Traffic Volume Prediction Using Machine Learning with Apache Spark

Reshma Vangala | Big Data Management Systems & Tools | January 14 2026

1. Abstract:

Accurate traffic volume prediction is important for traffic management and city planning. In this project, Apache Spark MLlib was used to build scalable machine learning models to predict traffic volume using historical traffic data, time-based features, weather data, and lag features. Exploratory Data Analysis (EDA) showed that time-based features such as hour of day and day of week affect traffic volume more than weather features.

Two regression models, Random Forest and Gradient-Boosted Trees (GBT), were developed using Spark ML pipelines. Model performance was measured using RMSE and R^2. The initial GBT model performed better than Random Forest with lower error and higher accuracy. To improve results, hyperparameter tuning was applied to the GBT model using 5-fold cross-validation in Spark.

Lag features, including traffic volume from the previous hour and the same hour on the previous day, were added to the tuned GBT model to capture real traffic patterns. The tuned GBT with lag features achieved the best performance with an RMSE of **291.4** and an R^2 of **0.978**, showing very high prediction accuracy, especially during peak traffic hours. Feature importance analysis confirmed that lag and time-based features were the most important predictors.

2. Introduction:

Traffic congestion is a major issue in cities, leading to delays, higher fuel use, and more pollution. By studying traffic patterns and predicting traffic volume, planners can manage roads more effectively.

This project uses the **Metro Interstate Traffic Volume** dataset from the **UCI Machine Learning Repository**, which includes hourly traffic, weather, and holiday data. The goal is to clean and prepare the data and use machine learning to predict traffic volume, showing how data analytics can help solve real-world transportation problems.

2.1 Data Source and Scope:

The dataset used in this project comes from the **UCI Machine Learning Repository**, specifically the Metro Interstate Traffic Volume dataset. It contains 48,204 hourly records of traffic volume on westbound Interstate 94 near Minneapolis Saint Paul, Minnesota, collected from 2012 to 2018. Each row represents a single hour of traffic observation, with associated weather and temporal information.

The dataset includes:

- **Temporal features:** date_time (timestamp), hour of day, day of week, month
- **Traffic-related feature:** traffic_volume (target variable: number of vehicles per hour)
- **Weather features:** temp (temperature), rain_1h, snow_1h, clouds_all, weather_main, weather_description
- **Holiday indicator:** holiday (US national and regional holidays, including Minnesota State Fair)

The dataset provides a comprehensive view of traffic patterns over multiple years, capturing both typical weekday/weekend behavior and seasonal variations.

This project focuses on analyzing the relationships between **time, weather conditions, and holidays** with traffic volume. Using big data techniques and **Apache Spark MLlib**, the goal is to build predictive models for traffic volume and gain insights into the key factors that influence traffic flow.

3. Data Acquisition and Cleaning/Preprocessing :

3.1 Data Acquisition and Initial Load :

The dataset used in this project was obtained from the UCI Machine Learning Repository, specifically the Metro Interstate Traffic Volume dataset. For the analysis, Apache Spark was used in a Databricks environment to handle large-scale data efficiently.

The dataset was loaded into a Spark DataFrame from the Databricks default database using the following command:

```
df = spark.table("default.metro_interstate_traffic_volume")
display(df.limit(10))
```

The display() function shows the first 10 rows, and loading the data into a Spark DataFrame enables efficient processing and analysis.

3.2 Data Cleaning and Preprocessing:

Before analysis and modeling, the dataset was cleaned and preprocessed to ensure data quality and suitability for machine learning tasks. The following steps were performed:

Step 1: Check for Missing Values

- The dataset was examined for missing values, and none were found.

Step 2: Check Schema

- The date_time column was verified as a timestamp type (nullable = true), suitable for extracting temporal features.

Step 3:Feature Extraction from date_time

- Additional features were created to capture temporal patterns:

```
from pyspark.sql.functions import hour, dayofweek, month
df = df.withColumn("hour", hour("date_time"))
df = df.withColumn("day_of_week", dayofweek("date_time"))
df = df.withColumn("month", month("date_time"))
```

Step 4:Create Holiday Indicator

- A new column is_holiday was created:

```
from pyspark.sql.functions import when, col
df = df.withColumn("is_holiday", when(col("holiday") != "None", 1).otherwise(0))
```

- This column has a value of 1 if the day is a holiday, and 0 otherwise.

Step 5: Categorical Feature Handling

- The weather_description column was dropped as it was redundant.
- The weather_main column was mapped to numeric indices for modeling. The mapping assigned **Clouds = 1, Clear = 2, Rain = 3, Drizzle = 4, Mist = 5, Haze = 6, Fog = 7, Thunderstorm = 8, Snow = 9, Squall = 10, and Smoke = 11.** This encoding allows the categorical weather data to be used directly in machine learning models.

These preprocessing steps ensure that all features are in a suitable format for building machine learning models using **Spark MLlib**, capturing both temporal and environmental influences on traffic volume.

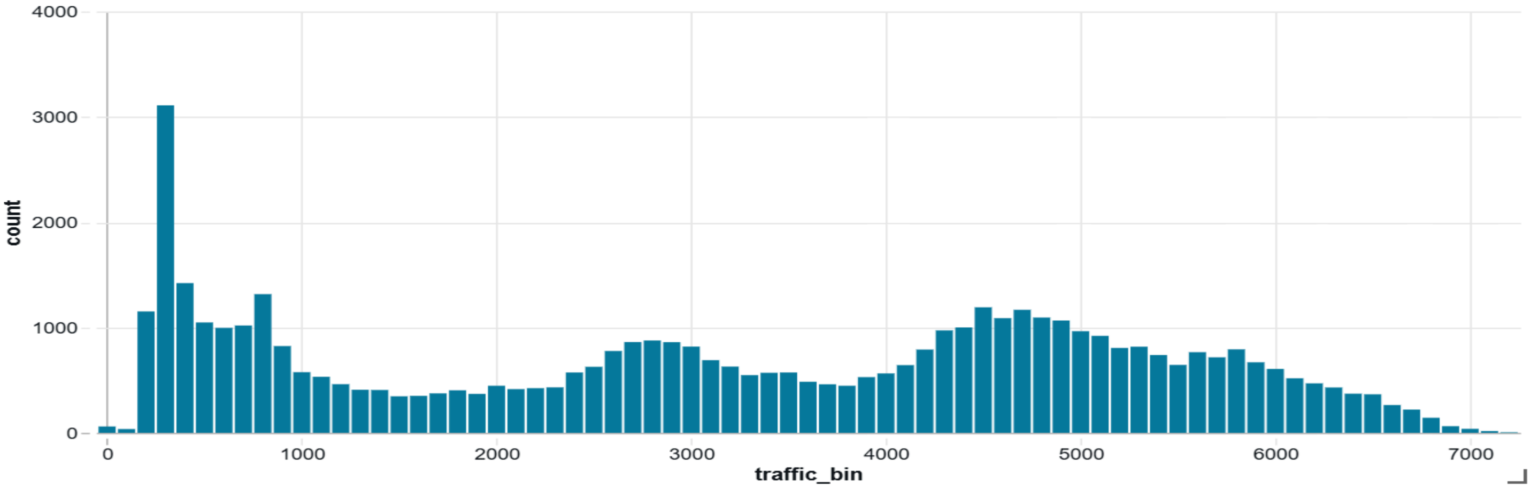
4. Exploratory Data Analysis (EDA)

4.1 Feature Selection

- **Target:** traffic_volume
- **Features:** is_holiday, weather_index, temp, rain_1h, snow_1h, clouds_all, hour, day_of_week, month

4.1 Distribution of Traffic Volume

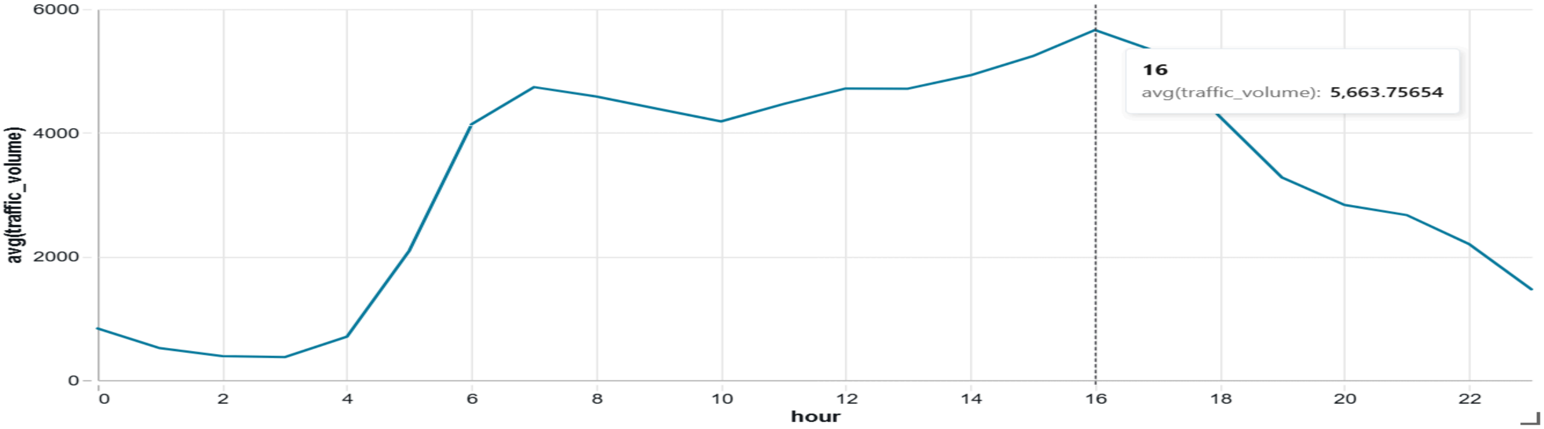
The distribution of traffic volume was analyzed using Spark's approxQuantile method. The estimated quartiles are **Q1 = 1,159**, **Median = 3,341**, and **Q3 = 4,928**, showing a wide range of traffic conditions.



- The distribution is bimodal, showing two main traffic patterns.
- A sharp peak at low volumes (0–1,000 vehicles) represents frequent low-traffic periods, likely late night or early morning.
- A broader peak between 4,000 and 5,500 vehicles represents regular high-traffic or commute hours.
- Traffic volume decreases after 6,000 vehicles, meaning extremely heavy traffic is less common.
- Overall, the data covers a wide range from very light to very heavy traffic conditions.
- Traffic volume was grouped into four quartile-based bins: Low, Medium, High, and Very High.
- Each bin has a similar number of records, showing the dataset is balanced and suitable for prediction modeling.

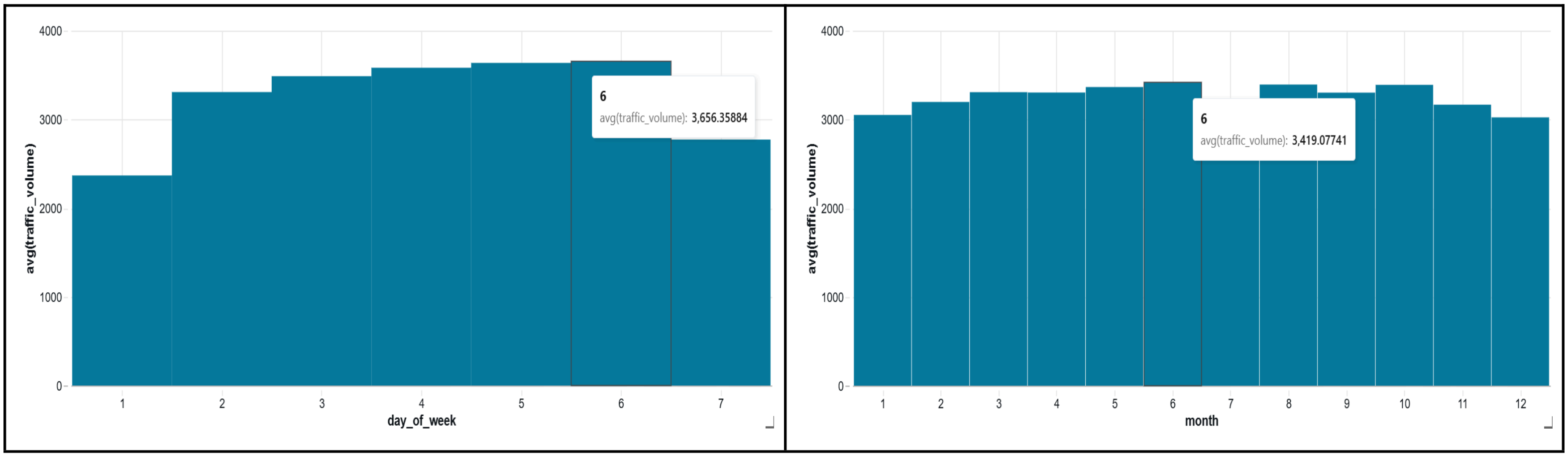
4.2 Time-Based Traffic Patterns

4.2.1 Average traffic per hour



Traffic volume shows clear patterns based on time. By hour of day, traffic is lowest during late-night and early-morning hours (midnight to 4 AM) and increases sharply after 5 AM. Peak traffic occurs between **3 PM and 6 PM**, corresponding to evening commuting hours, before gradually declining at night.

4.2.2 Average traffic per day of week and per month

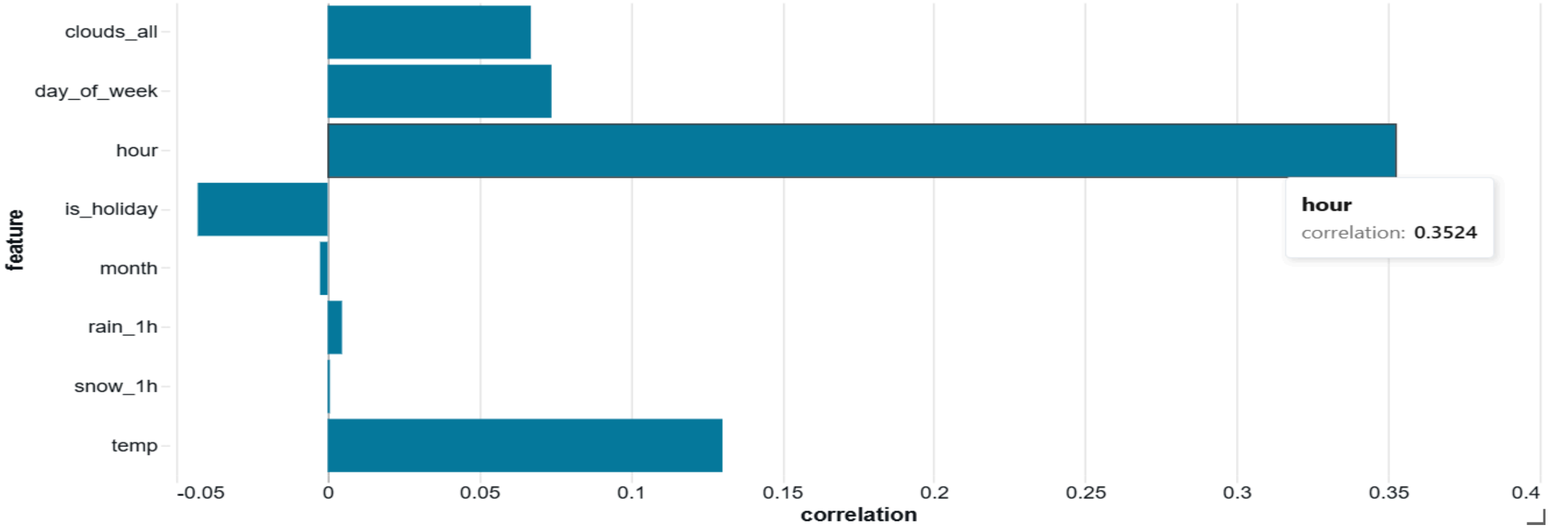


dayofweek() function returns values from **1 (Sunday) to 7 (Saturday).**Traffic volume varies by day of the week, with lower levels on weekends and higher volumes on weekdays, especially from Monday to Friday, indicating typical commuting behavior.

Average traffic per month

Monthly analysis shows moderate seasonal variation. Traffic volume is slightly higher during **late spring and summer months (May to October)** and lower during winter months, likely due to weather conditions and seasonal travel behavior.
These patterns confirm that **hour, day of week, and month** are important features for traffic volume prediction.

4.2.3 Correlation Analysis



The hour of day shows the strongest correlation with traffic volume, confirming that traffic is highly time-dependent. Temperature has a positive relationship, while weather factors such as rain, snow, and cloud coverage have little to no impact. Day of week has a small effect, whereas month shows almost no correlation. Holidays slightly reduce traffic volume. Overall, time-based features are the most important predictors of traffic volume.

5. Model Development and Predictive Modeling

Following the exploratory data analysis, the next phase of this project focuses on developing predictive models to estimate traffic volume. Based on the patterns and insights observed during EDA, relevant temporal and weather-related features were selected. Machine learning techniques were then applied using Apache Spark MLlib to train, test, and evaluate regression models for accurate traffic prediction.

5.1 Model Building Steps

1. **Data Splitting**: The dataset was divided into training and testing sets to ensure unbiased model evaluation.
2. **Label and Feature Selection**: traffic_volume was selected as the target variable, and relevant temporal and weather features were used as predictors.
3. **Vector Assembler**: All feature columns were combined into a single feature vector required for Spark ML models.
4. **Model Selection**: Two regression models were implemented: **Random Forest Regressor** and **Gradient-Boosted Trees (GBT) Regressor**.

5.1.1 Model Justification and Description

Why These Models Were Used:
Traffic volume data contains complex and non-linear patterns influenced by time, weather, and past traffic behavior. Tree-based ensemble models such as Random Forest and Gradient-Boosted Trees were selected because they can effectively capture these non-linear relationships, scale well with large datasets, and are well supported in Apache Spark MLlib.

Random Forest Regressor:
Random Forest builds multiple decision trees using different subsets of data and averages their predictions. This improves stability and reduces overfitting, making it a strong baseline model for traffic prediction.

Gradient-Boosted Trees (GBT) Regressor:
GBT builds trees sequentially, where each tree corrects errors made by previous trees. This allows the model to focus on difficult patterns and achieve higher accuracy, especially for time-dependent traffic data.

5. **Pipeline Construction**: Pipelines were created to streamline feature assembly and model training.
6. **Model Training**: Both models were trained using the training dataset.
7. **Prediction**: The trained models generated traffic volume predictions on the test dataset.
8. **Model Evaluation**: Model performance was assessed using RMSE and R² metrics.

5.2 Evaluation Metrics Explanation:

Root Mean Squared Error (RMSE) measures the average difference between predicted and actual traffic volumes. Lower RMSE values indicate better prediction accuracy.

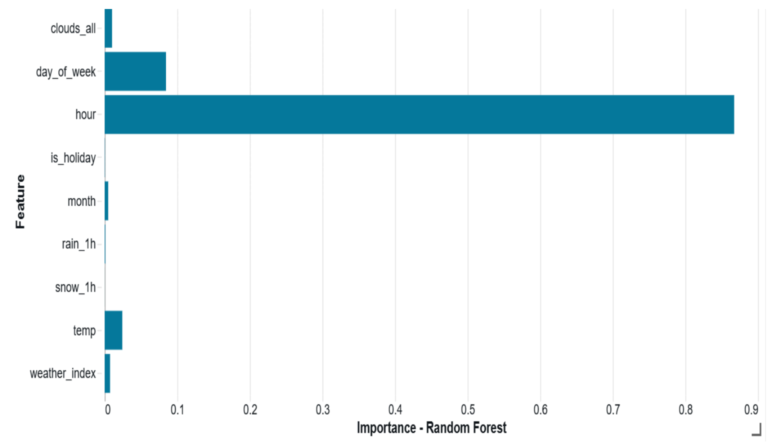
R² (Coefficient of Determination) represents how well the model explains the variability in traffic volume. Values closer to 1 indicate stronger predictive performance.

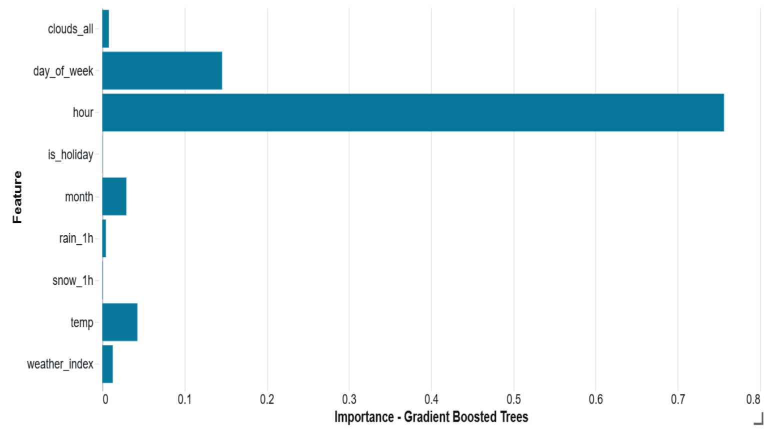
Model	**RMSE**	**R2**
Random Forest	600.90	0.91
Gradient-Boosted Tress	461.69	0.95

The Gradient-Boosted Trees (GBT) model outperformed the Random Forest model, achieving a lower RMSE and a higher R² score. This demonstrates that GBT is better at capturing complex, non-linear traffic patterns, making it the preferred model for predicting traffic volume.

5.3 Feature Importance - RF vs GBT

The importance of each feature was analyzed to identify the most influential variables for predicting traffic volume. Both models highlight **temporal features** as the primary drivers:





Random Forest (RF) Feature Importance:

- **Hour:** 0.867 – the most critical factor, reflecting daily traffic patterns.
- **Day of Week:** 0.085 – captures differences between weekdays and weekends.
- **Temperature:** 0.024 – minor effect.
- Other features (clouds, weather index, month, rain, holiday, snow) have very low importance.

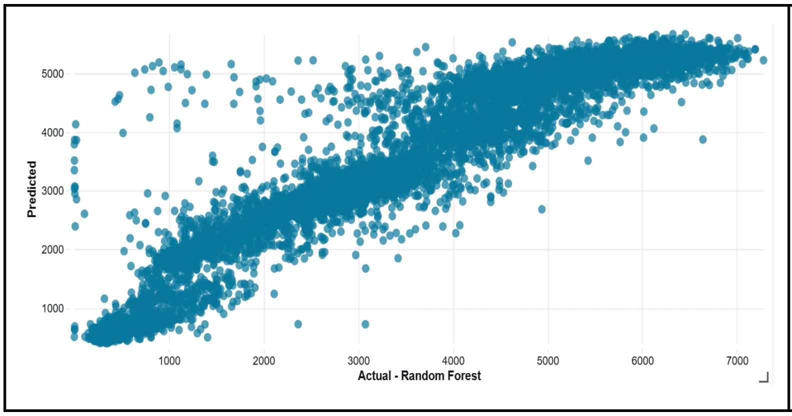
Gradient-Boosted Trees (GBT) Feature Importance:

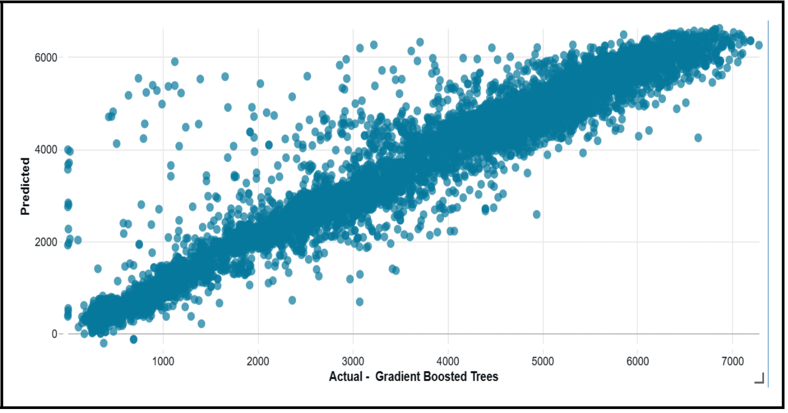
- **Hour:** 0.756 – again the dominant feature.
- **Day of Week:** 0.146 – second most important.
- **Temperature:** 0.043, **Month:** 0.029 – moderate influence.
- Weather and holiday indicators show minimal impact.

Key Insight:

Both models confirm that **hour and day of week are the most important predictors** of traffic volume, while weather and holiday-related features have a smaller role.

5.4 Actual vs Predicted Analysis - RF vs GBT



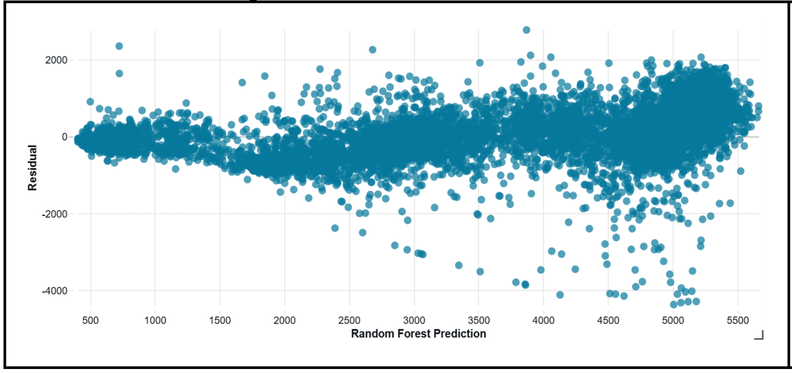


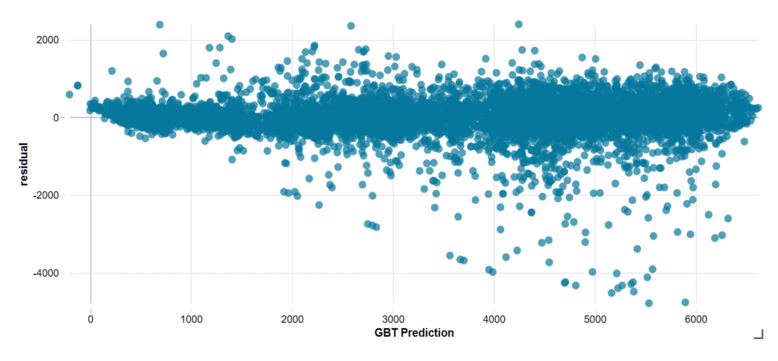
Random Forest (RF):
RF captures the overall trend but flattens at high traffic levels, showing difficulty in predicting very large values. It also has wider spread at low traffic, meaning less precise predictions.

Gradient Boosted Trees (GBT):
GBT follows the diagonal more closely, with better accuracy at low and medium traffic. It still slightly under-predicts at high traffic, but is more consistent than RF.

Key Insight:
GBT outperforms RF with tighter alignment to actual values, especially at medium and high traffic levels, making it the more reliable model.

5.5 Residual Analysis - Random Forest vs GBT





Residual Analysis - A residual plot shows the difference between the actual and predicted traffic volumes, helping to evaluate model accuracy and stability.

Random Forest Residual Plot:
Residuals increase with higher traffic volume, indicating reduced accuracy at high traffic levels.

Gradient-Boosted Trees Residual Plot:
Residuals show a more consistent spread compared to Random Forest, indicating better stability and prediction accuracy.

6. Optimized Gradient-Boosted Trees: Hyperparameter Tuning and Lag Feature Enhancement

The original GBT struggled with high traffic volumes and often under-predicted peak values. After hyperparameter tuning, the model became more accurate and stable: the RMSE was reduced, but the R² value remained nearly the same, but still had difficulty with extreme traffic levels.

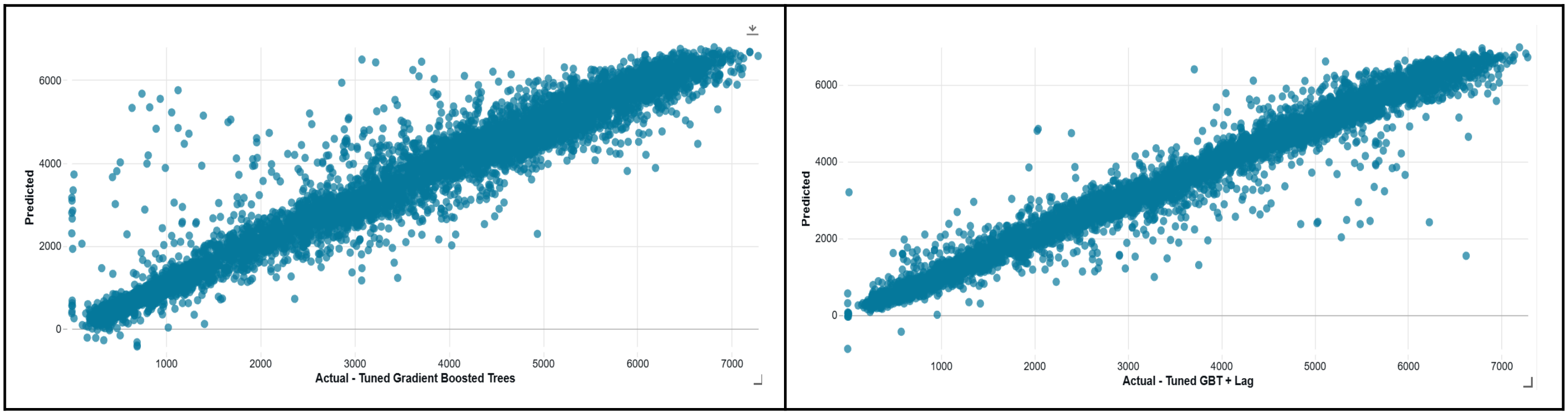
Adding lag features allowed the tuned GBT to use recent and past traffic patterns, helping it capture time-based trends more effectively. As a result, the Tuned GBT with lag features achieved the best performance, with the lowest RMSE (291.41) and highest R² (0.978), making it the most accurate model for traffic prediction.

Model	RMSE	R2
Tuned GBT	419.5	0.95
Tuned GBT with lag feature	291.41	0.978

6.1 Impact of Lag Features

Lag features let the model incorporate recent and past traffic, which is critical because traffic volume follows short-term trends and daily repetition.

Actual vs Predicted Scatter Plot

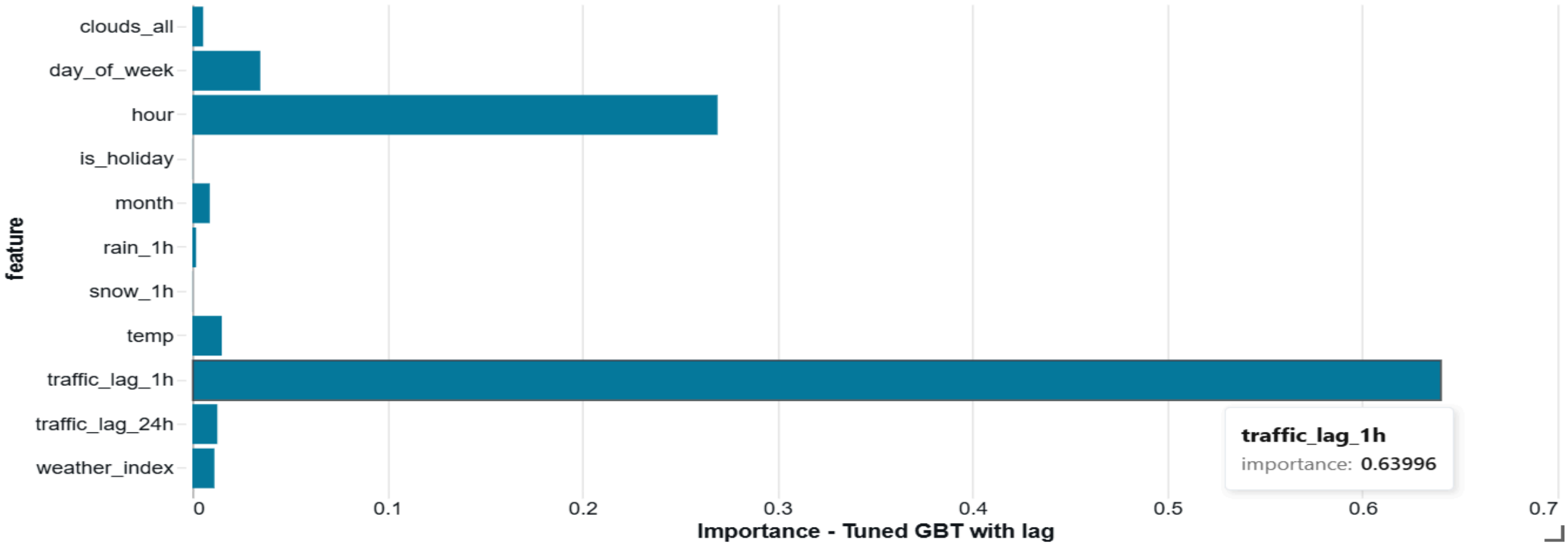


6.2 Accuracy and Stability

Hyperparameter tuning mainly improves the RMSE, meaning the average prediction error is reduced, but the overall accuracy pattern remains similar to the base GBT. The model behavior at high traffic levels does not change much and it still slightly under-predicts extreme values.

After adding lag features, the Tuned GBT + Lag model shows visibly tighter alignment with the diagonal, indicating real improvement in accuracy, not just error magnitude. Predictions become more consistent across all traffic levels, especially during peak periods.

Key Insight:
Tuning helps reduce error (better RMSE), but lag features drive the real accuracy improvement by capturing temporal patterns, leading to more reliable predictions, particularly at high traffic volumes.



6.3 Feature Importance

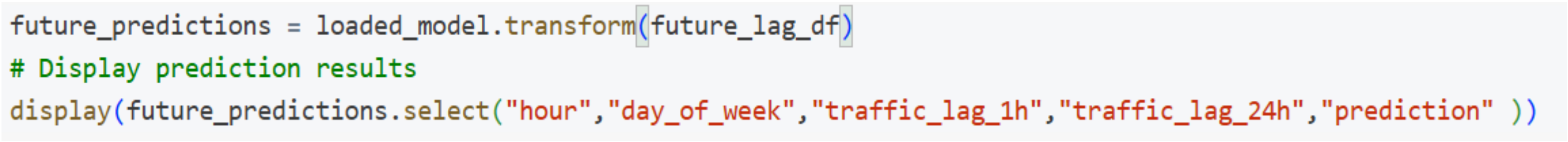
- The most important feature is **traffic_lag_1h**, showing that recent traffic strongly predicts current traffic.
- **Hour of day** is the second most important feature, confirming strong daily traffic patterns.
- **Day of week** also contributes, showing differences between weekdays and weekends.
- Weather-related features (temperature, rain_1h, snow_1h, clouds_all, weather_index) have very low importance, meaning they have little impact on traffic in this model.
- **Holiday** and **month** have minimal influence compared to time-based and lag features.
- Overall, traffic is mainly driven by **recent past traffic and time patterns**, not by short-term weather conditions.

Observation

While hyperparameter tuning improved the GBT, adding lag features gave the largest performance boost, making the tuned GBT with lag features the best model for accurate and reliable traffic volume prediction.

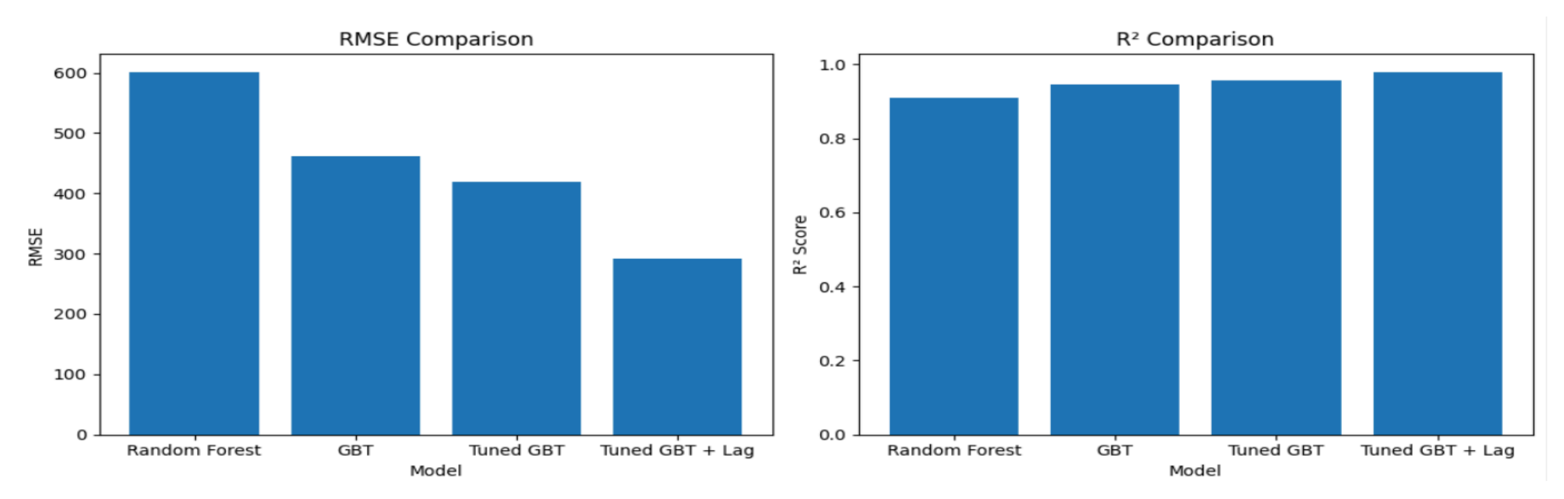
6.4 Future Traffic Prediction Using Tuned GBT with Lag Features

To validate the practical applicability of the proposed model, the tuned Gradient-Boosted Trees model with lag features was applied to simulated future data. The input included projected hour, day_of_week and historical traffic lag values. The model successfully generated future traffic volume predictions, confirming its suitability for deployment in real-time traffic monitoring systems.



```
future_predictions = loaded_model.transform(future_lag_df)
# Display prediction results
display(future_predictions.select("hour","day_of_week","traffic_lag_1h","traffic_lag_24h","prediction" ))
```

7. Final Conclusion



This project focused on predicting traffic volume using machine learning models implemented in Apache Spark MLlib. Exploratory data analysis revealed strong temporal patterns in traffic behavior, with hour of day and day of week emerging as the most influential features, while weather variables showed a relatively weaker impact.

Based on these insights, **Random Forest and Gradient-Boosted Trees** (GBT) models were developed and evaluated using RMSE and R^2 metrics. The initial GBT model outperformed Random Forest, demonstrating a stronger ability to capture complex, non-linear relationships in traffic data. Further improvements were achieved through hyperparameter tuning using 5-fold cross-validation, resulting in a tuned GBT model with an RMSE of 419.51 and an R^2 of 0.95. Actual-vs-predicted and residual analyses confirmed improved prediction stability, particularly during high traffic periods.

To further enhance performance, lag features representing traffic volume from the previous hour and the same hour on the previous day were incorporated into the tuned GBT model. These features enabled the model to capture short-term trends and daily traffic repetition more effectively. The tuned GBT with lag features achieved a significantly improved **RMSE of 291.41 and an R^2 of 0.978,** indicating excellent predictive accuracy. Feature importance analysis confirmed that lag variables were among the most influential predictors.

Overall, the Tuned Gradient-Boosted Trees model with lag features proved to be the most effective approach for traffic volume prediction, combining strong temporal modeling capability with high accuracy. This makes it well suited for real-world traffic forecasting and transportation planning applications.