Graduate Admission chances in America

Nicholas Gjuraj*
Northeastern University

Xiaoli Fang†
Northeastern University

Abstract

This project uses data to predict the likelihood of acceptance for graduate admissions. The study uses a dataset of previous applicants to graduate programs and applies a range of machine learning algorithms to model the factors most strongly associated with successful admission outcomes. This dataset forms the project's visualization. The metrics from these algorithms show us several critical factors, including undergraduate GPA, GRE scores, letters of recommendation, and so on, are highly predictive of admission outcomes. Furthermore, they show that certain combinations of these factors are particularly effective in predicting successful admission. These findings have important implications for graduate admissions committees, as they suggest that data-driven approaches can help to improve the accuracy and fairness of admissions decisions. Visualization tools will support two domain tasks: identifying the most promising candidates and understanding factors that affect their chances of admission. This tool will allow the committee to efficiently analyze candidate profiles and make informed decisions about admissions.

1 Introduction

The graduate school admissions process is a complex and time-consuming task in which members of a board evaluate a large number of candidate profiles. Admissions committees need to quickly identify the most promising candidates and understand the factors that influence their admissions decisions. The proposed visualization tool will allow committees to efficiently analyze candidate profiles and make informed decisions about admissions, as well as show trends and correlations among applicants to help them better understand the data and trends they are seeing. The tool will be designed for admissions committee members with varying degrees of familiarity with data analysis and visualization. The data that will be visualized includes; GRE scores (out of 340), TOEFL scores (out of 120), college grades (out of 5), Statement of Purpose and Letter of Recommendation Strength (out of 5), undergraduate GPA (out of 10), research Experience (0 or 1) and Admission Chance (from 0 to 1).

2 Related Work

Brink's *Selecting Graduate Student* [1] is less of a mathematical paper and more of an investigative and summary piece about reaching out to graduate schools to see what they do in order to select their students, and then breaking down those results into different categories to better understand the processes that these schools are using. Rather than drawing any conclusions and extrapolating the research himself, we are left with the ability to take these concepts presented to us to gain an understanding of what the other side is looking for specifically and use that to fine tune any visualization or model we create in order to better match up with what graduate schools might be more interested in to get a potentially better expression of our goals.

Attiyeh's *Testing for Bias in Graduate School Admissions* [2] is a study upon many graduate school applicants divided up by filed of study, broken down into individual characteristics in the same way our dataset is, and then applied weighted coefficients to all characteristics in order to predict whether or not an applicant gets into a graduate program in that particular field of study. There are no visualizations in this research paper, but there are many good ideas. There are many ways we can incorporate such concepts into our own project, like creating any kind of model to predict the outcome of a student, and then make some kind of visualization on that. It is filled with great ideas that can contribute invaluably to this project.

3 Use Case

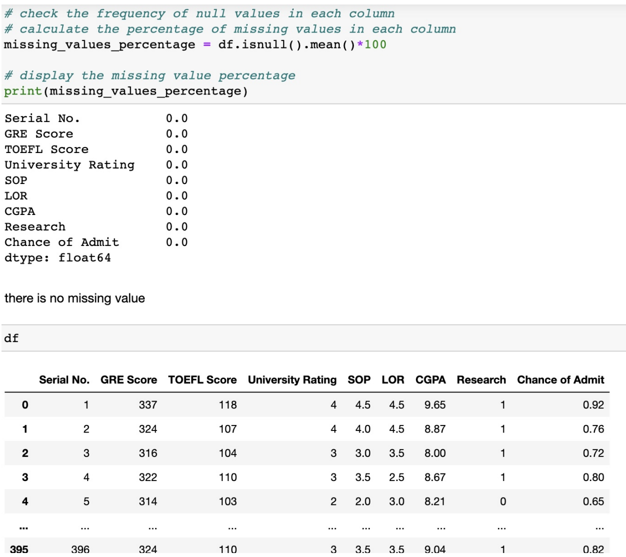
This visualization's intended users are members on a graduate admissions committee–we can expect them to be relatively well versed with technology because of their positions, though some more so than others. With this scenario in mind, we have two major use cases; trying to understand the relationships between certain metrics, and trying to predict scores based on other scores. A committee member can explore, say, the last 500 candidates and look through their scores and whether or not they decided to admit them, and using this visualization, understand what factors led them to choose the decision they made, which in turn allows them to make more efficient decisions when they know what scores they should be looking for. The second case arises in the scenario where a student is missing a score that might be considered pivitol in the decision making process. Based on their other scores, and the correlation between every other score they provided and the missing score, we can gather a sum of potential scores of the missing test based on the correlations between that test and the others, and average them together to get a potentially accurate reading on what that student might have scored on that test, and factor that into their chance of being admitted.

4 Data

The "Graduate Admissions" dataset was collected by Mohan S Acharya and is available on Kaggle at https://www.kaggle.com/mohansacharya/graduate-admissions.
There are no apparent biases or ethical considerations related to the dataset. The missing values in the "TOEFL Score" and "University Rating" attributes were replaced with the median and mode values, respectively. The outliers in the "GRE Scores" and "GPA" attributes were capped at the upper end of the interquartile range.
This dataset contains 500 rows and eight columns of data, including attributes such as GRE Scores, TOEFL Scores, university ratings, statement of purpose strength, letters of recommendation strength, GPA, research experience, and the chance of admission to a particular university.
Mohan S Acharya collected the dataset to predict the chances of admission for a student applying for a master's program. This dataset has no apparent bias or ethical considerations–the data is collected completely indiscriminately and reflects a wide array of students. However, it is important to note that there may be bias or ethical considerations in the dataset, even if they are not immediately apparent, so it is always important to thoroughly evaluate and explore the data to identify any potential issues. This dataset is inspired by the UCLA graduate student dataset. Test scores and GPA are in an old format so we need to be careful when

*e-mail: gjuraj.n@northeastern.edu
†e-mail: fang.xiaol@northeastern.edu

handling the information. During exploration, some missing values were identified in the ”TOEFL Score” and ”University Rating” attributes. Outliers were also identified in the ”GRE Scores,” and ”GPA” attributes. After downloading the dataset, we made two files, both csvs. One of them is the original one, and the other is the cleaned data.



```
# check the frequency of null values in each column
# calculate the percentage of missing values in each column
missing_values_percentage = df.isnull().mean()*100

# display the missing value percentage
print(missing_values_percentage)
```

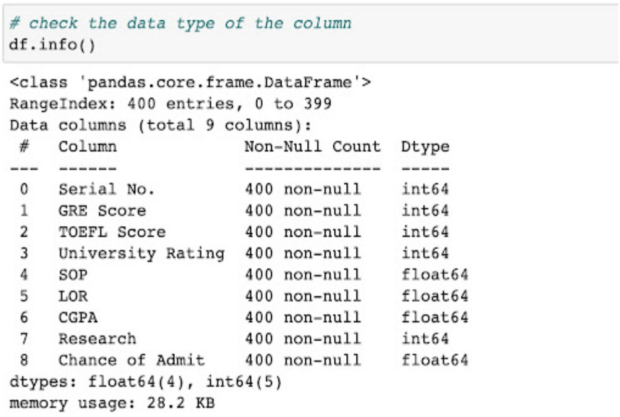
```
Serial No.           0.0
GRE Score            0.0
TOEFL Score          0.0
University Rating    0.0
SOP                  0.0
LOR                  0.0
CGPA                 0.0
Research             0.0
Chance of Admit      0.0
dtype: float64
```

there is no missing value

```
df
```

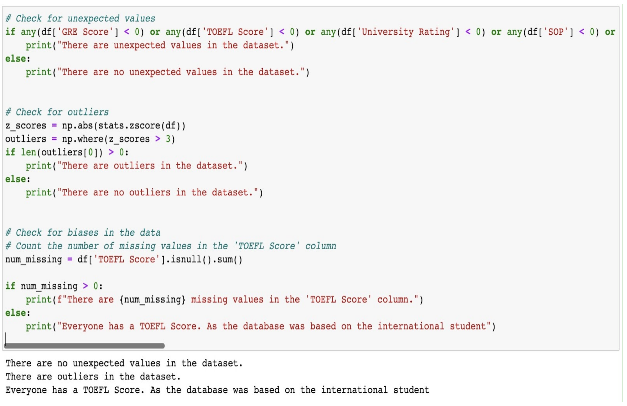
	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
0	1	337	118	4	4.5	4.5	9.65	1	0.92
1	2	324	107	4	4.0	4.5	8.87	1	0.76
2	3	316	104	3	3.0	3.5	8.00	1	0.72
3	4	322	110	3	3.5	2.5	8.67	1	0.80
4	5	314	103	2	2.0	3.0	8.21	0	0.65
...	...	...	...	...	...	...	...	...	...
395	396	324	110	3	3.5	3.5	9.04	1	0.82

Used jupyter notebook check for the cleaned and missing data.



```
# check the data type of the column
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Serial No.         400 non-null    int64
 1   GRE Score          400 non-null    int64
 2   TOEFL Score        400 non-null    int64
 3   University Rating  400 non-null    int64
 4   SOP                400 non-null    float64
 5   LOR                400 non-null    float64
 6   CGPA               400 non-null    float64
 7   Research           400 non-null    int64
 8   Chance of Admit    400 non-null    float64
dtypes: float64(4), int64(5)
memory usage: 28.2 KB
```



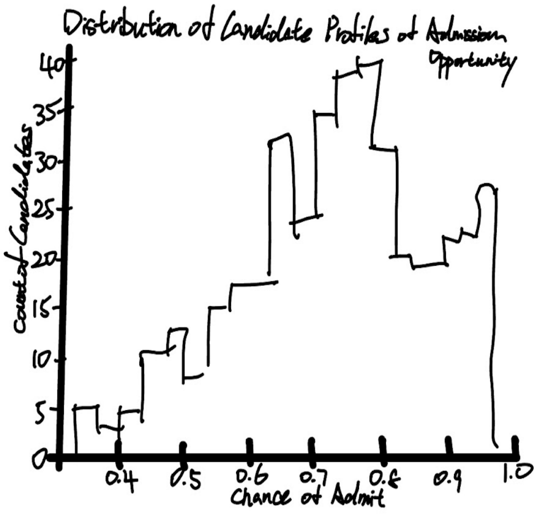
```
# Check for unexpected values
if any(df['GRE Score'] < 0) or any(df['TOEFL Score'] < 0) or any(df['University Rating'] < 0) or any(df['SOP'] < 0) or
    print("There are unexpected values in the dataset.")
else:
    print("There are no unexpected values in the dataset.")


# Check for outliers
z_scores = np.abs(stats.zscore(df))
outliers = np.where(z_scores > 3)
if len(outliers[0]) > 0:
    print("There are outliers in the dataset.")
else:
    print("There are no outliers in the dataset.")


# Check for biases in the data
# Count the number of missing values in the 'TOEFL Score' column
num_missing = df['TOEFL Score'].isnull().sum()

if num_missing > 0:
    print(f"There are {num_missing} missing values in the 'TOEFL Score' column.")
else:
    print("Everyone has a TOEFL Score. As the database was based on the international student")
```

```
There are no unexpected values in the dataset.
There are outliers in the dataset.
Everyone has a TOEFL Score. As the database was based on the international student
```

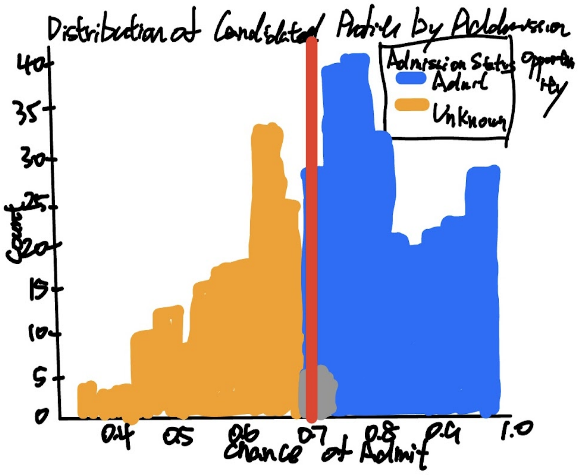
Unexpected values, Outliers checked.

5 Design Process

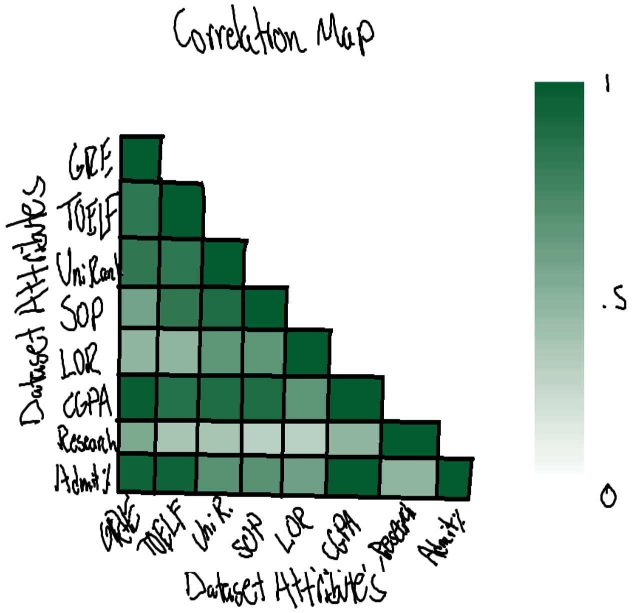
The first sketch was a rough sketch created during the ideation phase. It shows a basic histogram with a single quantitative variable on the x-axis and the count of candidates on the y-axis. The purpose of this sketch was to explore the initial idea of visualizing the distribution of candidate profiles by admission opportunity.



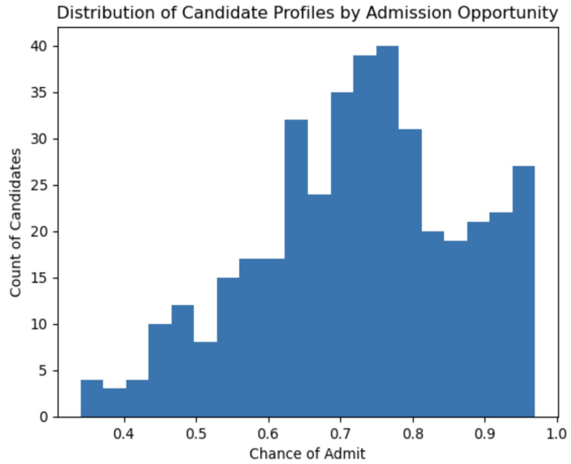
The second sketch is partial and contains more detail than the first. It added color coding to distinguish admitted candidates from rejected candidates and a threshold line representing the lowest chance of admission required for acceptance. The purpose of this sketch is to refine the initial idea further and explore possible visual encodings that can be used to enhance the visualization.



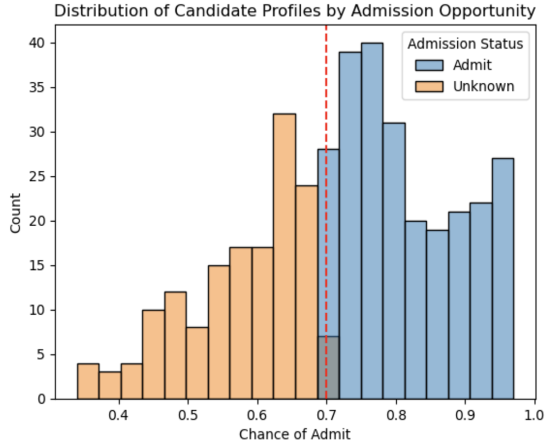
Another premise we figured was a good idea to look into was the correlations between our data. Was it more likely for a student who scored well in the TOEFL to also have research done? And, how does that correlate to their chance of being accepted? These types of questions pushed us to create a completely different, but also interesting visualization.



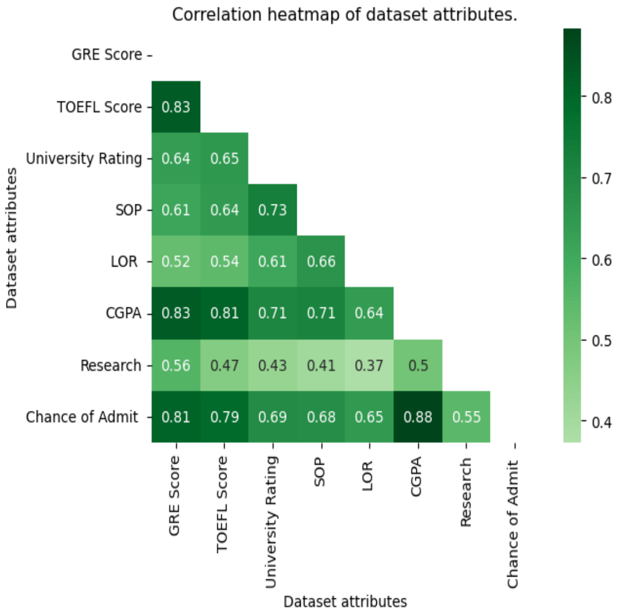
Now, we start to make it more professional, incorporating the same ideas and concepts of the first histogram.



Adding in color channel and new data marks in order to further explain our data as described in the second sketch, only now far more refined.



Now, with the alternate concept of correlations, we took the sketch and plotted out, and this time included numbers to get a better sense of how positively and negatively attributes were correlated.



The design process begins with identifying the motivation behind the project and providing admissions committees with visualization tools to help them make informed admissions decisions. The tool will help evaluate many candidate profiles and identify the most promising candidates based on attributes such as GRE scores, TOEFL scores, college grades, statement of purpose and strength of recommendation letters, undergraduate GPA, research experience, and chances of admission. Throughout the implementation of the Graduate School Admissions Visualization Tool, the design evolved based on iterative feedback and usability testing. The project relies on publicly available data from reputable sources and identifies and addresses any potential bias or ethical implications associated with the data. The team also takes steps to protect the privacy of individuals contained in the data, such as aggregating or anonymizing data where appropriate.

The initial design of the visualization tool consisted of a bar graph and admission numbers representing the admission chances of candidates based on their profiles and scatterplots of different database attributes like GPA and test scores and filtering options based on factors like research experience and university rank. In the usability test, participants were found to need help understanding the relationship between the scatter plot and the chances of admission represented in the bar chart. The final visualization design evolved from the identified attributes of the candidate profiles, which were used as variables to be plotted on the scatterplot. The original design consisted of a scatterplot where the x- and y-axes were not fixed, showing different attributes.

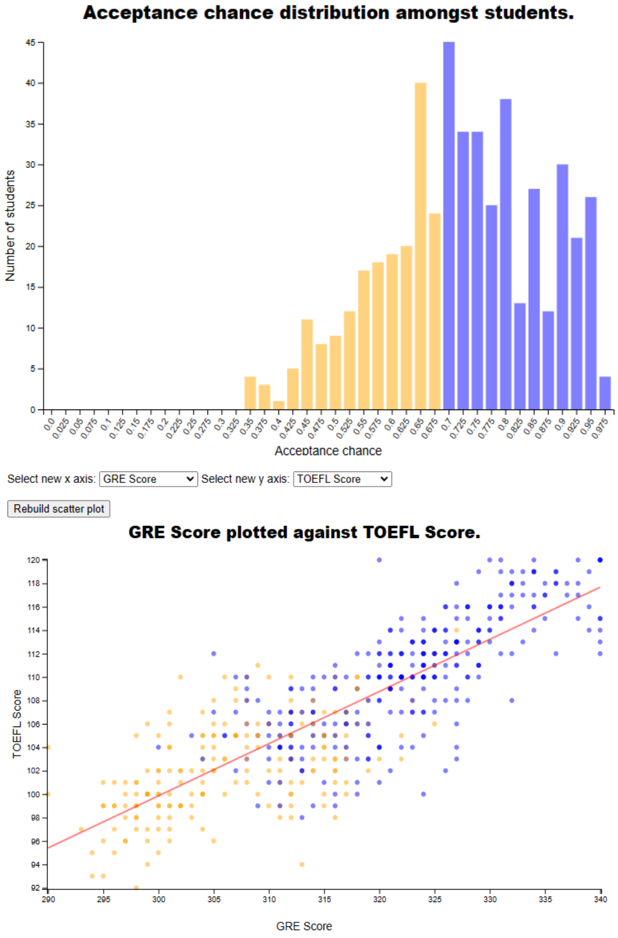
We planned to use a heatmap to show the correlation between variables and the likelihood of admission to graduate school. Ultimately, we gave up this solution and changed it to the scatterplot because it was difficult to determine the correlation between the bar chart and it. The professor suggested that we add a tooltip for the two graphs. However, we used scatterplots where users can change the axis to read different data to explore correlations themselves, making it more interactive.

In order to make the visualization more interactive and adaptable to different needs, a feature has been added that allows the user to select the variables plotted on the scatterplot x and y axes. A form is added to the HTML code with two drop-down menus for selecting x and y variables to enable this functionality. The options for each menu are populated based on the attributes available in the dataset. The form also contains a button that, when clicked, will rebuild the scatterplot with the new variables selected. Additionally, participants found it challenging to compare candidate profiles based on criteria such as GPA, test scores, or chances of admission. A comparison feature has been added that allows admissions committees to compare the profiles of admitted and rejected candidates to determine their differentiators.

JavaScript code handles the construction of the scatterplot and updates it based on user selections. The scatterplot was created using the D3.js library, a popular library for creating data visualizations. A scatter plot shows selected x and y variables on the axes and represents each candidate profile as a circle on the plot. The size and color of the circles can be customized to represent other attributes, such as admissions opportunities or research experience.

Overall, usability testing played an essential role in the evolution of the final visual design. Iterative feedback enabled the team to identify and address usability issues. The resulting visualization design evolved from a static scatterplot to an interactive tool that allowed users to select variables to be plotted on the scatterplot. The driving force for this evolution was making the tool more adaptable to different needs. The design process for the final visualization evolved from an initial concept into a functional and interactive tool that helps admissions committees effectively analyze candidate profiles.

6 Final Design



The final visualization consists of an interactive scatterplot showing candidate profiles and their chances of admission based on various attributes such as GRE scores, TOEFL scores, college grades, statement of purpose, and strength of recommendation letters. Scatterplots can be customized to display different attributes on the x and y axes, and users can select the variables they want to plot using a form with drop-down menus. The circles on the scatter plot represent each candidate's profile, and their size and color can be customized to represent other attributes, such as admissions opportunities or research experience. The bar chart shows the admission probability and the number of students admitted in previous years. By moving the mouse left and right, you can correspond to their data on the scatterplot.

To address the domain of assessing graduate admissions candidate profiles, the target user (in this case, the admissions committee) would use a tool like this:

1. They will access the tool through the link provided and navigate to the scatterplot page.

2. They will use the drop-down menus on the form to select the attributes they want to be plotted on the x-axis and y-axis of the scatterplot. For example, they might plot GRE scores on the x-axis and college grades on the y-axis. After selecting students with less than 50 percent admissions on the bar chart, the scatter plot will work, marking GRE and GPA (in the scatterplot, it has stated as CGPA)points of those rejected students. By placing the mouse over that point, we can also know other performance data of that student for the admissions committee to consider.

3. They will click the Rebuild Scatterplot button to update the scatterplot with the new variables.

4. They will analyze the scatter plots to identify the most promising candidates based on the chance of admission and other attributes such as research experience or strength of letters of recommendation.

5. They can further analyze the scatterplot by representing other attributes with a long number axis on the bar chart and using the compare function to compare candidate profiles.

Overall, the tool allows admissions committees to quickly compare and contrast candidate profiles based on their chances of admission and other attributes, making the admissions process more informed and efficient.

7 Discussion

Reflections on the final visualizer:

Our final visualizer is a dashboard containing several interactive visualizations. It allows users to explore and compare candidate profiles based on their admission opportunities and the factors that affect those opportunities. The tool also summarizes the candidate pool and highlights the most promising candidates. Overall, the tool is user-friendly and provides valuable insight into the admissions process.

Does the tool completely solve the domain problem?

Yes, the tool solves exactly the domain we intend to solve. It provides a visualization tool that helps committees quickly identify a school's top candidates and understand the factors influencing their admissions decisions.

Limitations of our work:

A limitation of our work is that chances of admission are based only on selected factors in our dataset. Other factors may affect the admission decision, such as whether you have participated in and won awards in major international competitions. These factors were not considered in our dataset. Additionally, the dataset is limited to many candidates and schools and may not represent the broader population.

Future improvements:

A potential future improvement would be to expand the dataset to include more candidates and schools than just collecting information on the Kaggle website, which will provide a more representative sample and increase the accuracy of your chances of admission. Another improvement includes other factors that influence admissions decisions, such as work experience. Finally, we can incorporate machine learning techniques to predict admissions opportunities based on a broader range of factors and provide more accurate predictions.

8 Usability Testing

To conduct a usability test of our graduate admissions visualization tool, we designed an experiment in which participants were asked to use the tool to complete specific tasks. At the same time, we collected data on their interactions and feedback. The purpose of this experiment is to identify any areas of the tool that are difficult to use and gather feedback on how we can improve the design of the tool.

Task 1: Compare and Contrast Candidate Profiles Based on Admission Chances

Participants were asked to choose two candidates from the list and compare their profiles based on chances of admission. They were instructed to make selections using scatterplots and filter options and explain their thought process when comparing. The data we collected included the time it took participants to complete the tasks, the accuracy of their selections, and any feedback they provided regarding the availability of scatterplots and filtering options.

Task 2: Understand the factors that affect the admission opportunities of selected candidates

Participants were asked to select a candidate and use bar charts and scatterplots to understand factors affecting their chances of admission. They were instructed to adjust the weights of different factors and observe the effect on their chances of admission. The data we collected included the time it took participants to complete the task, the accuracy of their adjustments, and any feedback they provided regarding the usability of bar charts and scatter plots.

We experimented with two other groups, each participating in the other group's tests. We recorded participant interaction and feedback data during the task.

After analyzing the usability testing results, we found that the bar chart option was generally easy to use and intuitive for participants. However, some participants needed help with heat maps and adjusting the weights of different factors. Based on this feedback, we made several design changes to improve the usability of bar chart and scatterplot, including adding more evident labels and improving the UI for adjusting weights.

In conclusion, the usability testing helped us identify areas for improvement in the graduate admissions visualization tool and allowed us to make design changes to improve its overall usability.

9 Conclusion

In this project, we aim to develop a tool to help prospective students and graduate school admissions officers predict a student's chances of being admitted to a graduate school based on their academic qualifications. We start with data exploration and cleaning, then use JavaScript and the D3 library for data analysis and visualization. We also created a predictive model using machine learning techniques and integrated it into the tool.

One of our significant contributions to this project is an interactive visualization tool that allows users to explore admissions data and predict their chances of admission based on their input. We also made sure that the tool was easy to use, intuitive, and easy to navigate. Additionally, we worked hard to ensure that this project's data analysis and modeling techniques were reliable and accurate.

Overall, the program provides prospective graduate and graduate admissions officers with valuable resources and saves time evaluating students' admission opportunities and making informed decisions about their academic and professional careers.

References

[1] W. J. Brink. Selecting graduate students. *The Journal of Higher Education*, 70(5):517–523, Sept. 1999. doi: 10.2307/2649224

[2] R. A. Gregory Attiyeh. Testing for bias in graduate school admissions. *The Journal of Human Resources*, 32(3):524–548, June 1997. doi: 10.2307/146182