Ultimate Data Science Challenge Report

Springboard Data Science
Ishmail Grady
August 2019

Part 1: Exploratory Data Analysis

The data used for this exercise is was login timestamps for users between January 1st, 1970 and April 13th, 1970. From this data we can gain insight into the login patterns and volumes. Chart C1 and C2 shows the number of logins a week a January and a week in April in 15-minute intervals. These graphs illustrate the different in login behavior during weekdays and the weekend. On week days, the login pattern is approximately symmetrical. Logins are at a minimum (near 0) around 7:30 – 8:00AM, then logins rise quickly and peak around 12:00PM, and finally logins fall quickly back to their minimum around 4:30-5:00PM. This pattern makes sense in the context of the traditional 9-5PM work schedule; the significant events in the daily cycle correspond to the routine of the morning commute, lunchtime, and the evening commute. Login behavior on the weekends appears to be slightly more erratic. Times of higher login frequency are skewed towards the late afternoon and evening hours. The peaks occur around 3:00AM on the weekends. This behavior may be influenced by an increase in nightlife activity.

Table T1 shows the average hourly rate of logins for each month (data for April is incomplete). The average hourly rate in March and April is significantly higher than the rates in January and February. This could be an indication that login frequency may experience seasonal affects with more activity occurring in warmer seasons.

Part 1: Experiments and Metrics Design

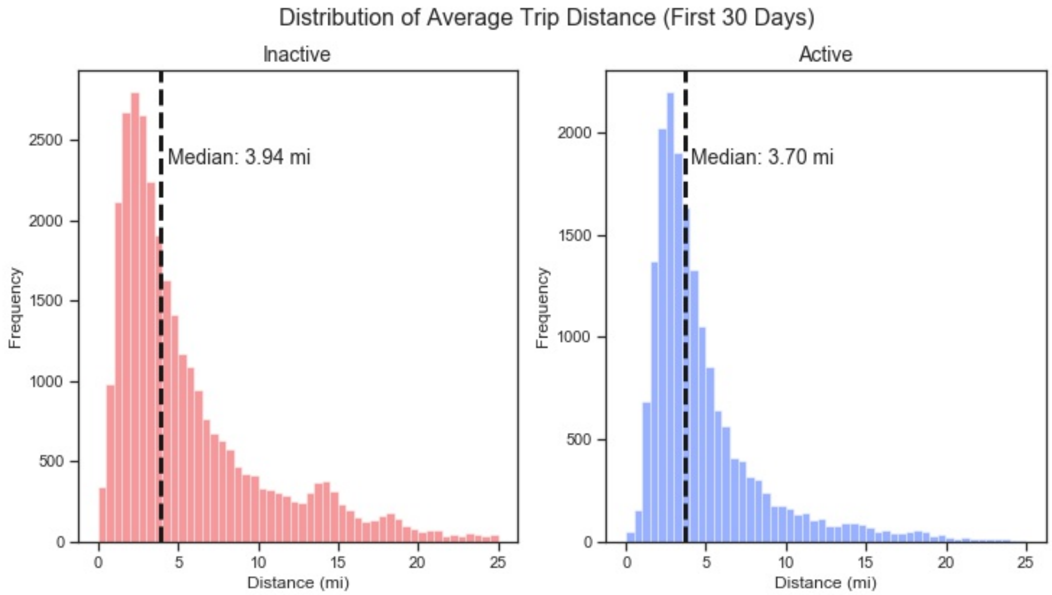
In order to best asses the results of the two-way toll experiment, the key metric of success should be the change in the mean number of trips that originate in one city and are completed in the other city per day, including only rides where the toll road was used on a weekend. This metric is good for this problem because it quantifies the specific activity of interest for accounts for confounding factors. A positive, negative, or no change in this metric will be able to accurately gauge the effect of the toll reimbursement.

To compare the effectiveness of the proposed change, an experiment should be conducted to run concurrently with a trial of the toll reimbursement policy. The data collection will run on consecutive weekends for as many needed to collect significantly more than amount needed for statistical significance. For each day, the number of trips originating in one city and completing in another city will be recorded. After data is collected, the metric will be calculated for the data from the experiment and historic value of the metric. A difference in means z-test will be used to test the hypothesis that the mean number of trips that originate in once city and complete in another is higher in the case where there is toll reimbursement.

In the case that the z-test indicates that the mean is higher with statistically significance, the change will be evaluated for practical impact. The impact on revenue and cost of toll reimbursement as well as other metrics related to traffic congestion and surges. If the increase in transportation is statistically and practically significant the recommendation should be given to implement the toll reimbursement. Otherwise, the experiment and data should be examined further to investigate why the change did not have the expected impact.

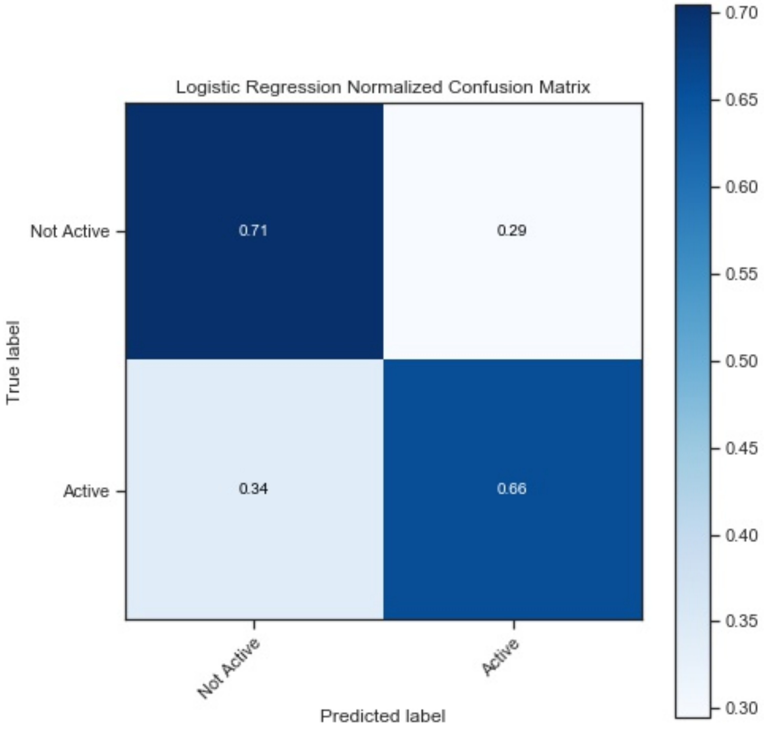
Part 3: Predictive Modeling

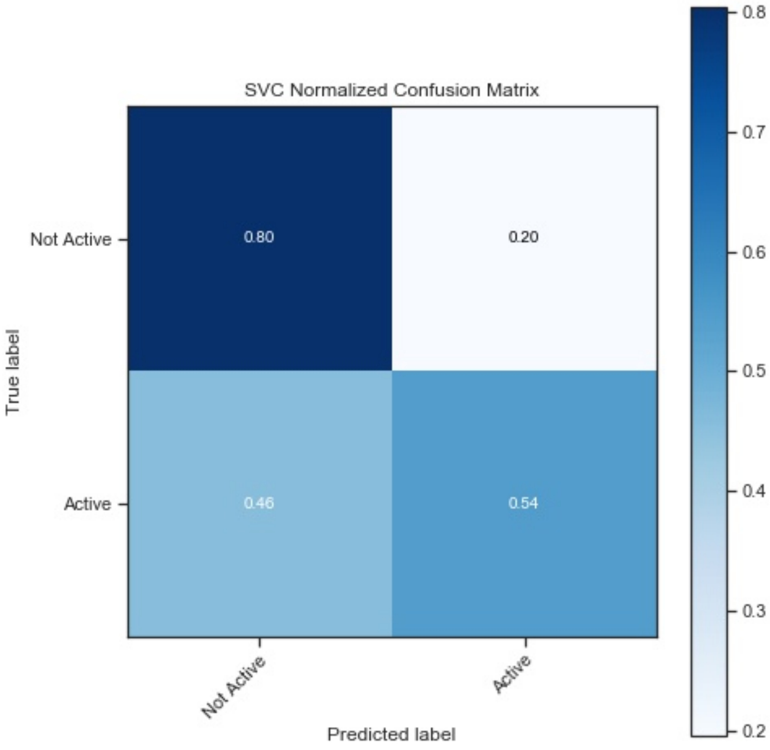
Of 50,000 users in dataset, 201 are missing values for their average rating by the driver and 8,122 users are missing their average rating of driver. Both of these features are skewed heavily towards higher values with the majority of users having a 5. Due to the skewed distribution, missing values were imputed used the median.



A user is considered active if they took a trip within 30 days of the latest date in the dataset. There are 31,690 inactive users and 18,310 active users. The chart above shows the distribution of average trip distance in the first 30 days of signup for active users and inactive users. The distribution is for both active uses and inactive users is skewed right and the median for inactive users is slightly higher.

Before training a predictive model. The numerical features were scaled, columns with skewed data were log transformed, and categorical variables were processed by one-hot encoding. Due to the imbalance in the response variable. The majority class, inactive users, was down-sampled by randomly sampling inactive users without replacement to match the minority class. Two models were trained, a logistic regression model and a linear support vector classifier. These models were used due to the ability to interpret the coefficients of the features. The metric that will be used to measure the performance of the model is the false positive rate. This was chosen under the assumption that falsely predicting that a user will continue to be active when they will not be would result in a missed opportunity to take actions to retain the customer and thus have biggest negative impact. A normalized confusion matrix for both models is shown below. The false negative rate for the SVC model was the lowest so this model was chosen as the final model. From this model we can extract the most important features. The top 3 features of the were indicator for living in the city of King's Landing, being an ultimate black user, and the number of trips in the first 30 days. All of these features have positive coefficients meaning that, according to the model, higher positive/true values increase the likelihood that a user will be active. To act on these results more research should be done to determine why retention in King's Landing is significantly higher than in other cities. Steps that also can be taken is target more users that would likely be ultimate black users and to have promotions to increase the number of trips a user takes in the first 30 days to increase retention.




SVC Normalized Confusion Matrix
True label
Not Active
Active
0.80
0.20
0.46
0.54
Predicted label
Not Active
Active
0.8
0.7
0.6
0.5
0.4
0.3
0.2