ORIE 4741 - Midterm Report

Audrey Yap, Sung Mo Kim

amy34, sk2346

1. Problem Statement

Taking a ride in an iconic yellow taxi is one of the most popular ways to get around New York City. However, this ride has its costs with time and money. Thousands of taxi trips occur in NYC each hour of the day all year round, even in the midst of the COVID-19 pandemic. Insights derived from the rides could help the large number of taxi customers figure out the best times to take a taxi and how much they expect to spend in both time and money.

Our project goal is to make accurate predictions regarding the time of travel and the cost of the travel based on basic information of a taxi trip such as the pick-up location, drop-off location, time of day, day of the week, and month. As a result of the prediction, a person can then decide whether it is worth it to them to take the taxi or a different mode of transportation.

2. Data

2.1 Dataset description

For our project, we are getting our data from NYC TLC Trip Record Data. The website has a dataset for each month from January 2009 to July 2021. The 18 features included in each of the datasets are the vendor that provided the record, number of passengers in the taxi, travel distance, pick-up and drop-off details such as location, date, and time, and charges made for the trip such as the fare amount, tip, and tolls. We made our dataset such that it only included trips during 2020.

2.2 Data cleaning

Before some necessary data cleaning, we realized it would be easier for us to have access to the time of travel in float figure as it would be an essential feature in our testing. We found the difference between the dropoff time and the pickup time (in the form of MM/DD/YYYY) and multiplied this by a factor that gave us precise numbers on the time of travel.

In each month, we dropped rows that had unknown boroughs as the pick-up or drop-off location since it indicated that the taxi went outside of the NYC area. The pick-up and drop-off locations are given as numerical ids in the original datasets, so we found the latitude and longitude of the locations based on the taxi zone lookup file from the website.

To make a collective dataset for the year of 2020, we randomly selected 10,000 rows from each month's dataset. We also noticed that there were points with negative numbers for the total amount of payment and promptly removed them from our dataset. We filled those with more randomly selected data points from the monthly report from the NYC TLC Trip Record Data.

Lastly, we split the data so that 80% was used as our train set and the 20% was used as our test set. Our features of the train set and the test set were the price variables such as the total amount, fare amount, tip amount, and the toll amount. Then, the data was shuffled randomly.

2.3 Data Analysis

(a) Correlation Heatmap

We first made a correlation heatmap(below) for the dataset's features to observe what affected the price of the payment the most. From the heatmap, we noticed the total amount was highly affected by the fare amount since the two will not vary much from each other unless there is a high amount of tolls passed or tips given. Another feature that stood out was the distance of the travel (r = 0.8362), which could be intuitively suggested. One surprising factor we observed was that the correlation between the time of travel and the price was lower than expected (r = 0.1356). The time of travel and the distance traveled also showed a low correlation (r = 0.1422). This led us to conclude that the volume of traffic in New York City was very high since these numbers were deviating from some basic intuitions that may be held.

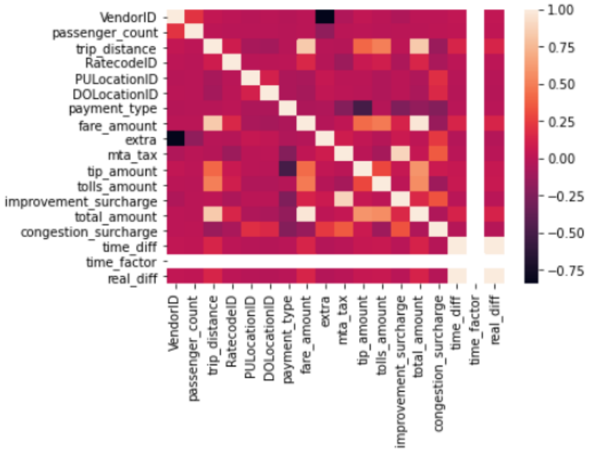
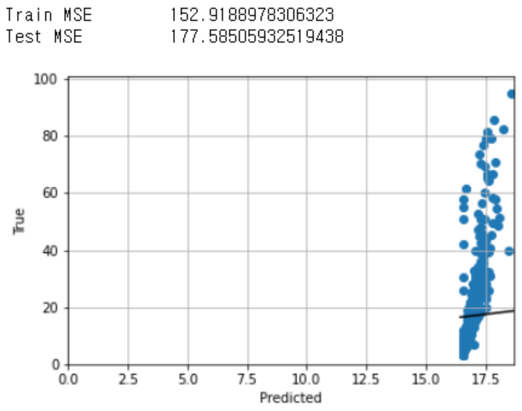


Figure 1. Correlation heatmap

(b) MSE and OLS

To confirm that our correlation figures were accurate, we decided to fit a linear model using the OLS function on two of the features mentioned above and calculated the mean squared error (MSE) for the train set and the test set.

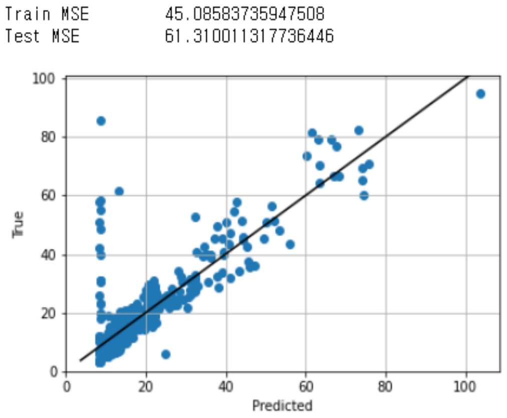
First, to check that the counterintuitive correlation factor was accurate, we fit a linear model on the time of travel with an offset term. Below is the result we obtained:



```
Train MSE        152.9188978306323
Test MSE         177.58505932519438
```

Figure 2. Linear fit for time of travel

The visualization is clearly representative of the expectations we had from the correlation heatmap we observed above. From Figure 2, we notice a large MSE for both the train MSE and test MSE and can observe that the predicted linear regression fits poorly. Once we had a linear fit model including the distance of travel (Figure 3), we got a much more desired result. There was clear correlation between the distance traveled and the price with some outliers observed at distance near zero.



```
Train MSE        45.08583735947508
Test MSE         61.310011317736446
```

Figure 3. Linear fit for distance of travel

3. Moving Forward

We want to do some more data visualizations specifically using a heatmap over the NYC area for both pick-up and drop-off locations to provide a better visual representation of which spots in NYC are most likely to have taxis around. We also want to look into some of the other features in the dataset to see if they can improve the models we use.

Additionally, we will test and compare other models to see if we can find a model that fits the data well. We will be comparing the models by comparing cross validation values as well as observing the different models' train and test MSEs.