Traffic Volume Model

Objective:

The notebook focuses on building machine learning models to predict **Total Traffic Volume** using a variety of features related to road characteristics, temporal factors, and vehicle data. The goal is to compare the performance of multiple models and select the best-performing one based on evaluation metrics.

Dataset Overview:

The dataset used contains:

- **Road Information**: road_name_encoded, location_encoded, suburb_encoded, speed_limit.
- **Temporal Features**: hour, day_of_week, month.
- **Speed Metrics**: average_speed, 85th_percentile_speed, maximum_speed.
- **Target Variable**: Total_Traffic_Volume.

Data Preprocessing:

1. **Feature Selection**:

a) The key features for the models are:
   i. **Road and Location Information**: Encoded columns for road_name, location, suburb.
   ii. **Temporal Features**: Time-based features such as hour, day_of_week, month were used.
   iii. **Speed Limit** and **Average Speed** were included to analyze their relationship with traffic volume.
   iv. **Target Variable**: Total_Traffic_Volume.

2. **Label Encoding**:

   - Categorical features like day_type was label-encoded to convert them into numerical format.

3. **Train-Test Split**:

   - The dataset was split into 80% training and 20% testing sets using train_test_split.

Models Used and Their Performance:

1. **Random Forest Regressor**:

   - **R^2 Score**: 0.9272
   - **Mean Absolute Error (MAE)**: 19.67
   - **Root Mean Squared Error (RMSE)**: 48.55

- **Summary**: The Random Forest Regressor was the top performer, explaining 92.72% of the variance in the target variable. It also had the lowest MAE and RMSE among all models.

2. **Gradient Boosting Regressor**:
    - **R^2 Score**: 0.6564
    - **MAE**: 60.38
    - **RMSE**: 105.47
    - **Summary**: Gradient Boosting had moderate performance. While its predictions were more accurate than some models, its R^2 score was significantly lower than Random Forest's, showing room for improvement.
3. **Support Vector Regressor (SVR)**:
    - **R^2 Score**: -0.0598
    - **MAE**: 85.66
    - **RMSE**: 185.25
    - **Summary**: SVR performed poorly with a negative R^2 score, indicating that it was not suitable for predicting traffic volume on this dataset without significant tuning.
4. **K-Nearest Neighbors (KNN)**:
    - **R^2 Score**: 0.9008
    - **MAE**: 25.14
    - **RMSE**: 56.66
    - **Summary**: KNN performed well, achieving an R^2 score close to Random Forest's performance. This model offers another strong option for predicting traffic volume, though its error metrics were slightly higher than Random Forest.
5. **Ridge Regression**:
    - **R^2 Score**: 0.0499
    - **MAE**: 104.13
    - **RMSE**: 175.40
    - **Summary**: Ridge Regression showed low predictive power, with a very low R^2 score. This suggests that Ridge Regression was not a good fit for this dataset.
6. **Lasso Regression**:
    - **R^2 Score**: 0.05
    - **MAE**: 104.10
    - **RMSE**: 175.39
    - **Summary**: Similar to Ridge Regression, Lasso performed poorly, showing low R^2 and high error metrics, making it an unsuitable model for this prediction task.
7. **ElasticNet Regression**:
    - **R^2 Score**: 0.0501
    - **MAE**: 104.10

- **RMSE**: 175.39
- **Summary**: ElasticNet, a combination of Ridge and Lasso, performed similarly to its individual components. It didn't significantly improve predictions over Lasso or Ridge Regression.

8. **Extra Trees Regressor**:

- **R^2 Score**: 0.9230
- **MAE**: 19.91
- **RMSE**: 49.93
- **Summary**: Extra Trees performed almost as well as Random Forest, with a slightly lower R^2 score and marginally higher RMSE. It can be considered a strong alternative to Random Forest.

9. **CatBoost Regressor**:

- **R^2 Score**: 0.9058
- **MAE**: 28.60
- **RMSE**: 55.23
- **Summary**: CatBoost showed solid performance, with an R^2 score above 90%. However, its MAE and RMSE were higher than those of Random Forest and Extra Trees, making it slightly less accurate overall.

Model Comparison:

Top Performing Models:

- **Random Forest Regressor** (R^2: 0.9272, MAE: 19.67, RMSE: 48.55) and **Extra Trees Regressor** (R^2: 0.9230, MAE: 19.91, RMSE: 49.93) emerged as the best models for predicting traffic volume. These models had the highest R^2 scores and the lowest error rates.

Moderate Performance:

- **K-Nearest Neighbors** also performed well with an R^2 score of 0.9008, but it had higher MAE and RMSE compared to the top models.
- **CatBoost** performed decently with an R^2 score of 0.9058, but its error metrics were not as competitive as Random Forest or Extra Trees.

Underperforming Models:

- **Support Vector Regressor (SVR)**, **Ridge**, **Lasso**, and **ElasticNet** performed poorly, with very low R^2 scores and high error metrics, indicating that these models are not suitable for this dataset.

Key Insights:

1. **Feature Importance**:

- Features such as road_name, location, suburb, and speed_limit were critical in predicting traffic volume. Temporal features like **hour** and **day_of_week** also significantly contributed to the model's ability to predict traffic volume.

2. **Top Models**:

    - **Random Forest** and **Extra Trees** are the most effective models for this dataset, both explaining over 92% of the variance in traffic volume. They are well-suited for complex data with non-linear relationships.

3. **Impact of Speed Metrics**:

    - Including **speed limit** and **average speed** as features allowed the models to account for traffic flow patterns, which further improved the accuracy of the predictions.

4. **Poor Performance of Linear Models**:

    - Linear models such as **Ridge**, **Lasso**, and **ElasticNet** did not perform well. These models were unable to capture the non-linear relationships within the dataset, leading to poor predictions.