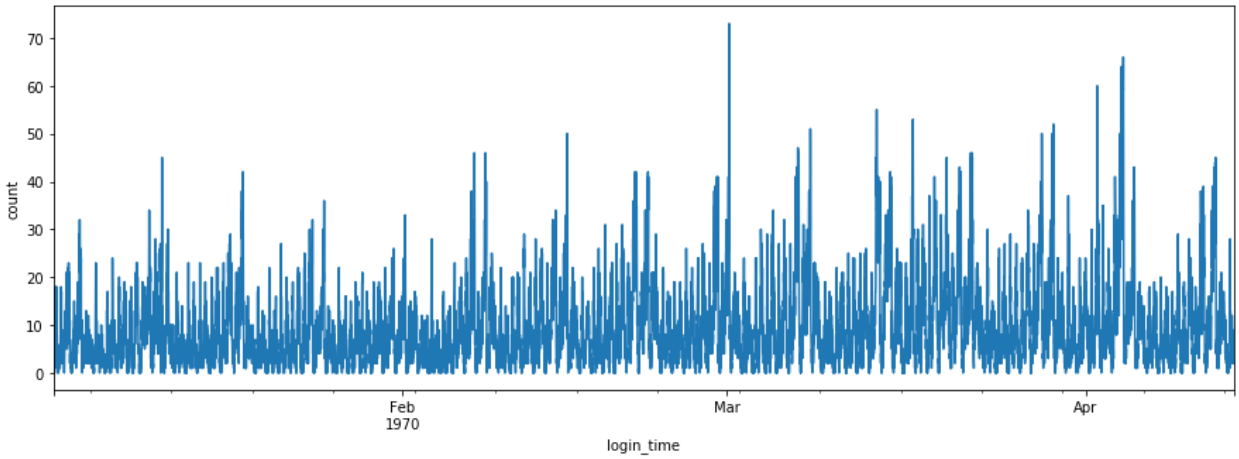
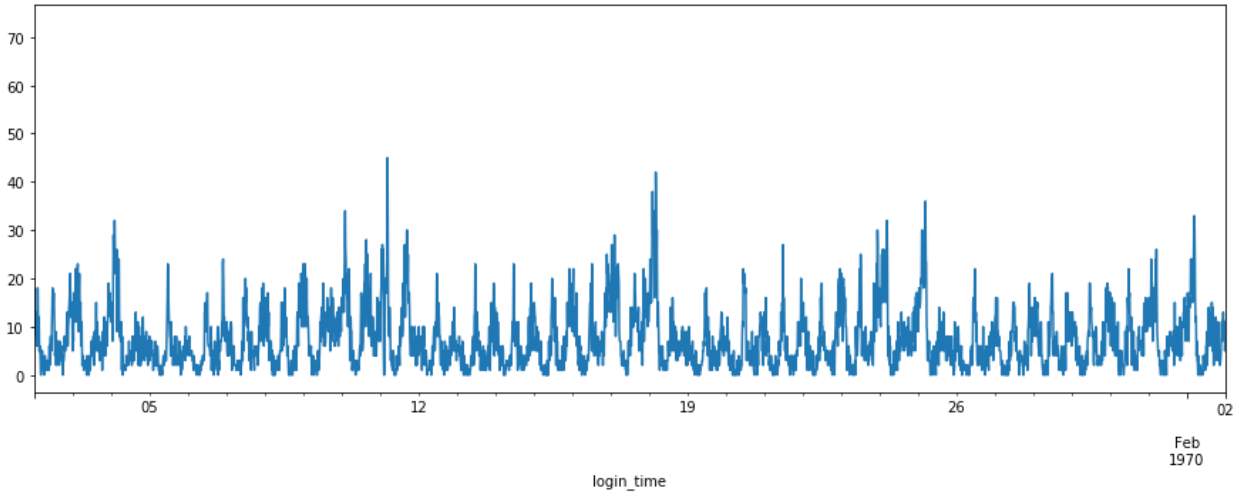
Part 1 - Exploratory data analysis

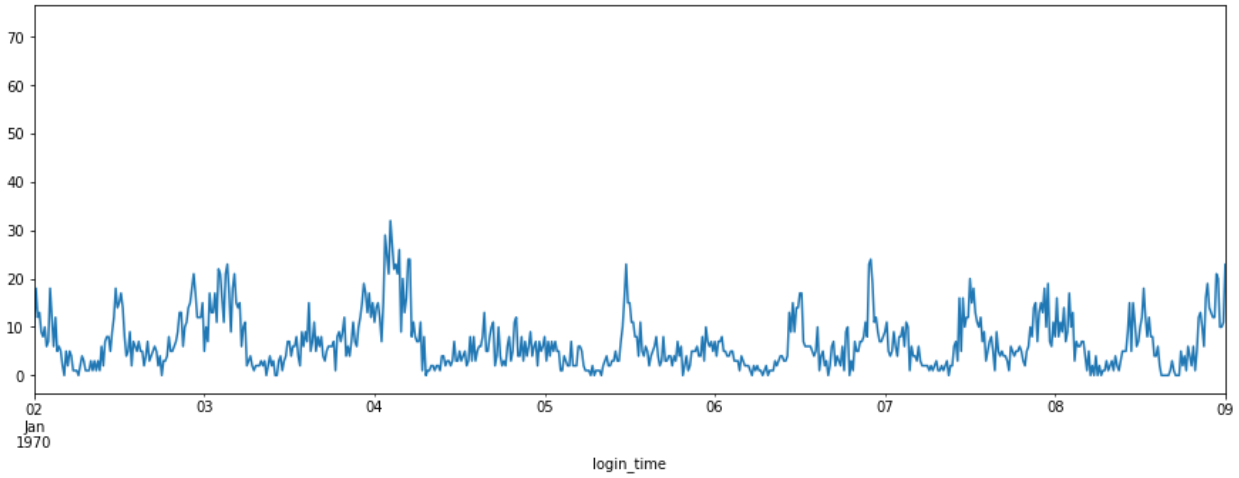
Full view (3-months):



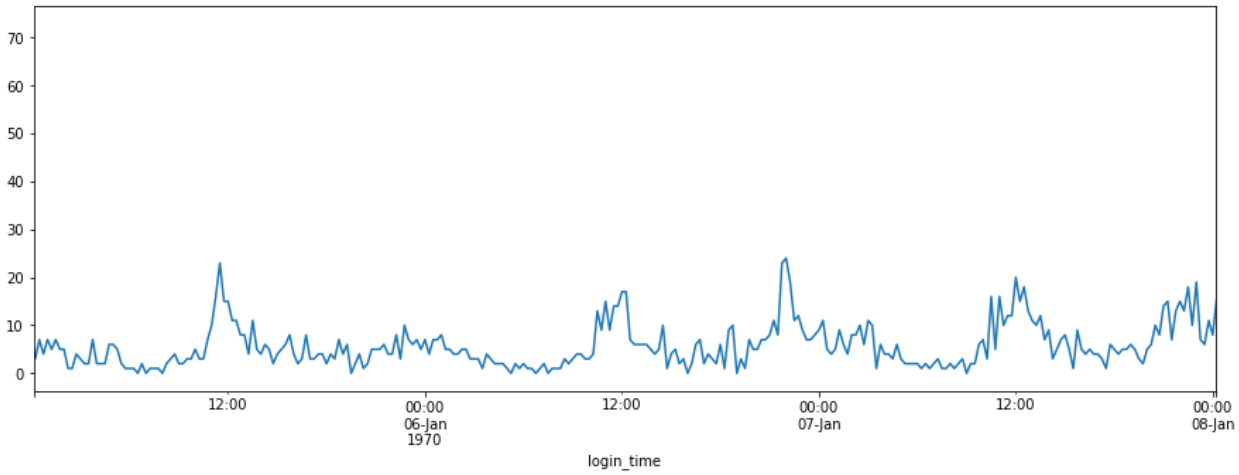
1-month view:



one week view:



One day view:



Interpretation:

There is a strong periodical pattern. Counts increase from weekday to weekend, peaked at weekends periodically. Peak time during one day would be around 12:00pm (noon time), or before 12am (before bedtime).

Part 2 - Experiment and metrics design

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?
   Logic flow:
   a. The problem is: driver partners tend to stay in the same city.
   b. Hypothesis for the reason is: the two-way toll between the cities might be a cause to keep drivers in the same city. Alternative Hypothesis (Ha): compensating toll-fee would promote the drivers cross bridges (increase the percentage of drivers to cross bridge), so that they would be active in both cities.
   c. What's measurable or say operationable?
      i. The location information, there are many ways to measure the difference.
      ii. Metrics:
         1. Percentage of drivers that cross the bridge in a certain period
         2. Cross bridge frequency for each driver
         3. Acceptance rate of driver for rider's request if the rider's destination is in another city.

2. Hypothesis testing:
   a. Experiment 1: A/B testing.
      i. We should randomly choose half drivers from each city, to compensate them with toll-fee, as the test group(treatment group), and the other half as a control group, which would not get compensated. Measure the cross city probability for each group based on location data. Sample size N is determined in this case by the driver number.
      ii. To set up the A/B test, we need to determine the following parameters:
         1. Baseline conversion rate: before the experiment, what is the cross bridge probability(what percentage of drivers cross bridge in a certain period, the period could be one day, or per week, or a month to reach a stable value). After randomly split the drivers, this value should be the same for each group to avoid selection bias for sample.
         2. Practical significance (minimum detectable effect): It depends on how much difference would be satisfied for the company. At least the profit should be larger than the fee-compensation.
         3. Statistical power: usually 0.8
         4. Significance level alpha: usually 0.05
         5. Calculate the practical significant needed accordingly to sample size (drivers) for experiment. We could also determine the practical significance according to desired profit.
         6. http://www.evanmiller.org/ab-testing/sample-size.html

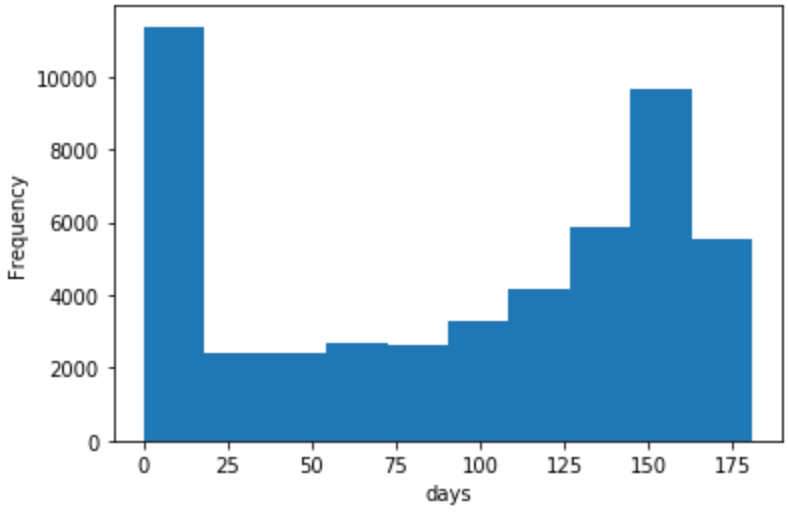
baseline	Practical	1-beta	alpha	N

	significance			
10%	3.85%	0.8	0.05	1000
20%	5.07%	0.8	0.05	1000
30%	5.78%	0.8	0.05	1000

7. After determined the sample size accordingly, we could start the experiment.
8. After completing the experiment, we need to calculate the confidence interval and determine whether we should reject the null hypothesis or not.
    a. We could use z-test: calculate the z-score according to the results, and see whether the p-value corresponding to the z-score is lower than 0.05, if it is, then we should reject the null hypothesis, saying compensate the toll-fee could improve drivers to cross-city to a significant practice level, and we should do it!

Part 3 - Predictive modeling

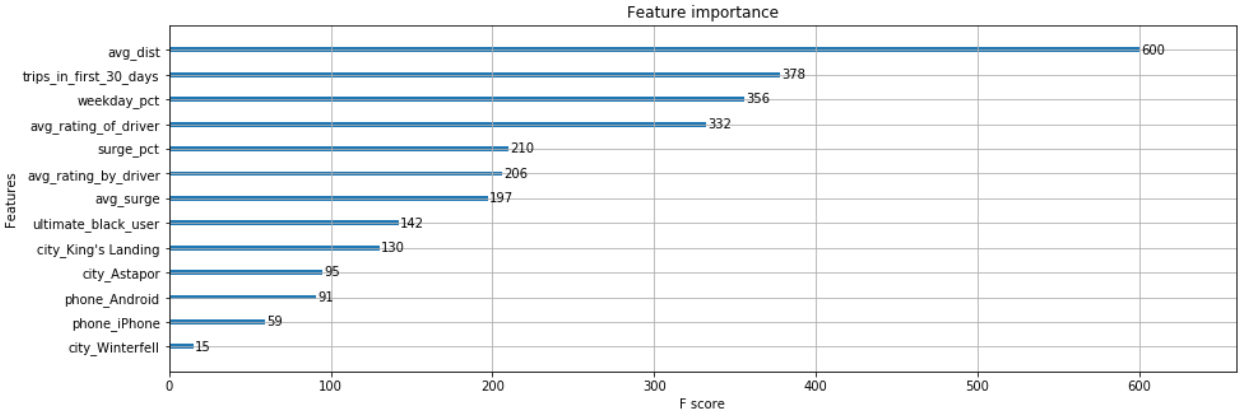
1. Data cleaning:
    a. Convert json file to pandas dataframe.
        i. Total 50000 drivers. 12 columns. Missing data in ‘avg_rating_of_driver ’, ‘phone’, ‘avg_rating_by_driver’.
    b. Get outcome from last-trip-date according to whether it took a trip in the last 30 days (last-trip-date after ‘2017-06-01’).
    c. Active user percentage is 0.3662 from total 50000 signup users.
    d. drop last-trip-date as the input matrix, this is the outcome.
    e. Convert data to dummy table: convert all categorical data to numbers (one hot encoder).
    f. Run train-test-split with test size = 0.3
2. EDA:
    a.



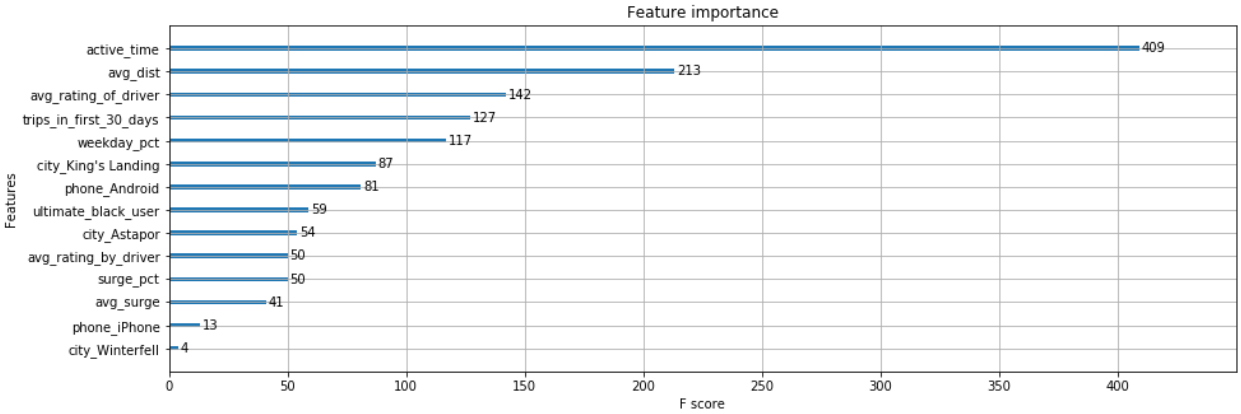
    b. There is a big fraction of user choose not to retain within 30 days.
    c. Then a stable losing rates. Another peak appeared around 150 days.

3. Run xgboost classification model with default hyperparameters to fit training sample. First model: without feature engineering
    a. Predict testing data
    b. Testing data validation:
    c. Testing accuracy is 0.787
    d. Testing confusion matrix:
    array([[8190, 1297],
    [1899, 3614]])
    e. Testing f1_score: 0.693

f. Feature Importance:



Second model: with feature: active_time, which is the total time the user retained, unit in days.



F1_socre > 0.94. Active_time is a most important indicator.

4. Insight:
   a. According to the model, the most import features are:
      i. **active_time:** how long the user retained
      ii. **avg_dist:** the average distance in miles per trip taken in the first 30 days after signup
      iii. **trips_in_first_30_days:** the number of trips this user took in the first 30 days after
      iv. **weekday_pct:** the percent of the user's trips occurring during a weekday
      v. **Avg_rating_of_driver:** the rider's average rating over all of their trips
      vi. **surge_pct:** the percent of trips taken with surge multiplier > 1

Active driver?	True	False
avg_dist	5.118977	6.188478

trips_in_first_30_days	3.349590	1.659167
weekday_pct	62.214604	60.181597
avg_rating_of_driver	4.593364	4.607018
surge_pct	9.128165	8.688548
avg_rating_by_driver	4.763121	4.786876

vii. According to the differences in the above table, we could see

1. the average distance per trip in the first 30 days is critical: if in the beginning, each trip is shorter, the user tend to retain.
2. If the users took more trips during weekday, they tend to retain. If the trips took during peak hours(when there is a surge), they tend to retain.
3. Also, the longer the user retained, the more likely they still retain.
4. All of the above information told us, the users that has inelastic demand, i.e., they have to use the ride, would retain, most likely they need the ride for work (during peak hours on weekdays, distance is shorter, and need to frequently took ride).

viii. Conclusion: Users retain because they have the need to commute. What we could do is, give discounts during the first month; attract those users that are likely to retain, by giving them discounts. Adding some promotion strategies, such as get a discount every 10 trips, receive cumulative incentives if they would keep active. For the users in the margin side, they are most worthy to advertise.