Power Pulse: Household Energy usage forecast

Domain: Energy and utilities

Name: Rajarajan N

A comprehensive report summarizing

Approach

This project follows a structured pipeline for data science and machine learning, comprising the following steps:

1.Data Loading & Exploration

- Loaded the dataset containing minute-level measurements of power usage in a household over several years.
- Explored the data types, structure, and volume.
- Conducted initial investigations:

  Time range and frequency of the data

  Power consumption patterns Relationship between different features (e.g., active power, voltage, sub-metering)

2. Exploratory Data Analysis (EDA)

- Time series plots of global active power to understand trends, seasonality, and anomalies.
- Correlation heatmaps to identify feature dependencies.
- Distribution plots for numeric variables.
- Boxplots and line plots for sub-metering analysis.

3. Handling Missing Data

- Identified missing timestamps and power usage values.
- Handled missing data using:
  - Handled based on average of weekend and weekday

4. Feature Engineering

Constructed meaningful features to enhance model performance:

- **Time-based features**:
  - Hour of the day
  - Date
  - Time
  - Weekday
  - Weekend
  - Peak hours
- **Rolling statistics**:
  - Gap rolling 60min
  - Daily avg

5. Post-Feature Engineering EDA

- Re-evaluated distributions of newly created features.
- Plots based on Hour of days
- Rolling avg based on over 60mins
- Average global active power by weekday and also by month
- Peak load prediction based on daily avg of global active power
- Anomaly detection using z score and Estimated daily CO2 emission

6. Outlier Detection and Transformation

- **Square Root and Log Transformations**:

  Used to reduce the effect of large values and make feature distributions more balanced.

- **Box-Cox Transformation**:

  Applied to features with only positive values to make them more normally distributed and reduce skewness.

- **IQR Method (after Box-Cox)**:

  For some features, even after Box-Cox, outliers remained.
  We used the **Interquartile Range (IQR)** method to detect and fix these:

  - Values outside the normal range (Q1 - 1.5×IQR to Q3 + 1.5×IQR) were treated or capped.

These steps helped clean the data and made it more suitable for modeling.

7. Feature Scaling

- Applied **StandardScaler** to normalize features, especially for models sensitive to feature magnitudes (e.g., Linear Regression).
- Ensured scaling was fitted only on training data to avoid data leakage and applied consistently to test data.

8. Feature Selection

- Performed feature selection using:

  Correlation analysis to remove redundant features

  Tree-based feature importance from Random Forest/XGBoost

9. Model Building and Training

Developed multiple regression models for comparison:

- **Linear Regression**: For baseline performance and interpretability.
- **Random Forest Regressor**: To capture non-linear relationships and interactions.
- **Gradient Boosting Regressor**: For sequential error correction and improved accuracy.
- **XGBoost Regressor**: A highly optimized boosting model that handles regularization and missing values effectively.

10. Hyperparameter Tuning

- Used **GridSearchCV** and **RandomizedSearchCV** with cross-validation to tune model hyperparameters such as:
    - `n_estimators`, `max_depth`, `learning_rate`, `subsample`, `min_samples_split`

11. Model Evaluation and Selection

- Evaluated models using:
    - **R^2 Score**: Goodness of fit
    - **RMSE** (Root Mean Squared Error): Penalizes large errors
    - **MAE** (Mean Absolute Error): Measures average prediction error
- Chose the best model based on a combination of:
    - High test R^2
    - Low RMSE and MAE
    - Generalization gap (train vs. test performance)

📊 Data Analysis

In this section, we explore various temporal and statistical patterns in household energy consumption to understand usage behavior, detect anomalies, and extract actionable insights. The following analyses were performed using the `global_active_power` and sub-metering features recorded at minute-level granularity.

1. Average Usage by Hour of the Day (All Features)

Calculated hourly averages for `global_active_power`, 'global intensity', 'voltage', 'global reaction power', `sub_metering_1`, `sub_metering_2`, and `sub_metering_3`.

Findings:

- Clear peak usage observed in the morning (6–9) and night (19–21) hours.

- Sub-metering usage shows appliance-specific patterns (e.g., kitchen appliances peak during breakfast/lunch times).

2. Daily Average of Global Active Power

Aggregated minute-wise data into daily averages of `global_active_power`.

Findings:

- Seasonal trends were observable.

- Weekends showed marginally higher average consumption.

3. Rolling Average over 60 Minutes

Applied a 60-minute rolling window to smooth high-frequency noise.

Insights:

- Captures micro-level behaviors.

- Helps in load forecasting at hourly resolution.

4. One Day Trend (Raw Global Active Power)

Visualized minute-level `global_active_power` for a single selected day.

Insights:

- Sharp morning and evening peaks indicate typical workday schedules.

- Sudden spikes suggest use of high-power appliances.

5. Smoothed Weekly Trend (7-Day Moving Average)

Computed a 7-day moving average to understand weekly consumption.

Findings:

- Smooth long-term trends help identify anomalies.

- Holidays and seasonal events clearly visible.

6. Average Usage by Weekday

Aggregated data by weekday.

Insights:

- Weekends generally show higher usage.

- Mondays and Thursdays tend to have the lowest usage.

7. Average Global Active Power by Month

Monthly aggregation performed to analyze seasonal effects.

Findings:

- Increase during winter (likely heating).

- Summer peaks may indicate cooling device usage.

8. Use Case: Peak Load Prediction (Visual Overlay)

Overlaid daily usage to find peak demand windows.

Observation:

- Consistent evening peak load between 6–9 PM.

9. Anomaly Detection Using Z-score

Applied Z-score normalization to detect outliers.

Use Case:

- Detect appliance faults, unauthorized usage, or errors.

Result:

- Some days flagged with unusually high readings.

10. Estimate of CO_2 Emissions Using Energy Usage

Estimated CO_2 using formula:

CO_2 (kg) = Energy (kWh) × Emission Factor (0.475 kg CO_2/kWh)

Outcome:

- Daily/monthly CO_2 footprints computed.

- Useful for environmental tracking.

11. Smart Grid Integration (All Sub-Metering)

Analyzed all sub-metering channels to simulate appliance-level insights.

Insights:

- Each sub-metering shows distinct patterns.

Value:

- Supports smart grid applications like demand-side management and dynamic pricing.

-AC/Heater (sub meter 3) is the highest consumer of electricity

Model Selection and Evaluation

Several regression models were trained and evaluated using performance metrics including RMSE (Root Mean Squared Error), MSE (Mean Squared Error), R-squared (Coefficient of Determination), and MAE (Mean Absolute Error). The goal was to balance accuracy, generalization, and interpretability.

Linear Regression

Train:

RMSE = 0.3442, R^2 = 0.8928, MAE = 0.2301

Test:

RMSE = 0.3424, R^2 = 0.8932, MAE = 0.2293

Random Forest Regressor

Train:

RMSE = 0.0133, R^2 = 0.9998, MAE = 0.0071

Test:

RMSE = 0.0309, R^2 = 0.9991, MAE = 0.0170

Random Forest Regressor (GridSearch Tuned)

Train:

RMSE = 0.0572, R^2 = 0.9970, MAE = 0.0376

Test:

RMSE = 0.0574, R^2 = 0.9970, MAE = 0.0376

Gradient Boosting Regressor

Train:

RMSE = 0.0340, R^2 = 0.9990, MAE = 0.0214

Test:

RMSE = 0.0345, R^2 = 0.9989, MAE = 0.0215

XGBoost Regressor

Train:

RMSE = 0.0352, R^2 = 0.9989, MAE = 0.0222

Test:

RMSE = 0.0357, R^2 = 0.9988, MAE = 0.0223

☑ Selected Model: Random Forest Regressor (GridSearch)

The selected model is the Random Forest Regressor tuned using GridSearch. This model offered a strong balance of high accuracy, stability, and interpretability.

Reasons for selection:

☑ $R^2 \approx 0.997$ on test data → Excellent generalization

☑ Very small gap between train and test scores → Indicates low overfitting

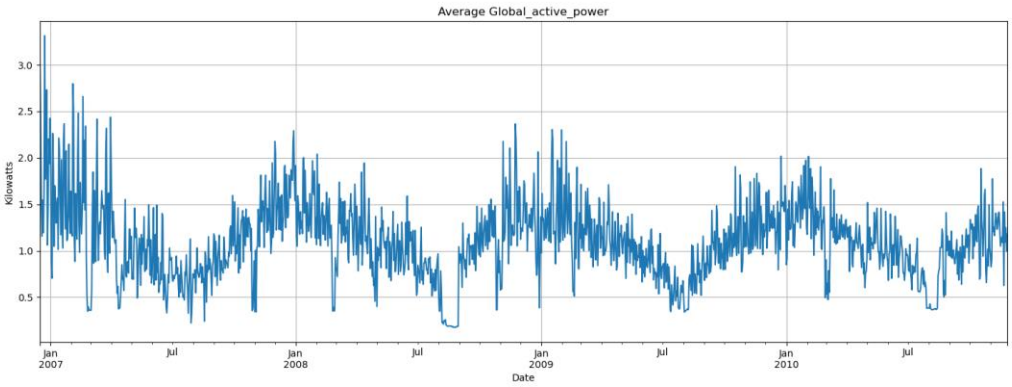
☑ Less complex than boosting models

☑ Easier to interpret using feature importance

☑ Faster to train and debug compared to Gradient Boosting or XGBoost

Insights and Recommendations

Average Global active power



📉 Recommendations

Investigate Early 2007 High Usage

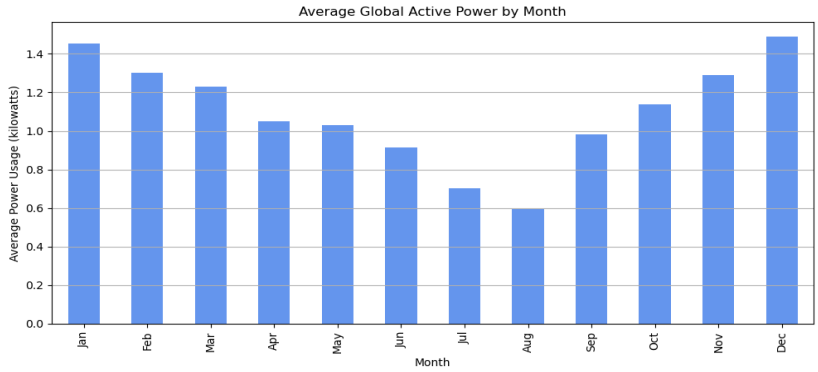
The sharp spike in power usage in early 2007 suggests either **anomalies or operational inefficiencies**.

Conduct a retrospective audit to identify root causes and avoid future occurrences.

Promote Seasonal Energy Conservation

Power usage appears to **fluctuate seasonally**. Encourage households to adopt **seasonal energy-saving practices** (e.g., insulation in winter, passive cooling in summer) to smooth out consumption peaks.

Average Global Active Power by Month



Recommendations

- Target Summer Months for Efficiency Campaigns

  Power usage is significantly lower in July and August, suggesting lower demand. Run maintenance or upgrades during these months to minimize service disruptions and prepare for peak seasons.

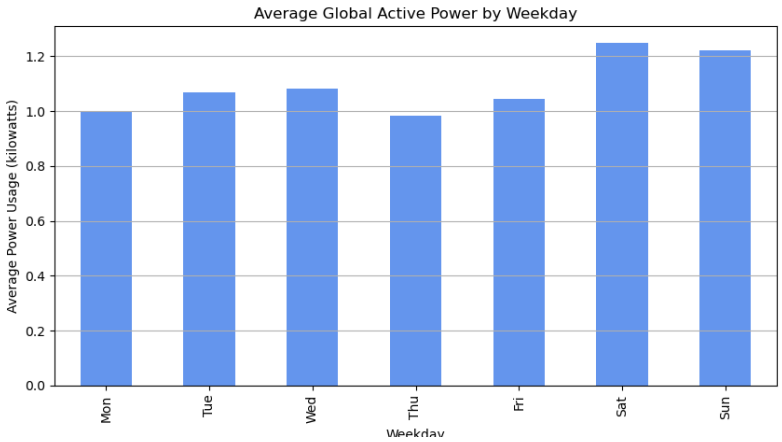
- Introduce Load Management in Winter and December

  The highest usage in December and January indicates heating appliances or festive energy consumption. Introduce incentivized off-peak usage and smart grid-based load balancing strategies to prevent overload.

- Seasonal Pricing Models

  Implement seasonal tariff adjustments (e.g., higher rates during winter peaks, discounts in summer) to encourage users to distribute their energy usage more evenly throughout the year.

Average Global Active Power by Weekday



Recommendations

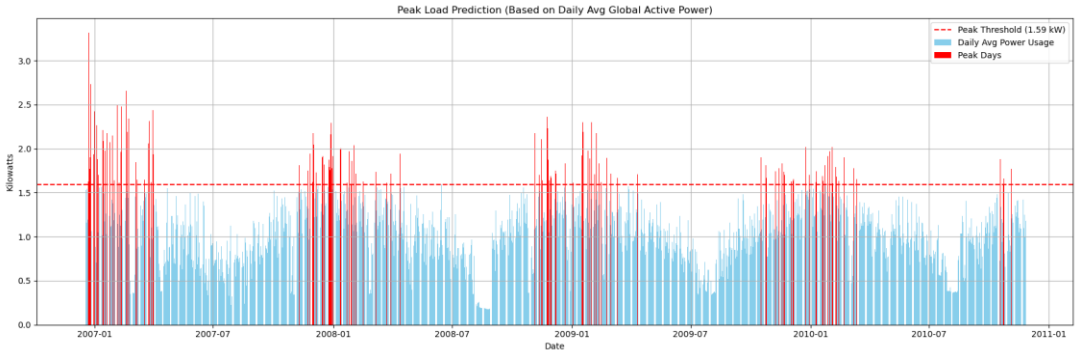
- **Optimize Weekend Load Scheduling**

  Power usage peaks on **Saturday and Sunday**. Encourage users or systems to shift non-essential tasks (e.g., laundry, water heating) to weekdays to balance the load and reduce weekend demand.

- **Smart Grid Load Balancing**

  Use **predictive analytics and smart grid technology** to prepare for high loads during weekends by pre-charging energy storage systems or adjusting renewable energy inputs accordingly.

Peak Load Prediction (Based on Daily Avg Global Active Power)



Key Points

- **Defined Peak Load Threshold**: A peak threshold of **1.59 kW** is established as the benchmark for identifying high-load days.
- **Frequent Exceedance During Early Years**: The years **2007 and 2008** show a **high density of red spikes**, indicating frequent peak load breaches.

Recommendations

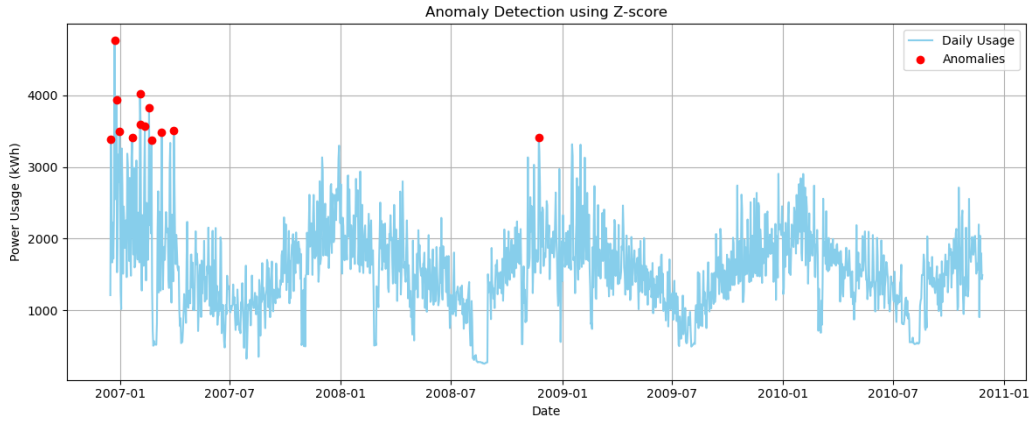
- **Investigate Early Peak Trends**

  Analyze causes for excessive peaks in **2007-08**, such as inefficient appliances, seasonal effects, or occupancy levels. This can guide mitigation strategies for similar future conditions.

- **Maintain Post-2009 Efficiency**

  Preserve the stability observed after 2009 by continuing **consumer awareness, smart metering, and peak-time tariffs**.

Anomaly Detection using Z-score



Key Points

- **Early Period Spikes**

  A **cluster of anomalies** is clearly visible in **early 2007**, where daily usage surpassed **4000 kWh**, significantly higher than average consumption.

- **Sharp Contrast with Later Periods**

  Post-2007, **few anomalies are detected**, suggesting either **behavioral changes**, better energy efficiency, or **data stabilization**.

Recommendations

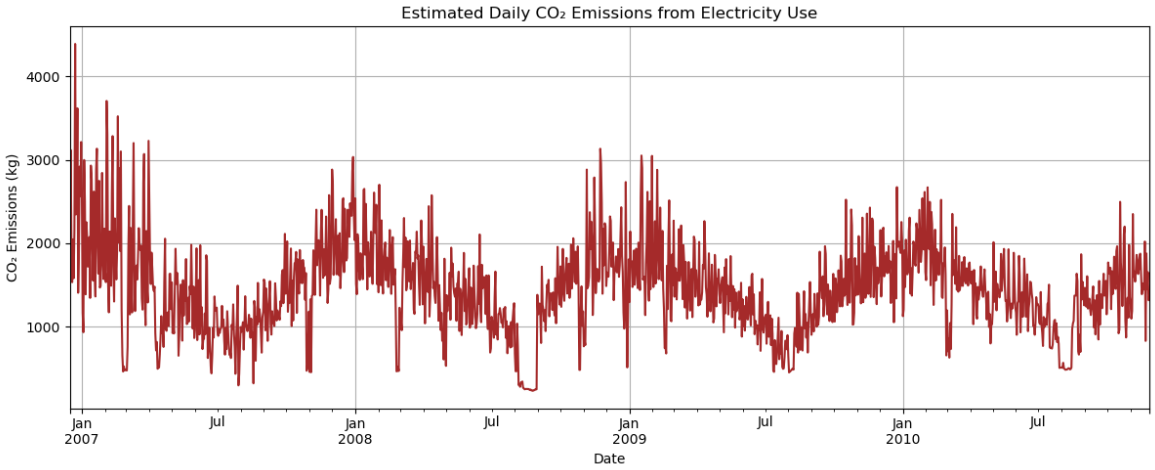
- **Root Cause Analysis of 2007 Anomalies**

  Investigate **external events, occupancy spikes, or malfunctioning devices** that could have caused extremely high usage days.

- **Monitor Ongoing Consumption Trends**

  Use similar anomaly detection approaches in real-time systems to **flag abnormal usage** early and take corrective actions.

Estimated Daily CO_2 Emissions from Electricity Use



Key Points

- **Early Emissions Were Very High**

  In early 2007, daily CO_2 emissions peaked at **over 4,000 kg**, indicating high energy usage or inefficiencies during that period.

- **Gradual Decrease and Stabilization**

  There's a **noticeable decline** in emissions after mid-2007, followed by **cyclical stability**—suggesting improved energy efficiency or reduced usage.

Recommendations

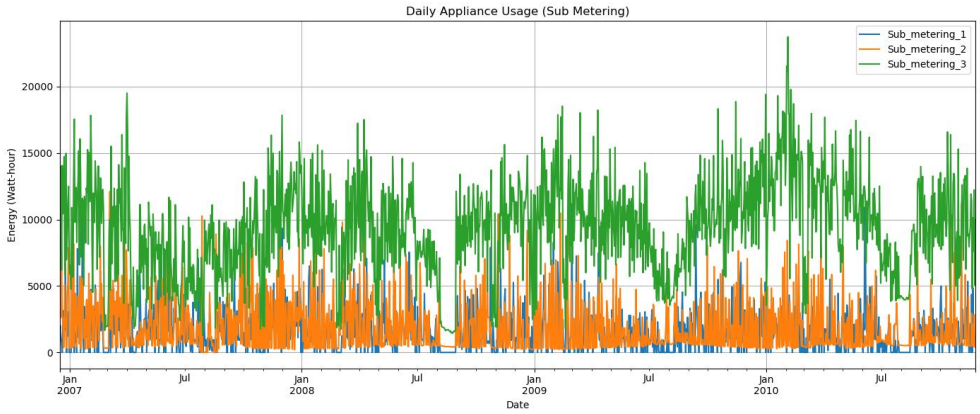
- **Continue Promoting Energy Efficiency**

  The decline over time shows improvement—**continue efforts** in using **efficient appliances** and **renewable sources** to maintain low CO_2 output.

- **Investigate Emission Spikes**

  Analyze periods with high spikes in emissions (like early 2007 and early 2009) to **identify operational or behavioral causes**.

Daily Appliance Usage (Sub Metering)



Key Points

- **Sub_metering_3 Dominates**

  Throughout the entire time period, **Sub_metering_3** consistently shows the **highest daily energy consumption**, peaking above **20,000 watt-hours**.

- **Seasonal and Cyclical Patterns**

  All three sub-meterings display **regular fluctuation**, potentially reflecting **seasonal appliance usage** (e.g., heaters, ACs, or cooking appliances).

Recommendations

- **Target Efficiency Improvements in Sub_metering_3**

  Since it contributes the **bulk of daily consumption**, it's ideal to audit appliances linked to this channel and consider **upgrades or usage optimizations**.

- **Investigate Missing Data Periods**

  Ensure **hardware reliability** by reviewing logging systems and performing **maintenance checks** on sub-metering devices.

- **Appliance-specific Consumption Forecasting**

  Apply **machine learning** models on sub-metered data to **predict appliance load** and **automate control strategies** for demand response.

Conclusion

This project provides an end-to-end pipeline for analyzing and forecasting household energy consumption.

Through robust data preprocessing, insightful visualizations, and model experimentation, the Random Forest model (with tuned hyperparameters)

emerged as the most reliable and interpretable solution for power usage forecasting. It offers high accuracy with minimal overfitting,

and its interpretability via feature importance makes it highly suitable for smart grid integration and future energy planning.

These insights can support smarter consumption strategies, peak load management, and environmental impact assessments,

thereby promoting energy efficiency and sustainability in smart homes and cities.