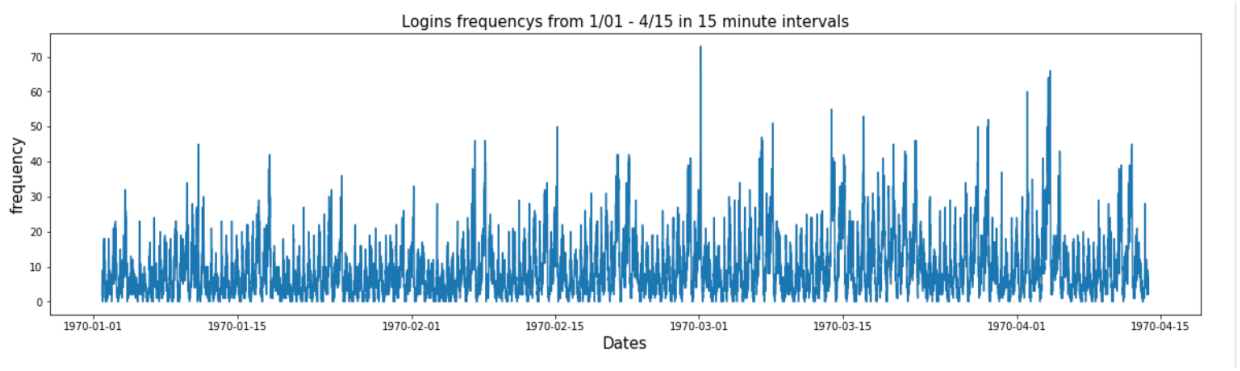
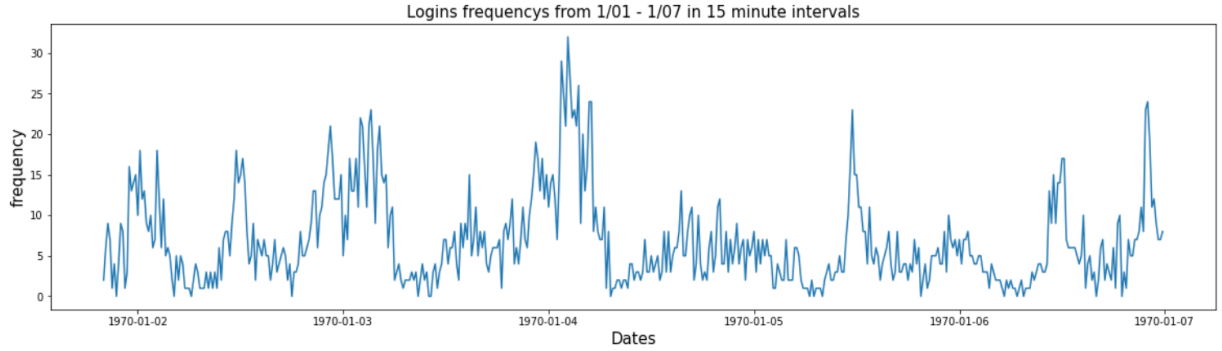
Ultimate data science challenge

Part 1 Exploratory data analysis



Cyclic pattern of logins repeats through whole data set



Zooming into a week of data increases in logins seen around midnight and noon of each day.

Part 2 Experiment and metrics design

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.

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?

The percentage of drivers who ride in the opposing city to the city they are based in would be the metric I use. I would use this as with one number I can get details about the drivers work load in both cities. I could also compare their initial percentages and changes to percentages easily.

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:

a) how you will implement the experiment

I would create two randomly selected groups of drives, one a control group that is not given toll fees and another group that is given paid toll fees. An equal amount of drives from each city would be used as the initial pool to create the two groups. I would select a window such as a quarter to be able to negate weekend and holiday trends from the data collected from the two groups. I would measure the percentages of each driver's rides in the opposing city from the two groups before the test and monitor that percentage throughout the test run.

b) what statistical test(s) will you conduct to verify the significance of the observation?

Depending on the size of the two groups a t-test of z-test can be performed to see if there is a statistically significant difference between the two groups' percentage of rides in opposing cities. This would be done assuming the two groups' percentages come from a normal distribution.

c) how you would interpret the results and provide recommendations to the city operations team along with any caveats.

To say there is a difference in mean percentage of ride in other cities between the two groups. After the test a p-value would be calculated to see if the null hypothesis that there is no difference between the two groups mean percentage of rides in opposing cities can be rejected. If the p-value is below .05 then we could confidently say the difference between the two groups mean is unlikely to be by chance.

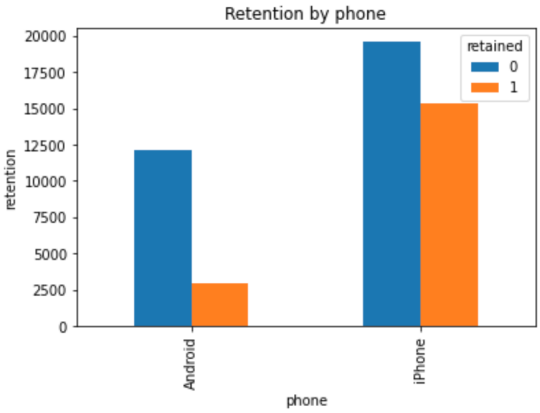
Part 3 - Predictive modeling

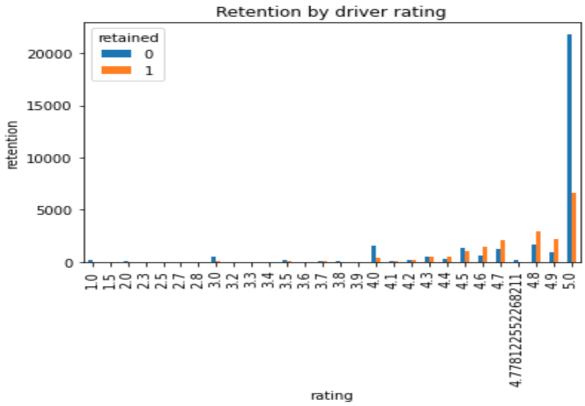
We would like you to use this data set to help understand what factors are the best predictors for retention, and offer suggestions to operationalize those insights to help Ultimate.

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.

> *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?*

36 % of the riders were retained in the observed dataset.

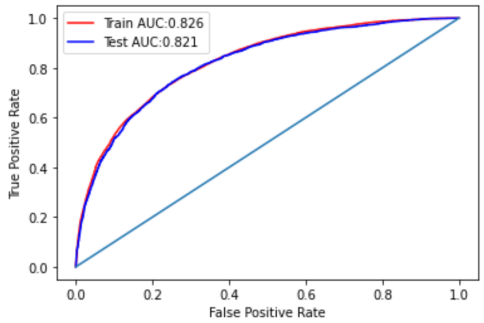




Distributions of retention for phone type shows clear bias towards iPhone users retaining at a higher rate, possibly UX differences here. Distribution of ratings by driver show if people who don't just give 5 stars are more likely to retain compared to people who just use 5 stars as default as these are drivers who probably give ratings at a higher frequency then short time users.

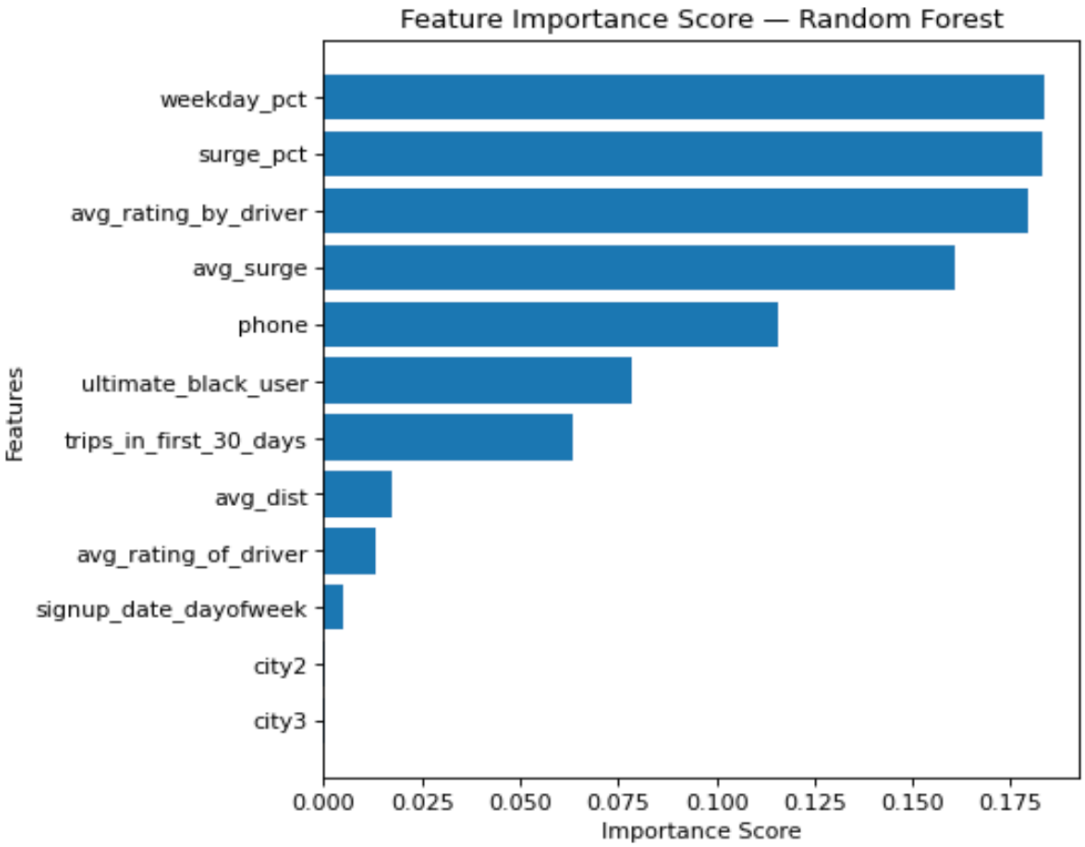
How valid is your model? Include any key indicators of model performance.

Initially a random forest model without tuning was tested before gridsearch was used to tune hyperparameters. Even in the final model the training and testing Area under the curve show there is bias in the model and tuning of model or adjusting of data set target distribution could help improve the model performance



	precision	recall	f1-score	support
0	0.76	0.89	0.82	6310
1	0.74	0.51	0.61	3687
accuracy			0.75	9997
macro avg	0.75	0.70	0.71	9997
weighted avg	0.75	0.75	0.74	9997

Retention is 1 in the classification report and is about the same between the two classes. The recall for retention which is the minority class is 51% which means we missed half of the actually retention users we wanted to correctly label. Actions to improve recall could be Over sampling to balance the classes, generating / collecting more descriptive features, or using a more complex model to better generalize over the test set.



The final models feature importance shows the weekday percent, surge percent, and avg_rating_by_driver of a user key to the models predictions . Again as half the retention predictions are off the value of these features are still in question. But using these features could help target users who are clearly using the service multiple times a week and likely commuting or using for social travel.