Traffic Conjestion Prediction – Modeling and Evaluation Report

1. Introduction

In this component, the objective was to build a predictive model to estimate traffic volume at different junctions in Pune city. Accurate traffic prediction is important for ride-hailing platforms like Uber, as it helps in better driver positioning, reducing waiting time, and improving overall operational efficiency.

The focus of this phase was on model building, evaluation, validation, and performance analysis using historical traffic and time-based features.

---

2. Data Splitting Strategy

Since traffic data follows a time-dependent pattern, a **time-based train–test split** was used instead of random splitting. This approach ensures that future data is not used to predict the past, thereby avoiding data leakage and making the evaluation more realistic.

The dataset was divided into:

- **Training set**: Used to train the models
- **Testing set**: Used to evaluate performance on unseen future data

---

3. Models Implemented

Three different regression models were trained and compared:

- **Linear Regression** – used as a baseline model
- **Random Forest Regressor** – to capture non-linear patterns
- **Gradient Boosting Regressor** – for improved predictive performance

Each model was trained using the same feature set to ensure a fair comparison.

---

4. Evaluation Metrics

The models were evaluated using the following metrics:

- **MAE (Mean Absolute Error):** Measures average prediction error in vehicle count
- **RMSE (Root Mean Squared Error):** Penalizes larger prediction errors
- **R^2 Score:** Indicates how well the model explains variance in traffic volume

Higher R^2 and lower MAE/RMSE values indicate better performance.

---

5. Model Performance Comparison

All three models performed well, with R^2 scores close to 1, indicating strong predictive capability.

- **Linear Regression** showed good performance and strong generalization
- **Random Forest** achieved the highest overall accuracy
- **Gradient Boosting** also performed well with stable results

Among them, **Random Forest slightly outperformed the others**, making it the best choice for this problem.

---

6. Cross-Validation Results

To verify that the models were not performing well by chance, **TimeSeriesSplit cross-validation (5 folds)** was applied.

The cross-validation results showed:

- Consistent R^2 scores across folds
- No major performance drops
- Stable behavior across different time periods

This confirms that the models are robust and reliable.

---

7. Overfitting and Underfitting Analysis

Training and testing R^2 scores were compared for all models.

The gap between training and testing performance was minimal in all cases, indicating:

- No overfitting
- No underfitting
- Good generalization to unseen data

All models demonstrated balanced learning behavior.

---

8. Hyperparameter Tuning

The best-performing model, **Random Forest**, was further optimized using **RandomizedSearchCV**.

After tuning:

- Performance changes were minimal
- No significant improvement was observed

This suggests that the original model was already well-tuned and close to optimal.

---

9. Feature Importance Analysis

Feature importance analysis revealed key insights:

- **Lag features (recent traffic history)** were the most influential predictors
- **Time of day** was the second most important factor
- Other features contributed less compared to historical traffic patterns

This confirms that traffic conditions tend to persist over short time horizons.

---

10. Junction-wise Performance Analysis

Model performance was also analyzed separately for each junction.

The results showed:

- Consistent prediction accuracy across junctions
- No major location-specific bias
- Strong adaptability to different traffic conditions

---

11. Final Conclusion

In this project, multiple machine learning models were developed and evaluated to predict traffic volume at Pune junctions. All models showed strong performance, but **Random Forest emerged as the best overall model**.

The most important insight is that **recent traffic conditions are the strongest predictor of near-future congestion**, followed by time-based patterns such as peak hours.

This model can be effectively used by Uber to build a real-time traffic prediction system, enabling better driver allocation **30–60 minutes ahead of demand peaks**, improving efficiency and customer experience.