# Will I Get In?: A Predictive Model of Graduate Admission at a US University

Amin Yakubu[1], Matthew Perrotta[1], Jyoti Ankam[2]

[1]Columbia University Mailman School of Public Health, Department of Epidemiology

[2]Columbia University Mailman School of Public Health, Department of Biostatistics

*Introduction*

Every year tens of thousands of students apply to American universities, and in the due process discover a plethora of unreliable sources to guide them in their decision-making processes. While most graduate school applicants are domestic students, US remains the top destination for international students as per a UNESCO report. Being an educational hub, US graduate school admissions committees try to select the most promising students. Every student that passes through the doors of a university has an official transcript that registers every academic aspect of their undergraduate experience. This includes their cumulative GPAs, standardized test scores (GRE and TOEFL), personal statements, prior research, and letters of recommendations. Universities, then, evaluate application packets based on rules or heuristics unknown to students, and release decisions. Owing to the tremendous amount of uncertainty, and the lack of information on the selection criteria which are generally university-specific, students tend to apply to a large number of universities based on hearsay or semi-informed opinions, resulting in confusion and a waste of resources for both the applicant as well as the university. Thus, in the current competitive scenario, students are often left pondering over this question: What exactly do graduate admissions committees look for in potential graduate students?

In this paper, we address this problem by developing a machine learning approach which enables applicants to make informed decisions by evaluating their chances of admission. American graduate school programs have been capturing incoming student data that can help for years. We were interested in analyzing graduate admission records in one particular US university to predict admission decisions. We used records from a hypothetical US university that was inspired by the UCLA Graduate Dataset and made available on Kaggle. The purpose of our study is to help students shortlist universities with their profiles. The predicted output would give them a fair idea about their chances for this university.

We used study variables which are considered most important during the application for Masters Programs. The dataset consists of 8 predictors and 1 response variable. The parameters included were: Serial No., Graduate Record Examinations (GRE) Scores (out of 340), Test of English Foreign Language (TOEFL) Scores (out of 120), University Rating (out of 5), Statement of Purpose (SOP), Letter of Recommendation (LOR) Strength (out of 5), Undergraduate cumulative

GPA (cGPA) (out of 10), and Research Experience (either 0 or 1). The response variable is Chance of Admit (ranging from 0 to 1). Of note, the test scores and GPA were in the older format. We analyzed our system from students' perspective, but it can be easily extended to university's perspective as well. In cleaning and pre-processing the data, we removed the variable Serial No. as this had no value in the analysis. The dataset contained no missing values and no linear dependencies.

*Exploratory Data Analysis/Visualization*

The dataset contains 500 observations and 7 predictors. Of these, cumulative GPA (cgpa), GRE scores (gre_score), TOEFL scores (toefl_score) are continuous and statement of purpose (sop), letter of recommendation (lor), university rating, and research are categorical. The continuous variables 'cgpa', 'gre_score', and 'toefl_score' all have a positive linear relationship with the outcome. This also appears to be the case with the categorical variables, with the response increasing on average with each increase in predictor value.

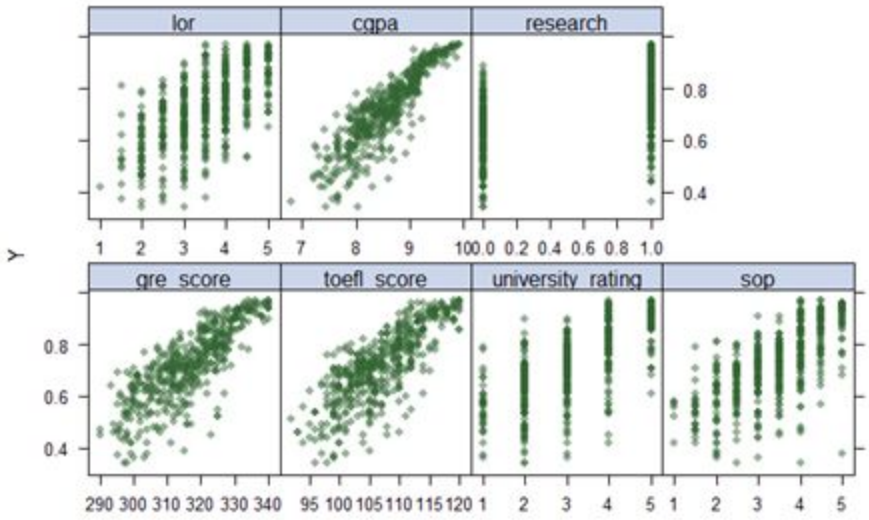


Figure 1. Predictors plotted against the response variable, 'chance_of_admit'

*Models*

The outcome of interest, 'chance_of_admit' is a continuous probability between 0 and 1. However, rather than categorizing it we decided to treat this as a continuous variable and therefore used linear regression methods for prediction. Although they may be biased from the truth (give a prediction outside the bounds of 0 and 1) linear models can still provide some useful information. The following methods were used; multiple linear regression, lasso, ridge, PCR,

elastic net, GAM, and Multivariate Adaptive Regression Splines (MARS) for predictive modeling.

The first model that was run was a multiple linear regression, following which the assumptions of linear regression were tested; linearity, independence, normality of residuals, and homoscedasticity.

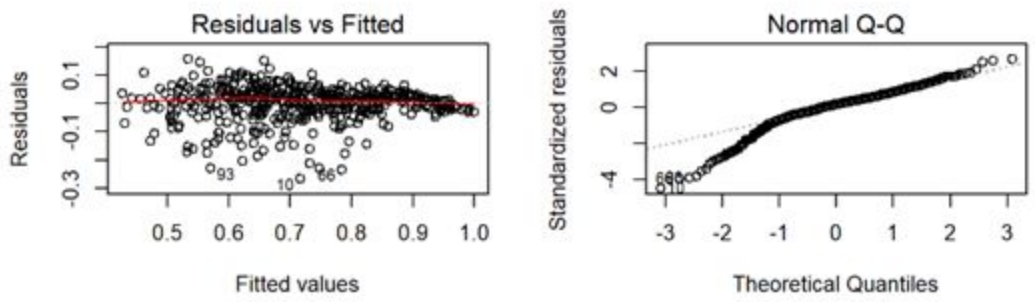


Figure 2. left plot: residuals vs. fitted values, checking homoscedasticity; right plot: Q-Q plot, checking normality

The data generally meets the assumptions of linearity, normality of residuals and homoscedasticity. In order to check independence of the predicators, a correlation plot was created. The plot displayed a slight collinearity between all predictors. This is expected as each predictor is a measure of academic prowess, and if one is high the other predictors are likely to be elevated. After running the multiple linear regression models, the variables 'toefl_score' and 'university_rating' were removed as their p values indicated they were not significantly associated to the chance of admittance. Also, we exclude these variables because adding them increases the variance without a corresponding decrease in the bias (Bias-variance tradeoff).

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -1.2757251 | 0.1042962 | -12.231747 | 0.0000000 |
| gre_score | 0.0018585 | 0.0005023 | 3.699832 | 0.0002401 |
| toefl_score | 0.0027780 | 0.0008724 | 3.184185 | 0.0015436 |
| university_rating | 0.0059414 | 0.0038019 | 1.562754 | 0.1187533 |
| sop | 0.0015861 | 0.0045627 | 0.347635 | 0.7282630 |
| lor | 0.0168587 | 0.0041379 | 4.074273 | 0.0000538 |
| cgpa | 0.1183851 | 0.0097051 | 12.198187 | 0.0000000 |
| research | 0.0243075 | 0.0066057 | 3.679753 | 0.0002592 |

Table 1. coefficient estimates for multiple linear regression

With these two variables removed, the models fitted in this project used gre_score, toefl_score, lor, cgpa, and research.

We used a repeated cross validation for 5 times with 10 folds. After each repetition of the cross-validation, the model assessment metric is computed using the RMSE. The scores from all repetitions are finally averaged to get a final model assessment score. This gives a more "robust" model assessment score than performing cross-validation only once.

In order to fit a lasso, ridge, and elastic net regression a tuning parameter (lambda) needed to be selected. This lambda was found using cross-validation methods. The parameter which provided the lowest cross-validation RMSE was selected for each model. The tuning parameters for lasso, ridge and elastic net are 0.00061929, 0.01142031, 0.1353353 respectively.

Amongst the models, GPA was found to be the most important factor based on the estimates. This could be due to the fact that a 1 (example from 2.0 to 3.0) unit increase in GPA is more significant in terms of academic success than a 1 unit increase in TOEFL score (120 to 121).

Linear regression modeling is one of the ways of demonstrating predictive relationships in data analysis. However, just like other methods, linear regression has certain limitations; one being it's high sensitivity to outliers. If most of the data lives in the narrower range on the x-axis, but there are one or two points far out, it could significantly swing regression results. In our study, we did not encounter any significant outliers. Secondly, simple linear regression is prone to overfitting such that the regression begins to model the random error (noise) in the data, rather than just the relationship between the variables. A model that has learned the noise instead of the signal is considered "overfit" because it fits the training dataset but has poor fit with new datasets. This most commonly arises when there are too many parameters compared to the number of records which fortunately wasn't the case in our study. To prevent overfitting we employed repeated cross-validation which allowed us to tune hyperparameters with only our original training set and allowed us to keep our test set as a truly unseen dataset for selecting our final model. Thirdly, linear regressions are meant to describe linear relationships between variables. So, if there is a nonlinear relationship, then we would get a bad model. In our study, our predictors demonstrated significant linear relationships and hence we did not encounter this limitation. Moreover, when the assumption of deterministic independent variables independent of random error is not entirely met, there may be a measurement error in the variables. We addressed this limitation above. Furthermore, the issue of endogeneity is not handled by simple linear regression, however, we did not encounter this limitation.

Even though the response variable was treated as continuous, we could have categorized it and used classification methods for prediction. An arbitrary cutoff could have been selected of 0.5 and a logistic regression used instead of a linear regression.

*Conclusion*

All the models performed very well. This is because the relationship between the variables might be truly linearly associated. Intuitively, having a high GPA, GRE scores, TOEFL scores increases the chances of admission into universities. It was clear about the global structure of the linearity assumption but various plot the regression methods confirmed the global linear structure.

Based off the below figure, the GAM model provides the best method of predicting acceptance into the graduate school (lowest median RMSE). GAM may have outperformed strong models such as the lasso and multiple linear regression models because the predictors may not have had a completely linear relationship with the outcome. GAM could have accounted for this by using splines on the continuous predictors. However, as mentioned, all models performed very well. If mean RMSE is considered as shown in table 2., the multiple linear model performs the best, again reflecting the linearity of the data.

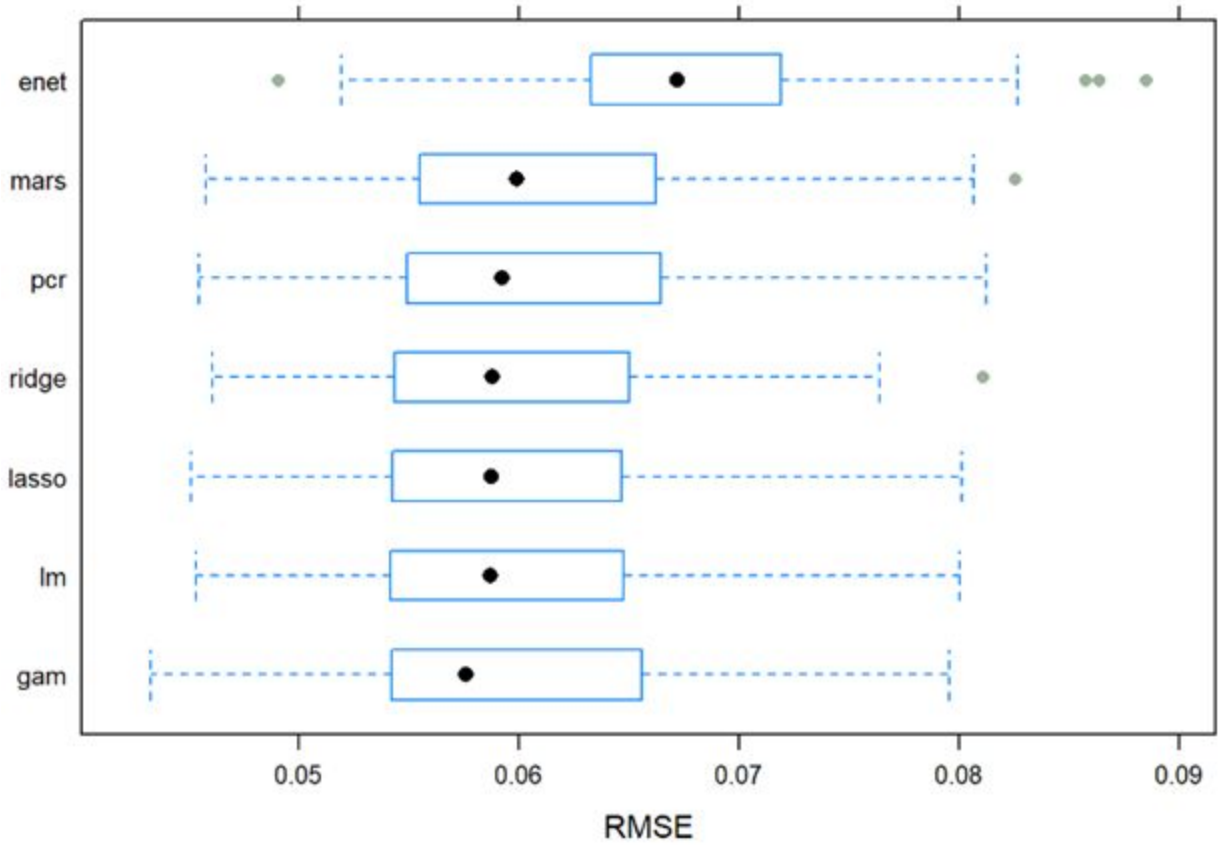


figure 3. Box and whisker plot of RMSE for each model

| Model | Mean RMSE |
|---|---|
| Lasso | 0.05988040 |
| Ridge | 0.06016358 |
| PCR | 0.06062494 |
| Lm | 0.05988188 |
| enet | 0.06760862 |
| mars | 0.06133136 |
| GAM | 0.05997482 |

table 2. Mean RMSE scores for each model