Predicting Applicant Admission Status for Georgia Tech's Online Master's in Analytics Program

Shawn Staudaher
Sam Houston State University
Conroe, USA
sstaudaher@shsu.edu

Jeonghyun Lee
Georgia Institute of Technology
Atlanta, USA
jonnalee@gatech.edu

Farahnaz Soleimani
Georgia Institute of Technology
Atlanta, USA
soleimani@gatech.edu

ABSTRACT

This work reports on progress made towards building an equitable model to predict the success of an applicant to Georgia Tech's Online Master's in Analytics program. As a first step, we have collected and processed data on 9,044 applications and have trained a predictive model with a ROC-AUC score of 0.81, which predicts whether an applicant would be admitted to the program. Our next steps will include using applicant data to model the successful completion of the Analytics program's three core courses, graduation, and finally job placement. In addition, we plan to expand our feature processing and incorporate techniques to ensure that our models do not discriminate based on demographic factors. In the long run, we hope that the results of this study can be used to improve the course contents, selection of offered courses, and prerequisite training, and even give guidance toward the selection of the applicants.

Author Keywords

Applicant success; machine-learning; predictive analytics.

CCS Concepts

•**Applied computing** → **Distance learning;**

INTRODUCTION

Georgia Institute of Technology's (Georgia Tech) online Master's in analytics (OMSA) has seen an exponential increase in applicants since the program's inception in the Fall of 2017, with 9,044 applicants to date. The workload for faculty and staff to process these applications is immense, and the Georgia Tech leadership has requested predictive models to help in their decision process.

The criteria that application reviewers use in their admissions decision is fundamentally different than a traditional master's level program. In a traditional program, the goal is to create a ranked list of the best candidates and then make offers until all positions are filled, but the goal of OMSA is to admit any candidate who would likely succeed in the program. This is more difficult particularly because the applicant pool is much more diverse than the on-campus pool and reviewers may not have personal experience instructing similar students. Time commitment expectations are different as many online students will take only a single course each semester. Additionally, application reviewers most likely will not have a personal relationship with the applicant, and therefore considerations of fit and research interest alignment with the instructors are less important.

Due to the above factors, using past results to predict the success of new applicants will be of value to the OMSA program. Through machine learning and statistical analysis, this study aims to predict student success in the OMSA program and identify the parameters with the most significant impact on program completion. Hence, the outcome of this research can be used to enhance program design. In the following section, we review relevant literature on the role of application materials and admission criteria in predicting program admission and student success in higher education.

RELATED WORK

A substantial body of research has investigated the predictive validity of some of the selective application materials that are required for graduate school admissions across different disciplines. These materials include objective variables such as standardized test scores and undergraduate GPA as well as subjective materials that reflect individuals' unique work experience or personality traits (e.g., personal essays, letters of recommendation). Among various types of application materials, the Graduate Record Examination (GRE) scores have been extensively used as one of the main criteria to make admission decisions for the graduate-level programs.

Previous research has revealed mixed results in terms of the predictive power of the GRE scores on graduate course performance. House et al. [3] assessed the significance of GRE quantitative, verbal, and analytical section scores in students' performance in graduate psychology courses. The results had a large variation across courses and the scores of each section of the GRE test. In general, the scores related to the analytical section of GRE showed the lowest predictive impact. The results also indicated that the overall GRE score is weakly predictive of student's grades in the considered graduate-level courses. Similarly, Miller[8] monitored how the GRE scores can predict the long-term research performance of the Ph.D. students admitted to the Physics program. The study found

L@S '20, August 12–14, 2020, Virtual Event, USA.

ACM ISBN 978-1-4503-7951-9/20/08 ...$15.00.
http://dx.doi.org/10.1145/3386527.3406735

that compared to the GRE scores, the undergraduate GPA has a stronger correlation with the students' graduate GPA.

On the other hand, Kuncel et al. [7] suggested that scores of the verbal and quantitative sections of the GRE exam can serve as good predictors of the first-year graduate GPA. Their findings also supported the predictive validity of GRE scores for both Master's and Ph.D. degree levels to estimate the faculty ratings of the students' performance. For this study, the authors used a comprehensive data set which consisted of a pool of around 10,000 databases assembled from a couple of previous meta-analysis studies. However, these findings need to be viewed with caution because the patterns might differ by discipline and the required skills and background in order to succeed could significantly vary between different programs.

In addition to the GRE, the Graduate Management Admissions Test (GMAT) is another important application material that is required by most of the MBA programs. Koys[5] investigated the level of significance of GMAT score and the undergraduate GPA to predict the academic performance of 75 non-US students who enrolled in three US business schools located in Bahrain, Czech Republic, and Hong Kong. The GMAT was found to have a stronger correlation with the MBA GPA than the undergrad GPA. The prediction results showed that the GMAT score was able to estimate around 41% of the variance in the student's MBA grades, affirming GMAT as a strong success predictor in the MBA programs.

Likewise, Kuncel et al.[6] conducted an analysis of a sample of 402 students in graduate business schools. They found that the combination of the GMAT scores and undergraduate GPA could predict students' first-year grades as well as their overall graduate GPA. Further analysis indicated that the GMAT scores were found to be a superior predictor than the undergraduate GPA. However, one of the main limitations of this work is that the dataset was assembled from various institutions and time periods. Therefore, further research is needed to take into account factors related to individual programs offered at different time points.

Some researchers have compared objective and subjective application materials. Halberstam and Redstone [2] tested whether any of the admission materials could predict the success of students attending a graduate-level speech-language pathology program. They collected data for a sample of 25 admission files and selected a set of objective variables and subjective materials. The objective variables consist of the applicants' age, native language, undergraduate major, overall GPA, and GPA for speech prerequisite courses. The subjective materials include personal essays, reference letters, and work experience. Although both GPAs displayed a strong correlation with the students' success measured as their graduate GPA, the GPA corresponding to speech prerequisite courses showed a higher level of impact on the learning outcome compared to the overall GPA. Among the subjective variables, only the rating related to the personal essay was found to be an important predictive of the graduate GPA.

Despite the relatively strong predictive power of objective and standardized test scores in measuring student success, prior research has suggested that it can be problematic to rely on single test scores in the admission process in higher education due to ethical concerns. For example, Nankervis [9] raised an issue of gender inequity in using SAT admission cutoff scores as part of the admission policy among state-supported four-year institutions. Given that there is an already existing performance gap in the SAT quantitative section between male and female high school students (i.e., differential validity), the author pointed out that employing cutoff scores would lead to a double standard, which could consequently increase the level of gender inequity.

Similar concern has been raised in the usage of the GRE scores. For example, Ji [4] collected data from 170 graduate students admitted at a private university and found a large discrepancy in the GRE scores among different ethnic groups including Asian, Hispanic, African American, and Euro-American. Also, Bleske-Rechek and Browne [1] performed time-variant analysis on those students who participated in the GRE test between 1982 to 2007. According to their results, men showed higher scores in both verbal and quantitative sections with a constant gap over time. Besides, a noticeable time-dependent score gap in the quantitative reasoning section was observed between different ethnicities. However,the authors suggested that, based on the increased distribution of diversity, GRE has not impeded in providing equal admission opportunities for various groups of applicants.

Overall, the existing literature has heavily focused on a few selective application materials, especially standardized test scores and GPA scores. However, this can be problematic given that application materials typically provide a wide variety of information about candidates. Therefore, for this project, we aimed to take a holistic approach by using machine learning techniques to build a robust model predicting the success of applicants to the OMSA program at Georiga Tech. We ask four research questions. (1) Can we predict whether an applicant will be admitted? (2) What application features predict admission? (3) Can we predict whether an applicant will complete the three core courses, graduate, and land a job in analytics? (4) Are our models demographically biased and can we correct them to avoid bias?

METHOD

Machine learning approaches including the classifier algorithms are implemented to build the models which predict whether an applicant would be admitted to the program. In the proposed models, the applicant's admission status for the OMSA program is considered as the model output variable (Y), and the model input variables (X) include the candidate's application data.

DATA COLLECTION AND PROCESSING

The available application data for the OMSA program at Georgia Tech consists of information for 9,044 applicants including 6,536 male candidates and 2,508 female individuals. Among the entire pool of applicants, 64% were admitted to the program and the remaining 36% were not accepted. The admitted group is formed of 71% male and 29% female applicants (Table 1).

Table 1. Distribution of Gender and First Time Graduate Student Status

			First Time Grad Student	
	Female	Male	Yes	No
Admitted	29%	71%	56%	44%
Not admitted	26%	74%	64%	36%
Grand Total	28%	72%	59%	41%

The required data to construct the matrix of the input variables include applicants' personal information such as gender and race, academic history (e.g., degrees earned), test scores (e.g., TOEFL, GRE), and supplemental information such as their background in computer programming. For constructing the output variable, the applicants' admission status is binary coded for whether an individual applicant was admitted to the OMSA program or not. The goal of data processing is to quantify the information presented in each candidate's application.

Numerical or categorical features are sufficient as-is, but more qualitative features such as the prestige of an undergraduate university or the tone of a recommendation letter, require preprocessing. For this analysis, natural language processing is employed in this study to quantify each applicant's statement of purpose and letters of recommendation. Natural language parameters are measured using the python package TextBlob's polarity and subjectivity, textstat's Flesch-Kinkaid Grade and Flesch Reading Ease, along with the number of words in the text, number of unique words, and the type token ratio. This allows the measurement of the tone and complexity of each text within an applicant's docket.

Currently, all other data is processed as the following: numerical data is scaled, and categorical data is split into k-1 indicator variables (where k is the number of categories). The processed data include 154 variables, with each falling into the broad classifications of demographics, previous education, self-description, letters of recommendation, and employment history.

STATISTICAL MODELS

As an initial step of the project, modeling was performed using python's scikit-learn library. The fitting was handled using a pipeline while the missing data was imputed using SimpleImputer and scaling was conducted by RobustScaler. In addition, a 3-fold cross-validation technique was implemented. Machine-learning algorithms including the Logistic Regression, Random Forest, Gradient Boosting, and ADABoost Classification were applied to establish the predictive models using a suite of hyper-parameters. Eventually, 1,110 models were trained based on five runs for each set of hyper-parameter and classification algorithm pairs.

The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) is computed from prediction scores as a performance measure of the statistical models. This value is found using the area under the ROC probability curve of the true positive rate against the false positive rate. In a classification problem, a higher ROC-AUC score implicates that the corresponding model performs better in distinguishing different classes of the output. Hence, a score equal to 1 indicate a perfect classifier while a score near to 0 means the model has a weak measure of separability.

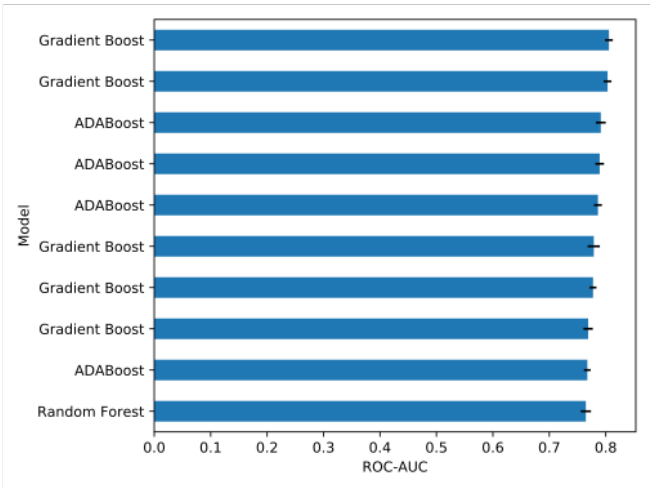

Figure 1. ROC-AUC scores for the top 10 scoring model and hyperparameter pairs. Note that models with different hyperparameter sets appear multiple times.

RESULTS

The results for the top 10 models and hyperparameter pairs are shown in Figure 1. All of the three models (Gradient Boost, ADABoost, and Random Forest) were found to perform better than logistics regression. Although, the tree-based algorithms performed well in predicting the admission status of the applicants, the gradient boost classifier using 1,000 estimators with a learning rate of 0.1 and an exponential loss function was found to have the highest average ROC-AUC score of 0.81 and standard deviation of 0.01.

Using the top performing classifier (Gradient Boosting), the relative importance of each feature was investigated, and the first 15 significant features are presented in Figure 2. According to this level of importance, the applicants' GPA played the most important role for being admitted to the program while the duration that the applicant spent in a college ranked the second. Besides, fulfilling a bachelor's degree, and having strong reference letters were detected as the next parameters with significant influence on predicting whether an applicant will be admitted to the OMSA program.

Contrary to the logistic regression model, the tree-based algorithms such as the Gradient Boosting do not construct coefficients for the final model. This type of model typically represents the entire data at the top (the root of the tree) by a single node that assigns a particular criterion for splitting the values of one of the input parameters in the model. Then, in an iterative process, additional nodes and leaves are added to the tree by testing different criteria for the remaining input parameters. Eventually, the features that are placed at the top of the tree (i.e., the applicant's college GPA) as the starting nodes are considered as the parameters with the most significant influence on forming the final decision model. With this

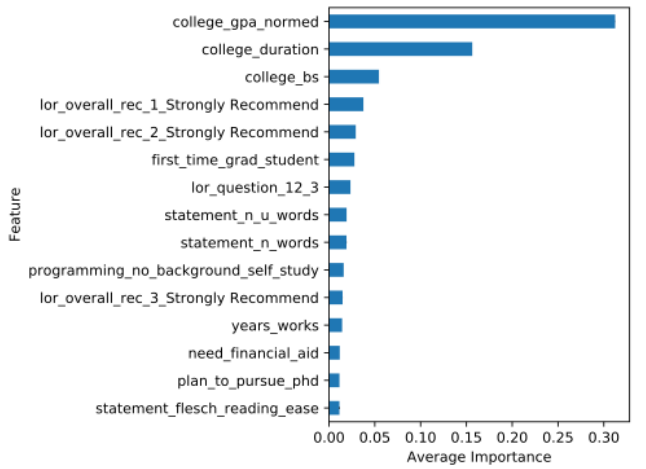

Figure 2. Feature importance in the top scoring model fit.

said, feature importance ranking provided in Figure 2 implicates the relative importance of the input parameters. However, correlation analysis is planned to be performed in the future phase of the project.

CONCLUSION AND FUTURE WORK

In this paper, we reported on the progress that we have made on building a robust model that can predict the admission status of applicants to the online Master's program at Georgia Tech. First, we were successfully able to process a massive amount of data from more than 9,000 candidates and reduce the dimensional complexity of the raw data. Through data processing, we identified 154 key variables to construct the matrix of the input variables for predictive modeling. We were also able to train the model and test model performance with reasonably high accuracy by using various machine learning algorithms. These results from our preliminary analyses suggest that data from graduate program applicants, including our OMSA applicant dataset, can provide a rich and promising basis so for applying machine learning techniques.

In the next steps of the project, we plan to expand the scope of our dataset and feature processing in several ways. Beyond whether an applicant was admitted to the OMSA program, we will use applicants' data to model the successful completion of the Analytics program. To do that, we will initially build models to predict students' grades in three core courses of the program. Then, based on those models' performance, we will construct models for predicting whether students drop out of the program, graduate on time, and land a job in analytics. Additionally, we plan to incorporate techniques such as Shapley Additive Explanation to ensure that our models do not discriminate based on demographic factors. Ultimately, we hope that our research will offer useful guidance for the OMSA program's admission process and help administrators make informed decisions contributing to program improvement.

REFERENCES

[1] April Bleske-Rechek and Kingsley Browne. 2014. Trends in GRE scores and graduate enrollments by gender and ethnicity. *Intelligence* 46 (2014), 25–34.

[2] Benjamin Halberstam and Fran Redstone. 2005. The predictive value of admissions materials on objective and subjective measures of graduate school performance in speech-language pathology. *Journal of Higher Education Policy and Management* 27, 2 (2005), 261–272.

[3] J Daniel House, James J Johnson, and William L Tolone. 1987. Predictive validity of the Graduate Record Examination for performance in selected graduate psychology courses. *Psychological Reports* 60, 1 (1987), 107–110.

[4] Chang-Ho C Ji. 1998. Predictive validity of the Graduate Record Examination in education. *Psychological Reports* 82, 3 (1998), 899–904.

[5] Daniel J Koys. 2005. The validity of the Graduate Management Admissions Test for non-US students. *Journal of Education for Business* 80, 4 (2005), 236–239.

[6] Nathan R Kuncel, Marcus Credé, and Lisa L Thomas. 2007. A meta-analysis of the predictive validity of the Graduate Management Admission Test (GMAT) and undergraduate grade point average (UGPA) for graduate student academic performance. *Academy of Management Learning & Education* 6, 1 (2007), 51–68.

[7] Nathan R Kuncel, Serena Wee, Lauren Serafin, and Sarah A Hezlett. 2010. The validity of the Graduate Record Examination for master's and doctoral programs: A meta-analytic investigation. *Educational and Psychological Measurement* 70, 2 (2010), 340–352.

[8] Casey W Miller. 2013. Admissions criteria and diversity in graduate school. *arXiv preprint arXiv:1302.3929* (2013).

[9] Bryan Nankervis. 2011. Gender Inequities in University Admission due to the Differential Validity of the SAT. *Journal of College Admission* 213 (2011), 24–30.