# Project Report: PowerPulse: Household Energy Usage Forecast

**Project Goal:**

To develop a highly accurate machine learning model to predict household energy consumption (Global Active Power) using historical, minute-by-minute data, and provide actionable insights into consumption drivers.

**Best Model Selected:** Extreme Gradient Boosting (XGBoost)

## 1. Approach and Methodology

The project followed a standard time-series machine learning pipeline, strictly maintaining the chronological order of data to ensure the training data chronologically precedes the test data.

### 1.1 Data Source

The dataset used was the Individual Household Electric Power Consumption dataset, containing 47 months of minute-by-minute data (over 2 million observations).

### 1.2 Data Splitting Strategy

Due to the time-series nature of the data, a sequential split was used to prevent data leakage:

- Training Set: The first 80% of the data (historical consumption)
- Testing Set: The last 20% of the data (future consumption)

## 2. Data Analysis and Preprocessing

The initial Exploratory Data Analysis (EDA) revealed several critical data quality issues that required meticulous cleaning:

### 2.1 Initial Cleaning and Conversion

- Missing Values: The original data contained '?' symbols representing missing values in all measurement columns (totaling 25,979 missing points). These were converted to NaN.
- Time Series Indexing: The separate 'Date' and 'Time' columns were combined, converted to a Datetime object, and set as the DataFrame's index.
- Imputation: Missing values were filled using a combination of **Forward Fill (ffill)** and **Backward Fill (bfill)** to maintain data continuity in the time series.

### 2.2 Feature Engineering

The model's accuracy was enhanced by deriving features from the raw data:

- Time-Based Features: Hour, DayOfWeek, Month, Year, and WeekOfYear were extracted to capture seasonal and cyclical consumption patterns.
- Lagged Feature: A **24-hour Rolling Mean (Rolling_Mean_24h)** of the target variable was created and lagged by one period.

### 2.3 Feature Scaling

All numerical features (e.g., Voltage, Global_intensity) were scaled using **MinMaxScaler** to normalize their range between 0 and 1.

## 3. Model Selection and Evaluation

Three distinct regression models were trained and evaluated on the test set.

### 3.1 Model Comparison Table

The models were assessed using the required project metrics: RMSE, MAE, and R-Squared (R^2).

| **Model** | **Root Mean Squared Error (RMSE)** | **Mean Absolute Error (MAE)** | **R-Squared (R2)** |
|---|---|---|---|
| Linear Regression (Baseline) | 0.0390 | 0.0234 | 0.9980 |
| Random Forest Regressor | 0.0400 | 0.0251 | 0.9979 |
| **XGBoost Regressor (Selected)** | 0.0356 | 0.0236 | **0.9984** |

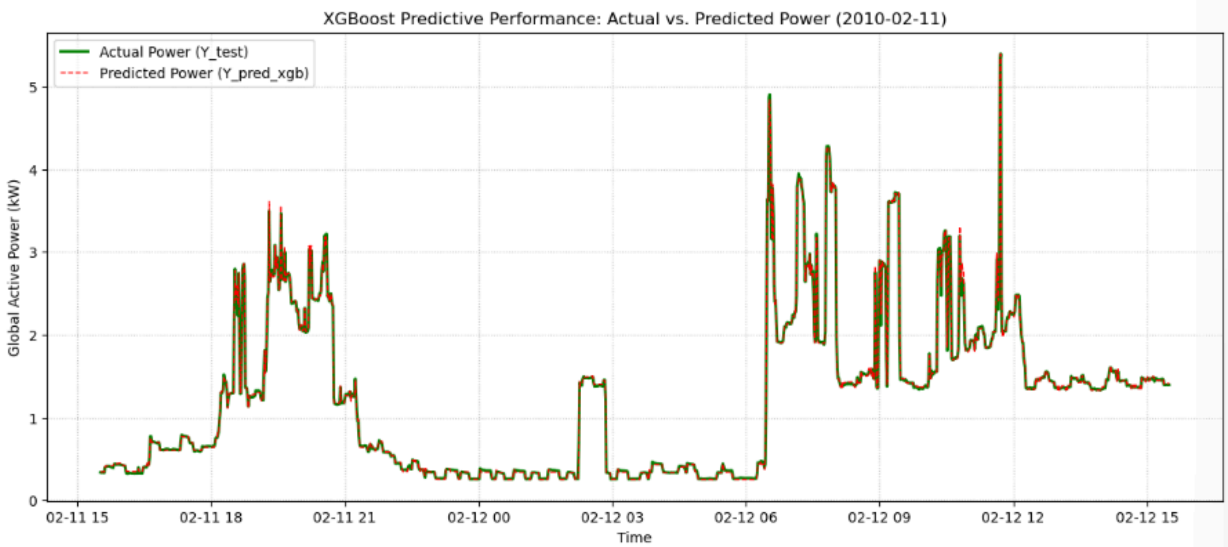
### 3.2 Selected Model and Performance

The **XGBoost Regressor** was selected as the final model due to its superior performance across all metrics, achieving the lowest RMSE (0.0356) and the highest explanatory power (R^2 of 0.9984).

### 3.3 Visualization of Predictive Performance

The plot below demonstrates the model's high accuracy on unseen test data.

XGBoost Predictive Performance: Actual vs. Predicted Power

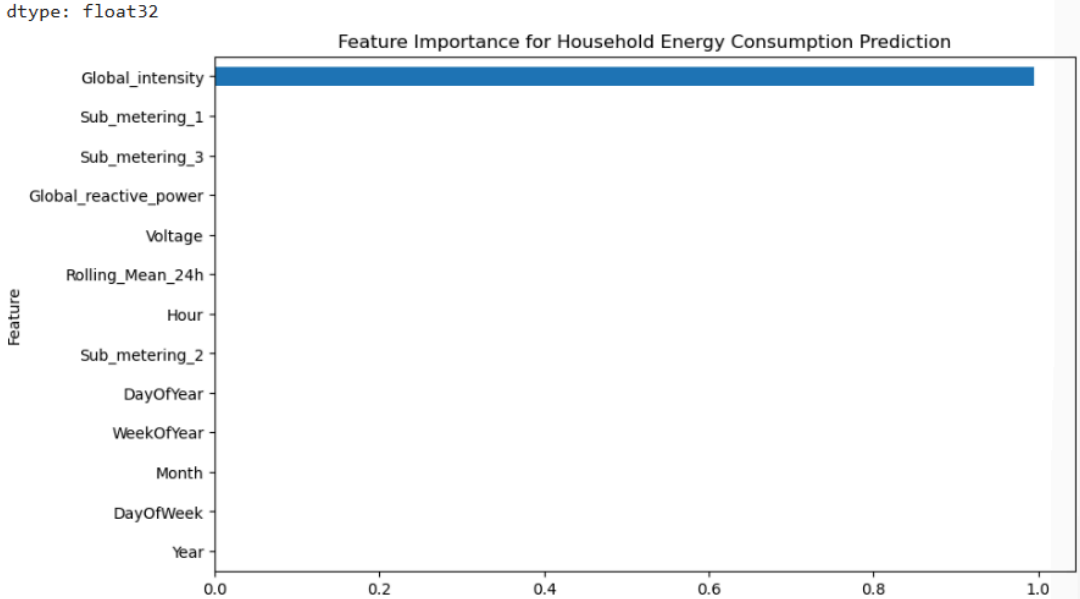



# 4. Insights and Recommendations

## 4.1 Feature Importance Analysis

The Feature Importance plot reveals the drivers of consumption:

```
Top 5 Most Important Features:
Global_intensity         0.995407
Sub_metering_1           0.001034
Sub_metering_3           0.000919
Global_reactive_power    0.000726
Voltage                  0.000691
dtype: float32
```



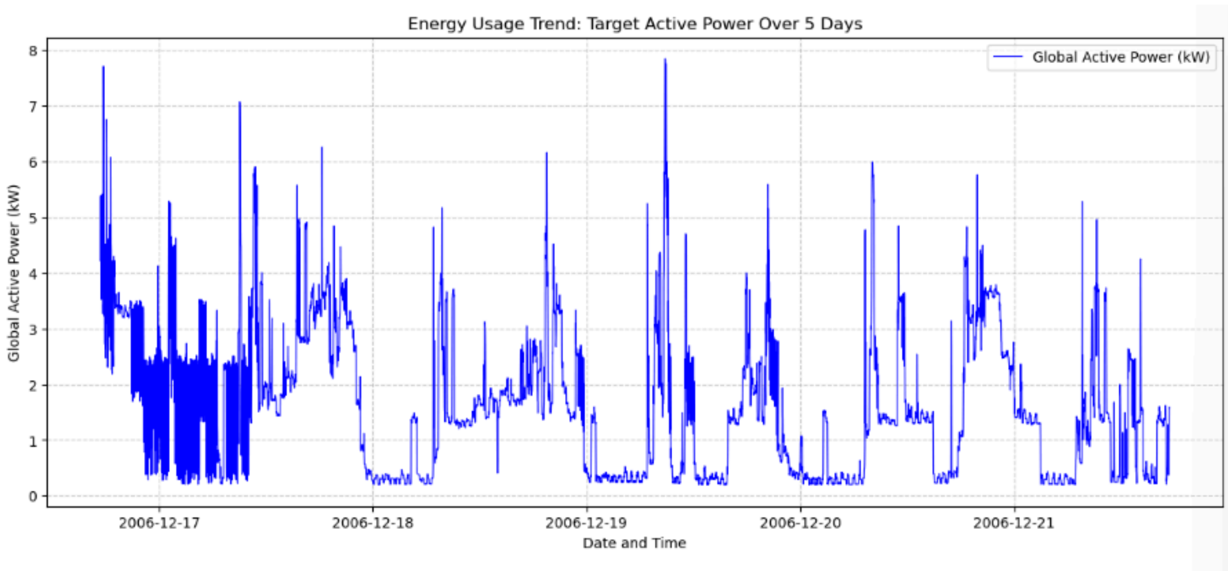
The most influential features are:

1. **Global_intensity:** ~99.5% Importance
2. **Sub_metering_1 & Sub_metering_3:** ~0.1% combined Importance

This overwhelmingly demonstrates that the total power consumed is primarily governed by the instantaneous current draw (Global_intensity).

### 4.2 Energy Trends Visualization

The Energy Trends plot illustrates the predictable cyclical behavior of consumption, showing clear daily peaks and low overnight usage.




### 4.3 Actionable Recommendations

Based on the model performance and feature insights, the following recommendations are made:

1. **Focus on Current Draw:** Future monitoring systems should prioritize high-frequency current sampling, as this metric provides the most immediate predictive value.
2. **Targeted Energy Management:** To reduce overall consumption, households should focus on the specific circuits tracked by Sub-metering 1 and 3.
3. **Real-Time Anomaly Detection:** The highly accurate XGBoost model can be deployed for real-time anomaly detection, signaling a fault or high-consumption event.