**GRADUATE ADMISSIONS**

Saurabh Burange Akhil Battu Rahul Ratra Mohit Varma

Student ID:801100584 Student ID: 801084290 Student ID: 801100033 Student ID: 801100038

[sburange@uncc.edu](mailto:sburange@uncc.edu) [abattu@uncc.edu](mailto:abattu@uncc.edu%20)  [rratra@uncc.edu](mailto:rratra@uncc.edu) [mvarma@uncc.edu](mailto:mvarma@uncc.edu)

Nisar Kalaria Jaya Sindhura Sadam Sanjay Tamvada Deepak Gorle

Student ID: 801100226 Student ID: 801134228 Student ID: 801091757 Student ID: 801135649

[nkalaria1@uncc.edu](mailto:nkalaria1@uncc.edu) [jsadam@uncc.edu](mailto:jsadam@uncc.edu) [stamvada@uncc.edu](mailto:stamvada@uncc.edu) [dgorle@uncc.edu](mailto:dgorle@uncc.edu)

Narendra Kumar Vankayala Aman Singh

Student ID: 801081957 Student ID: 801085305

[nvankay1@uncc.edu](mailto:nvankay1@uncc.edu) [asingh47@uncc.edu](mailto:asingh47@uncc.edu)

**ABSTRACT:**

Students are often worried about their chance of admission to graduate school. And they sometimes don't have an idea as to what university they should be applying for and what are their chances of admission to a particular university based on the scores they have. Our main goal is to build a model that helps in predicting the chances of admission of a student given his/her scores information.  Additionally, to help students in shortlisting universities with their profiles. The predicted output gives them a fair idea about their admission chances at a particular University. The dataset which we have taken into consideration consist of about 500 students and is taken from Kaggle website that provides us more valuable insights of student’s application such as their LOR’s and SOP’s, GRE and TOEFL score. These factors act as input to the model and based on these inputs, the model performs evaluation and predicts the universities which have a high acceptance rate based on students’ current profile.

**Keywords:**

Data Visualization, XLMiner, Classification, Prediction, KNN.

**1.INTRODUCTION**

On a daily basis, everyone tries to be more productive as we do live in a world of modernization where things tend to alter every minute. Every year the experienced ones and students start thinking for more pragmatic opportunities and they end up in a quandary to make a final decision on American Graduate programs. Furthermore, they commence with searching universities which are great in terms of education, campus activities and university architecture. In order to go in-depth of it, they lack proper guidance as to what should they aim for and what will be a prolific deal for them in the near future. There are a handful of sources available about the universities and their rankings which tends to stagger students. Moreover, they start afresh like every other university and end up in confusion as to what aligns with their interests. This is where complication starts building up and it’s too dicey for the students to make a final decision on it. Application process comprises of things like online test scores, transcripts, letters of recommendation (mostly 3), statement of purpose which includes the career interest of student and their ambitions, research interest, prior experience to the field in what they are applying for etc. Online test score includes GRE which is a compilation of analytical writing, verbal reasoning, quantitative reasoning, and language test scores such as TOEFL and IELTS. Universities then start evaluating students on disparate protocols that students don’t have an idea of and then release decisions. Basic requirements are 3 which are: Graduate record examination, Test of English as Foreign Language and International English Language Testing System. It depends upon university as which language score is taking as per the requirement like some use IELTS and some TOEFL. Now the most integral part comes where students are unaware of the certainty whether they are going to be selected or not; in other terms, they will receive an admission letter or not. This is a very critical situation where students need to take care as to which universities they are applying and which they should not. This includes the application cost as well and huge monetary investment in terms of location and studies. The cost of investment changes as they widen their options on universities. Students need to categorize universities based on investment and their chances of being selected. This goes with the possibility where what universities they should be targeting in order to get the specified result as getting admission is the important part. Basically, students should try to select universities based on where the prospect of admission is safe i.e. where there is more certainty of getting an admit. Multifold admissions help students to choose the university and study more about other universities at the same time. It’s good to have an option than to have only one admit which creates so many complications as in like the cost of investment, finalizing domains and aligning them with their interest. We are using the GRE and TOEFL values as predictors to check how they influence the chances of admit. We are generating visualizations that help in deciding certain factors in deciding the admission which includes research, CGPA. LOR and SOP.

**2. BUSINESS USE-CASE**

The main goal of the project is to prepare a model that will help students in predicting what universities to apply for based on their current scores. As we know tens of thousands of students around the globe apply for their dream universities as per their ambitions. There are about 5000 universities in America that provide varieties of concentrations in disparate domains. Few universities are famous for computer science research programs, few of them hold stronger ground in internships/Co-ops program and all of them have different parameters for providing admissions. Some programs focus more on the research work of students, few are interested in the professional experience and few of them look for their past academic achievements. Hence, while applying to different universities and programs various factors need to be considered such as TOEFL, GRE, IELTS Score, Statement of Purpose, Letter of recommendation, research papers and publications, professional experience and academic achievements, etc. Students need a predictive modeling tool with high accuracy that can help them in deciding and shortlisting their universities. A tool that can take their profile as input and predict the best university as per their profile score. Additionally, it can also help students in knowing what are their chances for a particular university. Considering all this thing in mind we developed a predictive model project that will help in classifying the output variable based on the input predictors. Using this predictive modeling output students will get a high level as to what universities they should be targeting and what are their chances of admission in their desired university.

**3. HYPOTHESIS QUESTIONS**

1. All the predictors are equally important to determine the Chance of Admittance.
2. GRE and TOEFL influence the Chance of Admittance.
3. Research and CGPA influence the Chance of Admittance.
4. SOP and LOR matter influence the Chance of Admittance.
5. University Rating and CGPA influence the Chance of Admittance.

**4. DATASET DESCRIPTION**

The Dataset contains many parameters as criteria upon which the universities select students for admission at the graduate level.

Following is a list and description of each parameter from the dataset –

GRE scores – The Graduate Record Examinations is a standardized test that is an admission requirement in almost all of the universities in the United States. The maximum score is 340.

TOEFL Scores – TOFEL is standardized for English proficiency that is an admission requirement in most of the universities in the United States and the United Kingdom. The maximum score is 120.

University Rating – University ranking academically compared to other universities and ratings from 1 to 5. 1 being the highest, 5 the lowest.

SOP – Statement of Purpose (SOP) is one the most part of graduate admission. It describes the admission committee who you are, why you are applying in that university, your career goals after completing education. Depending upon how strong the SOP is, it rated from 1 to 5. 5 being the highest, 1 the lowest.

LOR – Letter of Recommendation (LOR) is a letter written by someone who can affirm an individual’s work who is being recommended. It is a rating from 1 to 5. Stronger the LOR, the higher the rating number.

CGPA – CGPA is the Cumulative Grade Point Average, it takes into account all the semesters’ GPA to come with an Overall GPA. Each applicant holds a CGPA of his undergraduate degree.

Research - Whether the applicant has any Research Experience or not. 1 denotes positive, 0 denotes negative.

Chance of Admit – The likelihood of an applicant to get a positive decision for the admission considering the ratings in all the described parameters which are rated between 0 and 1.

**4.1 Pre-Processing of Data:**

Pre-processing of the data is a crucial part as the models will apply algorithms to the data, learn from the data and then predict outcomes. The data can contain a lot of noise, redundancy, inaccuracy, etc. and preprocessing the data means getting rid of these malevolent things and make the data close to being impeccable.

Some of the steps that we did while preprocessing the data are listed below:

1. Missing values: -

* We checked the whole data set and can see that there were no missing values.

1. Deleting unnecessary columns: -

* The serial number column which was not necessary for any modeling was removed from the data.

1. Removing outliers from the data: -

* Using box plots we identified 3 outliers in the data. We deleted the columns containing the outliers. The resultant data now had 497 columns.

1. Checking duplicate data: -

* A check was done to identify any duplicates/redundant data in the data set.

1. Creating dummy variables: -

* Dummy variables have been creating for the categorical variables in the data sets. This includes features like LOR, SOP and University Rating. The Research field is already in a binary form.

**5. DEMONSTRATING PROCEDURES**

We can apply different models in order to forecast and predict the information on pre-processed data to the dataset. All the hypothesis questions are being covered by utilizing and demonstrating procedures. Below are the following model techniques:

**1: Regression:** Regression is a statistical gauging method that attempts to ascertain a relationship between a variable with other sets of altering variables. It helps in defining the variable nature while other variables are modified in the operation.

**Advantages of Regression:**

1. Regression works great if the relationship between the covariates and response variable is linear.

**Disadvantages of Regression:**

1. It’s great for data analysis but it oversimplifies real world problems and shouldn’t be used for practical problems.

**2: K Nearest Neighbor:**

K nearest neighbor is the most basic but important classification algorithm in Predictive modeling. It is one of the supervised machine learning techniques and helps in pattern recognition, data mining and intrusion detection. KNN algorithm works on a phenomenon that assumes that similar things exist in close proximity. Or we can also say that similar things are nearer to each other.

**Advantages of KNN:**

1. KNN is robust to training data that is noisy
2. KNN works efficiently with large data

**Disadvantages of KNN:**

1. Choosing the value of K (nearest neighbor) may be tricky.
2. The testing stage is computationally expensive.
3. KNN is sensitive to noisy data, missing values, and outliers.

**3. Decision Tree:**

Decision tree works in a way such that classification and regression models are built in the form of a tree structure. The way it works is it breaks down a dataset into smaller chunks and that smaller chunks in the meanwhile are incrementally developed into a tree. They belong to a family of supervised learning in which they gain knowledge to classify and regress. The final result is a tree which contains selection nodes and leaf. The main intention behind creating a decision tree is that it helps in predicting the value of a goal variable, which makes it easy to make a decision.

**Advantages of Decision Tree:**

* 1. Simple and easy to interpret.
  2. Decision tree help to determine the best, worst and expected values.

**Disadvantages of Decision Tree:**

* + - * 1. Calculations may get complex.
        2. Small change in data can make a substantial difference in the overall tree

**6. BUILDING MODELS**

**Building Models:**

In this section, we build various models to analyze the hypothesis and get a result of it.

**Hypothesis 1:**

*All the predictors are equally important to determine the Chance of Admittance.*

To analyze this hypothesis, we use Multiple Linear Regression, K Nearest Neighbors and Regression Tree. The input variables or predictor variables are GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, Research, and the output variables are Chance of Admit. The data is pre-processed and dummies are created for SOP, LOR, University Rating and to cut down the redundancies we have removed University Rating\_5, SOP\_5, and LOR\_5 from the data. We have also partitioned the data into training and validation datasets in the ratio of 60% and 40%.

**Model 1: Multiple Linear Regression**

Multiple Linear Regression (MLR) model is used when we have two or more predictor variables for one single output variable.

The Multiple Linear Regression is run on the partitioned data with the input variables GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, Research and Chance of Admit as the output variable. We have used the best subset feature selection and below is the output of the best subsets we get after running the model.

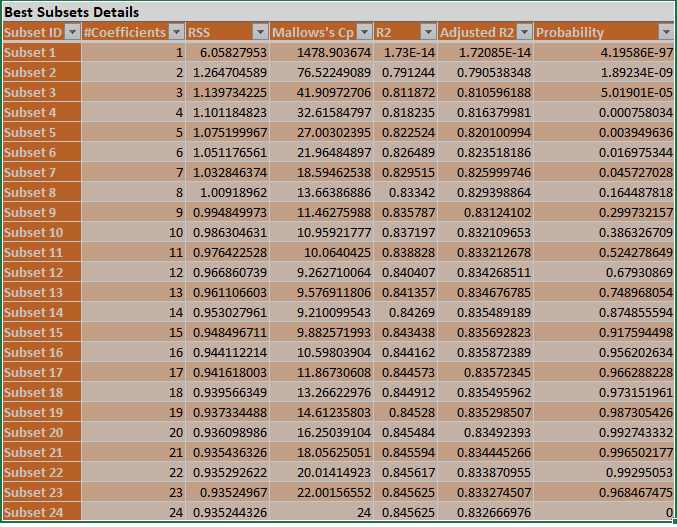


Fig 6.1

From the above fig 6.1, we choose subset 24 as the best subset as it has high R2 value and its Mallow’s Cp value is one more than that of the number of predictors which indicates that the subset is a better fit of the model. Multiple Linear Regression is again run on the subset 24 and the results of it are as follows.

A screenshot of a cell phone

Description automatically generated

Fig 6.3

A screenshot of a cell phone

Description automatically generated

Fig 6.4

The training and validation data are shown in the above pictures. The R2 values for training and validation are 0.845 and 0.776 respectively.

A screenshot of a cell phone

Description automatically generated

Fig 6.2

From the coefficient analysis table, Fig 6.2, based on the p-value we can say that CGPA, LOR\_1.5, Research, TOEFL Score have a great influence on predicting the chance of admittance because of their low p-value.

**Model 2: K Nearest Neighbors**

KNN Algorithm is used to classify or predict the values of the outcome variable. To classify or predict the records, the method relies on finding similar records that are trained with the training data and classify them accordingly. Here KNN is run using the same input variables as chosen in the above model and the K value is chosen as 10. The output of the model is shown below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Training: Prediction Summary** | | | |
|  |  | |  |
|  | **Metric** | | **Value** |
|  | SSE | | 0 |
|  | MSE | | 0 |
|  | RMSE | | 0 |
|  | MAD | | 0 |
|  | R2 | | 1 |
| Fig 6.5  **Validation: Prediction Summary** | | | | |
|  |  |  | | |
|  | **Metric** | **Value** | | |
|  | SSE | 1.092659403 | | |
|  | MSE | 0.005490751 | | |
|  | RMSE | 0.0740996 | | |
|  | MAD | 0.055713032 | | |
|  | R2 | 0.68717643 | | |

Fig 6.6

We got the best value for K as 8 as we chose the search for the best option in XLMiner. We can see that from the below table:

A screenshot of a cell phone

Description automatically generated

Fig 6.7

**Model 3: Regression Tree**

We use a Regression Tree when the output of the model is a continuous variable. The input and output variables are the same as mentioned in the above models. Using the tree we can make decisions using the rules. The regression tree is run on the partitioned data training (60%) and validation (40%).

The best-pruned tree is chosen to score the validation set and a full-grown tree is chosen as the tree to be shown. The tree generated by this model is shown below:

A close up of a map

Description automatically generated

Fig 6.8

From the above tree in Fig 6.8, we can say that CGPA, TOEFL Score and GRE Score are the top factors that influence the chance of admission.

A screenshot of a cell phone

Description automatically generated

Fig 6.9

A screenshot of a cell phone

Description automatically generated

Fig 6.10

**Hypothesis 1: Best Model**

We have run Multiple Linear Regression, K Nearest Neighbor and Regression Tree models and found that the error value for Multiple Linear Regression is least and also the R2 value is highest. We can see that in *Validation: Prediction summary* of MLR RMSE value is 0.062 which is less when compared to both KNN and Regression Trees. So by these, we can say that Multiple Linear Regression is the best model for this Hypothesis.

We can say that our hypothesis is not true, as we can see from figure 6.2 CGPA has the highest influence when compared to other predictors. This is because of its low p-value and high value for the coefficient.

**Hypothesis 2:**

*GRE and TOEFL influence the chance of admittance.*

To analyze this hypothesis, we use Multiple Linear Regression, KNN, and Regression Tree models. After the data is preprocessed, we use GRE Score and TOEFL Score as predictor variables and the Chance of Admit as the single outcome variable. We partition the data into 60% training data and 40% validation data.

**Model 1: Regression Tree**

In this model, we use the GRE Score and TOEFL Score as the only two predictor variables and Chance of Admit as the outcome variable. We partition the data in a 60-40 ratio into training and validation datasets. We select the best-pruned tree which prunes the validation dataset and we choose Full Grown Tree to visualize the tree. The tree generated is shown below:

A screenshot of a social media post

Description automatically generated

Fig 6.11

In the above tree Fig 6.11, we can say that the splits are generated by using only the GRE Score and TOEFL Score. The results of the model are shown below:

A screenshot of a cell phone

Description automatically generated

Fig 6.12

A screenshot of a cell phone

Description automatically generated

Fig 6.13

The prediction summaries for both training and validation datasets are seen above.

**Model 2: K Nearest Neighbor**

In this model, we use input and output variables similar to the previous model. We also put the value of K equal to 10 and run the model. We also partition the data into training and validation datasets in 60% and 40% ratios. The results of the model are shown below:

|  |  |  |
| --- | --- | --- |
| **Training: Prediction Summary** | | |
|  |  |  |
|  | **Metric** | **Value** |
|  | SSE | 0.56665 |
|  | MSE | 0.00190151 |
|  | RMSE | 0.043606308 |
|  | MAD | 0.023048098 |
|  | R2 | 0.906466845 |

Fig 6.14

|  |  |  |
| --- | --- | --- |
| **Validation: Prediction Summary** | | |
|  |  |  |
|  | **Metric** | **Value** |
|  | SSE | 1.790649621 |
|  | MSE | 0.008998239 |
|  | RMSE | 0.09485905 |
|  | MAD | 0.071721085 |
|  | R2 | 0.487344908 |

Fig 6.15

We choose to search for the best value of K while running KNN, so for this hypothesis, we get the best value of K as 9. This can be seen from the above table.

A screenshot of a cell phone

Description automatically generated

Fig 5.16

Model 3: Multiple Linear Regression

We run the Multiple Linear Regression using GRE Score and TOEFL Score as predictors and Chance of Admit as the outcome variable. We choose the best subset feature selection and run the model on the partitioned data 60% training and 40% validation. The subsets generated from this method are shown below:

A screenshot of a social media post

Description automatically generated

6.17

From this best subset details, we have to choose the best subset with high R2 value and Mallow’s Cp value which is one greater than the number of predictors. By looking at the above table we choose subset 3 which seems to be the best option. Selecting the subset and running the regression again on the data we get the final results of the model which should be the better fit of the model. The results are shown below:

A screenshot of a cell phone

Description automatically generated

Fig 6.18

A screenshot of a cell phone

Description automatically generated

Fig 6.19

From the coefficients table shown below, we have to find the predictor which has the most influence on the dataset based on its p-value.

A screenshot of a social media post

Description automatically generated

Fig 6.20

From the above table, Fig 6.20, we can say that the GRE Score has a higher influence than TOEFL Score when compared because of its very low p-value.

**Hypothesis 2: Best Model**

From the results of 3 models we ran for Hypothesis 2, we have to choose the best model which has a low error rate and high R2 value. By observing the values we can say that Multiple linear Regression has the lowest RMSE, SSE and MSE values and highest R2 value of 0.684. So we consider Multiple Linear Regression as the best model for Hypothesis 2.

We can say that our hypothesis is true, as both of them have positive coefficients in the Multiple Linear Regression. From the coefficient table in figure 6.20, we can say that the p-value of GRE is lower than TOEFL, so GRE has higher influence when compared to the later.

**Hypothesis 3:**

*Research and CGPA influence the Chance of Admittance.*

**Model 1: Multiple Linear Regression**

In this model, we use Research and CGPA as the two main predictors and Chance of Admit as the output variable. We have to choose the best subset feature selection and run the model on the partitioned data and we get the best subsets as:

A screenshot of a social media post

Description automatically generated

Fig 6.21

From the above table, Fig 6.21, we choose subset 3 as the best subset because of its high R2 value and Mallow’s Cp value which equals one more than the predictors. We then run the regression again on subset 3 to get a better fit model. The output of the resultant model is shown below:

A screenshot of a cell phone

Description automatically generated

Fig 6.22

A screenshot of a cell phone

Description automatically generated

Fig 6.23

From the coefficient analysis, we can say which predictor has a higher influence on the Chance of Admit.

A screenshot of a social media post

Description automatically generated

Fig 6.24

From the above table, by looking at the values in the coefficient table, we can say that CGPA has a higher influence than Research because of its low p-value.

**Model 2: Regression Tree**

We run the Regression Tree using Research and CGPA as the two predictors and Chance of admit as the output variable. We choose a best-pruned tree to score the validation dataset and visualize the full-grown tree to get the results. The tree generated is shown below:

A close up of a map

Description automatically generated

The nodes University Rating and CGPA are split across the tree to give the rules that determine the future data whether the candidate gets an admit or not.

A screenshot of a cell phone

Description automatically generated

Fig 6.25

A screenshot of a cell phone

Description automatically generated

Fig 6.26

From the above tables, we can see the prediction summary of the training and validation datasets.

**Model 3: K Nearest Neighbor**

We use University Rating and CGPA as the predictor variables and Chance of Admit as the outcome variable for this model. We also select the K value for this as 10 and choose to search for best K. The data is partitioned into training and validation datasets and then the model is built on the training dataset and tested on the validation dataset. The output of the above model is shown below:

|  |  |  |
| --- | --- | --- |
| **Training: Prediction Summary** | | |
|  |  |  |
|  | **Metric** | **Value** |
|  | SSE | 0.441028333 |
|  | MSE | 0.001479961 |
|  | RMSE | 0.038470259 |
|  | MAD | 0.021697987 |
|  | R2 | 0.92720238 |

Fig 6.27

|  |  |  |
| --- | --- | --- |
| **Validation: Prediction Summary** | | |
|  |  |  |
|  | **Metric** | **Value** |
|  | SSE | 1.188825631 |
|  | MSE | 0.005973998 |
|  | RMSE | 0.077291643 |
|  | MAD | 0.057454148 |
|  | R2 | 0.659644463 |

Fig 6.28

From the output, we get the best K value as 10 which has the least RMSE value. This can be seen from the figure below:

A screenshot of a cell phone

Description automatically generated

Fig 6.29

**Hypothesis 3: Best Model**

Comparing the above 3 models i.e., MLR, KNN and Regression Tree we have to choose a model that has a low error rate and high R2 value. From the above observations, we can say the error rate of KNN for Hypothesis 3 is highest among others, so it cannot be a good model. Comparing MLR and Regression Tree we can say that both the models have a very small difference in their error rates as well as the R2 values. Though Multiple Linear Regression has a slight edge over Regression Tree in terms of values and hence it is considered as the best model.

We can say that both Research and CGPA influence the Chance of Admittance, which makes the hypothesis true. From the coefficient table, we can see that both have positive coefficients and also CGPA has better influence than Research based on their p-value from Fig 6.24

**Hypothesis 4:**

*SOP and LOR influence the Chance of Admittance.*

**Model 1: Multiple Linear Regression**

We use SOP and LOR as the predictors and Chance of Admit as the output variables. We create dummies for SOP and LOR as they are categorical values and also to reduce the redundancies, we remove the Best subsets is chosen in the feature selection and run the model on the partitioned data. After running the model, we get the best subsets as shown below:

A screenshot of a cell phone

Description automatically generated

Fig 6.30

From this table, Fig 6.30, we choose subset 16 as it has high R2 value and Cp value is one more than predictors i.e., 15+1 =16. After running the model on the 16th subset, we get the following results.

A screenshot of a cell phone

Description automatically generated

Fig 6.31

A screenshot of a cell phone

Description automatically generated

Fig 6.32

From the coefficients table, we can get a better predictor among the SOP and LOR.

A screenshot of a cell phone

Description automatically generated

Fig 6.33

From the above table, Fig 6.33 we can say that SOP\_1.5, SOP\_2 have a relatively high influence on the Chance of Admit. So, if a candidate’s SOP is in the range 1.5 or 2 then his Chance of Admit is determined very accurately.

**Model 2: Regression Tree**

This model uses SOP and LOR as the predictor variables. As these are categorical variables, we generate dummies for each of these and also to reduce the redundancies we remove SOP\_5 and LOR\_5 from the dataset. Then we choose Best Pruned Tree to score the validation data and visualize the fully-grown tree. The tree generated is shown below:

A screenshot of a map

Description automatically generated

Fig 6.34

The validation and training prediction summaries are shown below:

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Fig 6.35

**Model 3: K Nearest Neighbor**

This model uses the same input and output variables as the above models. The data is partitioned into training and validation datasets. We choose the value of K as 10 and opted for the option of choosing the best value for K. The output of the model is shown below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Training: Prediction Summary** | | | | | |
|  | |  | |  | |
|  | | **Metric** | | **Value** | |
|  | | SSE | | 2.067910787 | |
|  | | MSE | | 0.006939298 | |
|  | | RMSE | | 0.083302449 | |
|  | | MAD | | 0.064817493 | |
|  | | R2 | | 0.658663689 | |
| **Validation: Prediction Summary** | | | | |
|  |  | |  | |
|  | **Metric** | | **Value** | |
|  | SSE | | 2.49811685 | |
|  | MSE | | 0.012553351 | |
|  | RMSE | | 0.112041738 | |
|  | MAD | | 0.084204801 | |
|  | R2 | | 0.28480016 | |

Fig 6.36

We can say that the best value for K is set to 9 as it has the least RMSE error value, which can be seen from the figure below:

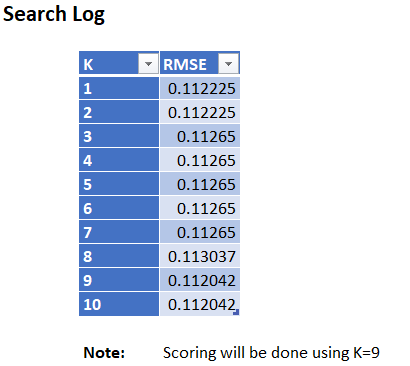


Fig 6.37

**Hypothesis 4: Best Model**

We have run MLR, KNN and Regression Tree for Hypothesis 4, we have to choose a model that has a lower error rate and high R2 value. From the above models, we can see that Multiple Linear Regression has a low error rate and best R2 value when compared to others. We can conclude that Multiple Linear Regression is the best model among the three.

We can say that both SOP and LOR have an influence on the Chance of Admit. From the coefficient table SOP and LOR doesn’t have any null values, so the hypothesis is true, this can be seen in Fig. 6.33.

**Hypothesis 5:**

*University Rating and CGPA influence the Chance of Admittance.*

**Model 1: Multiple Linear Regression**

We use University Rating and CGPA as predictors for the model while Chance of Admit being the outcome variable. We preprocess the University Rating and create dummies for it and also to reduce the redundancies we remove University Rating\_5 from the dataset. Now we run the model choosing the best subsets in the feature selection. The best subsets we get from the table are shown below:

A screenshot of a social media post

Description automatically generated

Fig 6.38

From the above best subsets, we choose subset 6 as it has the highest R2 value and Mallow’s Cp value is 1 more than that of the number of predictors. We select the subset 6 and run the Multiple Linear Regression again to get the best fit of the model with subset 6. The results are as following:

A screenshot of a cell phone

Description automatically generated

Fig 6.39

A screenshot of a cell phone

Description automatically generated

Fig 6.40

From the coefficients, we can get the predictor which best influences the output of the Model.

A screenshot of a social media post

Description automatically generated

Fig 6.41

From the above table, Fig 6.41 we can say that CGPA has higher influence when compared to University Rating as it has a low p-value than any other variable.

**Model 2: Regression Tree**

We use University Rating and CGPA as the predictor variables and Chance of Admit as the output variable. The University Rating column’s dummies are created and also redundant values are removed from the dataset. The data is partitioned into training and validation datasets. The tree generated is shown below:

A screenshot of a map

Description automatically generated

Fig 6.42

The training and validation prediction scores for this model are shown below:

A screenshot of a cell phone

Description automatically generated

Fig 6.43

A screenshot of a cell phone

Description automatically generated

Fig 6.44

**Model 3: K Nearest Neighbor**

In this model, we use the input and output variables similar to those of the above models. After partitioning the data into training and validation data. We set the value of K to 10 and choose the option to choose for best K which should give us the best K value for the dataset. The output of the model can be seen below:

|  |  |  |
| --- | --- | --- |
| **Training: Prediction Summary** | | |
|  |  |  |
|  | **Metric** | **Value** |
|  | SSE | 0.315975 |
|  | MSE | 0.001060319 |
|  | RMSE | 0.032562537 |
|  | MAD | 0.015710291 |
|  | R2 | 0.947844104 |

|  |  |  |
| --- | --- | --- |
| **Validation: Prediction Summary** | | |
|  |  |  |
|  | **Metric** | **Value** |
|  | SSE | 1.227438859 |
|  | MSE | 0.006168034 |
|  | RMSE | 0.078536835 |
|  | MAD | 0.05949466 |
|  | R2 | 0.648589666 |

Fig 6.45

From the output, we can see that the K value for the dataset is set to 10 which has the least RMSE value. By using that value we train and validate the data to get prediction summaries for the same and decide the feasibility of the model. The K value can be seen from the below figure:

A screenshot of a cell phone

Description automatically generated

Fig 6.46

**Hypothesis 5: Best Model**

Comparing the three models, we have to choose a model with a low error rate and high R2 value. As higher R2 values give better fit models, we will be choosing a model with higher R2 values. By looking at the values we can say that SSE, MSE and RMSE values of Multiple Linear Regression are low compared to others and the R2 value of Multiple Linear Regression is highest, so we can say that the best model for hypothesis 5 is Multiple Linear Regression.

From the results of the best model, we can say that both the CGPA and University Rating have an influence on the Chance of Admit. We can also say that CGPA has a higher influence on the coefficient value and p-value from figure, you can see this at 6.41.

**7. STRATEGIC RECOMMENDATION**

The strategic recommendation means what are the roadmaps down the line to make your model more robust more accurate and more flexible. There's always scope for improvement in anything you develop.

Hence as per our best knowledge, we have discussed below mentioned best practices that should be considered in the upcoming models.

1. Observe the original data and do the preprocessing like removing null, outliers, duplicate data, and unnecessary columns.
2. The selection of the dependent variable should be chosen mostly by considering the confusion matrix business use cases and hypotheses as we did.
3. Different visualizations across the attributes play an important role in determining the model accurately and most impact of variables like one we did for Research and CGPA as the two predictors and Chance of admit as the output variable. More the visualization plots better the Model.
4. Out of all the model build the one with less error rate and best subsets should be chosen, viz in our case, it was *Multiple linear regression* model which was most accurate with minimum error.
5. We should make sure that the model we developed is par excellence and is really performing outstandingly on the training data set with fewer errors and more focused on accuracy. Additionally, it shouldn’t be suffering from high variance. On the other hand, we should also make sure that both the test and training data set should not be highly biased. And if either of the cases persists in then we should do the following remedies mentioned below.
6. Add more data: This is one of the most important steps to make sure there are more precise results. if we can afford to have more data set it helps in reducing the variance and making the test and training data set less biased.
7. More Feature: This is certainly a good idea to always provide more features to consumers. As this adds to more flexibility and increase the overall usability and increase the competency of the product.
8. Have different models: We should always consistently look for different models as per our requirement. As the requirement varies from time to time. And it’s always good to think out of the box sometimes. For example, changing from linear regression to logistics or to neural model. As we know that some model is more suitable to some data sets while we must continuously monitor the performance and evaluate the market needs.

Note: It should always be kept in mind that model accuracy is not the only objective we should focus on. It’s harder sometimes to implement few models with high accuracy in production and so one should always look for something simpler lightweight and more productive algorithms

**8. CONCLUSION**

1. The main aim of this project is to predict the admit rate using the input variables such as GRE score, TOEFL score, University, SOP, LOR, CGPA, and research. The above hypothesis is evaluated based on the above modeling and visualizations done in XLminer and Python.
2. The highest chance of getting an admission for GRE and TOEFL score is 97.465% where the maximum GRE and TOEFL scores are 340 and 120. The least chance of getting an admission for GRE and TOEFL score is 42.5437232% where the minimum GRE and TOEFL scores are 290 and 92.
3. The highest chance of getting an admission for Research and CGPA is 99.47707558% where the max values are 1 and 9.92. The least chance of getting an admission for Research and CGPA is 36.9130486% where the min Research and CGPA are 6.8 and 0.
4. The highest chance of getting an admission for SOP and LOR is 0.9047598 where the max values are 5 and 5. The least chance of getting an admission for SOP and LOR is 0.68676025 where the min values are 1 and 1.
5. These are the hypothesis we get regarding the admission rate where a student can calculate his chance to admit and their requirements for their admission. Based on the above models which we had done here we used Multiple linear regression to calculate and is the best regression because the value of r^2 is greater than KNN and regression tree and the error rate is less compared to the other two.
6. The overall outcome of our analysis is to help the candidates to understand the minimum requirements to get into a university-based on their profile.

**9. REFERENCES**

[1]<https://www.kaggle.com/mohansacharya/graduate-admissions>

[2] Data Mining for business Analysis- Galit Shmueli, Peter C. Bruce, Mia L Stephens, Nitin R Patel

[3]<https://www.solver.com/k-nearest-neighbors-k-nn-classification-intro>

[4]<https://www.coursera.org/lecture/predictive-modeling-analytics/7-how-to-build-a-model-using-xlminer-fM9fY>

[5]<https://www.solver.com/using-multiple-linear-regression>

[6]<https://www.machinelearningapplications.com/classification/>

[7]<https://www.datasciencecentral.com/profiles/blogs/introduction-to-classification-regression-trees-cart>

[8]<https://towardsdatascience.com/graduate-admission-prediction-using-machine-learning-8e09ba1af359>

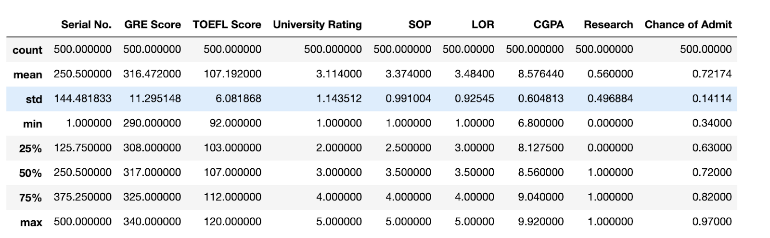
[9]<https://seaborn.pydata.org/generated/seaborn.heatmap.html>

[10]<https://seaborn.pydata.org/generated/seaborn.boxplot.html>

[11]<https://seaborn.pydata.org/generated/seaborn.violinplot.html>

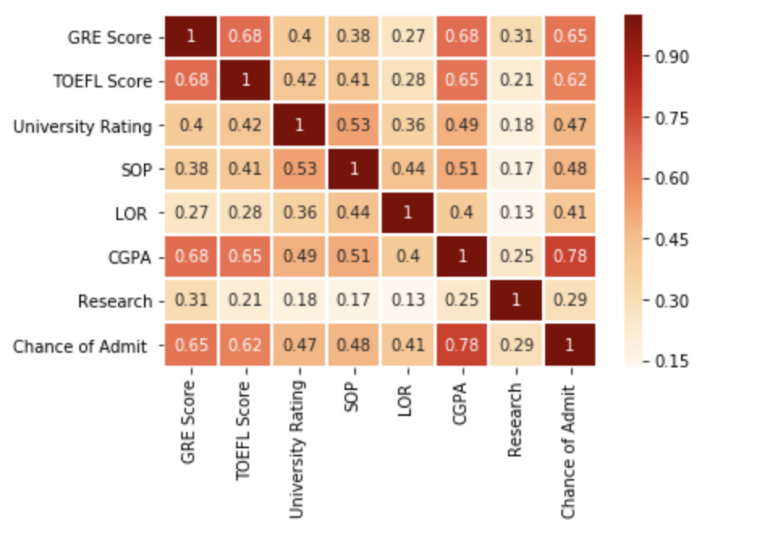
**10. APPENDIX**

**10.1 Description of data:**

****

The table above gives us a brief description of the data that is with us.

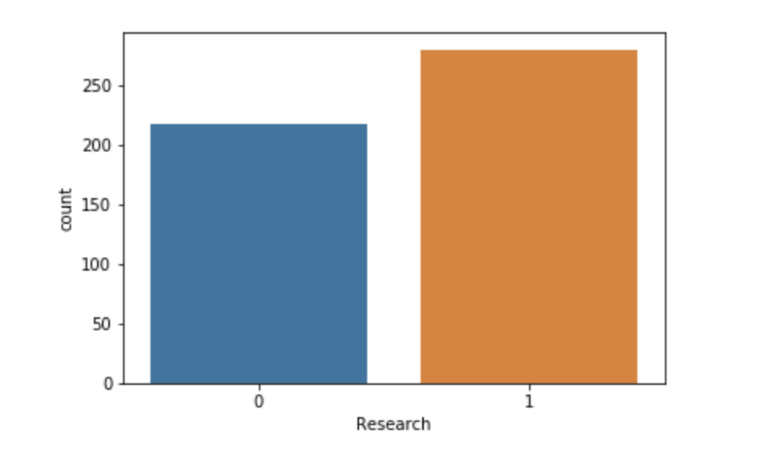
**10.2 Correlation matrix:**

****

The correlation matrix above shows the type of relationship every pair of features in the data.

We can conclude that CGPA and the chance of admission is the most highly correlated pair. GRE & TOEFL Score, GRE Score & chance of admit, GRE Score & CGPA, TOEFL Score & CGPA are some of the highly correlated pairs as well.

**10.3 Comparing the count of universities on the basis of research experience:**

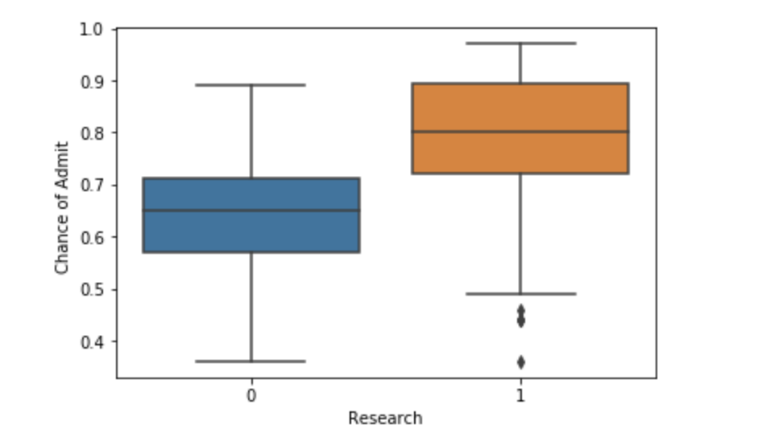
****

From the above graph, it can be concluded that the number of applicants who have research done is more than the applicants without the research.

Applicants with research experience: - 277

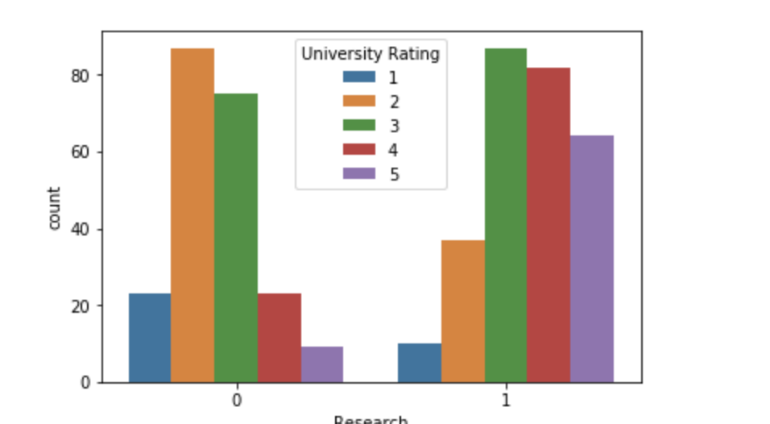
Applicants without research: - 220

**10.4 Chance of admittance on the basis of research experience:**

****

From the above box plots, we can see that the chance of admission for an applicant with research experience is greater than the applicants without the research experience. Mean for applicants with research experience is around 0.8 while whose without research is around 0.65.

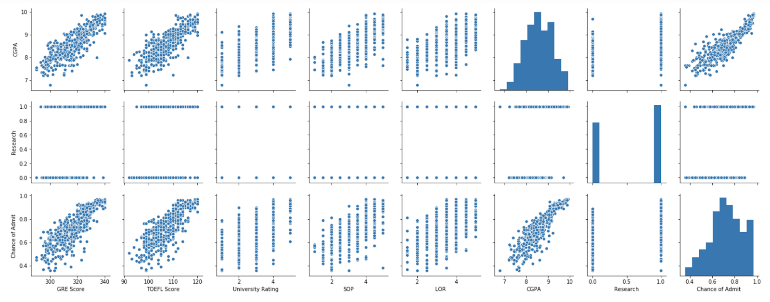
**10.5 University applications based on the research experience of applicant:**

****

The above count plot shows the trend of applying to the universities according to the research experience of the candidate. We can observe that the applicants without research experience tend to apply to universities with higher ratings. On the other hand, applicants with research experience tend to apply to universities with lower ratings.

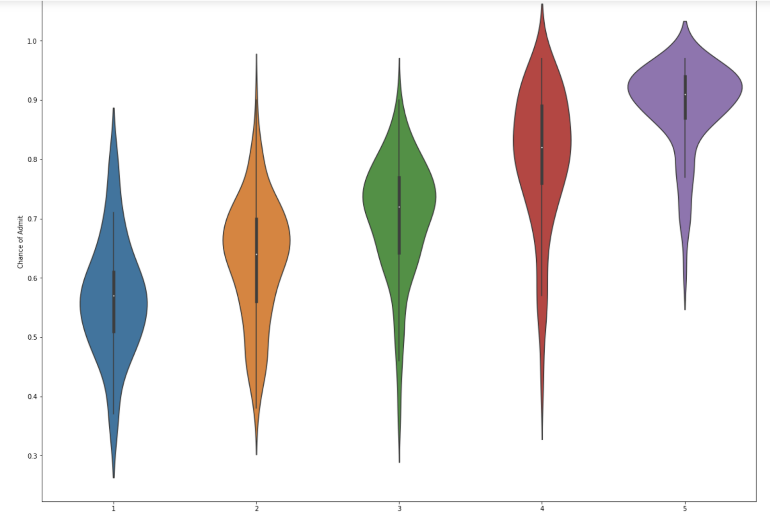
**10.6 Pair plots: -**

****

****

The above matrix shows every pair plotted against each other. We can observe that features like CGPA, TOEFL, GRE, and Chance of Admit are linearly related to each other.

**10.7 Chance of Admit on the basis of university ratings:**



From the above plot, we can conclude that the Chance of Admit in a higher rated college is less than the lower-rated college. College with rating 1 has a chance of admission at a mean of 0.55 whereas a college with rating 5 has a mean of 0.9.