University Admissions

sqz201

April 2019

1 Introduction

The data I have analyzed is from an author Mohan S. Acharya with information connected to the prediction of graduate admissions in Indian Universities. The dataset holds 9 attributes including: serial no., GRE Scores valued out of 340, TOEFL Scores valued out of 140, University Rating valued out of 5, Statement of Purpose Strength valued out of 5, Letter of Recommendation Strength valued out of 5, Undergraduate GPA valued out of 10, Research Experience 1 for True 0 for False, and Chance of Admit. There are approximately 400 rows, which each represent a different applicant.

I am interested in performing a linear regression on the dataset to determine regression best models the relationship between the predictor variables i.e. GRE Scores, Statement of Purpose Strength, Undergraduate GPA, Research Experience, on the outcome variable Chance of Admit.

Acknowledgements This dataset is inspired by the UCLA Graduate Dataset. The test scores and GPA are in the older format. The dataset is owned by Mohan S Acharya.

Inspiration This dataset was built with the purpose of helping students in shortlisting universities with their profiles. The predicted output gives them a fair idea about their chances for a particular university.

Choosing Predictor Variables I decided upon four predictor variables based on looking at correlations between those variables and chance of admit. For instance, I looked at the correlation between GRE Scores and Admit, which stands for Chance of Admit. The correlation between GRE Score and Admit was 0.80261.

I also looked at the correlation betweeen TOEFL Score and Admit. The correlation between TOEFL Score and Admit was 0.791594. Due to discovering

both exam score results high correlation with Chance of Admit. I decided to perform a correlation analysis on both exam score parameters to determine whether Multicollinearity was present. Based on the correlation between GRE Score and TOEFL Score we can see that there was multicollinearity present. In fact, they were highly correlated with eachother with a correlation value of 0.835977

Therefore I decided to choose either GRE Score or TOEFL score. And because GRE Score had a higher correlation with Admit initially, I chose solely GRE Score.

I also looked at the correlation betweeen Letter of Recommendation Strength out of 5 and Admit. The correlation between Letter of Recommendation Score and Admit was 0.669889. This indicated that it could be a potential predictor to use in my model.

I also looked at the correlation between Letter of Recommendation Strength out of 5 and Admit. The correlation between Letter of Recommendation Score and Admit was 0.669889. This indicated that it could be a potential predictor to use in my model.

I also looked at the correlation between Statement of Purpose Strength out of 5 and Admit. The correlation between Statement of Purpose Score and Admit was 0.675732. This indicated that it could be a potential predictor to use in my model. Due to discovering both supplementary strength results high correlation with Chance of Admit. I decided to perform a correlation analysis on both supplementary document parameters to determine whether Multicollinearity was present.

Based on the correlation between Statement of Purpose Strength and Letter of Recommendation Strength we can see that there was multicollinearity present. In fact, they were highly correlated with eachother with a correlation value of 0.729593.

As a result, I decided to choose between one of the supplementary document materials, and ended up deciding to choose Letter of Recommendation Strength because it was initially more highly correlated with Admit.

I looked at the correlation betweeen Research and Admit. The correlation between TOEFL Score and Admit was 0.553202. Due to discovering both research did not have a high correlation with the other predictor variables. I decided that it was valuable to include this in my regression.

Regression Based on the above, I chose the input variables: Research, Letter of Recommendation Strength, GRE Score for my model. The outcome variable was the Admissions Chance, denoted by the column 'Admit'.

I also added a constant coefficient, in order to make the regression model more precise. The resulting model can be described by the equation y = -1.7144 + 0.0071GRE Score + 0.0496LOR + 0.0278Research

The r-squared value of the model is 0.722. This indicates that 72.2 percent of the variation in y, the outcome variable can be explained by the model. The adjusted r-squared value of the model is also approxiamtely the same at 0.72. This adjusted version of R squared has adjusted for the number of predictors in the model. This value decreases when a predictor improves the model by less than expected by chance.

This model is using the least squares method.

Notes These are the notes from the model: [1] Standard Errors assume that the covariance matrix of the errors is correctly specified. [2] The condition number is large, 1.12e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Although I built the model to adjust for multicollinearity, it is clear that multicollinearity or other numerical problems exist in my model.

Biases There are likely biases in the scores of Letter of Recommendation, and Statement of Purposes, because these are subjective components, although they are rated on a scale out of 5. Additionally, there is likely not to be biases in exam scores because those are purely quantitative and cannot be tampered with unless the user incorrectly reports them. For the research predictor, since this is purely bivariate, either the applicant had research or didn't have research experience, there was likely not bias here. For