Introduction

The New York City MTA system serves millions of passengers daily. Given the number of passengers the MTA now faces significant challenges related to congestion which can lead to overcrowding, delays, and a diminished travel experience for customers. We aim to identify the factors that contribute to congestion in order to develop effective strategies to mitigate the issue. We will analyze historical data on ridership, weather conditions, and gas prices in an attempt to find patterns and relationships that contribute to congestion. Improved travel conditions can lead to reduced stress, increased productivity, and a better quality of life for commuters. Understanding the factors that contribute to congestion can inform strategic decision making and allocation of resources within the MTA.

Data

The dataset used in this project is a combination of several data sources. The main dataset, MetroCard History, contains information of the fare types and usage at various stations over specific date ranges. The additional datasets include a list of US holidays from 2004-2021, gasoline prices over the years in NY state, and lastly weather data which included total rainfall and snowfall within the date ranges specified in the metrocard history dataset. The final dataframe contains the columns:

- From Date
- To Date
- Remote Station ID
- Station
- Total Fares for the Week
- Contains Holiday
- Gas Price
- Total Snowfall
- Total Rainfall

Data Assumptions and Limitations

During the data cleaning process, certain assumptions were made to handle missing or null values. In our metrocard history dataset all null values were replaced with 0 under the assumption that no passengers used that specific fare type at the corresponding station during the given time period. This assumption simplifies the analysis but may not capture all of the nuances of missing data. Forward filling was used to impute missing values in the gas prices dataset, the assumption is that gas prices remain constant until the next recorded price change. While this approach provides a reasonable estimate, it may not account for short-term fluctuations or sudden price changes that occurred between the recorded dates. Regarding the weather data, the focus was primarily on snowfall and rainfall, as the occurrence of other weather events was relatively low and were most likely less significant for the analysis. However it is important to note that other weather factors, such as temperature or wind speed, may also influence ridership patterns and congestion levels.

It is important to recognize the assumptions made in our data cleaning step. The analysis and findings should be interpreted in light of these limitations, and future work may consider incorporating additional data sources or more advanced imputation techniques to address missing values. Despite these limitations the dataset provides a solid foundation for exploring the relationship between various factors and congestion in the MTA.

Methodology

In order to prepare the data for analysis we took several steps. First we dropped the Unnamed: 0 column as it was not relevant to the analysis. We then created dummy variables for the Contains Holiday column splitting the column into two new columns. We then used the Standard Scalar to help normalize our data for analysis. The resulting dataframe contained the the columns: 'From Date', 'To Date', 'Remote Station ID', 'Station', 'Total Fares for the Week', 'Gas Price', 'Total Snowfall', 'Total Rainfall', 'Contains Holiday_No', and 'Contains Holiday_Yes'.

Exploratory Data Analysis (EDA):

After we prepared our data we conducted EDA in order to gain insights into the relationships between the variables and identify any patterns or trends. Our EDA revealed interesting distributions for the features in our dataset.

The distribution of Total Fares for the Week is heavily right-skewed, with the majority of the totals falling between 0 and 100,000 as seen in figure 1. However there are some outliers on the higher end with a few weeks seeing total fares above 400,000.

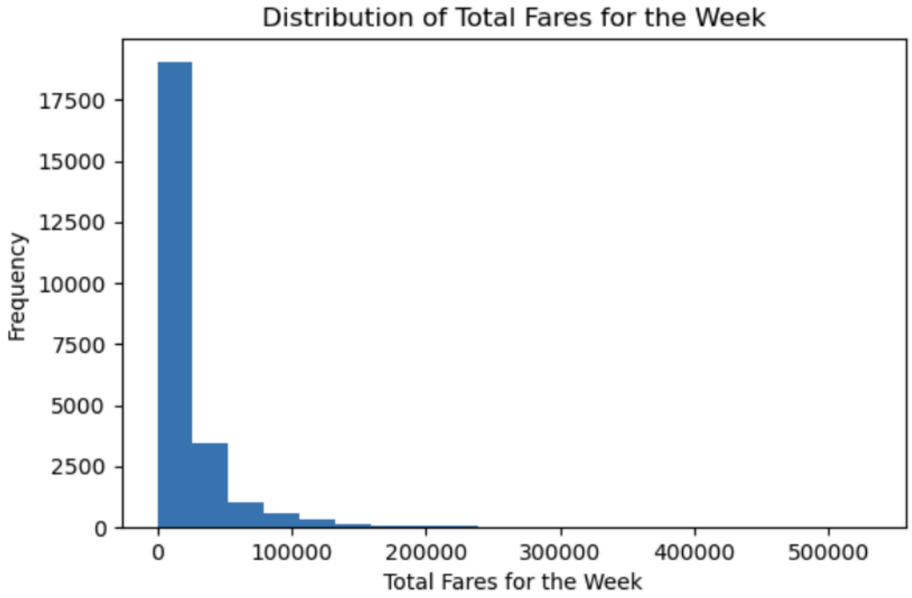


Figure 1: Distribution of total fares for the week

Gas prices seemed to cluster around certain points during the time period analyzed. Specifically around 2.2, 2.6, and 3.2. As seen in figure 2.

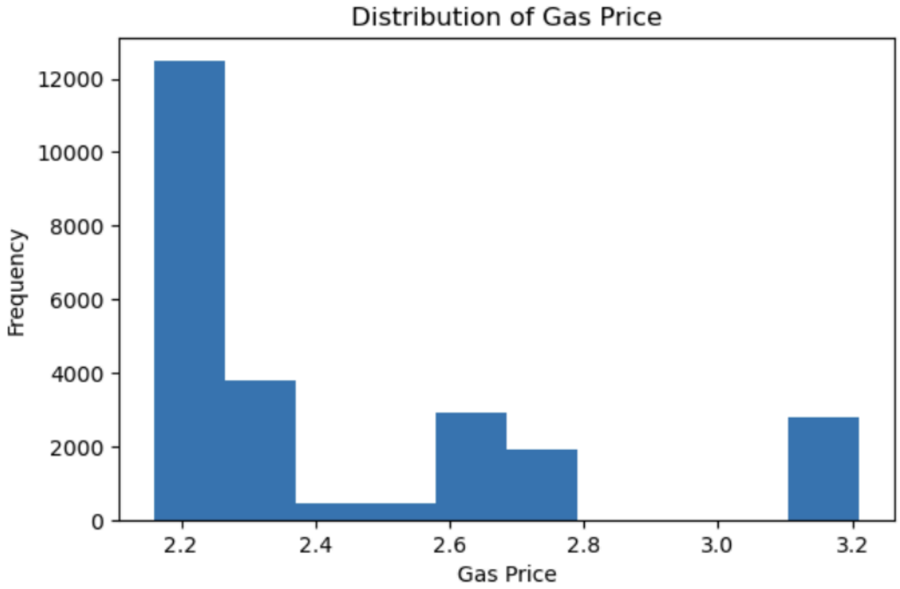


Figure 2: Distribution of Gas prices over the period 2020-2021

Lastly we look at the total snowfall and total rainfall for the time period. Both of these features followed a similar pattern to the features so far being heavily right skewed as seen in figure 3.

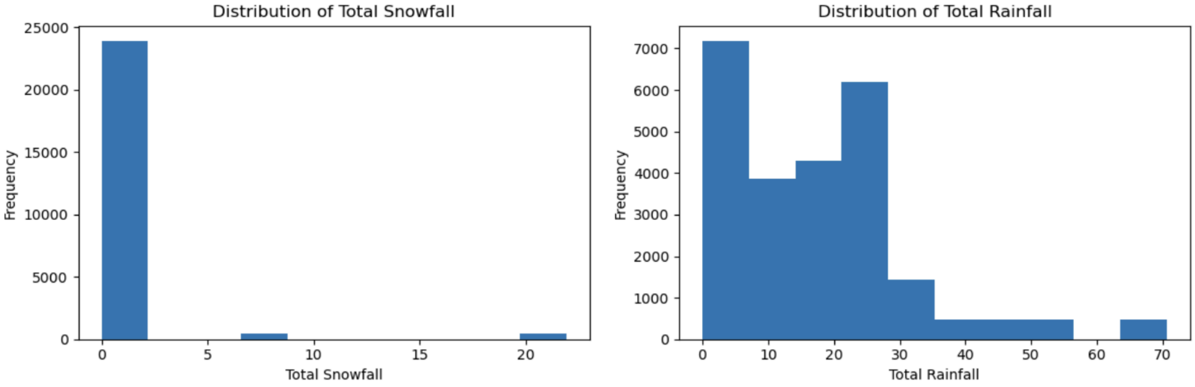


Figure 3: Distribution of snowfall and rainfall over the period 2020-2021

One feature of the EDA is that all of the features exhibit a right-skew. The right skew suggests that, for most periods, our features are relatively low. However, there are occasional weeks where these features take on unusually high values. These high-value outliers could have a significant impact on any model or analysis that uses these variables. These outliers present interesting points to investigate further. Understanding the circumstances that lead to unusually high fares could lead to valuable insights.

Our next step in EDA was to create a correlation matrix of numerical variables (Figure 4).

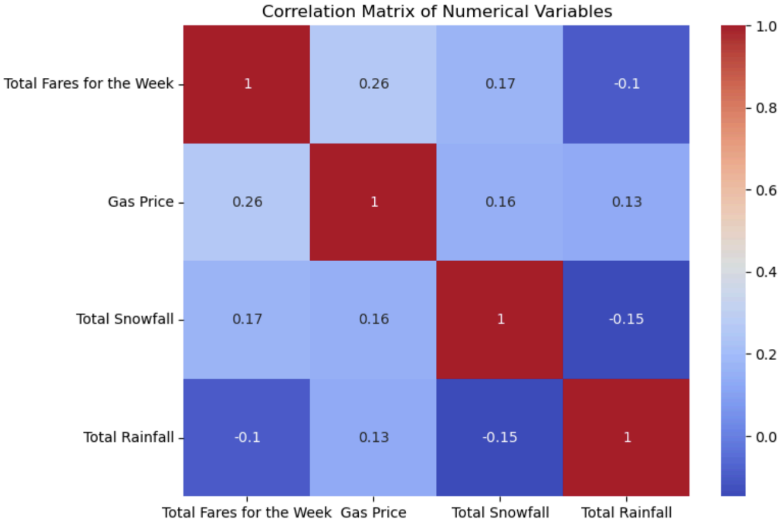


Figure 4: Correlation Matrix of numerical variables

The heatmap shows that the total fares have a moderate positive correlation with the gas prices (0.26) and a weak positive correlation with the total snowfall, as well as a weak negative correlation with the rainfall. This suggests that the weather and gas prices may have some influence on MTA ridership and fare revenue. Lastly we looked at the average total fares for weeks that contain a holiday and weeks that do not contain a holiday (Figure 5).

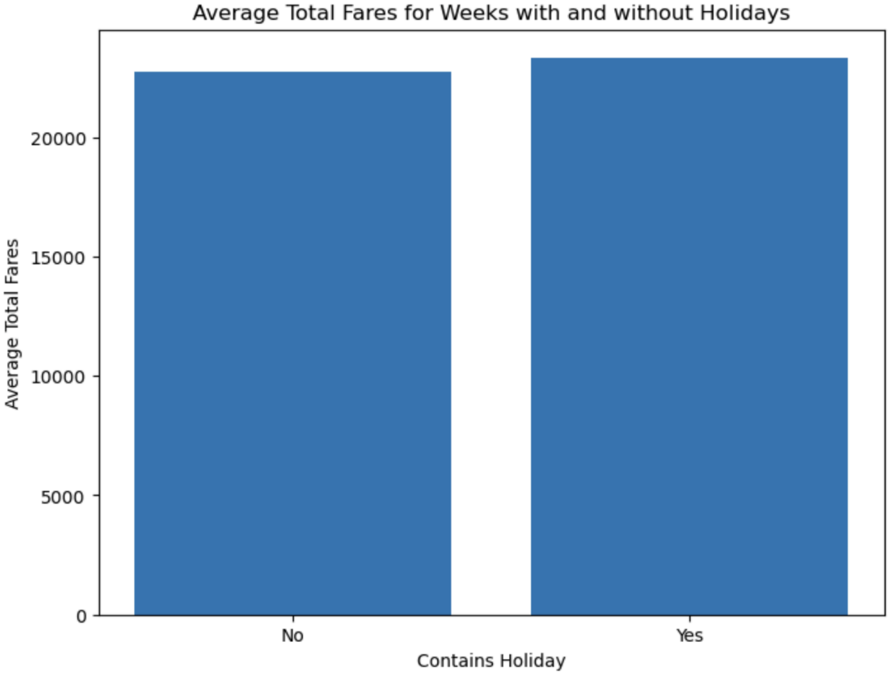


Figure 4: Average fare for weeks with and without a holiday

The bar plot clearly shows that the average total fares are higher during weeks that contain a holiday compared to weeks without holidays. This suggests that holidays have a positive impact on MTA ridership and fare revenue.

Model Development

In our model development step we looked at three models: Linear Regression, Random Forest, and XGBoost. These models were chosen to predict the total fares for the week based on the weather, gas prices, and holidays. In order to evaluate the model performance we used two metrics: Mean Squared Error (MSE) and R-squared (R^2). The evaluation of the linear model is as follows:

Linear Regression:

- MSE: 0.8450
- R-squared: 0.1021

These valued align with the observations we made during our EDA, suggesting there is not a strong linear relationship between the features and target feature.

Hyperparameter Tuning

To optimize performance of the Random Forest and XGBoost models, we performed hyperparameter tuning using grid search with cross-validation. The hyperparameters are as follows:

Random Forest:

- n_estimators: [50, 100, 200]
- max_depth: [None, 5, 10]
- min_samples_split: [2, 5, 10]

The best hyperparameters found for the Random Forest model were:

- n_estimators: 200
- max_depth: 10
- min_samples_split: 2

XGBoost:

- n_estimators: [50, 100, 200]
- max_depth: [3, 5, 7]
- learning_rate: [0.01, 0.1, 0.3]
- subsample: [0.5, 0.7, 1.0]

The best hyperparameters found for the XGBoost model were:

- n_estimators: 100
- max_depth: 3
- learning_rate: 0.1
- subsample: 1.0

The evaluation results for each model are as follows:

Random Forest (Best Model):

- MSE: 0.7204
- R-squared: 0.2345

XGBoost (Best Model):

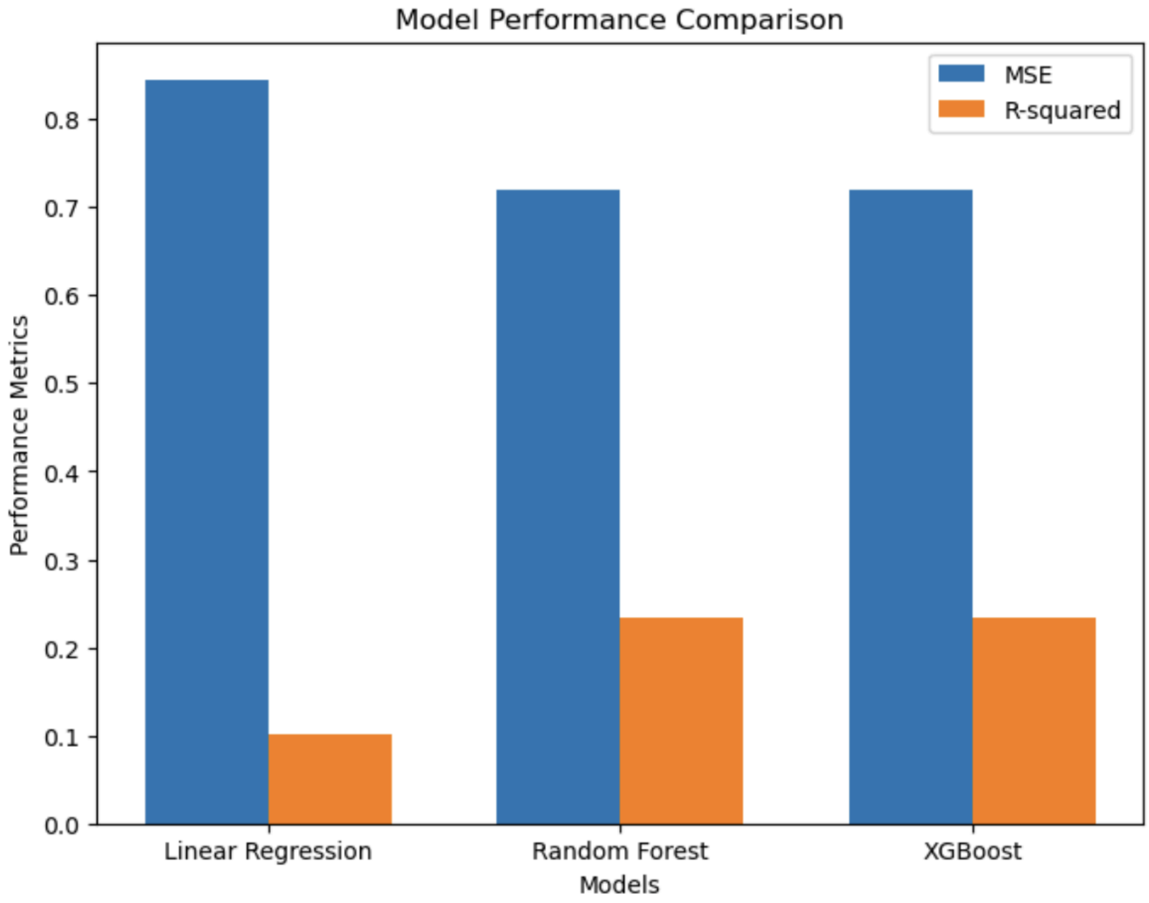
- MSE: 0.7202
- R-squared: 0.2347

Model Evaluation

Based on these results, we can observe that the Random Forest and XGBoost models outperformed the Linear Regression model in terms of both MSE and R-Squared. The XGBoost model achieved slightly better performance compared to the Random Forest. It is important to note that while the models showed some improvement over the baseline Linear Regression model, the R-squared values are all relatively low, indicating that there is still a significant amount of unexplained variance in the data. This suggests that there might be other important factors influencing the total fares that are not captured by the current set of features.

Results

We compare the models to identify the best performing model.



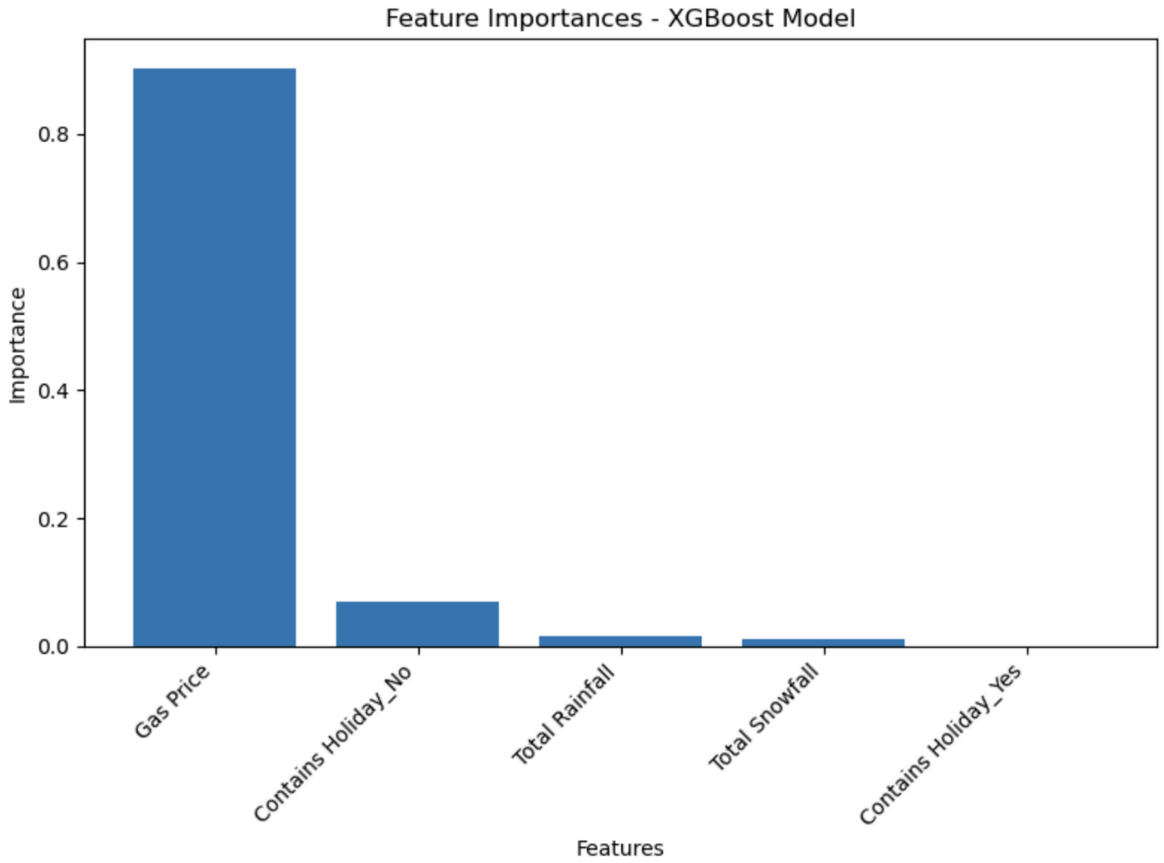
The XGBoost model achieved the best performance with the lowest MSE and the highest R-Squared. Closely followed by the Random Forest. The Linear Regression model had the poorest performance among the three models.

Based on the evaluation metrics, the XGBoost model was the best performing, The best hyperparameters for the XGBoost model were:

- n_estimators: 100
- max_depth: 3
- learning_rate: 0.1
- subsample: 1.0

We obtained these values through the grid search with cross-validation.

Given the performance of our models we look to gain insights into the importance of each feature in the XGBoost model.



Interestingly enough the feature importance plot shows us that gas prices and the absence of a holiday are the most influential features for predicting the total fares for the week with the absence of a holiday having a relatively lower importance than the gas price. Total snowfall and total rainfall have lower importance.

Conclusion

We aimed to investigate the factors influencing congestion in the NYC MTA system and provide recommendations to alleviate this issue. By analyzing historical data on ridership, weather conditions, gas prices, and the presence of holidays, we developed machine learning models to predict the total fares for the week, which served as a proxy for congestion levels.

Our analysis revealed several findings:

- Gas prices emerged as the most important feature in predicting total fares, with a significantly higher importance compared to the other features. This suggests that gas prices have a larger impact on MTA ridership and congestion when compared to the other features.

- The presence of holidays play a role in influencing total fares, with weeks containing holidays generally experiencing higher ridership and congestion compared to non-holiday weeks.

- Weather factors such as total rainfall and total snowfall had relatively minimal importance in the models, indicating that they may not be the primary drivers of congestion in the MTA system.

However, it is important to note the limitations of our models. Despite the use of techniques like XGBoost and Grid search, the models all achieved relatively low R-squared values, suggesting that there is still a significant amount of unexplained variance in the data. This indicated that there may be other important factors influencing congestion that were not captured by the current set of features.

Based on our findings, we would recommend:

- The MTA should consider implementing dynamic pricing that adjusts fares based on gas prices to encourage the use of public transport during times of higher gas prices.

- Special attention should be given to managing congestions during holiday periods, as these times seem to experience higher ridership. Strategies such as increasing service frequency or implementing congestion pricing could be helpful.

Future Work

To further enhance our understanding of congestion in the NYC MTA system and improve the predictive models for future research and analysis, we recommend:

- Explore other potential features that may influence congestion, such as economic indicators, population density, or major events happening in the city like sporting events or concerts. A wider range of relevant features could improve the explanatory power of the models.

- Analyzing the time series patterns of congestion, such as daily, weekly, or seasonal variations may help the MTA optimize resource allocation and service planning.

- Experimenting with other advanced machine learning models such as deep learning or ensemble methods could help to capture more complex relationships and improve predictive performance.

By continuing to refine our models, incorporating additional data sources, and exploring advanced analytical techniques, we can further improve our understanding of congestion in the NYC MTA system and develop more effective strategies to mitigate this issue. The insights gained from this project provide a foundation for ongoing research to enhance the efficiency and quality for commuters.