Justifying Higher Education: An Examination of Graduate Level Admission and their Qualifying Variables

Andrew Holland, Dario A. Rodriguez

Economics Department, Eller College of Management, University of Arizona

ECON 453 Data Analytics and Modeling: Quantitative Analysis for Economic Strategy

Christian Cox

May 1, 2024

Abstract

The purpose of this research paper is to investigate and identify any significant trends within a set of data that discerns predictors of admission into graduate school when observing different parameters. These parameters include the individual's GRE score, TOEFL score, their university rating, a statement of purpose (SOP), letters of recommendation (LOR), cumulative grade point average (CGPA), and whether they participated in research. Through analytical research, this research paper explores the inputs and the effect on those inputs on an individual's academic success. The authors of this paper (Holland, Rodriguez) have shown interest in attending graduate school after they finish their undergraduate studies. Insights gained from this study will help them, and other prospective students, in optimizing and influencing their chances of getting admitted to a graduate program of their choosing.

Background

The process of any graduate programs, from the application to the completion of this academic milestone, are characterized by their demanding nature, rigorous coursework, and highly competitive stressful environment with high expectations. Academic institutions typically seek individuals who demonstrate a combination of many things including academic resilience, relevant experience, desire to work in the subject matter, and cognitive ability. During the process of applications, there are many characteristics that institutions look for, but they tend to focus on most of those that encompass one's academic standing and cognitive ability. One of these measures is the Graduate Records Assessment (GRE) which is a standardized test that is part of the admissions process for many graduate schools (Wikipedia, 2024). Another measure for receiving an offer for admission to a graduate program is cumulative grade point average (CGPA), which is a numerical representation of a student's academic performance in their undergraduate studies. For both GRE scores and CGPA, when analyzed independently, there is a linear relationship that indicates that the higher this measure is (GRE or CGPA), the greater chance they have of being admitted into a program that they apply for (Colarelli, Monnot, Ronan & Roscoe, 2011) As of late, many academic institutions, particularly graduate schools of business, establish a habit of running statistical analyses of an individual's ability to predict classroom performance solely based on their grade records (Page, West, 1969). During our research, the collection of our findings we aim to address the following inquiries:

- Which, if any, are the most influential factors in graduate school admissions across different measure of academic performance?
- To what extent do academic credentials, including CGPA and standardized test scores, predict an individual's chance of being admitted?
- What role does research experience play in the admissions process? Is there any significance?
- During the graduate admissions process, does the rating of the university impact their chance of getting accepted?
- Are there any other underlying trends that this analytical analysis brought light to?

The admission process is multifaceted and is not indicative of just one or two academic indicators, but the higher your controllable indicators be, the higher chance you should have at getting admitted. In this paper, we aim to theorize a model that can provide insights to prospective students, to assist them in any way possible, so that they can purse higher education beyond their undergraduate education.

Literature Review

Consistent with the research done for this paper, previous research has explored the predictability of acceptance into a graduate program based on one's academic performance. These studies include machine learning approaches (Zhao, Chen, Xue &Weiss, 2023) as well as the comparison of regression models based on different measures of academic performance (Acharya, Armaan & Antony 2019). In the first study mentioned, the researchers focused only on the impact that the letter of recommendation (LOR) would have on an individual's acceptance. This study focused on examining the effectiveness of a machine learning approach for programs at the master's level. Their data was divided up to observe two key differences: those who did and did not have and LOR for this application. In summary, they were able to collect insights with their predicative models, constructed with different features, that underscored the significance of LORs in the admissions process. This differs from the research done for this paper because we are looking at the impact on a combination on different academic measures (and LORs and SOPs), while the study mentioned only focused on LORs. In the second study mentioned, the researchers created 4 different regression models (linear regression, support vector regression, decision tree regression, and random forest regression) to conclude which model works better to predict academic performance based on different measures of academic success, based on GPA and test scores. The model that was able to predict the best, based on data that they already collected, was the linear regression model with the random forest regression as a close second. This study focused on finding the best-fit model to predict admission rate. They concluded that higher test scores, higher GPA, and along with other factors, result in a greater chance of admission.

Objective & Hypotheses

From this data, we hoped to better understand what makes a student more likely to be admitted into a graduate program. This can be modelled with the general hypothesis in the form of:

$$H_0: There\ is\ no\ significant\ relationship\ between\ each\ predictor\ variable\ and\ admission\ chance$$
$$H_A: There\ is\ a\ significant\ relationship\ between\ at\ least\ one\ predictor\ variable\ and\ admission$$

We also wanted to see what qualities about a university (like rating) impact how many students get admitted into a program. During a preliminary look through the data, we found there is some kind of relationship between the CGPA of a candidate and their admissions chance. Thus, the interaction of GPA with admission chance and other variables are of interest to the research as well. One of hypotheses can be written as:

$$H_0: \beta_{GPA} = 0$$

$$H_A: \beta_{GPA} \neq 0$$

This hypothesis allows us to prove if there is a statistical significance between admission chance and GPA. Furthermore, we also want to test if there is interaction between different variables. These interactions may yield greater statistical significance for different variables and demonstrate the relationship that interaction has with admissions chance:

$$H_0: \beta_{Rating*SOP} = 0$$
$$H_A: \beta_{Rating*SOP} \neq 0$$

The relationship between the different predictor variables in the dataset may be nonlinear as well. Thus, we need to test the hypothesis of:

$$H_0: The\ relationship\ between\ GPA\ and\ admissions\ chance\ is\ linear$$
$$H_A: The\ relationship\ between\ GPA\ and\ admissions\ chance\ is\ nonlinear$$

Finally, we have one binary variable. We should test the impact this predictive binary variable has on the dataset as a control variable:

$$H_0: The\ predictor\ variable\ "Research"\ has\ no\ statistically\ signifcant\ effect\ on\ admission$$
$$H_A: The\ predictor\ variable\ "Research"\ has\ a\ statistically\ signfnificant\ effect\ on\ admission$$

Proposed Models

There are 7 different variables in the dataset. These include CGPA, GRE score, TOEFL score, university rating, statement of purpose strength, letters of recommendation, and if the candidate has research experience or not. The proposed models are based on preliminary analysis of these variables. Firstly, testing all variables in a linear model can display the impact of the data collected:

$$Admission = \beta_0 + CGPA\ \beta_1 + GRE\ \beta_2 + TOEFL\ \beta_3 + Rating\ \beta_4 + SOP\ \beta_5 + LOR\ \beta_6 + Research\ \beta_7$$

This linear model creates a baseline for all testing going forward. To derive further models, scatterplots of different variables were generated. The first of these new models is based on *Figure I*, the relationship between "Chance of Admission" and "CGPA". Based on this distribution, there could be an exponential relationship with GPA in the data. Thus, an exponential model should be tested:

$$\log(Admission) = \beta_0 + CGPA\ \beta_1$$

Another exponential model should also be tested, one with all the other predictive variables, to understand the marginal impact of GPA and its possible nonlinear relationship:

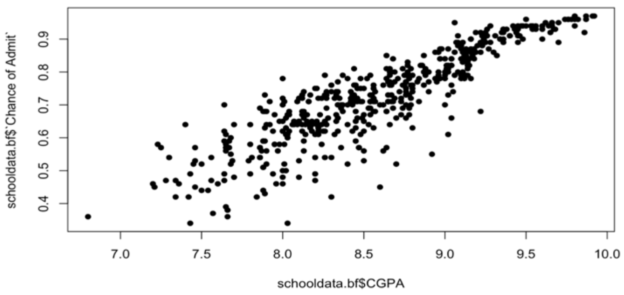


Figure I: Chance of Admission and CGPA score

$$\log(Admission) = \beta_0 + CGPA\ \beta_1 + Research\ \beta_2 + GRE\ \beta_3 + TOEFL\ \beta_4 + Rating\ \beta_5 + SOP\ \beta_6 + LOR\ \beta_7$$

A logit model would not make sense here, as the chance of admission is not expressed as a binary variable. The percentage chance of admission is already established in the model, and thus it's unnecessary to make a logit model. The next series of models, however, are different linear models to chart the impact of different variables. This would test for collinearity and display the marginal effect variables have with/without other predictive variables. This includes models like:

$$Admission = \beta_0 + CGPA\ \beta_1$$

$$Admission = \beta_0 + CGPA\ \beta_1 + GRE\ \beta_2$$

$$Admission = \beta_0 + SOP\ \beta_5 + LOR\ \beta_6 + Research\ \beta_7$$

Finally, in the event there is an interaction between two variables, a linear model with an interaction would be tested against the other models. This can take the form of:

$$Admission = \beta_0 + Rating * SOP\ \beta_1 + CGPA\ \beta_2 + GRE\ \beta_3 + TOEFL\ \beta_4 + LOR\ \beta_5 + Research\ \beta_7$$

Data

The dataset that we used for this research paper is based on data that was collected showing the individuals academic performance and their chance of getting admitted. This data set was collected from *Kaggle,* which is an online platform allowing its users to find datasets. Kaggle is also used to build AI models, publish datasets, work with data scientists/machine learning engineers, and participate in competitions to solve challenges present in data science. (Kaggle, Mahmoud Hassasn, 2022). When we downloaded the data from this database, we didn't need to clean much since everything we needed to observe was already present. The only data modification we did was standardizing the CGPA to a 4.0 scale. Since this data is based on Indian student's admission to a graduate program, their CGPAs were based on a 10.0 scale. We standardized this variable because the 4.0 scale is the most common GPA structure (Claybourn, U.S. News, 2022). To normalize the data into the 4.0 scale, we divided every datapoint by 10 and multiplied by 4. This then would chart an identical distribution when compared to the original CGPA. CGPA will now be referred to as GPA when undergoing analysis The data used can be described as low volume data since there are only 500 entries observing only 9 different variables. The variables observed can be categorized as: sequential integers, quantitative response variables, ordinal, and continuous.

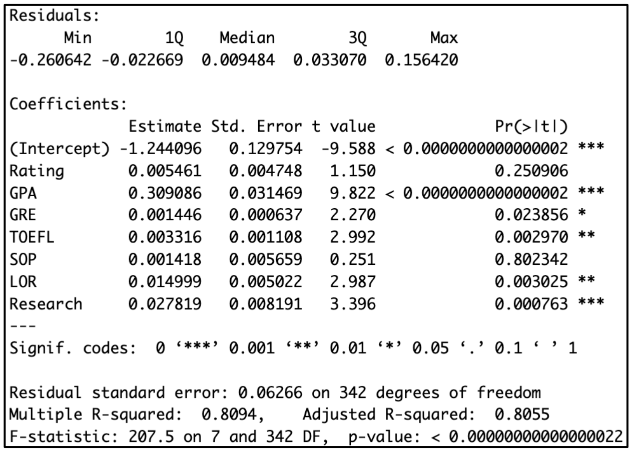
Empirical Methodology & Estimation Results

To begin analysis, the order of data records was randomized based on the arbitrary seed (828). From here, we split the data into testing and validation sets [350,150] for the application of the holdout method for robust model building. We then built 10 linear regression models and 2 exponential models. The linear models were tested against each other using their adjusted R^2 and the exponential models were tested against the best of the linear models and each other using the MSE, MAE, and MAPE measures. Of the 10 linear models tested, the first two proved to be the most robust with the highest adjusted R^2 of all the combinations. These models are as follows:

$$Admission = \beta_0 + GPA\,\beta_1 + GRE\,\beta_2 + TOEFL\,\beta_3 + Rating\,\beta_4 + SOP\,\beta_5 + LOR\,\beta_6 \\ + Research\,\beta_7$$

$$Admission = \beta_0 + Rating * SOP\,\beta_1 + GPA\,\beta_2 + GRE\,\beta_3 + TOEFL\,\beta_4 + LOR\,\beta_5 \\ + Research\,\beta_6$$

The interaction between the Rating and SOP predictive variables was found following the output of the above model, hereby referred to as “Model 1.” The model below it will be referred to as “Model 2” for the remainder. Model 1’s output (*Figure II*) illustrates lacking significance for the Rating and SOP predictive variables. Implementing an interaction between these two variables not only increases their relative significance, but also improves the adjusted R^2 of the model, now Model 2 (*Figure III*).

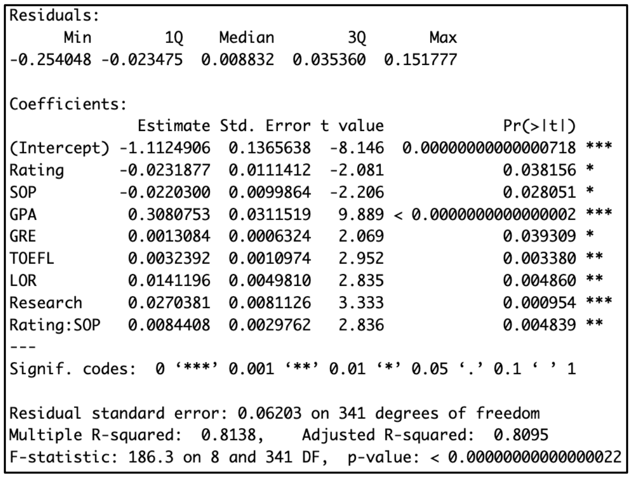


```
Residuals:
      Min        1Q    Median        3Q       Max
-0.260642 -0.022669  0.009484  0.033070  0.156420

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) -1.244096   0.129754  -9.588 < 0.0000000000000002 ***
Rating       0.005461   0.004748   1.150             0.250906
GPA          0.309086   0.031469   9.822 < 0.0000000000000002 ***
GRE          0.001446   0.000637   2.270             0.023856 *
TOEFL        0.003316   0.001108   2.992             0.002970 **
SOP          0.001418   0.005659   0.251             0.802342
LOR          0.014999   0.005022   2.987             0.003025 **
Research     0.027819   0.008191   3.396             0.000763 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06266 on 342 degrees of freedom
Multiple R-squared:  0.8094,    Adjusted R-squared:  0.8055
F-statistic: 207.5 on 7 and 342 DF,  p-value: < 0.00000000000000022
```

Figure II: Model 1 R Output



```
Residuals:
      Min        1Q    Median        3Q       Max
-0.254048 -0.023475  0.008832  0.035360  0.151777

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept) -1.1124906  0.1365638  -8.146  0.00000000000000718 ***
Rating      -0.0231877  0.0111412  -2.081             0.038156 *
SOP         -0.0220300  0.0099864  -2.206             0.028051 *
GPA          0.3080753  0.0311519   9.889 < 0.0000000000000002 ***
GRE          0.0013084  0.0006324   2.069             0.039309 *
TOEFL        0.0032392  0.0010974   2.952             0.003380 **
LOR          0.0141196  0.0049810   2.835             0.004860 **
Research     0.0270381  0.0081126   3.333             0.000954 ***
Rating:SOP   0.0084408  0.0029762   2.836             0.004839 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06203 on 341 degrees of freedom
Multiple R-squared:  0.8138,    Adjusted R-squared:  0.8095
F-statistic: 186.3 on 8 and 341 DF,  p-value: < 0.00000000000000022
```

Figure III: Model 2 R Output

To confirm the robustness of either both Model 1 and 2, each was tested against the validation set and fit on a scatterplot (*Figure IV, Figure V*).

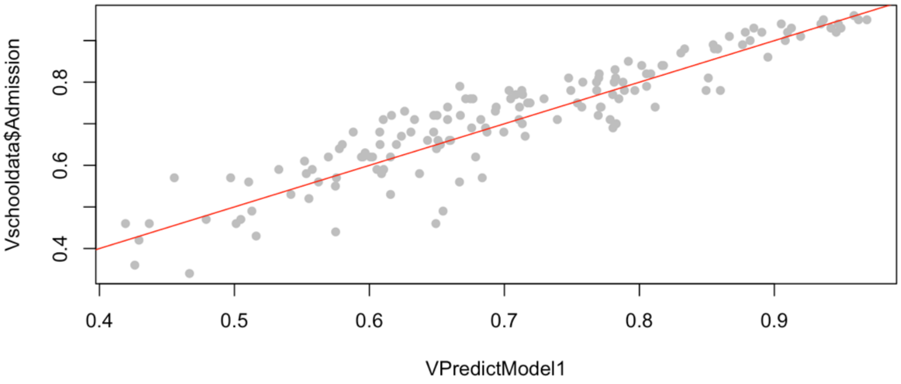


Figure IV: Model 1 Predictive Fit

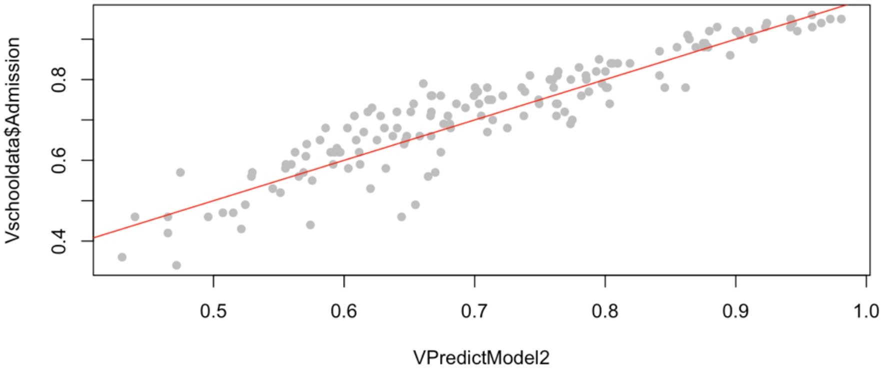
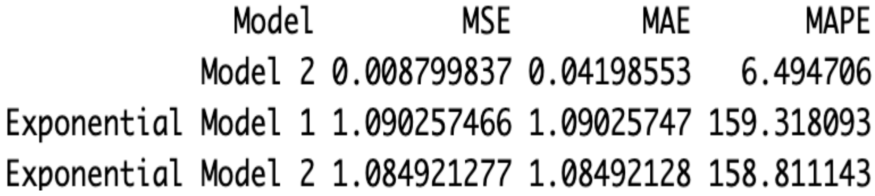


Figure V: Model 2 Predictive Fit

When considering the exponential models, they are based on the forms:

$$\log(Admission) = \beta_0 + GPA\,\beta_1$$

$$\log(Admission) \\ = \beta_0 + GPA\,\beta_1 + Research\,\beta_2 + GRE\,\beta_3 + TOEFL\,\beta_4 + Rating\,\beta_5 + SOP\,\beta_6 \\ + LOR\,\beta_7$$



```
                Model         MSE        MAE       MAPE
              Model 2 0.008799837 0.04198553   6.494706
Exponential Model 1 1.090257466 1.09025747 159.318093
Exponential Model 2 1.084921277 1.08492128 158.811143
```

Figure VI: MAE, MSE, MAPE Output

The first of the models was to simply test the marginal effect of GPA, however the second model, referred to as “E-Model 2”, was developed to explore the fit of the possible exponential fit of the GPA predictive variable. To see the robustness of E-Model 2 against the two linear models, the MAE, MSE, and MAPE were tested against each other. Based on the output (*Figure VI*), Model 2 becomes the most robust model against both the other linear model and the exponential model. Model 2 has the lowest MSE. MAE, and MAPE. This is expected, given the intuitive expectation that increased scores on all these predictor variables would lead to increased admission chance. There is an upward sloping relationship by some combination, and Model 2’s fit is the best of the models.

Summary and Implication

The research done for this paper was to observe and investigate the predictors of admission into a graduate program and potentially provide insights for prospective students to enhance their chances to pursue a higher level of education, beyond their undergraduate studies. Through thorough analysis, we were able to shed light on the complex, and sometimes stressful, process of graduate school admissions. First off, our research supported how significant measures of academic success – such as CPGA, GRE/TOEFL scores – are when predicating the probability of acceptance. In line with our analysis, higher score in these academic metrics resulted in greater chances of acceptance. Furthermore, we found that non-academic metrics, such as activity in research, SOPs, and LOPs also played crucial roles in the admissions process. Applicants with research experience, as well as higher rated SOPs and LORs, were significant factors that influenced their rate of acceptance. With a higher rate on those that had research experience, this suggests that graduate programs value applicants with practical skills and scholarly engagement. Additionally, graduate programs seeking higher quality of SOPs and LORs concludes that communication and personal branding and impact is a relevant factor. Despite these insights, our research comes with limitations. Since our analysis was based on a specific dataset of Indian graduate school applicants, generalizing this model could be difficult through different groups of applicants with different cultural or contextual factors. As a result of this, there is a necessity of research to validate our implications across diversified populations. Another limitation present within our research is that our analysis mainly focused on qualitative observations. By virtue of this, we overlooked potential qualitative observations (i.e., impact on interviews, applicant portfolios, etc.) of the admissions process. To conclude this paper, our study aimed to fundamentally support and generate an understanding of factors that influence graduate school admissions. By factoring in multiple measures of academic success, future applicants can utilize this information and tailor their applications and habits in undergraduate to increase their probability of acceptance.

References

Acharya, M. S., Armaan, A., & Antony, A. S. (2019). A comparison of regression models for prediction of graduate admissions. *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*. https://doi.org/10.1109/iccids.2019.8862140

Claybourn, C. (2022, July). *What students should know about the GPA scale | best colleges | U.S. news*. U.S. News. https://www.usnews.com/education/best-colleges/articles/what-students-should-know-about-the-gpa-scale

Colarelli, S. M., Monnot, M. J., Ronan, G. F., & Roscoe, A. M. (2011). Administrative assumptions in top-down selection: A test in graduate school admission decisions. *Applied Psychology*, *61*(3), 498–512. https://doi.org/10.1111/j.1464-0597.2011.00480.x

Hassasn, M. (n.d.). *What is a kaggle?*. Kaggle. https://www.kaggle.com/discussions/general/328265

Page, A. N., & West, R. R. (1969). Evaluating student performance in Graduate Schools of Business. *The Journal of Business*, *42*(1), 36. https://doi.org/10.1086/295163

Zhao, Y., Chen, X., Xue, H., & Weiss, G. M. (2023). A machine learning approach to graduate admissions and the role of letters of recommendation. *PLOS ONE*, *18*(10). https://doi.org/10.1371/journal.pone.0291107