CS542 Spring 2019 Final Project: Accepted or Not

Dayuan Wang
Boston University
wangdayu@bu.edu

Yunyu Zhang
Boston University
yzhang11@bu.edu

Shizhan Qi
Boston University
shizhan0@bu.edu

Abstract

There are more and more students decide to apply for graduate school, but how much do you know about the graduate admission? In our project, we used Linear Regression, Neural Network, Decision Tree and Random Forest methods to create a machine learning model that can predict your probability to get accepted by certain graduate school. In our experiment, we find out that Random Forest can give us the best model since it has the lowest mean square error in all the method that we used. In our dataset, we find out the GPA and University rating are the most significant predictors.

1 Introduction

The Github repo of the project source code:

`https://github.com/dwang1995/admission`

1.1 Project Introduction

Based on internet, of American adults, 9.3 percent of adults over 25 years old have a master's degree and 2 percent of Americans have a doctoral degree. Nowadays, there are more and more international students come to US for the graduate education [1]. Which means understanding the graduate admission is important for us. In this project, we are going to use the applier's information to create the machine learning model that can predict the applier's probability to get admitted into certain graduate school by knowing their key information, which could be really helpful for some of us since the application fee is expensive. At the same time, we want to learn what are the key aspects that the graduate admission really cares about. By understanding them could be more important and useful than constructing machine learning model.

1.2 Dataset Introduction

Our dataset is from Kaggle. There are 8 different features in our dataset. There is a unique ID for each of the data record. After the record ID, it is the column for the GRE score which is in the range from 0 to 340. The TOEFL score is the next and it is in the range from 0 to 120. Next column is the University Rating, which stands for the rating for your dream school and it is out of 5. Where 1 represents the top 10 universities, 2 represents the school's ranking in 10 to 20 and so on. 5 represents the school's ranking greater than 50. SOP is the next column and it stands for statement of purpose, it is the column in range 1 to 5, where 1 represent a weak statement of purpose essay and 5 represents a very strong statement of purpose. The column of LOR stands for letter of recommendation letter and it is using the same scale with statement of purpose, where 1 represents a weak recommendation letter and 5 represents a strong recommendation letter. Next column is CGPA which stands for undergraduate GPA or college GPA; it is in scale out of 10. The column of Research is next, which stands for the applier's undergraduate research experience. 0 represents no experience on research and 1 represent some experience on academic research. The last column is the change of admit; it is

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

the number that express applier's probability to get into certain university which is evaluated by the professionals.

2 Method and Implementation

In this project, we use linear regression, decision tree, random forest and neural network to make the models. Sklearn is used for linear regression, decision tree and random forest. The neural network is built by using keras library. The three layers Neural Network is initialized with weights' mean of 0.0 and the standard deviation of 0.05. Rectified linear unit (ReLU) is used as activation function in the input layer and hidden layer and sigmod is used in the output layer. To train and optimize the model, we use the Adam algorithm.

Because the all the variables in the dataset have different scales, for example, GPA is out of 10 and GRE is out of 340, we normalize our dataset first before the training. Then we use 2-fold validation methods to divide the dataset and we also divided the data into training and test set. After training the models with training set, we make the prediction with test sets and calculate the mean square error of each models.

Because our goal is to combine all four models as a one big prediction model, we compare each model with its mean square error, and assign weight to its prediction. Model with less mean square error will be assign with a larger weight. We use a normalization function to normalize the mean square error and then calculate the weights for each model. The final prediction is made by the following equation:

```
Prediction = w1*Linear Regression Prediction + w2*Decision Tree Prediction
             + w3*Random Forest Prediction + w4*Neural Network Prediction
```

Notice w1, w2, w3 and w4 are all normalized.

After building the model, we make a Python web application for users to see their chance to be accepted by a graduate school. A user can input his/her information, such as (GPA or GRE), then the web application will return a acceptance percentage to the user.

3 Results

3.1 Data visualization

For the data visualization part, we mainly used Python pandas dataframe, since pandas is a very useful package in data cleaning, modeling and exploration. It provides high performance and easy to use data structures in data analysis and data visualization and that's why we choose pandas. After reading the data, we made a histogram to have a basic idea about how the data looks like and the result plot is shown in Figure 1.

From the graph, we can see that the dataset contains 3 continuous variables, which are GPA, GRE score and TOEFL score, and 4 categorical variables, which are LOR, Research, SOP and university ranking. The outcome, which is the chance of admission, is also a continuous variable. We could notice that the variable Serial No. only represents the case number of each student which does not contain useful information for the prediction model. Based on that, we should omit this variable in the further analysis. This will also be proved in the correlation analysis in the next steps. After having a basic idea about how the data looks like, we calculate correlation matrix and the result is shown in Figure 2.

From the correlation matrix, we can see that the serial number is barely correlated with the chance of admission or other features. This means that we can delete this variable in the future analysis because it does not provide useful information.

After that, we generate a custom diverging colormap shown in Figure 3. From the graph, we can see that the features that are most positively correlated with the chance of admission are GPA, GRE score and TOEFL score. The features that are least positively correlated with the chance of admission are Research, LOR and SOP.

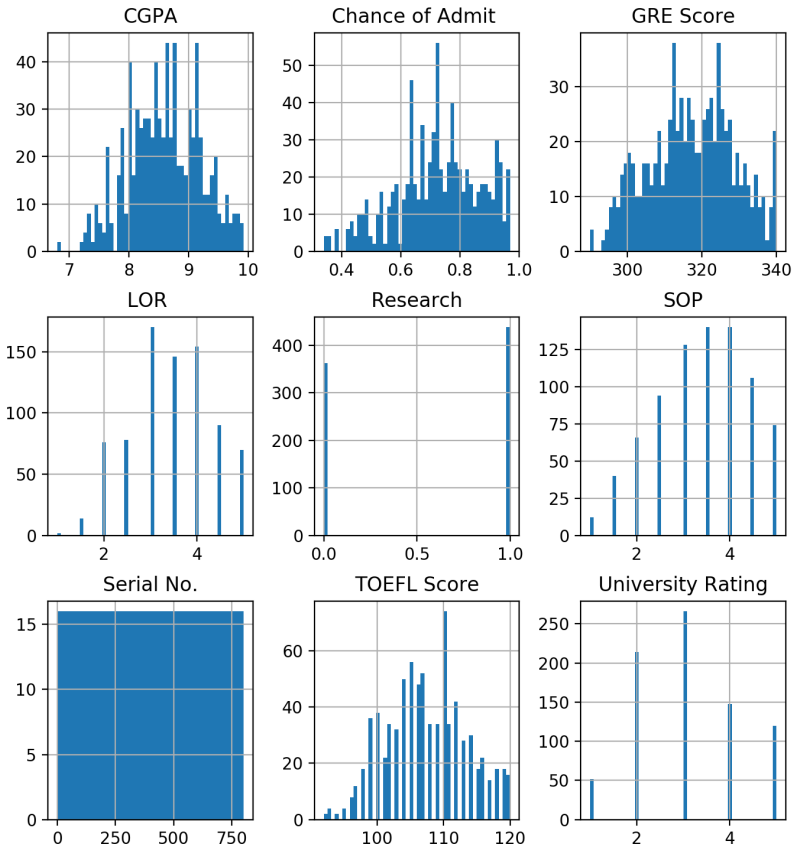


Figure 1: Dataset Histogram.

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
Serial No.	1.000000	0.002070	0.038419	0.017493	0.050507	0.044161	0.000863	-0.006010	0.006606
GRE Score	0.002070	1.000000	0.835977	0.668976	0.612831	0.557555	0.833060	0.580391	0.802610
TOEFL Score	0.038419	0.835977	1.000000	0.695590	0.657981	0.567721	0.828417	0.489858	0.791594
University Rating	0.017493	0.668976	0.695590	1.000000	0.734523	0.660123	0.746479	0.447783	0.711250
SOP	0.050507	0.612831	0.657981	0.734523	1.000000	0.729593	0.718144	0.444029	0.675732
LOR	0.044161	0.557555	0.567721	0.660123	0.729593	1.000000	0.670211	0.396859	0.669889
CGPA	0.000863	0.833060	0.828417	0.746479	0.718144	0.670211	1.000000	0.521654	0.873289
Research	-0.006010	0.580391	0.489858	0.447783	0.444029	0.396859	0.521654	1.000000	0.553202
Chance of Admit	0.006606	0.802610	0.791594	0.711250	0.675732	0.669889	0.873289	0.553202	1.000000

Figure 2: Correlation Matrix.

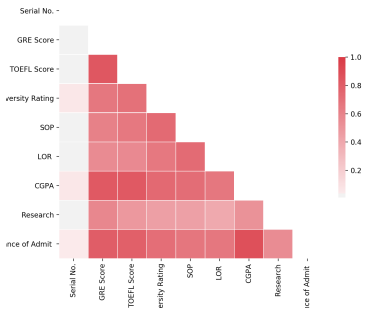
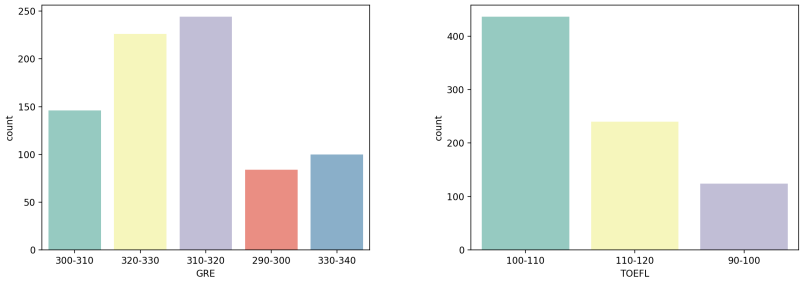


Figure 3: Diverging Colormap.



(a) GRE. (b) TOFEL.

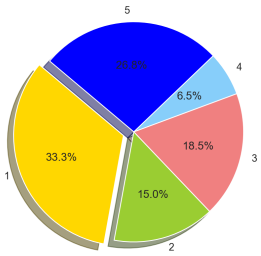
Figure 4: Frequency of test scores.

We also draw a bar plot about GRE and TOEFL score to see the frequency of test scores among the candidates in Figure 4. From the graph, we can see that the most frequent GRE score is in the range from 310-320. And the most frequent TOEFL score is in the range from 100-110.

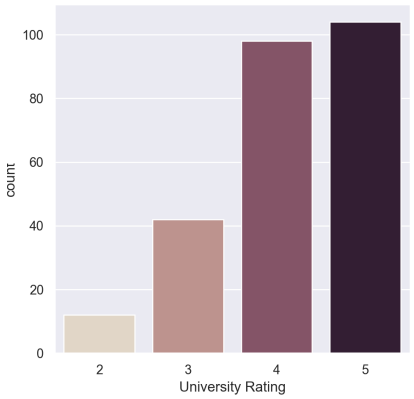
We want to see a comparison about GPA and GRE, so we draw a scatterplot by using seaborn package to get the comparison in Figure 5. The graph shows a positive linearity relationship between GPA and GRE, better GPA is correlated with better GRE score. Also, candidates with higher university ranking tend to have higher GPA and higher GRE score.

The pie plot in Figure 6(a) below shows the distribution of each students' university ranking. We could see a interesting points that the university ranked as 1 and university ranked as 5 does not have a big difference. From the bar plot in Figure 6(b) we can see among the candidates with high admission change, the most of them are coming from higher-ranked colleges. (Nearly a quarter of the data).

Figure 5: Relationship Between GPA and GRE.

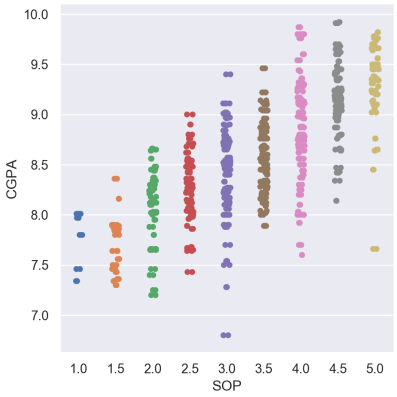


(a) University Ranking.

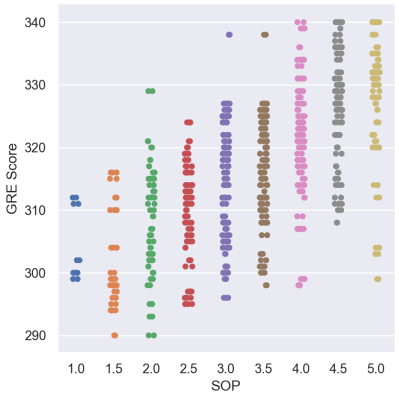


(b) University Rating.

Figure 6: Ranking and Rating.



(a) GPA and SOP.



(b) GRE and SOP.

Figure 7: GPA, GRE and SOP.

We made a comparison about GPA vs SOP as well as GRE vs SOP and the result plots can be seen in Figure 7. From the plot we could see better SOP associated with better GPA as well as GRE.

In university admissions, letter of recommendation also counts a great part to influence whether a student can be admitted or not, so we visualize it to see the trend and the plot is shown below in Figure 8. From the plot we found stronger letter of recommendation associated with higher chance of admission.

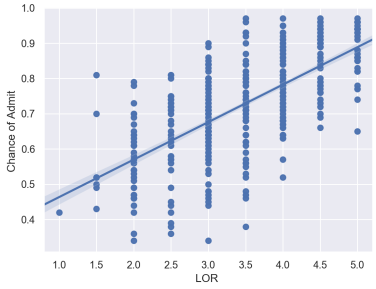


Figure 8: Recommendation Letter.

Figure 9: GPA and GRE.

Lastly, we draw a scatterplot to compare GPA and GRE in Figure 9. The graph shows a positive linearity relationship between GPA and GRE. Better GPA is correlated with better GRE score also candidates with stronger recommendation letter tend to have higher GPA and higher GRE score.

3.1.1 Mean Square Error and Assigned Weights of Each Model

Table 1: MeanSquare Error and Assigned Weights

Model	Mean Square Error	Assigned Weights
Random Forest	0.0017317715500000030	0.3456395296956547
Decision Tree	0.0019930000000000004	0.3274059585863088
Linear Regression	0.0043103926424645940	0.1656535019352298
NN	0.0043728402722308710	0.1613010097828067

This chart is sorted by the mean square error. From the table, we can see that Random Forest has the least mean square error.

4 Discussion

4.1 Model Comparison

In this project, we used linear regression, decision tree, random forest and neural network. Below is the comparison about these four models.

Linear regression attempts to find relationship between exploratory variable and dependent variable. If the model has multiple exploratory variables, it's called multiple linear regression. Linear regression is a useful and easy-to-use model. Since we have a small and clear dataset, the first model we have thought of was linear regression.

Decision tree is a model that straightforward in visualizations. There are many advantages of using decision tree. The internal workings are capable of being observed and thus we could use decision tree to reproduce work. We can use decision tree to analyze both numerical and categorical data. If the dataset is large, decision tree algorithm can also perform well. The disadvantage includes choose of each node (it can only choose the best result in each step rather than the whole step, which easily leads to local optimum rather than global optimum) and overfitting (we could use prune to somewhat get rid of it).

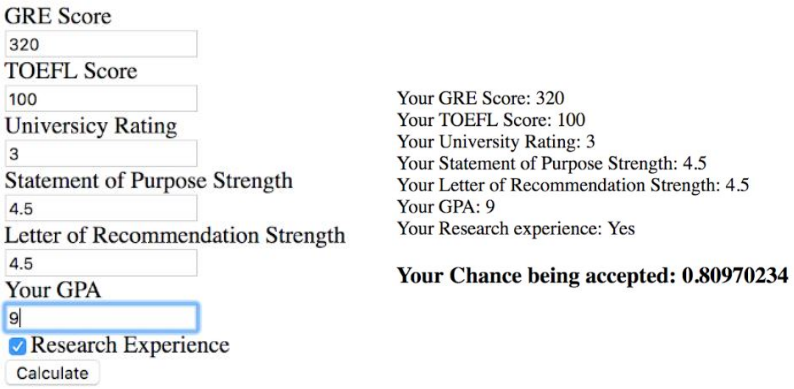
Random forest can decrease both error due to bias and error due to variance. It is simply a collection of decision trees and aggregated into one result. This can mitigate overfitting problem and won't substantially increase error due to bias. Thus, it is a better algorithm in avoid overfitting comparing to decision tree.

Based on this, if you want to make the model simple and easy to explain, and you also don't care a lot about the multi-collinearity and overfitting problem it's better to use decision tree than random forest.

Deep learning has been more and more popular nowadays. It is a subset of machine learning that achieves great power to represent world as nested hierarchy of concepts. In this project, we built three-layer neural network to train the model (input layer, hidden layer and output layer). Deep learning works extremely well in large dataset. Since our dataset is small, it might not be a good idea to use very complex deep learning model to train, that's why we simply use one hidden layer in our model building part. Basically, if the dataset is simple, and we build very complex neural network it can hardly receive well-performed result. If the dataset is simple and we use simple neural network to train, the result can be similar with machine learning models. That's why in our project, we would prefer using machine learning algorithms rather than deep learning.

4.2 Web application and Our Suggestions

This is the interface of our web application.



To see the which variables are the most significant predictors for the final prediction, we use linear regression summary as an example.

	coef	std err	t	P>|t|	[0.025	0.975]
GRE	0.0096	0.067	0.144	0.886	-0.121	0.141
TOEFL	0.1655	0.068	2.450	0.015	0.033	0.298
Rate	-0.1510	0.041	-3.718	0.000	-0.231	-0.071
Statement	0.1003	0.048	2.074	0.039	0.005	0.195
Recommendation	0.1104	0.042	2.633	0.009	0.028	0.193
GPA	0.9602	0.076	12.651	0.000	0.811	1.110
Research Exp	0.0006	0.017	0.038	0.970	-0.033	0.034

Omnibus:	5.918	Durbin-Watson:	2.065
Prob(Omnibus):	0.052	Jarque-Bera (JB):	3.919
Skew:	0.101	Prob(JB):	0.141
Kurtosis:	2.496	Cond. No.	19.4

We can see that both GPA and university rating have p-values of 0. So, GPA and university rating are the two most significant predictors. We can also see that GPA has a coefficient of 0.9602. This means that GPA have a huge impact on the result. The third most significant predictor is strength of recommendation letter with a p-value of 0.09. Thus, our suggestions for people who want to apply for graduate school are the following:

- Study hard and get a good GPA
- Apply for a reasonable school. Don't overestimate yourself
- Ask professors for a good recommendation letter.

5 Conclusions

In general, we train four different models with different methods. Although each method has different performances, we want to take advantages of each method. Therefore, we combine all methods and make a one prediction model. We find out that Random Forest can give us the best model since it has the lowest mean square error in all of the method that we used. In our dataset, we find out the GPA and University rating are the most significant predictors.

There are a few things we can do in the future. We could try out some more machine learning methods. By doing that it could give us some better machine learning model. On the other side, when we understand what is going on in graduate admission, maybe we can use our model to come up with the system to suggest some school that you get a better chance to get accepted. Currently, we also have some suggestion on the dataset. In our dataset, the research experience is just a boolean variable, but based on professor's suggestion, the research experience is never a boolean variable. If it became the number of researches in the past, it could become a significant predictor. These are the future works that we could do for our project.

References

[1] Baum, Sandy. & Steele, Patricia. (2017) Who Goes to Graduate School and Who Succeeds? *AccessLex Institute Research Paper* **17**-01.

[2] Acharya, M.S., Armaan, A. & Antony, A.S. (2019) A Comparison of Regression Models for Prediction of Graduate Admissions *International Conference on Computational Intelligence in Data Science* IEEE, 2019.