Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

Impact of Key Drivers on New York Traffic Flow: A Comprehensive Study

Zhaoxi Chen*

College of Computing, Data Science, and Society, University in Berkeley, California, United States

*Corresponding author: zhaoxichen0310@berkeley.edu

Abstract. This study aims to predict traffic volumes in New York City using integrated datasets from transportation, weather, and collision records. The author employed multiple machine learning models, including Random Forest Regressor, XGBoost, LightGBM, and an Ensemble Model combining Random Forest and LightGBM, to utilize their strengths. The data preprocessing involved cleaning, merging, and encoding categorical features, resulting in a comprehensive dataset that integrated traffic volume, collision, and weather data. The findings indicated that the LightGBM model achieved the highest accuracy with the lowest error rates (MSE: 1336.12, RMSE: 36.55, MAE: 26.35, R^2: 0.9788). While comparing with other studies, the models, particularly LightGBM, provided superior predictive performance. Despite the promising results, the limited dataset size and computational complexity posed challenges. Therefore, in the future, the research should focus on expanding the dataset, exploring advanced ensemble techniques, and evaluating the applicability of the model across different subpopulations to enhance the robustness and universality of traffic volume predictions.

Keywords: Traffic volume prediction; Urban traffic management; Random Forest Regressor; XGBoost; LightGBM.

1. Introduction

The issue of traffic congestion in New York City is so longstanding that New York is ranked as the second most congested in the United States and the third-worst globally [1]. New York owns extensive road network and high vehicle density, which would contribute to frequent congestion, delays, and accidents, and make efficient traffic management and prediction a critical concern for urban planners and city officials. Therefore, it is urgent to study ways to alleviate these challenges and improve traffic flow through advanced predictive models and integrated data analysis. In the present fields of predicting traffic volume, there are three models are largely preferred to be used, which are the Multilayer Perceptron Neural Networks (MLP-NNs) model, LightGBM model, and XGBoost model, to make traffic volume better predicted from the unique strengths of these diverse models. For instance, the MLP-NNs model could well capture complex non-linear patterns and automatically learn features; LightGBM is capable of providing efficient, scalable, and accurate predictions with low memory usage; and XGBoost can offer high performance, robustness, and effective handling of outliers and missing data with support for distributed computing.

Considering the advantages of the three models mentioned above and other models previously studied by the author, the research integrates transportation, weather, and collision data and uses various machine learning models, including Random Forest Regressor, XGBoost, LightGBM, and an Ensemble Model combining Random Forest and LightGBM to predict traffic volumes in New York to optimize public transportation, guide infrastructure improvements and relieve traffic congestion in NYC, with LightGBM performing the best.

2. Methods

2.1. Data

The nature of the data used in this project is multifaceted, encompassing a range of transportation, weather, and accident-collision-related indicators that collectively provide insights into traffic volume trends in New York City. The data comes from reputable sources, including the New York City Department of Transportation (NYC DOT) and the National Weather Service, ensuring the reliability and comprehensiveness of the datasets. The primary dataset, Automated Traffic Volume Counts, includes fields such as 'Borough', 'Yr' (year), 'M' (month), 'D' (day), 'HH' (hour), 'MM' (minute), 'Vol' (volume), and 'street' (street name). With over 1.6 million entries, it offers a robust foundation for analyzing traffic patterns over time. The second dataset, NYC Collisions, details traffic collision factors like the cause of the collision, types of vehicles involved, and the number of injuries or fatalities among pedestrians, cyclists, and motorists to help give valuable insights into the circumstances and impacts of traffic. The third dataset includes crucial meteorological variables such as 'PRCP' (precipitation), 'SNOW' (snowfall), 'TMAX' (maximum temperature), and 'TMIN' (minimum temperature). With around 17,500 entries, this dataset enables itself to analyze how weather conditions influence traffic volumes and collision rates.

2.2. Analytics Models

In this project, the author employed four machine learning models: Random Forest Regressor, XGBoost, LightGBM, and an Ensemble Model. The Random Forest Regressor constructs multiple decision trees to reduce overfitting and provide feature importance insights, though it can be computationally intensive. XGBoost, known for its speed and accuracy, uses gradient boosting to sequentially build trees to effectively reduce bias and variance, but careful hyperparameter tuning. LightGBM, optimized for large datasets, builds trees leaf-wise for better accuracy and efficiency, though it is sensitive to data preprocessing and complex to tune. Finally, the Ensemble Model combines predictions from Random Forest and LightGBM to leverage their strengths to enhance predictive performance and robustness while balancing their respective weaknesses.

2.3. Data Preprocessing

After making a decision to collect data from these three dataset, the author would go on processing the data. Data processing involved several key steps to ensure the datasets were clean, consistent, and ready for analysis. At first, unnecessary columns need to be dropped from each dataset to retain only the most relevant features. The date and time fields of the collision dataset were converted to appropriate datetime formats, and additional temporal attributes (year, month, day, hour, minute) were extracted. In addition, street names were standardized by converting them to uppercase for consistency. This preprocessing ensured uniformity across datasets, facilitating the merging process. The author took advantage of LabelEncoder to convert categorical features such as 'Borough', 'street', 'Contributing Factor', and 'Vehicle Type' to numerical one to make them suitable for machine learning models, and common attributes (year, month, day, hour, minute, borough, street) were used to merge to create a comprehensive dataset. Ultimately, this merging process resulted in a combined dataset with 320 entries, integrating traffic volume, collision, and weather data, and this integrated dataset, referred to as 'merge', includes a rich array of features to provide a comprehensive basis for analyzing the relationship between traffic volumes, collision occurrences, and weather conditions in New York City.

2.4. Experiment Design

The experiment aimed to predict traffic volumes using integrated data from transportation, weather, and collision records. The author prepared the data by cleaning and merging traffic, collision, and weather datasets, then encoded categorical features and extracted relevant time-based features. The author selected four models for evaluation: Random Forest Regressor for its ability to handle large

datasets and feature importance insights, XGBoost for its high accuracy and efficiency, LightGBM for its speed and scalability, and an Ensemble Model combining Random Forest and LightGBM to leverage their strengths. After splitting the data into training and testing sets, the author trained and validated each model using cross-validation and hyperparameter tuning, and evaluated them with metrics like MSE, RMSE, MAE, and R^2.

3. Experiment Result

The evaluation of model performance(see Table 1) across Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) reveals that the LightGBM model is better than the others. The LightGBM model had the lowest MSE (1336.12), RMSE (36.55), and MAE (26.35), along with the highest R^2(0.9788), indicating the highest accuracy and lowest error rates. The Ensemble model closely followed, with an MSE of 1338.84, RMSE of 36.59, MAE of 27.50, and R^2 of 0.9787. After hyperparameter tuning, the tuned Random Forest Regressor Model also performed well, improving from an initial MSE of 1968.65 and R^2 of 0.9687 to 1473.53 and 0.9766. In contrast, the XGBoost model had higher error rates, with an MSE of 2111.70, RMSE of 45.95, an MAE of 31.64, and a lower R^2 of 0.9665. Overall, the LightGBM model showed the best balance of high accuracy and low error across all metrics.

Table 1. The evaluation of model performance

Model	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)	R-squared Score
Random Forest Regressor Model 1	1968.65	44.37	32.69	0.9687
Random Forest Regressor Model 2 (Tuned)	1473.53	38.39	29.88	0.9766
XGBoost model	2111.70	45.95	31.64	0.9665
LightGBM model	1336.12	36.55	26.35	0.9788
Ensemble model	1338.84	36.59	27.50	0.9787

4. Discussion

Random Forest Regressor Model demonstrated robustness to overfitting by constructing multiple decision trees and averaging their results, so as to reduce overfitting and provide useful insights into feature importance. However, this model is computationally intensive, particularly with large datasets, and requires complex hyperparameter tuning to achieve optimal performance. To address these limitations, future work could employ techniques like Random Forest-Recursive Feature Elimination (RF-RFE) algorithm and Long-Short Term Memory (LSTM) network optimized by Bayesian Optimization Algorithm (BOA), whose features are selected from the original variables using the RF-RFE algorithm and used for training the LSTM network whose hyper-parameters are optimized by BOA, to enhance performance of the Random Forest Regressor Model [2].

XGBoost is known for its high accuracy and efficiency due to its gradient-boosting approach, which effectively reduces bias and variance. Nevertheless, it requires careful tuning of numerous hyperparameters and can overfit if not properly regularized. Improvements can be made by implementing an LSTM-XGBoost model, to analyze and address the issues of periodicity, stationarity, and anomalies in time series data, thereby enhancing the effectiveness of traffic flow prediction and achieving efficient traffic guidance and control [3].

LightGBM excels in handling large datasets with efficiency and speed by building trees leaf-wise, which is more memory-efficient than level-wise methods. However, it is sensitive to data preprocessing, requiring meticulous handling of categorical features and missing values, and has a complex hyperparameter tuning process. Therefore, to better predict the traffic volume in NYC, A SARIMA-GRU prediction model is proposed, which accounts for weekly periodicity by employing a strategy that combines the benefits of LightGBM and Gated Recurrent Unit (GRU) to generate features so that enhancing the representation capacity of the limited feature set [4].

The Ensemble Model, which combines Random Forest and LightGBM, showed improved predictive performance and robustness by leveraging the strengths of both models. While this approach reduces variance and enhances accuracy, it increases the complexity of implementation and requires more computational resources. Future research could explore stacking methods that use meta-learners such as the nonnegative lasso, nonnegative adaptive lasso and nonnegative elastic net, nonnegative ridge regression, et al. are suitable meta to combine predictions, which can further enhance performance and robustness, as discussed by Van Loon et al.

Comparing the results with other studies, it observed that the LightGBM model outperformed previous models used for traffic volume prediction. For instance, a study using the Multilayer Perceptron Neural Network (MLP-NN) achieved an R^2 score of 0.93(Navarro-Espinoza, et al., p. 5). In contrast, the LightGBM model achieved an R^2 score of 0.9788, indicating superior performance. This improvement can be attributed to the inclusion of comprehensive features such as weather and collision data. Similarly, another study using LightGBM and XGBoost Model as well reported average R^2 scores of 0.80 and 0.76, respectively [5]. Thus, the models performed better, likely due to extensive hyperparameter tuning and the integration of diverse datasets.

However, the research has certain shortcomings. The dataset used had only 320 entries after merging, which may limit the generalizability and preciseness of the results. For example, the Mean Squared Error (MSE) values of the models the research used are all over 1300, which can seem large. Hence, larger datasets are called for providing more robust insights and improving model performance. Furthermore, the complexity of merging and preprocessing multiple datasets poses challenges that need careful handling to ensure data consistency and accuracy. For future research aimed at better solving the traffic congestion problem in New York, expanding the dataset to include more entries and additional relevant features could enhance the models' predictive power. Investigating other advanced machine learning techniques, such as deep learning models, could also be a bright direction. Additionally, exploring the impact of different subpopulations, such as varying traffic patterns in different boroughs or during different weather conditions, could provide more granular insights and improve the applicability of the models in real-world scenarios.

5. Conclusion

All in all, this study demonstrates the effectiveness of integrating transportation, weather, and collision data to predict traffic volumes in New York City using advanced machine learning models. LightGBM achieved the highest accuracy and lowest error rates, highlighting its capability to handle large, complex datasets. However, each model employed in this study has its own set of advantages and disadvantages. In particular, while Random Forest Regressor provides robust feature importance insights and reduces overfitting through the construction of multiple decision trees, it is computationally intensive and requires complex hyperparameter tuning. XGBoost, known for its high accuracy and efficiency, effectively reduces bias and variance but also demands careful tuning and can overfit if not properly regularized. Despite the promising performance result, some limitations should be improved. The primary limitation is the dataset size, which, after merging, consisted of only 320 entries. This small sample size may affect the generalizability and precision of the results, as indicated by the relatively high Mean Squared Error (MSE) values. Additionally, the complexity of merging and preprocessing multiple datasets presents challenges that require meticulous handling to ensure data consistency and accuracy. Thus, future research should focus on expanding the dataset,

exploring advanced ensemble techniques, and evaluating the models' applicability across different subpopulations to provide more granular insights. Addressing these aspects will contribute to more accurate and reliable traffic volume predictions, aiding urban planners and traffic managers in making informed decisions to enhance traffic flow and reduce congestion in urban environments.

References

[1] Baghestani A, Tayarani M, Allahviranloo M, Gao H O. Evaluating the traffic and emissions impacts of congestion pricing in New York City. Sustainability, 2020, 12(9): 3655.

[2] Cheng W, Li J, Xiao H, Ji L. Combination Predicting Model of Traffic Congestion Index in Weekdays Based on LightGBM-GRU. Scientific Reports, 2022, 12(1).

[3] Laifa H, Khcherif R, Ghezalaa H H B. Train delay prediction in Tunisian railway through LightGBM model. Procedia Computer Science, 2021, 192: 981–990.

[4] Navarro-Espinoza A, López-Bonilla O R, García-Guerrero E E, Tlelo-Cuautle E, López-Mancilla D, Hernández-Mejía C, Inzunza-González E. Traffic flow prediction for smart traffic lights using machine learning algorithms. Technologies, 2022, 10(1): 5.

[5] Shang Q, Feng L, Gao S. A hybrid method for traffic incident detection using random Forest-Recursive feature elimination and long Short-Term memory network with Bayesian optimization algorithm. IEEE Access, 2021, 9: 1219–1232.