Data Challenge
for Ultimate

Table of Contents

Data Challenge
for Ultimate

Objectives:

The objectives for this report are to respond to Ultimate's request for...

1. Exploratory Data Analysis
2. Experiment and Metrics Design
3. Predictive Modeling

About the Data:

Ultimate has provided two data files:

1. *logins.json* file contains (simulated) timestamps of user logins in a particular geographic location.

2. *ultimate_data_challenge.json* to help understand what factors are the best predictors for retention (rider within the last 30 days)

Part 1: Exploratory Data Analysis:

Background:

Ultimate has requested the following be completed...

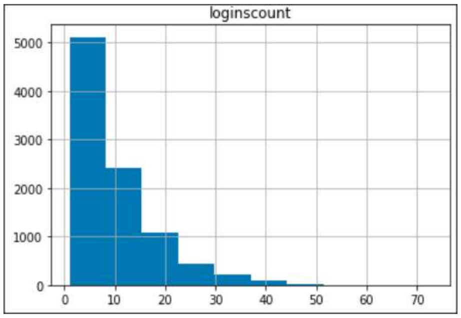
1. Aggregate the login counts based on 15 minute time intervals
2. Visualize and describe the resulting time series of login counts in ways that best characterize the underlying patterns of the demand.
3. Report/illustrate important features of the demand, such as daily cycles.
4. If there are data quality issues, report them.

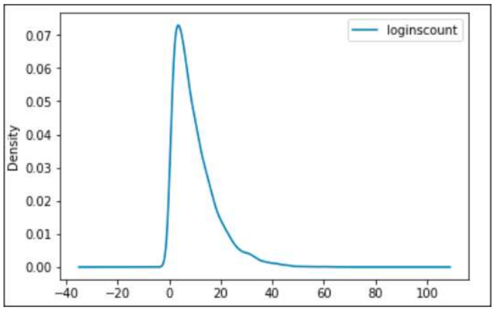
Data description and visualization for trends:

1. No missing data: all entries have a date time value. The only modification was to sort the values.

2. There are almost 9,400 fifteen minute intervals (equates to 2,350 hour groupings)

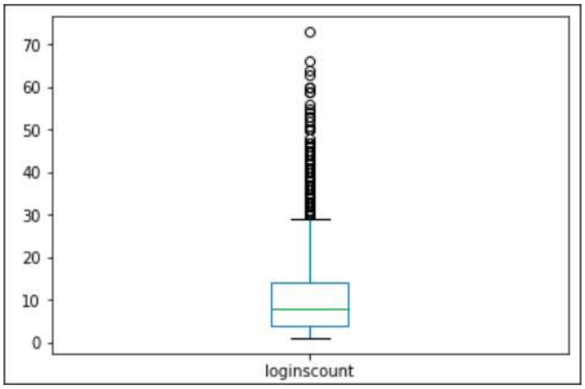
3. A LoginsCount histogram shows what appears to be an exponential distribution and further analysis via a density plot seems to confirm this.





4. The number of logins per 15 minutes period ranges from 1 to as high as 73 with a mean of around 10 logins.

5. The boxpot below highlights that there are a number of outliers...



Part 2: Experiment and Metrics Design:

Background:

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.

Question #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?

There are several possible measures that I'd consider for not paying tolls vs paying tolls...

Green = top metric to use (if feasible the first 3 metrics should all be utilized)

#	*Proposed Metric*	*Reason Metric is Proposed*	Notes
1	Increase in average duration of trip per driver	the longer the trip the higher the revenue/profit	adding paid access to a 2nd city should increase average duration of trips if this truly is an incentive for drivers (i.e. drivers may now recommend locations in the next city, when they might not have previously.
2	Increase in number of riders per driver	number of riders is used in the price calculation	
3	Increase in number of trips per driver	more trips = more revenue = more profit	to ensure number of trips is due to paying tolls, monitor trips that have tolls and trips that do not have tolls
4	Increase in number of tolls paid (if this was tracked prior)		Not a good measure – does not measure a variable that relates directly to profitability of the business.

Data Challenge
for Ultimate

Question #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 the experiment, how to implement, statistical tests to use, and interpreting the results.

Area...	*Description*
Experiment	Select a random subset of drivers (at least 25% of drivers) to receive toll reimbursement for a duration of at least 1 month and less than 4 months. For purposes of this response, assume there are no labor law violations for fair & equitable treatment of employees with this approach. If there are potential conflicts, modify this experiment to be a duration of time for all drivers to receive reimbursement and reconsider how to implement, statistical tests to use, and results interpretation.
Implementation	Select a time period where you have at least one month with and without a major holiday, assuming at least one city stages special events around major holidays, i.e. Christmas Tree lighting, fireworks on July 4th, summer ethnic festivals, etc... Validate that you have corresponding data to compare the results to, i.e. same month last year, equal time period prior to the experiment, etc... Be transparent – communicate to drivers that you are conducting this experiment and have a desire to increase their success (rides and income) as well as the companies success (we're all in this together mentality)
Statistical tests	By definition the Experiment and Implementation just described is employing an "*A/B Hypothesis test"* approach. *Produce visualizations* showing average durations for company overall, per driver, and per month within the experiment timeframe. *Clustering algorithms* – look for patterns, such as an increase in number of rides where the pick-up location or drop-off location is within a distance (< 3 miles?) close to the toll bridge. *Statistical Inference:* Complete a hypothesis test to assess if the H0 hypothesis of "Revenue increases (> 0.05) when tolls are reimbursed" is statistically significant from the H1 hypothesis of "Revenue does not increase significantly" Complete a similar HO, H1 hypothesis test using profit instead of revenue.
Interpreting results	A "home run" would be discovering that revenue (and profit) increased by a statistically significant margin from the prior periods (prior month and same month prior year), and that there is a clustering pattern for pick-up or drop-off locations. An "okay result" (i.e. deemed good enough to implement reimbursement of tolls) might be results that show a significant revenue increase, but not a large (or no) profit increase. The

	theory behind this result being that over time increased awareness among drivers could cause profit to increase and risk of profit decreasing is deemed very low. A “poor result” is no measurable change seen via the A/B test, visualizations, no clustering pattern, and no statistical significance. In this case, do not move forward with toll reimbursement as it presents too much risk (at least in the short term) to profit.

Part 3 Predictive Modeling

Background:

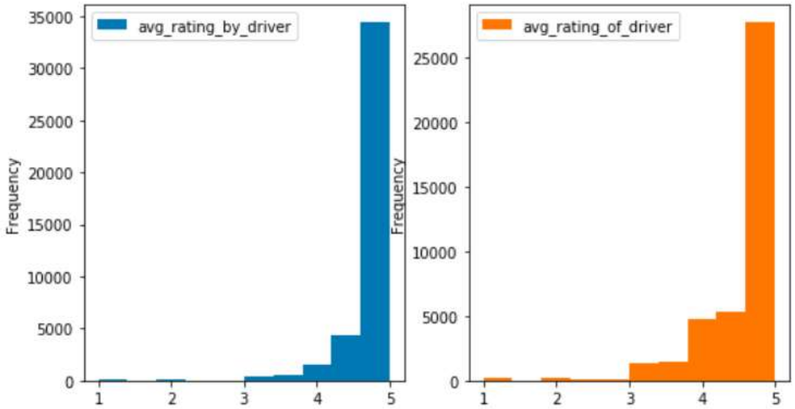
Ultimate is interested in predicting rider retention. To help explore this question, they have provided a sample dataset of a cohort of users who signed up for an Ultimate account in January 2014. The data was pulled several months later (June 2014) and they consider a user retained if they were “active” (i.e. took a trip) in the preceding 30 days.

They would like this data to be used to help understand what factors are the best predictors for retention, and offer suggestions to operationalize those insights to help Ultimate. They would like any code written for the analysis to be included and want the dataset deleted when the challenge is finished.

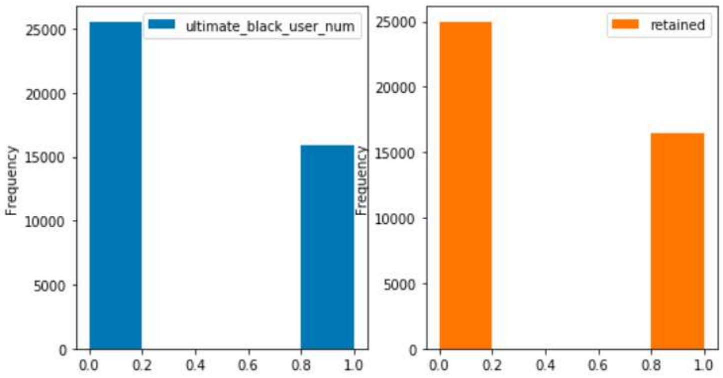
Cleaning, Exploration and Analysis

1. Of the 50,000 rows in the dataset, 8,555 have at least one missing value. Decided to delete these rows from the data because doing so still leaves 41,445 rows for visualization and machine learning.
2. Retention Rate: all riders signed up in January and the last date in the “last_trip_date” feature is July 1st. So a rider is considered “retained” if they took at least one trip in June.

   40% of riders who signed up in January were still retained in June (16,506 riders/41,445 signed up)
3. Histograms of avg ratings seem reasonable show a tendency towards drivers rating their riders high, and vice versa. This is a positive sign for overall company morale, professionalism of drivers, courtesy of riders, etc...



4. An interesting data point is that Histograms of “Retained” and “Ultimate Black” are very similar. This was reenforced later on in this analysis by machine learning results that listed Ultimate Black as a significant feature to predict retention.



Analyze the Features

Two models were employed (RFE and Feature Importance) to assess which features contribute the most to predicting retention; yielded results of...

1. For RFE: Ultimate Black, City, and Phone Type
2. For Feature Importance: Average Distance, Trips in First 30 Days, Average Rating By Driver, and Weekday Percentage

Predictive Model

Choosing a baseline: We know from prior analysis that the retention rate across the dataset is 40%. So, to do better than "prevalence", the model should acheive an accuracy higher than this. The two measures that will be used for each algorithm for scoring are accuracy and area-under-the-curve (AOC)

Appoach: The approach is to run three models (table below) using cross-fold validation, that are known for performing reasonably well in classification problems (in this study customer retained or not retained). The models are run twice...

1. once using all the features (except the sign up and last ride date)
2. second using only features that are the strongest contributors listed above under "Analyze the Features"

This intent of this approach is to ensure sufficient variety, while seeking to minimize risk of over-fitting the model.

Results/Validity: The table below shows the accuracy and area-under-the-curve scores for each model

# features used	*Algorithm*	*Accuracy Score*	*AOC Score*
10	LogisticRegression	68%	71%
10	KNeighborsClassifier	72%	77%
10	RandomForest	73%	79%
3	LogisticRegression	68%	71%
3	KNeighborsClassifier	71%	77%
3	RandomForest		

Based on the models used, it appears reasonable to expect a prediction accuracy of 70 – 75% for retention of riders.

Leveraging the Insights:

Ultimate could leverage these insights by...

1. Continue and seek to strengthen the use of the Ultimate Black program as it appears to be an important feature.
2. Share with drivers that the “avg rating by driver” that they provide is proving significant in helping to pick good riders. Consider education/training to help drivers improve their evaluation skills (i.e. are they potentially being too generous in their evaluation of riders)
3. Move forward with a test on the Toll Reimbursement program as soon as possible
4. Consider offering coupons/discounts to riders during their first 30 days as number of rides taken in this time period appears to be a significant contributor to retention.
5. Evaluate if there are additional features being tracked (or that should be tracked) to help improve the predictability of the retention model higher than 70 – 75% accuracy.

Links to Jupyter Data Exploring Notebooks (Python)

link to folder containing notebooks