A Machine Learning Approach to Predicting GRE Scores Using Graduate Application Information

Anonymous
Anonymous Institution
anonymous@anonymous.edu

ABSTRACT

GRE Aptitude Test scores have been a key criterion for admissions to U.S. graduate programs. However, many universities have lifted their standard test (e.g., GRE, SAT) requirements due to the COVID-19 pandemic, and many are unlikely to reinstate them after the pandemic ends. This change poses additional challenges in evaluating prospective students. In this paper, we examine the viability of applying machine learning models to predict applicants' GRE scores using the rest of their application materials. We utilize a diverse set of information from the student admissions application, including their undergraduate GPA, undergraduate major, and resume. Our work is based on a real-world dataset consisting of 814 MS in Computer Science and M.S. in Data Science applications submitted with GRE scores. We investigate two tasks that are of practical value: 1) predicting the GRE quantitative reasoning and verbal percentiles and 2) classifying the top 20% and bottom 20% of applicants based on GRE scores. The learned classification models for this second task can serve as a focus of attention tool for admission committees and aid in rendering scholarship and rejection decisions. We further identify and discuss the principal factors utilized by our models to better understand the relationship between the various application components and GRE performance. The findings also provide additional insight into what the GRE exam is really measuring.

Keywords

machine learning, graduate admission, GRE score prediction, supervised learning, educational data mining

1. INTRODUCTION

The Graduate Record Examination (GRE) is a cognitive abilities test administered by Educational Testing Service (ETS). The test consists of three components that measure applicants' competence in verbal reasoning (GRE-V), quantitative reasoning (GRE-Q), and analytical writing skills (GRE-A). Because of its objectivity, standardization, and predictive utility, GRE scores are used by many graduate schools as necessary criteria to admit qualified students. Indeed, an earlier study led by Norcross et al. showed that more than 90% of Ph.D. programs and over 80% of master's programs in the U.S. mandated GRE scores [16]. This is particularly the case in the science, technology, engineering, and mathematics (STEM) field. For example, estimated GRE scores for 2021 graduate programs across Stanford university range between 157-170 for GRE Verbal and between 155-170 for GRE Quantitative Reasoning. However, for a Masters in Engineering, the average Quantitative score is 167 [19].

Despite GRE's essential role in graduate admissions, many institutions were forced to lift the requirement due to the COVID-19 pandemic. The main concerns are the accessibility to physical test centers and potentially compromised scores from at-home tests. As a result, admission committees have to rely on more subjective materials such as the statement of intent (SOI) and letters of recommendation (LORs). Effective estimation of GRE scores could provide admission committees additional objective evaluations of the applicants.

This study aims to predict applicants' missing GRE scores using the rest of their application materials. The underlying assumption is that students' academic credentials and professional experiences are highly correlated to their performance on standardized tests. For example, applicants who have taken math courses (e.g., linear algebra or statistics) are likely to perform well in the quantitative section of the GRE test, and students' GPAs are in general positively correlated with their GRE scores. Table 1 presents the predictors utilized for this study. These features are extracted from student application materials and grouped into five categories: demographics, academic credentials, TOEFL performance (for international students), Math skills, and C.S. skills. We present the engineering of these features and their statistics in Section 3.

We built our predictive models using 800+ applications that include GRE scores from the M.S. in Data Science (MSDS) and M.S. in Computer Science (MSCS) programs at X[1] University. In this paper, we explore machine learning approaches for two GRE prediction tasks. First, we employ ridge regression to predict GRE-Q and GRE-V score per-

[1]name removed for blind review.

centiles. Next, we build classification models to identify the top 20% and bottom 20% of applicants based on their GRE-Q or GRE-V scores. The undertaking of the latter task is particularly valuable in reducing the workload of the admission committees given the continuous rise in graduate applications [7]. We further examine separate models for the international students leveraging their standard foreign language test scores (e.g., TOEFL).

Another motivation for this research is to study the principal factors in estimating applicant GRE scores. Given sufficient model predictive performance, the identified top predictors can be used for two purposes. First, they pinpoint the application components that are most relevant to applicant GRE abilities and, thus, can facilitate efficient and systematic applicant comparisons in the absence of GRE scores. Second, the principal predictors can help us identify and analyze potential racial and gender biases associated with GRE scores. A prior study has found that "the GRE is a better indicator of sex and skin color than of ability and ultimate success" [13]. Our study, however, suggests that gender and race are not top predictors in our models. We list and discuss our principal predictors in Section 5.4.

2. RELATED WORK

Investigating the validity of the GRE test and its predictive value in student performance in higher education is an active research area. Kuncel et al. launched extensive research examining the effectiveness of GRE in predicting performance at both the master and doctoral levels. Across nearly 100 studies and 10,000 students, their study found that GRE scores predict first year and overall GPAs well for both master's and doctoral students [10]. Young et al. investigated the validity of the GRE scores in the admission of Master of Business Administration (MBA) students using a sample of 480 admitted students[24]. They found that GRE-Q was the most influential predictor, followed by GRE-V and GRE-A, in predicting students' first term GPAs. Furthermore, they found that the three GRE test metrics are significantly more predictive than undergraduate GPAs.

Despite the value of standardized testing, some studies have found controversial evidence and, thereby, advocate the elimination of the GRE test. For example, Petersen et al. presented a multi-institutional study of GRE scores as predictors of STEM Ph.D. degree completion [17]. Their findings suggest that GRE scores are not a strong predictor of graduate school success and thus should not be considered the gold standard for admission. Likewise, Sealy et al. investigated the association between GRE scores and academic success among Ph.D. students' in biomedical sciences [18]. They concluded that the GRE scores were weak predictors of students' future academic success. In addition to questioning the predictive value of the GRE scores, researchers have also raised concerns about the test's fairness and implicit bias, considering the wide achievement gap in test scores between demographic groups [9, 15, 22].

It is worth noting that all the above studies focused on evaluating the validity of the GRE. Our study differs in that we investigate the possibility of predicting applicants' GRE performance using available application materials. This undertaking is valuable in two respects. First, as discussed above, there is controversy about the value of the GRE test. Consequently, some graduate programs may choose to eliminate the GRE requirements to attract a diverse pool of students. Estimated GRE scores based on student application materials can offer additional information in making admission decisions. Second, for programs that currently mandate GRE scores, accurate predictive models can facilitate lowering the requirement to make the GRE optional or just encouraged, to accommodate disadvantaged prospective students from rural and low-income backgrounds [9].

Table 1: Predictive Features

Demographics	**TOEFL Scores**	**CS Skills***
Gender	Listening	Python
Race	Speaking	Java
Age	Reading	C++
Permanent Country	Writing	Matlab
Native English Speaker		SAS
	Math Skills*	Database
Academic Credentials	Calculus	Microsoft Office
Undergraduate Major	Linear Algebra	Machine Learning
Undergraduate GPA	Statistics	Software
Months Since Degree		

* Each skill represents a binary feature indicating if the applicant's resume contains the keyword.

3. DATA AND PRE-PROCESSING

We based our study on 814 GRE-available applications submitted to the MS of Data Science (MSDS) and MS of Computer Science (MSCS) program at X[1] University. Of these, 300 are international applications. The features we collected from the structured application template can be classified into demographics, academic credentials, and TOEFL scores for international students. We also automatically extracted Math and CS skill information from applicants' resumes. Table 1 presents the specific features in each category. Note that the math and CS skill features are binary indicators for a given keyword. For example, the "C++" feature will take value 1 if the applicant's resume contains the "C++" keyword and 0 otherwise.

3.1 Feature Statistics

This section summarizes the distribution of features values for key demographic features and non-demographic features related to the applicants. These distributions are summarized in Figure 1. The Age chart shows that the vast majority (96%) of applicants are under 30, with approximately two-thirds of these under 24 and the remaining one-third between 24 and 30. We see that only a bit more than one-third (36%) of the applicants are female, which is consistent with the Computer Science and Data Science fields in the United States being male-dominated. In fact, the applicant pool has more gender diversity than one might expect based on U.S. Bureau of Labor Statistics for 2021 [21], which show that only 26% of the Computer and Mathematical Occupations in the U.S. are staffed by women.

The information in Figure 1 further shows that 84% of the applicants do not have English as their first language, which is mainly due to the large number of applicants who come from China and, to a lesser degree, India. In fact 67% of applicants have China or India listed as their permanent

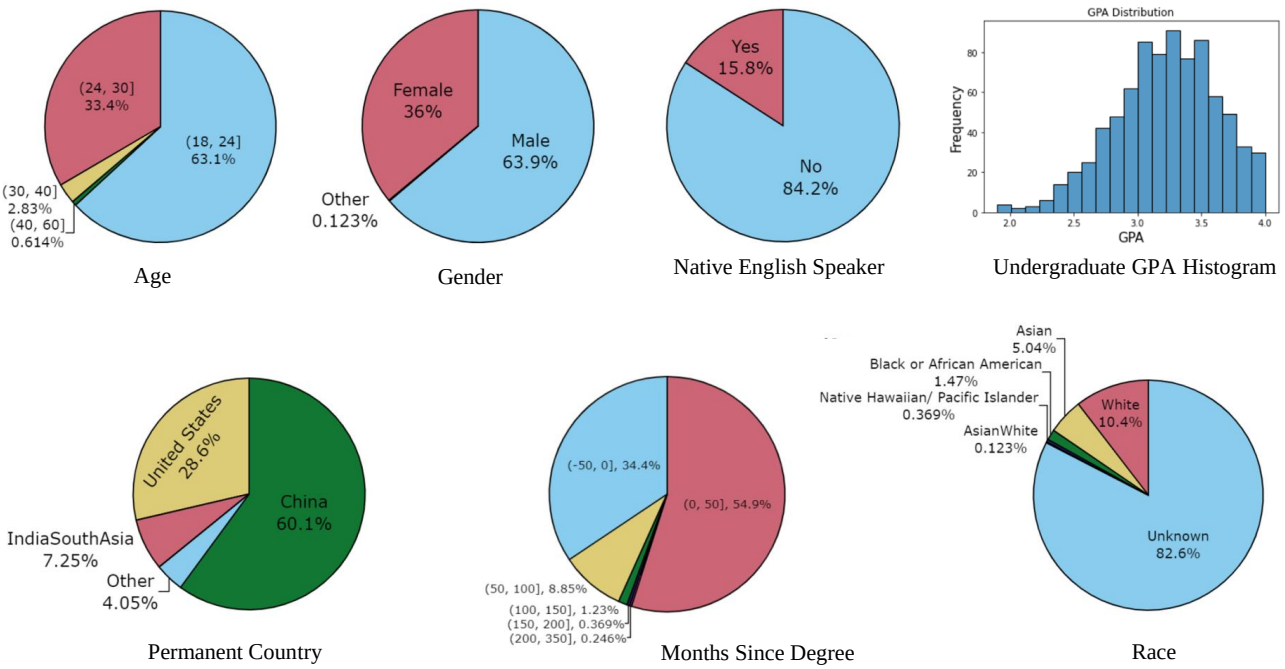

Figure 1: Statistics of Demographic and Academic Credential Features

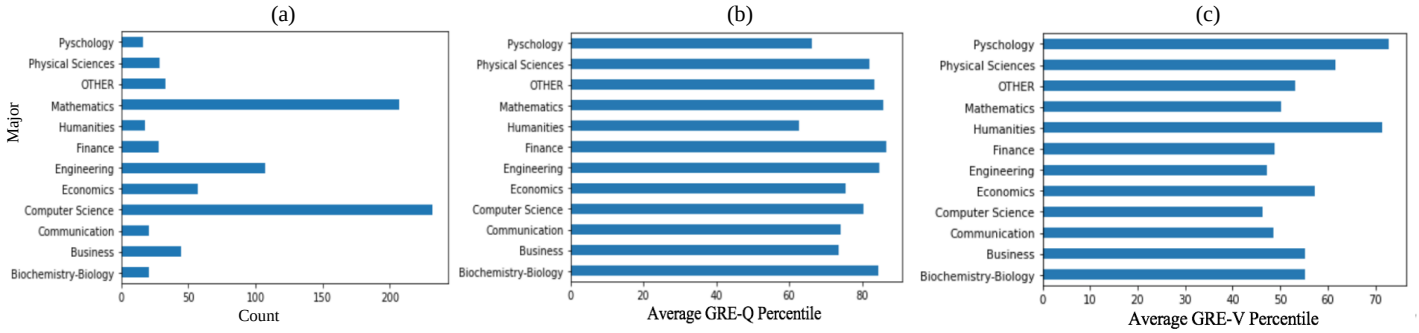

Figure 2: Applicant Count (a), Average GRE-Q (b) and GRE-V (c) Percentiles by Major

country, although we know that even more come from these countries since many subsequently become US citizens (i.e., many of the 26% that list the US as their permanent country were born overseas). The racial information provided by the applicants is not very helpful since that information typically (82%) is not specified. The Months since Degree information indicates the time from completion of the student's last degree to the time of submitting the application. For 34% of the students this time is negative, indicating that they are applying while still completing their undergraduate degree, while most of the rest (55%) have completed their last degree within roughly the last four years. In less than 10% of the cases has the student completed their last degree more than four years ago.

We expect the applicant's grade point average (GPA) to be one of the key features for predicting the GRE scores, so it is important to have a good understanding of the GPA distribution. We expect most of the applicants to have good grades since the university's general policy is to admit students only with an undergraduate GPA of 3.0 or better. The GPA distribution of the applicants generally conforms to this, although a substantial percentage of the applicants do have a GPA below 3.0 (28% are below 3.0 and 5% are below 2.5).

3.2 Applicant Prior Major Discipline

The applicant's academic major is a particularly important feature since it provides a great deal of information about what prior education the student received. Figure 2(a) shows the distribution of the applicants by (undergraduate) major category. Because the MSCS and MSDS degrees are technical, one would expect most of the undergraduate degrees to be from STEM fields, and many should be specifically in a

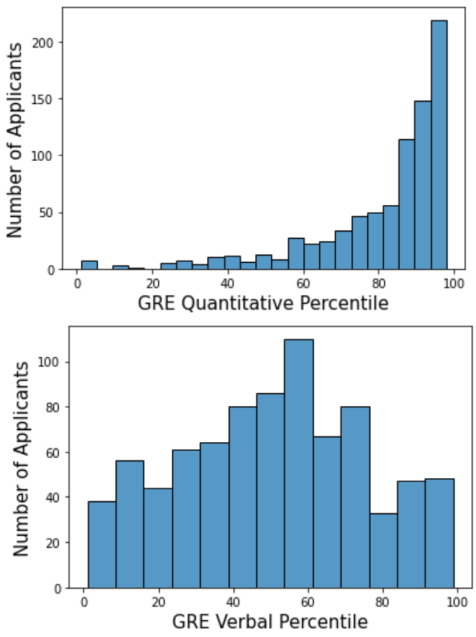

Figure 3: GRE-Q (left) and GRE-V (right) Percentile Histograms

Computer Science related area since most applicants to the MSCS program should have some form of Computer Science degree. These values in Figure 2(a) generally adhere to this, but there are a non-trivial number of non-STEM (e.g., humanities, communication, economics, business) applicants.

One might expect the applicants with non-STEM degrees to have, on average, lower GRE quantitative scores and perhaps higher GRE verbal scores. This is largely what is observed in Figure 2(b) and Figure 2(c). Humanities majors have the lowest GRE-Q score, followed by psychology majors, and then economics, communications, and business majors. The psychology quantitative scores may seem anomalous, but psychology is not always considered a STEM discipline and sometimes is considered a social science. Conversely, the STEM disciplines are associated with consistently high GRE-Q scores. Overall, the biggest surprise may be the very high GRE-Q scores for Finance majors. The patterns with the GRE-V scores tend to show that STEM majors yield low GRE-V scores and non-STEM fields yield higher GRE-V scores, but in many cases the fields perform similarly. The standouts are that psychology, humanities, and to a lesser degree economics, yield the highest GRE-V scores. Perhaps most surprising is that the communications majors have relatively low GRE-V scores.

3.3 Distributions of GRE Score Percentiles

The distribution of the class variable is of special importance, and for that reason, we take a close look at the distribution of the GRE quantitative and verbal distributions by percentile. These are shown in Figure 3. Note that the GRE-Q applicant percentile scores are heavily asymmetric and skewed toward the 100% percentile. This outcome is expected since applicants to the MSCS and MSDS degrees should have very high GRE-Q scores. It is worth noting, however, that while low scores are rare, they do exist (8% of applicants are below the 50th percentile), and in fact extend all the way down to the very bottom. The distribution of GRE-V scores is far more traditional. It is much more symmetric and evenly distributed than the GRE-Q scores. Note that these different distributions have implications for the associated regression and classification tasks. Guessing the average value will yield much better regression results for the quantitative scores than the verbal scores, since the quantitative scores are much more tightly concentrated. On the other hand, the nature of the problem will make it much harder to beat the strategy of always guessing the average value. Similarly, identifying the top 20% of GRE-Q percentile scores will not be easy since so many applicants have scores close to this border, but identifying the bottom 20% will not have this issue and therefore may be more achievable.

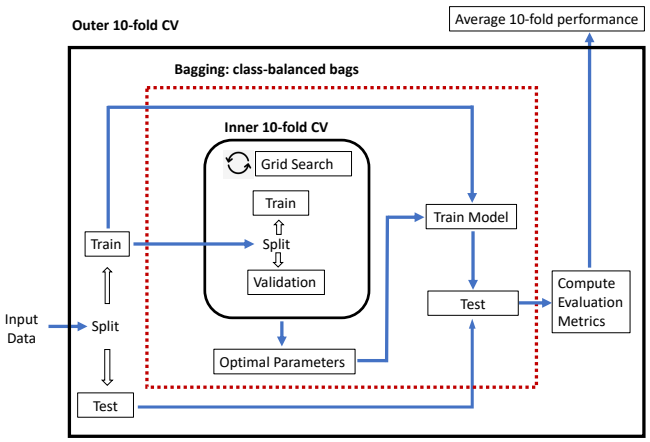

Figure 4: Model Training and Evaluation Architecture

4. METHODOLOGY

In this section, we first illustrate our model training and evaluation framework. We next introduce each regression and classification model and its optimal hyperparameter values used in model training. These values were selected via grid search [3] using a nested 10-fold cross-validation on the training data. The unspecified hyperparameters use the default values in Python's *scikit-learn* library, the software we used to implement our models. Lastly, we describe how we address the imbalanced training data.

4.1 Experimental Design

Figure 4 illustrates our model training framework. We evaluate each model's performance using a 10-fold (outer) cross-validation. The process involves randomly splitting the data into ten disjoint groups (i.e., folds) of approximately equal size. Subsequently, each model is trained ten times using the i^{th} $(i = 1, 2, ..10)$ fold as the test data, and the remaining nine folds as the training data (T_i). For each evaluation metric, we report a model's performance as the mean of the ten out-of-sample scores on the 10 test folds, indicated by the upper right box in Figure 4. The red box indicates the "bagging" technique used to address data imbalance in training the classification models.

We further apply a nested cross-validation to facilitate hyperparameter selection. To this end, we partition training data T_i in each outer iteration i into ten folds and conduct a grid search [3] on a set of algorithm-specific parameters. The optimal parameter set (P_i) for T_i is chosen to produce the highest average AUC (Area Under the ROC Curve) [5] score on the 10 test folds. Finally, we report the performance of the model trained using T_i and P_i.

4.2 Regression Models

Linear regression models assume a linear relationship between the observed feature vector $\mathbf{x}$ and the target variable y. Formally,

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d = \mathbf{w}^T \mathbf{x}$$

where $\mathbf{x} = (1, x_1, x_2, \ldots, x_d)^T$ are the observed features, y is the real-valued target, and $\mathbf{w} = \{w_0, w_1, \ldots, w_d\}^T$ are the model parameters. Training a linear regression model entails minimizing the mean squared error (MSE) defined as

$$E(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{w}^T \mathbf{x}_n - y_n)^2 \quad (1)$$

where N is the number of training samples.

We experimented with three linear regression models in predicting the GRE-Q and GRE-V percentiles. These models applied different regularization techniques to prevent overfitting.

4.2.1 LASSO Regression

LASSO (Least Absolute Shrinkage and Selection Operator) regression [20] encourages simple, sparse models by adding the L_1-norm of model complexity to the cost function defined in (1). Specifically, LASSO minimizes

$$E(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{w}^T \mathbf{x}_n - y_n)^2 + \lambda \parallel \mathbf{w} \parallel_1$$

where $\parallel \mathbf{w} \parallel_1 = \sum_{j=1}^{d} |w_j|$ and λ is the trade-off hyperparameter controlling the regularization strength. A desirable property of LASSO regression is shrinkage and feature selection. In particular, regression coefficients for less predictive features will be shrunk toward zero and variables with non-zero regression coefficients are the principal predictors.

In our study, we applied grid search to optimize the trade-off hyperparameter λ. Depending on the specific tasks, the optimal λ values varied from 0.12 to 0.4.

4.2.2 Ridge Regression

Ridge regression [8] follows the sample regularization principle as LASSO except that it applies a L_2-norm penalty instead of L_1-norm. Thus, ridge regression minimizes

$$E(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{w}^T \mathbf{x}_n - y_n)^2 + \lambda \parallel \mathbf{w} \parallel_2$$

where $\parallel \mathbf{w} \parallel_2 = \sum_{j=1}^{d} |w_j|^2$ and λ is the trade-off hyperparameter controlling the regularization strength.

Ridge regression is desirable because the quadratic penalty term makes the cost function strongly convex, engendering a unique minimum and a closed-form solution. We further explored kernelized ridge regression with polynomial and RBF kernels. The best models were achieved using the RBF kernel with gamma = 0.01 and polynomial kernels with degrees ranging from 2 to 7.

4.2.3 Elastic Net

Elastic Net combines LASSO and ridge regression to exploit the advantages of both approaches. In addition, the elastic net encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together. The cost function of elastic net is defined as

$$E(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{w}^T \mathbf{x}_n - y_n)^2 + \lambda \left(\frac{(1-\alpha)}{2} \parallel \mathbf{w} \parallel_2 + \alpha \parallel \mathbf{w} \parallel_1 \right.$$

where α controls the trade-off between L_1 and L_2 regularization, and $\alpha = 1$ is equivalent to LASSO, and $\alpha = 0$ is equivalent to Ridge. Through grid search, we found that the optimal λ ranged from 0.06 to 0.38 for our models, and α ranged from 0.7 to 1.

4.3 Classification Models

We experimented with five classification models to identify the top 20% and bottom 20% of applicants based on their GRE-Q and GRE-V percentiles.

4.3.1 Decision Tree

A decision tree (DT) [14] employs a tree structure to model a decision-making process, in which each internal node performs a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a decision (e.g., class label). An essential task in building a DT model is to select the optimal attribute for each internal node. *Information Gain* and *Gini impurity* are the two popular metrics used to maximize the purity of data in the after-split branches. We trained our DT model using the *GINI impurity* criterion. The optimal maximum tree depth ranged from 3 to 5.

4.3.2 Random Forest

A Random forest [1] model is a collection of decision trees, each of which is trained with a randomly sampled subset of training instances and attributes. A *Random Forest* model strives to achieve robust and superior performance by ensembling predictions from multiple decision trees. In our study, we searched for the optimal hyperparameter values for the number of decision trees, minimum samples at a leaf node, and the minimum number of samples required to split an internal node. Depending on individual models, their values ranged from 40 to 280, 1 to 5, and 2 to 12, respectively.

4.3.3 Logistic Regression

Logistic regression (LG) [12] classifies a binary dependent variable (i.e., label) using a linear combination of the one or more existing independent variables (i.e., features). Formally,

$$P(y = 1|\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x})$$

where $\mathbf{x} = (1, x_1, x_2, \ldots, x_d)^T$ are the dependent variables, $y \in \{0, 1\}$ is the dependent variable, and $\mathbf{w} = (w_0, w_1, \ldots, w_d)^T$

Table 2: Regression Results of GRE Quantitative Score Percentile Prediction

Group	Model	Training			Test		
		MAE	RMSE	R^2	MAE	RMSE	R^2
Without TOEFL (814 samples)	Lasso	9.12	12.67	0.38	9.56	13.25	0.31
	Ridge	9.09	12.52	0.39	9.59	13.17	0.31
	Kernel Ridge	8.42	11.56	0.48	9.66	13.28	0.30
	Elastic Net	9.10	12.63	0.38	9.54	13.22	0.31
	Average	**8.93**	**12.35**	**0.41**	**9.59**	**13.23**	**0.31**
With TOEFL (300 samples)	Lasso	6.95	9.98	0.44	7.27	10.36	0.37
	Ridge	6.68	9.62	0.48	7.43	10.60	0.34
	Kernel Ridge	6.80	9.70	0.47	7.36	10.32	0.38
	Elastic Net	6.93	9.97	0.44	7.26	10.36	0.37
	Average	**6.84**	**9.82**	**0.46**	**7.33**	**10.41**	**0.37**

Table 3: Regression Results of GRE Verbal Score Percentile Prediction

Group	Model	Training			Test		
		MAE	RMSE	R^2	MAE	RMSE	R^2
Without TOEFL (814 samples)	Lasso	17.82	21.73	0.24	18.42	22.40	0.17
	Ridge	17.90	21.78	0.24	18.49	22.43	0.17
	Kernel Ridge	17.65	21.50	0.26	18.62	22.63	0.16
	Elastic Net	17.86	21.76	0.24	18.45	22.42	0.17
	Average	**17.81**	**21.69**	**0.25**	**18.50**	**22.47**	**0.17**
With TOEFL (300 samples)	Lasso	12.07	14.69	0.51	12.91	15.56	0.40
	Ridge	11.89	14.52	0.52	12.67	15.38	0.41
	Kernel Ridge	11.55	14.07	0.55	12.76	15.35	0.41
	Elastic Net	11.97	14.61	0.52	12.81	15.45	0.41
	Average	**11.87**	**14.47**	**0.53**	**12.79**	**15.44**	**0.41**

are the model coefficients. σ denotes the sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$ which maps a number $z \in (-\infty, +\infty)$ to $(0, 1)$. We experimented with different regularization parameter (α) and the models were trained with $\alpha = 1$.

4.3.4 Neural Network

Neural networks (NN) [11] are computational models inspired by the architecture of neurons in a biological brain. The nodes are divided into sequential layers with learnable weights connecting the nodes across adjacent layers. The first and last layers represent the model's input and output, respectively, and the middle ones are hidden layers. In this study, we used a network with two hidden layers, which had 32 and 16 neurons, respectively. We applied ReLU activation and trained our models using the Adam optimizer with a batch size of 64.

4.3.5 XGBoost

XGBoost [2] is a homogeneous ensemble method in which the base learners are generated from a single machine learning algorithm, exploiting the concept of "adaptive boosting" [6]. Unlike the traditional boosting technique, which adjusts the penalty weight for each data point before training the next learner, XGBoost fits the new learner to residuals of the previous model and then minimizes the loss when adding the latest model. The process is equivalent to gradient descent converging to a local optimum. Our XGBoost models were trained with a learning rate of 0.2. The optimal number of iterations ranged from 5 to 50, depending on each specific classification task. Similarly, the values for maximum depth ranged from 1 to 7.

4.4 Addressing Imbalanced Data

Since our classification tasks aim to identify the top 20% and the bottom 20% of applicants, the training data is imbalanced with a class ratio of 1:4. Learning directly from imbalanced data often leads to unsatisfactory performance in the minority class when the machine learning algorithms strive to minimize the global loss. To this end, we employed the "bagging" technique [23]. In particular, k "bags" of balanced datasets were created where each bag contained all minority class instances and an equal number of randomly sampled majority class instances. Each subset of majority instances were sampled with replacement from the entire majority population. Consequently, k sub-models were trained using these balanced "bags" of data. The final model predictions were made by using a majority vote on the sub-models' decisions. We selected optimal k as a hyperparameter for each classification model.

5. RESULTS

In this section, we first present the experimental results of our regression and classification models. We then analyze the principal predictors identified by our models.

5.1 Regression Results

Table 2 presents our models performance in predicting GRE-Q percentiles. The average mean absolute errors (MAEs) for

Table 4: Quantitative Percentile Classification Performance Comparison

Group	Model	Overall Accuracy	Recall	Specificity	Precision	F-1	AUC
Without TOEFL (814 samples)	Top 20% vs. Rest						
	DT	0.61	0.60	0.61	0.30	0.40	0.67
	RF	0.56	0.69	0.52	0.29	0.40	0.67
	LG	0.58	0.67	0.55	0.29	0.41	0.65
	NN	0.55	0.70	0.51	0.28	0.40	0.65
	XGBoost	0.61	0.61	0.62	0.30	0.41	0.67
	Average	**0.58**	**0.65**	**0.56**	**0.29**	**0.40**	**0.66**
	Bottom 20% vs. Rest						
	DT	0.77	0.72	0.78	0.45	0.55	0.82
	RF	0.76	0.82	0.74	0.44	0.58	0.84
	LG	0.79	0.79	0.78	0.48	0.59	0.85
	NN	0.79	0.80	0.78	0.48	0.60	0.83
	XGBoost	0.79	0.80	0.79	0.49	0.61	0.85
	Average	**0.78**	**0.79**	**0.77**	**0.47**	**0.59**	**0.84**
With TOEFL (300 samples)	Top 20% vs. Rest						
	DT	0.69	0.58	0.72	0.40	0.47	0.70
	RF	0.63	0.64	0.62	0.35	0.46	0.68
	LG	0.67	0.68	0.66	0.39	0.50	0.72
	NN	0.60	0.71	0.57	0.34	0.46	0.70
	XGBoost	0.67	0.60	0.69	0.39	0.47	0.69
	Average	**0.65**	**0.64**	**0.65**	**0.37**	**0.47**	**0.70**
	Bottom 20% vs. Rest						
	DT	0.80	0.69	0.83	0.52	0.59	0.86
	RF	0.74	0.79	0.73	0.43	0.56	0.88
	LG	0.82	0.73	0.85	0.56	0.63	0.87
	NN	0.79	0.69	0.81	0.49	0.57	0.84
	XGBoost	0.81	0.76	0.82	0.52	0.62	0.87
	Average	**0.79**	**0.73**	**0.81**	**0.50**	**0.59**	**0.86**

the general and international groups are 9.59 and 7.33, respectively. Given that GRE score-percentile mappings are not uniform and have large gaps [4], we believe an MAE within 10% can be used as effective estimations. For example, the score difference between 89% and 79% of GRE-Q percentiles is only 4 (i.e., 167 vs. 163). Thus, we believe all four machine learning models can be used to provide a reasonable estimation of an applicant's performance in GRE quantitative reasoning, and Elastic Net demonstrates a marginal advantage over the others three models.

Table 3 presents the models performance in predicting GRE-V percentiles. The MAE for the general and international groups are 18.5% and 12.79%, respectively. The GRE scoring guide [4] shows that the gaps in verbal percentiles are in general similar to the quantitative ones. Thus, our experimental results suggest that our current features may not be sufficient in providing effective estimations for an applicant's performance in GRE verbal reasoning. Perhaps this is not too surprising, however, since the two degrees in question rely more on quantitative abilities than on verbal abilities, and the application materials submitted may respond to that. We discuss potential improvements in Section 6.

5.2 Quantitative Classification

Table 4 presents the performance of our machine learning models in identifying the top 20% and bottom 20% of applicants based on GRE-Q percentiles. Our first observation is that it is much harder to correctly classify the top 20% compared to the bottom 20%. This is evidenced by the substantially higher overall accuracy and AUC scores in the latter task. Specifically, for the without-TOEFL group, the average overall accuracies for these two tasks are 58% and 78%, respectively; the average AUC scores are 0.66 and 0.84, respectively. Furthermore, this trend is consistent across both without- and with-TOEFL cohorts. In Figure 3 we observed that the distribution of GRE-Q percentiles is notably skewed towards the right. As a result, it is not surprising that the top 20% applicants are more challenging to identify than the bottom 20%.

Our second observation is that the models' performances are consistently better for the international (i.e., with-TOEFL) group. One explanation could be the increased model expressiveness from the additional TOEFL features. Another underlying factor could be that most of our international students are from one country (i.e., China) and, thus, the data is more likely to form an i.i.d. distribution, which is the fundamental assumption for most classifier-learning algorithms.

Lastly, for the general group, five machine learning models offered similar performances in terms of AUC scores in both classification tasks. Compared to other methods, XG-

Table 5: Verbal Percentile Classification Performance Comparison

Group	Model	Overall Accuracy	Recall	Specificity	Precision	F-1	AUC
Without TOEFL (814 samples)	Top 20% vs. Rest						
	DT	0.75	0.52	0.82	0.46	0.49	0.75
	RF	0.66	0.56	0.68	0.34	0.42	0.70
	LG	0.71	0.59	0.74	0.40	0.47	0.74
	NN	0.70	0.59	0.73	0.39	0.47	0.74
	XGBoost	0.72	0.53	0.78	0.41	0.46	0.73
	Average	**0.71**	**0.56**	**0.75**	**0.40**	**0.46**	**0.73**
	Bottom 20% vs. Rest						
	DT	0.63	0.58	0.64	0.29	0.38	0.65
	RF	0.57	0.63	0.56	0.23	0.37	0.63
	LG	0.61	0.61	0.60	0.28	0.38	0.68
	NN	0.62	0.56	0.63	0.27	0.37	0.65
	XGBoost	0.70	0.54	0.74	0.34	0.42	0.68
	Average	**0.63**	**0.58**	**0.63**	**0.28**	**0.38**	**0.66**
With TOEFL (300 samples)	Top 20% vs. Rest						
	DT	0.81	0.80	0.82	0.54	0.65	0.87
	RF	0.80	0.88	0.78	0.52	0.65	0.88
	LG	0.79	0.86	0.77	0.50	0.64	0.88
	NN	0.73	0.95	0.67	0.44	0.60	0.88
	XGBoost	0.81	0.88	0.79	0.53	0.66	0.89
	Average	**0.79**	**0.87**	**0.77**	**0.51**	**0.64**	**0.88**
	Bottom 20% vs. Rest						
	DT	0.77	0.77	0.77	0.46	0.58	0.86
	RF	0.76	0.85	0.74	0.46	0.60	0.86
	LG	0.80	0.79	0.80	0.51	0.62	0.88
	NN	0.79	0.84	0.78	0.50	0.62	0.87
	XGBoost	0.77	0.82	0.76	0.47	0.60	0.87
	Average	**0.78**	**0.81**	**0.77**	**0.48**	**0.60**	**0.87**

Boost demonstrated a more balanced performance between the minority and majority classes. For the international group, logistic regression led the performance in predicting the top 20% of applicants in AUC score, which offered 68% and 66% Recall and Specificity, respectively. In predicting the bottom 20% of applicants, RF, LG, and XGBoost are all excellent choices depending on the desired trade-off between the Recall and Specificity.

5.3 Verbal Classification

Table 5 presents the performance of our machine learning models in identifying the top 20% and bottom 20% of applicants based on GRE-V percentiles. For the general group, we observe that it is easier to identify the top 20% (average AUC: 0.73) than the bottom 20% of applicants (average AUC:0.66). This trend is opposite to the GRE-Q results. Figure 3 shows that the GRE-V percentile follows a close to normal distribution skewed slightly to the left. Thus, it is not surprising that the top 20% of applicants are easier to identify.

We further observe that the models' performances are significantly better for the international students. This outcome is consistent with what we observed in the above GRE-Q results, and we believe the same underlying factors could have contributed to the discrepancies.

In terms of model performance, LG seems to provide the highest practical value for the general group. Specifically, LG delivered a balanced performance of (Recall:0.59, Specificity: 0.74) and (Recall:0.61, Specificity: 0.61) for the top 20% and bottom 20% tasks, respectively. For the international group, all models can provide practical value. The choice would depend on the desired trade-off between the accuracies of the two classes.

5.4 Analysis of Principal Predictors

This section presents some interesting findings on statistically significant features identified by our regression models. Specifically, we examined the p-value associated with each variable in the LASSO, ridge regression, and Elastic Net models and studied the common predictors from all three models satisfying p-value <0.05. (i.e., 95% confidence interval). Our findings are summarized as follows.

- Software skills (Python, Matlab, Statistics) and Undergraduate GPA are positively correlated with GRE-Q performance, while MS Office is negatively correlated. The finding is consistent with our experience with reading MSDS/MSCS applications. Applicants that highlight MS Office as one of their key skills in their resume tend to be compensating for a lack of software skills, and generally lack the STEM background necessary to succeed in our programs.

- Our models identified four essential features that are positively correlated with GRE-V percentiles: Native English Speaker, Undergraduate GPA, Machine Learning, and Linear Algebra. The first two are reasonable factors, but the latter two are less intuitive. It is only natural that students who are raised in English-speaking countries will, on average, perform best on the GRE-V exam; however, this highlights the bias inherent in such an exam and careful thought should be given to using the associated scores in a responsible manner. This is especially true if the type of English verbal abilities measured by the exam are not essential for the program the student is applying to. In these cases, perhaps the TOEFL exam is more appropriate and can suffice.

- With p-value<0.05, the models did not discover any statistically significant features that are negatively correlated to GRE-V performance. If we relax the condition, we found Statistics and Software are the two negative predictors for verbal reasoning. One explanation for this is that high achievers at GRE-V tend to major in humanities and are less capable of STEM skills (i.e., math and software)

- For the international group, the models found that all TOEFL components played essential roles in predicting GRE performance except for TOEFL Speaking. In particular, TOEFL Reading and Listening are top positive predictors for both GRE-Q and GRE-V, and TOEFL Writing was an additional top predictor for GRE-V. Our findings suggest that TOEFL components can perhaps take on the role of the GRE assessment if it is available.

- It is worth noting that Undergraduate GPA was not an essential predictor for the international group even after converting all grading systems to a 4.0 scale. One possible explanation is the discrepancies in grading standards across global universities.

6. CONCLUSION AND FUTURE WORK

In this paper, we studied the viability of predicting applicants' performance on GRE quantitative and verbal reasoning using demographic and quantitative information from their application materials. Our experimental results suggest that features extracted from standard application templates can provide practical estimations for GRE-Q percentiles but fell short for the GRE-V estimations. Future work will be to determine how to look at textual materials such as students' statement of intent (SOI) and letters of recommendation (LORs) for more predictive signals. For example, we conjecture that the quality of SOI, which can be estimated via automated means, is predictive for the GRE verbal/writing scores. We also evaluated the ability to identify high-achieving (i.e., top 20%) and less-achieving (i.e., bottom 20%) applicants. To this end, we employed five classification models and the experiment showed convincing results.

We further performed principal predictor analysis and shared our findings. These findings provide insight into the application components that are most relevant to the GRE scores, and may help us to prioritize these components. It also highlighted that the GRE-V score itself has a potential (perhaps obvious) bias towards those who were born in an English-speaking country, and showed that the TOEFL could be a valid replacement. It also showed that automatic extraction of pre-specified keywords on an applicant's resume can provide useful information for making admission decisions.

Although the performance of our approach is modest, our results demonstrate great promise in adopting machine learning approaches to estimate applicants' missing GRE scores using their available application information. Our classification models can help identify top and bottom candidates and, thus, serve as a focus of attention (FOA) tool for an admission committee to render rejection and scholarship decisions. We are highly motivated to continue this work because of the continued rise of MSDS/MSCS applicants at X[1] University and we hope to use ML to help decrease the workload of the admissions committees.

7. REFERENCES

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

[2] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[3] M. Claesen and B. De Moor. Hyperparameter search in machine learning. *arXiv preprint arXiv:1502.02127*, 2015.

[4] ETS. Gre general test interpretive data. `ets.org/s/gre/pdf/gre_guide_table1a.pdf`, 2020.

[5] T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[6] Y. Freund, R. Schapire, and N. Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[7] Higher Education Today. Graduate enrollment and completion trends over the past year and decade. `https://www.higheredtoday.org/2021/11/10/trends-graduate-enrollment-completion/`, 2020.

[8] A. E. Hoerl and R. W. Kennard. Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970.

[9] J. C. Hu. Online gre test heightens equity concerns, 2020.

[10] N. R. Kuncel, S. Wee, L. Serafin, and S. A. Hezlett. The validity of the graduate record examination for master's and doctoral programs: A meta-analytic investigation. *Educational and Psychological Measurement*, 70(2):340–352, 2010.

[11] J. Li, J.-h. Cheng, J.-y. Shi, and F. Huang. Brief introduction of back propagation (bp) neural network algorithm and its improvement. In *Advances in computer science and information engineering*, pages 553–558. Springer, 2012.

[12] S. Menard. *Applied logistic regression analysis*, volume 106. Sage, 2002.

[13] C. Miller, K. Stassun, et al. A test that fails. *Nature*, 510(7504):303–304, 2014.

[14] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown. An introduction to decision tree

modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(6):275–285, 2004.

[15] D. A. Newman, C. Tang, Q. C. Song, and S. Wee. Dropping the gre, keeping the gre, or gre-optional admissions? considering tradeoffs and fairness. *International Journal of Testing*, 22(1):43–71, 2022.

[16] J. C. Norcross, J. M. Hanych, and R. D. Terranova. Graduate study in psychology: 1992–1993. *American Psychologist*, 51(6):631, 1996.

[17] S. L. Petersen, E. S. Erenrich, D. L. Levine, J. Vigoreaux, and K. Gile. Multi-institutional study of gre scores as predictors of stem phd degree completion: Gre gets a low mark. *PLoS One*, 13(10):e0206570, 2018.

[18] L. Sealy, C. Saunders, J. Blume, and R. Chalkley. The gre over the entire range of scores lacks predictive ability for phd outcomes in the biomedical sciences. *PloS one*, 14(3):e0201634, 2019.

[19] C. Swimmer. Stanford gre scores. `https://magoosh.com/gre/stanford-gre-scores/#:~:text=Estimated%20Stanford%20GRE%20scores%20for,vary%20by%20program%20and%20emphasis`, 2020.

[20] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[21] U.S. Bureau of Labor Statistics. Labor force statistics from the current population survey. `https://www.bls.gov/cps/cpsaat11.htm`, 2021.

[22] S. E. Woo, J. LeBreton, M. Keith, and L. Tay. Bias, fairness, and validity in graduate admissions: A psychometric perspective. *Preprint*, 2020.

[23] B. W. Yap, K. A. Rani, H. A. A. Rahman, S. Fong, Z. Khairudin, and N. N. Abdullah. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets, 2014.

[24] J. W. Young, D. Klieger, J. Bochenek, C. Li, and F. Cline. The validity of scores from the gre® revised general test for forecasting performance in business schools: Phase one. *ETS Research Report Series*, 2014(2):1–10, 2014.