Graduate School Acceptance Prediction

Kevin Dowdy

Abstract

Many of the top Graduate programs in the United States have acceptance rates that are less than 20%. People decide to pursue higher education for many reasons but the numbers suggest that only 1 in 5 will be selected for the program of their choice. The hope of this paper is to analyze the criteria that admissions committees use to make their decision on who makes the cut. Future applicants can use this project to make informed decisions when they begin their process to have accurate predictions for whether they would be admitted or not. For this analysis I use supervised learning models to predict the Chance of Admit for a given candidate.

Introduction

In this paper I will be using Supervised learning techniques to predict how likely it is that an applicant with a particular set of characteristics will be admitted to graduate school. I am using a moderate sized dataset with data points that are found in applicants' records for graduate schools including their GRE scores, undergraduate GPA, the strength of their letter of recommendation and more. The target for the supervised model will be a value between 0.00 and 1.00 that will represent admission likelihood. As part of my analysis I will be using multiple methods to find predictions and compare their accuracy to identify the best model for finding graduate school acceptance predictions. Using classification I will represent scores above 0.5 as ‘Yes’ which means that the person will get in and scores below 0.5 as ‘No’ which means that the student will not get into the school. I will use regression to train the model and predict the Chance of Admit by the value alone.

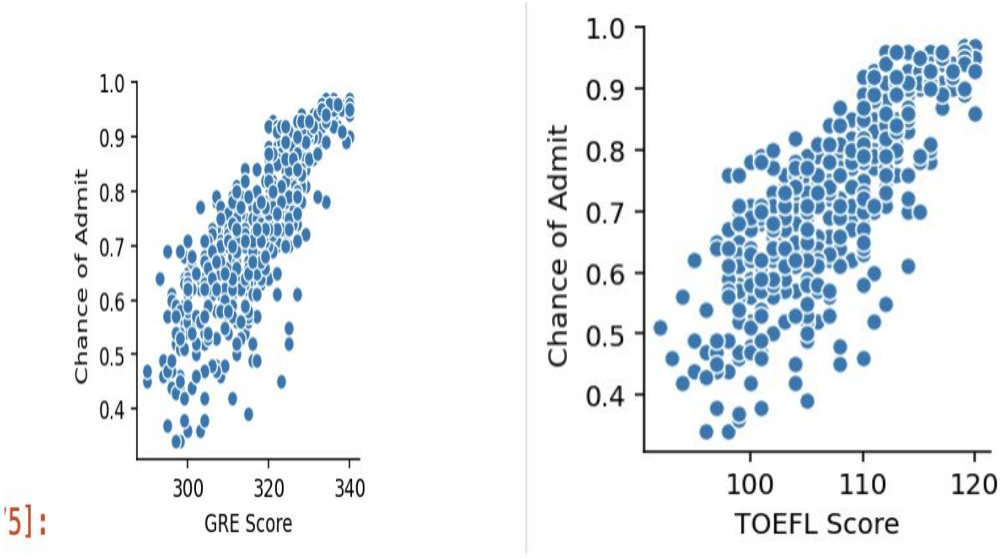
Related Works

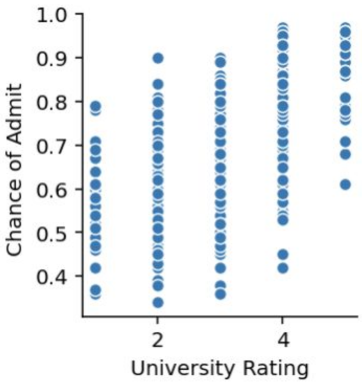
The dataset I will be using in this project has been used in other cases to mine data about different aspects related to graduate school admissions. Researchers on Kaggle used the data to determine how important GRE scores were to the overall admission prospects of a candidate and during my analysis of the data I will be able to identify the correlation between the GRE score and the chance of admission.

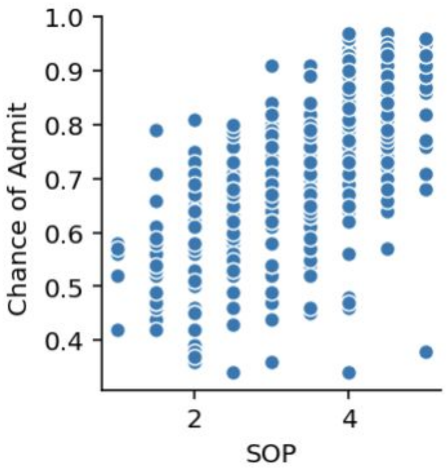
Dataset

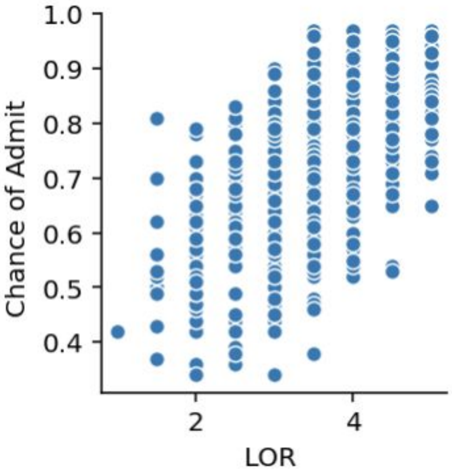
The dataset has 499 rows and 9 columns including the target attribute and the data's index value. For the preprocessing stage of my analysis I will be dropping the indexes as they have no significant impact on the accuracy of the model. Using the Pandas and Sklearn packages I will load and split the data into a training set, testing set and validation set.

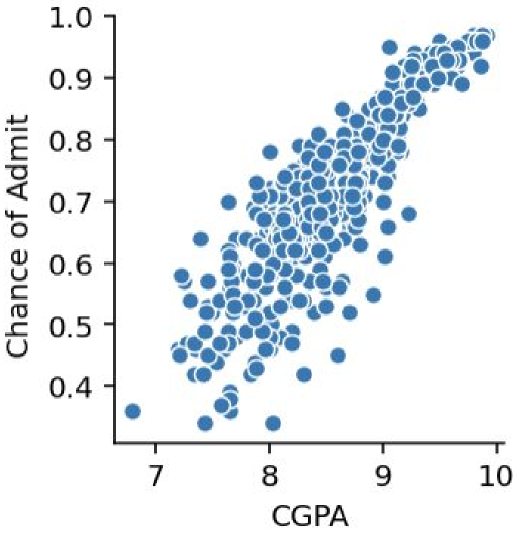
Statistics:

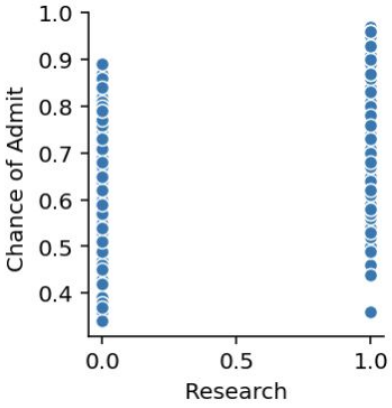




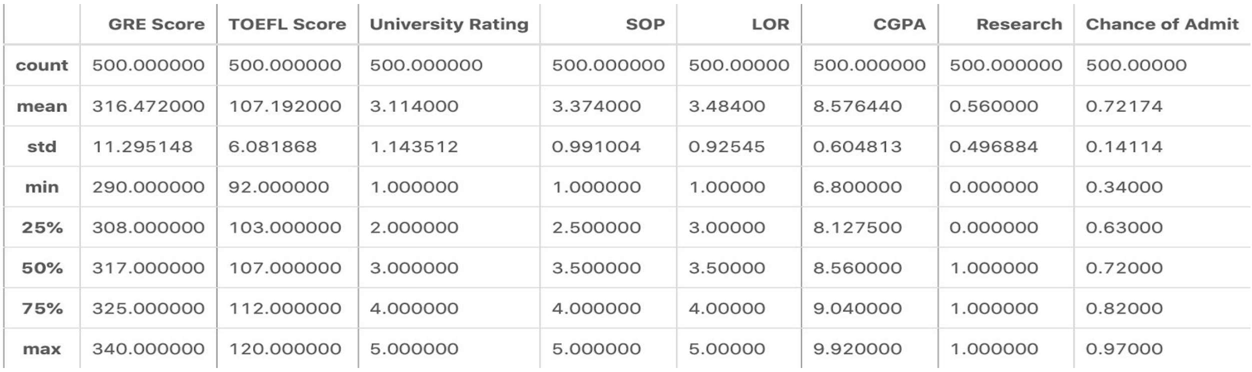




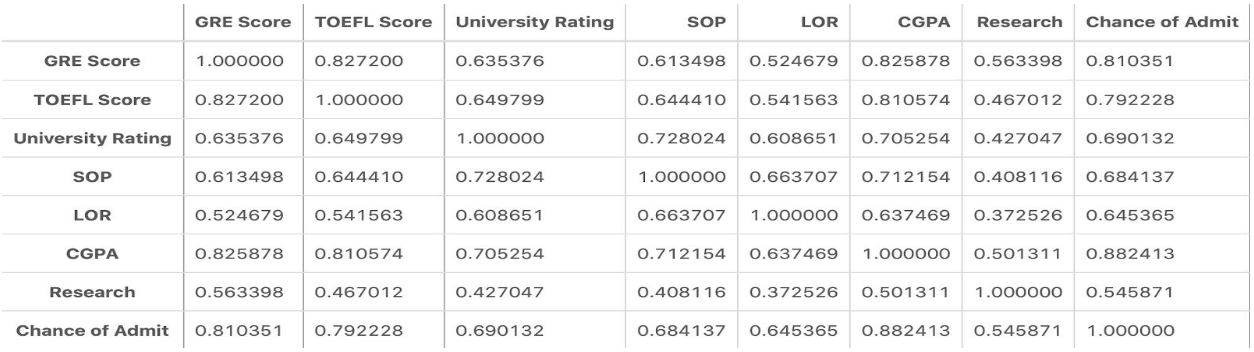




The initial details about the data revealed that there is a very high correlation between the Chance of Admit and the GPA and GRE scores, respectively. This means that those attributes will be important for the model when determining the predicted Chance of Admit.



	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
count	500.000000	500.000000	500.000000	500.000000	500.00000	500.000000	500.000000	500.00000
mean	316.472000	107.192000	3.114000	3.374000	3.48400	8.576440	0.560000	0.72174
std	11.295148	6.081868	1.143512	0.991004	0.92545	0.604813	0.496884	0.14114
min	290.000000	92.000000	1.000000	1.000000	1.00000	6.800000	0.000000	0.34000
25%	308.000000	103.000000	2.000000	2.500000	3.00000	8.127500	0.000000	0.63000
50%	317.000000	107.000000	3.000000	3.500000	3.50000	8.560000	1.000000	0.72000
75%	325.000000	112.000000	4.000000	4.000000	4.00000	9.040000	1.000000	0.82000
max	340.000000	120.000000	5.000000	5.000000	5.00000	9.920000	1.000000	0.97000



	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
GRE Score	1.000000	0.827200	0.635376	0.613498	0.524679	0.825878	0.563398	0.810351
TOEFL Score	0.827200	1.000000	0.649799	0.644410	0.541563	0.810574	0.467012	0.792228
University Rating	0.635376	0.649799	1.000000	0.728024	0.608651	0.705254	0.427047	0.690132
SOP	0.613498	0.644410	0.728024	1.000000	0.663707	0.712154	0.408116	0.684137
LOR	0.524679	0.541563	0.608651	0.663707	1.000000	0.637469	0.372526	0.645365
CGPA	0.825878	0.810574	0.705254	0.712154	0.637469	1.000000	0.501311	0.882413
Research	0.563398	0.467012	0.427047	0.408116	0.372526	0.501311	1.000000	0.545871
Chance of Admit	0.810351	0.792228	0.690132	0.684137	0.645365	0.882413	0.545871	1.000000

The details about the dataset's features and the correlation between the features can also be useful in determining which model to use for the predictions. Before creating the model I scaled the data using the MinMaxScaler from the Sklearn.preprocessing module. Then added a bias column of all ones to my training data through the Sklearn implementation of the models.

Methods

I will be using Linear Regression and Logistic Regression on the dataset to find the best model. Although linear regression is used to find a continuous variable and logistic regression is used to find a categorical variable it will be interesting to see which techniques lead to the best results especially because this dataset is flexible enough to be used for both.

Linear Regression Model

The linear regression model is used to find a continuous variable based on the input and finds the best fit line to make predictions. The independent variables need to be correlated to a dependent variable which I confirmed earlier.

Logistic Regression Model

The logistic regression model is used to find the category that an input would belong to and the variables do not need to have any type of correlation between them.

Results & Findings

The difficult part about comparing the two models' accuracies is that both are using different methods to test the accuracy but I can compare how well they score themselves. After training the models I used their score functions to determine their accuracy. The Logistic model scored itself at 94.6% and the Linear model scored at 84.1%.

The Logistic model the scoring is based on the accuracy of the predicted values compared to the true values of the testing set. The equation for the accuracy is

$$\texttt{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

The Linear model’s score is based on it’s Coefficient of Determination where the equation is

$$R^2 = \frac{SSR}{SST}$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

$$SST = \sum_i (y_i - \bar{y})^2$$

As mentioned before, the variance in the methods being used to test the models makes it difficult to compare them and therefore skews the data results.

Conclusion & Future Works

After training both models I identified that the model built using Logistic Regression was the more accurate model when their scores were compared. In recognizing that this model was better it should also be noted that it was significantly better because most of the specificity was removed from the results. The Chance of Admit which could be any percentage was reflected over a smaller space of 0 or 1 so a 52% chance and 99% chance both became a Yes which may represent the real life circumstances of a candidate. In the future I would adjust the bounds of what constitutes a ‘Yes’ or ‘No’ for the Chance of Admit. I predict that the scores of the Logistic model would then resemble the Linear model’s score.

Citations

[1] Muniz, H. Graduate School Acceptance Rates: Can You Get In?, 2016

[2] Mishra, S. Unsupervised Learning and Data Clustering

[3] Mohan S Acharya, Asfia Armaan, Aneeta S Antony: A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019

[4] Kapil, D. Stochastic vs Batch Gradient Descent

[5] Khan, T. Predicting Graduate Admissions Using Regression