MACHINE LEARNING PROJECT
Traffic Flow Volume Prediction

Ashish Telukunta, G31033685

CSCI 6364 MACHINE LEARNING, Prof. DAVID TROTT

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Abstract

Traffic congestion is a pervasive urban issue, and developing an accurate traffic flow volume prediction system is crucial. This project aims to create a robust system using Machine Learning (ML) models, including Linear Regression, SVM, Random Forest Regression, and XGBoost, alongside Neural Network models such as GRU, LSTM, BI-LSTM, and CNN-LSTM architectures. I will consider a comprehensive set of input features, including weather, temperature, holidays, and other relevant factors, coupled with rigorous data preprocessing, feature engineering, and exploratory data analysis.

The Metro Interstate Traffic Volume dataset is utilized for model development, and model performance will be analyzed using metrics like MAE, RMSE, MAPE, and R2. I will also assess the effectiveness of incorporating lagged, cyclic, and rolling mean statistics in the prediction process to inform model selection. Building on the literature, which highlights the potential of ML models and LSTM networks for traffic flow predictions, this project seeks to incorporate additional temporal and environmental factors, refining existing systems and providing more accurate traffic volume predictions, ultimately contributing to the development of efficient and sustainable transportation systems.

This project could potentially guide traffic management strategies and support the development of intelligent transportation systems

Introduction

The proliferation of modern mobile and vehicular communication technology has promoted the development of Intelligent Transportation Systems (ITS), resulting in an exponential increase in the number of vehicles on the road each year [3]. Traditional methods of mitigating traffic congestion involve changing road and urban infrastructures. However, redesigning the city structure to improve traffic patterns and reduce congestion is expensive and time-consuming. Consequently, dynamic route planning, optimizing road allocations, and employing modern technologies to better understand traffic patterns have become crucial tasks for successfully reducing traffic congestion. Forecasting future traffic status based on historical data offers a promising approach to alleviate traffic congestion and optimize traffic distributions.

Various techniques for traffic flow prediction have been explored in recent years, including naive, parametric, and non-parametric models. While traditional parametric methods such as the ARIMA model have been widely used and effective in predicting traffic flow, they often fall short when dealing with irregular traffic patterns [1]. Non-parametric methods, including deep neural networks, have shown superiority in handling traffic forecasting challenges. Recurrent neural networks (RNN), specifically Long Short-Term Memory (LSTM) networks, have demonstrated their advantages in modeling and predicting traffic flow [2].

This project aims to develop a traffic volume prediction system with a focus on LSTMs and Gated Recurrent Units (GRU) to address the limitations of traditional models and further improve the accuracy of traffic flow predictions. By incorporating a comprehensive set of input features, such as weather, temperature, holidays, and other relevant factors, I will evaluate my model on the Metro inter-state traffic dataset. This project seeks to advance the field of traffic flow volume prediction and contribute to the development of more efficient urban transportation systems.

Methodology

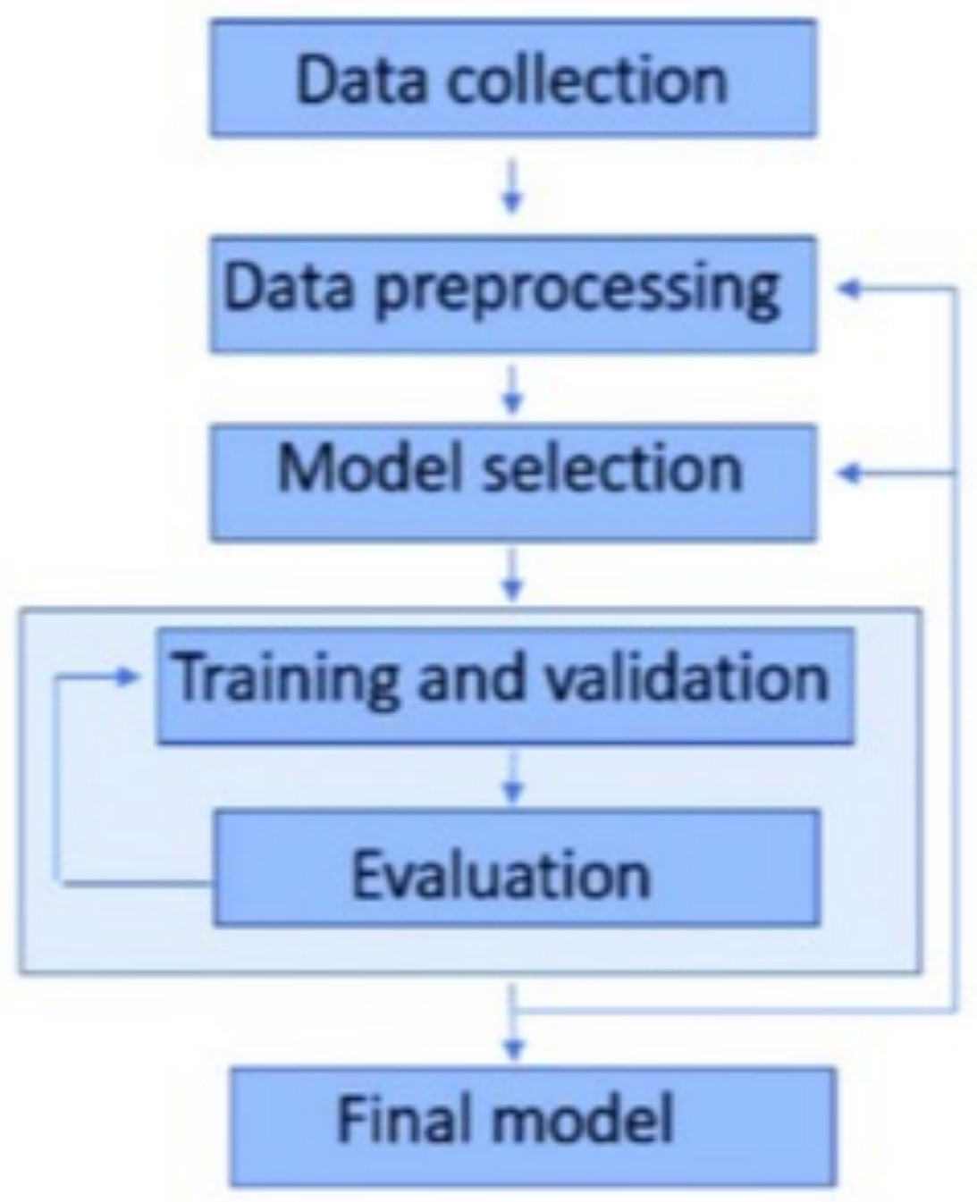
Data Cleaning: To maintain the quality and reliability of the dataset, I first removed duplicates based on the date time feature. This resulted in a cleaner representation of the data and eliminated redundant records. I have also identified and removed outliers present in the 'temp' and 'rain_1h' attributes by applying a cutoff of 1.5 times the Interquartile Range (IQR). This step prevented extreme values from negatively affecting the model's performance. After data cleaning, 7,629 records were removed, leaving a total of 40,575 records for further analysis.

Feature Engineering and Data Augmentation: To enhance dataset and extract valuable insights, I performed feature engineering and data augmentation. I extracted time-related features from the date time column, binarized holiday, rain, and snow attributes, and created categorical features for season, day of the week, and hour. I have also generated 1 to 6-hour lagged features for rain, snow, temp, and cloud cover percentage to provide information on recent trends. Cyclical dependencies for hour, day of a week, and months were encoded using sine and cosine transformation to encode the cyclical nature of the attribute by mapping the values onto a circle, and applied rolling mean statistics to temp, clouds cover percentage, rain, and snow features. Finally, I used mean target encoding to create features for temperature, rain, and cloud cover percentage.

Exploratory Data Analysis: I conducted EDA to gain a better understanding of the impact of various factors on traffic volume. By analyzing the influence of rain, seasonal trends, weekday vs. weekend, yearly variations, and time of day, I was able to identify valuable insights for feature selection and model development.

Feature Selection: Based on the results of EDA and Lasso regression, which penalizes the sum of absolute values of the coefficients, giving more insight into what attributes could me more important. Based on these insights and general assumptions for traffic flow, I have selected a set of features for the model selection and training stage. I included lagged features, rolling mean statistic features, mean target encoded features, and cyclical features while excluding certain features to maintain the model's relevance and accuracy.

Model Training and Evaluation: Trained and evaluated multiple machine learning and neural network models to identify the best performing model for traffic volume prediction. The machine learning models included Linear Regression, Random Forest Regression, XGBoost, and Support Vector Machine (SVM). The neural network models trained were Gated Recurrent Unit (GRU), LSTM with a single layer (LSTM-1L), LSTM with two layers (LSTM-2L), Bidirectional LSTM (BI-LSTM), and Convolutional Neural Network-LSTM (CNN-LSTM). To assess the performance of these models, MAE, RMSE, MAPE, and R2 scores are used.


Data collection

Data preprocessing

Model selection

Training and validation

Evaluation

Final model

Results

I evaluated the quality of predicted traffic flow using the MAE (Mean Absolute Error), RMSE (Root Mean Square Error), MAPE (Mean Absolute Percentage Error), and R^2 scores. MAE measures average absolute errors, RMSE emphasizes larger errors, MAPE expresses average errors as percentages, and R2 score indicates the regression model's goodness of fit compared to a baseline. Based on the data gathered and prediction curves, I have chosen MAPE score to be more representative of the Model Accuracy and hence we have chosen it as a base to evaluate and compare model performance. .

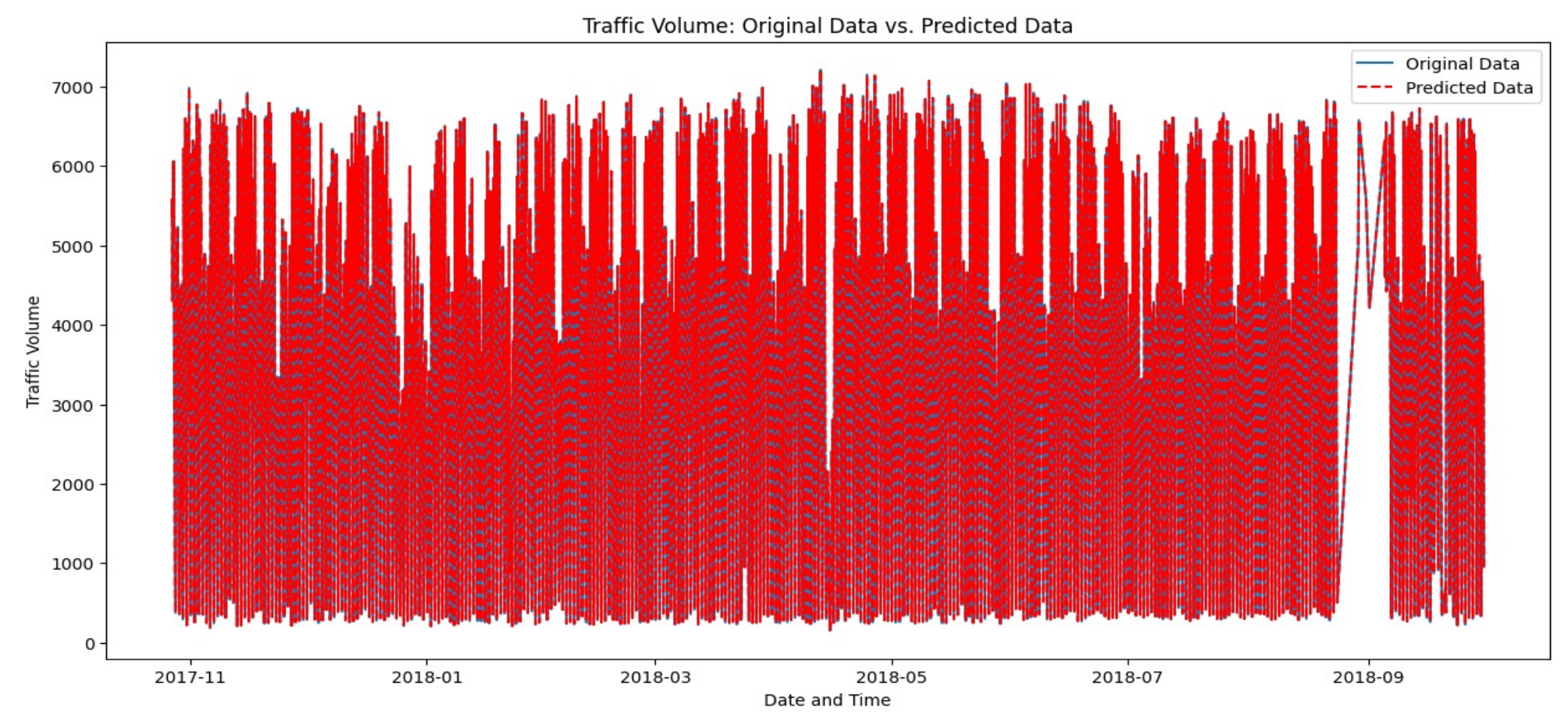
ML Model	MAE	RMSE	MAPE	R2 Score
Linear Regression	0.0562	0.0736	0.2661	0.9262
Random Forest Regression	0.0234	0.0377	0.0763	0.9806
XGBoost	0.0227	0.0362	0.0748	0.9799
Support Vector Machine (SVM)	0.0558	0.0719	0.2303	0.9294

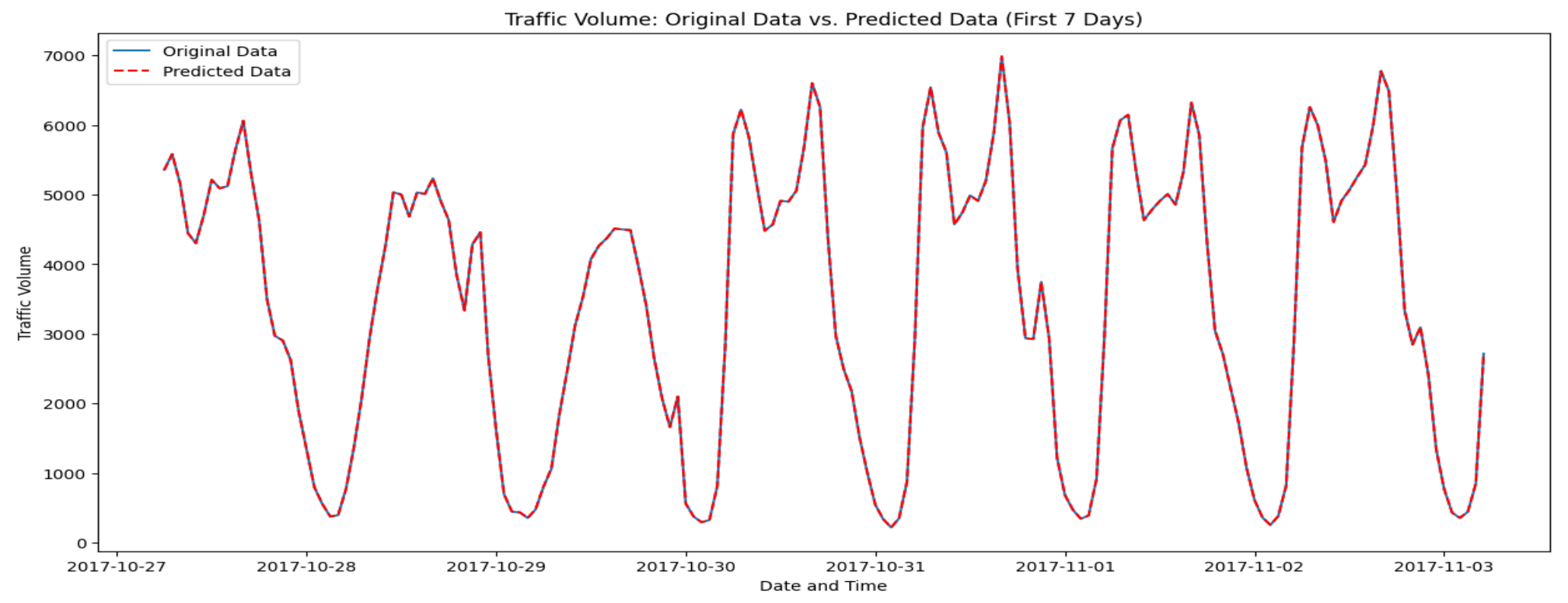
Table 1: Experiment Results from Machine Learning Models

NN Model	MAE	RMSE	MAPE	R2 Score
GRU	4.1370	6.3518	0.0027	1.0
LSTM – 1L	1.7612	2.2150	0.0012	1.0
LSTM – 2L	13.583	15.258	0.008	0.9999
BI-LSTM	4.8124	5.6245	0.0025	1.0
CNN-LSTM	34.61	38.52	0.17	0.0189

Table 2: Experiment Results from Neural Network Architectures

Focusing on MAPE values, XGBoost (0.0748) and Random Forest Regression (0.0763) are the top-performing ML models, while LSTM-1L (0.0012) and GRU (0.0027) show the best performance among NN models. The CNN-LSTM model has a significantly higher MAPE value (0.17), indicating poorer performance. Overall, the results suggest that XGBoost, Random Forest Regression, LSTM-1L, and GRU models are effective choices for traffic flow volume prediction, with LSTM-1L having the lowest MAPE value, making it the best model among the evaluated options.





Conclusion

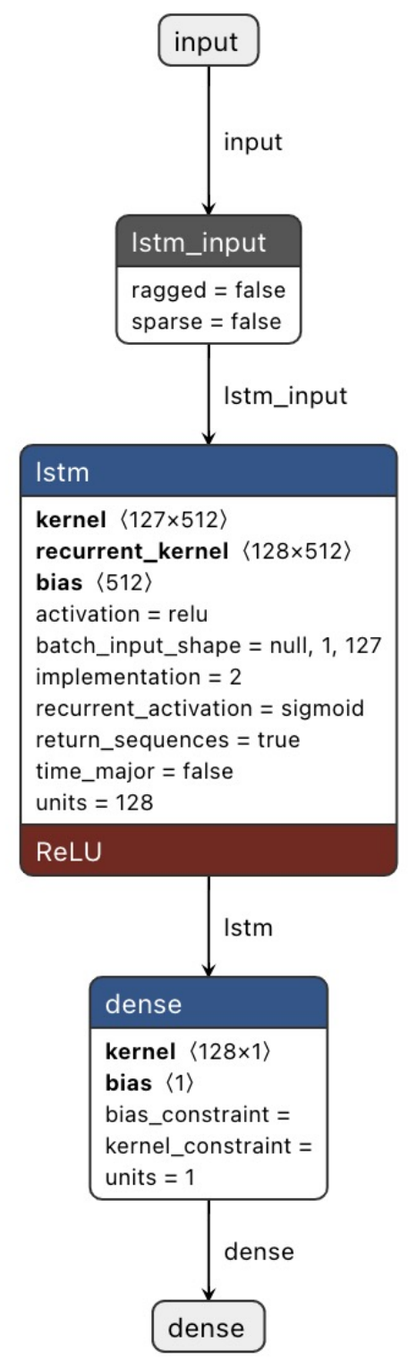
In conclusion, the **single-layer LSTM with a dense layer** provided the best performance across all metrics, closely followed by the BI-LSTM model.

When compared to traditional ML models, both LSTM and BI-LSTM demonstrated superior performance in traffic flow volume prediction. Although BI-LSTM captures both forward and backward dependencies within time series data, the inclusion of lagged, cyclic and rolling mean statistics may have supplied sufficient information, rendering the simpler LSTM model adequate for achieving the desired performance.

A thorough analysis of the data has proven to be instrumental in my case, enabling to select the most suitable neural network models, which outperformed all the tested ML models.

The traffic flow volume prediction system I developed has numerous potential use cases, including optimizing real-time traffic management, informing infrastructure planning, enhancing emergency response, empowering traveler information, enabling sustainable urban planning, and improving traffic simulation and modeling.

Future research could focus on enhancing accuracy through continued advancements in deep learning techniques, exploring multimodal approaches, and incorporating advanced attention mechanisms. Ultimately, the study demonstrates the importance of thorough data analysis and the advantages of using neural network models for traffic flow volume prediction, which can significantly improve urban transportation management and planning.



References

[1] Lokesh Chandra Das. (2023). Traffic Volume Prediction using Memory-Based Recurrent Neural Networks: A comparative analysis of LSTM and GRU.

[2] Fu, R., Zhang, Z., & Li, L. (2016). Using LSTM and GRU neural network methods for traffic flow prediction. In *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)* (pp. 324-328).

[3] Vlahogianni, Eleni & Karlaftis, Matthew & Golias, John. (2014). Short-term traffic forecasting:Where we are and where we're going. Transportation Research Part C: Emerging Technologies.43. 10.1016/j.trc.2014.01.005.