PREDICTIVE MODELING FOR CHANCE OF ADMISSION

UCLA Graduate Dataset

By: Junran Cao, Ricardo Diaz, Jeremiah Ogunla, and Osemekhian Ehilen

DATA SET

- Obtained from Kaggle. Inspired by the UCLA Graduate data set. Created by several Indian students in 2019

- Data of 500 applicants were collected

- 7 predictor variables and 1 outcome variable

Features used:

Test scores, personal statement, university rating, recommendation letter, CGPA, research experience

Predicted value: Chance of admission

PREPROCESSING AND CLEANING

- Binning ‘GRE score’ and ‘TOEFL score’ into 4 groups
- Making ‘University rating’ categorical
- ‘’Restricting” the data - identifying the valid range of data values
- Interquartile range detection for checking outliers
- Checking for missing values

VISUALIZATION

"Visualization helps us to better understand the data."

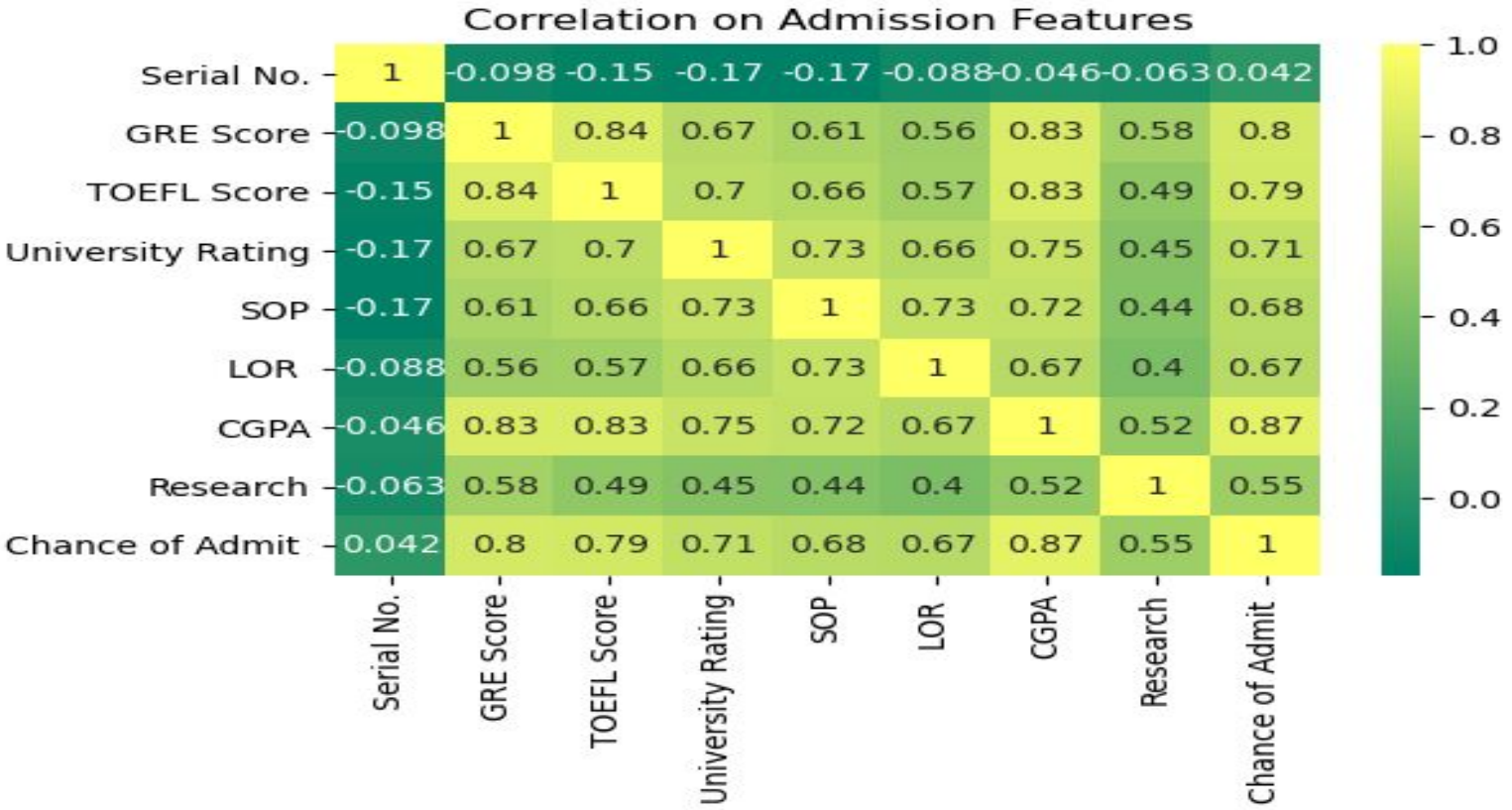
A Step-by-step approach

Heat map: correlations on the admission features

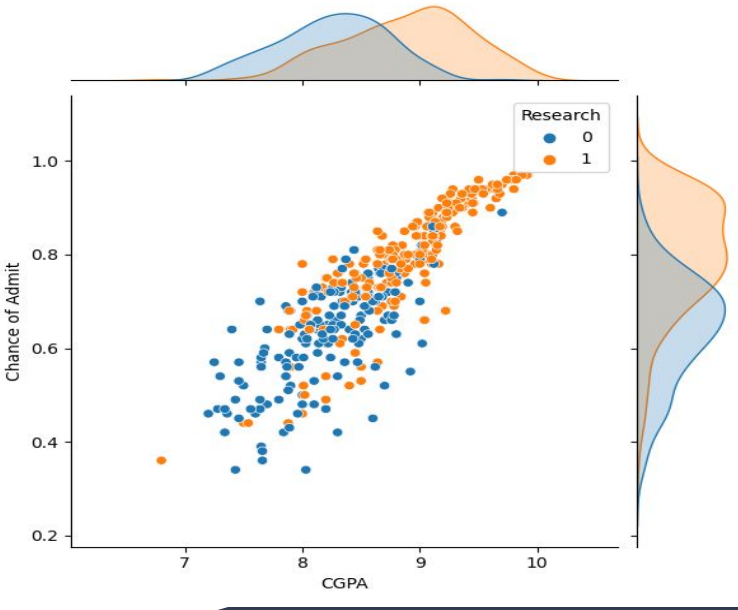
Joint plots: relationship between the admission features and 'Chance of admit'

Box plots: relationship between 'Research', 'University Rating', and 'Chance of admit' through their quartiles

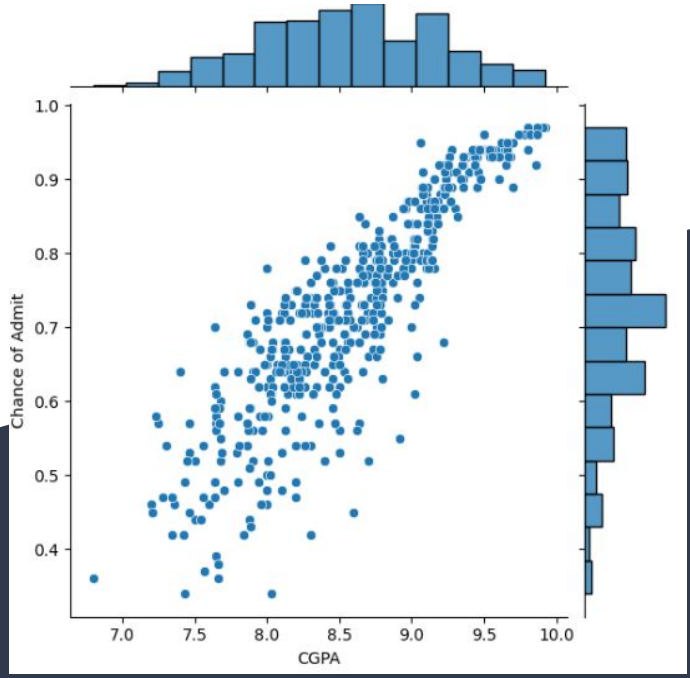
Visualization - Heat map



Visualization- Joint Plots

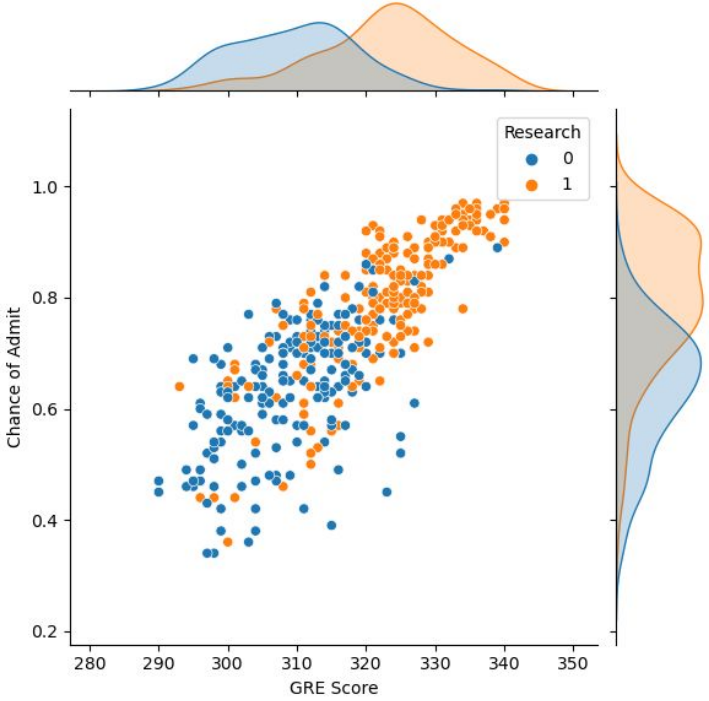


In addition to higher ‘CGPA’, having research work increases the chance of admission

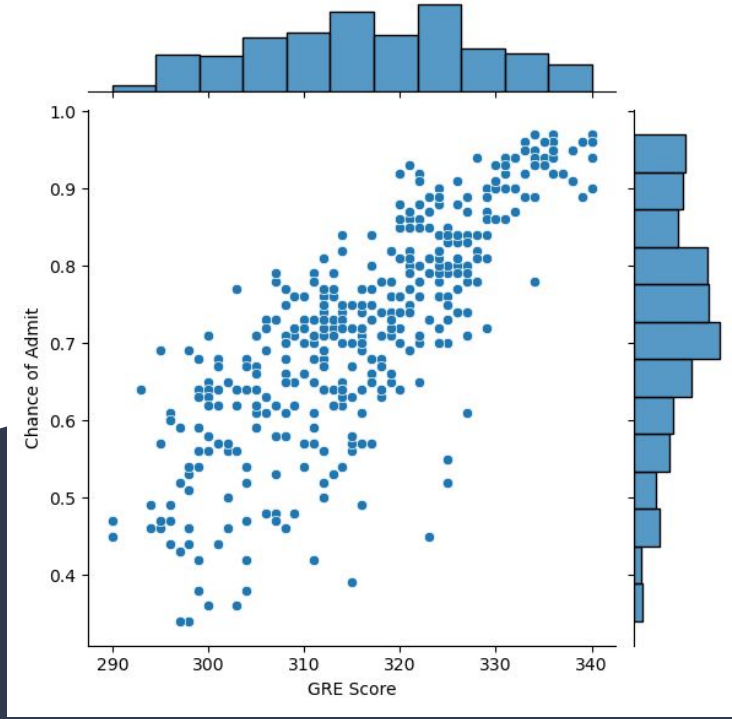


‘CGPA’ is highly positive correlated to the Chance of Admit

Visualization- Joint Plots

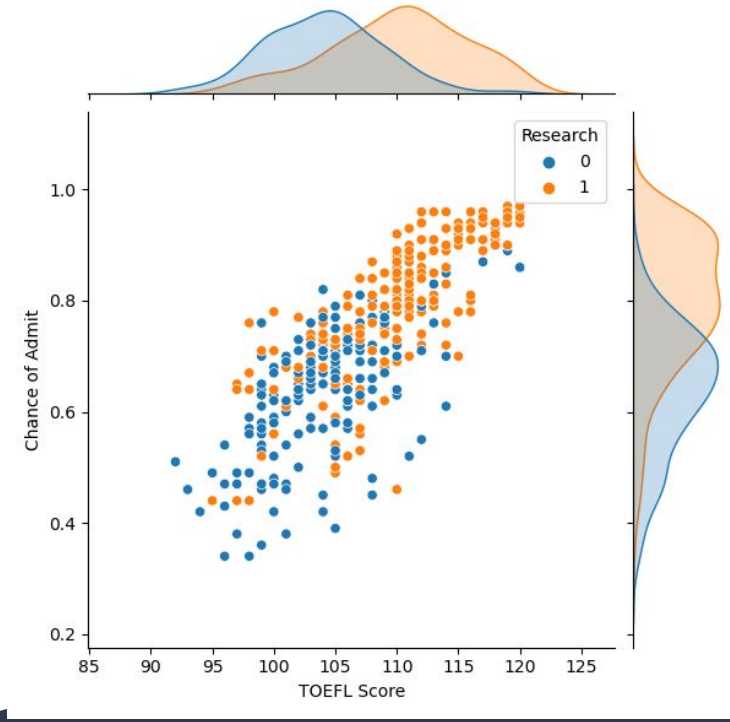


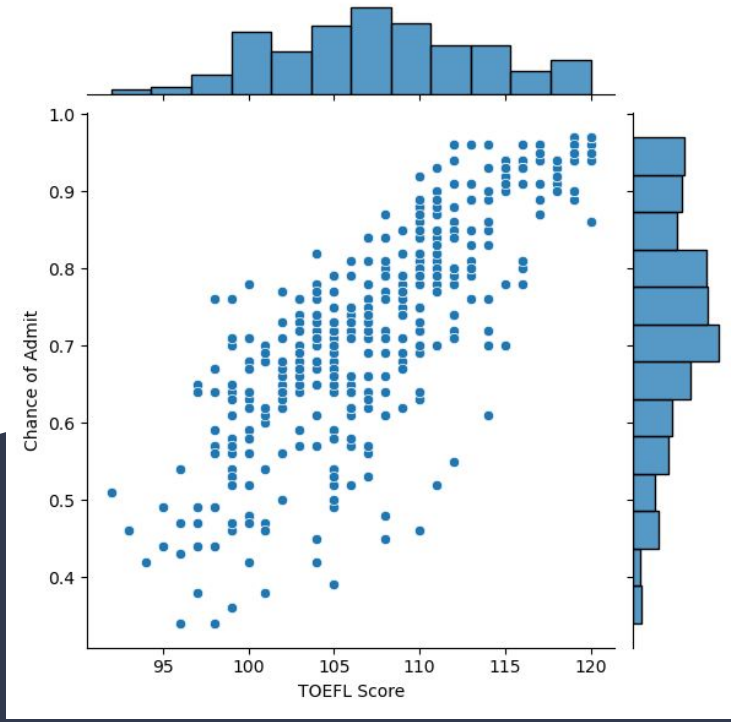
In addition to higher ‘GRE Score’, having research experience increases the chance of admission



‘GRE Score’ is highly positive correlated to ‘Chance of Admit ‘

Visualization- Joint Plots

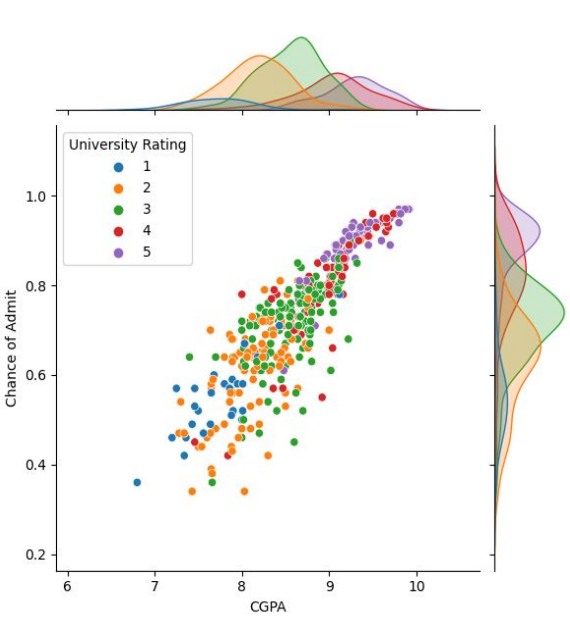


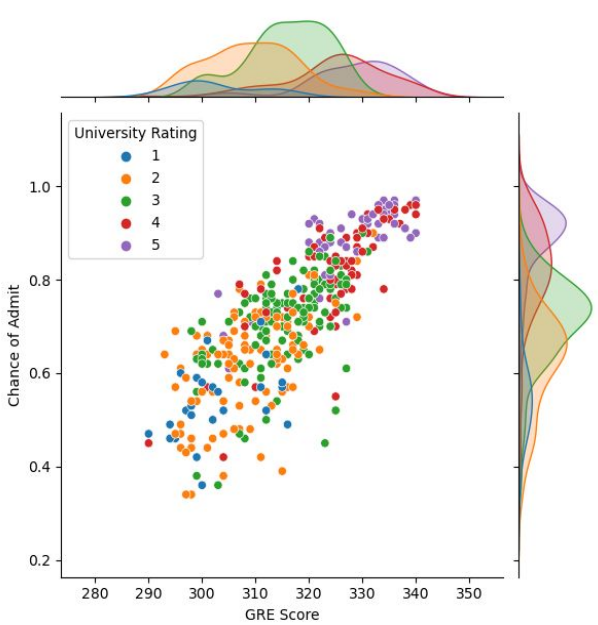


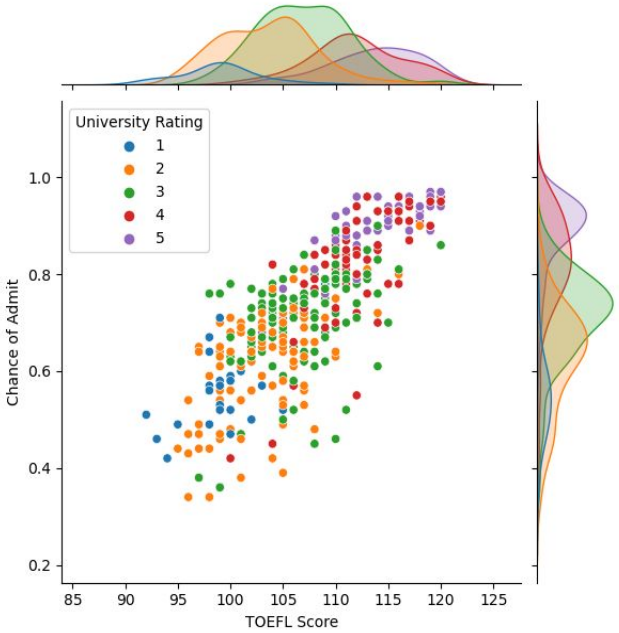
In addition to higher ‘TOEFL Score’, having research experience increases the chance of admission

‘TOEFL Score’ is positive correlated to ‘Chance of Admit’

Visualization- Joint Plots

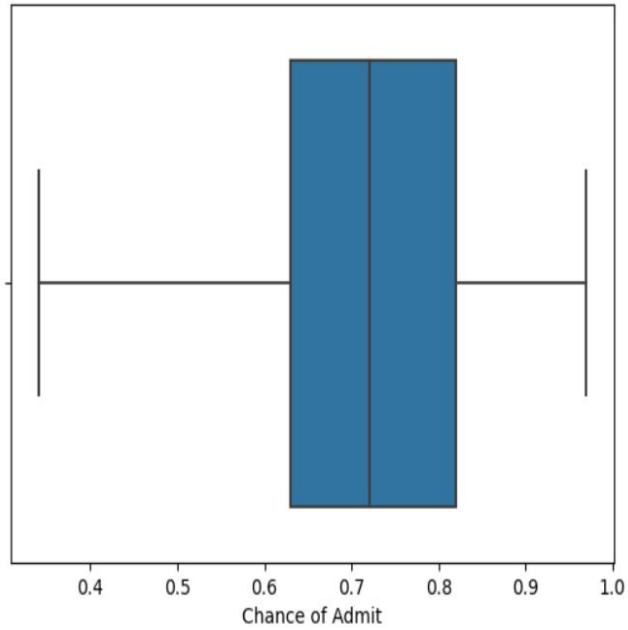




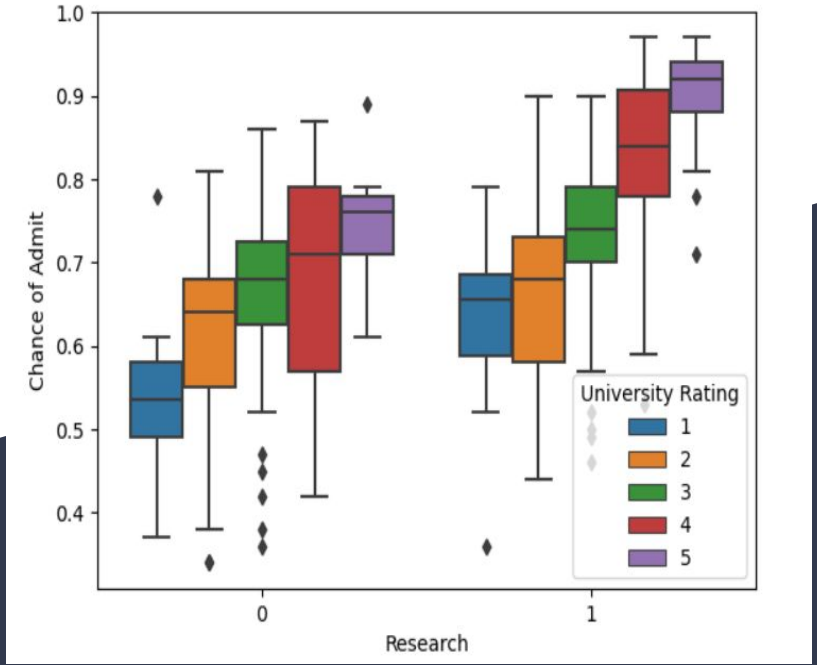


The highest ‘University Rating’ has the highest ‘Chance of Admit’

Visualization – Box plotS

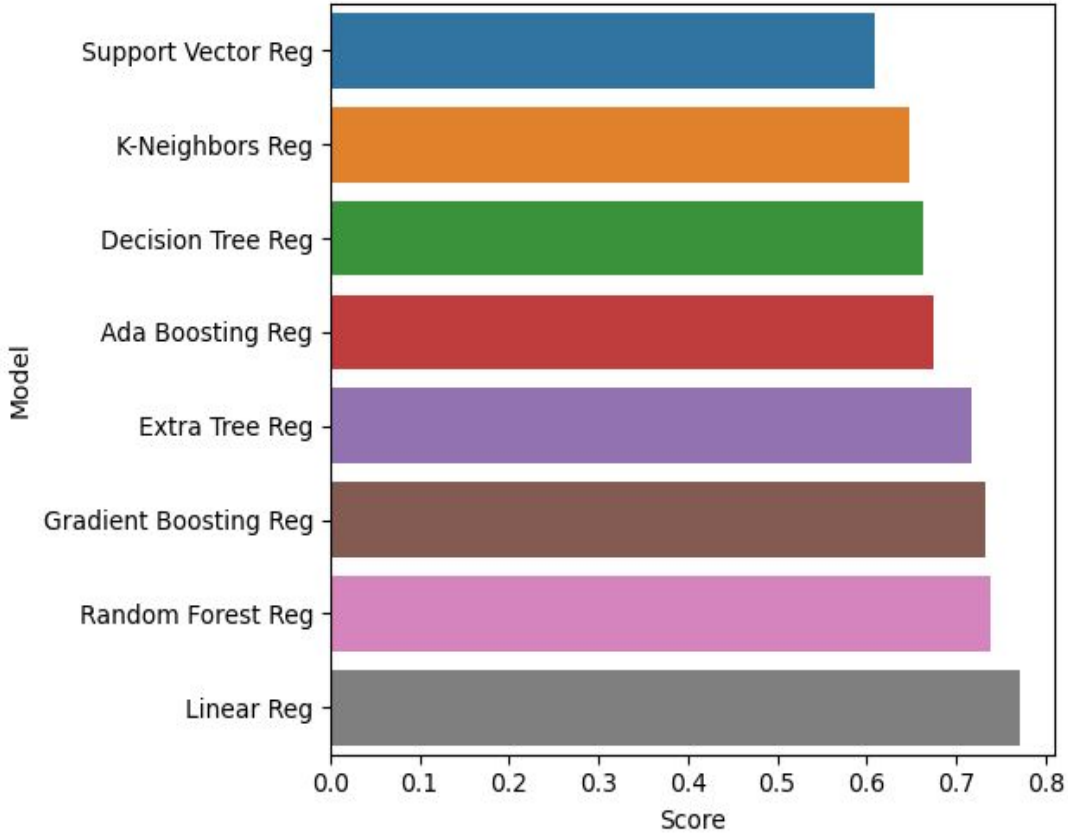


The boxplot shows the distribution of our target variable (‘Chance of Admit’)

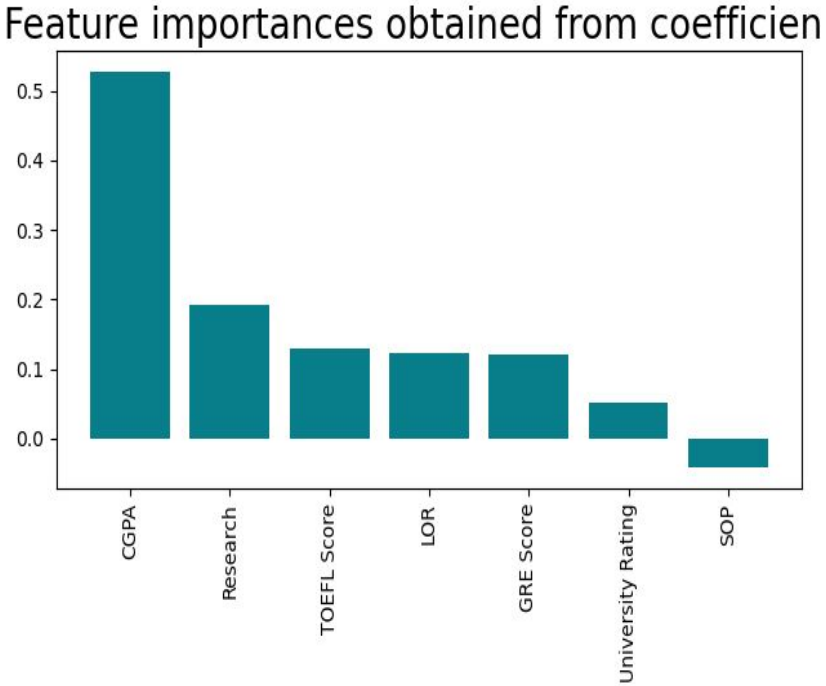


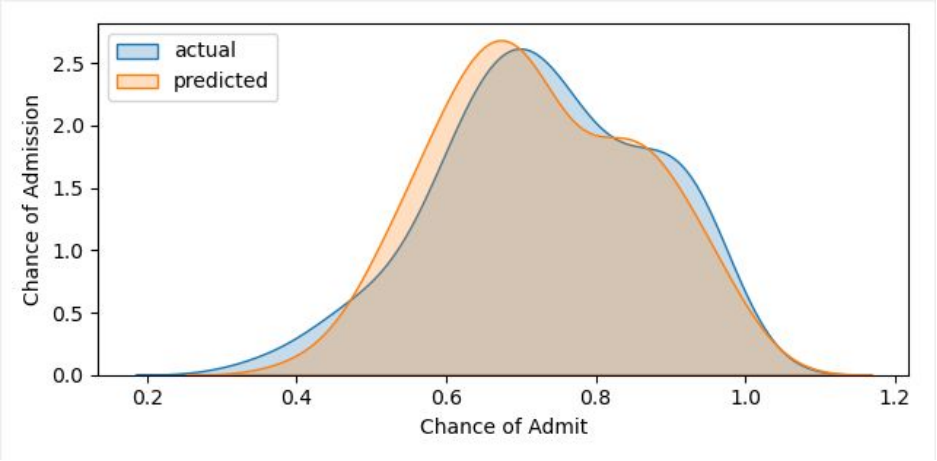
It can be noticed that with a higher university rating and having a research experience increases the chance of admission

CROSS VALIDATION – Identifying the best model



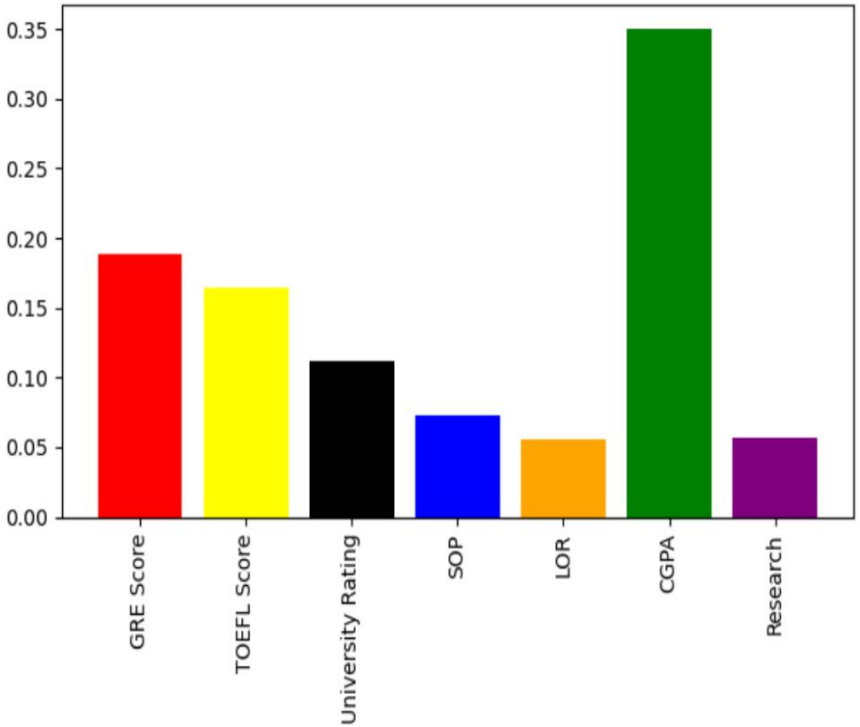
LINEAR REGRESSION – Results





Linear Regression Score (R Square) : 78.398%

RANDOM FOREST REGRESSION – Results



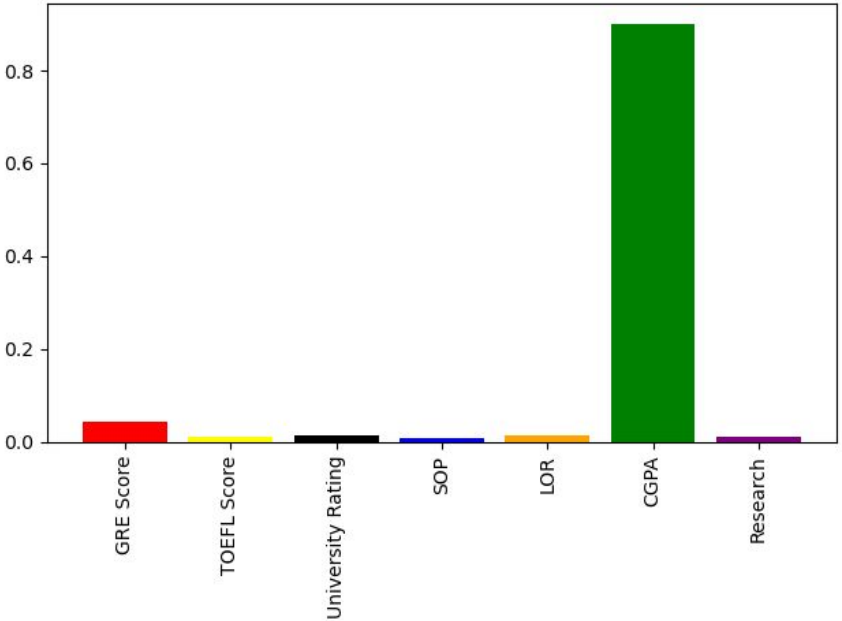
Regressor score: 78%
MSE: 0.42%

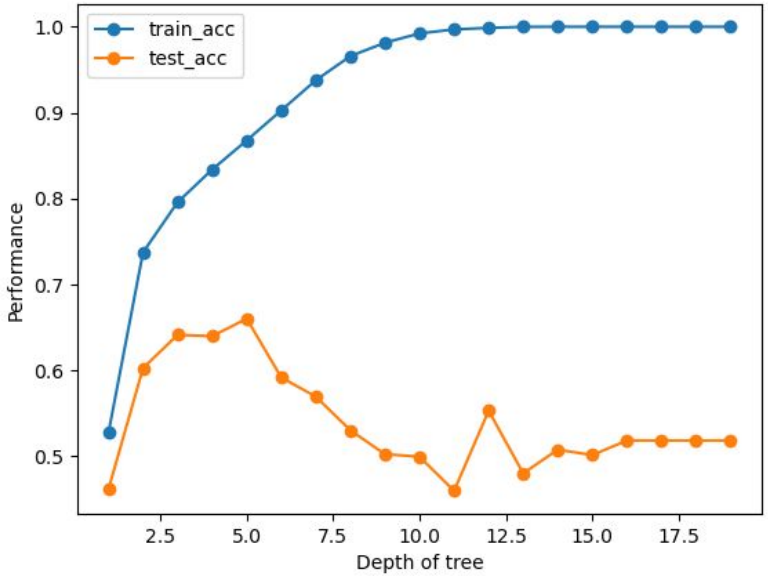
Best parameters:

Using: RandomizedSearchCV

RandomForestRegressor(n_estimators=80, min_samples_split=2, min_samples_leaf=1, max_features='sqrt', max_depth=10,bootstrap=True)

DECISION TREE REGRESSION





Not totally in line with our EDA, lower regressor score (69%)

CONCLUSION

- The three machine learning models showed us that College CGPA has the most important feature in predicting the chance of admission to the master's programs in the US, meaning that schools value academic criteria (standardized test scores and GPA) more than non-academic criteria (letter of recommendation, statement of purpose, research, and university rating)

- The linear regression model gave the highest R square value of 78.39%.
- Testing the scaled independent variables did not give us a significant increase in the score on our modeling.
- Linear regression was the best model.
- Voting Regression with the the three models gave us lower score compare to the RF and LR model.

THANK YOU