Ultimate Challenge

By Muluemebet
June, 2020

Part 1: Exploratory data analysis

There is a daily cycle in the number of logins. Most logging occurs from 21 to 2hr and from 11 to 12hr. There was an exceptionally large login on March 01, 1970 from 04:30 to 4:45 hr (see the Figure-1 to Figure-3).

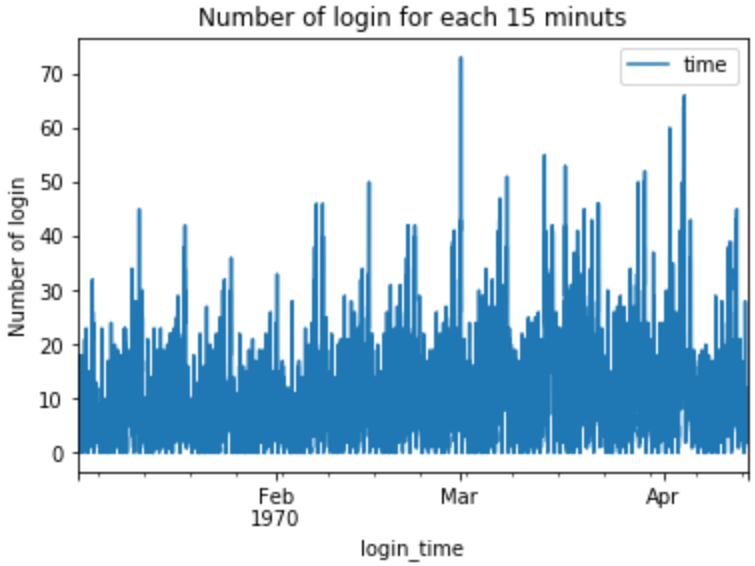

Figure -1

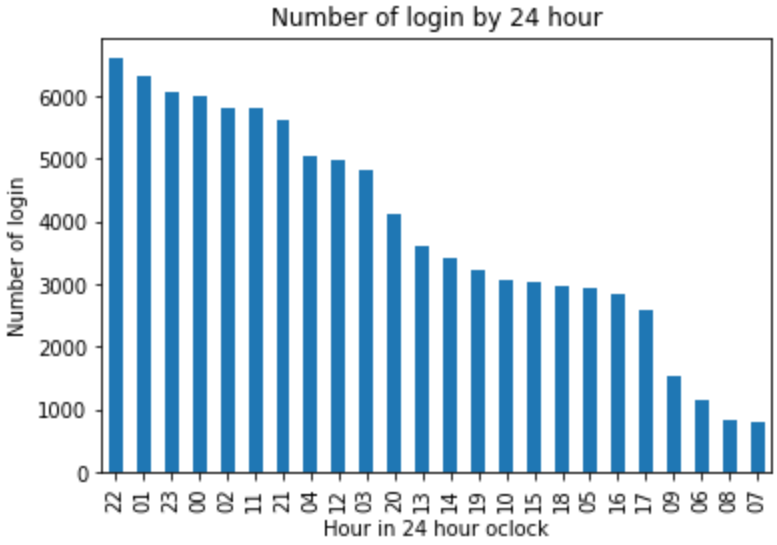

Figure -2

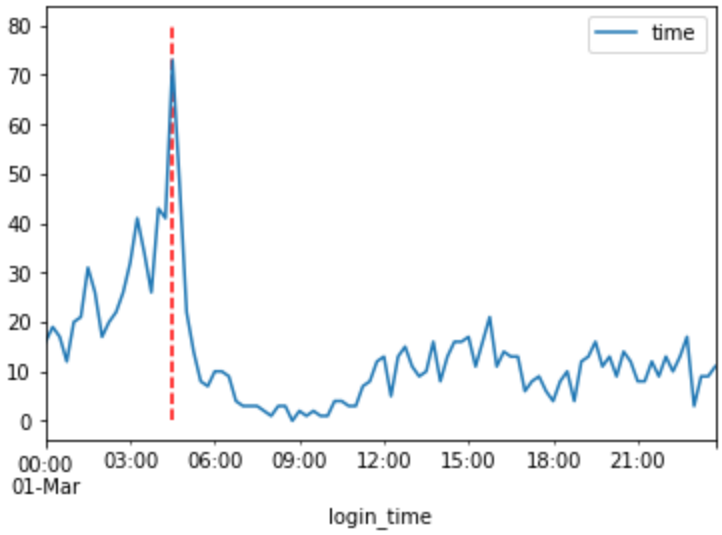

Figure -3

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?
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 b. what statistical test(s) you will conduct to verify the significance of the observation c. how you would interpret the results and provide recommendations to the city operations team along with any caveats.

Solution:

The key measure of success of this experiment is the number of times drivers availability to their neighbouring city is measured in terms of the number of times they cross the toll bridge for each day.

Steps to make experiment

Step_1 Define research questions and variables: The objective of the experiment is to study the effect of reimbursing all toll costs of drivers on drivers' availability in both Gotham and Metropolis. The dependent variable is the number of driver's trips between the two cities. The

independent variable is whether the driver is told to be reimbursed or not. Other confounding variables such as gender, driving experience, age, distance from driver's residence location to the neighbouring city and trip distance may affect drivers presence in both cities.

Step_2 Write the hypothesis:

H0= Reimbursement of toll cost doesn't affect the number of trips of a driver between the two cities.

H1: Reimbursing toll cost increases the number of driver's trip between the two cities.

Step_3 Design experimental treatments:

The treatment groups are drivers who are told that they will get reimbursement for all toll costs. The control groups are drivers with no information about reimbursement. Then, count the number of times a driver crosses the bridge(both inbound and outbound) during the study period. This information can be collected from toll receipt or invoice. Since the experiment will be conducted for a number of days, a driver will have a number of records. Only the aggregated value, that is, the average number of trips during the study period will be used for the analysis. The mean of the average number of trips for controlled and treated groups will be computed separately. Choose test statistics based on the distribution of the data(could be z-test or t-test), set a significant level to reject the null hypothesis and then compute the test statistics. Finally, compute the p-value which is the probability of obtaining a value of test statistic that is at least as extreme as what was observed under the assumption the null hypothesis is true. If p value is above the significant level, accept the null hypothesis otherwise reject the null hypothesis and accept the alternative hypothesis.

Step_4 Assign subjects to treatment groups

To make the sample more representative, a randomized block design is used. According to this method, first the drivers will be grouped into a character they share(gender,driving experience, age, distance from driver's residence location to the neighbouring city and city of residence) and then randomly assigned to a controlled and treatment group.

Part-3: Predictive modeling

The variables avg_rating_of_driver, phone and avg_rating_by_driver have missing values. All of the signup was performed during the month of January whereas the last trip date went from January to July. 69.22 % of the users had trips in the first 30 days of their signup(see Figure-4). A model was made to predict user retention and to determine the most important variable. The variable ‘avg_surge’ has the largest coefficient and is assumed to be the most important predictor variable.

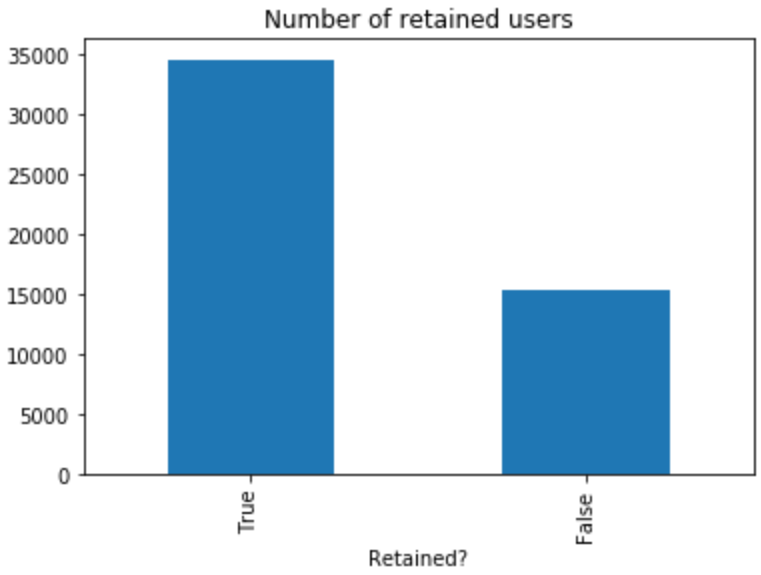

Figure-4

Modeling: The outcome variable is not directly provided from the data. We need to create it from the data. We are interested to determine whether or not a user will be active in their sixth month on the system. To assume the user is active within six months of signup, his/her last trip date has to be at most the end of the six month. Thus, the day between signup and last trip date is an important variable to be computed. If this value is less than 6 months, we can assume that the user is active. But if the value is greater than six months, it is difficult to assume that the user is not active. This is because the user can have the last ride after seven months of signup with other rides in between. Another variable we need to look at is, trips within 30 days. If we know that the user has trips within 30 days and his/her last trip is after the seventh month, we can still say that this person is active. To say a person is not active within six months, his/her first trip has to be after the six month of signup or not riding at all. There is no such information from the data. Let us assume that users who didn't ride within 30 days and their last trip was after six months are not active(which could not be true in reality if users had their first ride between the first month(after 30 days) and the sixth month). The six month period is assumed to have 180 days(6*30) for simplicity.

This problem can be modeled using classification techniques. The outcome variable can be labeled as active or not active depending on the computed time length between signup and last trip date as well as trip information within the first 30 days. From the given data, almost all users ride within six months of signup(see Figure-6). Thus, it is difficult to make a model with this approach without having non- active users.

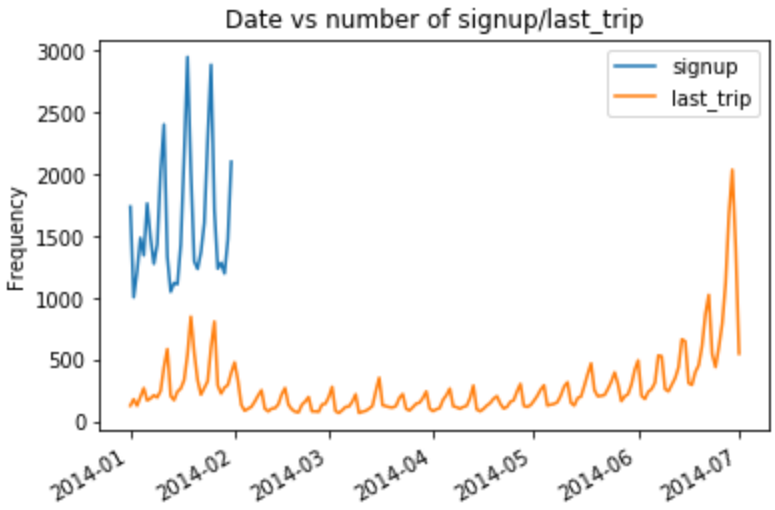

Figure-5

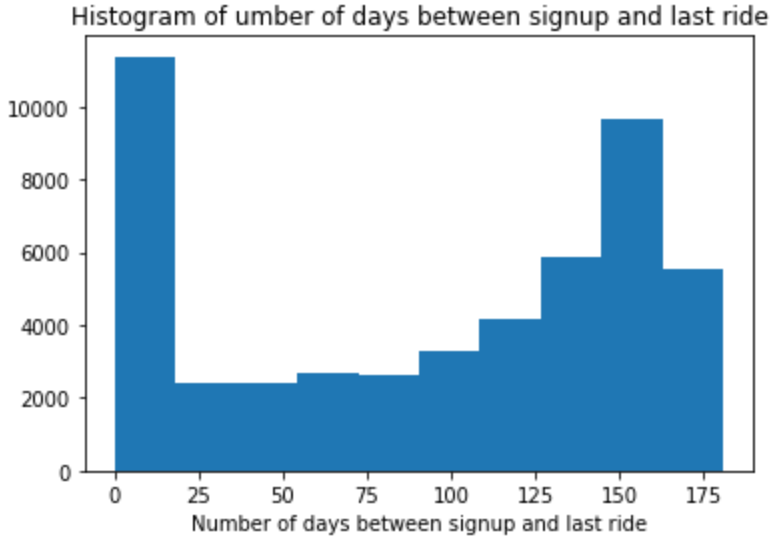

Figure-6