edX Capstone Graduate Admissions Data CYO Project

Gary Hanson

5/24/2021

Section 1: Introduction

What do graduate schools look for when determining who should be admitted to their elite ranks? That is the question that every student, both domestic and international, wishes they knew the answer to so they could create the perfect resume and give them the best chance at securing entry to their dream school.

The goal of this project and analysis is to help answer that question, and create a predictive model that could be used to give an accurate prediction of admittance into graduate school.

The data set I will use contains 400 observations of 7 variables representing 400 international students who applied and were accepted or not to various US graduate schools. A description of the variables can be found below.

- admit = student's admittance status (1 = admitted, 0 = denied)
- gre = student's gre score (range 220 - 800)
- gpa = student's undergraduate gpa (range 2.26 - 4.00)
- ses = student's socioeconomic status (1 = high income, 3 = low income)
- rank = student's undergraduate school ranking (1 = high prestige, 4 = low prestige)
- Gender_Male = student's gender (1 = male, 0 = female)
- Race = student's Race (1 = Hispanic, 2 = Asian, 3 = African-American)

Before diving in to develop a model to predict admittance based on these various factors, I first had to analyze the data to look for any known correlations and to get an idea of the spread of the variables. I also had to clean up the data set to prepare for analysis by removing the admit column, then creating train and test sets of the data. Once the data was fully prepared, I then used various techniques and methods to come up with the optimal algorithm, and compared the results of the prediction to the actual values in the test set to calculate the accuracy rate. Finally, I compiled the various accuracies and methods into a table to select the optimal method for the predictive algorithm.

Section 2: Analysis

Data Cleaning

Before preparing the data and exploring the values, I first investigated to see if there were any missing data points or any improperly formatted values. Upon quick inspection, I noted that the data set was complete, and accurately compiled. This allowed me to very quickly move on to the data exploration phase.

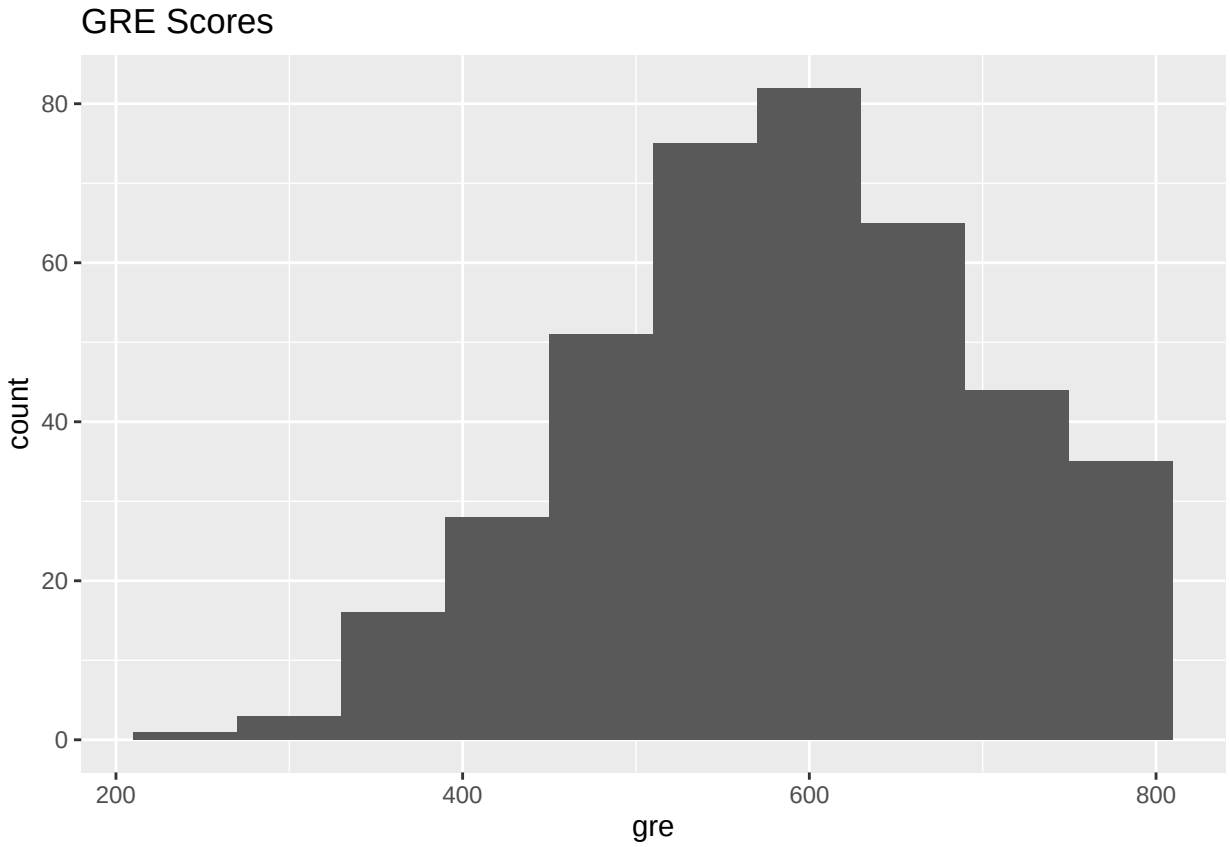
Data Exploration

The very first thing I wanted to determine, was the overall average admittance rate for the 400 students. Overall, the students in the set had an admittance rate of 0.3175.

Next, I wanted to look some of the variables to get an idea of their range and spread. Starting with the GRE (Graduate Record Examination) scores, I wanted to get a quick summary of the data which can be shown below:

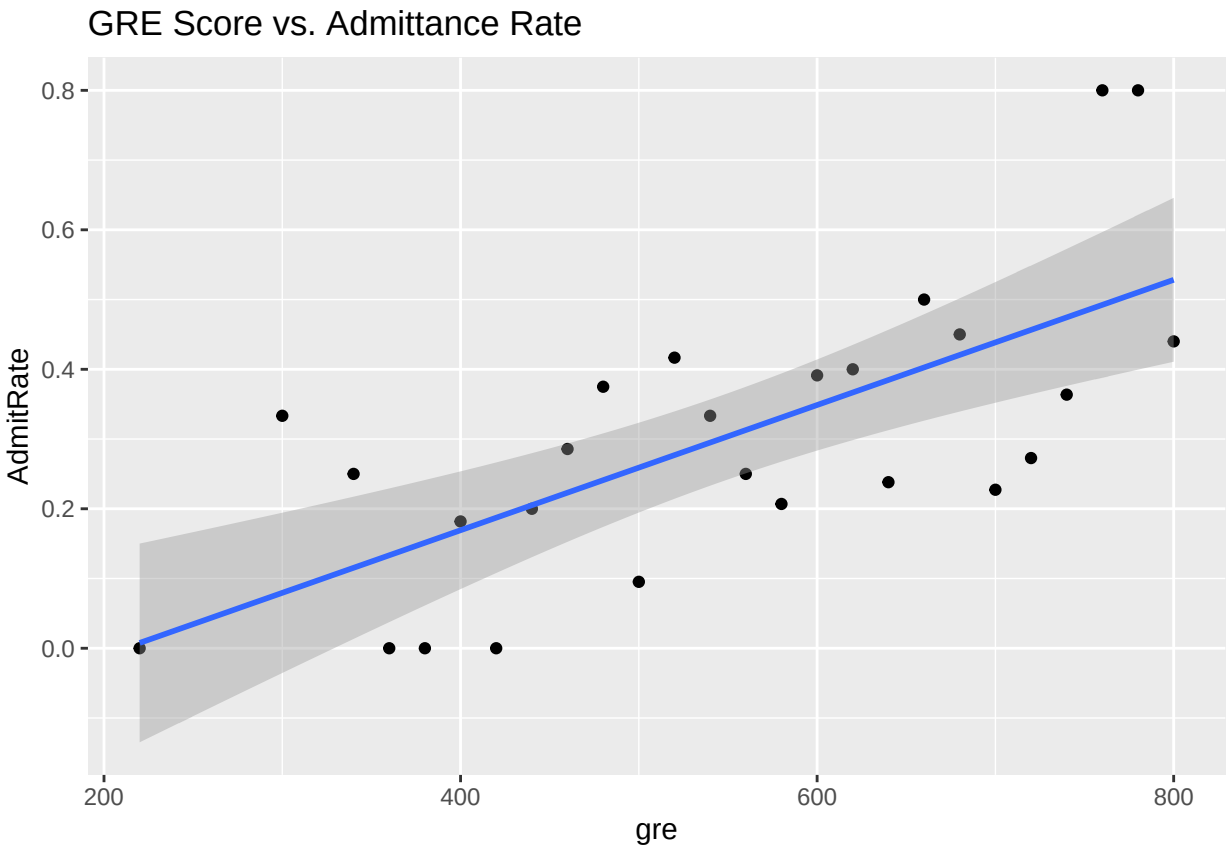
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   220.0   520.0   580.0   587.7   660.0   800.0
```

With this information, we can get a fast idea of the range of the values within this variable, but to better visualize the scores and how they breakdown, we can create a histogram of the data and see if it is roughly normally distributed. When creating the histogram, the most accurate graph used bins of width 60.

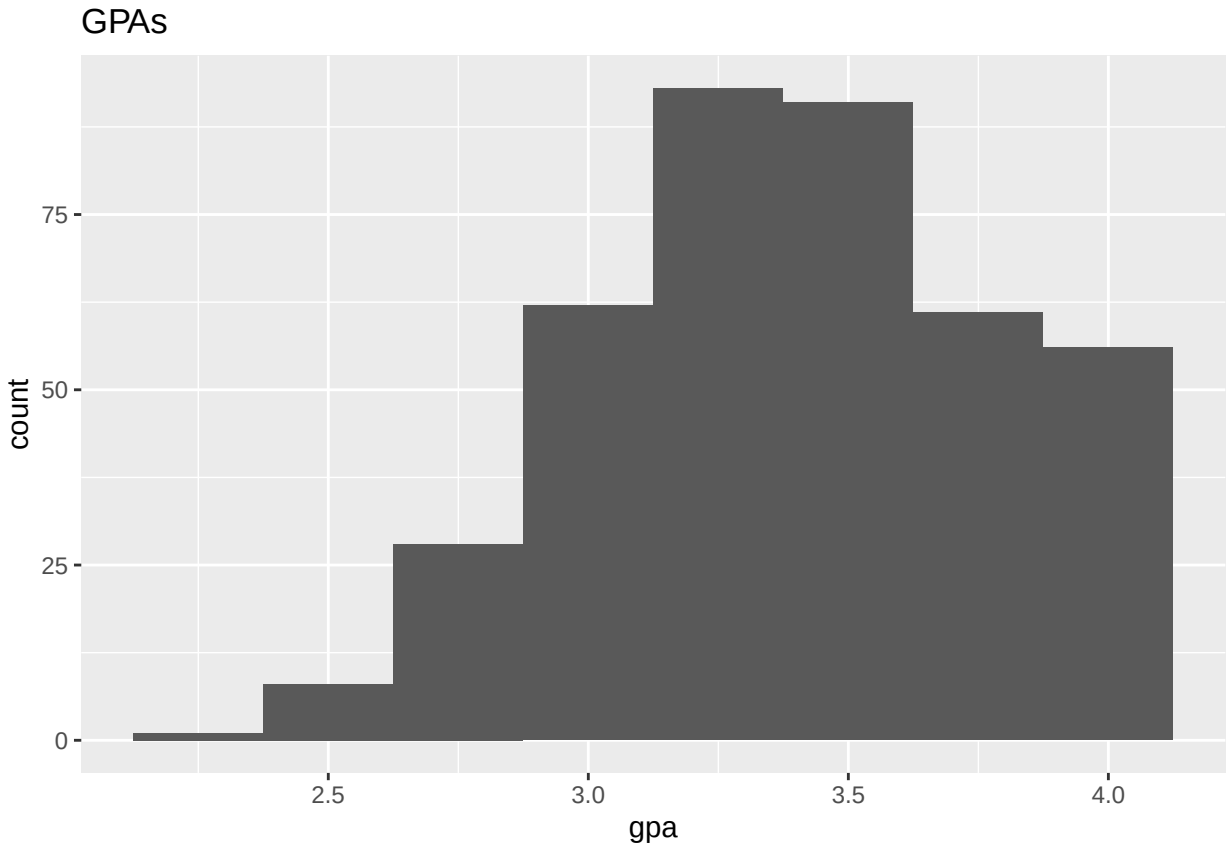


From this view, we can determine that the GRE scores are close to normally distributed with a slight left skew and a center close to 600. (Confirmed with our 5 number summary showing the median as ~580.)

I also wanted to see if there was a strong correlation between the GRE score of the student and their admittance rate. To display this visually, I first grouped the like GRE scores together, then calculated the percentage in each of those groups that were admitted. I then plotted the grouped GRE scores compared to their respective admittance rates. In this graph, you can see a slight positive correlation and notice that as GRE scores increase, generally the admittance rate increases as well.



I next wanted to analyze the only other variable that is fully in the loci of control of the student. The student's GPA. When running the 5 number summary of the GPA variable, we see that it ranges from 2.26 to 4.00 with a median of 3.395. To better see this, we can create a histogram with bins of .25 and see the spread of these values.



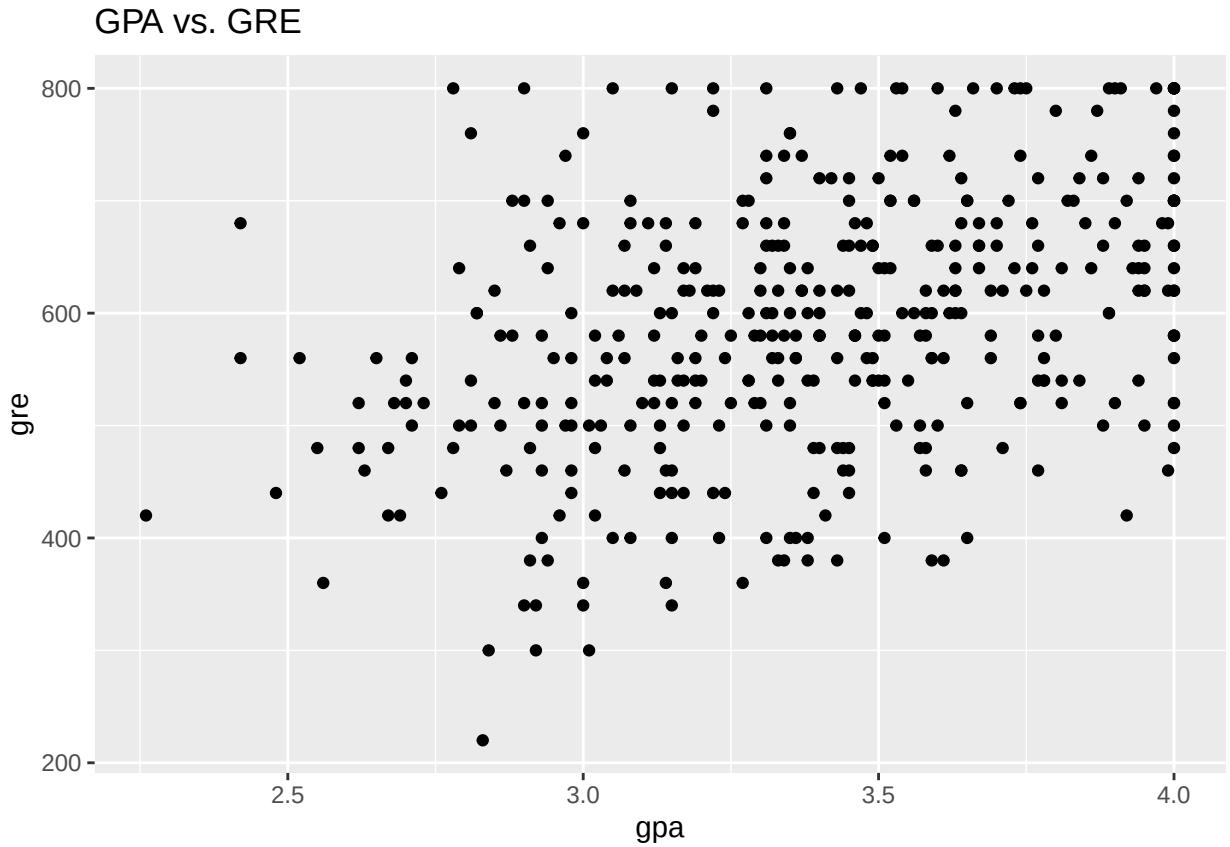
One thing that I noticed after viewing this data next to the GRE data, both sets are roughly normally distributed with a slight left skew. Which, when thinking using common sense, those typically applying to graduate school are those with higher GPAs and higher GRE scores and this data set confirms those thoughts.

To see if the GPA though had an effect on admittance rates, I grouped the GPAs to the nearest .5, then calculated the admittance rates for those individual groups.

```
## # A tibble: 4 x 2
##     gpa AdmitRate
##   <dbl>     <dbl>
## 1   2.5     0.25
## 2   3       0.244
## 3   3.5     0.337
## 4   4       0.402
```

It is clear to see that as the GPA increases, the admittance rate increases as well, with a dramatic difference in admittance rates for those with a roughly 2.5 GPA and those with a 4.0 GPA.

Further exploration of the GPA and the GRE scores show that there is some positive correlation between the two. The correlation between the two is `0.3842659` and this can be seen visually on the scatter plot below.



The only other variable in the data set that the student has any control over, is the rank of the undergraduate school they attend. When completing a similar analysis to that of GPA above, we can see that the rank (or prestige) of the undergraduate school the student attends is a significant factor in the admittance of the student into graduate school. For those students attending the most prestigious undergraduate schools, we see an admittance rate of over 50%, whereas for students attending the least prestigious schools, the admittance rate drops down below 18%!

```
## # A tibble: 4 x 3
##    rank     n AdmitRate
##   <dbl> <int>     <dbl>
## 1     1    61     0.541
## 2     2   151     0.358
## 3     3   121     0.231
## 4     4    67     0.179
```

Finally, the rest of the the data points are demographic values that the student has no control over. I will show below that some do tend to have an effect on admittance, but I will not be using them in my modeling as they are factors that students cannot change or alter. These data points however would help other research questions regarding the equity of student admissions.

Socioeconomic Status - Students with higher status are admitted at higher rates than those from lower status.

```
## # A tibble: 3 x 5
##     ses     n AdmitRate AvgGPA AvgGRE
##   <dbl> <int>     <dbl>  <dbl>  <dbl>
## 1     1   132     0.348   3.38   587.
```

```
## 2     2   139     0.309   3.41   597.
## 3     3   129     0.295   3.38   578.
```

Gender - The admittance rates for females and males are very similar, with females having a slight advantage.

```
## # A tibble: 2 x 5
##   Gender_Male     n AdmitRate AvgGPA AvgGRE
##         <dbl> <int>     <dbl>  <dbl>  <dbl>
## 1           0   210     0.329   3.40   587.
## 2           1   190     0.305   3.38   589.
```

Race - The admittance rates vary significantly between races, with Hispanic students getting admitted much more frequently than Asian students, despite the Average GPA and GRE scores for these groups being relatively similar.

```
## # A tibble: 3 x 5
##    Race     n AdmitRate AvgGPA AvgGRE
##   <dbl> <int>     <dbl>  <dbl>  <dbl>
## 1     1   143     0.371   3.38   591.
## 2     2   129     0.271   3.36   594.
## 3     3   128     0.305   3.43   578.
```

Data Preparation

Now that I have investigated the various data points in the data set we have, I want to pull some of the variables out that I will use in the analysis and create the train and test data sets. I will use the train set to train the algorithm and then the test set to test it for accuracy.

First, I am going to pull out the demographic values (ses, Gender, and Race) in order to train the model. I am doing this because they are immutable factors of a student and in an ideal world should not be used in determining admittance to a graduate school. Although I know that the disparities exist, they should not be a part of this algorithm as that would normalize the situation and perhaps give admissions offices the idea that it is expected. I will discuss this more in the conclusion of the report.

Because this leaves us with just 3 variables used to predict the outcome, I am forgoing doing a principal component analysis because there are only a few variables and the data set is relatively small. If this was a much larger data set, or if there were many more variables, I would perform the PCA to determine which variables contributed most to the variance, and use just those said variables to then create my algorithms.

When creating the partition for the train and test sets, I will use 75% of the data to train the algorithm, then the remaining 25% will be our test set. This should provide enough observations to train an accurate model while also giving us enough of a test set to get a good reading on it's accuracy. I will also separate out the admit variable into a separate factor so the model is trained solely on the three variables that students can control (GPA, GRE, and undergraduate school ranking)

Once the train and test sets were created, I also had to pull out the admit variable as a separate factor value in order to run the training function using the caret package. I created a separate admit factor for both the train and test sets.

Analysis

Finally, I am now ready to use various methods to train the algorithm and hopefully develop a very accurate predictive model.

I will be trying several methods to see which will give the most accurate prediction. I will also be combining the several methods in an ensemble prediction at the end to see if that would improve accuracy.

Overall Average The very first method I will try is by randomly guessing using the known average of the total set. This will likely not produce an accurate result, but would be the easiest formula to produce and so is worth testing out. I know the average of the set is roughly `0.3175`. So, if I create a random factor using that average, and compare it to the test set admissions, I get an accuracy of `0.59`. This is not very accurate at all, as predicted, and so further analysis is warranted.

Logistic Regression - GLM Next, I will try Logistic Regression. Using the caret package and the train/predict functions, I can quickly use the method "glm" to train the model, then create predictions on the test set, then calculate the accuracy. The accuracy of this model is `0.74` and you can see how this compares to the first comparison in the table here.

```
## # A tibble: 2 x 2
##   Method                Accuracy
##   <chr>                    <dbl>
## 1 Overall Average Guess     0.59
## 2 GLM                       0.74
```

We can see that this is significantly better than simply guessing, but I need to explore other methods to see if we can get an even more accurate prediction. Following similar processes, I tested the following other machine learning methods: LDA, QDA, Loess, KNN, and Random Forest. The accuracies of those methods can be seen in the table below.

```
## # A tibble: 7 x 2
##   Method                Accuracy
##   <chr>                    <dbl>
## 1 Overall Average Guess     0.59
## 2 GLM                       0.74
## 3 LDA                       0.74
## 4 QDA                       0.75
## 5 Loess                     0.72
## 6 KNN                       0.72
## 7 Random Forest             0.6
```

Up until this point, the most accurate method is the QDA (or Quadratic Discriminant Analysis) method. This is a variant of LDA in which an individual covariance matrix is estimated for every class of observations.

The last method I will try is to create an ensemble of the previous methods and will basically use whatever the different methods average out to to make my prediction.

When putting together the ensemble, I get an accuracy of `0.73`. Unfortunately, this is not as accurate as the QDA method, and we can see the various accuracies in the final Result table below.

```
## # A tibble: 8 x 2
##   Method                Accuracy
##   <chr>                    <dbl>
## 1 Overall Average Guess     0.59
## 2 GLM                       0.74
## 3 LDA                       0.74
## 4 QDA                       0.75
## 5 Loess                     0.72
## 6 KNN                       0.72
## 7 Random Forest             0.6
## 8 Ensemble                  0.73
```

Section 3: Results

Based on the analyses above, I can conclude that the `QDA` method is the optimal method for predicting admission based on the few factors we used. The accuracy of that model on the test data was `0.75`.

When using the QDA model on the entire data set (with the caveat that this might be inflated due to overtraining as we trained with 75% of this data), we get an overall accuracy of `0.715`.

Although this was the most accurate model I was able to produce, it is still far from being very accurate and very predictive. Ideally, I would be able to produce a model with at least 85% accuracy. But, there were some limitations I faced with the data set given and this model (or others) could be greatly improved with tuning or with further research and data acquisition.

Section 4: Conclusion

Overall, with just the 3 data points used (GPA, GRE, and Undergraduate School Ranking), I was able to develop a predictive model that would be accurate ~ 75% of the time. This would be significantly more helpful for students than if they were to simply guess or use anecdotal evidence from their peers or colleagues. This could also impact how students are working throughout their undergraduate careers and might also impact how students are selecting their undergraduate places of study. They can see that their GPA, their GRE score, and the ranking of their institution all play major factors in admissions offices decisions to admit (or not) international students.

As mentioned earlier in the report, there are many more factors that admissions offices consider before making a decision, and those things are not easily quantifiable or readily available to use in machine learning algorithms. Those things could include interviews with alumni of the graduate program, various personal essays the students could be submitting, letters of recommendation from their professors, etc. Admissions offices therefore are in a difficult position of coming up with objective ways of judging many subjective pieces of evidence. On top of those things, there are also other factors that shouldn't have an impact on admissions rates, but most definitely do as can be evidenced by various databases around the country and around the world. Gender, Race, and Socioeconomic status should not play a role in admissions, but as I noted above, there were some significant differences between acceptance rates for various Races and for various Socioeconomic Status.

Graduate programs however, may be able to use this data to make any adjustments to their admissions practices and policies to ensure an equitable distribution of the students they admit, and in the United States, it has been ruled and affirmed that these other demographic data points can be used to offer equitable admission to historically underserved or underadmitted students groups. So, it is hard to determine based on the given data set whether these differences in admission rates are corrections to previous inequalities, or if they are evidence of the inequalities themselves and should be reviewed by the graduate institutions to rectify any inequities.

Moving forward, to improve on the accuracy of this model and make a more useful prediction, I would do a few things. I would first do some research and try to find other data points that could be included in the data (i.e. scores on personal essays, rubric evaluations from interviews, etc.) and use more data points to help refine the prediction. I would also use tuning to select the best parameters for the various models as this might drastically improve the performance of the models using the data I already have. Overall, this project and this data set can offer invaluable insight into the factors that most effect student admission into graduate school programs and in terms of machine learning, can demonstrate some of the limitations that exist with smaller data sets and fewer variables compared to some of the larger data sets that we have used.