Problem Understanding

ZeeMee is a Silicon Valey based social network startup that let connects among students who are looking for same colleges. The dataset was created by ZeeMee. I chose this dataset because when I was doing research to get an admission into US universities, I used this app. And later I found this dataset from ZeeMee Mini-Hackathon by MatrixDS and it caught my interest.

The goal here is to predict the following task by using the dataset about the student behavior:

- Perform descriptive statistics and visualization to understand the data well.
- Build a prediction model to predict if the student will enroll in a specific college or not.
- Performing exploration of important features for the prediction model.

The goal of the prediction model is to predict **final_funnel_stage**. The funnel is a series of steps that a student moves through on their way to actually showing up to class. The stages are thought of in the following progression:

- **Inquired:** Expressed interest in the college on the zeemee app
- **Applied:** Filled out some part of an application from the college
- **Application_Complete:** Completed an application from the college
- **Accepted:** Accepted by the college
- **Deposited:** Paid a deposit to the college
- **Enrolled:** Enrolled in class at the college

The prediction of interest for this project is to focus on identifying students that **enroll (funnel stage Enrolled or Deposited)**. This is a **binary prediction**, either the student **does or does not enroll**. Use the two csv files in the data folder to build your model.

Data Understanding

The dataset has 18838 instances and 19 features with a file size of 1996 KB. The zeemee_data.csv file in the data folder is used to build the model.

The dataset contains **19 features** which are described below.

1. **college:** The college of interest that a particular student is following
2. **public_profie_enabled:** If the student has made their zeemee profile public
3. **going:** If the student has stated (in a non-binding way) on the zeemee app that they are going to the college
4. **interested:** If the student has stated they are interested in the college on the zeemee app

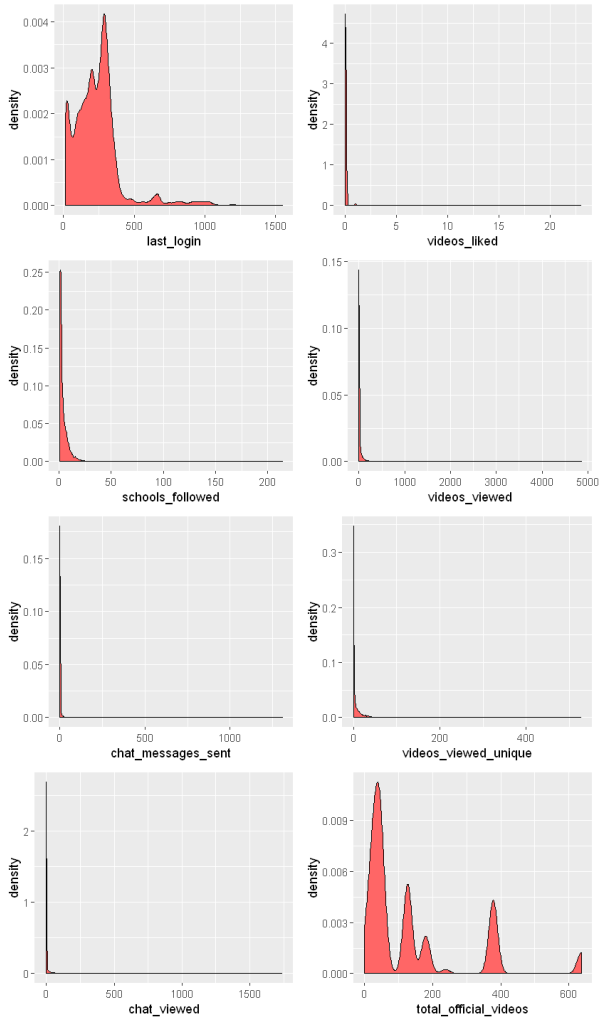
5. **start_term:** Which term the student is projected to begin class
6. **cohort_year:** Which year the student is projected to begin class
7. **created_by_csv:** If the students zeemee account associated to the college as part of a batch upload
8. **last_login:** Number of days since the last login
9. **schools_followed:** Number of schools followed on the zeemee platform
10. **high_school:** Which high school the student attends
11. **transfer_status:** If the student is transferring from another college
12. **roomate_match_quiz:** If the student filled out a ZeeMee provided quiz to match with a roomate at the college of interest
13. **chat_messages_sent:** Number of messages sent
14. **chat_viewed:** Number of chats viewed
15. **videos_liked:** Number of videos liked
16. **videos_viewed:** Number of videos viewed
17. **videos_viewed_unique:** Number of unique videos viewed
18. **offical_videos:** Number of videos produced by the college of interest
19. **engaged:** If the student is engaged with the college on the zeemee app
20. **final_funnel_stage:** What stage in the enrolment process did the student end

Findings from the categorical columns of Training data

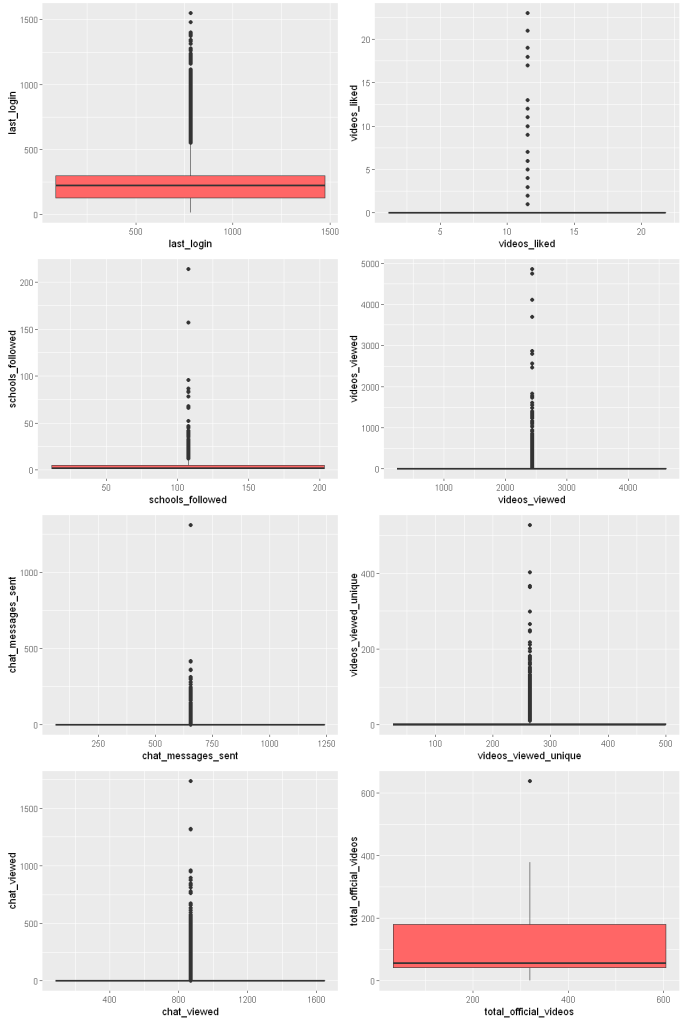
- **college** = Seems well balanced. However, it seems *'college 2'* has greater demand.
- **public_profile_enabled** = Very imbalanced. *'true'* is dominating. Most of the students have made their zeemee profile public.
- **going** = Very imbalanced. 'undecided' is dominating. Most of the students have stated on the zeemee app that they *haven't decided* wheather they are going to the college or not.
- **interested** = Very imbalanced. 'True' is dominating. Most of the students have stated they are interested in the college on the zeemee app.
- **start_term** = Very very imbalanced. 'fall' is dominating. Almost all the students are projected to begin class in the *'Fall' semester*.
- **cohort_year** = Very very imbalanced. '2019' is supreme. Almost all the student have projected to begin class in *'2019'*.
- **created_by_csv** = imbalanced. 'false' is dominating. Most of the student's zeemee account *is not associated* to the college as part of a batch upload.

- **high_school** = Very very imbalanced. 'unknown' and 'other' is dominating. Most of the students don't willing to provide information about which high school they attended.
- **transfer_status** = Very imbalanced. 'false' is dominating. Most of the students are not transferring from another college.
- **roommate_match_quiz** = Imbalanced. 'false' is dominating. Most of the students filled out a ZeeMee *did not provided quiz* to match with a roomate at the college of interest.
- **engaged** = Well balanced. However, a higher number of students are *not engaged* with the college on the zeemee app.
- **final_funnel_stage** = Imbalanced. 'inquired' is dominating. This is the target. It seems that most of the students used ZeeMee for 'inquiry' purpose. Very few students have deposited or enrolled

Distribution of the numerical columns using numerical feature were visualized.



Describing the numerical columns for every box plots

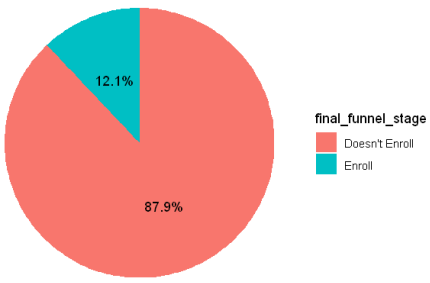


Findings from the numerical columns of Training data

- **last_login** = Right skewed. Contains outliers. On average the last login of the student is 239 days before. Where 75% of the students last login is 299 days before.
- **schools_followed** = Highly right skewed. Contains lots of outliers. On average most of the students followed 4 schools on the zeemee platform. 75% of the students followed 1-5 schools.
- **chat_messages_sent** = Highly right skewed. Contains lots of outliers. Though the average chat messages sent is 1, 75% of the students haven't sent any single message.
- **chat_viewed** = Highly right skewed. Contains lots of outliers. Though the average chat views is 12, 50% of the students didn't give any single view.
- **videos_liked** = Highly right skewed. Contains lots of outliers. Almost none of the students gave like to the videos.
- **videos_viewed** = Highly right skewed. Contains lots of outliers. 50% of the students did not viewded videos for a single time.
- **videos_viewed_unique** = Highly right skewed. Contains lots of outliers. 50% of the students did not viewded unique videos for a single time.
- **total_offical_videos** = A zigzag distribution. On average 135 videos were produced by the college of interest.

Data Preparation

After loading data different data types of the features were observed and missing values form **last_login** feature was omitted. There are 12 categorical feature and 8 numerical feature and these features were separated in **cat_cols** and **num_cols.** Target column **final_funnel_stage** was transformed into two categories; Enroll and Does not enroll. From the pie chart we can observe from the ratio of the target column that it is a highly imbalanced data where 12.1% students enroll and 87.9% does not enroll. The dataset was split into test(20%) and training(80%) for deployment.



Modeling

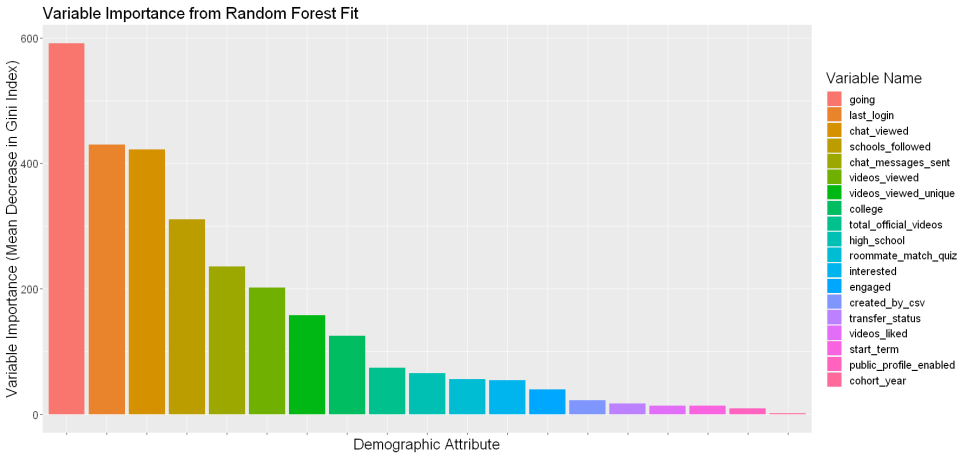
Two models were used to for the prediction; Logistic Regression and Random Forest. Logistic Regression is used because the classification is a binary task and it is a very power algorithm and less prone to overfitting. Another algorithm is used to compare the performance of the prediction. Random Forest works well on imbalance data which is the case in this context.

Evaluation

From confusion matrix for logistic regression we can see that the accuracy of the model is 95% with Kappa value of 0.71, sensitivity 0.64 and specificity 0.99. The AUC score for the model is 0.93 and area under the curve is 0.169.

For Random Forest, the accuracy of the model is 95% with Kappa value of 0.74, sensitivity 0.67 and specificity 0.99. The AUC score for the model is 0.93 and area under the curve is 0.169. From the graph we see the that **going, last_login and chat_viewed** have the most impact on the prediction column.

Deployment



Both model performed good for the prediction form the train set with a similar accuracy. From the graph we see the that going, last_login and chat_viewed have the most impact on the prediction column which means that students who stated in the app that they are going to a particular university logged into the app frequently and chats viewed in the app, are most like to enroll in a particular university.