TRAFFIC VOLUME ANALYSIS AND PREDICTION USING MACHINE LEARNING

Submitted by: Sufia Haseen | B.Tech – IT, 1st Year

Abstract

This project aims to predict traffic volume using weather and time-related features. Predicting traffic can help in better traffic management, reducing congestion, and improving commute planning. The project uses machine learning techniques to learn from historical data and forecast future traffic conditions

sufia H
hasee.sufia@gmail.com

ABOUT

This project explores real-world traffic volume data collected from a major highway and analyzes how factors such as time, weather, and environmental conditions affect traffic flow. Using Python for exploratory data analysis (EDA) and a machine learning model (Random Forest Regression), this project aims to predict traffic volume based on various input features. The project was developed as a self-learning initiative to understand data science concepts practically

DATASET OVERVIEW

- Source: UCI Machine Learning Repository – Metro Interstate Traffic Volume Dataset
- Size: 48,204 rows × 9 columns
- Features used: date_time, hour, dayofweek, weather_main, temp, rain_1h, clouds_all, traffic_volume

KEY COLUMNS USED

- temp: Temperature (in Kelvin)
- rain_1h: Rain volume in the last hour (mm)
- clouds_all: Cloudiness percentage
- hour: Extracted from datetime
- dayofweek: Extracted from datetime
- traffic_volume: Target variable (what we want to predict)

TOOLS USED

- Programming Language: Python
- Libraries: pandas, matplotlib, seaborn, scikit-learn
- Platform: Google Colab

DATA PREPROCESSING

- Removed missing/null values
- Extracted useful time features like hour and dayofweek
- Converted temperature from Kelvin to Celsius
- Classified rain and cloudiness into categories
- Selected relevant numeric features for correlation and modeling
- Cleaned unrealistic temperature values

```
            date_time  traffic_volume    temp  rain_1h  clouds_all  hour  \
0 2012-10-02 09:00:00            5545  288.28      0.0          40     9
1 2012-10-02 10:00:00            4516  289.36      0.0          75    10
2 2012-10-02 11:00:00            4767  289.58      0.0          90    11
3 2012-10-02 12:00:00            5026  290.13      0.0          90    12
4 2012-10-02 13:00:00            4918  291.14      0.0          75    13

   dayofweek
0          1
1          1
2          1
3          1
4          1
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48204 entries, 0 to 48203
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   date_time       48204 non-null  datetime64[ns]
 1   traffic_volume  48204 non-null  int64
 2   temp            48204 non-null  float64
 3   rain_1h         48204 non-null  float64
 4   clouds_all      48204 non-null  int64
 5   hour            48204 non-null  int32
 6   dayofweek       48204 non-null  int32
dtypes: datetime64[ns](1), float64(2), int32(2), int64(2)
memory usage: 2.2 MB
None
```

Figure 1: Sample of cleaned dataset after preprocessing

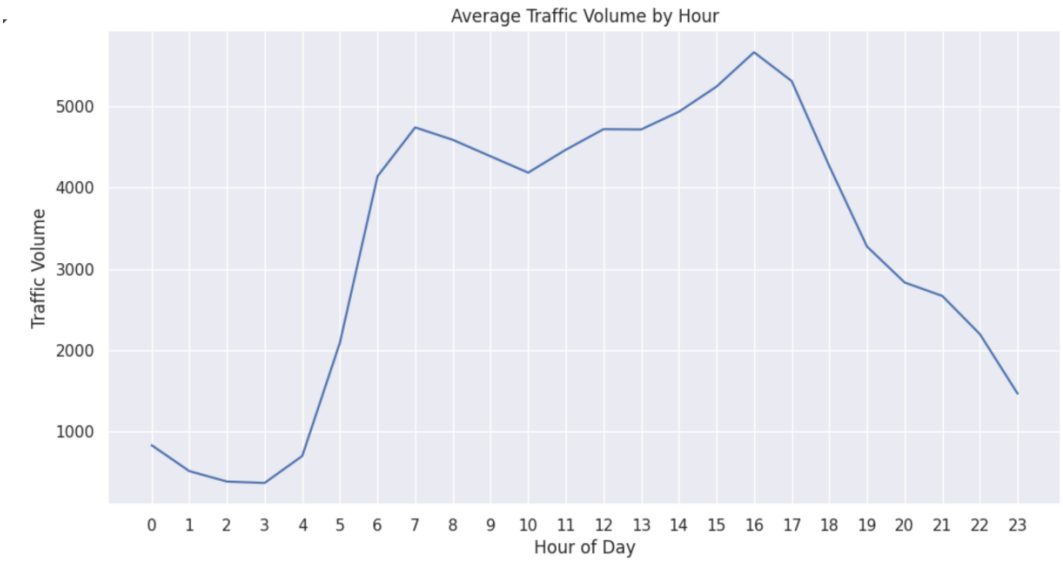
EXPLORATORY DATA ANALYSIS (EDA)

- Created multiple visualizations to understand traffic trends
- Correlation heatmap was plotted to understand relationships between variables
- Found strong correlation between hour and traffic_volume
- rain_1h showed minimal impact on traffic

Visualizations used:

- Bar plots for weather and time-based patterns
- Scatter plot for temperature vs traffic
- Correlation heatmap
- Feature importance plot

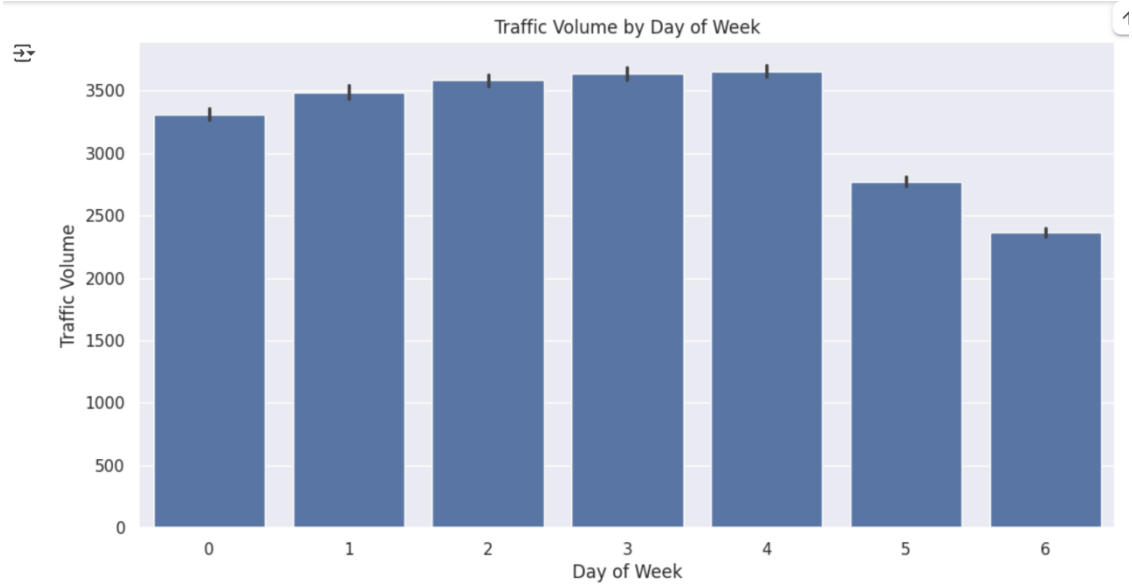
1) HOUR VS TRAFFIC VOLUME



2) DAY OF WEEK VS TRAFFIC

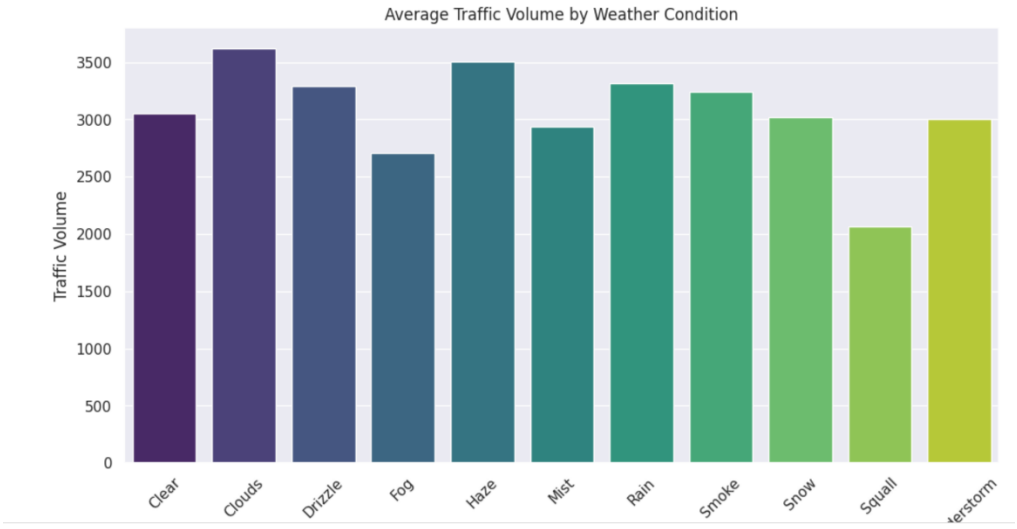
(0=Monday, 6=Sunday)

Higher traffic on weekdays compared to weekends



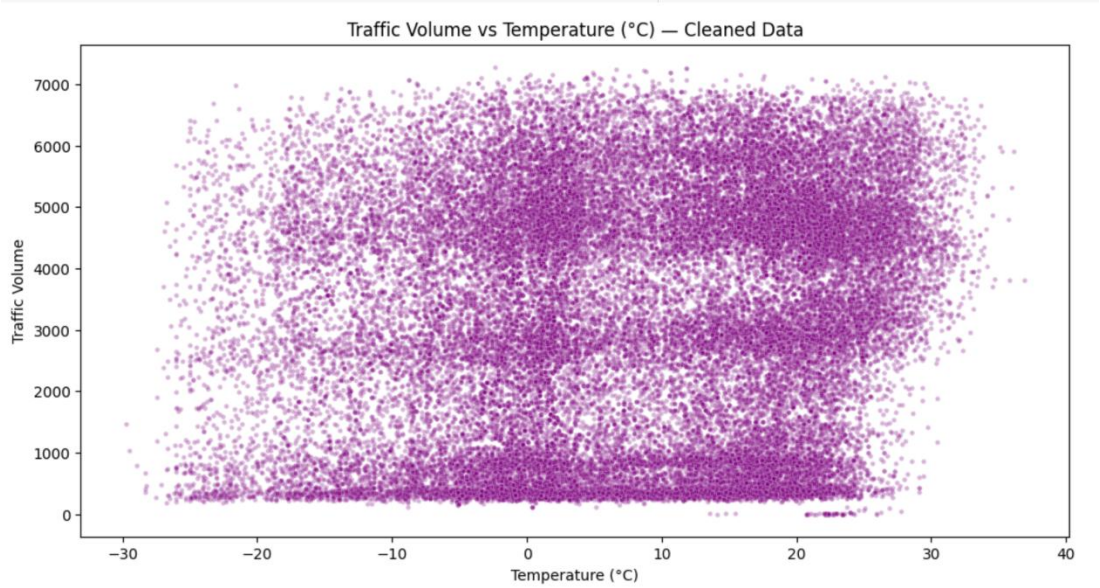
3) AVERAGE TRAFFIC VOLUME BY WEATHER CONDITION

More traffic is observed in cloudy weather



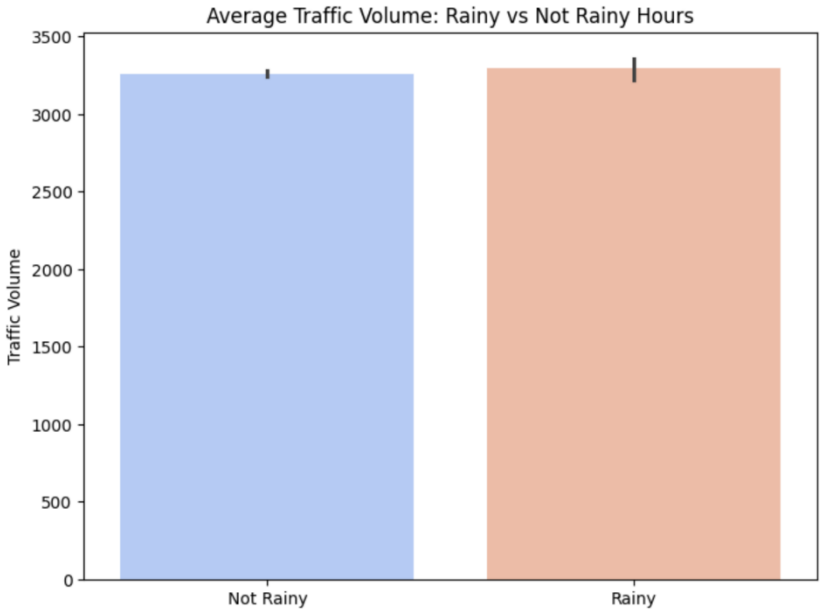
4) TRAFFIC VOLUME VS TEMPERATURE

The traffic volume is scattered and it visualizes how traffic changes with temperature



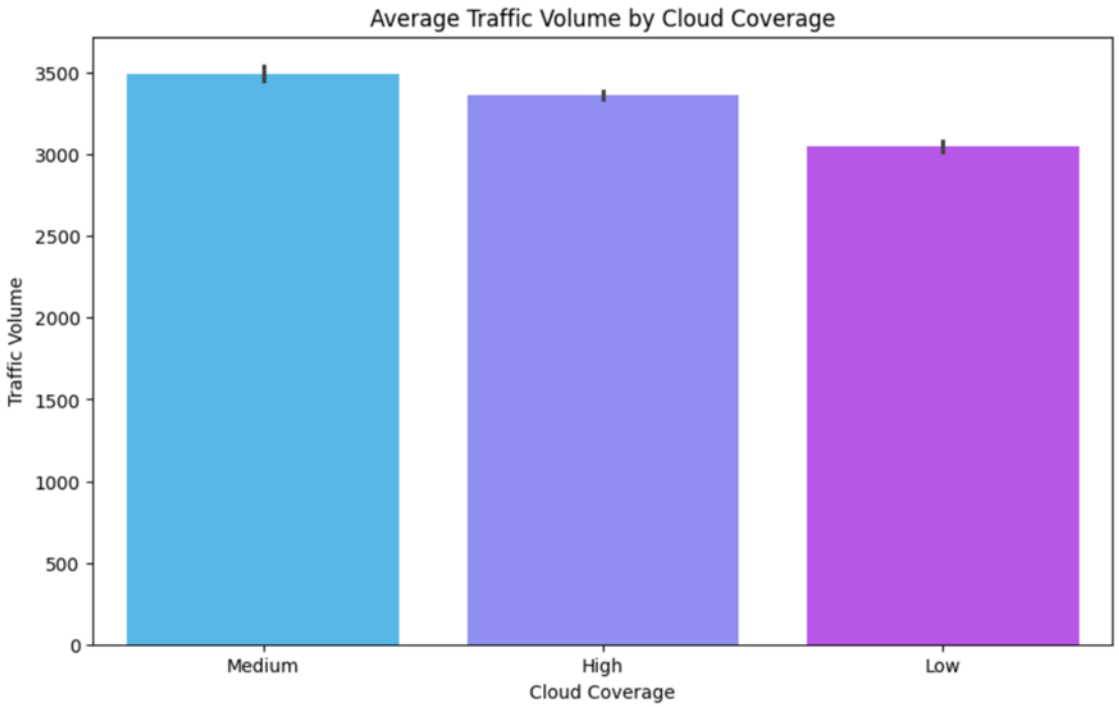
5) RAINY VS NOT RAINY (bar plot)

Shows minimal impact of rain on traffic



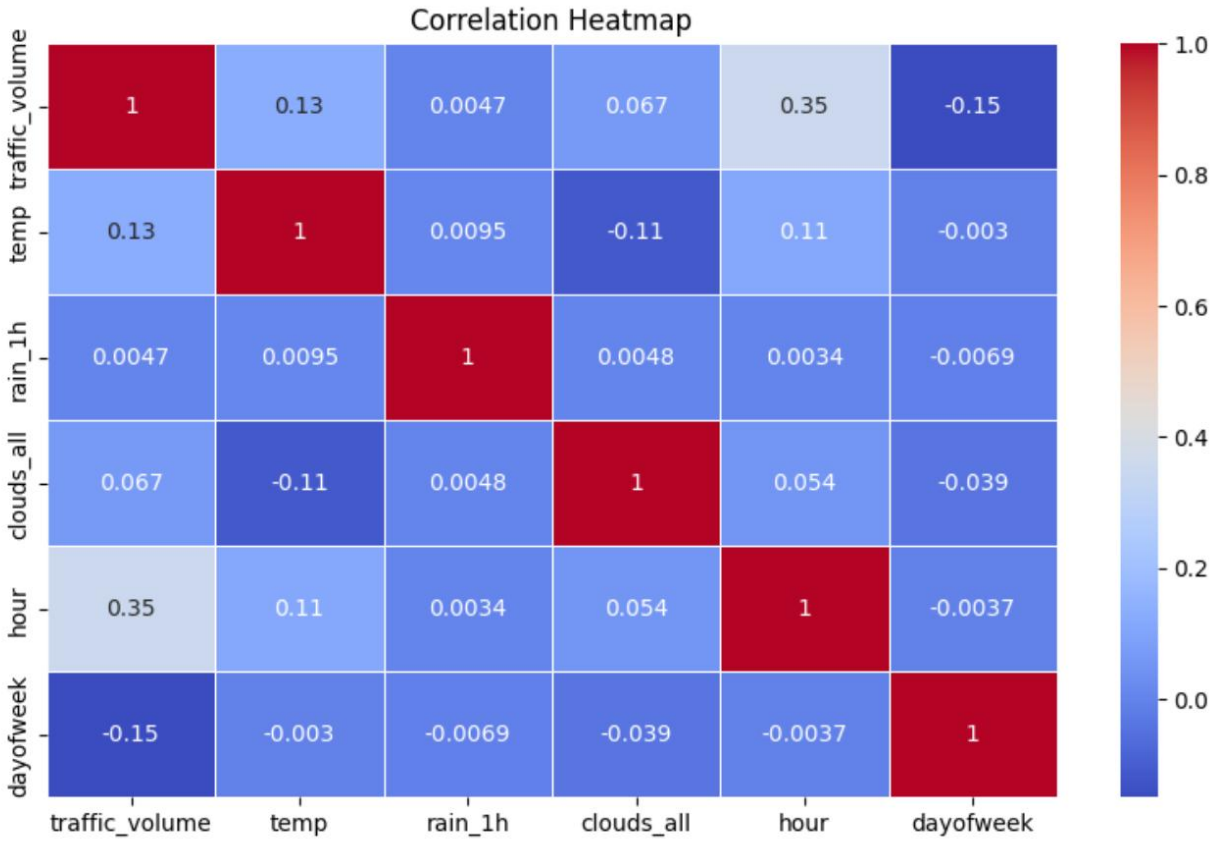
6) CLOUDINESS VS TRAFFIC

Traffic volume is higher on low cloud days.



7) CORRELATION HEATMAP

Strong correlation seen between hour and traffic volume.



MODEL BUILDING USING ML

Linear Regression

I tried a basic linear regression model to predict traffic volume

Results:

- MAE: 1666.87
- RMSE: 6640.28
- R^2 Score: -10.12

The model performed poorly, suggesting traffic patterns aren't linear.

Random Forest Regression

Used Random Forest Regressor with default parameters (100trees)

Results:

- MAE: 255.90
- RMSE: 464.98
- R^2 Score: 0.95

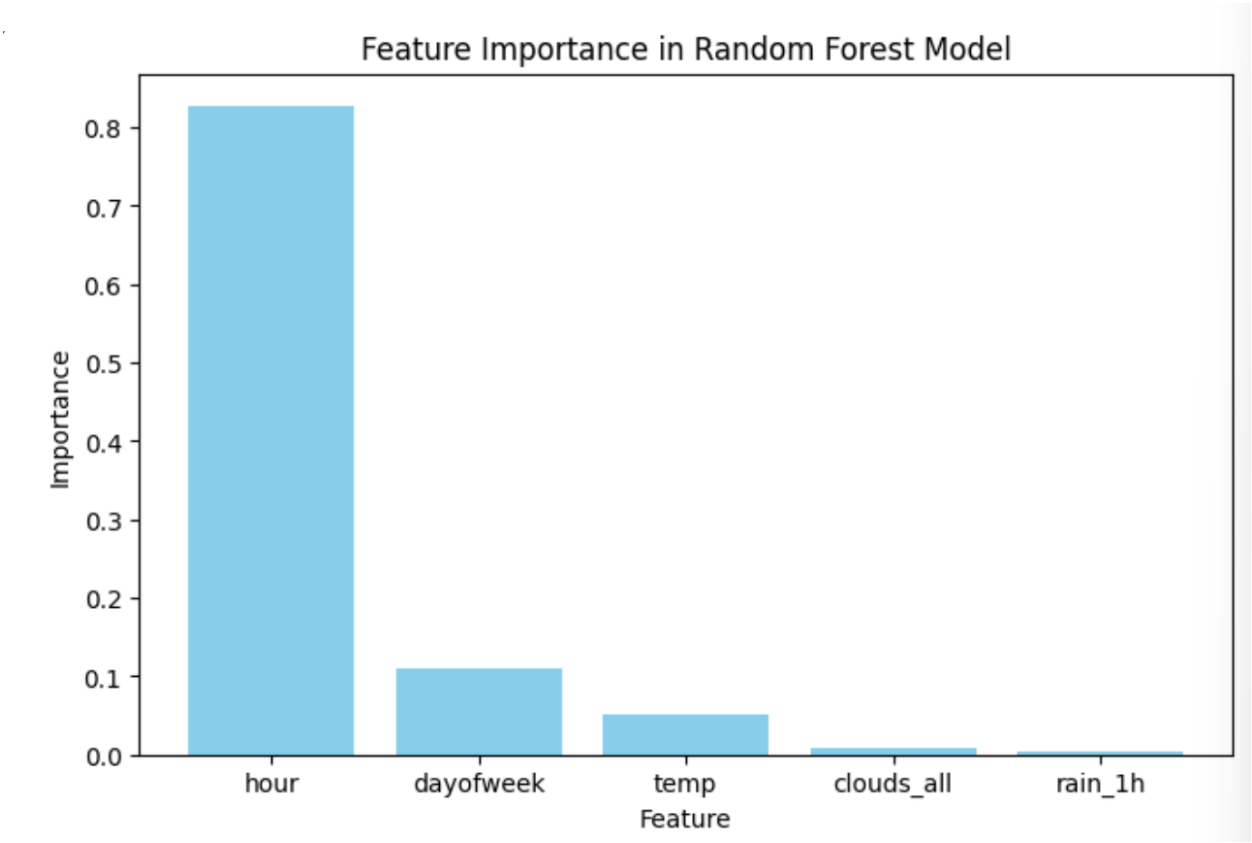
This model performed very well, learning complex traffic patterns and relationships in the dataset.

FEATURE IMPORTANCE

- The most important feature for prediction was **hour**, followed by **dayofweek** and **temperature**
- The least important was **rain_1h**

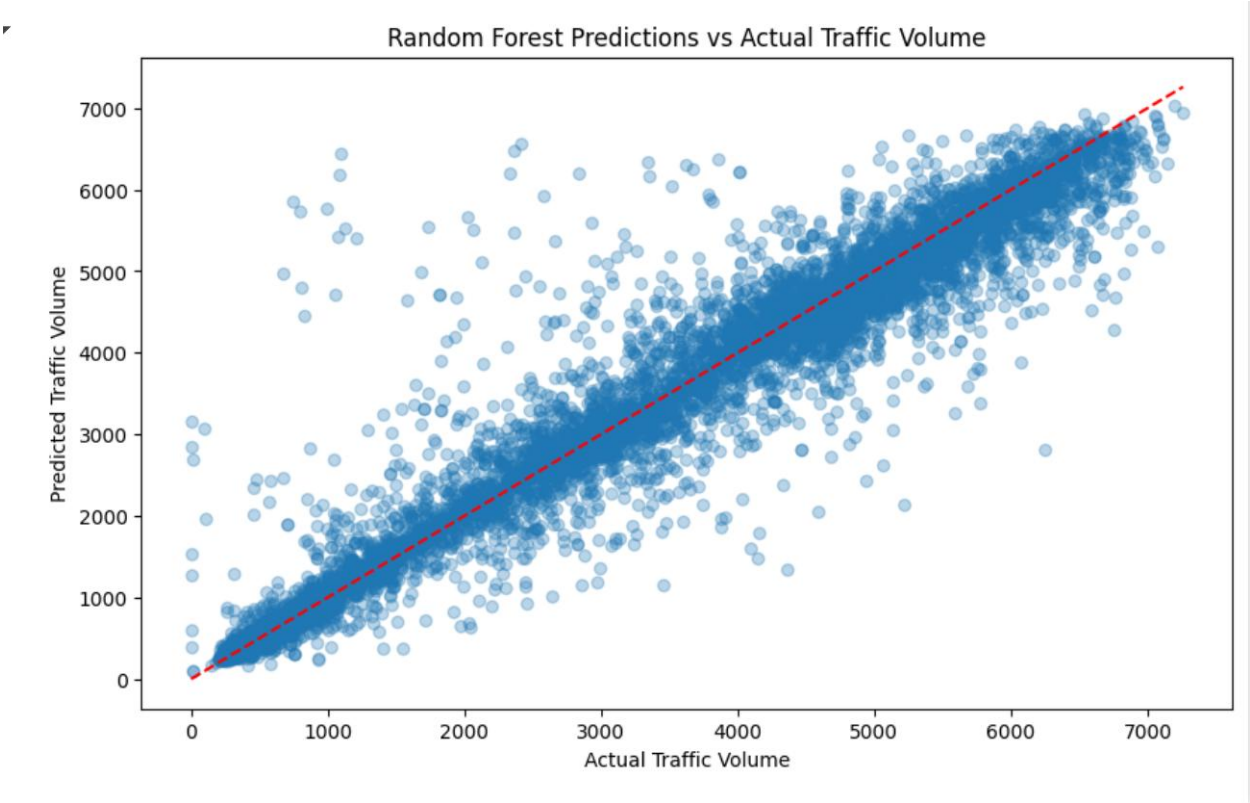
A bar chart of feature importances was generated using the model's feature_importances_ attribute

The most important feature impacting the traffic is depicted in the graph below



MODEL EVALUATION AND VISUALIZATION

A scatter plot comparing actual vs predicted traffic volume showed most points clustered along the diagonal — indicating high accuracy.



CONCLUSION

This project successfully built a traffic volume prediction model using weather and time data. The Random Forest model achieved excellent accuracy, making it suitable for real-world applications.

KEY TAKEAWAYS:

- Time of day is the most important predictor of traffic
- Random Forest significantly outperforms Linear Regression for this dataset
- Visualizations played a key role in identifying patterns and evaluating model performance

FUTURE TAKEAWAYS

- Deploy the model as a web app using Streamlit
- Include more features like holidays, traffic events, or road type
- Try other ML models such as XGBoost or LightGBM
- Perform hyperparameter tuning for better results