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.

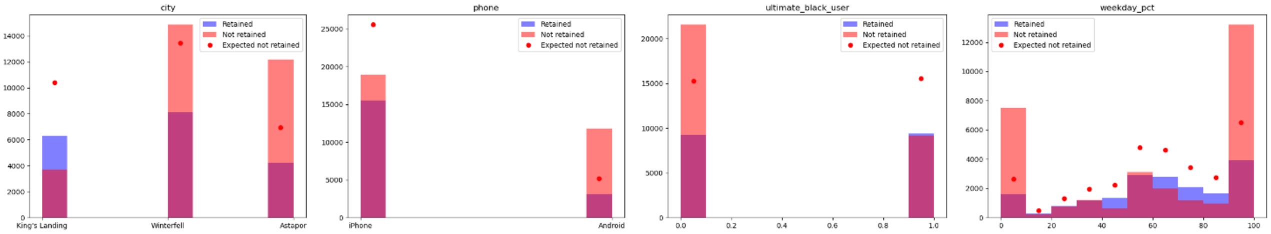
We would like you to use this data set to help understand what factors are the best predictors for retention, and offer suggestions to operationalize those insights to help Ultimate.

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.

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?

After confirming the reasonableness of the timestamps and variable ranges, I removed or imputed null values (to facilitate modeling), noting that drivers with null "rating_by" and/or "rating_for" values had retention rates well below the mean and imputing accordingly.

After this, I plotted histograms of each variable's retention & non-retention rates. Here are a few striking examples. The "expected" non-retention rate, estimated by a proportion of the retention rate, is shown as a red dot. Red bars that do not reach the red dot indicate better-than-expected retention, and vice-versa.



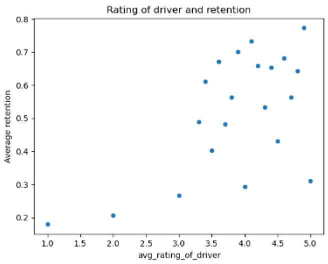
2. Build a predictive model to help Ultimate determine whether or not a user will be active in their 6th month on the system. Discuss why you chose your approach, what alternatives you considered, and any concerns you have. How valid is your model? Include any key indicators of model performance.

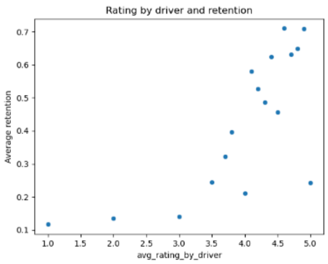
After performing the data cleaning mentioned above, I also one-hotted the city variable and scaled the ratings values to between 0 and 1, putting all of our data between 0 and 1 (including the label).

I applied logistic regression and a gradient boosted classifier, with the latter showing better performance. After grid searching with cross validation to identify optimal hyperparameters, I was able to correctly identify 87% of churned drivers with only a 20% false-positive rate.

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).

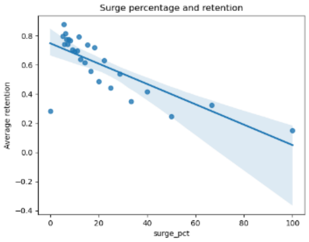
1. The most predictive features are ratings (by and of the drivers), surge percentage, and weekday percentage.
2. Higher ratings (both ways) tended to correlate to higher retention, but a very strange finding is that drivers with whole-number average ratings (either way) had much lower retention than drivers with non-whole-number ratings.





3. Drivers in King's Landing had much higher retention. This could be due to driving conditions there, economic conditions, or other policies. It warrants further investigation.

4. Higher weekday percentage tended to correlate to higher retention, but drivers who drove exclusively on either weekdays or weekends had the lowest retention of all.
5. Higher participation in surge events correlated to lower retention (as well as complete abstinence from surge events). This could be because of greater driver frustration, or it could be because drivers who just needed cash quickly and never intended to drive for the long term tended to chase surges.



6. iPhone use is a big indicator of retention. Android users had much more churn. This might be due to the app experience, or to a difference in Android vs iPhone customer profiles.