Data 1030 - Final Project Report
Michael Harder
Brown University
https://github.com/Michael-Harder/NYC_Citi_Bike_-Analysis
10/3/2019

Citi Bike Investigation

In an effort to not only reduce traffic, carbon emissions, and roadwear, but also improve public health in New York City; public officials launched a public bike share transportation alternative known as Citi Bike. Citi Bike became officially operational in 2013 and has been expanding ever since. Servicing Manhattan, Brooklyn, Queens, and Jersey City provides its share of logistical challenges. Applying machine learning techniques to publicly published Citi Bike data on their website, we can better understand their riders' habits.

NYC Open Data has published datasets containing Citi Bike station data as well as Citi Bike trip data; the latter representing 30 million Citi Bike trips from 2013 to present. These datasets have been made public via various avenues. A classification algorithm will allow us to identify how Citi Bike annual subscription riders differ from 24-hour pass or 3-day pass users. By successfully predicting whether or not a trip was conducted by a subscription rider or an everyday customer we can potentially find interesting insights into how their behaviors differ by investigating the model's parameters. For the purposes of this project I will be looking into Citi Bike trip data over the last two years, from August, 2017 to August, 2019. The features included in the data set are: trip duration (in seconds), start time and date, start station name, end station name, station ID, station latitude and longitude, bike ID, user type (Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member), gender (0 = unknown, 1 = male, 2 = female), and birth year. Because we want to classify if s trip was made by a customer or a subscriber, the target variable will be user type. Please note, the trip start time and end time columns were formatted as "yyyy-mm-dd hh:mm:ss.ssss" in order to work with this data I trimmed the string to get time and converted it into seconds from the start of the day. This process converted a string into a float and will allow it to be properly preprocessed later. Additionally, I dropped the following features: start station ID, end station ID, and bike ID. These were internal ID numbers that would not be useful for the purpose of this project.

EDA:

EDA revealed how unbalanced the dataset was. Just over 92% of the data set is made up of subscribers. This leaves a baseline accuracy for when the time comes to model. This also reveals interesting information that CitiBike trips are predominantly made by subscribers. Reference figure 1 below:

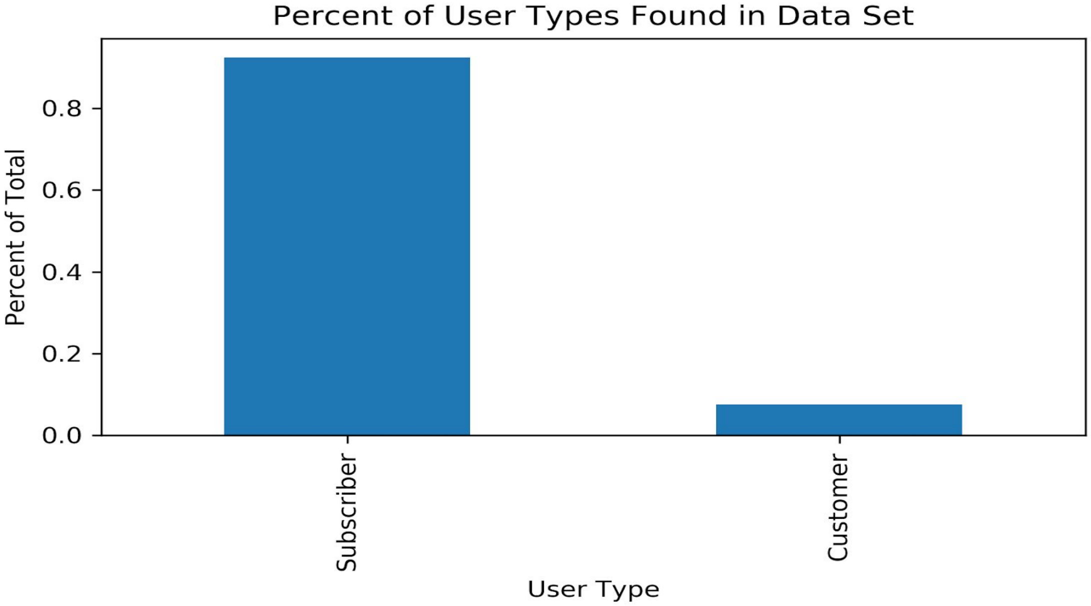

Figure 1: displays the percent of trips made in the data set by usertype

Additionally, it became clear that categorical features contained less information for everyday consumers than for subscribers. This makes intuitive sense as subscribers would likely need to fill out more personal information than everyday customers. An example of this is shown in figure 2 below:

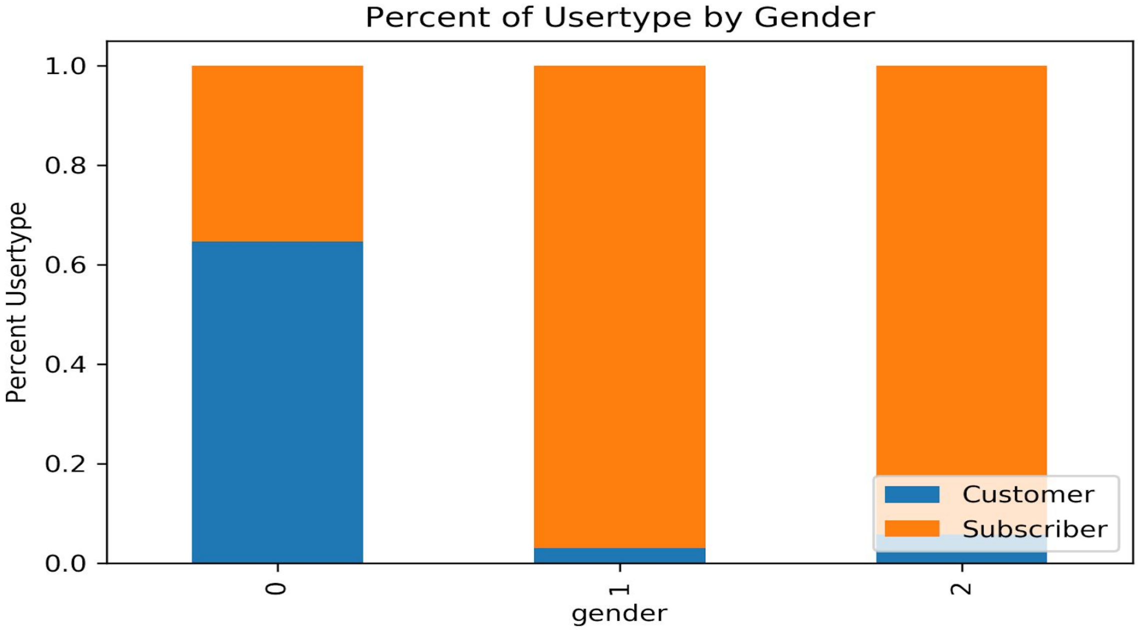

Figure 2: displays the percent of user types by gender (0 being unknown, 1 being male, 2 being female).

EDA also showed their could be a potential relationship between start trip station and trip duration. Start stations would affect the potential trip to be taken as stations are not evenly distributed across the city. Furthermore, users may inherently use stations differently; stations close to subway stops may be used for commuters while stations closer to tourist areas like Central Park may be used for leisure. This potential relationship is shown in figure 3 below:

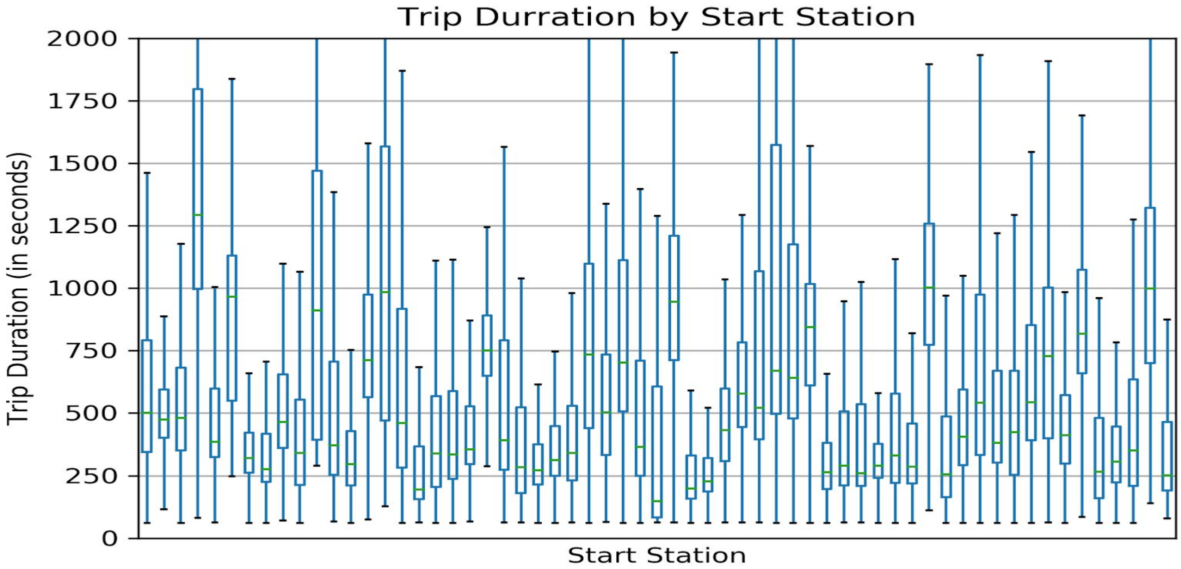

Figure 3: displays trip duration by start station. This is a busy chart but it shows how different start stations tend to have different trip durations.

Trip duration also had an interesting relationship with user type, our eventual target variable to classify. Subscribers tend to take shorter strips than customers; this is shown in figure 4:

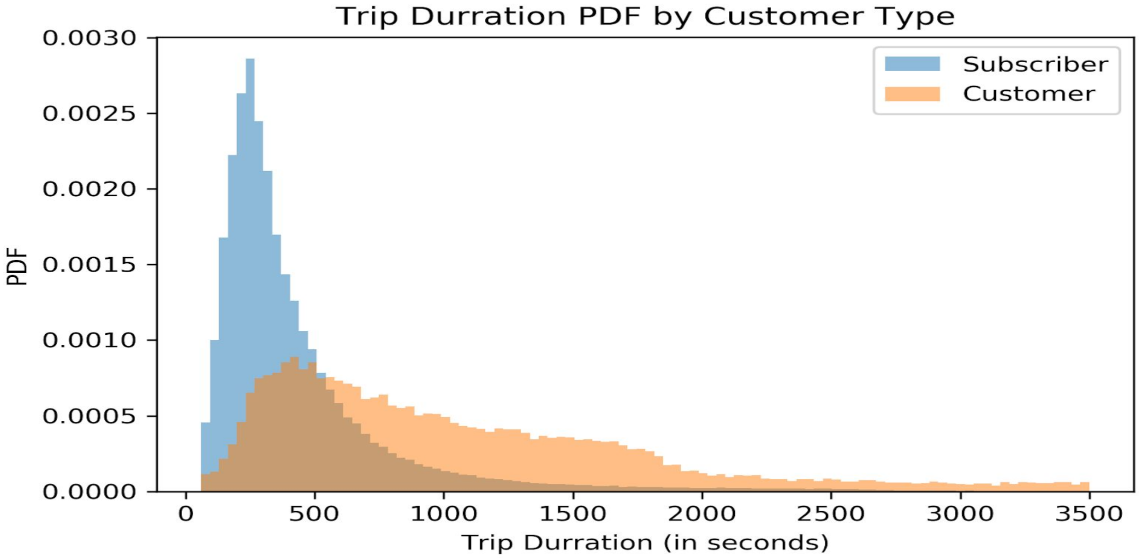

Figure 4: displays the PDF of tripdurration broken out by customer type.

Finally, through the EDA process it also became clear that subscribers and customers tend to start and end their trips at different times of the day. This relationship is shown below in figure 5:

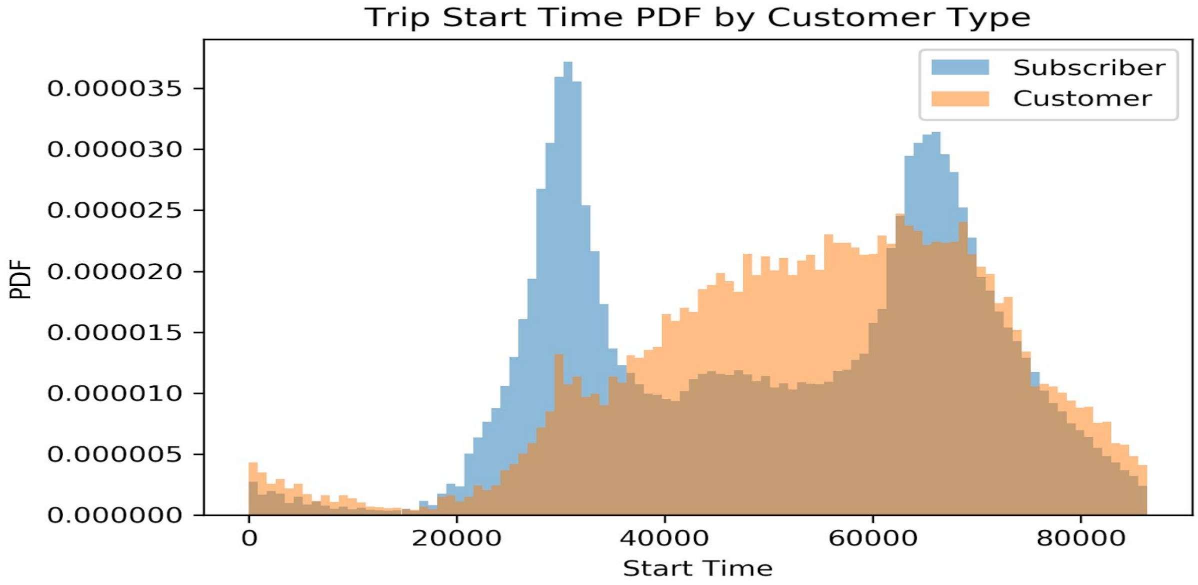

Figure 5: displays the PDF of trip start time in seconds from the start of the day (from 12:00am) by user type. Subscribers tend to conduct their trips in the morning and again in the evening, resembling commuter behavior.

Methods:

Before testing any models I had to address missing data. One feature contained missing data, birth year. Birth year accounts for 1% of data and during EDA there was no apparent relationship between birth year and user type. The following features were dropped birth year, start station latitude, end station latitude, start station longitude, end station longitude. Start and end station makes the others duplicative.

Stratified k-folds were used to test different models. Because this is a classification problem the following models were tested with stratified k-fold splits: logistic regression, SVC, and random forest. The parameter to tune for logistic regression was our ‘C’ value. C was set at [0.1, 1.0, 10, 100]. These values of C performed well, the boundaries were not bottoming out. The parameters to tune for SVC were our ‘C’ value and our gamma value. These were both set at [1.e-03, 1.e-02, 1.e-01, 1.e+00, 1.e+01, 1.e+02, 1.e+03, 1.e+04]. These parameter ranges also worked well as the boundaries were not bottoming out as well. Lastly, the parameters to tune for random forest were the min number of splits and the max depth. These were set to be the range between 2 and 25 counting by 5 and the range between 1 and 30 counting by 5, respectively. These parameters also seemed to do well in these ranges as the models were not bottoming out on the boundaries. The stratified k-folds function used for all models first split the data into training and test data. Then a pipeline is applied to preprocess the data and apply the appropriate ML algorithm. Then we set up our parameters, prepare a gridsearch, and then finally apply k-fold cross validation.

In order to measure uncertainty in splitting the models were run multiple times with different random seeds. Both random forest and logistic regression were run with 10 different random seeds and SVC was run with 5 different (due to the lack of computing power on my laptop). By applying these loops we mitigate the concern with uncertainties due to splitting and due to non-deterministic ML methods like random forest.

These models were then evaluated by the average accuracy score of the model across the multiple runs with random seeds in comparison to the baseline and with reference to the standard error provided again by the multiple runs with random seeds. After evaluating all three models, random forest provided the highest accuracy score and lowest standard error. This appeared to be the model of choice. However, before jumping to conclusions I wanted to test these models on a data frame with a 50/50 split of subscribers and customers as my original data was so unbalanced.

After creating a new data frame of a random sample without replacement of the original data containing an equal number of rows for both user types I underwent the same process and came to the same conclusion; random forest provided the highest accuracy score with the lowest standard error. In the original data frame test random forest had an accuracy score of over 95% and a standard error of roughly 0.003 with a base accuracy score of 92%. When testing on the balanced data frame the random forest had an accuracy score of over 83% and a standard error of roughly 0.013 with a base accuracy score of 50%.

Results:

- Random forest
  - Original data frame: test accuracy = 0.9515 +/- .0031
  - Balanced data frame: test accuracy = 0.8262 +/- 0.0131
- Logistic Regression
  - Original data frame: test accuracy = 0.9438 +/- 0.0025
  - Balanced data frame: test accuracy = 0.8240 +/- 0.0181
- SVC
  - Original data frame: test accuracy = 0.9412 +/- 0.0025
  - Balanced data frame: test accuracy = 0.83 +/- 0.02

The results of the models were close in many ways. In order to run my SVC I could only run 5 random seeds and I needed to reduce the amount of the data in the data frame to be evaluated by SVC to a 1% sample without replacement from the original data frame for 22,794 observations. The limitations of my computing power restricted the analysis that could be done with SVC. Therefore as SVC was not decerinably better from the other models I moved on from it first.

Deciding between random forest and logistic regression required a more precise calculation. By using the following formula:

(average model test accuracy - base accuracy) / standard error

Random forest was marginally better than logistic regression for both my original data frame and the balanced data frame. These scores are shown below (a larger number is preferable).

- Original data frame:
  - RF_score = 8.8834
  - LR_score = 8.2084
- Balanced:
  - RF_score = 24.8261
  - LR_score = 17.9212

After landing on random forest as my model of choice I then evaluated the random forest regressor with permutation feature importance. The results of this analysis are below in figure 6:

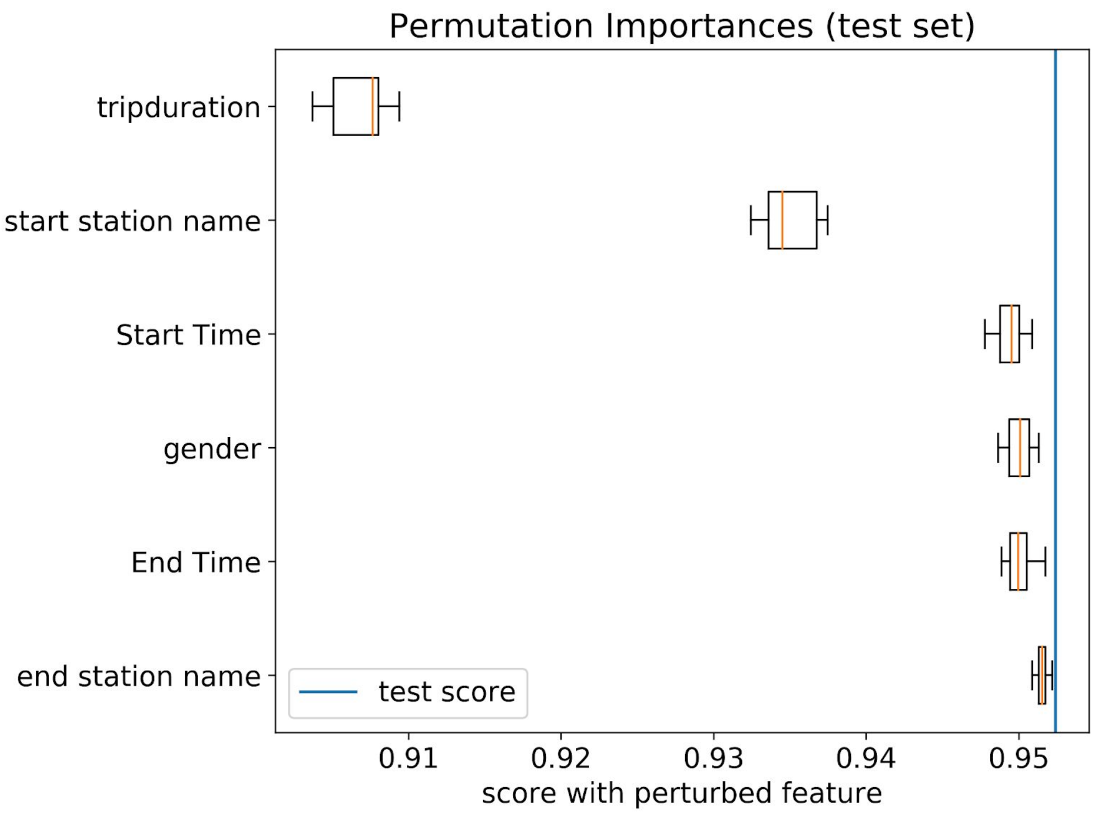

Figure 6: displays the permutation feature importance. Trip duration being the most important feature for the classification by a wide margin; followed by start station name. The remaining features have less of an impact on the classifier.

As a result we know that trip duration and start station name are the most important features in determining if a trip was made by a subscriber or an every day customer. This provides interesting insights into Citi Bike's customer habits. First looking at trip duration, our most important feature we can see that the two user types use Citi Bike differently. Subscribers appear to use the bikes more out of utility or convenience, with short trips. Next looking at start trip station (our second most important feature); it appears that subscribers start their trips at certain stations. In order for Citi Bike to retain these subscribers it would benefit them to make sure these stations have enough bikes stationed for their subscribers. It also raises questions about where subscribers potentially live and work in the city. It may interest Citibike to use this information to identify the neighborhoods their subscribers frequent vs. where their customers frequent and try to find differences in these areas of the city. There are many different directions to go from their; if you can identify neighborhoods where subscribers use Citi Bike you can potentially find look alike neighborhoods where Citi Bike should have more subscribers than they do and ultimately grow their customer base.

Outlook:

This modeling approach could be improved in the following ways. First, I could not properly apply the MCAR test for my missing data. There could be a more rigorous way to handle the feature with missing values (birth year). I tried to leave the missing values in and treat them as another category in the one-hot-encoder; however, I could not get it to run with the missing values. This would be an improvement on my model.

Additionally, with access to more computing power SVC could have been tested and properly compared to the other models. More computing power would have also allowed for a larger random sample to be used with replacement from the original dataset; this would have benefited all models. The data frame tested for random forest and logistic regression contained 22,794 observations (just 3% of my data set). The balanced data frame only contained 1,746 observations.

Smaller considerations to improve include the following. Experiment with other models like XGBoost. Citi Bike publicly updates this data set monthly; in some future state this model could be deployed each month on the updated data instead of my static data set I extracted.

References:

Motivate International, Inc. “Citi Bike System Data.” *Citi Bike NYC*, Citi Bike, https://www.citibikenyc.com/system-data.

John, Alexa St. “Citi Bike Expanding to the Bronx Six Years After New York Launch.” *The Wall Street Journal*, Dow Jones & Company, 16 July 2019, https://www.wsj.com/articles/citi-bike-expanding-to-the-bronx-six-years-after-new-york-launch-11563309799.

Roundedup. “Of Bikes and Cabs - NYC.” *Kaggle*, Kaggle, 15 Aug. 2017, https://www.kaggle.com/roundedup/of-bikes-and-cabs-nyc.