Ultimate Challenge

Part I

After loading the data and converted to csv file, I counted the logins with 15 minutes interval from 1970-1-01 to 1970-4-15. I plotted the data and it looks like this:

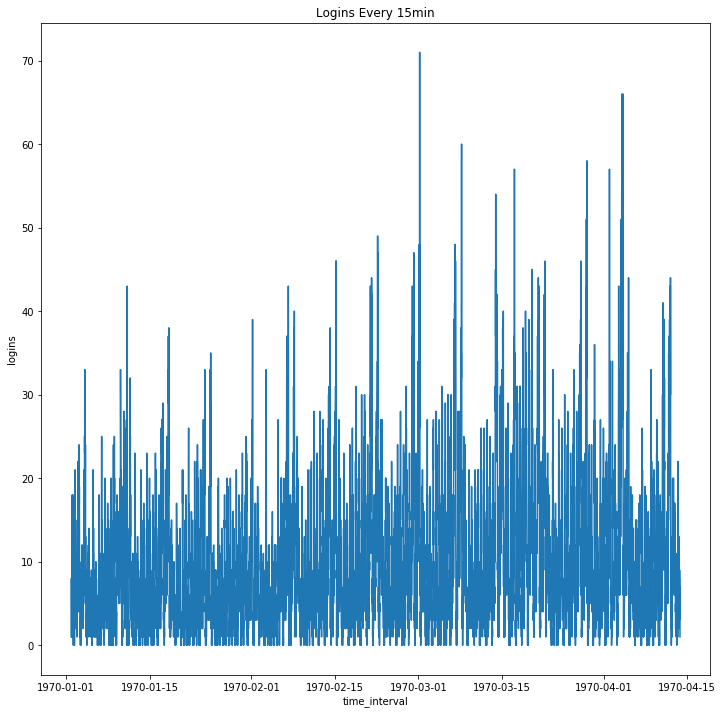


Figure 1. Logins Every 15 minutes

From the above graph, the logins are increasing from January to March. What I am also interested in is daily pattern of the logins. As a result, I filtered the dataset with logins greater or equal to 30. From these data, I grouped the peak logins into 4 groups that correspond to each stage of the day with 6 hours apart. What I discovered is that most of the highest logins reached when it is at early morning (12AM to 6AM). The following graph illustrated my findings:

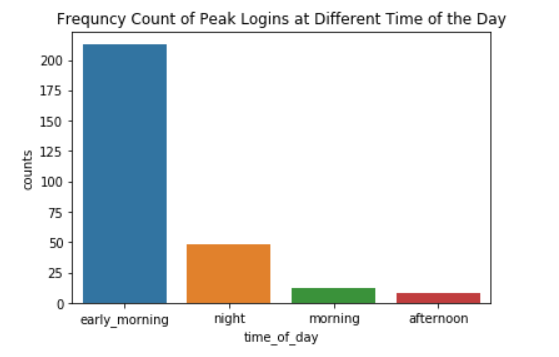


Figure 2. Frequency Count of Peak Logins at Different Time of the Day

Part II

1. I would choose the reimburse amount as the key measure metric in this experiment because the amount of reimburse increase significantly, that mean that more driver partners are traveling between the two cities. If not, that mean that the plan is not effective in encouraging driver partner to be available in both cities.

2. I would design a A/B testing. What I would do is reimburse toll costs for some day of the week and collect toll costs at other days of the week. I will utilize these data to perform A/B testing and figure whether the changes are significantly effective. If the result is effective, then I will recommend to the city to either change to the new plan, think of new plan, or stay with the old plan.

Part III

*Data Wrangling*

After loading the data and transform it into data frame, the dataset contains 50000 entries with 12 columns in total. I noticed there are missing values from columns “avg\_rating\_of\_driver”. “avg\_rating\_by\_driver”, and “phone”. For column “avg\_rating\_of\_driver” and “avg\_rating\_by\_driver”, I filled the null values with computed mean from each column. For the “phone” column, I simply dropped the rows with null values because I cannot compute unique phone number and there are only about 300 missing values out of 50000 entries. After cleaning the dataset, the result dataset contains 12 columns and 49604 entries. I also created a target variable that identify whether a client is active or not for the first month based on their trips in first 30 days. As a result, the final dataset for analysis contains 13 columns and 49604 entries.

*Data Visualizations*

In this part of the project, I mainly looked at the difference between active user and non-active user. The graph below illustrated the percentage of active and non-active users for the first month across all cities:

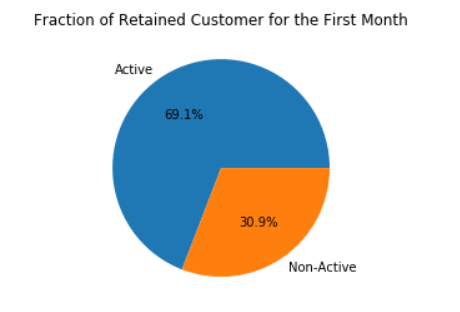


Figure 3. Fraction of Retained Customer for the First Month

The second graph illustrated active and not active client at each city:

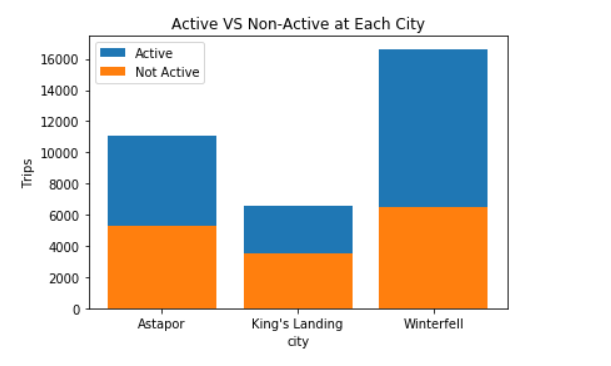


Figure 4. Active VS Non-Active User at Each City

The last graph depicts trips traveled at each city:

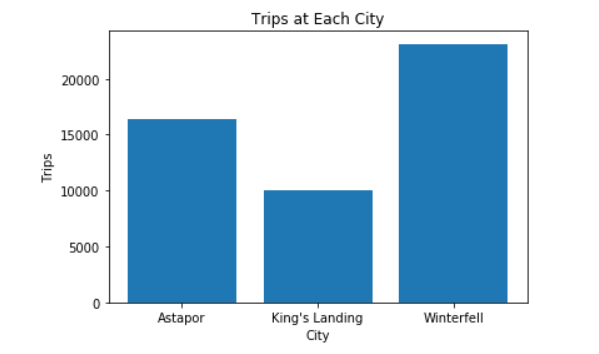


Figure 5. Trips at Each City.

*Machine Learning*

In this part of the project, I am going to feed the data into machine learning algorithms: Decision Tree, Logistic, and Random Forest to classify whether clients will be active after 6 months. Since the model I used can only take in numerical values, I change all the Boolean to 0 and 1. I also calculated the client’s lifetime on system from “signup\_date” and “last\_trip\_date” and assigned it as “delta\_days” on new column. Following that, I dropped the columns “signup\_date” and “last\_trip\_date”. I spilt the data into training and testing set. I fitted the train data on Decision Tree to find the feature importance. This is the result:

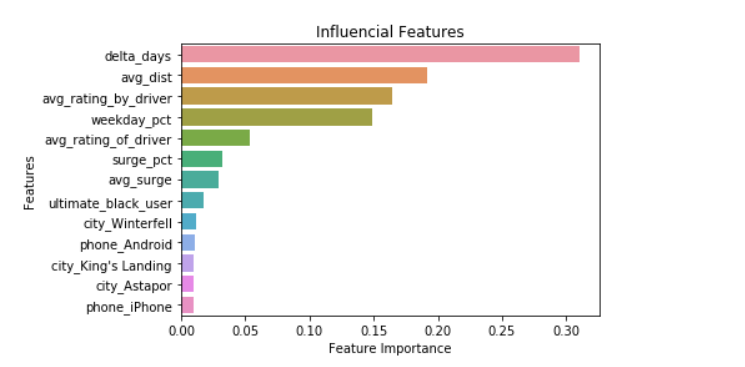


Figure 6. Influential Features

I computed the AUC score for Decision Tree model, and it is about 0.726. I also tried Logistic Regression model, before turning the AUC score is about 0.638 and after tunning it is about 0.634. The last model I tried is Random Forest model. The initial result has an AUC score of 0.878. After tunning, the AUC score went up to about 0.891. I also plotted the ROC curve illustrating that the best model to used is Random Forest following by Decision Tree; Logistic Regression is the worst model to use out of the three I chose.

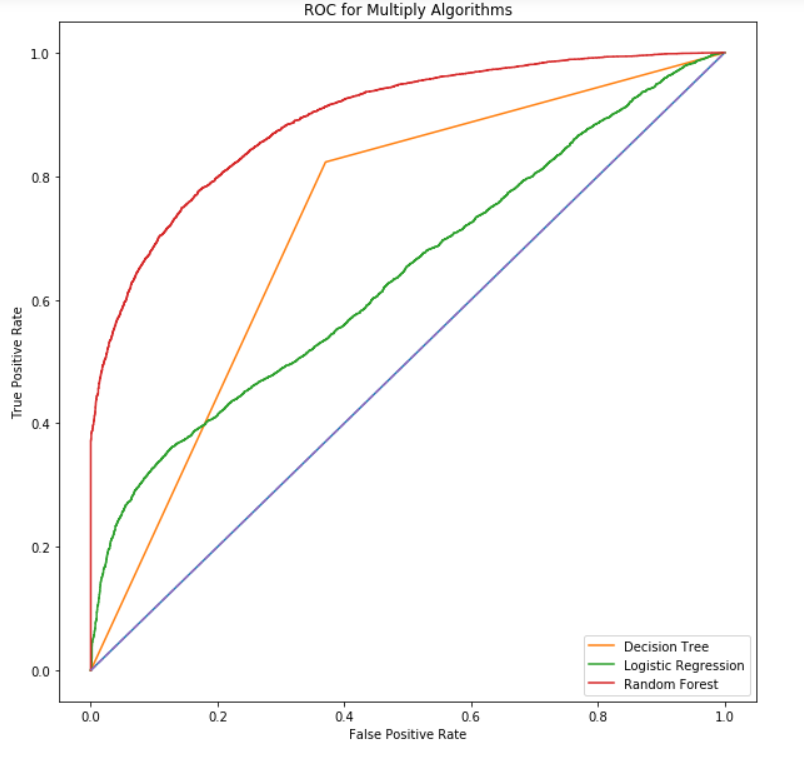


Figure 7. ROC for Multiply Algorithms

*Conclusion*

In conclusion, Random Forest is the best model to predict whether clients will be retained after six months. The top 5 most significant features that influences the models are “delta\_days”, “avg\_dist”, “avg\_rating\_by\_driver”, “weekday\_pact”, and “avg\_rating\_of\_driver”. Going forward to increase the retention rate I would recommend that the company could improve driver’s service. The company could also provide discount to clients periodically to increase “delta\_days”.