Ultimate Technologies Challenge Report

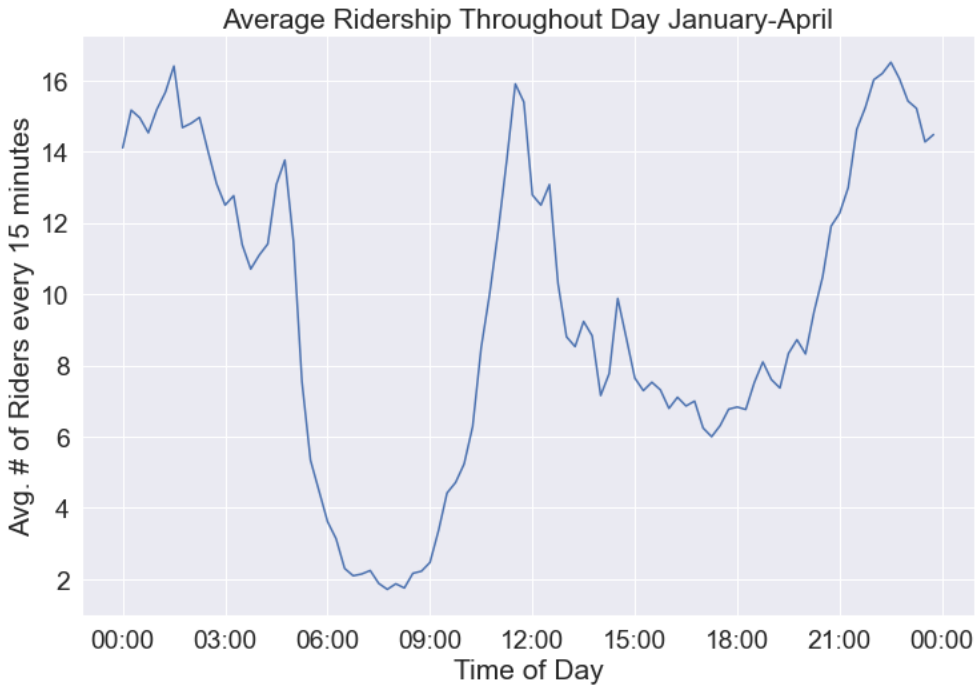
By Aisling Casey – July 2, 2021

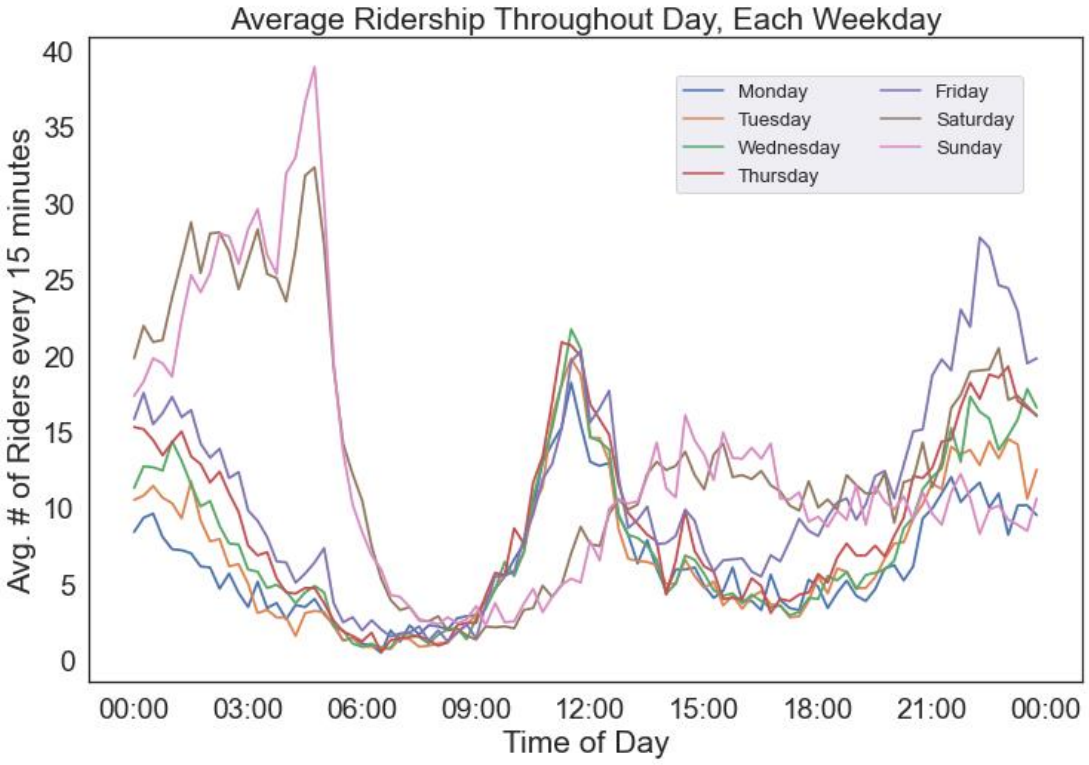
Part 1: Exploratory Data Analysis

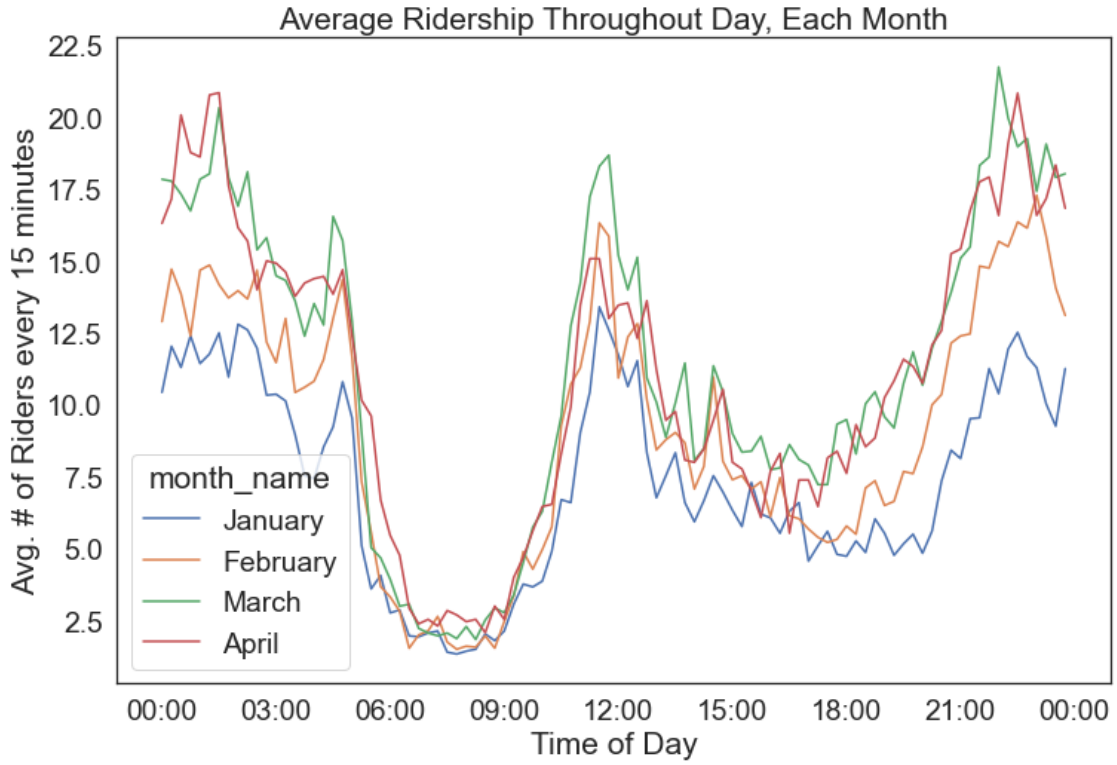
The various cyclical trends in the data are shown in the three graphs at the end of this section and described below; note that the average number of riders refers to the average taken every 15 minutes.

- **Daily:** Across all days, there is a demand minimum at 8am with an approximate average value of 2 rides every 15 minutes.
- **Weekdays** (Sunday Night->Friday Afternoon): A demand high of approximately 20 rides occurs at noon. There is an afternoon dip to about 5 rides, followed by an increase in activity at night and into early morning between 8 and 16 average rides.
- **Weekend** (Friday Night->Sunday Afternoon): A demand high occurs early in the morning around 5am, between 32 and 38 rides. There is no noon high as there is for the weekday, but rather a high of 15 rides at 3pm.
- **Monthly**: Roughly, January sees lower overall average demand throughout an average day, particularly during the late night/early morning hours. February sees slightly more, and March and April see the most (keeping in mind that the data for April is only available for about the first half of the month). The biggest difference in average daily value between the months occurs at 11pm, with an average of 12.5 rides in January compared to 21 in March and April.

No data quality issues to report other than the year being set to 1970 for all logins, which is mostly likely inaccurate. But this error did not affect the data exploration whatsoever.




Average Ridership Throughout Day, Each Weekday
Avg. # of Riders every 15 minutes
Time of Day
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
0
5
10
15
20
25
30
35
40
00:00
03:00
06:00
09:00
12:00
15:00
18:00
21:00
00:00


Average Ridership Throughout Day, Each Month
Avg. # of Riders every 15 minutes
Time of Day
month_name
January
February
March
April
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
00:00
03:00
06:00
09:00
12:00
15:00
18:00
21:00
00:00

Part 2: Experiment & Metrics Design

1. I would choose the metric of average proportion of company driver partners in both cities at various times throughout the week. I would choose this metric because ultimately what the city wants is drivers from both companies to be available to customers in both cities at any given time. This metric should control for time spent in traffic, average trip distance, city attractions and any other variables that might differ between cities.
2. a. To test the hypothesis that reimbursing toll costs to drivers will encourage driver partners to be available in both cities, there would need to be a control and an experimental period. In the control period, the costs of the toll would not be reimbursed; in the experimental period, they would. During both periods, the proportion of company driver partners in both cities would be measured at 30-minute intervals. The data would need to be collected over at least two weeks (one week for each trial period) to allow for changes in rider demand throughout the day and week. The exact number of weeks needed is not clear – it would depend on the expected size of the discrepancy between the experiment and control periods, which could be determined through pilot data collection. This pilot data could be collected over two days, and then used to approximate the necessary sample size.

   b. I would implement a paired t-test to determine if the means of the data from the control and experimental periods were statistically significant.

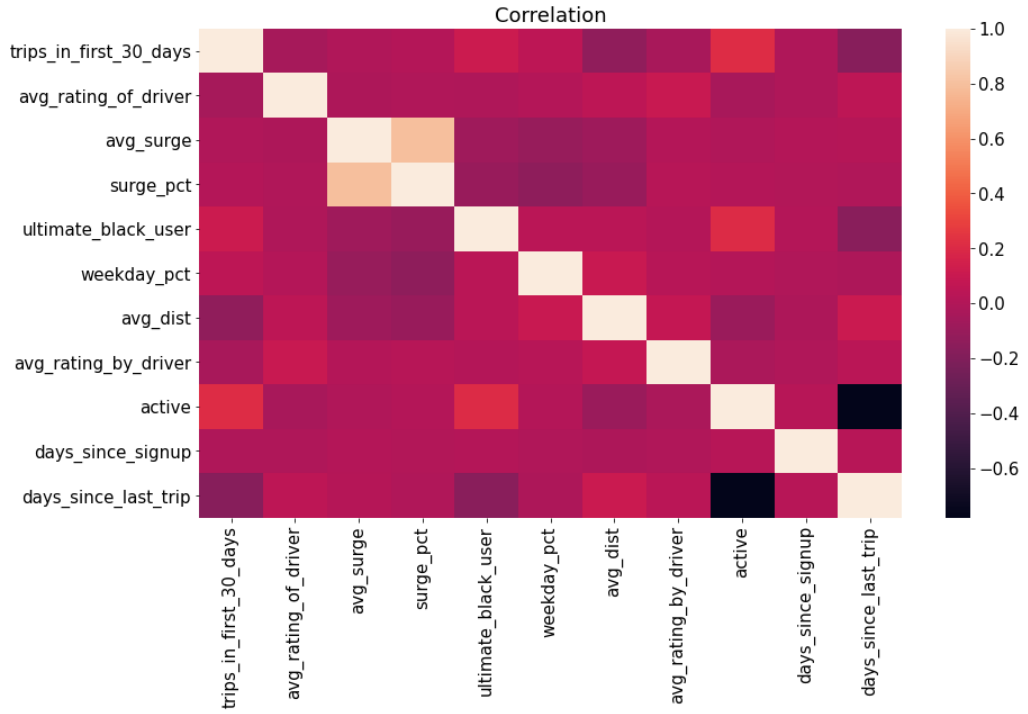
   c. After collecting the data and calculating the results of the paired t-test, if there was a statistically significant difference between the means of the trial periods, I would recommend that the city operations team implement the proposed change; but only if the predicted drop in toll revenue was worth having the observed increase in driver availability in both cities.

Part 3: Predictive Modeling

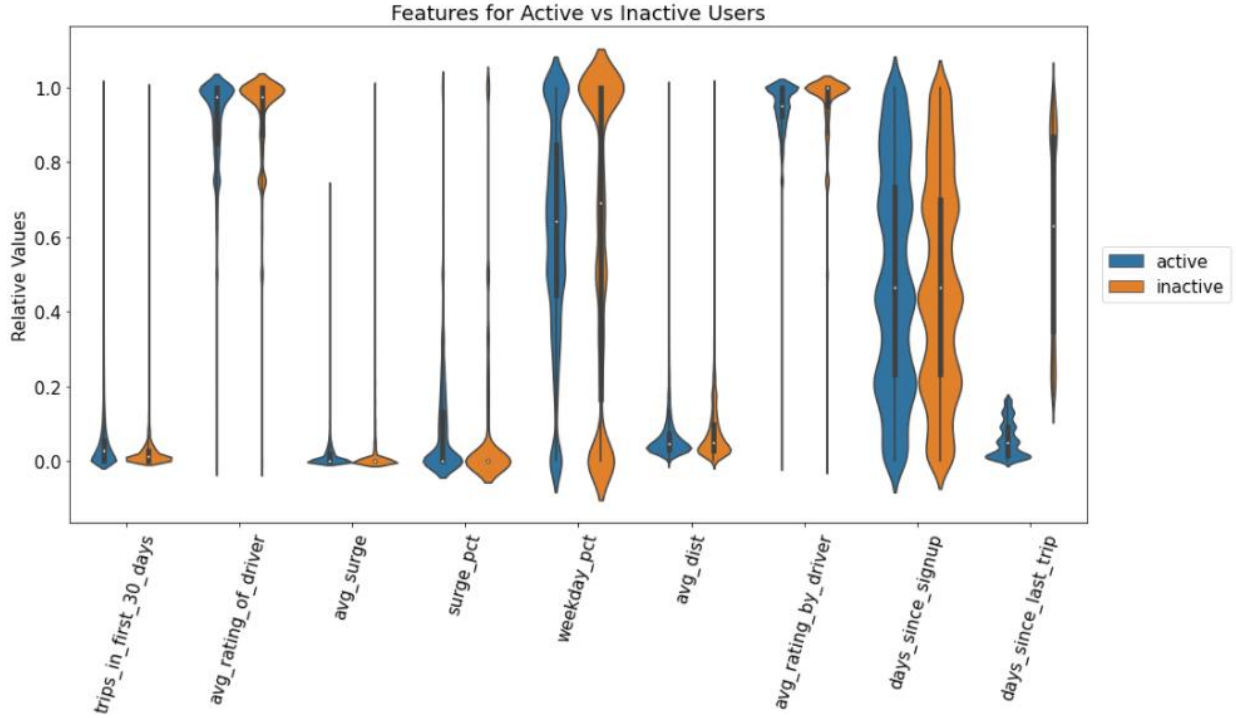
1. **Data Cleaning** - The following steps were taken to clean the data:
   a) A target variable column of active vs. inactive was created, based on whether the user had taken a ride in the previous 30 days
   b) Outliers replaced with next highest value in average distance, # of trips in first 30 days and average surge columns
   c) Any rows missing 'phone type' were dropped (396/50000 columns) while any missing rating values were filled in with the median value for the column.

   Exploratory Data Analysis - The following insights were gleamed from exploratory analysis:

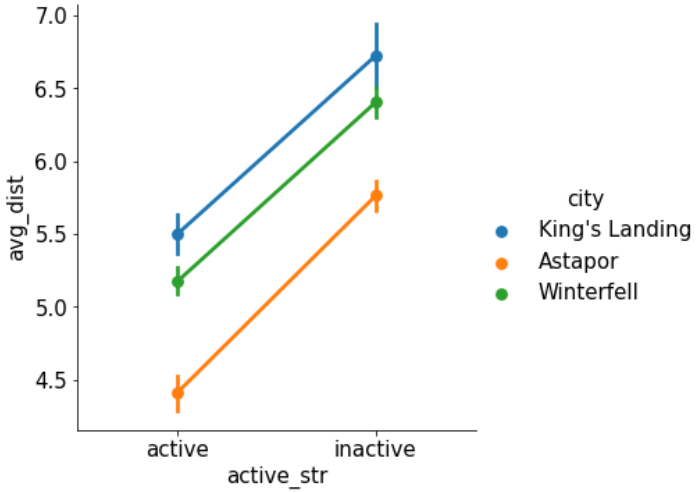
   a) 37.64 % of users in this cohort are active while 62.36 % of users in this cohort are inactive.
   b) As seen in the correlation heatmap below, most variables had a low correlation with one another; the exceptions were 'average surge' and 'surge percent', as well as 'active account' and 'days since last trip'. Since the 'days since last trip' feature was used to determine if an account was active or not, this strong negative correlation makes sense. But it also means that including it in the model would cause data leakage. So, this feature was removed from the data set before the model was created.



c) The violin plot below shows the feature distributions for the active users in blue and inactive in orange. To create this graph, the feature values were min/max scaled, resulting in their common range of 1 to 0. Other than 'days since last trip' and 'weekday percentage' the distributions for each feature look quite similar between the two groups. This means that for the model to distinguish the groups, it must rely on interactions between variables.



d) In the point plot below, we start to see how the latent interaction between variables may be useful in classification. When the mean value of average distance is plotted for each group and stratified by location, the groups become easier to distinguish.

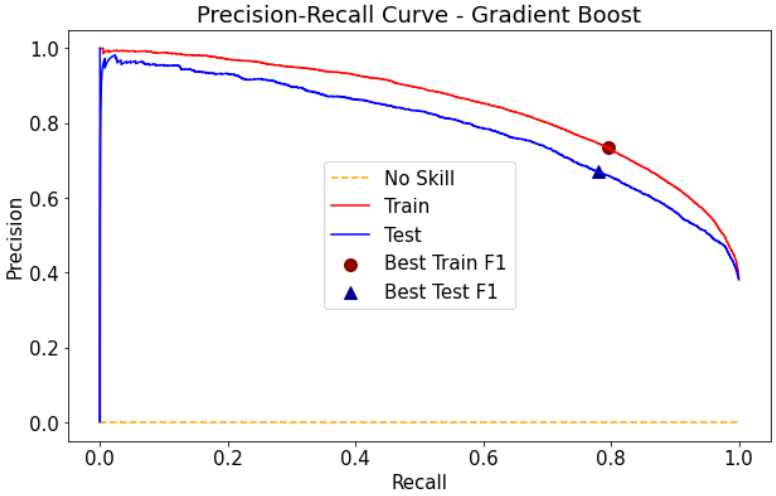


Preprocessing & Modeling-

a) 'Days since last trip' column was dropped.
b) Categorical variables were dummy encoded.
c) The data was split into training & test sets.
d) Min/Max scaling was fit on the training data and applied to both the testing & training sets.
e) Randomized search was used to determine optimal hyperparameters for each model.
f) Random forest, gradient boost and logistic regression models were trained and evaluated.

2. **Predictive Model:**

The final model selected was a gradient boosting model; its parameters are detailed in the model metrics file provided in the reports folder of this project's directory. The model had a training set ROC-AUC of 0.86, and a training set precision-recall AUC of 0.72, which is depicted in the graph below. It was important to consider the precision-recall AUC, since the classes are imbalanced.



The gradient boost model performed much better than the logistic regression model, which had a ROC-AUC of 0.79 on the test set. This was unsurprising after the EDA showed the individual features to have no obvious distinction between groups on their own; differences were more likely to be seen in interactions between variables, which tree models would be better at detecting.

The gradient boosting model performed marginally better than the random forest model, which had a ROC-AUC of 0.85 on the testing set. The random forest overfit on the training set, as it had a ROC-AUC of 0.93, hence its slightly worse performance on the test set.

For the final classification on the testing set, a probability threshold of 38.5% was set as the threshold; this probability yielded the highest f1 score on the training set. The f1 metric was chosen to make this decision because it better characterizes performance on imbalanced classes than a metric like accuracy. This probability threshold yields the classification matrix on the test set seen below. It has a false positive rate of 20.4%, a false negative rate of 24.6% and an accuracy of 78%. This is a reasonable performance and depending on whether the company would rather avoid false negatives more than false positives or vice versa, the threshold could be set accordingly.

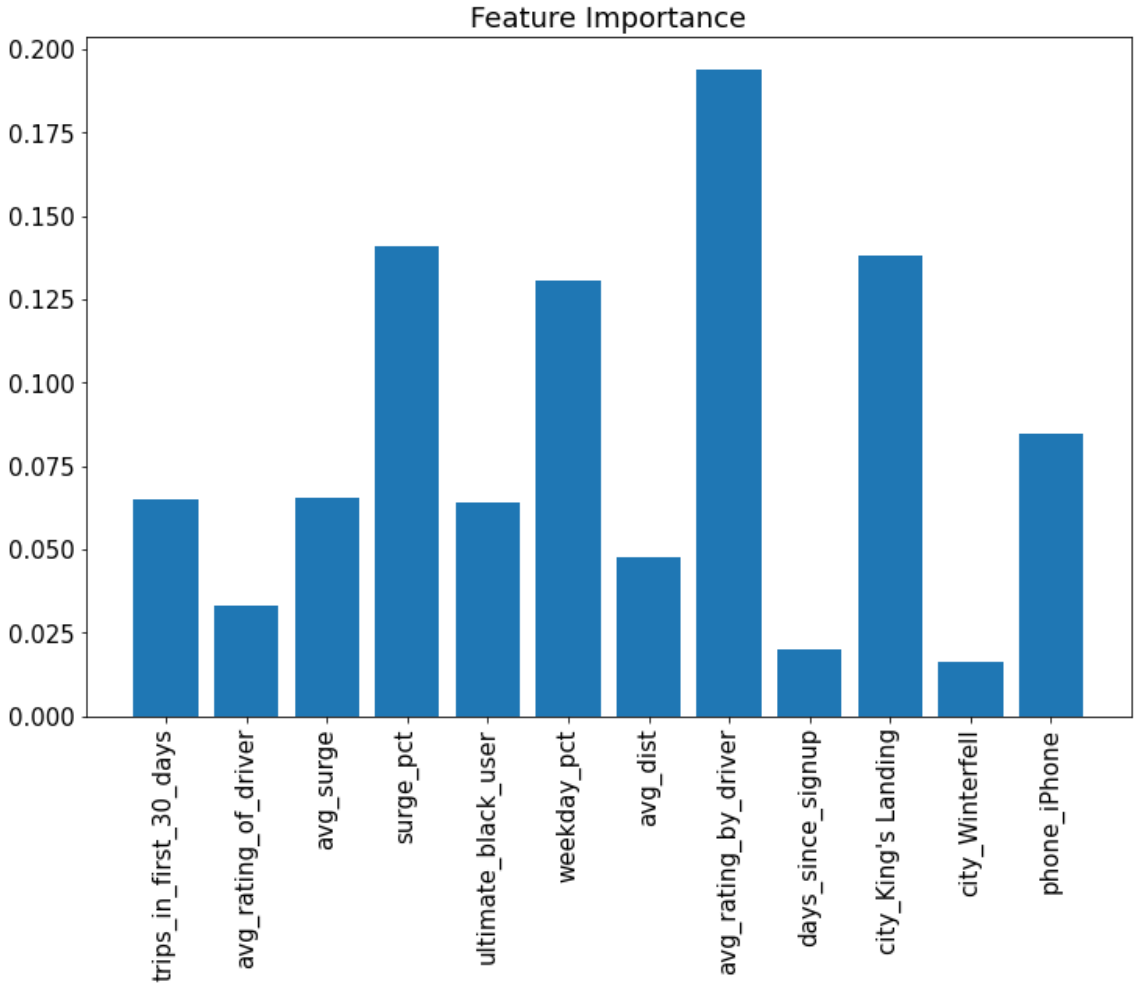
Classification Matrix

	Predicted Inactive	Predicted Active	Support
Actual Inactive	6163	1587	7750
Actual Active	1147	3504	4651

3. **Insights**:

The feature importance graph below shows the relative importance of each feature in the final gradient boosting model. Surprisingly, 'average rating by driver' was most important in distinguishing the two groups, followed by 'surge percentage', 'weekday percentage' and which city the user signed up for the app in, which all had similar importance values.

There is a confound in many of the features to consider. The active users take more trips on average than inactive users, so they have more chances of earning outlier values – more chances to score lower ratings as riders (i.e. really not getting along with their driver), more chances to run into large surges, more chances to take longer trips. So while 'average rating by driver' and 'surge percentage' are important to the model, I think that 'weekday percentage' and 'location of signup' are more insightful features, because they are more telling of rider profiles than just their app usage. Indeed, an unsupervised learning analysis to generate rider profiles might be even more insightful to the company than this predictive model in understanding long-term rider retention, and who to target with marketing.


Feature Importance
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
trips_in_first_30_days
avg_rating_of_driver
avg_surge
surge_pct
ultimate_black_user
weekday_pct
avg_dist
avg_rating_by_driver
days_since_signup
city_King's Landing
city_Winterfell
phone_iPhone