Shortlisting and Forecasting Graduate Program admissions using Statistic Modelling

Chelsea QianWen Shen

22 December 2020

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 guidance, however, many students are still unsure about the requirements and qualifications that most institution values. Hence, this study used statistical techniques such as Logistic Regression to forecast and analyze based on past admission data which aims to help students shortlists their choices of universities based on their performance. The result indicates 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.

Keywords: Admission, Master, Graduate, Logistic Regression, Forecasting

Introduction

Graduate programs evaluate the candidacy of each individual based on many factors, such as GPA, research experience and multiple well-recognized graduate admissions 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 suit their ideal interests but also required to maintain their best academic performances.

During the application process, students can be indecisive due to a large number of options of institutions that offers graduate programs. It is important that students are well-informed when making their choices. 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 assist students in evaluating their schools of choice, and are as expected given a general fair idea of their chance of admission.

Although performances of all aspect are considered by the graduate institution, certain requirements are valued more important than others. Our model concluded that factors such

as CGPA, GRE, and Research experience are more significant than the others, students who have excellent achievements on these three measurements had a higher 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 modelling process, and finally, Section 5 discusses our results and findings Reference section containing all the sources is at the end.

Section 2: Data

The dataset is retrieved from Kaggle created by Mohan Acharya. There is a total of 500 observations on 9 features representing a student's general performance. The author sampled responses from students of selected universities located in India. The respondents are selected from the Engineering faculty who has pursued post-graduate educations 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.

Table 1 will show the first 5 observations of the dataset.

Table 1: Table 1

Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
1	337	118	4	4.5	4.5	9.65	1	0.92
2	324	107	4	4.0	4.5	8.87	1	0.76
3	316	104	3	3.0	3.5	8.00	1	0.72
4	322	110	3	3.5	2.5	8.67	1	0.80
5	314	103	2	2.0	3.0	8.21	0	0.65

Since the dataset is collected and randomly generated based on the response from the engineering faculty of selected universities in India. Universities could have different standards and admission requirements for students outside of India. Therefore limited perspective is one of the weaknesses of this dataset. Hence, analysis of this dataset should only be viewed as a minimum guideline to individuals.

2.2 Response

Chance of Admit, renamed as Admission is chosen as the response variable since we are trying to shortlisting and forecasting the probability of a student being admitted to a particular school given his or her general performance. It is a continuous variable ranged between 0 and 1 and has the below statistics

- Minimum: 0.34
- Median: 0.72
- Mean: 0.7212
- Max: 0.97

2.3 Predictors and Covariates

There are many co-variates in this dataset, such attributes are as follows:

- 1. **Serial Number**

  - This is the identification number for each student.

- 2. **GRE Score**

  - Renamed as GRE
  - Graduate Record Examinations
  - Student's score on the Graduate Record Examinations which is (out of 340)

- 3. **TOEFL Score**

  - Renamed as TOEFL
  - Student's score on the Test of English as a Foreign Language Exam (out of 120)

- 4. **University Rating**

  - Renamed as Rating
  - Rating of Undergraduate University
  - Categories of 1 to 5
  - 5: Top rated
  - 1: Lowest rated

- 5. **Statement of Purpose (SOP)**

  - Representing the strength of student's statement of purpose
  - Scale of 1 to 5
  - 1: lowest strength
  - 5: highest strength

- 6. **Letter of Recommendation (LOR)**

  - Representing the strength of student's letter of recommendation
  - Scale of 1 to 5
  - 1: lowest strength
  - 5: highest strength

- 7. **CGPA**

  - Numeric representation of student's cumulative grade (out of 10)

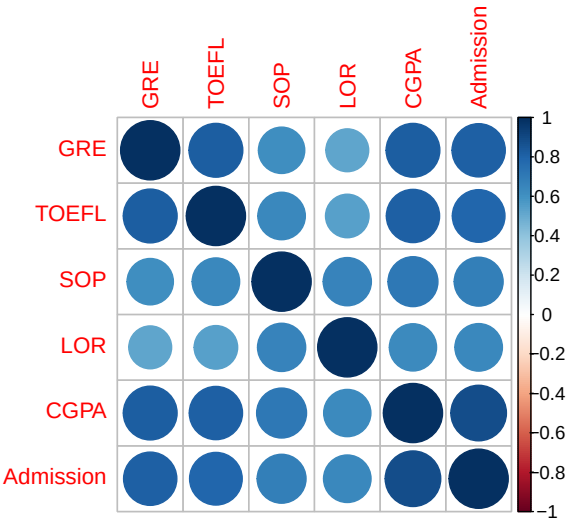
- 8. **Research Experience**

  - Renamed as Research
  - Representing whether student has research experience or not in the past
  - 0: Inexperienced
  - 1: Experienced

2.4 Data Visualization

Figure 1 shows a correlation plot that visualizes the correlation between all variables excluding 'Serial No' because it is an identifier. 'Research experience' and 'University Rating' are also excluded because they are categorical, which will be discussed later. This plot can help us selecting predictors for our initial model.

Figure 1: Correlation plot



According to Figure 1, variables that represent academic performance such as GRE, TOEFL and CGPA are highly-correlated. GRE, TOEFL and CGPA are strongly correlated with the chance of admission, whereas SOP and LOR and moderately correlated with Admission.

After visualizing the numeric variables, we will now look at the categorical variables 'Research experience' and 'University Rating'. Figure 2 shows the box plot between Admission and Research. A pattern can be found that the groups of students in the data-set with past research experience had higher chances of admission, and vice versa.

Figure 2: Box−plot between Research and Admission

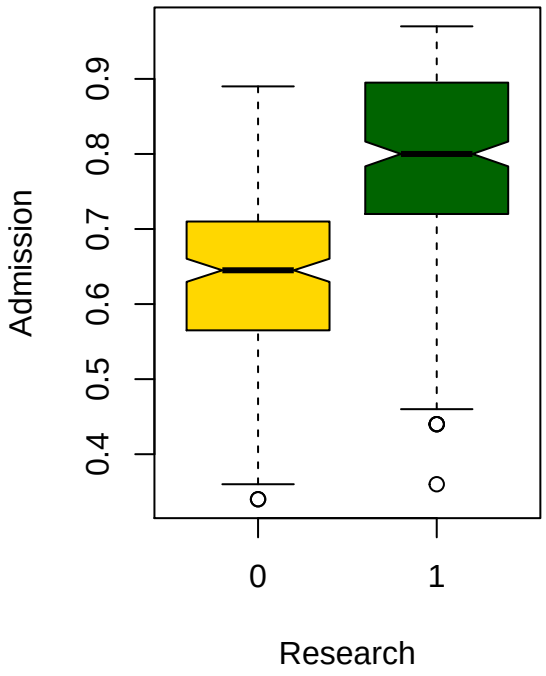
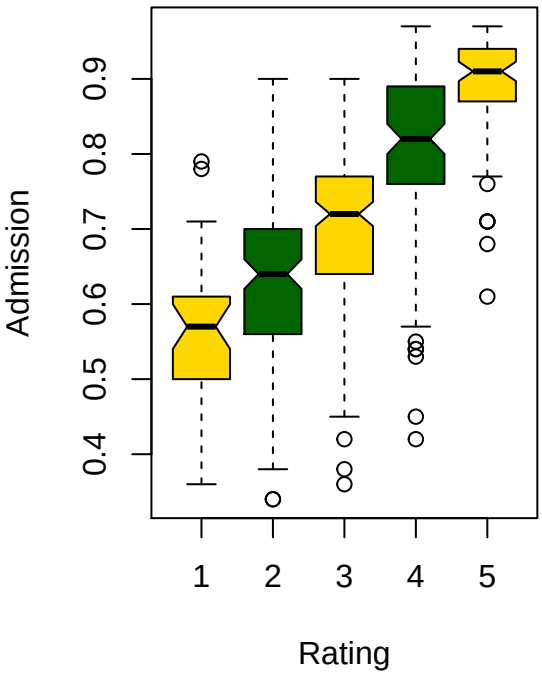


Figure 3: Box−plot between University Rating and Admission



Moreover, Figure 3 shows the box plot between Admission and University Rating. A pattern can also be found that responses from higher-rated undergraduate universities had a higher chance of admission.

Combining our interpretations from Figure 1, 2, and 3, although it does not show causality, but we can still see that all of the variables had various extent of correlations with Admission, therefore, all variables will be considered when modeling.

Section 3: Model

Although the Admission data given in the dataset is a continuous variable, but in reality, admission is a binary result, a student is either admitted or rejected. Hence, the response variable will be dichotomized as "admitted" (represented by 1) and "rejected" (represented by 0). Furthermore, to avoid un-levelled data division, we will use the median 0.72 to classify instead of 0.5. After the data is dichotomized, logistic regression will be used to predict the information because the dependent variable is now binary, observations is not from repeated measurements nor matched data. Hence, for such kind of dataset, logistic regression is most commonly used to analyze the binary response.

Logistic Regression is a statistical model that uses the logistic function to model our binary response variable.

Equation for **Model 1** is :

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

Dataset is being divided into a training set and a test set, and since the original dataset contains only 500 observations which is a relatively small dataset, hence we will use 90% of the data for the training set to fit the prediction model, while test set contains 10% of the data to estimate the accuracy our model.

Section 4: Model Process and Results

The modelling process and results fitted using **R studio** are shown in the tables below

Table 2: Model Summary for Model 1

	Intercept	P-value
(Intercept)	-71.4376173	0.0000000
GRE	0.1052887	0.0018347
TOEFL	0.1375249	0.0152723
Rating2	1.4356020	0.2223492
Rating3	1.3560609	0.2389974
Rating4	1.0032265	0.4181254
Rating5	0.6951522	0.6327127
SOP	0.4431404	0.1701897
LOR	0.2884966	0.3214277
CGPA	2.2646931	0.0009631
Research1	1.2020367	0.0023190

Table 2 shows the coefficients and p-value of the two model. A smaller p-value indicates that there is stronger evidence that the effect of the particular predictor is large on the chance of admission. By setting the threshold to 0.05, it shows that only GRE, TOEFL, CGPA and Research are significant with CGPA being the most significant variable. The estimated coefficients also make sense intuitively, such that all variables have positive coefficients.

Reducing insignificant variables is a method to improve model accuracy. According to Table 2, GRE and TOEFL have very small coefficients, but GRE is more significant according to its p-value. Hence, we will try to fit a new model by removing the TOEFL predictor from model1.

Equation for **Model 2** is :

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

Table 3: Model Summary for Model 2

	Intercept	P-value
(Intercept)	-68.4363250	0.0000000
GRE	0.1354047	0.0000217
SOP	0.5525418	0.0773567
Rating2	1.2992519	0.2272737
Rating3	1.2590349	0.2342812
Rating4	1.0178410	0.3773002
Rating5	0.7194508	0.6068162
LOR	0.2012598	0.4724782
CGPA	2.5165286	0.0001360
Research1	1.1461786	0.0029219

In order to check which model has better performance, an ROC curve will be used to graph the performances of our model. ROC (Reciever Operating Characteristics) curve is a evaluda-tion metric for checking a model's performance which visualizes the results of our prediction.. It can help us to understand how much the model is capable of distinguishing between the classes. The area under the curve known us AUC is valued between 0 and 1, larger area under the curve indicates that our model prediction is more accurate.

Figure 4: ROC curve for Model 1

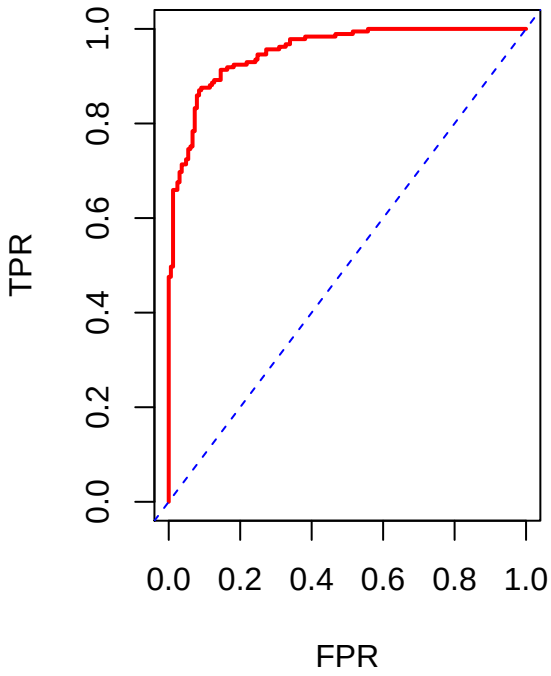
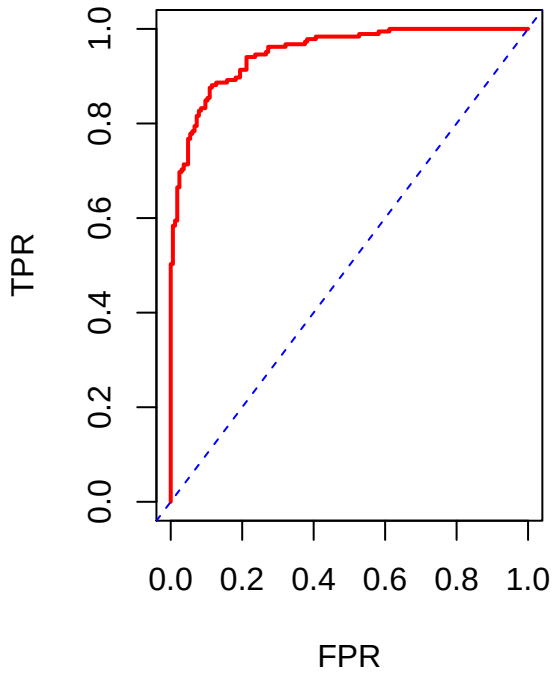


Figure 5: ROC curve for Model 2



After comparison, the difference between Figure 4 and Figure 5 is not significant. Hence we will double check this result with our test data using confusion matrices.

A confusion matrix is a table that is commonly used to describe the performance of a model with similar idea behind the ROC curves, we will construct a confusion matrix on our **test data** to test the accuracy of our model and also verify our result provided by the ROC curves.

Table 3 below shows the confusion matrices for the first models. We can use the calculate the accuracy of our result by dividing the number of prediction that was correct by the total number of test data. After computing, the model accuracy is **83%**.

Table 4: Confusion Matrix for Model 1

Var1	Var2	Freq
0	FALSE	61
1	FALSE	15
0	TRUE	10
1	TRUE	64

Table 5: Confusion Matrix for Model 2

Var1	Var2	Freq
0	FALSE	61
1	FALSE	13
0	TRUE	10
1	TRUE	66

After removing the TOEFL variable, the accuracy yield by our test data with confusion matrix increased to **84%**, with a slight improvement. Therefore, model 2 is selected to predict a student's probability of admission to Graduate programs. Predictors kept in model 2 are GRE score, Statement of Purpose (SOP), Letter of Reference (LOR), CGPA and Research experience.

Lastly, among these predictors, the most significant GRE score and CGPA, with positive coefficients. Since the coefficients in the logistic model provide the change in the log odds of the outcome for a one-unit increase in the predictor variable. We can see that from Table 4:

- For every unit increase in GRE score, the log odds of admission increases by 0.135.
- For every unit increase in CGPA, the log odds of admission increases by 2.52.
- Students who has past research experiences are likely to be admitted by exp(1.146) times than students without.

Section 5: Discussion

In this paper, we have classified the response data 'Admission' into binary response and fitted models using logistic regression with the best accuracy of 84%. The model is constructed and built to help students to forecast their chance of admission based on their various academic performances. We have also visualized our model accuracy with ROC curves, and verified with confusion matrix using test dataset. Our model results have shown that achieving excellence in the measurements above is a good sign of a higher chance of admission, which also makes sense intuitively. The most significant predictor accordig to our model's p-value is GRE score, followed by CGPA and lastly research experience.

One of the caveats and also the weaknesses of this paper is a limited perspective. The origin of the dataset is mentioned in section 2, therefore, evidence and model result generated by this dataset is optimal when forecasting from an Indian perspective pursuing graduate studies in a US institution. It is unlikely that graduate institutions disregard cultural and regional differences and have equal admission requirements for undergraduate universities all around the world, hence it does not give a valid inference about the whole population and the accuracy of our model could be different when performing in the large world.

Other weaknesses of this paper is that the response variable was dichotomized into two outcomes using the median to reflect the admission result in reality, however, this could lead to some information loss, and regression discontinuity which could problems when estimating for causal effect.

With consideration of the limitations, the model can still provide students with a minimum guideline when preparing for their graduate admissions. Certain explanatory variables carried more weight according to our model results such as CGPA and GRE score and Research experience. Based on our model, research experienced student with high CGPA and GRE score has a higher chance of admission.

For future work, other than finding a more representative data, another opportunity to improve the paper, specifically our model is to create dummy variables for the categorical variable such as Undergraduate University Rating and treat each factor as it's own when modelling. For example, when reducing the predictors when modelling, we could reduce only certain categories of the rating variable instead. Other statistic models other than logistic regression could also be used to forecast and select the model with best accuracy. Also, many other factors that could affect a student's chance of admission are not considered such as extracurricular achievements.

Github

Code and data supporting this analysis is available at: https://github.com/choshi123123/admission

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.

Mohan S Acharya. Graduate Admission 2, version 1.URL https://www.kaggle.com/mohansacharya/graduate-admissions.

Hosmer, D. & Lemeshow, S. (2000). Applied Logistic Regression (Second Edition). New York: John Wiley & Sons, Inc.

Long, J. Scott (1997). Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage Publications.

UCLA, Statistical Consulting Group: LOGIT REGRESSION | R DATA ANALYSIS EXAMPLES. URL https://stats.idre.ucla.edu/r/dae/logit-regression/.

Andrew Gelman and Yu-Sung Su (2020). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.11-2. https://CRAN.R-project.org/package=arm

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Taiyun Wei and Viliam Simko (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77 http://www.biomedcentral.com/1471-2105/12/77/