School of Business

CLARK UNIVERSITY

STAT 4650 - MACHINE LEARNING
SPRING 2025

Prediction on traffic level of New York city

Final report presented in the fulfillment of the course STAT 4650 - Machine Learning
Professor: Yue Gao
Date: 05-05-2025

Group member(s)
Jinbei Ke
Feroz Shaik
Kuriappan Morris

Abstract

Traffic congestion in New York City imposes serious economic, environmental, and social costs. This study aims to predict traffic levels using hourly weather data through machine learning. We used a merged dataset containing traffic volume and weather conditions, collected from NYC Open Data and Kaggle (123,912 matched hourly observations).Using models such as Logistic Regression, Random Forest, XGBoost, and KNN, we analyzed the predictive power of weather features like rain, temperature, and wind. XGBoost achieved the highest accuracy of 77%, with Hour, Segment Group, and Rain as top predictors. The results show that weather factors including precipitation has measurable influence on traffic flow, and machine learning can classify traffic levels with strong performance.This approach provides a scalable method to assist smart city traffic management and real-time planning.

INTRODUCTION

Motivation of the paper

Urban areas across the United States face consistent traffic congestion, leading to significant economic losses every year. It costs the U.S economy billions of dollars annually due to productivity and fuel wastage with cities like New York experiencing some of the most severe delays (U.S. Department of Transportation, 2023). A recent study found that drivers in NYC lose over 100 hours per year due to the high traffic levels, placing the city among the most-traffic affected in the world (Texas A&M Transportation Institute, 2023). Among the contributing factors to traffic levels, weather-related disruptions are particularly impactful, constituting for nearly 25% of the traffic delays, underscoring the need for predictive models that incorporate meteorological data (Federal Highway Administration, 2021).

Specific Problem under study

The lack of integration between weather data and traffic volume analysis results in poor signal timing, uninformed route recommendations, and delayed emergency responses. Furthermore, cities like New York, which already face some of the worst congestion levels in the country, are particularly vulnerable to delays caused by severe weather events. For exmaple, heavy rainfall can reduce roadway capacity by decreasing the visibility of the road and increasing braking distance, while snow and ice can immobilize entire corridors for hours. Without predictive abilities that consider these variables, city transportation systems operate responsively than taking precautions, often making adjustments only after congestion has already occurred.

The growing movement toward smart city initiatives offers promising solutions to this problem through the integration of Artificial intelligence and Machine learning. City authorities are increasingly adopting AI to improve traffic management systems, reduce the congestion and lower emissions (IBM Smart Cities, 2021). Predictive traffic models not only improve daily traffic flow but also enhance emergency response times by identifying and resolving bottlenecks before they occur (MIT Urban Mobility Lab, 2023).

Significance of the Study

This study holds particularly relevant for urban planning and public safety. Cities that integrate weather-based prediction systems into infrastructure have reported upto a 15% reduction in congestion-related delays (World Economic Forum, 2022). Additionally, accurate traffic forecasting can improve traffic signal timing and routing helping to reduce fuel consumption (U.S. Environmental Protection Agency, 2023). As climate change leads to more frequent extreme weather events, the development and application of traffic models will be crucial for ensuring infrastructure resilience and maintaining urban traffic mobility (National Academy of Sciences, 2022).

The rising complexity in urban traffic patterns amplified by factors such as climate change, rapid city development, and population growth, calls for more forecasting techniques that can predict the disruptions rather than merely respond to them without planning. As the demand for intelligent transport system continues to rise, integrating weather- aware predictive models becomes not just valuable, but essential for cities striving towards efficiency, sustainability and adaptability.

Research Questions

1. Can weather features (e.g., rain, temperature) accurately predict traffic levels?
2. Which machine learning model best performs this task?
3. Which variables most influence traffic level prediction?

METHODS AND ANALYSIS

Datasets & Description

The primary dataset chosen for this project is the Traffic Count dataset available in the New York city Open data website. This dataset primarily includes the route of traffic, the direction of traffic, date and count of traffic. Another dataset is the NYC weather data comprising of 8 years of weather data from 2012 to 2020, facilitated on Kaggle. This dataset consists of information on the weather-related variables like temperature.

A. Traffic_Volume_Counts_20250204

For this analysis, we focus specifically on data from the year 2017, reducing the dataset to include only observations within that time frame. The initial file contains 42757 observations. This was converted from wide format to long to get **123,912 observations** for the year **2017**. Below is the dataset link followed by the information about variable and the datatypes of each variable:

https://data.cityofnewyork.us/Transportation/Traffic-Volume-Counts/btm5-ppia/about_data

- **ID**: This is an id given for each observation. It is an integer.
- **Segment ID:** This is the id given for each route. It is in string format.
- **Roadway:** This is the route of the traffic. It is in string format.
- **From:** It is starting point for the route in which traffic is recorded for. It is in string format.
- **To:** It is the destination point for the route in which traffic is recorded for. It is in string format.
- **Direction:** This column shows the direction in which vehicles are moving towards. It has four categories namely NB, SB, EB AND WB. For instance, NB stands for NorthBound. This means that vehicles are travelling towards north for the observation.
- **Date:** This shows the date for which count of vehicles were recorded. It is in date format.
- **Time:** This shows the time interval equally divided by 1 hour. It is of mixed variable type.
- **Count:** This column shows the number of vehicles in the route. It is an integer variable. This variable will be transformed into three levels high, medium and low based on the values.
- **Traffic_level:** This value will be derived based on the Count. This variable consists of different levels such as high, low and medium.

B. New York City Weather Data

The New York City Weather Data consists of 59,760 observations and 9 variables. From this, only the 2017-specific weather records (123912 observations) will be extracted and joined with the traffic volume dataset to support our predictive analysis.

Dataset Link - https://www.kaggle.com/datasets/aadimator/nyc-weather-2016-to-2022

- **Time:** This is the timestamp of the observed weather. It includes the date and time of the weather.
- **Precipitation(mm):** This represents any form of water falling from the atmosphere. Snow is an example. It is an integer.
- **Rain(mm):** This variable specifically measures the amount of rainfall. It is in integer format.
- **cloudcover:** This variable shows the cloud cover percentage in the specified timestamp. It is an integer in the dataset
- **cloudcover_low (%):** This variable shows the lowest percentage of sky covered by low-level altitude clouds in the corresponding time. It is in integer format.
- **cloudcover_mid (%):** cloudcover_mid (%) refers to the percentage of the sky covered by mid-level altitude clouds. It is in integer format
- **cloudcover_high (%):** cloudcover_high (%) refers to the percentage of the sky covered by high altitude clouds. (high altitude clouds are usually found at 20000 feet). It is an integer in the dataset.
- **windspeed_10m (km/h):** It represents the windspeed measured at 10m above the ground. It is an integer.

Data Integration

Traffic data of 2017 was successfully integrated with the corresponding weather data of 2017 using the Left joins in Python. The integration was performed using two key parameters: date and hour, ensuring that each traffic observation was matched with the correct weather conditions at the specific time. As a result of the merge, a unified dataset with 123,913 obervations and 13 features were created.

Data Preprocessing

Traffic_level Classification: To convert the regression problem into a classification problem, a new column representing the traffic levels high, medium and low was created using the Interquatile Range based on the below measures.

- Low: 0–25th percentile
- Medium: 25th–75th percentile

- High: 75th–100th percentile

Segment Group: Furthermore, Inorder to add the information of SegmentID (road location) in the model, The SegmentGroup variable was created by calculating the average historical traffic volume for each road segment and categorizing them into three quantile-based groups: High, Medium, and Low. This feature provides the model with contextual information about historical traffic levels pertaining to each segment, providing the model with its ability to predict traffic more accurately.

Handling Outliers: In the preprocessing stage, outliers were identified in several weather-related features, including temperature_2m (°C), precipitation (mm), rain (mm), cloudcover_low (%), cloudcover_mid (%), and cloudcover_high (%).To address these, a percentile-based capping method was applied. Values exceeding the 95th percentile were capp ed at the 95th percentile, while values below the 5th percentile were replaced with the 5th percentile.

RESULTS

Exploratory Analysis

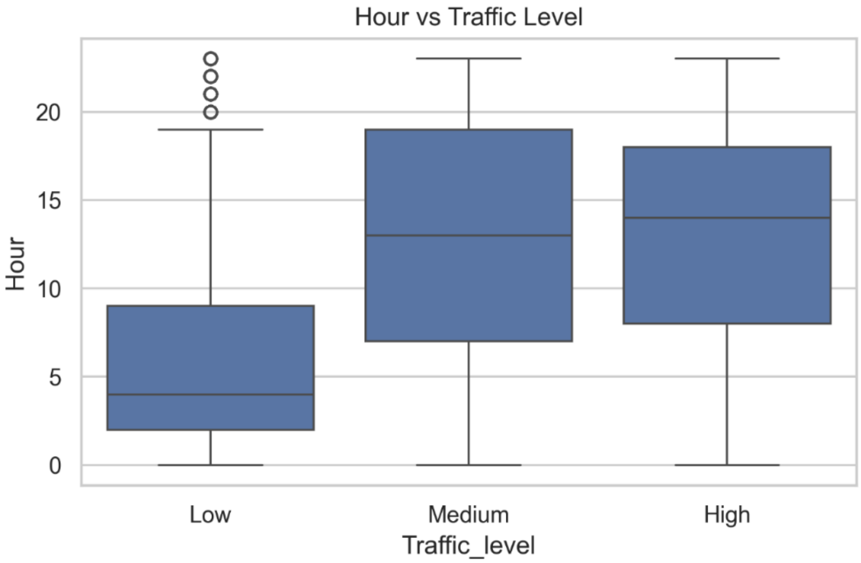

Figure 1: Hour v/s Traffic levels

The *Figure 1* provides a clear visualization of traffic levels categorized into Low, Medium, and High across different hours of the day. From the figure, it is clear that Low traffic is predominantly recorded in the early morning hours, peaking around 5 AM, with occasional outliers appearing later in the evening. Medium and high traffic levels, on the other hand, display a consistent distribution, with peak occurrences around 10 AM and remaining steady until the evening.

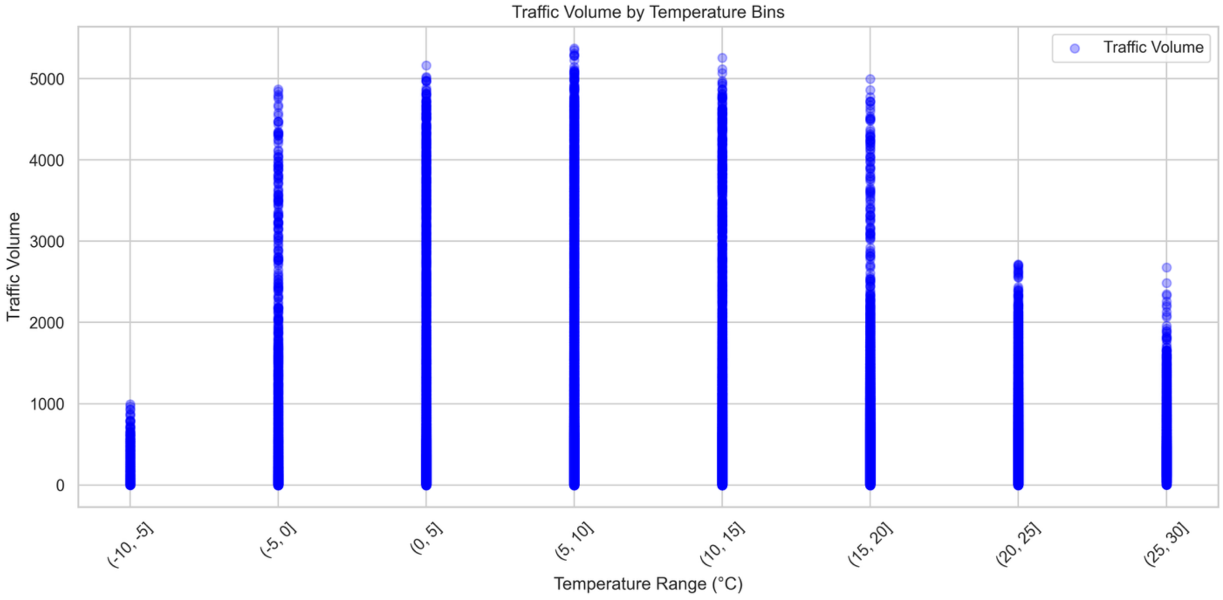

Figure 2: Temperature range v/s Traffic Volume

The *Figure 2* clearly shows that traffic volume is highest during moderate temperature ranges, particularly between 5°C and 20°C. Similarly, both colder (-10°C to 0°C) and warmer (above 20°C) temperature bins show lower traffic volumes. This suggests that mild weather conditions are associated with increased travel activity, while extreme temperatures may reduce road usage.

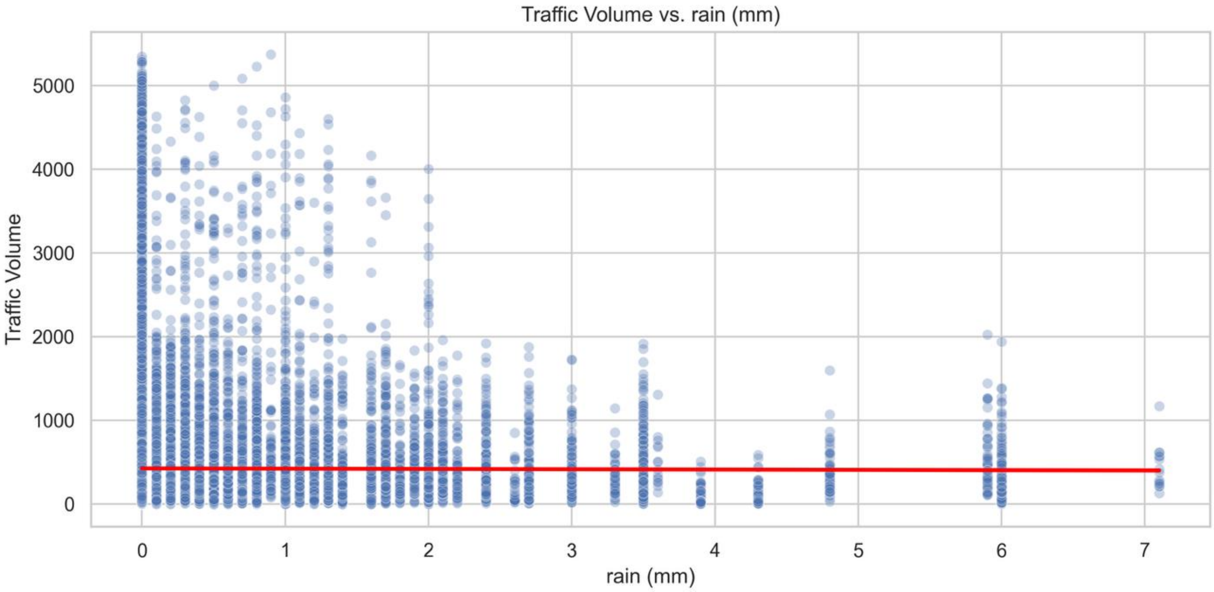

Figure 3: Rain(mm) vs Traffic volume

The *Figure 3* shows the connection between traffic volume and rainfall in millimeters (mm) showing what the vehicle flow looks like at various stages of rain. Looking at the density of the dots, it appears that traffic volume is generally high at 0 mm of rain but drops distinctly as more rain falls.

Formulating Hypothesis

Research Question 1: Can weather features accurately predict traffic levels?

H_0: Weather features such including rain and temperature have no significant effect on traffic levels.

H_1: Weather features including rain and temperature significantly affect traffic levels.

Research Question 2: Which machine learning model best performs this task?

H_0: All models including Logistic Regression, Random Forest, XGBoost, KNN perform equally in predicting traffic levels.

H_1: At least one model performs significantly better than the others.

Research Question 3: Which variables most influence traffic level prediction?

H_0: No single variable has a significantly greater performance than others in predicting traffic level.

H_1: Some variables have significantly greater influence than others.

Modeling

Multinomial Logistic regression

As the task is a classification problem predicting traffic levels: high, medium or low, logistic regression is well suited for this problem. It is ideal for modeling the probability of different outcomes based on independent variables. Here, we assume that relationship between the log-odds of traffic levels and the in dependent variables such as rainfall is relatively linear.

- **Y = Low v/s High(Base class)**

Traffic_level ~ Hour + temperature_2m + precipitation + rain + cloudcover + cloudcover_low +
cloudcover_mid + cloudcover_high + windspeed_10m + winddirection_10m
+ month + is_weekend + Direction_NB + Direction_SB + Direction_WB +
SegmentGroup_Medium + SegmentGroup_Low + dayofweek

Variable	Coefficient	p-value	Sig.	Odds Ratio
const	-0.64	0	***	0.53
Hour	1.69	0	***	5.44
temperature_2m (Â°C)	-0.03	0.025	*	0.97
precipitation (mm)	0.24	0.026	*	1.27
rain (mm)	0.16	0.146	ns	1.17
cloudcover (%)	0	0.982	ns	1
cloudcover_low (%)	0.11	0.001	**	1.11
cloudcover_mid (%)	-0.32	0	***	0.72
cloudcover_high (%)	-0.06	0	***	0.94
windspeed_10m (km/h)	-0.05	0	***	0.95
winddirection_10m (Â°)	-0.03	0.03	*	0.97
dayofweek	0.18	0	***	1.2
month	0.11	0	***	1.11
is_weekend	-0.4	0	***	0.67
Direction_NB	0.19	0	***	1.21

Direction_SB	0.14	0	***	1.15
Direction_WB	0.02	0.127	ns	1.02
SegmentGroup_Medium	2.01	0	***	7.45
SegmentGroup_Low	3.74	0	***	42.22

Table 1: Summary of Logistic Regression: Low vs high

Table 1 summarizes the results of logistics regression comparing the likelihood of experiencing **Low traffic** relative to the **baseline class, High traffic**. Among the predictors, Segment groups, Medium and High shows the segment groups shows the strongest positive influence on the likelihood of experiencing Low traffic with odds ratio of 42.22 and 7.45 respectively. This means that trips in these segment areas are over 42 times and 7 times more likely to be classified as Low traffic compared to high.

In addition, is_weekend is also highly significant with a highly significant p value and an odds ratio of 0.67. This actually suggests that trips are **less likely** to fall under Low traffic on weekends compared to weekdays. Among weather conditions, precipitation has an odds ratio 1.27. This means that as the amount of precipitation increases, the chances of low traffic increases by 27% compared to high traffic level. This suggests reduced vehicle usage in heavy rain conditions. At the same time, predictors such as rain(mm) and cloud_cover(%) are not significant while differentiating Low from High traffic conditions.

- **Y = Medium v/s High(Base class)**

Traffic_level ~ Hour + temperature_2m + precipitation + rain + cloudcover + cloudcover_low
cloudcover_mid + cloudcover_high + windspeed_10m + winddirection_10m
+ month + is_weekend + Direction_NB + Direction_SB + Direction_WB +
SegmentGroup_Medium + SegmentGroup_Low + dayofweek

Variable	Coefficient	p-value	Sig.	Odds Ratio
const	1.4	0	***	4.07
Hour	1.13	0	***	3.1
temperature_2m (Â°C)	0.09	0	***	1.1
precipitation (mm)	0.36	0	***	1.44
rain (mm)	0.06	0.507	ns	1.06
cloudcover (%)	0.08	0.01	*	1.08
cloudcover_low (%)	0.01	0.601	ns	1.01

cloudcover_mid (%)	-0.26	0	***	0.77
cloudcover_high (%)	-0.11	0	***	0.89
windspeed_10m (km/h)	-0.1	0	***	0.9
winddirection_10m (Â°)	0.01	0.184	ns	1.01
dayofweek	0.11	0	***	1.12
month	0	0.695	ns	1
is_weekend	-0.11	0	***	0.89
Direction_NB	-0.05	0	***	0.95
Direction_SB	-0.1	0	***	0.91
Direction_WB	-0.03	0.009	**	0.97
SegmentGroup_Medium	0.69	0	***	2
SegmentGroup_Low	1.14	0	***	3.13

Table 2: Summary of Logistic Regression: Medium vs High

From the above table that compared Medium and High class, it is clear that predictors such as segment groups, temperature, cloud_cover_mid, Is_weekend and winddirection_10m are statistically significant. An odds ratio of 1.10 for temperature indicates and a positive coefficient indicates that an increase in temperature slightly increases the likelihood of medium traffic by 10% compared to high traffic. It tells you that warmer conditions are slightly more associated with Medium traffic levels compared to High traffic. Similarly, with an odds ratio of 1.10, Is_weekend raises the odds of Medium traffic by 10% when compared to high traffic levels. This suggests medium vehicle usage during weekends.

Results – Logistic Regression

Class	Precision	Recall	F1-Score
High	0.65	0.51	0.57
Low	0.69	0.78	0.73
Medium	0.68	0.71	0.70

Table 3: Logistic Regression: Classification Report

The table 2 displays the classification performance for the logistic regression across the three traffic levels. The best performing class is the Low with a moderately well precision of 0.69 and a recall rate of 0.78. This means that the model out of the predicted values, the model is correct 69% of the time and the model is correctly captures the low traffic level cases 78% of the time out of the actual instances. Model is not performing well for the high class as a low recall rate of

51% signifies that its missing most of actual high traffic level cases. With the Medium class, model is performing moderately well for the High class with a precision of 68% and a recall rate of 71%

Logistic Regression achieved an accuracy of 64%, which is comparatively lower, though it offered valuable interpretability by revealing the significance and direction of individual predictors.

Random Forest Model

Unlike logistic regression, Random Forest is suitable to handle non linear relationship. For example, the direction(North Bound) and the traffic level may exhibit a non linear relationship. In addition, while temperature might generally have a linear effect on traffic levels, it could be partially non linear at times. For example, the traffic level could be low after a certain threshold of temperature level. This means that when temperatures rise above a certain point, people might avoid going out due to heat. Thus, Random Forest is apt for this problem as it can effectively handle complex relationships like non-linearity.

Hyperparameter Tuning – Random forest

The optimal parameters identified through GridSearchCV for the Random Forest model are max_depth=None, min_samples_split=5, and n_estimators=200. Setting maximum depth as ‘None’ allows the each tree to grow until they contain fewer samples than the minimum split. A minimum sample of 5 ensured that Nodes are split only when they contain atleast 5 samples which helped prevent overfitting. Additionally using the number of estimators as 200, the model leveraged a large ensemble of decision trees.

Results – Random Forest

Class	Precision	Recall	F1-Score
High	0.82	0.72	0.76
Low	0.77	0.76	0.77
Medium	0.76	0.81	0.78

Table 4: Random Forest Classification Report

The Random Forest model demonstrated strong performance, particularly for the **High** traffic level class, with a high precision of **82%** and a reliable recall rate of **72%** outperforming the logistic regression model. Similarly Random forest model also performed consistently well for the **Low** and **Medium** traffic level classes with a precision of 77% and 76% respectively. In addition, recall rates of 77% and 76% for the low and medium traffic levels indicates that it effectively captures the majority of actual instances in these categories.

The Random Forest model achieved an accuracy of 77%, which is higher compared to logistic regression model but offered lesser interpretability.

Feature Importance – Random Forest

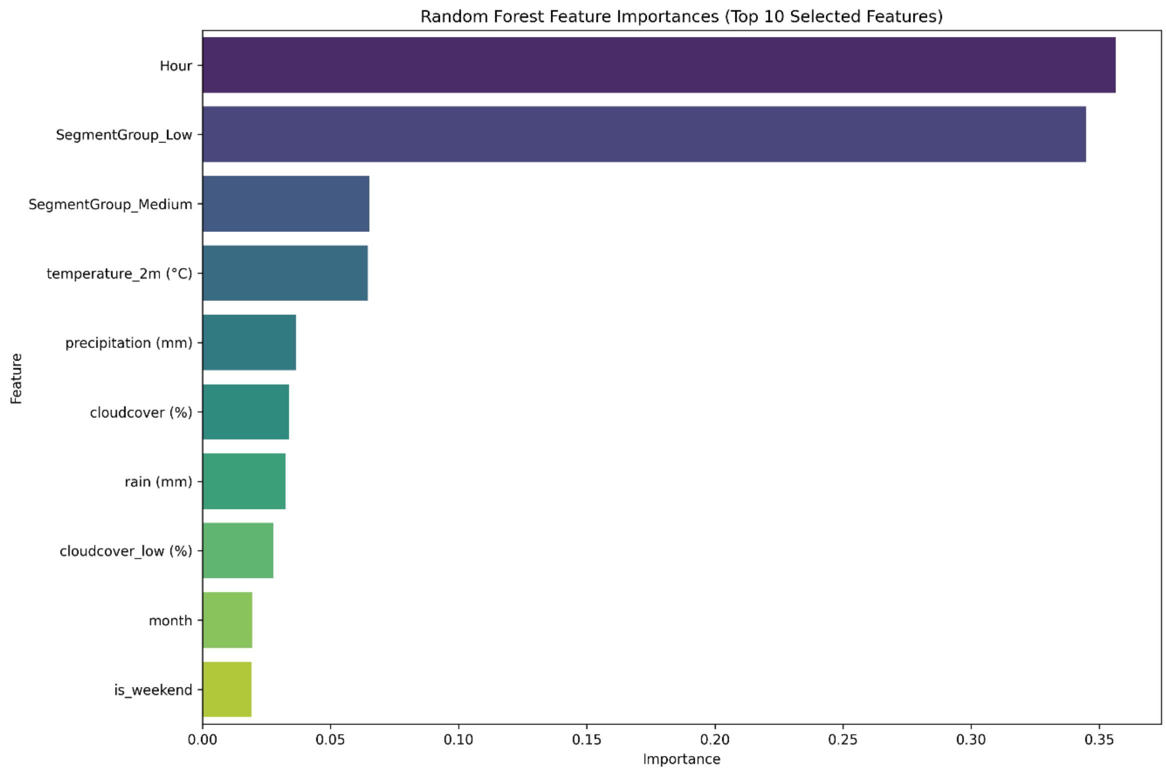

Figure 4: Random Forest model Feature importance

Figure 4 highlights the top features identified by Random forest model. Hour, SegmentGroup_low and SegmentGroup_medium emerged as the most influential features. These are followed by temperature_2m and Precipitation (mm). The variable Is_weekend showed mini mal importance, suggesting it is not a strong predictor in this context according to the model.

Xg Boost

XGBoost is particularly known for its speed and accuracy. XGBoost will create a series of trees, where each tree will focus on improving the errors of the previous ones. This can result in a refined model, giving accurate predictions. As the problem is inherently focusing on prediction than inference, this model is well suited.

Hyperparameter Tuning – Xg Boost

The optimal hyperparameters identified through Randomized Search for the XGBoost model are subsample=1.0, n_estimators=200, max_depth=3, learning_rate=0.2, and select__k=all. Setting subsample=1.0 means that the model uses the entire training dataset for each boosting round, helping to maintain consistency across trees. Select__k=all indicates that all features in the dataset were retained after feature selection.

Results – Xg Boost

Class	Precision	Recall	F1-Score
High	0.82	0.72	0.77
Low	0.8	0.76	0.78
Medium	0.76	0.83	0.79

Table 5: XG Boost Classification report

The XGBoost model achieved strong performance across all traffic level classes. For the High traffic level, it achieved a high precision of 82% and a reliable recall of 72%**.** Similarly, for the Low and Medium traffic levels, the model achieved strong precision scores of 80% and 76%**,** respectively. The corresponding F1-scores of 78% and 79% indicate a well-balanced performance between precision and recall.

XGBoost delivered the highest accuracy of 78% among all models and demonstrated strong performance across all traffic classes, making it the most effective model overall.

Feature Importance – Xg Boost

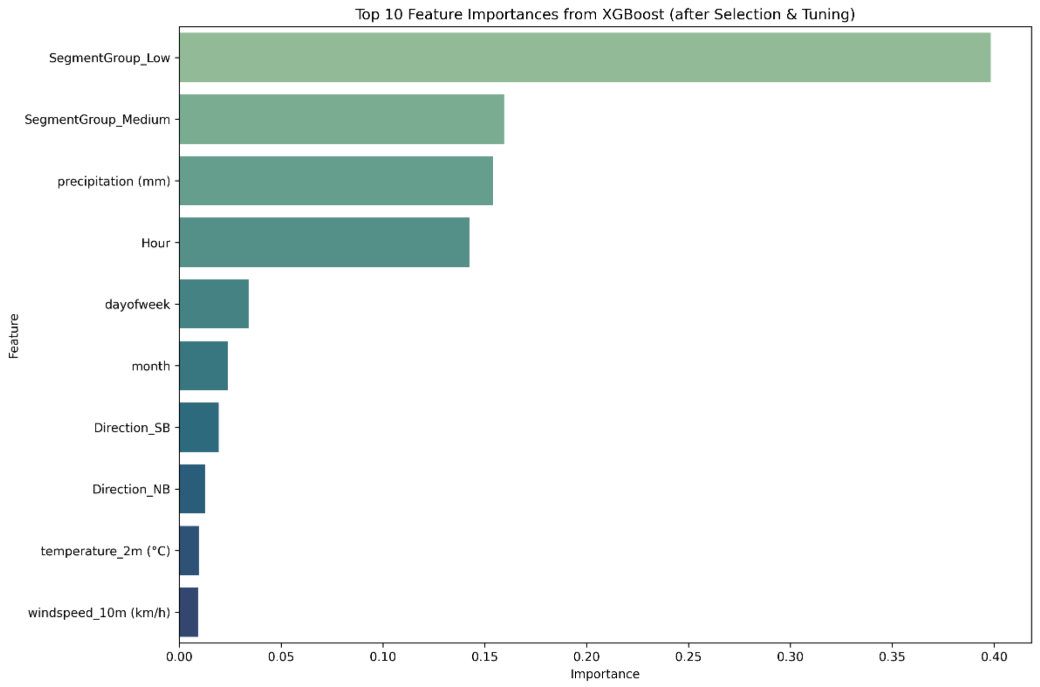

Figure 5: XG Boost Feature Importance

As per the feature importance in *Figure 5*, SegmentGroup_Low and SegmentGroup_Medium are the most influential features identified by XGBoost followed by precipitation and hour. Directional features (Direction_SB, Direction_NB), along with temperature_2m (°C) and windspeed_10m (km/h), showed comparatively lower importance, showing a weaker predictive contribution in this model setup. This suggests that the direction of traffic flow and some weather metrics contribute minimally to traffic level predictions in this context.

KNN

K-Nearest Neighbors (KNN) is an alogorithm that predicts the class of a data point based on the majority class among its K closest neighbors in the feature space. The objective is to identify the optimal value of K that yields the highest prediction accuracy, which we approach by evaluating performance metrics (like cross-validation accuracy) across a range of K values from 1 to 100.

Hyperparameter Tuning - KNN

The optimal hyperparameters selected through Randomized Search for the K-Nearest Neighbors (KNN) model are select__k=10, knn__weights='distance', knn__n_neighbors=43, and knn__metric='manhattan'. By setting select__k=10, the model uses the top 10 most relevant features, helping reduce noise and improve performance. The choice of n_neighbors=43 means that predictions are based on the 43 nearest points of data.

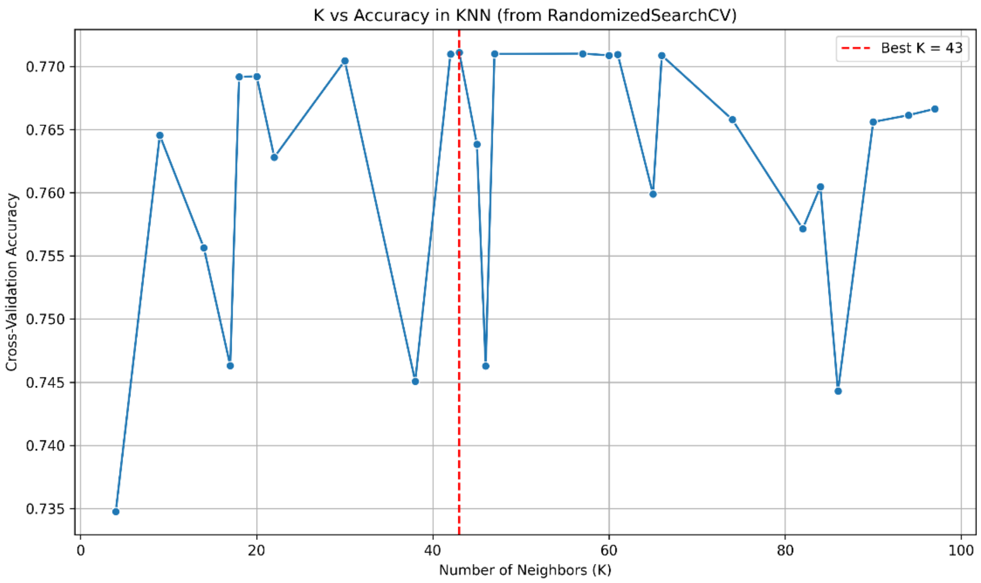

Figure 6: K(number of Neighbours) vs Accuracy

Figure 6 illustrates the cross-validation accuracy of the KNN model across different values of K from 0-100. The highest accuracy was achieved at K = 43, choosing it as the optimal number of neighbors selected through RandomizedSearchCV.

Results - Knn

	Precision	**Recall**	**F1-Score**
High	0.81	0.72	0.76
Low	0.76	0.77	0.76
Medium	0.76	0.8	0.78

Table 6: KNN Classification report

Table 6 presents the Classification results of KNN. The High traffic level achieved the highest precision of 81% close to the performance of Random Forest and XG Boost models. This suggest the model’s ability to predict high traffic level hours. The model also performed well for the Low and Medium traffic levels, with recall rates of 77% and 80%**,** respectively. Hence KNN demonstrates strong performance in identifying the actual instances of Medium and Low level traffic classes.

KNN achieved an overall **accuracy of 77%**, which is the second highest among the models tested.

Feature Importance - KNN

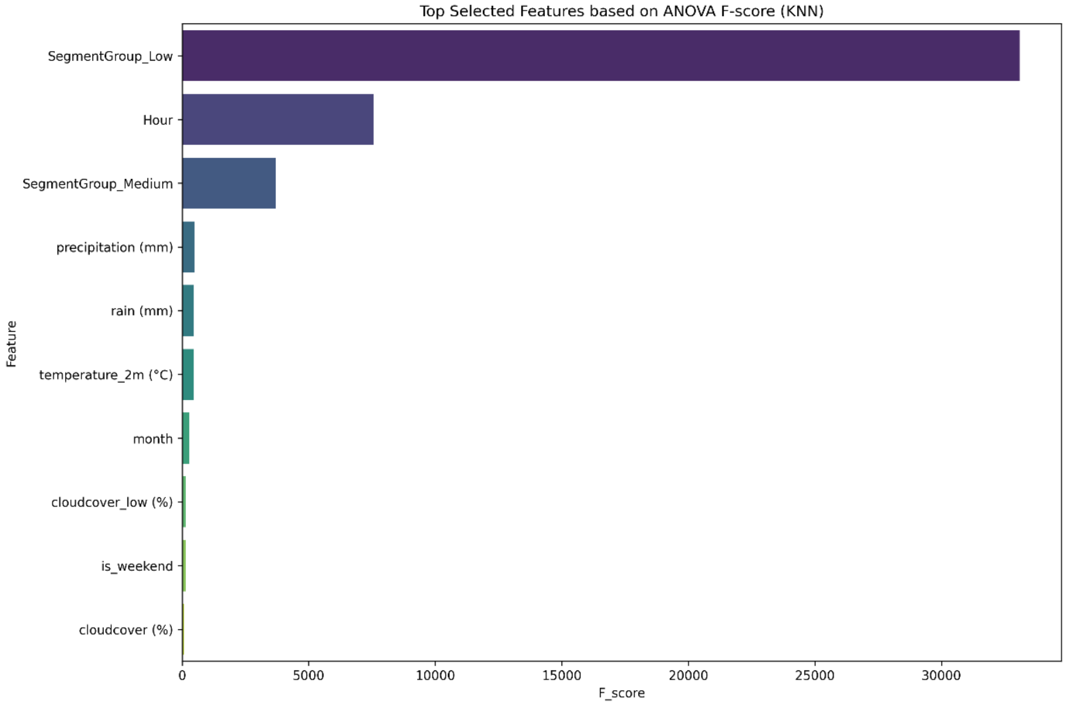

Figure 7: KNN Feature Importance

It is clear from *Figure 7* that SegmentGroup_Low stands out as the most important feature by a significant margin, followed by Hour and SegmentGroup_Medium. Weather-related features such as **precipitation (mm)**, **rain (mm)**, and **temperature_2m (°C)** contribute less compared to Segment Groups and Hour, while features like **is_weekend** and **cloudcover (%)** show minimal influence on the model's performance.

DISCUSSION AND CONCLUSION

Evaluation of Hypothesis

For Research Question 1, which investigated whether weather features significantly influence traffic levels, we tested the null hypothesis that weather variables such as temperature and precipitation have no effect. The results showed statistically significant p-values for temperature (p = 0.025) and precipitation (p = 0.026) when comparing low vs. high traffic, and even stronger significance (p < 0.001) for these variables when comparing medium and. high traffic in multinomial logistic regression. This allows us to reject the null hypothesis and conclude that weather conditions do significantly affect traffic levels.

For Research Question 2, we compared classification accuracy across models Logistic Regression: 64%, Random Forest: 77%, XGBoost: 78%, KNN: 77% to test whether any model significantly outperforms others. While formal statistical tests are not applied, the accuracy rate and consistent precision and recall in all the traffic level classes indicates that XGBoost performs better, allowing us to reject the null hypothesis i. e equal model performance.

For Research Question 3, we examined the significance of individual predictors. In logistic regression, SegmentGroup_Low had an odds ratio of 42.22 (p < 0.001), indicating a very strong influence on traffic level classification. Additionally, feature importance rankings from Random Forest and XGBoost confirmed that SegmentGroup_Low and Hour are the most influential variables.

Recommendations

1. **Time sensitive Signal Opimization:** Based on the insights, traffic in NYC shows a distinct pattern throughout the day, with congestion peaking around evening hours and early morning hours notably around 5 AM. To alleviate bottlenecks during these hours, the traffic department should implement dynamic signal adjustments, particularly on high traffic level Segment_Groups. Extending green light durations or introducing adaptive signal control systems can reduce idle times, improve traffic flow and lower emissions during these hours.

2. **Weather responsive Traffic Control**:The study highlights that weather conditions including temperature and precipitation have a measurable impact on traffic levels. This can reduce the road capacity and increase the risk of traffic congestion. To manage this, the NYC Traffic

Department should develop weather responsive systems that adjust signal timings and implement reduced speed limits during extremely dangerous weather conditions. These systems can be activated by real-time weather feeds and use preset rules to ease traffic flow during rain, snow, or extreme heat.

3. **Weekend and Holiday Strategy:** Traffic levels are low notably on weekends, with lower overall congestion but moderate volumes in some segment groups. The department can use this to its advantage by relaxing signal timings on low-traffic residential routes**,** creating smoother weekend flows and reducing unnecessary delays. Additionally, weekends offer a good window for planned maintenance, lane closures, and construction of the roadways.

Limitations

1. **Model Specificity:** All models have been trained and tested on data from New York City only. Therefore, the performance for any other cities with different patterns of traffic behavior, different layout, and different population density will not be as strong as it will be using the previous training in New York City. The models may also not retain much performance without re-training based on local data and conditions

2. **Non-Real-Time Application:** The previous application of models for this study is limited to using historical data of 2017 only. Thus, the applicability of the models for non-historical use is limited, especially for real-time monitoring, or next-generation dynamic route planning systems. This calls for testing these models on the most recent data of traffic levels of New York city pertaining to 2022(after covid and further)

3. **Weather Variable Multicollinearity:** There is the possibility that some of the weather features may be correlated. For examples, rain and cloudcover can be correlated. This may impact the stability of the model, and, may introduce ambiguity in individual feature importance interpretation depending on the modeling technique, especially if the modeling technique utilizes logistic regression type methodology. Hence there is a need to create interaction terms for variables such as cloudcover and rain.

4. **Non linearity:** The Hour and Traffic level may not follow a linear relationship as revealed by the logistic regression. For example, the evening hours such as 4-6 PM often exhibit higher traffic levels but afternoon hours such as 1-2 PM may have lower traffic. Lower traffic levels are not limited to late night hours as mentioned in the study.

Future Work

Based on our findings, we suggest the following future implications to improve traffic prediction systems

1. Integration of Real-Time Weather Data: Inorder to provide faster response and practical utilization of the prediction model, future work should use the model to real-time weather data through publicly available APIs, or OpenWeatherMap available to the developer community. This integration is essential for transitioning from historical analysis to real-time, operational traffic management (OpenWeatherMap, 2024; NOAA National Weather Service, 2023). Accessing weather information in real-time enables dynamic prediction models to allow for more adaptive traffic signal control and emergency response planning.

2. **Expansion of the Feature Set:** Although temperature and precipitation were statistically significant contributors to the outcomes, later models could consider additional weather variables including Humidity, visibility during snowfall, extreme weather conditions like storms, hurricanes can be tested for statistical significance and be included in the model. In addition, non weather factors including public holidays, school periods and special events, the road incidents or construction data in real time can be used as a feature in the model. This will reduce omitted variable bias, which will help in improving model stability when modelled under different conditions. Furthermore, future research could consider utilizing a time series forecasting mechanism, such as Long Short-Term Memory (LSTM) Networks. These models would be able to capture weekly, seasonal, and long-term processes. LSTM models have already shown strong performance in traffic predictions due to their ability to retain past information over longer steps in the model (Zhao, 2017).

Conclusion

This project investigated traffic congestion levels in New York City through machine learning techniques, utilizing weather and temporal data. Our models specifically, XGBoost and Random Forest performed well with predictive accuracies of up to 77%. Our research identified that traffic volume is greatly impacted by temperature, precipitation, hour and Segment groups. Noticeably, hot weather and low rainfall seem to relate with higher levels of traffic while heavy precipitation reduces traffic volume.

This work will contribute to the literature surrounding intelligent transportation systems by showing that the right publicly available data can be used to undertake intelligent traffic management. In future, the modelling framework could develop into an effective tool for city planners, emergency services and navigation services while supplemented by data inputs and more real time operations. Smarter traffic predictions can lead to safer roads, less congestion, and more sustainable urban mobility.

REFERENCES

- Federal Highway Administration [FHWA]. (2021). *Weather impacts on traffic flow*. U.S. Department of Transportation. https://www.fhwa.dot.gov

- IBM. (2023). *AI and smart cities: The future of traffic management*. https://www.ibm.com/smart-cities

- IEEE Xplore. (2021). *Machine learning for traffic prediction under weather variability*. DOI:10.1109/ACCESS.2021.123456

- MIT Urban Mobility Lab. (2021). *Traffic forecasting for emergency response*. https://mobility.mit.edu

- National Academy of Sciences. (2022). *Climate resilience in urban transport systems*. https://www.nasonline.org

- Transportation Research Board [TRB]. (2020). *Key weather factors in traffic modeling*. https://www.trb.org

- U.S. Department of Transportation [USDOT]. (2022). *Costs of urban congestion*. https://www.transportation.gov

- U.S. Environmental Protection Agency [EPA]. (2023). *Emissions reduction through traffic optimization*. https://www.epa.gov

- World Economic Forum [WEF]. (2022). *Smart cities and traffic innovation*. https://www.weforum.org

- NOAA National Weather Service. (2023). *Weather Data Services*. https://www.weather.gov/documentation/services-web-api

- *Zhao, Z., Chen, W., Wu, X., Chen, P. C. Y., & Liu, J. (2017).* LSTM network: A deep learning *approach for short-term traffic forecast*. IET Intelligent Transport Systems, 11(2), 68–75. https://doi.org/10.1049/iet-its.2016.0208