Project Report
Foundations of Data Science

Kunal Relia
N14724679
Kbr263

Question 1

What did you propose to do? What is the motivation/background?

Universities of USA receive a lot of applications because of their reputation. And New York University is no different. It's Tandon School of Engineering has attracted more than 40,000 applications for their master program since Fall2007. This has led to a huge dataset of all the applications. And like any other Data Science student, I saw an opportunity here. I proposed to analyze the applications and clean the data for them. I aimed at removing all the data inconsistencies and incompleteness making the data analysis ready. As for statistical analysis, I analyzed various parameters like number of applications received, percentage of students given the admission and various other parameters. Along with this, I proposed to do linear regression on data and implement K-Nearest Algorithm to determine if an applicant can get an admission offer based on various factors like test scores.

The motivation for this project was the various affiliations that School of Engineering has undergone in the last decade which led to a "dirty" dataset. The dataset is a mix of data for each semester of the last decade stored into spread sheets. The spread sheets are the only source of data now with no access possible to previous software database. Hence, cleaning the data and coming up with interesting analysis would have been a good hand-on practice in the real world of data science where the Office of Graduate Admissions was a real time customers which explained the background and various parameters of the dataset.

Question 2

Explain the data you used and model in detail.

I have used the following data set:

Attribute Name	Attribute Type	Description
ADMIT_TERM	String	Term to which applied for
Decision	String	Decision to the application
Admission_Decision	String	Decision to the application – Admit or Reject only
Acad_plan	String	Department applied to
ACAD_PROG	String	Master or PHD
GRE_Verbal	Numeric	GRE verbal score
GRE_Quant	Numeric	GRE quantitative score
Total.Score	Numeric	GMAT Score
TOEFL_Total	Numeric	Total TOEFL Score
IELTS Score	Numeric	Final IELTS band
Transcript1 GPA	Numeric	UnderGrad GPA
Transcript1 School Name	String	UnderGrad School Name
Stage of Process	String	Current stage of the application
Submitted Date	Date	Date on which application is submitted
Decision Publish Date	Date	Date on which decision on application was posted
Scholarship_Amount	Numeric	Scholarship awarded
First_Condition	String	Conditional admit or not
COUNTRY OF CITIZENSHIP	String	Country of citizenship of the applicant
Deposit Paid	String	Whether admission offer was accepted or not
Academic_Load	String	Part-Time or Full-Time

Along with the above table, I have used the following table which contains only numeric columns. This table was used for all the analysis purposes:

Attribute Name	Attribute Type	Description
ADMIT_TERM	Numeric	Term to which applied for
Admission_Decision	Numeric	Decision to the application – Admit or Reject only
Acad_plan	Numeric	Department applied to
ACAD_PROG	Numeric	Master or PHD
GRE_Verbal	Numeric	GRE verbal score
GRE_Quant	Numeric	GRE quantitative score
Total.Score	Numeric	GMAT Score
TOEFL_Total	Numeric	Total TOEFL Score
IELTS Score	Numeric	Final IELTS band
Transcript1 GPA	Numeric	UnderGrad GPA
Submitted Date	Numeric	Date on which application is submitted (EPOCH time)
Scholarship_Amount	Numeric	Scholarship awarded
COUNTRY OF CITIZENSHIP	Numeric	Country of citizenship of the applicant
Academic_Load	Numeric	Part-Time or Full-Time

I initially loaded all the data from the csv files and cleaned it. Cleaning was done and an all numeric dataframe was created for doing analysis. Then, I used Linear Regression to find how fit was the model. I did linear regression on Admission_Decision & Scholarship_Amount. Then, using the K-nearest neighbor algorithm, I tried to find out the admission_decsion & the scholarship_amount that one would earn. Finally, evaluation of KNN was done using the confusion matrix and its accuracy and its sensitivity and specificity.

Question 3

What did you end up doing?

Mostly I was able to stick to the plan that I had made. To sum up, I cleaned the data, did linear regression on it followed by KNN which was in turn evaluated by the use of confusion matrix (accuracy, Precision,ReCall,etc.).

- Most importantly, it was discovered that the scholarship amount is highly dependent on the admission decision, test scores & GPA. This was proved using Linear Regression.
- Also, we desire a higher Specificity here instead of sensitivity. This is because a wrong prediction that one will get admission is not a good thing. Imagine first getting an admission and then getting a rejection apologizing for the earlier 'wrong' decision. Rather exact opposite case is something that is not so bad. Here, False Negative is better than a False Positive. This desire was validated during the evaluation stage.

Question 4

What if anything did you change about your approach and why?

I followed the path that was initially thought of. But en route, a few changes were necessary to be made in order to improve the overall efficiency.

- Though initially not thought of, I created a data frame consisting of important parameters required for predictive analysis. All these parameters were converted into numeric format so that regression can be done.
- I added added ielts, GMAT, Admission_Decision and Acad_Load for improved analysis.
- It was expected that admission was only dependent on test scores & GPA. But it turned out various other factors like resume and SOP do play an important role in determining the admission decision. The expectation was held wrong by the regression analysis.

Question 5

What visualization(s) have you included? Explain what is conveyed in the visualization and why.

**All the visualizations are shown in the HTML file with the corresponding explanation **

Question 6

What evaluation method did you propose?

I chose to use the confusion matrix to evaluate the work done. This was chosen so as to directly get the accuracy of the model used. Further, other characteristics of the model such as Precision, ReCall, Sensitivity, specificity, etc. were also obtained by the use of confusion matrix.

The accuracy obtained is pretty high. Also as previously discussed, we have a high specificity too.

Question 7

How did your model perform according to this evaluation?

There were two models being evaluated:

1) Admission Decision
    a. Linear Regression proved the non-dependency amongst the parameters to determine the admission decision. Graph plots discussed testify this fact.
    b. KNN on the predictive model is of least importance given the above analysis.

    Therefore, much evaluation could not be done.

2) Scholarship Amount
    a. Linear Regression showed less support for entire dataset but a very strong regression on semester wise dataset.
    b. KNN had a good accuracy.

    The model was highly accurate. This was testified by the high accuracy obtained on the confusion matrix. Further, the corresponding graph too testifies that the model did well.

Question 8

Based on your results what conclusions do you draw?

The biggest drawback of analyzing the data was the change in GRE scoring pattern. This meant that the older applicant data was virtually not so useful. But the good learning curve was the fact that analyzing each and every semester's data was better than selecting the whole dataset at once. Further, contrary to the expectations, the test results & GPA do not significantly contribute to the admission decision. As discussed, other factors like resume, SOP & LORs do have a significant impact which were not included in my analysis due to the absence of relevant data. Finally, I can conclude that scholarship amount is determined by parameters like admission decision (pretty obvious), GRE scores & UnderGrad GPA. Resume, SOP & LORs don't seem to contribute over here.

Question 9

Based on your results what further studies would you do or are warranted?

As discussed in the conclusion, it is highly certain that factors like Resume, SOP & LORs do have a weightage when an admission decision is made by the department. Hence, such parameters can be weighed in when doing the admission analysis and we can hence, have a stronger model which can rightly predict the admission decision with higher accuracy.

Also I agree that there can be a more complex and more accurate model available that can be used. Hence, rather than doing linear regression and using KNN, it will be interesting to find a different model and evaluate its performance.