Data Science Challenge: Crack the Traffic Code: Build Smarter Predictions!

Problem Statement

In this challenge, you are tasked with developing a predictive model for **traffic volume forecasting** using a real-world dataset. Your focus will be on **data preprocessing, feature engineering, and model training** to achieve the best possible performance on the **test dataset**.

Like other standard machine learning competitions, you **need to submit predictions on a test dataset**, also your goal is to demonstrate strong **data preparation skills and achieve a low RMSE on the test data** while following best practices in model development.

Dataset: Download the dataset using the below link:

Train Data: https://docs.google.com/spreadsheets/d/e/2PACX-1vQKu1w3cJAnm1R-1or-jBLnU7SS48s6_u2ndxt7PyIThnw1ID-q1ewUZ1sm-xj6umpG23xLtq_4CSuM/pub?output=csv

Test Data: https://docs.google.com/spreadsheets/d/e/2PACX-1vTTvl5JV9ajz5vvQg0dWg798tvDjyR9OrigvYg_617dJojhBgF1mjMhowpefMFLrjYrVSXCFObx1H2u/pub?output=csv

Sample Submission: https://docs.google.com/spreadsheets/d/e/2PACX-1vTOtpdwwLU64wlMH_c1ImWfMVOCv1F_k0ihyzfN_EQXSKujNhCBbjaQ73ego3sIK2YWitN8yaLkFESK/pubhtml

Dataset Description

You will be provided with a dataset containing:

- **Timestamp-based traffic volume data**
- **Environmental conditions (e.g., weather, temperature, etc.)**
- **Other influential factors affecting traffic flow**

Your Task

1. Load and Explore the Data

- Read and inspect the dataset to understand its structure.

2. Data Preprocessing

- Handle missing values, outliers, and any inconsistencies.
- Convert timestamps into meaningful time-based features (hour, day of the week, etc.).

3. Feature Engineering

- Generate new features that could improve model performance.
- Handle categorical variables appropriately.
- Scale/normalize numerical features as needed.

4. Train a Predictive Model

- Choose an appropriate regression model or time series model to forecast traffic volume.
- Justify your model choice and optimization strategy.

5. Generate Predictions on Test Data

- Apply your trained model to the test dataset to predict traffic volume.
- Submit your predicted values for the test dataset.

6. Evaluate Performance

- Your submission will be evaluated using Root Mean Squared Error (RMSE) based on actual test labels.

7. Documentation

- Clearly explain your approach, preprocessing steps, and feature engineering choices in a Jupyter Notebook.

Challenge:

Submit your predicted CSV file with a single column named **'Traffic_Vol'** to " **https://challenge.astrikos.xyz:3443/** " to see your ranking on the leaderboard. You have a total of 10 submission attempts, so focus on fine-tuning your existing model to improve the RMSE.

Submission Guidelines

📌 **Submit a .zip file containing below files:**

- **Python Notebook (.ipynb) containing** your **data preprocessing and feature engineering** steps, your **trained model and RMSE on the training dataset.**
- **A CSV file containing your predictions for the test dataset.**

Fill the below form for submission: (you can only fill the form once)

Submission Form: https://forms.gle/rWamnUCQ24bfkzrN9

📌 **Evaluation Criteria**

Criteria	Weightage
Data Cleaning & Preprocessing	20%
Feature Engineering & Selection	30%
Model Performance (based on RMSE)	30%
Clarity & Explanation	20%

⚡ **Think like a real-world data scientist**—focus on meaningful preprocessing and feature engineering to improve prediction quality. Good luck!