Shortlisting and Forecasting Graduate Program admissions using Statistic Modelling

Chelsea QianWen Shen

2020-12-06

Abstract

As graduation and admission season coming up, many prospective graduates are interested in pursuing graduate studies, and there are many websites or consultancies that provide general guidences, however, many students are still unsure about the requirements and qualifications that most institution values. Hence, this study utilized statistic techniques such as Logistic Regression and Causual Inference to forecast and analyze based on past admission data which aims to help students shortlisting their choices of universities based on their performance. The result will indicate how likely they are going to be admitted, and based on the result, students will have a better understanding of whether their choice is a safe one or not.

Introduction

Graduate program evaluate the candidacy of each individual based on many factors, such as GPA, research experience and multiple well-recognized graduate admission exam scores such as TOEFL(Test of English as a Foreign Language) and GRE(Graduate Record Examinations) . Therefore, students are not only busy selecting universities that suits their ideal interests, but also required to maintain their best academic performances.

During the application process, students can be indecisive due to the large number of options of institution that offers graduate program. It is important that students are well-informed when making their choices, and hence, in this paper, data gathered from various past students profile are processed and trained with Logistic Regression model which provides explanatory results. The results are aiming to asist students evaluating their schools of choice, and are as expected given a general fair idea of their chance of admission. Further details are explained in the following sections: Section 2 discusses the datasets that we use; Section 3 introduces our model; Section 4 shows the results of our modeling process; and finally Section 5 discusses our results and findings. Appendices additionally contain additional information about (not sure yet), and reference section containing all the sources is at the end.

Section 2: Data

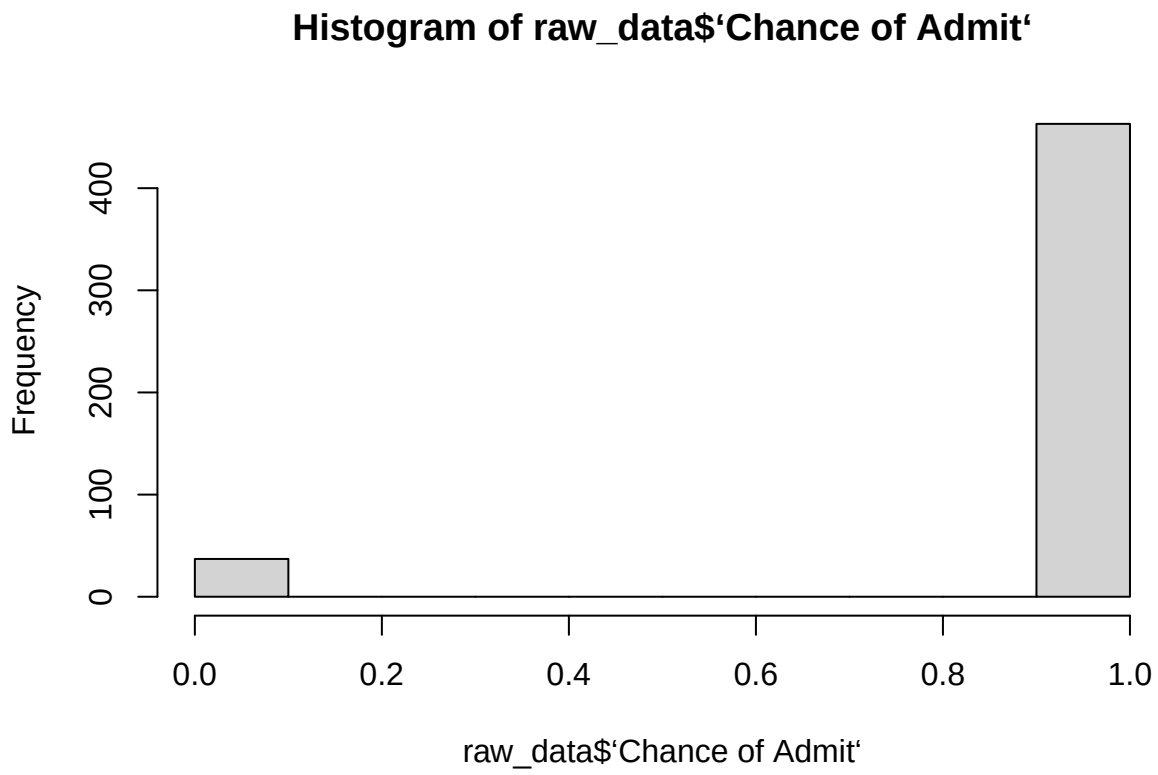
The dataset is retrieved from Kaggle created by Mohan Acharya. The author sampled responses from students of a selected University located in India. The respondents are selected from the Engineering faculty who has pursued a post-graduate education in the United States.After, random data generation was performed using the collected students response. Hence, This dataset is created using students response and also with the help of random generation inspired by UCLA Graduate Dataset.

Since the dataset is collected and randomly generated based on response from engineering faculty of a university in India, therefore the perspective is limited, hence, analysis on this dataset should only be viewed as a minimum guideline to individuals.

There are total of 500 observations on 9 features:

- Serial Number: This is an identifier
- GRE Score: Graduate Record Examinations (out of 340)
- TOEFL Score: Test of English as a Foreign Language (out of 120)
- University Rating: Rank of University (Scale of 1(Top) to 5(Bottom))
- Statement of Purpose (SOP): (Scale of 1 (low strength) to 5 (high strength))
- Letter of Recommendation Strength: (Scale of 1 (low strength) to 5 (high strength))
- CGPA: (out of 10)
- Research Experience: (0 (Unexperienced) or 1 (Experienced))
- Chance of Admit: Chance of the students being admitted (0(Admission probability less than half) or 1 (Admission probability more or equal to half))

Chance of Admit is chosen as the response variable in this paper which represents the the probability of the student being admitted to a particular school given his or her general performance.



The dataset is very inbalanced, Hence we will use Logistic Regression to model the data.

Section 3: Model

Logistic Model will be used to analyze on this data.

Our model equation is :

$$Pr(Admission = 1) = \text{logit}^{-1}\left(\alpha_{a[i]}^{GRE} + \alpha_{e[i]}^{TOEFL} + \alpha_{s[i]}^{SOP} + \alpha_{d[i]}^{LOR} + \alpha_{e[i]}^{CGPA} + \alpha_{s[i]}^{Research} + \alpha_{d[i]}^{Rating}\right)$$

```
knitr::kable(table1)
```

	Intercept	P-value
(Intercept)	-52.4091916	0.0000003
GRE	0.0359491	0.3048276
TOEFL	0.1323475	0.0827822
Rating	-0.2027210	0.4847678
SOP	-0.5835761	0.0713495
LOR	0.8367268	0.0153604
CGPA	3.7169350	0.0000069
Research	0.1009753	0.8580592

Section 4: Results

After conducting logistic regression, we see that the most dominant factor affecting our chance of admission is CGPA and the Letter of Reference strength. Hence, students should focus on their CGPA and try to get a better reference letter in order to get into the university that they want.

Section 5: Discussion

After analyzing on the data, the most dominant factor affecting our chance of admission is CGPA and the Letter of Reference strength.

Appendix

Reference

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

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