ECSE 411 Final Project - Graduate Admission Data Analysis

Linjiang Ke
lxk334
Case School of Engineering
Computing and Data Science Department

ABSTRACT

The U.S. University admissions process has always been a challenging and daunting task. Various indicators such as test scores, extracurricular activities, essays, and GPA all play a decisive role in admission to U.S. graduate schools. The purpose of this study is to analyze the difficulty of admission to the University of California, Los Angeles (UCLA) graduate schools by analyzing the relationship between various indicators to determine what exactly influences the difficulty of admission and informs the development of machine learning-based predictive models, and finally attaching a comparison of GRE and TOEFL scores. I found that GPA scores, TOEFL scores, and GRE scores are the main components that affect admissions. Despite the generalizability of the data set and potential biases in the data collection, I believe the study can provide valuable insights for prospective students and universities to make more informed decisions in the graduate admissions process.

Keywords: Graduate Admission, Statistical Analysis, Machine Learning, Decision-making

1 Introduction and Background

In recent years, the admissions process for graduate programs in the United States has become increasingly competitive. Universities typically consider a variety of factors when evaluating applicants, including test scores, undergraduate grades, research experience, and the strength of letters of recommendation and personal statements. Therefore, understanding the relationship between these factors and the probability of admission can help prospective students optimize their applications and increase their chances of being accepted into their desired programs. Therefore, in this study, I used multiple data analysis and statistical methods to analyze a graduate admissions dataset with data on various application metrics for over 400 applicants to UCLA. Among these are: GRE scores, TOEFL scores, undergraduate GPA, college ratings, statement of purpose (SOP) strength, letter of recommendation (LOR) strength, and research experience. Our goal is to analyze the dataset to discover trends and patterns in the admissions process and develop predictive models to estimate applicants' chances of admission based on given characteristics, identify the most significant factors influencing an applicant's probability of being admitted, and evaluate the accuracy of linear regression, decision trees, and random forest models.

This graduate admissions dataset has been widely used in the fields of data analysis and machine learning to explore aspects of the admissions process, such as the impact of research experience or the correlation between test scores and admissions chances. In recent years, there has been a growing interest in using machine learning and data mining techniques to inform and optimize the admissions process. By understanding the underlying patterns and relationships between various factors, universities and applicants alike can make more informed decisions in an increasingly competitive graduate admissions environment. In this study, I will explore trends and patterns in the data by conducting a comprehensive analysis of graduate admissions datasets using descriptive statistics and data visualization techniques. This research will provide valuable insights for prospective students and universities to make more informed decisions in the graduate admissions process.

The goals of the study will be influenced by many reasons:

1. Selection bias: The data set is limited to applicants applying to UCLA, which may not be representative of the broader graduate applicant pool. This may limit the generalizability of the study results.
2. Missing or incomplete data: If certain factors are not recorded in the dataset, or if data are missing for certain applicants, this may affect the accuracy of the analysis and predictive models, and some of the data may be very significantly biased.
3. Temporal factors: The dataset is a snapshot in time and may not reflect recent changes in the graduate admissions process or trends in applicant characteristics.

2 The Simple Study

2.1 Independent variables

This dataset contains over 400 applicants to the University of California, Los Angeles (UCLA) graduate program. The independent variables are the following:

1. GRE Scores (Out of 340): The Graduate Record Examination (GRE) is a standardized test that is widely accepted by many universities in the graduate admissions process.
2. TOEFL Scores (Out of 120): The Test of English as a Foreign Language (TOEFL) is a standardized test that measures an applicant's English proficiency. This score is particularly important in admissions because universities usually require a minimum TOEFL score for admission
3. University Rating (1-5): This variable represents the rating or reputation of the applicant's undergraduate institution, from 1 (lowest) to 5 (highest). A higher rating may indicate a more competitive academic environment and stronger preparation for graduate study.[1]
4. Statement of Purpose (SOP) Strength: The Statement of Purpose (SOP) is a personal essay submitted by applicants detailing their motivation for pursuing graduate study and outlining their research interests and career goals. This variable measures the strength of the applicant's SOP on a scale of 1 (lowest) to 5 (highest).
5. Letter of Recommendation (LOR) Strength: Written by a professor, internship supervisor, or other professional, it can illustrate the applicant's academic ability, research experience, and potential for success in graduate studies.
6. GPA (Out of 10): An indicator of academic performance during undergraduate studies, normally, a higher GPA usually indicates the better academic performance
7. Research Experience (0 or 1): Research experience can provide valuable skills and demonstrate an applicant's potential to contribute to their field of study.

2.2 Independent variables

The dependent variable in the dataset is the Chance of Admit, which represents the probability of an applicant being admitted to the graduate program, ranging from 0 to 1.

3 Statistical Analysis

In this research, I will show the possible admission outcomes by comparing the relationships between the different independent variables and how the independent variables affect the dependent variable.

3.1 Data Processing

I will start by importing the libraries and data and showing the first 5 pieces of data so that it will be more intuitive to see the type of data. Then I can get the correlation, covariance, variance, and mean in the data, which is just a number here for now to facilitate a better understanding of the data.

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
0	1	337	118	4	4.5	4.5	9.65	1	0.92
1	2	324	107	4	4.0	4.5	8.87	1	0.76
2	3	316	104	3	3.0	3.5	8.00	1	0.72
3	4	322	110	3	3.5	2.5	8.67	1	0.80
4	5	314	103	2	2.0	3.0	8.21	0	0.65

Figure 1: First 5 details of the data

Then we delete the Serial number because it is useless in the following analysis. The other indexes are as follows:

	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
GRE Score	131.644555	58.216967	8.778791	7.079699	5.747726	5.699742	3.318690	1.313271
TOEFL Score	58.216967	36.838997	4.828697	4.021053	3.095965	2.998337	1.481729	0.685179
University Rating	8.778791	4.828697	1.308114	0.845865	0.678352	0.509117	0.255232	0.116009
SOP	7.079699	4.021053	0.845865	1.013784	0.660025	0.431183	0.222807	0.097028
LOR	5.747726	3.095965	0.678352	0.660025	0.807262	0.359084	0.177701	0.085834
CGPA	5.699742	2.998337	0.509117	0.431183	0.359084	0.355594	0.155026	0.074265
Research	3.318690	1.481729	0.255232	0.222807	0.177701	0.155026	0.248365	0.039317
Chance of Admit	1.313271	0.685179	0.116009	0.097028	0.085834	0.074265	0.039317	0.020337

Figure 2: First 5 details of the data

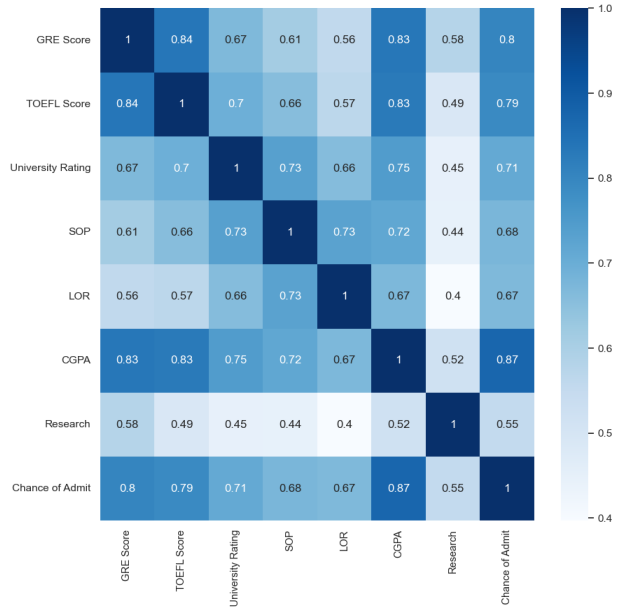

Figure 3: Heatmap of the Correlation

```
Mean Value:

GRE Score            316.807500
TOEFL Score          107.410000
University Rating      3.087500
SOP                    3.400000
LOR                    3.452500
CGPA                   8.598925
Research               0.547500
Chance of Admit        0.724350
```

Figure 4: Mean of Each Variable

```
Variance Value:

GRE Score            131.644555
TOEFL Score           36.838997
University Rating      1.308114
SOP                    1.013784
LOR                    0.807262
CGPA                   0.355594
Research               0.248365
Chance of Admit        0.020337
```

Figure 5: Variance of Each Variable

Then we can find that GPA (CGPA), GRE Score, and TOEFL Score are the 3 most important factors that will influence the Chance of Admit. So the following analysis I will focus mainly on these three parts of the analysis.

3.2 GPA, GRE, and TOEFL Score Analysis

For the effect of the GPA, GRE Score, and TOEFL Score factors on the Chance of Admit, I used the same method in my study, i.e., using a density plot to show how sparse the data are, and then using a scatter plot to show whether they have a linear equivalence relationship, etc. The details are presented and analyzed as follows:

3.2.1 Influence analysis based on GPA. For the impact of GPA on the result of the admission, this should be the most important. Because each university has a different concept of GPA, the University in the United States generally has 4 points out of 10, but overseas colleges and universities have a 5-point system, 7-point system, 1-point system, etc. Therefore, here according to the conversion of the unified 10 points out of 10, if the score is 4.0/4.0 then it is 10 points here. So it is important to note that if want to predict the probability of admission, you need to convert to 10 points in advance.

3.2.2 Influence analysis based on GRE Score. For the impact of the GRE Score, we can see based on the data that the GRE is the second most important indicator after GPA if you want to get admission to UCLA. It is true that GRE is an indispensable part of graduate school admissions, we can see the distribution of students' GRE scores according to the density figure of the specific GRE scores and see the relationship between the admission rate and the high GRE scores according to the linear analysis figure.

3.2.3 Influence analysis based on TOEFL Score. For the impact of the TOEFL Score, as we can see, the TOEFL test is in third place as an important and significant indicator. Admittedly, as above, I used the same visualization method where we can analyze the specifics of TOEFL scores based on density figure and see the relationship between acceptance rates and high TOEFL scores based on a linear analysis figure

In the above three analyses, they all have some factors that can be easily influenced to cause bias, their analysis methods are similar and they all have a positive impact on the final results, so here the bias will apply to the above points:

1. Measurement error: There may be errors in the calculation of reported GRE scores, TOEFL Score, GPA, or admissions chances, which may introduce noise in the analysis and may weaken the relationship between observed variables
2. Confounding variables: Because only the three most relevant variables are analyzed here, but other factors of admission are also very important. This is because the other variables are not only relevant to the Chance of Admit, they are just as relevant to GRE, TOEFL scores, and GPA scores, for example, Stanford's GPA must not be the same as the GPA of other less reputable institutions.
3. Non-linear relationships: The relationships compared above may not be linear, which may limit the explanatory power of linear regression models. Non-linear models or data transformations may need to be considered in analyses with larger data volumes to better capture the relationships.
4. Scope limitation: The data set may only include applicants with relatively high scores in each category, which may limit the range of scores observed and lead to an underestimation of the relationship between each score and admissions opportunities. This is especially important since most UCLA majors have strict minimum requirements for GRE and TOEFL scores.

3.3 Variable of Research Experience

I then compared the independent variable of whether the student had research experience. There are many people who misunderstand research experience as a decisive factor for admission to top schools, and it is even used by many study abroad application agents as a gimmick to make money, but after a simple comparison, you can see that this thing is actually untrue. It is important to note that the data here does not determine admission or non-admission, so it is not possible to accurately analyze whether research or non-research can affect admission results.

3.4 Variable of University Rating

A student's undergraduate background also plays a very important factor in graduate school admissions. As mentioned above, university rating has a positive effect on all other independent variables, and bias is also mentioned in the section above. Here I use bar charts and box plots to visualize the impact of undergraduate school reputation and quality on students' applications to UCLA.

3.5 GRE and TOFEL Score Relation

This section is a comparison of the GRE and TOEFL Score. The future can be extended into anti-cheating projections as needed because after 2020 with the ETS online test, GRE, and TOEFL scores falsified seriously, this data is not affected. GRE and TOEFL score level gap is too large will be in admissions officers seriously lose authenticity, so as a candidate, also need to be careful not to test if the gap is too large. So, for this section, I used box figure that can reflect both data and gaps as well as scatter plots to visualize.

But this analysis might take bias of Heterogeneous test-taker populations, which means GRE scores apply to all applicants, whereas TOEFL scores are primarily relevant for non-native English speakers. As a result, the populations of test-takers for each exam may differ in systematic ways, potentially introducing bias into the comparison.

3.6 Machine Learning Prediction

To predict the admission probability using ML, I used three methods, namely Linear Regression, Decision Tree and Random Forest models.

3.6.1 Linear Regression. Linear regression is a simple and widely used statistical method for modeling the relationship between a dependent variable (in this case, Chance of Admit) and one or more independent variables. It assumes a linear relationship between the dependent and independent variables and estimates the coefficients of each independent variable to minimize the sum of squared differences between the predicted and actual dependent variable values.

It has the advantage of requiring less computational resources and training time than more complex models while working well when the relationship between the independent and dependent variables is linear. The disadvantages are that the assumed linear relationship between the dependent and independent variables may not always hold and is sensitive to outliers, which can significantly affect the coefficients, while it may perform poorly if there are interactions or nonlinear relationships between the variables, which also could cause bias.

3.6.2 Decision Tree. A decision tree is a nonlinear hierarchical model that recursively divides the data into subsets based on the values of the input features with the goal of minimizing impurities (e.g., Gini index or entropy) in the resulting subset. The final tree structure represents a series of decisions that can be used to predict the target variable.

The advantages are that it allows the modeling of nonlinear and complex relationships between variables and requires minimal data pre-processing, as it can handle categorical variables and missing values. The disadvantages are that it is easy to overfit, especially when the tree is very deep, resulting in poor generalization to new data, and it is prone to instability, as small changes in the data may result in a completely different tree structure.[2]

3.6.3 Random Forest. Random forest is an integrated learning method that builds multiple decision trees and combines their predictions to improve overall model performance. It introduces randomness into the tree construction process by using a random subset of features and bootstrap samples of the training data to build each tree at each split.

It has the advantage of reducing the risk of overfitting compared to a single decision tree, as it averages the predictions of multiple trees while allowing the modeling of nonlinear and complex relationships between variables. The disadvantage is that it is more computationally expensive and slower to train compared to linear regression or individual decision trees.[3] Also may perform poorly in the analysis of high-dimensional, sparse data or data sets.

4 Result

4.1 Top 3 Variances of Admission Result

The following graphs show the relationship between CGPA (GPA) and Chance of Admit, GRE Score and Chance of Admit, and TOEFL Score and Chance of Admit respectively.

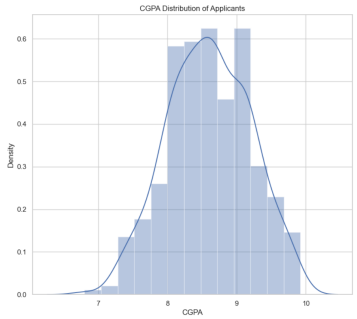

Figure 6: GPA Density

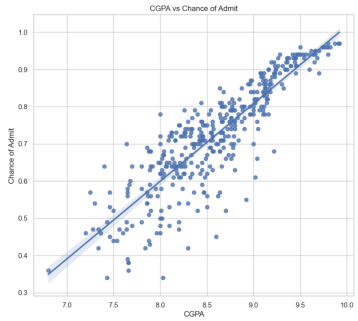

Figure 7: Linear Regression of GPA and Chance of Admit

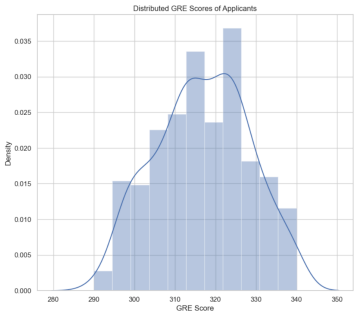

Figure 8: GRE Score Density

Figure 9: Linear Regression of GRE Score and Chance of Admit

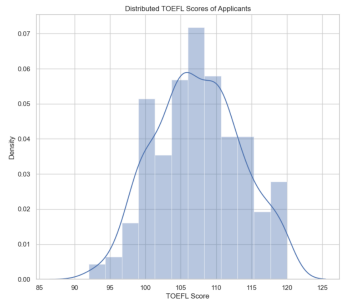

Figure 10: TOEFL Score Density

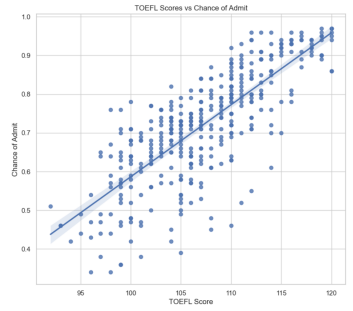

Figure 11: Linear Regression of TOEFL Score and Chance of Admit

We can see from the above graph that GPA has a very strong correlation with admissions chances and is one of the most linearly correlated variables. All three graphs show that the higher the results of these three independent variables, the better the student's chances of being admitted. And the point distribution of "TOEFL Scores" is more spread out, which means that TOEFL is not as important as one might think, if it is compared to the other two.

4.2 Comparison Result of Research Experience

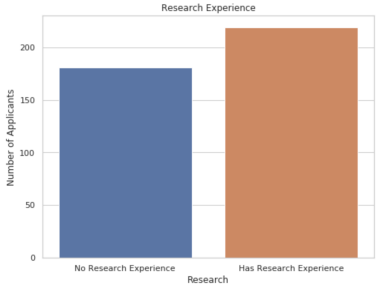

Figure 12: Bar Chart of Research Experiences

We can see that students do not all have research experience, or even almost close to it, so the theory put forward by the agent is not very reliable, while UCLA, as a very good school, is far more difficult to admit than most American universities, so it needs to be analyzed according to the specific application.

4.3 University Rating Result

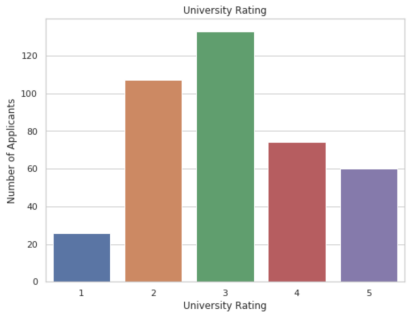

Figure 13: Chart of University Rating

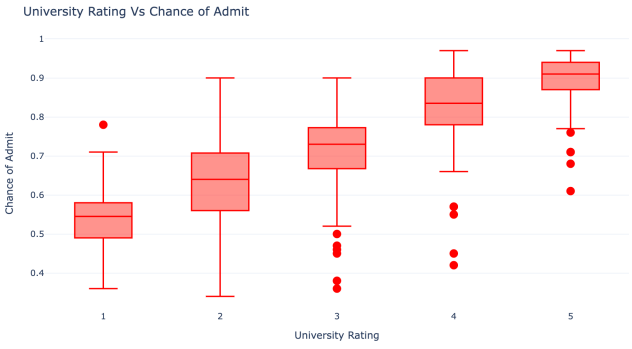

Figure 14: Box Plot of University Rating VS Chance of Admit

As we can see, the applicants' backgrounds are quite diverse. But applicants from universities with significantly higher scores are more likely to get offers and score in the 5 range much higher than in the other categories.

4.4 Result of GRE and TOFEL Score Relation

Figure 15: Box Plot of GRE Score VS TOEFL Score

Similarly, we can see a very strong positive correlation between TOEFL and GRE scores, with applicants who have high GRE scores having higher TOEFL scores, but there are some outliers who have very high TOEFL scores but no good TOEFL scores, which can be understood because the GRE is diverse. There are no cases of low TOEFL scores but close to perfect GRE scores, this data is relatively clean, and it is not possible to analyze the cheating situation, but I believe the data will be very significantly different after 2020.

4.5 Result of Machine Learning Methods

The achieved accuracy scores are 82.1208259148699%, 73.02776364703219%, and 82.03511985595362% respectively. Random Forest or Linear Regression both could be used for predicting admission success because they performed almost the same. But the bias here is obvious, and if the future model has a large enough amount of data, I think it is possible that the Random Forest will have a better performance effect, and it is only necessary to train the data using the Random Forest in the future.

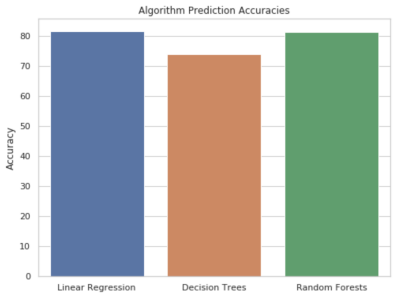

Figure 16: Bar Chart of 3 Methods

5 Conclusion

In summary, the purpose of this study was to analyze the difficulty of admission to the University of California, Los Angeles (UCLA) Graduate School by examining the relationship between various indicators such as GRE scores, TOEFL scores, and other relevant factors. I applied three machine learning models - linear regression, decision tree, and random forest. The results showed that GPA scores, TOEFL scores, and GRE scores were the main factors affecting admission chances. Also, it is recommended to use the random forest model to make similar predictions, whose complex nonlinear approach can better capture the relationship between the input features and the target variables. However, it is important to note potential biases in the dataset, such as selection bias, range restrictions, and confounding factors, which may affect the generalizability of the findings. It is also important to note that the applicant's undergraduate background is not as important compared to research experience. Based on these findings, I recommend that prospective students focus on improving their GPA, GRE, and TOEFL scores to maximize their chances of being accepted into a UCLA graduate program. Additionally, universities can use the insights from this study to fine-tune their admissions criteria and make more informed decisions about applicants.

For future work, I suggest first expanding the diverse dataset to increase the generalizability of the findings. Secondly more advanced machine learning techniques such as support vector machines, neural networks, or gradient boosting could be explored to further improve the predictive performance of the models, but

this is all based on a larger amount of data. By addressing these issues, future research can further refine our understanding of the graduate admissions process and help develop more effective and fair admissions strategies for students and universities.

REFERENCES

[1] Луговий, Володимир, Олена Слюсаренко, and Жанна Таланова. "University rating & development: challenges and opportunities for Ukraine." Education: Modern Discourses 2 (2019): 60-77.

[2] Quinlan, J. Ross. "Learning decision tree classifiers." ACM Computing Surveys (CSUR) 28.1 (1996): 71-72.

[3] Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.