PowerPulse: Household Energy Usage Forecast

**Approach:**

**1. Data Understanding and Exploration**

* Begin by loading the dataset to review its structure.
* Check for missing values, duplicate entries, and the types of data present.

**2. Data Preprocessing**

* Address missing data with suitable techniques.
* Resolve any issues with placeholder values.
* Adjust data types as required.
* Identify outliers using methods such as Interquartile Range (IQR) and Z-score.
* Manage outliers appropriately.

**3. Feature Engineering**

* Develop new features including Hour, Month, Year, Weekend, Weekday, Is\_Night, Residual\_Power, Rolling\_Mean\_10min, Rolling\_Mean\_1h, Rolling\_Mean\_24h, TimeofDay , Season, carbon\_emission, Energy\_Deviation, Anomaly.

**4. Exploratory Data Analysis**

* Examine trends, outliers, and feature skewness.
* Visualize the relationships between variables.

**5. Model Selection and Training**

* Train various regression models using a 0.1% sample from the complete dataset.

**6. Evaluation**

Assess and compare the performance of models using metrics such as Mean Absolute Error (MAE), R² Score, and Root Mean Squared Error (RMSE).

**Data Analysis:**

***1. Data Understanding and Preparation***

* **Data Loading and Examination:** The dataset was loaded from the UCI repository. It includes features related to individual household electric power consumption. The initial examination involved checking the structure of the data, which consists of both features (X) and targets (y), merged into a single DataFrame.
* **Initial Data Cleaning:** The data underwent initial cleaning steps such as combining date and time into a single datetime column, dropping the original date and time columns, and ensuring no duplicate rows existed. Missing values and placeholders marked as '?' were identified and replaced with NaN, followed by the removal of all rows with missing data.

***2. Data Type Conversion***

* **Data Type Adjustments:** Key columns that represent power consumption metrics were converted from object to float types to facilitate numerical analysis. This ensures accuracy in the subsequent statistical and machine learning processes.

***3. Outlier Detection***

* **Outlier Identification:** Outliers were visualized using box plots for key continuous variables such as global active power and voltage. The identification was based on statistical methods like the Interquartile Range (IQR) and Z-score.

***4. Exploratory Data Analysis (EDA)***

* **Trend Analysis:** The data was explored to identify trends and patterns. Visualizations such as line plots were used to understand hourly, weekly, and monthly trends in energy consumption. This helped in uncovering insights such as peak hours, days, or months which could significantly impact energy use forecasts.
* **Correlation Analysis:** Pearson’s correlation coefficients were calculated to understand the relationships between different features. This helped in identifying which variables have a significant influence on energy consumption and informed feature selection for model building.

***5. Feature Reduction***

* **Dropping Redundant Features:** Based on the correlation analysis, features that did not contribute significantly to the power consumption prediction, or that were highly correlated with others, were dropped. This step helps in reducing model complexity and avoiding multicollinearity.

***6. Final Dataset Preparation***

* **Dataset Finalization:** The cleaned and preprocessed data was saved as 'Final\_Data.csv', which will be used for model training and evaluation. This dataset excludes unnecessary features, focusing only on those that are relevant to forecasting energy usage.

**Model Selection & Evaluation:**

#### **Data Preparation**

The dataset, comprising 20 lakh records, was split such that 30% (approximately 600,000 samples) was used for model training and evaluation. The feature set (X) and target variable (y, representing Global\_active\_power) were defined, with the data subsequently divided into training and testing subsets. A 10% test split was used to evaluate model performance, ensuring that training and validation conditions mimicked real-world scenarios as closely as possible.

#### **Model Selection**

Several regression models were selected to address the forecasting problem, acknowledging different strengths and capabilities of each model:

* **Linear Regression**: A baseline for comparison with other models.
* **Random Forest**: Known for its high accuracy in regression problems through ensemble learning.
* **Gradient Boosting**: Boosts the prediction accuracy by focusing sequentially on incorrect predictions.
* **Neural Networks (MLP Regressor)**: Capable of capturing complex nonlinear relationships between variables.
* **XGBoost**: An efficient and scalable implementation of gradient boosting.
* **K-Nearest Neighbors**: Uses feature similarity to predict energy values.

#### **Model Training**

Each model was trained using the training dataset (X\_train, y\_train). The training process involved fitting each model to the data, adjusting parameters where necessary to optimize performance.

#### **Model Evaluation**

Model evaluation was conducted by predicting energy usage both on the training set and on the unseen test set. The evaluation metrics used were:

* **Mean Absolute Error (MAE)**: Represents the average error made by the model in predicting energy usage.
* **Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)**: Measure the average of the squares of the errors, with RMSE providing a scale-sensitive measure of errors between actual and predicted values.
* **R² Score**: Indicates the proportion of variance in the dependent variable that is predictable from the independent variables.

#### **Performance Visualization**

Actual vs. Predicted values were plotted for both the training and testing sets to visually assess the accuracy of each model. These plots provide intuitive insight into the model's prediction capability and error distribution.

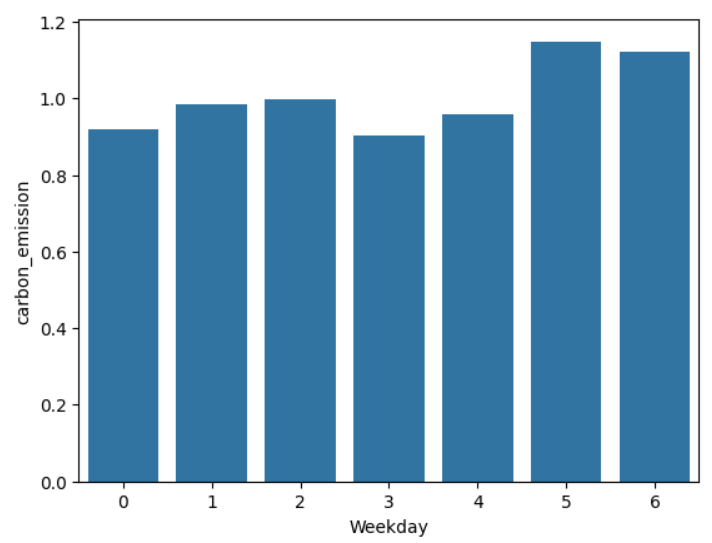
#### **Feature Importance**

For models supporting feature importance analysis (e.g., Random Forest, Gradient Boosting, and potentially Linear Regression through coefficients), the importance of each feature in predicting the target variable was computed and visualized. This analysis helps in understanding which features most significantly impact household energy consumption, guiding future data collection and feature engineering efforts.

#### **Results Compilation**

The performance metrics for each model were compiled into a DataFrame, facilitating a straightforward comparative analysis of how each model performs in terms of error metrics and variance explanation.

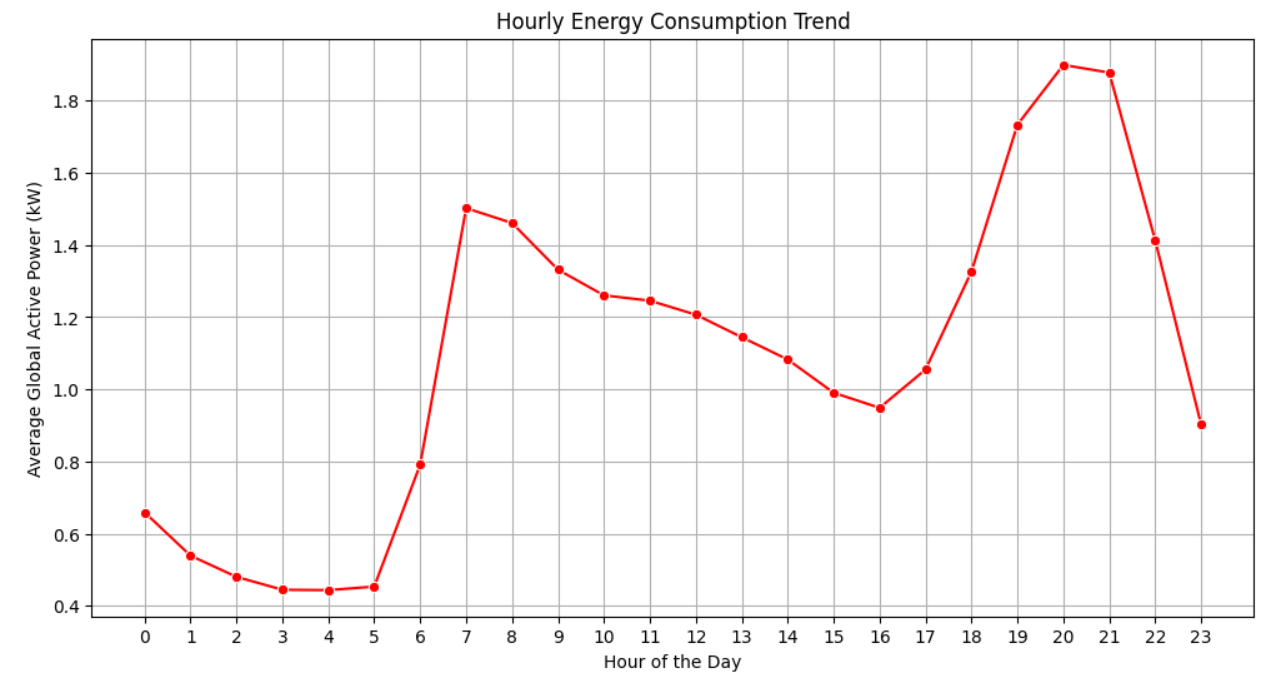
**Insights:**



**Recommendations:**

* **Energy Provider:** Optimize grid operations and integrate renewable sources to manage higher weekend demand and reduce emissions.
* **Energy Consumer:** Monitor weekend energy usage, adopt energy-saving habits, and use smart appliances to minimize carbon footprint.

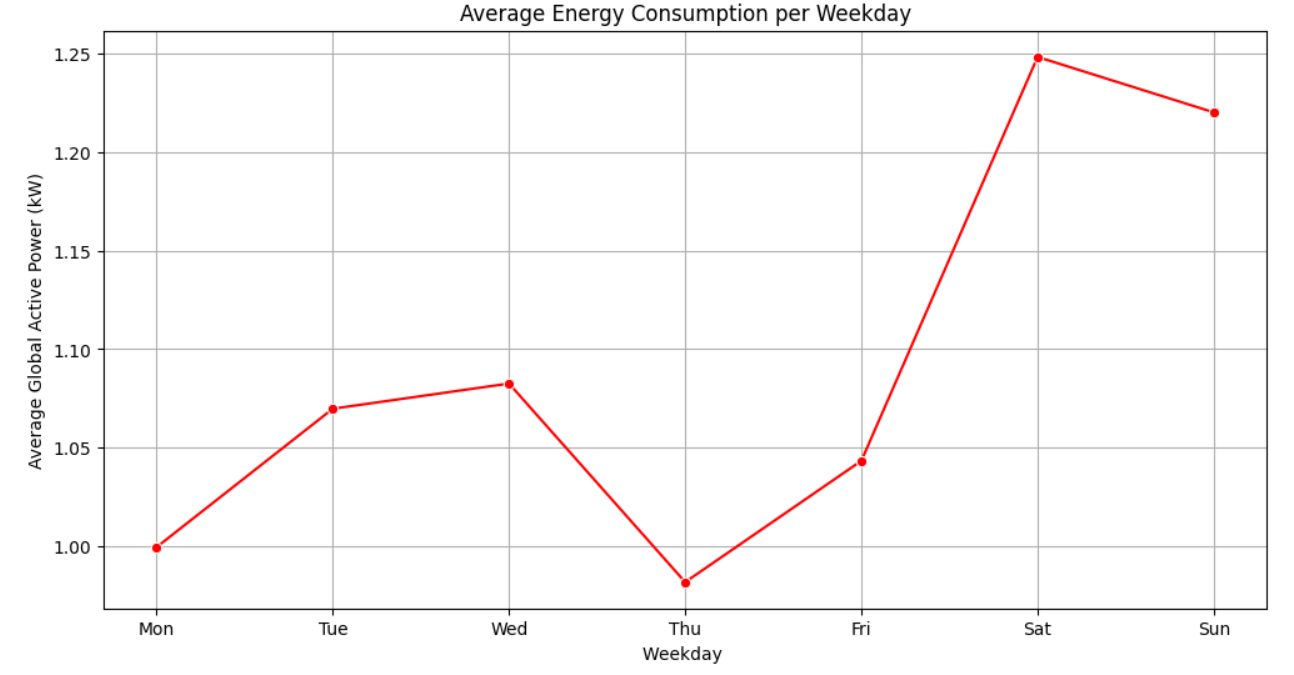
Energy consumption peaks around 7-9 AM and 6-9 PM, indicating high usage during morning and evening routines.



### **Recommendations:**

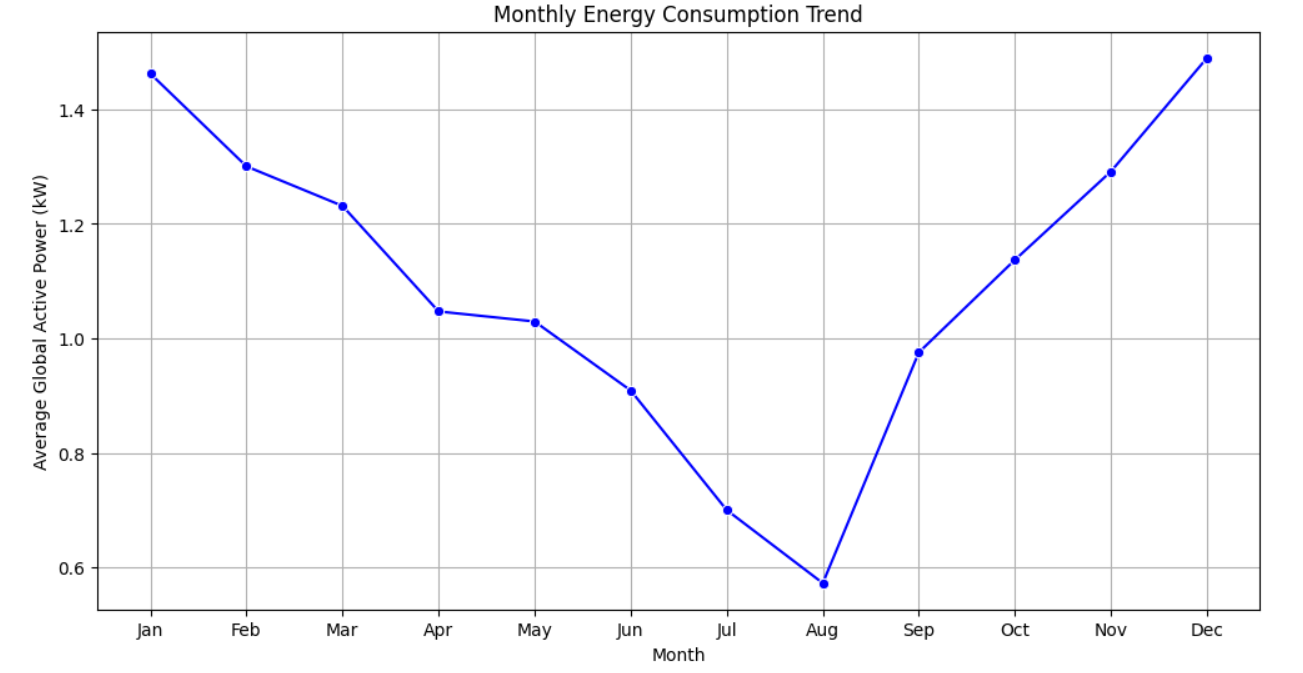
* **Energy Provider:** Manage peak-hour demand by optimizing grid performance and promoting off-peak incentives.
* **Energy Consumer:** Shift non-essential energy use to off-peak hours to reduce costs and prevent grid strain.

Energy consumption is relatively low during weekdays, with a significant increase on weekends, indicating higher household activity and energy use during Saturdays and Sundays.

**Recommendations:**

* **Energy Provider:** Focus on weekend load management and encourage demand response programs to balance peak usage.
* **Energy Consumer:** Plan high-energy tasks on weekdays to reduce costs and prevent unnecessary weekend power surges.

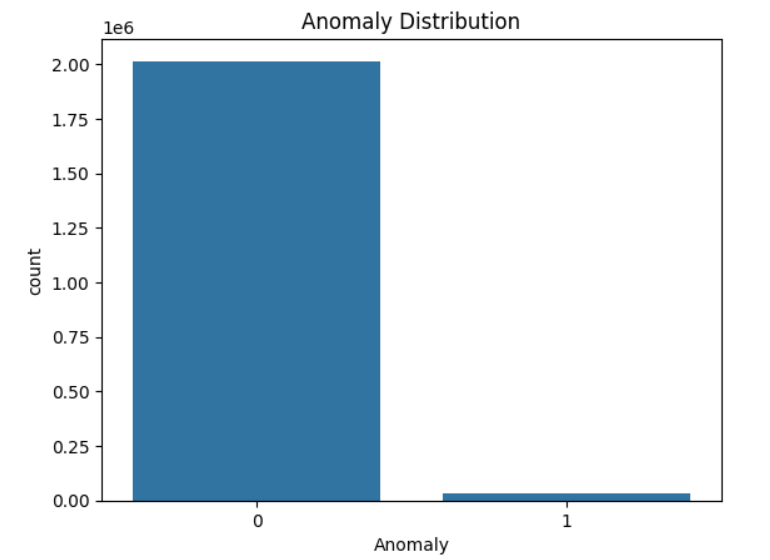
Energy consumption is highest in **January** and **December**, likely due to increased heating or holiday activities, while it significantly drops during **June to August**, indicating reduced usage during summer months.



### **Recommendations:**

* **Energy Provider:** Prepare for higher energy demand during winter months and plan infrastructure accordingly to avoid supply shortages.
* **Energy Consumer:** Optimize energy usage in winter by improving insulation and using energy-efficient appliances to reduce costs.

The anomaly distribution indicates that most data points are normal, with only a small percentage flagged as anomalies, suggesting rare occurrences of unusual energy consumption pattern.

**Recommendations:**

* **Energy Provider:** Implement real-time monitoring and alerts to investigate and respond quickly to anomalies, minimizing potential risks.
* **Energy Consumer:** Regularly monitor energy usage to identify and address unusual spikes that may indicate faulty appliances or unauthorized energy consumption.