Impact of Traffic Congestion on Taxi Trip Durations and Fare in Different Areas of New York City

Authors:
JaladurgamNavya,jnavya@kent.edu,
Mukthasree Vengoti, mvengoti@kent.edu,
Aravind Reddy Thummala, athumma3@kent.edu

Date: 5/08/2024

MS Data Science, Kent State University

Title:

Impact of Traffic Congestion on Taxi Trip Durations and Fares in Different Areas of New York City.

Description of the research question:

How does traffic congestion affect taxi trip durations and fares in various neighborhoods of New York City, and what are the spatial and temporal variations in congestion levels across different times of day and locations?

Discussion about why the solution to the problem would be valuable:

Better Transportation Planning: Understanding how traffic jams affect taxi rides can help city planners and leaders make better transportation systems. This could mean fewer traffic backups and smoother trips for everyone.

Improved Taxi Services: Taxi companies and drivers can use this information to plan better routes, avoid congestion and provide faster and more reliable service for passengers.

Helps Passengers:
Passengers can make smarter decisions about when they can take taxi rides, to avoid traffic delays and save money and time.
.

Beneficial for Urban Living: People who are living and working in busy areas can better understand how traffic affects their daily lives. This information can assist them in making decisions to better their lives and their businesses efficiency.

Overall, answering the research question we can understand the impact of traffic congestion on taxi trips can lead to better transportation, happier passengers.

Discussion about previous attempts to solve the problem and what you learned from them:

We used a Linear regression model to train the data. When we used this model, found that the actual values of total fare and predicted values of total fare which has high difference values between the those two after predicting with the test data.

Decision Tree Regression:

The decision tree performs well when compared to linear regression. It has less difference between the actual and predicted total fares. But, after using the decision tree there are a few difference between the actual and predicted total fares.

We learned that one can utilize Linear Regression when features exhibit a linear relationship with the dependent variable, but this may not be the case for most situations occurring. On the other hand, Decision Tree Regression belongs to non-linear models and allows one to capture more intricate dependencies between features and responses. Decision trees can capture non-linear patterns in the data, which might explain why you observed less difference between the actual and predicted total fares compared to linear regression.

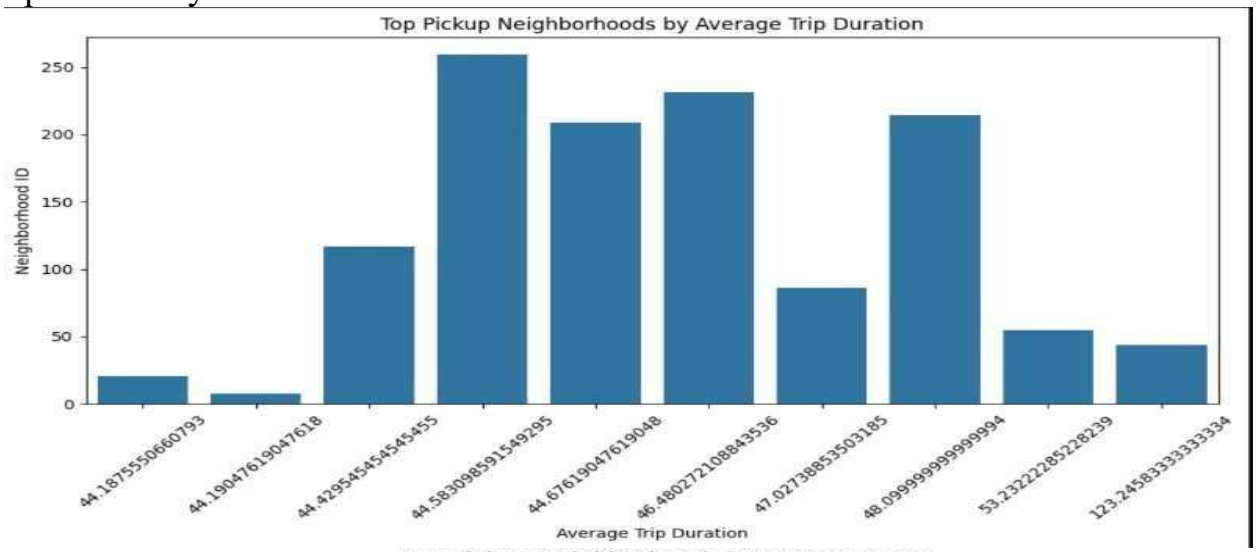
Discussion about the methods your team used to solve the problem:

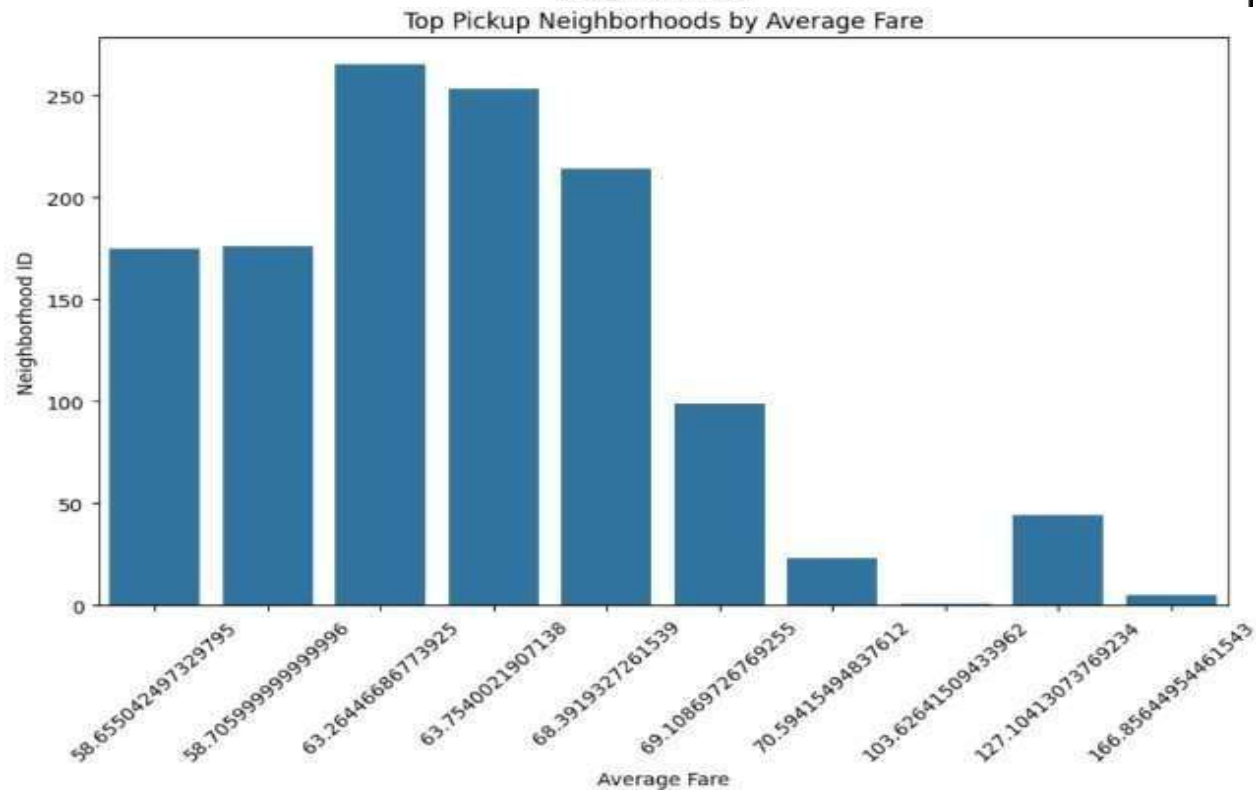
Even after using decision tree regression, we still have a difference between the predicted and actual values. So, we used another modeling technique which is called Random Forest. In this model the difference between the actual and predicted total fare values are less when compared to linear and decision tree.

Discussion of the results:

Initial statistical descriptive statistics:

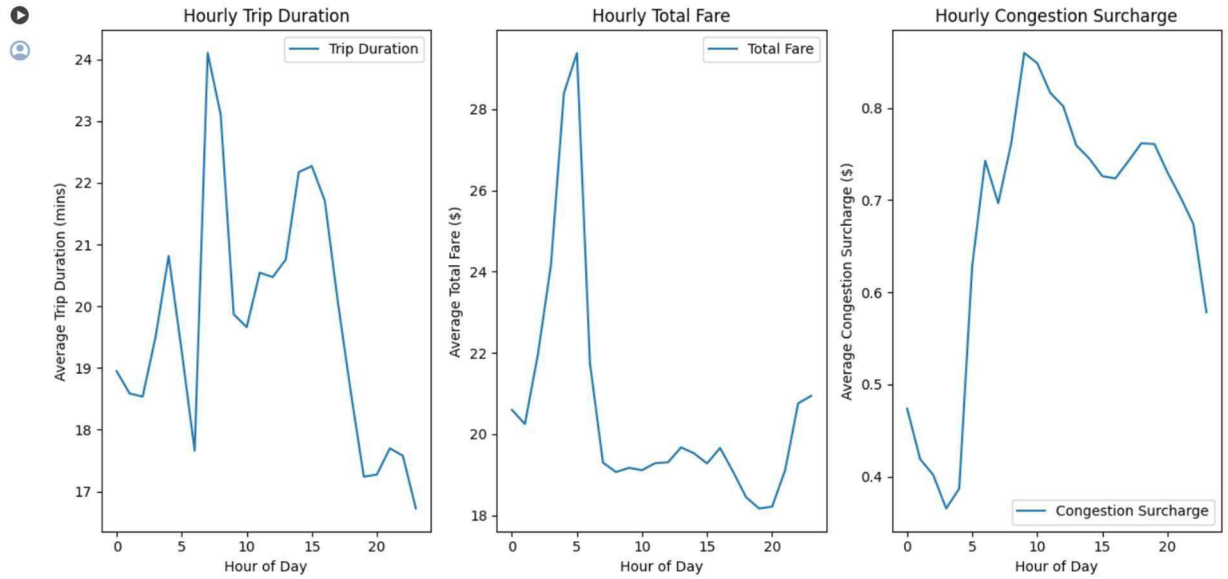
Spatial Analysis**:**





This graph highlights the neighborhoods that tend to have longer ride durations on average, which could be influenced by factors such as their distance from popular destinations, traffic patterns. The neighborhoods where ride fares tend to be higher on average.

Temporal Analysis:



Hourly Trip Duration:

The first subplot displays the average trip duration in minutes for each hour of the day.

It helps in understanding how the duration of taxi trips changes throughout the day, indicating potential peak hours of travel.

Hourly Total Fare:

The second subplot shows the average total fare in dollars for each hour of the day. This shows how fares fluctuate based on the time of day.

Hourly Congestion Surcharge:

The third subplot illustrates the average congestion surcharge in dollars for each hour of the day. It shows the impact of congestion on taxi fares, showcasing when congestion surcharges are typically higher or lower.

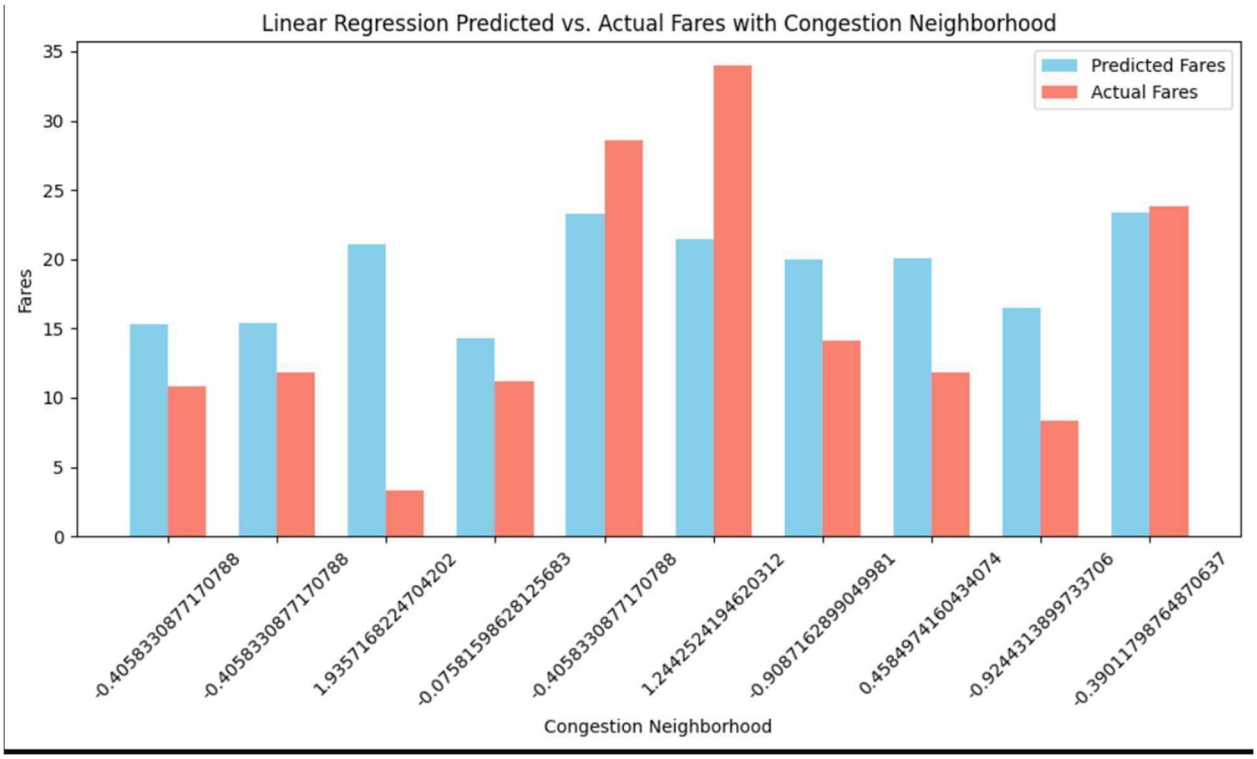
Correlation with congestion surcharge:

Trip duration and trip distance have very little to no relationship with congestion surcharge. Total fare has a moderate positive relationship with congestion surcharge, suggesting that higher fares may be associated with slightly higher congestion surcharges, but the correlation is not very strong.

```
Correlation matrix:
                     trip_duration  trip_distance  total_fare  \
trip_duration             1.000000       0.003197    0.155800
trip_distance             0.003197       1.000000    0.013947
total_fare                0.155800       0.013947    1.000000
congestion_surcharge      0.023612       0.000196    0.165170

                     congestion_surcharge
trip_duration                    0.023612
trip_distance                    0.000196
total_fare                       0.165170
congestion_surcharge             1.000000
```

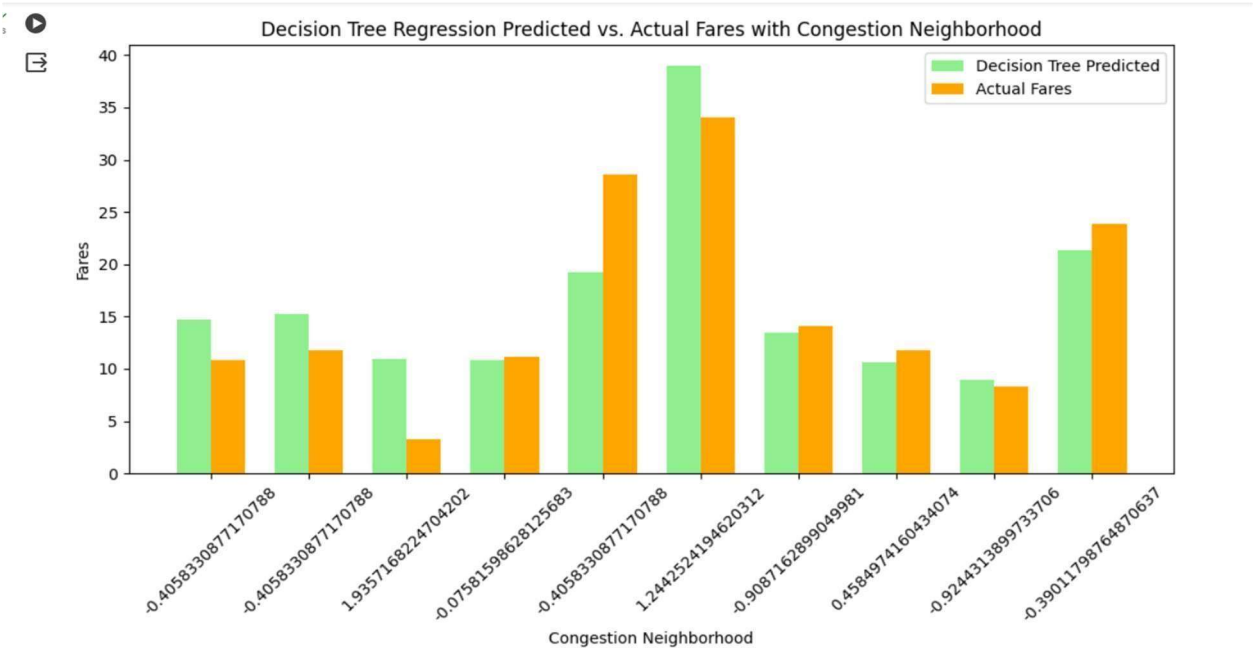
Performing initial modeling Visualizations:
Linear Regression modeling result:



In x-axis are congestion neighborhood locations these locations are scaled by using the scaling method and Y- axis has fares.

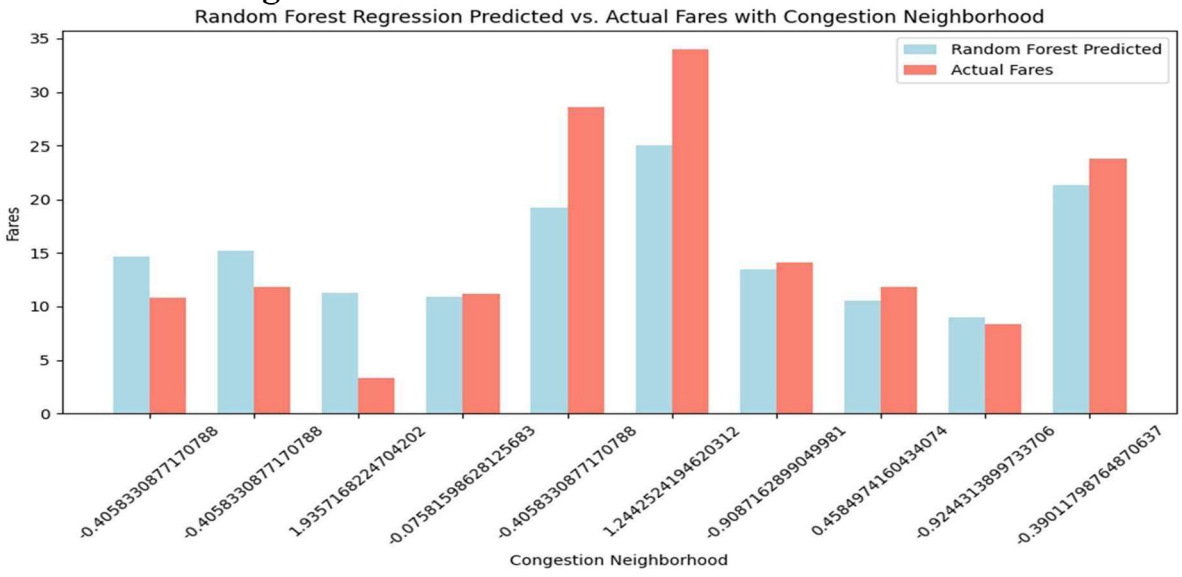
The above plot values predicted by the linear regression model. The blue bars are predicted values and other bars are actual fares. By seeing this plot on these locations, the fares will increase due to congestion. These are the top ten locations.

Decision Tress Modeling Result:



This plot is about decision tree models predicted the fares. The green bars are predicted decision tree values, and remaining orange bars are actual fares.

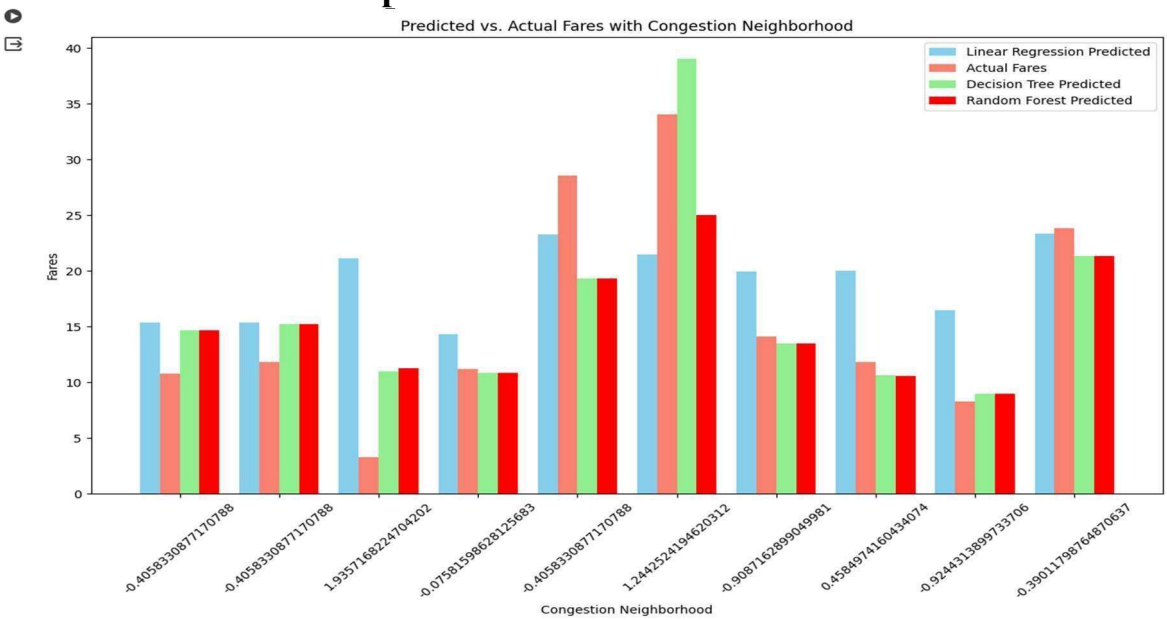
Random Forest Regression Results:



The plot shows the results visually, each bar in the graph represents a specific Congestion Neighborhood value. Where X-axis represents congestion neighborhood and Y-axis represents the fares. The red bars indicate the actual fares, and the blue bars indicate the predicted fares by random forest model.
Our results display the top 10 prediction observations made by random forest for taxi fares.

By comparing the heights of blue and red bars for each Congestion neighborhood value, we can see how good the Random Forest Regression model was at guessing the real fares. If the blue and red bars are about the same height, it means the model's predictions were accurate for that Congestion Neighborhood value. However, if there is a significant difference between the blue and red bar heights, it shows that model's predictions deviated from the actual fares for that specific Congestion Neighborhood value.

Overall results of the experiment:



The plot shows the overall results of our modeling which shows the comparison of all the three modeling techniques we used i.e, Linear regression, Decision Tree, Random Forest.

The plot represents x-axis as Congestion Neighborhood, y-axis as Taxi Fares, the blue bar indicates Linear Regression predicted fare values, the orange bar indicates actual fares, The green bar indicates Decision Tree Regression predicted fare values and finally, red bars indicate Random Forest Regression predicted fare values.

In general, the lowest MSE indicates a better accurate model. From our observations, the Random Forest model is better compared to other models.

Discussion about how successfully solved the problem:
Model quality Metrics:

1. **Linear Regression Metrics:**

```
Linear Regression Metrics:
Mean Squared Error: 260.97
R-squared: 0.07
Mean Absolute Error: 9.18
```

```
Linear Regression Predicted vs Actual Fares with Congestion Neighborhood:
Congestion Neighborhood: -0.4058330877170788, Predicted: 15.33, Actual: 10.80
Congestion Neighborhood: -0.4058330877170788, Predicted: 15.36, Actual: 11.80
Congestion Neighborhood: 1.9357168224704202, Predicted: 21.10, Actual: 3.30
Congestion Neighborhood: -0.07581598628125683, Predicted: 14.30, Actual: 11.16
Congestion Neighborhood: -0.4058330877170788, Predicted: 23.30, Actual: 28.55
Congestion Neighborhood: 1.2442524194620312, Predicted: 21.44, Actual: 34.00
Congestion Neighborhood: -0.9087162899049981, Predicted: 19.94, Actual: 14.10
Congestion Neighborhood: 0.4584974160434074, Predicted: 20.03, Actual: 11.80
Congestion Neighborhood: -0.9244313899733706, Predicted: 16.49, Actual: 8.30
Congestion Neighborhood: -0.39011798764870637, Predicted: 23.36, Actual: 23.81
```

MSE: Higher values indicate poorer performance.
R- squared: Indicating a weak level of explanatory power.
MAE: Higher values indicate lower accuracy.

2. **Decision Tree Regression Metrics:**

```
Decision Tree Regression Metrics:
Mean Squared Error: 190.72
R-squared: 0.32
Mean Absolute Error: 5.32
```

```
Decision Tree Regression Predicted vs Actual Fares with Congestion Neighborhood:
Congestion Neighborhood: -0.4058330877170788, Predicted: 14.67, Actual: 10.80
Congestion Neighborhood: -0.4058330877170788, Predicted: 15.24, Actual: 11.80
Congestion Neighborhood: 1.9357168224704202, Predicted: 10.95, Actual: 3.30
Congestion Neighborhood: -0.07581598628125683, Predicted: 10.83, Actual: 11.16
Congestion Neighborhood: -0.4058330877170788, Predicted: 19.28, Actual: 28.55
Congestion Neighborhood: 1.2442524194620312, Predicted: 39.00, Actual: 34.00
Congestion Neighborhood: -0.9087162899049981, Predicted: 13.50, Actual: 14.10
Congestion Neighborhood: 0.4584974160434074, Predicted: 10.64, Actual: 11.80
Congestion Neighborhood: -0.9244313899733706, Predicted: 8.98, Actual: 8.30
Congestion Neighborhood: -0.39011798764870637, Predicted: 21.31, Actual: 23.81
```

MSE : Lower values indicate better performance.
R-squared: An R-squared of 0.32 means that 32% of the variance in the fare amount is explained by the model.
MAE: Lower values indicate better accuracy.

Random Forest Regression Metrics:

```
Random Forest Regression Metrics:
Mean Squared Error: 143.28
R-squared: 0.49
Mean Absolute Error: 4.93
```

```
Random Forest Regression Predicted vs Actual Fares with Congestion Neighborhood:
Congestion Neighborhood: -0.4058330877170788, Predicted: 14.64, Actual: 10.80
Congestion Neighborhood: -0.4058330877170788, Predicted: 15.22, Actual: 11.80
Congestion Neighborhood: 1.9357168224704202, Predicted: 11.27, Actual: 3.30
Congestion Neighborhood: -0.07581598628125683, Predicted: 10.87, Actual: 11.16
Congestion Neighborhood: -0.4058330877170788, Predicted: 19.29, Actual: 28.55
Congestion Neighborhood: 1.2442524194620312, Predicted: 24.98, Actual: 34.00
Congestion Neighborhood: -0.9087162899049981, Predicted: 13.45, Actual: 14.10
Congestion Neighborhood: 0.4584974160434074, Predicted: 10.59, Actual: 11.80
Congestion Neighborhood: -0.9244313899733706, Predicted: 8.98, Actual: 8.30
Congestion Neighborhood: -0.39011798764870637, Predicted: 21.34, Actual: 23.81
```

MSE: Lower values indicate better performance.
R-squared: Indicating a moderate level of explanatory power.
MAE: Lower values indicate better accuracy.
The Random Forest Regression model is performing the best among the three models based on the provided metrics. It offers the lowest errors and the highest explanatory power. And the outputs for each model also help to find which model performs best in prediction of fares in comparison with the actual fares. The outputs of the three models show the difference between actual and predicted fares, the lower the difference the better the performance of the model. In conclusion, the Random Forest Regression performs best. Thus, we have successfully solved the problem.

High Congestion Neighborhoods: 82,182,18,153,191,9,31 these neighborhoods experience significantly longer trip durations and higher total fares, indicating the impact of congestion on both travel time and cost. This suggests that congestion patterns vary spatially within the city, influenced by factors such as traffic flow, road infrastructure, and time of day.

Deploying the solution in the real world to create value for someone:
The solution helps city planners and leaders about effects of traffic congestion to make better transportation system. Taxi drivers or companies will use this information to make reliable services for passengers. Intern passengers can save money and time with this information. More helpful to people who are living in busy areas can make better decisions based on the information to make their lives more efficient.

Appendix A:

Statement of goals achieved by the team:

As a team we achieved a goal of understanding how traffic jams impact how long taxi rides are and how much they cost in different parts of New York City. We also investigated how traffic congestion changes throughout the day and in different areas.

Appendix B: Statement of goals achieved by each person individually.

Jaladurgam Navya: Achieved the goal of understanding the impact of traffic congestion affect taxi trip durations and fares across various neighborhoods of New York City by using congestion analysis.
Mukthasree Vengoti : Achieved the goal of understanding the spatial and temporal patterns in congestion levels.
Aravind Reddy Thummala:
Achieved the goal of developing predictive models to estimate taxi fares by considering factors such as congestion levels, trip distance, pickup time, and neighborhood.

Github link: https://github.com/navya56789/Navya188/tree/main
Colab: :
https://colab.research.google.com/drive/1seVt0yvBbKO5VHZdKb1acOD_nbN2XCap?usp=sharing