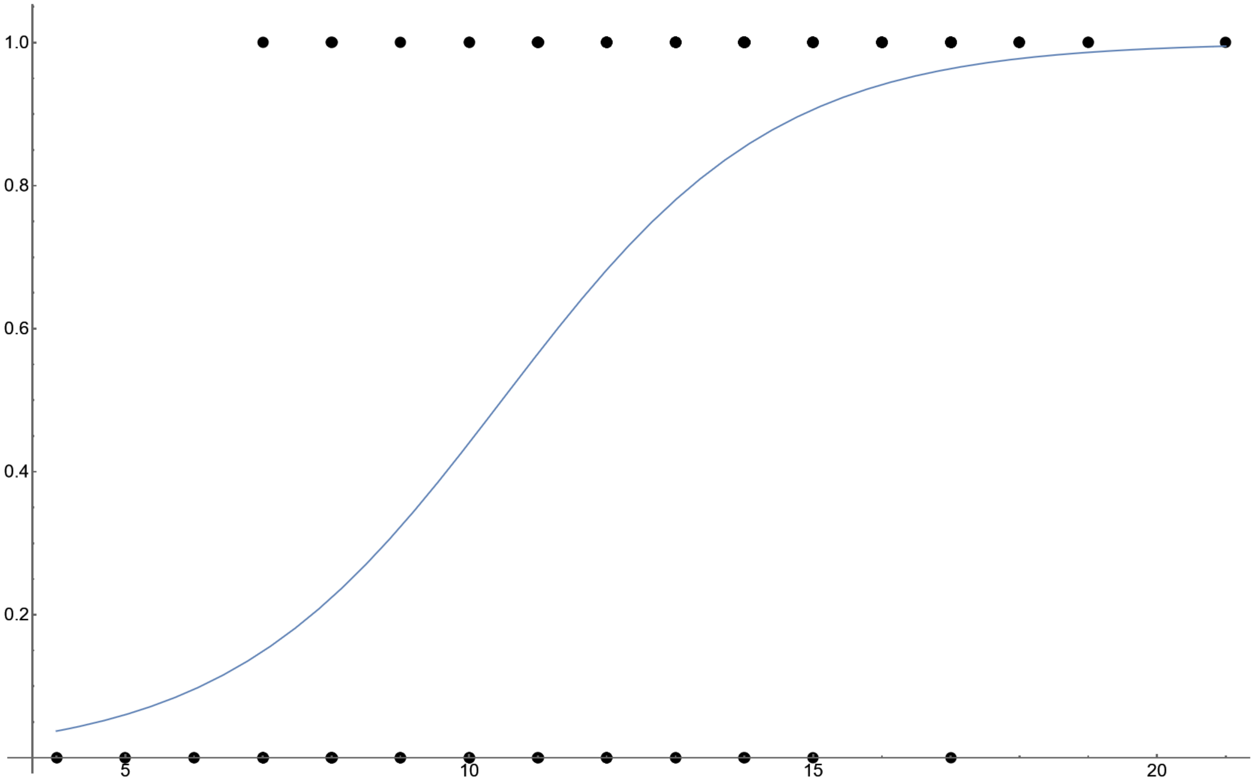
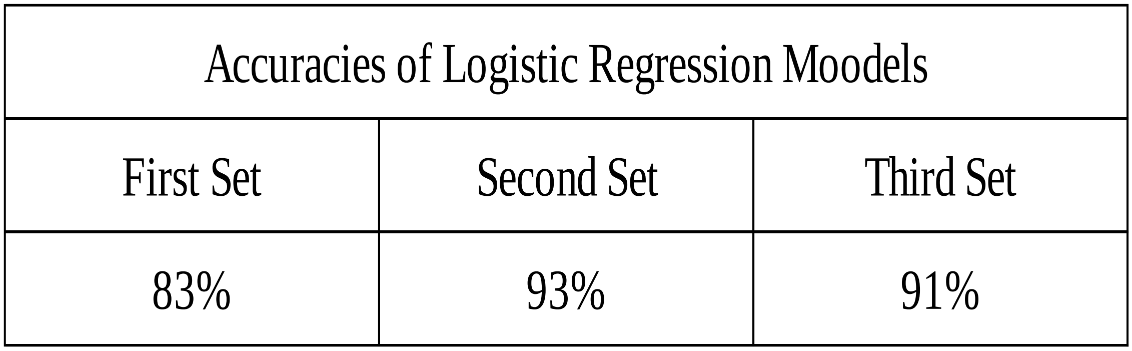


Figure 6. Plot of Logistic Model for the Student Behavior (Single Variable)



Accuracies of Logistic Regression Moodels		
First Set	Second Set	Third Set
83%	93%	91%

Figure 7. Accuracies of the First Three Sets

Accuracies of Logistic Regression Moodels			
Type \ Group	Academic Performance	Psychological Characteristics	Student Behavior
Single Variable	91%	71%	83%
Multivariable	96%	73%	88%

Figure 8. Accuracies for Specific Groups

Discussion & Conclusions

Looking at Figure 8, we can see that our model is better at predicting success when looking at a student's academic performance. In all cases using multivariable logistic regression gave us a more accurate outcome than using singular variable logistic regression. Using the psychological characteristic gives us the worst prediction of whether a student will pass or fail the course. If we wanted to best predict if a student would pass it is best to go off on their academic abilities. Advisors should encourage the student to try to increase their GPA and the number of credits they earn to better their chances of completing the course. Now, this is a lot easier said than done, telling somebody to "just increase their GPA", we can see that student behavior predicts outcomes with 88% accuracy which is still good. It would be much more beneficial and realistic to tell students to attend more workshops, campus events, faculty meetings, advisor meetings, and so on. This is also an easier task to complete for the student while also being an effective way of increasing their odds of passing. It is also more encouraging to hear from an advisor that it will be almost as effective to go to many workshop events, and meetings, and be active in the university's events as would increasing one's GPA.

Focusing on a student's behavior could be the best way to help them complete the course. While doing this the academic performance should be closely looked at but, possibly treated as a secondary source. In general, encouraging a student to become more academic involved both grade and participation wise is what will be best for them. Psychological characteristics are okay to focus on in the background of everything but according to our model and data, it is not as good of a predictor or determiner of whether a student will pass the course. A student may find it less stressful and have a better experience focusing just a little more on participating in university events and advising meetings than stressing over their GPA, though it is still a major and important role in most people’s academic success.

References

https://stats.stackexchange.com/questions/281162/scale-a-number-between-a-range

https://www.umassd.edu/collegenow/about/#:~:text=An%20academic%20support%20program,your%20fullest%20at%20the%20university.