Ultimate Inc. Take Home Challenge

Part 1: Please see my Jupiter notebook.

Part 2: Experiment and Metrics Design

There would be two ways to track the effectiveness of this campaign. One would be to track the locations of the drivers using smartphone apps, and see the percentage of time during working hours that they spend in either city. Another way to track effectiveness would be to calculate how many drivers get reimbursement for the tolls during time they spent driving during their working hours, however this may not be effective because this variable will automatically jump from zero to something greater than zero, and will therefore be statistically significant and not informative since it is basically guaranteed. Location should be tracked.

One thing to keep in mind is that working hours will differ depending on the city that the driver is predominantly driving in, due to the different circadian rhythms of the cities. Unless the driver does not need sleep, their activity will probably remain unchanged during the week even with toll reimbursement in place. Therefore, since people are very active on the weekends usually, and assuming driver location can be tracked, we propose tracking the location of the driver during the weekend as the variable of interest for the experiment.

As for statistical testing, bootstrap testing can be used, but a parametric test using the normal distribution will be better. Under the Central Limit Theorem, given a sufficiently large sample size, the sampling distribution of the mean will approach the normal distribution. We can use the normal distribution over the t-distribution, despite not knowing the population variance, because we will have a high number of degrees of freedom in a large urban population. We will employ a z-test to find the difference in means.

Steps of experiment:

1. Take a random sample of no less than 100 drivers
2. Calculate the percentage time that they spend in each city on the weekends
3. Implement the toll reimbursement policy for a set period and notify the drivers
4. After the period is complete, sample the same drivers and calculate the percentage time they spend driving in each city on the weekends, again.
5. Using a Z-test for proportions, at a chosen confidence level (99%), test the samples under the following setup

    H0: There is no difference in the mean proportions of times spent in each city

    Ha: There is a statistically significant difference in the mean proportions of times spent in each city

It is important that this experiment be conducted on normal weekends that are not holidays and with no major events unless said events are ordinary, to prevent from artificially inflating the results. Confidence Intervals should be generated for the mean proportions as well in order to characterize the change in proportions over the test, if any. Also note that the higher the sample size, the greater chance of their being a statistically significant difference detected, even if that statistical significance does not translate to practical significance. The results of the experiment should be interpreted in that light.

If the percentage of time spent in the non-dominant city goes up with statistical significance, then the toll reimbursement has been effective in driving this effort. As for prediction of the profit results, there must be a comparison in the end between the total cost of the toll reimbursement vs the increased profit from having drivers available in both cities. The toll reimbursement could be effective and still lose money for the company. It is critical that this final step is taken.

Part 3: Predictive Modeling

Data Preparation

Data was imported into Jupyter from the file ultimate_data_challenge.json. Rows were deleted that did not contain information on the Phone feature. Missing values in the average_rating_by_driver and average_rating_of_driver were filled in using the average value of that column.

EDA

Several charts were generated for variables of interest, of which two are presented in this report. We have chosen the violin plot since we have a binary target variable, and also many data points in the dataset. The first shows the average distance per trip in the first 30 days of app use, by whether or not the user is still active or not.
We can see that the average for inactive users is slightly higher, though active users have a large spike. The second chart shows the average rating given by the driver to the rider within the first 30 days, by activity.
Inspecting the chart we can see that active users have a lower overall rating than inactive, which is counterintuitive.

Predictive Models

Random Forest and Logistic Regression were employed because of the outputs of each model. Besides aiming for accurate classification, Random Forest outputs features importance for each feature, which will give us insight into the most impactful variables. We also need to see the effect of the variable on the odds of being an active user. The coefficients of the Logistic Regression model do just that for us.

Both models gained about 75% accuracy after hyper parameter tuning. While this is good, it could be improved on in the future with several things: 1. More data, 2: Different model runs 3. Finer hyper-parameter tuning. Putting the predictive accuracy aside, the models yielded important insights for the company. Combining the insights from both models, we have developed a list of recommendations for the company to maximize revenue.

Recommendations

1. Offer incentives to riders to use the app on the weekends more than the weekdays.
2. Incentivize riders to take more trips during their first 30 days on the app.
3. Encourage drivers not to artificially inflate the ratings they give to users, especially if the drivers can tell that they've never done this before.

We are confident that these recommendations will have a positive impact on revenue for the company and will increase the total number of active users for the long term.