ACN Final Project

Aaron Nadell

5/11/2020

Abstract:

In this project, I will examine the key factors that impact someone's grades in the math courses they take using the "MathPlacement" dataset. I will use the clean the data before using the step function to find an optimal generalized linear model to explain students' grades in the math course. The generalized linear model will explain the relationships between PSATM, Size, and Placement Scores as well as the recommended courses taken. These variables explain the probability of getting a desirable grade relatively well compared to a model that incorporates all the explanatory variables, so there is little need for more criteria to place students.

Introduction:

Standardized tests and GPAs are critical to college acceptance for students living in the U.S. They are used to gauge the proficiency of incoming students in most curriculae taught in American public and private schools. They are generally a good representations of how well a student performs in a typical classroom setting or at the collegiate level and can be used to compare the learning and ability of students across schools. Higher scores on these standardized tests indicate better competency, and students that score better on these tests might be accepted into more selective colleges and attend harder classes. These tests as well as other characteristics of the school like class size or student rankings speak to the ability of incoming students.

In addition to performances on standardized tests, the recommendations received from the professors about what class to take can inform a prospective student as to what grade they might hope to acheive. Direct placement testing by the college for mathematics might provide further information as to whether an incoming student is likely to succeed in a given course.

This project seeks to analyze 13 explanatory variables using generalized linear modeling to explain whether an incoming student will achieve a satisfactory grade in the course that they chose. These variables relate to standardized tests, GPAs, class size, and recommendations. We will choose the best 4 explanatory variables that explain whether students will succeed.

Methods:

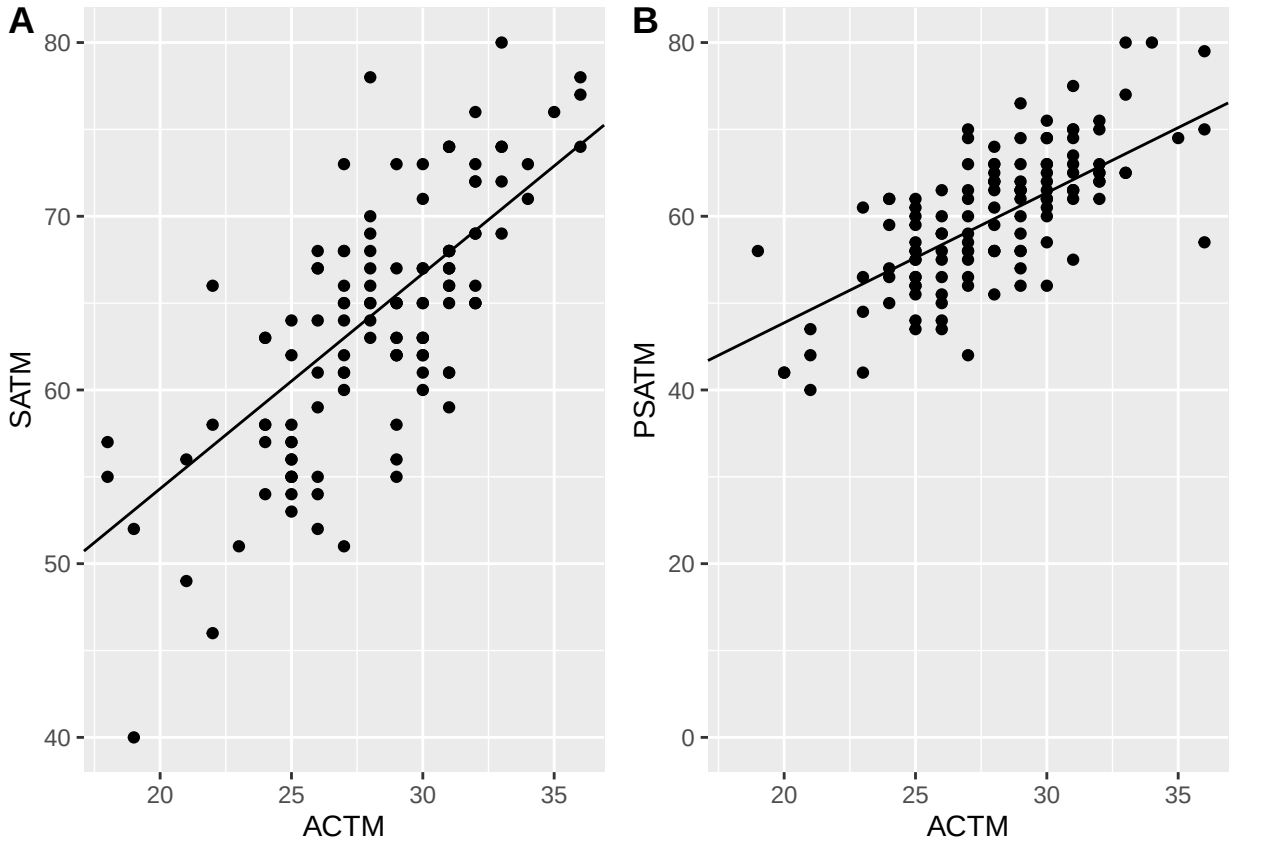
Data Cleaning

I will begin by reformatting the data so that a "B" and above is marked as a 1 to represent a successful learning experience of the incoming college student. I am marking incompletes and withdraws as a 0 because the students were likely behind on work and unable to achieve success in the class. I also removed GPAadj values of 0 because they seem to be outliers and initial GPAvalues of 0 likely mean that the adjusted GPAs might be 0. This also seems unlikely given that they would not be accepted into college with a GPA of 0 and is likely because their schools did not record their GPA. I will also be doing the same for SATM, PSATM, and ACTM as students who scored a zero likely did not try or would not be accepted into a college.

I then imputed the data using the missForest package to quickly build a new dataset with approximated values for NAs. This package does not provide extra information about the data it only provides completion of the dataset using artificial distributions and imputing iteratively for missing values while minimizing the normalized root mean squared error. This function works automatically for continuous variables, so I was forced to remove categorical variables other than Grade like gender.

Results

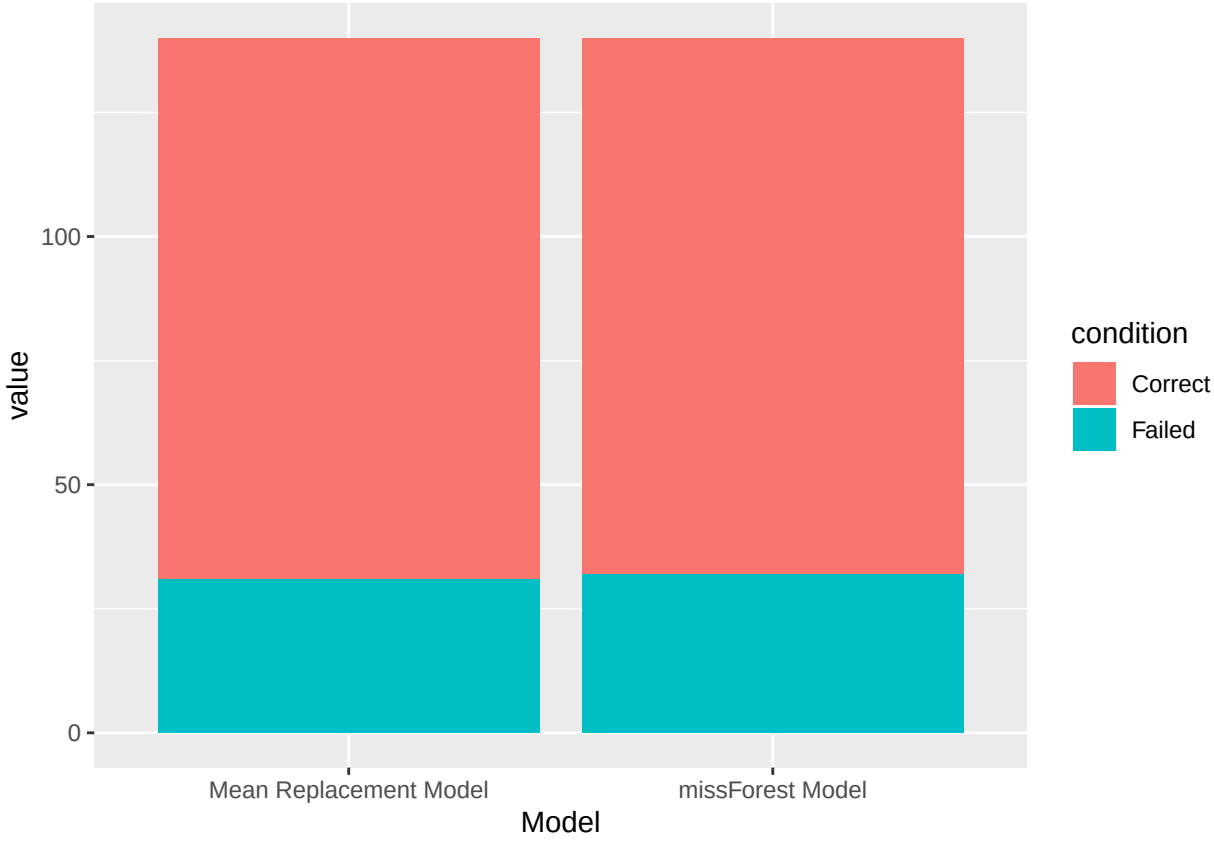
```
figure1
```



Checking Assumptions

It appears that there is a high correlation between SATM and ACTM as well as a high correlation between ACTM and PSATM which violates a condition for logistic regression. Logistic regression requires there to be very little correlation between independent variables so it seems we should only include one of these in our final model, however all other assumptions required for generalized linear modeling are met because the dependent variable is binary, the independent variables are independent of each other, and there is a large sample size (n>300).

To compare the quality of my missForest model, I also created a separate dataframe and replaced the NA values with the means for the continuous variables, but I removed the NA values for the categorical variables.



This barplot shows how each of the models performed on the Clean dataset. I'm partial to the missForest model because eventhough it was trained on the dataset created by the missForest function it was able to predict comparably to the mean Replacement method.

```
Table1
```

```
##
##
## |                              |Results (N = 4) |
## |:-----------------------------|:---------------|
## |**missForest AICs **          |      |
## |   Total:           |318.81          |
## |   SATM             |317.46          |
## |   Rank             |318.6           |
## |   Placement Score  |334.98          |
## |   Recommends       |339.88          |
## |**mean replacement AICs**     |      |
## |   Total:           |313.65          |
## |   PSATM            |314.41          |
## |   Size             |316.08          |
## |   Placement Score  |326.89          |
## |   Recommends       |331.08          |
## |**Deviances for missForest**  |      |
## |   SATM             |317.46          |
## |   Rank             |318.6           |
## |   Placement Score  |334.98          |
## |   Recommends       |339.88          |
## |**Mean Replacement Deviances**|      |
```

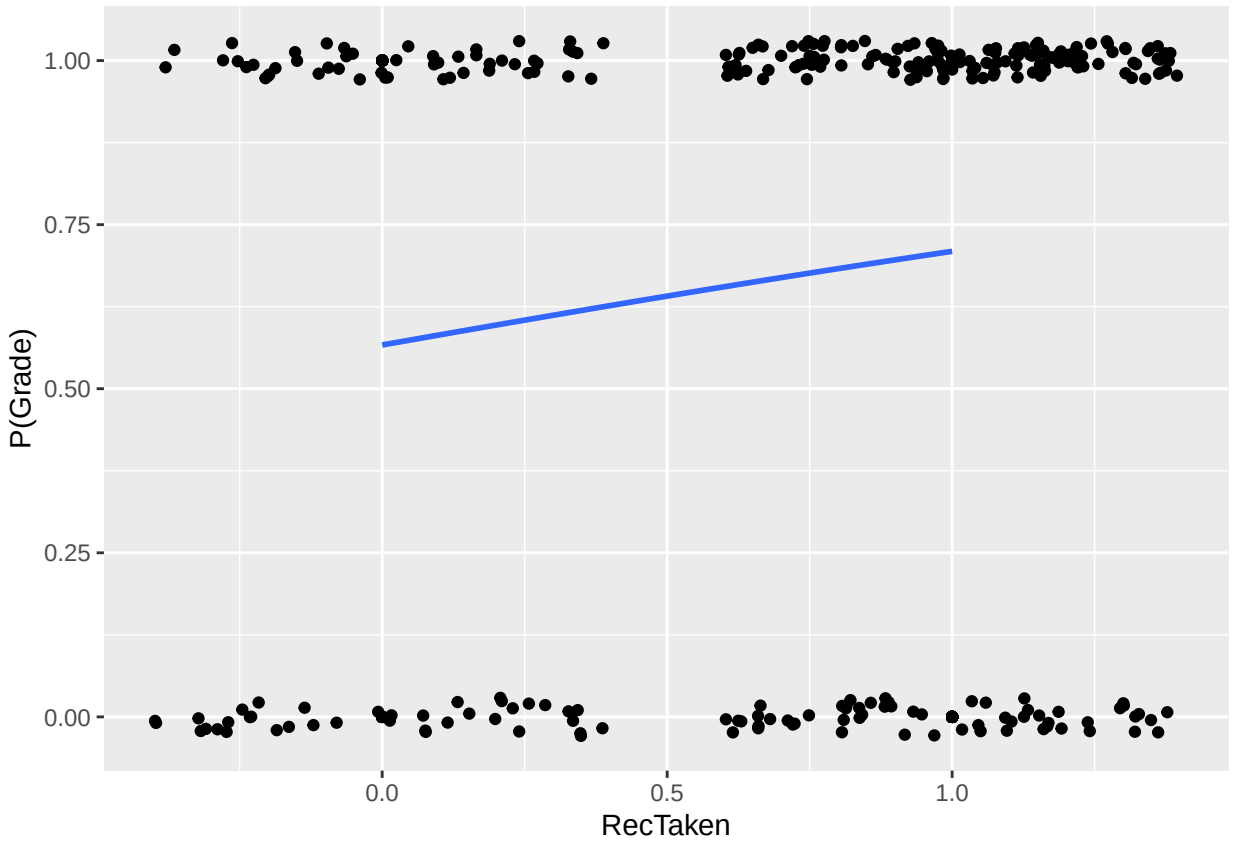
```
## |   PSATM             |294.41          |
## |   Size              |296.08          |
## |   Placement Score   |306.89          |
## |   Recommends        |323.08          |
```

As we can see in the missForest Model the four explanatory variables that best explain whether or not a grade was satisfactory were PSATM, Size, PlcmtScore, and Recommends. It appears that the log(odds) of a satisfactory grade change by 5.848e-03 for each point higher on the PSATM in relation to the other variables within the model. The log(odds) of a satisfactory grade change by 5.211e-04 for each point higher on the Size in relation to the other variables within the model, and the log(odds) of a satisfactory grade change by 2.531e-02 for each point higher on the Placement Score in relation to the other variables within the model. The log(odds) of a satisfactory grade changed negatively with respect to any recommendation received by the professors. The Likelihood ratio test (LRT) for the full model showed a p-value of 0.9999 which provides very weak evidence of a relationship between the explanatory variable and the odds of obtaining a satisfactory grade. However, the missForest Model yielded a p-value of 0.0865 which provides a moderate relationship between a satisfactory grade and our four variables.

When we subdivide the data based on the recommends we get the following AICs, and different coefficients.

```
Table2
```

```
##
##
## |                 |Results (N = 4) |
## |:----------------|:---------------|
## |**R01 AIC **     |      |
## |   AIC |8.6686          |
## |**R12 AIC**      |      |
## |   AIC |21.282          |
## |**R2 AIC**       |      |
## |   AIC |34.032          |
## |**R4 AIC**       |      |
## |   AIC |9.0036          |
## |**R6 AIC**       |      |
## |   AIC |21.282          |
## |**R8 AIC**       |      |
## |   AIC |14.462          |
```



We can see in this graph that taking the recommended course improves the probability of achieving a satisfactory grade with the p-value of 0.0348, however it has an LRT value of 0.319. This indicates that while it does some explaining in the model, it is not enough to prove a relationship between Grade and RecTaken.

Discussion:

In the unsubdivided models of the missForest Model, we obtained different coefficients. What this means is that the odds of obtaining a satisfactory grade change based on what courses were recommended to the incoming students. The 95% confidence intervals for R01, R12, R4, R6 and R8 are wide enough as to be uncertain whether PSATM, Size, and PlcmtScore are reliably increasing or decreasing the odds of obtaining a satisfactory grade. The values of the coefficients in these models are highly dependent on what the values of the other variables are doing. In the R2 model, the odds of being accepted is a 1.1 to 15.5 times higher with each point increase in PSATM with respect to the other variables Size and PlcmtScore.

Limitations:

The ability of standardized testing to measure how well a student performs limits this study because certain students might take one test and not the other which limits the completeness of this dataset. Additionally, students might be stressed out by standardized tests and perform better own low-stress assignments like homework or a simple placement test. There were also many blank entries for the Grades in this dataset which limit the information provided.

Conclusions:

I was able to create a model that provided a strong relationship between Size, PSATM, Recommended, and Rank to determine what course a student should take. The successful predictions of the model suggest that we have sufficient criteria to predict what courses incoming students should take. There was also little evidence to support that taking the recommended course improves the chances of obtaining a satisfactory grade. The results of this study might improve the placement of incoming freshman by allowing colleges to focus on the criteria that best predict their student's success.

References:

Stekhoven DJ. Using the missForest Package. 2011. https://stat.ethz.ch/education/semesters/ss2012/ams/paper/missForest_1.2.pdf