**Ultimate Technologies Project Report**

Part 1 ‑ Exploratory data analysis

First of all, I read the json file into dataframe and check whether there is missing value. The answer is that there is no missing value in this file. Secondly, I aggregate the number of logins based on 15 minutes interval. When I plot the autoregression function, the plot shows that dataset has seasonal pattern in every 50 lags (about every 12 hours). The pcf plot suggests that the demand will go down in every 6 hours and then go up in the next 6 hours. Thirdly, I decompose the data. The “trend” plot indicates that there is neither upward nor downward trend in the data. In other worlds, that data is stationary. Moreover, the “seasonal” plot shows that there is a strong seasonal pattern in the data (as described above).

Part 2 ‑ Experiment and metrics design

1. I would measure the increase of the number of rides in Gotham during daytime and that of Ultimate Metropolis at night. The reasons why I choose this measurement are as the following. The ultimate goal of this experiment is to encourage driver partners to be available in both cities (i.e. drivers go in between these two cities). If the experiment succeeds, the rides in both Gotham and Ultimate Metropolis would increase. Therefore, I would measure the increase of the number of rides in both cities. Regarding weekends, as there are reasonable activities in both cities, driver partners would probably stay in their own cities instead of going to the other one, hence I think there is no need to measure the number of driver partners going to another city.
2. A) first, I will keep the tolling system as usual and collect the number of rides in both cities, Gotham during daytime and Ultimate Metropolis at night. In the following week, I will change to the desire tolling system and collect the same data.

B) I will conduct two tails t test to verify the significance of observation. The null hypothesis is the average daily rides before changing the tolling system is the same as that after changing the tolling system. The alternative hypothesis is average daily rides before changing the tolling system is smaller than that of after changing the tolling system Based on the result of p value, I will be able to determine whether to reject or not reject the null hypothesis.

C) If p value is smaller than 0.05, then we have enough evidence to reject the null hypothesis. In other words, it is statistically significant to conclude that the average daily rides before changing the tolling system is not the same as that after changing the tolling system. If F value is negative, we would be able to conclude that average daily rides before changing the tolling system is smaller than that after changing the tolling system, and vice versa.

Part 2 ‑ Predictive model

1. 37.6% of the observed users are retained.
2. First of all, this is logistic regression problem because we are trying to predict whether a user will be active or not. Hence, I used SK-learn logistic regression model with gridsearch method to train and predict the dataset. It turned out the accuracy score is 0.72. Furthermore, I used a decision tree model (SK-learn random forest) to train and predict the dataset, the accuracy score increased to 0.9992. This model may have overfitting problem as the accuracy score is so high.
3. The feature importance analysis for the random forest model indicates that when a user taking his/her last trip is the most important feature for prediction, with a significant margin. This insight of analysis tells the company that once a user leaves for a while, he/she will never come back. Therefore, encouraging a user repeat order in short interval is vital for the business. Moreover, necessary actions should be taken to re-engage the users if the company finds out that some users have left for a certain period.