Logistic Regression

Nick Collins

UMass Dartmouth

April 24, 2022

Abstract-

Applying to college can be extremely competitive. Since colleges are getting thousands and thousands of applications, they only want to select the best candidates that will hopefully stay all four years and graduate from that same college. With linear regression we can investigate these variables and see exactly which ones are helpful in predicting the success of a certain student. This isn't to say that linear regression will tell us exactly which variables are useful and which are not. But we will get some insight into which ones prove more of a factor in the completion of a preliminary college course.

Introduction-

Logistic regression is a powerful tool used to show relationships between a binary variable and other variables. In this case we want to compare the binary variable “Completed Course” (0 for no, 1 for yes) with the other variables included in the spreadsheet. The goal is to get an idea of exactly which variables are likely to predict a student's success or failure in a current college. This is an important task for colleges to do because it can save them money in the long run by not having to collect so much data on different variables. It can also help them more accurately predict which students will stay in their program and graduate.

Methods-

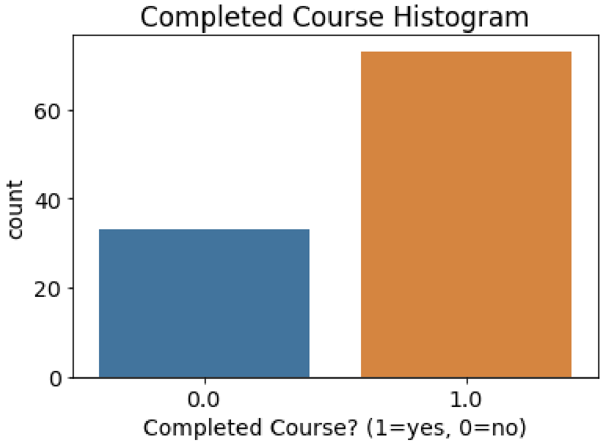
The data for this project was 106 rows with 31 columns. These columns included many different variables, but the majority of the columns could be classified into four major groups. These groups are academic performance, personal characteristics, physiological characteristics and student behavior. We want to use these variables to make a prediction on what best represents whether a student will drop the course or be a good fit for the school. This will be done through the use of logistic regression.

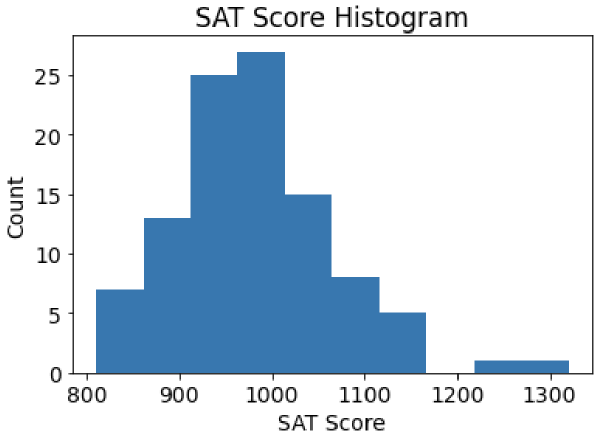
Logistic regression is used to describe data and show the relationship between a binary variable and other variables. This binary variable is the completed course variable. It is either a zero or one, where zero means the student did not complete the course and one shows that the student completed the course. This will produce some output that can show us which variables are more influential in predicting the success or failure of a college student. The results for logistic regression were produced using the glm function in R.

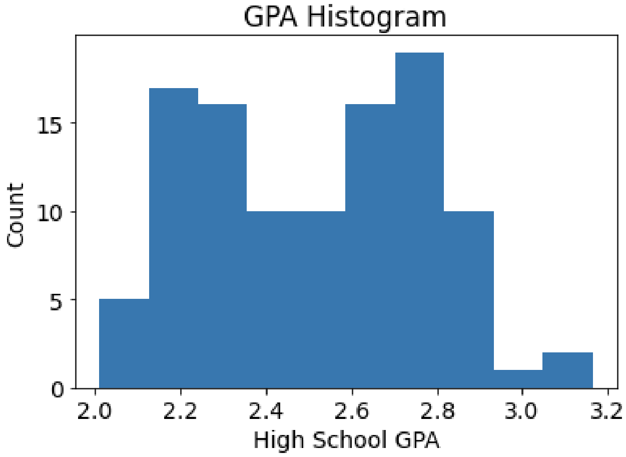
Before I could move onto doing analysis, I had to clean the data up a bit and make a new “predictor” variable. There were some missing values in the data, so I didn’t include them in calculations where the missing values played a role. The new variable I had to create was called the predictor variable. This was the summation of the following variables “Completed summer bridge”, “Completed campus event requirement”, “Completed community service requirement”, “Number of faculty adviser meetings attended”, “Number of peer mentor meetings” and “Number of workshops attended”. This variable had a range from two to nineteen. The higher the number for this variable the more active the student was in participating in extracurricular activities related to school. We would expect more and more students to have completed the course as the predictor variable increases.

Results-

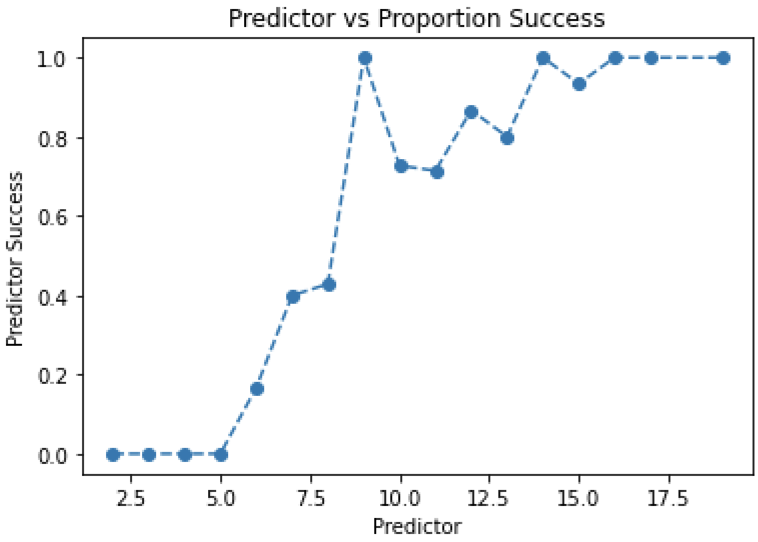
Looking at some simple graphs of the data can give us some great insight into how the data is acting. Plotting the distribution of the binary variable shows there are more completed courses than non-completed. I also plotted the SAT scores for the students so we can see the breakdown of the students in that area, because I know SAT scores can be a big deciding factor in some school's opinion. Another plot included was the high school GPA for the students. Like the SAT score, GPA is often regarded as a big indicator in whether someone is accepted into a college. It was good to see the distribution for these two histograms spread out and not shifted to one side with outliers.







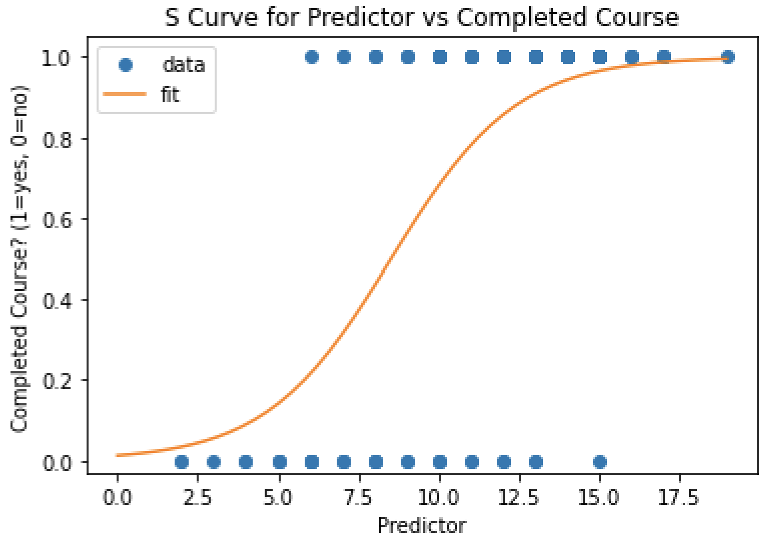
The variable I mentioned earlier that I created was the predictor variable. The table below shows the range for the predictor variable, the count, the success and the success proportion. The count is the number of times the predictor variable shows up. The success column is the number of for each predictor. And lastly the proportion success is the success/N. This gives us the proportion that we can plot against the predictor variable. We also can see this plot below. This plot is a good representation of how well the predictor variable helps to predict if a student will make it through the course. We would expect to see the proportion of success increase as the predictor increases. The line graph does fluctuate a bit near the nine to twelve range, but this can be explained by some other factors. Just because a student didn't attend workshops or faculty meetings doesn't mean they will fail. If a student was doing well, they might not attend as frequently as a student who is struggling.



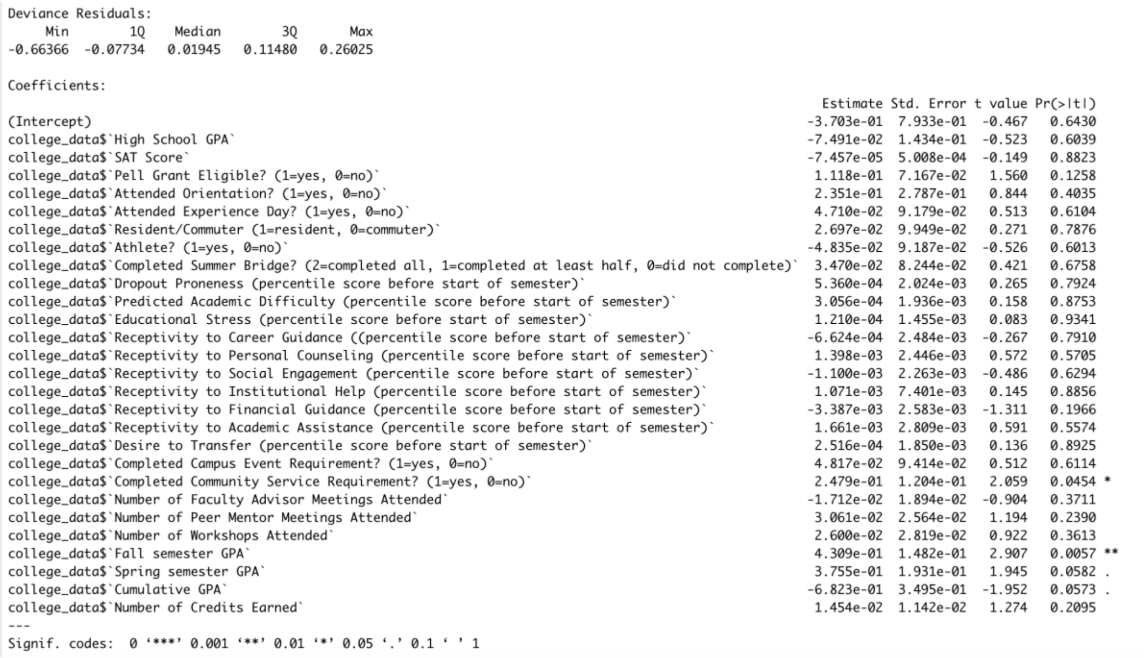
Predictor	N	Success	Proportion Success
2	2	0	0/2 = 0
3	1	0	0/1 = 0
4	2	0	0/2 = 0
5	3	0	0/3 = 0
6	6	1	1/6 = .167
7	5	2	2/5 = .4

8	7	3	3/7 = .429
9	3	3	3/3 = 1
10	11	8	8/11 = .727
11	7	5	5/7 = .714
12	15	13	13/15 = .867
13	10	8	8/10 = .8
14	9	9	9/9 = 1
15	15	14	14/15 = .934
16	3	3	3/3 = 1
17	3	3	3/3 = 1
19	1	1	1/1 = 1

Another plot that includes the predictor variable is the s curve where the predictor is plotted against the binary outcome variable. After plotting these values, we can add an s curve to see where the outcome starts to change from fail to pass. In the graph below we see this trend occur from the predictor value range six to eleven. This is a decent overlap for this predictor value because it only ranges from two to nineteen in the whole dataset.

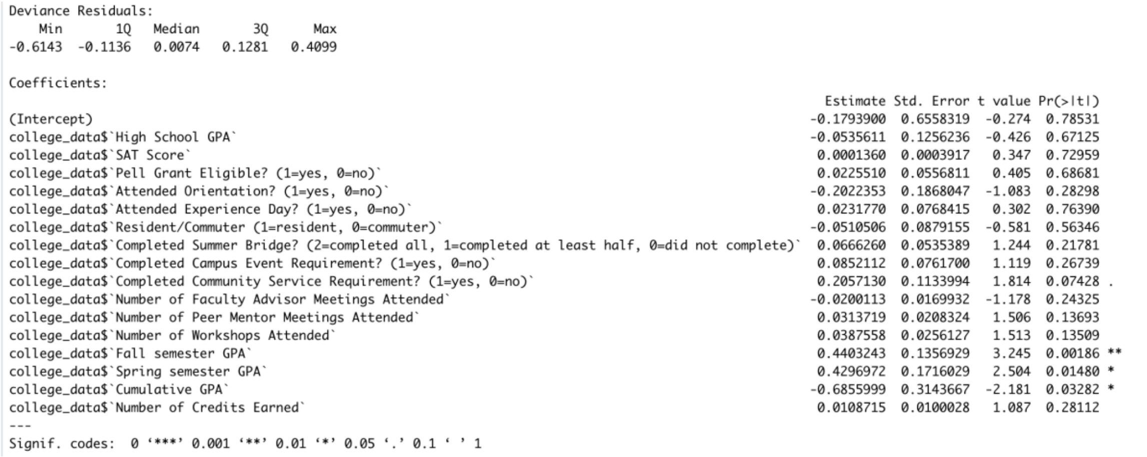


The last thing was to perform some logistic regression. This was done using the glm function in R. This takes our outcome variable (Completed Course) and uses it to try to fit a linear model with the variables the user puts in. In the first logistic regression I used almost every variable that was included in the excel spreadsheet. There were some variables that had to be dropped because they were missing values and were throwing off the output. The second logistic regression was done with fewer variables, where I only tried to include the ones I thought would be most important. In both we see the right most column have a p value and if the p value is of a certain threshold, it is marked significant. The significant codes are included at the bottom of both images, so we can see just how influential a certain variable can be according to logistic regression.



```
Deviance Residuals:
     Min        1Q    Median        3Q       Max  
-0.66366  -0.07734   0.01945   0.11480   0.26025  

Coefficients:
                                                                                                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                                                              -3.703e-01  7.933e-01  -0.467   0.6430
college_data$`High School GPA`                                                                           -7.491e-02  1.434e-01  -0.523   0.6039
college_data$`SAT Score`                                                                                 -7.457e-05  5.008e-04  -0.149   0.8823
college_data$`Pell Grant Eligible? (1=yes, 0=no)`                                                         1.118e-01  7.167e-02   1.560   0.1258
college_data$`Attended Orientation? (1=yes, 0=no)`                                                        2.351e-01  2.787e-01   0.844   0.4035
college_data$`Attended Experience Day? (1=yes, 0=no)`                                                     4.710e-02  9.179e-02   0.513   0.6104
college_data$`Resident/Commuter (1=resident, 0=commuter)`                                                 2.697e-02  9.949e-02   0.271   0.7876
college_data$`Athlete? (1=yes, 0=no)`                                                                    -4.835e-02  9.187e-02  -0.526   0.6013
college_data$`Completed Summer Bridge? (2=completed all, 1=completed at least half, 0=did not complete)`  3.470e-02  8.244e-02   0.421   0.6758
college_data$`Dropout Proneness (percentile score before start of semester)`                              5.360e-04  2.024e-03   0.265   0.7924
college_data$`Predicted Academic Difficulty (percentile score before start of semester)`                  3.056e-04  1.936e-03   0.158   0.8753
college_data$`Educational Stress (percentile score before start of semester)`                             1.210e-04  1.455e-03   0.083   0.9341
college_data$`Receptivity to Career Guidance ((percentile score before start of semester)`               -6.624e-04  2.484e-03  -0.267   0.7910
college_data$`Receptivity to Personal Counseling (percentile score before start of semester)`             1.398e-03  2.446e-03   0.572   0.5705
college_data$`Receptivity to Social Engagement (percentile score before start of semester)`              -1.100e-03  2.263e-03  -0.486   0.6294
college_data$`Receptivity to Institutional Help (percentile score before start of semester)`              1.071e-03  7.401e-03   0.145   0.8856
college_data$`Receptivity to Financial Guidance (percentile score before start of semester)`             -3.387e-03  2.583e-03  -1.311   0.1966
college_data$`Receptivity to Academic Assistance (percentile score before start of semester)`             1.661e-03  2.809e-03   0.591   0.5574
college_data$`Desire to Transfer (percentile score before start of semester)`                             2.516e-04  1.850e-03   0.136   0.8925
college_data$`Completed Campus Event Requirement? (1=yes, 0=no)`                                          4.817e-02  9.414e-02   0.512   0.6114
college_data$`Completed Community Service Requirement? (1=yes, 0=no)`                                     2.479e-01  1.204e-01   2.059   0.0454 *
college_data$`Number of Faculty Advisor Meetings Attended`                                               -1.712e-02  1.894e-02  -0.904   0.3711
college_data$`Number of Peer Mentor Meetings Attended`                                                    3.061e-02  2.564e-02   1.194   0.2390
college_data$`Number of Workshops Attended`                                                               2.600e-02  2.819e-02   0.922   0.3613
college_data$`Fall semester GPA`                                                                          4.309e-01  1.482e-01   2.907   0.0057 **
college_data$`Spring semester GPA`                                                                        3.755e-01  1.931e-01   1.945   0.0582 .
college_data$`Cumulative GPA`                                                                            -6.823e-01  3.495e-01  -1.952   0.0573 .
college_data$`Number of Credits Earned`                                                                   1.454e-02  1.142e-02   1.274   0.2095
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-0.6143  -0.1136   0.0074   0.1281   0.4099  

Coefficients:
                                                                                                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                                                              -0.1793900  0.6558319  -0.274  0.78531
college_data$`High School GPA`                                                                           -0.0535611  0.1256236  -0.426  0.67125
college_data$`SAT Score`                                                                                  0.0001360  0.0003917   0.347  0.72959
college_data$`Pell Grant Eligible? (1=yes, 0=no)`                                                         0.0225510  0.0556811   0.405  0.68681
college_data$`Attended Orientation? (1=yes, 0=no)`                                                       -0.2022353  0.1868047  -1.083  0.28298
college_data$`Attended Experience Day? (1=yes, 0=no)`                                                     0.0231770  0.0768415   0.302  0.76390
college_data$`Resident/Commuter (1=resident, 0=commuter)`                                                -0.0510506  0.0879155  -0.581  0.56346
college_data$`Completed Summer Bridge? (2=completed all, 1=completed at least half, 0=did not complete)`  0.0666260  0.0535389   1.244  0.21781
college_data$`Completed Campus Event Requirement? (1=yes, 0=no)`                                          0.0852112  0.0761700   1.119  0.26739
college_data$`Completed Community Service Requirement? (1=yes, 0=no)`                                     0.2057130  0.1133994   1.814  0.07428 .
college_data$`Number of Faculty Advisor Meetings Attended`                                               -0.0200113  0.0169932  -1.178  0.24325
college_data$`Number of Peer Mentor Meetings Attended`                                                    0.0313719  0.0208324   1.506  0.13693
college_data$`Number of Workshops Attended`                                                               0.0387558  0.0256127   1.513  0.13509
college_data$`Fall semester GPA`                                                                          0.4403243  0.1356929   3.245  0.00186 **
college_data$`Spring semester GPA`                                                                        0.4296972  0.1716029   2.504  0.01480 *
college_data$`Cumulative GPA`                                                                            -0.6855999  0.3143667  -2.181  0.03282 *
college_data$`Number of Credits Earned`                                                                   0.0108715  0.0100028   1.087  0.28112
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion-

One of the best plots we see is the predictor variable vs the proportion success. This plot shows a steady increase in the proportion of successful students as the predictor variable increase also. We see one slight outlier at predictor variable nine, but there are only three occurrences for this number and all three passed the course, so the success rate is 1. Overall, the predictor variable was a fairly good estimate on whether a student would pass or fail the course. It was only a fairly good estimate because if a student was doing well they might not go to optional extra help meetings. This isn't the case for all students but can influence the accuracy of the predictor variable. Another factor in this variable is the summation of three counter variables and three binary variables. Binary variables are only zero or one, while the other variables being added can range from zero to nine. For example, a student could score zero in very variable besides the number of workshops attended. If he/she attended nine workshops but nothing else the predictor variable would have the student likely to pass the course. With large values like this compared to binary variables could lead to some flawed results.

The s curve of predictor vs completed course also gives us some good results on what ranges a student will pass or fail the course. From range two to six it is unlikely that the student will complete the course. Then we see the curve from six to eleven start to have some students completing the course while others are failing. The last range is eleven to nineteen. This range has the most students passing the course, which makes sense and we would expect to see this.

The logistic regression gave us some insight into what variables are going to be influential in the overall success of a student. However, there were many variables of all kinds in this dataset so we would expect the results to differ a bit. With so many variables it is tough to see if one impacts the output more than another certain variable. This is because the p values are high with so many variables making up the logistic regression. We see some rudimentary results from the glm function I used for the logistic regression.

References-

MTH 332 Mathematical Statistics Spring 2022

https://mth332.files.wordpress.com/2022/01/mth_332_spring_2022.pdf

Applied Logistic Regression by: David W. Hosmer and Stanley Lemeshow

https://mth231.files.wordpress.com/2020/11/applied-logistic-regression-hosmer.pdf