This report is to outline the findings from trying to understand University of Rochester student engagement with the Gwen M. Greene Center.

Data

Data has been collected from Handshake's dataset. In summary four datasets and their respective features that has been used is outlined below:

- Applications:
    - Username, Applications ID, Employer Industry Name, 'Postings Apply Start Date', 'Postings Expiration Date Date', 'Student Educations Cumulative Gpa', 'Applications Created At Date'
    - N = 51845
- Career Fair:
    - 'Student Attendees Username', 'Career Fair ID', 'Career Fair Session Session Start Date', 'Career Fair Student Registration Start Date', 'Career Fair Student Registration End Date', 'Career Fair Session Attendees Checked In? (Yes / No)', 'Career Fair Session Attendees Checked In At Date', 'Career Fair Session Attendees Created At Date', 'Career Fair Session Attendees Pre-Registered At Date', 'Career Fair Session Attendees Pre-Registered? (Yes / No)'
    - N = 4264
- Students:
    - 'Students Username', 'Students Gender', 'Work Authorization Name', 'School Year Name', 'Career Interests: Career Clusters Name', 'Majors Name', 'Educations Cumulative Gpa', 'Documents Count'
    - N = 21259
- Appointments:
    - 'Student Username', 'Staff Member ID', 'Appointments Checked In? (Yes / No)', 'Appointments Drop-in? (Yes / No)', 'Appointments Start Date Date', 'Appointment Type Length (Minutes)', 'Appointment Type Name', 'Appointment Categories Name', 'Created By Created At Date'

Before starting the analysis, all data has been joined together down to a student level. The code for this can be found in "Student Dataset.ipynb". N = 13,215 is the final data count after merging and is called "all_data_numeric.csv" in the repository. The Jupyter notebook contains comments for each step to understand what has been done for preparing the final data.

Methods

The goal of the analysis was to define student engagement and to explore if there are any insights that we can find that is driving student engagement.

i) Defining Student Engagement

We define student engagement with the career center via three indicators:

- Jobs: Whether the student applied to a job

- Career Fair: If the student attended at least one career fair
- Appointments: If the student came to the career center for drop in hours or met with an advisor

In the final dataset you can find these three variables as binary attributes. In the future, it would be much beneficial to provide them with weights that represents relative importance. This will allow us to create a final student engagement variable.

ii) Statistical Analysis

To understand the data basic exploratory analysis and statistical tests has been done. This can be found in script/Stats_Tests.ipynb

A correlation matrix has been used to find dependability between each of the variables to understand the trend with respect to each other.

Statistical models such as t-test and Mann-Whitney tests for each feature to find differences between the engaged and not engaged groups. In the notebook, the means of the two groups has been printed along with Cohen's d effect size. The higher the Cohen's d value, the higher the effect or difference sizes. A negative cohen's d means that the not engaged group is higher than the engaged group. The two groups are engaged and not engaged changes according to the three engagement variables ('Engaged_Fair', 'Engaged_Appointment' and 'Engaged_Jobs')

iii) Predicting Student Engagement

Some of the basic Machine Learning models (Logistic Regression, Support Vector Machines, Decision Tree) was applied to predict student engagement from the following features:

'US Citizen','School Year Name','Educations Cumulative Gpa','Documents Count','Engaged_Fair', 'Engaged_Appointment','Engaged_Jobs'

The same model was run three times where the dependent/predicting variable was switched between 'Engaged_Fair', 'Engaged_Appointment' and 'Engaged_Jobs'. Of course, the dependent variable was not included as an independent variable during their respective models. Details on how to run the code and model architecture can be found in script/README.md and the code is in script/ml.py.

To draw more interpretation from the machine learning model, logistic regression weights has been plotted which can be found in the "figs" directory. This allows us to understand the relative importance of each of the features with respect to each other including their association with the prediction variable (engagement).

Results

i) Exploratory and Statistical Findings

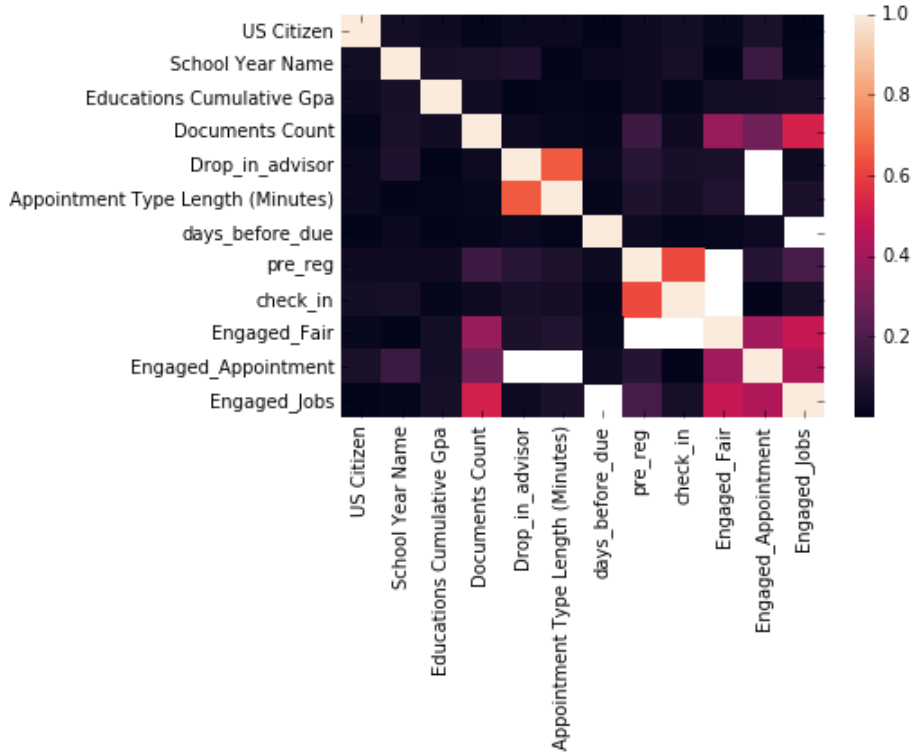


Fig 1a) Correlation Matrix between all the variables from the mega dataset

Fig 1b) Correlation Matrix between features used in the Machine Learning models

Table I: Statistical Tests between Engaged Jobs and Not Engaged Jobs

Feature name	t-test p-val	MW p-val	Engaged Mean	Not Engaged Mean	Cohen's d Effect size
US Citizen	2.04e-11	1.09e-11	0.64	0.71	-0.16
School Year	6.4e-177	1.43e-168	3.63	2.68	0.71
GPA	6.63e-09	5.00e-09	3.49	3.32	0.12
#Documents	0	0	5.03	0.13	0.98
Engaged Fair	0	1.63e-307	0.52	0.12	0.93
Engaged Appointments	4.25e-211	5.57e-199	0.72	0.36	0.77

Table II: Statistical Tests between Engaged Appointment and Not Engaged Appointment

Feature name	t-test p-val	MW p-val	Engaged Mean	Not Engaged Mean	Cohen's d Effect size
US Citizen					
School Year					
GPA					
#Documents					
Engaged Fair					
Engaged Appointments					

Table III: Statistical Tests between Engaged Fairs and Not Engaged Fairs

Feature name	t-test p-val	MW p-val	Engaged Mean	Not Engaged Mean	Cohen's d Effect size
US Citizen	1.69e-16	9.76e			
School Year					
GPA					
#Documents					
Engaged Fair					
Engaged Appointments					

ii) Predicting Student Engagement

a. Model Predictions

Target variable: Engaged_Jobs

class proportion: 0.3424144609425436

MODEL: KNN

Avg test acc:0.919 Avg train acc:0.923 Avg f1 acc:0.889 Avg f1 acc:0.889

MODEL: LOGISTIC

Avg test acc:0.901 Avg train acc:0.901 Avg f1 acc:0.841 Avg f1 acc:0.841

MODEL: TREE

Avg test acc:0.918 Avg train acc:0.918 Avg f1 acc:0.890 Avg f1 acc:0.890

Target variable: Engaged_Fair

class proportion: 0.25861846352485474

MODEL: KNN

Avg test acc:0.793 Avg train acc:0.801 Avg f1 acc:0.570 Avg f1 acc:0.570

MODEL: LOGISTIC

Avg test acc:0.786 Avg train acc:0.787 Avg f1 acc:0.508 Avg f1 acc:0.508

MODEL: TREE

Avg test acc:0.781 Avg train acc:0.788 Avg f1 acc:0.465 Avg f1 acc:0.465

Target variable: Engaged_Appointment

class proportion: 0.4839251129761136

MODEL: KNN

Avg test acc:0.688 Avg train acc:0.698 Avg f1 acc:0.672 Avg f1 acc:0.672

MODEL: LOGISTIC

Avg test acc:0.668 Avg train acc:0.668 Avg f1 acc:0.590 Avg f1 acc:0.590

MODEL: TREE

Avg test acc:0.693 Avg train acc:0.694 Avg f1 acc:0.670 Avg f1 acc:0.670

b. Logistic Regression Weights

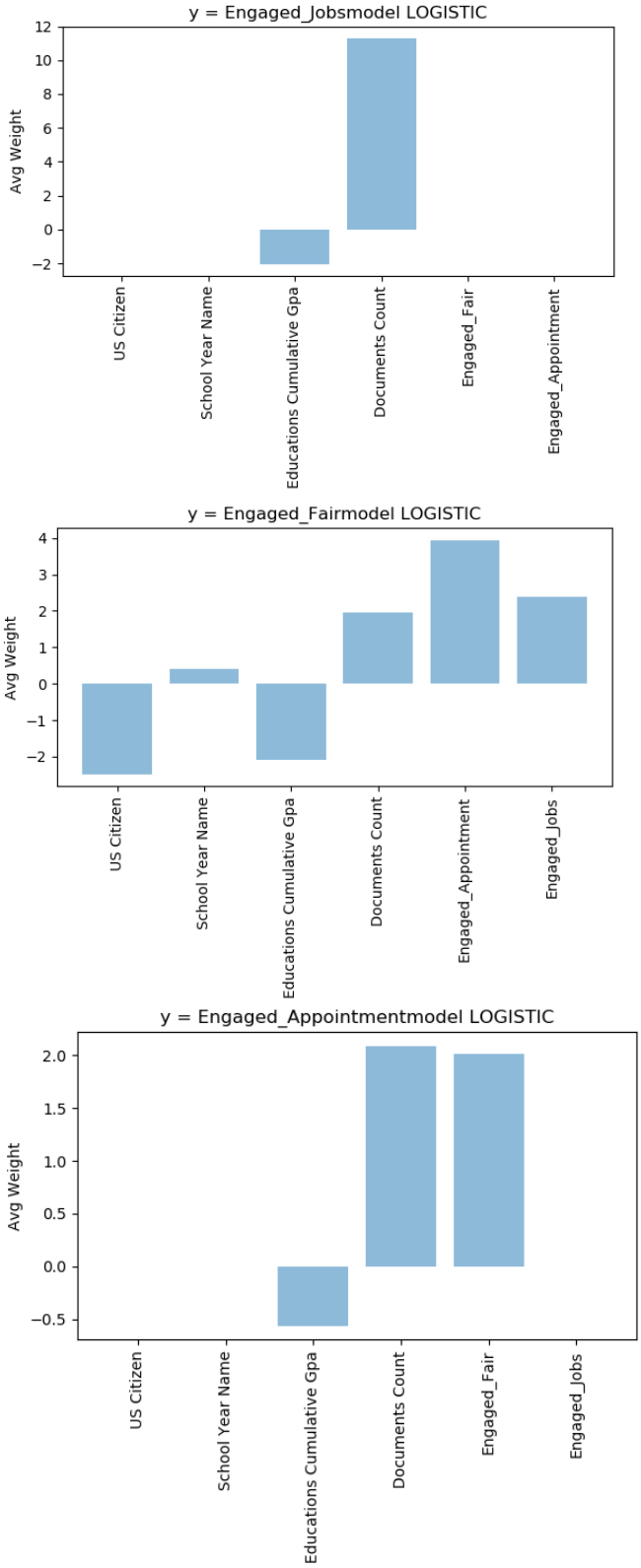


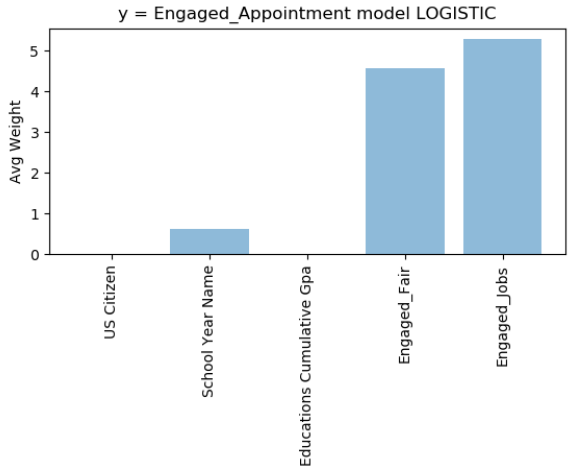
Fig 2a,b,c): Weights of Logistic Regression for each feature in the three models

Discussion

Finally, there are certain key insights that we find from the tests and analysis. In this section all those are accumulated, and explanations are given with respect to findings from statistical analysis and the machine learning models. Here are the following conclusions from the three major engagement models:

- Engaged Jobs
    - We can achieve the best prediction rates across all three machine learning models
    - #Documents is the top predictor for engaged jobs
        - Highest weight with positive association from Fig2a of Logistic regression weights
        - Statistical significance (p-value < 0.05) with high Cohen's d effect size (0.98) showing that it is high differentiating factor and that engaged group contains high #Documents
        - Correlation matrix shows high correlation with Engaged Jobs in both Fig1a and Fig1b
    - GPA even though has the next highest weight in Fig2a, cohen's d effect size (small effect size: 012) from Table I and a low correlation from correlation matrix in Fig1 does not show much promise, for its contribution in differentiating engagement in jobs
- Engaged Fair
    - Second best prediction rates, <78% accuracy
    - Engaged Appointments is the top predictor
        - Highest logistic regression weights in Fig 2b), association is also positive which means that higher Engagement in Appointments lead to higher Engaged Fair
        - Statistical significance (p-value < 0.05) with high Cohen's d effect size (0.79)
        - Correlation matrix in Fig 1 shows the most correlation of Engaged Fair with Engaged Appointments
    - Engaged Jobs and #Documents have similar contributions and predicts more Engaged Appointments
        - Could be because both variables are pretty closely and positively correlated, so the model may provide one or the other more relative importance during the 10-fold cross validation tests
        - Both Engaged Jobs and #Documents are statistically significant and have pretty high Cohen's d effect size (0.64 and 1.05)
        - Correlation matrix also shows decent correlation with Engaged Jobs
- Engaged Appointments
    - Lowest Prediction rates across all three models, <66% accuracy
    - Engaged Fair and #Documents are the top predictors here
        - Both have relatively similar and the highest regression weights with a positive association to engaged Appointments
        - Table III shows that both Engaged Fair and #Documents are both statistically significant with decently high Cohen's d effect size (0.67 and 0.45)

        - Correlation matrix also shows that they have decent correlations even though #Documents is slightly less correlated than Engaged Fiar
    - Engaged Jobs may have some association
        - Did not show up or had no weight for logistic regression weights
            - This happens as Engaged Jobs and #documents is highly correlated so the logisitic regression model ignores Engaged Jobs. In the figure below, where Engaged Appointments model was run without #documents we find Engaged Jobs to have the highest weight



        - We find statistical significance in Table III along with high Cohen's d effect size of 0.72
        - Correlation Matrix in Fig1 also shows pretty good (higher than #Documents) correlation with Engaged Appointments

Limitations and Future Work

- Getting the data ready took a considerable amount of time since, most handshake data is categorical and running analysis is difficult on that. Identifying some sort of scaling or conversion to the features will be greatly beneficial. For e.g. we converted school year to a numerical scale, with higher number representing more seniority.
- A lot of the features had to omitted as a lot of the features that was generated or included in the mega dataset was depended on the presence of one of the Engagement variables. For e.g "early apply" was a feature that accounted for how early the student applied for the job with relation to the due date. For this feature to be used/be present it was necessary that the student had applied to once job at least, hence, it was not possible to include that in the Engaged Jobs model. Such features were omitted. In the future, may when different models are run features like these could be added or discarded according to the type of model being run. For instance, "early apply" could be used in the Engaged Appointment or Engaged Fair models
- The weighted student engagement variable as discussed previously in the Methods section could also be used.
- For this analysis the Events data that accounts for all events (recruitment events, any hall talks etc.) could be included to do a similar analysis in the future.

- Student duplicates could be handled with more care. Since, students progress with their school years over the years, it could be beneficial to use a time series model to take that into considerations. In this analysis the students most recent school year is considered even though the data contains averaged out activity from their previous years. A way to tackle this problem could be to use a time series analysis where we analyze how the student progresses over their four years. For e.g in this case the y variable could be the number of jobs they apply to in a given year. Hence, each student will have four datapoints and the general trend or progression could be predicted for all students.