DSC 550-T301  
Final Paper:

Regression Analysis on Graduate Admissions Data

Vilius Smilinskas

08/07/20

**Abstract**

By 2018, 2.5 million new jobs are estimated to need advanced degrees. The Bureau of Labor Statistics estimates that jobs requiring master’s degrees will grow by 18% from 2008 to 2018.[1] Thus it is important for the potential student to know what is necessary to be accepted to a Graduate program. The goal of the project is to prepare data on graduate admissions, train and compare models. Then the model could be utilized for a website or application assisting students as a guide for graduate admissions.

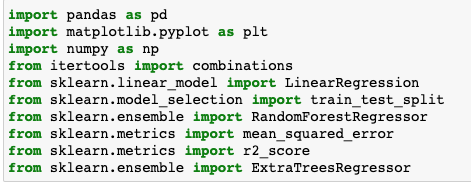


Figure 1: Libraries Utilized in the project.

**Data Exploration**

**Libraries:**

For the project the python programming language was utilized due to rich variety of libraries built for data cleaning, wrangling and management:

* **pandas:** Utilized for the DataFrame object, allowing for storage and manipulation of structured data.
* **matplotlib.pyplot:** A visualization library used for graphing and plotting of data.
* **numpy:** Short for Numerical Python the library contains many useful mathematical functions, also contains the array object.
* **itertools.combinations:** Provides for easy implementation of finding combinations of values.
* **sklearn:** Library filled with implementations of models, metrics and tools to efficiently set up the modeling process.

**Variables:**

Although not comprehensive the data includes many necessary independent variables as contributing factors in ones admission to a graduate program. Standardized testing is often utilized to gauge the readiness of a student for a competitive program. The prestige of the school and program attributes to the difficulty of being accepted. Written sections of admission process’ allow for a potential student to stand out through their ambition and personality. Additional metrics on scholarly performance are represented through the prior Grade Point Average (GPA), and prior research experience highlights the students with knowledge of scientific process’. All contributing to the dependent variable, the probability of Admittance. The distributions for the variables can be seen in Figure 2.

* GRE Scores (out of 340) - Scores in standardized testing.
* TOEFL Scores (out of 120) - Scores in standardized testing.
* University Rating (out of 5) - Rating of the School applied to.
* Statement of Purpose and Letter of Recommendation Strength (out of 5) -  
  Rating of the Written Application. Half intervals are possible (Ex. 2.5)
* Undergraduate GPA (out of 10) - GPA during undergraduate studies for applicant.
* Research Experience (either 0 or 1) - Boolean True or False if applicant has research experience.
* Chance of Admit (ranging from 0 to 1) - A probability of being admitted to program.

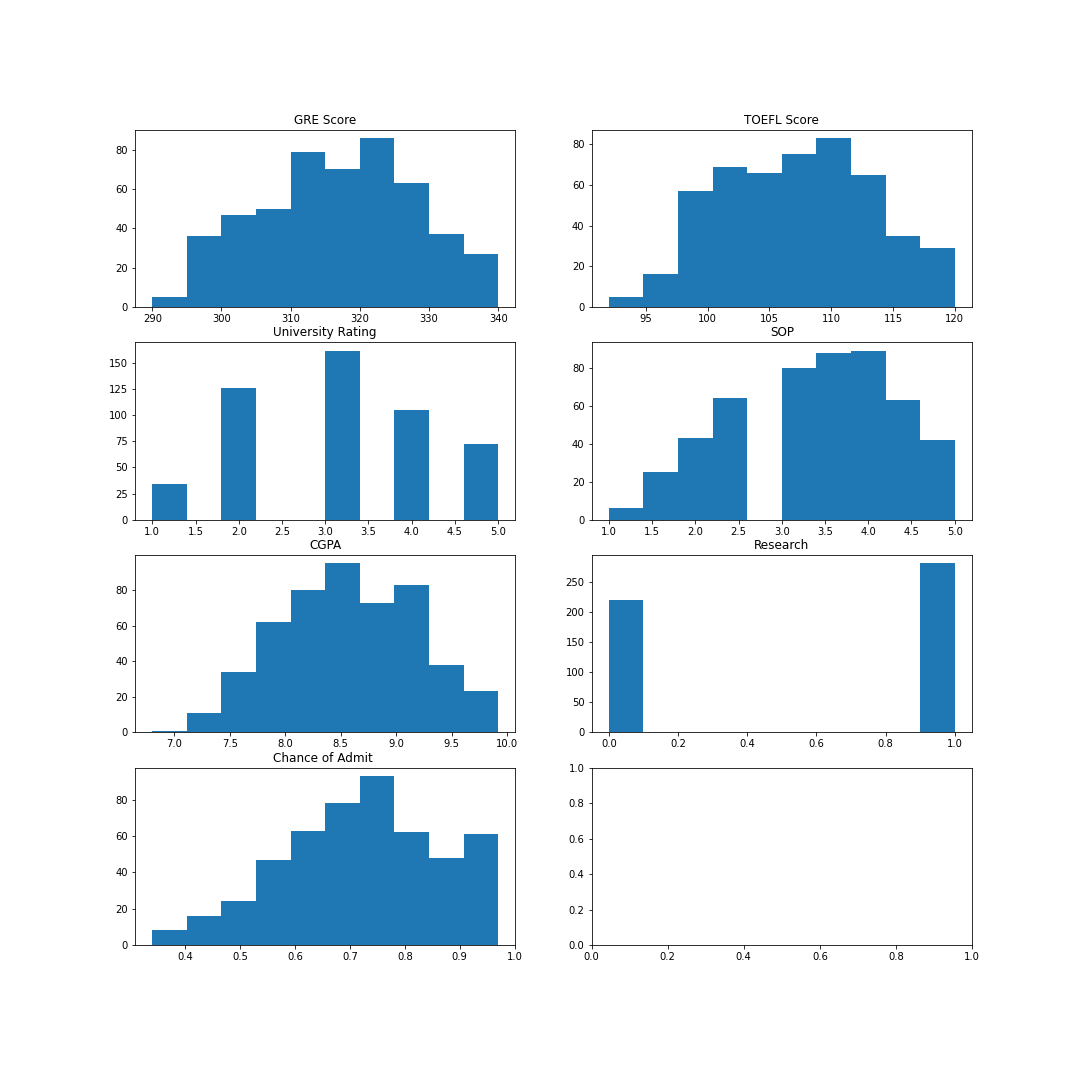


Figure 2: Variable Histograms

Among the variables, distributions are generally, approximately normal. with the exception of the binomial Research variable and a few right skewed graphs.

**Relationships**

As the aforementioned variables are often contributing factors of successful applications we can expect strong correlation among each other. Which is incredibly useful when it comes to the predictions, the more we can know on the strength of the model the better our final product will be.

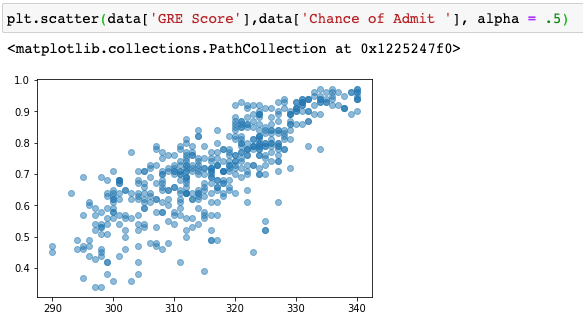


Figure 3: GRE vs Admission Scatter

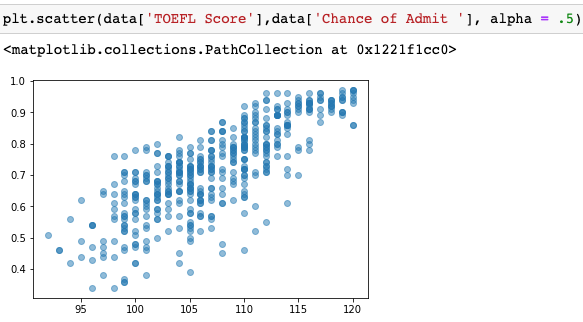


Figure 4: TOEFL vs Admission Scatter

As seen above the values between the two variables form a stretched line going in the positive direction. Indicating a positive relationship. In Figure 5 it can be seen that the correlation value of GRE and Chance of Admittance is .810351 indicating a very strong positive relationship. Similarly the TOEFL and Chance of Admittance correlation is a very strong positive relationship measured at .792228. The lowest correlation value among the dependent variables and the independent variable is for research. Although the relationship is still strong at .545871 it is not as influential as the other values.

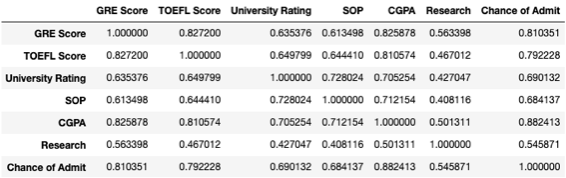


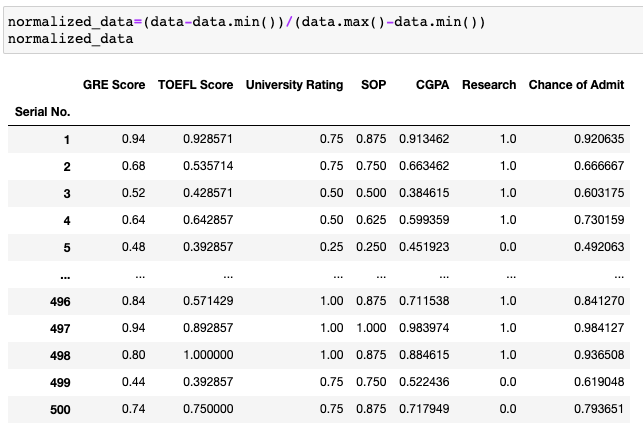
Figure 5: Correlation Table

**Data Preparation**

The dataset provided by Mohan Acharya [2] was very easy to implement, and required minimal cleaning. Utilizing the pandas library the data was loaded into a DataFrame object and the entire data set was normalized, due to the DataFrame supporting matrix multiplication, the entirety was completed in a single line of code which can be seen at the top of Figure 6. Additionally the data set had to be split between the dependent variables and the independent variables. Isolating the independent variables allows us to keep the the information separate from being used in the predictive model, which would negatively affect our accuracy. The split set with the variables are then split again into a Training Set (80%) and a Testing Set (20%) by utilizing the train\_test\_split function from the sklearn library.

**Data Analysis**

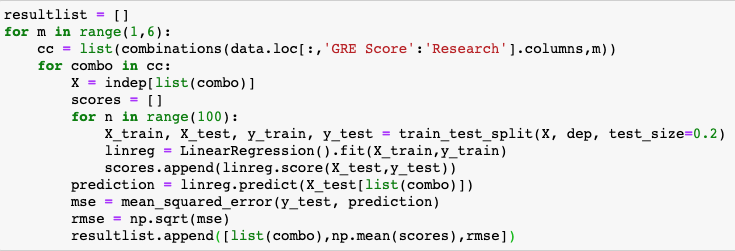
Figure 6: Normalized Data

Estimating the probability of admission can be accomplished by the use of regression. In this project we used three different types: Linear Regression, Random Forest Regression and Extra Trees Regression.

Firstly through Linear regression many possible options were available to form our model. The dependent variable can be estimated through the use of just one dependent variable or many. The combinations of the independent variables were many, they were obtained through the use of the itertools library that provided the combinations() function. By passing the column names from our data set we were returned a set of all possible combinations. We then took those columns and built regression models with each one, collecting a 100 values of each result and taking the mean as the final values for the dependent variable combination. That allowed for our values not to be dependent upon the split of the training and testing sets. The code for the regression can be seen in below.

Figure 7: Code for Linear Regression

From the analysis we selected the top combination of Explanatory Variable, it yielded the highest R^2 Values and 2nd lowest Root Mean Square Error (RMSE). Utilizing this combination we compared it to the other two regression models. Utilizing the sklearn library the implementation of both was simple. Setting up a regression object, fitting the data to the object and then utilizing it to predict upon the test data.

**Conclusion**

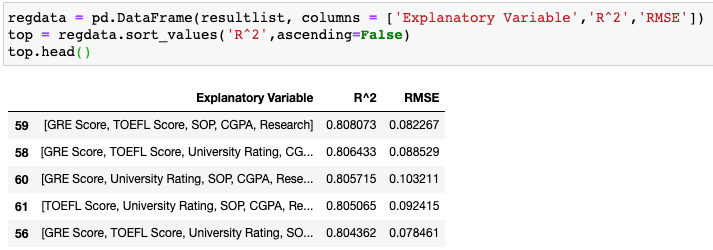


Figure 8: Top 5 Linear Regression Models

Each one of the regression models performed well. The Linear one took the most amount of effort because of the decision to look for the best possible combination. The R^2 value still performed the best and thus the work was worth it, the RMSE for each one of the regressions was around the same space. The results can be seen in Figure 11. In an extensions of the project a possible way to improve upon the model would be to combine the regressors together to supplement each other. Additionally, other methods could be used to estimate the change of admittance. Lastly work could be done to find the mean value of the last 2 regressors through the randomization of training and testing sets.

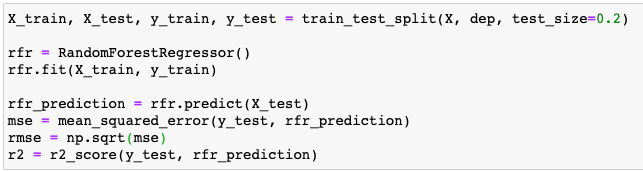


Figure 9: Random Forest Regression Model

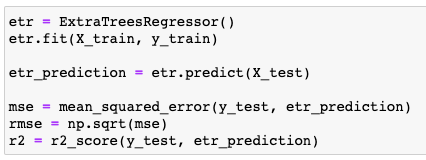


Figure 10: Extra Tress Regression Model





Figure 11: Regression Results

**Resources**

[1] The Commission on the Future of Graduate Education. (n.d.). The Path Forward The Future of Graduate Education in the United States. Retrieved August 07, 2020, from <http://www.fgereport.org/rsc/pdf/CFGE_report.pdf>

[2] Acharya, M. S. (2018, December 28). Graduate Admission 2. Retrieved August 07, 2020, from https://www.kaggle.com/mohansacharya/graduate-admissions