# Using data visualization and machine learning methods to analyze graduate school admission rate

*Team 22 Haotian Liu, Yifei Zhao, Yifu Yuan, Yujie Guo*

Abstract

In this final project, our group explored the graduate school admission dataset from Kaggle (https://www.kaggle.com/datasets/akshaydattatraykhare/data-for-admission-in-the-university). We aim to find out how those features can affect the graduate school admission of a student. We used different models to do the regression training and predictive analysis.

## 1. Introduction

Our group members are Seniors and Juniors at WPI. We need a plan after graduation. Continuing study in graduate school is a great idea, and we are all curious about how to get in. We found a data set for graduate school admissions on Kaggle.

In the dataset, multiple features can affect our graduate school admission. However, we are still determining how each feature will impact our graduate school admission and which one or multiple features will play a significant role. Hence, our central target for this project has two goals: First, answering what the relationship of each feature with the final admission chance is. Train each model and select the most precise one in prediction.

### 1.1 Dataset combing and explanation

The dataset includes multiple pieces of information: GRE score, TOEFL score, university rating, SOP (Statement of Purpose), LOR (Letter of Recommendation), CGPA, research, and the chance of admission. In this dataset, 400 entries are included.

### 1.2 Evaluation scales of each feature

GRE Scores: out of 340
TOEFL Scores: out of 120
University Rating: out of 5
SOP: Statement of Purpose
LOR: Letter of Recommendation Strength (out of 5)
Undergraduate GPA: out of 10
Research Experience: either 0 or 1
Chance of Admit: ranging from 0 to 1

## 2. Data Visualization and analysis

First, we used a pair plot (Fig1) to understand the data distribution and correlation between each data feature with the admission chance. We used the correlation analysis from the *Sklearn* python package to give the correlation coefficient table (Table1) and correlation heat map (Fig2), which depicts how strongly each feature correlated with the admit chance. The above graphs and chart show that CGPA is the most correlated parameter contributing to admission. Other parameters show high correlations: GRE Scores, TOEFL Scores, and University Rating. Our following analysis will focus on those four features.

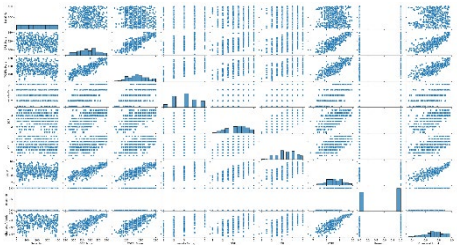

Fig1 Pair plot

| Feature | Coefficient with Admission |
|---|---|
| GRE Scores | 0.802610 |

| | |
|---|---|
| TOEFL Scores | 0.791594 |
| University Rating | 0.711250 |
| SOP | 0.675732 |
| LOR | 0.669889 |
| CGPA | 0.873289 |
| Research | 0.553202 |

Table1 correlation coefficient

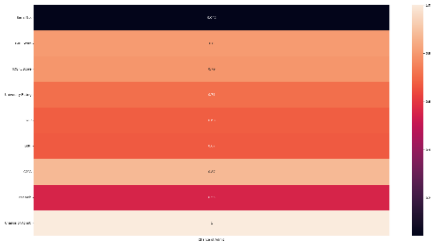

Fig2 correlation heat map

## 3. Model selection, training, and test

Visualizing the data is far from enough, so we will conduct regression analysis and use machine learning methods: Linear Regression, Polynomial Regression, Lasso Regression, Logistic Regression, Random Forest, and KNN to make a prediction.[1] The following paragraphs will interpret each model used and give a rational analysis for each of them.

### 3.1 Lasso Regression

After thoroughly exploring the relationship between features and admission chances, we are facing choosing which type of models to begin analyzing. Since we only need some of the features to create a good model. In other words, we must select predictors that help produce the best model. From the machine learning algorithms, we have a few algorithms we can use to help us. They are forward selection, backward selection, best subset selection, ridge regression, and lasso regression. Evaluating each model individually, forward, backward, and best subset selections are made. The selection process will be significantly slower if the total number of models is vast.[2] Ridge and lasso regression are excellent methods because they act as an extension of the linear model. By adding terms that penalize less useful predictors, ridge and lasso regression can force their coefficients to close to zero if they are not less related to the model.[3] Hence, they require less computational power (no need to go through all models). Lasso works better if there are a few significant predictors and many less significant ones. Ridge, however, works better if there are many significant predictors. Hence, we use Lasso regression to find the best model and calculate its score.

When selecting the best lambda values for our lasso regression, we are surprised that the best value is almost 0 (Fig 3).

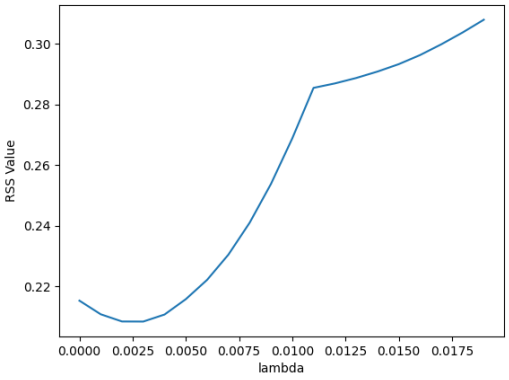


Fig3 lambda vs RSS value

Value equals zero means there should be no penalty for any terms. In other words, we should keep the model as a linear regression model if we want to build the best model.

### 3.2 Linear Regression

We start by fitting a linear regression model. The reason for picking a linear model is that those essential features show, from the pair plot, a linear relationship with Admit Chance.

We used K-fold cross-validation methods; Fig3 and Table 2 show the linear regression model fit results and the test score details.

The all-feature linear regression (Fig4) test score (coefficient of determination R^2) shown in

table 2 is 0.808, so we need further study to analyze each vital correlation feature and use other models to predict.

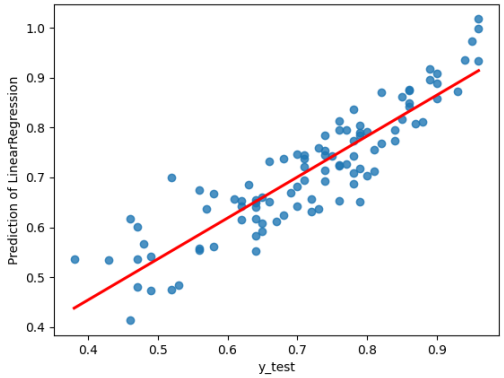


Fig4 all feature linear regression

| | |
|---|---|
| Score for Test Data | 0.8082208515051801 |
| Mean Absolute Error | 0.04593455753180755 |
| MSE | 0.0035045721595943395 |
| RMSE | 0.05919942702082799 |

Table2 all feature test results

The following graphs and tables show those four strong correlations feature linear regression models and test results.

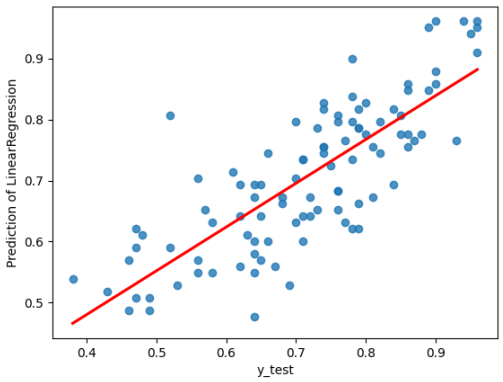


Fig5 GRE Score linear regression

| | |
|---|---|
| Score for Test Data | 0.625619346345464 |
| Mean Absolute Error | 0.064453644517201 |
| MSE | 0.0068414320648829885 |
| RMSE | 0.08271294980160597 |

Table3 GRE Score results

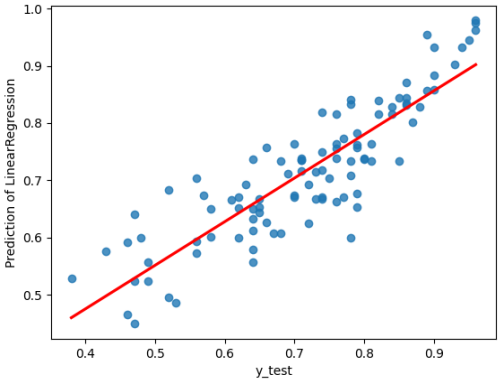


Fig6 CGPA linear regression

| | |
|---|---|
| Score for Test Data | 0.765141842088592 |
| Mean Absolute Error | 0.05004704773967004 |
| MSE | 0.004291797977673068 |
| RMSE | 0.06551181555775315 |

Table4 CGPA Score results

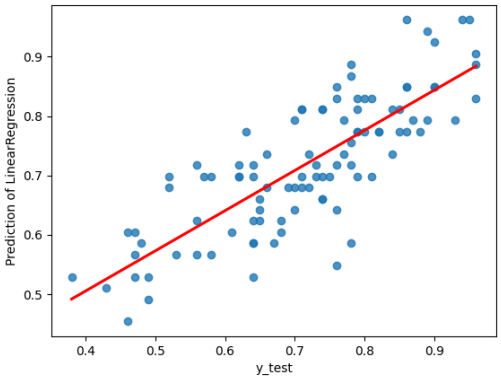


Fig7 TOEFL Score linear regression

| | |
|---|---|
| Score for Test Data | 0.6428885975973616 |
| Mean Absolute Error | 0.06610533721484149 |
| MSE | 0.006525853767505812 |
| RMSE | 0.08078275662234988 |

Table5 TOEFL Score results

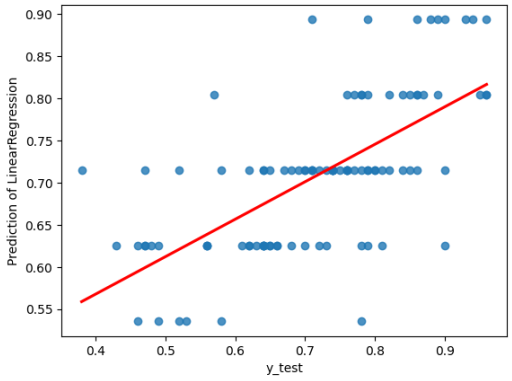


Fig8 University rating linear regression

| | |
|---|---|
| Score for Test Data | 0.42733730788747704 |
| Mean Absolute Error | 0.07545999227277118 |
| MSE | 0.010464838035664243 |
| RMSE | 0.10229779096180056 |

Table6 University rating results

From the above graphs and tables, it is easy to see that the linear regression shows acceptable performance in multiple data fits but doesn't fit each feature well, test scores, GRE: 0.626, CGPA: 0.765, TOEFL: 0.643, and University rating: 0.427, are not that acceptable.

### 3.3 Polynomial Regression

We did Polynomial Regression tests for each crucial feature selected from the data set. From the linear regression model, the test scores are not that satisfied. Polynomial regression is more flexible than linear regression. It is less biased and has a higher variance than the linear regression model.[4]

Since it is more flexible, we expect to end up with a better model using Polynomial Regression rather than linear regression.

The change in test score value after adding higher degree polynomials to each feature's model shows in graphs (From 9 to 12), which suggests that higher-degree polynomials will only increase or even decrease the model's accuracy in a negligible amount. In other words, polynomials are not crucial in this model apart from the first degree.

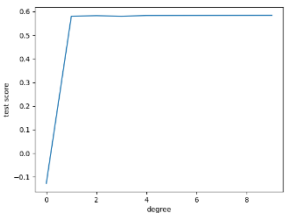


Fig9 GRE Score

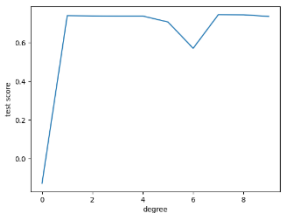


Fig10 CGPA

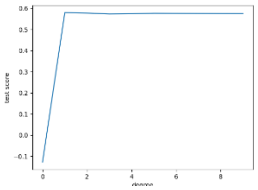


Fig11 TOEFL Score

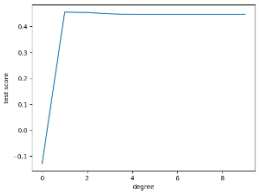


Fig12 University Rating

The following graphs (From 13 to 16) and table 7 show the model fit results. Take GRE Score as an example; Fig 13 shows the model plot when the highest degree of polynomials is 9. The model has an average R squared score of 0.649 from 10-fold cross-validation, slightly better than the linear regression model. This validates what figure 9 has suggested. Increasing the number of degrees of polynomials will not make the model significantly better. Other features performed a similar situation.

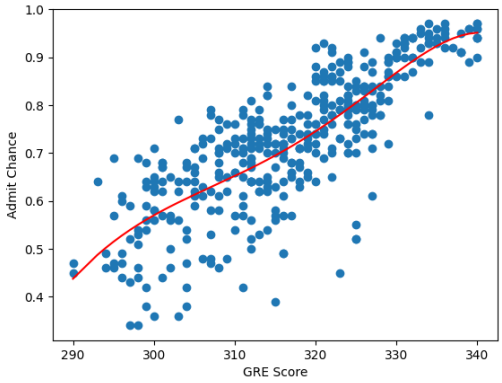


Fig13 GRE Score poly model fit

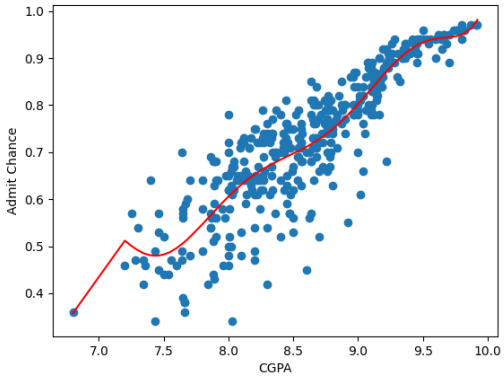


Fig14 CGPA poly model fit

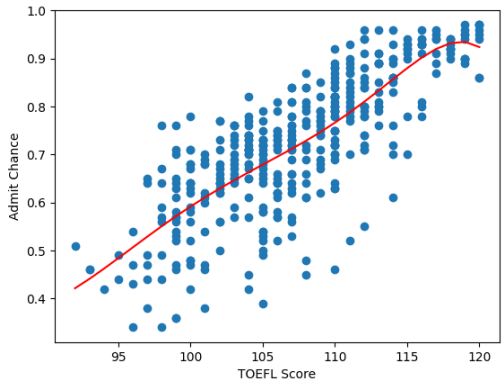


Fig15 TOEFL Score poly fit

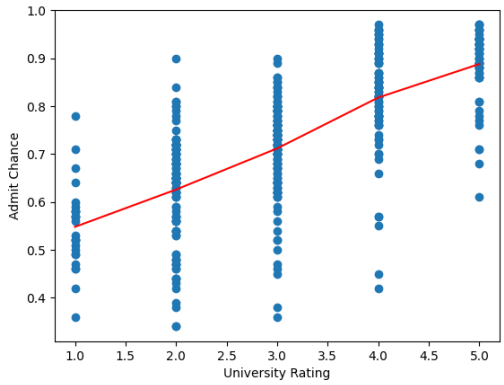


Fig16 University Rating poly fit

| Feature | Poly degree | Test score |
|---|---|---|
| GRE Score | 9 | 0.648912 |
| CGPA | 7 | 0.773106 |
| TOEFL Score | 1 | 0.629942 |
| University Rating | 1 | 0.508291 |

Table7 Feature, Poly degree, and Test score

### 3.4 Random Forest

After conducting linear regression and polynomial regression analysis, the test score is still around 0.8; we decided to use other algorithms to improve the prediction.

Random Forest can be a better choice for our dataset because it can be used in scenarios with many features and high accuracy requirements and does not require dimensionality reduction or feature selection.[5] A random forest can make predictions about the mean value of previously observed labels, and the range of predictions it can make is bounded by the highest and lowest labels in the training data.[6] In addition, it can judge the importance of features and the mutual influence between different features. It is not easy to overfit and is relatively simple to implement, which is excellent for our dataset.[7]

We used the *stage_predict*() function provided by *Sklearn,* which can measure the validation error at each stage of training to find the optimum *n_estimator*; besides that, we used cross-validation to validate the *n_estimator* it shows that increasing the size of the *n_estimator* can improve the accuracy of the model on the training set and the test set because increasing the number of trees can reduce bias and variance. It can also be found that the model does not overfit the training data as the complexity increases. Naturally, the random forest will stabilize after some *n_estimator*. Since there is no benefit to adding weak tree estimators, we can choose around 800.

The following graph (Fig 17) and table (Table 8) show that Random Forest has more accuracy (test score is 0.8402) than Linear regression in all feature predictions. However, from table 9, for each feature, the test score does not show significant improvement.

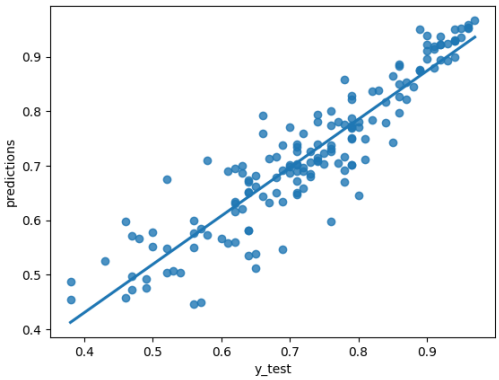


Fig17 All feature Random Forest

| Score for Test Data | 0.8400200049390012 |
|---|---|
| Mean Absolute Error | 0.040658583333332964 |
| MSE | 0.003033298196666646 |
| RMSE | 0.05507538648676599 |

Table 8 all features Random Forest results

| Feature | Test Score |
|---|---|

| | |
|---|---|
| GRE Score | 0.6216338590373303 |
| CGPA | 0.7259695081091487 |
| TOEFL Score | 0.6487262516454298 |
| University Rating | 0.4603595520094591 |

Table 9 Random Forest each feature

3.5 KNN

We conducted KNN on our data set, and in the application, the cross-validation method is usually used to select the optimal K value.[8] It can also be known that the general K value is relatively small. We will choose the K value in a smaller range. The one with the highest accuracy rate on the verification set will be determined as the final algorithm hyperparameter K.

But with the carefully choose of K, the test results have not improved significantly. The test results are in Table 10.

| | |
|---|---|
| Score for Test Data | 0.7524761956878625 |
| Mean Absolute Error | 0.04875 |
| MSE | 0.004523249999999999 |
| RMSE | 0.06725511132992049 |

KNN test results

3.5 Logistic Regression

Because Logistics Regression is a classification model and often used for binary classification[9], we created a new predictor called "well prepared" based on the chance of admission. The well-prepared has either a value of 0 or 1. It is an indicator of whether a person is ready for admission based on other predictors. In this section, we will try to predict whether a person is well prepared to apply to a specific college for a master's degree. Logistic regression will create a model that predicts the probability of responses based on predictors. Also, the probability will always be meaningful since the output will always be between 0 and 1.[10] From the lasso regression, we found out that all predictors are meaningful. In the logistic regression, we will also include all the predictors.

The result is delightful, and we achieved almost 100% accuracy. The actual accuracy for test data when splitting into five folds is 98.75%. The test error is 1.25%.

4 Conclusion

From the above study, we can conclude that the Random Forest performs best in multiple feature prediction conditions, with a test score of 0.840. We will use other methods and data sets with more features in future research to get the *n_estimator*.

For single-feature prediction, the Polynomial Regression gives the best results (Table 7) but is only slightly improved compared with the Linear Regression; it may be because the data set performs a robust linear relationship, and the degree reached the boundary at the very begging. The Random Forest did not provide a satisfying prediction when it only had a single feature. Because using a smaller number of input features reduces the similarity between the individual trees but also reduces the tree's complexity and, thus, a reduction in the tree's strength.

With a new predictor based on Admit chance imported, we conducted Logistic regression. The test results are highly idealistic, and future work needs to be done to interpret these results.

5. References

[1] https://towardsdatascience.com/a-practical-introduction-to-9-regression-algorithms-389057f86eb9

[2] Smith, G. Step away from stepwise. *J Big Data* **5**, 32 (2018). https://doi.org/10.1186/s40537-018-0143-6

[3] Articles - Model Selection Essentials in R Penalized Regression Essentials: Ridge, Lasso & Elastic Net http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/153-penalized-regression-essentials-ridge-lasso-elastic-net/

[4] https://towardsdatascience.com/polynomial-regression-bbe8b9d97491

[5] Chen, RC., Dewi, C., Huang, SW. *et*

*al.* Selecting critical features for data classification based on machine learning methods. *J Big Data* **7**, 52 (2020). https://doi.org/10.1186/s40537-020-00327-4
[6] A limitation of Random Forest Regression https://towardsdatascience.com/a-limitation-of-random-forest-regression-db8ed7419e9f
[7] Model Complexity & Overfitting in Machine Learning. May 29, 2022 by Ajitesh Kumar https://vitalflux.com/model-complexity-overfitting-in-machine-learning/
[8] Cross-validation using KNN. Deepak Jain https://towardsdatascience.com/cross-validation-using-knn-6babb6e619c8
[9] Why Is Logistic Regression a Classification Algorithm? Log odds, the baseline of logistic regression, explained. Written bySparsh Gupta Published on Jun. 09, 2022 https://builtin.com/machine-learning/logistic-regression-classification-algorithm
[10] McLeod, S. A. (2019, May 20). What a p-value tells you about statistical significance. Simply Psychology. www.simplypsychology.org/p-value.html