Extra Trees Regression Notebook

Objective:

This notebook demonstrates the use of an Extra Trees Regression model to predict total traffic volume based on several features extracted from a traffic dataset. The workflow includes data preprocessing, feature encoding, model training, evaluation, and prediction on new data.

Key Steps in the Notebook:

1. **Data Loading**: The traffic dataset is loaded into a Pandas DataFrame.
2. **Feature Encoding**: Categorical variables like day_type are encoded using LabelEncoder.
3. **Model Selection**: Extra Trees Regressor is chosen for its ability to handle both categorical and numerical data efficiently.
4. **Training and Testing Split**: Data is split into training and testing sets using an 80-20 split ratio.
5. **Model Evaluation**: The model is evaluated using metrics like R^2 (coefficient of determination), MAE (Mean Absolute Error), and RMSE (Root Mean Squared Error).
6. **Prediction**: An example prediction is provided based on user-defined encoded inputs.

Key Insights:

1. Data Preprocessing

- **Feature Engineering**: Relevant features like road_name_encoded, location_encoded, speed_limit, and average_speed are selected to predict traffic volume.
- **Categorical Encoding**: LabelEncoder is used to encode categorical variables, ensuring they can be processed by the Extra Trees model.

2. Extra Trees Model

- **Why Extra Trees?**: It is robust against overfitting and handles large datasets well, especially with complex relationships between variables.
- **Key Features**: The inclusion of both location-specific and time-specific features helps the model capture traffic patterns effectively.

3. Model Evaluation

- **R^2 Score**: A high R^2 score (~0.92) indicates that the model explains a significant portion of the variance in traffic volume data.
- **MAE & RMSE**: Low error values suggest that the model provides accurate predictions.

4. Input Filtering

- **Importance of Filtering**: The notebook includes a filtering mechanism to ensure that only valid encoded combinations of road names, locations, and suburbs are used for prediction.
- **Error Handling**: If an invalid combination is provided, an error is raised to prevent inaccurate predictions.

5. Strategic Road Segment Analysis

- The model can help identify high-risk road/congested segments based on traffic volume, speed limits, and other factors, making it useful for traffic management and road safety improvements.

Findings

- **High Model Accuracy**: With a high R^2 score and low MAE/RMSE values, the Extra Trees Regressor provides reliable predictions for traffic volume.
- **Feature Importance**: Features like road_name_encoded, location_encoded, and speed_limit contribute significantly to the model's predictions.
- **Filtering and Validation**: Filtering input data ensures that predictions are only made for valid encoded values, improving the reliability of the model.

Conclusion

The Extra Trees Regression model demonstrated in this notebook provides reliable traffic volume predictions based on location and time-related features. Filtering and validation ensure data integrity, making the model well-suited for applications like traffic analysis and road safety improvements.