A

Project Report

On

Graduate Admission Prediction

Using

Machine Learning

By

Ms. Vaibhavi Ganesh Rao

Ms. Bhagyashree Rajendra Atre

Mr. Narendra Singh Rathore

Guided by

Mr. Pranav Jaipurkar

Fergusson College Pune

[2020-21]

ACKNOWLEDGEMENT

Any accomplishment requires efforts of many people and this project is no different. I am highly obliged and grateful to all those who helped me in the development of this project and provided all encouragement that has helped me in this project work.

It is my pleasure to present the project report entitled **“Graduate Admission Prediction Using Machine Learning”** which has been a through join efforts. Since the time we have started working on this project, we have thoroughly been engrossed with technological exploration of field.

I take this opportunity to thank my respected project Guide **Mr. Pranav Jaipurkar** without whose support, help and encouragement, this endeavor of ours wouldn’t be materialized. Finally, we thank to all teaching and non-teaching staff that associated directly/indirectly in the execution of this project work.

LIST OF FIGURES

LIST OF GRAPHS

LIST OF TABLES

TABLE OF CONTENTS

ABSTRACT

Student admission problem is very important in educational institutions. In today's education world there are many numbers of students who want to pursue higher education after engineering or any graduate degree course. Higher education in the sense, some people want to do MTech through GATE or through any educational institute entrance examination and some people want to do MBA through CAT or through any respective educational institute entrance examination and some people want to do Masters in abroad universities. We are focusing on only the students who want to pursue their higher education in Indian universities. This project addresses machine learning models to predict the chance of a student to be admitted to a master's program. This will assist students to know in advance if they have a chance to get accepted. The machine learning models are multiple linear regression, random forest regression and multiple linear regression with PCA. This project should consider all the crucial factors which plays a vital role in student admission process and compare their performance to select the best performing model.

1 INTRODUCTION

In the world of competitions, the real challenges are faced by the students themselves. Time to time, they have their entrance tests and they are under pressure to get admission for their graduation. In this process of getting admission, the students take all kinds of risks. The biggest one of those risks is applying in institutions for graduate admission and waiting for them without applying in sufficient number of other institutions of higher or lower rating. Under such circumstances, if not selected then there is a huge wastage of time and resources. Also, if selected, but in a university of low rating in spite of the fact that the student deserves admission in comparatively high-rated institutes, poses a problem. Therefore, due to the lack of proper prognosticators, students opt for either very ambitious institutions or very low-ranked institutions. In order to encounter such onerous things, chance-estimate prediction of graduate admission comes into play.

In this project, we present a Machine Learning based approach where the data is trained on a range of values, from stellar profiles to mediocre ones. In this, the machine learning model is developed in which parameters necessary for the admission purpose like GRE Score, TOEFL Score, University Rating, Statement of Purpose and Letter of Recommendation Strength, Undergraduate GPA and Research Experience are taken into consideration. A sample profile is tested against all the three models defined earlier in order to understand the performance of each model. We aim to bring students closer to their university of choice through a robust evaluation of their profiles.

1 LITERATURE REVIEW

This section includes the literature review of previous research on the assessment of student enrolment opportunities in universities. A great number of researches and studies have been done on graduation admission datasets using different types of machine learning algorithms. One impressive work by Acharya et al. [1] has compared between 4 different regression algorithms, which are: Linear Regression, Support Vector Regression, Decision Trees and Random Forest, to predict the chance of admit based on the best model that showed the least MSE which was multilinear regression.

In addition, Chakrabarty et al. [2] compared between both linear regression and gradient boosting regression in predicting chance of admit; point out that gradient boosting regression showed better results.

Gupta et al. [3] developed a model that studies the graduate admission process in American universities using machine learning techniques. The purpose of this study was to guide students in finding the best educational institution to apply for. Five machine learning models were built in this paper including SVM (Linear Kernel), AdaBoost, and Logistic classifiers.

Waters and Miikkulainen [4] proposed a remarkable article that helps in ranking graduation admission application according to the level of acceptance and enhances the performance of reviewing applications using statistical machine learning. The main objective of the project was to develop a system that can help the admission committee of the university to take better and faster decisions. Logistic regression and SVM were used to create the model, both models performed equally well and the final system was developed using Logistic regression due to its simplicity. The time required by the admission committee to review the applications was reduced by 74% but human intervention was required to make the final decision on status if the application.

Sujay [5] applied linear regression to predict the chance of admitting graduate students in master's programs as a percentage. However, no more models were performed.

Bibodi et al. [6] used multiple machine learning models to create a system that would help the students to shortlist the universities suitable for them also a second model was created to help the colleges to decide on enrolment of the student. He composed two different predictive models regarding graduate admissions:

1. A statistical analytical model based on naive Bayes that filters (selects) out universities that are suitable for the students based on their marks and other biographical information.
2. A machine learning classification model powered by random forest, decision tree, naive Bayes, SVM-linear, and SVM-radial algorithms that can be deployed by universities for selecting the deserving students for their admission programs.

Ghai [7] developed an American graduate admission prediction model that enables students to choose the apt University by foretelling whether he/she will be admitted there or not.

Roa et al. [8] developed a College Admission Predictor System in the form of a Web application by taking the scores obtained by the candidate and his/her personal information as input, and the possible admissions in colleges are predicted as output.

2 METHODOLOGY

Cross-Industry Standard Process (CRISP) methodology (Azevedo, 2008) as shown in Figure 2-1 was followed in this project.

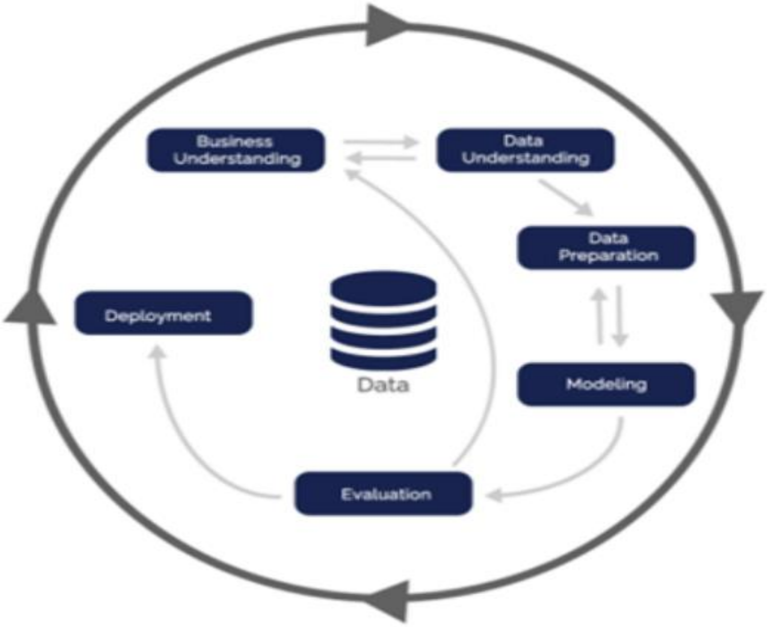


Figure 2-1 CRISP

- **Business Understanding**: Initially good amount of time was spent on understanding the problem statement by understanding the concerns of students regarding the current admission process, the objectives of the project were defined in this process.

- **Problem Understanding**: Initially first we have to spend some time on what are the problems or concerns students having during their pre admission period and we should set the solutions to those problems as objectives of this project.

- **Data Understanding**: Data required for the project was provided by our guide. Different features of the data were analysed based on their importance and relevance. Data-set would be explained in more detail.

- **Data Preparation**: Data should be cleaned that is removing the noise in the data and filling the missing values or extreme values and finalising the attributes/factors which will have crucial importance in student admission process.

- **Building Models**: Several ML models have to be developed using various machine learning algorithms for admission to a particular university and the user interface has to be developed to access those models.

- **Evaluation**: Developed models are evaluated according to their accuracy scores and performance.

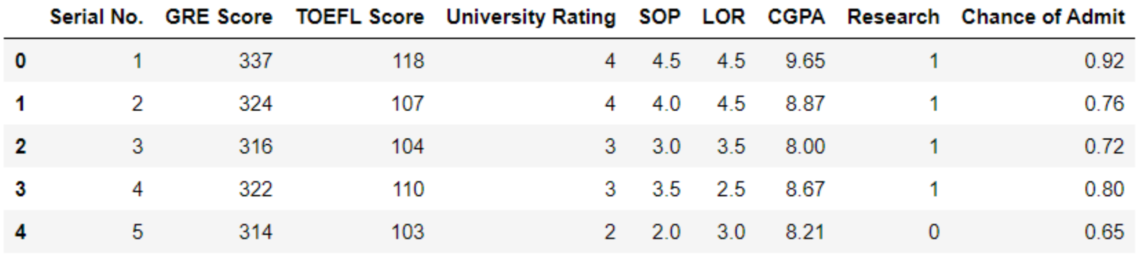
3 IMPLEMENTATION

3.1 DATASET

This section describes, in brief, the data that has been used for the project. The dataset presented in this project is related to educational domain. This dataset was provided by our guide as .csv file. Admission is a dataset with 500 rows that contains 7 different independent variables as shown in Table 3-1 which are:

- Graduate Record Exam (GRE) score. The score will be out of 340 points.
- Test of English as a Foreigner Language (TOEFL) score, which will be out of 120 points.
- University Rating (Uni.Rating) that indicates the Bachelor University ranking among the other universities. The score will be out of 5.
- Statement of purpose (SOP) which is a document written to show the candidate's life, ambitious and the motivations for the chosen degree/ university. The score will be out of 5 points.
- Letter of Recommendation Strength (LOR) which verifies the candidate professional experience, builds credibility, boosts confidence and ensures your competency. The score is out of 5 points
- Undergraduate GPA (CGPA) out of 10
- Research Experience that can support the application, such as publishing research papers in conferences, working as research assistant with university professor (either 0 or 1).

Table 3-1 First five entries of dataset with all parameters

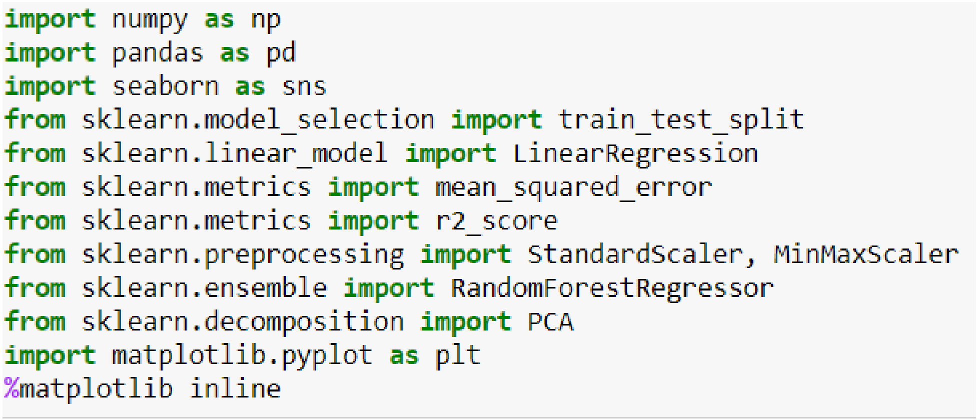


	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
0	1	337	118	4	4.5	4.5	9.65	1	0.92
1	2	324	107	4	4.0	4.5	8.87	1	0.76
2	3	316	104	3	3.0	3.5	8.00	1	0.72
3	4	322	110	3	3.5	2.5	8.67	1	0.80
4	5	314	103	2	2.0	3.0	8.21	0	0.65

A unique feature of this dataset is that it contains equal number of categorical and numerical features. The data has been collected and prepared typically from an Indian student’s perspective. However, it can also be used by other grading systems with minor modifications. In the dataset the GRE scores, TOEFL scores, university rating, statement of purpose strength, letter of recommendation strength, CGPA, research experience and chance of admit, being the target variable. So, there are six features that are continuous with only one feature, research experience, as categorical. One dependent variable can be predicted which is chance of admission, that is according to the input given will be ranging from 0 to 1.

3.2 LIBRARIES

Python's large standard library, commonly cited as one of its greatest strengths, provides tools suited to many tasks. For Internet-facing applications, many standard formats and protocols such as MIME and HTTP are supported. It includes modules for creating graphical user interfaces, connecting to relational databases, generating pseudorandom numbers, arithmetic with arbitrary-precision decimals, manipulating regular expressions, and unit testing. The libraries imported in this project are shown in the Figure 3-1 below.



```
import numpy as np
import pandas as pd
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from sklearn.ensemble import RandomForestRegressor
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
%matplotlib inline
```

Figure 3-1 Libraries imported in this project

- **NumPy**- It is fundamental for scientific computing with Python. It supports large, multidimensional arrays and matrices and includes an assortment of high-level mathematical functions to operate on these arrays. NumPy arrays are stored at one continuous place in memory unlike lists, so processes can access and manipulate them very efficiently. This behaviour is called locality of reference in computer science

- **Pandas**-It is built on top of NumPy, offers data structures and operations for manipulating numerical tables and time series. It is a tool data scientist will use again and again.

- **Matplotlib**-It is a 2D plotting library that can also generate data visualizations, such as histograms, power spectra, bar charts and scatter plots. Matplotlib produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms.

- **Seaborn-** It is a library that expands upon the functionality of Matplotlib, using enhanced graphics to make heatmaps and other dynamic visualizations.

- **Scikit-learn**-It is a machine learning library built on NumPy, SciPy and Matplotlib that implements classification, regression and clustering algorithms including support vector machines, logistic regression, Naive Bayes, random forests and gradient boosting.

4 DATA PRE-PROCESSING

4.1 DATA CLEANING

Data cleaning refers to detecting and correcting student records which have inaccurate or corrupt values resulting in “dirty” data. Sometimes this is caused simply by values being incorrectly manually entered by a worker, e.g., a GPA of 40.1 instead of 4.01. Such records must be verified and corrected, or when the data cannot be verified, the record can be dropped from the analysis. A different issue is when two sets of data have the same information but separate representations.

The data cleaning process has several key benefits to it:

1. This eliminates major errors and inconsistencies which are unavoidable when dragging multiple data sources into one dataset.

2. Having data cleaning software will make everyone more effective as they will be able to get easily from the data what they need.

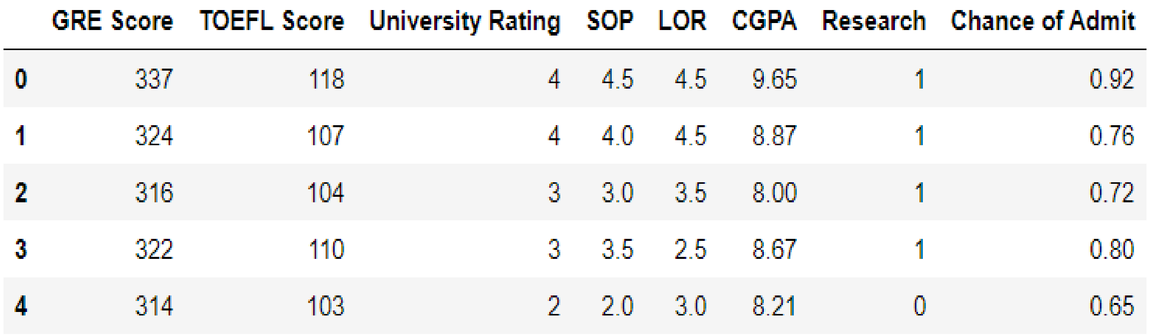
3. Fewer mistakes mean happy clients, and less unhappy workers.

4. The ability to chart the various functions, and what your data is supposed to do and where it comes from your data.

4.1.1 Removing the Unwanted Column

In this dataset we do not need the serial no. column so we drop that column as shown in the Table 4-1

Table 4-1 Dataset after removing the serial no. column



	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
0	337	118	4	4.5	4.5	9.65	1	0.92
1	324	107	4	4.0	4.5	8.87	1	0.76
2	316	104	3	3.0	3.5	8.00	1	0.72
3	322	110	3	3.5	2.5	8.67	1	0.80
4	314	103	2	2.0	3.0	8.21	0	0.65

4.1.2 Checking the Null Values

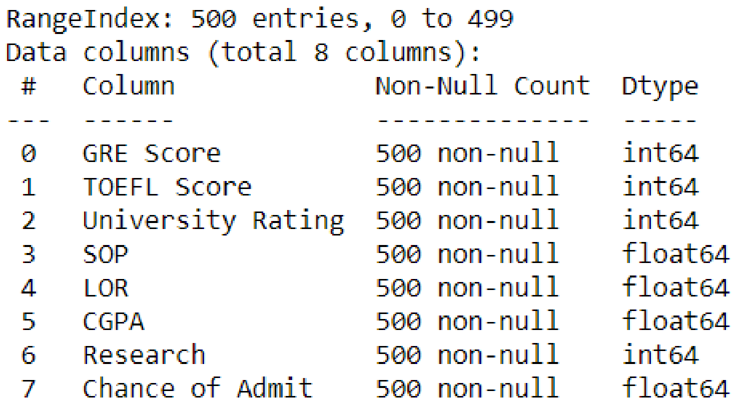
While making a Data Frame from a csv file, many blank columns are imported as null value into the Data Frame which later creates problems while operating that data frame. Pandas isnull() method are used to check and manage NULL values in a data frame.

Syntax: Pandas. Isnull(“Data Frame Name”) or DataFrame.isnull()

Parameters: Object to check null values for dataset

Return Type: Dastaframe of Boolean values which are True for NaN values

Table 4-2 Null values



```
RangeIndex: 500 entries, 0 to 499
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   GRE Score          500 non-null    int64
 1   TOEFL Score        500 non-null    int64
 2   University Rating  500 non-null    int64
 3   SOP                500 non-null    float64
 4   LOR                500 non-null    float64
 5   CGPA               500 non-null    float64
 6   Research           500 non-null    int64
 7   Chance of Admit    500 non-null    float64
```

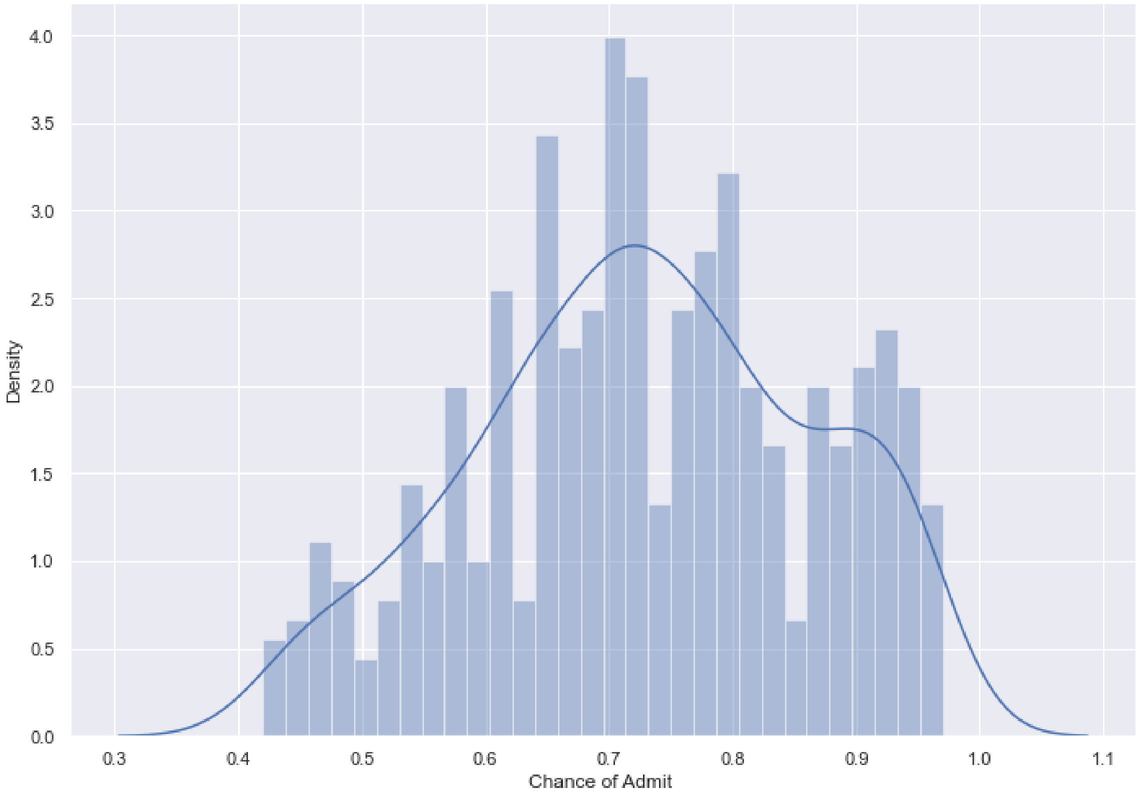
As we can see in **Error! Reference source not found.** the dataset does not consist of any NULL values, hence there are no missing values.

4.1.3 Outliers

Outliers are data values that differ greatly from the majority of a set of data. To find the outliers, there are many methods that can be used, such as: scatterplot, histogram and boxplot.

There are many different methods to deal with outliers, such as:

- Remove the case
- Assign the next value nearer to the median in place of the outlier value
- Calculate the mean of the remaining values without the outlier and assign that to the outlier case



Graph 4-1 Distplot of Chance of Admit

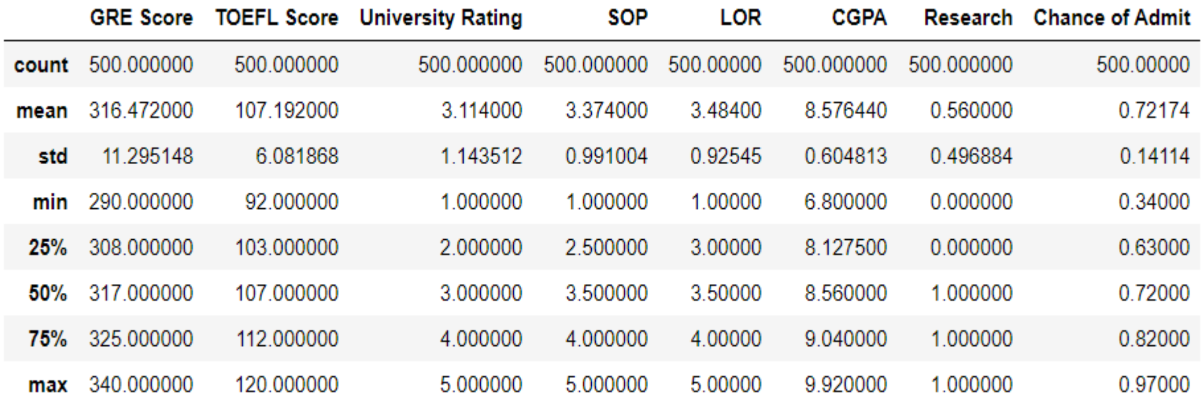
Outlier Detection and Removal Outliers are observations in a dataset that don't fit in some way. In given data set after plotting distplot we get to know that the data is normally distributed as shown in Graph 4-1.

There are no missing values and outliers because we analysed the data, so for this data there is no need to fill the missing values and deal with outliers. If there are any missing values and outliers we can fill (or) drop using the fillna method and drop method and we can also standardize the data using the min-max scaler, if necessary.

4.2 DESCRIPTIVE SUMMARY

Descriptive summary provides numerical measures of some important features which describe a dataset. Some features are presented in Table 4-3 which shows mean, median (std), minimum value(min), first quartile (25%), second quartile (50%), third quartile (75%) and maximum value(max).

Table 4-3 Descriptive summary

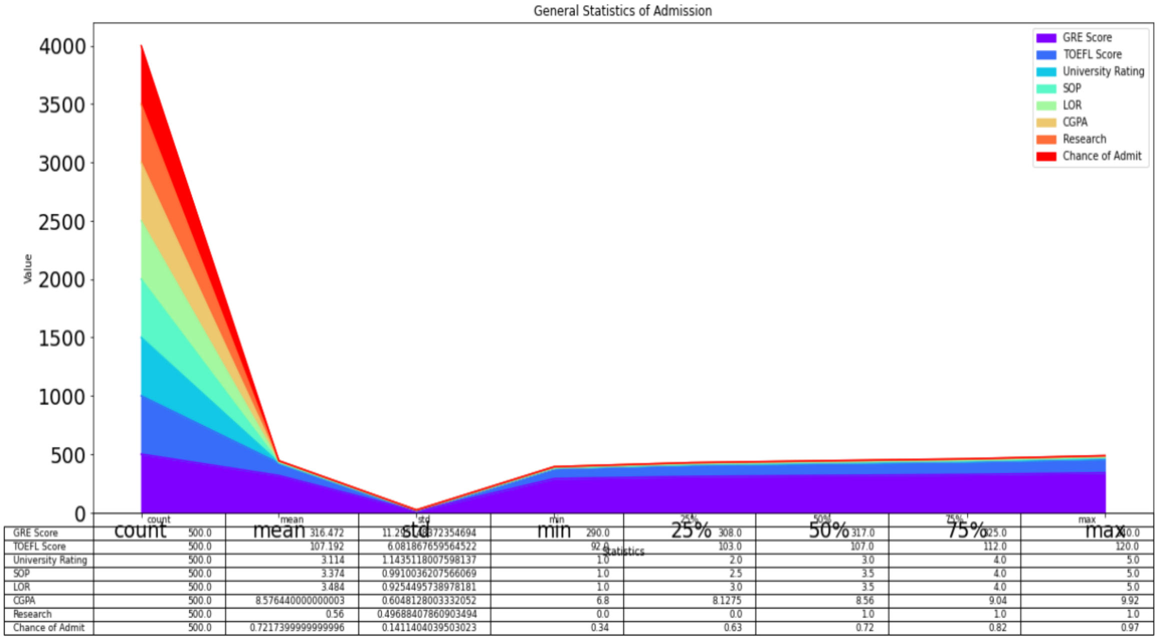


	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
count	500.000000	500.000000	500.000000	500.000000	500.00000	500.000000	500.000000	500.00000
mean	316.472000	107.192000	3.114000	3.374000	3.48400	8.576440	0.560000	0.72174
std	11.295148	6.081868	1.143512	0.991004	0.92545	0.604813	0.496884	0.14114
min	290.000000	92.000000	1.000000	1.000000	1.00000	6.800000	0.000000	0.34000
25%	308.000000	103.000000	2.000000	2.500000	3.00000	8.127500	0.000000	0.63000
50%	317.000000	107.000000	3.000000	3.500000	3.50000	8.560000	1.000000	0.72000
75%	325.000000	112.000000	4.000000	4.000000	4.00000	9.040000	1.000000	0.82000
max	340.000000	120.000000	5.000000	5.000000	5.00000	9.920000	1.000000	0.97000

Note that the first and third quartile are found using the following equations where N is equal number of data points.

First Quartile (Q1) = (N+1) × 0.25

Third Quartile (Q3) = (N+1) × 0.75



	count	mean	std	min	25%	50%	75%	max
GRE Score	500.0	316.472	11.295148372354694	290.0	308.0	317.0	325.0	340.0
TOEFL Score	500.0	107.192	6.081867659564522	92.0	103.0	107.0	112.0	120.0
University Rating	500.0	3.114	1.1435118007598137	1.0	2.0	3.0	4.0	5.0
SOP	500.0	3.374	0.9910036207566069	1.0	2.5	3.5	4.0	5.0
LOR	500.0	3.484	0.9254495738978181	1.0	3.0	3.5	4.0	5.0
CGPA	500.0	8.576440000000003	0.6048128003332052	6.8	8.1275	8.56	9.04	9.92
Research	500.0	0.56	0.49688407860903494	0.0	0.0	1.0	1.0	1.0
Chance of Admit	500.0	0.7217399999999996	0.1411404039503023	0.34	0.63	0.72	0.82	0.97

Graph 4-2 Descriptive Summary

We can also display the descriptive summary in pictorial way. The Graph 4-2 shows the descriptive summary shown in Table 4-3 for our better understanding. In most cases the describe table is sufficient for us to get the valuable information about the data.

5 DATA VISUALIZATION

Visualizations are useful for getting a quick sense of the information contained in the data and the distribution. Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

After analysing the data, we will be able to know what the features and labels are, so from the above data, the label we have to consider is Chance of Admission and then we have to consider the parameters that influence or play a major role in Chance of Admission. We can get to know certain features that are more affected by the visualization (or) analysis. Below are some of our data visualizations.

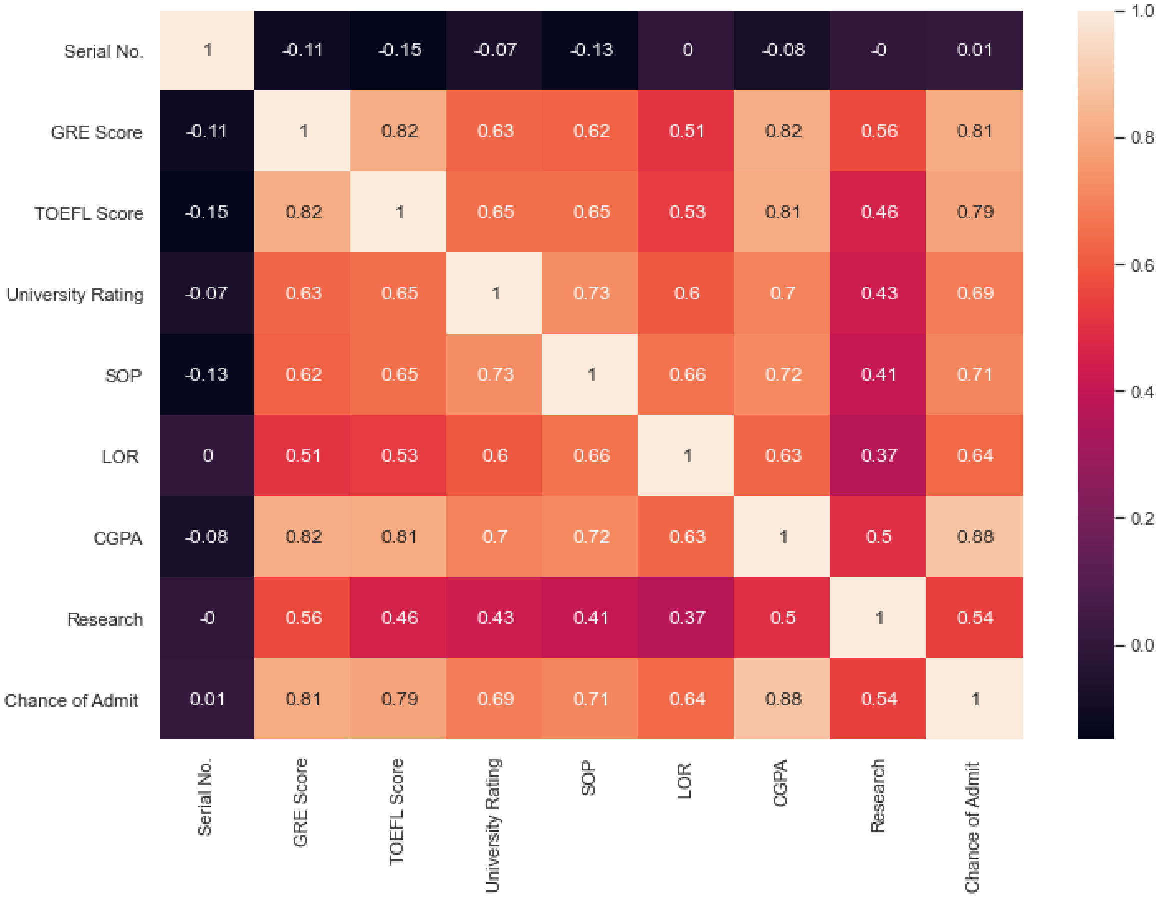
5.1 CORRELATION MATRIX

The correlation matrix is a matrix structure that helps the programmer analyze the relationship between the data variables. It represents the correlation value between a range of 0 and 1. The positive value represents good correlation and a negative value represents low correlation and value equivalent to zero (0) represents no dependency between the particular set of variables.

It is important because it helps to understand

- Understand the dependence between the independent variables of the data set.
- Helps choose important and non-redundant variables of the data set.
- Applicable only to numeric/continuous variables.

A heat map (or heatmap) is a data visualization technique that shows magnitude of a phenomenon as colour in two dimensions. The variation in colour may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space. The heatmap which is used in for the correlation is plotted below as shown in Graph 5-1.



Graph 5-1 Heatmap for correlation

To fit a model, we select those features which have a high correlation with our target variable Chance of Admit. By looking at the correlation matrix we can see that CGPA has a strong positive correlation with Chance of Admit (0.88). In this case we don't have any negative correlation with Chance of admit.

An important point in selecting features for a model is to check for multi-co-linearity. Multicollinearity is a huge issue that exists whenever an independent variable is highly correlated with one or more independent variables in a multiple regression equation. If VIF is > 10, high multicollinearity is found. This problem can lead to unstable regression model. In other words, any slight change in the data will lead to a huge change in the coefficients of the multiple linear regression model.

In conclusion, there is no multicollinearity problem since all the values are less than 10. This also leads to the fact that our regression model is stable.

The features GRE Score and TOFEL Score have a correlation of 0.82. These feature pairs are strongly correlated to each other.

The features GRE Score and CGPA have a correlation of 0.82. These feature pairs are strongly correlated to each other.

We should not select both these features together for training the model.

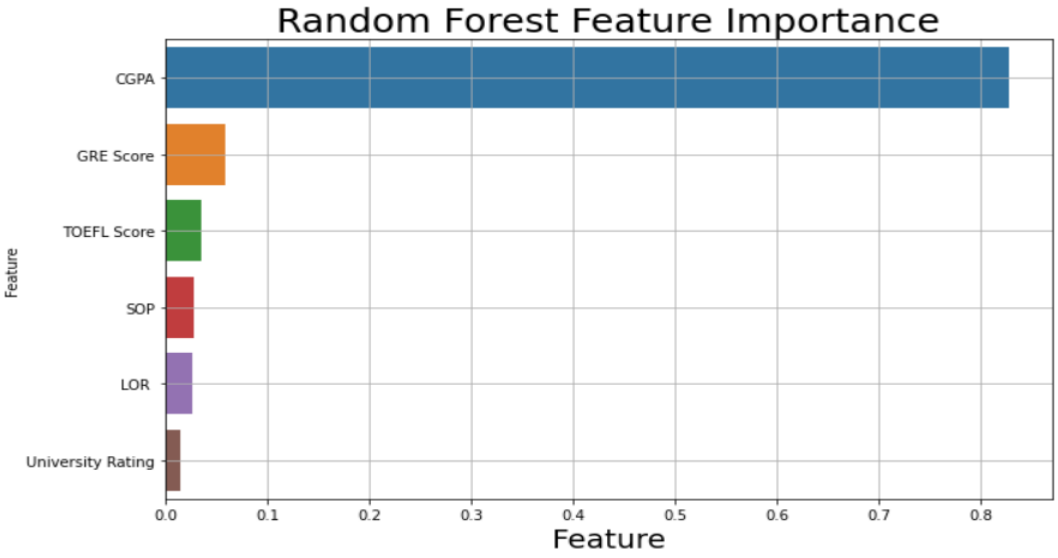
5.2 FEATURE SELECTION

Random forest regressor is used for feature selection according to the feature importance returned by the trained model on the whole dataset. The random forest regressor assigns the feature importance scores based on their qualification for being the best split at each step in the ensemble of decision trees.

To find the importance range of the independent variables, Random Forest classifier can be used. The higher the value, the more important it is.

The results in Graph 5-2 show that the most important variable is CGPA as it has the highest ranking among all other variables. And the second highest variable is GRE.

Out of 7 features, 6 features are selected for model development and further analysis as per random forest regressor’s feature importance scores.



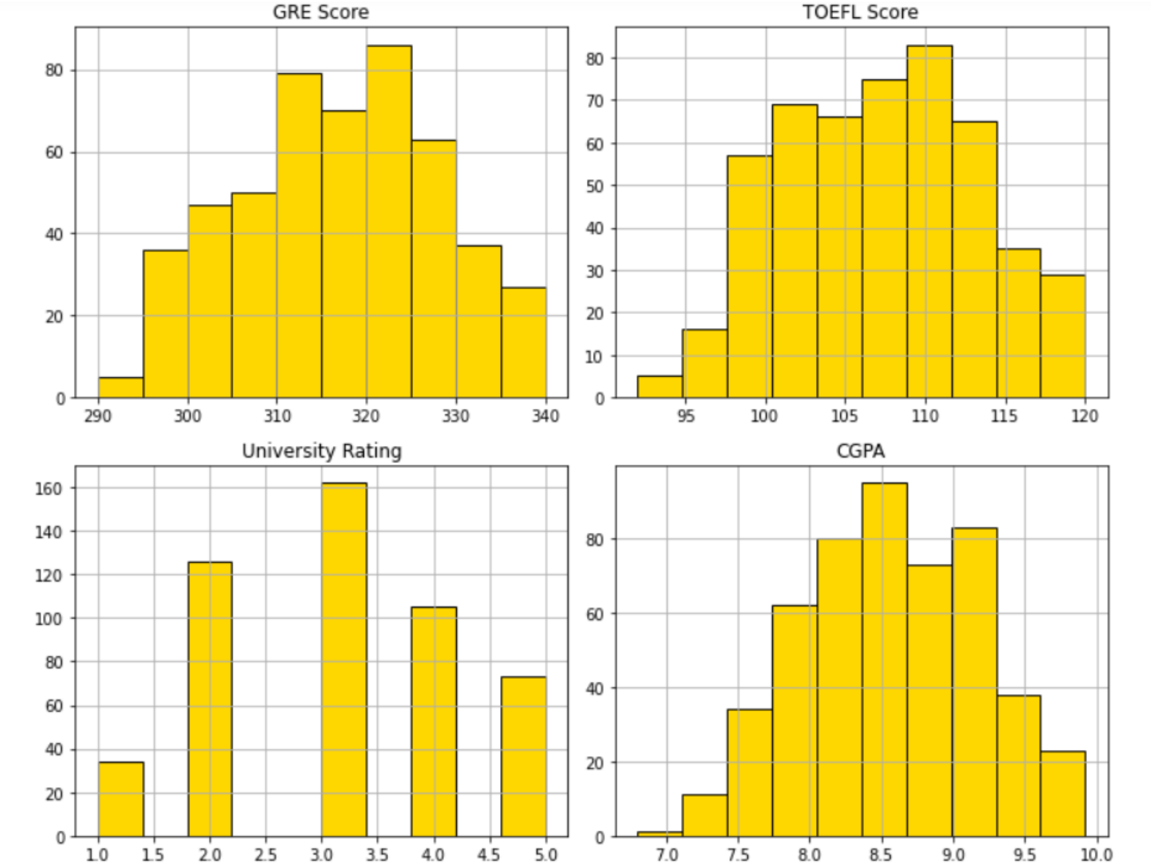
Graph 5-2 Bar plot showing features and their importance scores given by random forest regressor

5.3 VARIOUS PLOTS

5.3.1 Histogram

A histogram plot is used to present the frequencies of continuous numbers and to show the distribution of the data selected. Outliers and skewness can be predicted from a histogram along with some other features. Skewness measures how much a graph is asymmetric.

The histogram graphs of the most important independent variables are presented also according to the importance test as shown in Graph 5-3.



Graph 5-3 Histogram of important features

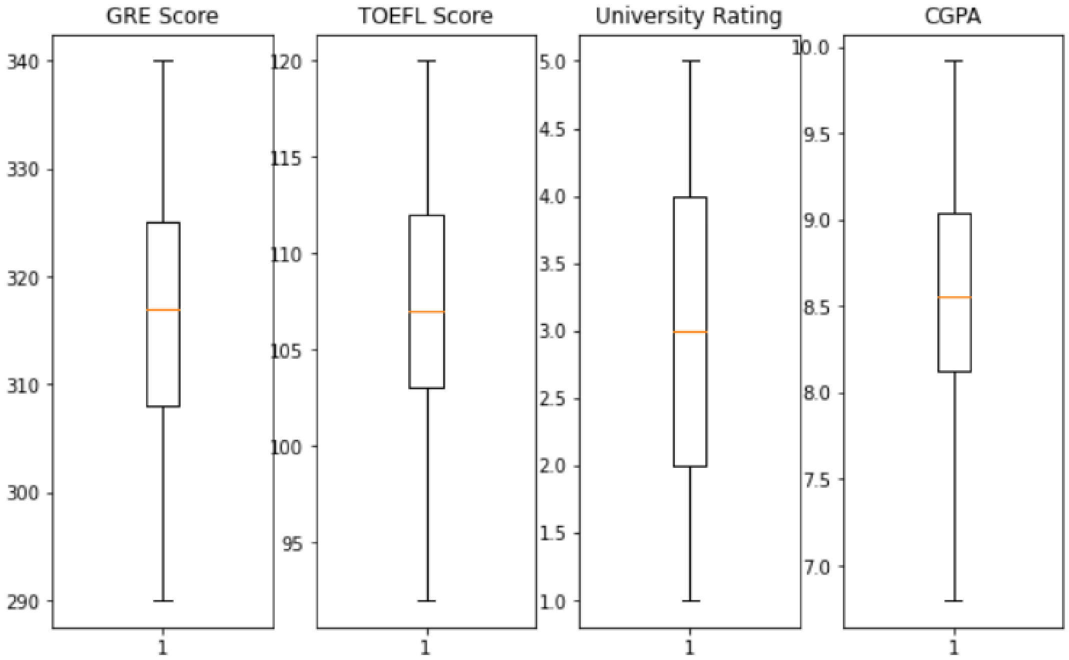
The histogram of CGPA, the most important independent variable has skewness −0.0283553 to the left. The histogram of GRE with skewness −0.04 to the left. The histogram of TOEFL Score is to the left.

5.3.2 Box Plot

The method to summarize a set of data that is measured using an interval scale is called a box and whisker plot. These are maximum used for data analysis. We use these types of graphs or graphical representation to know:

- Distribution Shape
- Central Value of it
- Variability of it

A box plot is a chart that shows data from a five-number summary including one of the measures of central tendency. It does not show the distribution in particular as much as a stem and leaf plot or histogram does. But it is primarily used to indicate a distribution is skewed or not and if there are potential unusual observations (also called outliers) present in the data set. Boxplots are also very beneficial when large numbers of data sets are involved or compared.

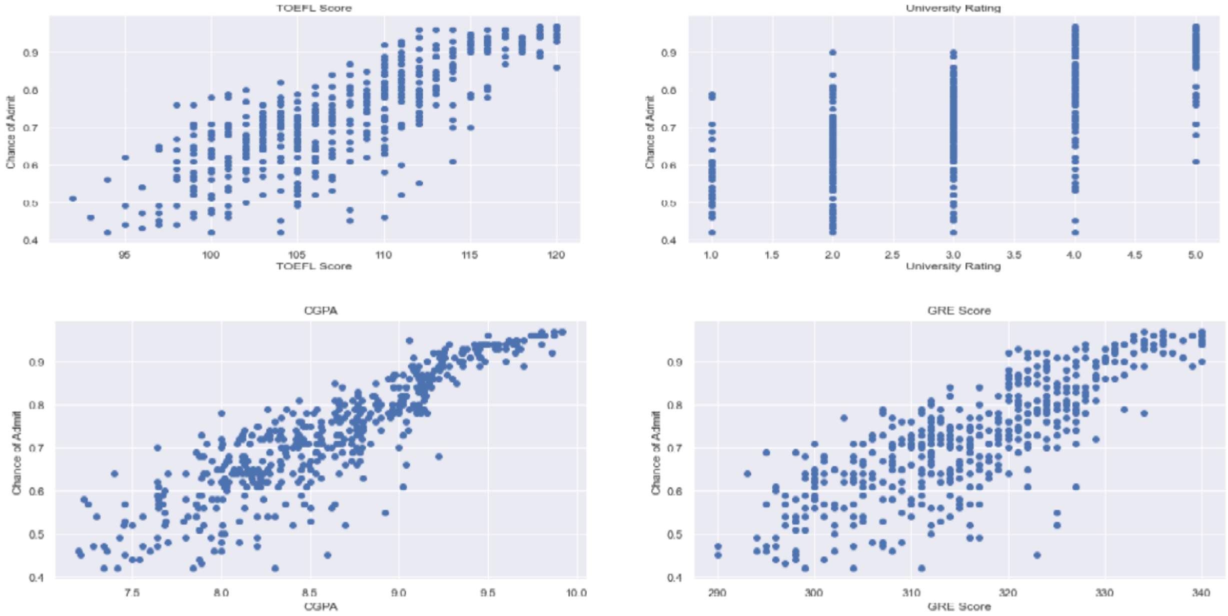


Graph 5-4 Box plot for independent variable

In the boxplot, the middle part of the plot represents the first and third quartiles. The line near the middle of the box represents the median. The whiskers on either side of the IQR represent the lowest and highest quartiles of the data. The ends of the whiskers represent the maximum and minimum of the data as shown in the Graph 5-4.

5.3.3 Scatter Plot

Scatter plots are used to plot data points on horizontal and vertical axis in the attempt to show how much one variable is affected by another. Each row in the data table is represented by a marker the position depends on its values in the columns set on the X and Y axes. A third variable can be set to correspond to the colour or size of the markers, thus adding yet another dimension to the plot.



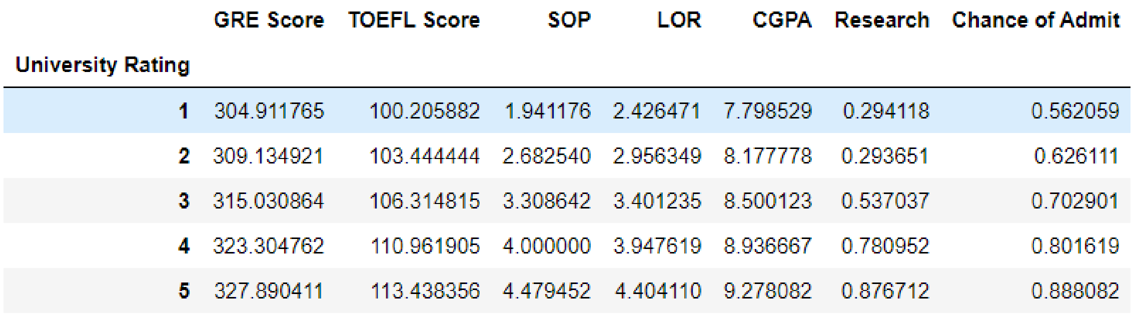
Graph 5-5 scatter plot for some variables

The Graph 5-5 shows the scatter plot for the independent variables.

5.3.4 University Rating Correlation Analysis

In order to find out how the university rating affects this model; we group the data on the basis of university rating. The Table 5-1 shows the grouping of various parameters. From this data we can get to know the average GRE Score, TOEFL Score, LOR and SOP required for the university of particular ranking.

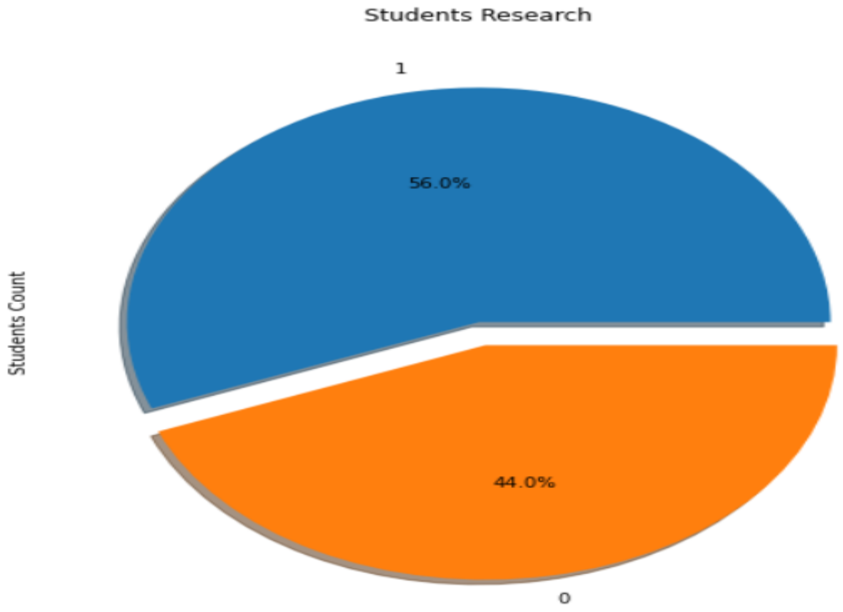
Table 5-1 Grouping on the basis of university rating



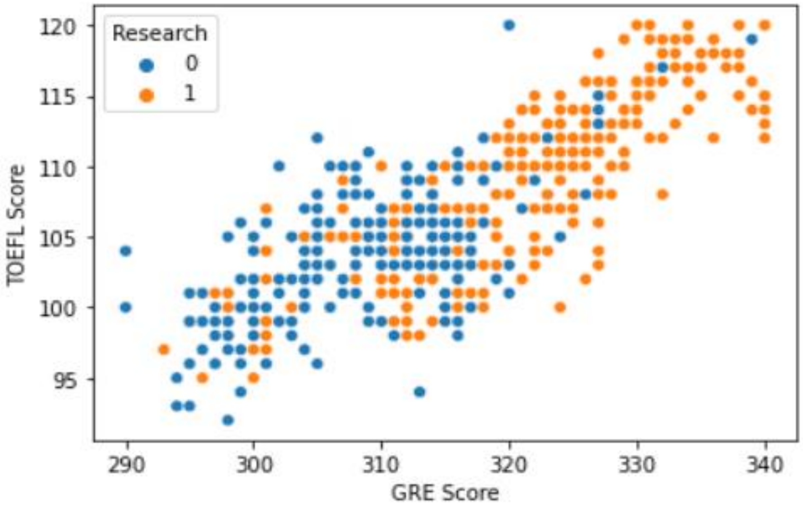
	GRE Score	TOEFL Score	SOP	LOR	CGPA	Research	Chance of Admit
University Rating							
1	304.911765	100.205882	1.941176	2.426471	7.798529	0.294118	0.562059
2	309.134921	103.444444	2.682540	2.956349	8.177778	0.293651	0.626111
3	315.030864	106.314815	3.308642	3.401235	8.500123	0.537037	0.702901
4	323.304762	110.961905	4.000000	3.947619	8.936667	0.780952	0.801619
5	327.890411	113.438356	4.479452	4.404110	9.278082	0.876712	0.888082

5.3.5 Importance of Research to get Admission

Out of 500 students the total number of students having research is 280 and those not having research is 220. The Graph 5-6 shows that 56% of students have done research. We can say that only better students can get a chance for doing research. Doing research adds practical knowledge and increases the student's skill of working with groups or teams.



Graph 5-6 Pie diagram showing students have done research



Graph 5-7 Scatter plot for research

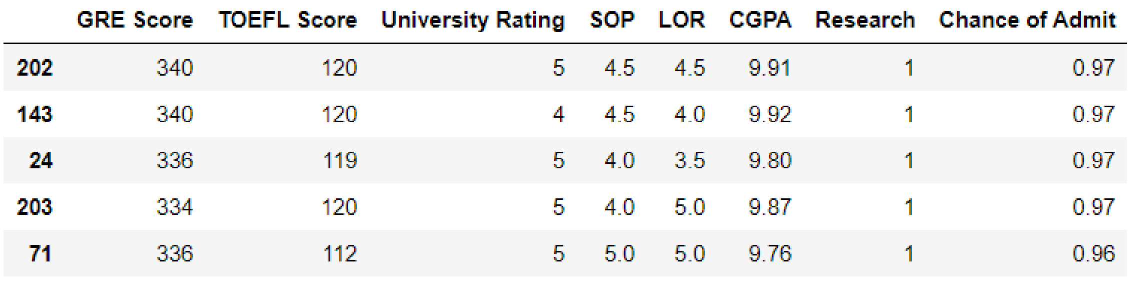
In Graph 5-7 we can see that students who have done research do have good TOEFL and GRE Score. Also, the students who have high TOEFL and GRE Score have high chance of getting admission.

Therefore, the chance of admission increases for those who have done research.

5.3.6 Score for 90% chance of admission

Another important thing we need to analyse is the scores of various parameters needed for 90% chance of admission.

Table 5-2 Determining the chance of admit



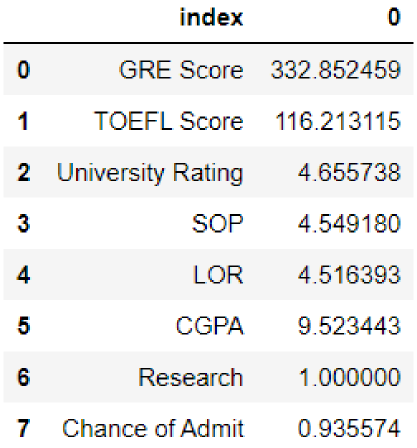
	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
202	340	120	5	4.5	4.5	9.91	1	0.97
143	340	120	4	4.5	4.0	9.92	1	0.97
24	336	119	5	4.0	3.5	9.80	1	0.97
203	334	120	5	4.0	5.0	9.87	1	0.97
71	336	112	5	5.0	5.0	9.76	1	0.96

The Table 5-2 shows that the maximum chance of admission is 0.97.

The Table 5-3 shows the scores required for getting 90% chance of admissions. For having a 90% chance to get admission one should have GRE Score = 332.8, TOEFL Score =

116.2 and CGPA = 9.52. If students get more scores than mentioned above then they will have good chance of admission.

Table 5-3 Scores required for 90% chance of admit



	index	0
0	GRE Score	332.852459
1	TOEFL Score	116.213115
2	University Rating	4.655738
3	SOP	4.549180
4	LOR	4.516393
5	CGPA	9.523443
6	Research	1.000000
7	Chance of Admit	0.935574

In order to prevent the biasness and the overfitting of the model scaling was done using StandardScaler and MinMaxScaler.

As a result of the above visualization and data analysis, the features given have a high impact on the probability of admission, so these features are considered only.

6 BUILDING MODELS

Once the data visualization is done, we have to do predictive modelling for this purpose first we divide the data into train part and test part. The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. The train test split technique can be used for classification and regression problems to test machine learning algorithms. The procedure takes the given dataset and splits it into two subsets:

- Training dataset: it is used to train the algorithm and fit the machine learning model.
- Test dataset: Using the input element from the training data, the algorithms make predictions.

The model is first to fit on the available data with known inputs and outputs. It is then run to make predictions on the rest of the data subset to learn from it. The dataset is consistently shuffled and split into training, and test set such as 85% of the 500 instances is present in the training set, and remaining 15% of the 500 instances are present in the test set.

a) Multiple Linear Regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable. The equation for the multiple linear regressors looks as follows:

y = b0 + b1 *x1 + b2 * x2 + ... + bn * xn

Here, y is dependent variable and x1, x2,..,xn are our independent variables that are used for predicting the value of y. Values such as b0,b1,…bn act as constants.

Multiple regression analysis is also used to assess whether confounding exists. Since multiple linear regression analysis allows us to estimate the association between a given independent variable and the outcome holding all other variables constant, it provides a way of adjusting for (or accounting for) potentially confounding variables that have been included in the model.

b) MLR With PCA

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

This lower dimension version of the data can be useful to train models more efficiently where a large number of attributes can slow the training or increase noise or bias. However, an important contribution is visualizing high dimensional data which is still an active area of research. By transforming the data into two or three-dimensional space, it can be visualized in traditional plots. This can help to detect clusters, outliers, and the general distribution of the data. This technique is applied after the data has been transformed, scaled, and imputed.

Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analysing data much easier and faster for machine learning algorithms without extraneous variables to process.

So, to sum up, the idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.

c) Random Forest Regression

The Random Forest Regression [9] is a type of additive model that makes predictions by combining decision from a sequence of base models as shown in Figure 6-1. Random forest is a machine learning algorithm which is a combined effect of classification and regression and other tasks which operate by erection of decision trees at training time and outputs the class that is the mode of the classes or mean value of individual trees. Each base model is a Decision Tree and the result of the

Random Forest model is the cumulative output of the Decision Trees. This technique of using multiple models to obtain better predictive performance is called model ensembling.

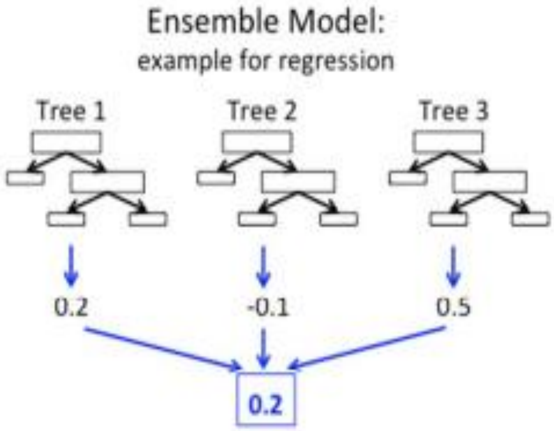


Figure 6-1 Random Forest as a combination of multiple Decision Trees

In Random Forests, all the base models are constructed independently using a different subsample of the data. This algorithm creates forests within number of decision reads. Therefore, the more data is available the more accurate and robust results will be provided. The random forest model is very good at handling tabular data with numerical features, or categorical features with fewer than hundreds of categories. Unlike linear models, random forests are able to capture non-linear interaction between the features and the target. Random Forest method can handle large datasets with higher dimensionality without over fitting the model. In addition, it can handle the missing values and maintains accuracy of missing data.

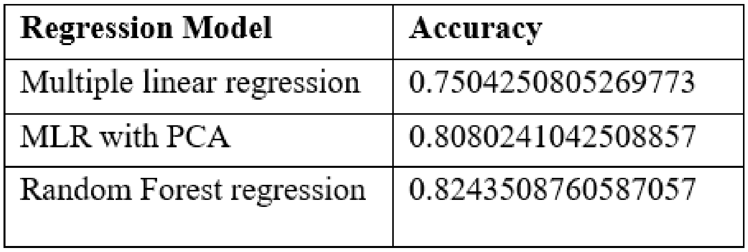
7 MODEL EVALUATION

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Model evaluation is done using evaluation metrics. Evaluation metrics are the measures by which we can rate and understand the performance of a machine learning model.

7.1 ACCURACY

Accuracy is a method for measuring a classification model’s performance. It is typically expressed as a percentage. Accuracy is the count of predictions where the predicted value is equal to the true value. It is binary (true/false) for a particular sample. Accuracy is often graphed and monitored during the training phase though the value is often associated with the overall or final model accuracy.

Table 7-1 Accuracy for various models



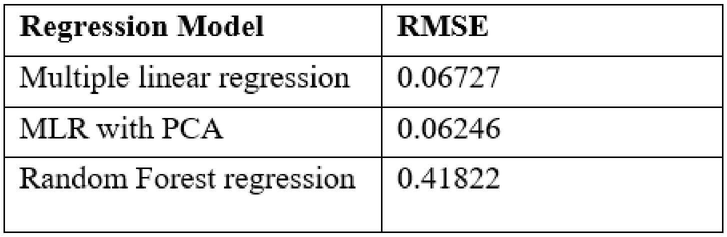
Regression Model	Accuracy
Multiple linear regression	0.7504250805269773
MLR with PCA	0.8080241042508857
Random Forest regression	0.8243508760587057

Accuracy for the various models performed is given in Table 7-1. So, the random forest regression models show the accuracy of 82% but it does not mean that this model is doing a great job. In an imbalanced case, accuracy is a misleading metric. For this data set, simply predicting graduated for all student samples should give an accuracy of 79%, which is far above random guessing. This model would actually be useless for the problem despite having a good score. With imbalanced data sets, models can quickly develop a bias towards the majority class while ignoring type 1 and type 2 errors. Accuracy is not used to evaluate the models for these reasons.

7.2 RMSE

The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. The RMSE represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences. These deviations are called residuals when the calculations are performed over the data sample that was used for estimation and are called errors (or prediction errors) when computed out-of-sample. The RMSE serves to aggregate the magnitudes of the errors in predictions for various data points into a single measure of predictive power. RMSE is a measure of accuracy, to compare forecasting errors of different models for a particular dataset and not between datasets, as it is scale-dependent. RMS is the square root of the average of squared errors.

Table 7-2 RMSE scores for various models



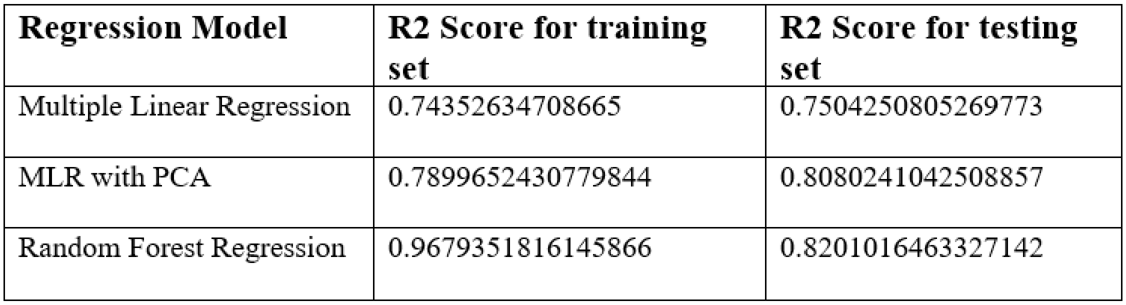
Regression Model	RMSE
Multiple linear regression	0.06727
MLR with PCA	0.06246
Random Forest regression	0.41822

The Table 7-2 shows the RMSE scores for various models. RMSE is least for random forest and highest for multiple linear regression. Hence, random forest model has a good performance.

7.3 R SQUARED SCORE

R-squared is a statistical measure that represents the goodness of fit of a regression model. The ideal value for r-square is 1. The closer the value of r-square to 1, the better is the model fitted. R-square is a comparison of residual sum of squares with total sum of squares. Total sum of squares is calculated by summation of squares of perpendicular distance between data points and the average line. The value of r-square always increases or remains same as new variables are added to the model, without detecting the significance of this newly added variable (i.e., value of r-square never decreases on addition of new attributes to the model). As a result, non-significant attributes can also be added to the model with an increase in r-square value. The Table 7-3 below shows the R2 scores for various models. The higher the values the better will be the performance of the model. So the random forest regression model have better performance.

Table 7-3 R2 scores for various models



Regression Model	R2 Score for training set	R2 Score for testing set
Multiple Linear Regression	0.74352634708665	0.7504250805269773
MLR with PCA	0.7899652430779844	0.8080241042508857
Random Forest Regression	0.9679351816145866	0.8201016463327142

7.4 MEAN ABSOLUTE ERROR

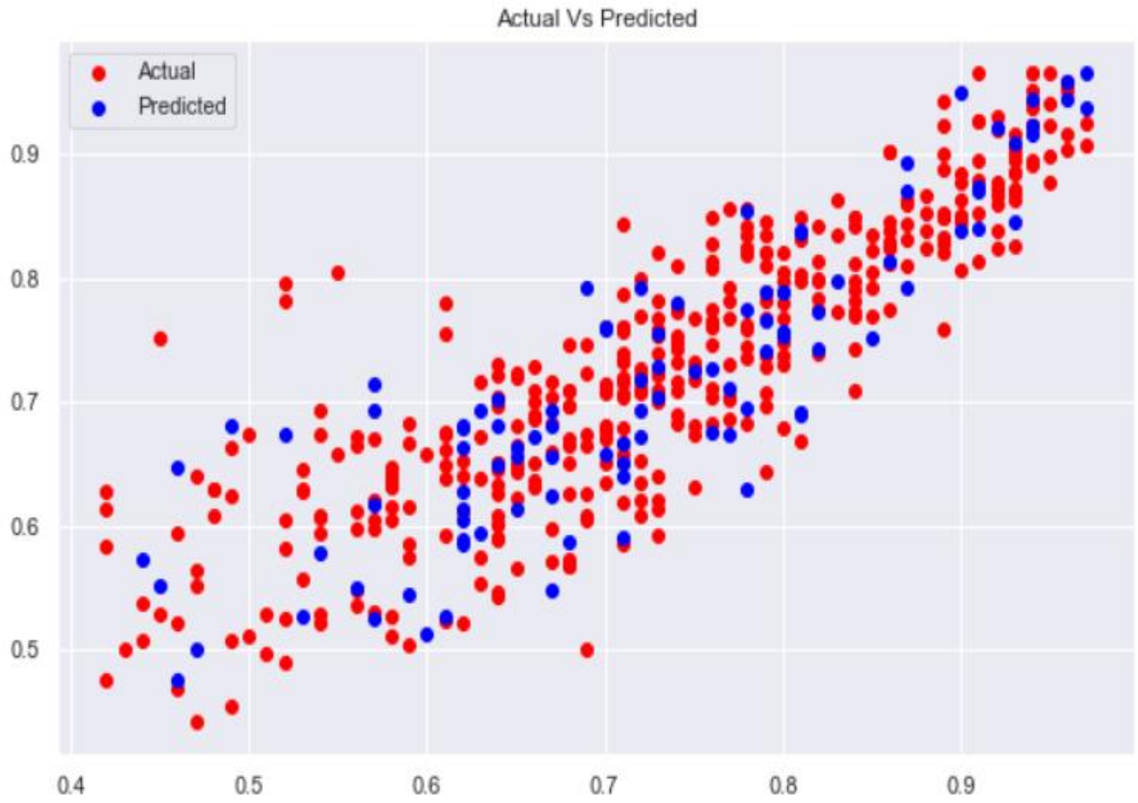
Given any test data-set, Mean Absolute Error of your model refers to the mean of the absolute values of each prediction error on all instances of the test data-set. Prediction error is the difference between the actual value and the predicted value for that instance.

The MAE value for the random forest regression model is 0.3077086171390197 which is a small value. Hence, we can say that this model has a good performance.

7.5 ACTUAL VS PREDICTED VALUE

In statistics, the actual value is the value that is obtained by observation or by measuring the available data. It is also called the observed value. The predicted value is the value of the variable predicted based on the regression analysis.
The difference between the actual value or observed value and the predicted value is called the residual in regression analysis. The nature of regression fits for the gradient boosting regressor model is visualized in the form of fitting diagrams.
There are three natures of regression fits for evaluating model performance: –

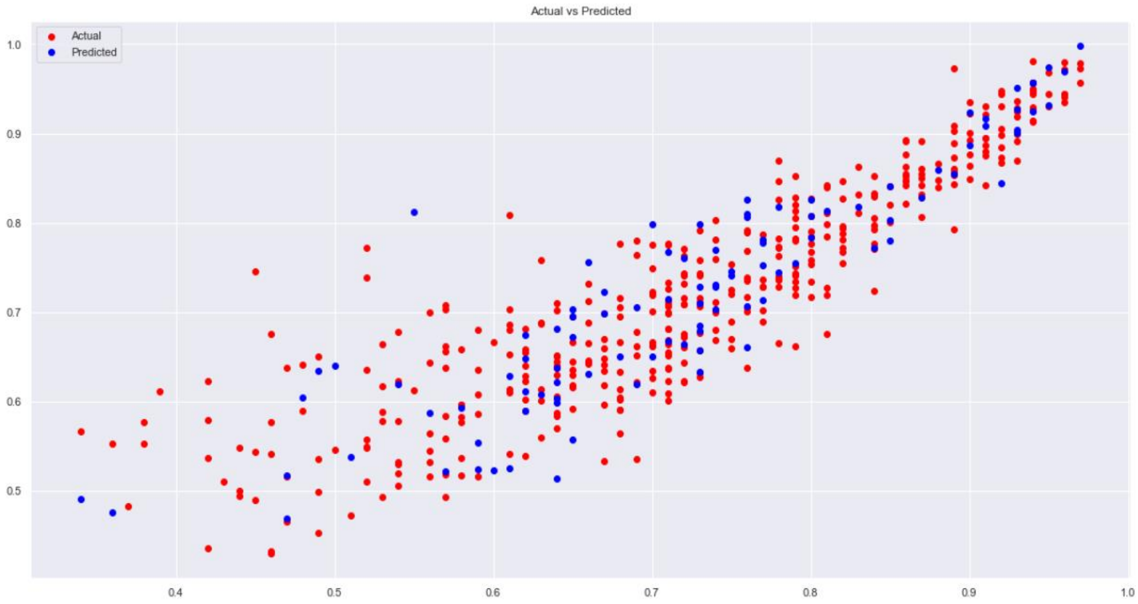
- Under Fit: This refers to the nature of the fit in which the model under-performs with both the training set instances and test set instances.
- Perfect Fit: This refers to the nature of the fit in which the model works excellently with both the training and test set instances.
- Over Fit: This refers to the nature of the fit in which the model works excellently with only the training set instances and under-performs with the test set instances.



Graph 7-1 Plot for MLR model

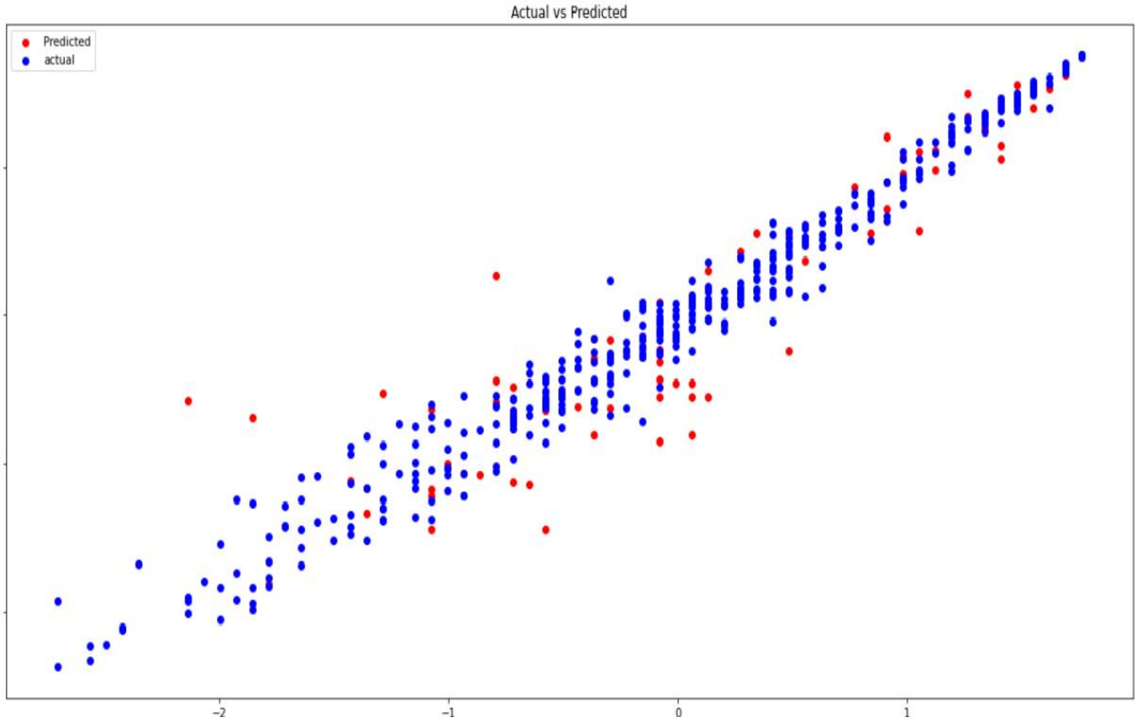
The Graph 7-1 shows the actual and predicted values for multiple linear regression model. It shows the over fitting of the model.

The Graph 7-2 shows the actual and predicted values for multiple linear regression with PCA model. It also shows the over fitting to some extent.



Graph 7-2 Plot for MLR with PCA model

The Graph 7-3 shows the actual and predicted values for random forest regression model. It can be interpreted that this regression fit is close to a perfect fit.



Graph 7-3 Plot for Random Forest Regression model

8 CONCLUSION

In this project, machine learning models were performed to predict the opportunity of a student to get admitted to a master's program. The machine learning models included are multiple linear regression, random forest and multiple linear regression with PCA.

After evaluating all three models on the dataset, we compare the performances to find out which model predicts better. Based on evaluation metrics we can say that the random forest regression is good followed closely by MLR with PCA. As compared with the MLR with PCA it can be used by the students for evaluating their chances of getting shortlisted in a particular university with an average accuracy of 82%. Random forest regression may be used in order to learn better the relations between different variables; still, the results achieved are appropriate and reliable for the chosen fields only. In other words, the use of the random forest regression model cannot be general.

The main limitation of this project is we developed models based solely on data from Indian Students, we considered only few universities with different rankings. More information relating to new colleges and courses can be added to the curriculum in the future. The system may also be modified to a web-based application by making node-red modifications. To solve the problem, it is possible to test other classification algorithms if they have high accuracy score than the current algorithm. Finally, students can have an open source Machine Learning model which will help the students to know their chance of admission into a particular university with high accuracy.

REFERENCES

[1] M. S. Acharya, A. Armaan, and A. S. Antony, “A comparison of regression models for prediction of graduate admissions,” ICCIDS 2019 - 2nd Int. Conf. Comput. Intell. Data Sci. Proc., pp. 1–5, 2019

[2] N. Chakrabarty, S. Chowdhury, and S. Rana, “A Statistical Approach to Graduate Admissions’ Chance Prediction,” no. March, pp. 145–154, 2020.

[3] N. Gupta, A. Sawhney, and D. Roth, “Will i Get in? Modeling the Graduate Admission Process for American Universities,” IEEE Int. Conf. Data Min. Work. ICDMW, vol. 0, pp. 631–638, 2016.

[4] A. Waters and R. Miikkulainen, “GRADE : Graduate Admissions,” pp. 64–75, 2014.

[5] S. Sujay, “Supervised Machine Learning Modelling & Analysis for Graduate Admission Prediction,” vol. 7, no. 4, pp. 5–7, 2020.

[6] Bibodi, J., Vadodaria, A., Rawat, A. and Patel, J. (n.d.). Admission Prediction System Using Machine Learning.

[7] https://pdfs.semanticscholar.org/39b2/cd2a11ebdeb4d31c761527195e06a7136314.pdf

[8] Roa, Annam Mallikharjuna, et al. “College Admission Predictor.” Journal of Network Communications and Emerging Technologies (JNCET) www.jncet.org 8.4 (2018)

[9] Random Forest Regression,Turi Machine Learning Platform, https://turi.com/learn/userguide/supervisedlearning/random_forest_regression.html