Data Analysis Interview Challenge

Part 1 - Exploratory data analysis

Conclusion when looking at the 15 min intervals is that there seems to be the highest number of logins during:

-Night time and after 5 has the largest amount of logins and the lowest amount of logins are during the day time

Conclusion when looking at the daily intervals is that there seems to be the highest number of logins during:

-Saturdays have the largest logins, Mondays have the lowest logins, and logins increase as it gets closer to April

Part 2 - Experiment and Metrics Design

1. Ultimately success can be defined in this scenario by looking at the one measurable factor across both cities, the toll costs. Since riders tend to be exclusive to each city, with each city's riders only operating either during the day or during the night, we could loosely correlate the time of the toll as well to where the riders are coming from as well.
2. a. The metrics we would use are the mean number of tolls incurred increasing after the reimbursing of the toll costs go into effect for each set interval of time throughout the day. We would first collect data from the two-way toll regarding how many toll costs were incurred before the change for 24 hours, and then after the proposed change for 24 hours.
   b. We can use a bayesian A/B testing in order to see whether or not there is a significant difference before the proposed change and after the proposed change in the mean toll costs incurred across the duration of a 24 hour day.
   c. We can conduct this test with the null hypothesis that there is no difference before and after the proposed change, and if the p-value is low enough we can reject that null hypothesis and claim that there exists a correlation between the proposed change and the amount of toll costs incurred throughout the day.

Part 3 - Predictive modeling

1. Conclusion: Roughly 3/4 of the riders were considered retained, but at six month mark only 1/3 of them were considered active/retained. There also didn't seem to be any correlation between the amount of trips taken in the first 30 days versus how likely they were to remain active after 6 months.
2. I decided to use a boosted decision trees classifier as there were a lot of potential features involved and I was interested in seeing how they would all fit in to predict

whether riders were considered active at the 6-month mark. Since it was a decision of which features came up as important to this prediction, I followed it up with a feature importance test to see how each feature played into the prediction. At the moment I have a 78.58% accuracy score for how valid each prediction is at the moment. I had initial concerns of removing and including features due to how they visually presented when the data was initially graphed through the bar chart and scatter plot, as a few features showed that they didn’t seem to correlate much with whether or not the riders were retained at the 6 month mark. I also was interested in whether a fully connected neural network would perform better at prediction, but I felt as though the data needed feature selection first before I would approach it via deep learning, and hence the use of boosted trees.

3. The important finding that resulted from this was that there were clearly features that were incredibly more important than others when it came to rider retention in the 6th-month:

   city, #1
   trips_in_first_30_days, #4
   signup_date, #8
   avg_rating_of_driver, #6
   avg_surge, #7
   phone, #1
   surge_pct, #2
   ultimate_black_user, #1
   weekday_pct, #3
   avg_dist, #5
   avg_rating_by_driver, #1

   Seems like the most important features are: city, phone, ultimate_black_user, and avg_rating_by_driver. Which means that the city the rider is from, which phone they use, whether or not they are an ultimate_black_user, and what their average rating was, and the surge percent mattered more than other other feature provided. So much so that removing all other features lowered the validity of the prediction by less than two percent. This can be used to focus more on what about these features leads to the riders driving into the 6-month mark, such as what is it about the city they come from that causes them to lean towards riding long term, or perhaps marketing the ultimate black service they have in order to increase overall rider retention.