Jacob Kurstedt

Ultimate Data Challenge

**Part 1 ‑ Exploratory data analysis:**

Biggest Takeaways:

1. The most logins occur on the weekends with Saturday having the most logins by day over the 4-month period.

2. Logins were most frequent between the hours of 9pm-3am.

3. March the most logins by about 10,000 with February having the second most logins

The reason I think these were the biggest takeaways is because a lot of people go out on the weekends and people who are drinking are not going to drive since it is reckless thus they will call for a ride increasing the amount of logins on the weekends.

**Part 2 ‑ Experiment and metrics design:**

**QUESTIONS:**

1. What would you choose as the key measure of success of this experiment in encouraging driver partners to serve both cities, and why would you choose this metric?

2. Describe a practical experiment you would design to compare the effectiveness of the proposed change in relation to the key measure of success. Please provide details on:

a. how you will implement the experiment.

b. what statistical test(s) you will conduct to verify the significance of the observation.

c. how you would interpret the results and provide recommendations to the city operations team along with any caveats.

**ANSWERS:**

1. The metrics I would choose to measure the success of the experiment would be the distance and cost per ride. If there is a significant increase in the payout that the drivers get to traveling between cities compared to being exclusive to one city, I feel that drivers would be more inclined to want to drive between cities. Although getting reimbursed for all toll costs is a good incentive for the drivers, I feel that some drivers still might just be exclusive to the city they are in if the payout for traveling between cities is not worth it to them. Also, drivers driving the distances between the two cities will have to pay for gas more often so they will want to know that they will be getting paid enough to cover that. With that being said, if the distance and cost per ride increase it will show how successful the experiment was.

2. I will collect data on all the rides that are happening and keep track of the distance, cost, starting location, end location, date, toll cost, toll (if they went through a toll). I will then split the data up into rides that were between cities and rides that were within one city. Then I will calculate averages for cost and distance and find out the average cost per mile of each ride (cost / distance). I will perform two sample t-test, chi-squared test, and null hypothesis test to find the significance of my observations. I will interpret my results by creating graphs to show the difference in costs and distance between rides between both cities and rides within one city. Then present my findings in PowerPoint form**.**

**Part 3 ‑ Predictive modeling:**

**QUESTIONS:**

1. Perform any cleaning, exploratory analysis, and/or visualizations to use the provided data for this analysis (a few sentences/plots describing your approach will suffice). What fraction of the observed users were retained?

2. Build a predictive model to help Ultimate determine whether or not a user will be active in their 6th month on the system. Discuss why you chose your approach, what alternatives you considered, and any concerns you have. How valid is your model? Include any key indicators of model performance.

3. Briefly discuss how Ultimate might leverage the insights gained from the model to improve its long-term rider retention (again, a few sentences will suffice).

**ANSWERS:**

1. I first had to fill in missing values for avg\_rating\_of\_driver and av\_rating\_by\_driver I filled these in using the mean. I then changed signup\_date and last\_trip\_date into datetime objects. Also dropping NA values in the phone column. I found that most people had iPhones, Winterfell had the most people using Ultimate, and most people were not ultimate black users. Lastly, out of the 49604 users, 36.66% (18180/49604) of them were retained.
2. My approach was to try different models to see which one preformed the best. I performed a Logistic Regression Model, Random Forest Model, and Gradient Boosting. Logistic Regression had an accuracy of 0.64, Random Forest had an accuracy of, 0.71, and Gradient Boosting preformed the best out of the three with an accuracy score of 0.75 so I choose to use Gradient Boosting further. After tuning the parameters for Gradient Boosting it had an accuracy score of 0.75 and the AUC score was 0.79. After parameter tuning, I looked at the feature importance and found that avg\_rating\_by\_driver, avg\_distance and weekday\_percentage were the three most important features.
3. In order to improve long term rider retention, I would recommend that Ultimate makes sure it is able to retain drivers with higher ratings since it was the most important feature and I believe this is so because the riders feel safer with drivers who have high ratings. Next, I would recommend that Ultimate tries ways to encourage riders to take longer rides since the average distance was the second most important feature. Lastly, having riders take more trips during the weekdays whether that be lowering fares during the weekday or encouraging people to use Ultimate to get to work could be some ideas Ultimate could try to implement.