Predicting Student Performance in a Master of Data Science Program using Admissions Data

Yijun Zhao
Computer and Information Science Department
Fordham University
New York, NY
yzhao11@fordham.edu

Qiangwen Xu
Computer and Information Science Department
Fordham University
New York, NY
qxu47@fordham.edu

Ming Chen
Computer and Information Science Department
Fordham University
New York, NY
mchen177@fordham.edu

Gary M. Weiss
Computer and Information Science Department
Fordham University
New York, NY
gaweiss@fordham.edu

ABSTRACT

Predicting student success in a data science degree program is a challenging task due to the interdisciplinary nature of the field, the diverse backgrounds of the students, and an incomplete understanding of the precise skills that are most critical to success. In this study, the applicant's future academic performance in a Master of Data Science program is assessed using information from the admission application, such as standardized test scores, undergraduate grade point average, declared major, and school ranking. Simple data analysis methods and visualization techniques are used to gain a better understanding of how these variables impact student performance, and several classification algorithms are used to induce models to distinguish between students that will perform very well and those that will perform very poorly. Historical admissions and grading data are used to perform these analyses and build the classification models. The analyses and predictive models that are generated provide insight into the factors that identify good and poor candidates, and can aid in future admissions decisions.

Keywords

Admission decision making, Master's program, data science, learning assessment, machine learning.

1. INTRODUCTION

Data mining methods are now in widespread use in many industries, from healthcare[10] to business[15]. Data mining is increasingly applied to education [3][8][14] and includes many diverse applications, all of which fall under the area of educational data mining (EDM). A particular focus of such applications is the college admissions process and its effectiveness, since this process directly affects the reputation of the institution as well as its financial well-being. Examples of work in this area include predicting college admissions yield [5], student retention[11], and enrollment management[1]. Another related area of EDM relates to predicting student performance. One such study used student personal and social factors, along with academic performance data, to identifying poor performers early on[2], while another study used similar information to predict third semester academic performance [13]. One more study used student course data during the semester (attendance, homework scores, etc.) to predict the student score on the end of the semester examination[18].

In this paper we investigate the problem of identifying a good admissions strategy for a Master's of Science program in Data Science (MSDS), so that the students that are admitted into the program will perform well. This problem is generally related to the EDM admissions topic, but also to the topic of predicting student performance. This problem is interesting, and distinctive, for a variety of reasons. One reason is that the vast majority of applications of data mining to college admissions deals with undergraduate admissions. That admissions process is very different from the process for our MSDS program, since undergraduate admissions is controlled by full-time admissions professionals, whereas admissions for our MSDS program is controlled by faculty with little time to devote to admissions, and who lack specialized admissions training. This is true for most graduate programs, except for possibly the large professional schools (e.g., law, medicine) that may admit many more students and have deeper resources. Determining admission to MSDS programs is especially challenging since it is an interdisciplinary field that attracts applicants from diverse backgrounds, and because MSDS programs were introduced only recently and hence have limited historical knowledge to leverage. Furthermore, even experts in the area do not fully understand exactly which undergraduate skills are most critical to success, so it is hard to know which students to admit or reject.

Yijun Zhao, Qiangwen Xu, Ming Chen and Gary Weiss "Predicting Student Performance in a Master's Program in Data Science using Admissions Data" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 325 - 333

The goal of this data mining task is to determine if, using only information available in the admissions application, a student will perform very well in the program, and hence deserve merit-based aid, or perform very poorly and hence should not be admitted. At this time only structured data is utilized in order to simplify the classification task. Thus external recommendations, personal statements, and resumes are not considered. However, there is still a wealth of information that is available, which includes prior degrees and associated grades, the name and country of the prior educational institutions, standardized test scores such as the GRE (Graduate Record Examination) and TOEFL (Test of English as a Foreign Language), and personal information about the applicant such as age, nationality, work history, and whether they ask for merit-based financial aid.

The purpose of this study is not just to identify which students will perform very well or very poorly, but to better understand the relevant factors. Thus, the predictive model that we build will most likely not be used for automated decision making, but instead will be used to educate the admissions committee about which factors are most relevant for success in the program. As mentioned earlier, this is especially important for the MSDS degree because the applicants have such different backgrounds and because the degree is relatively new.

A practical issue that impacts this study is that because the offered degree was launched only a few years ago, the data is quite limited. Compounding this issue is the fact that we do not have outcomes for students who are accepted but do not attend the university, and worse yet, we cannot know anything about how students who are rejected from the program would perform. One of our long term goals is to fully utilize this unlabeled data to improve the admissions process. This is discussed later in this paper as future work.

The rest of the paper is organized as follows. We present the details of our dataset in Section 2. The design of our experiments and associated methodology are presented in Section 3. Section 4 presents our experimental results and predictive factor analysis. We conclude and suggest future work in Section 5.

2. THE DATA

This section describes the data utilized in this study. Section 2.1 describes the data at a high level and includes some summary statistics, while Section 2.2 describes the features included in each application record. Section 2.3 then describes the distribution of feature values for key features, while Section 2.4 describes how these feature values relate to student performance in the MSDS program.

2.1 Overview

The data in this study is extracted from the application data provided by each applicant to Fordham University's MSDS program. The application process is completely electronic, so the underlying data was already in electronic form. Much of this data is structured data (e.g., student GRE scores) that is already parsed and readily available for extraction. However, other information, such as the student's required statement of purpose and external recommendations, are unstructured text and would require substantial effort to exploit. This study is limited to structured data.

The data set is comprised of 826 applicant records. Of this total, 503 (60.9%) applicants were accepted into the program and 323 (39.1%) were rejected. Of the 503 accepted applicants, 132(26.2%) enrolled in the program while 371 (73.8%) did not enroll. Since only students who enrolled have grade information, the main analyses presented in this paper are based on only 132 records. Note that the data used in these analyses depends on our current admissions strategy, since it is possible that some of the students who were denied admission into the program could have performed well in the classes. The best we can do with respect to the population of students who were denied admission is to compare their characteristics with those of the students who enrolled and performed poorly; if it turns out that the factors used to determine admission into the program differ from those that tend to predict good performance, then the current admissions strategy should be modified. The characteristics of each student population are explored and compared in Section 2.3.

2.2 Features and Feature Generation

The features that are extracted from the student applications and used in this study are listed in Table 1, along with sample values. The first three features describe the Graduate Record Examination (GRE) standardized test scores and are encoded using the score percentile. The fourth feature describes the Test of English as a Foreign Language (TOEFL) total score. The next field specifies the number of months from the time of completion (or projected completion) of the last degree to the time the current application was submitted. For students who plan to start the MSDS program immediately following the graduation from their current program, this value is typically a negative six months. Student age is at the time of application and marital status is single, married, divorced, domestic partner, or blank (unspecified). Gender is either male or female and citizenship specifies the country of citizenship. The next six features relate to the last degree program (i.e., school) that the student attended. They include the student's GPA (Grade Point Average), major and degree, the country that the school resides in, the primary language of instruction, and the school ranking. The MSDS GPA uses a 4-point scale, and is based on the student's performance after enrolling in the program. This attribute is utilized to generate the class value, as described in Section 3.1.

School rank is the only feature in Table 1 that is not a feature from the student application. Instead, the ranking is generated from the school name via a multi-step process. The first step involves matching the school name against the US News and World Report "Best Global Universities" ranking (`usnews.com/education/best-global-universities`), which includes 1500 universities from eighty countries and is based on academic research performance and global and regional reputation. If a match is found, then this global ranking is used; otherwise the US News and World Report "Best Colleges" ranking is searched (`usnews.com/best-colleges`). This is restricted to colleges in the United States and includes separate rankings for national universities (major research institutions), liberal arts colleges, and regional

Table 1: Data Set Features and Sample Values

#	Feature Name	Sample Value
1	GRE Verbal %	52
2	GRE Quantitative %	95
3	GRE Writing %	34
4	TOEFL Total	105
5	Months since Degree	6
6	Student Age	22
7	Marital Status	Married
8	Gender	M
9	Citizenship	China
10	School GPA	3.7
11	School Major	Chemical Engineering
12	School Degree	BS
13	School Country	China
14	School Language	Mandarin
15	School Rank	85
16	MSDS GPA	3.5

colleges and universities. If a match is found in one of these rankings, then the ranking is converted to a global ranking by adding 1200 if the match was for a national ranking, and 1400 if the match was for a regional ranking. This process of assigning a global ranking is a very rough heuristic method, but generally provides reasonable values. If a school is not found on any of these rankings then a global ranking of 9999 is used.

There are a number of features that are available from the original application information but are not used in this study and do not appear in Table 1. For example, the Total TOEFL score is included but the four TOEFL subscores are not included, since preliminary analysis indicated that these subscores did not provide much benefit. Additionally, if the applicant attended multiple institutions of higher education, then information for more than one school was provided. However, since providing additional schools for only some applicants would substantially complicate the analysis, this information was dropped, so only the most recent degree granting school was included.

2.3 Distribution of Feature Values

In any applied data mining study, it is important to understand the data. In this section, we provide information about the distribution of feature values. Since the focus of this study is in identifying students who will perform well or poorly in the program, we begin with the feature distribution of the students who enrolled in the MSDS program. This information is provided in Figure 2.2. The figure provides a good overview of the demographics of the applicants: males outnumber females by a ratio of almost 2 to 1, nearly 90% are single, and based on citizenship, about 74% are foreign nationals, while 21% are US citizens, and 5% are permanent residents. Clearly the MSDS program attracts a large international contingent. As expected, most students are young, although about 6% are over 30, suggesting that they likely have substantial industry experience. Overall, more than 80% are within two years of their last degree.

The TOEFL scores, which are only required for international students who have not completed two years of instruction at an English-language university, show that most students, but not all, have good English language skills. According to the testing agency, the average TOEFL score is 84, and any such score is generally considered good. For the MSDS program, a score of 80 or above is generally required, and hence our admitted students tend to have good English language skills.

A feature that is critical to the admissions decision is the student's prior major discipline. The program is geared towards students who have substantial mathetmatics background and at least some experience in computer science and programming. While computer science and mathematics majors are thought to have an advantage, students in any science or quantitative discipline are encouraged to apply. Figure 2.2 shows the distribution of major over all applicants, and further shows the number in each major that were admitted and rejected. The statistics show that the largest number of admitted students have a background in computer science or a highly related field, with mathematics and statistics a close second.

Although this study focuses on enrolled students, it is useful to understand the characteristics of the students who were rejected from the program, or were admitted and did not enroll, and how they compare to students who did enroll. Displaying this information graphically for all three populations would take up too much space, so the key observations are summarized below.

- Applicants who were rejected are much more likely to have a GPA under 3.0 (41%) than those who enrolled (14%) or were admitted and did not enroll (17%).
- Applicants who were rejected are more likely to have a degree from an institution not ranked in the top 2000 (30.7%) than those who enrolled (25.8%) or were admitted but did not enroll (20.2%).
- Applicants who enrolled were about 6% more likely to have completed a graduate degree (21.2%) than those who were admitted but did not enroll (15.1%) or were rejected (14.6%).
- Applicants who enrolled were less likely to be female (35.6%) than those who were admitted but did not enroll (47.4%) or were rejected (39.6%). Female applicants who are admitted are less likely to enroll than their male counterparts.
- Foreign nationals made up 74% of enrolled students, 79% of those admitted who did not enroll, and 77% of those rejected. As might be expected, foreign nationals who are admitted are somewhat less likely to attend.
- The age profile does not vary much between those applicants who enroll and are rejected. However, applicants who enroll are much more like to be older and between the ages of 24 and 30 (33%) than those were who are admitted but do not enroll (22%).
- Those who apply more than two years after completing their last degree are more likely to be rejected from the program (33.3%) than those who either enroll (17%) or are admitted and do not enroll (19%).

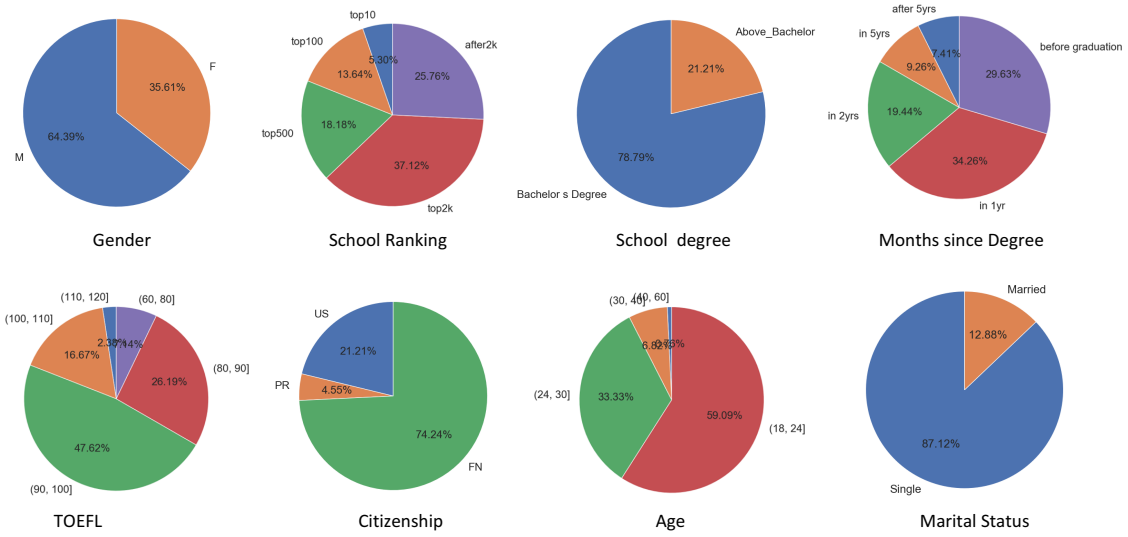

Figure 1: Feature Statistics

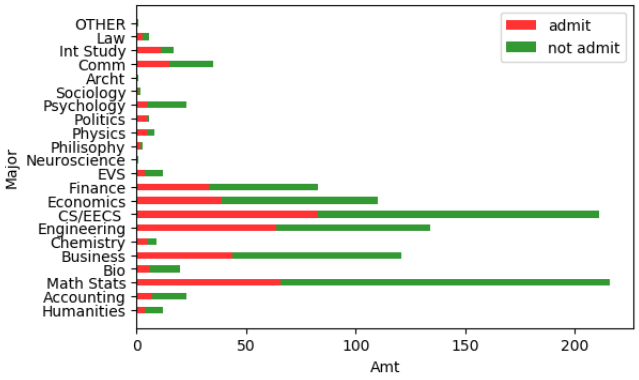

Figure 2: Major Distribution

Table 2: Mean Values for MSDS Performance Groups

Feature	Bottom 20%	Middle 60%	Top 20%	All
GRE Verbal %	42.5	48.6	57.0	49.4
GRE Quantitative %	<u>79.0</u>	81.9	<u>82.8</u>	81.6
GRE Writing %	32.5	31.2	34.0	32.0
TOEFL Total %	96.1	96.1	96.6	96.2
Foreign National %	<u>59.1</u>	74.4	<u>85.7</u>	74.2
Graduate Degree %	<u>18.2</u>	19.5	<u>28.6</u>	21.2
Married %	22.7	9.8	14.3	12.9
Female %	<u>22.7</u>	39.0	<u>35.7</u>	35.6
Months Since Degree	<u>12.8</u>	13.6	<u>22.8</u>	15.4
Age	25.5	24.7	24.6	24.8
School Rank	1005	1005	1082	1005
School GPA	<u>3.17</u>	3.29	<u>3.53</u>	3.32

2.4 Features and MSDS Performance Groups

In Section 2.3 the distribution of feature values was examined for the population of enrolled students, and then the key differences between the features for the three admission categories (enrolled, admit but not enrolled, rejected) was analyzed. In this section, we compare the feature values with respect to student performance in the MSDS program to provide insight into the factors that influence student performance. As will be discussed in Section 3.1, our focus in this study is to identify the students who enroll in the MSDS program that will perform in the top 20% and the bottom 20%. Thus, in this section, we examine the feature values for three performance groups: the bottom 20%, middle 60%, and top 20%. In order to simplify the comparison, the mean values of numerical features are considered. Table 2 provides the relevant information. The features values that differ substantially between the bottom and top 20%, and we believe are of predictive value, are underlined.

The three GRE test score percentiles in Table 2 show the expected trend: the scores improve as we move up the performance groups. The one exception is that there is a slight dip in the GRE writing score when moving from the bottom group to the middle group, but even in this case the writing scores for the top group outperform the bottom group. Our admissions committee normally places the most weight on the quantitative score and generally does not consider the writing score. What is most interesting is that the difference in the quantitative scores between the bottom and top 20% is only 3.8% (79.0% vs. 82.8%), even though quantitative abilities are generally thought to be critical for data scientists. The modest difference may reflect the fact that the GRE exam only tests fundamental mathematical skills. The TOEFL score barely differs between the three performance groups; however, this may not be surprising since the mean values are quite high, indicating that most students have more than sufficient English language skills.

There is an obvious pattern with respect to the percentage of foreign nationals—the percentage of foreign nationals increases from 59.1% for the bottom 20% to 85.7% for the top 20%. This marked difference occurs even though we showed in Section 2.3 that the percentage of foreign nationals is relatively constant across the three admissions categories. It is worth mentioning that many of the foreign nationals completed their undergraduate education in the United States. The data also shows that a higher percentage of students in the top 20% previously earned a graduate degree. This may seem intuitive, but since these degrees are generally in very different disciplines, the relationship is not obvious. Such students will have demonstrated the prior ability to complete graduate work and perhaps the maturity associated with this is a significant factor. A graduate degree is considered quite favorably in the admissions process, and also enables a student to compensate for a low undergraduate GPA.

Students in the top 20% are less likely to be married than those in the bottom 20%, but the trend is not consistent through the middle 60%, so we tend not to place too much weight on these differences. The students in the top 20% are more likely to be female than those in the bottom 20% (35.7% versus 22.7%) and this suggests that there is a real gender difference. The explanation for these gender differences is not obvious, but at the undergraduate level we have observed that academically weak female students tend not to major in Computer Science due to the societal pressure that already discourages them from majoring in scientific and technical disciplines.

The next two features show that higher performing students tend to have more time since the granting of their last degree (about one more year), but are still approximately the same age. This could reflect the fact that students who immediately proceed from an undergraduate degree to the MSDS program have not thought as deeply about their desire to become a data scientist and hence may not be as committed. The school rank does not differ significantly between the performance categories, suggesting that the reputation of the prior school is not a key factor in student performance in the MSDS program. Finally, there is a very clear trend that the higher the GPA in the prior degree, the higher performing the student. This is perhaps the most obvious indicator of future achievement and the values support that: the prior GPA of those that are in the top 20% of the MSDS program is 3.53 versus 3.17 for those in the bottom 20%.

3. EXPERIMENT METHODOLOGY

This section describes the experiments related to predicting student academic performance in the MSDS program. Section 3.1 precisely defines the problem as a classification problem. Section 3.2 provides a brief description of the eight classification algorithms utilized in this study. The details concerning the design of the experiments are provided in Section 3.3.

3.1 Problem Formulation

We are primarily interested in identifying the applicants that will perform very well and will have GPAs within the top 20% of enrolled MSDS students, or will perform poorly and fall within the bottom 20%. The reason for this is that we want to deny admission to those who we anticipate will perform in the bottom 20% and may want to provide merit-based aid to those we expect to perform in the top 20%. Note that this does not mean we only deny admittance to the bottom 20%, since admission will already be denied to those who do not meet our general admissions requirements (e.g., GPA above 3.0, TOEFL above 80, etc.). We therefore build two classification models: one that distinguishes the top 20% from the bottom 80% and one that distinguishes the bottom 20% from the top 80%. The minority class is always considered the positive class. The performance of these two models is described in Section 4.

3.2 Classification Algorithms

This section provides brief descriptions of the established machine learning algorithms that are employed in this study. A heterogeneous ensemble approach is also described.

3.2.1 Logistic Regression

Logistic Regression [12] is a type of generalized linear model (GLM) that studies the association between a categorical response variable Y and a set of independent (explanatory) variables $X = \{X_1, X_2, \ldots, X_n\}$. In particular, the Y variable is first modeled as a linear function of X, and then the numerical predictions of Y are transformed into probability scores using a sigmoid function. In a binary classification task, the scores indicate a corresponding instance's likelihood of belonging to the positive class. Thus, a cutoff (usually 0.5) can be established as a decision boundary to further categorize the instances into the more likely class.

3.2.2 Support Vector Machines (SVM)

SVM [6] performs classification tasks by constructing a decision boundary in a multidimensional space that separates instances of different class labels. *SVM* strives to maximize the distance between the hyperplane and the data points of both classes. Maximizing the margin distance reinforces that future data points can be classified with more confidence. *SVM* is capable of transforming the data into a higher dimensional space using various kernel functions to enhance data separability. In this study linear *SVM* is used to facilitate risk factor analysis.

3.2.3 Decision Trees

A *Decision Tree* [16] model uses a tree structure to model the data in which each leaf node corresponds to a class label and attributes are represented as the internal nodes of the tree. Each branch represents a potential value of its parent node (i.e., an attribute). The major challenge in building a *Decision Tree* model is to choose the attribute for each node in each level. In our study we use the *Gini Index* as our criterion for attribute selection.

3.2.4 Random Forest

Random Forest [4] is a collection of decision trees, where each tree is trained with a subset of training instances and a subset of attributes. By pooling predictions from multiple decision trees, *Random Forest* reduces the variance of each individual tree and achieves a more robust and superior performance.

3.2.5 Neural Network

A *Neural Network* [9] is a computational model that is inspired by the way biological neural networks in the human brain process information. It consists of an input layer, one or more hidden layer(s), and one output layer. The adjacent layers are connected by transferring the values in one layer to a new set of values in the next layer with a set of weights and an activation function. "Training" is the process of adjusting the network weights using a back propagation algorithm to achieve the highest consistency (i.e., cross entropy) between the model outputs and the true class labels.

3.2.6 Naive Bayes

A *Bayes classifier* belongs to the family of probabilistic generative models. The algorithm differs from discriminative models in that, instead of finding a functional form, it models the probability distributions of the data. In a binary classification task, predictions are set to the larger of $P(y = i|X)$ where $i \in \{0, 1\}$ and $X = \{x_1, x_2, \ldots, x_d\}$. A *Naive Bayes* classifier further assumes that features are independent of each other given the class, which simplifies the evaluation of P(X|y=i) to $\prod_{j=1}^{d} p(x_j|y = i)$.

3.2.7 K-Nearest Neighbor (KNN)

KNN is an effective classification algorithm that does not require pre-training of a model. Classification decisions are based on a majority vote on k empirically observed instances that are most similar to the instance in question. The resemblance is typically measured by a distance metric such as Euclidean distance operated on the attributes describing the two instances.

3.2.8 Ensemble Learner L

In addition to individual machine learning algorithms, we explored ensemble techniques [7] to integrate information from different classifiers. Ensemble learning is a family of algorithms that seek to create a "strong" classifier based on a group of "weak" classifiers. In this context, "strong" and "weak" refer to how accurately the classifiers can predict the target variable. Ensemble learning has been proven to have improved and more robust performance than a single model. Specifically, multiple base classifiers are built for the original classification task with the training data. A meta-learner L is constructed by combining the outcomes from the base classifiers to improve predictive accuracy. In this study we combine the predictions from the base classifiers using an unweighted majority vote and our base learners consists of seven single models described in Sections 3.2.1 - 3.2.7.

3.3 Experiment Design

All experiments in this study utilize 10-fold cross validation. In addition to reporting overall predictive accuracy, the results in Section 4 and Table 3 report the performance on the positive/minority class via the sensitivity metric, which is also known as recall and true positive rate, and the performance on the majority/negative class via the specificity metric, which is also known as true negative rate. For both classification tasks, there is class imbalance since the ratio of the positive to negative class is approximately 1:4. Bagging is used to address this class imbalance; at training time five bags of balanced training data are created where each bag consists of all minority-class examples and an equal number of randomly selected majority-class examples. The class for each test example is based on a majority vote of the five models built using the data from each bag.

The parameters of the models are selected experimentally using the training data using a grid search. Both the training and test accuracies are reported in Table 3. Specifically, for the SVM model, the trade-off parameter $C = 0.1$. For the KNN algorithms, the number of nearest neighbors $k = 3$. For the neural networks model, we used a 3-layer architecture with (128, 256, 512) nodes in identifying the bottom 20% of the students, and a two-layer architecture with (128, 256) nodes in identifying the top 20% of the students. For the rest of the algorithms, including the depth of the decision tree, the number of trees in the random forest, etc., we applied the default parameters provided by the Python scikit-learn package.

4. EXPERIMENT RESULTS

This section presents the results of the classification experiments. The accuracy results for identifying the top and bottom performing students are presented, as are the top predictors for identifying these two populations.

4.1 Analysis on Performance Measures

Table 3 presents all of the performance results for the two classification tasks. This analysis focuses exclusively on the performance on the test data. The results in the table show that *Random Forest* and the ensemble learner L achieve the two best overall predictive accuracy values for both classification tasks. For the tasking of identifying the bottom 20% of students, L achieved an 86% overall accuracy compared to 83% for random *Random Forest*. When these results are broken down into performance on the bottom 20% and the rest, L achieved results of 90% and 83%, respectively, versus 91% and 75% for *Random Forest*. It should be noted, however, that although *Decision Tree* has only the third best overall performance, it has the best performance at identifying the bottom 20% of the students (94% versus 91% for *Random Forest* and 90% for L). However, *Decision Tree* performs very poorly at classifying the remaining 80%, with a specificity of 65%.

For the classification task of identifying the top 20% of students, *Random Forest* delivered an overall accuracy of 86%, while L achieved an overall accuracy of 85%. When these results are broken down into performance on the top 20% and the rest, *Random Forest* achieved results of 94% and 79%, respectively, versus versus 92% and 79% for L. In this case *Decision Tree* again did very well when just evaluated on the minority class, with a performance of 94% for the top 20%, equalling the performance of *Random Forest* on this population. Note that since *Random Forest* is a collection of decision trees, it belongs to the family of *homogeneous* ensemble methods. Thus, we conclude that ensemble learners are the best machine learning models for the two classification tasks.

4.2 Analysis on Predictive Features

An additional motivation of our research is to identify the top predictors for the successful and struggling students.

Table 3: Performance Comparison Over Eight Models

	Bottom 20% vs. Rest						Top 20% vs. Rest					
	Test			Training			Test			Training		
Models	Bot20	Rest	Overall	Bot20	Rest	Overall	Top20	Rest	Overall	Bot20	Rest	Overall
SVM	0.74	0.62	0.68	0.85	0.69	0.77	0.68	0.54	0.61	0.75	0.58	0.67
Decision Tree	0.94	0.65	0.80	0.95	0.80	0.87	0.94	0.67	0.80	0.96	0.80	0.88
Random Forest	**0.91**	**0.75**	**0.83**	**0.96**	**0.94**	**0.95**	**0.94**	**0.79**	**0.86**	**0.96**	**0.85**	**0.90**
Logistic Regression	0.71	0.63	0.67	0.84	0.74	0.79	0.71	0.63	0.67	0.86	0.75	0.81
KNN	0.93	0.63	0.78	1.00	0.82	0.91	0.90	0.58	0.73	0.98	0.70	0.84
Naive Bayes	0.83	0.54	0.68	0.91	0.58	0.74	0.58	0.58	0.58	0.72	0.67	0.70
Neural Network	0.34	0.80	0.57	0.39	0.83	0.61	0.52	0.53	0.53	0.59	0.57	0.58
Ensemble (L)	**0.90**	**0.83**	**0.86**	**0.96**	**0.92**	**0.94**	**0.92**	**0.79**	**0.85**	**0.96**	**0.87**	**0.91**

Table 4: List of Top 10 Predictive Features in Identifying Bottom 20% of Students

Predictors of the Bottom 20% vs. Rest Models				
Rank	**SVM**	**Logistic Regression**	**Random Forest**	**Decision Tree**
1	Economics[1]	Economics[1]	GRE Verbal %	GRE Verbal %
2	China[2]	Environmental Studies[1]	Months since Degree	Months since Degree
3	Communications[1]	US[3]	GRE Quantitative %	GRE Writing %
4	Environmental Studies[1]	CS/EECS[1]	GRE Writing %	Economics[1]
5	Psychology[1]	Business[1]	School Rank	GRE Quantitative %
6	CS/EECS[1]	Communications[1]	Student Age	School Rank
7	Applied Math/Stats[1]	FN[3]	Overall GPA	Business[1]
8	Masters[4]	Biochemistry/Biology[1]	Psychology[1]	FN[3]
9	FN[3]	Bachelors[4]	Economics[1]	Overall GPA
10	Bachelors[4]	Architecture[1]	FN[3]	Student Age
Predictors of the Top 20% vs. Rest Models				
Rank	**SVM**	**Logistic Regression**	**Random Forest**	**Decision Tree**
1	Business[1]	Business[1]	Overall GPA	Months since Degree
2	Engineering[1]	International Studies[1]	GRE Verbal %	GRE Verbal %
3	Overall GPA	Bachelors[4]	GRE Writing %	Overall GPA
4	CS/EECS[1]	US[3]	Student Age	Student Age
5	Bachelors[4]	Chemistry[1]	GRE Quantitative %	GRE Quantitative %
6	International Studies[1]	Humanities[1]	School Rank	School Rank
7	China[2]	Accounting[1]	Months since Degree	TOEFL Total
8	Accounting[1]	Finance[1]	Business[1]	Business[1]
9	United States[2]	Applied Math/Stats[1]	TOEFL Total	Engineering[1]
10	US[3]	Engineering[1]	CS/EECS[1]	GRE Writing %

1: School major

2: Country of last school

3: Citizenship code. Values include PR (permanent resident), FN (foreign), and US.

4: Last school degree.

The findings will help the admission committee to focus on more effective rubric measures and assign merit-based financial aid. Table 4 presents the top-10 predictors for the four classification algorithms: *Linear SVM*, *Logistic Regression*, *Random Forest*, and *Decision Trees*. These algorithms are selected because the rankings of predictive features are well-defined. In particular, for linear models the importance of a feature is proportional to the magnitude of its coefficients, while for tree-based models the ranking follows the order of the attributes used to partition the data (i.e., the attribute used to split the root node has highest rank).

Our first observation is that the *SVM* and *Logistic Regression* models rely heavily on applicants' background data including their undergraduate major, level of education, and country of origin. On the other hand, *Decision Trees* and *Random Forest* models utilize quantitative attributes such as GRE quantitative/verbal/writing scores, overall GPA, student age, and undergraduate school rankings. The superior performance of the *Random Forest* model compared to other standalone algorithms suggests quantitative measures are more reliable metrics in predicting a student's potential success in the MSDS program.

Our next analysis involves distinguishing the positive and negative predictors among the highly ranked predictive features. To this end, we resort to the magnitude of positive and negative weights provided by the linear classifiers (i.e., *SVM* and *Logistic Regression*) together with our first-hand experience in overseeing our MSDS program. Our findings suggest that students with an undergraduate major in Business, Economics, International Studies, Humanities, and Communications are poor candidates for an MSDS program, while applicants with Computer Science, Electrical Engineering, (Applied) Mathematics or Statistics backgrounds are more likely to succeed in the program. High GRE scores, Overall (undergraduate) GPA, and School Ranking, are positive indicators for success. We find these discoveries of important practical values because of the interdisciplinary nature of a data science program. Because data science programs attract students from diverse backgrounds, our studies suggest that a solid mathematics, computer science, or engineering background is essential for a student to be highly successful in an MSDS program.

5. CONCLUSION

Graduate admissions is a challenging task because it is generally controlled by faculty that have other responsibilities and priorities, and have limited training in the admissions process. The admissions process for a graduate data science program has even more challenges because it is interdisciplinary, most students do not have undergraduate backgrounds in data science, and the degree program has not existed long enough so that there is significant institutional knowledge about what applicants make the best (and worst) data science students. Thus this is an area that can benefit from data mining. The results in this paper show that mining a combination of admissions application data and student performance data can help to identify those students who are likely to do well, as well as those that are likely to struggle in their studies.

The results in this study demonstrate that our models can effectively identify both top students, who could then be offered the merit-based aid that is allocated to the MSDS program, and the bottom-performing students, who then could be denied entry into the program. The results show that our best-performing algorithms can achieve an accuracy of about 90% when identifying either the top or bottom performing students. We feel that these results are sufficiently strong that it is reasonable to take action based upon them.

The data analyses conducted in this study, as well as the examination of the features that are most important for some of the classification models, both provide valuable insight into the factors that influence success in the data science program. The key conclusions are summarized below. However, in viewing these, it is important to understand that these conclusions are based on the performance of enrolled students, so those students with weak backgrounds (e.g., very low GRE scores) will have already been excluded. Our analysis shows that the GRE quantitative score and, to a lesser extent, the GRE verbal score, do impact performance in the program, but only to a modest degree— perhaps to a lesser extent than we expected. The TOEFL score has almost no impact (partially due to the fact that all admitted students have satisfactory TOEFL scores). However, the GPA from the last degree, which is usually an undergraduate degree, has a very strong impact on performance in the program. The major associated with the last degree also plays a significant role, with computer science, mathematics, and engineering degrees positively impacting success in the program, while business, communications, economics, psychology, and humanities degrees negatively impacting performance. Students who are foreign nationals, female, hold prior graduate degrees, or who have been out of school for more than a year also tend to perform well.

There are many ways in which this study can be extended. The most straightforward is to utilize more data. Given that the current MSDS program is thriving, we expect in a few years there will be substantially more labelled data, as well as more unlabelled data since many students will either be rejected or will choose not to enroll. We believe that we can leverage this unlabelled data to improve the results via the use of semi-supervised learning algorithms. During this study, we tried to leverage the existing unlabelled data, but there simply was not enough to have a significant impact given the diversity of the applicant pool. We also tried to use our domain knowledge to form data subgroups to fine-tune the model. We hypothesize that students with different backgrounds warrant different treatment, but currently we have too little data and too many feature values to leverage this information. Exploring algorithms such as SVM+ [17], which facilitates learning with heterogeneous data, can further improve the efficacy of our models. Finally, we are very interested in extending this work to other STEM graduate programs, including the MS in Cybersecurity and MS in Computer Science degree programs that currently reside in our department.

6. REFERENCES

[1] C. M. Antons and E. N. Maltz. Expanding the role of institutional research at small private universities: A case study in enrollment management using data

mining. *New directions for institutional research*, 2006(131):69–81, 2006.

[2] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider. Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, 113:177–194, 2017.

[3] B. Bakhshinategh, O. R. Zaiane, S. ElAtia, and D. Ipperciel. Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*, 23(1):537–553, 2018.

[4] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] L. Chang. Applying data mining to predict college admissions yield: A case study. *New Directions for Institutional Research*, 131:53–68, 2006.

[6] C. Cortes and V. Vapnik. Support vector machine. *Machine learning*, 20(3):273–297, 1995.

[7] T. G. Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[8] A. Dutt, M. A. Ismail, and T. Herawan. A systematic review on educational data mining. *Ieee Access*, 5:15991–16005, 2017.

[9] L. V. Fausett et al. *Fundamentals of neural networks: architectures, algorithms, and applications*, volume 3. Prentice-Hall Englewood Cliffs, 1994.

[10] H. C. Koh, G. Tan, et al. Data mining applications in healthcare. *Journal of healthcare information management*, 19(2):65, 2011.

[11] S. Lehr, H. Liu, S. Kinglesmith, A. Konyha, N. Robaszewska, and J. Medinilla. Use educational data mining to predict undergraduate retention. In *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*, pages 428–430. IEEE, 2016.

[12] S. W. Menard. *Applied logistic regression analysis*. Number 04; e-book. 1995.

[13] T. Mishra, D. Kumar, and S. Gupta. Mining students' data for prediction performance. In *2014 Fourth International Conference on Advanced Computing & Communication Technologies*, pages 255–262. IEEE, 2014.

[14] P. Nithya, B. Umamaheswari, and A. Umadevi. A survey on educational data mining in field of education. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 5(1):69–78, 2016.

[15] D. L. Olson, Y. Shi, and Y. Shi. *Introduction to business data mining*, volume 10. McGraw-Hill/Irwin Englewood Cliffs, 2007.

[16] J. R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[17] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22.5:544–557, 2009.

[18] S. K. Yadav, S. Pal, and B. Bharadwaj. Data mining applications: A comparative study for predicting student's performance. *International Journal of Innovative Technology and Creative Engineering*, 1(arXiv: 1202.4815):13–19, 2012.