Predicting Transportation Preference of New Yorkers based on Weather Patterns

Machine Learning for Cities
Project Report

Group Members
Ajayrangan Kasturirangan - ak10196@nyu.edu
Akshay Shetty - as16477@nyu.edu
Sharvari Deshpande - sd5270@nyu.edu
Shantanu Anikhindi - saa9213@nyu.edu
Devang Dave - dld9783@nyu.edu

Introduction

Description

Weather anomalies imply a feeling of uncertainty among commuters. The weather change has revealed behavioral shifts in relation to a person's decision to utilize numerous or single modes of transportation during his commute. Weather can have a significant impact on transportation in New York City, as it can affect road and sidewalk conditions, as well as the operation of public transportation systems. For example, heavy snowfall can make roads and sidewalks slippery and difficult to navigate, leading to traffic delays and disruptions to public transportation services. Similarly, extreme heat can cause train tracks to expand and contract, leading to slower train speeds and potential delays. Additionally, severe weather events such as hurricanes and tropical storms can cause widespread power outages and flooding, which can severely disrupt transportation systems. Overall, weather can have both positive and negative effects on transportation in New York City, depending on the specific weather conditions and the transportation mode being used. The study will forecast the movement between various public modes of transport of individuals throughout their everyday commutes over the course of different weather seasons. The project involves collecting data on the weather patterns in New York City and the transportation choices of its residents. This data could then be used to develop a model that predicts how the weather may influence a person's transportation preference. The model could be used to help city planners and transportation providers better understand and plan for the transportation needs of New Yorkers based on the weather.

Motivation

New York City has a complicated network of public transit routes, notably subways. By understanding how weather influences people's transportation choices, transportation authorities and policymakers could make more informed decisions about infrastructure investment, traffic management, and other transportation-related issues. This could lead to more efficient and effective use of transportation resources and improve transportation experiences for individuals living in or visiting New York City. It may also help local governments promote public safety by allocating police officers based on predicted footfall. Additionally, anticipating demand can help the concerned authorities adjust transit services to avoid overcrowding and congestion.

Current Research

Existing research on the effects of weather on public transportation has mostly focused on the effects of average weather conditions on total public transit ridership. Our project aims to get a better understanding of the relationship between daily weather and public transportation. Research has investigated the effect of weather on transportation, looking at how it impacts the performance of transport systems, including infrastructure performance, road capacity and vehicle speed, and the disruptions caused by bad weather. For example, Koetse and Rietveld (2009) studied the influence of weather on infrastructure performance, while Kyte et al. (2001) and Smith et al. (2004) examined the effects of weather on road capacity and vehicle speed. Hofmann and O'Mahony (2005) investigated the disruptions caused by adverse weather on transport systems.

Several studies have been conducted on the influence of weather on passenger travel behavior. These studies have examined the impact of weather conditions on travel demand (Cools et al., 2010), modal shift (Heinen et al., 2010; Koetse and Rietveld, 2009), and route and destination choices (Cools et al., 2010). These studies have covered different weather elements and have targeted various transportation modes and systems. Studies have looked at the relationship between temperature and transportation mode, as well as the effects of weather on traffic volume and the use of active transportation modes like cycling. For instance, Muller et al. (2008) studied the link between seasonal temperature fluctuations and shifts in transportation mode, while Tang and Thakuriah (2012) examined the association between vehicle traffic volume and cold weather. Maze et al. (2006) investigated the effects of snow and heavy wind on traffic. Additionally, research has focused on the impact of weather on active transportation modes, including cycling (Nankervis, 1999; Bergström and Magnusson, 2003; Brandenburg et al., 2004).

In contrast to private and active transportation modes, the impact of weather on public transit has gotten comparatively little study. One reason for this could be that the effects of weather on public transportation are more complex than on other forms of transportation. Another reason could be a result of the limited availability of data. For example, Guo et al. (2007) and Stover and McCormack (2012) studied the effects of average weather conditions on aggregate public transit ridership. Existing studies on the impact of weather on public transit do not account for daily weather fluctuations and ridership, making it difficult to establish clear connections between weather and public transportation demand. These studies also do not examine the effects of weather on individual public transit passengers' travel behavior. Understanding the influence of weather on public transportation is important for developing better transit management strategies and improving resilience to weather fluctuations.

Data Analysis

Data Collection

For the project, several datasets of different modes of transportation in NYC were utilized. Different modes of transportation like cycling, metro and taxi (for-hire vehicle) were considered. The project analyzes pre-pandemic data for the years 2018, 2019.

1) Weather data
The datasets were gathered from Visual Crossing Weather API Platform. The dataset contains all the weather information along with details on the meteorological conditions of NYC on the particular day. Some of these include precipitation, temperature, humidity, and snow.
Source: Historical Weather Data of New York City

2) Subway Turnstile Entry and Exit Data
The datasets were collected from MTA Open Data NYC. The dataset contains information on entries, exits and dates.
Source: MTA Turnstile Data

3) Citi bike Data
The datasets were collected from the Official Website of Citi Bike. The dataset contains information on the number of trips, start time, end time, date.
Source: CitiBike Data

4) Taxi Data
FHV taxi data includes the yellow taxi data and green taxi data. The dataset was collected from the TLC (Taxi and Limousine Commission) NYC Open Data. It contains information on date, time, and passenger count.
Source: Taxi Trip Records

Data Cleaning

Data cleaning is an essential step in the process of developing a machine learning model. It involves identifying and removing irrelevant, inconsistent, or incomplete data, as well as correcting any errors or anomalies that may be present in the data. The goal of data cleaning is to improve the performance of the machine learning algorithms that will be applied to the data, by ensuring that the data is accurate, consistent, and complete. This can be achieved by applying various techniques, such as outlier detection, missing value imputation, and data transformation, to the data. Additionally, it is important to maintain the relevance and integrity of the data throughout the cleaning process, so that the resulting data accurately reflects the underlying phenomena being studied and can be used to generate meaningful and reliable insights. In order to replicate to maintain this standard we have performed the following transformations to our datasets.

1) Weather data (Features)

To generalize and ameliorate the raw data, the hourly data was grouped into six-hour intervals (0, 4, 8, 12, 16, 20) and the mean values of the continuous variables were calculated for each interval. This approach assumes that weather changes happen gradually, rather than discreetly,

and that the mean values of the surrounding intervals can provide a fairly accurate estimate of the weather conditions at a given time. By aggregating the data in this way, the number of null values is reduced and the data is regularized, making it more suitable for analysis and modeling.

2) MTA Turnstile Entry and Exit data

The MTA data available for this project is provided at a four-hour frequency, and the turnstile counters do not reset to zero at a fixed frequency. To make the data more usable, the following steps were taken. The data was grouped by station, control area name, unit, subunit/channel/position, time, and description, to identify each turnstile uniquely at each four-hour interval. The difference between the current and following turnstile counter values was treated as the number of passengers passing through the turnstile. If this value was negative, it indicated that the counter had been reset. Outliers and negative values were handled by imputing them with median values, based on their context. This process ensured that the data was accurately and consistently captured, allowing for more accurate analysis and modeling.

3) Citi Bike Ridership Data

The citibike data used in this project was clean and well-organized, but it was available at a very fine-grained level, with each trip recorded individually. This made the cleaning and aggregation process laborious and resource-intensive. To address this issue, the data was aggregated at the hour-bin level one month at a time, while making a copy of already-processed data. This allowed for more efficient data handling and enabled the necessary information to be captured for analysis and modeling. The number of trips was calculated for each hour-bin, providing a summary of the citibike usage patterns over time.

4) Yellow and Green Taxi Ridership Data

The Green and Yellow taxi data was recovered from the TLC data which was in parquet format. The total number of rows for green taxi were more than 7 million for 2 years and yellow taxi was more than 160 million records which was a challenge to filter data with csv. To handle this issue every 2 months of data was downloaded, filtered and made a dataframe to the required format and appended at the end to get a complete dataframe.

Exploring the trends

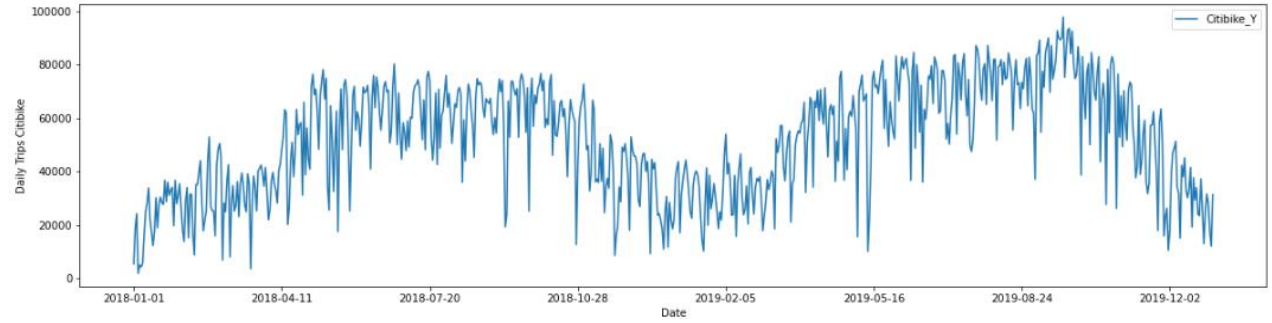

Figure 1: Trend for Citibike data

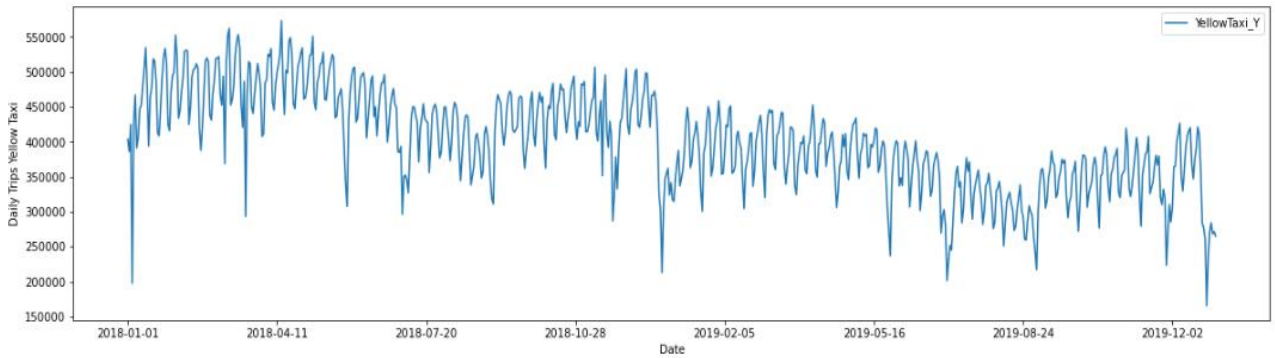

Figure 2: Trend for Yellow Taxi data

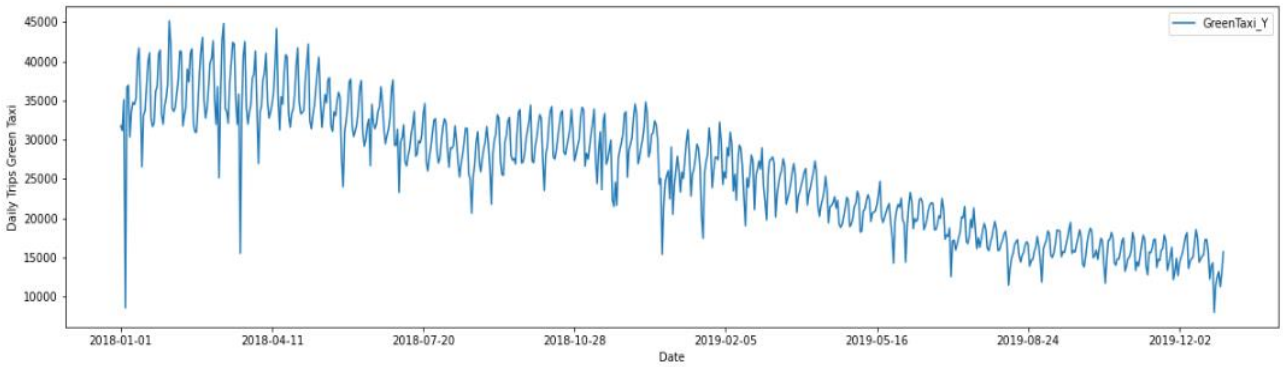

Figure 3: Trend for Green Taxi data

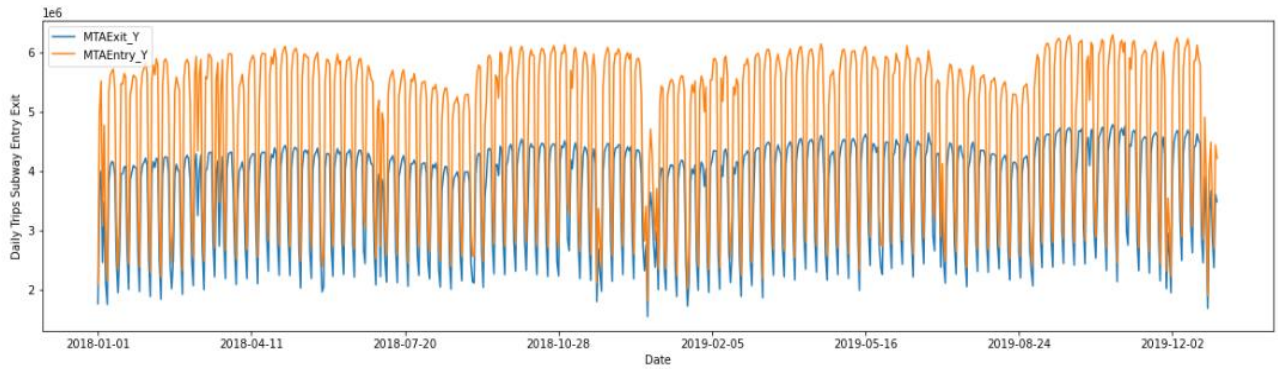

Figure 4: Trend for MTA Entry/Exit data

Once trends have been identified in ridership data for NYC citibike, MTA, green and yellow taxi, they can be related to a machine learning model in several ways. First, the trends can be used to train and test the machine learning model. For example, a trend has been identified showing an increase in ridership on citibikes during the summer months, this information could be used to train the model to predict increased ridership during the summer.

The general trends for taxi show a decrease in ridership over the two years more so for green taxis than the yellow. The MTA data is more consistent throughout the two years in ridership. For example, if a trend is identified showing a strong correlation between the number of riders on and the weather, this information could be used to incorporate a feature representing the weather into the machine learning model. Because we see a difference in trends for various transportation modes, we observed that there was a possibility of overfitting or underfitting a particular model in for different Y variables of a model.

Anomaly Detection

To determine the anomalies in the weather and transportation pattern we predicted the Gaussian Mixture model on the existent data counts of the transportation modes considering all modes of transportation together without considering the weather data. We found various anomaly patterns for different models run on the data, out of which the Gaussian mixture model gave us a better density of scores with extreme anomalies.

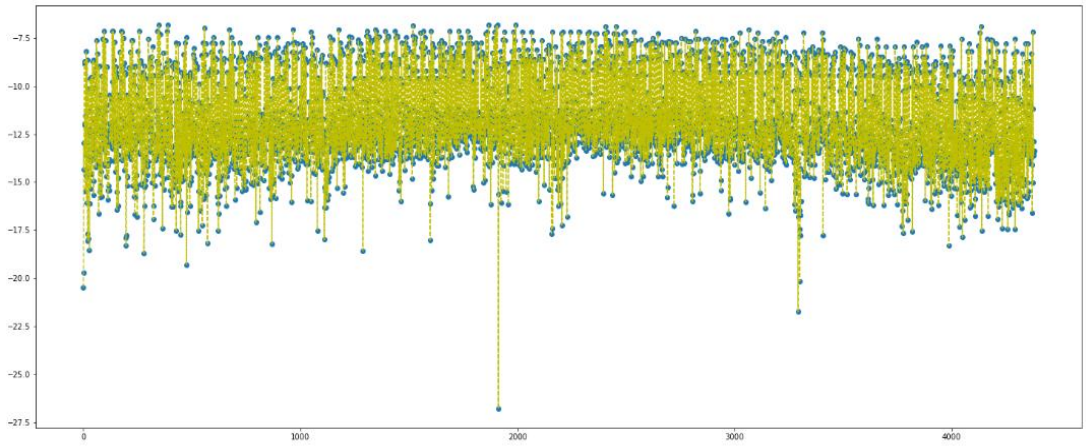

Figure 5: GM anomaly score graph

We found the following extreme anomalies observations with the Gaussian mixture scores

1) For the date of 15th November 2018, there was a pre seasonal snow in NYC that spiked the number of counts for people using MTA subways over other modes of transport.

2) The new years on 1st Jan 2018 ridership for all modes of transport saw a spike for the hour bins of 0 and 4.

3) The days after and before of 4th of July, 2019 saw an increase and in ridership for late night hours bins

In reference to the graph of Gaussian Mixture scores shown above, we considered all anomaly points less than -15 to be considered as an extreme anomaly which we took into account to calculate the outlier factor within a certain range.

Forecasting Methodology

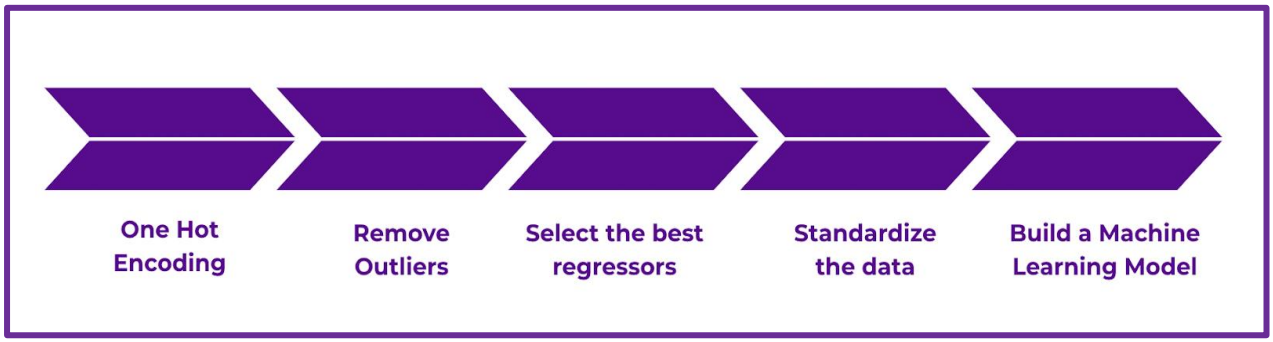

Figure 6: Flow chart of the methodology

1) One Hot encoding

One-hot encoding was used to categorize the hour-bin, day-of-the-week, and monthly variables in this project. These variables were introduced to capture the seasonal nature of the target variables, but they do not have an ordinal nature, as they are cyclic in nature. One-hot encoding ensures that each value of these categorical variables is treated as an independent variable, allowing the machine learning algorithms to properly process and analyze the data. This approach allows for more accurate and meaningful analysis of the data, by ensuring that the relationships between the variables are properly represented.

2) Remove Outliers

In order to improve the generalizability of the model and prevent it from being skewed by extreme weather events, a density-based outlier filtering method known as Local Outlier Factor (LOF) was introduced. This method allows for the detection of outliers based on their context, rather than their absolute values. For example, high levels of snow in May would be considered an outlier, even if the absolute value is less than a typical snow day in January. The amount of outlier removal is controlled using an outlier factor, which is a parameter in the model. This approach helps to ensure that the model is not influenced by extreme weather events, allowing it to make more accurate and reliable predictions.

3) Selecting the best regressors

Having a large number of features in a machine learning model can have several disadvantages. These disadvantages can be particularly relevant in our case, as we have more than 42 features in our model. One disadvantage is that having a large number of features can make the model more complex and difficult to train, which can lead to longer training times, higher memory requirements, and reduced performance on unseen data. Another disadvantage is that having a large number of features can increase the risk of overfitting, where the model performs well on the training data but poorly on new data. This can lead to inaccurate and unreliable predictions, and can make the model less useful for real-world applications. Finally, having too many features can also make it difficult to interpret the results of the model, as the relationship between the features and the target variable may be complex and hard to understand. To address these disadvantages, we chose to use only the n regressors with the best n r squared values as a parameter in our model.

4) Standardize the data

Standardizing data is important in machine learning when we have many features with different measurement units because the model will typically perform better when all of the features are on the same scale. This is because many machine learning algorithms that we plan to use some form of distance or similarity measure to compare observations and make predictions, and these measures are often sensitive to the scale of the features. For example, if precipitation is measured in inches and temperature is measured in Fahrenheit, the model may give more weight to the precipitation because it has a smaller scale. This can lead to inaccurate and unreliable predictions. Standardizing the data ensures that all of the features are on the same scale, which can improve the performance of the model. Additionally, standardizing the data can also make it easier to compare the importance of different features, since they will all be on the same scale.

5) Building a Machine Learning Model

We used two approaches to the same problem by treating this as a regression problem and a classification problem as well. In order to implement these two approaches, we did the following:

a. Regression Based

To evaluate the performance of a regression model, we calculate the model's accuracy by allowing for a margin of error of 15%. This means that we consider the model to be accurate if the predicted values are within 15% of the true values. This is a common approach for evaluating regression models, as it allows us to assess the model's performance while taking into account the inherent uncertainty and variability in the data. By allowing for a margin of error of 15%, we are indicating that we are willing to tolerate a certain degree of error in the predictions made by the model. This can help us to determine whether the model is performing well or poorly, and can guide our efforts to improve the model's performance.

b. Classification Based

To create a baseline for a classification method, we binned the target variables into 30 equal size bins and allowed for a margin of error of 1 while evaluating the model's accuracy. This is a common approach for creating a baseline, as it allows us to compare the performance of our model to a simple, standardized benchmark. By allowing for an error of 1, we are indicating that we are willing to tolerate a certain level of misclassification when evaluating the model's accuracy. This can help us to determine whether our model is performing significantly better or worse than the baseline, and can guide our efforts to improve the model's performance.

Models Used

Our model included both regression and classification analysis, with daily foot track count as our goal. Each of our analysis includes the following:

Regression Methods

1) Linear Regression

Linear regression did not perform well when predicting the outcome of our data because the relationship between many of the variables appeared to be non-linear. This is a common issue with linear regression, as it assumes that the relationship between the variables is linear. When the relationship is non-linear, linear regression can produce inaccurate and unreliable

predictions. One possible solution to this problem is to transform the variables to other scales, such as logarithmic or exponential scales. However, this can be time-consuming and may not be necessary if we use other methods that are better suited to non-linear relationships. Additionally, transforming the variables on a large scale may be redundant if we use other methods that can handle non-linear relationships without the need for transformation.

2) Support Vector Regressor

The polynomial kernel only worked well for higher degrees but resulted in excessive overfitting. Additionally, using penalty terms to try to reduce overfitting often led to a reduction in the model's accuracy on the training data. These issues made it difficult to use support vector regression effectively for our data, so we chose to explore other methods instead. This allowed us to avoid dealing with the challenges associated with support vector regression and focus on methods that were more suitable for our data and objectives.

3) SGD Regressor

As we fine-tuned the hyperparameters for our stochastic gradient descent regressor, we observed that the model's performance reached a plateau at a certain level of accuracy. This indicated that the model had reached its optimal performance and that further improvements were not possible. To try to improve the model's accuracy, we reduced the learning rate, which allowed the model to make more fine-grained adjustments to its parameters. However, this made the training process very time-consuming, as the model had to process a large number of iterations to converge on the optimal solution. This made it difficult to use the stochastic gradient descent regressor effectively for our data, and led us to consider other methods that might be more efficient and effective.

4) Decision Tree Regressor

The decision tree regressor was an effective tool for modeling complex functions and achieving higher out-of-sample accuracy. We used tree-based methods and techniques to generalize our models, and the MTA Exit model used a Decision Tree regressor for its final output. Overall, this approach helped us achieve better results.

5) Random Forest Regressor

Trying a Random Forest Regressor gave the best results for our decision tree model. This improved performance for all of the targets we tested: MTA Entry, Citibike, Yellow Taxi, and Green Taxi.

6) Gradient Boosting Regressor

We were tempted to try a gradient boosting regressor, but we soon realized the computational cost of this approach made it difficult to achieve good results. Gradient boosting is an additive process, which can be computationally expensive.

7) Gaussian Process Regressor

To account for seasonality in Citibike data, the decreasing trend in Yellow and Green Taxi data, and the high level of noise in the data, we tried fitting kernels to a Gaussian process regressor. Although this approach yielded very high in-sample accuracy (99%), the out-of-sample accuracy was much lower (35% at most). This suggests that the model may not have generalized well to new data.

Classification Methods

1) *Decision Tree Classifier*

The decision tree classifier performed well for many of the target variables, but we chose to use a random forest model instead because it offered better accuracy and we didn't mind sacrificing interpretability.

2) *Random Forest Classifier*

We trained all of our target variables using a random forest classifier and were able to achieve out-of-sample accuracy between 0.7 and 0.8, with very little overfitting. This suggests that the random forest model was effective at accurately predicting the target variables without memorizing the training data.

Challenges

1) We tried polynomial regression because the relationships between variables were not linear. In some cases, the complexity of a polynomial regression model was too high for it to be practical to use due to computational limitations. Even though we chose the best correlators to reduce the number of features, the data used to train the model was very large and complex to be computationally tractable.
2) Interpreting the decision tree with many features was challenging, as the tree was complex and difficult to understand. Additionally, although random forests gave good accuracy, it was not very interpretable because of being an ensemble method.
3) For tackling Overfitting in our model, we used feature selection to select the best correlators and grid search cross-validation to evaluate the model and used early stopping to halt training when the model began too overfit.

Results and Interpretation

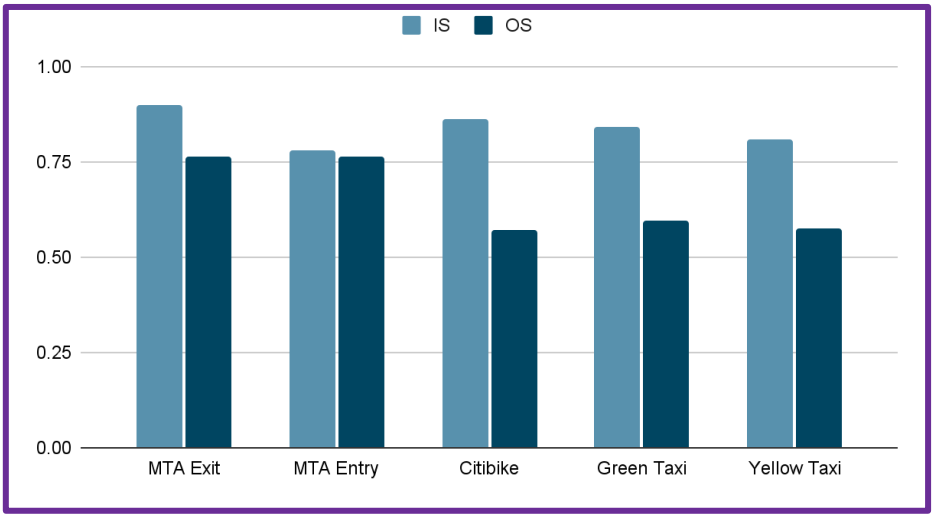

Figure 7: Accuracy for Regression based Approach

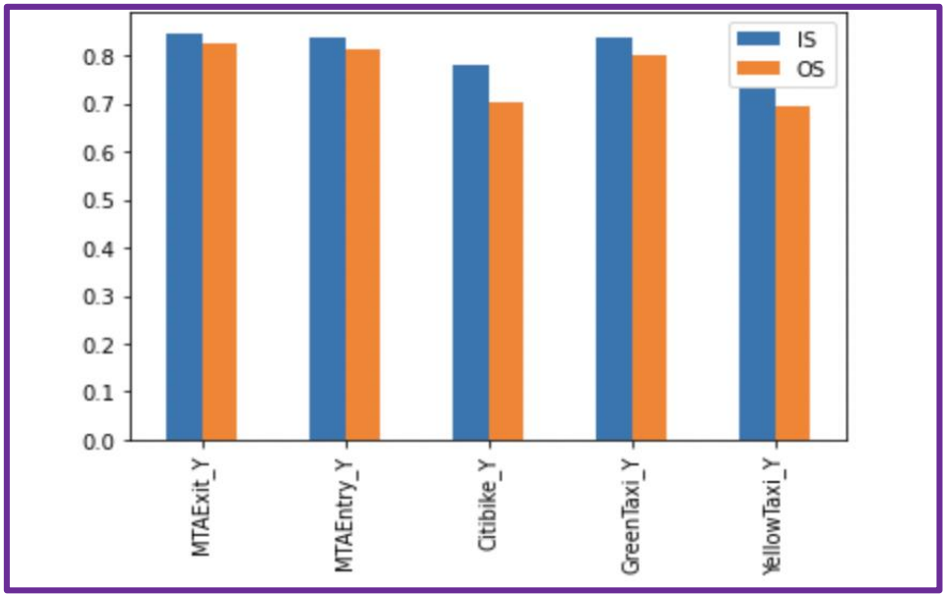

Figure 8: Accuracy for Classification based Approach

For a regression-based approach, the in-sample accuracy is much higher than out-of-sample accuracy. The MTA Exit and Entry model was consistent and had an accuracy of around 80%. Green taxi and yellow taxi had a continuous downward trend and there was comparatively lesser level of noise. The accuracy was around 60%. There was a very high level of noise in the citibike. And the accuracy was approximately 60%. There was a significant amount of overfitting observed in the Citibike model. Similarly, both green and yellow taxi observed overfitting. For a classification-based approach, the in-sample accuracy and out-of-sample accuracy is comparable. The accuracy was above 80% for MTA Exit, Entry, Green and Yellow Taxi. For the Citibike model, accuracy is approximately 70%. The data was already generalized and fit into bin and therefore overfitting observed was much lesser.

Limitations of Analysis

1) We have binned all our variables into 30 groups with an error of 1 for our classification model. The main limitation of using binned data in a machine-learning model is that it can lead to loss of information because some details and nuances in the original data are lost. Binning can make it more difficult for the machine learning model to learn from the data and potentially lead to less accurate predictions.

2) We have allowed a 15% error margin in our regression model which can cause some predictions to be inaccurate, which can be problematic if the predictions are used for decision-making purposes. Additionally, we haven't evaluated classification models with a similar error margin making it more difficult to compare the performance of different models.

Future Scope

1) Currently, our model predicts overall footfall for various modes of transportation at the city level. We could now create models for predicting ridership at each subway station and taxi or citibike demand at a zip code or census tract level. Increasing the granularity of our study could be useful for several reasons. For example, it could help transportation agencies better plan for the capacity of their subway systems and make sure that there are enough trains and personnel available to accommodate the expected number of passengers. This could improve the overall efficiency and reliability of the subway system. Additionally, knowing the expected number of people at a given stop could also be useful for subway riders, who could use the information to plan their trips and avoid crowded trains.

2) In addition to weather data, we can analyze the change in demand due to sporting events. Demand for various modes of transportation in New York City likely increases because many people attend sports games and use the subway or taxis for transiting to and from the stadium or arena. Additionally, a sporting event in the city can attract more visitors and tourists.

Conclusion

The results of this study demonstrate that weather patterns can have a significant impact on transportation preferences among New Yorkers. By analyzing data on weather conditions and transportation mode choices, we were able to identify clear trends and relationships between weather and transportation preferences. Our findings suggest that certain weather conditions, such as extreme temperatures and precipitation, can influence the likelihood of individuals choosing certain modes of transportation. The use of the New York City subway is affected by weather conditions. For example, extreme temperatures and precipitation impact the number of people using the subway, as some individuals choose to avoid using public transportation in inclement weather. Additionally, adverse weather conditions cause disruptions and delays on the subway, which impacts ridership. Citibike, New York City's bike-sharing system, is also affected by weather conditions. There is a seasonal trend in Citibike usage, with higher ridership during the warmer months and lower ridership during the colder months. This is likely due to the fact that people are more likely to choose cycling as a mode of transportation when the weather is pleasant, whereas extreme temperatures and precipitation can make cycling less attractive. Similarly, the taxi trend depicts that it has been decreasing over time which indicates that the ridership for taxi is becoming less popular. These results can be used to inform transportation planning and management strategies in New York City, helping to improve the efficiency and resilience of the city's transportation systems.

Bibliography

Bergström, A., Magnusson, R., 2003. Potential of transferring car trips to bicycles during winter. Transport. Res. Part A Pol. Pract. 37 (8), 649–666.

Brandenburg, C., Matzarakis, A., Arnberger, A., 2004. The effects of weather on frequencies of use by commuting and recreation bicyclists. Adv. Tourism

Cools, M., Moons, E., Creemers, L., Wets, G., 2010. Changes in travel behavior in response to weather conditions: do type of weather and trip purpose matter?

Guo, Z., Wilson, N., Rahbee, A., 2007. Impact of weather on transit ridership in Chicago, Illinois. Transport. Res. Rec. J. Transport. Res. Board 2034, 3–10.

Heinen, E., van Wee, B., Maat, K., 2010. Commuting by bicycle: an overview of the literature. Transp. Rev. 30 (1), 59–96.

Hofmann, M., O'Mahony, M., 2005. The impact of adverse weather conditions on urban bus performance measures. In: Proceedings of the 8th International

Koetse, M.J., Rietveld, P., 2009. The impact of climate change and weather on transport: an overview of empirical findings. Transport. Res. Part D Transp.

Kyte, M., Khatib, Z., Shannon, P., Kitchener, F., 2001. Effect of weather on free-flow speed. Transport. Res. Rec. J. Transport. Res. Board 1776, 60–68.

Maze, T.H., Agarwal, M., Burchett, G.D., 2006. Whether weather matters to traffic demand, traffic safety, and traffic operations and flow. Transport. Res. Rec. J. Transport. Res. Board 1948, 170–176.

Muller, S., Tscharaktschiew, S., Haase, K., 2008. Travel-to-school mode choice modeling and patterns of school choice in urban areas. J. Transp. Geogr. 16, 342–357.

Nankervis, M., 1999. The effect of weather and climate on bicycle commuting. Transport. Res. Part A Pol. Pract. 33 (6), 417–431.

Stover, V.W., McCormack, E.D., 2012. The impact of weather on bus ridership in Pierce County, Washington. J. Public Transport. 15 (1), 95–110.

Tang, L., Thakuriah, P.V., 2012. Ridership effects of real-time bus information system: a case study in the City of Chicago. Transport. Res. Part C Emerg.

Contribution

Extracting and cleaning MTA Data, Developed Classification based ML Models and evaluation metrics - Ajayrangan Kasturirangan
Extracting and preparing weather data, Developed MTA Regression Models - Akshay Shetty
Extracting and cleaning citibike data and building regression models for the citibike data - Sharvari Deshpande
Extracting Yellow and green taxi data, data analysis, detecting anomalies and clustering patterns - Shantanu Anikhindi
Developed Yellow and Green Taxi Regression Models - Devang Dave