Jared R. Lieberman
MA Department of Revenue
Data Scientist Coding Challenge
6 January 2018

Project Description

Using the NYC public taxi data, the project intends to analyze data from the three different types of rides (yellow taxis, green taxis, and for-hire rides) and predict which borough a ride will end in. I used data from January 2018 for each ride type. Because there were millions of rides, I took a randomized subset of each (20,000 rides). Thus, the data alone provide an insight into an independent analysis of each, rather than a comparison between the three ride types. However, the models effectively represent the original data.

Exploratory Analysis

I began by analyzing how many trips start and end in each borough. I matched each pick-up and drop-off location ID to the borough zoning data also provided on the website. Thus, I had which boroughs each ride began and finished in. The results are shown in the table below, which was created in exploratory_analysis.py. For yellow taxis, the greatest number of rides start and end in Manhattan. For green taxis, the greatest number of rides start in Brooklyn and end in Manhattan. For FHV's, the greatest number of rides start and end in Manhattan. Besides New Airport, the least number of rides for each type of vehicle start and end in Staten Island. As noted, because I am using a constant, rather than proportional, subset of the data, the ride type counts can only be compared to itself, rather than to other ride types.

Further steps to take in the exploratory analysis could be looking at how long each ride takes depending on where it started and where it ended. For example, trips to EWR (Newark Airport) probably have a longer duration.

	EWR	Queens	Bronx	Manhattan	Brooklyn	Staten Island
Yellow Taxi	3 | 36	1196 | 918	26 | 130	18167 | 17744	251 | 817	0 | 3
Green Taxi	0 | 3	5910 | 5831	856 | 1215	6527 | 7176	6674 | 5716	4 | 3
For-Hire Vehicle	9 | 104	3193 | 3254	1789 | 1746	9576 | 8799	5209 | 5216	159 | 168

Summary of Findings

I chose to describe through the project the patterns of each of the three ride types. Thus, I although the task asked for one model, I created three total for each ride type. I believe this is the best way of achieving the goal of creating policy-relevant predictions on where NYC taxis and for-hire rides end. I chose to use Random Forest algorithm for the data because I wanted a supervised approach to the classification problem. The entirety of creating the models can be found in random_forest_classifier.py.

Each of the datasets go through preprocessing in order to remove null data, standardize the time data, set borough names to numbers. They are represented as: ('EWR' : 1, 'Queens' : 2, 'Bronx' : 3, 'Manhattan' : 4, 'Brooklyn' : 5, 'Staten Island' : 6).

The process of creating Random Forest models are then performed. The features and labels, pick-up borough, are first separated. The features for yellow and green taxis are pick-up time, passenger count, and which borough a ride starts in. The features for for-hire rides are the car's dispatching base number, pick-up time, and which borough a ride starts in. Next, the training and test data are split. The training data is 70% of the data, and consequently 30% is test data. Next, grid search is used to find the best parameters on n_estimators and max_features. The parameter n_estimators represents the number of trees in the forest. The parameter max_features represents the size of the random subsets of features to consider when splitting a node.

A classifier is then created and fitted to the training data. Next, the accuracy is taken on the testing data. Each model performs extremely well on the training data (between 96% and 99% accuracy). The is more discrepancy on the testing data. The yellow taxi model performs at about 85% accuracy, the green taxi model performs at about 80% accuracy, and the for-hire rides perform at about 65-70% accuracy. I believe the accuracies would increase using the entirety of the data, rather than a subset of 20,000 of each dataset.