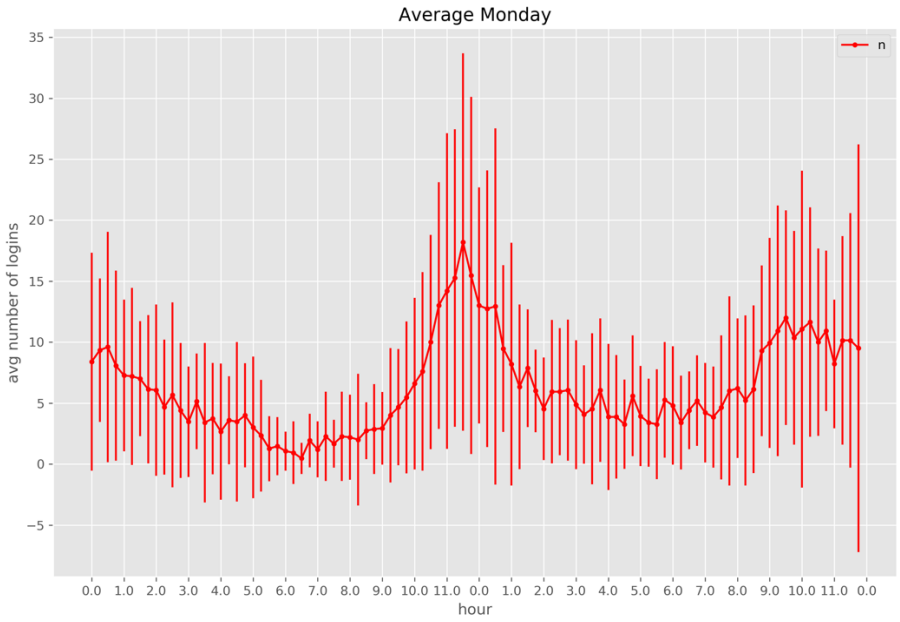
Ultimate Inc. Data Challenge Write-up

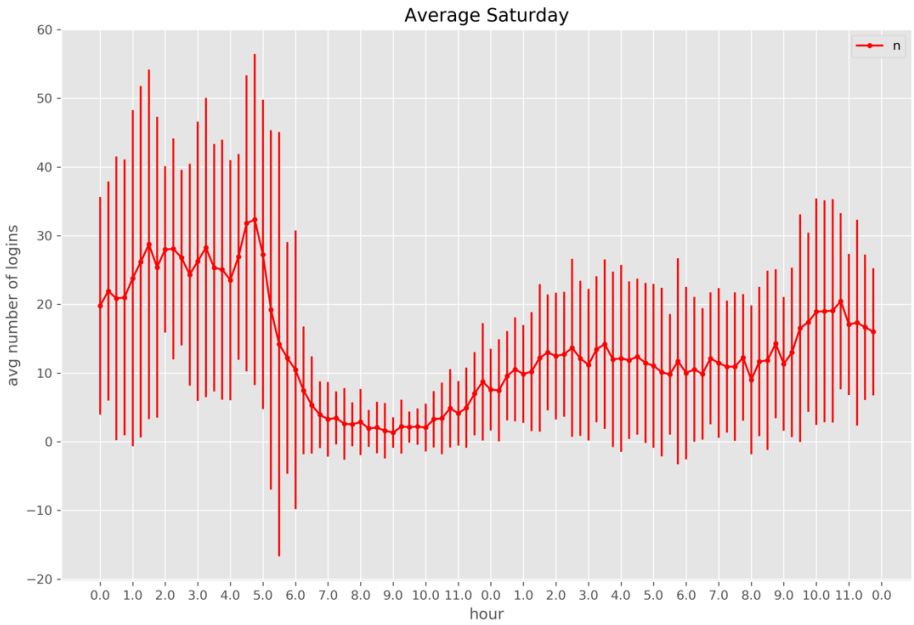
Sam Fisher

Exploratory Data Analysis

By visualizing the time series of average logins per 15 mins for each weekday with 95% confidence, we were able to find a distinctive rhythm to weekly ridership.

- Weekdays have a peak around noon and a peak around midnight.
- The size of the midnight peak modulates upward from Monday to Friday, from about 12 users per 15 minutes on Monday to about 30 users per 15 minutes on Friday.
- Weekends have a distinctly different rhythm, peaking in the wee hours of the morning and gradually rising throughout the day.





Experiment and metrics design

Metric Design:

We'll call our metric the "two-city driver" status. It will be a Boolean value representing whether or not a driver crosses the bridge on at least *x*% of the days that they drive in a given two-week period. An a priori guess for the *x* threshold variable is 33%. Exploratory analysis of the current frequency of bridge crossings could help define an optimal value for this. We can calculate this metric using the raw GPS data or higher-level features that may already exist in Ultimate's database.

Experiment Design:

Each driver's two-city driver status is a Bernoulli trial. Their probability of their status after the intervention will be strongly conditional on their status before the intervention. We will select a random sampling of drivers and measure their two-city driver status in two week periods before and after the intervention of the reimbursement incentive. The success of the experiment be determined by a McNemar's test, this will determine the statistical significance of the intervention's effect on two-city drivership. We have chosen McNemar's test because we are interested in a change within a given sample before and after some intervention.

To measure the effect size, we will calculate a 90% confidence interval for the true percentage change in two-city drivership after the intervention. To accomplish this, we will treat the portion of two-city drivers before and after the intervention as realizations of a binomially distributed random variable and calculate 95% confidence intervals for them. We can then take the minimum and maximum of the changes between these regions to determine a 90% confidence interval for the effect size. Knowing this effect size will be important to the business decision on the final cost/benefit analysis for the incentive program.

Please see the ipython notebook included for a simulated realization of this experiment.

Predictive Modeling

Data Cleaning:

There were a few missing values for ratings and phone type. Since these features were predictive, we decided to take the subset of ~41,500 users for which there was complete data. Imputing the mean value for rating features would have artificially shrunk their variance.

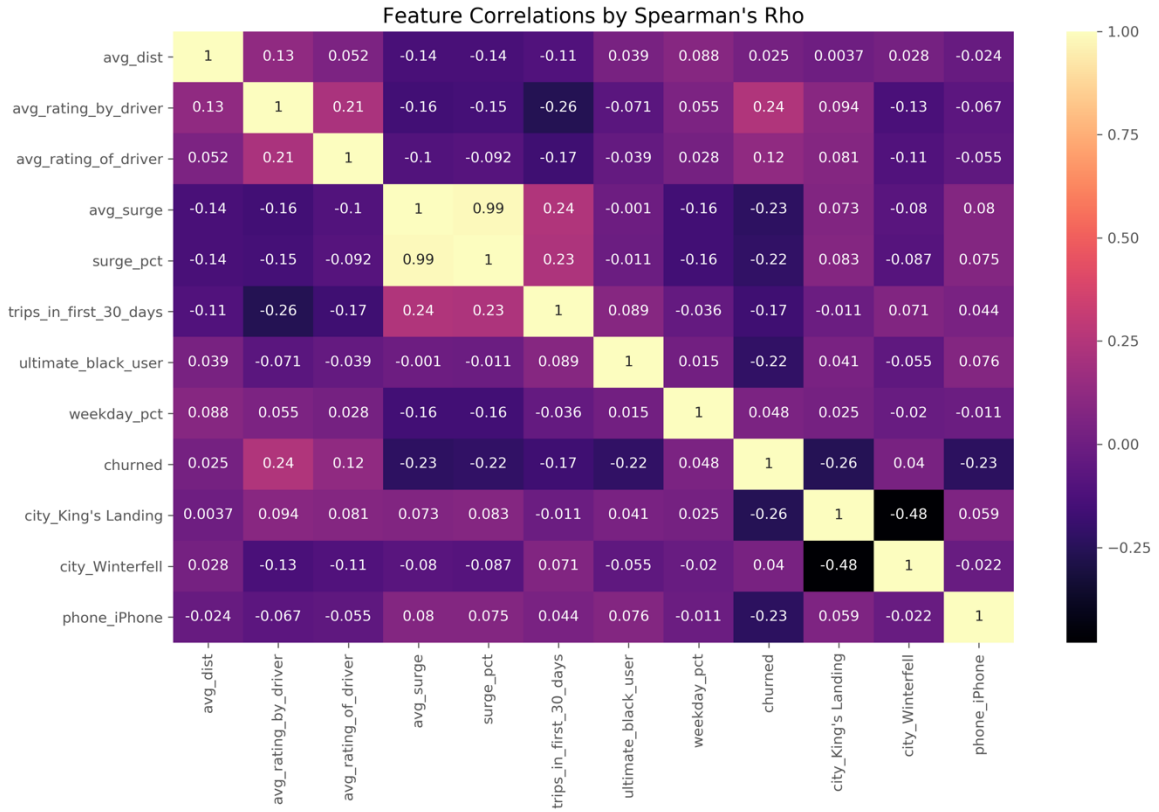
Feature Engineering:

We created a binary outcome for churn by subtracting the last ride datetime from the date '2014-07-01' and determining if the time-delta was greater than 30 days. About 41% of the sample were retained. This makes predicting for this label a fairly balanced classification problem.

We also created one-hot encodings for the phone type and city categorical variables, dropping one category for each in the final dataframe.

Exploratory Analysis:

We created a heatmap of the matrix of the correlation values in order to investigate collinearity of the features and do some feature selection for a logistic regression model.



- The percentage of trips taken with a surge multiplier and the average surge multiplier over all trips are highly correlated. Since avg. surge is slightly more predictive, we'll use that in logistic regression
- More trips in the first 30 days reduces the likelihood of churn. Taking an ultimate black in the first 30 days reduces the likelihood of churn. These two factors are somewhat positively correlated with one another.
- Being in King's landing makes it substantially less likely that you will churn.
- Android users are substantially more likely to churn.
- Higher average ratings by the driver and of the driver result in a lower likelihood of churn. These are significantly correlated with one another, we'll take the more predictive one- ratings by the driver.
- Average distance does not seem to have much bearing on the outcome, though random forest will prove us wrong about this later!

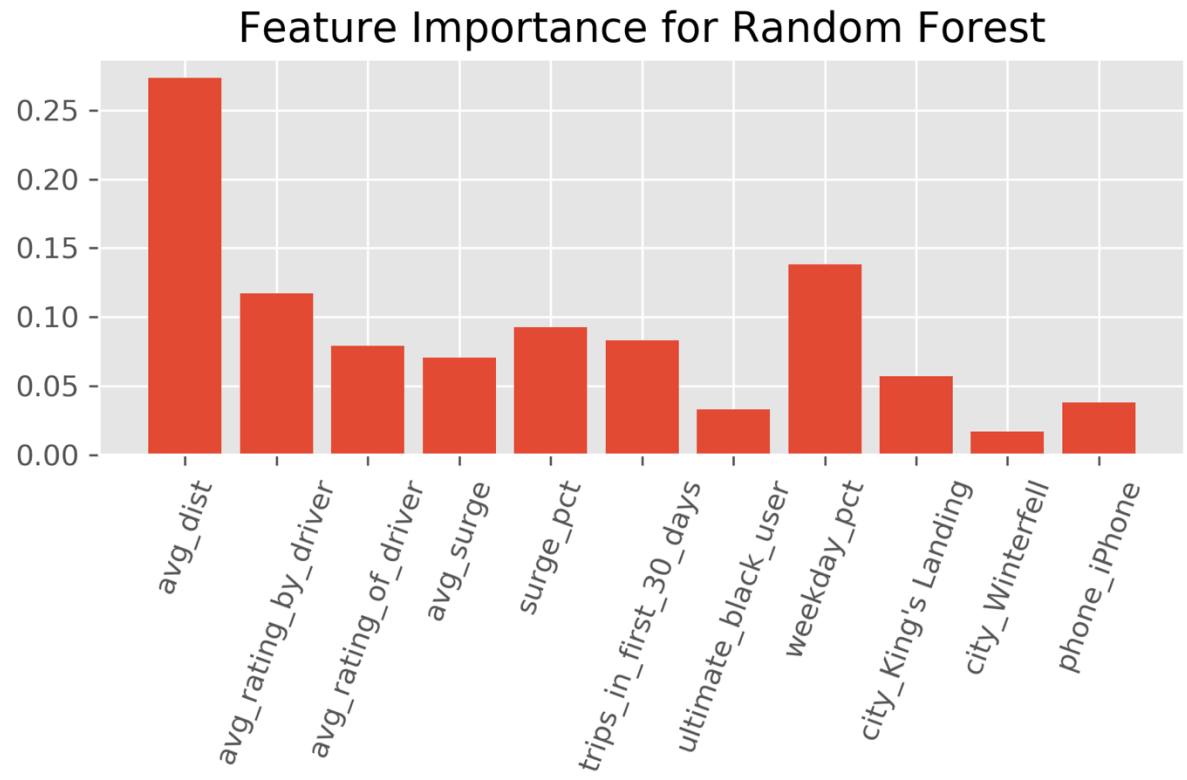
Just under 50% of users are in Winterfell, 20% in King's Landing, 30% in Astapor. 100% are located in the world of Game of Thrones. About 30% of users are on Android, 70% on iPhone.

Predictive Modeling:

We implemented two models for the churn prediction, logistic regression and random forest optimized with grid search cross validation. We selected these models as disparate hypotheses about the decision boundary. Logistic regression forms a simple, linear boundary, whereas random forest might fit some non-linearities. Because there are more than 40,000 samples and 11 features, we knew that the data was sufficiently tall and wide to power the random forest. To choose between models, we took the greater overall f1 score on a test set of 33% of the data.

For the logistic regression model, we excluded the features 'avg_dist', 'avg_rating_of_driver', and 'surge_pct' because they had strong collinearity with other features. Logistic regression had an f1 score of 0.7 and tended to overpredict in favor of churn.

For the random forest, we included all of the features and searched for the best number of trees over the grid [20,40,60,80,100], taking the best mean accuracy in 5-fold cross-validation. The optimal random forest had an f1 score of 0.76 on the test set and outperformed logistic regression in precision, recall, and f1 for both classes for both test and training sets.



Recommendations:

We recommend that Ultimate run the random forest classifier described above once per day to predict for customers who will churn. If a customer is classified as positive, we suggest trying churn intervention emails or discounts. Additionally, we recommend that ultimate consider the relative feature importance in this model when targeting potential users to acquire to on the other end of the customer funnel.