Modeling Graduate Student Success Utilizing Admissions Data: A Pilot Study

Rebecca Daniels, Brandon M. McConnell

Department of Industrial Systems Engineering
NC State University, Raleigh, NC

Presented at the 2023 NC State Summer Undergraduate & Creativity Symposium, Raleigh, NC.

Background

Graduate admissions processes are highly competitive and decisions affect program quality by way of student success. Applicants come from various educational backgrounds and degree programs; the process of selecting the top candidates becomes a multifaceted challenge.

Admissions committees and Directors of Graduate Programs (DGPs) grapple with the added complexities of diverse grading standards across students' prior institutions. Graduate programs seeking interdisciplinary candidates from diverse degree programs and majors face challenges for the faculty in aligning academic backgrounds and providing sufficient support for students to succeed. Relying heavily on standardized tests (e.g., GRE) may limit the committee's understanding of a student's full potential.

To address these challenges, universities have turned to data analysis and mathematical modeling to identify factors impacting the success and assisting in tailoring processes. At Fordham University, the Master of Data Science program explored a wide variety of regression models to assist their DGP and admissions committees with the admissions process and faculty with understanding expected student knowledge.[1] Epoka University conducted a similar study at the undergraduate level using decision trees to predict success in a three-year program.[2]

Objectives

There are many interconnected objectives that will assist in the understanding and support of graduate students. The primary goal of this research is to establish a repeatable workflow to predict graduate student success. By analyzing the outcomes of the models, staff and faculty can be better equipped to support students and understand the strengths and challenges that students may face in the course material. This knowledge can inform staff and faculty of necessary changes to the curriculum and allow for the programs to be tailored to accommodate students better and maximize student potential.

Another objective is to support admissions committees. This research is not designed to replace the committees but aid in understanding what data matters for student success. One way to do this is to determine the most relevant exams that students should take and what students need to take the exams.

By analyzing a student's background education, the research aspires to explore the impact of their prior schools on their overall success in their graduate school program. This may help admissions recognize the expertise of different institutions and their programs.

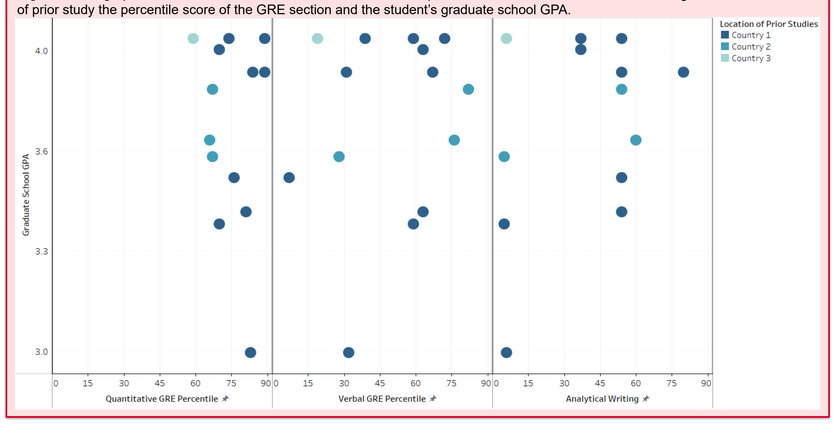
Data

The data used in this study was provided by DGPs for students of specific degree programs. The admissions data included all the predictors as well as other information that was excluded from this study. The target values of graduate GPA and student grades for specific course were provided separately. The predictors and target values were matched based on the individual student.

The predictors used in the modeling for this pilot study are the GRE quantitative percentile scores, prior GPA, location of prior studies, and major. The target value associated with these predictors is the GPA values for the grades of a specific course required in the degree program requested by the DGP's.

- Predictors
  - GRE Scores
  - Prior School
  - Location of Prior studies
  - Prior GPA
  - Degree
  - Major
  - English Test scores
- Target Values
  - Graduate GPA
  - Course Grades

Figure 1. The graph is broken into three sections based on the three parts of the GRE test. The dots are showing a student's location of prior study the percentile score of the GRE section and the student's graduate school GPA.



Methods

Data
- Application Data
- Course Grades
- Import and build dataframes

Cleaning
- Data cleansing with Python & OpenRefine
- Remove blank rows and unusable columns
- Cluster like data in specific columns to use

Exploration
- Interrogate data using Tableau
- Hypothesize possible relationships between predictors and targets

Modeling
- Develop mathematical models with Python
- Encode categorical data

Compare
- Analyze results from random forest and linear regression models

Random Forest

- Decision tree classifiers
- Harder to interpret results
- Can handle non-linear relationships between predictors and target
- Missing data does not impact results of algorithm
- Averages predictions of multiple trees reduces risk of overfitting
- Provides an importance table for individual features

Both

- Handles both numerical & categorical data
- Supervised machine learning algorithms
- Evaluation through common regression metrics

Linear Regression

- Linear relationships only between predictors and target
- Does not capture non-linear patterns
- Linear coefficients directly show predictor impact on target
- Must have complete data
- Complexity may cause overfitting the model
- Less taxing on computation at larger scales

Results

Table 1 below shows performance metrics for the two models used in this study, linear regression and random forest, for a 10-fold cross-validation. The metrics included are the average Root Mean Squared Error (RMSE), the standard deviation of the RMSE, and the 95% confidence interval.

The Mean RMSE shows the average error of the 10 predictions. The smaller the value the better the model performed. The standard deviation of the RMSE indicates the variability between the 10 folds of the cross-validation. The lower the value of the standard deviation the more consistent the results were. The 95% confidence interval provides a range of values that the true RMSE of each model is likely to be between. This is a combination of the prior information.

By analyzing these and other metrics, the stability and accuracy of the predictions allow for the informed decision making about the most suitable model to move forward with. Since the confidence intervals overlap, there is not enough data currently to infer one of the models is better than the other.

Model	Mean RMSE	Standard Deviation	95% Confidence Interval
Linear Regression	1.125022	0.744963	(-0.335106, 2.585150)
Random Forest	0.837099	0.555371	(-0.251428, 1.925625)

Table 1. The table shows the mean, standard deviation, and confidence interval for the RMSE for the linear regression and random forest models

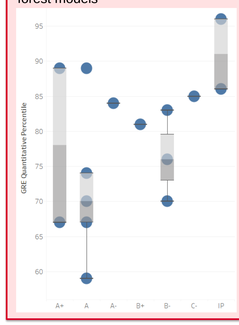


Figure 2 (left). The chart shows box and whisker plots comparing GRE quantitative percentile scores for different grades in the class. If there was more data, a prediction could be made about how well students will do in the class based on the GRE score.

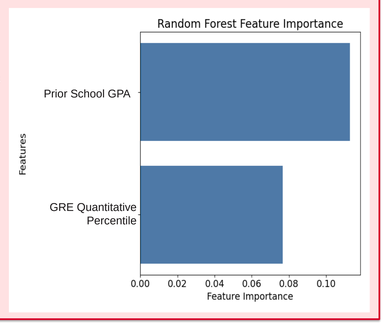


Figure 3 (right). The graph shows the top two features and their importance in the random forest model. This shows that currently the prior school GPA is the most important feature to the random forest model.

Next Steps

To continue this research, there are critical steps that need to be taken. The most important of which is to enrich the datasets by gathering more data for both the admissions and target variable. This will improve accuracy and robustness of the model.

As a pivotal aspect of this research is to investigate the correlation between predictors and targets, exploring a wider variety of predictors is paramount to understanding what predictors influence success beyond the few predictors used in the pilot study.

Exploring alternative regression techniques on the expanded dataset will permit selecting the most suitable approach to the specifics of this data.

A close relationship and meetings with the DGPs is required for the continued success of this research. The first component of these meetings is to allow the DGPs to provide feedback and keep the process on track and yielding useful results to them. The other component is to provide the DGPs with the most up-to-date results and potential implications for student success.

These steps are not the only ones necessary to propel the research forward but will help open the door to new insights and understandings of what it takes for a graduate student to succeed.

References

(1) Zhao, Y., Xu, Q., Chen, M., & Weiss, G.M. (2020). Predicting Student Performance in a Master's Program in Data Science using Admissions Data. *Educational Data Mining*
(2) Cengiz, N., & Uka, A. (2014). Prediction of Student Success Using Enrollment Data. *Educational Data Mining*.

Acknowledgement

Research funding provided by the Edward P. Fitts Department of Industrial Systems Engineering in partnership with the Office of Undergraduate Research.

Modeling Graduate Student Success Utilizing Admissions Data: A Proof of Concept

Rebecca Daniels, Brandon M. McConnell

Department of Industrial & Systems Engineering, NC State University, Raleigh, NC

Assessing student success is critical in university graduate school admissions. This is inherently difficult due to the heterogeneous applicant pool, widely varying grading standards across international universities, and other factors. This research is a pilot study to develop mathematical models that leverage admissions data to predict student success in graduate school for a limited set of graduate programs. The goal is a repeatable workflow to support (not replace) graduate admissions committees. Admissions data is matched with academic performance in key courses selected by Directors of Graduate Programs (DGPs). Initial supervised learning models include random forest decision trees and regression to analyze the data and produce useful models, including program-specific scoring models for international school-degree pairs. Initial data exploration led to the construction of multiple models to answer several key questions from DGPs regarding program admissions policies. The findings will highlight factors that contribute to student success and frame potential decisions regarding the admissions process, curriculum changes, and support to improve the likelihood of student success. This proof of concept is a foundation for future research in the study of graduate student success. This research provides a workflow to exploit additional data, DGP feedback, and additional models to support robust decision-making regarding admissions and support for graduate students.

Presented at the 2023 NC State Summer Undergraduate & Creativity Symposium, Raleigh, NC.