Data Analysis Interview Challenge

Part 1 - Exploratory data analysis

The attached logins. json file contains (simulated) timestamps of user logins in a particular geographic location. Aggregate these login counts based on 15minute time intervals and visualize and describe the resulting time series of login counts in ways that best characterize the underlying patterns of the demand. Please report/illustrate important features of the demand, such as daily cycles. If there are data quality issues, please report them.

The time series data containing the timestamps of user logins with 93142 observations (rows). we found 877 duplicated values and we dropped them. The time series length is from January to mid April1970.

The plotting of time series based on 15 minutes intervals (aggregate) did not help us to better visualize the underline pattern. Time series decomposition splits the series into several components. This helped us to better visualize the underline pattern. we found that:

- the time series tend to be multiplicative with the constantly increasing amplitude
- the series has a weekly seasonal pattern.
- the trend is upward with a slight decrease at mid-April (end of the series)
- the Noise is pronounced, this variability could impact model building and cannot be explained by the model.

Next, plotting the time series based on 15-minute intervals (aggregate) by day, month, week_day revealed that:

- day: users are more active on the first 13 days of the month, and moderately so between the 14th and 27th day, they are less active at the end of the month
- months: users are more active in the first 3 months and less active in April.
- week_of_day: users are more active from Sunday to Monday and less active from Tuesday to Thursday.

Finally, we assessed the stationarity of the time series using Augmented Dickey-Fuller Test (ADF test) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests. in fact, A Stationary time series is one whose statistical properties like mean, variance, covariance does not vary or change over time. according to ADF test, the series is stationary and, KPSS test showed that the series is not stationary.

Part 2 Experimental and Metric design

The present section aims to conduct an experiment in encouraging drivers' partners to serve neighboring cities (Gotham and Metropolis). the cities are separated with a two-way toll, and having complementary circadian rhythms, day active versus nigh active, except the weekend. **the goal is to this encourage driver partners to become more available in both cities**.

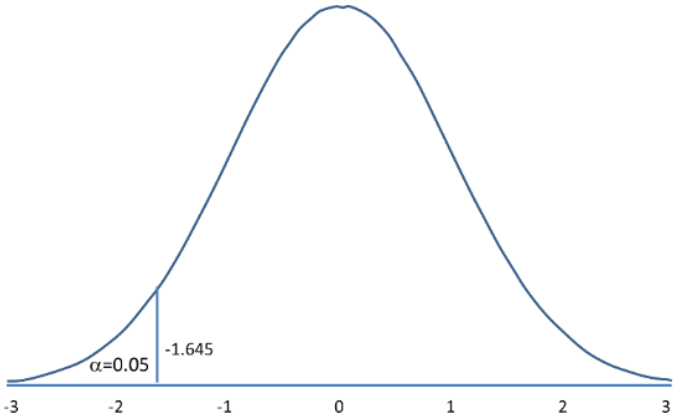
What would you choose as the key measure of success of this experiment in encouraging driver partners to serve both cities, and why would you choose this metric?

Given the context, the key measure of success would to tract the **average of toll collected during the weekend only.** There is some bias that exist during weekdays therefore, we can only consider tolls collected on the weekend. We chose this metric, because it would allow us the measure the effectiveness of the action taken that: will the reimbursement of toll fees encourage drivers to work for both cities?

Describe a practical experiment you would design to compare the effectiveness of the proposed change in relation to the key measure of success.
a. how you will implement the experiment

- get the list of all drivers
- shuffler the list
- randomly select drivers for group A (control group) and group B (experimental group)
    - no reimbursement to drivers in group A
    - reimbursement to group B drivers
- State the hypothesis
    - The null hypothesis: the average toll of group A is lower than that of group B.
    - The alternative hypothesis: the average toll of group A is not lower than that of group B.
- choose the significance level alpa
- Calculate required sample size
- Collecting information's until there is enough observations
    - Type of information's to collected (driver name, address , trip distance, specified the group (control group (A) or experimental group(B)), specify the day(weekend or weekday), and toll (city A-B)/(cityB-A).
    - NB: group A do not pay toll, but group B pay toll
    - NB. If possible, it is preferable to collect as much information as possible during the experimentation that can help you later on to deepen the objective of the study.
- run the test with group A and B (you could implement the central limit theorem to sure make that samples are normally distributed)
- Calculate the mean, standard deviation, z score and the p-value

b. what statistical test(s) you will conduct to verify the significance of the observation
We are going to conduct A/B test with 2 sample z test (sample size should be greater than 30) , according to the null hypothesis, it would be one tail Z test (lower tail test).



Rejection Region for Lower-Tailed Z Test (H_1: $\mu < \mu_0$) with α =0.05
The decision rule is: Reject H_0 if $Z \leq 1.645$.

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

Interpretation of result will be guide by the P-Value where:

- If p-value <= alpha: significant result, reject null hypothesis (Ho: the average toll of group A is lower than that of group B), it implied that the reimbursement of toll will not encourage drivers to serve both cities.
- If p-value > alpha: not significant result, do not reject null hypothesis, which mean that the reimbursement will encourage the drivers to serve both cities.

The caveat of this experiment is that, in all tests of hypothesis, there are two types of errors that can be committed. In fact, **correct decision** is when we do not reject true null hypothesis, **Type I error** is when we reject true null hypothesis and **Type II error** is when we fail to reject false null hypothesis.

Part 3 - Predictive modelling
Context:

Ultimate is interested in predicting rider retention. we consider a user retained if they were "active" (i.e. took a trip) in the preceding 30 days. The goal is to understand what factors are the best predictors for retention, and offer suggestions to operationalize those insights to help Ultimate.

A-1/Perform any cleaning, exploratory analysis, and/or visualizations

We first performed data acquisition and wrangling which involves gathering raw data and preparing it for processing and analysis. Here, we explored and performed data cleaning. The raw dataset is 50000 rows and 12 columns. The missing data were present in 3 variables (avg_rating_of_driver, phone, avg_rating_by_driver), we found duplicated observations (n=8).

Next, we did an exploratory data analysis where you explore features to identify some important characteristic of the dataset. After the identification of active users, the dependent variable was 37.608 % of active users and 61.392% non-active users. The exploratory data analysis reveals that, the most active user (approximately 18%) sign up in Winterfell city as compared to other cities, the most active user (approximately 33 %) sign up with Iphone. The average rating is between 1 and 5, its distribution of reveals that the probability to be scored 5 is high as compared to 4 and 3. Been scored one or two are less likely to occur and, are viewed in boxplot as outliers (natural outlier, not due to measurement or entry). More, outliers are visualized in the boxplot of average surge, surge_pct, average distance.

A-2/ What fraction of the observed users were retained?

Retained user ("active") is those who took a trip in the preceding 30 days. We found that 37.608 % was active.

B-1/Build a predictive model to help Ultimate determine whether or not a user will be active in their 6th month on the system.

We started with data preprocessing where we replaced missing values (the rating feature with the average value, and phone was set in binary value (1, 0)). We labeled encoded categorical features. We dropped signup date and last trip date. We could had used Z-score method or Winsorization method for outlier treatment, since we are going to use a tree base algorithm, there are robust to outliers.
Our target variable has two outcomes, so it is a binary classification. We used a linear classifier (logistic regression) and tree base algorithm (because are not sensitive to noisy data or outliers and usually provide good performance).

Logistic regression, random forest and XGBoost was the algorithm used. To improve the quality of models (being able to perform to unseen data), we performed the cross validation and split the data into train and test set (75/25). With an imbalance dataset (approximately 38% over 62%), we considered F1-score as the metric for model evaluation and selection. We also look at the value of area under the curve.
After training and evaluating each model, from the classification report, we found that:

- logistic regression: F1_score is 0.42, AUC is 0.677
- random forest: F1_score is 0.66, AUC is 0.807
- tune random forest: F1_score is 0.67, AUC is 0.839
- XGBoost : F1_score is 69 , AUC is 0.844
- tune XGBoost: F1_score is 69, AUC is o.841

The model that best predict rider retention is XGBoost. The F1_score and the area under the curve is good as compared to other. Larger the area under the curve, better the model of distinguish both class.

C/ Briefly discuss how Ultimate might leverage the insights gained from the model to improve its longterm rider retention (again, a few sentences will suffice).

The XGBoost feature of importance relieved that the average distance is the 1st , weeday_pct is the 2nd and trips_in_first_30_days was the best features in predicting the rider retention. Ultimate could focus on these features and develop an incentive to leverage the improve the rider retention.