Part 2 - Experiment and metrics design

The neighboring cities of Gotham and Metropolis have complementary circadian rhythms: on weekdays, Ultimate Gotham is most active at night, and Ultimate Metropolis is most active during the day. On weekends, there is reasonable activity in both cities.

However, a toll bridge, with a two-way toll, between the two cities causes driver partners to tend to be exclusive to each city. The Ultimate managers of city operations for the two cities have proposed an experiment to encourage driver partners to be available in both cities, by reimbursing all toll costs.

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?

- The key measure of success would be the location of the driver partners on weekdays. If drivers are in Metropolis during the day and Gotham at night, then that would mean that the tolls are no longer an impediment to drivers working in both cities.

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

1. I would first collect data as to where Gotham based drivers are during the daytime and where Metropolis based drivers are at night. I would gather data on average wait times in each city during peak hours.

2. I would give a sample size of drivers that are based in each city, the ability to get tolls reimbursed and a way to track them. Then I would collect data of the location of the Gotham based partners during the daytime and the Metropolis based drivers during the nighttime.

3. Finally, I would compare the data sets to see if the toll reimbursement is incentivizing the drivers to work in both markets and to what degree.

4. If the analysis is favorable, I would increase the toll reimbursement program so that both markets have the driver partners they need.

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

1. I would use kmeans to do cluster analysis comparing where the partners from each city cluster during peak times of each city before and after reimbursement.
2. I would use Ordinary Least Squares regression to compare wait times before and after toll reimbursement in each city during peak times.

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

1. If the cluster data shows that partners are clustering in a city's peak time, that would be proof that the tool reimbursement has worked.
2. The wait times should be reduced with more partners in a city during peak demand times.
    a. The caveats:
        i. Don't want to saturate a market.
        ii. The home market cannot have increased wait times.

Part 3 - Predictive modeling

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?
    a. Plots and analysis can be seen on the Jupyter Notebook.
    b. According to the data, 36.6 % of customers were retained for 30 days. Only .17% were retained for 6 months.

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.
    a. I approached this as a classifier problem. I created a binary column ('retained_180') which was based on if the difference between the last ride and the sign-up date was at least 6 months. The data was very low, only .17 % of customers used the service. Because of this, I used SMOTE to increase and balance the data. I tried several classifier models with Logistic Regression ultimately delivering the best results. The most important feature was the 'sign-up date'. This seemed trivial to me so I removed it from the data and ran it again. The model improved so I kept removing the most important feature until the model stopped improving. I then did that with the least important feature until the model stopped improving. Finally, I did a grid search to see if different parameters yielded better results but it did not.

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).
    a. Based on the data, the top 4 most important features were:
        i. Astapor
        ii. Avg_dist
        iii. King's Landing
        iv. Surge_pct

Looking at the data, the Average distance between 6 month customers and not, the mean distance is not that different. The surge pct is higher with customers that were retained than not. Customers that were retained for 6 months used the service more in the first 30 days.

My recommendation to the company would be to incentivize customers to use the service in the first 30 days.

The fact that the surge % is higher with retained customers could indicate several things.

1. Prices are too low for what customers perceive to be a fair price. Thus thinking that the service will be inadequate.
2. Loyalty is built during busier times. Advertising in high traffic areas might have benefits.