NYC Taxis

Audrey Yap and 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. The taxis have been used in movies featuring NYC, and most travellers associate these taxis with the city. So, it comes to no surprise to know that thousands of taxi trips occur in NYC each hour of the day all year round. Even though Uber and Lyft have become more prominent in the transportation field and the COVID-19 pandemic causing less people to leave their homes, there were still thousands of trips recorded. 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. We decided to include trips from that time since we expect the price has been affected by the pandemic. In order to provide a report that is relevant, we chose 2020 as that is when the prices would be changing the most. We also needed to show what is happening over a full calendar year.

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 drop-off 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. While looking through the datasets, we also noticed that there were multiple points that contained negative numbers for the total amount of payment and removed them from the dataset.

To make a collective dataset for the year of 2020, we randomly selected 10,000 rows from each month's dataset. 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.

3. EDA

3.1 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

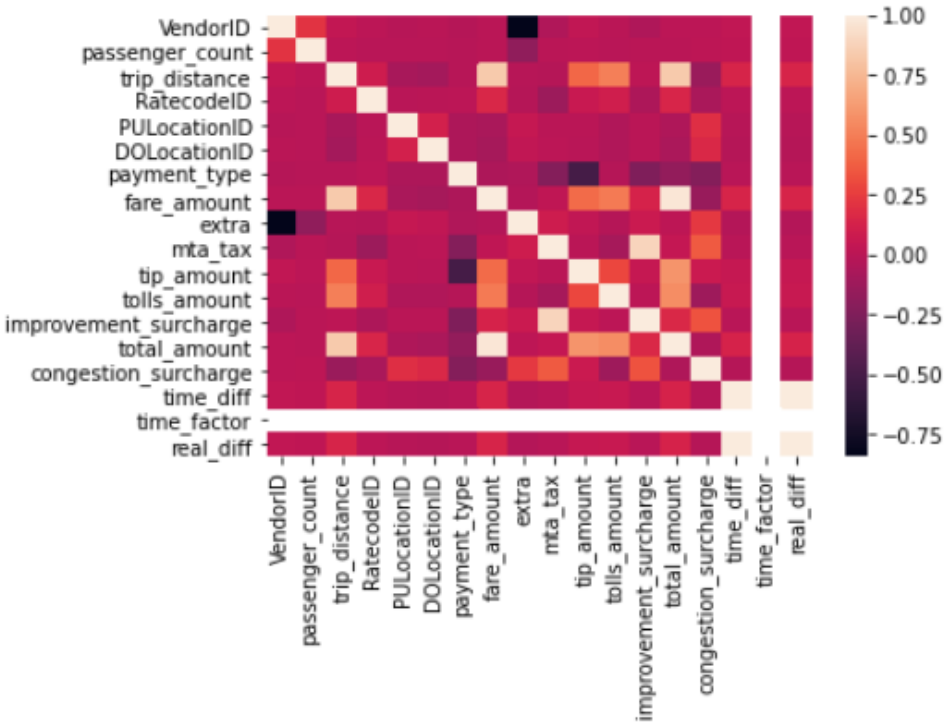

Figure 1. Correlation heatmap

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.

3.2 Months and price

Figure 2 shows the average price for a taxi ride in each month of 2020 from our cleaned data set. We can see a massive drop from March's average to April's average due in part to the United States' lockdown for the COVID-19 pandemic. During this time, many were laid off of their jobs or forced to work remotely from their homes. So, the prices of taxi rides must have been lower in order to get the demand. Since the months all have different average prices, we will further investigate how useful having the months in the model is.

Figure 2. Average Monthly Price Line graph

4. Models & Other Techniques

4.1 Feature Engineering

The month for each ride entry was taken from the pickup time. The values from that process gave the integer values of the month. We interpreted the months as nominal values rather than integers. So, we used one-hot encoding since it is either that month or not that month.

We also made a boolean encoding for the stages seen in Figure 2. So, we have one group for January, February, and March. The second group is April and May, and the last group is the remaining months. We chose to do those groupings since the months in each group have similar prices.

Another boolean encoding we made was for time. There is an extra surcharge for those taking a taxi between 4 p.m. to 8 p.m., so we grouped them together. And all the other times are in a separate group.

4.2 MSE and OLS

In our previous report, we did 2 least squares regressions that had time and distance as factors for the price (Figure 3 and 4). They were done separately to see the relationship between the variables and the price. We found that the time spent on the trip did not correlate with the price of the trip. To see how having the months included affected the MSE and OLS functions, we included the distance as it plays the biggest role in how the price of a ride is determined.

```
Train MSE       153.58131635309158
Test MSE        151.00757557489453
```

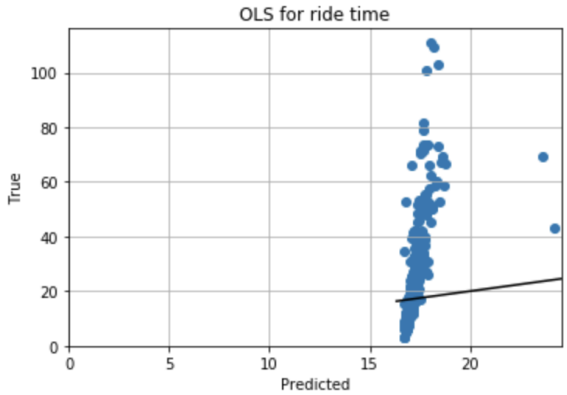

Figure 3. OLS for ride time

```
Train MSE       45.459540317770305
Test MSE        32.30076182631869
```

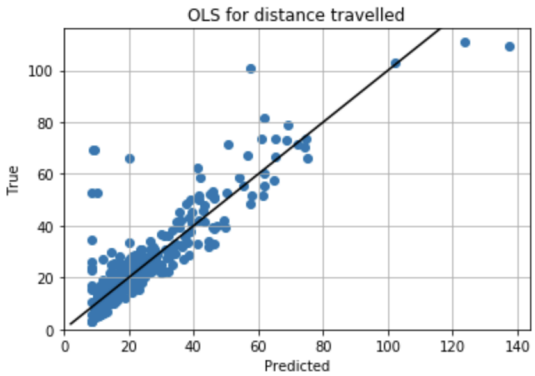

Figure 4. OLS for distance travelled

```
Train MSE       45.023794917346876
Test MSE        31.86228748990579
```

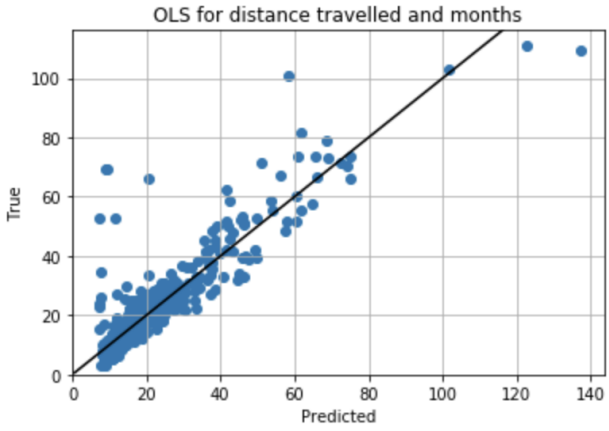

Figure 5. OLS for distance travelled and months

Figure 5 shows that including the months in the model does improve it. We see that both the training set MSE and the testing set MSE from Figure 5 are both lower than the MSE's presented in Figure 4. There is not a very noticeable difference between the OLE graphs in the two figures which may be because there has not been a giant change in prices for rides. As observed before, there is only around a $3 difference from March into April.

When we use the grouped months, we would see that the train and test MSE are very slightly larger than that of using the one-hot encoded months. However, the difference is negligible.

Accounting for the surcharge during 4 p.m. to 8 p.m. seemed to be somewhat useful as it shows a decrease in the train MSE and test MSE from Figure 5 to Figure 6. The changes we see in adding more variables to our model are small due in part to the weak correlation between the total amount paid and all the variables. We suspect this might be because of the tip amount which is varied as it depends on the customer.

```
Train MSE        44.94264283968819
Test MSE         31.750376611033985
```

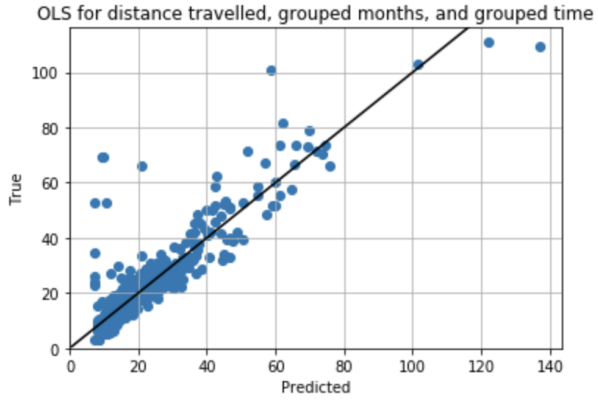

Figure 6

We chose to only do linear models because there are flat rates given for the distance travelled. So, it does not make sense to use an exponential or a quadratic model. We account for the change in price in months by adding in either the one-hot encoded or boolean encoded (grouped) features.

5. Fairness

Fairness is not an important criterion to consider when choosing a model for this project. This project is mainly focused on yellow taxis trip times and prices during 2020. There are flat rates given for prices, so the prices vary on the distance traveled rather than the demographics of the rider. It is also in the best interest of the taxi drivers to have the pickup and drop-off times be close together so that they can do more rides within the day.

6. Conclusion

We can draw from our results above that the information offered by the TLC is accurate. It could be told from the correlation charts that the amount of fare is related most closely to the distance, not the time of travel. The data cleaning was also very accurate in that it clearly reflected the number of travels appropriately during the time of the pandemic. The MSE and the OLE charts agree with the results and the conclusions. Overall, the initial prediction we made on the project was correct and was precisely described by our visualizations.

For further insights, we wanted to work on a heatmap to see where in NYC the pickup locations were. We also had planned to do cross validation, but because of time constraints, health constraints, and lack of a third partner, we could not finish these tasks in time.