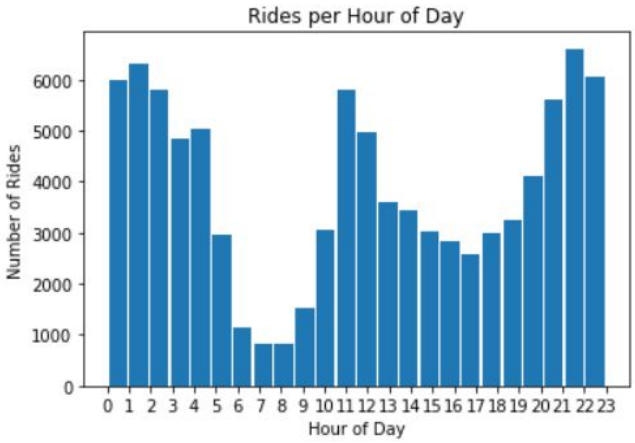
Ultimate Data Science Challenge

Jorge Londono

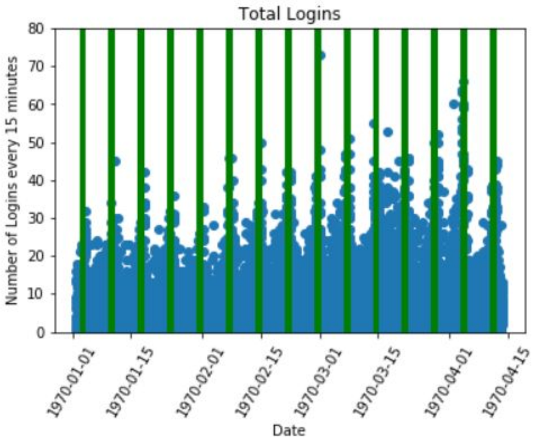
Part 1 - Exploratory data analysis

Hours of the day when service is mostly used



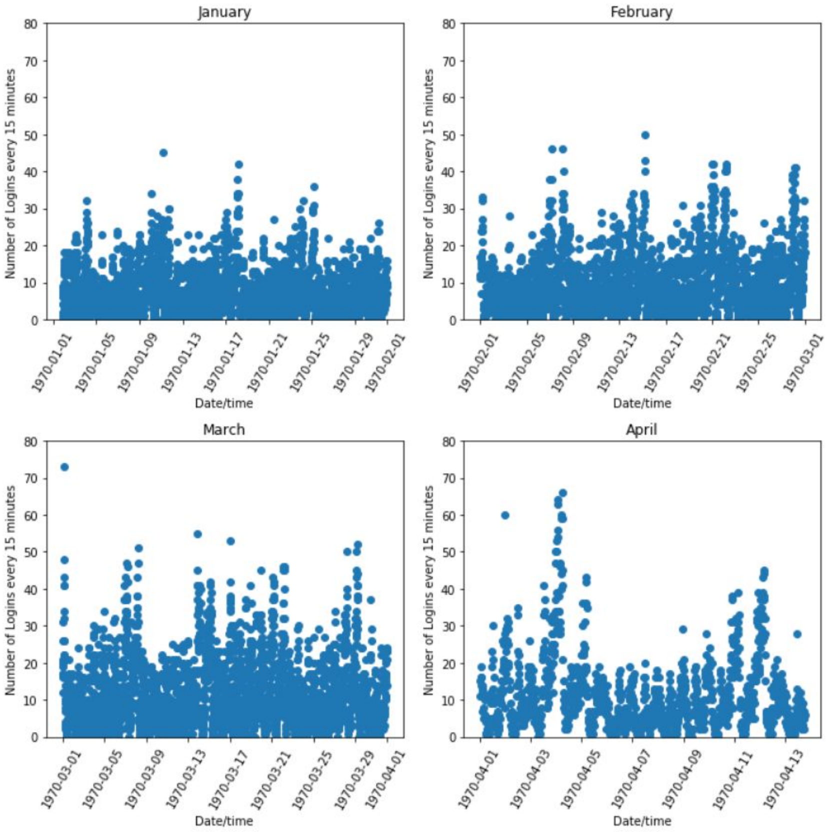
Rides are mostly used very late at night, as well as midday. The service is least used in the early morning and around rush hour.

How Weekends (Green Lines) affect service use



Service spikes during Weekends.

Monthly Usage



The service is growing steadily monthly, and there are no patterns aside from the aforementioned weekend service spike.

Part 2 - Experiment and metrics design

Metric

The best metric to evaluate the success of this experiment would be 2 different numbers:

1. The percentage increase of ridership at night in Metropolis
2. The percentage increase of ridership at daytime in Gotham

This metric would be a great evaluator, because we already know that the citizens from Gotham are active at night, and the citizens from Metropolis are active during the day. If we see that there is an increase of ridership during the times when that city should not be active, it should be safe to assume that it comes from the citizens of the other city.

Experiment

The experiment would take 3 full months to complete. The independent variable in this case would be the exclusion of the toll bridge and the dependent variable would be the amount of daily rides in both of the cities from the time frames of 12:00 - 18:00 on weekdays for Gotham and 18:00-24:00 on weekdays for Metropolis from the 3 month before. Something that would have to controlled is the natural increase or decrease in ride activity that occurs between the current 3 month period and the previous 3 month period. For example, if the months we are testing are during winter, there might be a natural decrease in rides than in fall months due to the weather, and that has to be taken into account.

The null hypothesis being: There is no difference in mean daily rides in the periods before and during the tolls being removed. The alternative hypothesis would be that there would be a difference in the mean daily rides in the periods before and during the tolls being removed. I would then resample via bootstrapping the daily rides during those 60 or so days the experiment took place and find the mean. I would do the same for the control group. I would then find the difference between those two means, and I would do this 10,000 times. Eventually I would be able to calculate the 95% confidence interval, and if the lower 2.5% and higher 97.5% range of values don’t include the number 0, then I can reject the null hypothesis.

If the null hypothesis was rejected, I would recommend that the city operations team extend the reimbursement program indefinitely. If we fail to reject the null hypothesis, I would recommend the reimbursements stop.

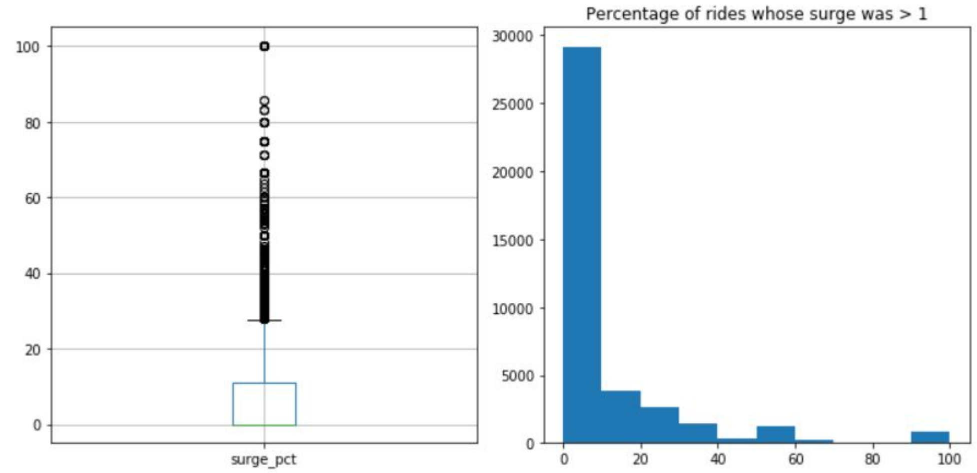
Part 3 Predictive Modeling

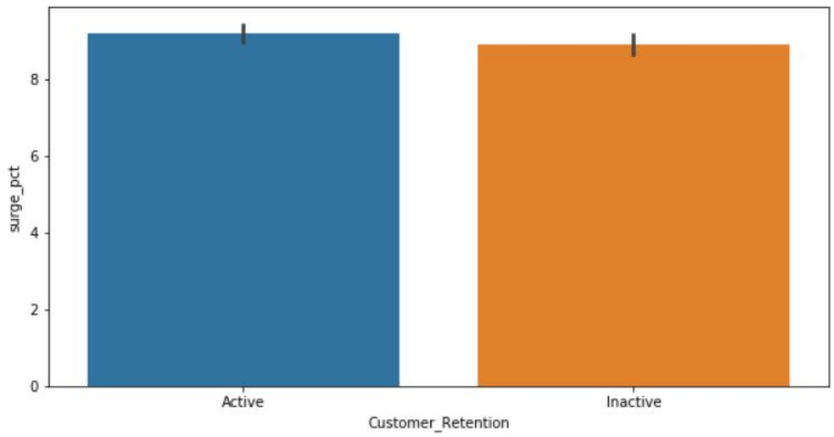
Cleaning

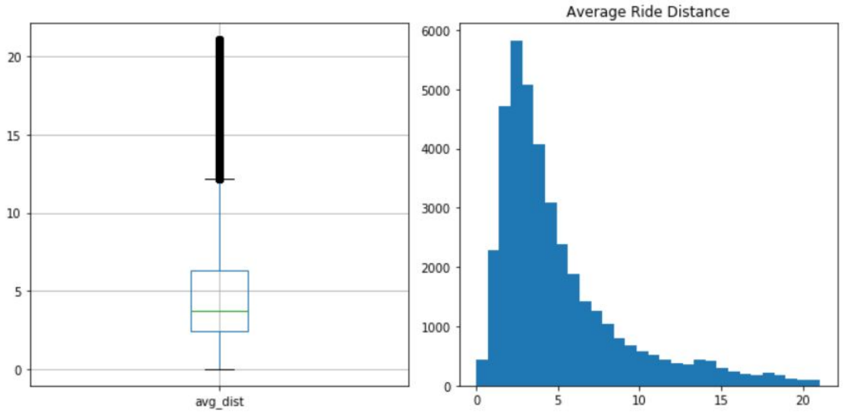
To get this data prepared, I removed all the observations that had missing values since it didn't really make sense to use replace them with other values such as the mean for the numerical values. Additionally, outliers were removed for ’number of trips in the first 30 days’, and ‘average distance’. Then I created a new column for the target variable called ‘Costumer Retention”. If there was activity by the account after June 1st, there would be a value listed as ‘Active’ . If not, it would show as ‘Inactive’.

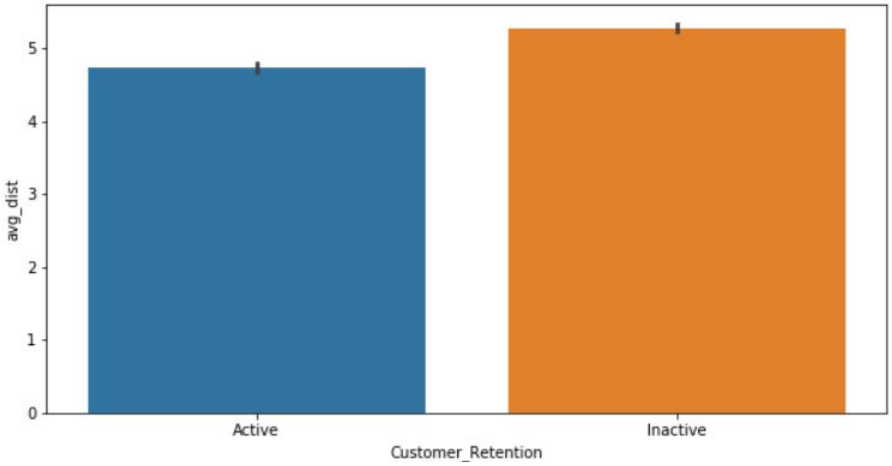
Exploratory Data Analysis

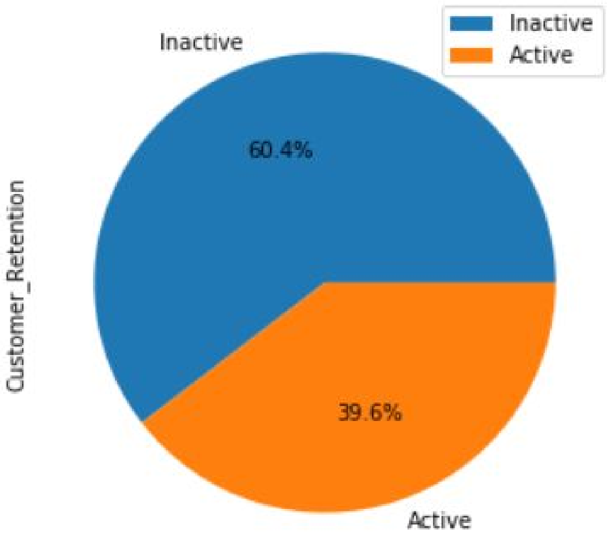
To check what the features look like, I went ahead and created some plots to look at the spread and to compare the means of the two categories.







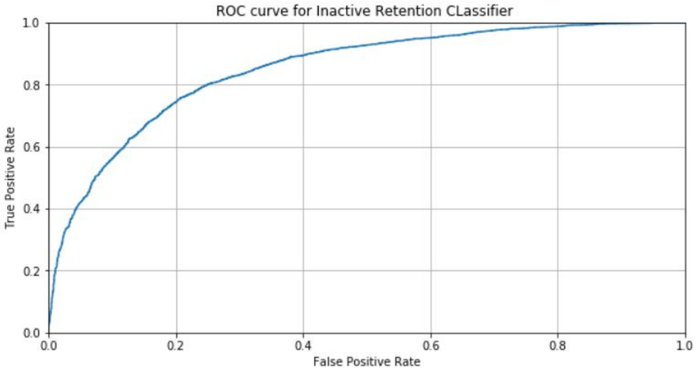


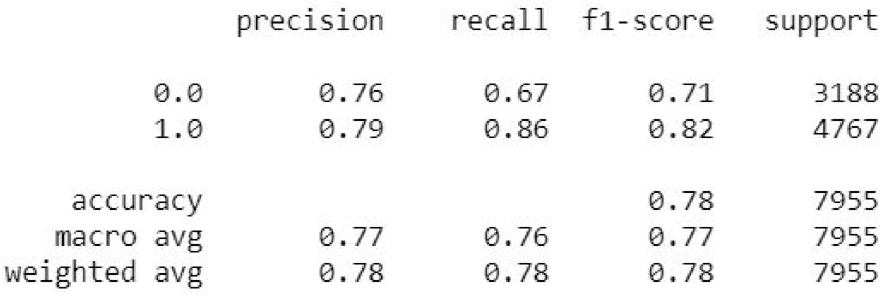


Machine Learning

Since the purpose of the model is to predict whether a customer will be classified as either one of two labels, this will be a classification machine learning model. Ensemble models work better therefore I will test Random Forests and Gradient Boost. While other classification models such as SVC can be applied to this situation, they usually don't perform as well as the aforementioned. I will be using trying a logistic regression model as a baseline. A problem That could present itself, is that there might not be enough data to build a model that won't overfit. About 25% of the data was removed in the cleaning process. Therefore, I will employ cross validation to not only tune the hyper-parameters, but to train the model in an optimal way.

Using the AUC score as the key metric for comparing scores, Gradient Boost performed the best, with 0.85.



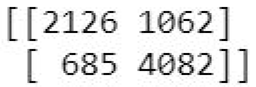


```
              precision    recall  f1-score   support

         0.0       0.76      0.67      0.71      3188
         1.0       0.79      0.86      0.82      4767

    accuracy                           0.78      7955
   macro avg       0.77      0.76      0.77      7955
weighted avg       0.78      0.78      0.78      7955
```

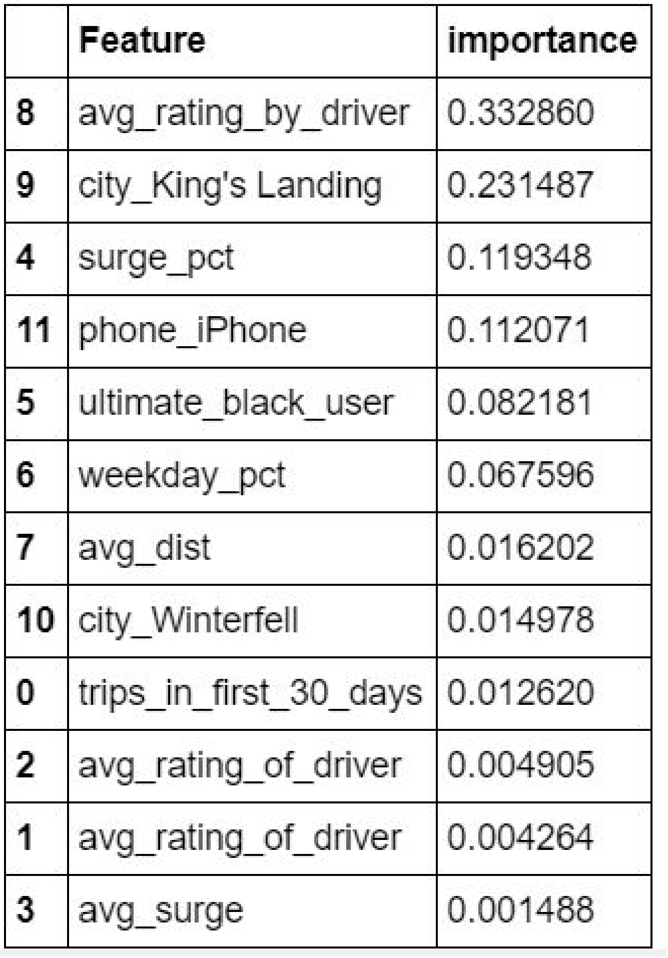
Confusion Matrix



```
[[2126 1062]
 [ 685 4082]]
```

Recommendations

When looking at the most important features that make up the predictive model, we can see what is most likely to classify someone as ‘active’ or ‘inactive’. The average rating by the driver can highlight a negative experience with a customer, therefore better training or communication on how to improve relations between the two can increase retention. Likewise, percentage of surge trips shows that the threshold for surge pricing could be set too low, and that by making it higher, more customers are likely to stay active. Finally, the type of phone someone is using could hint that there are problems with one of the versions of the app, and there needs to be an investigation to see if performance can be improved.



	Feature	importance
8	avg_rating_by_driver	0.332860
9	city_King's Landing	0.231487
4	surge_pct	0.119348
11	phone_iPhone	0.112071
5	ultimate_black_user	0.082181
6	weekday_pct	0.067596
7	avg_dist	0.016202
10	city_Winterfell	0.014978
0	trips_in_first_30_days	0.012620
2	avg_rating_of_driver	0.004905
1	avg_rating_of_driver	0.004264
3	avg_surge	0.001488