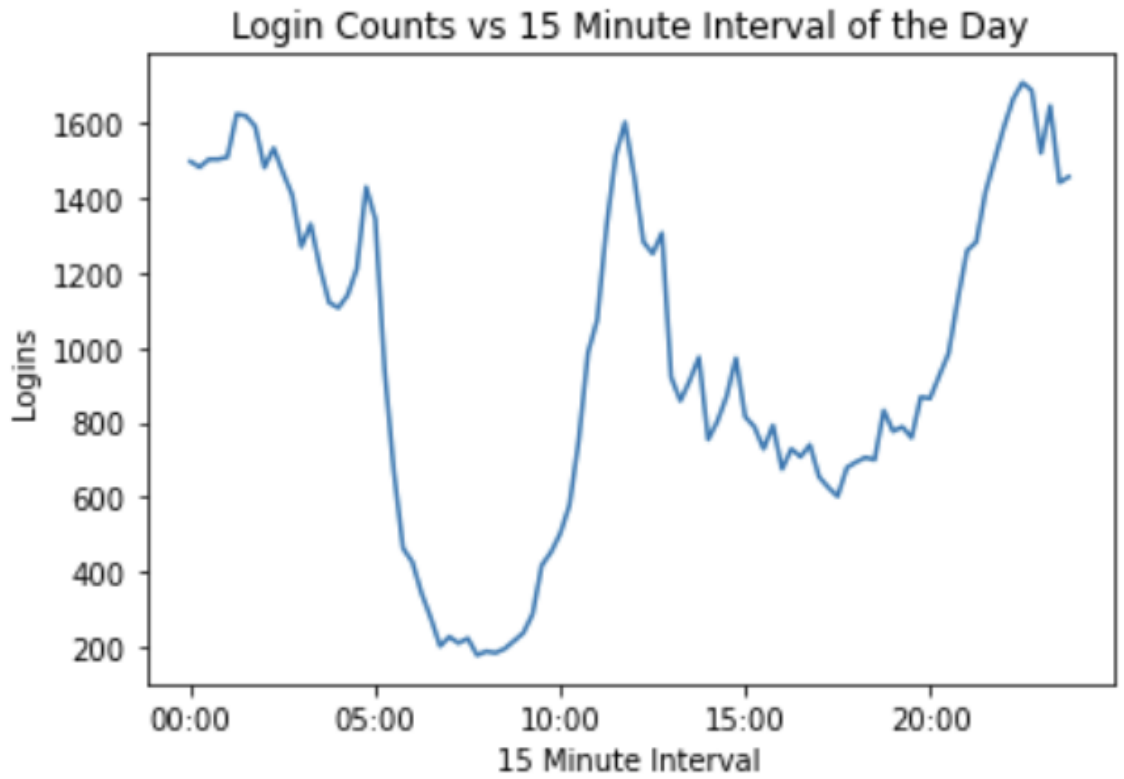
Ultimate Data Challenge
Edward Seymour

PART 1

The logins.json file was found to be cleaned and well organized. There are no data quality issues to report. After assuring that the data was clean, I removed the dates and grouped the logins by 15 minute intervals. I then got counts for each interval. Below is a graph showing the number of logins at each 15 minute interval throughout the day.



As you can see, there is quite a large drop off in the morning hours between the hours of 5:00 and noon. 5:00 in the morning and noon are some of the times that see the highest number of logins. Another drop in logins occurs in the afternoon and early evening, increasing again until the early hours of the morning. The average number of logins in any 15 minute interval is 970.

PART 2

The key measure for success in determining the effectiveness of this proposed change is to see whether increasing individual driver times in their respective alternate cities brings in more profit.

To measure this criteria, I would use an A/B test. In this test, half of the drivers from each city would be split up into groups, one group would have their tolls reimbursed (Group A) and the other half would continue to be exclusive to their respective cities (Group B). The groups will then continue driving, one group serving both cities and the other serving only one, for 6 weeks. After the 6 week period, the results can be investigated.

To investigate the results, a couple factors should be considered. First, how much time the drivers in Group A spent in their alternative cities and compare this to Group B. Did it actually increase the time they drove in the other city? If not, there might be another factor stopping the drivers from serving both cities. A T-test can be used to determine the significance of the results. Finally, the revenue should be considered, as it is the main key for success. The total profit from Group A can be compared against the profit from Group B. Again, A T-test can be used to determine whether the increase in profit is statistically significant.

If the increase in profit is significant, I would suggest moving forward with the proposal for all drivers. If not, I would investigate other ways of encouraging drivers to serve both cities, such as strategic surge pricing.

PART 3

Of the users who signed up in January 2014, about 37% are considered to be retained, i.e. they took a ride in the last thirty days.

After determining that ratio, I began cleaning the data. I filled in missing values in the ratings columns with their respective averages. I dropped rows that were missing phone data since it was only about 0.5% of the data. Finally, I looked at the distributions of the variables to ensure there were no erroneous values.

After cleaning, I performed PCA on a few of the variables. They all appeared to explain about the same amount of variance between all of them. I also looked at the correlation between the variables which showed very low auto-correclation. Next, I split the data into test and train sets and got a baseline accuracy based on a dummy classifier using the majority class (not retained) as the strategy for prediction. The baseline accuracy was around 62%.

The model that I decided to go with was a random forest classifier due to its power as an ensemble method and my familiarity with it. I performed a randomized search to determine the best parameters for the pipeline, which also includes a scaler and PCA. The features trips_in_first_30_days, avg_surge, ultimate_black_user, avg_rating_by_driver were found to be the most important to the model. The final model performed with an accuracy 77% on the test data, performing better than our dummy classifier in that metric.

Based on the important features found by the model, I would suggest focusing attention and promotion on ultimate black users. Focusing on this customer base will give a higher likelihood that the money is being spent on users that are more likely to provide returns on that investment.