Graduate Admissions Analysis

Karthik Rangaraj   
Student ID:801135834  
[krangar2@uncc.edu](mailto:krangar2@uncc.edu)

Sriram Gunja  
Student ID: 801136225  
[sgunja@uncc.edu](mailto:sgunja@uncc.edu)

Sriganesh Lokesh   
Student ID: 801135650  
[slokesh1@uncc.edu](mailto:slokesh1@uncc.edu)

Herleen Kaur Sanhotra   
Student ID: 801080524  
[hsanhotr@uncc.edu](mailto:hsanhotr@uncc.edu)

Kanika Sani  
Student ID: 801151445[ksaini1@uncc.edu](mailto:ksaini1@uncc.edu)

Snehal Kekane  
Student ID: 801136357  
[skekane@uncc.edu](mailto:skekane@uncc.edu) Peter Kim   
Student ID: 800590688[*pkim11@uncc.edu*](mailto:pkim11@uncc.edu)

Roa Locklear  
Student ID: 800590688[*rlockl10@uncc.edu*](mailto:rlockl10@uncc.edu)

An Luong  
Student ID: 800590688[*aluong1@uncc.edu*](mailto:aluong1@uncc.edu)

*Abstract*— The aim of the paper is to develop a machine learning model to help students during their graduate admissions process by checking if their profile will match the requirements of the university which they are planning to apply or which they have already applied to. The predicted output gives them a fair idea about their admission chances in a university. In recent years, the number of students applying for university has increased drastically, making the traditional review process tedious. In order to overcome this problem, universities can use this model to filter out low scoring or high scoring applications.

In this paper, we have found out correlations between all the variables and how they impact the dataset. Various visualizations are plotted to validate the hypothesis we generated. For building this model, we have used XLMiner and the Graduate Admissions dataset from Kaggle.

Keywords—Data Cleaning, Data Visualization, Data Preprocessing, Machine Learning, Classification, Evaluation

# Introduction

The graduate admissions is a crucial step for both the student and the university, both striving for the best. The admissions process will play a key role in controlling the quality of higher education. We are considering classification models in combination with variable selection and data pre-processing. One of the main goals of Education Data Mining(EDM) is to reduce and filter out the variables required to assess a student's performance to predict his chances of getting into a university.

The different components which are used by the university’s admissions committee for evaluating the student’s profile are:

* *Graduate Record Examinations (GRE)*

The GRE General Test, is an examination conducted by ETS, is a major requirement for all the major universities in the United States of America. It is mainly used by the universities to filter out bad applications. A candidate with a very low GRE score will be rejected, while a student with exceptional GRE will also be subject to evaluation. With a minimum score of 280, students are expected to get a minimum of 150/170 in the quantitative section to apply for graduate studies in any engineering field. A low score (3.5/5 or less) in analytical writing can be looked as a cause of concern.

A GRE score will help a student to gain admission to a good university in case the student has come from an unrecognized school or university. Even with an excellent academic track record, it is difficult to convince the committee with a low GRE score, as it will lead to suspicion regarding the student’s undergraduate college.

* *TOEFL*

TOEFL is an English proficiency test which is also conducted by ETS, is a prerequisite for international students for graduate admissions in the United States. The student must clear the out of score set by the university, but the actual score does not matter. In some cases, when the TOEFL is quite low and the rest of application is strong, the candidate is given admission with a condition to take an extra English course in order to satisfy the master’s degree.

* *CGPA*

Grade represents the academic performance of a student in his previous University. Different University have different grading criteria and different Scale. A University might be more stringent or lenient in awarding grades than other University. The most intuitive way is to calculate GPA as the ratio of grade and topper’s grade. Given the dataset, we can easily calculate this ratio as topper grade is also available to us.

* *Statement of Purpose*

A statement of purpose is where you tell the admissions committee why you're interested in a graduate program, the kind of work you plan on doing when you're there, and why you should be doing that work.

* *Letter of Recommendation*

A letter of recommendation is a letter, memorandum, or online form in which a writer (usually a person in a supervisory role) evaluates the skills, work habits, and achievements of an individual applying for a job, for admission to graduate school

* *Research Experience*

Since the application pool for a top school is generally of very high quality, numerical scores (like GRE, GPA) will in most cases not be enough to set one candidate apart from the other. It is therefore the candidate’s research background that might hold the key to him/her being chosen over someone else. An applicant’s research background can be inferred from the internships and projects done by him/her (mentioned in the SOP and/or resume) and research papers published in journals and conferences.

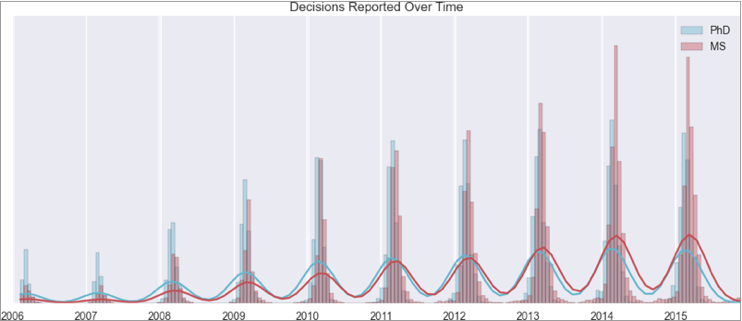
* *University Rating*

This mainly determines the quality of the undergraduate university from which the applicant has graduated. During the admissions process, the graduate admissions board tend to look at the undergraduate university in case there are many applicants which a similar profile. Applicants from a well-known university have an edge over other applicants. This rating scales from 0 to 5.

* *Chance of Admit*

This is the target variable which we are trying to predict. Based off the above-mentioned criterion values, the graduate admissions board comes to a decision of granting an admission to the applicant. Since this is a binary classification problem, the outcome is either 0 or 1.

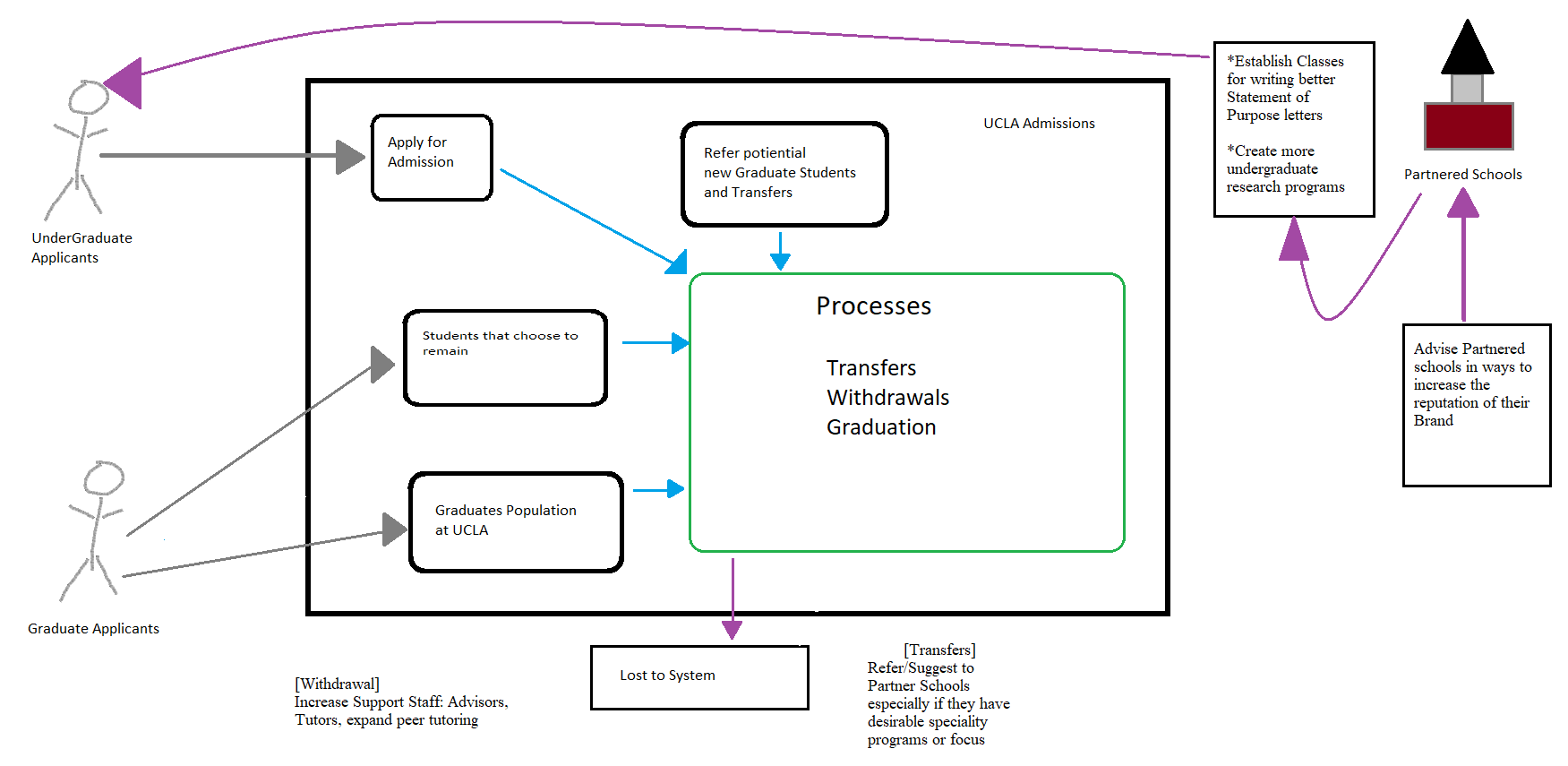
We will visualize the trend in the graduate admissions process, we will make use of a Kernel Density Estimate Plot to show how it has been affecting the decision rates.



From the above plot[11] we observe that MS admissions have drastically increased from 2013.

# business use – case

The dataset which we have chosen is Graduate Admissions dataset from Kaggle[1]. We will build a model which will predict the chance of admittance based on the scores given to GRE, TOEFL, Statement of Purpose, Letter of Recommendation, and the rating of the university to which the student has graduated from. From here, we can figure which among of the features highly influences the student’s chance of admittance to the respective university. Using UCLA Admissions, the business use case[2] we created will allow for the visualization of how the built model can be used by the users[3].The actors in our business use case include “partnered schools,” “undergraduate applicants,” and “graduate students.” There are many use cases in the system that include cases where something is entering the system which are depicted by rounded rectangles, cases that are leaving the system which are depicted by a purple line pointing to “Lost to System,” cases that are happening outside the system which are depicted by sharp rectangles. The arrows show the association between the actor and the use cases. For example, actor “Graduate Applicants” are shown to be able to join use case “Graduates Population at UCLA” and potentially entering the system by having the gray arrow (meaning entering) pointed from the actor to the rounded rectangle use case[4]. From the use case “Graduates Population at UCLA,” the students can remain in the school until graduation, which is shown by the blue arrow (meaning remaining inside) pointing to the rounded rectangle of use cases having “Graduation.” The exit of the system is followed by a purple arrow (meaning exiting or remaining outside) pointing outside the system towards the sharp rectangle use case “Lost to System.” Another example would be actor “Graduate Applicants” that enters the rounded rectangle use case as “Transfers,” but could also mean use case “Transfers” that exits the system as shown by a purple arrow pointing to the sharp rectangle use case, “Lost to System.”



# hypothesis questions

1. Strong SOP & LOR scores have a higher chance of getting a prospective student an admission offer.
2. Research experience matters more than any other criteria for consideration for admission.
3. The rating of the undergraduate university has no impact on whether a student receives an offer of admission to Graduate school.
4. Higher CGPA corresponds with a higher chance of an offer of admission
5. GRE score is a stronger predictor than TOEFL on receiving an offer of admission.

# dataset

The dataset contains several parameters which are considered during the graduate admissions process.

1. GRE Scores (out of 340)
2. TOEFL Scores (out of 120)
3. University Rating (out of 5)
4. Statement of Purpose (out of 5)
5. Letter of Recommendation (out of 5)
6. Undergraduate GPA (out of 10)
7. Research Experience (either 0 or 1)
8. Chance of Admit (ranging from 0 to 1)

*4.1 Pre-processing the data*

Data preprocessing is a data mining technique that involves transforming raw data into understandable format. Data acquired from real-world scenarios will have lot of inconsistent and incomplete features. Data preprocessing is used to resolve such a situation.

The following are the data preprocessing techniques which we have used in our dataset.

1. *Removing Duplicates*

* There were around 400 duplicate values in the merged dataset. These duplicate values lead to overfitting of the model and hence were removed from the merged dataset.

1. *Feature scaling Numerical Data*

* The “Chance of Admit” column was feature scaled into 2 outcomes – 0 and 1 based on the threshold value (which is the mean of chance of admit).

1. *Scaling Columns*

* We have rescaled the data using Normalization for the entire dataset

# demonstrating procedures

Once the dataset has been preprocessed, we apply various machine learning techniques to prove our hypothesis. Once after applying all these models, we choose the best model for each hypothesis by providing relevant details using XLMiner.

1. *Linear Regression*

Linear Regression is a linear method to model the relationship between the independent variables and a dependent variable[7]. Multiple regression is used to predict a relationship between multiple variables. It allows the users to identify if there are correlations between various variables. However, it is not a good choice if a given dataset have a large array of predictor variables.

1. *Logistic Regression*

The idea of Logistic Regression is to find the relationship between the features and probability of particular outcome. Also known as Logit regression, this is a mathematical model used to estimate the probability of an event occurring using previously collected data. Logistic regression is used when the dependent (target) variable is categorical in nature[8].

1. *Decision Trees*

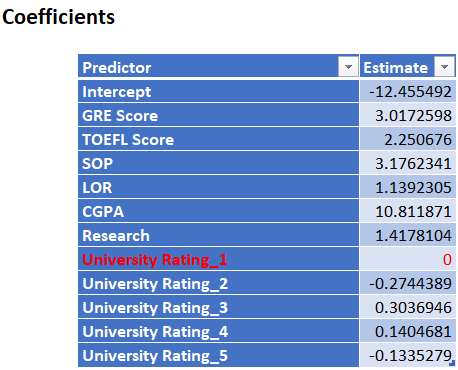
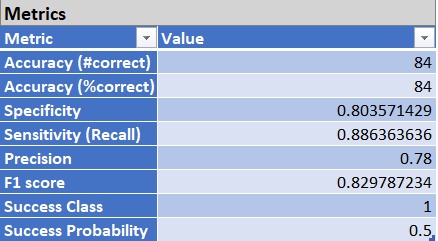
Decision trees is a supervised learning method used for classification and regression. Decision trees learn from the data to approximate the sine curve with the set of if-else rules. The deeper the tree, the more complex the decision tree becomes and better fits the model[9]. Decision trees are used to classify membership of cases or objective in the classes of a categorical dependent variable (i.e. target) based on measurements from the predictor variables. Classification Trees are both flexible and interactive making them a popular choice in data analysis.

# building the model

Now, we have constructed our models and have used them in evaluating our hypotheses.

1. *Logistic Regression:*

Using logistical regression, we will examine the coefficients to see how much “weight” to attribute to each predictor. A higher coefficient value is indicative of that predictor being more important when compared to the other predictors[10].



**HYPOTHESIS 1:** (*Strong SOP & LOR scores have a higher chance of getting an admission offer*)



Looking at this pictures of the coefficients we can see that the SOP value of (3.17) when compared to the LOR value of (1.13) would seem to indicate that the “Statement of Purpose” is more important that the “Letter of Recommendation” that a student would be required to submit with their application. However, taken together one can assert that the hypothesis is true.

**HYPOTHESIS 2:** (*Research experience matters more than any other criteria for Admission*)

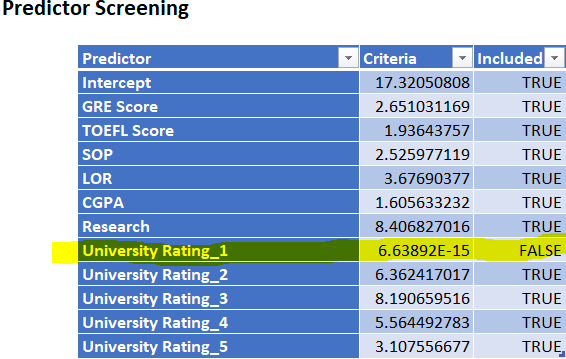
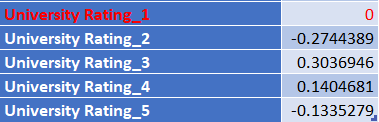


Looking at the value of Research (1.41) we can conclude that we cannot support the claim that research is the single most important predictor when it comes to admissions. While prior research experience has an impact, it is impossible to support the hypothesis as stated, it must be regarded as false.

*Binning***:**

Or Discretization is the process of turning numerical variable into categorical variables. This also allows one to turn a range of numerical values into a discrete “bin” that captures that range. For example, a range of Ages as 18~25 could be put into Age\_1 as a bin.

**HYPOTHESIS 3:** (*The rating of the undergraduate university has no impact on whether a student receives an offer of admission to Graduate school*)



For University rating we created a bin for each rating. Each bin represents an integer for University 1,2,3,4,5. Looking at this chart, we can determine that University ratings of 3 or 4 are the most significant. Bins 2 and 5 have negative values indicating that they have very little impact on the chance of admission to the university.

Looking at the mode of the unaltered dataset we can see that the University rating of 3 is the most common, which might explains why of all the bins it has the highest weight of (0.3036).  
  
Predictor screening is performed automatically by the algorithm, based on its determination, “University Rating\_1” was dropped, hence what we see that is bin has a value of (0).

**HYPOTHESIS 4:** (*A higher CGPA corresponds with a higher chance to be offered a chance of Admittance*)



The rating of CGPA effectively outweighs all the other predictor combined. Which would serve to indicate that is the single most important criterion for being offered a chance of Admission. However, let us examine the behavior of the predictors on the other models before making a judgement. While it seems to support the hypothesis, this metric would suggest we could postulate an even stronger hypothesis.

**HYPOTHESIS 5:** (*GRE predictor provides more insight than TOEFL; TOEFL should be dropped as a predictor*)

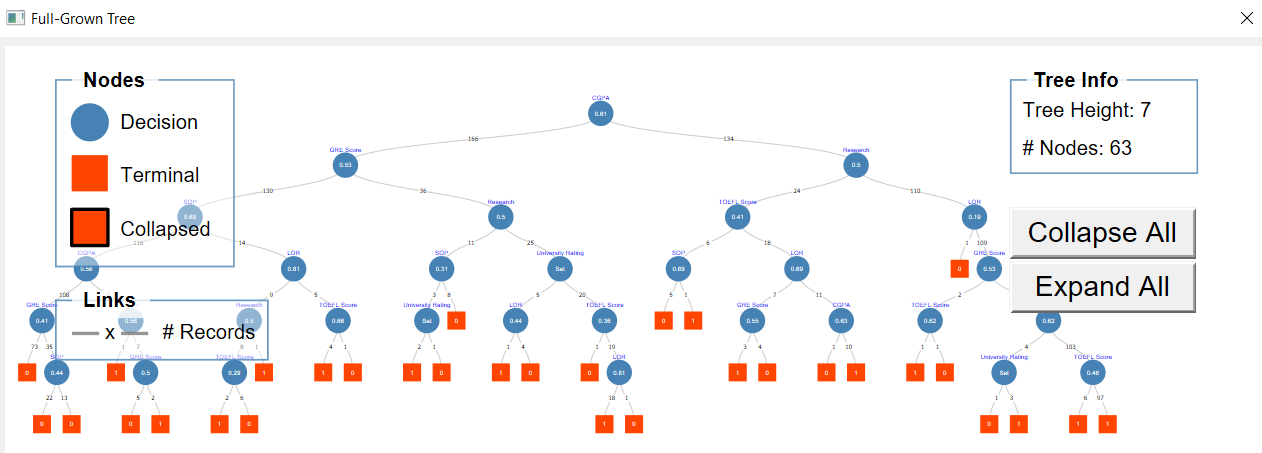


Comparing the values seen here, GRE with (3.017) versus TOEFL (2.250) we can see that both predictors are valuable. We can assert that GRE is the more valuable between the two. However, TOEFL is worth keeping especially considering the already low dimensionality of the data set.

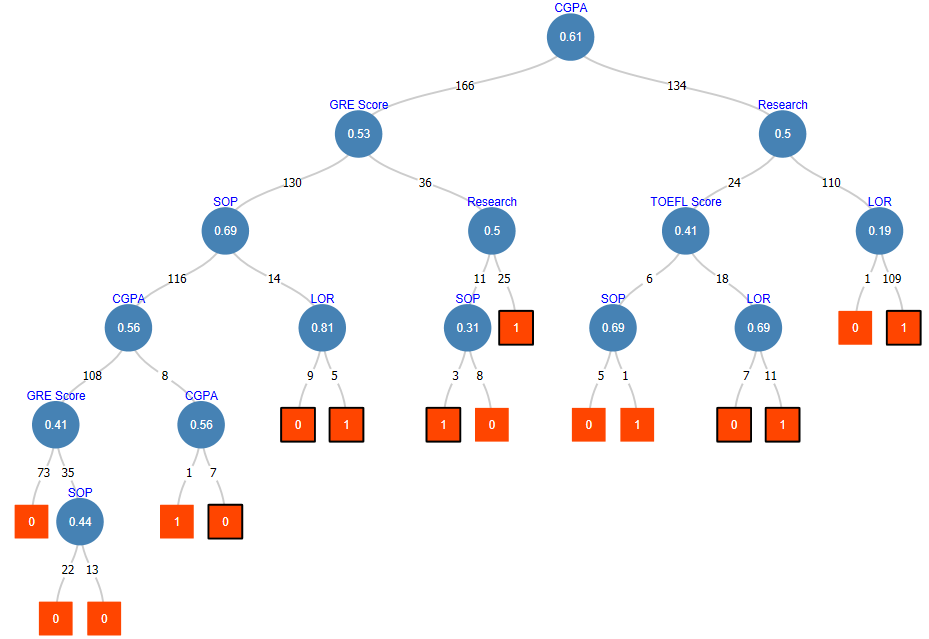
**Conclusion:** Logistic Regression most strongly supports hypothesis 4 (CGPA). The weight of CGPA outweighs all the other predictors combined. This is the best model in support of [H4] being true.

1. *Classification tree*

Below figure represents the decision tree for the dataset.



**HYPOTHESIS 1:** (*Strong SOP & LOR scores have a higher chance of getting an admission offer*)

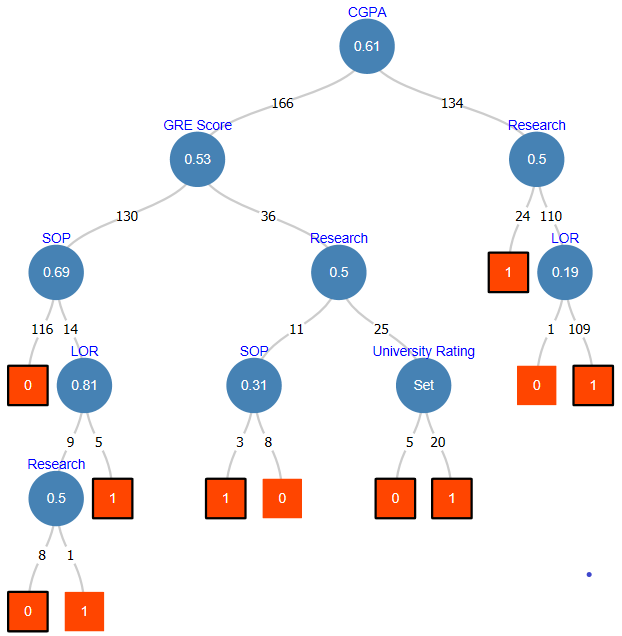


This chart is a partial tree which highlights all the decision points involving SOP and LOR.  
However to evaluate these predictors we will instead use the Feature Importance instead.



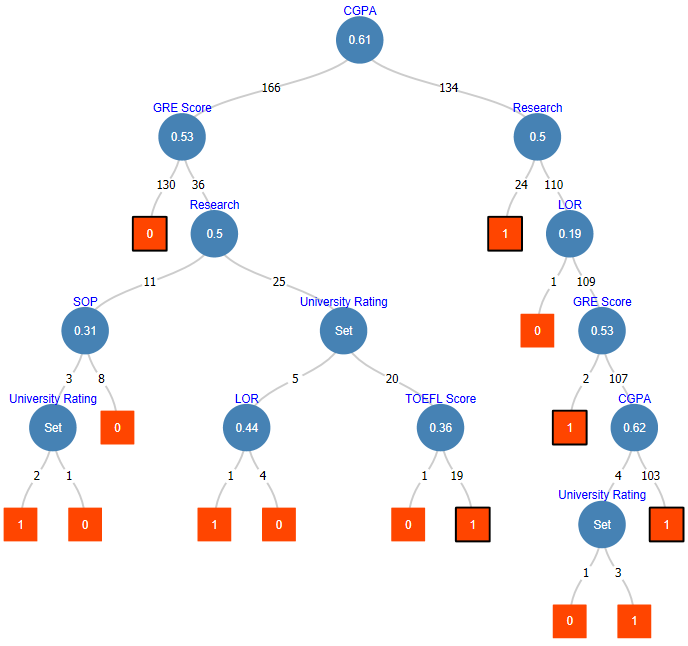
This clip from Feature Importance would indicate that while they have some merit, we would expect them to be rated higher than (.5) to support the claim that they are of significant import when determining a chance to admit. Based on this we can say that while SOP and LOR have a measurable impact, they cannot be said to be of critical importance.

**HYPOTHESIS 2:** (*Research experience matters more than any other criteria for Admission*)



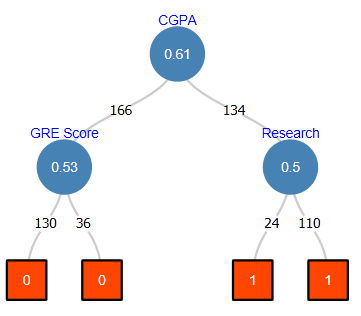
The value of Research (0.35) means we cannot support the claim that research is the most important predictor towards admission. While it has an impact, this model refutes the claim.

**HYPOTHESIS 3:** (*The rating of the undergraduate university has no impact on whether a student receives an offer of admission to Graduate school*)



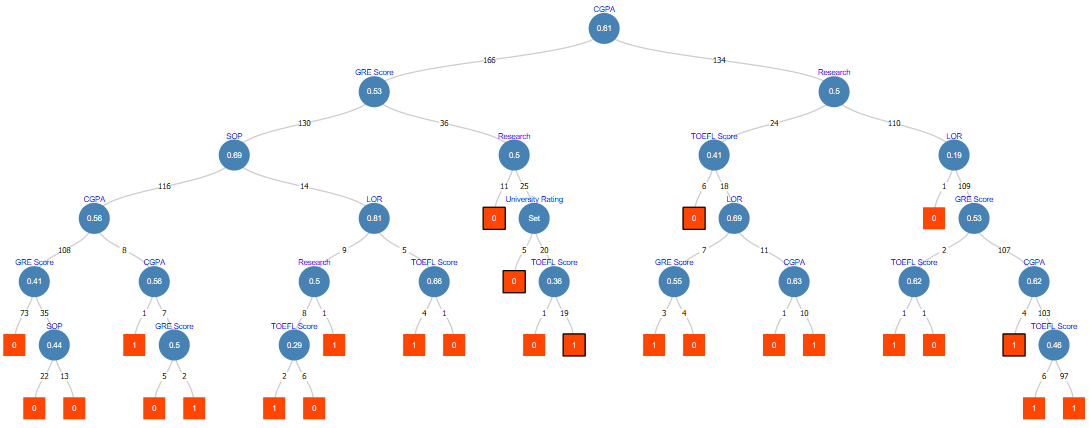
Here we can see that since University rating isn’t 0, it clearly has some effect towards the odds of admittance. This evidence refutes the Hypothesis and instead we must conclude that University rating has some influence on the chance of Admittance.

**HYPOTHESIS 4:** (*A higher CGPA corresponds with a higher chance to be offered a chance of Admittance*)



We can see that with a rating of (0.793) it surpasses our threshold of (0.5), further it has the highest rating of all given predictors, indicating that there is a strong relationship between CGPA and the odds of being admitted. This information lets us uphold the claim that a higher CGPA corresponds to a higher chance of Admittance.

**HYPOTHESIS 5:** (*GRE predictor provides more insight than TOEFL; TOEFL should be dropped as a predictor*)



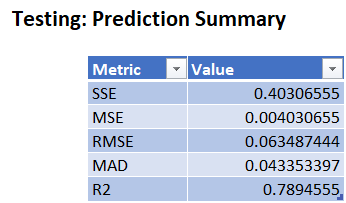
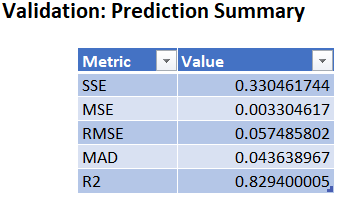
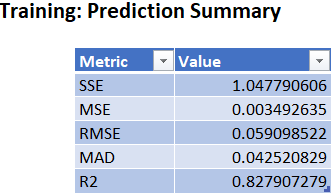
Here we can see that while GRE’s score of (0.563) is greater than our threshold of (0.5). While TOFEL score of (0.39) is significantly below the threshold. Its impact is not insignificant. If we were working with a dataset that had a very large array of predictors, we could decide that TOEFL does not provide any data that isn’t already being provided by GRE. However, since there are only seven predictors in the dataset; dropping one would prove detrimental.

**Conclusion:** Using our threshold value of (.5) we can see that this model most strongly holds [H2] as false. With only a rating of (.35) it is the clearest case of [H2] being held false. The is the Best Model to prove [H2] FALSE.  
In this model we have the predictor for GRE rated as (.56), whereas, TOEFL only has a rating of (.39) this is the strongest case in which the [H5] holds true. This is the Best Model to hold [H5] as TRUE.

1. Linear Regression

For our model we relied on standard linear regression.

As there were several different scales used by the variables it was necessary to first normalize the data, so that all the variables are using the same scale.



**HYPOTHESIS 1:** (*Strong SOP & LOR scores have a higher chance of getting an admission offer*)



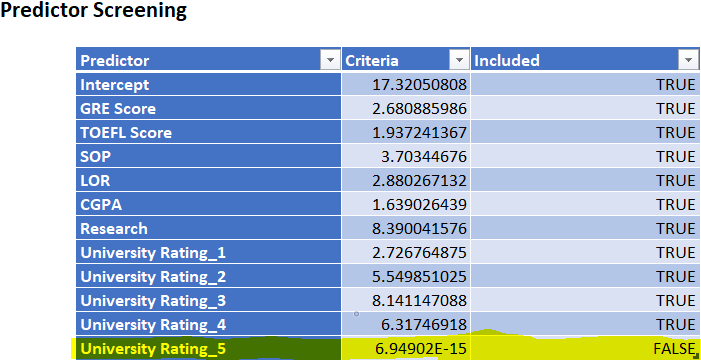
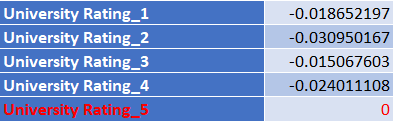
Here we see that SOP has a negative coefficient indicating that it does not seem to have much of an impact on admittance whereas LOR is positive it appear to have a reduced influence when compared to the first two model even accounting for the normalization used in this model. The result is that only LOR seem to have any import meaning that original hypothesis does not hold true in this model.

**HYPOTHESIS 2:** (*Research experience matters more than any other criteria for Admission*)



The value of Research (0.0296) is the smallest non-negative rating on this scale. Thus, according to this model instead of being the most important predictor, we can say it is the least.

**HYPOTHESIS 3:** (*The rating of the undergraduate university has no impact on whether a student receives an offer of admission to Graduate school*)



One can see that again we needed to use binned values. As with the first model we’ve explored we can see that the algorithm we used to model for Linear regression automatically precluded the use of the “University Rating\_5” bin.

We can see that except for “University Rating\_5” all the bins have negative coefficients. This would indicate that the University rating has no bearing on the odds of being admitted, which would allow us to assert that this Hypothesis holds true. Further, one could assert, using this model as a basis, that University Rating should be dropped as a predictor.

**HYPOTHESIS 4:** (*A higher CGPA corresponds with a higher chance to be offered a chance of Admittance*)



Based on this metric CGPA (0.2993) has the largest value in absolute terms compared to all the other predicators on this scale. Which allows us to affirm this hypothesis as true. Further we could go so far as to claim a student’s CGPA is the most important predictor to account for when determining the chance of an Admission offer.

**HYPOTHESIS 5:** (*GRE predictor provides more insight than TOEFL; TOEFL should be dropped as a predictor*)



With a GRE value of (0.0976) compared to TOEFL (0.1086) we cannot claim this hypothesis holds true. Instead base on this model, we must declare it as false and we could declare that both GRE and TOEFL must be kept as valid predictors.

**Conclusion:** In this model SOP (-.011) has a negative weight, whereas, LOR has the smallest non-negative weight of all the predictors. This is the Best Model for proving that [H1] is FALSE.

When University rating was binned all the weights became negative. Further, “University Rating\_5” was simply dropped. This is the Best Model for holding [H3] as TRUE.

# strategic recommendation

Not every feature in dataset in the dataset has equal contribution towards the prediction. Some features may have positive correlation and some features might degrade the results too. Therefore, feature engineering is done to build a model only using important and significant features. This will eliminate the usage of poor features and enhance the results. So, feature engineering can be done for further improvements.

The project can also be further enhanced by checking for bias. More accurate and true results can be obtained if the bias is identified and removed.

The project can be implemented on Microsoft Azure to create a better interface for the non-technical people to use the project for universities.

# conclusion

Based on the findings from the visualization and modeling, the hypotheses has been successfully studied. The following inferences have been drawn:

1. From Linear Regression model, Hypothesis 1 can be best proven wrong. This means that, having strong SOP and LOR scores doesn’t mean there is a higher chance of getting admission.
2. The feature “Research” is not the most important feature to predict whether a student will get admission. We can conclude this best from the decision tree model.
3. We can clearly see from the linear regression model, that University rating has no influence on chance of admit. Thus our 3rd Hypothesis holds true.
4. CGPA and Chance of Admission are highly positively correlated. This means that higher the CGPA, better the chance of admit. Thus, Hypothesis 4 holds true and can be best explained by the Logistic Regression model.
5. GRE score has a greater influence on chance of admit than Toefl score does. We can understand this by looking at the decision trees. Thus, Hypothesis 5 holds true and is strongly supported by the decision tree model.

# references

[1] M. S. Acharya, “Graduate Admission 2,” *Kaggle*, 28-Dec-2018. [Online]. Available: https://www.kaggle.com/mohansacharya/graduate-admissions. [Accessed: 10-Oct-2019].

[2] "Guidelines: Business Use-Case Model", *Sceweb.uhcl.edu*, 2019. [Online]. Available: https://sceweb.uhcl.edu/helm/RationalUnifiedProcess/process/modguide/md\_bucm.htm#Different%20Categories%20of%20Business%20Use%20Cases. [Accessed: 10- Oct- 2019]

[3] "Use Case Template | Business Mentor", *Westernmotodrags.com*, 2019. [Online]. Available: http://westernmotodrags.com/use-case-template. [Accessed: 10- Oct- 2019]

[4] "Use-Case Diagram", *studylib.net*, 2019. [Online]. Available: https://studylib.net/doc/8107761/use-case-diagram. [Accessed: 11- Oct- 2019]

[5] L. Chaou, “Top 16 types of chart in Data Visualization”, *Medium,* 22- Jul-2019. [Online]. Available: <https://towardsdatascience.com/top-16-types-of-chart-in-data-visualization-196a76b54b62> [Accessed: 20-Oct-2019]

[6] “XLMiner Data Visualization”, *Frontline Systems Inc.* [Online]. Available: https://appsource.microsoft.com/en-us/product/office/WA104370643?tab=Overview [Accessed: 21-Oct-2019]

[7] S. Swaminathan, “Linear Regression - Detailed View,” *Medium*, 18-Jan-2019. [Online]. Available: https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86. [Accessed: 27-Oct-2019].

[8] S. Swaminathan, “Logistic Regression - Detailed Overview,” *Medium*, 18-Jan-2019. [Online]. Available: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc. [Accessed: 28-Oct-2019].

[9] A. Chakure, “Decision Tree Classification,” *Medium*, 07-Jul-2019. [Online]. Available: https://towardsdatascience.com/decision-tree-classification-de64fc4d5aac. [Accessed: 30-Oct-2019].

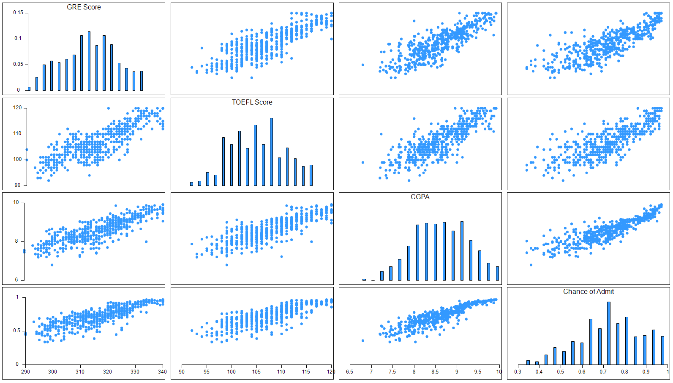
[10] M. B. Editor, “How to Interpret Regression Analysis Results: P-values and Coefficients,” *Minitab Blog*. [Online]. Available: https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients. [Accessed: 03-Nov-2019].

[11] <https://debarghyadas.com/writes/the-grad-school-statistics-we-never-had/>

# appendix

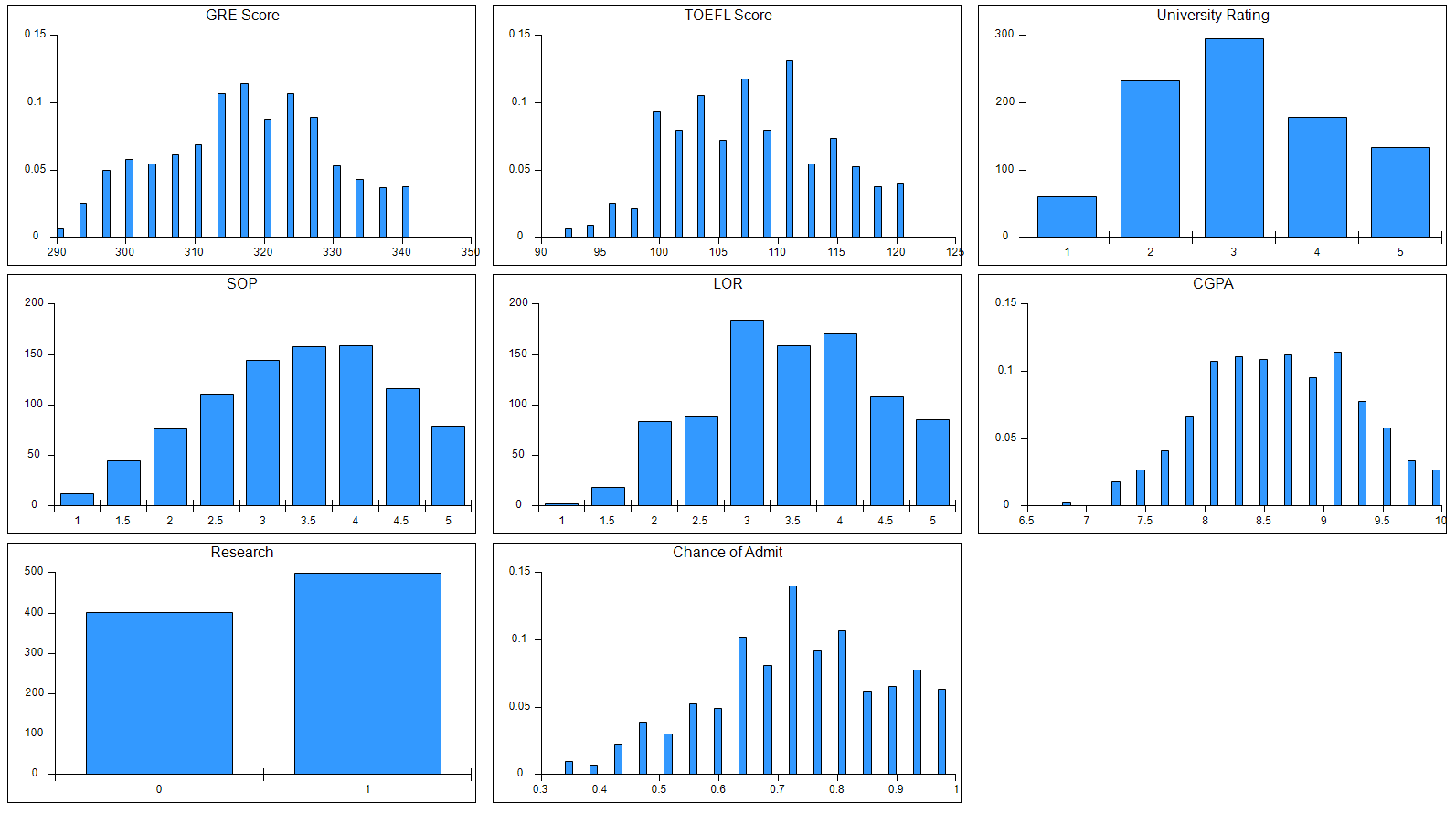
*10.1 Visualizations*

1. *Scatter Plot*



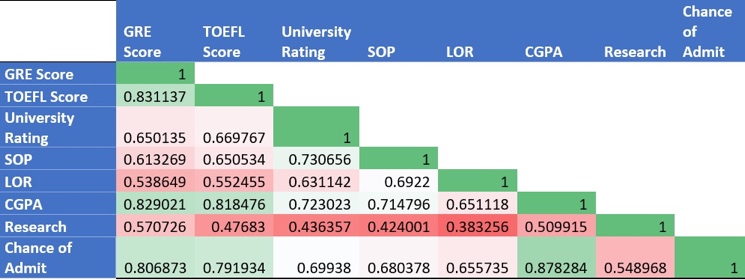
The plot[5] evidently shows that no two variables are negatively correlated, which means that if the value of one variable increases, the value of the other variable also increases. For instance, the chance of admitting increases with higher CGPA which is true in the real-world where a student with higher academic CGPA has a higher chance of admit. The same goes for the GRE and TOEFL scores as well.

1. *Variable Plot*



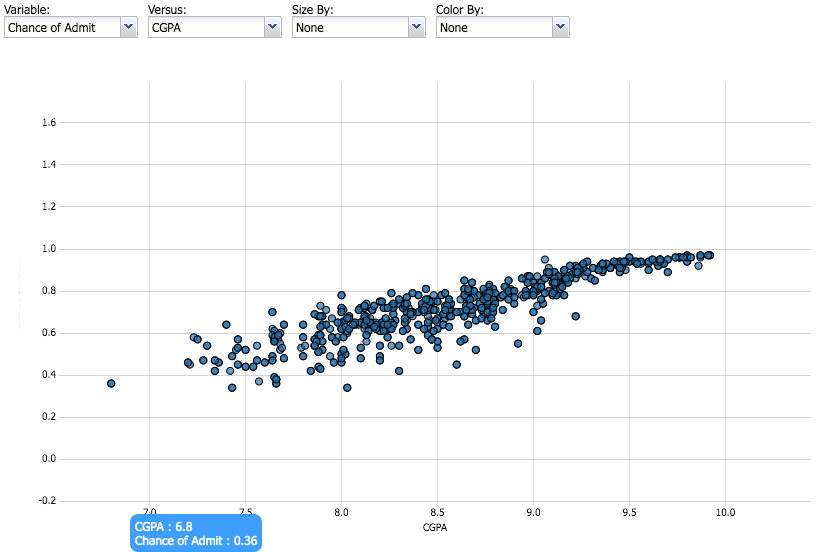
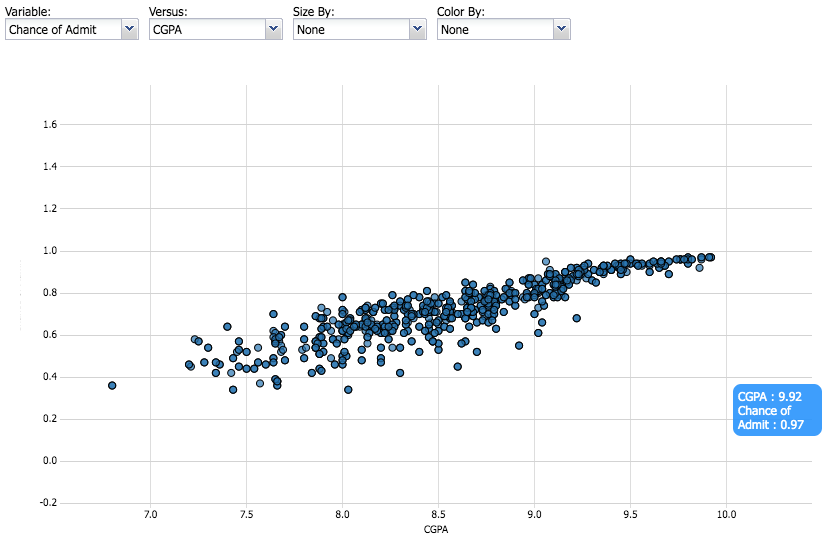
The above variable plot[6] shows the distribution of various attributes in the dataset. It can be inferred that the GRE score and TOEFL score data are normally distributed and SOP data looks left-skewed. It can also be seen that there are almost 300 universities with a rating of 3. Also, there are almost 500 students who have taken part in research activities. It can also be concluded that there are many students with CGPA between 8 and 9.

1. *Correlation Table*



On creating a correlation table[6] for the dataset, it can be seen that ‘Chance of Admit’ and ‘CGPA’ are highly correlated and ‘LOR’ and ‘Research’ are least correlated.

1. *Scatter Plot of CGPA vs Chance of Admit*

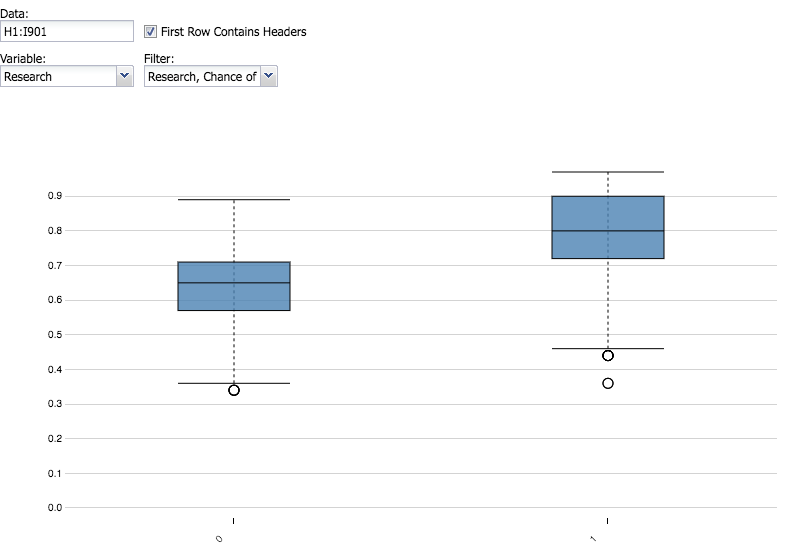


Here we have CGPA on the X-axis and Chance of admit on the Y-axis.

From the above scatterplot[6], two values have been taken from the scatter plot, one which shows that students with high CGPA (9.92) has high Chance of admit (97%) in comparison to the one where low CGPA (6.8) has a lower chance (36%).

From both, correlation matrix[6] and scatterplot we can infer the high positive correlation between the two variables CGPA and Chance of Admit.

1. *Box Plot*



Here, we have plotted a box plot[6] of Research versus Chance of admit. On X-axis we have Research (=0 or 1) and on Y-axis is the Chance of Admit (ranging from 0 to 1).

Here we can observe from the plot[5] that the distribution of box plot for Research=1(right plot) is above that of the left plot. It means that students who have done research have a higher chance of admission as compared to those who have not done any research. From the 3rd quartile of the left plot, we can see that students without research still have about 71% chance of getting an admit. But if we compare medians of both plots, the median of students without Research (left plot) is lower as compared to the median of students with Research (right plot). We can also notice some outliers in both plots (one in left and two in the right plot).