Determining What Impacts the Classification Between Morning and Evening Taxi Rides in NYC

Maira Asmat

Abstract

The goal of this project was to classify NYC taxi rides as occurring either during the morning rush hour or the evening rush hour based on a dataset obtained from the NYC taxi and limousine commission. I decided on using a random forest model as it had the highest accuracy of the models I tested, and interpreted the feature importance from that model.

Design

I recently read a NYT article that discussed how taxi activity in NYC declines from 4 to 5 pm as the taxi drivers change shifts at that time, and that the mayor was looking into perhaps changing that time because 4 – 5pm is when many people are getting out of work. Thus, I wanted to determine the difference in how the features impacted the morning rush hour from 6 – 10 versus the evening rush out from 4 – 8.

Data

The main dataset used was from February 2019, chosen because February tends to be a month where there are less tourists that could impact the data. I obtained it from the official NYC government taxi and limousine commission, and it initially included over 4 million rows. 100k of those were randomly sampled and used in modeling. The features included passenger count, trip distance, pickup location ID, dropoff location ID, payment type (dummy), fare amount, extra charges, mta tax, tip, toll, total amount paid, congestion surcharge, and datetimes of dropoff and pickup.

Algorithms

The data was baselined using a logistic regression model, and based on the business use case, a decision tree and random forest model were also used. All three models were used initially on the raw data, then their parameters were tuned using GridSearchCV. The data was validated and scaled before being used in the tuned models. On the tuned logistic regression, the final 5-fold CV scores on 20 chosen candidates were:

```
Accuracy: 0.8175
 Precision: 0.9806
 Recall: 0.7153
 F1: 0.8272
 F-beta: 0.7562
 ROC-AUC: 0.8466
```

On the tuned random forest, the final 5-fold CV scores on 200 chosen candidates were:

```
Accuracy: 0.8274
 Precision: 0.9645
 Recall: 0.7446
 F1: 0.8404
 F-beta: 0.7802
 ROC-AUC: 0.8508
```

Tools

- Exploratory data analysis in pandas and numpy
- Modelling in sklearn
- Visualization in Matplotlib

Communication

All visuals are embedded within the slides, which are attached in the GitHub repository and will be presented.