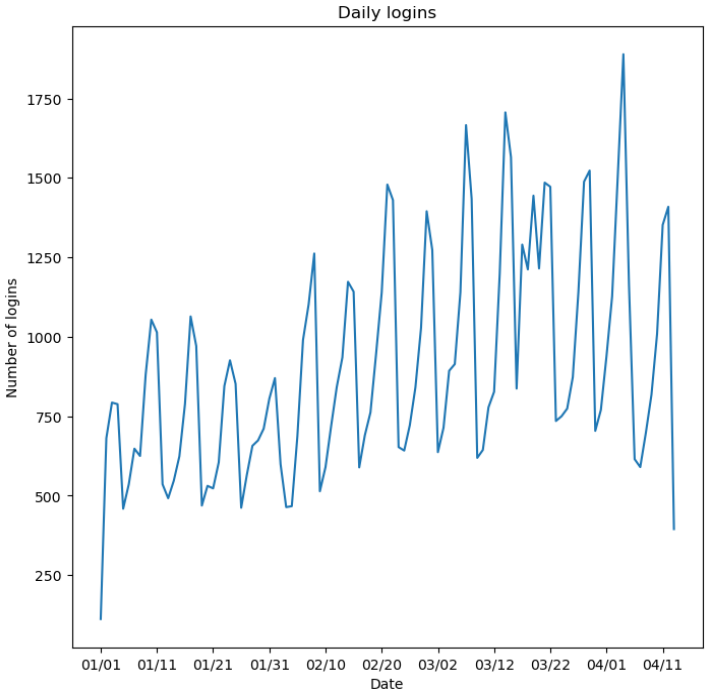
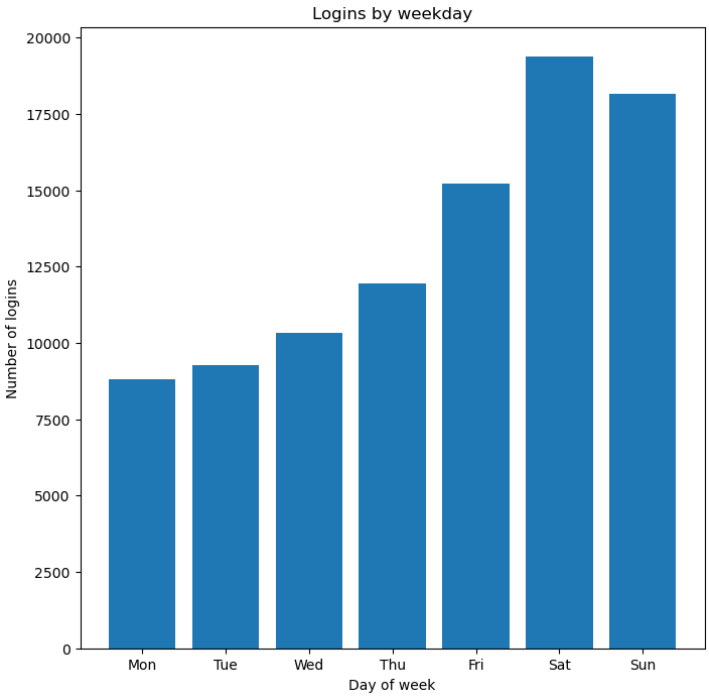
Exploratory Data Analysis

To begin with, we have a simple plot of login demand over time, aggregated by day.

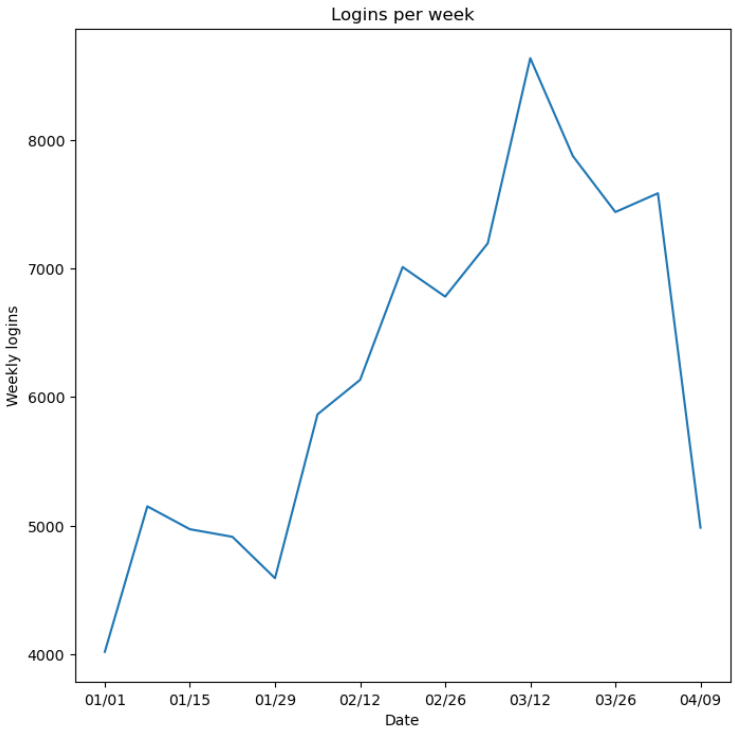


Although demand does appear to vary over time, there is also an evident cycle that looks as if it might be weekly. In fact, there is a clear difference in demand based on day of the week.



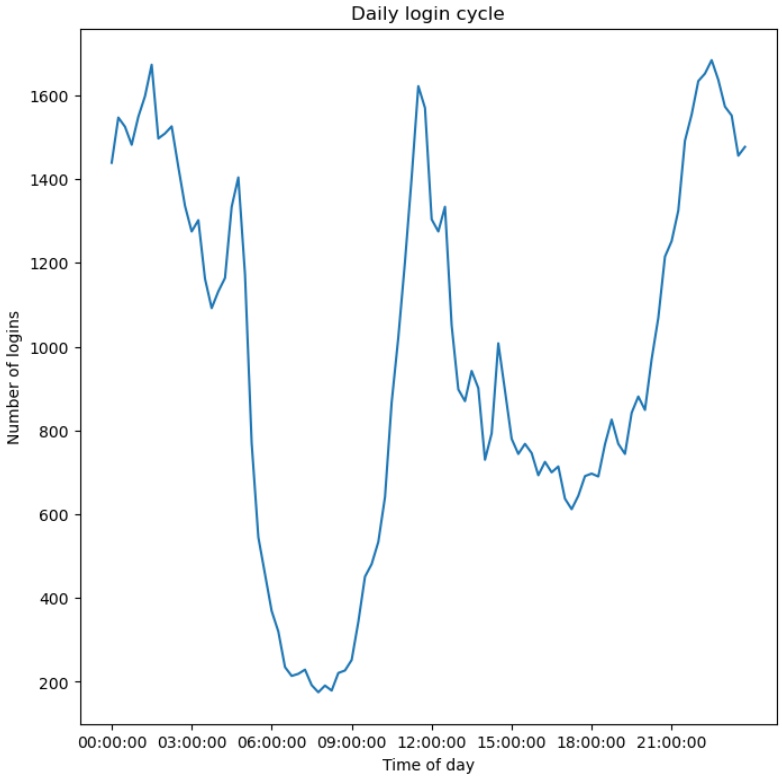
Demand for the service appears to be far greater on the weekends than at the beginning of the week, with more than double the logins on Saturday and Sunday as on Monday and Tuesday.

Once the login data are aggregated by week, we can see the demand over time with the cycle removed.



Now, it is quite clear that the highest demand occurred in March, peaking around the week of the 12th. The logins near the very ends of the data collection period are the lowest, hinting that the data around those times may be incomplete. It is possible that the causation goes in the other direction, and data collection began and ended at times of low demand, for instance at the beginning and end of the service.

One last feature of note is the daily login cycle.



The peak login times appear to occur around both noon and midnight, with very low demand in the early morning hours, and a smaller but significant dip in the early evening.

Experiment and Metrics Design

1. The key measure of success for the toll-paying experiment could be the proportion of time spent by each driver in either one of the two cities–say, Gotham. Proportions near 0 or 1 would be an indication of an unsuccessful experiment, while proportions near 0.5 would be ideal. So, the quantity we'd be trying to maximize could be the absolute distance of the average proportion from 0.5:
   abs(Time in Gotham/ Total time - 0.5).
   If necessary, we could adjust the ideal ratio to account for the total activity in Gotham and Metropolis, if they are not exactly even. For instance, if Gotham is three times the size of Metropolis, the ideal ratio could be 0.75.
2. We could use a two-sample z-test for proportions to see if the proportion of time spent in Gotham was significantly different for drivers given the toll reimbursement than not. We would randomly assign our drivers to be part of the experimental group (that gets reimbursed) and the control group (with no reimbursement). We would check if the percent was closer to 0.5 after the reimbursement. If the difference was statistically significant, we could check whether the difference was sufficient to increase profits enough to offset the reimbursement costs. If not, then the new protocol might not be worth implementing.

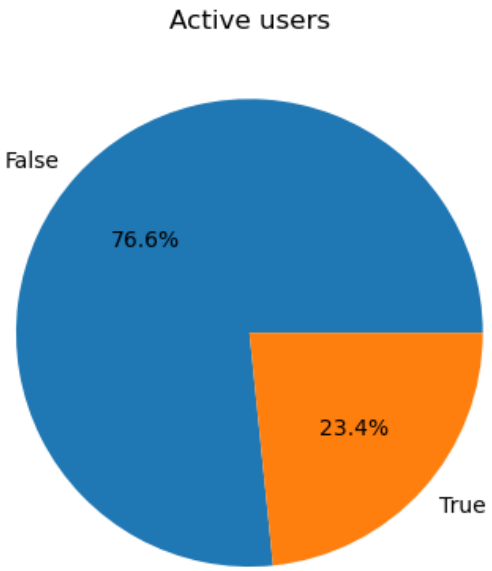
Predictive modeling

To prepare the data, I converted the dates into Datetimes. The goal was to figure out which users had taken a ride in the last 30 days, 6 months after their start date. Since a month is not an exact unit of time, and since all the users in the dataset started in January, I found the "six months out" time by keeping the day and year but changing the month to July. This was convenient since January and July both have 31 days. To generalize the process to other months, I'd create a six-months-out function that would compensate for some months being shorter. Alternatively, I could ask stakeholders whether 150 days from the signup date is a satisfactory approximation for 30 days before 6 months.

The target feature, "active," is a boolean column that is True if the user has taken a ride within 30 days of the 6-month point.

During the EDA process, I created pie charts of each categorical feature and found the summary statistics of each numerical feature.

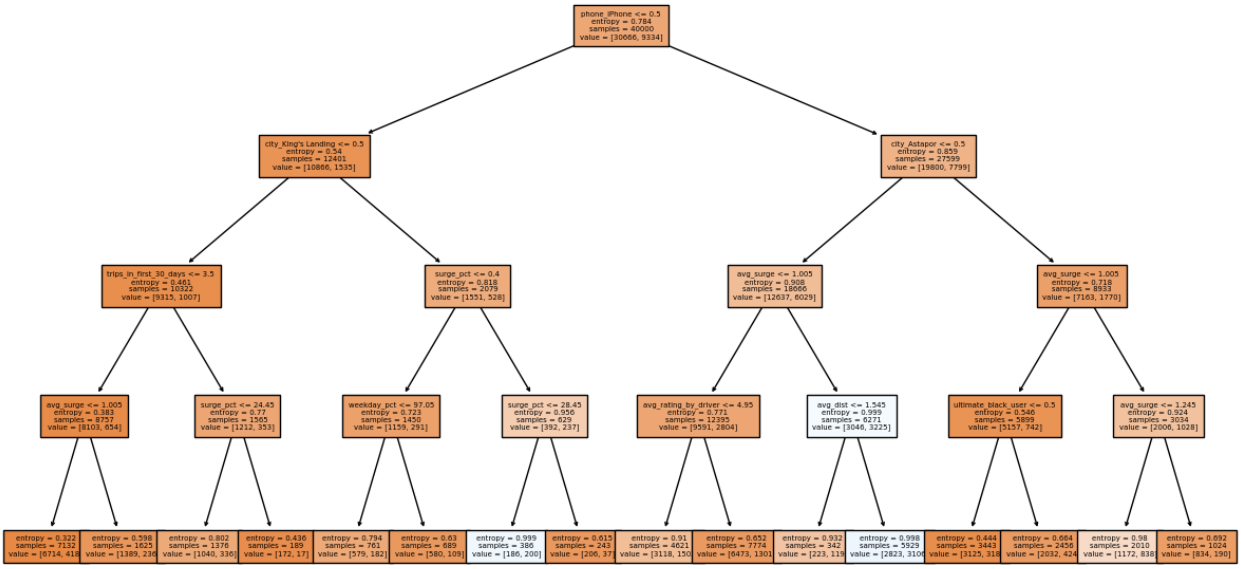
Most importantly, here is a pie chart of active users as of six months from each user's start date.



It can be seen that only 23.4% of our users are active after 6 months.

I tried a number of models to predict which users would remain active to the extent specified, including K-Nearest Neighbors, with hyperparameters tuned by a Random Search, and two types of decision tree–one with hyperparameters calculated tuned by a Random Search, and one created with the purposes of being easy to visualize and interpret. The difference in predictive ability for each of these three models was negligible, and well within expected variation due to chance. The f1-score accuracy on the test set of each model was between 0.77 and 0.80. I did attempt a Support Vector Classification model, but the model did not fit well, with far lower accuracy scores than the others.

Here is an image of the interpretable decision tree, with the maximum number of features for each split set to 1, and the maximum tree depth set to 5, which is the most that create a legible tree.



With the insights gained from the chosen model, the company should be able to predict whether a given customer is likely to stay active on the service after only the first month. For customers predicted to be unlikely to continue use, Ultimate could target a retention campaign towards those customers. They could use some sort of incentive, such as a discount on future purchases, or a reward for so many uses of the service within a month.