Probability of Entering your Dream Graduation Program

Sean Xu

DATA1030 Fall22 S01

Hands-on Data Science

Dec 6th 2022

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

Introduction

Problem:

- What are factors deciding whether you are admitted to your dream program?

Importance:

- What can you do to improve the probability?

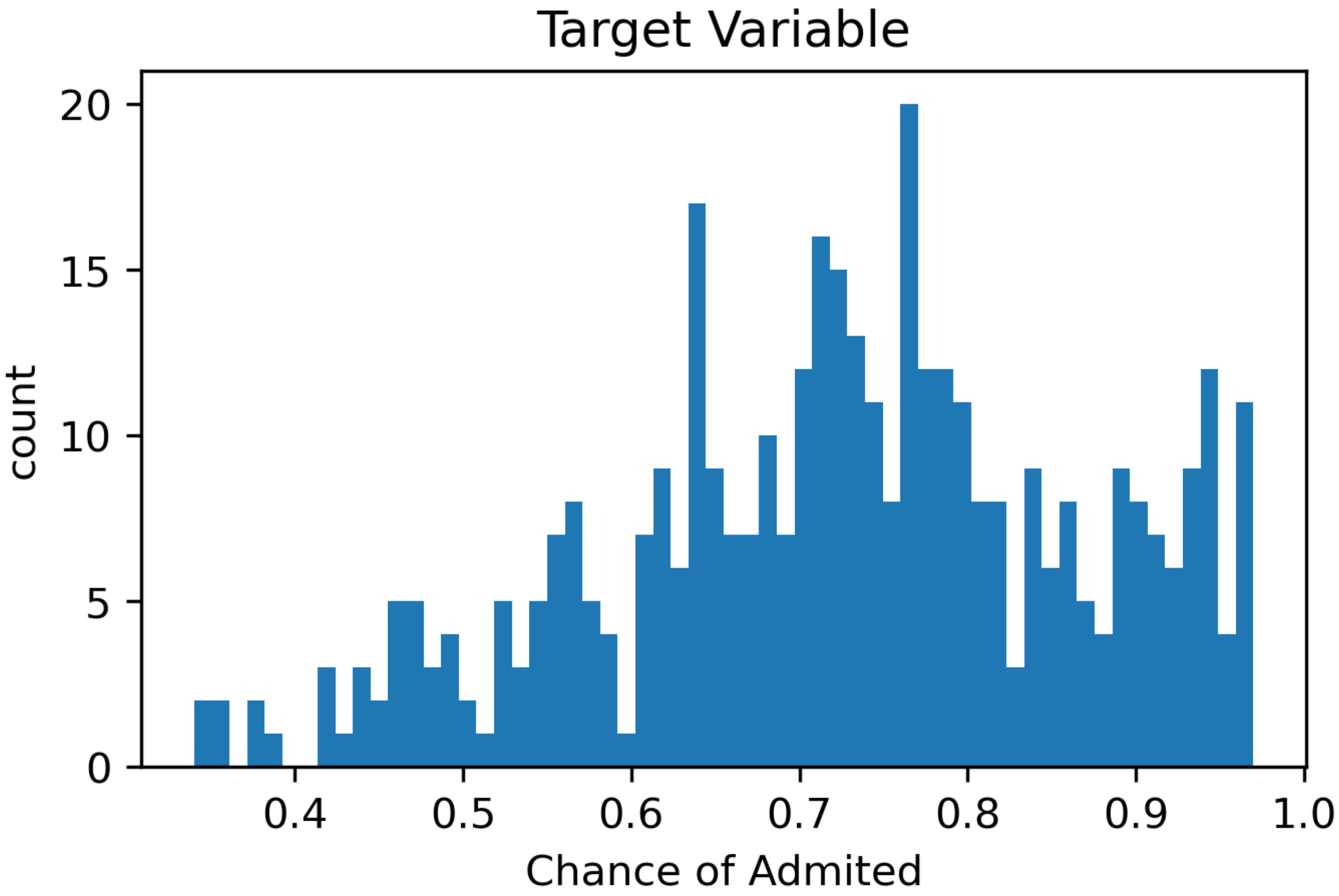
Recaps

- Target variables: Chance of Admitted

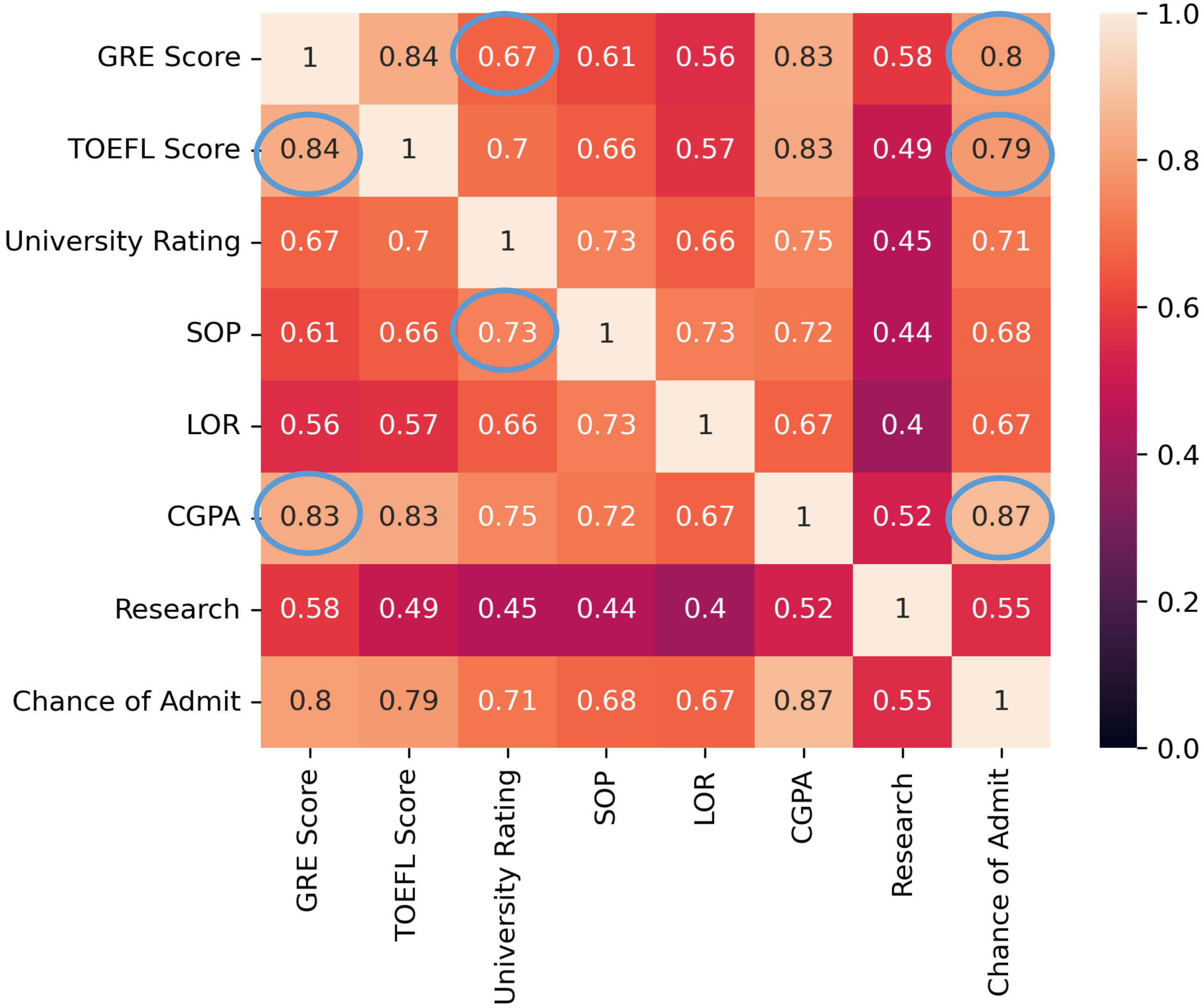
- Regression: Probability (0% - 100%)

- 400 columns x 8 rows
- 7 features 1 target variables

- Kaggle: UCLA Database



Recaps



Not so surprising

- V.S. target variables

Highest three:
GPA (0.87) GRE (0.8) TOFEL (0.79)

- Hard skills

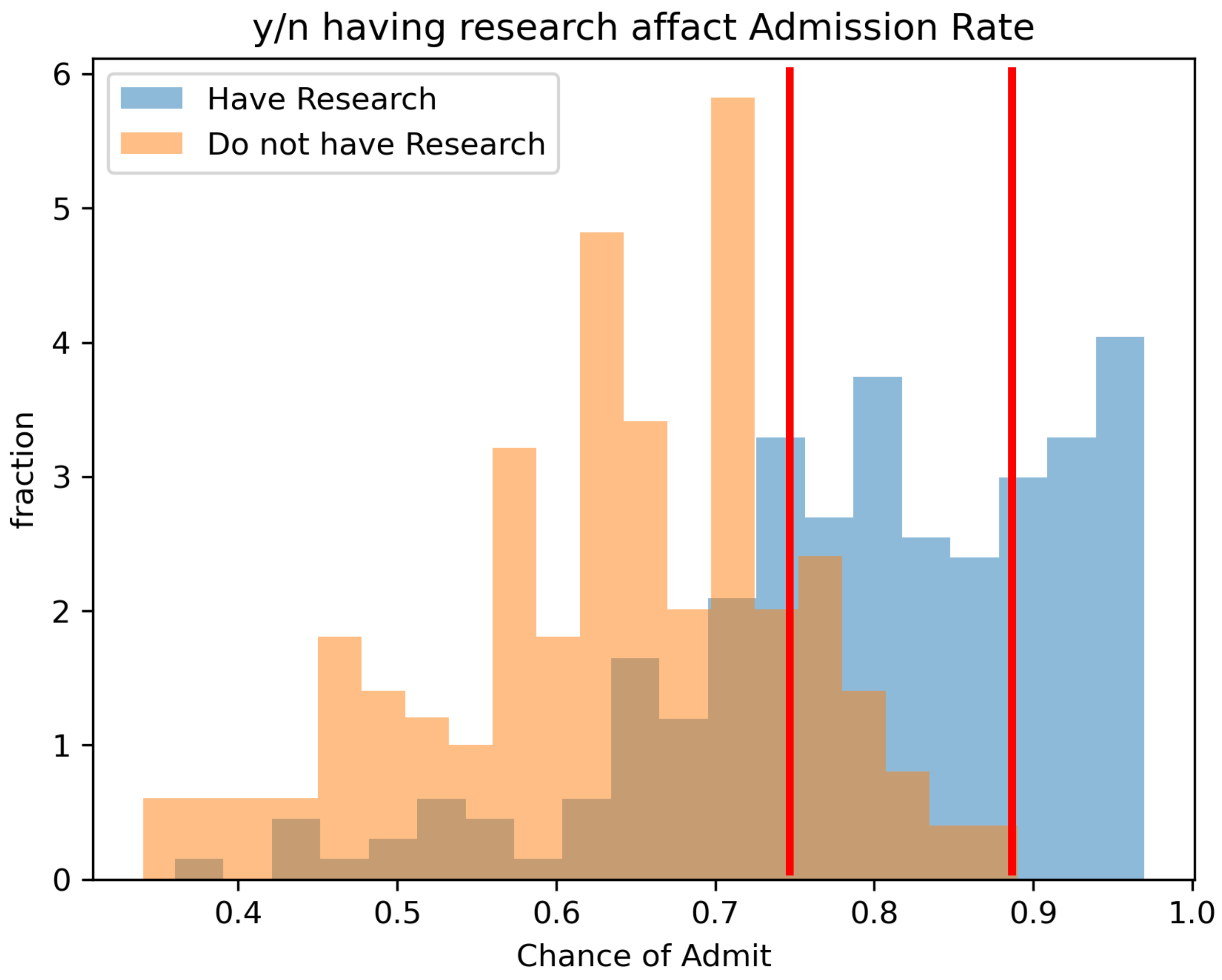
GPA and GRE and Tofel are all over 0.8

But what surprised me was
University ranking with SOP and GRE

Recaps Search vs Admission

- Most of the people do not have research has less than 75% chance of admit

- If you want chance of admit > 90% need research



CV Pipelines

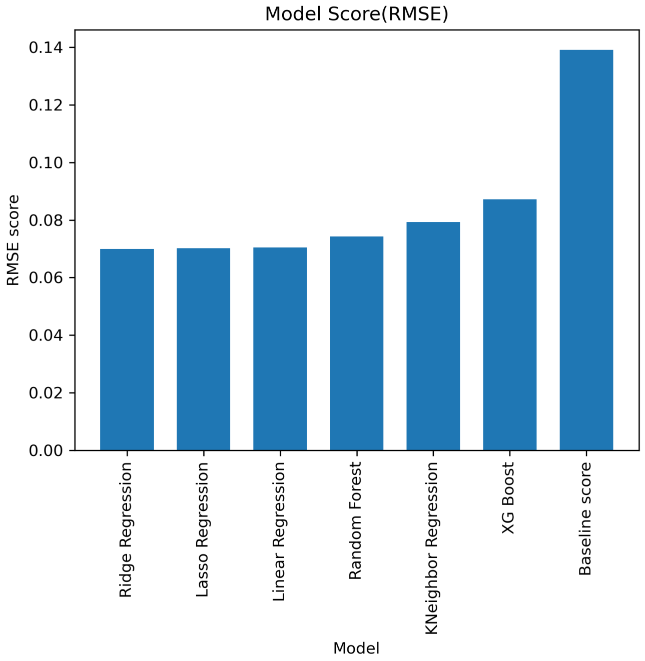
- No missing value
- Simple Split, 5 random state
- Use MinMax Scaler to continuous variables (GRE, TOFEL, and GPA)
- Use Onehot Encoder to categorized variables

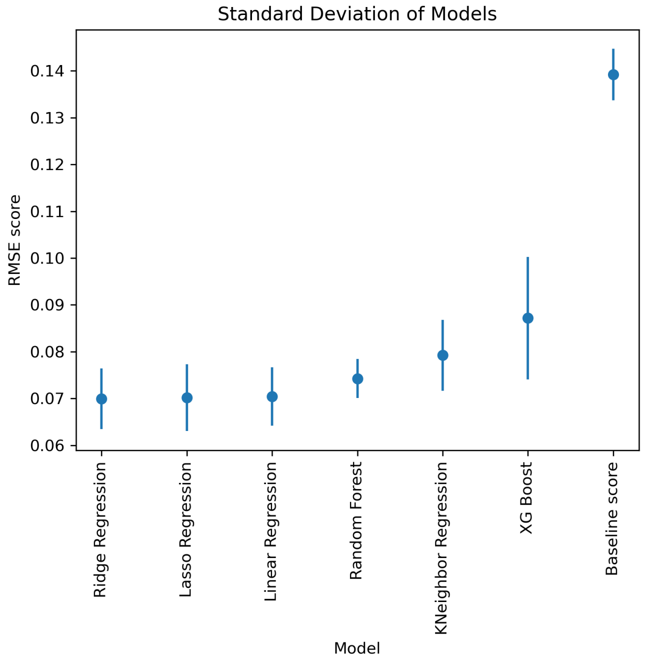
CV Pipelines & Cross Validation

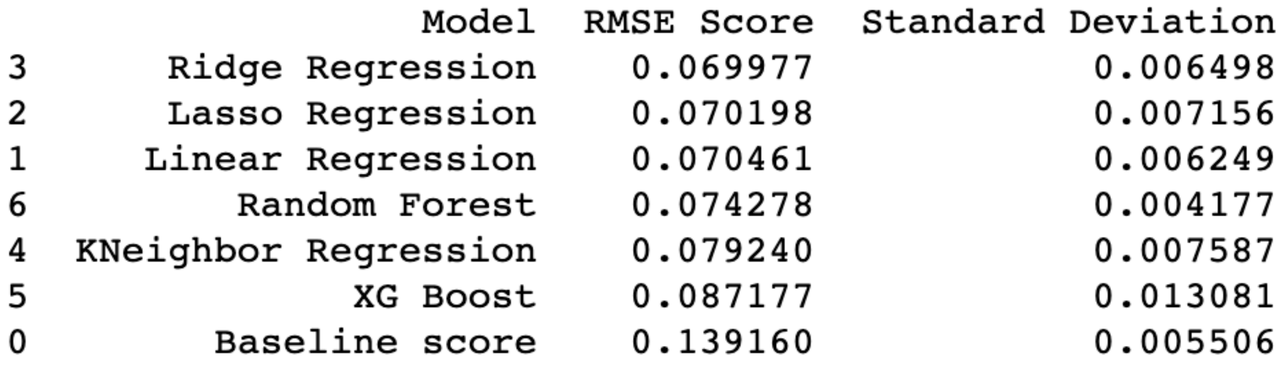
- 6 Regression models
- Record the best score for each models in each random state
- Loop throughout 5 random State to find mean
- Use RMSE

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

Result (Test Score)





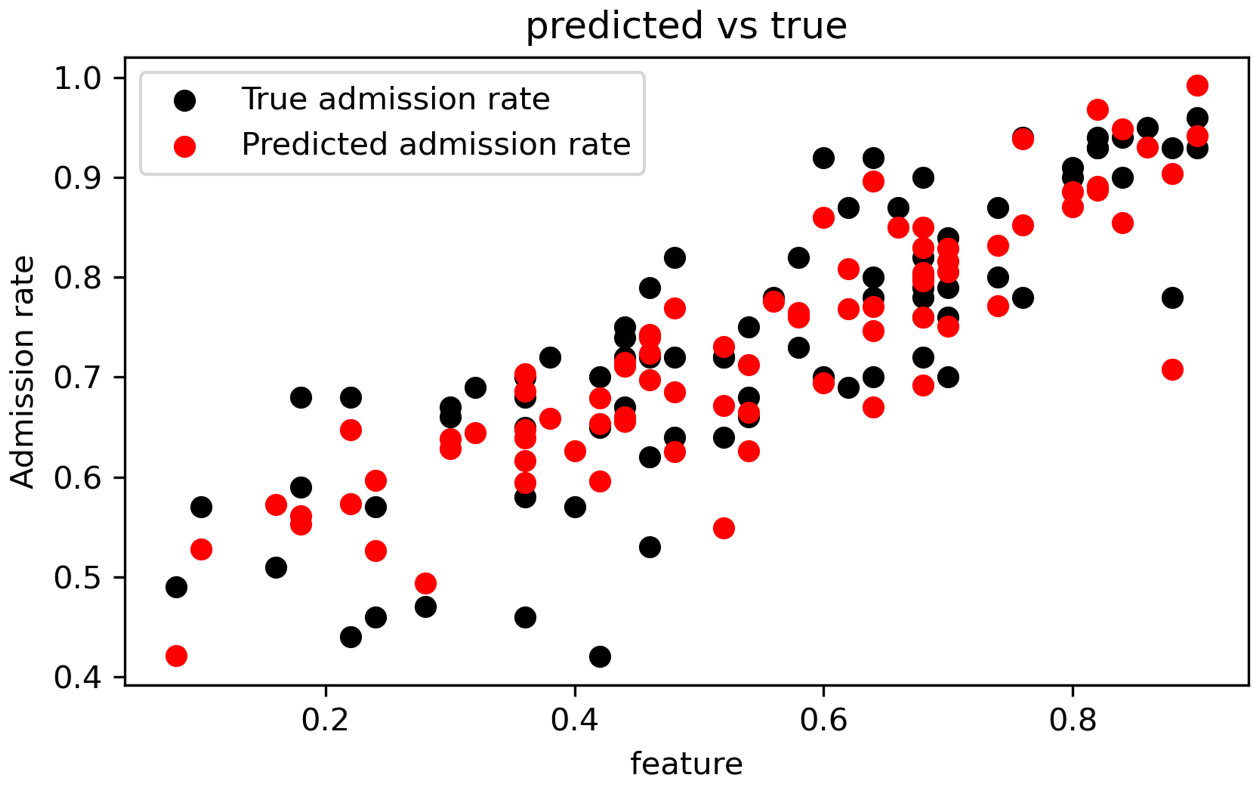


```
                   Model  RMSE Score  Standard Deviation
3       Ridge Regression    0.069977            0.006498
2       Lasso Regression    0.070198            0.007156
1      Linear Regression    0.070461            0.006249
6          Random Forest    0.074278            0.004177
4   KNeighbor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

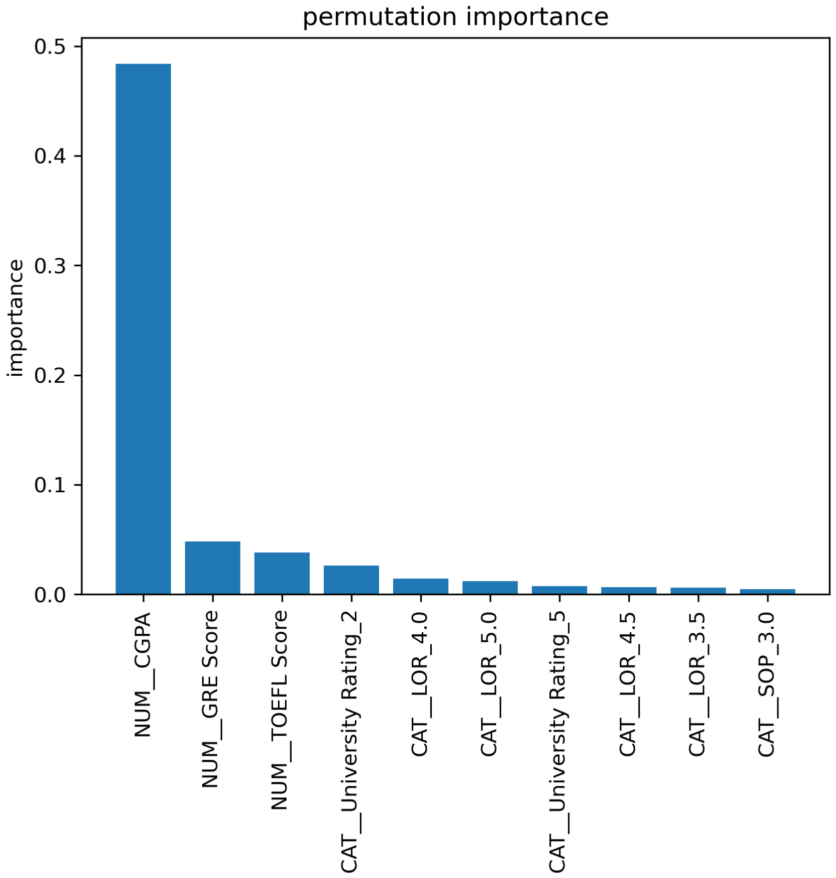
Result
(Model inspection)

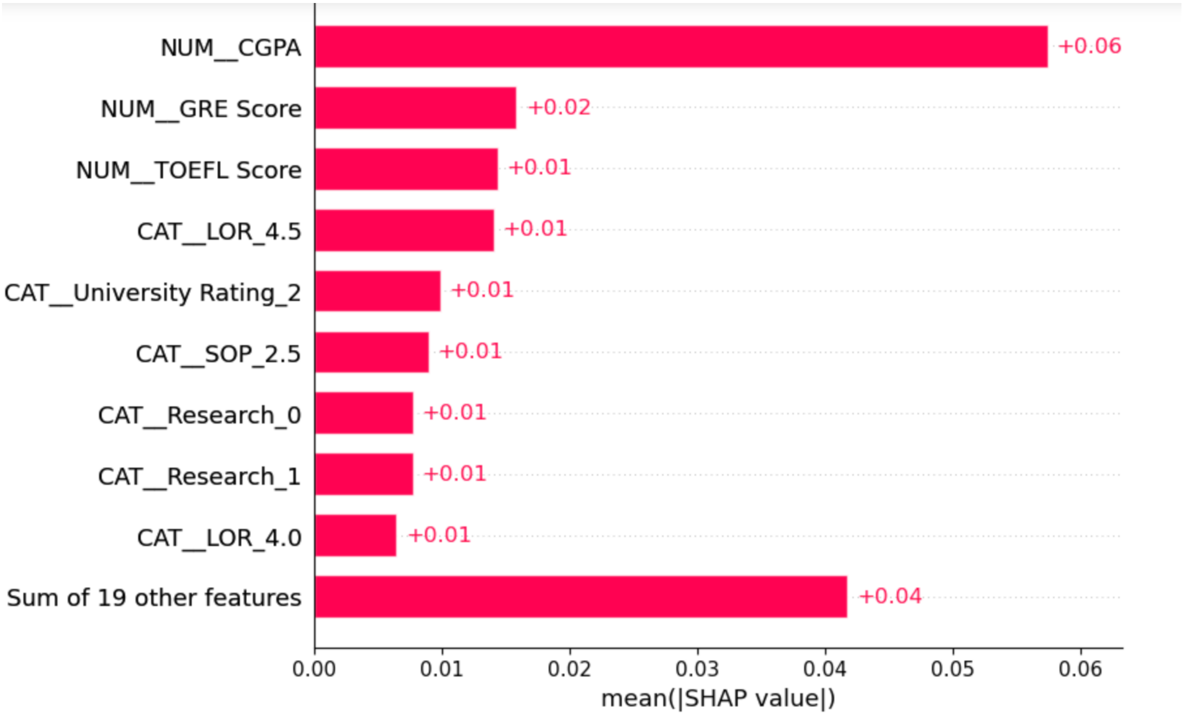
- Predicted Chance of Admission is not so far from the data points

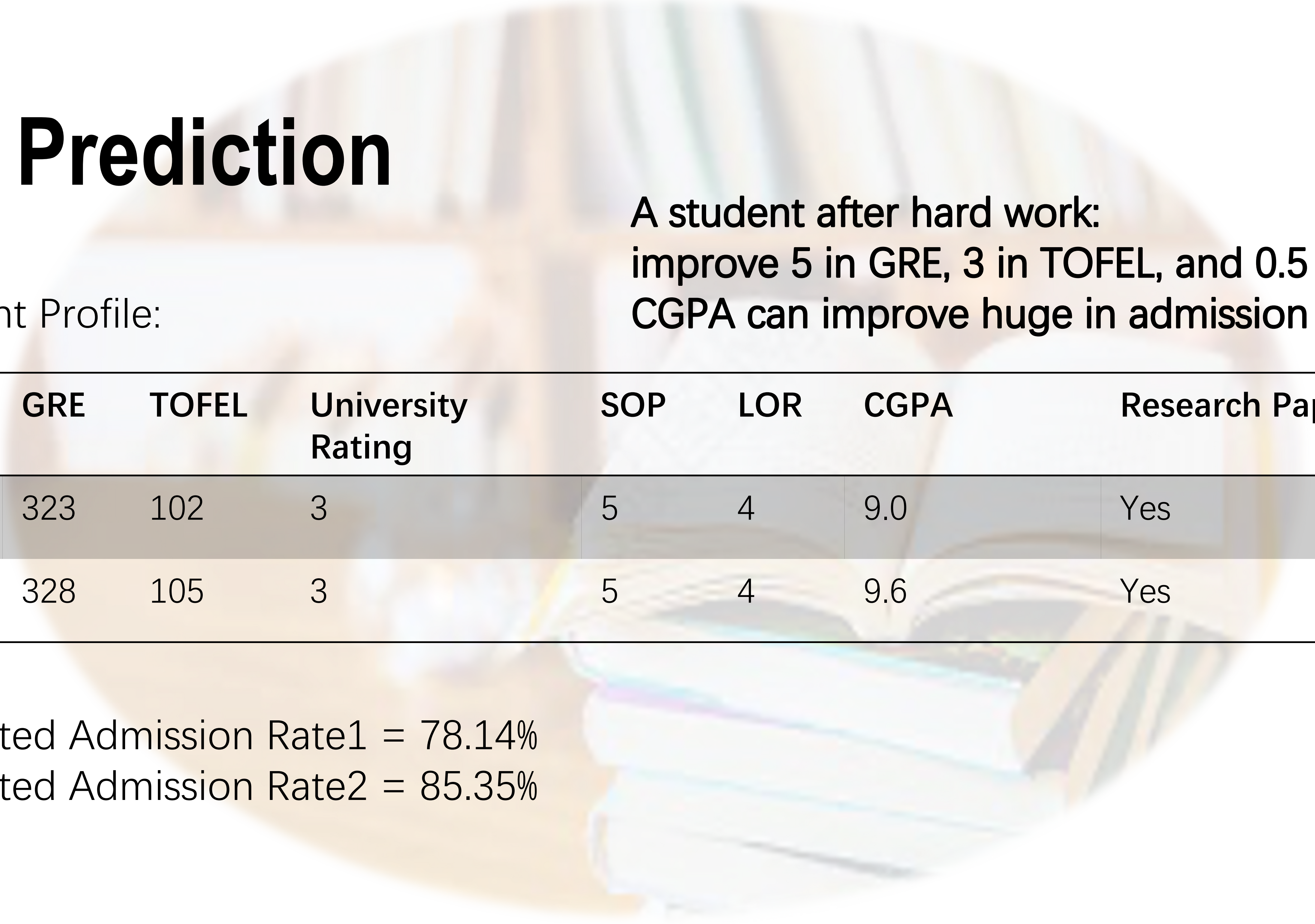
- Generally in a line



Result (Feature importance)







My Prediction

A student after hard work:
improve 5 in GRE, 3 in TOFEL, and 0.5 in CGPA can improve huge in admission rate

Student Profile:

NO.	GRE	TOFEL	University Rating	SOP	LOR	CGPA	Research Paper
1	323	102	3	5	4	9.0	Yes
2	328	105	3	5	4	9.6	Yes

Predicted Admission Rate1 = 78.14%
Predicted Admission Rate2 = 85.35%

Outlook

- Try other models like LGBM
- The size of data is not so big
- Try to decrease std
- Tune number more