University Admissions Data Analysis

Introduction

The goal of this final project is to select, explore and describe different aspects of data using combination of graphics and models. The dataset we have chosen for this project is University Admissions data from Kaggle datasets. This dataset contains information like GRE and TOEFL scores, Rating of University, SOP (Statement of Purpose) document of Student, LOR (Letter of Recommendation) of Student, CGPA (Cumulative Grade Point Average), Research conducted by Student and Chance of Admit. We will further use this dataset to answer questions like

1. What variable significantly affects the Chance of Admit: GRE, TOEFL, Rating, SOP, LOR, CGPA, Research?
2. Which data values are negatively affecting the response variable? (Outlier detection)
3. What effect does the statistically insignificant predictors have on the model? How does the model improve after removing them from the model selection process?

Dataset Description

The admissions dataset consists of 9 columns namely Serial No., GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, Research, Chance of Admit.

GRE Score: values ranging from 260 to 340, TOEFL Score: values ranging from 0 to 120, University Rating: 1 to 5, SOP: 1 to 5, LOR: 1 to 5, CGPA: 1 to 10, Research: 0 or 1, Chance of Admit: 0 to 1.

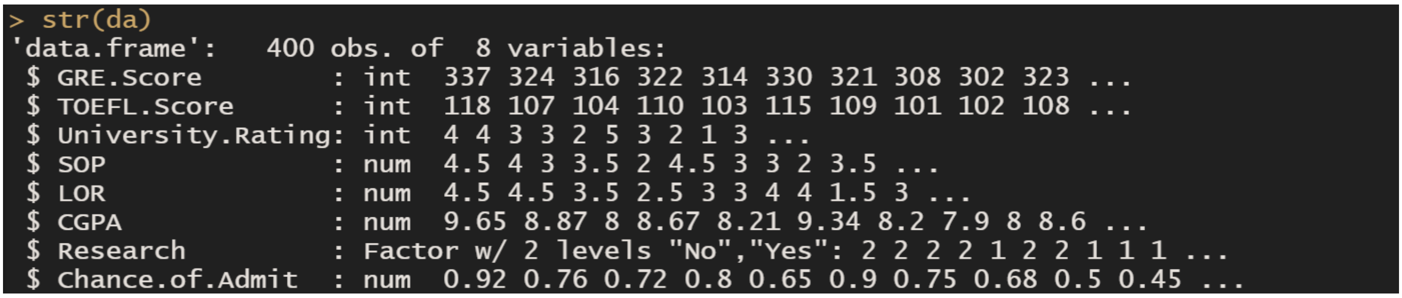
Approach

We first start the analysis of data by checking the following assumptions. Linearity, Constant Variance, Normality, Independence. We then continue with fitting a backward model which selects the best model with all significant predictors.

We further categorise the response variable into 0 and 1 to perform logistic regression. After fitting the model, we check for the misclassification error of the model and the accuracy. We then perform random forest classification on the model so that we can find the most significant variable of the model and compare it with the logistic regression.

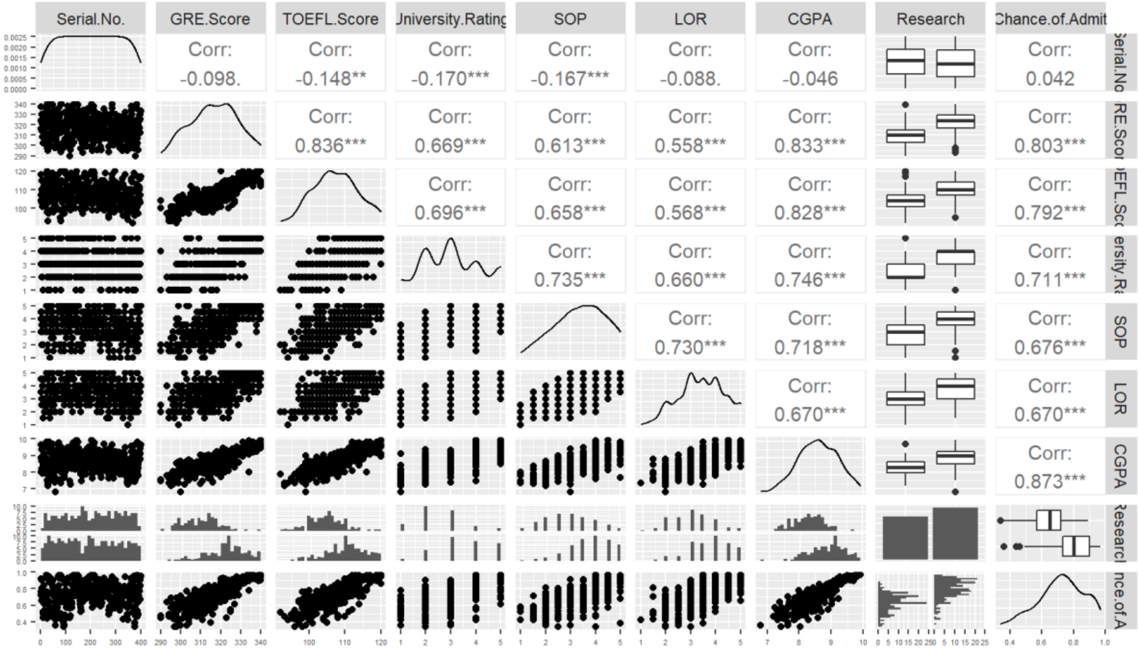
Data Pre-processing

In the dataset, we have a column Serial Number which is just a serial number of each student and it is the least significant variable of the dataset. So, we remove the column from the data and then we convert the values of Research which are 0 and 1 to No and Yes, so that it can be easily understood by the user and then we convert the column as a factor. We then have the structure of the pre-processed dataset:



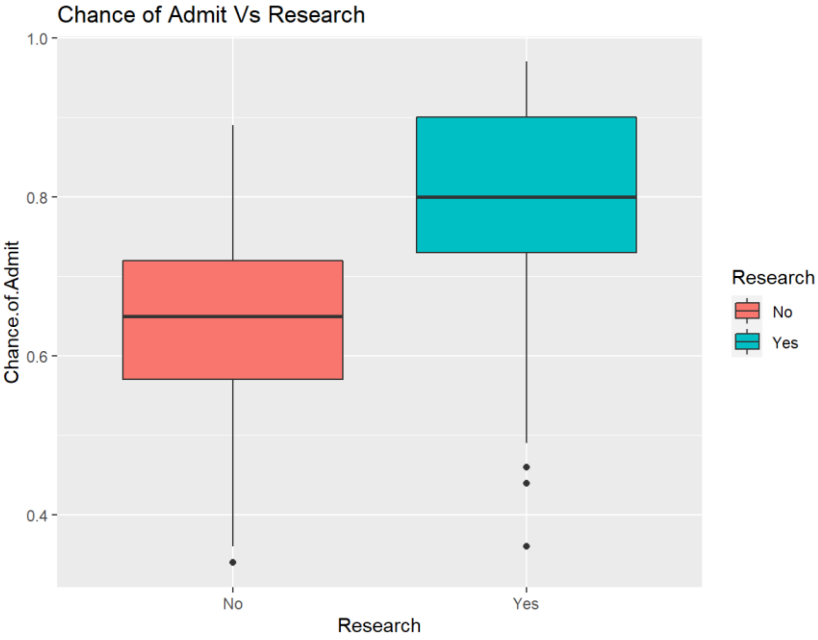
```
> str(da)
'data.frame':	400 obs. of  8 variables:
 $ GRE.Score        : int  337 324 316 322 314 330 321 308 302 323 ...
 $ TOEFL.Score      : int  118 107 104 110 103 115 109 101 102 108 ...
 $ University.Rating: int  4 4 3 3 2 5 3 2 1 3 ...
 $ SOP              : num  4.5 4 3 3.5 2 4.5 3 3 2 3.5 ...
 $ LOR              : num  4.5 4.5 3.5 2.5 3 3 4 4 1.5 3 ...
 $ CGPA             : num  9.65 8.87 8 8.67 8.21 9.34 8.2 7.9 8 8.6 ...
 $ Research         : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 2 1 1 1 ...
 $ Chance.of.Admit  : num  0.92 0.76 0.72 0.8 0.65 0.9 0.75 0.68 0.5 0.45 ...
```

Correlation matrix of variables



As we can see from the matrix, least correlation with the response variable (Chance of Admit) is Serial Number as it has no predictive power in the model. The next least variables are SOP and LOR. The highest correlation value is from CGPA (0.873), second highest value is from GRE Score (0.803).

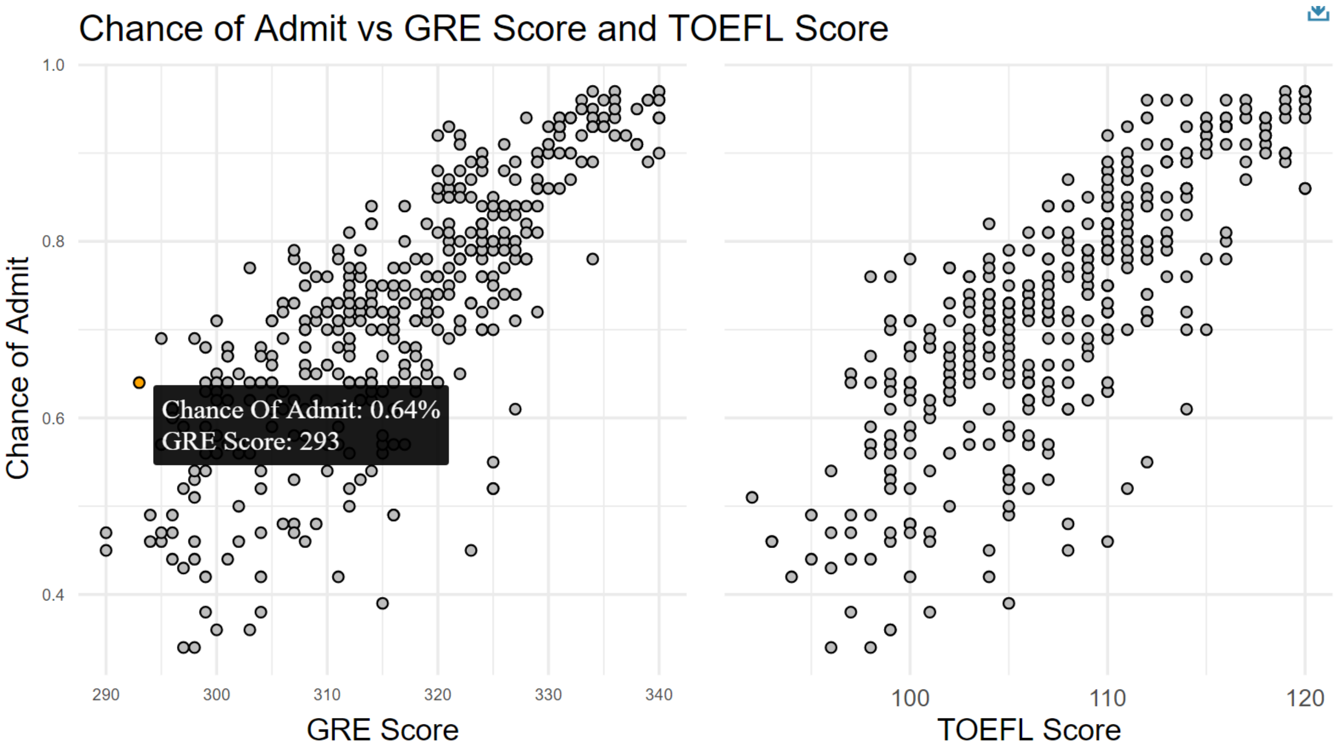
Boxplot for Chance of Admit vs Research



As we can see that the Research variable after converting it to a factor, in the box plot if the Research value is 1 i.e., Research exists for that student then there is a higher chance for the student to get admitted into the university. We can also see that there are few outliers in Research=1 box, it means that even when Research existed in certain students, their chance of getting an admit was pretty low.

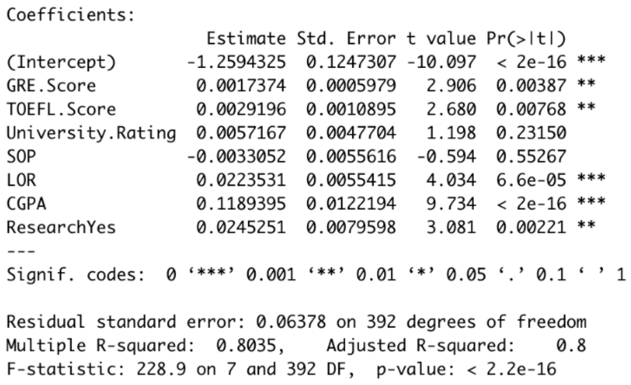
Interactive Scatter Point Plot

We have created an interactive scatter point plot for Chance of Admit vs GRE Score and TOEFL Score using Ggiraph package and ggplot2, We connected the two plots using giraffe() and show them side by side for clear understanding of the user.



The points when hovered display the Chance of Admit percentage and the GRE or TOEFL Score for that point. We will also upload a video displaying the interaction with the plot.

Linear Regression

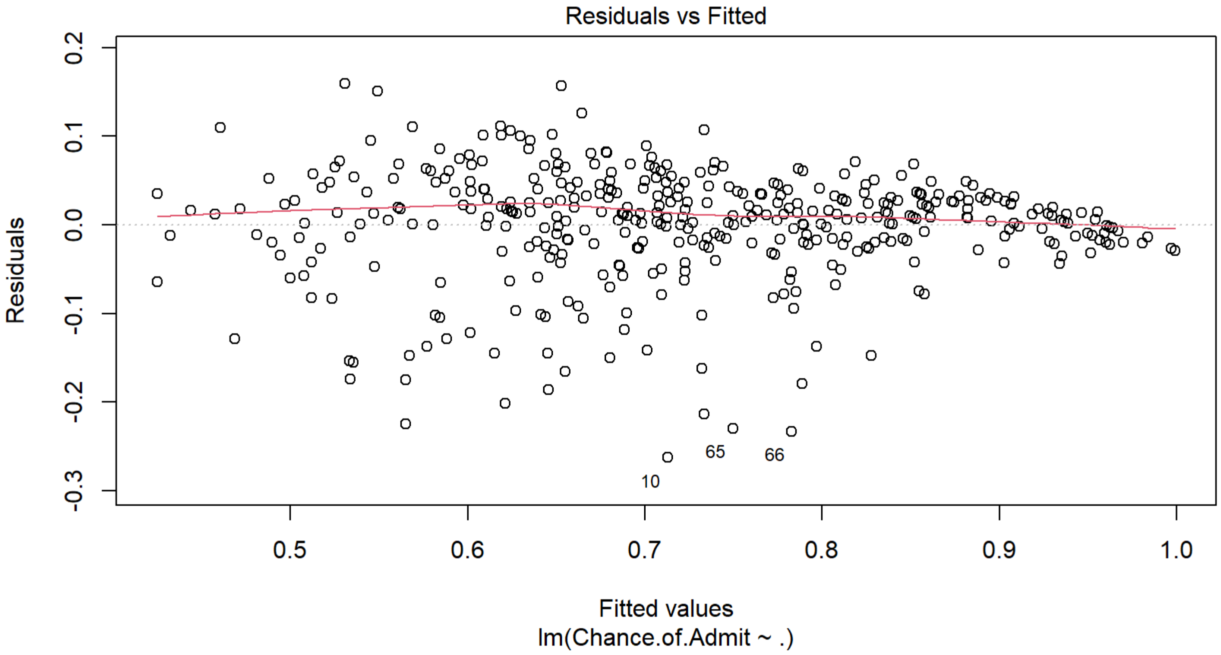


```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -1.2594325  0.1247307 -10.097  < 2e-16 ***
GRE.Score          0.0017374  0.0005979   2.906  0.00387 ** 
TOEFL.Score        0.0029196  0.0010895   2.680  0.00768 ** 
University.Rating  0.0057167  0.0047704   1.198  0.23150    
SOP               -0.0033052  0.0055616  -0.594  0.55267    
LOR                0.0223531  0.0055415   4.034  6.6e-05 ***
CGPA               0.1189395  0.0122194   9.734  < 2e-16 ***
ResearchYes        0.0245251  0.0079598   3.081  0.00221 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06378 on 392 degrees of freedom
Multiple R-squared:  0.8035,	Adjusted R-squared:    0.8 
F-statistic: 228.9 on 7 and 392 DF,  p-value: < 2.2e-16
```

From the summary of the model, we can see the significant variables are GRE Score, TOEFL Score, Letter of Recommendation, Cumulative Grade Point Average and Research. University Rating is generally considered as an important variable in terms of admission for student however according to our research, it has a minimal impact on university admissions.

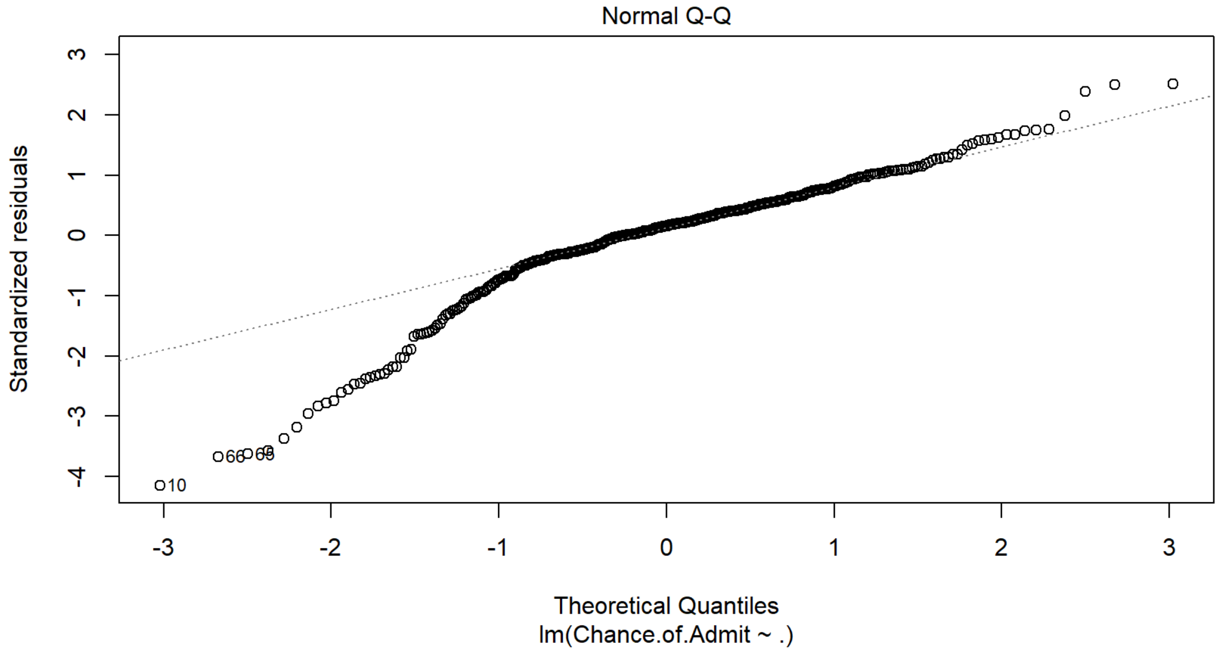
The diagnostic plots are as follows:

Residuals vs Fitted plot



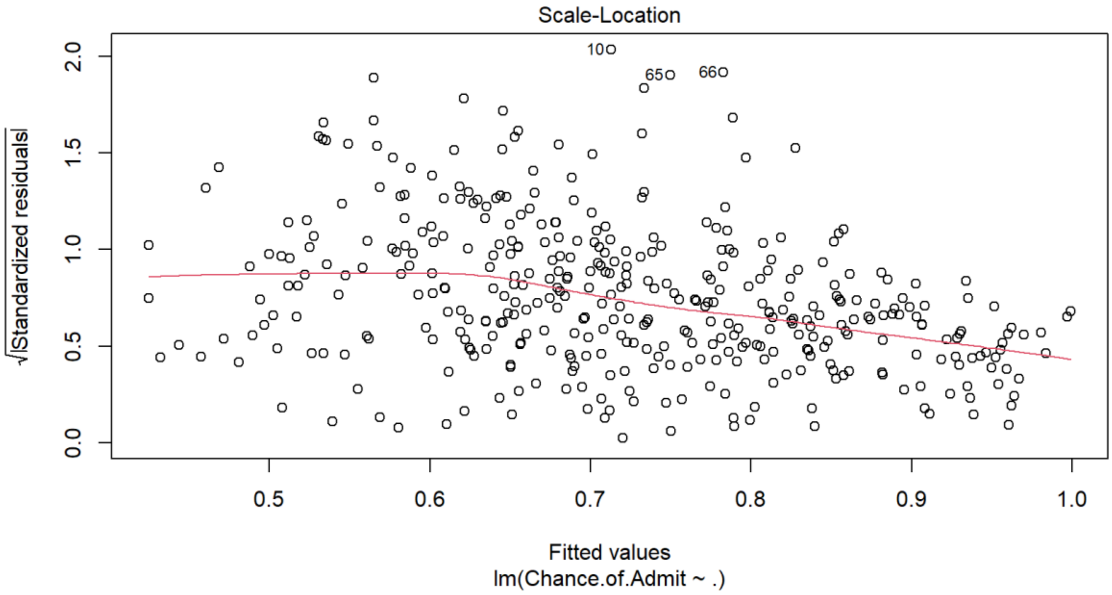
The residual vs fitted plot is used to check the linearity assumption. Random fluctuations of the residual values about zero indicates linearity and systematic patterns in the non-zero residuals indicates non-linearity. From the above graph, we can assume that linearity assumption is valid.

Normal Q-Q plot



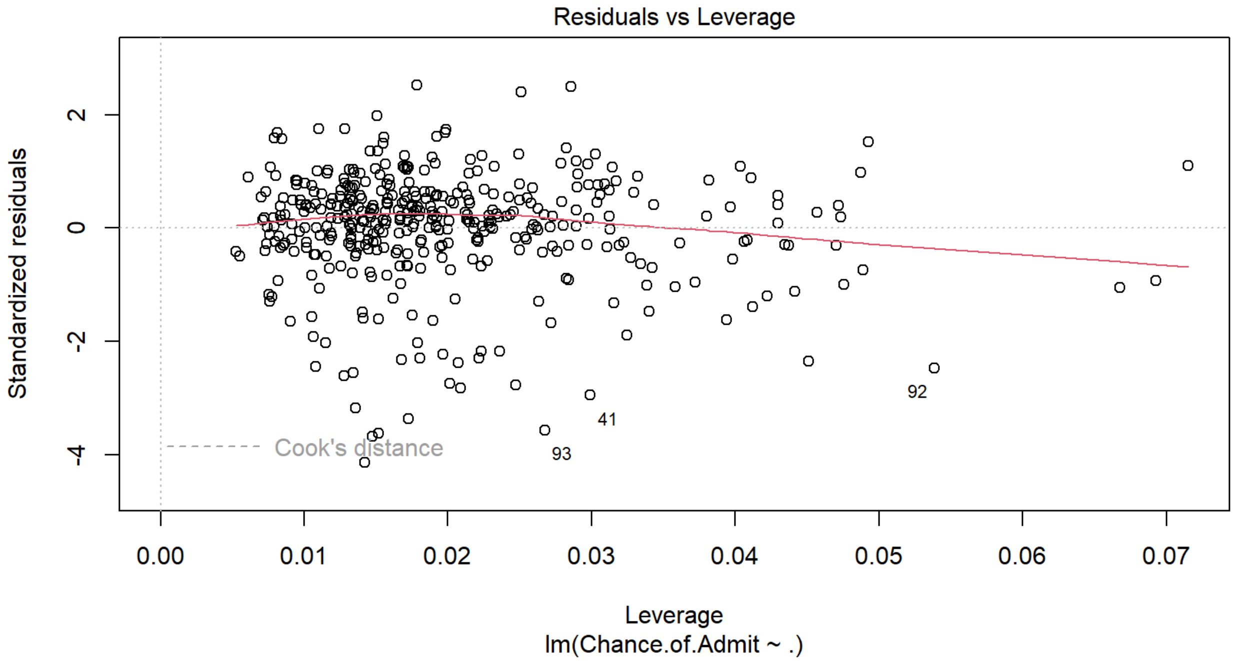
The Normal Q-Q plot is used to check the normality assumption. The residuals should mostly follow a straight line. From the above plot we can see that most of the points follow a straight line and hence we can assume that the plot shows normality.

Scale Location Plot



This plot can be used to check the equal variances assumptions. This plot shows if the residuals are spread equally along the ranges of predictors with a horizontal line. In the plot, points are equally spread but there are few outliers and absence of horizontal line violates equal variance assumption holding heteroscedasticity. To reduce this problem, outliers can be removed or log or square root transformation of fitted values can be applied.

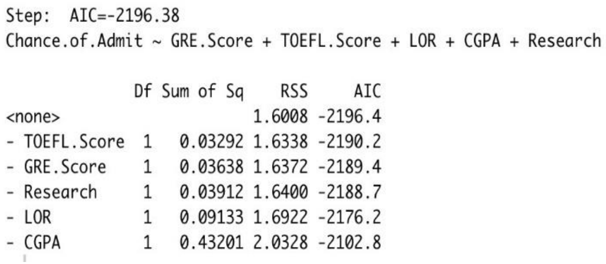
Residuals vs Leverage plot



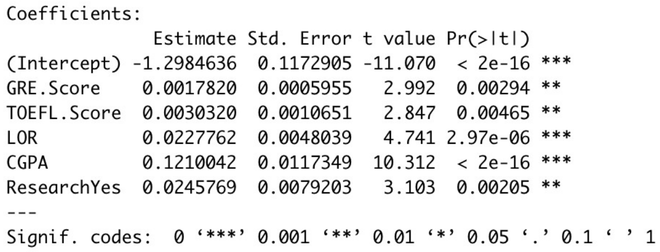
The residuals vs leverage plot can help us to find outliers and influential points if they exist. Influential points are generally located at the upper right corner or lower right corner. Cook's distance can be used to determine influential points. From the above graph, data doesn't show any influential points because all the points are well inside the Cook's distance.

To fit a better model and remove less significant variables, we fit a backward model for the dataset.

The best step AIC is -2196.38. The variables University Rating and SOP are removed as they are less significant variables. The best fit model is as follows:



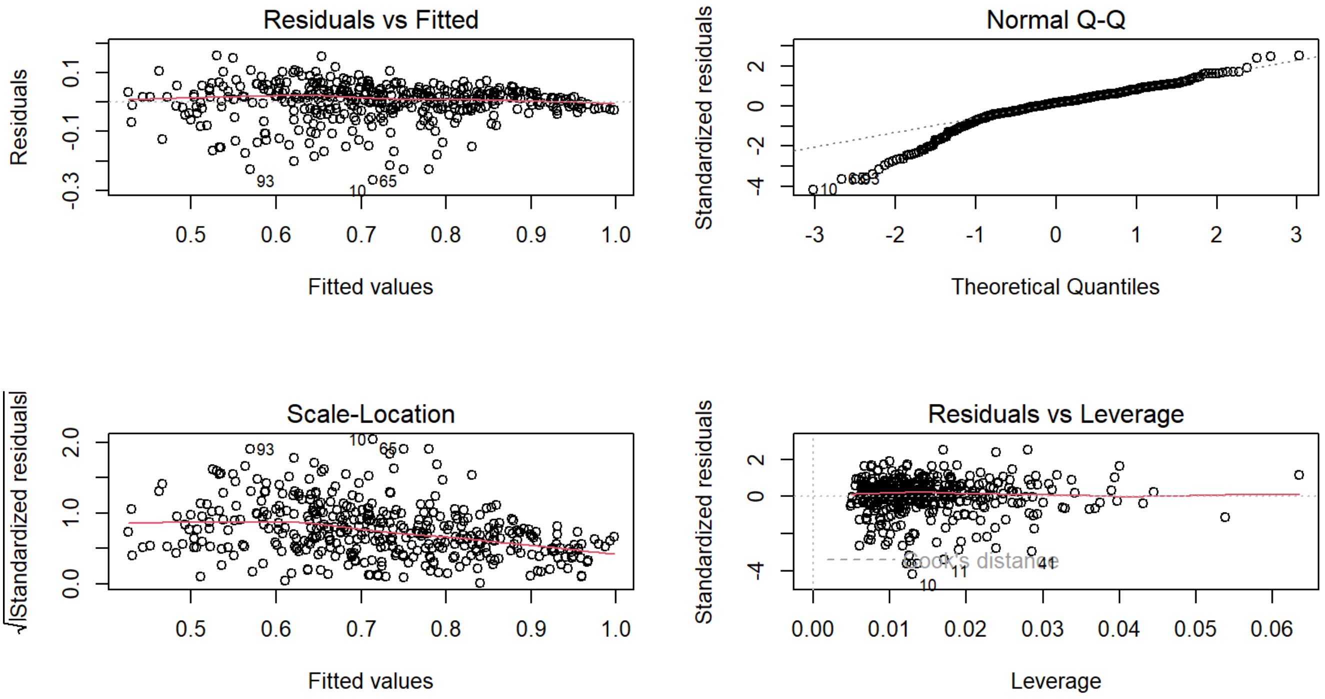
```
Step:  AIC=-2196.38
Chance.of.Admit ~ GRE.Score + TOEFL.Score + LOR + CGPA + Research

              Df Sum of Sq    RSS     AIC
<none>                     1.6008 -2196.4
- TOEFL.Score  1   0.03292 1.6338 -2190.2
- GRE.Score    1   0.03638 1.6372 -2189.4
- Research     1   0.03912 1.6400 -2188.7
- LOR          1   0.09133 1.6922 -2176.2
- CGPA         1   0.43201 2.0328 -2102.8
```



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.2984636  0.1172905 -11.070  < 2e-16 ***
GRE.Score    0.0017820  0.0005955   2.992  0.00294 ** 
TOEFL.Score  0.0030320  0.0010651   2.847  0.00465 ** 
LOR          0.0227762  0.0048039   4.741 2.97e-06 ***
CGPA         0.1210042  0.0117349  10.312  < 2e-16 ***
ResearchYes  0.0245769  0.0079203   3.103  0.00205 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see from the summary, all the predictors are significant with the response variable.

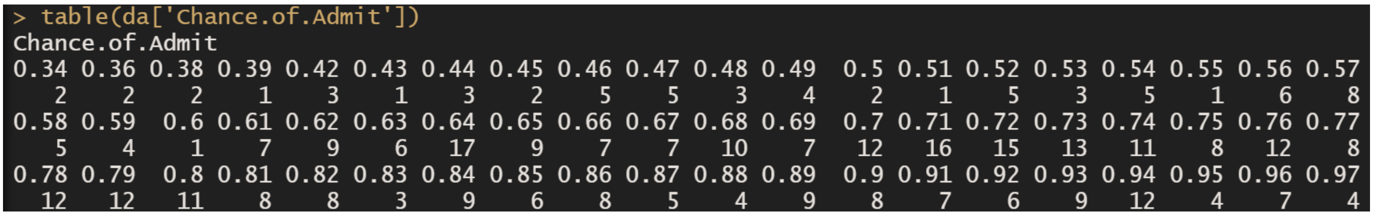
The backward model diagnostic plots are as follows:



From the backward fit model diagnostic plots, Residuals vs fitted plot holds linearity assumption, Q-Q plot holds Normality assumption, Scale-location holds equal-variances and Residuals Vs Leverage plot fits a lot better compared to linear model diagnostic plot of Residuals vs Leverage plot. In that plot we can see that at the end the line moves away from the center and in the backward fit model the line is aligned to the center.

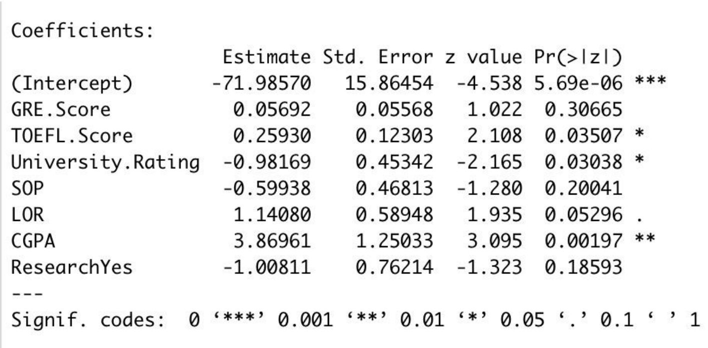
Logistic Regression

The next model we fit for the admissions data is the logistic regression model. For the model we had to categorise the Chance of Admit column. We can see from the figure that the response variable is not categorised

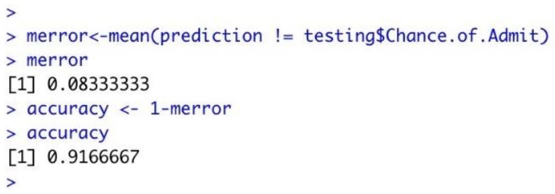


```
> table(da['Chance.of.Admit'])
Chance.of.Admit
0.34 0.36 0.38 0.39 0.42 0.43 0.44 0.45 0.46 0.47 0.48 0.49  0.5 0.51 0.52 0.53 0.54 0.55 0.56 0.57 
   2    2    2    1    3    1    3    2    5    5    3    4    2    1    5    3    5    1    6    8 
0.58 0.59  0.6 0.61 0.62 0.63 0.64 0.65 0.66 0.67 0.68 0.69  0.7 0.71 0.72 0.73 0.74 0.75 0.76 0.77 
   5    4    1    7    9    6   17    9    7    7   10    7   12   16   15   13   11    8   12    8 
0.78 0.79  0.8 0.81 0.82 0.83 0.84 0.85 0.86 0.87 0.88 0.89  0.9 0.91 0.92 0.93 0.94 0.95 0.96 0.97 
  12   12   11    8    8    3    9    6    8    5    4    9    8    7    6    9   12    4    7    4 
```

So, we categorised the data by taking the values below 0.5 as No and > 0.5 as Yes and convert the column as factor. Then we fit a logistic model for the data and the summary is as follows:



```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)      -71.98570   15.86454  -4.538 5.69e-06 ***
GRE.Score          0.05692    0.05568   1.022  0.30665    
TOEFL.Score        0.25930    0.12303   2.108  0.03507 *  
University.Rating -0.98169    0.45342  -2.165  0.03038 *  
SOP               -0.59938    0.46813  -1.280  0.20041    
LOR                1.14080    0.58948   1.935  0.05296 .  
CGPA               3.86961    1.25033   3.095  0.00197 ** 
ResearchYes       -1.00811    0.76214  -1.323  0.18593    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

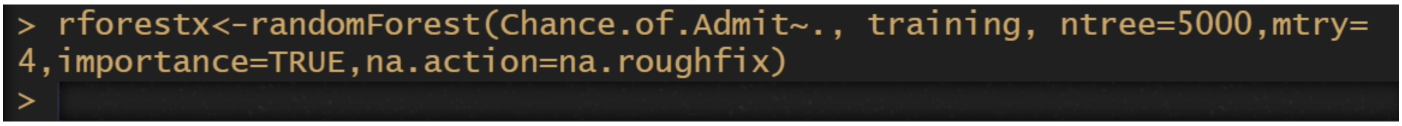


```
> 
> merror<-mean(prediction != testing$Chance.of.Admit)
> merror
[1] 0.08333333
> accuracy <- 1-merror
> accuracy
[1] 0.9166667
> 
```

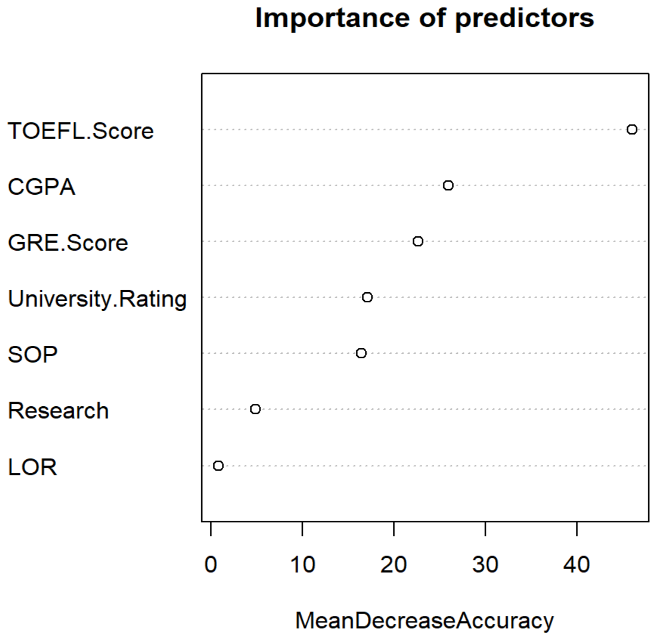
The significant variables of the logistic regression model are CGPA, University Rating and TOEFL Score with CGPA being the highest. Then we found out the misclassification error of the model was 0.08 and the accuracy and the values was 0.92 which is pretty good.

Random Forest Classification

To find out the most significant predictor in the data we built a random forest model for the data with 5000 trees and 4 random variables and then we found the most important variable using the varImpPlot() function and plots of Mean decrease in accuracy and Mean decrease in Gini Index.



```
> rforestx<-randomForest(Chance.of.Admit~., training, ntree=5000,mtry=4,importance=TRUE,na.action=na.roughfix)
> 
```



As we can see from the varImpPlot function the most important variable from the random forest model is TOEFL Score as it has high value in mean decrease accuracy plot and it affects the model negatively if it is removed. The conclusion we can draw from this model is that the most important variable is not CGPA as shown in the logistic regression but it is TOEFL Score.

Comments on the data

A good GPA score is an indicator of how well a student has performed in their field of study. The type of courses a student has taken in the past, how they connect to the course they are applying for, and how well they performed in it are all factors that the university looks at depending on the field for which the student has applied for admission. A student's complete transcript will give a clear picture of their academic potential based on these elements. A foreign applicant to a university in a nation where English is the official language must submit a TOEFL score. The GRE score used to be a significant factor in admissions. Due to the recent epidemic that has affected the entire world, many colleges have cancelled the GRE exam requirement for university applications and have begun to view the TOEFL as the only way to evaluate a student's performance in English language. This is the reason why our data and models predicted TOEFL and CGPA to be the most significant variables that affect the student’s chance of admitting into a university.

Conclusion

The data pre-processing and various models we used to fit the data and the analysis have answered the questions we discussed about in the Introduction part of the report.

1. For the most significant variable we got TOEFL Score from the Random Forest model as the most important predictor for the dataset.
2. For the data values which have almost no effect on the response variable was Serial Number and SOP variables which had very less correlation with the response variable.
3. In the backward fit model, we eliminated variables that are statistically insignificant in the model, the variables were SOP and University Rating. The plots were more aligned to the center and did not deviate as much as they did before, which made the model entirely significant.

The packages used for this analysis were ggplot2, tidyverse, Ggiraph, mlbench, randomForest, caret, GGally.

References

KHARE. Data for Admission in the University. Retrieved December 7, 2022, from https://www.kaggle.com/datasets/akshaydattatraykhare/data-for-admission-in-the-university

Ggiraph Make 'ggplot2' Graphics Interactive. Retrieved December 11,2022, from https://cran.r-project.org/web/packages/ggiraph/index.html