Data Analysis Interview Challenge

This is your chance to wow us with creative and rigorous solutions! Please include your code at the end of your submission, or in a separate file. We also accept incomplete solutions.

Part 1 - Exploratory data analysis

The attached *logins.json* file contains (simulated) timestamps of user logins in a particular geographic location. Aggregate these login counts based on 15minute time intervals, and visualize and describe the resulting time series of login counts in ways that best characterize the underlying patterns of the demand. Please report/illustrate important features of the demand, such as daily cycles. If there are data quality issues, please report them.

> This exercise is titled eda.ipynb and is in the "ultimate_challenge" folder

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?

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

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

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

Note: The two cities of Gotham and Metropolis are not in the provided dataset; however, you do not need this information to answer Part 2.

Part 3 - Predictive modeling

Ultimate is interested in predicting rider retention. To help explore this question, we 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; we consider a user retained if they were “active” (i.e. took a trip) in the preceding 30 days.

We would like you to use this data set to help understand what factors are the best predictors for retention, and offer suggestions to operationalize those insights to help Ultimate.

The data is in the attached file ultimate_data_challenge.json. See below for a detailed description of the dataset. Please include any code you wrote for the analysis and delete the dataset when you have finished with the challenge.

Exercise is in the retention.ipynb file in the ultimate challenge folder

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?

   36%

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.

   Xgboost used for large dataset with mixed features needing easy evaluation. Accuracy 78%, ROC AUC Score 76%

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).

Ultimate may be interested in digging into the differences in service/competition in King's Landing versus the other cities. King's is the most predictive. Anything that might be emulated in other cities could be useful. Also, iPhone users seem less likely to be retained. Perhaps the iPhone app is less convenient to use and improvements could be made there.

Data description

- **city:** city this user signed up in

- **phone:** primary device for this user

- **signup_date:** date of account registration; in the form ‘YYYYMMDD’

- **last_trip_date:** the last time this user completed a trip; in the form ‘YYYYMMDD’

- **avg_dist:** the average distance in miles per trip taken in the first 30 days after signup

- **avg_rating_by_driver:** the rider’s average rating over all of their trips

- **avg_rating_of_driver:** the rider’s average rating of their drivers over all of their trips

- **surge_pct:** the percent of trips taken with surge multiplier > 1

- **avg_surge:** The average surge multiplier over all of this user’s trips

- **trips_in_first_30_days:** the number of trips this user took in the first 30 days after signing up

- **ultimate_black_user:** TRUE if the user took an Ultimate Black in their first 30 days; FALSE otherwise

- **weekday_pct:** the percent of the user’s trips occurring during a weekday