Grad School Admission Analysis

Group 2

STAT-515 | Prof. Dr. Tokunbo Fadahunsi

Sai Goutham Kumar Akula (G01355914)

Namrata Reddy Palle (G01284937)

Gowtham Mukkara (G01353489)

Abstract:

Choosing the best graduation college with the available scores is one of the most important aspects of graduating from high school. For our final project we selected a dataset which gives us the admission chances based on the scores obtained in several competitive exams, we initially created several visualizations in R-Studio to visually represent the data for different variables and compared the variables with the chance of admission to determine what all factors are responsible for one to get an admission. We created and analyzed different regression models like linear, logistic, and random forest to find the best fitting model with maximum accuracy and least error possible.

Introduction:

The graduate admission process is the set of procedures and policies used by universities to select graduate students. The process can vary significantly between universities and may also vary depending on the type of graduate program to which one is applying. The major depending variables are GRE score, TOEFL score, CGPA, Letters of recommendation, Statement of Purpose, Research experience, using all these variables we can find a potential college based on the university ratings and calculate the chance of admit to a specific university. As an international student, we are pretty familiar with how hectic the college admission process will be and I would be easier to select certain universities if we have data visualizations available for us to choose the colleges based on the scores.

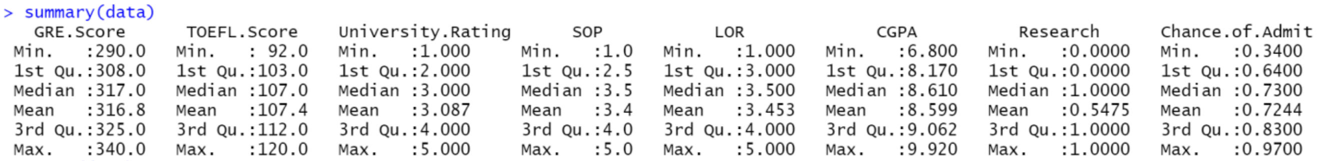
The chance of admission is mostly dependent on the specific university requirements and many of the universities have different priorities. However, in general, a high GPA and high standardized test scores will give you a better chance of being admitted into the university. There are many universities that use a holistic approach to review applications and consider many different factors. The most important factor that they consider is the applicant’s academic record. Other factors that are considered include the applicant's letters of recommendation, standardized test scores, essays, and extracurricular activities.

About Dataset:

This dataset was chosen among many others because of the various possible visualizations and analysis. It contains 400 rows and 8 variables which is enough to create and analyze linear models. The variables include GRE (graduate record examination) score, TOEFL (Test of English as a foreign language) score, University rating, SOP (statement of purpose) rating, LOR (letter of recommendation) rating, CGPA (cumulative grade point average in undergrad), Research experience (0 or 1) and the output that is Chance of admission, this dataset is selected from Kaggle website.

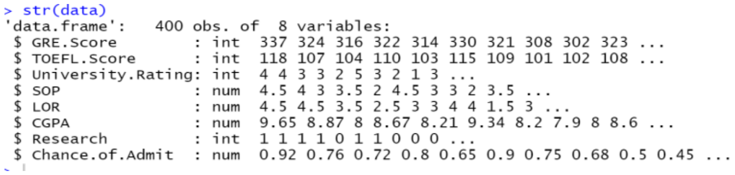
Summary Statistics:

The summary of all the variables was calculated and is displayed in the below figure by using the summary function.



```
> summary(data)
   GRE.Score      TOEFL.Score    University.Rating      SOP             LOR             CGPA          Research      Chance.of.Admit
 Min.   :290.0   Min.   : 92.0   Min.   :1.000     Min.   :1.0   Min.   :1.000   Min.   :6.800   Min.   :0.0000   Min.   :0.3400
 1st Qu.:308.0   1st Qu.:103.0   1st Qu.:2.000     1st Qu.:2.5   1st Qu.:3.000   1st Qu.:8.170   1st Qu.:0.0000   1st Qu.:0.6400
 Median :317.0   Median :107.0   Median :3.000     Median :3.5   Median :3.500   Median :8.610   Median :1.0000   Median :0.7300
 Mean   :316.8   Mean   :107.4   Mean   :3.087     Mean   :3.4   Mean   :3.453   Mean   :8.599   Mean   :0.5475   Mean   :0.7244
 3rd Qu.:325.0   3rd Qu.:112.0   3rd Qu.:4.000     3rd Qu.:4.0   3rd Qu.:4.000   3rd Qu.:9.062   3rd Qu.:1.0000   3rd Qu.:0.8300
 Max.   :340.0   Max.   :120.0   Max.   :5.000     Max.   :5.0   Max.   :5.000   Max.   :9.920   Max.   :1.0000   Max.   :0.9700
```

The structure of the data is displayed in the below figure.



```
> str(data)
'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         : int  1 1 1 1 0 1 1 0 0 0 ...
 $ Chance.of.Admit  : num  0.92 0.76 0.72 0.8 0.65 0.9 0.75 0.68 0.5 0.45 ...
```

Data Cleaning:

One of the most important tasks before plotting visualizations is to clean the data properly. All the missing values and null values were omitted. The dplyr package was used to edit the data. A column 'serial number' was omitted as it just gives us the number of columns available.

Data Visualizations:

Data visualizations are visual representations of data. They can take many forms, including charts, graphs, and maps. For our dataset we selected the following plots to visually represent and compare the data wherever possible.

To visually represent how chance of admissions is varied for different scores of GRE is shown below, this plot is created in R-Studio by using ggplot function aesthetics are defined for the plot and the plot type is given as geom_point for a scatter plot to be created. Once, the plot is created the points are filled as color based on the chance of admit and using scale_color_gradient function.

The scatter plot created clearly shows us that with increase in the GRE score there is more chance to get an admission.

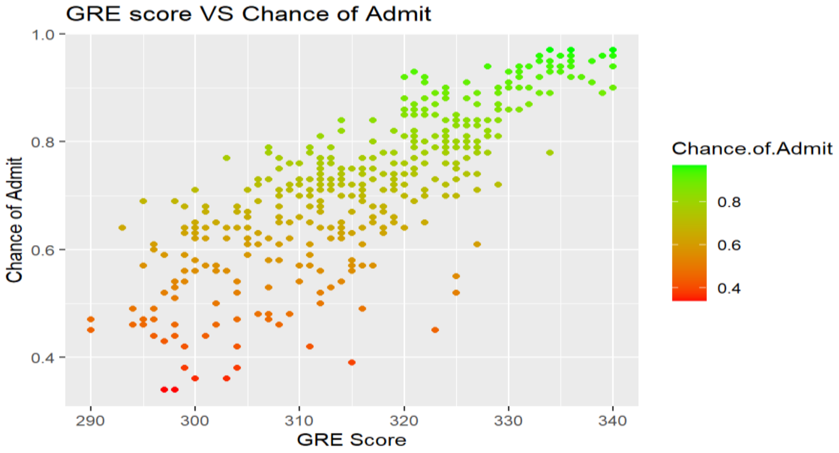


Figure 1

For the CGPA and Chance of admit a scatterplot is created using the same process as mentioned above and the below plot is obtained.

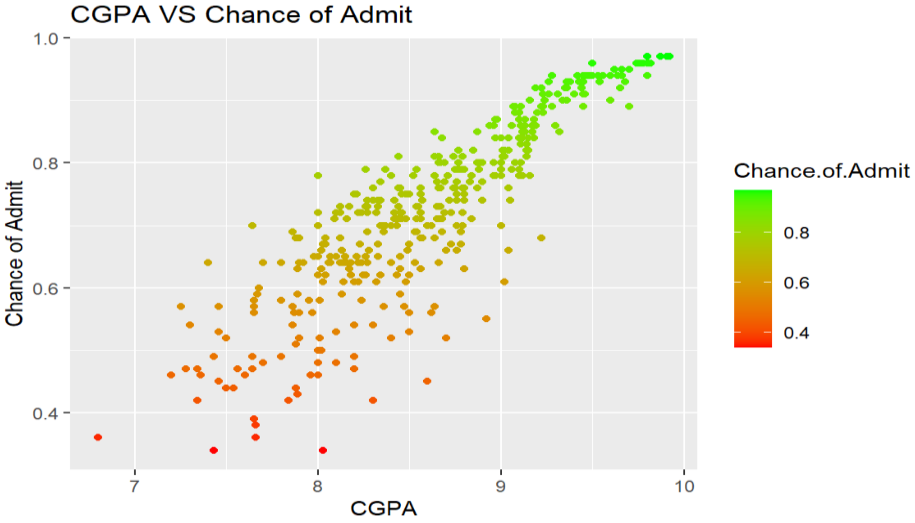


Figure 2

From figure 2 we can conclude that with the increase in CGPA the chance of admission increases.

To show the relation between TOEFL score and chance of admission a barplot is created using ggplot and geom_bar functions, fill parameter is defined as TOEFL score to visually show the change in the scores. Using labs and ggtitle functions labels are defined and the plot obtained is shown below in Figure 3 which is further analyzed.

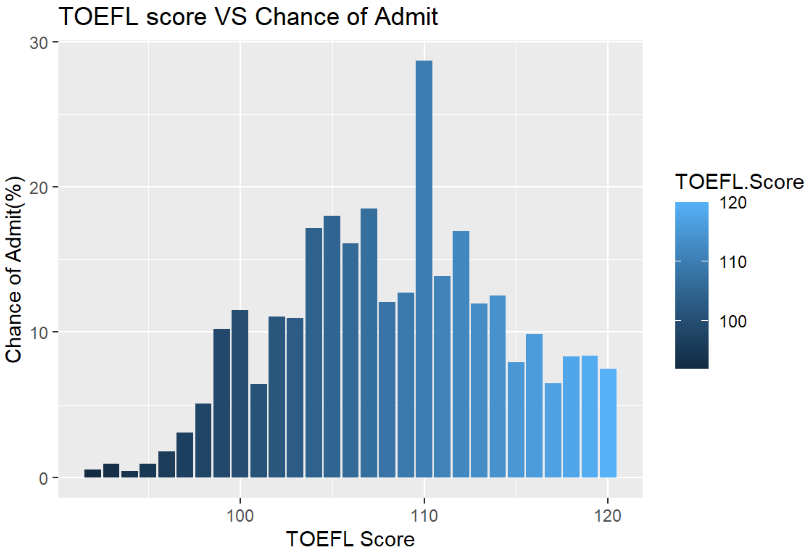


Figure 3

From the above bar plot, we can conclude that an average TOEFL score of 105-110 I observed to have maximum number of admits, it proves that TOEFL score is a significant parameter. But having the maximum of TOEFL score doesn’t confirm admission in most of the cases where other variables are considered in the admission process.

To show the relation between “University rating” and “Chance of Admission” two plots are created, the first plot initially gives us the information of how the Universities are rated on a scale of 5 and the count of universities for different rating, for this we used a histogram to represent the data using geom_histogram function. The histogram obtained is shown in figure 4.

For the relation of University rating with Chance of admit a bar plot is created using geom_bar function, below is the bar plot in figure 5.

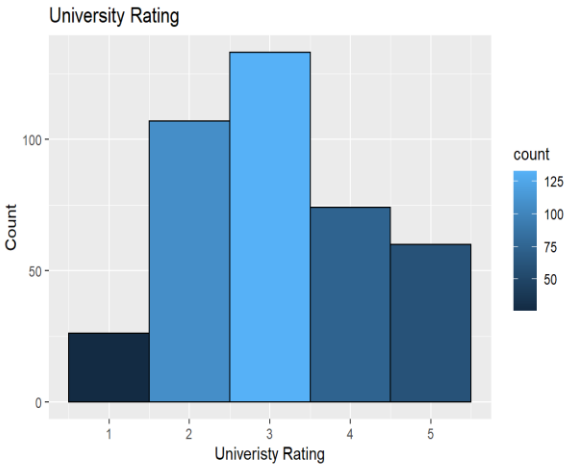


Figure 4

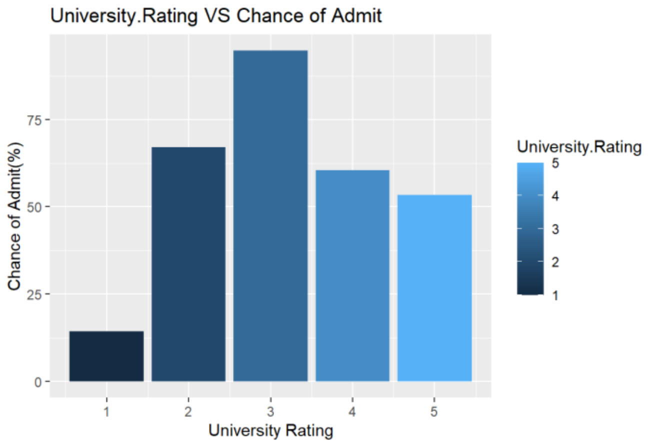


Figure 5

From figure 4 we can conclude that maximum number of universities are rated as 3.

From figure 5 we can confirm that as the Universities with a rating of 3 have more than 75% chance of getting admitted.

For the relation between research and chance of admit two plots were created one is to show the count of research experience and one to show the relation with chance of admission, first plot is histogram as shown in figure 6 ad in figure 7 a box plot is created to show the relation between research experience and chance of admission.

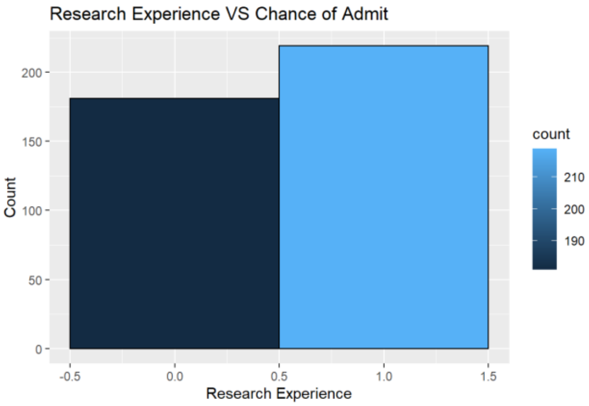


Figure 6

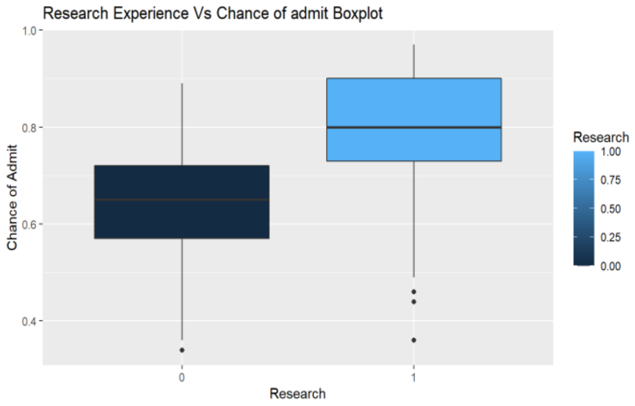


Figure 7

From figure 6 we can confirm that the candidates with research experience are applying more to the graduate programs.
From figure 7 we can conclude that the chance of admission increases for candidates with research experience.

To show the relation between LOR and Chance of admission a bar plot is created as shown in figure 8.

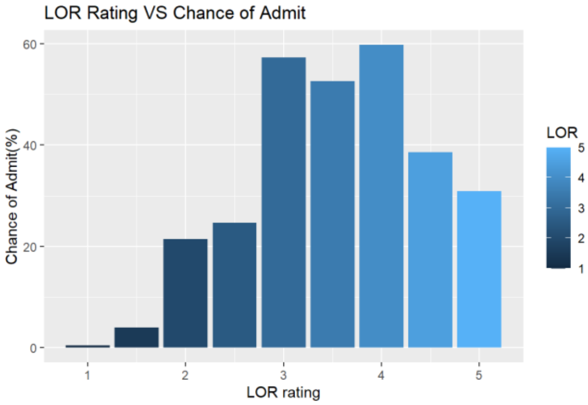


Figure 8

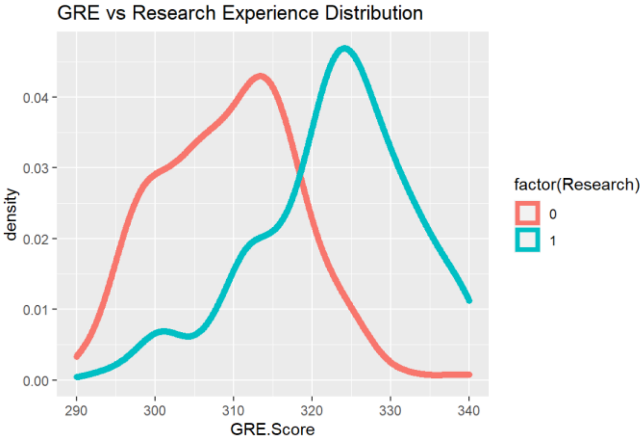


Figure 9

From Figure 8 we can confirm that the average rating of a LOR should be between 3-4 for the best chance of admission.
In figure 9 we have also plotted a density plot for GRE score and Research Experience which shows us that candidates with research experience achieved more GRE scores compared to the one's without any experience.

Correlation plot: The correlation plot shows the correlation between the two or more variables. A correlation plot (Figure 9) is created to show the relation between the chance of admit and all the other variables.

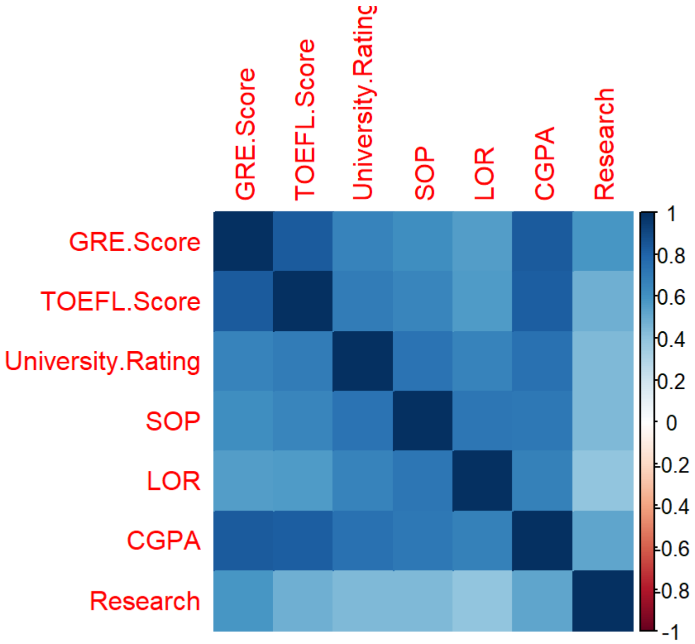


Figure 10

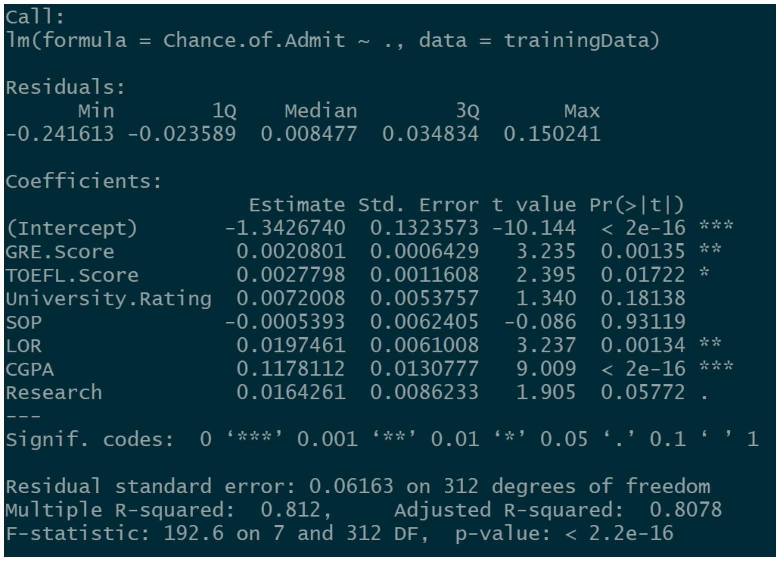
Regression Models and Predictive Analysis

Some of the models chosen for further analysis are multiple linear regression, random forest, and logistic regression.

Multiple Linear Regression Model

A multiple linear regression model is used to predict the value of a dependent variable using two or more independent variables. In this case, the Chance of admission is the dependent variable and the GRE score, TOEFL score, LOR, SOP, etc. are the independent variables. According to the summary of the multiple linear model, the intercept is -1.3426740. The residual standard error was found to be 0.06163 on 312 degrees of freedom which is very good for a regression model.

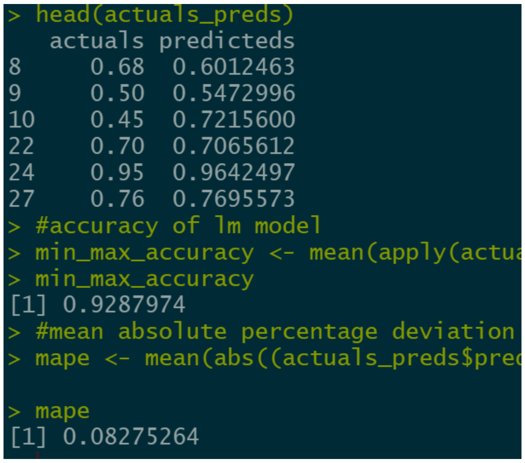
The CGPA, LOR, and research seem to affect the multiple linear regression model the most.



```
Call:
lm(formula = Chance.of.Admit ~ ., data = trainingData)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.241613 -0.023589  0.008477  0.034834  0.150241 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -1.3426740  0.1323573 -10.144  < 2e-16 ***
GRE.Score          0.0020801  0.0006429   3.235  0.00135 ** 
TOEFL.Score        0.0027798  0.0011608   2.395  0.01722 *  
University.Rating  0.0072008  0.0053757   1.340  0.18138    
SOP               -0.0005393  0.0062405  -0.086  0.93119    
LOR                0.0197461  0.0061008   3.237  0.00134 ** 
CGPA               0.1178112  0.0130777   9.009  < 2e-16 ***
Research           0.0164261  0.0086233   1.905  0.05772 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06163 on 312 degrees of freedom
Multiple R-squared:  0.812,	Adjusted R-squared:  0.8078 
F-statistic: 192.6 on 7 and 312 DF,  p-value: < 2.2e-16
```

Figure 11

After splitting the data into testing and training, Using the predict function we have calculated some predicted observations. A min-max accuracy for the model is calculated and found to be 92.87%. A mean absolute deviation was also calculated which was found to be 8.27%.



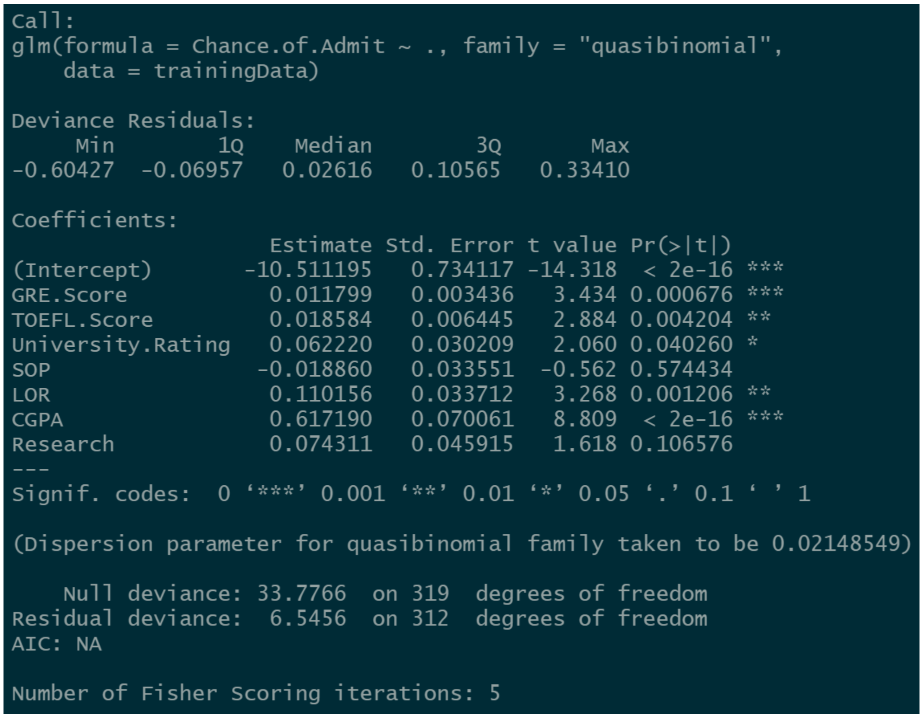
```
> head(actuals_preds)
   actuals predicteds
8     0.68  0.6012463
9     0.50  0.5472996
10    0.45  0.7215600
22    0.70  0.7065612
24    0.95  0.9642497
27    0.76  0.7695573
> #accuracy of lm model
> min_max_accuracy <- mean(apply(actua
> min_max_accuracy
[1] 0.9287974
> #mean absolute percentage deviation
> mape <- mean(abs((actuals_preds$pred

> mape
[1] 0.08275264
```

Figure 12

Logistic Regression Model

The logistic regression model is used to understand the relationship between independent and dependent variables by estimating probabilities using a logistic regression equation. The family used is quasibinomial to avoid non-numeric warnings. The intercept was found to be -10.511195 and the residual deviance was 6.5456 on 312 degrees of freedom.

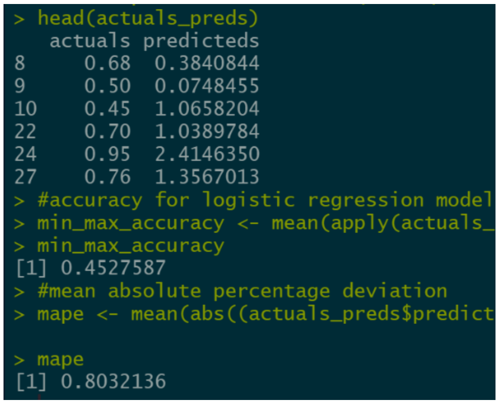
The CGPA and LOR seem to affect the chance of admission most in the logistic regression model.



```
Call:
glm(formula = Chance.of.Admit ~ ., family = "quasibinomial",
    data = trainingData)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.60427  -0.06957   0.02616   0.10565   0.33410

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -10.511195   0.734117 -14.318  < 2e-16 ***
GRE.Score           0.011799   0.003436   3.434 0.000676 ***
TOEFL.Score         0.018584   0.006445   2.884 0.004204 **
University.Rating   0.062220   0.030209   2.060 0.040260 *
SOP                -0.018860   0.033551  -0.562 0.574434
LOR                 0.110156   0.033712   3.268 0.001206 **
CGPA                0.617190   0.070061   8.809  < 2e-16 ***
Research            0.074311   0.045915   1.618 0.106576
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 0.02148549)

    Null deviance: 33.7766  on 319  degrees of freedom
Residual deviance:  6.5456  on 312  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Figure 13

Using the predict function some values were calculated and compared to the actual values. The min-max accuracy was found to be 45.27% and the mean absolute percentage deviation was 80.32%. The multiple linear regression model seems to be better than the logistic regression model.

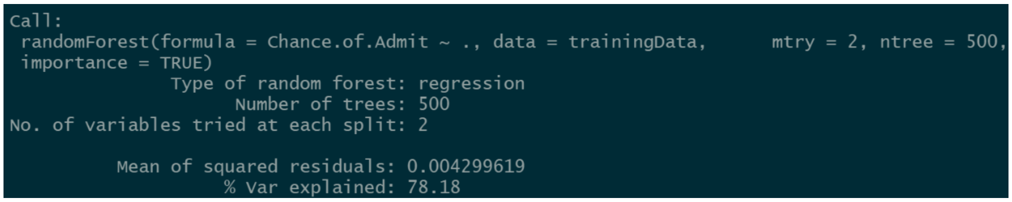


```
> head(actuals_preds)
   actuals predicteds
8     0.68  0.3840844
9     0.50  0.0748455
10    0.45  1.0658204
22    0.70  1.0389784
24    0.95  2.4146350
27    0.76  1.3567013
> #accuracy for logistic regression model
> min_max_accuracy <- mean(apply(actuals_
> min_max_accuracy
[1] 0.4527587
> #mean absolute percentage deviation
> mape <- mean(abs((actuals_preds$predict

> mape
[1] 0.8032136
```

Figure 14

Random Forest Model

The random forest is a supervised machine learning algorithm that is best for classification and regression problems. It is most useful when calculating the variable importance of the independent variables which gives us further insight into the dependency. A random forest model was calculated for the chance of admission using 500 decision trees and 2 variables were tried at each split. The mean squared residual error was found to be 0.00429.



```
Call:
 randomForest(formula = Chance.of.Admit ~ ., data = trainingData,      mtry = 2, ntree = 500,
 importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 0.004299619
                    % Var explained: 78.18
```

Figure 15

An importance plot was also calculated and CGPA, GRE Score, and TOEFL Score were found to be the most important variables which affect the chance of admission.

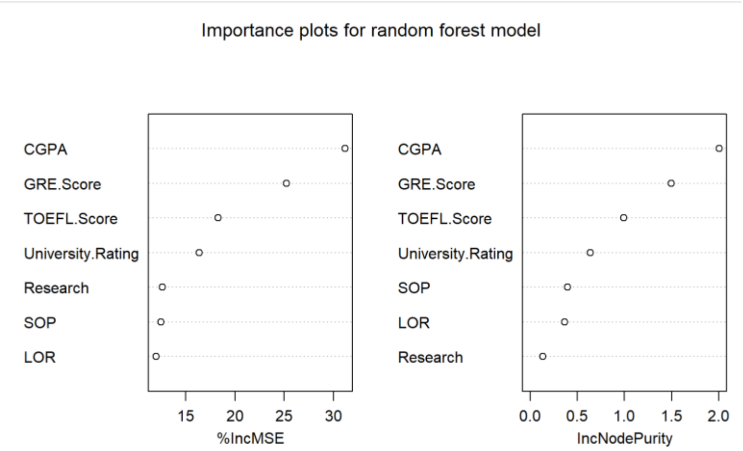
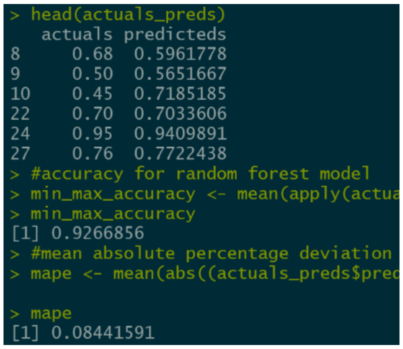


Figure 16

The predict function was used and some predicted values were calculated and compared with the actual values. The min-max accuracy was found to be 92.65% and the mean absolute percentage of deviation was 8.48%. The random forest model seems to have a similar accuracy like the multiple linear regression model. There is only a 0.2% difference in the accuracy for the two models.



```
> head(actuals_preds)
   actuals predicteds
8     0.68  0.5961778
9     0.50  0.5651667
10    0.45  0.7185185
22    0.70  0.7033606
24    0.95  0.9409891
27    0.76  0.7722438
> #accuracy for random forest model
> min_max_accuracy <- mean(apply(actua
> min_max_accuracy
[1] 0.9266856
> #mean absolute percentage deviation
> mape <- mean(abs((actuals_preds$pred

> mape
[1] 0.08441591
```

Figure 17

Conclusion

Based on the regression models we now know that multiple linear regression model is the best model. It has an accuracy of 92.87% and an error of 8.27% which is good. We now know that CGPA most affects the chance of admission from the three regression models.

References:

[1] Dataset: Mohan S Acharya, Asfia Armaan, Aneeta S Antony : A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019 https://www.kaggle.com/code/filibertogh/starter-graduate-admissions-1a5b7963-3

[2] Tidyverse: Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). “Welcome to the tidyverse.” *Journal of Open Source Software*, **4**(43), 1686. doi: 10.21105/joss.01686.

[3] gplot2 Package: Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

[4] Tidyverse: Wickham H, François R, Henry L, Müller K (2022). dplyr: A Grammar of Data Manipulation. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

[5] Tidyr: Wickham H, Girlich M (2022). tidyr: Tidy Messy Data. https://tidyr.tidyverse.org, https://github.com/tidyverse/tidyr.

[6] Random Forest: Liaw A, Wiener M (2002). “Classification and Regression by randomForest.” R News, 2(3), 18-22. https://CRAN.R-project.org/doc/Rnews/.

[7] Corrplot: Wei T, Simko V (2021). R package 'corrplot': Visualization of a Correlation Matrix. (Version 0.92), https://github.com/taiyun/corrplot.

[8] *Correlation Analyses in R - Easy Guides - Wiki - STHDA*. (2020, October 4). STHDA. http://www.sthda.com/english/wiki/correlation-analyses-in-r