Al Alipour
November 27, 2021

Data Analysis Interview Challenge

Part 1 - Exploratory Data Analysis

Visualizations created based on aggregating login counts for 15-minute time intervals indicate that user logins illustrate a daily pattern (please refer to the Jupyter Notebook for this part of the project for visualizations). They also indicate that weekends influence the user logins considerably as well. Accordingly, the daily pattern observed for most of the days is that login counts peak to their maximum at around midnight and from there plummet to their lowest at around 6 am. User login counts then rise again to a high level at around noon, followed by another drop to their minimum at around 6 pm. From 6 pm, login counts rise again to reach their maximum at around midnight. Considering weekends, login counts are normally higher during weekends with the weekly maximum being around late hours of Saturday to early hours of Sunday.

Part 2 - Experiment and Metric Design

Q1. Key measure of success for this experiment would be the average percentage of trips each driver from either city completes that has a pick up location, destination, or both in the other city. I would choose this metric because it would show if the drivers still prefer to ride trips in their own city or if reimbursing toll costs has made a meaningful difference in encouraging the drivers to ride trips to, from, and inside the other city.

Q2. a) Given the available budget and resources, I would select a number of drivers from each city and first acquire data on their current performance with respect to the key performance metric. This would be their performance on the metric defined in the answer to Q1 for the past several weeks. I would then implement toll cost reimbursement for these drivers for the same number of weeks and see how their performance changes accordingly.

Q2. b) With the data collected, I would see if there has been an important change in the performance of the drivers from pre- to post-implementation periods. I would then perform bootstrapping to see if the p-value of the change is statistically significant.

Q2. c) If the change in drivers' performance is significant, I would make a recommendation to the city operations team that reimbursing toll costs should probably encourage the drivers to serve both cities. I would also let them know on the probable level of change. Moreover, I would inform them about the limitations of the experiment. For example, the experiment has been conducted among a limited number of drivers so only a selected number of drivers have had the possibility of being reimbursed for their toll costs. This may have provided unique opportunities and advantages for these drivers that may no longer exist when toll cost reimbursement is implemented for all drivers. Thus, the drivers may no longer be as motivated to ride in the other city as they were when only a group of them had this advantage. Another caveat would be that pre- and post-implementation periods of toll reimbursement during the experiment may have been affected by different factors that made them not necessarily exactly the same. For example, post-implementation period may have been in a time of the year that naturally attracts the drivers to ride more frequently in the other city compared to the pre-implementation period. This may be due to better weather and road conditions, quieter streets, or higher demand.

Part 3 - Predictive Modeling

Q1. At first, data were studied using common approaches for gaining an initial understanding. Correlations were also studied between different features. Data were further studied with respect to missing values. Multiple visualizations were created to see if the missing values could be imputed and whether the missingness was random or systematic (please refer to the Jupyter Notebook for this part of the project for visualizations). Given the definition of a retained user, data were analyzed to find out what fraction of the users had taken a trip in the preceding 30 days to the last date for which data was available. Accordingly, 37.6% of the users met the definition of a retained user.

Q2. In order to predict whether a user will be active in their sixth month, multiple machine learning algorithms were analyzed. Please note that for the sake of brevity none of the models underwent hyperparameter tuning. The algorithms tested included logistic regression, random

forest, gradient boosting, and support vector classifier. Another factor considered during predictive modelling was regarding missing values. Three features of the data included missing values, and among them phone type and average rating by driver missed data on only a small number of observations. Thus, the rows associated with missing values on these two features were simply dropped. The other feature including missing values was average ratings given to drivers. This feature included missing values for over 8,000 observations out of the 50,000. Imputing the missing values was not possible for this feature given the available data and information. Thus, each machine learning algorithm was tested twice: once by dropping this feature and using all of the observations and once by dropping the observations associated with missing values on this feature and using all of the features.

Among the tested algorithms, gradient boosting had the best overall performance. Gradient boosting also had the best performance in terms of correctly classifying a retained user. However, random forrest had the best performance in terms of correctly classifying an un-retained user. Therefore, some of the concerns about the gradient boosting model include its ability in correctly classifying an un-retained user as well as the fact that the model has an overall accuracy of 0.79 and 0.78 on the first and second test sets which should be improved for example through hyperparameter tuning among other possible methods. Please note that model rankings did not change between the first and second test sets as average ratings given to drivers was not among the most important features.

Q3. Feature importance analysis based on the gradient boosting model indicates that currently average ratings by drivers followed by the city of King's Landing, surge percentage, and weekday percentage are the most important features in predicting whether a user will be active in their sixth month. Thus, Ultimate can focus on these factors to improve its long term rider retention. For instance, as the riders who get better ratings from the drivers tend to become retained users, Ultimate can study what are the characteristics of these riders and provide services to attract more riders of this type. Another example would be putting more focus on King's Landing as riders from this city tend to become retained users as well. Also, Ultimate can study why riders from King's Landing have a higher tendency than the other cities to become retained users and try to provide the same conditions in the other cities as well if that is possible. There are many other insights that can be gained from the model, which are beyond the scope of the current project.