Traffic Congestion Analysis and Collision Prediction in NYC

Yang Liu

Professor Stanislav Sobolevsky

Applied Data Science

May 1st, 2023

Abstract

This project report gives details about the use of the integrated framework for traffic time series analysis, visualization, and network analysis, coupled with machine learning techniques, using the NYC context. In this case, the researcher studies the traffic condition in NYC by looking at core components that influence traffic such as events, and time. The goal is to investigate how traffic congestion in the city affects the efficiency of travel, businesses, and people's safety. In this study, the research relies on secondary data derived from the NYC DOT Traffic DATA**,** and so on. The integrated framework for the traffic time series analysis shows components that matter in a city on the issue of traffic congestion. Ultimately, the information is useful to city planners who have to make decisions related to infrastructure development and expansion. Besides, this report gives important details to people operating in traffic management. From the report, it is possible to filter out major contributors to traffic congestion in NYC.

Introduction

In New York City (NYC), traffic congestion has been a major concern. Given the adverse effects of traffic congestion, the city's environment, economic, and social life are hindered. The adverse effects on the city's transport system mean slow economic activities, which affects people's livelihoods. People moving from one side of the city to the other have to experience delays and inconveniences associated with the slow system. Since this is an issue that can be solved, understanding its patterns of occurrence is imperative. With an understanding of collision and traffic congestion patterns, a mitigation plan can be developed to help avert the issue. The use of machine learning models has been presented as a potential solution to predicting collision patterns. Variables such as economic growth and population play an instrumental role in shaping the use of machine learning models.

The city has been able to compile a sizable amount of data on traffic patterns and congestion locations through NYC DOT Traffic DATA. Combining this data with machine learning techniques enables the early detection and prediction of probable collisions. Through proactive action to mitigate potential accidents, authorities can lower the risk of injuries and fatalities thanks to this predictive analysis.

Additionally, data-driven traffic congestion analysis can assist city planners in determining which areas need infrastructure upgrades like road enlargements or modifications to traffic flow. These upgrades can assist in reducing traffic congestion, enhance traffic flow, and lessen the detrimental effects of traffic delays on the economy.

Literature review

The majority of earlier research has focused on comprehending the patterns of traffic congestion and collisions in the metropolis. The use of machine learning-based approaches has been an effective tool for predicting such patterns [1]. For instance, Liu et al. [2] proposed a random forest algorithm to construct a traffic congestion status prediction model using input variables such as weather conditions, time periods, special road conditions, road quality, and holidays. The

final result showed that the proposed algorithm achieved a prediction accuracy of 87.5%. Mahmuda Akhtar [1] summarizes three types of AI-based methodologies for real-time traffic congestion predictions, consisting of probabilistic reasoning, shallow machine learning, and deep machine learning. Probabilistic models [3][4], while generally simple, can become complex when various factors that influence traffic congestion, such as weather, social media, and events, are taken into account. Shallow machine learning methods employ traditional ML algorithms for traffic congestion studies, including artificial neural networks [5], regression models [6][7], decision trees [8], and support vector machines [9]. Jiwan et al. [6] constructed a multiple linear regression analysis (MLRA) model by utilizing weather data and traffic congestion data, which underwent preprocessing using Hadoop. Tseng et al. [13] utilized support vector machines (SVM) to estimate travel speed for real-time traffic congestion prediction. In recent years, deep machine learning algorithms have gained popularity due to their ability to access large datasets. Ma et al. [10] and Sun et al. [11]

Research Question: What analytical methods could be employed to gain a better understanding of the correlation between traffic congestion and collisions?

Purpose of Research: To develop more accurate predictions and recommendations that improve public safety and reduce traffic delays by using different analysis techniques

Data

I used these three data 'NYC DOT Traffic DATA, Motor_Vehicle_Collisions, and Traffic Congestion Data'.

To start with, I begin by selecting the appropriate data from NYC DOT Traffic DATA, including Motor_Vehicle_Collisions, Traffic Congestion Data, and Taxi Data. Once the data is downloaded, I remove any missing or incomplete data and check for any unusual or anomalous data points. As part of this process, I also filter out 76 million rows, leaving us with 9 million rows that pertain exclusively to the year 2022.

Methods

1. EDA Analysis

1.1 Data presentation

The data is sourced from traffic flow data of a certain city, and its basic presentation is shown in Table 1.

Table 1. Partial data presentation

DATA_AS_OF	ID	SPEED	TRAVEL_TIME	STATUS	LINK_ID	LINK_POINTS	BOROUGH	LINK_NAME
2022-01-01 00:03:10	345	7.45	567	-101	4620314	40.85526,-73.918591 40.85266,-73.92085 40.8500...	Bronx	MDE S HARLEM RIVER PARK - GWB W AMSTERDAM AVEN...
2022-01-01 00:03:10	325	9.32	569	-101	4329472	40.75829,-73.997531 40.7605,-74.0032 40.762060...	Manhattan	LINCOLN TUNNEL E SOUTH TUBE - NJ - NY
2022-01-01 00:03:10	324	8.69	627	-101	4329473	40.7578106,-73.996801 40.7604506,-74.003221 40...	Manhattan	LINCOLN TUNNEL E CENTER TUBE NJ - NY
2022-01-01 00:03:10	440	50.95	151	0	4329483	40.5264504,-74.27001 40.52568,-74.267851 40.52...	Staten Island	WSE S TYRELLAN AVENUE - 440 S FRANCIS STREET
2022-01-01 00:03:10	329	0.00	0	-101	4329508	40.75766,-73.99687 40.7604,-74.00328 40.76197,...	Manhattan	LINCOLN TUNNEL W CENTER TUBE NY - NJ
...	...	...	...	...	...	...	...	...

From Table 1, it can be seen that the last three columns of the data contain textual information with a small amount of useful data. To simplify modeling, they were directly extracted instead of being converted into one-hot encoding. The fourth column of the data, link_id, is an identifier that has no practical significance for data analysis, and the ID information is already included in the first column. Therefore, link_id was removed and not considered an analytical variable.

1.2 Correlation analysis

To perform correlation analysis on the data after removing redundant features, the Pearson correlation coefficient was used to measure the correlation between the data, and its calculation formula is shown in Equation 1. The heatmap of the analysis results is shown in Figure 1.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left((X-\mu_X)(Y-\mu_Y)\right)}{\sigma_X \sigma_Y}$$

$$\rho_{X,Y} = \frac{N\sum X - \sum X \sum Y}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}} \quad (1)$$

$$\rho_{X,Y} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{\left(\sum X\right)^2}{N}\right)\left(\sum Y^2 - \frac{\left(\sum Y\right)^2}{N}\right)}}$$

The Pearson correlation coefficient (Px,y) between two continuous variables (X, Y) is defined as the covariance cov(X, Y) between them divided by the product of their respective standard deviations (σX,σY). The coefficient always takes a value between -1.0 and 1.0, with variables close to 0 indicating no correlation and variables close to 1 or -1 indicating a strong correlation.

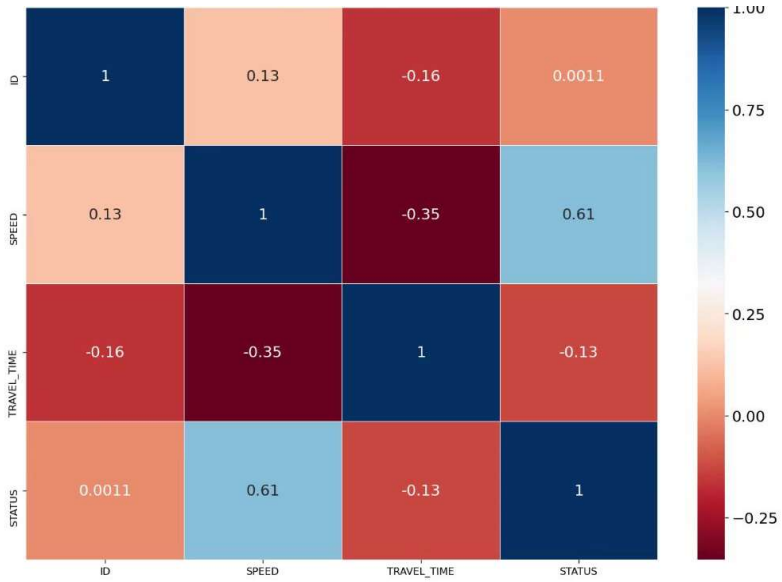

Fig 1. A heatmap based on the Pearson correlation coefficient

The SNS. pair plot method plots a scatter plot between each pair of variables and a histogram for each variable on the diagonal, as shown in Figure 2.

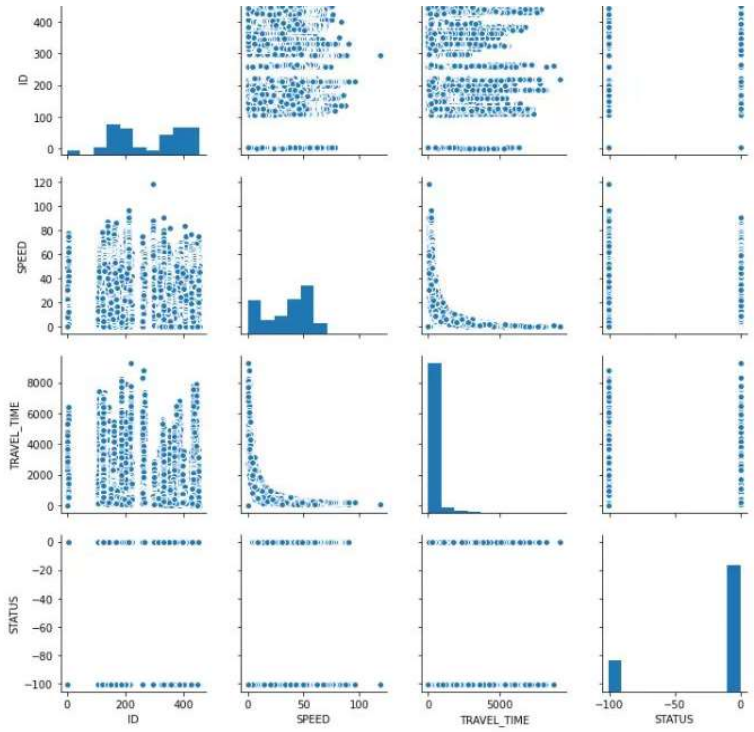

Fig 2. Variable grid plot

1.3 Conditioned Analysis

The primary objective of conducting a conditioned analysis was to establish a comprehensive understanding of the multifaceted relationships between numerous factors in traffic accidents,

such as the number of persons killed or injured, the vehicle types involved, and the contributing factors. This in-depth comprehension of the interplay between these factors can be invaluable for policymakers, urban planners, and transportation engineers who aim to develop targeted interventions, implement safety measures, and improve overall road safety.

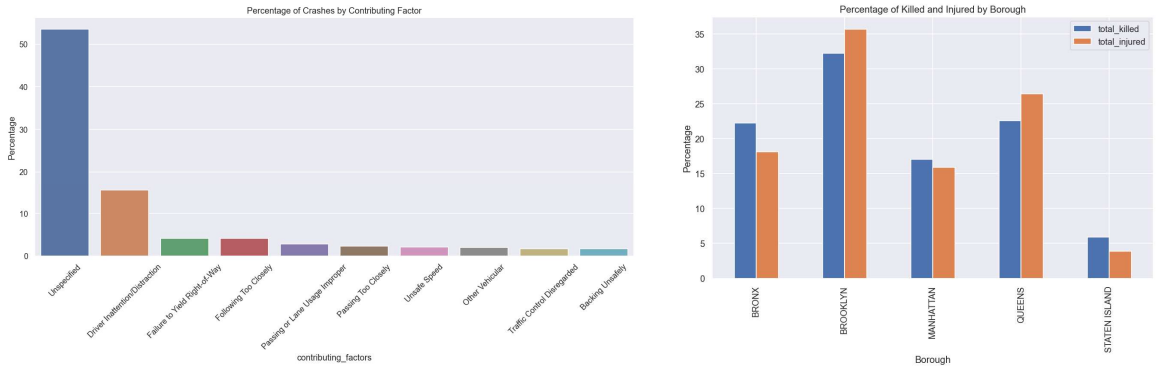

Fig 3. Conditioned Analysis of Contributing Factors

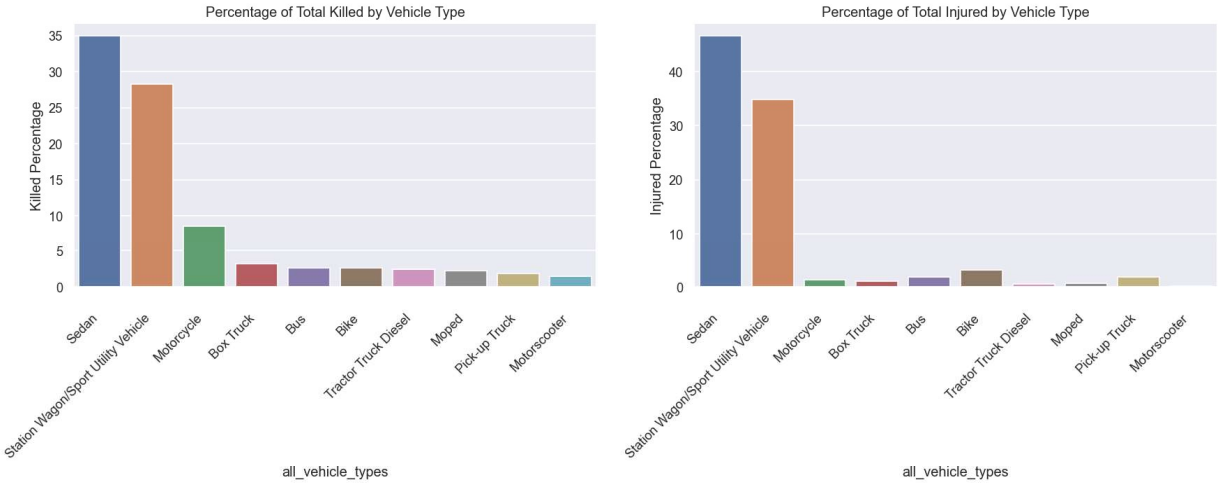

Fig 4. Percentage of Injured and killed By Vehicle Types

Table 2. Percentage of crashes for top 2 contributing factors

Contributing_factors	Crash_count	Percentage
Driver Inattention/Distraction	30417	15.521332
Failure to Yield Right-of-Way	8154	4.160862

Table 3. Percentage of the total killed and injured by top 2 vehicle types

Vehicle_types	Killed_percentage	Injured_percentage
Sedan	34.944238	46.607953
Station Wagon/Sport Utility Vehicle	28.252788	34.889053

My analysis revealed (*Table 2*) that driver inattention accounted for 15 percent of traffic crashes, emphasizing the importance of focusing on driver awareness, education, and training as part of any comprehensive road safety strategy. Moreover, I discovered that sedans were the most common vehicle type involved in accidents, which implies that targeting safety enhancements specifically for sedans could yield substantial benefits.

1.4 Data distribution display

The described method in the Pandas library is a function used for calculating descriptive statistics, which can be applied to data frames (DataFrames) or Series data. It can calculate basic statistical information about the data, including count, mean, standard deviation, minimum, 25%, 50%, 75%, and maximum values. The describe() method provides a quick way to understand the distribution and statistical characteristics of the data. It can help us understand information about the data's distribution, anomalies, and missing values, thereby better understanding the features and issues of the data set. Table 4 shows the basic statistical information obtained using the describe() method, and Figure 5 uses a violin plot to display the distribution of the data.

Table 4. Data distribution display

	ID	SPEED	TRAVEL_TIME	STATUS
count	1.489071e+06	1.489071e+06	1.489071e+06	1.489071e+06
mean	2.802970e+02	3.495245e+01	2.664035e+02	-2.010815e+01
std	1.195744e+02	2.033074e+01	4.895966e+02	4.033096e+01
min	1.000000e+00	0.000000e+00	-2.200000e+01	-1.010000e+02
25%	1.720000e+02	1.739000e+01	8.000000e+01	0.000000e+00
50%	2.980000e+02	4.163000e+01	1.380000e+02	0.000000e+00
75%	3.870000e+02	5.157000e+01	2.350000e+02	0.000000e+00
max	4.530000e+02	1.180600e+02	9.279000e+03	0.000000e+00

Data Distribution

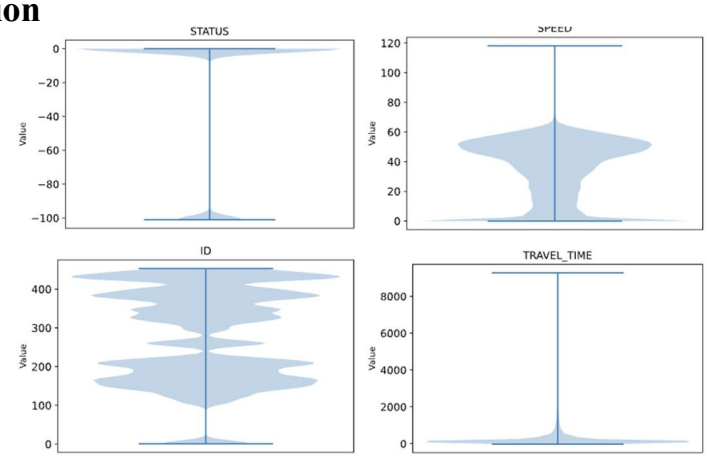

Fig 5. Violin plots of variable distributions

2. Regression model for predicting vehicle speed

I built random forest, support vector regression, and linear regression models to perform regression analysis on the data. The target variable was the vehicle speed, and the other features were used as input for the regression models. First, the data was standardized using the formula shown in Equation (2). The standardized data was then divided into training and validation sets in an 8:2 ratio. The MAE, MSE, and R2, were used as evaluation metrics for the models, and their formulas are shown in Equations (3) to (5).

$$x = \frac{(x - x_{mean})}{x_{std}} \quad (2)$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|\left(\hat{y}_i - y_i\right)\right| \quad (3)$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y}_i)} \quad (5)$$

The evaluation metrics obtained from the three models are shown in Table 5, The comparison chart of predicted results is shown in Figure 6.

Table 5. Evaluation metrics for the model's

model	MAE(km/h)	RMSE(km/h)	R^2
Random Forest	0.3155	1.4019	0.9965
Support Vector Regression	6.744	90.981	0.7790
LinearRegression	219.110	12.13	0.4678

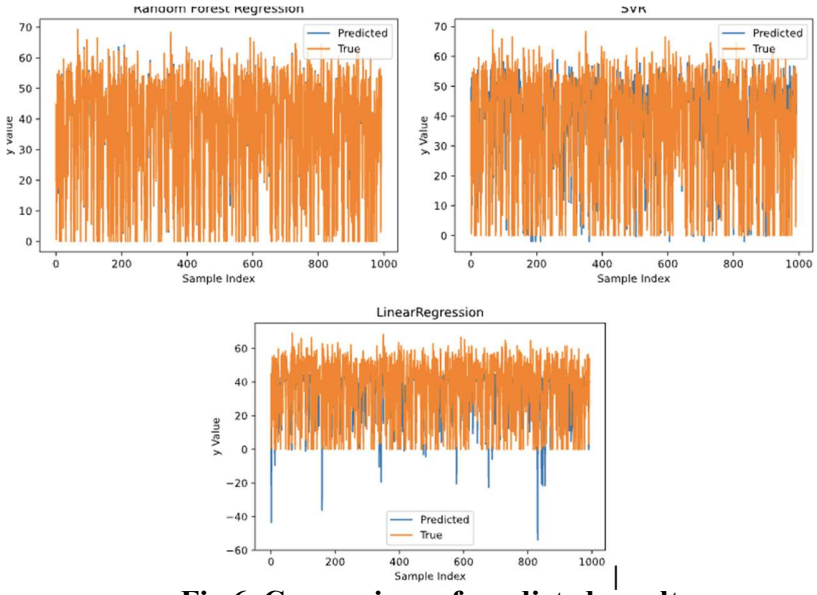

Fig 6. Comparison of predicted results.

Based on Table 5 and Figure 6, it can be seen that the random forest model achieved the best performance in regression prediction.

From this part, The traffic flow data of a certain city was analyzed and processed, and redundant features were removed. The Pearson correlation coefficient was used to obtain the correlation

between variables, with the highest correlation being 0.61 between vehicle speed and status. Regression prediction models for vehicle speed were constructed based on random forest, support vector regression, and linear regression. The models were evaluated using MAE, MSE, and R2. The evaluation results showed that the random forest model had the best fitting effect for vehicle speed, with MAE, MSE, and R2 being 0.3155, 1.4019, and 0.9965, respectively, which were better than the other models.

3. Time Series

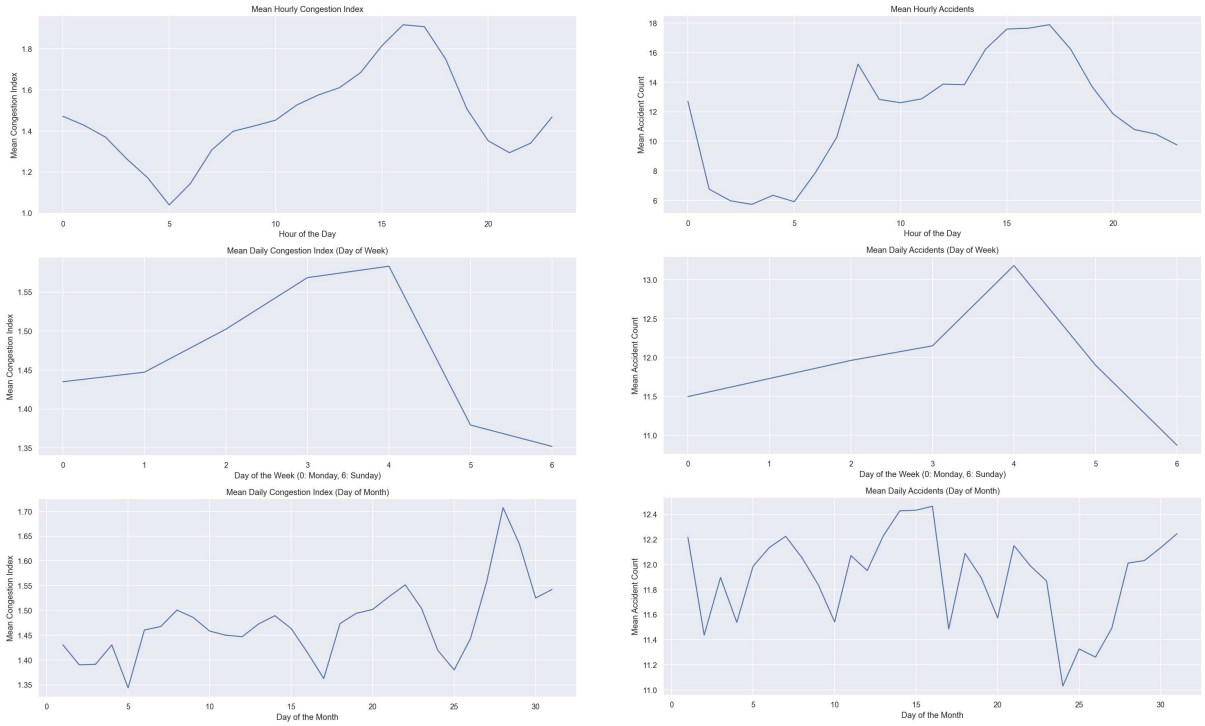

Fig 7. Mean Hourly Congestion Index(Left) and Mean Hourly Crashes(Right)

The congestion index (*figure 6*) analysis aimed to explore the impact of traffic congestion on accident occurrences. By understanding the relationship between traffic congestion and accident rates, I can design strategies to alleviate congestion and subsequently reduce the frequency of accidents, leading to safer and more efficient transportation systems. My findings indicate that the patterns of hourly and daily congestion indexes closely resemble those of mean hourly and daily accidents, suggesting a potential positive correlation between traffic congestion and accident rates. A more in-depth investigation of this relationship could have critical insights for traffic management, infrastructure planning, and congestion mitigation strategies.

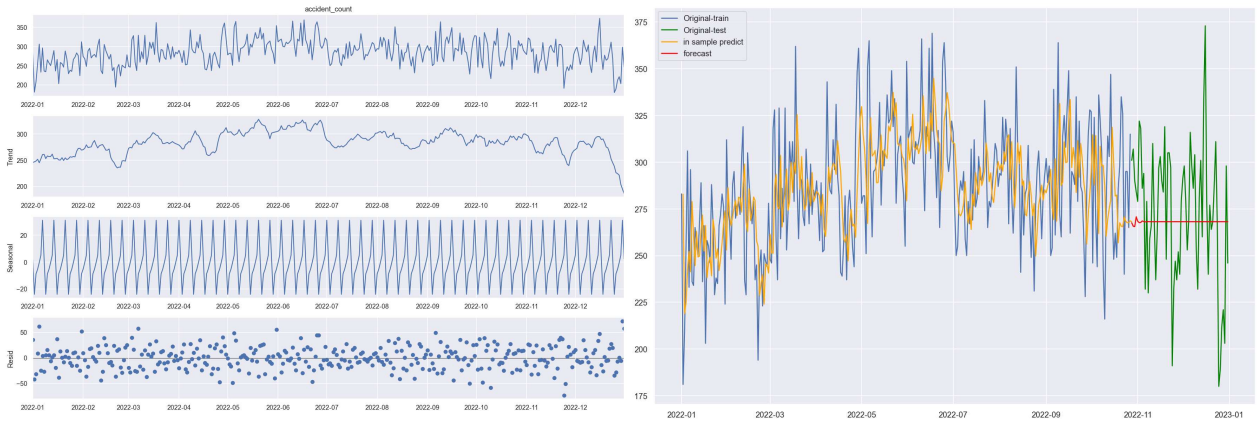

Fig 8. Time Series Decomposition (Left) and ARIMA Prediction (Right)

The time series analysis (*figure 8*) using the Autoregressive Integrated Moving Average (ARIMA) model aimed to predict daily crash counts over time and identify temporal patterns within crash data. I employed an ARIMA model with parameters (P=3, D=1, Q=0). The series displayed a prominent 7-day periodicity, suggesting a weekly cycle in the crash data. Although the extracted trend was not overtly apparent, the residuals exhibited an underlying cyclical pattern. Utilizing the ARIMA model (P=3, D=1, Q=0), I achieved a Mean Absolute Error (MAE) of 29.72 and a Mean Squared Error (MSE) of 1354.28.

4. Visualization

The following visualization depicts car collisions that occurred in New York in 2022. The visualization is divided into two sections: zip codes and taxi zones. The taxi zones are more detailed, providing higher precision in the data. The top 5 zip codes with the most collisions are 11207, 11208, 11236, 11203, and 11368, while the top 5 taxi zones are 76, 61, 37, 216, and 39. The East New York region, which includes zip codes 11207, 11208, and 11236, and taxi zone 76, has a higher probability of car accidents. Moreover, Brooklyn has a higher probability of collisions compared to Manhattan. In 2022, most areas in New York experienced 400-800 car accidents.

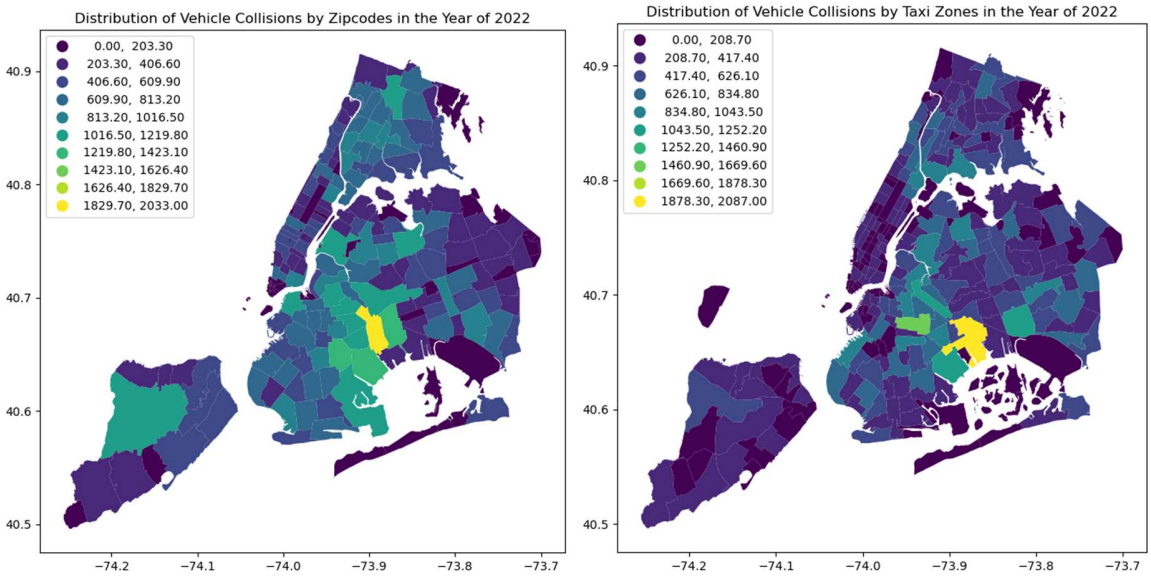

Fig 9. Vehicle collisions count in each zone

5. Network Analysis

The taxi ridership data includes pick-up and drop-off locations which enable us to perform the network analysis. It is helpful for us to study the flow of cars in New York. The analysis identified the top 5 taxi zones based on degree centrality, betweenness centrality, closeness centrality, and PageRank. Zone 137, which is located in Kips Bay, Manhattan, has the highest degree of centrality, indicating that it is well-connected to other zones. Furthermore, it has a high closeness centrality, meaning it is geographically closer to other zones. Zone 92, located in Flushing, has the highest betweenness centrality, indicating that it plays a crucial role in facilitating communication between other parts of the city. Finally, while PageRank measures the influence of each zone, the top 5 taxi zones show an equal level of importance.

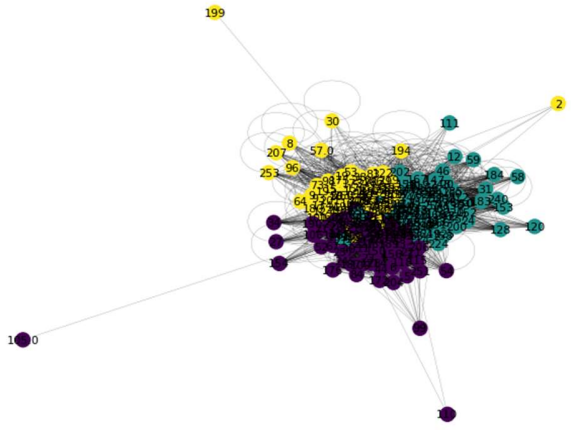

Fig 10. network graph

6. Pipeline

Predicting traffic congestion can help manage traffic flow, enhance public safety, and minimize the detrimental effects of traffic congestion on the economy in NYC. In this part, I will discuss a pipeline for analysis and prediction of traffic congestion using data from the Department of Transportation (DOT). The dataset contains information on travel speeds, travel time, status, and other related variables for different links in the transportation network.

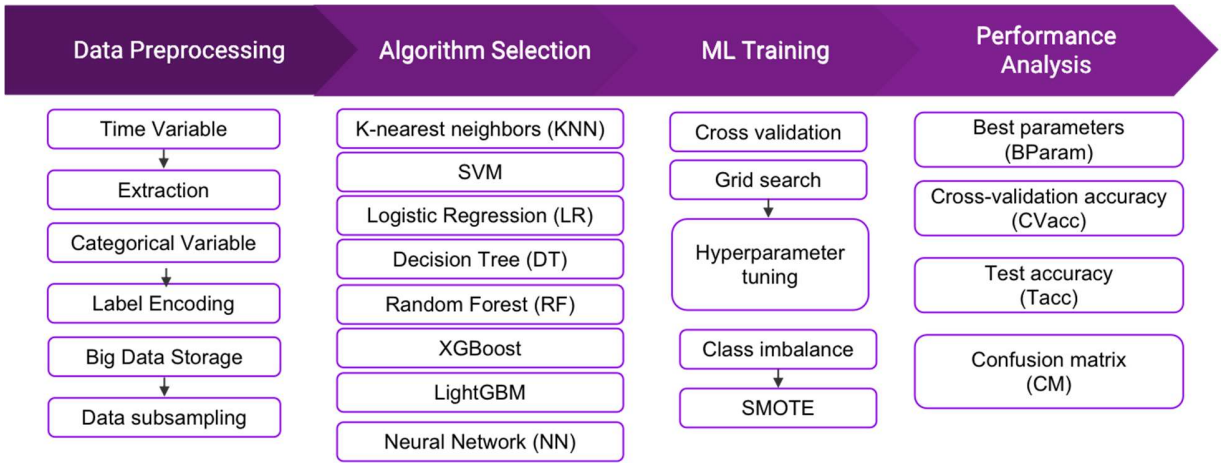

Fig 11. The pipeline of modeling

I begin by reading the original dataset, which is approximately 28.7 GB. Then I convert the "DATA_AS_OF" column to DateTime format and filter the data based on the year 2022. To make the dataset more manageable, I randomly select a small subset of rows (0.02% of the original dataset) for further analysis. I preprocess the data by converting the "DATA_AS_OF" column to DateTime format, encoding the 'STATUS' column, and extracting 'HOUR', 'DAY_OF_WEEK', and 'MONTH' from the DateTime column. I also encode categorical columns using LabelEncoder.

I split the data into training and testing sets, with 80% for training and 20% for testing. To handle a class imbalance in the training data, I use SMOTE. I then train a series of classical ML and DL classifiers using GridSearchCV for hyperparameter tuning and obtain the best parameters, cross-validated accuracy, and test accuracy for these models. The algorithm includes K-nearest neighbors, SVM, logistic Regression, decision tree, random forest, XGBoost, LightGBM, and multi-layer neural network, where random forest, XGBoost, and LightGBM are three tree-based boosting algorithms.

I display the classification report for these models to evaluate their performance. Additionally, I plot a confusion matrix for these models to visualize how well they perform in predicting the 'STATUS' of traffic. In conclusion, the tree-based boosting ML algorithm achieves the best generalization performance in both cross-validation accuracy and test accuracy.

Table 6. Model evaluation for congestion prediction of different ML methods

Algorithm	Best Parameters	Cross-validation accuracy	Test accuracy
K-nearest neighbor	{'n_neighbors': 13}	0.9370	0.9471
SVM	{'C': 1, 'kernel': 'rbf'}	0.9408	0.9521
Logistic regression	{'C': 1}	0.9339	0.9320
Decision tree	{'max_depth': 7, 'min_samples_split': 10}	0.9383	0.9471
Random forests	{'n_estimators': 100}	0.9509	0.9622
XGBoost	{'learning_rate': 0.1}	0.9528	0.9471
LightGBM	{'learning_rate': 0.1}	0.9553	0.9597
Neural networks			0.9043

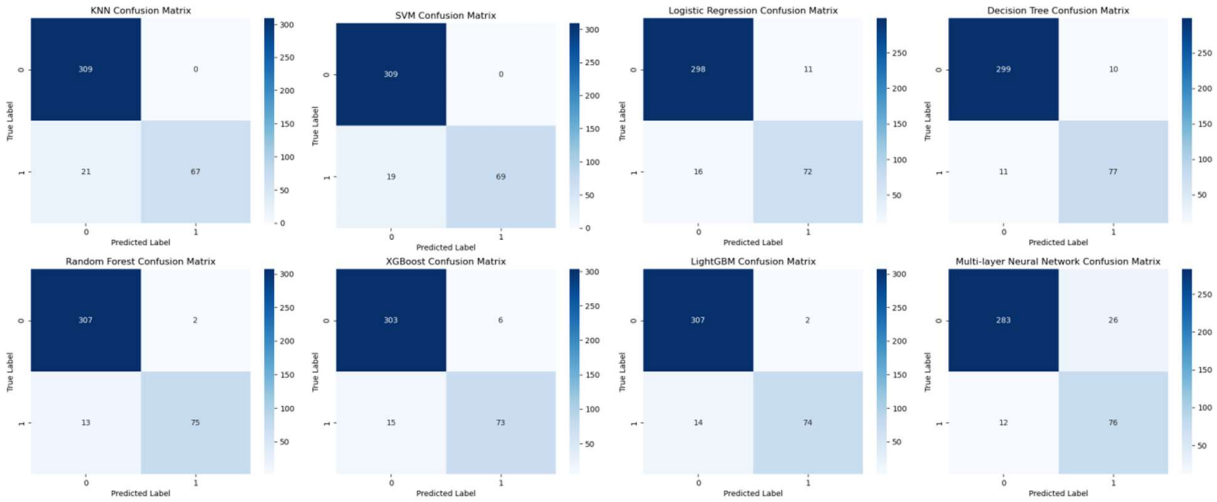

Fig 12. Confusion matrix of algorithms

Throughout this presentation, I have demonstrated how to load, preprocess, visualize, and analyze a large dataset of traffic speeds. I trained a series of classical ML and DL classifiers to predict the 'STATUS' of traffic, achieving a pretty high test accuracy. Further work could include testing other classification algorithms or incorporating additional features to improve the

predictive performance.

Result / Limitations

My analysis of traffic congestion and collision prediction in NYC yielded several important findings. After filtering through the data, I focused on NYC DOT Traffic DATA, Motor_Vehicle_Collisions, and Traffic Congestion Data, as they offered a more comprehensive set of variables. I performed data cleaning, excluding variables related to traffic congestion and the most recent year.

The EDA methodology analysis revealed a comprehensive understanding of the multifaceted relationships among the factors involved in traffic crashes. I identified the highest percentage of casualties by vehicle type, with passenger cars having the highest percentage. Additionally, my findings showed that the East New York area had a higher probability of traffic collisions, with zone 137 having the highest centrality, indicating its strong connections to other zones. The 92nd district in Flushing had the highest median centrality.

I used Visualization to examine the distribution of traffic congestion throughout New York City, with results showing the top 5 zip codes with the highest number of collisions were 11207, 11208, 11236, 11203, and 11368. The top 5 taxi zones were 76, 61, 37, 216, and 39, with taxi zone 76 falling within the East New York area. In addition, I conducted a time series analysis (Figure 8) using an autoregressive integrated moving average (ARIMA) model to predict the daily number of collisions over time and identify any time patterns within the collision data. My pipeline analysis revealed that the tree-based boosting ML algorithm displayed the best generalization performance in terms of cross-validation accuracy and testing accuracy. For research purposes, I utilized the random forest model to provide more accurate predictions and recommendations for improving public safety and reducing traffic delays.

Regarding the limitations, the variables and topics I examined are dynamic and subject to change over time. Therefore, it is necessary to update my analysis frequently in order to evaluate its ongoing applicability and development. My study focused on incidents involving motor vehicles and excluded the investigation of other accidents involving motorcycles, bicycles, or pedestrians. This emphasizes the necessity of my findings and further limits their applicability to other situations.

Conclusions

Analyzing traffic congestion and collision prediction in NYC using techniques such as time series analysis, Correlation Analysis, Conditioned Analysis collision visualization, and network analysis, combined with the power of machine learning, can provide valuable insights for policymakers and city planners to develop more accurate predictions and recommendations that improve public safety and reduce traffic delays.

References

[1] Akhtar, M., & Moridpour, S. (2021). A review of traffic congestion prediction using artificial intelligence. *Journal of Advanced Transportation*, *2021*, 1-18
[2] Liu, Y., & Wu, H. (2017, December). Prediction of road traffic congestion based on random forest. In *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*(Vol. 2, pp. 361-364). IEEE.
[3] J. Zaki, A. Ali-Eldin, S. Hussein, S. Saraya, and F. Areed, "Time aware hybrid hidden Markov models for traffic Congestion prediction," International Journal on Electrical Engineering and Informatics, vol. 11, no. 1, pp. 1–17, 2019.
[4] A. Daissaoui, A. Boulmakoul, and Z. Habbas, "First specifications of urban traffic-congestion forecasting models," in Proceedings of the 27th International Conference on Microelectronics (ICM 2015)., pp. 249–252, Casablanca, Morocco, December 2015.
[5] Y. Xu, L. Shixin, G. Keyan, Q. Tingting, and C. Xiaoya, "Application of data science technologies in the intelligent prediction of traffic Congestion," Journal of Advanced Transportation, 2019.
[6] L. Jiwan, H. Bonghee, L. Kyungmin, and J. Yang-Ja, "A prediction model of traffic congestion using weather data," in Proceedings of the 2015 IEEE International Conference on Data Science and Data Intensive Systems, pp. 81–88, Sydney, NSW, Australia, December 2015.
[7] S. Jain, S. S. Jain, and G. Jain, "Traffic congestion modeling based on origin and destination," Procedia Engineering, vol. 187, pp. 442–450, 2017.
[8] Z. Chen, Y. Jiang, D. Sun, and X. Liu, "Discrimination and prediction of traffic congestion states of urban road network based on spatiotemporal correlation," IEEE Access, vol. 8, pp. 3330–3342, 2020.
[9] F.-H. Tseng, J.-H. Hsueh, C.-W. Tseng, Y.-T. Yang, H.-C. Chao, and L.-D. Chou, "Congestion prediction with big data for real-time highway traffic," IEEE Access, vol. 6, pp. 57311–57323, 2018.
[10] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang, "Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction," Sensors, vol. 17, no. 4, p. 818, 2017.
[11] S. Sun, J. Chen, and J. Sun, "Traffic congestion prediction based on GPS trajectory data," International Journal of Distributed Sensor Networks, vol. 15, no. 5, Article ID 155014771984744, 2019.
[12] R. Ke, W. Li, Z. Cui, and Y. Wang, "Two-stream multichannel Convolutional neural network for multi-lane traffic speed prediction Considering traffic volume impact," Transportation Research Record: Journal of the Transportation Research Board, vol. 2674, no. 4, pp. 459–470, 2020.
[13] W. Zhang, Y. Yu, Y. Qi, F. Shu, and Y. Wang, "Short-term traffic flow prediction based on spatiotemporal analysis and CNN deep learning," Transportmetrica A: Transport Science, vol. 15, no. 2, pp. 1688–1711, 2019.
[14] T. Bogaerts, A. D. Masegosa, J. S. Angarita-Zapata, E. Onieva, and P. Hellinckx, "A graph CNN-LSTM neural network for short and long-term traffic forecasting based on trajectory data," Transportation Research Part C: Emerging Technologies, vol. 112, pp. 62–77, 2020.