College Acceptance

Analysis & Predictions

Anit Mathew & Ritwik Katiyar

Summary

Our objective was to conduct a data analysis and search for characteristics that are crucial for forecasting the likelihood of admission. Our goal was to forecast both the GRE results and the chance of admission. Our findings indicate that a student's GRE and cumulative GPA have the biggest effect on their chances of being admitted. Additionally, we were able to create a model that could forecast GRE scores based on TOEFL results, GPA, and if the student had any prior research experience. Subsequently, we developed a model that could predict the likelihood of admission based on research experience, GRE scores, the quality of letters of recommendation cumulative GPA, and university rankings.

Introduction

Predicting student admission to a masters degree program is a challenging task due to the diverse backgrounds of the students, and an incomplete understanding of the precise skills that are most critical to success. In this study, the chance of a student admission is assessed using information from the admission application, such as standardized test scores, undergraduate grade point average, Statement of purpose rankings, research paper ranking, etc. Simple data analysis methods, data visualization and machine learning algorithms techniques are used to gain a better understanding of how these variables impact the chance of admission.

The dataset contains several parameters which are considered important during the application for Masters Programs. The parameters included are:

- GRE Scores (out of 340)
- TOEFL Scores (out of 120)
- University Rating (out of 5)
- Statement of Purpose (SOP) and Letter of Recommendation (LOR) Strength (out of 5)
- Undergraduate GPA (out of 10)
- Research Experience (either 0 or 1)
- Chance of Admit (ranging from 0 to 1)

The following table represents the first five rows of our data set:

Serial.No.	GRE.Score	TOEFL.Score	University.Rating	SOP	LOR	CGPA	Research	Chance.of.Admit
1	337	118	4	4.5	4.5	9.65	1	0.92
2	324	107	4	4.0	4.5	8.87	1	0.76
3	316	104	3	3.0	3.5	8.00	1	0.72
4	322	110	3	3.5	2.5	8.67	1	0.80
5	314	103	2	2.0	3.0	8.21	0	0.65
6	330	115	5	4.5	3.0	9.34	1	0.90

The queries that we would be researching are as follows:

1. Is the mean chance of getting an admission greater than 60%
2. How strongly are the features correlated to each other
3. Which factors are the most important in predicting the chance of admission?
4. Constructing a model that can predict the chance of admission
5. Developing a model that can predict the GRE scores

Methodology

The data was collected by Mohan S Acharya, Asfia Armaan, Aneeta S Antony and was downloaded from Kaggle. This dataset is inspired by the UCLA Graduate Dataset. Link: Graduate-Admissions Dataset

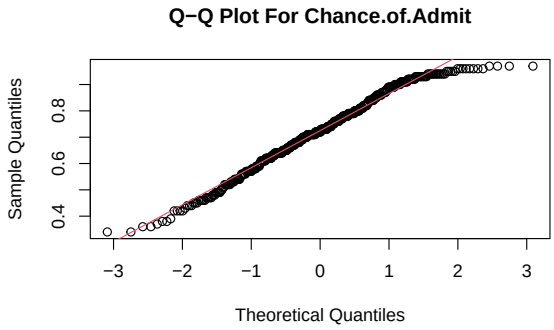
1. Data Cleaning: The data contained almost 400 records. Data will be cleaned by checking for null values and dropping the Serial No. column as it is irrelevant.
2. Then, we will be checking the distribution of all variables to check for missing values and abnormalities
    - Start by creating histograms
    - Generate qq plots for our target variables
3. We will run a one-sample t-test to determine if the mean chance of admission for a student that applies is greater than 60%
4. We could then check for the level of correlation between the variables by using a correlation matrix.
5. We will be conducting Random Forest Regression to check the importance of predictor variables.
    - The model will be first assessed to make sure that we are obtaining the best results possible.
6. Moving forward, we will be running a multiple linear regression model to predict the chance of admission.
    - The model will be optimized to have the best model scores possible using various methods.
7. Finally we will be running another multiple linear regression model to predict GRE scores.
    - Once again the model will be optimized to insure optimal results.

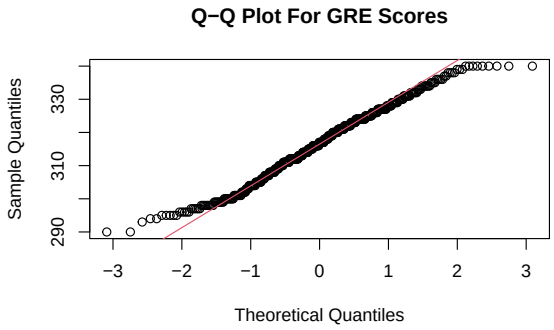
Results

Before running any tests we first need to prob our data. To make sure that there are not any missing observations or anomalies in order to check for that we can we can simply plot frequency histograms for each variable.

[The plots have been moved to the appendix, check Figure 1 in the appendix]

We cannot plot a histogram for the variable 'research' because it is a binary variable. This means that the observations for 'research' are 0's (meaning no research experience) and 1's (meaning they have research experience). Similarly, there is no need to construct a plot for 'University Rating' because the ratings are between 1-5 integer values. We can see from the plots that most variables seem to have no real issues that would require any cleaning or data transformations. Furthermore, the chance of admissions and GRE scores are the target variables for our model. Hence, we need to check for normality for those variables. Which we can archive by constructing a quantile-quantile (q-q) plot.





Since the majority of our data points seem to be on the q-q line, asides from the data points at the end. We can assume that the data for the chance of admission as well as the GRE Scores are normally distributed and we run statistical tests as well as machine learning models on our data.

One Sample t-test

To check if the mean chance of admission for any student that applies is greater than 60% (or 0.6) we can run a simple t-test. Since we are assuming that the chance of admission is greater than 60%; that would be our alternate hypothesis.

H_o: $\mu = 0.6$
H_a: $\mu > 0.6$

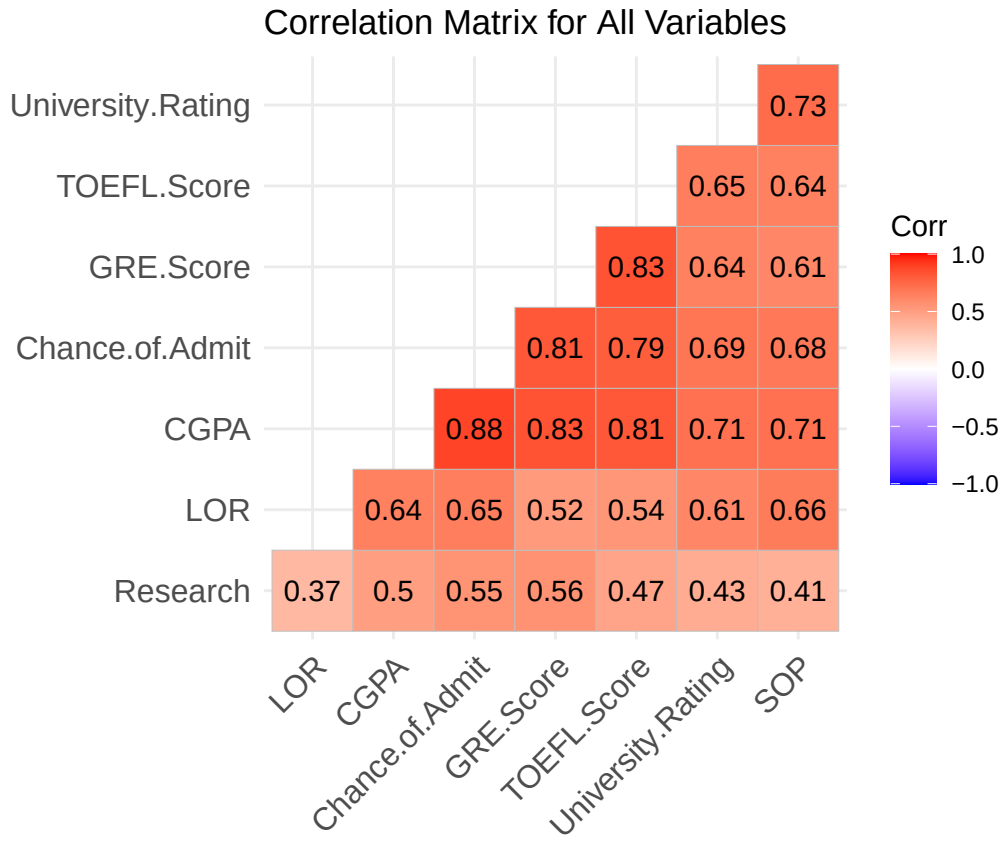
```
data:  data$Chance.of.Admit
t = 19.287, df = 499, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0.6
95 percent confidence interval:
 0.7093386 0.7341414
sample estimates:
mean of x
  0.72174
```

If we assume an $\alpha = 0.05$ We Reject the null if p-value is < 0.05.

Since our p-value is less than 0.05 we can reject the null and conclude that there is significant evidence to suggest that the mean chance of admission for any student that applies is greater than 60% based on the output we can also see that the mean of our sample is 72.14% with a confidence interval of 95% being between 70.9% and 73.4% chance of admission.

Correlation Matrix

To check for correlations between the variables and if they have any correlations or patterns that can be useful to us. We can construct a correlation matrix. Furthermore, we can look for any multicollinearity our data may have.



We can see that most of correlations are positive and that the highest correlation is between GPA and chance of admission. We can also observe that most of the strong correlations are between GPA and other variables. Meaning that GPA probably is a very important factor when it comes to determining chance of admission. We can also see that almost all variables have some form of corelation to one another meaning that we would need to watch out for multicollinearity in our models. To get a better understanding of which variables have more importance compared to another we can build a Random Forest model.

Random Forest

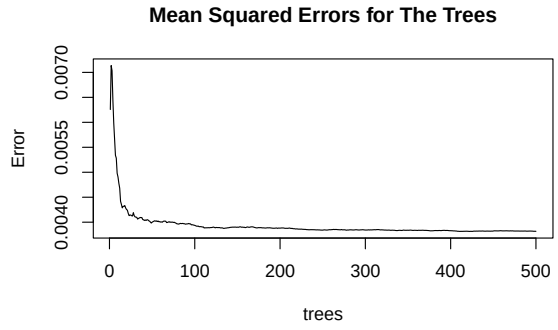
Random forest is a machine learning algorithm that takes the output from multiple different decision trees to reach a specific conclusion. Without getting into specifics a decision tree is a type of machine learning algorithm that constructs a flow chart that it then uses to predict a target variable. However, using a decision tree results in high variance since no two decision trees results are the same. To combat

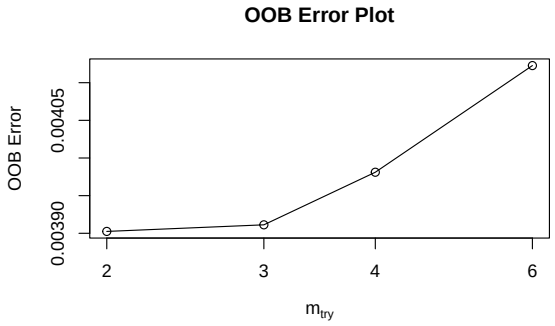
that random forest algorithm bootstraps the samples from the data set and then builds decision trees for each bootstrapped sample. Then it takes the average of the trees to formulate results. We can then use these results in order to generate a plot that shows the importance of a variable (Importance in this case refers to the variable's ability to help estimate the target value).

```
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 0.003817467
                    % Var explained: 80.8
```

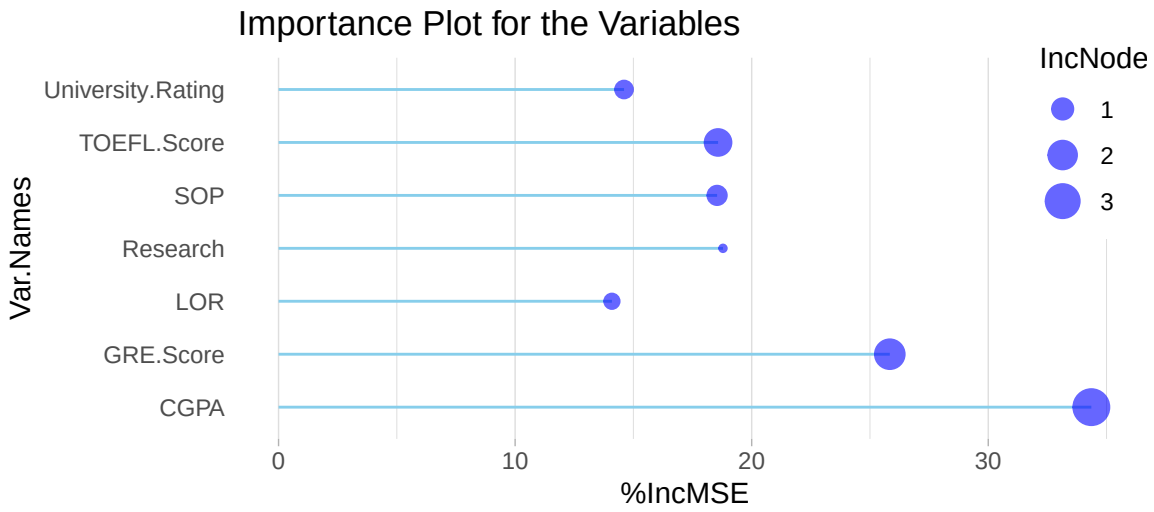
Based on the random forest model's scores we notice that our percent variance explained is almost 81%. We can construct a plot for all the mean squared errors for every decision tree that was constructed by the model.





In the plot to our left we can see that our highest mean squared error is under 0.01. And the mean squared errors seem to approach almost 0 as more trees were constructed. This means that we are getting accurate results. However, we could try and optimize our random forest and make sure that the number of trees constructed is the right amount. We can use the tune function to assess our random forest model.

Based on the plot on the right we can see that the lowest out-of-bag (OOB) error is achieved by only using two randomly chosen predictors at each split; this means that our parameter of 300 trees seems to be working well enough and we do not need to remove or add more trees.



The above plot showcases that GPA is highly important in order to determine the chance of admission; followed by the GRE score, whether the applicant possessed research experience, and TOEFL score respectively. Moreover, to predict the chance of admissions we can use a Multiple Liner Regression Model.

Multiple Linear Regression

Given that we are already aware of possible multicollinearity issues within our predictors. Before we even run a model we can try and optimize our results. One way to do that is through subsetting.

[Subsetting plots in the Appendix as Figure-2]

From the top left plots for the adjusted R-squared we can determine that we need six variables in order to get the best results R-squared results. By following the top right plot we can conclude that those variables are GRE-score, TOEFL Score, Uni.rating, LOR, CGPA and Research. If we run a model we get the best R-squared value of 0.8197. Since we are concerned with multicollinearity we can determine the Variance Inflation Factor (VIF) for our model.

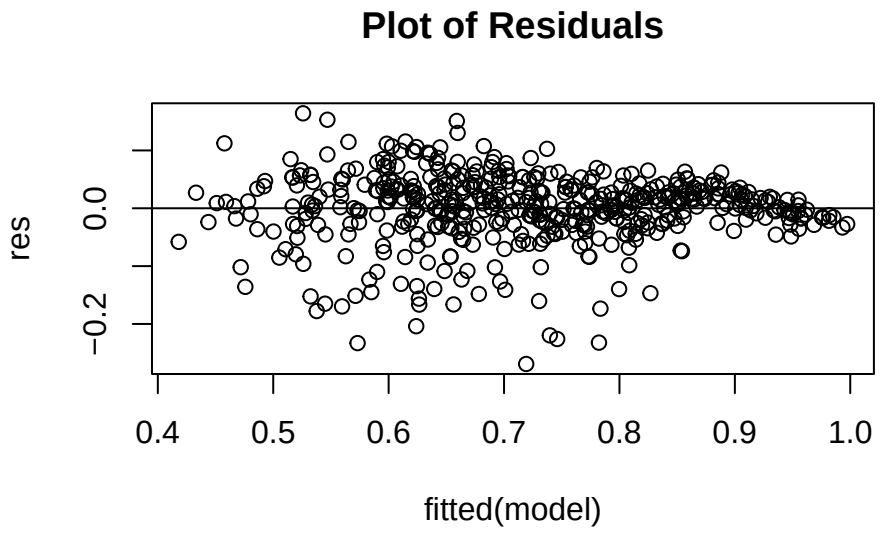
	x
GRE.Score	4.459541
TOEFL.Score	3.867976
LOR	1.853078
CGPA	4.619554
Research	1.493400
University.Rating	2.265898

Some variables have a high VIF. Granted the VIF is lower than the standard value of five, we could still do better. So we can try to determine the variables by another subset plot that gives the lowest Bayesian Information Criterion(BIC) score. From the subsetting plots (Figure-2) we can look at the bottom left plot and ascertain that we need five variables to get the lowest BIC score. We can then

determine the name of the variables by looking at the bottom right plot, which tells us that we do not need the TOEFL score. if we now build our model we get the following results.

	x
GRE.Score	3.520089
LOR	1.851740
CGPA	4.187533
Research	1.490232
University.Rating	2.225758

We can notice that our VIF score has dropped for all the variables. Although, we do end up sacrificing our adjusted r-squared value by about 0.09. However, with reduced multicollinearity, our results should be more reliable. We can now run a residual plot to see if our model needs any more changes.



We notice that there is clustering as the x-value approaches one, which results in a cone shape. It happens beacuse our model is bound to the values between zero and one since we are trying to predict the chance of admission. Also beacuse our mean for the chance of admission is high, the model struggles to form much variance as it approaches one. Even though there is a pattern we can understand that the pattern is simply due to the limitations of our data. There does not seem to be anything wrong outside of that so we can move on with our model and make further conclusions.

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.3011448  0.1042504 -12.481  < 2e-16 ***
GRE.Score          0.0025978  0.0004499   5.774 1.37e-08 ***
LOR                0.0176304  0.0039827   4.427 1.18e-05 ***
CGPA               0.1284335  0.0091642  14.015  < 2e-16 ***
Research           0.0233705  0.0066544   3.512 0.000485 ***
University.Rating  0.0079488  0.0035337   2.249 0.024926 *
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0605 on 494 degrees of freedom
Multiple R-squared:  0.8181,    Adjusted R-squared:  0.8162
F-statistic: 444.3 on 5 and 494 DF,  p-value: < 2.2e-16
```

Based on our model our least squared line can be given as:

$$Adm. = -1.301 + 0.0025 * (\beta_{GRE}) + 0.018 * (\beta_{LOR}) + 0.13(\beta_{CGPA}) + 0.023(\beta_{Res.}) + 0.007(\beta_{Uni.}) + \epsilon$$

The Residual standard error given by our model is 0.0605. Also, the final Adjusted Multiple R-squared score is 0.8162. We can also notice that all our predictors are highly significant asides from university ranking; which is significant as long as the alpha value is greater than 0.025.

After much hard work and analysis, we are now ready to use our model for predictions. To test our model we decided to elect the help of a student named John Doe.

GRE.Score	LOR	CGPA	Research	University.Rating
322	4	8.8	0	5

We can see that John Doe is a little above average. His scores were selected by looking at the general mean and giving him greater scores than the mean values. Using our model to predict his chance of admission we get:

```
The Chance of Admission Given by Our Model: 0.7758143
```

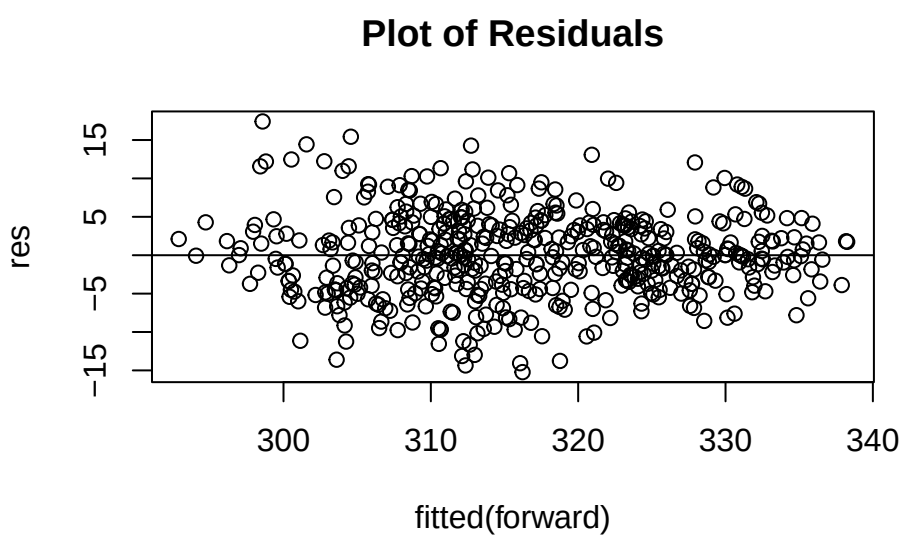
For our student, John Doe our predicted chance of admission from our model is 0.7758 or 77.58%. This makes sense since he is slightly above average and compared to the mean his predicted result is higher. Let's now try to predict his GRE scores before he has even taken the GREs.

Gre Scores Prediction: Multiple Linear Regression

Now that most of our queries regarding the chance of admission have been fulfilled. We will be disregarding the chance of admission and university rating because we are assuming that the student has yet to apply and hasn't taken their GRE test yet. For this model, we decided to use forward stepwise multiple linear regression to determine the model with the best statistical values. Since multicollinearity was such an issue in the previous model we decided to check for it first here.

	x
TOEFL.Score	2.958096
CGPA	3.089320
Research	1.355091

Based on the results we notice that the VIF scores are generally low so we can move on to looking at the residuals from the model.



Looking at the residuals we notice that the values seemed to be pretty spread out with no real pattern or a cone line shape. We can now take a look at the summary data of our model

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 165.47635    4.68915  35.289  < 2e-16 ***
TOEFL.Score   0.80023    0.06809  11.753  < 2e-16 ***
CGPA          7.35994    0.69968  10.519  < 2e-16 ***
Research      3.74180    0.56405   6.634 8.58e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.378 on 496 degrees of freedom
Multiple R-squared:  0.7746,    Adjusted R-squared:  0.7733
F-statistic: 568.3 on 3 and 496 DF,  p-value: < 2.2e-16
```

The least-squares line given by the model would hence be:

$$GRE = 165.47635 + 0.80023 * (\beta_{CGPA}) + 7.35994 * (\beta_{TOEFL}) + 3.74180(\beta_{Res.})$$

The Residual standard error given by our model is 5.378. Also, the final Adjusted R-squared score is 0.7733. We can notice that all predictors used are significant in predicting the GRE scores.

We are once again ready for predictions using our student John Doe, who in this scenario hasn't yet taken his GRE we can use the same GPA and research experience as before. We can assign him a new TOEFL score that is also slightly above the mean.

TOEFL.Score	CGPA	Research
110	8.8	0

```
The GRE score Predicted by Our Model: 318.269
```

The mean GRE score is 318.269 and keeping consistent with the trend the model has predicted John Doe's GRE score to be slightly above average. Meaning that our model seems to be performing well.

Limitations

We noticed that there is clustering as the x-value approaches one, which results in a cone shape. It happens because our model is bound to the values between zero and one since we are trying to predict the chance of admission. This is a limitation of our dataset which we cannot do anything about. Along with that, there is a strong multicollinearity among the features for which we selected the best model that generated the least multicollinearity by using subsetting. However, we cannot eliminate the multicollinearity completely.

Conclusion

- Based on our one-sided t-test, we determine that the mean chance of getting an admission for any student is greater than 60%.
- Based on the Correlation matrix, we can determine that the variables mostly have strong positive correlation among themselves. The highest correlation is between CGPA and chance of admission.
- Based on Random Forest regression, we can conclude that the features which have the highest influence on the chance of admission are CGPA and GRE.
- Further, our multiple linear model can predict the chance of acceptance with more than .80 of adjusted R-square. Our model also displays low VIF scores.
- And, for predicting the GRE Scores, our model has an R-square value of 0.77, which means that both the models are dependable.

Appendix

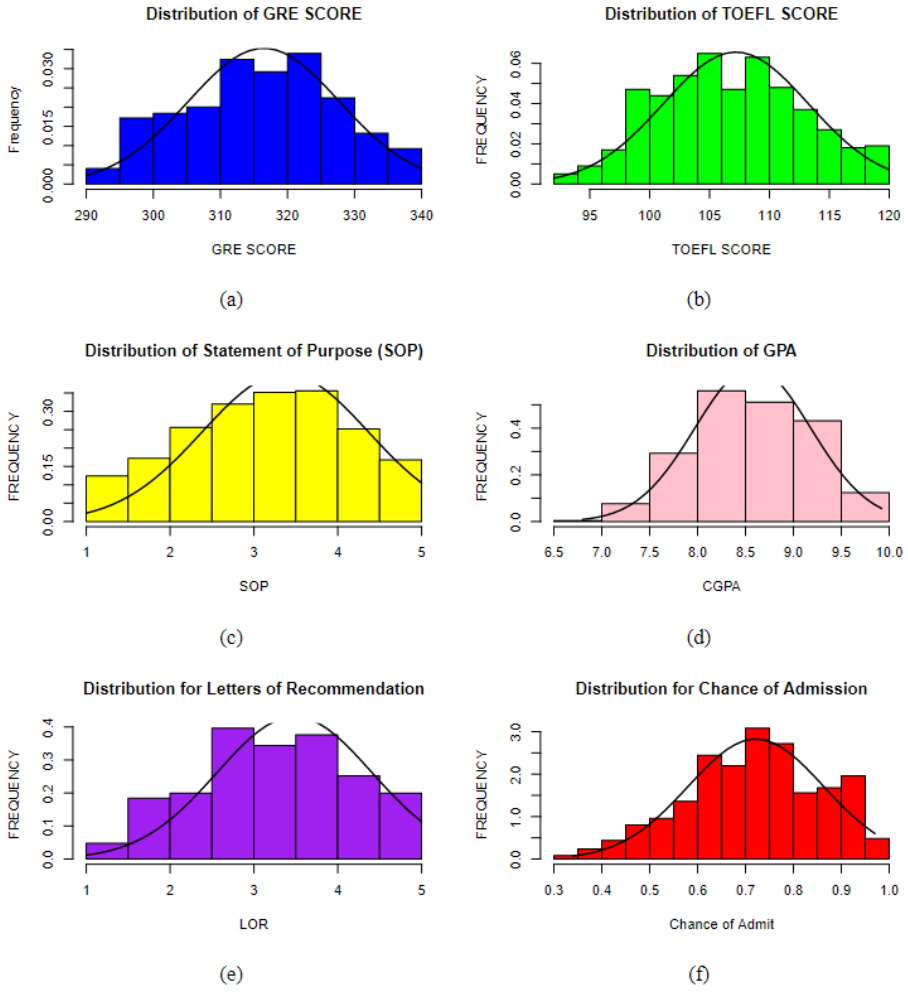
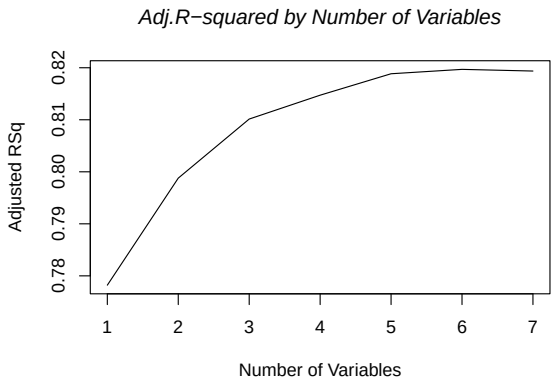
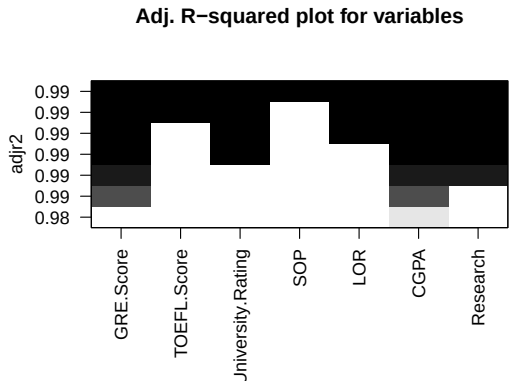


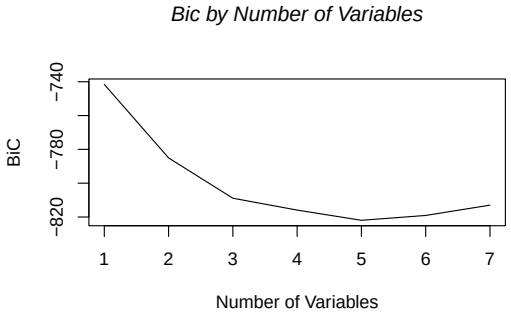
Figure 1: Frequency Distribution Histograms



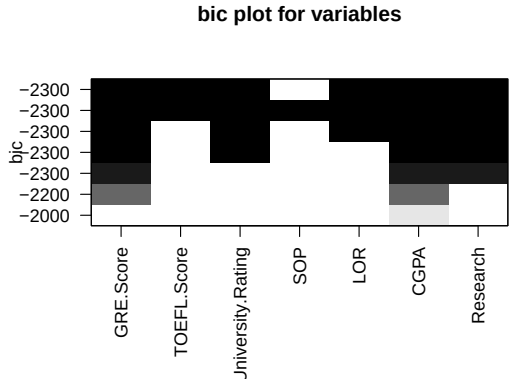
(a) Line Plot for Adj R-Squared



(b) Scale Plot for Adj R-Squared



(c) Line Plot for BIC



(d) Scale Plot for BIC

Figure 1: Subsetting Plots

References

[1] Aneeta S Antony Mohan S Acharya Asfia Armaan. *Graduate Admissions*. www.kaggle.com/datasets/mohansacharya/graduate-admissions. Accessed: December 17th 2022.

[2] Zach. *Random Forest Methodology*. www.statology.org/random-forest-in-r. Accessed: December 17th 2022. 2020.