Data Science In The Wild Project Final Report

Shangjing Tang, Henry Wu, Yimin Ou, Alan Michael Hsieh
Cornell Tech
New York, NY, 10044

Abstract

To address the uncertainties students face during the application process, our goal is to develop a reliable model that can help them better gauge their competitiveness and offer guidance for self-improvement. We utilize self-reported data from students applying to psychology programs, which we have gathered from the GradCafe website[1] and matched misspelled or abbreviated school names data with GhatGPT API. We will employ Fbprophet, Multivariate Regression and other analysis to reveal the following insights: (i) trends and patterns concerning shifts in application status, (ii) students' academic performance on each program, and (iii) impacts of students' standard test score on their admission outcomes.

1 Background

1.1 Motivation and Problem Setup

Historically, students have encountered difficulties due to the uncertainty inherent in the application process. To address this, we aim to provide a dependable resource to support future applicants as they navigate upcoming application cycles. Our tool is not intended to predict acceptance decisions but rather to assist users in understanding their relative standing among their peers, drawing from past experiences and offering advice on how to bolster their qualifications.

For this project, we will employ data science, feature analysis and other machine learning methodologies using self-reported student data gathered from the website known as The GradCafe. As a prominent platform for real-time graduate school admissions updates, The GradCafe enables applicants to submit their application statuses and admission results, fostering an interactive space for students to discuss and share their experiences. To gather the necessary data, students must input information such as the institutions they are applying to, program names, degree types, application cycles, demographic details, notifications (rejection, acceptance, interview), notification dates, GRE scores, GPAs, and any additional comments.

In this project, our final curated dataset includes crucial features such as the institution, major, degree, entry term, application decisions, academic performance, and student comments. Our primary focus is on the following key areas:

(i) Examine historical data to identify trends and patterns concerning shifts in application status (interviews, rejections, acceptances) and project these changes for the years 2021, 2022 and 2023.

(ii) Analyze students' GPA and GRE scores in each program, providing insights into the academic performance and aptitude of students.

[1]GradCafe: https://www.thegradcafe.com/

(iii) Perform multivariate regression study based on GPA and GRE Scores and analyze the impact of GPA and GRE scores on admissions outcomes in psychology programs, revealing their significance while considering the varying importance across different institutions.

2 Dataset

2.1 Dataset used and collected

We aim to extract and analyze data related to psychology graduate school applications from The GradCafe, a popular online platform where students voluntarily share their experiences and application information. The data spans from 2006 to 2023 and includes details such as institution, program, degree, entry term, application decisions, academic performance, test scores, and student comments. To achieve this, we use web scraping techniques with the help of Python libraries like Requests and Beautiful Soup.

The data collection process involves web scraping The GradCafe platform for relevant information. We use the Python Requests library to fetch the web pages and Beautiful Soup library to parse the HTML content. The code iterates through 1,484 pages of the platform's search results, containing 40 results per page, to collect the required data.

By using web scraping techniques, this project successfully extracts valuable information related to psychology graduate school applications from The GradCafe platform. The resulting dataset can be used for various analyses, such as identifying trends in application decisions, understanding the impact of different factors on admission, and providing insights for prospective students.

2.2 General information about this dataset

General information about this dataset comprises:

Size: The dataset contains approximately 200,000 data points spanning 21 different majors. Of these, 59088 data points pertain specifically to psychology, which is our main area of interest.

Features included: The dataset consists of key features such as university, program, degree type, entrance semester, decision (acceptance, rejection, or interview), decision date, GPA, GRE Verbal, GRE Quantitative, GRE Writing, GRE Subject, application status, date added, and notes or comments from the applicants.

Biases: Since the data is based on self-reported information from students, it may be subject to potential biases and inaccuracies. Students could unintentionally or intentionally misrepresent certain details, and the dataset might not be completely representative of the entire applicant pool due to self-selection biases.

2.3 Data Processing

During our data processing steps, we noticed that many entries contained misspelled or abbreviated school names. For example, some students entered "CMU" instead of "Carnegie Mellon University" or misspelled it as "Carnegi Melon University". Therefore, we developed a method to accurately match all misspelled and abbreviated names to their corresponding school names.

To do this, we first created a dictionary in Python containing the names of all US universities and colleges. We first used an exact matching method to compare the names in the GradCafe file with the names in our dictionary. Then, we applied a reverse inclusion method to identify any inputted school names that matched those in our dictionary.

Matching abbreviations such as "CMU" proved challenging, as it is difficult to determine whether "PU" refers to Purdue University or Princeton University. After the steps above, we were only able to match around 65% of the data. In order to improve the matching result, we also use ChatGPT API to detect the name and match with our established dictionary. After openAI detection, we were able to accurately match approximately 77% of the data.

3 Analysis & Result

3.1 Predicting number of application status changes per day

3.1.1 Model Setup: Prophet Univariate Time Series Forecasting Algorithm

To analyze the historical trends in application status changes and predict future timelines, we used the Prophet univariate time series forecasting algorithm developed by Facebook[2]. Prophet is a forecasting tool that fits non-linear trends in time series data using yearly, weekly, and daily seasonality, along with holiday effects. We chose this algorithm because it accommodates multiple seasonality. We assumed that each post on GradCafe is submitted by a unique user.

In our dataset, users posted their application status with the decision date and in the following types: interview, accepted, rejected, waitlist, and others. To analyze the historical trend, we count the number of status updates per date from 2006 to 2020, fit the model with each type of status update and make predictions. Notes that we did not analyze the classes: "waitlist" and "other" because of a lack of data points.

3.1.2 Prediction of number of application status changes

In this section, we show the prediction results of the number of status changes for the following types of application status: all, interview, accepted, and rejected.

Table 1 shows the MSE of each type of application status. Figures 1, 2, 3, and 4 compare the ground truth and model predictions of different types of application status, including all, interview, accepted, and rejected.

For the slack season, the model makes a prediction that is obviously wrong because the dataset lacks data points in the slack season. Hence, in the following analysis, we perform interpolation to our dataset so that all dates in the time series have value either from the dataset or interpolated. We expected better prediction accuracy in the slack season with interpolation.

For the peak season, the predicted value is roughly the mean of actual values. In other words, the model cannot fit the extremely high-frequency change of the number of status changes in peak season. To fix the issue, in the following analysis, we calculate five days mean (i.e., $y(t) = (x(t) + x(t\text{-}1) + \ldots + x(t\text{-}4)) / 5$) of the original dataset and use the five days mean to fit the model. We expect that we can successfully make prediction to five days mean dataset.

Application status	All	Interview	Accepted	Rejected
Mean square error	654.02	141.54	46.63	99.32

Table 1: Mean Square error of the prediction

[2]Prophet: https://facebook.github.io/prophet/

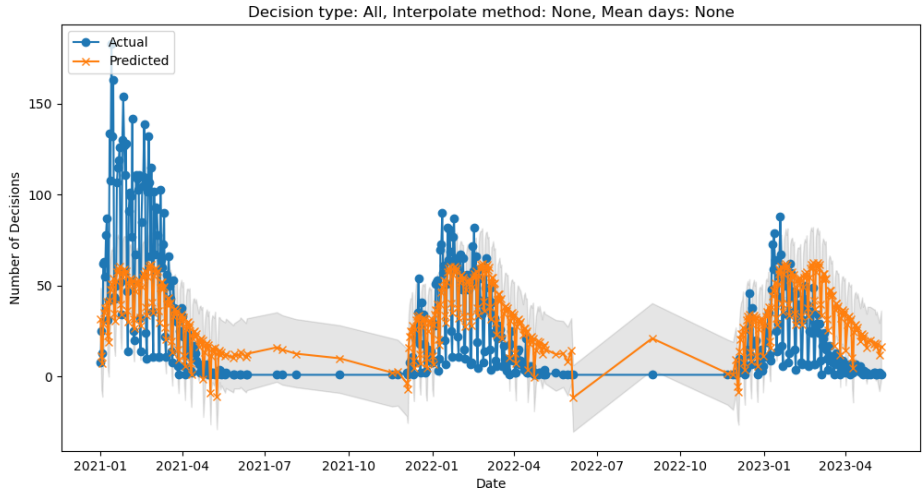

Figure 1: Prediction versus actual number for all type of decisions

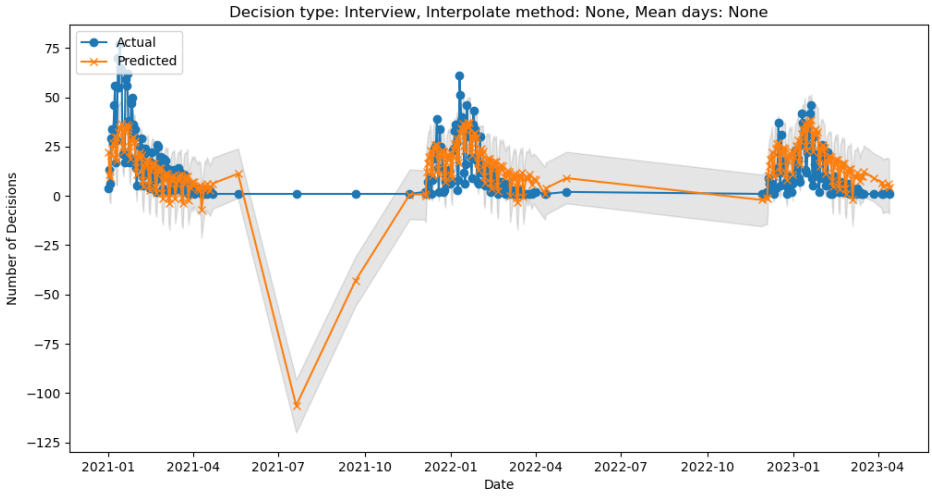

Figure 2: Prediction versus actual number for 'Interview' decisions

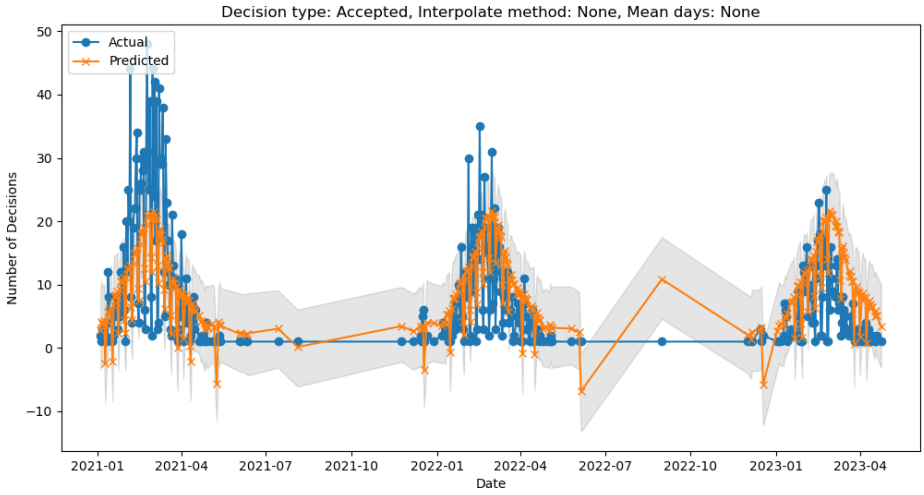

Figure 3: Prediction versus actual number for 'Accepted' decisions

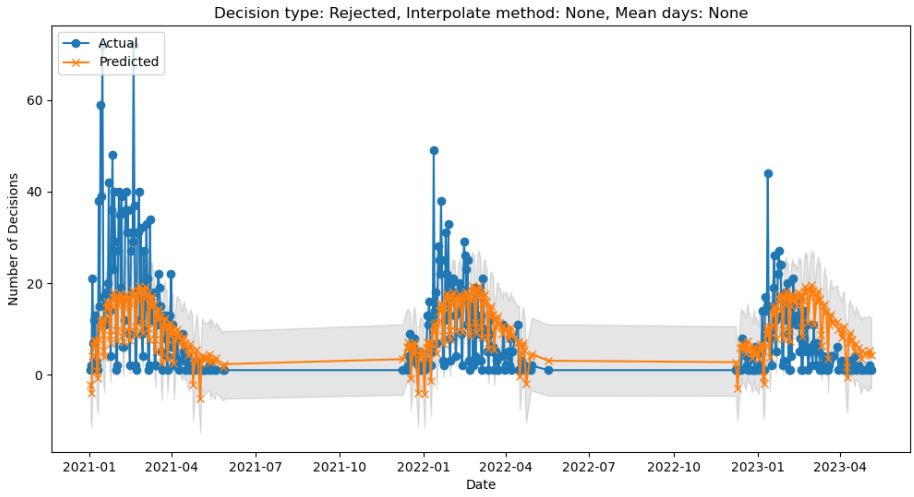

Figure 4: Prediction versus actual number for 'Rejected' decisions

3.1.3 Prediction with different interpolation methods

In this part, we show the prediction results of the number of status changes by performing the following interpolation methods to the test dataset: linear, backfill, and quadratic, and compare it to the prediction without interpolation.

Table 2 compares the MSE of different interpolation methods. Figures 5 shows the original prediction result, and 6, 7 and 8 offer the result by linear, backfill, and quadratic interpolation, respectively.

After performing each interpolation method, the mean square error is much higher than the original result. And all interpolation methods give nearly the same prediction result. However, by comparing the figure of the original prediction and the prediction after interpolation, we

observed that the prediction for slack seasons become much more accurate after interpolation. Therefore, we can conclude that decreasing the prediction accuracy of peak season is a tradeoff for better prediction accuracy during the slack season for the model. And we believed that it is necessary to perform interpolation for the dataset to avoid prediction errors in the slack seasons, and we interpolated data with the backfill method in the next section because it is the most straightforward approach.

Interpolation method	No interpolation	linear	backfill	quadratic
Mean square error	654.02	777.46	777.83	772.53

Table 2: Mean Square error of different interpolation methods

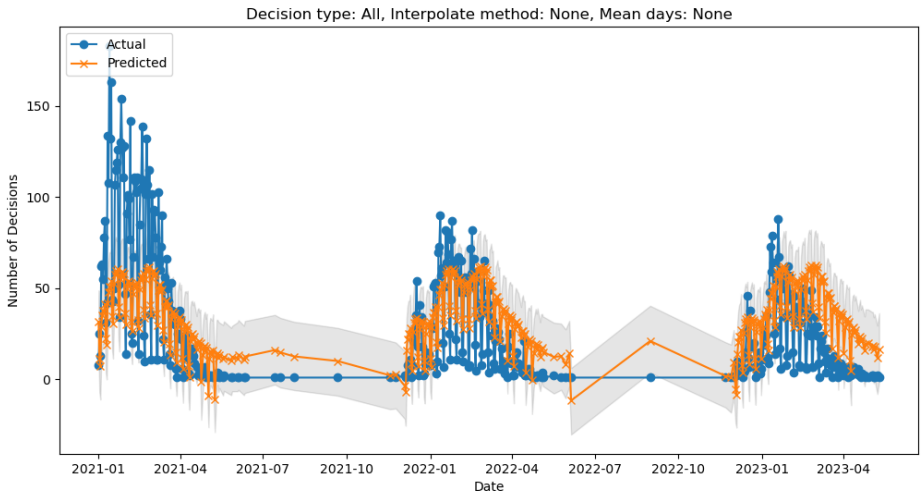

Figure 5: Original prediction result

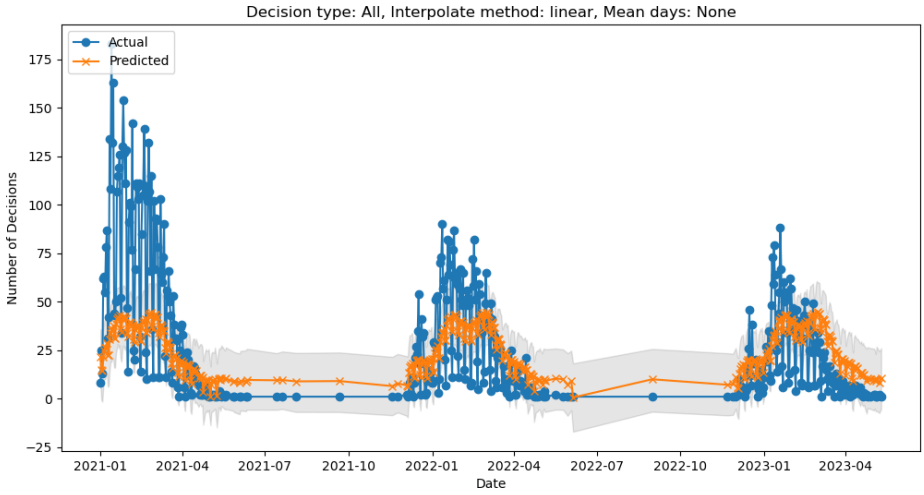

Figure 6: Prediction result with linear interpolation

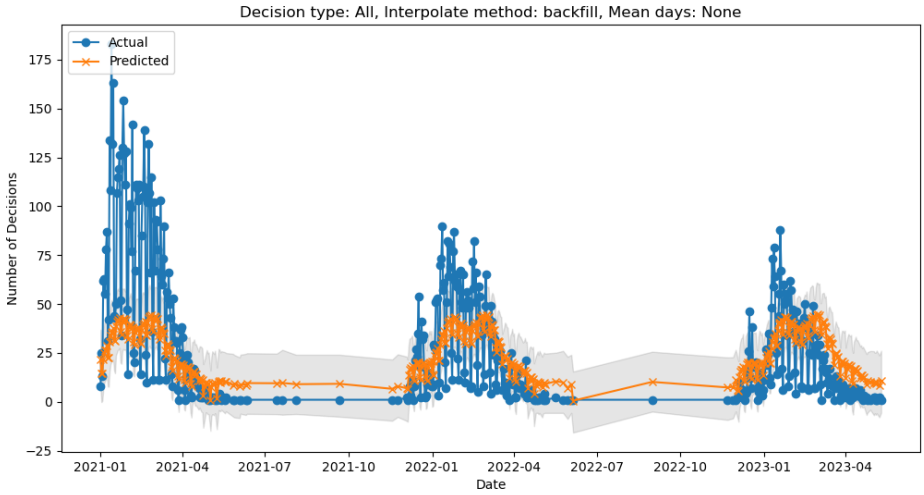

Figure 7: Prediction result with backfill interpolation

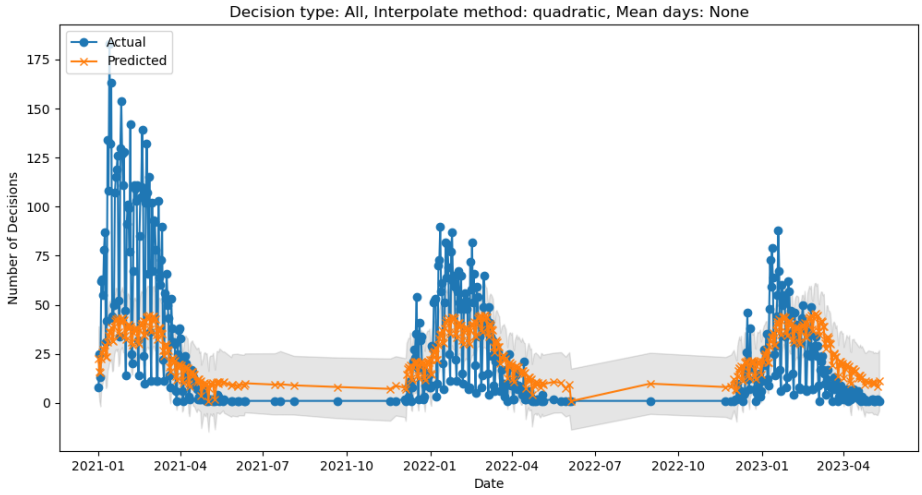

Figure 8: Prediction result with quadratic interpolation

3.1.4 Prediction with taking five days mean.

In this section, we make a prediction of the number of status changes per day that taking a five-day mean (i.e., y(t) = (x(t) + x(t-1) + ... + x(t-4)) / 5) and compare it with the prediction that do not taking a five-day mean. Notes that we apply backfill interpolation to both datasets before fitting the model. Table 3 compares the MSE of prediction with and without a five-day mean. Figure 9 shows the prediction with a five-day mean, and Figure 10 shows the prediction without a five-day mean. The table shows that the MSE decrease remarkably if taking a five-day mean. By comparing the prediction curve, we observed that the prediction is accurate from 2022 to 2023. However, it still cannot model the unprecedented large number of decisions in 2021.

	Five-day mean	No five-day mean
Mean square error	487.55	777.83

Table 3: MSE of prediction with and without a five-day mean

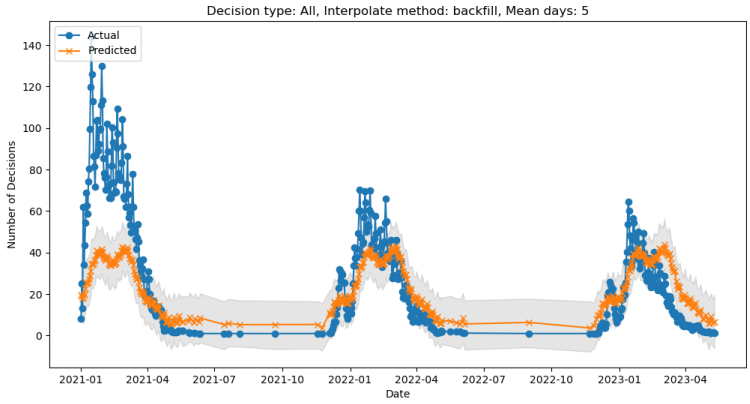

Figure 9: Prediction with a five-day mean

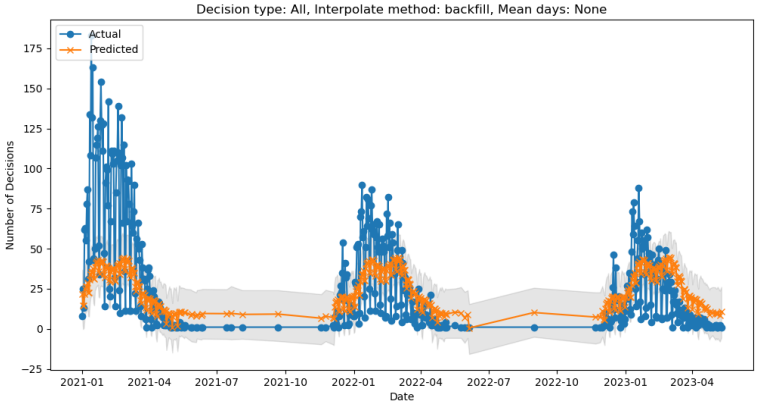

Figure 10: Prediction without a five-day mean

3.2 GPA and GRE Analysis

We have analyzed a dataset containing the GPA and GRE scores of accepted applicants in psychology programs across different institutions. Our analysis aims to classify schools based on the average GPA and GRE scores, shedding light on the academic performance and standardized test scores of students attending these schools.

The classification criteria are as follows:

Great Schools: Schools with a high average GPA (≥ 3.8) and high average GRE scores (Verbal and Quantitative ≥ 160).

Mid-level Schools: Schools with a moderate average GPA (≥ 3.5) and moderate average GRE scores (Verbal and Quantitative ≥ 155).

Worse Schools: Schools that do not fit into the 'Great' or 'Mid-level' categories.

Before classification, we remove data points where the school has less than 10 entries and where the GPA is greater than 4. This step ensures that our analysis is based on reliable and accurate data, eliminating any outliers or schools with insufficient data for a meaningful comparison.

After processing the dataset and applying the specified thresholds, our analysis shows the following results:

- Number of Great Schools: 22
- Number of Mid-level Schools: 86
- Number of Worse Schools: 28

Using GPA and GRE scores as the primary factors for classification can be a reasonable approach to differentiate schools based on students' academic performance. The classification criteria are as follows:

Great Schools: Schools with a high average GPA (≥ 3.8) and high average GRE scores (Verbal and Quantitative ≥ 160). Examples include Harvard University, Stanford University, and Yale University. These schools typically have students with outstanding academic achievements and strong aptitude, reflecting their rigorous admission standards and competitiveness.

Mid-level Schools: Schools with a moderate average GPA (≥ 3.5) and moderate average GRE scores (Verbal and Quantitative ≥ 155). Examples include the University of Michigan, University of Virginia, and University of Florida. These schools still provide excellent education and opportunities while being somewhat more accessible to a broader range of students.

Worse Schools: Schools that do not fit into the 'Great' or 'Mid-level' categories, often with lower average GPA and GRE scores. Examples include Eastern Michigan University, California State University - Fresno, and University of Southern Mississippi. However, this does not necessarily mean that these schools offer a subpar education. They might cater to students with different strengths and learning experiences or focus on other aspects of education beyond standardized test scores and GPA.

GPA is a widely-used metric to evaluate students' academic achievements, while GRE scores provide a standardized measure of their aptitude in various fields. Therefore, a combination of these two factors can offer a comprehensive understanding of students' capabilities.

3.3 Multivariate Regression Study Based on GPA and GRE Scores

To prepare students for the specific school they want to be closer to the needs of the school, in this analysis, we conducted a multivariate regression study to explore the relationship between GPA, GRE scores, and admissions outcomes in psychology programs across different universities. The aim was to identify the extent to which GPA and GRE scores impact the decision of acceptance or rejection.

We began by removing missing values and focusing on the relevant variables, namely GPA, GRE Verbal, GRE Writing, and GRE Quantitative scores. We transformed the Decision variable into a binary outcome, with "Accepted" coded as 1 and "Rejected" coded as 0.

And, we performed a multivariate regression analysis using the Ordinary Least Squares (OLS) method. The regression model included GPA, GRE Verbal, GRE Writing, and GRE Quantitative scores as independent variables and the binary admission decision as the dependent variable. We also added a constant term to the model for more accurate estimates.

The results of the regression analysis provided valuable insights into the impact of GPA and GRE scores on admissions decisions. We examined the coefficients of the independent variables to determine their significance. A positive coefficient indicated that an increase in the variable was associated with a higher likelihood of acceptance, while a negative coefficient indicated the opposite.

Based on the analysis, we found that both GPA and GRE scores had a statistically significant impact on admissions outcomes. A higher GPA and better GRE scores were associated with a higher likelihood of acceptance. However, the magnitude of the effect varied across institutions.

At most colleges, GPA is the most influential characteristic, suggesting that at these schools, academic performance has the greatest impact on admission likelihood. For some schools (e.g., Chicago School of Professional Psychology, Wright Institute, etc.), the most influential

feature is GRE writing scores, which may reflect the emphasis these schools place on students' writing and critical thinking skills.

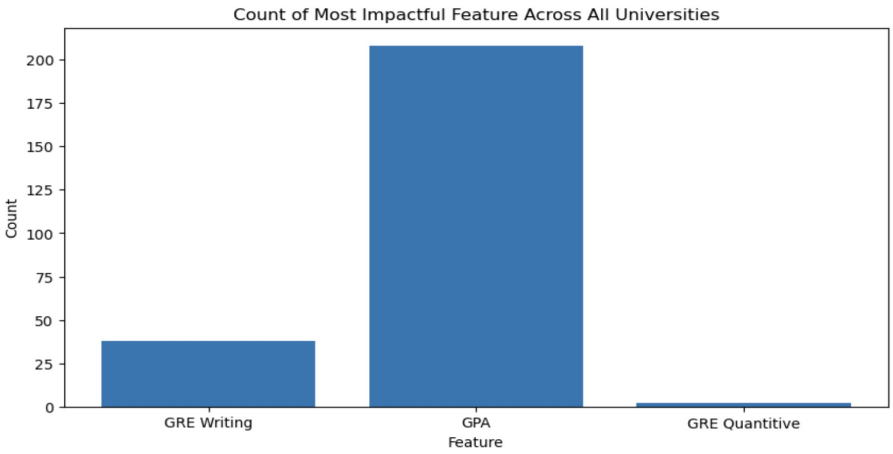

Figure 11: Most important feature on admission by schools

One interesting finding is that GPA is the most impact feature for Ivy League admissions, but its influence varies. Princeton, Brown, and UPenn place greater emphasis on GPA, while Columbia, Harvard, and Yale weigh other aspects more heavily. Cornell falls in the middle. Students should focus on both academics and other application elements to strengthen their profiles.

```
University: brown university, Max Impact Feature: GPA, Value: 0.6322984973860563
University: columbia university, Max Impact Feature: GPA, Value: 0.05185894733811572
University: cornell university, Max Impact Feature: GPA, Value: 0.2272191387764071
University: harvard university, Max Impact Feature: GPA, Value: 0.06668720098732867
University: university of pennsylvania, Max Impact Feature: GPA, Value: 0.5143702583472057
University: princeton university, Max Impact Feature: GPA, Value: 1.0619335769061808
University: yale university, Max Impact Feature: GPA, Value: 0.06969870965659354
```

Figure 12: Ivy Leagues results

Another interesting part is that the values of each feature vary considerably across universities. This reflect the admissions criteria and preferences of different schools. For example, some schools may put more emphasis on academic performance, while others may place more weight on test scores or other factors. In particular, the University of Alabama at Birmingham has a GPA impact score of 25, far exceeding other schools, which indicate that the school places a particularly high emphasis on academic achievement "among candidiates".

But, these result only provide a perspective on characteristics and do not fully represent the admissions criteria of each university. In fact, the college admissions process usually considers a variety of factors, like a student's extracurricular activities, letters of recommendation, application essays, and more. Thus, while these data provide valuable insights, they cannot be used alone to determine the likelihood of admission.

3.4 Conclusion

3.4.1 Scraping and processing data

Among the 21 different majors data we scrapped from GradCafe, we decided to narrow our scope to focus on the Psychology major which contains 59,088 data points. Since our major concern about the dataset is the misspelling and the abbreviation, we match the universities given by students in 'gradcafe' to a standardized list of all universities. We are using several strategies to match the names, including direct comparison, substring inclusion, and abbreviation matching. However, for many abbreviation and misspelling that are too different from our standardized list, the above methods won't help us match the name and yields a initial 65% matching rate for our data. We then decided to use the ChatGPT API to further predict the original school name given a specific prompt and restrictions. In the end, we were able to match around 77% of data from our dataset.

3.4.2 Predicting status changes per day

The model can fit the seasonal change and make a prediction. However, the model cannot handle the rapid value change in peak season, and a significant error occurs while modeling the slack seasons.

We can improve accuracy in the slack seasons by interpolating the missing value in the dataset, but the overall mean square error also increases.

Though the model cannot make accurate prediction to the peak season with high frequency change, the model is capable to make prediction to the dataset taken five days mean.

In the future, we can explore better metrics to evaluate the prediction accuracy to replace the mean square error metric.

3.4.3 GPA and GRE Analysis

In summary, our analysis of the dataset allowed us to classify psychology programs into three categories based on average GPA and GRE scores: Great Schools, Mid-level Schools, and Worse Schools. These classifications provide insights into students' academic performance and aptitude. The combination of GPA and GRE scores offers a comprehensive understanding of students' capabilities, making it a reasonable approach for differentiating schools.

3.4.4 Multivariate Regression Study Based on GPA and GRE Scores

Our study on multivariate regression offers insights into the correlation between GPA, GRE scores, and admission outcomes in psychology programs. It emphasizes the significance of academic performance and standardized test scores in the admissions process, while also acknowledging the impact of other factors in the decision-making process.

Contributions

Data scraping: Henry Wu, Alan Michael Hsieh

Data processing: Shangjing Tang

Predict number of application status changes per day: Henry Wu

GPA and GRE Analysis: Alan Michael Hsieh, Yimin Ou

Multivariate Regression Study Based on GPA and GRE Scores: Yimin Ou

References

Github link: https://github.com/hsinyuwu576/dsw_final/tree/main

"Univariate Time Series Using Facebook Prophet." Section, https://www.section.io/engineering-education/univariate-time-series-using-facebook-prophet/.