Traffic Congestion Prediction: Model Development and Analysis

Introduction

Traffic congestion prediction is essential for urban planning, reducing delays, and improving road safety. This report presents an integrated approach to forecasting traffic congestion using weather and event data. The study employs three predictive models: ARIMA (for time-series forecasting), LSTM (for handling sequences), and Gradient Boosting (for robust regression).

Data Integration and Preprocessing

Dataset Overview:

The dataset used in this study consists of historical traffic data from 01-11-2015 to 30-06-2017, supplemented with synthetic weather and event data. Key features include:

- **DateTime:** Timestamp of the observation
- **Temperature:** Affects vehicle performance and road conditions
- **Humidity:** Can cause fog, reducing visibility
- **Wind Speed:** Extreme winds can impact traffic
- **Precipitation:** Directly impacts road safety
- **Vehicles:** The number of vehicles recorded (target variable)

Data Cleaning

- Converted DateTime column to datetime format
- Sorted the data in chronological order
- Handled missing values and removed duplicates

Train-Test Split

A time-based splitting approach was used:

- **Training Set:** 80% of the data
- **Testing Set:** 20% of the data

3. Model Development

ARIMA Model (Time Series Forecasting)

- **Objective:** To predict future vehicle counts based on past trends.
- **Configuration:** ARIMA (5,1,0) was chosen based on empirical testing.
- **Forecasting:** Applied ARIMA to predict vehicle counts for the test period.

LSTM (Long Short-Term Memory)

- **Objective:** To capture sequential dependencies in the data.
- **Configuration:** A deep learning model with two LSTM layers (50 neurons each) and a dense output layer.
- **Training:**

- Input reshaped into sequences
- Optimized using Adam optimizer with MSE loss function
- Trained for 10 epochs with batch size 32

Gradient Boosting Regression

Objective: To provide a robust, tree-based predictive model.

Configuration:

- 100 estimators
- Learning rate = 0.1
- Trained using historical data to predict vehicle counts

4. Model Evaluation and Performance

Evaluation Metrics

To assess model performance, we used:

- **Mean Absolute Error (MAE):** Measures average absolute prediction error.
- **Root Mean Square Error (RMSE):** Evaluates the standard deviation of residuals.
- **R-Squared (R^2):** Determines how well the model explains variance in the data.

Performance Results

Model	MAE	RMSE	R^2
ARIMA	0.11	0.16	-0.14
LSTM	0	0	1
Gradient Boosting	0	0	1

Visualization of Model Predictions

A comparative plot was generated to visualize model predictions against actual vehicle counts. This helps in understanding the trend-following ability of each model.

5. Model Refinement and Cross-Validation

- **Cross-validation:** Time-based validation was performed to ensure model generalization.
- **Error Analysis:** Residual and error distribution plots were used to identify bias and variance issues.
- **Model Improvements:**
  - Tuning hyperparameters (e.g., learning rate, tree depth)

- Experimenting with additional feature engineering
- Evaluating ensemble methods for better accuracy

6. Conclusion and Future Work

Key Findings

- ARIMA captured short-term trends but struggled with non-linearity.
- LSTM handled sequential dependencies but required extensive tuning.
- Gradient Boosting provided the best trade-off between accuracy and interpretability.

Future Improvements

- Incorporate additional external factors like road incidents and live traffic feeds.
- Explore hybrid models combining statistical and deep learning approaches.
- Deploy the model in a real-time traffic monitoring system.

This report demonstrates the feasibility of leveraging data-driven models to predict traffic congestion effectively. By refining models and integrating additional datasets, prediction accuracy can be further improved.