Jeremiah Ogunla

Introduction to Data Mining – DATS 6103 Fall 2021

Individual report on Predictive Modeling For Chance Of Admission

Introduction

The project focused on predicting the chances of admission to a US graduate school using data from UCLA the data is available publicly on Kaggle.

The project was divided into 4 sections:

- Data preprocessing
- Data Visualization
- Machine learning modeling
- GUI (pyqt5)

Individual work

I worked on part of the Data preprocessing, Data visualization and machine learning algorithm.

Importing Dataset

For the Data preprocessing I work on integrating the Kaggle API into our project such that we do not need to have the dataset downloaded and stored locally on our machine or online such as on google drive or GitHub, but we worked directly from where the dataset is hosted on Kaggle (kaggle, n.d.).

Heat map and boxplot

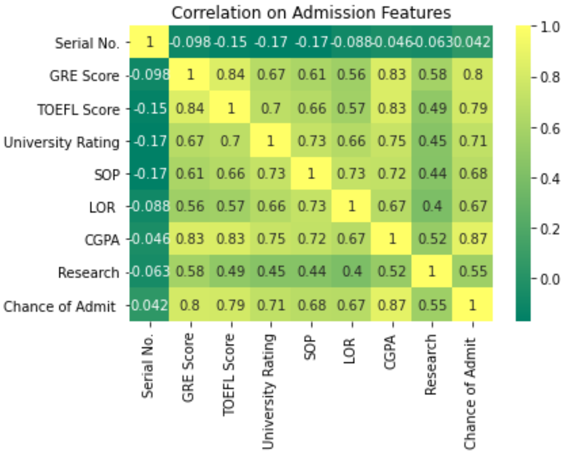


Figure 1

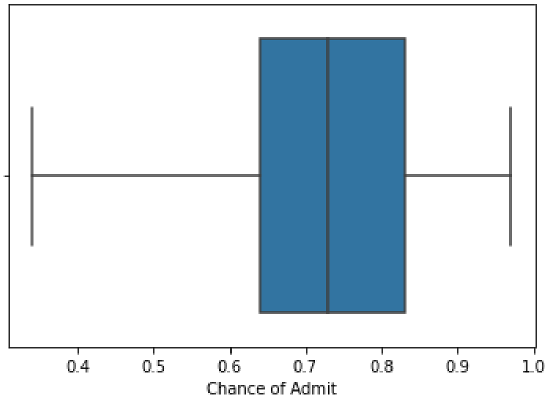


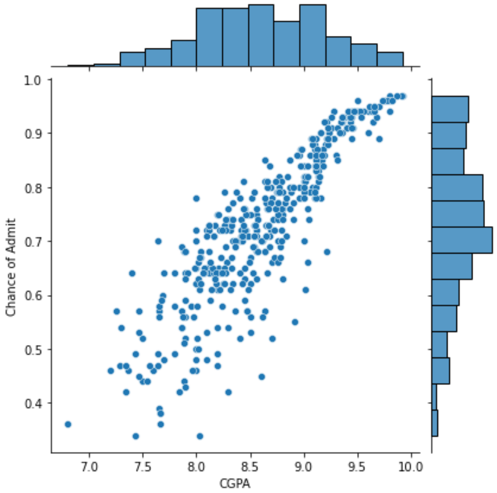
Figure 2

In the figure 1 the heat map shows the correlation between our variables ranging from -1 to +1, and the color of the map also visually shows the extent of the correlation from deep yellow (+1) to deep green (-1), the closer to zero means there is no linear trend between the two variables. The closer to 1 the correlation is the more positively correlated they are. A correlation closer to -1 is similar, but instead of both increasing, one variable will decrease as the other increases.

CGPA has the highest correlation to chance of admit with a positive correlation of 0.87, followed by GRE score, TOEFL Score with a correlation of 0.8 and 0.79 respectively.

The boxplot in figure 2 shows how the chances of admission distribution are spread and skewed across the various quartiles, the distribution is a normal distribution the lower quartile of the chance of admission is approximately 0.64 while the median score is 0.73 and the upper quartile is 0.83.

Pair plots CGPA



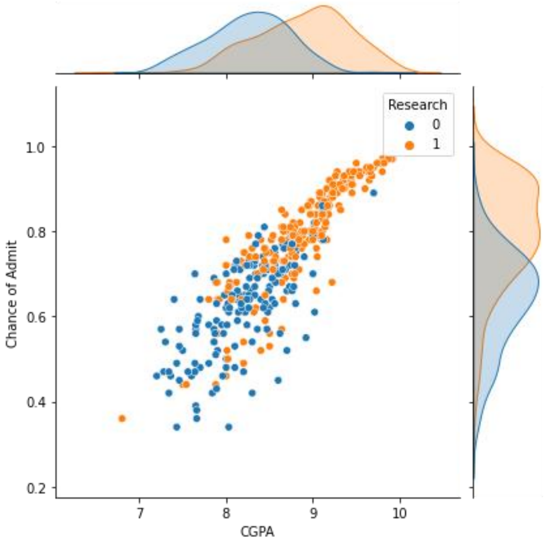
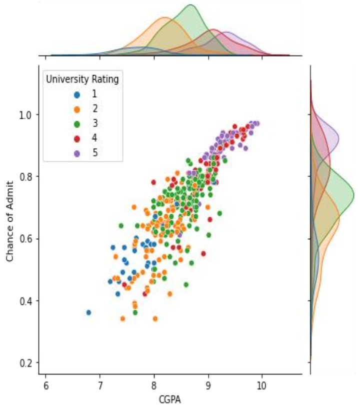


Figure 3 *Figure 4*

In figure 3 the pair plot between our dependent variable (Chance of admit) and independent variable (CGPA) shows a positive relationship between the two variables student with higher CGPA have higher chances of being admitted compared to those with lower CGPA, also we can see that the CGPA has a normal distribution, we went further to see the influence on research on this two variable in figure 4 and from the result 0 representing those without research skills and 1 representing student with research skills, student with higher CGPA and research skills have higher chances than those without a research experience.



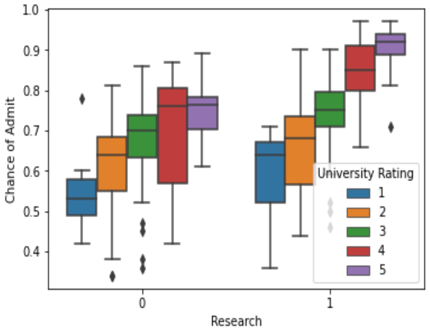
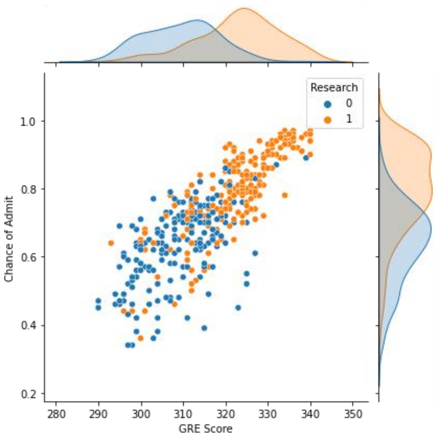


Figure 5 Figure6

We also used boxplots to explore the relationship between ‘Research’, ‘University Rating’, and ‘Chance of admit’. From Figure 6, we saw that applicants with research experience tend to have a higher chance of admission, with their university rating taken into account.

Pair plots CGPA



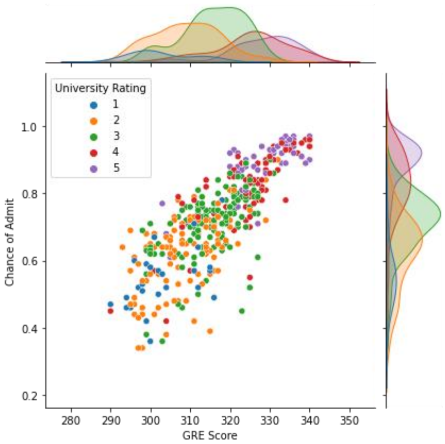
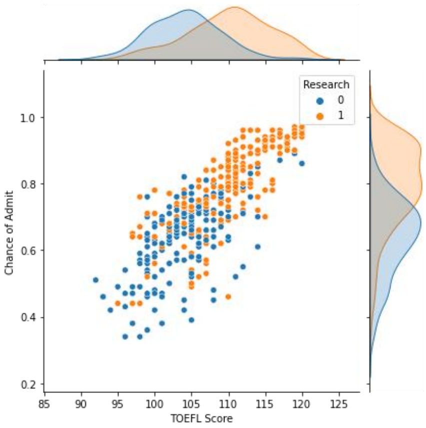


Figure 7 *Figure 8*

In figure 7 the pair plot between our dependent variable (Chance of admit) and independent variable (GRE Score) shows a positive relationship between the two variables student with higher GRE Score have higher chances of being admitted compared to those with lower GRE Score, also we can see that the GRE score has a normal distribution, we went further to see the influence on university rating on this two variable in figure 8 student with higher GRE score and good university rating have higher chances than those from a lower university rating .

Pair plots TOEFL Score



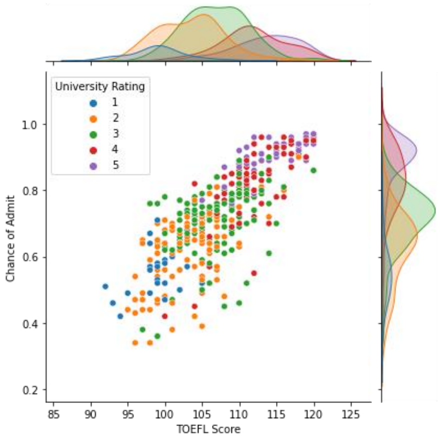


Figure 9 *Figure 10*

In figure 9 and 10 the pair plot between our dependent variable (Chance of admit) and independent variable (TOEFL Score) shows a positive relationship between the two variables student with higher TOEFL Score have higher chances of being admitted compared to those with lower TOEFL Score, also we could see that having research experience and a higher university rating increases the chances of the student admission.

CONCLUSION:

From the data visualization we can see that that the CGPA has the highest correlation to the chance of admission, and this is the most important feature in predicting the chance of admission to the master's programmed, other important independent variable are GRE score and TOFEL as they also have strong positive correlation to the chance of admission.

I was able to learn the importance of Exploratory data analysis and visualization process of data mining as a very crucial tool in model building, from the EDA we were able to get an overview of our data, know what variables are important for our model building in determining the dependent and independent variables.

Bibliography

kaggle. (n.d.). Retrieved from https://www.kaggle.com/docs/api