Ultimate Take Home Challenge

Part 1

See Data Exploration Notebook

Part 2

1. The most direct measure of success would be to see how much the driver partners had to be reimbursed. A reimbursement would mean that that specific driver partner had traveled to the city that they would not have operated in previously due to the toll. A less direct but possibly more meaningful measurement would be to see if the driver partner activity from each city flattened out over time. The reimbursement of the toll should mean that the driver partners can operate in either city whenever they want, meaning that their schedules do not necessarily have to follow the circadian rhythm of the city they live in anymore.
2.
    a. I would design an experiment where the driver partners' time in each city was directly measured via gps/geo-tagging. I would give a small subset of driver partners the toll reimbursement deal for the duration of the experiment and compare the average time spent in each city for that subset to the time spent in each city of the rest of the driver partners.
    b. Mostly I would just compare the averages between groups. If you needed to go further than that you could always have multiple different subsets of driver over a longer amount of time.
    c. Interpretation of the results of the experiment could be done with automated reports on a daily, weekly, monthly etc. basis. You could also implement an internal, live dashboard where you could see a live count of the driver partners available in each city.

Part 3

See Predictive Model Notebook.

1. Data wrangling consisted of converting data from json to a pandas dataframe, creating a target label column by comparing the signup_date and last_trip_date, one-hot-encoding categorical features, and handling NA values by filling them in with the average value of the rest of the rows for that feature. Approximately 36.5 users in the training dataset were retained.

2. For this task I chose to use a random forest model. I chose this type of model for its accuracy, relatively quick training time, and its ability to provide features importances after model training. Once it is learned what features are important, one may want to do more feature selection/engineering and then move on to a deep learning approach. Using a random forest model I was only able to achieve accuracies in the high 70s.
3. According to the random forest model, the top most important features were ‘avg_dist’, ‘weekday_pct’. One could fathom then, that Ultimate may want to more carefully target users who commute during the week and have a relatively large distance to cover.