Gradient Boosting Regression

Introduction

This notebook performs a detailed analysis of traffic volume data using various machine learning regression models, focusing on predicting traffic volumes based on features such as road name, location, speed limits, and others. The goal is to identify the best model for traffic volume prediction on strategic road segments.

Data Preprocessing

- **Features**:
    - **Categorical Features**: road_name, location, suburb, etc. (encoded as numerical values).
    - **Numerical Features**: speed_limit, average_speed, maximum_speed, hour, etc.
    - **Temporal Features**: datetime, day_type (weekend vs weekday), time_of_day (morning, afternoon, etc.).
- **Target**:
    - **Traffic Volume** (Total_Traffic_Volume): The predicted variable representing the total number of vehicles on a given road segment at a specific time.

Model Selection

Three primary models were tested to predict traffic volume:

- **LightGBM (LGBM)**
- **XGBoost (XGB)**
- **LSTM**

Reasons for Preferring LightGBM

- **Performance**: LightGBM provided superior performance in terms of model accuracy, speed, and scalability. The R^2 score for LGBM was consistently higher across all test datasets, with lower Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).
- **Speed**: LightGBM is designed to handle large datasets efficiently and outperformed XGBoost in terms of training time, especially on high-dimensional traffic data. LGBM utilizes histogram-based algorithms, which reduce computational overhead.
- **Memory Efficiency**: LGBM has lower memory consumption compared to XGBoost, which allowed it to handle the large traffic dataset with ease.
- **LSTM**: Initially considered due to its strength in time-series data, LSTM was computationally expensive and slower during both training and prediction phases. The traffic dataset did not heavily rely on sequential patterns, making Gradient Boosting models more appropriate.

- **XGBoost**: Popular boosting algorithm but slower in training on large datasets.
- **Extra Trees**: Although Extra Trees provided decent accuracy, it was less interpretable compared to LGBM. Feature importance analysis in LGBM helped in identifying the key factors affecting traffic volume prediction.

Model Tuning

The notebook includes hyperparameter tuning for LGBM and XGBoost, where parameters such as learning rate, number of estimators, and maximum depth were optimized. LightGBM's early stopping feature was employed to prevent overfitting.

Results

The results showed that LGBM outperformed XGBoost and Extra Trees in terms of both predictive power and computational efficiency. LGBM was selected as the primary model for further analysis and deployment.