Analysis of Effects of College Students Performance on Graduate School Admission

Authors: Phil Lin, Jiasui Qin

Summary of Research Questions

1. **What aspects of student's college performance can increase the graduate school admission rate the most?**
   We aim to find which features contribute more to the admission rate. Students who are applying for graduate schools could then have a better idea of what they should pay more attention to.

   Result: Among the 7 features, the most important three features are Undergraduate GPA, GRE Score, and TOFEL Score.

2. **How closely are the exam scores (TOEFL, GRE, and CGPA) related to the admission rate, or other aspects of student performance?**
   We aim to find the difference between taking the two exams and how the exam scores are related to the admission rate. And we also aim to find the correlation between all these aspects of student academic performance.

   Result: the higher the score is, the larger the chance of admit. However, chance of admit varies widely for mediocre students. The three scores have high correlation with each other, and the research experience plays the least role in graduate school admission process.

3. **How does the rating of the college affect the admission rate?**
   We aim to see the relationship between the college ratings and the admission rate. It helps to find out whether colleges with high ratings have lower admission rates.

   Results: In general, colleges with high ratings tend to have higher admission rates. But the correlation is not strong.

Motivation and Background

As we are becoming rising juniors, it is the time that we start to plan on our future professional career or academic career. While many students might choose to work after they graduate, it is also a popular choice to go to graduate school to deepen the knowledge in a more specific academic area. We want to dive into the data about graduate school admission so that we can

give advice and suggestions to students who are still deciding which exam to take, or which part of the student performance is more helpful to put more work in. If we can answer these questions, students can see their goals clearly and work motivationally.

Dataset Description

Our data, found in https://www.kaggle.com/mohansacharya/graduate-admissions, is scraped from *Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019.* The dataset contains several parameter which are considered important during the application for Master's Program. The parameters include:

1. GRE scores
2. TOEFL scores
3. University Rating
4. Statement of Purpose
5. Letter of Recommendation Strength
6. Undergraduate GPA
7. Research Experience
8. Chance of Admit

Methodology (algorithm or analysis)

1. **Research question one**
   For our first research question, we will train a linear regression model with the chance of admission as the target data and all other features as independent variables. We will first split the data into testing dataset, validation dataset, and training dataset for training and testing purposes. To standardize input values for training, we make each feature have a mean of 0 and a std deviation of 1. We also center the target feature, which is admission rate, to make it have a mean of 0. As stated in our first challenge goal, we will use grid search to find the optimal hyper-parameters of our model, which is specifically the bias term in Lasso regularization. By using Lasso regularization, we can then extract the most relevant attributes to help us determine what qualities universities put a high value on.

2. **Research question two**
   For the second research question, we will use a new library, Plotly, to create several box plots on GPA and two test scores versus the admission rate as well as a heat map to visualize how each attribute contributes to the admission rate.

3. **Research question three**

For the third research question, we will train a new linear regression model with the rating of the college and the chance of admission as independent variable and dependent variable. We then plot the regression line to make analysis on how the chance of admission changes with different ratings of colleges.

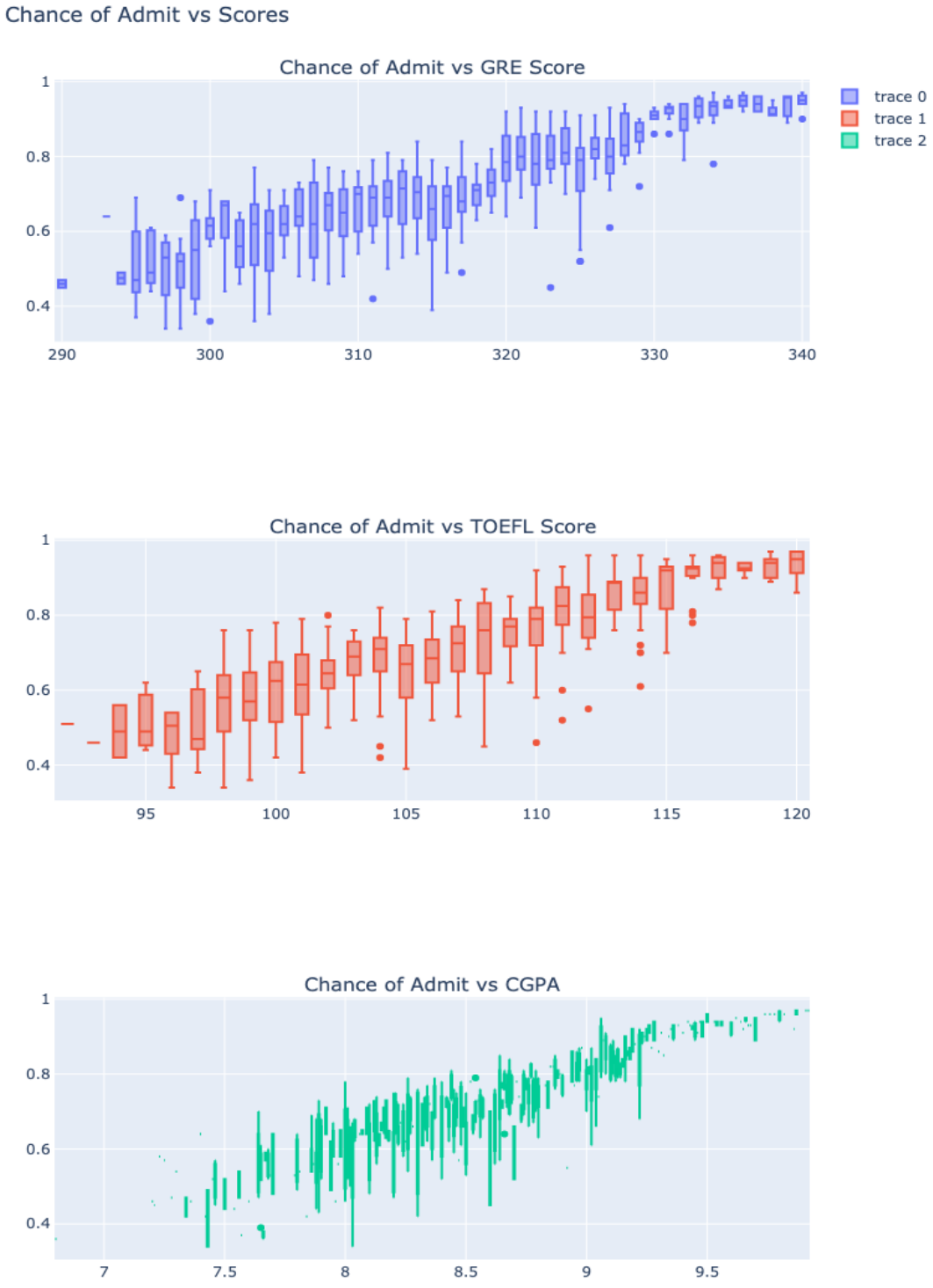
Results

1. **What aspects of student's college performance can increase the graduate school admission rate the most?**
   For alphas ranges from 0.01 to 0.09 with 0.01 increase for each value, the grid search test scores returns [0.78948907 0.76546255 0.72738043 0.68199966 0.62791971 0.56333294 0.48698761 0.40219235 0.31016021]. Therefore, the best alpha found by grid search is 0.01. This shows that the Lasso regression model does not need a high penalty to avoid overfitting because we have enough training points and features are not complicated. After building the model with this alpha, we get the weight of each feature:

   'GRE Score' - 0.02033514468007994
   'TOFEL Score' - 0.018411024857731103
   'University Rating' - 0.0015769490017739094
   'Statement of Purpose' - 0.0
   'Letter of Recommendation' - 0.012094787000636133
   'Undergraduate GPA' - 0.06900818262310371
   'Research' - 0.010516212430885429

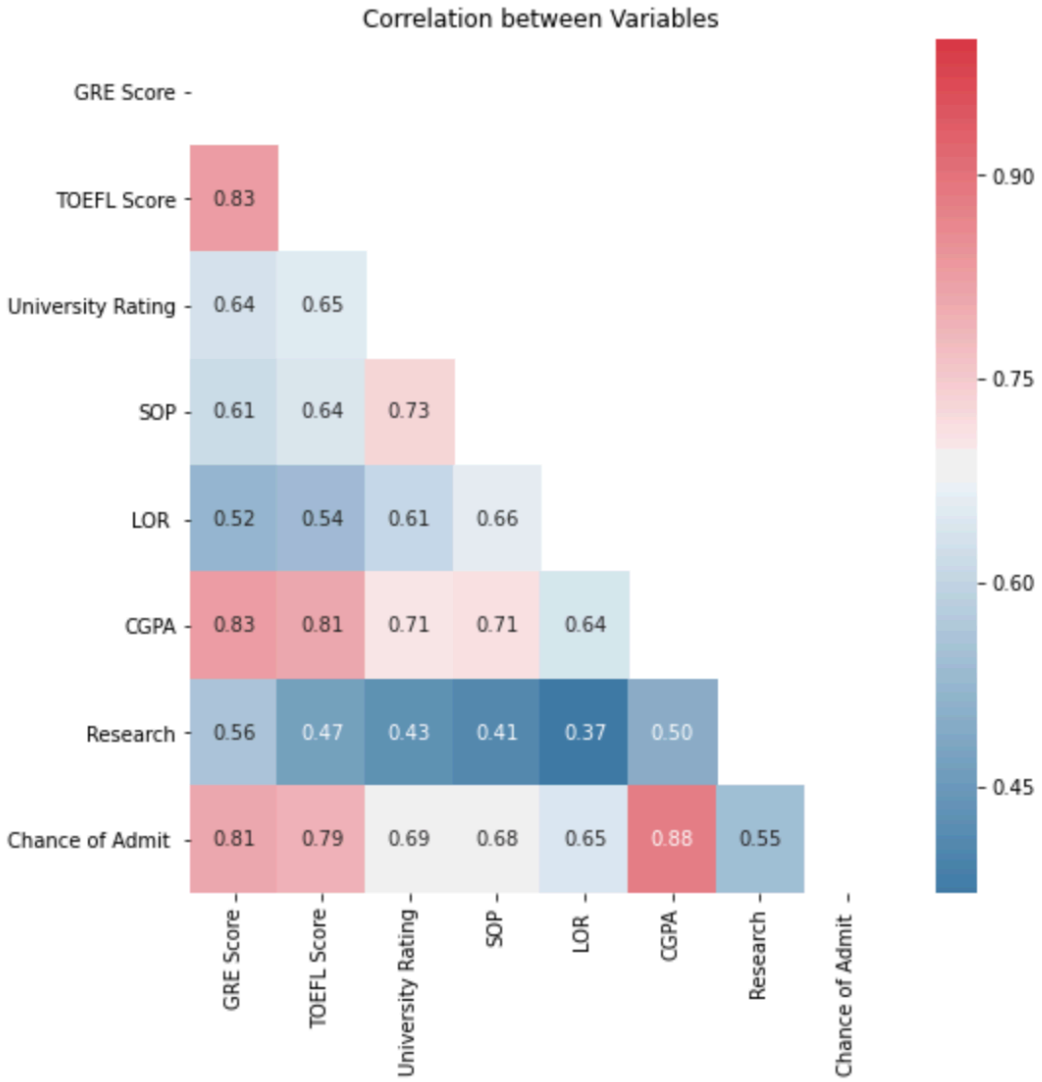
   In conclusion, the rank of weights for the seven features is: Undergraduate GPA > GRE Score > TOFEL Score > Letter of Recommendation > Research > University Rating > Statement of Purpose. The weight of GPA is more than three times of the weight of GRE score and the weight of TOFEL score. We also noticed that the weight of statement of Purpose and the weight of university Rating are extremely low, meaning that they take a less important role in achieving a high admission rate. Therefore, students can focus more on their undergraduate gpa score as well as test scores of GRE and TOFEL.

2. **How closely are the scores (TOEFL, GRE, CGPA) related to the admission rate, or other aspects of student performance?**



We created three box plots to visualize the distribution of scores and the relationship between scores and admission rate. From the plots above, we can clearly see that the

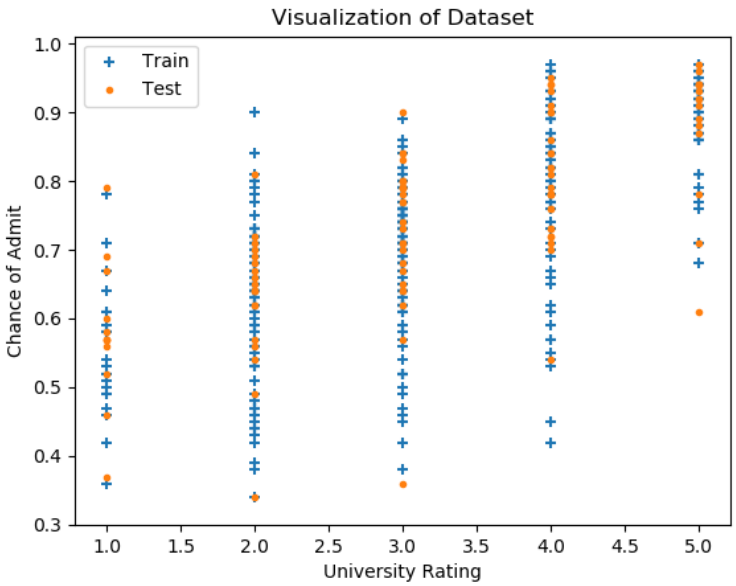
higher the scores, the more possible for the student to receive the offer. The pattern is the same among the three type of scores. However, we can see that the chance of admit varies largely for mediocre students. For example, among the students who have 315 in GRE, the highest admission rate is 0.79 and the lowest is 0.39. This pattern also appears in the other two plots. And the students who have nearly the perfect scores have very high chance of admit.

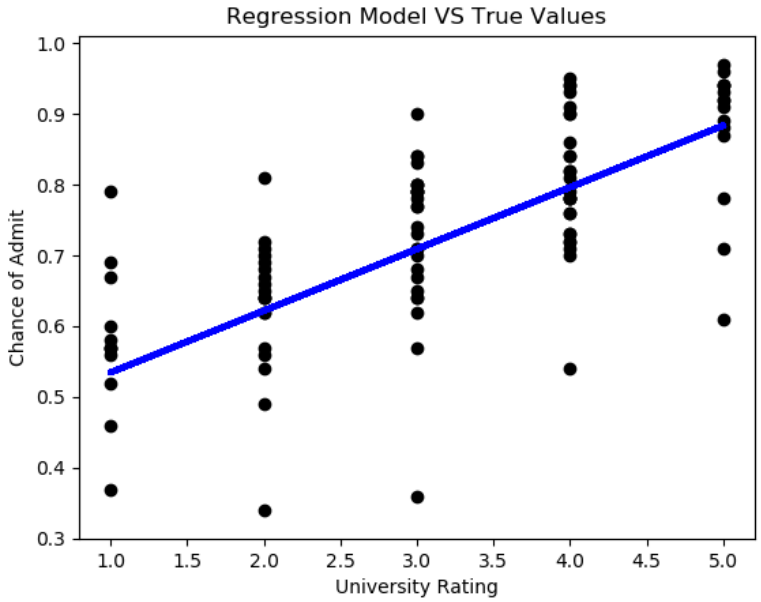


This heatmap helps us answer the second part of the question. We calculate the correlations of aspects of student academic performance in pairs. From the heatmap we can easily visualize that the three most important aspects of performance that helps increase admission rate are CGPA, GRE score, and TOEFL score. This result matches with our machine learning model, which suffices our expectation. In addition, we can see that the correlations between these three aspects of performance are high as well, which tells us that students who do well in school work are also good at standardized testing. We have also observe that the research experience plays a relatively small role in the

admission process. Therefore, we suggest students to focus more on their school work and standardized testing if they plan to apply graduation schools.

3. **How does the rating of the college affect the admission rate?**
   The general visualization of the dataset and the plot of the trained regression model are:





The range of admission rates spreads widely. It ranges from a really low value to a high value no matter what university rating is. We also got a 0.484 model score for the trained regression model. This shows us that university rating and the chance of being admitted are not strongly correlated to each other. However, we can still see a general trend of increasing in the second graph. This shows that the chance does slightly increase while university rating increases.

Challenge Goals

We achieved both of the challenge goals below. We built Lasso regression model and tuned hyper-parameter by using GridSearch. We also used a new library, Plotly, to make various plots in research question two.

1. **Machine Learning:** we will be using different models and parameter tuning skills to find the best model that predicts the data.

2. **New Library:** we will be using more advanced visualization libraries like plotly for our data analysis.

Work Plan Evaluation

1. **Parameter Tuning (1 hour):**
   We will be using grid-search algorithm to find the hyper-parameters that could possibly result in the best model.

   Evaluation: It took about 2 hours for us to find the best parameter to train our model. We thought that using grid search would be really fast because it helps find the parameter automatically. However, we didn't realize that we still needed to specify a range of values for grid search algorithm to choose. Finding this proper range took us some time. Besides, to get a better idea of how each parameter performs, we spent some time on calculated the RMSE for each parameter.

2. **Building model (1 hour):**
   We will build a linear regression model with Lasso regularization for research question 1 and another regression model for research question 2.

   Evaluation: Our estimate is accurate. As soon as we found the best parameter for the model, building models were super-efficient using Sklearn.

3. **Plotting (1 hour):**

We will plot two graphs: a heatmap and a box plot. We will be using plotly as our choice of library. We need to first read the documentations to understand the syntaxes. The deepest colors in our heatmap should represent the same features that we find by our machine learning model.

Evaluation: The plotting took much longer than we expected, because we needed to read through numbers of documentations which took the most of the time. And we first generated the plots using plotly.express, but the plots in this format cannot be exported as static images. And we went back to the documentation and realized we could use plotly.graph_objects to achieve the same goals.

4. **Data Analysis (1.5 hour):**
   We will observe the results that we find in the model and the plots and then write down our observations in concise and clear language. In our conclusion, we will be giving advice and suggestions to students who still now know how to improve themselves academically to be successful in the competition of graduation school applications.

   Evaluation:
   The data analysis process is the easiest part of the project, because the visualization helps a lot for us to observe the pattern and trends.

Testing

1. **Research question one**
   For research question one, we split the total 500 data points into 361 training points, 64 validation points, and 75 testing points. After using training data and validation data to train the model, we then used testing data to test the accuracy of our model. The mean squared test error of our model is only 0.04986396861727547. Besides, to test that grid search works well in tuning the best hyper parameter, we also trained models with different alphas and calculated each root mean square error of the validation data. For alphas ranges from 0.01 to 0.09, we found RMSE of validation to be 0.054108, 0.056441, 0.061655, 0.068786, 0.077308, 0.086661, 0.096632, 0.107021, 0.117653. It proves that 0.01 is indeed the best alpha to choose.

2. **Research question three**
   For research question three, we also split total data points into 80% training data and 20% testing data. The model score we got is 0.48389479123306256. This proves that university rating has a very small effect on admission rate.

Collaborations

Plotly documentations :

1. https://plotly.github.io/plotly.py-docs/generated/plotly.graph_objects.Box.html
2. https://plotly.com/python/box-plots/
3. https://plotly.com/python/static-image-export/

Heatmap documentations:

1. https://seaborn.pydata.org/generated/seaborn.heatmap.html
2. https://seaborn.pydata.org/generated/seaborn.diverging_palette.html
3. https://numpy.org/doc/stable/reference/generated/numpy.triu_indices.html
4. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html
5. https://stackoverflow.com/questions/33158075/custom-annotation-seaborn-heatmap