Final Report

The Probability Of Entering Your Dream Graduation Program

Xinyang Xu

DATA1030 Fall22 S01 Hands-on Data Science

Instructor: Andras Zsom

December 9, 2022

Github: https://github.com/seanxxy0528/Data-project.git

I. Introduction

Whether one can enter a graduate program is a mystery; Sometimes, your classmates who seem not as good as you may be admitted to a program you dream while sometimes you can be admitted to programs that you think you must be denied. In this case, I want to have a logical prediction of whether one can enter their dream program, what is the probability of entering it, and, more importantly, what one can do to improve their probability of being admitted. It is more reliable than some admission consulting companies that give decisions just based on their experience. My target variable is the Chance of Admitted, a regression variable because it indicates probability. After dropping the "Serial No." which is not meaningful, there are 3200 data points, containing 400 rows with 8 columns, with 7 features and 1 target variable.

In all features, GRE Scores, TOEFL Scores, and Undergraduate GPA are continuous variables because they are random numbers in a range, having clear minimum and maximum; University Rating (1-5), Statement of Purpose (1.0 – 5.0), and Letter of Recommendation Strength (1.0 – 5.0) are Ordinary variables because they are all in ranked, while having different levels; Research Experience is categorized variable because the result is 0 or 1 indicating yes or no. This data set is well documented, I found this in Kaggle, from the UCLA database. Most of the people use it as basic EDA, and some people use it to tryout different machine learning algorithm, only a few of them tunes value for that. I found that most of the people have best scores for linear regression, and also some for random forest.

II. Exploratory Data Analysis

In the EDA session I first visualized the target variables. By looking at the graph below, I found that there are not many outliers in Chance of Admit. The mean of it is 0.724 (72.4%), higher than I expected. The standard deviation is 0.1426, and all data points are in the range of 0.34 – 0.97.

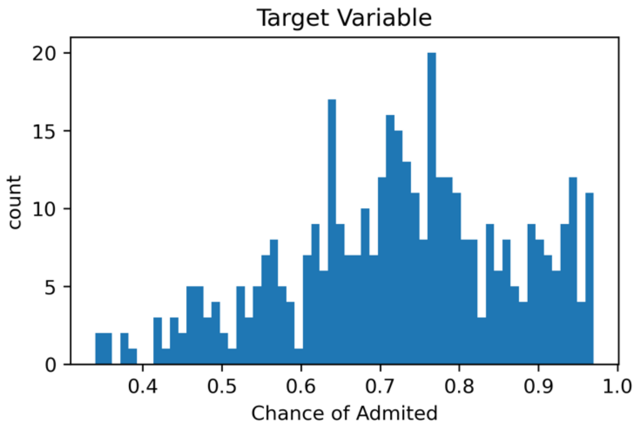
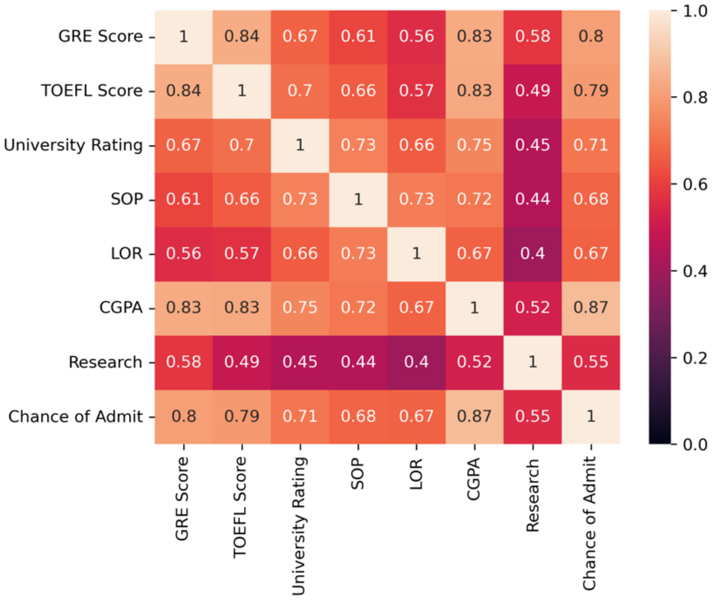


Figure 1. The layout of my target variables, chance of admitted

Then, I did a heatmap to show a broad correlation as shown below. Closer to 1, the more relative they are. Not surprisingly, GRE, Tofel, and GPA are three figures that have the highest correlation to the target variable, around 8, and the lowest score is research, 0.55. Hard skills are correlated with each other which make sense because people good at study will be good at GRE, TOFEL and GPA. But what surprised me was the correlation of university ranking with others. We normally think that the higher the school rank is, the higher GRE they should be, so these two variables should have a strong correlation. In the contrast, Statement of Purposes should focus more on individuals rather than school ranking. But the result of correlation is the opposite. The school ranking is more correlated to SOP rather than GRE.



	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
GRE Score	1	0.84	0.67	0.61	0.56	0.83	0.58	0.8
TOEFL Score	0.84	1	0.7	0.66	0.57	0.83	0.49	0.79
University Rating	0.67	0.7	1	0.73	0.66	0.75	0.45	0.71
SOP	0.61	0.66	0.73	1	0.73	0.72	0.44	0.68
LOR	0.56	0.57	0.66	0.73	1	0.67	0.4	0.67
CGPA	0.83	0.83	0.75	0.72	0.67	1	0.52	0.87
Research	0.58	0.49	0.45	0.44	0.4	0.52	1	0.55
Chance of Admit	0.8	0.79	0.71	0.68	0.67	0.87	0.55	1

Figure 2. The Heatmap of features with target variables

After finding GPA and GRE are two of the most correlated variables to my target, I did a scalar plot of both to the Chance of Admit because they are continuous variables. I found that they are both kindly in linear correction to the Chance of admission. The slope of CGPA is nearly 45 degrees, while the GRE one has a shallower slope here. That means an increase of the same percentage of CGPA will affect the target more than GRE did.

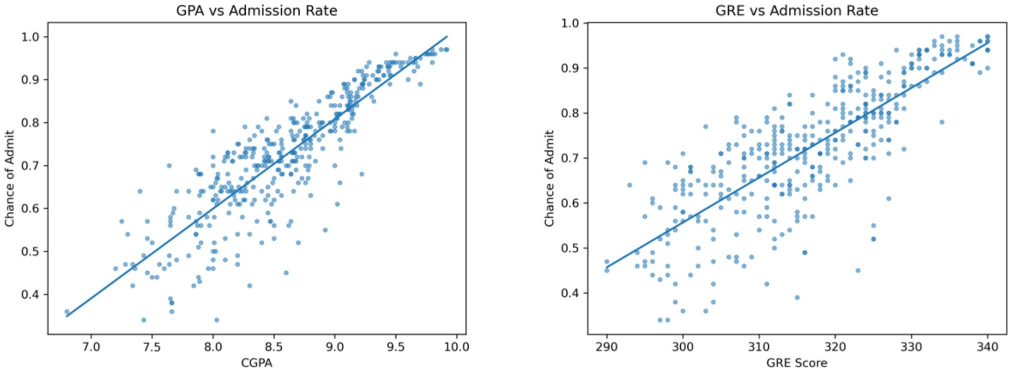


Figure 3&4. The scalar plot of CGPA and GRE against target variables

For ordinary variables, I use a boxplot graph because a bar chart is not good with many categories. For a LOR, I found that increasing LOR will effetely increase the mean of Chance of Admit. However, there is a huge increase in the mean when we increase LOR from 4.0 to 4.5, while there isn't much change when we increase LOR from 4.5 to 5.0. That means that after a certain level of strength of the recommendation letter, it will be less effective. Also, If we want a chance of admission over 90%, we need a LOR strength of more than 3. For SOP, it is quite similar to LOR. But the biggest difference there is that it has more outliers. Compared to LOR, which nearly has no outlier, even if you have a maximum of 5.0 in SOP, there is one data point that has a chance of being admitted lower than 40%. It also happened when SOP equals 2.0, 3.0, and 4.0.

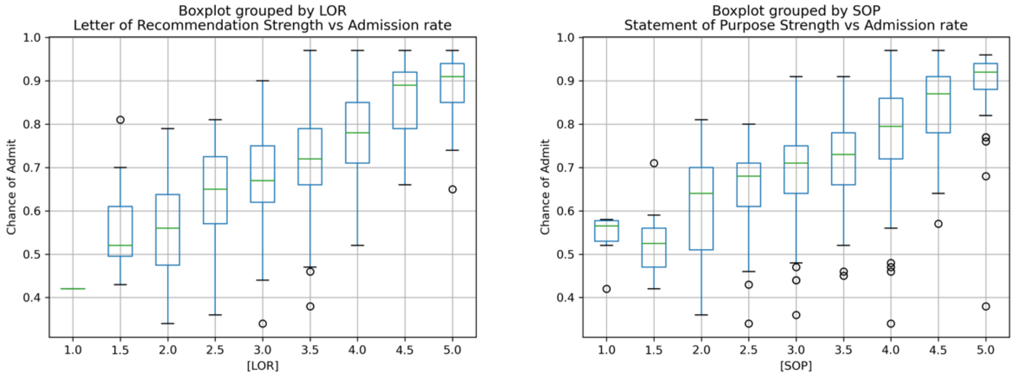


Figure 4&5. The boxplot of LOR and SOP against target variables

For categorized variables, the change in admitted rate based on whether or not we have research, I found two facts. One is that most people who do not have research have less than a 75% chance of admission, and the second is that if you want a chance of admission over 90%, you need to have research.

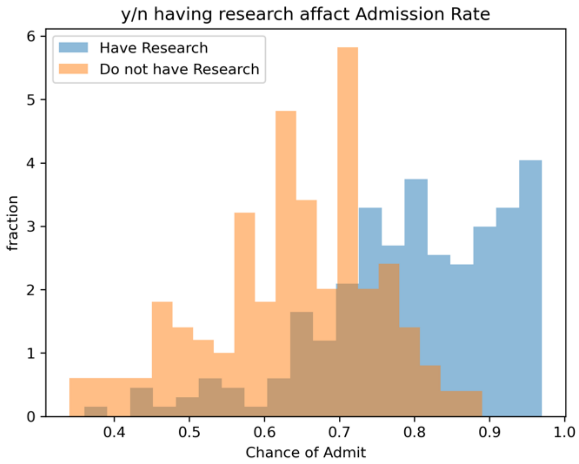
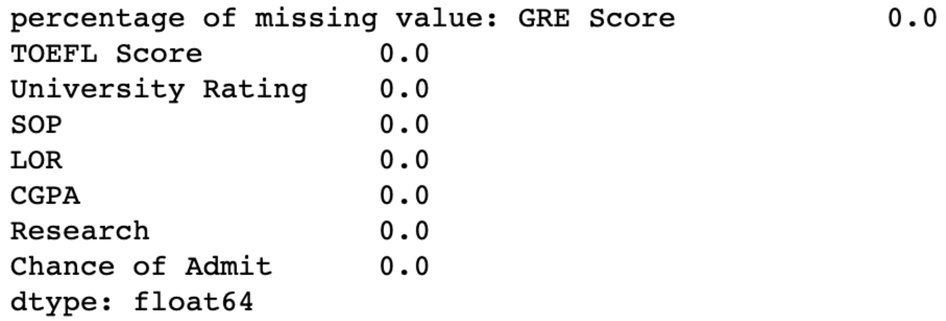


Figure 6. Whether or not having research with target variables

III. Methods

i. Splitting and Preprocessing (Pipeline begins)

I initially want to use stratified splitting based on university rating, as I think university rating is kind of imbalanced. However, because it is a feature instead of target variable, I listened to the comments of just use simple splitting there. I split the data set to train set, validation set and test set at a ratio of 60%, 20%, 20% because it is not a big data set. My data set is IID because there are no group features inside, and it is not time-series data. Because it is a well-organized and designed dataset, there are no missing values inside.



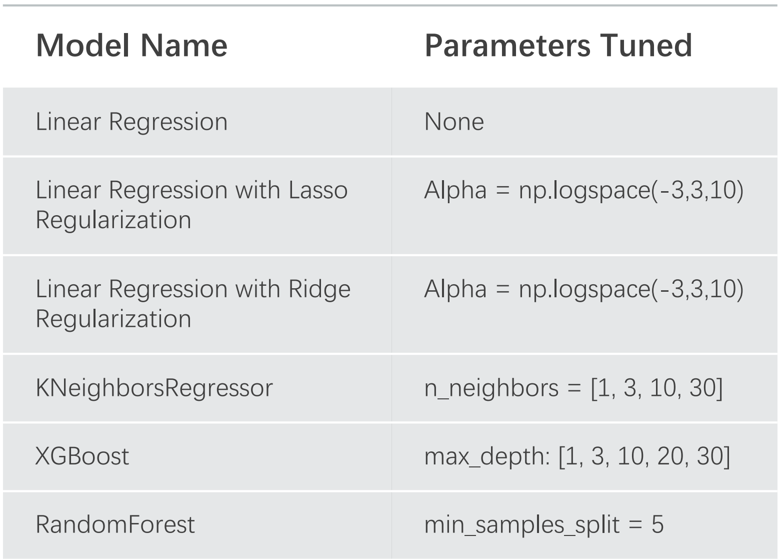
```
percentage of missing value: GRE Score            0.0
TOEFL Score          0.0
University Rating    0.0
SOP                  0.0
LOR                  0.0
CGPA                 0.0
Research             0.0
Chance of Admit      0.0
dtype: float64
```

Figure 7. The fraction of missing values

I loop through 5 different random state with the number of 42 * (1-5) to know the uncertainty of splitting. In the preprocessing process, I initially just use minmax scaler to TOFEL, GRE and CGPA, and use one-hot encoder to other features. But after my presentation, I found that I should use ordinary encoder to SOP, LOR and University rating as they are ordinary variables which clearly have a rank and use one-hot encoder only to research. In this case, I add the ordinary encoder and use three different preprocessors here.

ii. Models and Metrics (Pipeline continues)

After preprocessing, the algorisms are introduced to the pipeline now. I use six different machine learning algorithms including three linear regression models and three non-linear regression models. The models are shown in the lists: Linear Regression, Linear Regression with Lasso Regularization, Linear Regression with Ridge Regularization, KNeighborsRegressor, XG Boost, and Random Forest. It is a regression problem so I should choose between the RMSE score or the R2 score, and I choose RMSE because my data set didn’t include many features, and it is accurate. As I have five different random state, I will first find the best score for each models under each random state, then I will find the mean of the best scores, as well as the standard deviation of the scores. In this case, I can compare them and find the best model here. I tune the model parameters in the following chart.



Model Name	Parameters Tuned
Linear Regression	None
Linear Regression with Lasso Regularization	Alpha = np.logspace(-3,3,10)
Linear Regression with Ridge Regularization	Alpha = np.logspace(-3,3,10)
KNeighborsRegressor	n_neighbors = [1, 3, 10, 30]
XGBoost	max_depth: [1, 3, 10, 20, 30]
RandomForest	min_samples_split = 5

Figure 8. The tuned parameters for different model

For linear regression, we can not tune parameters here. For lasso and ridge parameters, I decide to make alpha the same, in order to make fair comparison. I take ten values from -3 to 3 because I think it is enough numbers here. For three other models, I have tried multiple numbers and find the parameters which have relatively high RMSE scores.

IV. Results

i. Best Model selection

Below are graphs and charts about the mean of the best RMSE scores, and the standard deviation of each models in different random state.

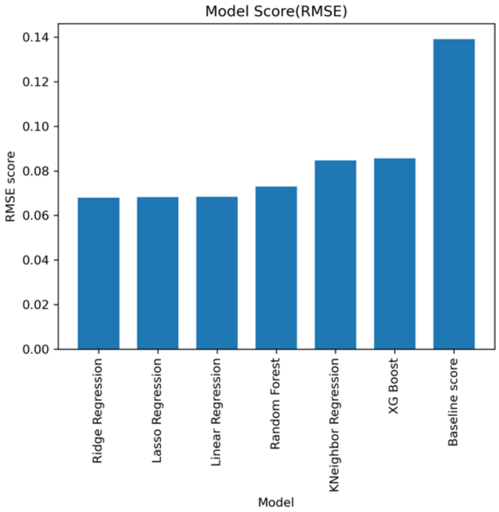


Figure 8. The visualized RMSE score for each model

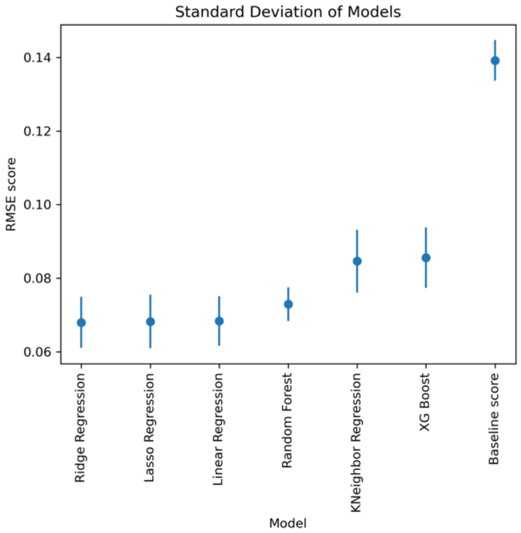
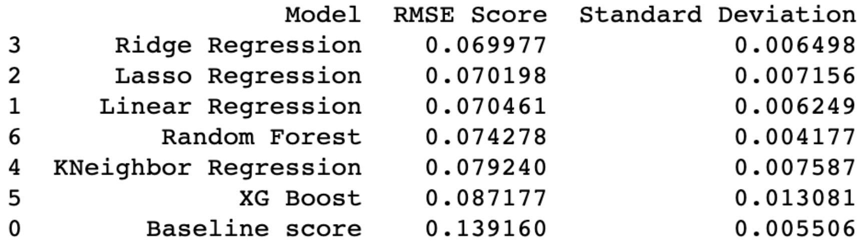


Figure 9. Visualized standard deviation for each model



```
                  Model  RMSE Score  Standard Deviation
3      Ridge Regression    0.069977            0.006498
2      Lasso Regression    0.070198            0.007156
1     Linear Regression    0.070461            0.006249
6         Random Forest    0.074278            0.004177
4  KNeighbor Regression    0.079240            0.007587
5              XG Boost    0.087177            0.013081
0        Baseline score    0.139160            0.005506
```

Figure 10. The Exact RMSE score and std for each model

By looking at the visualized RMSE score there, we can clearly see that all models perform much better than the baseline model. The RMSE for the baseline model is 0.139 which is much higher than the least performed model (XG Boost)'s 0.871. As a result, three linear models nearly perform as good as each other. All of the models expect XG boost has a good perform on standard deviation which means the uncertainty. As **Linear regression with Ridge regularization** has the best RMSE score with comparative good standard deviation, I will choose it as my best model there.

ii. Feature importance

After knowing that Linear regression with Ridge regularization is the best model here, I then use three different ways to show the importance of each features.

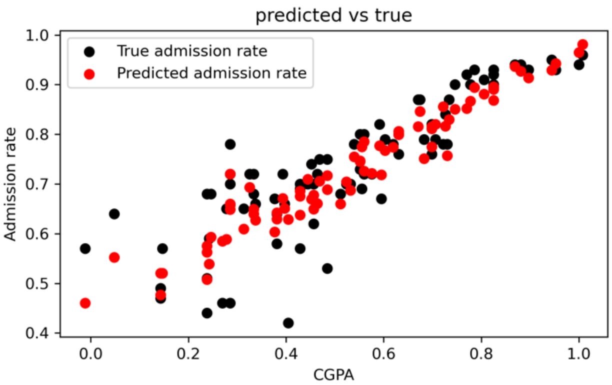


Figure 11. The model inspection of CGPA against Admission rate

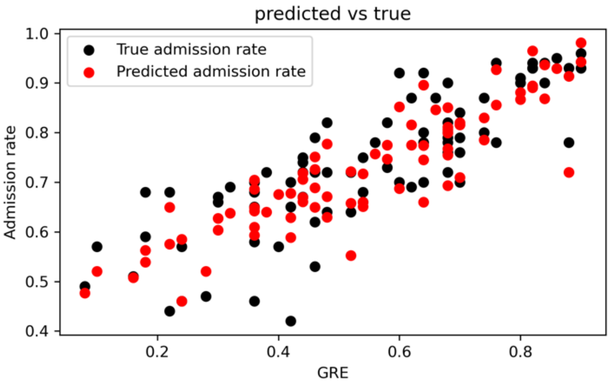


Figure 12. The model inspection of GRE against Admission rate

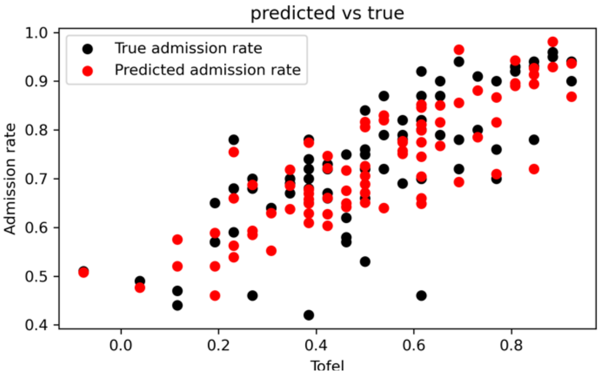


Figure 13. The model inspection of TOFEL against Admission rate

For model inspection, I use three continuous features, which is GPA, GRE, and TOFEL scores, as the three of the factors to predict the admission rate. By looking at the graph, we can clearly see that CGPA here has the best predicted points compared to GRE and TOFEL. The predicted red points are generally in a linear line, and fits the black point (true rate). It isn't works so well with TOFEL, as the points here are little bit distant than the other two.

Figure 13. The model inspection of TOFEL against Admission rate

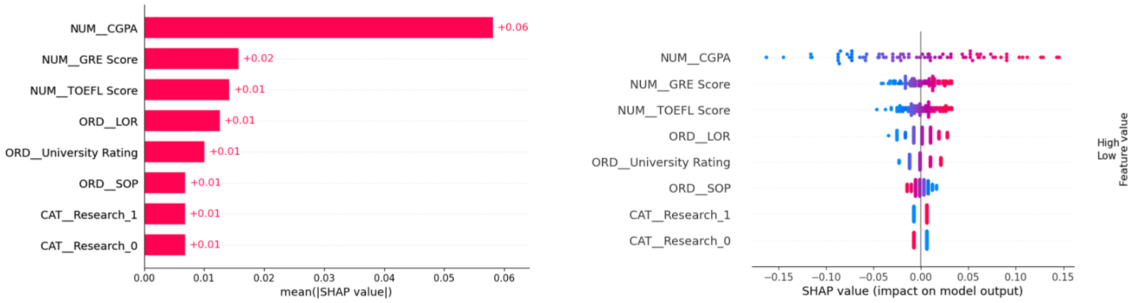


Figure 13&14. The SHAP value of the features

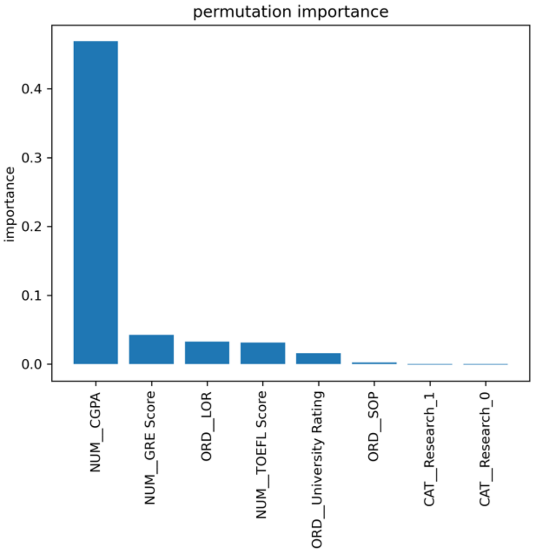
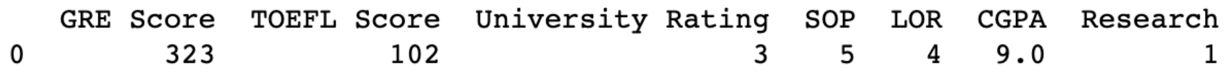


Figure 13&14. The permutation importance

There is not much to say about the feature importance. CGPA is definitely the most important features for our model both for the SHAP values and for the permutation importance. It is not so surprising. But what surprise me is that I think school ranking should be an important factor when we consider whether one can be admitted. But due to the data result, it didn't. Another interesting point here is that the LOR, the strength of the letter of recommendation pays also an roles on the admission rate. It is one of the top five features, even attribute more than the university rating. That also surprised me a lot.

iii. Predicted result and Conclusion

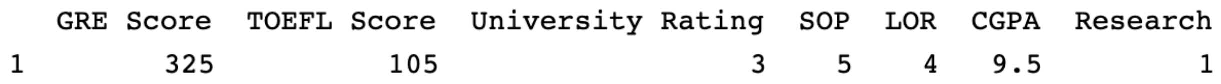
I think this project is meaningful because now I know the factors which will affect the admission rate. In this case, I am like a consultant, which can tell student what to do next in the future study in order to achieve their admission goals. In the final part, I create a new student profile listed below



```
   GRE Score  TOEFL Score  University Rating  SOP  LOR  CGPA  Research
0        323          102                  3    5    4   9.0         1
```

Figure 15. The current student profile

While the university ranking can’t be changed by yourself, and I think it is hard to increase LOR or whether you can have research, so I remain these data. So I want to try to increase some scores there to see what will be the predicted admission rate increase.



```
   GRE Score  TOEFL Score  University Rating  SOP  LOR  CGPA  Research
1        325          105                  3    5    4   9.5         1
```

Figure 16 After improvement

By improve 2 points in GRE, 3 points in TOFEL, and 0.5 in GPA, finally the student can have an increase of amount 8% of chance of admitted, from 78.14% to 85.35%.

V. Outlook

In the last part of report, I will talk about how I can improve in my project. Firstly, I think the data set is comparatively small, only contains 400 rows. That might lead some bias in the machine learning processes. If I have more time, I can maybe find other databases and add them together. Secondly, I can use more models such as LGBM in the model selection part. This might change my best model to choose. Last but not least, in the tuning process, I can use more parameters to have a RMSE score even higher.

Word Count: 1973

VI. Reference

Acharya, Mohan S. “Graduate Admission 2.” Kaggle, 28 Dec. 2018, https://www.kaggle.com/datasets/mohansacharya/graduate-admissions.

Nilanml. “How to Get into Graduate School ?” Kaggle, Kaggle, 5 Feb. 2019, https://www.kaggle.com/code/nilanml/how-to-get-into-graduate-school.

sreshta140. “Chances of Getting into My Dream University.” Kaggle, Kaggle, 14 June 2020, https://www.kaggle.com/code/sreshta140/chances-of-getting-into-my-dream-university.

Problem set 1-10 by myself