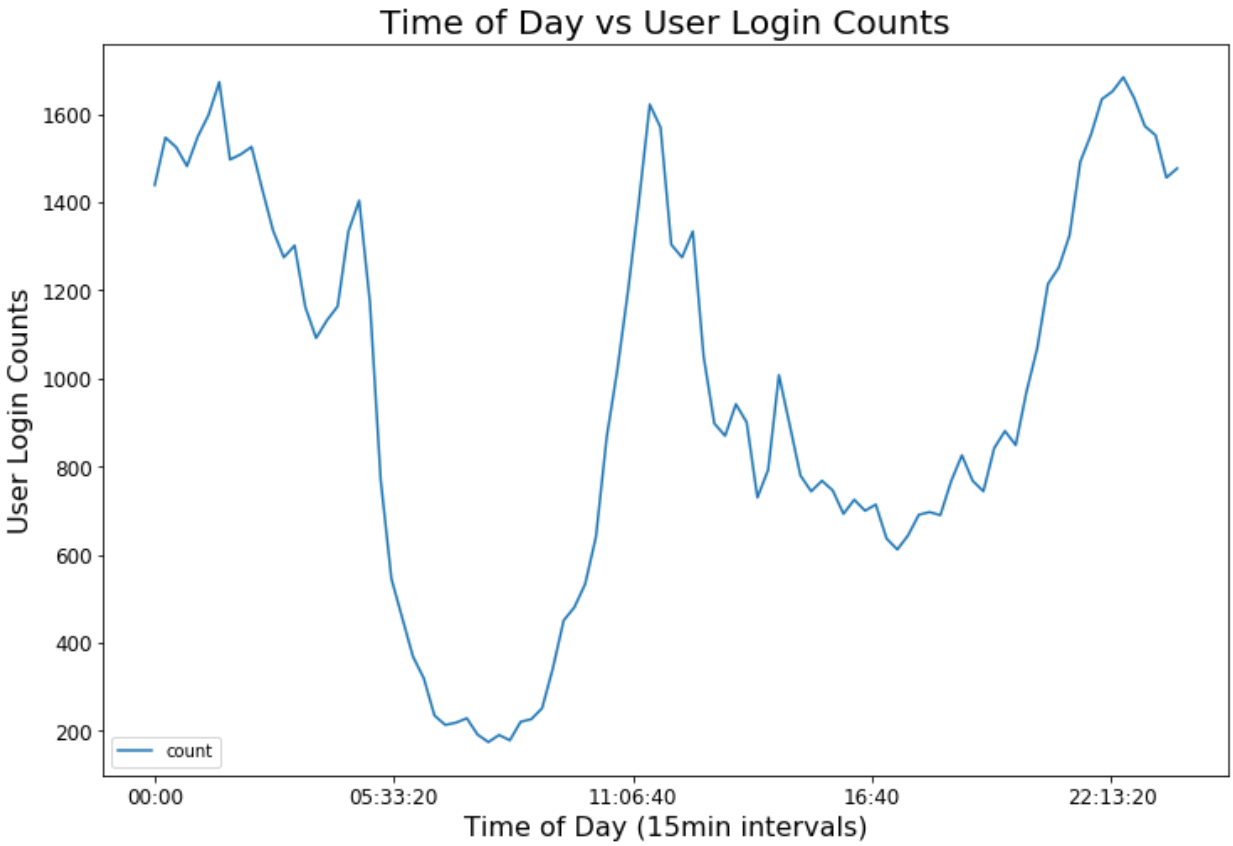
Solution to the Ultimate Inc. Data Science Challenge
By Dovid Burns

Part 1) Exploratory Data Analysis



The above chart clearly shows the daily patterns of user activity. We see that there are two different low user periods daily and two high user periods. The lowest user period based on login counts is from around 5:30am to 10am. The next relative low period is from 2:00pm to 4:15pm. The largest peak usage time is from 8:00pm to 5:00am. There is also a small peak of usage daily from 10:45am to 12:45pm.

Aggregating by week, we see a general trend of increased usage each week from January to April. Additionally, counting all the logins in this period by day of the week, we see that as the days go from Monday through Friday, the number of logins goes up each day starting at 8,823 to 15,2008. The weekend has the most logins with Sunday having 18,167 and Saturday having 19,377.

Part 2) Experiment and Metrics Design

1) I would choose a significate increase in the percent of drivers who respond to calls across the bridge as a key measure of success. I believe this is a good metric because it would show if this experiment created a tangible increase in cross city driver partners.
2)

a) I would start by labeling drivers by where they usually drive. I would then take a random sample of 10% of each of these groups of drivers. These randomly chosen would be told in secret of the full toll reimbursement. The other 90% of the drivers would be unaware that the experiment that was happening. We would then record the number of times that the bridge was crossed for all drivers, which direction they went, and the day of the week.

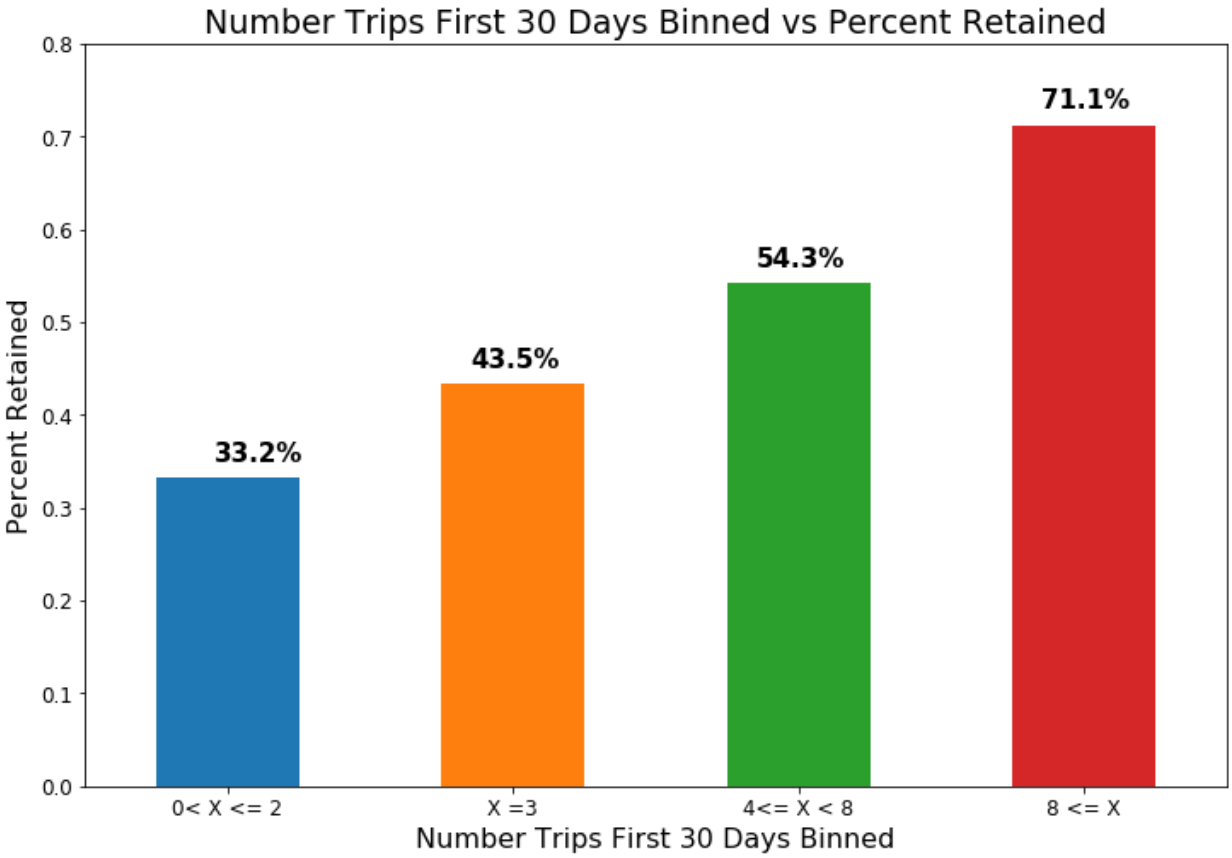
b) I would run a t-test to verify that the assumed differences between the control and experiment groups are statistically valid.

c) I would want to look at if there is a major difference between toll crossers on the weekend verses during the week. If the results showed a big increase in bridge crossers, we would need to try to quantify this increase in benefit to the citizens of both cities while also analyzing the total cost to the city. For example, it might make sense to only have this free toll on during the weekdays, when there might only be one free toll a day necessary to cause benefit. It is also possible that during the weekdays, this extra incentive did not have much of an effect but on the weekends, when both cities have activity the free polls had the most benefit.

Part 3) Predictive Modeling

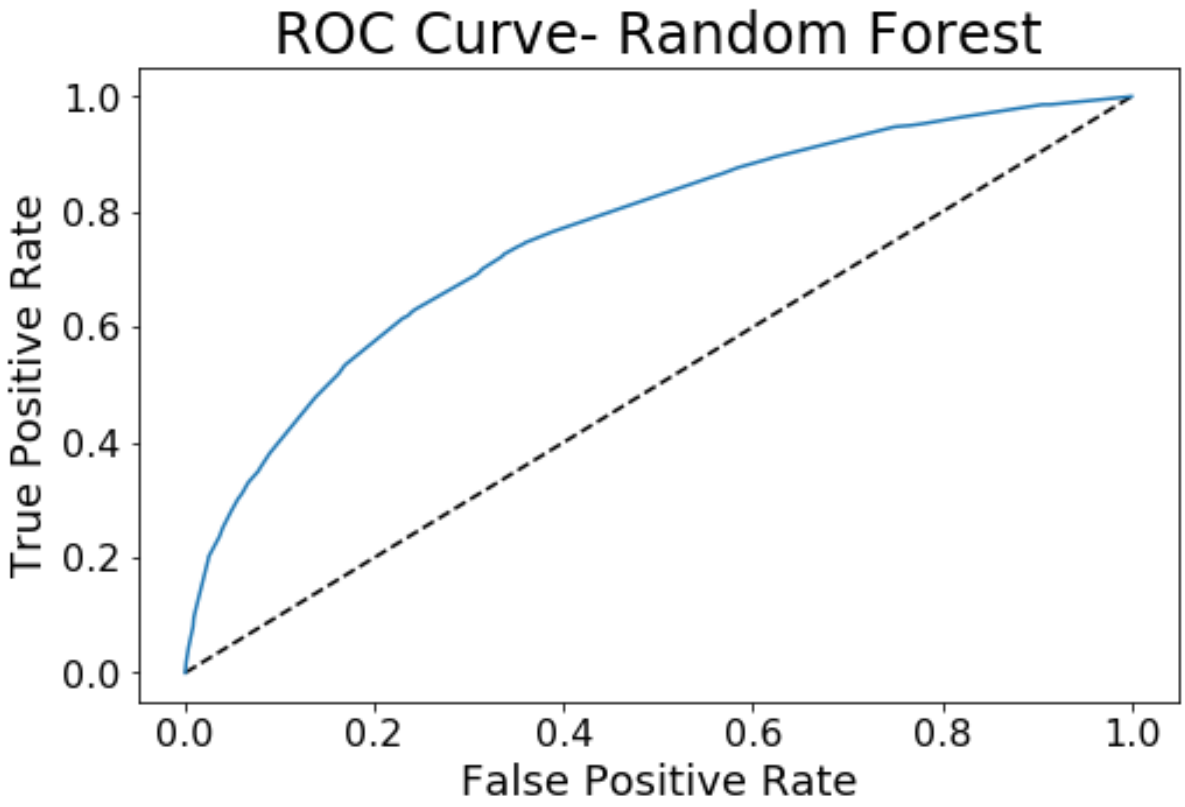
1) Looking at the data with the info and describe method, we did not find any obviously incorrect data. We did however find some missing data in the avg_rating_by_driver, avg_rating_of_driver and phone columns. Since these gaps were relatively small and the reason for the missing data was unknown I dropped these rows. A simple analysis of all the variables connection with active users showed many features did not have any effect of a user being retained in the 6th month. After performing data cleaning, we found 40.1% of the customers were still active/ retained in the 6th month. The statistically significant and strongly predictive feature were the city a user signed up in, the primary phone device, if they took an Ultimate Black in the first 30 days and the number of trips a customer took in the first 30 days.

Feature	Percent Retention Difference Max	Correlation with Retention: p-value
Ultimate Black User	22%	< 0.001
City	37.8%	< 0.001
Phone Preference	25.8%	< 0.001
Binned Trips in First 30 Days	37.9%	< 0.001



2) I built a Random Forest Classifier using the top predictive features listed above. I built chose a Random Forest Classifier as it reduces overfitting to the training data. It also runs very fast and I have had lots of success in the past with this type of classifier. I ran a cross validated grid search to optimize the machine learning algorithm for AUC. The grid search returned the default settings. To check validity of the model we looked at the following metrics:

Metric	**Score**
Accuracy on Training Data	71.5%
Accuracy on Test Data	71.2%
F1- Score on Test Data	0.597
AUC on Test Data	**0.759**



3) Ultimate can run this model on customers who have been with them for a month to predict who will still be active in six months. Using this prediction, they could spend more targeting customers who are unlikely to have long term retention. Additionally, they will know who they don't need to target because they are likely to have long term retention. The most important features that we found that Ultimate can try to affect most easily are the number of rides in the first 30 days and users who took an Ultimate Black in the first 30 days. This indicates that by

pushing new users to take more rides or offering discounts on the Ultimate Black rides during the first month could increase long-term rider retention.