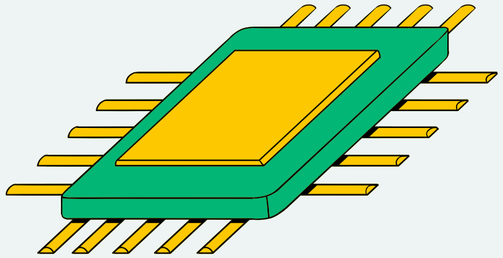
MACHINE LEARNING GROUP-16

PRESENTATION

PRESENTED BY:

K RITWIK
ABHILASH
TYAGARAYAN
VIVEK



PROBLEM STATEMENT

The project aims to predict traffic volume on metro interstate highways using historical data and machine learning models. like random forest and decision trees. The data includes weather attributes, holidays, and temporal features. The main objective is to create a robust regression model that effectively captures patterns in traffic volume to assist in planning and management.

INNOVATIVE IDEA

We have generated weekly and hourly how the traffic volume is changing from the dataset which doesn't have them intially.

We also generated how the traffic volume varies on a holiday and non holiday days

We have categorised the main_weather into 12 categories of boolean type

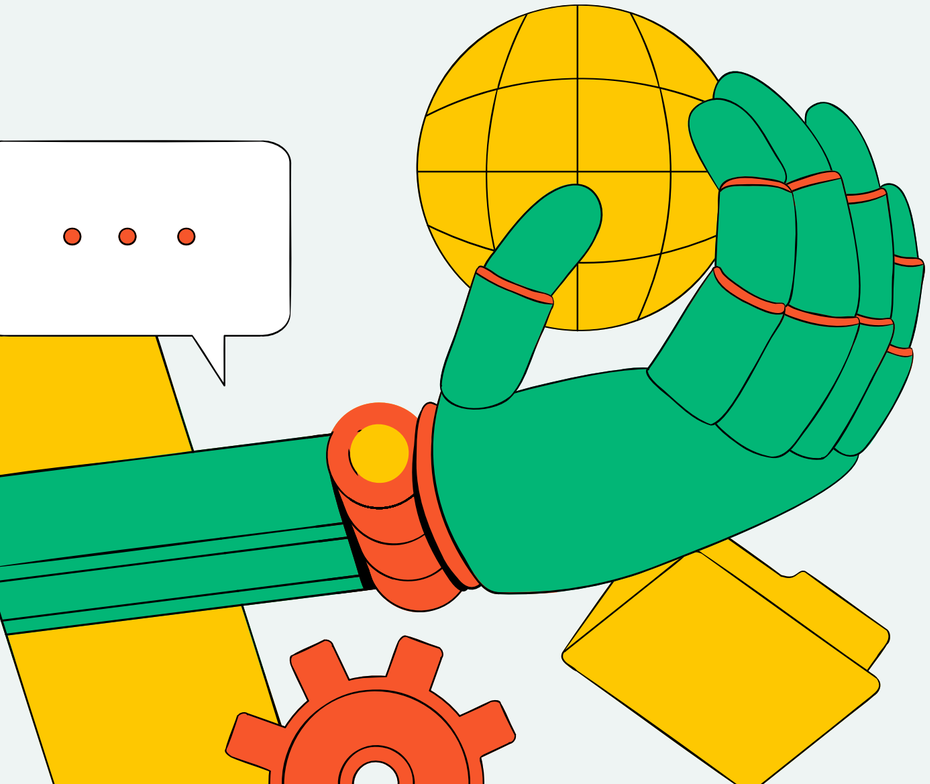
DATA COLLECTION AND PRE-PROCESSING

- Outlier removal: Dropped extreme values for temperature, rain, and snow.
- Conversion of temperature to Celsius.
- Creation of additional features: holiday_bool (binary column for holidays) and hour/day-related features.
- One-hot encoding of categorical variables (weather_main).
- Grouping data by date_time to retain maximum values for identical timestamps.

MODEL BUILDING - RANDOM FOREST

- Mechanism: Combines predictions from multiple decision trees trained on random subsets of data to improve robustness and reduce overfitting.
- Key Features:
- Uses bagging (Bootstrap Aggregating) for sampling.
- Aggregates results from all trees to provide the final prediction.
- Performance metric: Mean Squared Error (MSE).
- Strengths:
- Handles complex relationships in data.
- Reduces variance compared to single decision trees.

MODEL BUILDING- DECISION TREE



- Mechanism: Builds a single tree by recursively splitting data based on features that minimize the Mean Squared Error (MSE).
- Key Features:
- Finds the best feature and threshold at each node to split data.
- Grows branches until a stopping criterion is met (e.g., maximum depth, or minimum samples per leaf).
- Performance:
- Simpler than Random Forest but more prone to overfitting, especially on noisy datasets.
- Used as a baseline model for comparison with Random Forest.

EVALUATION METRICS

R2 VALUE

Random Forest: 0.91
Decision Tree:0.84

MEAN SQAURED ERROR

Random Forest: 334637.13
Decision Tree:624364.32

MEAN ABSOLUTE ERROR

Random Forest: 334637.13
Decision Tree:569.97

VISUALISATIONS

SCATTER PLOT

- Compare actual vs. predicted traffic volumes

RESIDUALS PLOT

- Display residuals (errors) to assess prediction accuracy.

RESIDUALS HISTOGRAM

- Visualize residual distribution to check for normality.

PRECISION-RECALL CURVE

- Evaluate trade-off between precision and recall.

ROC CURVE

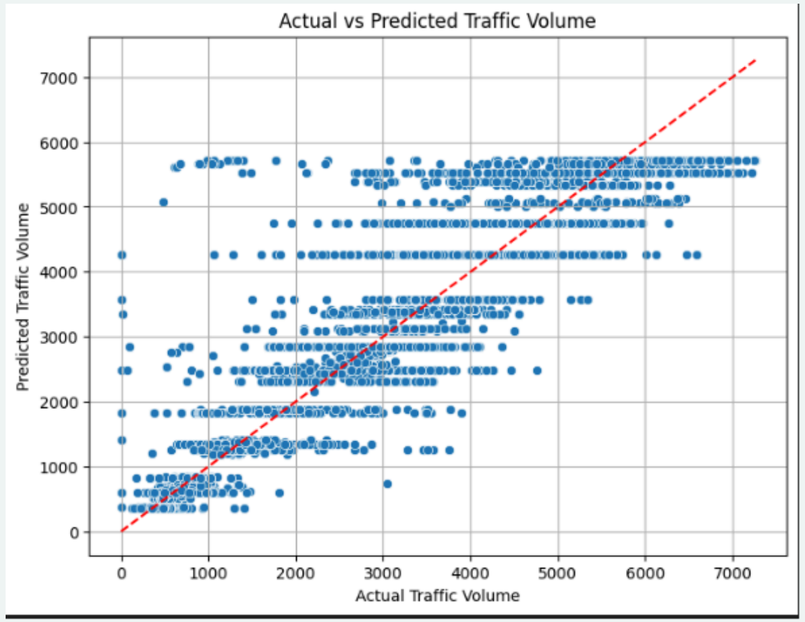
- Assess model's classification ability by plotting True Positive Rate (TPR) vs. False

Q-Q PLOT

- Compare data distribution to theoretical normal distribution

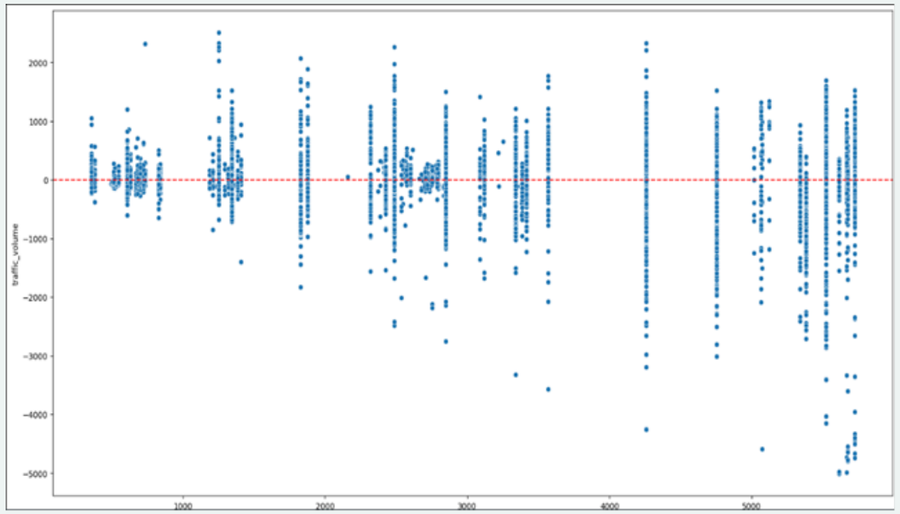
SCATTER PLOT

Most points aligned closely to the diagonal (perfect prediction line), indicating good accuracy for mid-range traffic volumes.



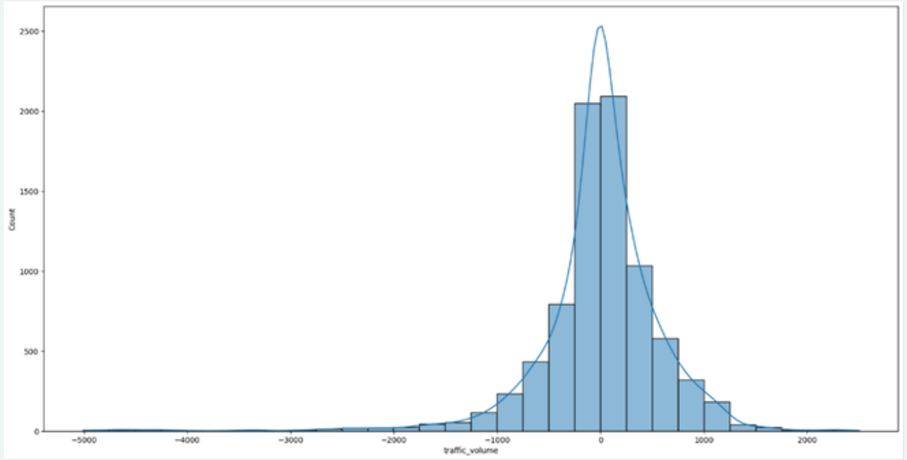
RESIDUALS PLOT

Residuals were centered around zero, indicating minimal bias. Random dispersion suggested the models captured the main relationships in the data.



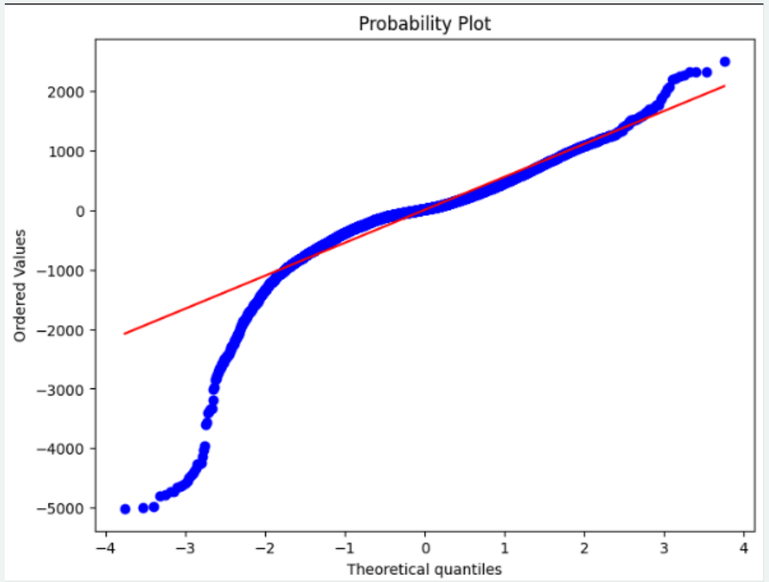
RESIDUALS HISTOGRAM

A central peak around zero indicated reasonable performance for most predictions.



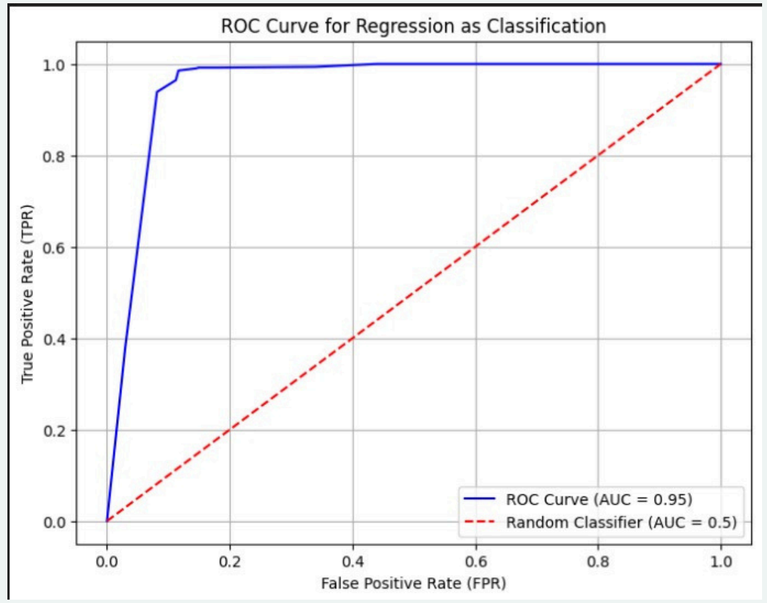
Q-Q PLOT:

Middle quantiles aligned well with the reference line, suggesting normality for central values.



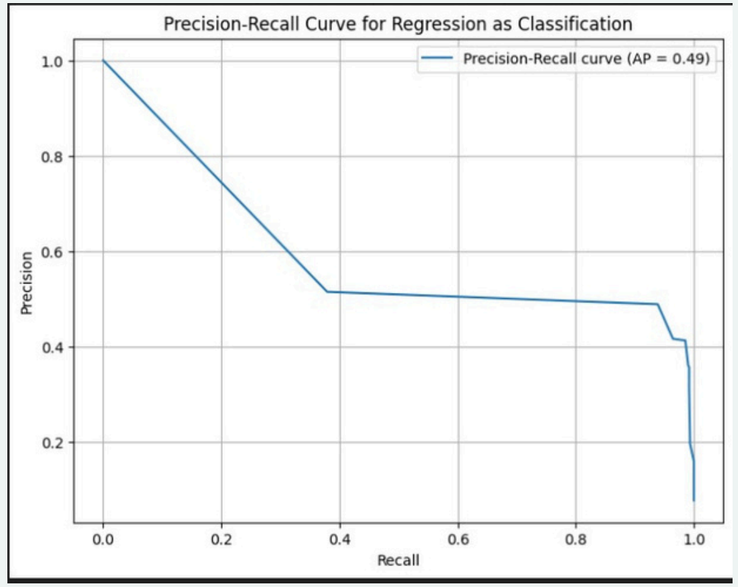
ROC CURVE

A steep initial rise and minimal deviation from the top-left corner highlighted strong classification ability.

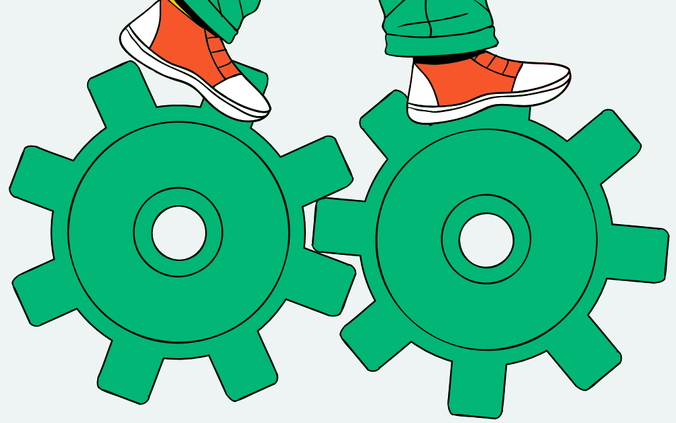


PRECISION-RECALL

High precision at low recall showed the models' ability to correctly identify true positives but with decreasing performance for high recall.



INDIVIDUAL CONTRIBUTIONS



VARUN TYAGARAYAN

Decision Tree model building along with results of Decision Tree

ABHILASH

Random Forest model development along with Data pre-proessing

RITWIK KARTHIKEYA

Data collection and Data pre-processing involving outlier removals , etc

VIVEK PIDUGU

Results , graphs and comparision of model metrics visually .