Data analysis: Graduate Admission 2 Dataset

1 Introduction

International students are common in the US, especially in Master's programs. According to data coming from the UCSD graduate admission division(UCSDGraduateDivision, 2021), in 2021 fall 50% of admitted students in Master's programs are international students. When they apply, they have to submit a series of documents to certify that they are excellent enough to earn diplomas from these famous universities. This is an extremely competitive process so they prepare a lot of documents such as high scores in TOEFL and GRE, high-quality research experience, recommendation letters from famous professors, and so on. Among these documents, we want to research whether they really can improve their admission chance and whether we can predict their admission chance just from the documents they provide. We hope the research can provide some orientation to international students who want to enter famous universities in the US to get their Master's degree.

2 Dataset

The Graduate Admission 2 dataset (Mohan S Acharya, 2019) contains information on various factors that may affect a student's chances of being admitted to a graduate program. The dataset includes information on the student's undergraduate GPA(CGPA), research experience, TOEFL and GRE score, statement of purpose(SOP) and letter of recommendation strength(LOR), the rating of their undergraduate institution, and admission chances. The data can also help to identify potential barriers to graduate school admissions and inform efforts to improve the admissions process.

3 Data Analysis

3.1 Basic Information and Metrics

In this part, we want to show some basic information and metrics about our dataset. We have 8 metrics in total. Seven documents provided by applicants, GRE score(range from 260-340), TOEFL score(range from 0-120), University rating of applicants' undergraduate school(1-5), Statement of Purpose(range from 1-5), Letter of Recommendation Strength(range from 1-5), GPA(range from 0-10), Research Experience(0 for no and 1 for yes). From the above, we can calculate the chance of admission for each applicant. **Table 1** shows some basic statistical metrics of factors in our dataset. Because the research experience is binary, we think it is meaningless if we show it in this way. We will utilize it in other parts of our research.

3.2 Correlations

In this part, we want to represent the correlations between each pair of factors in our dataset. We calculate the correlation coefficients of them and **Fig 1** shows them by a heatmap. We find that the CGPA, TOEFL, GRE are most relative to the chance of admission. Among them, the CGPA is obviously higher than the other two factors. Also, we also can know that the correlations between each two of the three factors are also strongly relative. However, whether an applicant has research experience is relatively uncorrelated to the chance of admission. Also, the research experience is weakly correlated to any other factors. The results represent that if students focus on applying for Master's programs in the US, they may focus more on their academic grades. High grades in various examinations are much more useful to persuade

Table 1: Basic statistical metrics of Graduate Admission 2.

	mean	std	min	25%	median	75%	max
GRE	316.62	11.37	290	308	317	325	340
TOEFL	107.29	6.07	92	103	107	112	120
University rating	3.10	1.14	1.00	2.00	3.00	4.00	5.00
SOP	3.39	0.99	1.00	2.50	3.50	4.00	5.00
LOR	3.47	0.91	1.00	3.00	3.50	4.00	5.00
CGPA	8.58	0.60	6.8	8.14	8.57	9.05	9.92
Chance of Admit	0.72	0.14	0.34	0.64	0.73	0.82	0.97

dream schools to admit them. However, those who lack research experience should not be worried too much. The research experience may not severely affect your chance of admission.

3.3 Differences w/ and wo/ research experience

In this part, we want to research the distributions of each pair of factors with and without research experience. From the above, we know that the research experience seems relatively not correlated to the chance of admission. Also in our dataset, the research experience is binary(applicants have or do not have it). Based on these, we want to research whether there are some differences between the two types of applicants. Firstly, we draw overall pair plots(**Fig 2**) which represent distributions of all pairs of factors with and without research experience. Then we pick out some interesting pairs we want to focus on. In section 3.2, we mention that the CGPA, TOEFL, and GRE are the three most relevant factors to the chance of admission. **Fig 3** separately represent their distributions with and without research experience.

3.3.1 TOEFL vs Chance of Admission

In **Fig 3a)**, we can find that the yellow points tend to distribute slightly above the blue points, which means that when applicants with the same TOEFL scores, universities tend to choose those with research experience.

3.3.2 GRE vs Chance of Admission

In **Fig 3b)**, The yellow points are more concentrated at the top-right while blue points are more concentrated at the bottom-left. This may be because the GRE and research experience can improve each other to a certain degree. In GRE, there are many vocabularies related to scientific work. These vocabularies may help a lot when applicants try to publish essays and applicants with research experience will have more opportunities to know these words which may help them to get higher scores on GRE more easily. What's more, when the GRE score is from 290 to 310, the difference in the chance of admission between applicants with and without research experience is very slight. Specifically, when the GRE score is lower than 300, applicants with research experience may even have a slightly lower chance of admission than applicants without it. This may certify that the research experience has a positive effect on the application only when applicants can get a relatively high score on GRE. And also it may show that research experience may not be able to make up for a relatively low GRE score. Applicants should pay more attention to their GRE examination. It is very important and correlated to their chance of admission while research experience is just the icing on the cake.

3.3.3 CGPA vs Chance of Admission

In **Fig 3c)**, same as **Fig 3b)**, the yellow points tend to distribute at the top-right while blue points tend to locate at the bottom-left. Applicants with the same CGPA and those who have research experience tend to have more chances to enter their dream schools. However, we can observe that the distribution of yellow points tends to be closer to its re-

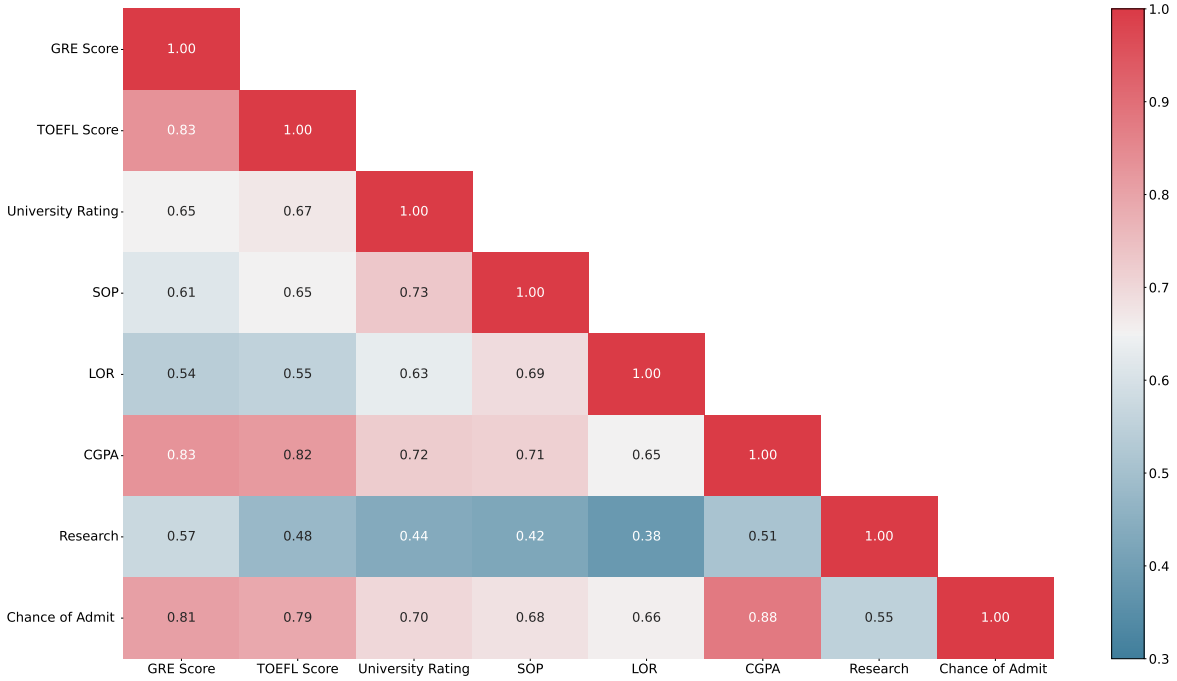

Figure 1: Correlation coefficients of each pair factors.

Figure 2: Pair plots of Graduate Admission 2.

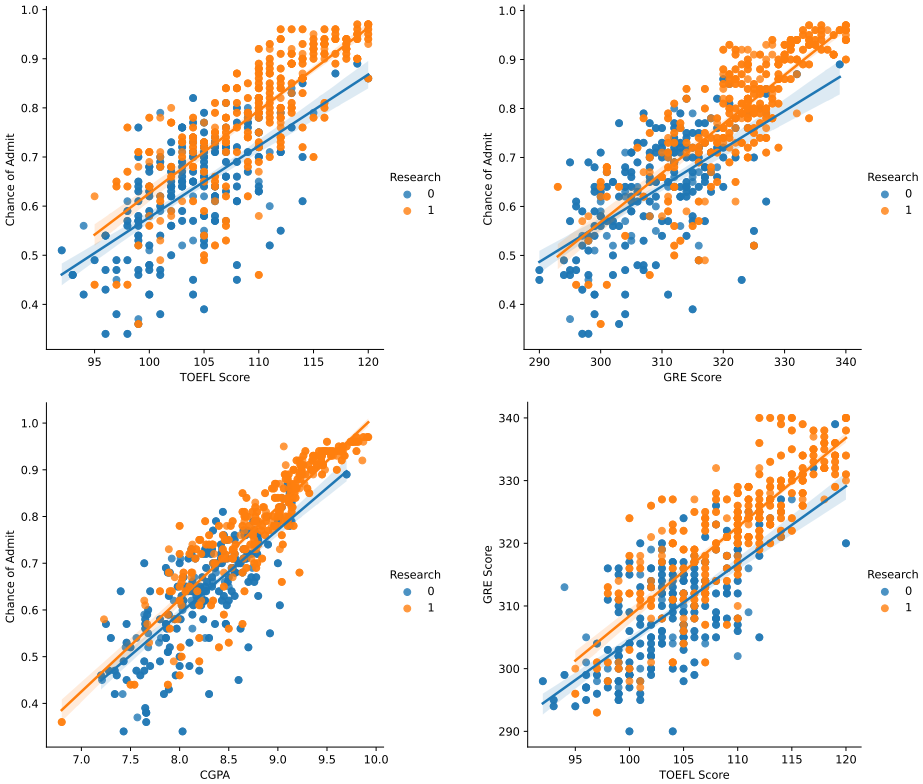

Figure 3: a) top-left: TOEFL VS Chance of admission **b)** top-right:GRE VS Chance of admission **c)**bottom-left: GPA VS Chance of admission **d)**bottom-right: TOEFL VS GRE

gression line than blue points. This may show that if someone has research experience and he or she wants to predict his or her application results by GPA, the predictions will be much more stable and believable. The research experience indeed can provide a positive effect on someone's application process.

3.3.4 TOEFL vs GRE

In **Fig 3d)**, we show the relationship between TOEFL and GRE. It can obviously find that applicants with research experience are easier to having relatively higher scores both on TOEFL and GRE. This may also certify that research work indeed can help applicants to their English proficiency. Because a lot of famous publications are written in English, applicants have to read, write, and understand in English. These abilities are also required by TOEFL and GRE.

3.4 Summary

In this section, we first briefly introduce the dataset and show some basic metrics. Then we conduct research on their correlations. Firstly, we utilize a heatmap(**Fig 2**) to show the correlation coefficients of each pair of factors. We find that the CGPA, TOEFL, and GRE are much more relative to the chance of admission. However, the research experience seems relatively not correlated. Furthermore, we conduct research to find the effect of research experience by separately drawing the distributions of factors with and without it. We find that the positive effect of research experience can be observed only when applicants get relatively high scores on academic examinations, especially on CGPA and GRE. The research experience is more like the icing on the cake, so applicants should stick to focusing on their grades because research experience can not compensate for their bad grades.

4 Methodology

From the statistical analysis of the data, we have seen the correlation between the chances of admission and different factors, so it is possible for us to use statistical learning methods that can capture the such relationship and

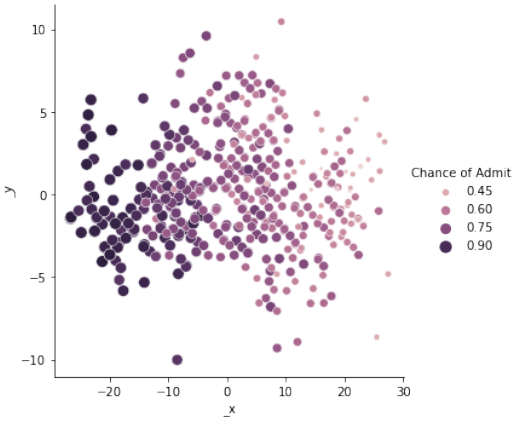

Figure 4: PCA of features

predict the chances of admission given unseen data. In this section, we explore various methods that can map the multivariate vector composed of the factors above into the distribution of admission chances and use such mapping to predict admission chances. We first select 80% samples from the dataset randomly as our train set and the rest as the test set. We also use the following evaluation metrics to evaluate the performance of our model on the test set:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - Y\right)^2} \tag{2}$$

In addition, we perform principal components analysis to visualize our features: **Fig 4)** show that the features with higher chances of admit distribute differently from the features with lower chances of admit, which indicates that we may find a way to correctly recognize such differences.

4.1 Model 1 - Linear Regression

Linear regression is a widely used statistical analysis method that uses regression analysis in mathematical statistics to determine the quantitative relationship of interdependence between two or more variables. In particular, given a data vector X, the model can be described as follow:

$$f(X) = w^T X + b \tag{3}$$

where w^T is a vector of coefficient and b is a bias parameter. In order to calculate these parameters, we first define our loss function as follow:

$$\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - y_i\right)^2 \tag{4}$$

the goal is to find w^T and b to minimize the function above and the least square method is a direct way to find a solution when the dataset is relatively small and the method can be described as follow:

$$W^* = \underset{W^*}{argmin}\left(y - XW^*\right)^T\left(y - XW^*\right)$$
$$W^* = \left(X^T X\right)^{-1} X^T y \tag{5}$$

where $W^* = (w, b)$. Using this model, we get the MSE of 0.0043 and R^2 of 0.7679 on the test set.

4.2 Model 2 - DecisionTree

The core concept of the Decision Tree is that similar inputs will produce similar outputs. First, select a feature from the training sample to partition the subsets, then select the next feature in each subset to continue to partition subsets according to the same rules, and repeat until all the features are used up. At this time, a tree with decision nodes and leaf nodes is obtained. For the sample to be predicted, according to the value of each feature, select the corresponding nodes and match them one by one until a leaf node. The average output of the samples from each leaf node provides output for the sample to be predicted. By minimizing the MSE between the output of the leaf node and the true value, we can find the partition boundary of each node as follow:

$$\underset{j,s}{min}\left[\underset{c_1}{min}\sum_{x_i \in R_1(j,s)}\left(y_i - c_1\right)^2 + \underset{c_2}{min}\sum_{x_i \in R_2(j,s)}\left(y_i - c_2\right)^2\right] \tag{6}$$

where j is the partition variable and s is the optimal partition point, while traversing j, the optimal partition point s that minimizes the Eq. (6) can be found. We use this model to get the MSE of 0.0018 and R^2 of 0.9095 on the test set.

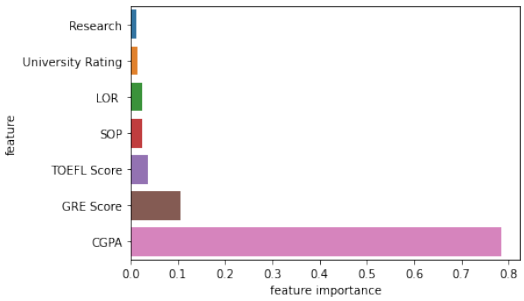

Figure 5: Importance of each feature

4.3 Model 3 - RamdomForest

Since the decision tree will be troubled by a large variance of data, various methods were proposed to improve the performance of decision tree, and one of the widely used models is RamdomForest(Breiman, 2001). A RamdomForest is a classifier containing multiple decision trees each decision tree can obtain a prediction result by randomly selecting samples and features, and the regression prediction result of the whole forest can be obtained by averaging the results of all trees. The model can also give estimates of what variables are important in the regression. We use this model to get the MSE of 0.000693 and R^2 of 0.9628 on the test set. We also use the model to generate the importance of each feature. From **Fig 5)**, we can observe that the CGPA is the most significant factor that affects the chances of being admitted, which is also consistent with our common prior knowledge that CGPA is relatively important.

4.4 Model 4 - Lasso Regression

Lasso regression is a kind of linear regression model that has the ability of selecting features and adjusting complexity. The model is the linear regression added with L1 regularization and can be described as follow:

$$\beta = \underset{\beta}{argmin} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \tag{7}$$

$$\|\beta\|_1 = \sum_{i=1}^{N} |\beta_i| \tag{8}$$

Therefore, when utilizing the least square method, the regularization term will penalize all the parameters and might be down to 0 when λ is large, which can be used as a way of feature selection. We use this model to get the MSE of 0.00447 and R^2 of0.8008 on the test set.

4.5 Model 5 - Ridge Regression

Ridge regression is another kind of linear regression model similar to Lasso regression using L2 regularization and have the following form.

$$\beta = \underset{\beta}{argmin} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2 \right\} \tag{9}$$

$$\|\beta\|_2 = \sum_{i=1}^{N} |\beta|^2 \tag{10}$$

Therefore, the L2 regularization will scale down the result of least square methods in proportion and the parameters will close to zero if the λ is large. We use this model to get the MSE of 0.00348 and R^2 of 0.8449 on the test set.

4.6 Model 6 - Support Vector Regression

The goal of SVR(Support Vector Regression(Ma et al., 2003)) is to minimize the distance between the hyperplane and the farthest sample point, so that data can be fitted by using the hyperplane. In SVR model, a band will be created on both sides of the linear function. For all samples falling into the band, no loss is calculated. Only those outside the band are included in the loss function. Then the model is optimized by minimizing the width of the band and the total loss. Therefore, the goal of the model can be described as follow:

$$\underset{W,b}{min} \frac{1}{2} \|W\|^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) \tag{11}$$

$$s.t. \begin{cases} y_i - WX - b <= \epsilon + \xi_i \\ WX + b - y_i <= \epsilon + \xi_i^* \\ \xi_i, \xi_i^* >= 0 \end{cases} \tag{12}$$

where ξ_i, ξ_i^* are slack variables that measures the distance between the marginal points and the hyperplane. We use this model to get the

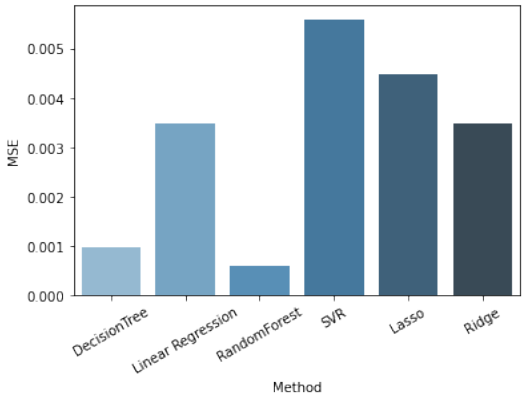

Figure 6: MSE Results of Different Models on Test Set

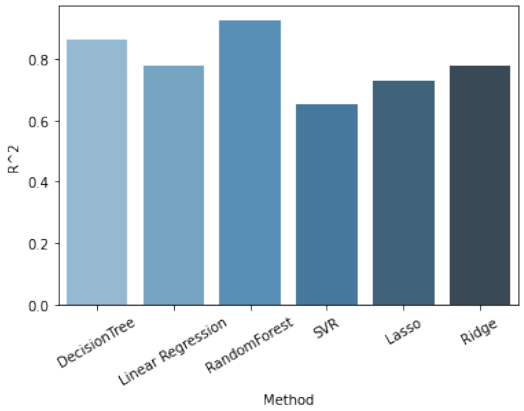

Figure 7: R^2 Results of Different Models on Test Set

MSE of 0.00559 and R^2 of 0.7508 on the test set.

5 Results and Analysis

In the Data Analysis section, we research the correlations between the factors in our dataset. The CGPA, TOEFL, and GRE are strong relative to the chance of admission while research experience is weakly relative. In the experiment of methods in predicting the chances of admission, the results in **Fig 6)**, **Fig 7)** show that RamdomForest generates the best result while the SVR generates the worst. It's reasonable because RamdomForest uses ensemble learning from multiple decision trees, which is compatible with our low-dimension data. In addition, the linear regression methods can't capture nonlinearity in the data and will result in overfitting when fitting the features that are not relative to the chances of admission. In conclusion, applicants should focus on their grades cause research experience is just compensation when they have good academic grades.

References

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

Ma, Junshui, James Theiler, and Simon Perkins. 2003. Accurate on-line support vector regression. *Neural computation*, 15(11):2683–2703.

Mohan S Acharya, Asfia Armaan, Aneeta S Antony. 2019. A comparison of regression models for prediction of graduate admissions. IEEE International Conference on Computational Intelligence in Data Science 2019.

UCSDGraduateDivision. 2021. Total applications, admits, new with percent of international, urm, and women - 10 year trend. `https://grad.ucsd.edu/about/grad-data/admissions.html`.