Data Analysis Interview Challenge

Ultimate is interested in predicting rider retention. The Ultimate company have provided a sample dataset of a cohort of users who signed up for an Ultimate account in January 2014. The data was pulled several months later; the Ultimate company consider a user retained if they were "active" (i.e. took a trip) in the preceding 30 days.

Part 1 – Exploratory data analysis

User-logins increase during late night 9PM to 4AM and at middle of the day: 11AM to Noon:

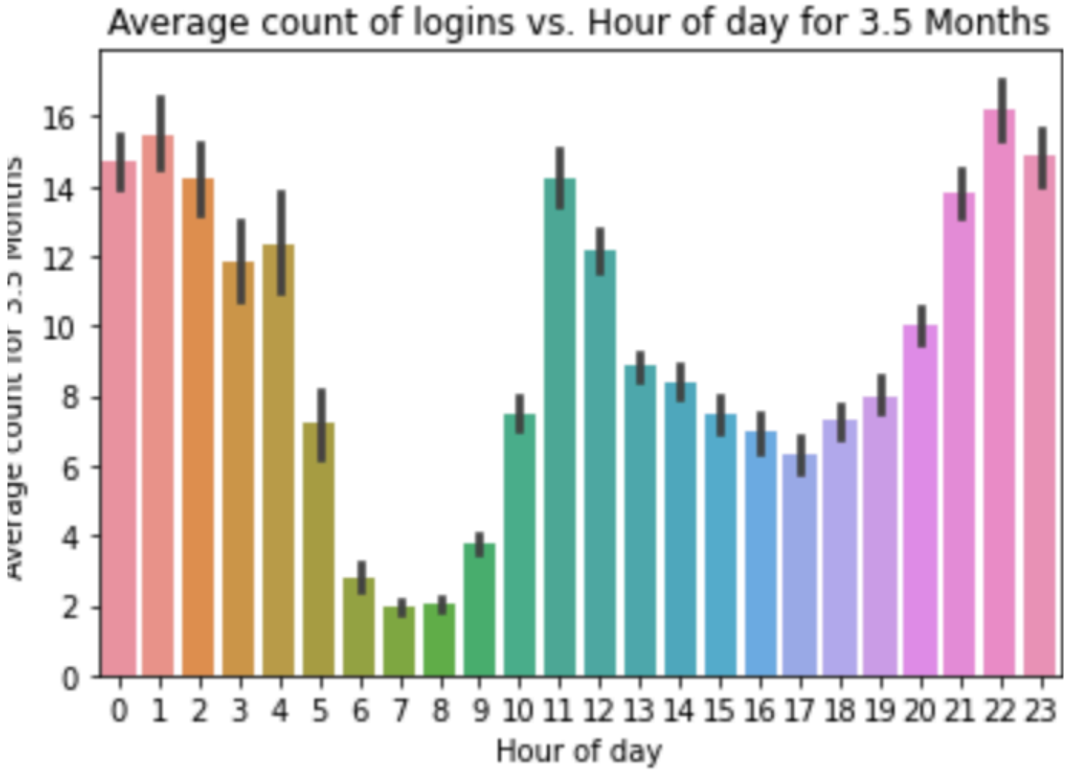
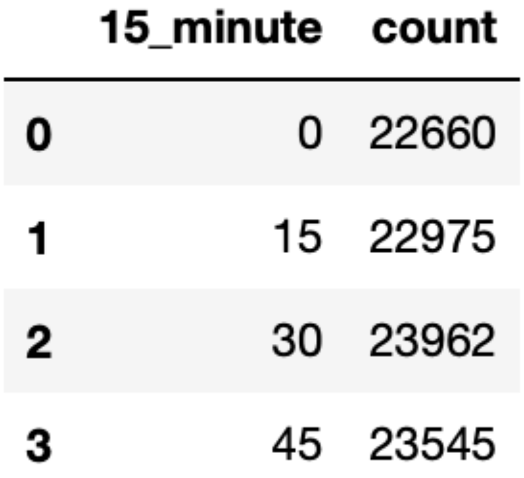


Fig. 1. Mean count of logins versus Hour

There is not much change for 15 minute intervals:



	15_minute	count
0	0	22660
1	15	22975
2	30	23962
3	45	23545

Fig. 2. User count for each 15 minutes in an hour

Part 2 - Experiment and metrics design

The dataset contains the feature "avg_dist," which may be interpreted to be the average distance a driver travels per trip per day. The feature, avg_dist, is computed by taking the mean of all the trips traveled in a given day. If drivers are available for both cities, then the average distance traveled will increase per day. Let the distance from the Gotham city center to the Metropolis city center be D1. If the driver exceeds D1 per day, then avg_dist > D1, thus the driver serves both cities. If the driver travels less than D1 distance per day, then it is unlikely that the driver serves both cities.

The Null Hypothesis is avg_dist < D1 and the Alternative Hypothesis is that avg_dist > D1. The key measure of success is to reject the Null Hypothesis. The degrees of freedom equal N-1(N, sample size is the number of trips taken by the driver per day). To implement the experiment, the driver needs to record the distance of each trip in a given day. The scientist must find out what is D1, and perform a t-test in which the Null Hypothesis: avg_dist - D1 < 0, and the Alternative Hypothesis: avg_dist - D1 > 0. To increase the accuracy of the test, the recorded avg_dist must be computed for more than one day, and more than one driver should participate in the test. The scientist must report whether or not the Null Hypothesis is rejected. If the Null Hypothesis is rejected, then the scientist may recommend that the experiment is a success. To ensure a definite success or failure of the experiment, many drivers and many days need to occur. The caveat is that the drivers are sampled from the entire population of drivers and the number of days is also a sample.

Part 3 - Predictive modeling

36.62 percent of users were retained. We used XGBoost as well as a neural network. The accuracy for XGBoost training data was 0.9698 and for the test data the accuracy was 0.9583. For the neural network, the confusion matrix accuracy was computed to be 0.9580. The features most correlated with 'retained' were city_1 (Winterfell), phone_1 (iPhone), and trips_in_first_30_days.