Model Development, Evaluation & Refinement Report

Traffic Congestion Prediction

-Harini Mukesh

Objective

The objective of this component is to develop robust predictive models to forecast hourly traffic congestion levels at different road junctions. These models leverage historical traffic patterns combined with temporal, weather, and event-related features. Accurate congestion prediction can support traffic management, reduce congestion, and improve decision-making for urban mobility and ride-sharing platforms such as Uber.

Dataset Description

The analysis uses an integrated dataset that combines:

- Hourly traffic volume data at multiple junctions
- Weather variables including temperature, humidity, precipitation, and wind speed
- Event indicators representing public holidays and major public events

The dataset covers the period from **November 2015 to June 2017**, with the number of vehicles per hour as the target variable.

Model Selection

Two regression models were selected for traffic congestion forecasting:

- **Random Forest Regressor (Baseline Model):**
  Chosen for its robustness, ability to capture non-linear relationships, and interpretability using feature importance.
- **Gradient Boosting Regressor (Refinement Model):**
  Implemented to improve performance by sequentially correcting prediction errors from prior models.

These tree-based models were selected due to their stability and effectiveness on structured time-series data.

Training and Validation Strategy

A **time-based train–validation split** was used to preserve the temporal structure of the data:

- Training set: All observations before **January 2017**
- Validation set: Observations from **January 2017 onward**

This approach ensures the evaluation reflects real-world forecasting, where future data is predicted using historical patterns.

Feature Engineering

The following features were used for model training:

- Temporal features: hour of day, day of week, month, weekend indicator
- Weather features: temperature, humidity, precipitation, wind speed
- Event indicator: presence of special events
- Lag features: traffic volume from the previous hour (lag_1h) and previous day (lag_24h)

Lag features were particularly important for capturing short-term and daily traffic dependencies.

Evaluation Metrics

Model performance was evaluated using:

- **Mean Absolute Error (MAE)**
- **Root Mean Square Error (RMSE)**
- **R-squared (R^2)**

These metrics collectively assess prediction accuracy, error magnitude, and explanatory power.

Baseline Model Performance (Random Forest)

The Random Forest model achieved the following performance on the validation dataset:

- **MAE:** 3.56
- **RMSE:** 6.02
- **R^2:** 0.946

These results indicate that the model explains approximately **95% of the variance** in traffic volume, with a low average prediction error.

Cross-Validation Results

Time-series cross-validation was performed using five rolling splits to test model robustness.

- **Cross-validation MAE scores:**
  5.68, 2.71, 2.62, 3.08, 2.75
- **Average CV MAE:** 3.37

The similarity between validation MAE (3.56) and average CV MAE (3.37) confirms that the model generalizes well and does not suffer from overfitting.

Feature Importance Analysis

Feature importance analysis revealed that:

- **Traffic volume from the previous hour (lag_1h)** is the most influential predictor
- **Daily periodicity (lag_24h)** and **hour of day** further contribute to prediction accuracy
- Weather variables have a smaller but non-negligible impact

These findings align with real-world expectations of traffic behavior.

Model Refinement

To enhance prediction accuracy, a **Gradient Boosting Regressor** was trained as a refinement model.

- **Gradient Boosting RMSE: 5.88**

This represents an improvement over the Random Forest baseline (RMSE 6.02), demonstrating that iterative refinement successfully enhanced model performance.

Conclusion

This component successfully developed, evaluated, and refined predictive models for hourly traffic congestion forecasting. The models showed high predictive accuracy, robustness under time-based cross-validation, and improved performance through refinement. The analysis confirms that short-term traffic history and temporal patterns are the strongest drivers of congestion, while weather and events provide secondary influence. These models provide a strong foundation for future traffic forecasting and urban mobility planning.