**UBER Analytics – Wait Time Analysis**

*Project done by Guilherme Osorio for the University of San Diego (USD), Knauss School of Business*

**Introduction – Goals & Model Choice**

The goal of this analysis is to inform Uber about how to manage wait times in the Uber Express Pool service. The experiment conducted which gave origin to the dataset used in the analysis explored different variables based on waiting times of 2 or 5 minutes. To measure the success of the waiting time used, the Data Science team decided to look deeper into the ‘rider\_cancellations’ variable. Therefore, the analysis will culminate with the design of a machine learning algorithm that can be used to predict rider\_cancellations based on the relevant variables in the dataset. Due to the target variable ‘rider\_cancellations’ being numerical, the most efficient approach to take is to build a linear regression model, and that is what this analysis accomplishes.

**Uber Express VS Uber Pool**

For a better understanding of this analysis, it is necessary to understand the Uber services in question – Pool and Express. In both services, riders share their ride with other riders, for a cheaper price. Express differentiates from Pool by also requiring riders to have to walk a short distance before the ride (to the meet up spot with the driver), and after the ride (to reach their destination). This results in Uber being able to charge an even lower price for the service.

**Pre-Regression Analysis**

The pre-regression analysis consisted of some exploratory analysis. As can be seen in Figure 1 and Figure 2, the ‘period\_start’ variable seems to change consistently through a pattern which cycle seems to represent a week. A binary variable named ‘weekend’ was created to capture this pattern. The value is 1 if the day is a weekend, and 0 if not. Next, the data was explored to look for dubious relationships between the independent variables to find potential multicollinearity issues through a correlation matrix. It was decided to remove one variable from each pair with a high correlation (above 0.7). Additionally, a scatter plot for most possible independent variables in the dataset was built to compare each variable with the rider cancellations (Figures 5 through 10). As all variables seemed to have at least a mild positive correlation with the target variable, we can assume that they have at least a mild correlation.

**Interpreting the Model**

After some linear regression assumptions were tested, it was time to fit the data to the model. Based on the pre-regression analysis, the variables which ended up being used were the treat variable, the trips\_pool variable, the total\_matches variable, and the weekend variable. All the variables used were statistically significant, with a p-value under 0.05. Based on the variables’ coefficients, a population model for the target variable can be designed: rider\_cancellations = 32.74 \* treat + 0.04 \* trips\_pool + 0.06 \* total\_matches – 14.18 \* weekend – 61.03. So, by increasing the wait time from 2 to 5 minutes, an increase of 32.73 rider cancellations are expected, an increase of 1 Pool trip increases cancellations by 0.04, an additional match is expected to increase rider cancellations by 0.06 and being a weekend day is expected to decrease rider cancellations by 14.18. The Adjusted R-squared of 74.8% shows us that 74.8% of the variability in the rider\_cancellations variable can be explained by the independent variables used in the model.

To check for linearity and homoscedasticity, a residual plot was built (Figure 11). Since the points are randomly scattered around 0, we can assume that those 2 assumptions are met in the model built.

**Customer Experience + Efficiency**

To capture how the customer experience is affected by the increase in wait times, four variables should be analyzed - rider\_cancellations, trips\_pool, trips\_express and total\_matches. To capture how efficiency is affected, the three variables – total\_matches, total\_double\_matches, and rider\_cancellations – should be analyzed. By analyzing results from the Mann-Whitney test, it is concluded that only one of those 5 variables does not vary from treatment to control group – total\_double\_matches.

**Conclusion / Final Thoughts**

The regression model predicts with statistical significance that increasing the wait time to 5 minutes creates a significant increase in the number of rider cancellations. As this is a key measure in terms of both customer experience and efficiency, the model seems to indicate that Uber should not increase wait times. Uber cannot afford to have its reputation damaged by providing poor customer experience in a crowded transportation market. However, Uber’s Uber X service seems to provide an alternative to Pool and Express and prevent Uber from losing customers to competitors. Therefore, I believe Uber should analyze the pros from increasing wait times and decide if those pros outweigh the cons from this analysis. Finally, there seems to be an increase in absolute rider cancellations during commuting hours. By analyzing the Mann-Whitney test, we cannot reject the hypothesis that the variable commute differs by treatment. Therefore, the wait time does not seem to be a key factor to this variable.

**VISUALIZATIONS**

**Visualization #1 – Line plot of evolution of rider cancellations**

Chart, histogram

Description automatically generated

**Visualization #2 – Scatter plot of evolution of rider cancellations**

Chart, scatter chart

Description automatically generated

**Visualization #3 – Bar plot of rider cancellations by group**

Logo

Description automatically generated with low confidence

**Visualization #4 – Bar plot of rider cancellations by commute and groupChart, bar chart

Description automatically generated**

**Visualization #5 – Scatter plot of commute VS rider cancellations**

Shape

Description automatically generated

**Visualization #6 – Scatter plot of trips pool VS rider cancellations**

Chart, scatter chart

Description automatically generated

**Visualization #7 – Scatter plot of trips express VS rider cancellations**

Chart, scatter chart

Description automatically generated

**Visualization #8 – Scatter plot of total driver payout VS rider cancellations**

Chart, scatter chart

Description automatically generated

**Visualization #9 – Scatter plot of total matches VS rider cancellations**

Chart, scatter chart

Description automatically generated

**Visualization #10 – Scatter plot of total double matches VS rider cancellations**

**Chart, scatter chart

Description automatically generated**

**Visualization #11 – Residual plot of linear regression modelChart, scatter chart

Description automatically generated**

**Visualization #12 – Regression plots for trips pool variable**

**Chart, scatter chart

Description automatically generated**

**Visualization #13 – Regression plots for total matches variable**

**Chart, scatter chart

Description automatically generated**