Chance of master’s degree Admission Project

Discover if you are ready to be accepted for master’s degree.

Proposed By:

Bashnona Gamal

Antony Samir

Mina Alphons

Kholod Abdelnasr

Youssef Tarek

Andrew Amgad

Supervisor

Dr. Sherine Rady

TA. Mohamed Ashraf

Table of Contents:

1 Introduction

1.1 Project Overview

Chance of master's degree admission project is a solution for master's degree students who are willing to apply in science related fields in the United States of America. In American Universities, there are a lot of different criteria that are being taken into consideration to assess the applied students. These criteria include TOEFL score, GRE Test Score, recommendation letters and much more factors. We are helping master's degree students to predict how likely they were to be accepted according to their scores in different criteria. Therefore, they can enhance themselves and get better scores according to the most important criterion.

1.2 Problem Definition

There are two main problems that we aim to provide solution for in this project. The first problem is that students who are willing to apply for master's degree do not always know what are good scores that qualifies them to be accepted in the United States Universities. The second problem is that students do not know which scores are more important than others therefore, they do not know what score they must invest time enhancing them.

1.3 Objectives

The main objective of the project is to help students to be capable of knowing how likely they will be accepted based on their scores therefore, they consider enhancing their scores. For example, retaking a TOEFL or GRE Exam.

The project aims to identify all the relationships between the scores and chance of being accepted in the United State Universities as a result, student can invest time to enhancing specific scores according to which is more important. We also want to help undergraduate students who are willing to apply for master's program in the US to concentrate on specific factors to better utilize their time.

1.4 Dataset Description

We are using a dataset from Kaggle.com website which contains 500 Indian students records. Each record contains 8 variables as stated in the table below. The records contain 500 students GRE score, TOEFL score, University Rating, SOP score, LOR score, CGPA score, and research experience. The following table describes each column, its description, and the possible values of change of being accepted according to the dataset owner description.

Column Name	Column Description	Possible Values
GRE Test Score **(Graduate Record Examination)**	GRE test measures student's verbal, and quantitative reasoning. It also measures critical thinking and analytical writing.	Out of 340
TOEFL Score **(Test of English as a Foreign Language)**	TOEFL test measures the English language ability of non-native English speakers who are willing to apply for English Universities	Out of 120
University Rating	University Ranking for the student.	Out of 5
SOP Score **(Statement of Purpose)**	SOP is long essay issued by the university which states the introduction about the applicant and the purpose of the application.	Out of 5
LOR **(Letter of Recommendation)**	LOR is written by someone who can recommend the student according to previous experience and performance.	Out of 5
CGPA **(Cumulative Grade point Average)**	CGPA measures the performance of applicant throughout the academic courses and semesters.	Out of 10
Research Experience	Indicates if the applicant knows how to make research or not.	Either 0 or 1
Chance of Admit	Indicates how likely student is to be accepted according to all the previous attributes.	Ranging from 0 to 1

2 Observation

2.1 Data Visualization

Visualizing the data gives more insights about it and helps understanding it. In this section we are visualizing all the dataset with its columns to get familiar with the data and extract information based on visualization techniques.

2.2 Correlation Matrix

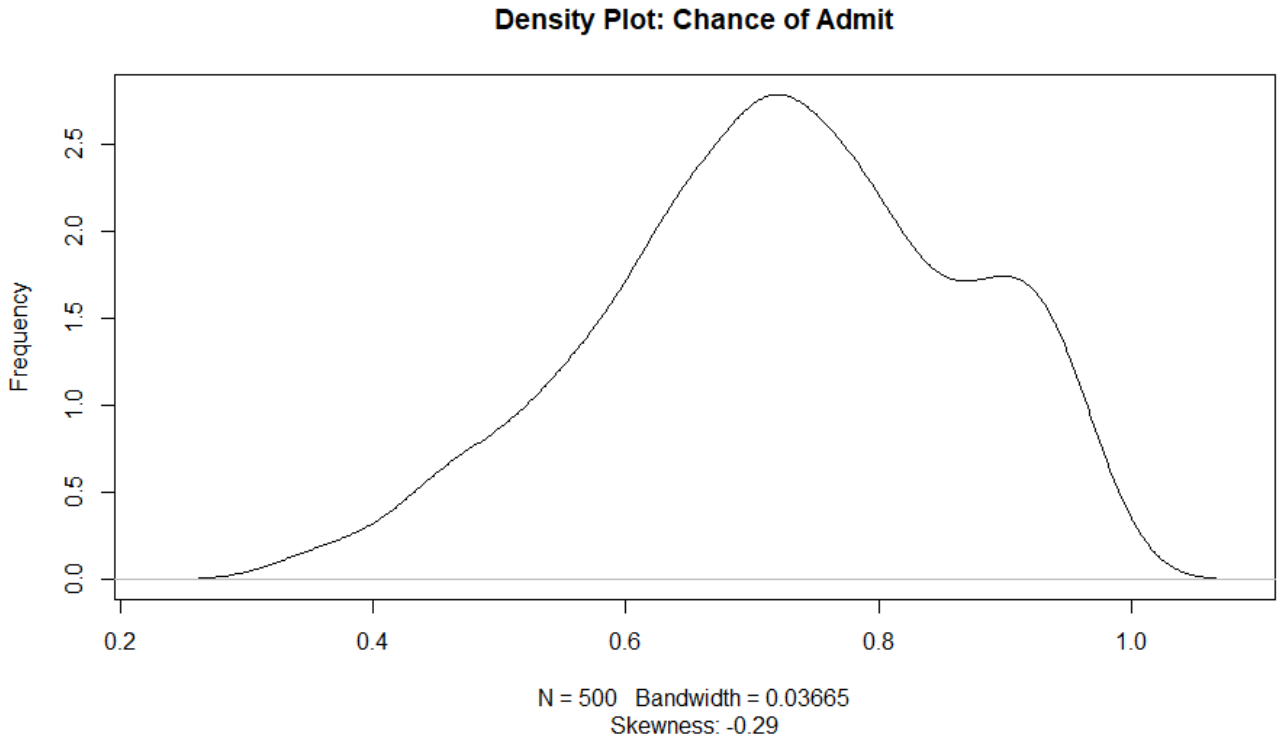
Correlation shows the relationship strength between two continuous variables. It shows the linear dependency between the two variables through calculating the correlation coefficient. Correlation matrix is applied for the dataset to study the relationship between all the variables. The below table describes students' chance of acceptance dataset.

	GRE	TOEFL	UR	SOP	LOR	CGPA	Research	COA
GRE	1.0	0.8	0.6	0.6	0.5	0.8	0.6	0.8
TOEFL	0.8	1.0	0.6	0.6	0.5	0.8	0.5	0.8
UR	0.6	0.6	1.0	0.7	0.6	0.7	0.4	0.7
SOP	0.6	0.6	0.7	1.0	0.7	0.7	0.4	0.7
LOR	0.5	0.5	0.6	0.7	1.0	0.6	0.4	0.6
CGPA	0.8	0.8	0.7	0.7	0.6	1.0	0.5	**0.9**
Research	0.6	0.5	0.4	0.4	0.4	0.5	1.0	0.5
COA	0.8	0.8	0.7	0.7	0.6	**0.9**	0.5	1.0

As Shown in the correlation matrix table, there is a very strong relation between the chance of being accepted in USA University and cumulative GPA of the student, which means the higher the student GPA the better chance of acceptance. In addition, it shows that the Research skills is the least important. GRE and TOEFL share the same relation with the chance of acceptance, they are equal in importance.

2.3 Data Population

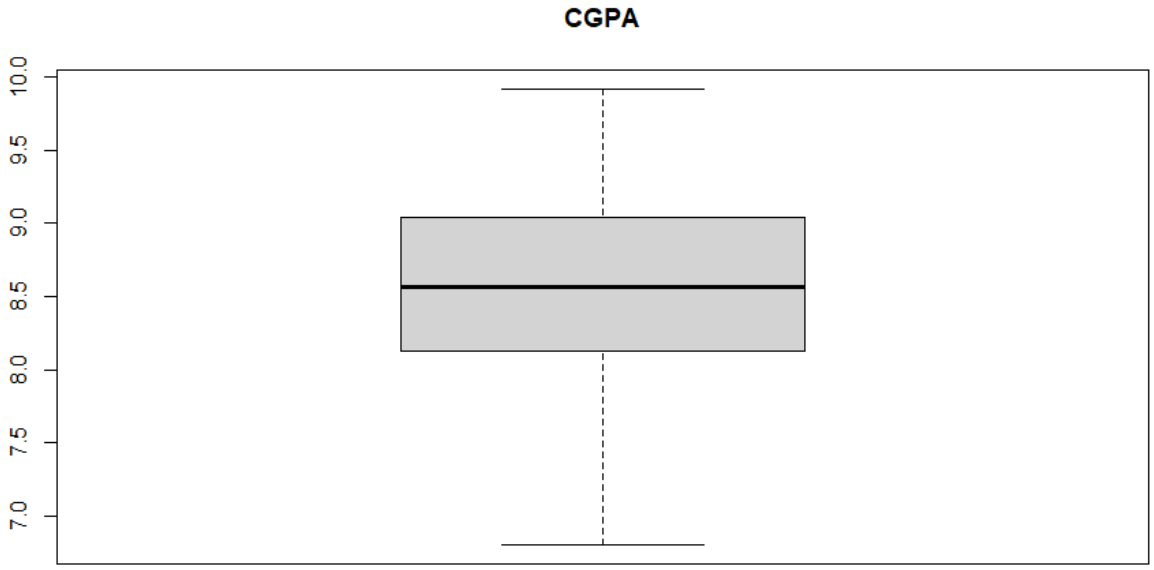
Data population help us to determine if the chance of admit is close to normal distribution or not. A normally distributed variable is a very powerful key to make sure that there are no skews in the database and good predictions are made. As a result, we will have reliable model that we can depend on in decisions. The following diagram is the density plot for chance of admit column in the dataset.



The density plot shows that chance of acceptance is close to normal distribution with 0.29 left skew. This means that the curve is very close to normal. We also noticed that this density plot is bimodal as it has two peaks.

2.4 Box Plot Diagram

Box plot is a very important diagram which visualize the median, first quartile, third quartile, minimum, maximum, and the most important attribute: the outliers. The following diagram is the boxplot for the cumulative GPA column. It is plotted specifically with CGPA column since this column has higher correlation with the chance of admit column as stated in the correlation section. Therefore, outliers in this column must be removed.

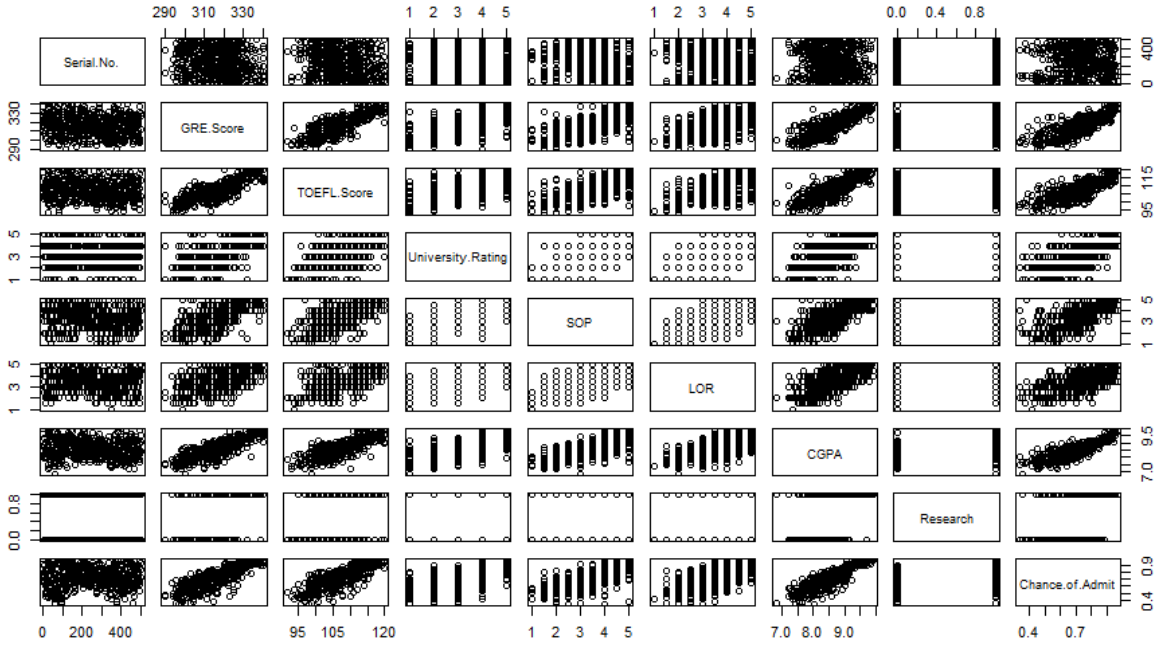


Outlier rows:

As Shown in the previous boxplot diagram, CGPA column has no outliers which will not affect the result with false positives. The distribution of the column is close to normal with median equal to 8.560 and minimum of 6.800 and maximum of 9.920. This attribute is very important affecting the result because of its direct relationship with the chance of admit column.

2.5 Variables Relationships

We study the dataset through all the relationships between its variables. By identifying the relationship, we gain insights, better decision making, and know the linear dependency between its variables. The following diagram shows all the relationships between all variables. This diagram is more likely to be the big picture focusing on the importance of relationships and by drawing them separately helped us studying the connection between attributes.

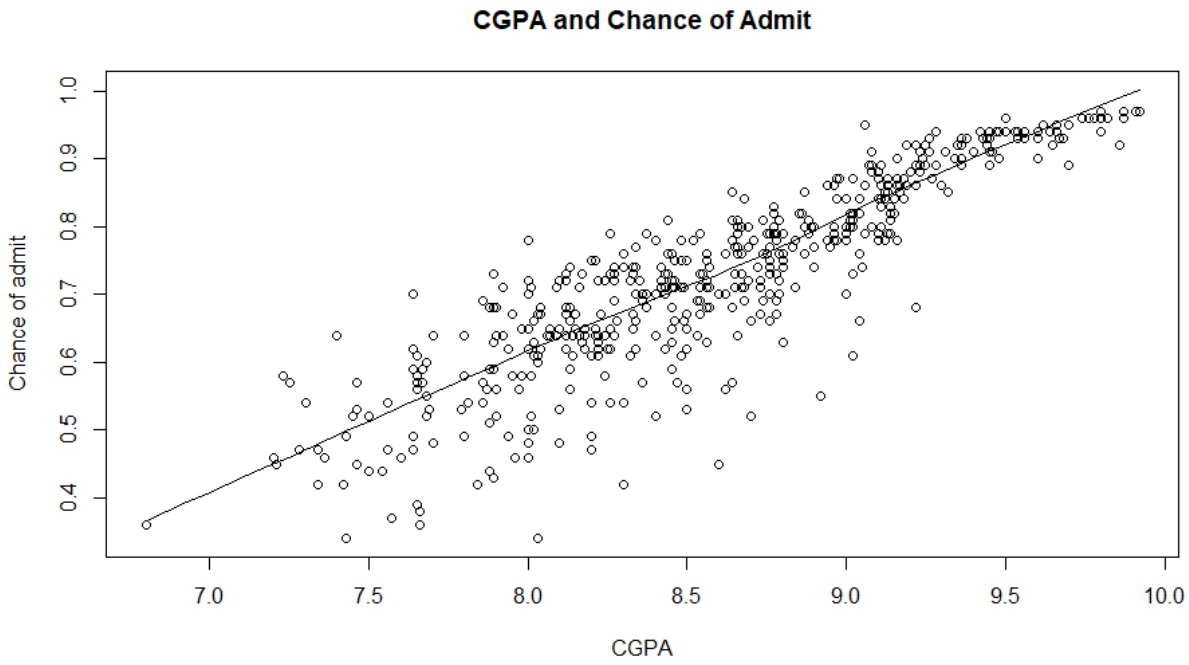


In variables relationships diagram, it shows each relationship between all columns, we get a grasp on how the data related to each other. Obviously, there are linear relation between some columns. For instance, TOEFL score and GRE score are linearly dependent on each other, this means that students who get high GRE score have a good chance to have good TOEFL scores. In addition, Students with higher CGPA has very good TOEFL and GRE scores. On the other hand, CGPA and research skills are linearly independent.

2.6 Scattered Plot

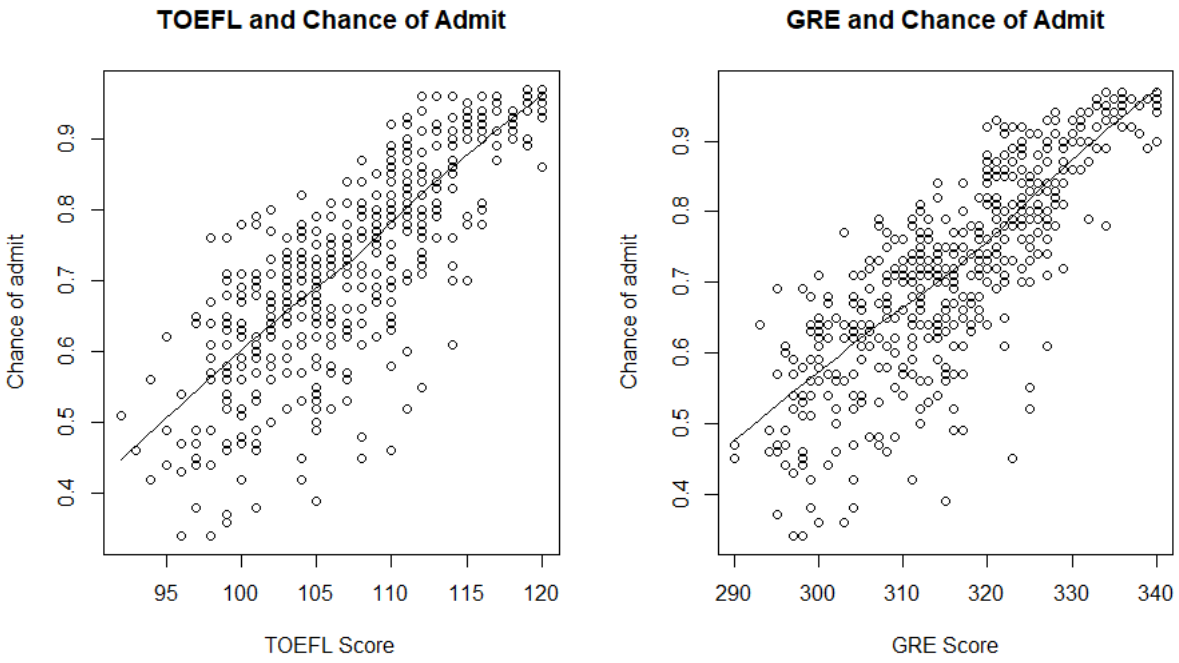
Scattered Plot visualize the relationship between two variables. It is very useful to identify whether the variables related to each other by any mean or not. Using the scattered plot, we know how the predicator variable is important to the response variable. In the previous subsection, all relationships have been determined and visualized. In this section we will discuss some of the important relationships with more details.

CGPA and Chance of admit



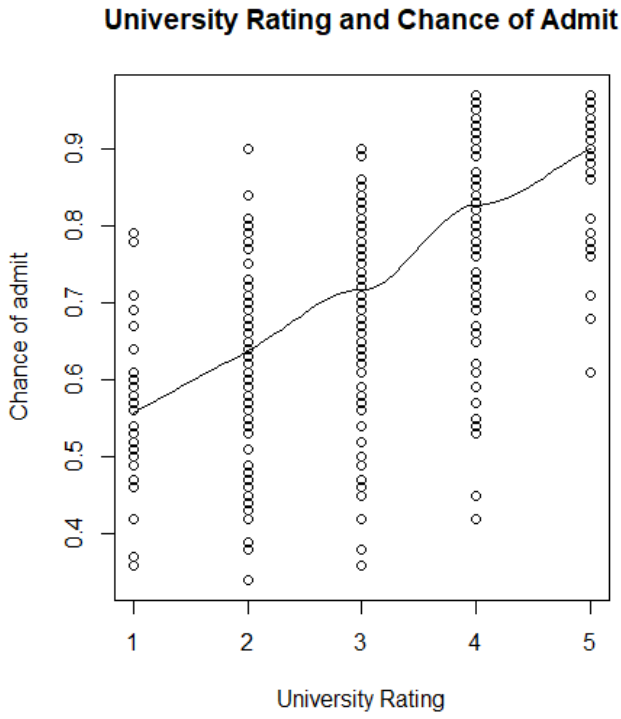
This diagram shows that there is a strong linear relationship between the cumulative GPA and the chance of being accepted. Therefore, the higher the student CGPA the better for his chance. CGPA is the main important score for students, and it has highest correlation with the chance of being accepted.

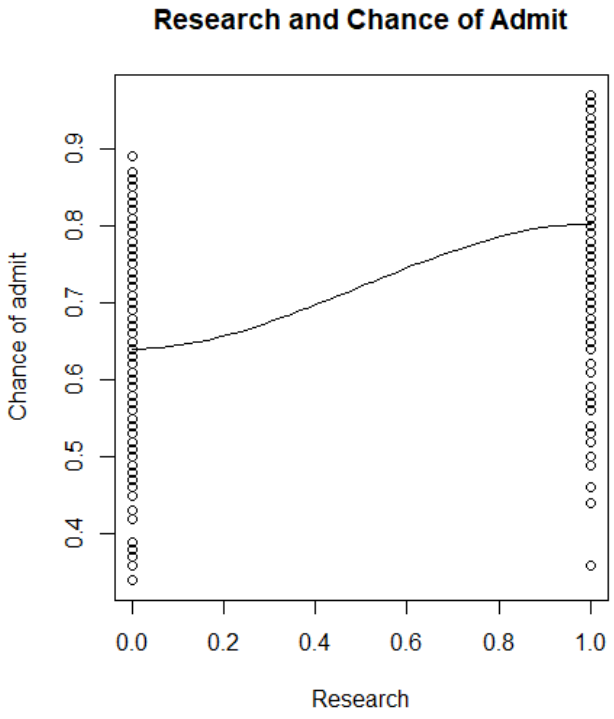
TOEFL and GRE with respect to chance of admit



TOEFL and GRE have a linear relationship with the chance of admit. The plot also shows that the higher the TOEFL and GRE the better chance for being accepted. However, the range of chance of being accepted is wide. For example. The chance of students who got in TOEFL 105 ranges from less than 0.4 and 0.8. This means that the dependency is not very strong, but these ranges is decreased while achieving higher TOEFL and GRE scores.

University Rating and Research skills with respect to chance of admit

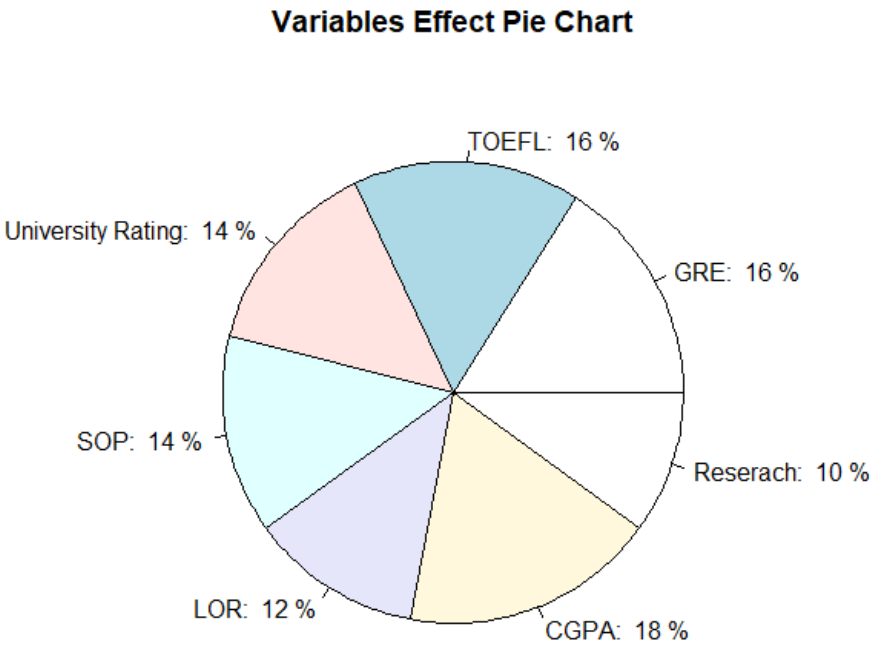




University rating and Research skills have very poor linear dependency with the chance of being accepted in USA universities. This means that whether you have research skills or not, your chance varies from 0 to 0.9 which is approximately the normal range. These two variables do not add too much to the student as they have very weak dependency with the response variable that will be predicted.

2.7 Variables Effect Pie Chart

Pie Chart visualizes how each variable is important to the chance of being accepted. We noticed that CGPA has the highest percentage which means it is the most important factor. GRE and TOEFL come in the second stage with 16% importance. After that, university rating and statement of purpose shares 14% among other variables. At the end, the research skills and letter of recommendations have the least priority affecting the chance of being accepted.



3 Data Cleaning

Chance of master's degree admission dataset is already cleaned. The columns values have already been clear and clean. All the ranges are in the same ranges stated by the dataset owner in Kaggle website. A check has been made for outliers in the important columns, the columns with high linear dependency with the chance of admit column, and it has no outliers. This is done by checking for the boxplot for each and determined by the equation Q3 + 1.5 * IQR for the maximum and Q1 – 1.5 * IQR for the minimum.

4 Dataset Preparation

The chance of master's degree admission dataset has 500 rows with 8 variables. The number of rows is not too much however, 500 records are a good number to start building the model and predict reliable results. Therefore, the 500 rows have been divided to 80% (400 rows) for training the model and 20% (100 rows) to test the model.

5 Data Analytics Technique

5.1 Multiple Linear Regression

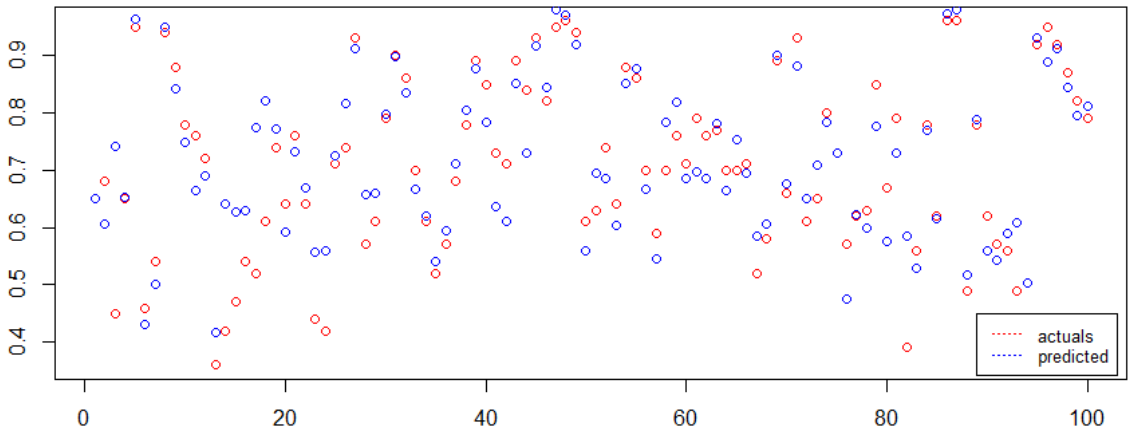
Multiple Linear Regression is a data analytics technique which uses more than one variable called predictors to predict a specific variable called response variable. It uses the below function for predictions:

$$y_0 = B_0 + B_1 x_1 + B_2 x_2 + \cdots + B_n x_n$$

Where n is the number of variables we have. In master's degree database we have 8 variables that means n = 8. By using this technique, we can predict the chance of being accepted in USA Universities using all the other variables. Multiple regression technique is used in this project because the dataset has continuous values, and because of having multiple variables that affect the chance of being accepted. The following equation contains all the coefficients for all the variables including the intercept value.

$$\begin{aligned} y = {} & -1.4067512 + 0.0023085(GRE) + 0.0030805(TOEFL) + 0.0069311(UR) \\ & - 0.0007643(SOP) + 0.0116529(LOR) + 0.1164724(CGPA) \\ & + 0.0199933(Research) \end{aligned}$$

5.2 Multiple Linear Regression Visualization



The previous diagram shows both the actual values of the 100-test case and their predictions using the model we developed. As stated in the legend, the red color dots used for the actuals and the blue for the predicted values. If the 2 vertical points are close t each other, it indicates that the predicted value is too close to the actual value.

6 Model Performance

The model is tested using two accuracy methods which are correlation accuracy and min-max-accuracy. The correlation accuracy is trying to find the relationship between two variables, it identifies the linear relationship between them. It ranges from 0 to 1.0. Obviously, the higher the correlation the better accuracy value. Furthermore, the min/max accuracy is also calculated, it simply gets the mean of minimum of the actuals and predictions over the maximum of the actuals and predictions. Below is the formula that calculates the min-max-accuracy:

$$Accuracy = mean(\frac{\min(actuals, predictions)}{\max(actuals, predictions)})$$

By using both methods and testing our model we got the accuracy stated in the below table

	Accuracy
Correlation Accuracy	87.83141%
Min-max-accuracy	92.50631%

7 Conclusions

In this paper, we presented the chance of master's degree admission problem, and how to help students predict their case before going into the process that takes time. Our conclusion is that the most important factor to have high chance of acceptance is to have a high CGPA because it has high correlation with the chance of being accepted in the US universities. TOEFL and GRE are important too where they come in the second stage after the CGPA. Furthermore, we concluded that having research skills and good university rating do not affect student chance comparing with other factors. Multiple linear regression model is developed to predict the chance of being accepted using all attributes. The performance of the model is calculated in two ways which are the correlation accuracy and min-max-accuracy. Finally, we are encouraging all students who are willing to apply for being a master's student, to concentrate their CGPA during the bachelor. Our studies also shows that students who get high CGPA get good scores in GRE and TOEFL.

8 References

[1] How multiple linear regression *works*. (n.d.). Investopedia. https://www.investopedia.com/terms/m/mlr.asp

[2] How to give color to each class in scatter plot in R? (n.d.). Stack Overflow. https://stackoverflow.com/questions/7466023/how-to-give-color-to-each-class-in-scatter-plot-in-r

[3] Scatter plots - R base graphs - Easy guides - Wiki - STHDA. (n.d.). STHDA - Acc*ueil. https://www.sthda.com/english/wiki/scatter-plots-r-base-graphs*

[4] *Quick-R: Combining plots*. (n.d.). Quick-R: https://www.statmethods.net/advgraphs/layout.html

[5] Linear regression with R. (n.d.). https://r-statistics.co/Linear-Regression.html

[6] *Linear regression - A complete introduction in R with examples*. (2019, October 22). ML+. https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/

[7] *Graduate admission 2*. (n.d.). Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/mohansacharya/graduate-admissions