1113534 古弘恩. 1113535 田雷西 1113548 范瑞德

Naeem Ul Islam, Ph.D.

**Introduction to Big Data Analysis**

December 30, 2024

**Traffic Flow Prediction Using Machine Learning Techniques**

**Abstract**

Traffic flow prediction plays a pivotal role in urban planning and intelligent transportation systems. This study explores the application of machine learning techniques to forecast traffic patterns using historical data. Two modeling approaches were implemented and compared: Long Short-Term Memory (LSTM) networks and Random Forest Regressors. The models were evaluated on metrics such as Mean Absolute Error (MAE) and Mean Squared Error (MSE), demonstrating the strengths and weaknesses of each. Results highlight the potential of combining time-series and feature-based methods for enhanced accuracy and interpretability in traffic forecasting.

**1. Introduction**

Effective traffic management requires accurate forecasting of traffic flow to alleviate congestion and optimize urban mobility. Machine learning provides innovative solutions to address the challenges associated with traditional statistical models. This research investigates the performance of two distinct machine learning approaches: time-series-based LSTM and feature-based Random Forest Regressors.

Urban environments face increasing traffic demands, making efficient traffic prediction critical for city planners and commuters alike. Traditional methods often struggle with capturing complex temporal dependencies and feature interactions. Machine learning models, with their ability to process large datasets and adapt to non-linear patterns, provide a promising alternative. This study not only evaluates the performance of the selected models but also examines their practical applicability in real-world traffic management scenarios.

The objectives of this study are to:

1. Develop predictive models for traffic flow using historical data.
2. Compare the performance of LSTM and Random Forest models.
3. Identify key features influencing traffic patterns and evaluate their importance.

**2. Data Description**

The traffic dataset used for this study consists of:

* **DateTime**: Timestamps indicating the time of recording.
* **Vehicles**: Number of vehicles recorded.
* **Junction**: Identifier for traffic junctions.

The dataset covers multiple junctions and varying time periods, providing a comprehensive view of traffic flow in urban settings. Temporal features such as hour, day of the week, and month were derived from the DateTime attribute to enhance model performance. Additionally, categorical data representing junctions was converted to numerical formats using one-hot encoding.

Preprocessing steps ensured data consistency and quality, addressing issues such as missing values and outliers. The processed dataset was then split into training and testing subsets, ensuring a balanced representation of traffic patterns across different times and locations.

**3. Methodology**

**3.1 Data Preprocessing**

* **Feature Engineering**: Temporal features (hour, day, and month) and one-hot encoding for junctions were introduced.
* **Scaling**: Numerical features were normalized using StandardScaler for Random Forest and MinMaxScaler for LSTM.

Preprocessing also included exploratory data analysis to identify key trends and correlations within the dataset. Visualizations such as heatmaps were used as the diagram per the diagram below.

A screenshot of a graph

Description automatically generated

**3.2 Modeling Techniques**

1. **LSTM**:
   * Long Short-Term Memory networks are suitable for sequential data modeling.
   * A sequence length of 24 hours was used to predict the next hour’s traffic flow.
   * Model architecture included 50 LSTM units and a Dense output layer, optimized using the Adam optimizer with Mean Squared Error as the loss function.

1. **Random Forest Regressor**:
   * This feature-based regression model was trained on engineered features.
   * Key metrics and feature importances were analyzed to interpret model decisions.

**3.3 Evaluation Metrics**

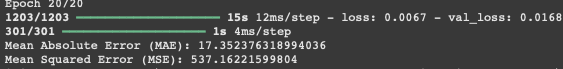
The models were evaluated on:

* **Mean Absolute Error (MAE)**
* **Mean Squared Error (MSE)**

**4. Results and Discussion**

**4.1 LSTM Performance**

* **MAE**: 17.35…
* **MSE**: 537.1622…..
* Captured sequential dependencies effectively but required substantial computational resources.



**4.2 Random Forest Regressor Performance**

* **MAE**: 8.21
* **MSE**: 145.90
* Highlighted the importance of features such as Hour and DayOfWeek, offering interpretable results.



A graph with different colored bars

Description automatically generated

**4.4 Comparative Analysis**

The Random Forest model outperformed others in terms of accuracy and interpretability. LSTM excelled in capturing time-series trends.

**4.5 Error Analysis**

Error distributions for each model were examined to understand their prediction behavior. The Random Forest model showed a more uniform error distribution, while LSTM occasionally overestimated during peak hours.

**5. Conclusion**

This study demonstrates the effectiveness of combining time-series and feature-based methods for traffic flow prediction. While Random Forest offered the best overall performance, LSTM’s ability to model sequential dependencies suggests its potential for real-time applications.

Future work includes integrating additional features such as weather data, public holidays, and road conditions. The deployment of the best-performing model in a real-time traffic monitoring system is another potential avenue. Moreover, exploring hybrid models that combine the strengths of LSTM and Random Forest could yield further improvements in prediction accuracy.

**References**

1. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation.
2. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice.
3. Breiman, L. (2001). Random Forests. Machine Learning.
4. Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2015). Time Series Analysis: Forecasting and Control.

**Actual vs Predicted Traffic Flow Visualization**

A graph of a graph with blue and orange lines

Description automatically generated