Ultimate Technologies Take Home Challenge

**Exploratory Data Analysis:**

When aggregating & plotting the number of logins by hour of the day, we can see that the more logins are either occurring late at night between 9:00 PM – 2:00 AM or in the middle of the day around 11:00 AM. When aggregating & plotting the number of logins by day of the week, we can see that more logins are occurring at the end of the week/during the weekend between Friday-Sunday. Finally, when aggregating & plotting the number of logins by month, we can see that logins are gradually increasing between January & March.

**Experiment & Metrics Design:**

**Question 1:**

I would choose customer wait time as the metric to be used in determining measure of success of the experiment to encourage driver partners to serve both cities. I would choose this metric because a reduced customer wait time would indicate that driver availability has increased within a city & that there is a better overall distribution of drivers to serve all customers as well.

**Question 2:**

When it comes to implementing an experiment to test the effectiveness of toll reimbursements to drivers on customer wait time, I would first begin by establishing a time frame. Within that time frame, I would alternate between giving & not giving drivers toll reimbursements & collect data on customer wait times under each event.

Using the data collected, I would generate bootstrap samples for both datasets (customer wait times with toll reimbursements & customer wait times without toll reimbursements) & calculate a distribution of the average wait times for both datasets. Next, I would use a hypothesis test with the null hypothesis being that there is not a significant difference between average wait times of both datasets. I would evaluate what % of the average wait time distribution of the reimbursed tolls dataset falls outside of the 95% confidence interval of the average wait time distribution of the non-reimbursed tolls dataset. This will allow me to determine how likely it is that reimbursing tolls to drivers affected the average wait time of customers.

When interpreting the results of the experiment, I would focus on the cities individually as well as together along with specific days/times. We know that the ride demands fluctuate in each city during the week with Metropolis being the more active city during the day & Gotham being the more active city at night. With that in mind, I would want to see if average wait time decreased in Metropolis during weeknights & decreased in Gotham during the day. This would show us if more drivers are providing rides in these area during those times despite lower demand. I also want to see if wait times decreased within those cities during their busy times as well as this would show us if more drivers are going to provide more rides in the other city during peak demand times. If average customer wait times are decreasing without a loss in profits or customer satisfaction due to changing driver availability, I would recommend that Ultimate reimburse drivers for tolls.

**Predictive Modeling:**

Overall, 37.61% of observed users were retained/active. Data wrangling steps taken included filling in missing values that were found in the avg\_rating\_of\_driver, phone, & avg\_rating\_by\_driver columns. I then created the boolean variable ‘active’ to show if a user was still active or not based on whether they had taken a ride in the last 30 days. Finally, I plotted multiple variables I thought would be correlated with whether a user was still active or not to see if they actually were. ‘City’, ‘trips\_in\_first\_30\_days’, & ‘ultimate\_black\_user’ were variables that showed significantly different values between active & non-active users.

To build a model to predict whether a user is still active or not, I decided to use a Random Forest Classifier as this is an effective model when it comes to labeling binary values such as ‘active’ accurately. The standard ‘out-of-box’ random forest model had an accuracy score of 74.87% on the testing dataset but only identified 3,448 out of the 5,623 users that were still active correctly. I then implemented a tuned Random Forest Classifier model to see if it could do a better job of identifying users that were still active. The tuned random forest model had an accuracy score of 78.45% on the testing dataset & was able to identify 3,664 of the users that were still active correctly. I then compared the feature importances within both models to see which variables were more important in identifying the users that were still active correctly.

Ultimate can leverage the insights gained from my analysis to improve long-term rider retention by focusing in on the variables shown to be most important in keeping users active. Through my exploratory data analysis & model results, Ultimate will see that focusing their attention & efforts towards specific cities, ultimate black users, getting riders to take more trips in the first 30 days, & surge percentages, they will be able to retain more users.