Graduate Admission

Priyen Dang
pdang2@uncc.edu
801102016

Pratik Rajni Ashani
pashani@uncc.edu
801075870

Riddhi Jagdish Patil
rpatil14@uncc.edu
801076080

Alzarrio Rolle
arolle1@uncc.edu
800747319

Rahel Paul Fargose
rfargose@uncc.edu
801076463

Venkat Siddharth
vvennela@uncc.edu
801077224

Aishwarya Madadi
amadadi@uncc.edu
801077193

Jinal Butani
jbutani@uncc.edu
801077913

Amit Rajesh Patil
apatil19@uncc.edu
801077053

Raj Hitendra Jani
rjani@uncc.edu
801077311

Sankeerthan Burugula
sburugul@uncc.edu
801084135

***Abstract*— The goal of this paper is to do an in-depth analysis of the admission process in American Universities. The designated audience and beneficiaries of this paper will be students seeking Graduate study opportunity in the United States of America. It is always very difficult to assess which universities will be the best choice and a lot of factors matter while making the final decision. Our purpose is to do a deep study and provide the end-users(students) with information which will not only help them with the admission process but also help to make the decision of selecting the university easier. We have observed that students have generic questions while applying for a Graduate School and there are many distributed forums who give answers to these questions but that is based on personal experience or hearsay. To clear all the misconceptions, and make the admission application process clear, we are analyzing the dataset for graduate students which we selected from the Kaggle website. The dataset has data for more than 500 students. This will allow us to develop models that provide insight into student application behavior and university decision patterns, in turn making an improvement to the overall admission process.**

Keywords—education pattern, grading system, target group, Application process, decision pattern.

I. Introduction

The United States of America is one of the most popular countries in the world which is well known for the technological advent and the quality of education provided by the universities. The facilities, quality of professors and libraries provided by most of the universities in the United States have an unmatched level. Due to this, students from all over the world apply to these universities in order to give their career a much-required boost.

Since most of the students pursuing Graduate education are international/not native students, the fees associated with the application are very high, with little or no scope of error or reapplication.

Also, keeping into mind that each country has its own grading system and education pattern, it becomes difficult to adapt to the United States application system comparing it with respective countries education application pattern. We have observed that maximum Graduate admissions reject occur due to missing document, missing transcript or lack of knowledge of the what are basic requirements to secure an admission into a university.

Also, there many websites which give partial information based on student experience, but it does not provide a definitive answer to most of the questions.

In this paper, we create visualizations from the data and try to analyze the different parameters which play a major role in deciding the admission to the university. For example, we are comparing the GRE and TOEFL values and determining the average chance of admission if the student scores in that range of GRE and TOEFL. We analyze our system from a student's perspective, but it can be easily extended from the perspective of the universities as well.

The primary objective of this paper is to analyze the data set and apply data mining methodology which will help determine admission chances, making the admission process simpler for students applying for Graduate Studies. The application process is tedious and takes a lot of time, energy and is a multi-million business in the US. The target group for this paper is all the students/parents/aspirants who are planning to apply and pursue their graduate studies in America and help them as there is very less information available. To achieve this, we have done data preprocessing to remove the outliers. Next, we have applied various Regression models to analyze the data. We have used Multiple linear Regression, Regression Tree and KNN Regression for the above purpose. These models are been evaluated by mean square error, mean absolute error and r2 score.

II. Dataset

The graduate admission applications data was collected from Kaggle. The dataset was used for predicting the chances of admission into universities at a graduate level. It will help the students in shortlisting universities according to the candidate's profile. The output predictions that are derived using various data processing techniques give a fair idea about the chances of admission in a university. The dataset contains several parameters that can be considered very important and concurrently relevant regarding master's Programs. There is a total of 9 columns: Serial No., GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, Research, Chance of Admit. The parameters can be understood as follows:

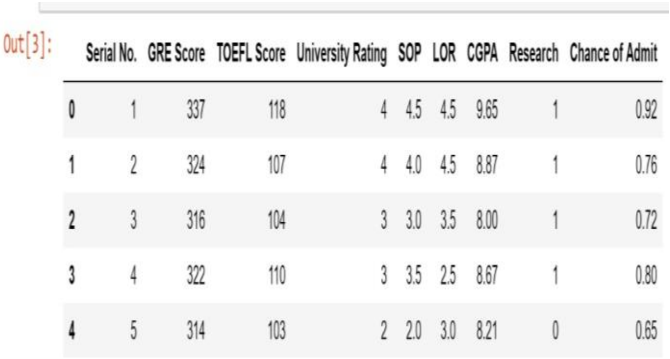
XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

- GRE Scores:- In most the graduate schools in the USA, the Graduate s Record Examinations(GRE) is a standardized test that is a requirement for admission. The candidate's performance is evaluated based on 2 sections which are classified as quantitative and verbal. The maximum GRE score one can get is 340.
- TOEFL Scores:- TOEFL is an abbreviation for Test of English as a foreign language. It is a standardized test for non-native speakers to enroll in a US university. This test grades a student based on the performances in 4 sections namely listening, speaking, reading and writing.
- University Rating:- This attribute is based on the reputation of the university. A university can have at the most a rating of 5.
- Statement of Purpose:- The SOP is one of the most important parts of a candidate's application that will tell the admissions committee who the candidate is really, what has influenced his/her career path so far, professional interests and where he/she plans to go from here. Based on how strong the SOP is there is a value given to this attribute out of 5.
- Letter of Recommendation:- The value of this attribute is based on how strong the candidates LOR is. It is a value out of 5. In a LOR the writer assesses the qualities, characteristics, and capabilities of the candidate who is being recommended. This attribute is related to admission in graduate school or scholarship eligibility.
- Undergraduate GPA:- This is the grade what individuals earn during undergraduate School.
- Research Experience:- This attribute is a binary value of either 0 for having no research experience or 1 for having some research experience. Majority of candidates have some research experience thus having a value of 1.
- Chance of Admit:- It Depends on all the previous parameters and the value is in the range of 0 and 1.

Before preprocessing the dataset has a total of 500 records. It will be further preprocessed to achieve the final dataset.

A. *Pre-Processing of Data*

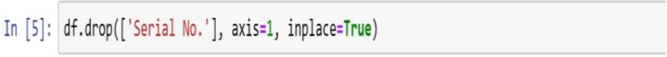
The data was thoroughly analyzed, and various functions were used for pre-processing the dataset. The dataset was imported into python data frame. "pandas.read_csv()" was used for importing the data.



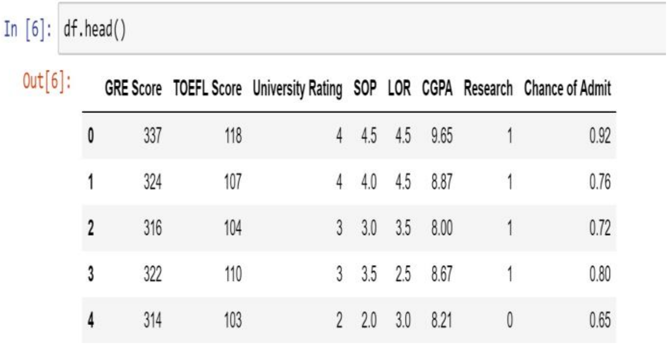
```
Out[3]:
   Serial No.  GRE Score  TOEFL Score  University Rating  SOP  LOR  CGPA  Research  Chance of Admit
0           1        337          118                  4  4.5  4.5  9.65         1             0.92
1           2        324          107                  4  4.0  4.5  8.87         1             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
```

Figure 1: - Graduate Admission dataset import

After importing the dataset, the Serial No. column was dropped.



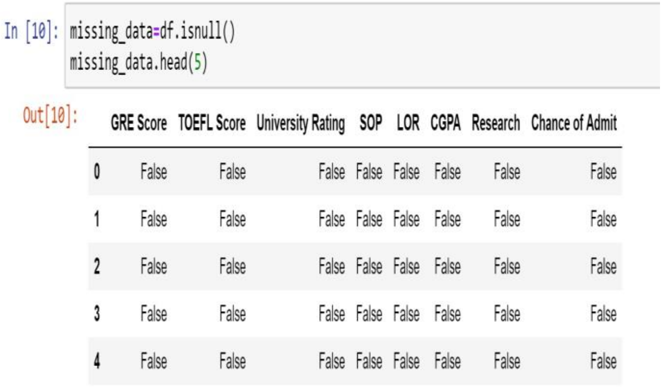
```
In [5]: df.drop(['Serial No.'], axis=1, inplace=True)
```



```
In [6]: df.head()

Out[6]:
   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
```

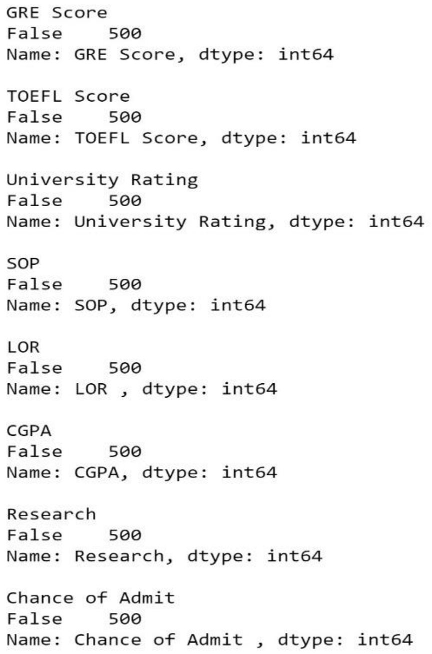
Figure 2: - Dropping serial column

Then it was checked for null values using the "isnull()". All the columns were checked, and the results were false for all the columns proving that there are no null values in the dataset.



```
In [10]: missing_data=df.isnull()
         missing_data.head(5)

Out[10]:
   GRE Score  TOEFL Score  University Rating    SOP    LOR   CGPA  Research  Chance of Admit
0      False        False              False  False  False  False     False            False
1      False        False              False  False  False  False     False            False
2      False        False              False  False  False  False     False            False
3      False        False              False  False  False  False     False            False
4      False        False              False  False  False  False     False            False
```

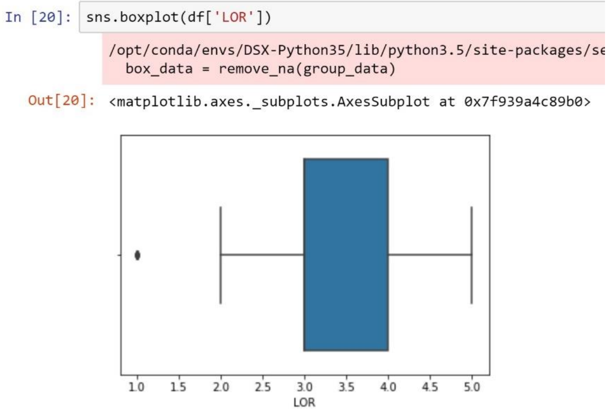
Figure 3: - Missing data handling



```
GRE Score
False    500
Name: GRE Score, dtype: int64

TOEFL Score
False    500
Name: TOEFL Score, dtype: int64

University Rating
False    500
Name: University Rating, dtype: int64

SOP
False    500
Name: SOP, dtype: int64

LOR
False    500
Name: LOR , dtype: int64

CGPA
False    500
Name: CGPA, dtype: int64

Research
False    500
Name: Research, dtype: int64

Chance of Admit
False    500
Name: Chance of Admit , dtype: int64
```

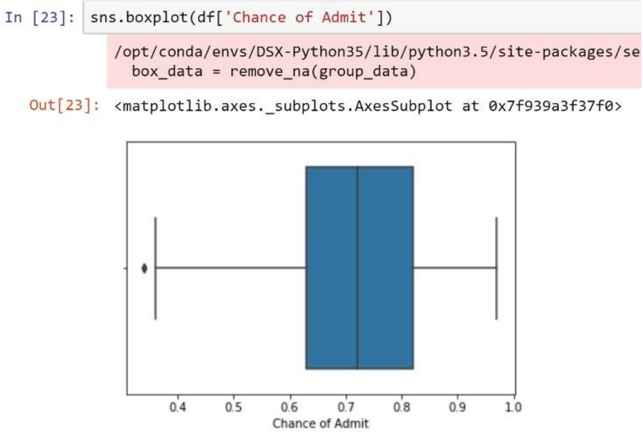
Figure 4: - Missing data handling result

Next, the dataset was checked for outliers. This was done using box plots. Using box plot visualizations outliers were found in two variables namely LOR and Chance of admit.



```
In [20]: sns.boxplot(df['LOR'])

         /opt/conda/envs/DSX-Python35/lib/python3.5/site-packages/se
           box_data = remove_na(group_data)

Out[20]: <matplotlib.axes._subplots.AxesSubplot at 0x7f939a4c89b0>
```

Figure 5: - Identifying outliers in LOR

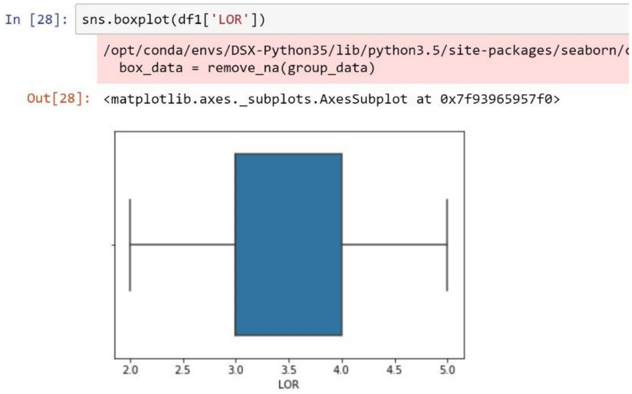


```
In [23]: sns.boxplot(df['Chance of Admit'])

         /opt/conda/envs/DSX-Python35/lib/python3.5/site-packages/se
           box_data = remove_na(group_data)

Out[23]: <matplotlib.axes._subplots.AxesSubplot at 0x7f939a3f37f0>
```

Figure 6: - Identifying outliers in Chance of Admit

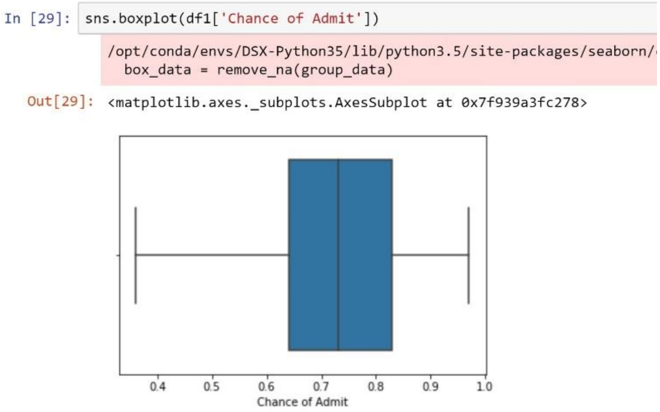
Finally , the outliers were removed from the dataset. To remove the outliers, interquartile rule was used. The following rule was considered :-

df < (Q1 – 1.5 * IQR)) |(df > (Q3 + 1.5 * IQR)



```
In [28]: sns.boxplot(df1['LOR'])

         /opt/conda/envs/DSX-Python35/lib/python3.5/site-packages/seaborn/c
           box_data = remove_na(group_data)

Out[28]: <matplotlib.axes._subplots.AxesSubplot at 0x7f93965957f0>
```

Figure 7: - Removing outliers from LOR



```
In [29]: sns.boxplot(df1['Chance of Admit'])

         /opt/conda/envs/DSX-Python35/lib/python3.5/site-packages/seaborn/c
           box_data = remove_na(group_data)

Out[29]: <matplotlib.axes._subplots.AxesSubplot at 0x7f939a3fc278>
```

Figure 8: - Removing outliers from Chance of Admit

The dataset was refined with proper values and no outliers. The final dataset has total of 486 records.

III. BUSINESS USE

The objective for this project is to prepare a dataset to help Indian students in shortlisting the universities based on their profile. Every year thousands of students apply to the universities across the United States. There are over 2000 universities in the United States with different majors and different streams. Deciding the suitable university which will match the profile and interest of the student is a key challenge for them. There are various factors involved in deciding the university such as entrance exams like GRE and TOEFL scores, university ranking, statement of purpose, letter of recommendation, undergraduate GPA and research work. Our data mining project will classify the output variable based on the predictors used. Using the values for different parameters mentioned we will be able to predict the chances of admission. The outcome of this project will help us to gather insights on the average rating of admit and help students classify themselves as to which tier universities, they have better chances respectively.

IV. Data visualization

A. Scatter Plot: -

A scatter plot is used to represent two dimensions on each axis. It can be extended to three dimensions by adding color. They are used to represent the association or co-relation between two numerical variables and find out clusters of observations if any.[11]

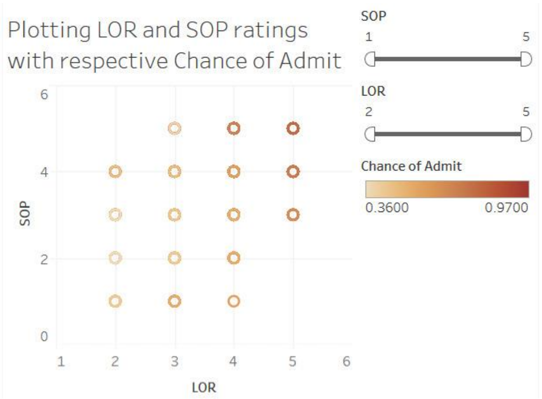


Figure 9: - This visualization contrasts with our second hypothesis to examine the LOR and SOP rating combinations against the chance of admit.

B. *Tree Map: -*

The tree map displays data in nested rectangles. The dimensions define the structure of the tree map and measures define the size or color of the individual rectangle. The rectangles are easy to visualize as both the size and shade of the color of the rectangle reflect the value of the measure.[12]

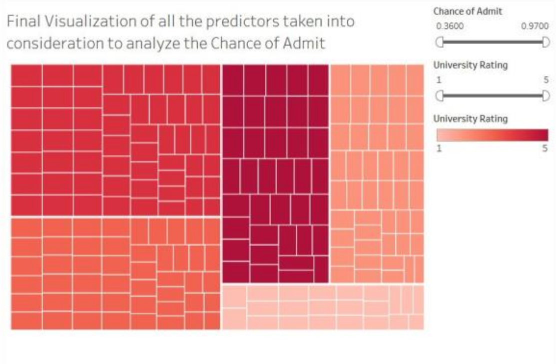


Figure 10: - This visualization depicts the chance of admit based on all the predictor variables, i.e. GRE score, TOELF score, SOP and LOR ratings, CGPA, Research work.

C. *Bubble Chart: -*

A bubble chart is a variation of a scatter chart here every data point is replaced with the bubbles and the size of bubbles indicates the additional dimension of the data. Like scatter chart, bubble chart does not use a category axis, both horizontal and vertical axes are value axes.

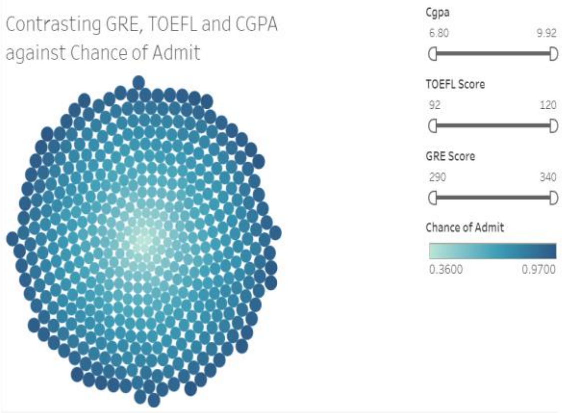


Figure 11: - This visualization provides us insights as to our preliminary hypothesis which takes the GRE, TOEFL and CGPA as the predictors and provides the Chance of Admit as response variable

D. *Line Chart: -*

Line Charts are primarily used for showing time-series data. They can be used to show changes in values of quantitative data over a short or a long period of time, e.g. data about changes in chances of admit with change in examination scores.[11]

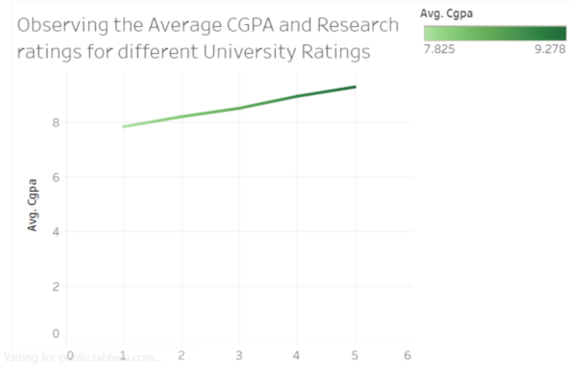


Figure 12: - This demonstrates the average CGPA score in contrast to different tier universities.

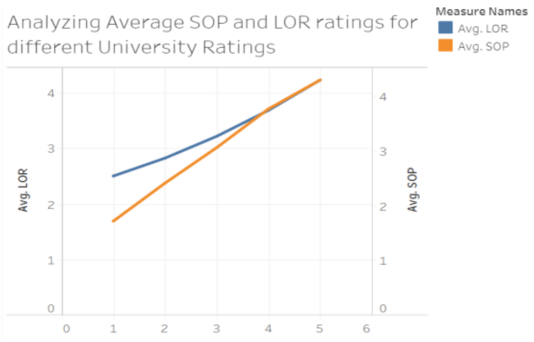


Figure 13: - This visualization contrasts with our second hypothesis to examine the LOR and SOP rating combinations against the chance of admit.

V. ALGORITHMS

A. *Multiple linear regression*

Multiple linear regression is the one form of linear regression analysis. Multiple linear regression is a statistical technique that uses explanatory variables to predict the output of a response variable. Basically, the Main goal of the linear regression is to model the linear relationship between the independent (explanatory) variables and dependent(response) variable. One can identify outlier with the use of Multiple linear regression.[10]

B. *K-nearest neighbors*

KNN is an algorithm which is used to solve classification and regression problems. KNN stores all available cases and classifies new cases based on distance functions. It's a classifier algorithm where the learning is based on how similar a data/set is in comparison with others. To calculate the distance, we use Euclidean distance.[8]

C. *Regression Tree*

A regression tree/decision tree is a process which is built through as binary recursive partitioning. This is an iterative process that splits the data into partitions or branches. This process continues splitting each partition into smaller groups as the method moves up each branch. Since the target variable does not have classes, we fit a regression model to the target variable using each of the independent variables.[9]

VI. HYPOTHESES

A. *How important are GRE, TOEFL scores and CGPA to determine chance of getting an admit?*

B. *Does higher rating of SOP and LOR lead to greater chance of admit?*

C. *What are the chances of getting an admit from the university considering all variables(i.e GRE Score, TOEFL Score, University rating, CGPA, Research)?*

VII. MODEL EVALUATION

After deciding on the hypotheses, a total of three modeling techniques were applied on each of the hypotheses. Various attributes were considered as input and output variables and then results were generated for each of the models. Finally, all the three models were compared, and the best model was selected between the three. Following is the detailed explanation of all the different modelling techniques that were applied for each of the hypothesis:

A. *Hypothesis 1*

How important are GRE, TOEFL scores and CGPA important in getting an admit?

The final number of records used for the creation of model was 486.The data was divided into training, validation and testing datasets. There were 243 records(50%) in training, 146 records(30%) in validation and 97 records(20%) in testing dataset.

Independent(Input) variables: -

1. CGPA
2. GRE
3. TOEFL

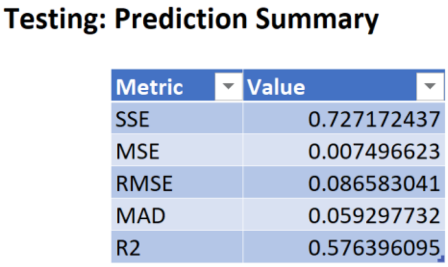
Dependent(output) variable: -

1. Chance of Admit

Model Building: -
Three algorithms were performed on the above hypothesis and they are: -

1. Multiple linear Regression
2. Regression Tree
3. K-NN

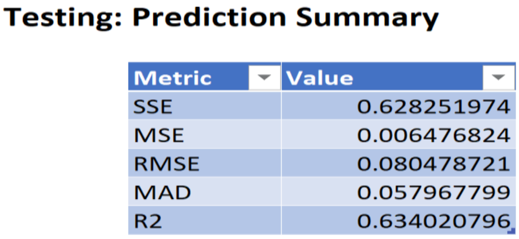
Best Model: -



Testing: Prediction Summary

Metric	Value
SSE	0.727172437
MSE	0.007496623
RMSE	0.086583041
MAD	0.059297732
R2	0.576396095

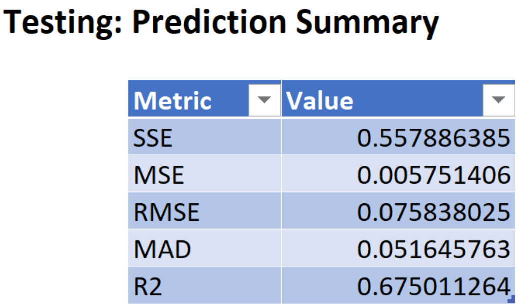
Figure 14: - KNN Testing Prediction summary



Testing: Prediction Summary

Metric	Value
SSE	0.628251974
MSE	0.006476824
RMSE	0.080478721
MAD	0.057967799
R2	0.634020796

Figure 15: - Regression Tree Testing Prediction summary



Testing: Prediction Summary

Metric	Value
SSE	0.557886385
MSE	0.005751406
RMSE	0.075838025
MAD	0.051645763
R2	0.675011264

Figure 16: - Linear Regression Testing Prediction summary

We build three models for this hypothesis namely; KNN, Regression Tree and Multiple Linear Regression. The best

model was chosen based on higher R2-score and lower SSE, MSE and RMSE values.
From the testing prediction summary for all three models we can see that Multiple Linear Regression model has the least RMSE value (0.0758) when compared to KNN and regression tree. Multiple Linear Regression also has the highest R2 score (0.6750) amongst the three. In addition to the above, Multiple Linear Regression model has lower values of sum of squared error (SSE) and Mean squared error (MSE) when compared to KNN model and Regression tree model. Hence, for this hypothesis Multiple linear Regression is the best model amongst all three models.

B. *Hypothesis 2*

Does higher rating of SOP and LOR lead to greater chance of admit?

The final number of records used for the creation of model was 486.The data was divided into training, validation and testing datasets. There were 243 records(50%) in training, 146 records(30%) in validation and 97 records(20%) in testing dataset.

Independent(input) variables: -

1. SOP
2. LOR

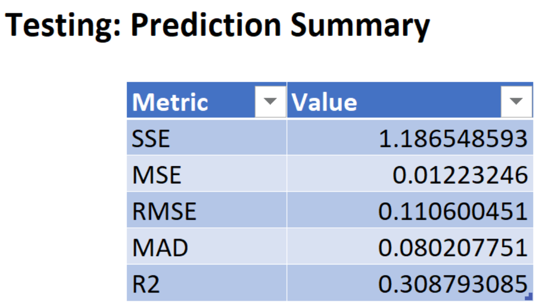
Dependent(output) variable: -

1. Chance of admit

Model Building: -
Three algorithms were performed on the above hypothesis and they are: -

1. Multiple linear Regression
2. Regression Tree
3. K-NN

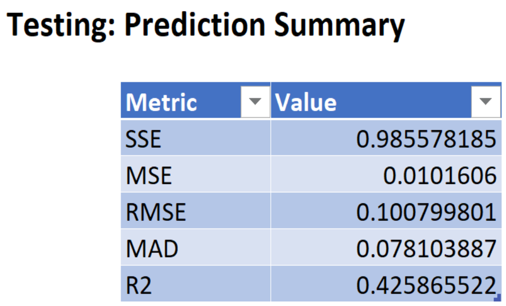
Best Model: -



Testing: Prediction Summary

Metric	Value
SSE	1.186548593
MSE	0.01223246
RMSE	0.110600451
MAD	0.080207751
R2	0.308793085

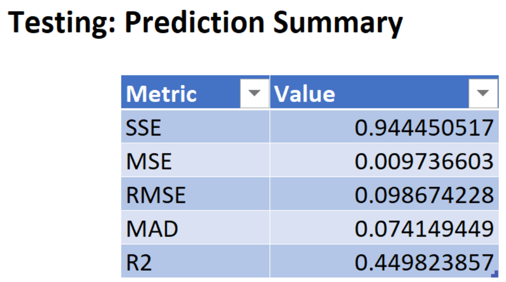
Figure 17: - KNN Testing Prediction summary



Testing: Prediction Summary

Metric	Value
SSE	0.985578185
MSE	0.0101606
RMSE	0.100799801
MAD	0.078103887
R2	0.425865522

Figure 18: - Regression Tree Testing Prediction summary



Testing: Prediction Summary

Metric	Value
SSE	0.944450517
MSE	0.009736603
RMSE	0.098674228
MAD	0.074149449
R2	0.449823857

Figure 19: - Linear Regression Training Prediction summary

Three models were built for this hypothesis namely; KNN, Regression Tree and Multiple Linear Regression. The best model was chosen based on higher R2-score and lower SSE, MSE and RMSE values.
From the testing prediction summary for all three models we can see that Multiple Linear Regression model has the least RMSE value (0.0986) when compared to KNN and regression tree. Multiple Linear Regression also has the highest R2 score (0.4498) amongst the three. In addition to the above, Multiple Linear Regression model has lower values of sum of squared error (SSE) and Mean squared error (MSE) when compared to KNN model and Regression tree model. Hence, for this hypothesis Multiple linear Regression is the best model amongst all three models.

C. *Hypothesis 3*

What are the chances of getting an admit from the university considering all variables(i.e GRE Score, TOEFL Score, University rating, CGPA, Research)?

The final number of records used for the creation of model was 486.The data was divided into training, validation and testing datasets. There were 243 records(50%) in training, 146 records(30%) in validation and 97 records(20%) in testing dataset.

Independent(input) Variables: -

1. GRE Score
2. TOEFL Score
3. University rating
4. SOP
5. LOR
6. CGPA
7. Research

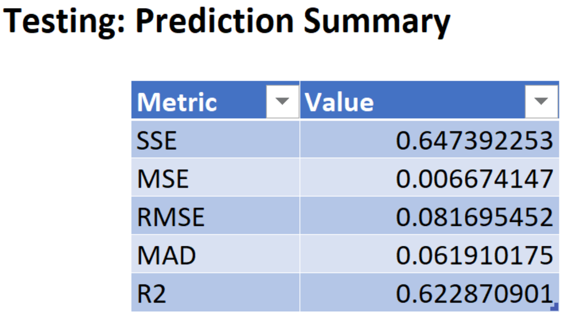
Dependent(output) variable: -

1. Chance of admit

Model Building: -
Three algorithms were performed on the above hypothesis and they are: -

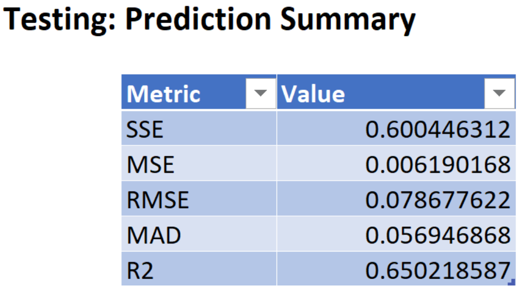
1. Multiple linear Regression
2. Regression Tree
3. K-NN

Best Model: -


Testing: Prediction Summary

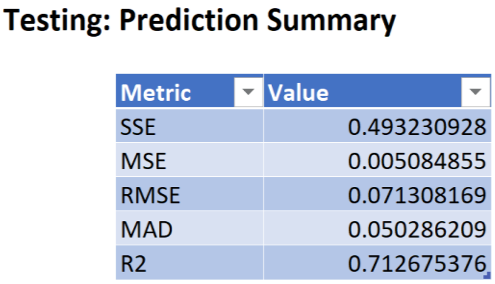
Metric	Value
SSE	0.647392253
MSE	0.006674147
RMSE	0.081695452
MAD	0.061910175
R2	0.622870901

Figure 20: - KNN Testing Prediction summary


Testing: Prediction Summary

Metric	Value
SSE	0.600446312
MSE	0.006190168
RMSE	0.078677622
MAD	0.056946868
R2	0.650218587

Figure 21: - Regression Tree Testing Prediction summary


Testing: Prediction Summary

Metric	Value
SSE	0.493230928
MSE	0.005084855
RMSE	0.071308169
MAD	0.050286209
R2	0.712675376

Figure 22: - Linear Regression Training Prediction summary

Three models were built for this hypothesis namely; KNN, Regression Tree and Multiple Linear Regression. The best model was chosen based on higher R2-score and lower SSE, MSE and RMSE values.

From the testing prediction summary for all three models we can see that Multiple Linear Regression model has the least RMSE value (0.0713) when compared to KNN and regression tree. Multiple Linear Regression also has the highest R2 score (0.7126) amongst the three. In addition to the above, Multiple Linear Regression model has lower values of sum of squared error (SSE) and Mean squared error (MSE) when compared to KNN model and Regression tree model. Hence, for this hypothesis as well Multiple linear Regression is the best model amongst all three models.

VIII. CONCLUSION

By carefully examining and analyzing the results obtained upon performing various modelling techniques and visualizations on the dataset of Graduate Admissions, we have cross-examined with our hypothesis and made the following conclusions:

• Consequently, the LOR and SOP ratings to secure a better chance of admit, i.e. more than 0.70, are (3,4), (3,5), (4,4), (4,5) or (5,5). As the university rating increases, the SOP and LOR ratings need to be either 4 or 5 to be safe.

• From our analysis, it is also clear that the research work plays an important role to secure an admission into a higher rated university. The difference in the average research work from a tier 3 university and a tier 4 university is 0.25 which explains how impactful and important it is towards the admission.

• The most important data to have obtained is that students having a GRE and TEOFL score more than 317 and 109 have the Chance of Admit averaging about 0.76 coupled with a CGPA of 8.65 or more coupled with Research work.

• Finally, from the results obtained, we have concluded that most of the students have their admissions into tier 2 and tier 3 universities.

The outcome of our research and analysis is to help students to analyze and chalk out an approach that aids them to understand the possibilities and requirements of their admissions into the Graduate Programs.

IX. STRATEGIC RECOMMENDATION

Upon completion of our analysis and pondering over solutions to our hypotheses, we have come up with some strategic recommendations that could provide some boost to the purpose of this paper and enhance the possibilities of future development.

1. One such recommendation is to have more tuples in the dataset. As of now, our dataset contains only 500 records and after pre-processing, the count is reduced to 486. This makes our analysis difficult as we are pushing towards a real-time response from our hypotheses and the chance of error is more as compared to when we have thousands of records.
2. Another major improvement that can be done is to have an interesting and pivotal attribute like work experience. In recent times, work experience has become an important factor in securing an admission into universities and it has become a crucial factor for judging the admission chance in respective university tiers.
3. The final scope of development that could be patched is categorizing the admissions into their respective programs. As not all students are admitted into the same program, having a dimensional view of several graduate programs will help us to improvise the output in accordance to their respective programs. This will aid students to prepare for the requirements or know their possibility of admission into various categories of degrees into universities of different tiers.

REFERENCES

[1] M. G. Haug. (2019). *Measure of association*. Available: https://www.britannica.com/topic/measure-of-association

[2] G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," in *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, 2013, pp. 1-7.

[3] H. Gulati, "Predictive analytics using data mining technique," in *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2015, pp. 713-716.

[4] C. Song, "Research of association rule algorithm based on data mining," in *2016 IEEE International Conference on Big Data Analysis (ICBDA)*, 2016, pp. 1-4.

[5] G. Shmueli, P. C. Bruce, and N. R. Patel, "Data mining for business analytics : concepts, techniques, and applications with XLMiner," (in English), 2016.

[6] M. S. Acharya. (2019, March 12, 2019). *Graduate Admissions*. Available: https://www.kaggle.com/mohansacharya/graduate-admissions

[7] J. Phillips. (2014, April 2, 2019). *Analytics 3.0: The Era of Impact*. Available: https://www.sas.com/content/dam/SAS/en_us/doc/event/The-Era-of-Impact-127837.pdf

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10-18, 2009.

[9] I. Frontline Systems. (2019, April 20, 2019). *Regression Trees*. Available: https://www.solver.com/regression-trees

[10] K. J. Preacher, P. J. Curran, and D. J. Bauer, "Computational Tools for Probing Interactions in Multiple Linear Regression, Multilevel Modeling, and Latent Curve Analysis," *Journal of Educational and Behavioral Statistics*, vol. 31, no. 4, pp. 437-448, 2006/12/01 2006.

[11] L.-S. Tsay. (2019, February 21, 2019). *Data Visualization using XLMiner*. Available: https://uncc.instructure.com/courses/89875/files/5048964?module_item_id=1594412

[12] T. Point. (2019, April 23, 2019). *Tableau - Tree Map*. Available: https://www.tutorialspoint.com/tableau/tableau_tree_map.htm

[13] B. Leiva. (2016, April 23, 2019). *Creating Scatter Plots in Tableau*. Available: https://www.thedataschool.co.uk/borja-leiva/creating-scatter-plots-tableau/