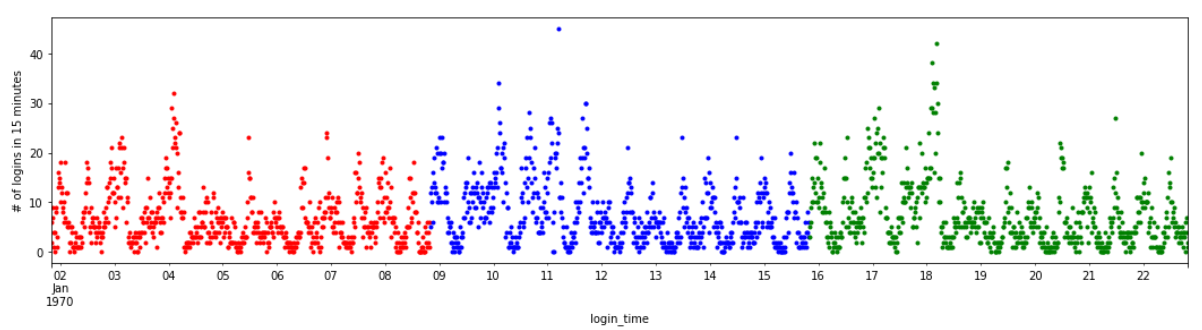
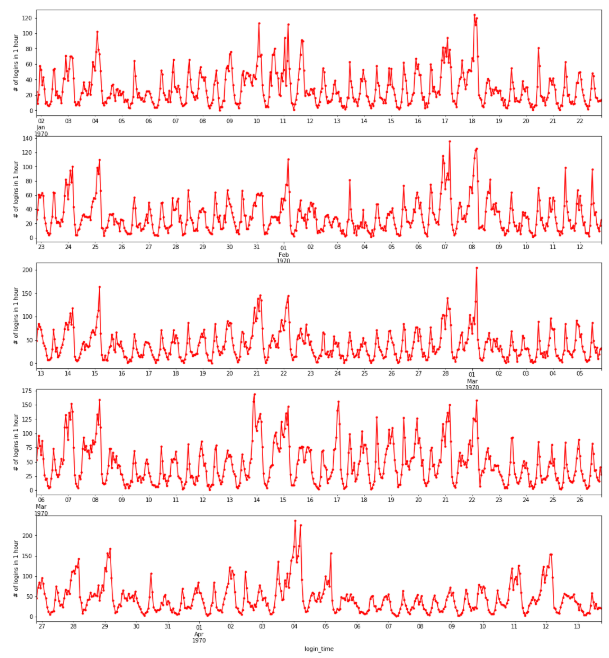
Take home challenge—Ultimate

Part 1 ‑ Exploratory data analysis

* Resampled login file data to 15 minutes. The data shows apparent daily cyclic patterns. Below shows first 3 weeks of data as an example.



* Resampled data to 1 hour smoothed the data a little more and clearly shows 2 peak login time periods in one day. One peak appears around mid-night and the other peak appears in the middle of the day. See plots below.



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?

**Ans**: I would choose the change in number of trips between the two cities as the metric. The assumption is that reimbursing all toll costs would increase traffic between the two cities.

1. 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:
   1. how you will implement the experiment

**Ans**: Choose two adjacent time periods.

Control group would be the first time period that is long enough (say 3 weeks) and collect the data how much in-between city trips have happened during that time period on each day.

Treatment group would be the 2nd time period with the same time length, with all bridge toll costs reimbursed for weekdays. Collect the data again on how much in-between city trips have happened on each day.

Time periods chosen need to avoid special events, holidays, spring/summer/winter break, etc.

1. what statistical test(s) you will conduct to verify the significance of the observation

**Ans**: I will conduct 2-sample t-test to test the difference of the mean daily in-between city trips.

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

**Ans**: Significance of the statistical test at pre-determined alpha level (propose 5%) would indicate that the measure of removing bridge tolls is expected to increase in-between city traffics by the amount shown as 95% confidence intervals of the mean difference. Communicate such intervals to city operations team. If it is determined (probably by city operations team and others) that the benefits from increase of in-between traffic is greater than the loss of toll income to the city, it can be recommended to run a pilot phase of such an experiment in much longer time period to validate the results from previous short term experiment. This is to address the caveats of small sample size (3 weeks) and not paired time experiment—it would be better to randomly select two groups of driver partners with one group paying the toll and the other group getting toll reimbursed, in the same time period; however such an experiment is not believed to be practical.

Part 3 – Predictive modeling

Date cleaning/wrangling/EDA:

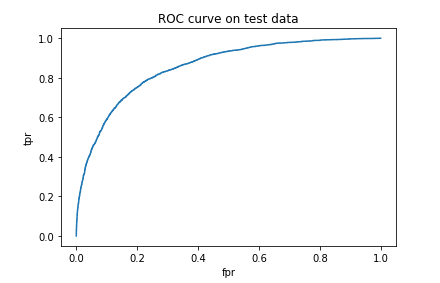
* Provided Json file has NaN in it and cannot be read into Pandas directly. Found a workaround to solve the issue.
* Used summary, value counts, box plots to check data. Did not find data issues.
* There are small amount of missing data. # impute missing values. Impute avg\_rating\_by\_driver and df.avg\_rating\_of\_driver by their respective median ratings, and impute phone by 'Unknown'.
* Generated label for retained drivers based on whether they made any trip in the last 30 days (on and after 2014-06-01).
* **There is 37.6% of drivers labelled as retained drivers**.
* Used crosstab and scatter plots to get some idea on relations between all predictor variables and the retained driver label. It appeared that ultimate\_black\_user, phone, and city could have some predictive power.
* Used one-hot-coding on phone and city to prepare for modeling.

Modeling:

* Used Random Forest classifier. It is known with its flexibility in dealing with non-linearity, and predictive performance.
* Set aside random 20% of data for model evaluation.
* Used rest 80% of data as training.
* Used 5-folder cross-validation.
* Used randomized search to tune hyper-parameters. Hyper-parameters tuned include:
  + "n\_estimators"
  + "max\_depth"
  + "max\_features"
  + "min\_samples\_split"
* Used ROC\_AUC score as metric.

Predictive Results:

|  |  |  |
| --- | --- | --- |
|  | On cross-validation data | On unseen test data |
| ROC\_AUC score | 0.851 with std 0.004 | 0.856 |



Confusion Matrix on test data: Accuracy = 78.98%

|  |  |  |
| --- | --- | --- |
|  | Predicted NOT retained | Predicted retained |
| Actual NOT retained | 5413 | 841 |
| Actual retained | 1261 | 2485 |

Feature Importance Plot below shows that avg\_rating\_by\_driver, surge\_pct, weekday\_pct are among the factors that have most impact on the prediction.

