Goergen Institute for Data Science Masters Admissions

Sung Beom Park, Joseph Smith, Jiecheng Gu, Xiaoen Ding

Abstract

This project uses the University of Rochester's applicants data, the National Clearinghouse data, and the school rankings data (provided by the U.S. News & World Report) to create meaningful visualizations and plots to answer the most important questions the department wants to know. Ultimately, we convey and suggest ideas that will help raise the accepted yield of the GIDS program.

I. Introduction

The Goergen Institute for Data Science is the University of Rochester's interdisciplinary data science hub that offers a Bachelor of Arts (BA), a Bachelor of Science (BS), a Master of Science (MS), and the Advanced Certificate in data science. For this study, the focus is specifically on the Master of Science program. The department was particularly interested in understanding the types of institutions and programs that students are choosing to attend instead of our program using the applicants data, enrollment reporting data, and the school rankings data. In addition, assisting the data science department in marketing/future recruitment and building machine learning models to predict which applicants will potentially accept our offer were tasks we were interested in exploring.

II. Data Set description

The data sets that we were given to work with include admissions data for the Goergen Institute of Data Science since 2015, clearinghouse enrollment data of those who applied to the University of Rochester, and a data set of college rankings. In order to prepare the data sets for analysis there were several pre-processing steps that needed to be completed. Firstly, we dealt with missing values and removed unnecessary columns. It was also important to encode the admissions decisions numerically and ensure that the college names are the same across all three data sets. In order to accomplish this it was necessary to create a dictionary of over 100 unique search and replace values.

III. Exploratory analysis

To better understand the applicant pool and the application cycle, we performed exploratory analysis on the application data. The line plots of application creation per day show the number of creations in days for 2020 and 2021 as seen on Fig.1 and Fig.2. The plot is shown by different birth countries and only the three major countries are listed. The Chinese applicants tend to create their application earlier in November, 2021. The Indian and American applicants tend to create their application starting from December.

On Fig.3 and Fig.4, we can see the submission number per day for applicants from the three major countries. The number of submissions is more related to the deadlines comparing to the holidays. We have the highest submission number on deadlines and a couple of days before it. The pattern of submission for applicants from the three major countries is close. However, there was a small group of Chinese applicants submitted their application after the second deadline in 2020 when Covid-19 first break out.

Also, we can see most applicants finish their application on the day they create it from Fig.5. From Fig.6, we can see half of the applicants submit their application in two weeks and nearly 86% of the applicants finish the application in 50 days. If we divide applicants into their birth country in Fig.7, we can see that it takes Indian applicants a shorter time to complete the application. The average time needed for submission for applicants from different country is shown in the TABLE I.

On the other hand, we have some applicants who never finish their application. From Fig.8, we can tell that 22% of applicants never submit their application. Among the three countries, 41% of applicants

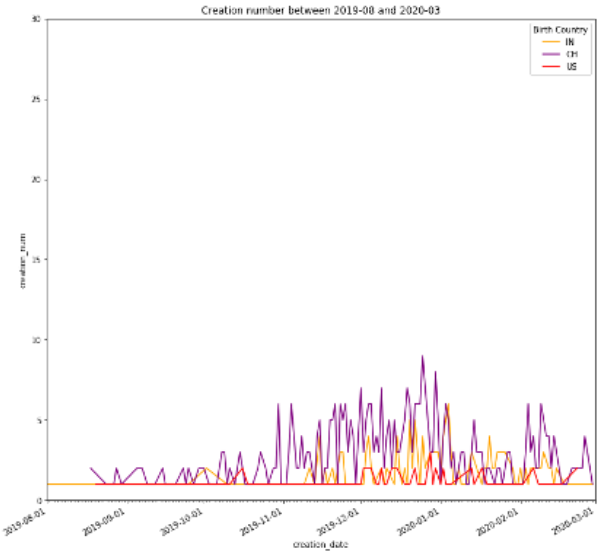

Fig. 1. Creation Per Day of 2020

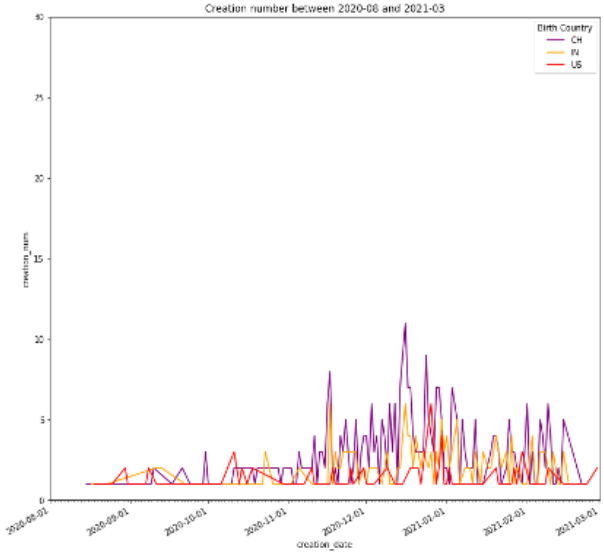

Fig. 2. Creation Per Day of 2021

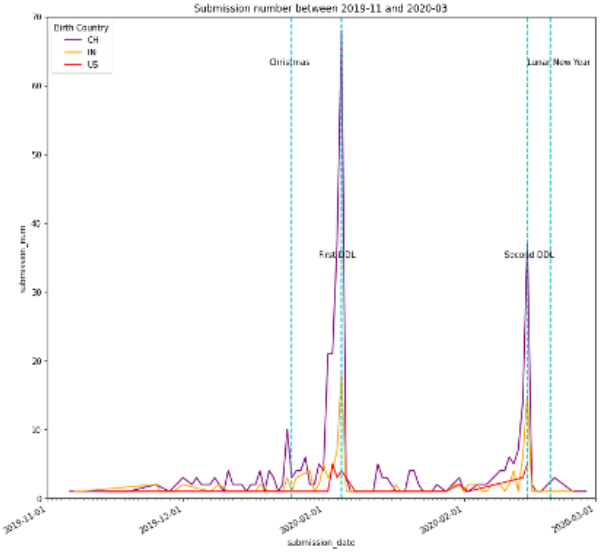

Fig. 3. Submission Per Day of 2020

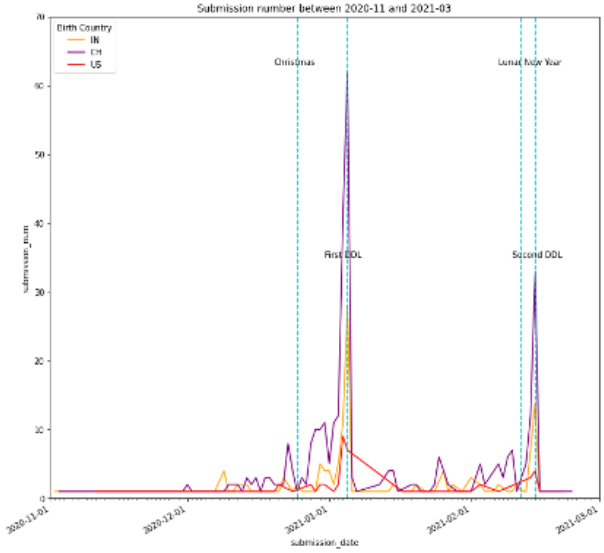

Fig. 4. Submission Per Day of 2021

TABLE I
AVERAGE APPLICATION DAY BY BIRTH COUNTRY

Country	Mean(days)	Median(days)
All	24	14
US	30	18
CH	26	19
IN	15	8

from the United States did not finish their application, which is the highest in the three countries. This might be problematic, so we take a deeper look into these applicants and try to see their demographics.

From Fig.9 we can see that most of the applicants of the United States who never finished their

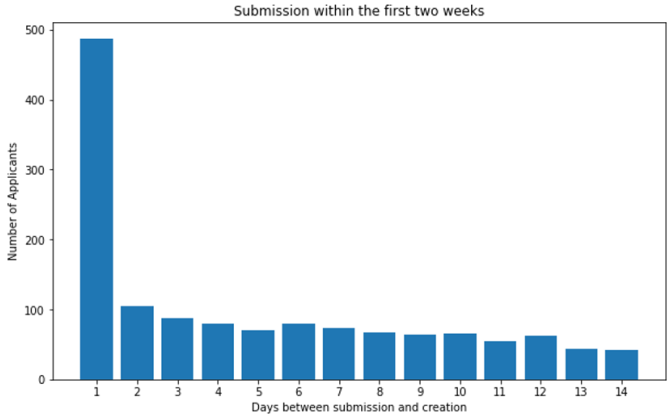

Fig. 5. Submission within the First Two weeks

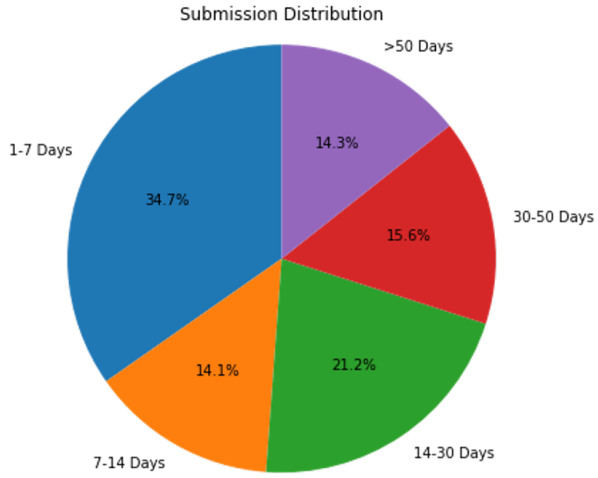

Fig. 6. Submission Distribution

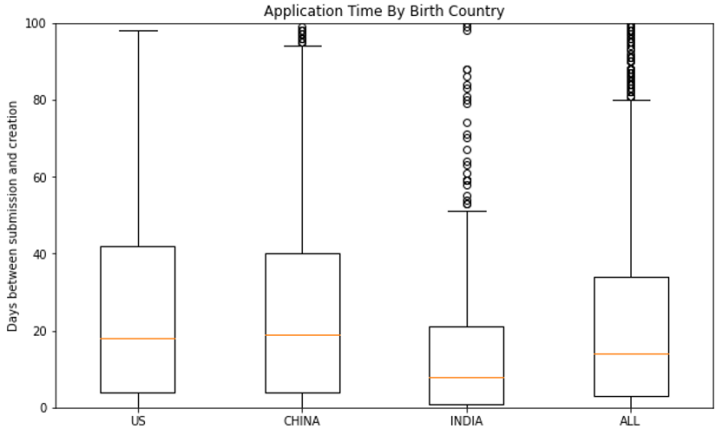

Fig. 7. Application Time by Birth Country

application are between the age of 21 to 30. And 69% of them are male. Also, 97% of the applicants are not employed and 96% of the applicants are not a current student by the time they start their application.

We can see histogram of the number of schools students list on their application as seen on Fig.10. The count is largely dominated by the group 'NA' which simply just means that they did not list another school that they applied to. This is followed by the count of one.

Furthermore, we wanted to explore the one other school that the students did list on their application, to see the most popular schools and location (state). As seen on Fig.11, we can see that this count is dominated by the University of Rochester. We suspect that this is due to the fact that students are applying to other departments, such as the Simon School of Business. Other than the University of Rochester, the most frequently listed schools were Columbia University, New York University, and the University of Southern California.

While on the topic of the most popular schools that the students did list and apply to, we can also look at the most popular states of these schools. It is not surprising that the pie chart as seen on Fig.12 is largely dominated by the state of New York as the top 3 schools of the most popular schools that students applied to are located in NYS. Following the state of New York, the states with the most

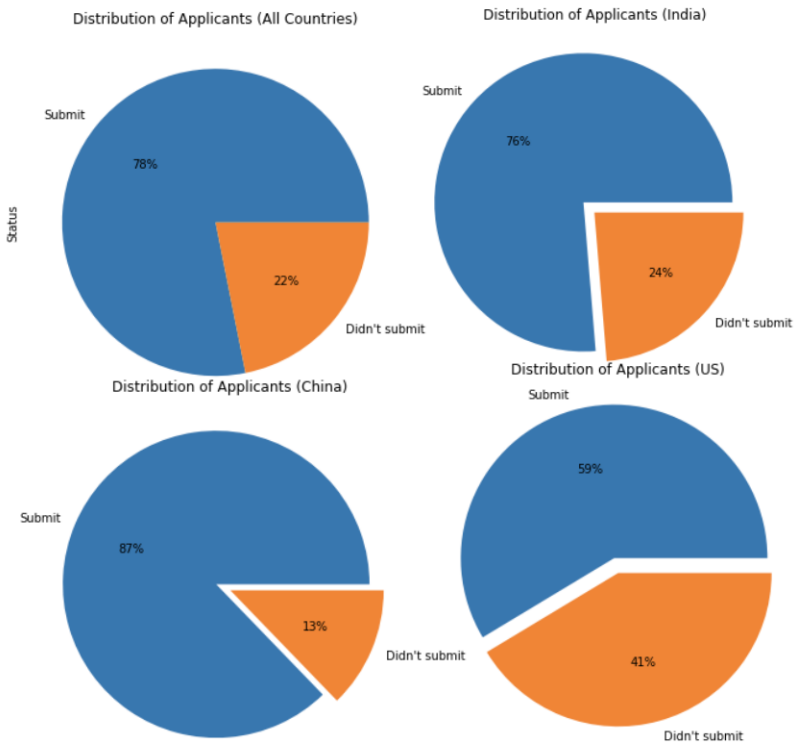

Fig. 8. Distribution of Applicants Who Never Submit

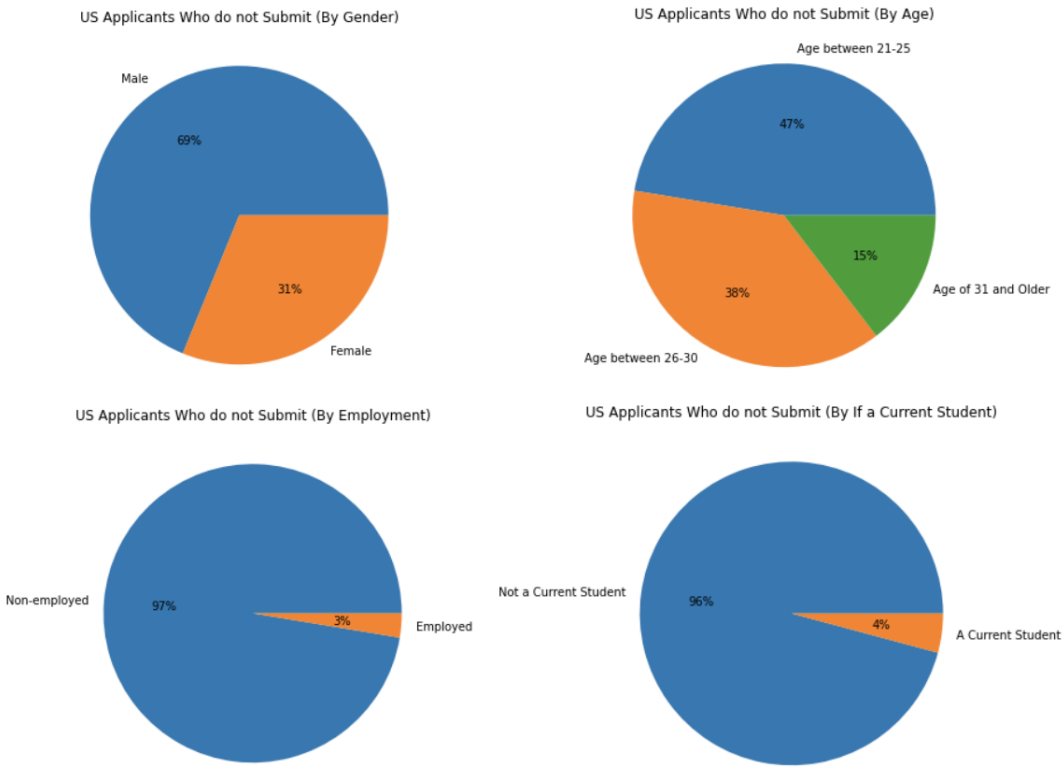

Fig. 9. Applicants from the United States Who Never Submit

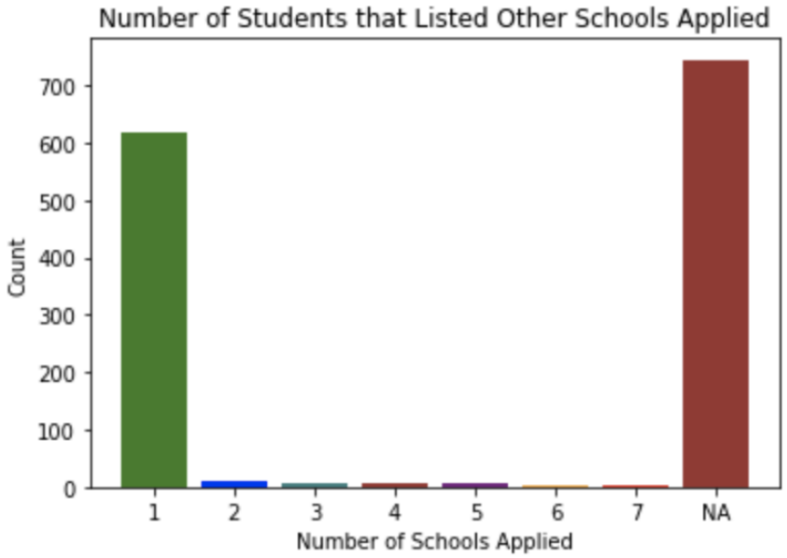

Fig. 10. Number of Students that Listed Other Schools Applied

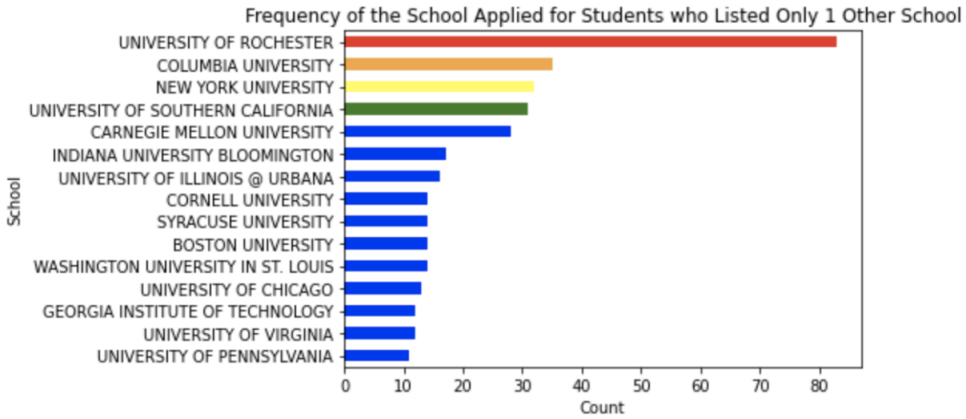

Fig. 11. Frequency of Schools Students Applied for

frequent applications are California, Pennsylvania, Illinois, and Massachusetts.

IV. Statistical Tests

The difference between the average previous college rank of a student who was denied by University of Rochester versus one who was accepted but denied admittance is statistically significant. The t-statistic was found to be 3.49582 and a p-value of 0.00072. In Fig.13 below you will see box plots for both populations where the rank of University of Rochester is indicated by the black arrow. Accompanying these box plots are bar plots displaying the most common schools within the population.

The difference between the average enrolled college rank of a student who was denied by University of Rochester versus one who was accepted but denied admittance is statistically significant. The t-statistic was found to be 3.66607 and a p-value of 0.00028. In Fig.14 below you will see box plots for both populations where the rank of University of Rochester is indicated by the black arrow. Accompanying these box plots are bar plots displaying the most common schools within the population.

The difference between the average GPA of a student who was denied by University of Rochester versus one who was accepted but denied admittance is statistically significant. The t-statistic was found

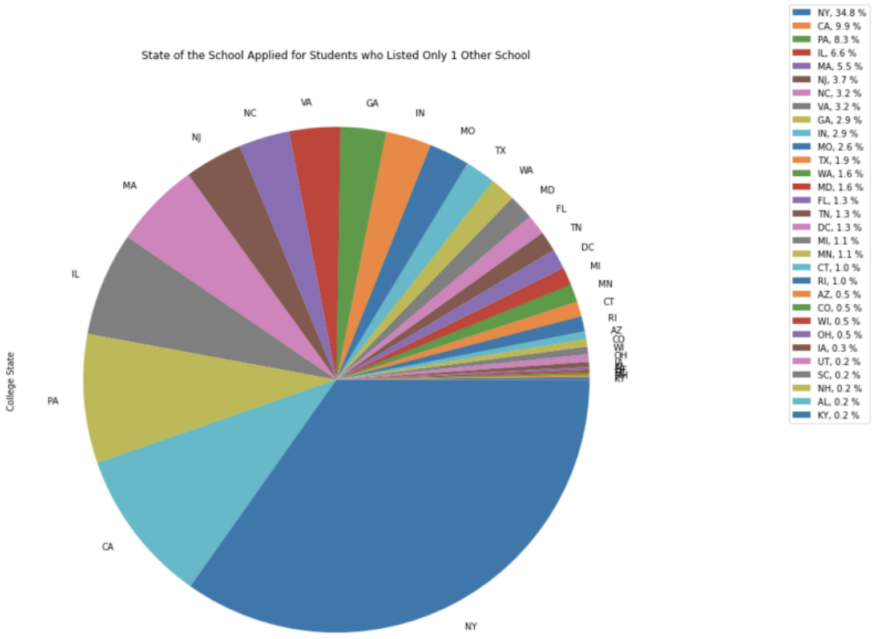

Fig. 12. State of Schools Students Applied for

Fig. 13. Previous School Rank of Applicants

to be -12.0666 and a p-value of 7.48e-29. The difference between the average GPA of a student who accepted an offer from University of Rochester versus one who was accepted but denied admittance is not statistically significant. The t-statistic was found to be 0.47724 and a p-value of 0.63378.

The difference between the average GRE quantitative percentile of a student who was denied by University of Rochester versus one who was accepted but denied admittance is statistically significant. The t-statistic was found to be -3.5295 and a p-value of 0.00047. The difference between the average GRE quantitative percentile of a student who accepted an offer from University of Rochester versus one who was accepted but denied admittance is not statistically significant. The t-statistic was found to

Fig. 14. Enrolled School Rank of Applicants

be -0.49467 and a p-value of 0.62149.

In Fig.15 below you will see box plots for both populations where the rank of University of Rochester is indicated by the black arrow. Accompanying these box plots are bar plots displaying the most common schools within the population.

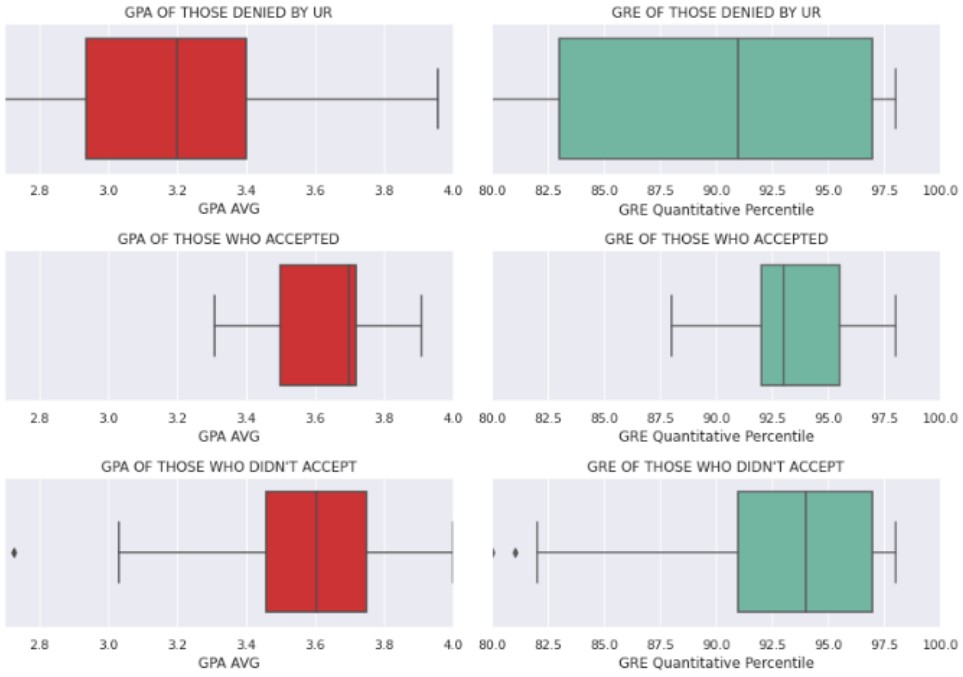

Fig. 15. GPA and GRE Quantitative Percentile of Applicants

V. Model development

Based on exploratory data analysis above, we aim to use classification models including Decision Tree, Logistic, Naive Bayes, K-nearest Neighbours (KNN), and Random Forest to predict if admitted applicants would accept our offer or not.

A. Feature Engineering

Firstly, we select applicants who had the offer based on the decision types. Then we delete all the PhD applicants from admitted students because this project focuses on the applicants with an academic degree that is lower than PhD. The target variable of our classification models is whether applicants accepted our offer or not. So we label those accepted our offer as 1 and those declined our offer as 0.

Based on previous data analysis and statistics tests, we choose Tuition Scholarship Percentage, Age at App. Submission, Completion Time (time difference between creating and submitting the application), GPA (4.0 Scale), Birth Country as independent variables of our models. There are null values in the column of Tuition Scholarship Percentage, and we replaced these values with 0 because these null values refer to applicants who did not receive a scholarship. As for GPA, some applicants listed more than one GPA at different institutions or with different degrees. We only select the first GPA they wrote down.

To reduce the cardinality of continuous and discrete variables, we binned these independent variables based on their percentiles. As shown in TABLE II, Tuition Scholarship Percentage is binned into four groups: No tuition. 20%-30% tuition, 40%-45% tuition, and larger than 45% tuition (notice that there is no range between 30% to 40% because no one received a scholarship in that range in our data set). Age at Application Submission is binned into four groups: 18-22, 23-25, 26-30, and larger than 30. Completion Time is binned into five groups: 1-7 days, 8-14 days, 15-30 days, 31-49 days, >=50 days. GPA is also binned into five groups: <3.0, 3.0-3.5, 3.5-3.6, 3.6-3.7, >3.7. Birth Country is binned into four groups: US, IN, CH, and other countries.

TABLE II

Binning of independent variables

Feature	Bin
Tuition Scholarship Percentage	No tuition, 20%-30% tuition, 40%-45% tuition, >45% tuition
Age at App. Submission	18-22, 23-25, 26-30, >30
Completion Time	1-7 days, 8-14 days, 15-30 days, 31-49 days, >=50 days
GPA	<3.0, 3.0-3.5, 3.5-3.6, 3.6-3.7, >3.7
Birth Country	US, IN, CH, OTHERS

B. Clustering

After binning the independent variables, we use clustering to explore patterns of admitted students. Since our features are categorical features after binning, we decided to use K-modes clustering instead of K-means clustering. As shown in Fig.16, the optimal number of clusters with Elbow method should be 5, 6, or 7. In order to better represent the applicants from three major countries, the number of clusters is set to 7.

Table III shows the centroid, the accept rate, and the melt rate of applicants in each cluster. We can see from Table III that there are four clusters in which applicants' birth country is China because Chinese applicants are dominated in our data set. Group 2 has the lowest accept rate and a relatively high melt rate, indicating that applicants who had a very high GPA (above 3.7) and submitted their applicant within more than 50 days are highly likely to decline our offer, even though they received

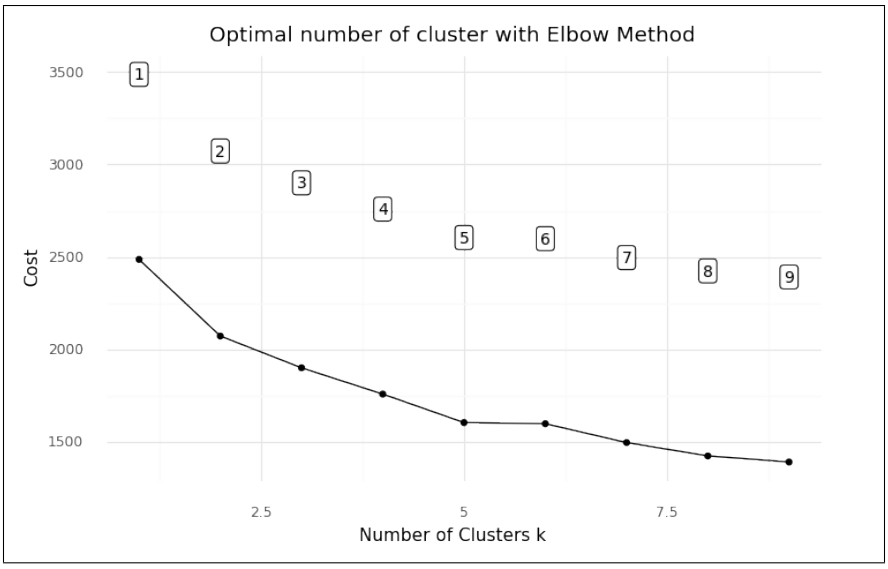

Fig. 16. Optimal number of cluster with Elbow Method

TABLE III
CENTRIODS OF K-MODES CLUSTERING

Group	Tuition Percentage	Age	Time (days)	GPA	Country	Accept Rate	Melt Rate
1	40%-45%	18-22	8-14	3.0-3.5	CH	49.2%	36.5%
2	40%-45%	18-22	>=50	>3.7	CH	26.6%	54.2%
3	20%-30%	18-22	15-30	3.5-3.6	CH	33.3%	31.3%
4	20%-30%	18-22	1-7	3.6-3.7	CH	28.0%	43.3%
5	20%-30%	18-22	31-49	3.0-3.5	US	35.5%	45.5%
6	40%-45%	23-25	1-7	3.6-3.7	US	58%	58.6%
7	>45%	18-22	1-7	>3.7	IN	45.9%	23.5%

very high scholarship. Group 1 also received very high scholarship, but their GPA is lower than Group 2 and they spent shorter time in their application. This group has a relatively high acceptance rate and a low melting rate. Applicants in Group 6 are older than those in other groups, indicating that they may have worked for a while. These people have the highest acceptance rate and melting rate. As for Group 7, applicants in this group received the highest scholarship (50% tuition scholarship) and have a high acceptance rate and a low melting rate. In conclusion, scholarship, age, completion time, and GPA can influence applicants' decision.

C. Oversampling and Undersampling

We also notice that there is a class imbalance in our data set. As shown in Fig.17, over 70% admitted students declined our offer while only about 30% admitted students accept our offer. To solve this problem, we use SMOTEENN to adjust the class distribution of our data set. SMOTEENN is a sampling method of combining over and undersampling using SMOTE and Edited Nearest Neighbours.

D. Model selection

The classification models that we use in this capstone are Decision Tree, Logistic, Gaussian Naive Bayes (GNB), K-nearest Neighbours (KNN), and Random Forest Model. We split our data set in 80:20 ratio. 80% of our data set goes is used for training the models and 20% is used for testing the models. The metrics the we use to evaluate and caompare the models are the accuracy score and the recall score.

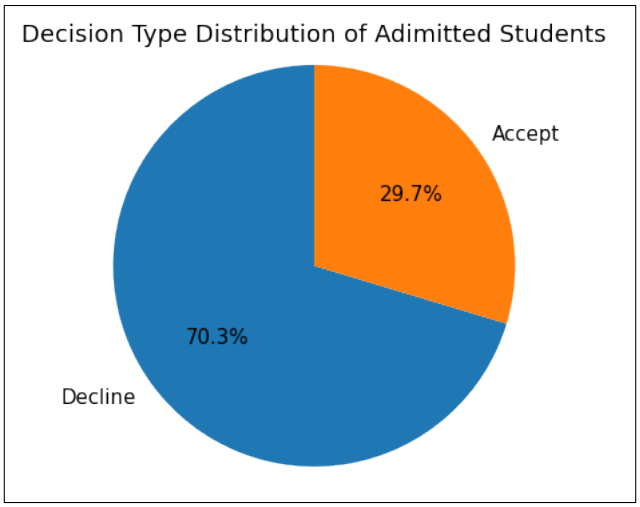

Fig. 17. Decision Type Distribution of Admitted Students

VI. Performance and Results

A. Model Evaluation

In the capstone, we used two methods, including label encoding and one hot encoding, to encode our binned features before feeding data into classification models. Fig.18 shows the accuracy scores of our models. We can see that the best model so far is Naive Bayes Model with one hot encoding and its accuracy score is 0.753, followed by Decision Tree (0.727) and KNN (0.727). But because of the class imbalance, the recall scores of our models are low (< 0.3), which is shown in Fig.19. As discussed in the section above, we used SMOTEENN to adjust the class distribution. Fig.20 and Fig.21 show the accuracy scores and the recall scores of our models respectively. It is clear that the recall scores significantly increased after sampling while the accuracy score declined a little. The accuracy score of Decision Tree decreased the most from 0.727 to 0.55, while its recall raised the most from 0.166 to 0.66.

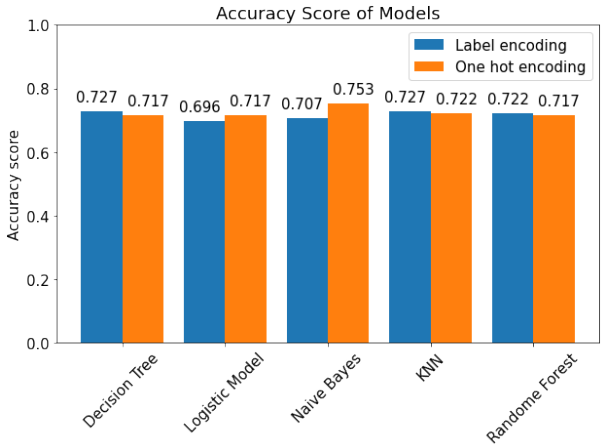

Fig. 18. Accuracy Score of Models

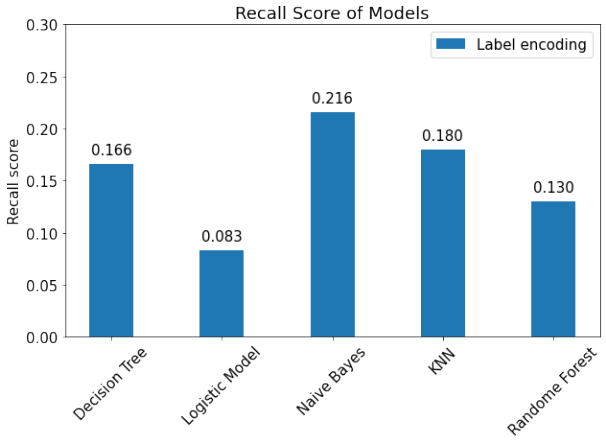

Fig. 19. Recall Score of Models

B. Feature Analysis

We also want to see how each feature affects the acceptance rate. TABLE IV shows the results of Logistic Model. According to the coefficients of the features, both Tuition scholarship percentage and

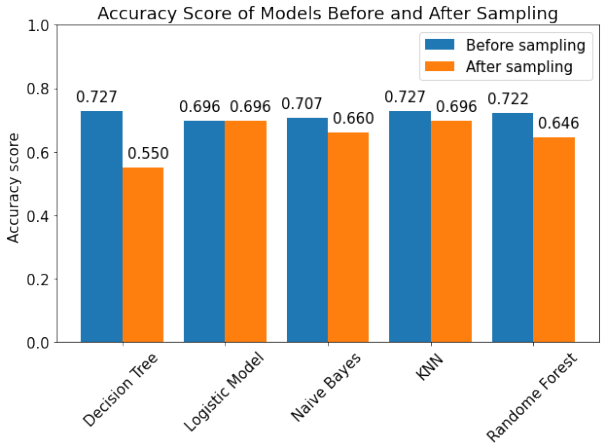

Fig. 20. Accuracy Score of Models Before and After Sampling

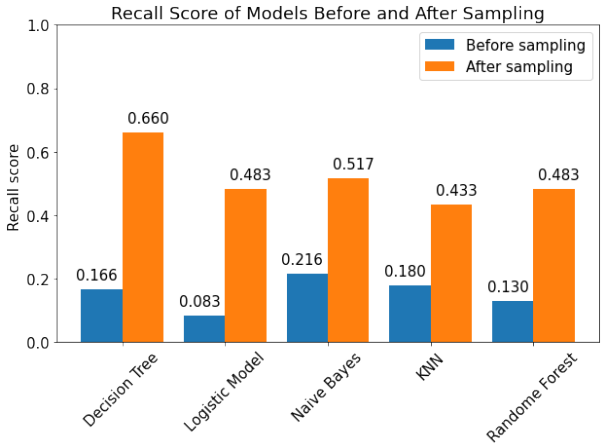

Fig. 21. Recall Score of Models Before and After Sampling

Age have a positive impact on the acceptance rate, while Completion time and GPA have a negative impact. As we labeled international students as positive numbers (TABLE I), the negative coefficient of Birth country means that being born in countries outside the U.S. could have a negative impact on the acceptance rate. Moreover, according to the p-value, Age, GPA, and Birth country are statistically significant since their p-values are all less than 0.05.

TABLE IV
RESULTS OF LOGISTIC MODEL

Feature	Coefficient	P-value
Tuition Scholarship Percentage	0.129	0.111
Age at App. Submission	0.300	0.001
Completion Time	-0.102	0.055
GPA	-0.326	0.000
Birth Country	-0.148	0.046

VII. CONCLUSION AND NEXT STEPS

By completing the necessary exploratory data analysis and model development, we are able to provide some insightful information and suggestions for the department to potentially increase the accepted yield of the GIDS.

- There are submission peaks before the two deadlines. So we suggest reminding applicants to submit their applications before the deadlines, and don't need to remind people before holidays (Christmas and Lunar New Year) because there are actually no noticeable submission peaks during holidays.
- American applicants tend to procrastinate and not submit their application compared to other groups. So we suggest urging American applicants to submit their applications.
- A long completion day, defined as an application open for longer than fifty days, will most likely lead to a lower acceptance rate. So when considering if an applicant is qualified for the program, consider their academic background as well as their completion day.

The following steps will still focus on data analysis and model improvement. As for data analysis, since we lack the ranking data of universities outside the U.S., we cannot compare the universities that

international students graduated from and the University of Rochester. As for model improvement, Grid Search would be an excellent method to find the best parameters of models. Also, we can add more features to the models, such as if applicants have worked before, if they have failed a course, and the ranking of their previous schools.

ACKNOWLEDGMENT

We would like to give a special thank you to professor Ajay Anand, Lisa Altman, and Gretchen Briscoe for providing us with an intriguing problem to solve, the data set, and for leading/guiding us throughout this semester in helping us successfully completing this project.