Predicting Household Electric Power Consumption

Name: Ritesh Khatri

University id: 2548382

Assignment Due Date: 09-10/2026

Module leader: Siman Giri

Tutor: Robin Tuladhar

Abstract

Purpose: The goal of this report is to predict household global active power using regression techniques.
Dataset: The dataset used is Individual Household Electric Power Consumption, containing 2,075,259 records and 9 features.
UNSDG Link: This aligns with UN SDG 7 (Affordable and Clean Energy) by analyzing household energy consumption patterns.
Approach: The methodology includes Exploratory Data Analysis (EDA), building regression models including a Neural Network, hyper-parameter optimization, feature selection, and final model comparison.
Key Results: Models were evaluated using MAE, RMSE, and R^2. Ridge Regression and Random Forest achieved high predictive performance with minimal error.
Conclusion: The final models demonstrate accurate predictions, highlight influential features, and show the benefit of tuning and feature selection.

Table of Contents

1. Introduction

1.1 Problem Statement

The objective of this project is to develop predictive models that accurately estimate household Global Active Power, a continuous target variable, based on electrical measurements. Accurate prediction of energy consumption can help in energy efficiency, load management, and planning in smart homes.

1.2 Dataset

The dataset, Individual Household Electric Power Consumption, is sourced from the UCI Machine Learning Repository. It contains 2,075,259 observations over nearly four years, with 9 features including Global Active and Reactive Power, Voltage, Global Intensity, and sub-metering values for kitchen, laundry, and water heating appliances. The dataset aligns with UNSDG 7 by providing insights into household energy usage patterns.

1.3 Objective

- Predict Global Active Power accurately using multivariate features.
- Identify the most influential predictors.
- Compare classical and neural network regression models.
- Demonstrate the impact of hyperparameter tuning and feature selection.

2. Methodology

2.1 Data Preprocessing

- Missing values were handled using appropriate imputation strategies.
- Datetime conversion was performed for proper time-series analysis.
- All numeric variables were checked for consistency and scaled where required, particularly for linear models.
- Outliers and unusual values were inspected, especially in sub-metering and intensity features.

2.2 Exploratory Data Analysis (EDA)

EDA was performed to understand distributions, ranges, central tendencies, and relationships among variables:

- Distribution: Sub-metering features were highly skewed, with many zero values. Global Intensity and Global Active Power showed wider ranges.
- Correlation: Global Active Power was strongly correlated with Global Intensity, suggesting high dependency. Voltage had moderate influence, while other sub-meterings contributed less individually.
- Visualizations: Histograms, scatter plots, line plots of daily averages, and correlation heatmaps were used to identify trends, patterns, and potential influential features.

Key insights from EDA:

1. Global Intensity is the most influential predictor.
2. Sub-meterings provide appliance-level usage information.
3. Voltage is relatively stable but has small fluctuations that may impact predictions.

2.3 Model Building

Neural Network

- A regression neural network (MLP) with 2 hidden layers was built.
- Activation functions were ReLU with a continuous output layer.
- The network was trained using MSE as the loss function and Adam optimizer.
- Resampling strategies and validation splits were applied to avoid overfitting.

Classical Regression Models

Two classical models were considered:

1. Linear Regression (Ridge)

o Alpha parameter tuned using GridSearchCV.

o Selected features were top 5 correlated with the target.

2. Random Forest Regressor

o Hyperparameters tuned using RandomizedSearchCV.

o Selected features based on feature importance from the best estimator.

o Sampling strategies were applied during tuning to reduce computation time due to dataset size.

2.4 Model Evaluation

Models were evaluated using:

- MAE: Measures average magnitude of errors.
- RMSE: Provides error in the same units as the target and penalizes larger errors.
- R^2 score: Measures goodness-of-fit of the model.

Observations:

- The Neural Network achieved high predictive performance but required longer training times.
- Ridge Regression and Random Forest achieved excellent accuracy, with Ridge slightly outperforming RF on the test set.

2.5 Hyperparameter Optimization

•Ridge Regression: Optimal alpha = 0.01, CV R^2 = 0.9985

•Random Forest: Best parameters: n_estimators=100, max_depth=5, min_samples_split=2, min_samples_leaf=1, CV R^2 = 0.9977

- Sampling was applied during Random Forest tuning to improve computation efficiency.

2.6 Feature Selection

- Ridge Regression: Top 5 correlated features selected: Global Intensity, Sub-metering 3, Sub-metering 1, Sub-metering 2, Voltage.
- Random Forest: Top 4 features selected based on feature importance: Global Intensity, Global Reactive Power, Voltage, Sub-metering 1.

Feature selection improved model interpretability and reduced complexity without significant loss in predictive performance.

---

3. Results and Conclusion

The analysis of daily average global active power consumption over the period from 2007 to 2010 reveals clear patterns and fluctuations in energy usage. The data indicates both short-term variations and longer-term trends, with periods of higher consumption interspersed with noticeable drops. These patterns provide insight into the dynamics of global energy demand and highlight the temporal variability in power usage. Understanding these trends is crucial for optimizing energy management and predicting future consumption behaviors.

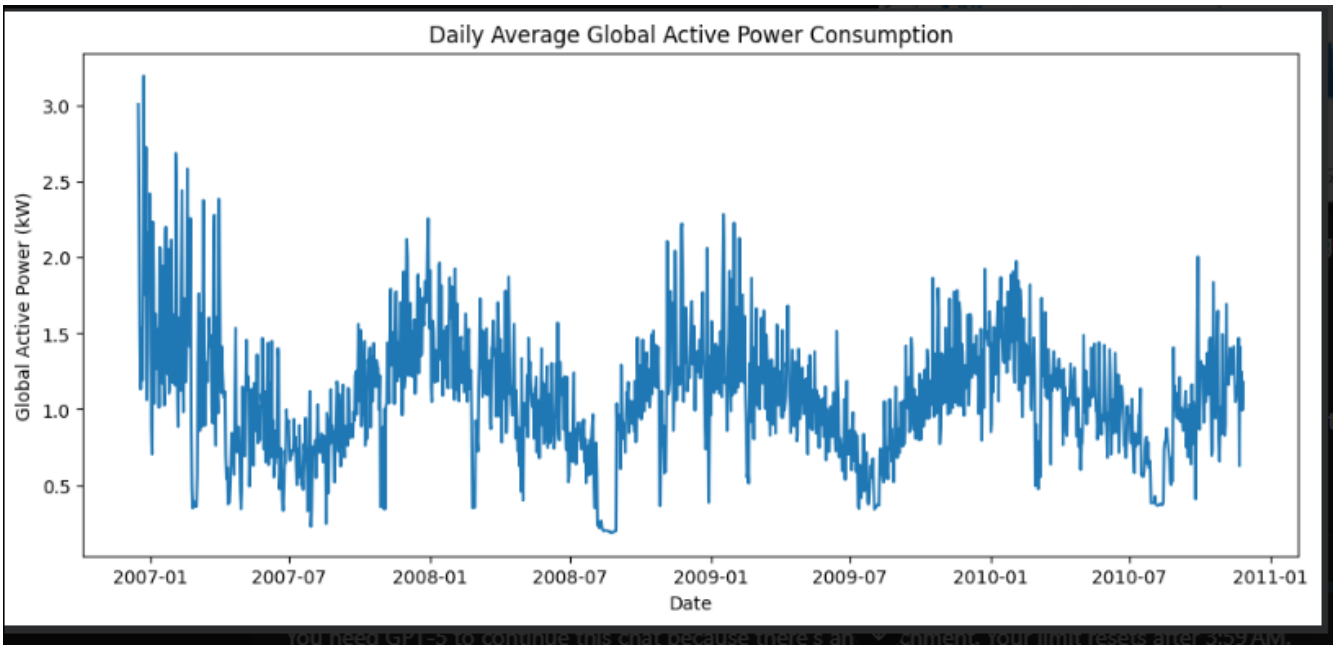

Figure 1: Daily Average Global Active Power Consumption

This line graph illustrates the fluctuations in household electricity usage (measured in kilowatts) over a four-year period, from late 2006 through late 2010.

Key Observations

- **Seasonality (Cyclical Patterns):** The most striking feature is the recurring "wave" pattern. You can see distinct peaks occurring around **January** of each year (2007, 2008, 2009, and 2010). This suggests higher energy consumption during winter months—likely due to heating and increased indoor lighting—and lower consumption during the summer months (July/August).
- **Volatility:** The data shows high daily variance. Even within a specific season, power consumption swings sharply between roughly **0.5 kW and 2.5 kW**, reflecting the day-to-day unpredictability of household activities.
- **Outliers and Anomalies:**
    - **Late 2006:** There is an unusually high spike at the very beginning of the dataset, exceeding **3.0 kW**.
    - **Late 2008:** There is a noticeable "flat" period where consumption drops significantly and stays very low for a short duration, which could indicate a period of absence (like a vacation) or a data collection gap.

- **Overall Trend:** Despite the seasonal swings, the baseline for power consumption remains relatively stable over the four-year period, generally oscillating between a floor of **0.5 kW** and a ceiling of **2.0 kW**.

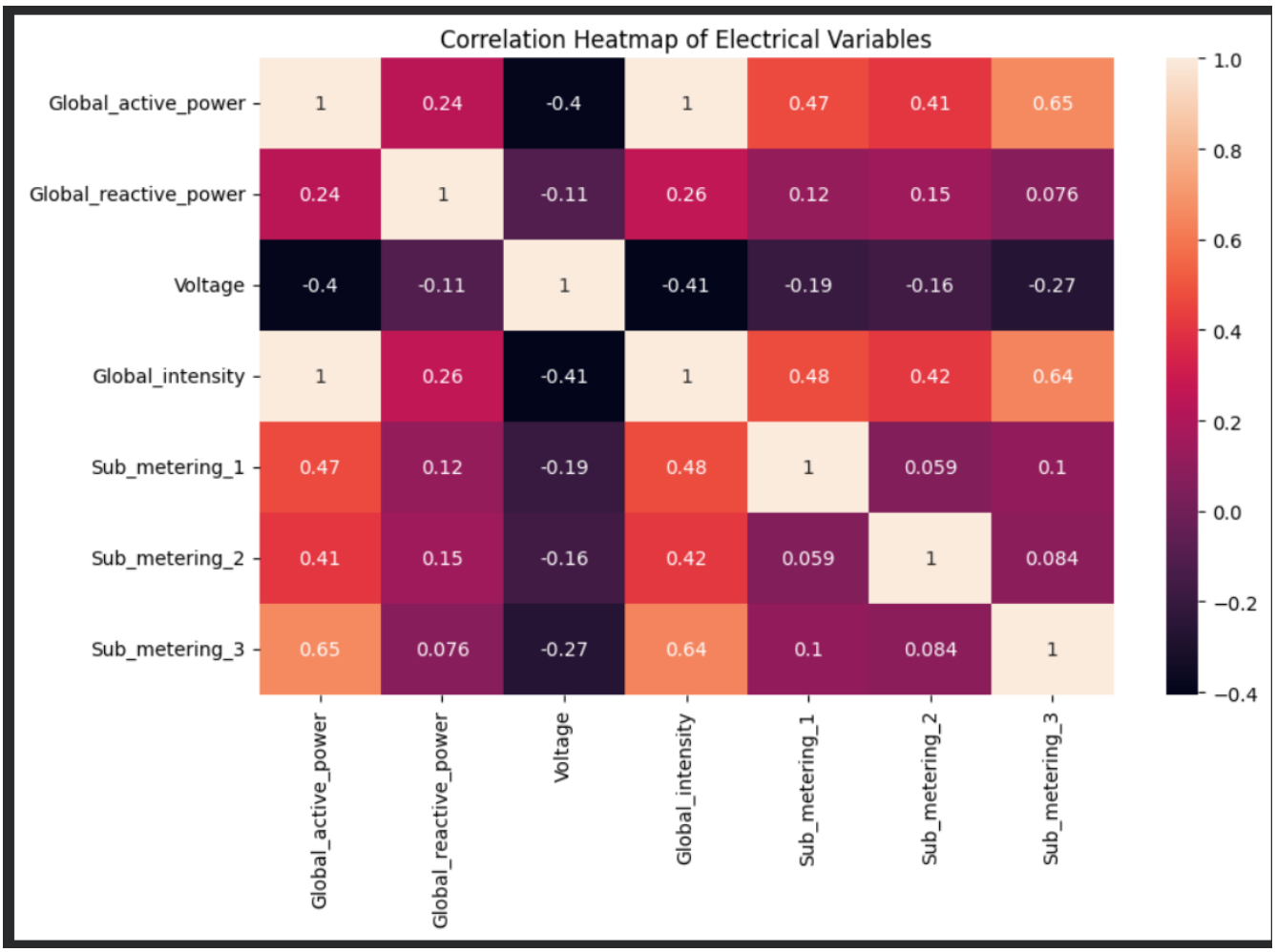

Figure 2: Correlation Heatmap of Electrical Variables

A correlation coefficient ranges from **-1 to 1**. A value of **1** indicates a perfect positive relationship, **-1** indicates a perfect negative relationship, and **0** indicates no linear relationship.

Key Findings from the Heatmap

- **The "Perfect" Correlation (1.00):**
    - **Global_active_power** and **Global_intensity** have a correlation of **1**. This makes sense physically because active power is directly proportional to current (intensity) when voltage is relatively stable. They are essentially measuring the same phenomenon.
- **Strong Positive Relationships:**
    - **Sub_metering_3** shows the strongest link to **Global_active_power (0.65)** compared to the other sub-metered areas. This suggests that whatever equipment is on sub-meter 3 (often electric water heaters or air conditioners) is a major driver of total energy use.

- **Sub_metering_1** and **Sub_metering_2** have moderate correlations with active power (**0.47** and **0.41** respectively).

- ## Negative Correlations:
  - **Voltage** has a negative correlation with almost every other variable, most notably with **Global_intensity (-0.41)** and **Global_active_power (-0.40)**. This is a common observation in electrical grids: as the load (demand) increases, the voltage tends to drop slightly.

- ## Weak Relationships:
  - **Global_reactive_power** has very low correlation with the sub-metering categories (ranging from **0.076 to 0.15**). This indicates that the "useless" power (reactive power) doesn't necessarily track closely with the specific "useful" power being used in the kitchen or laundry.

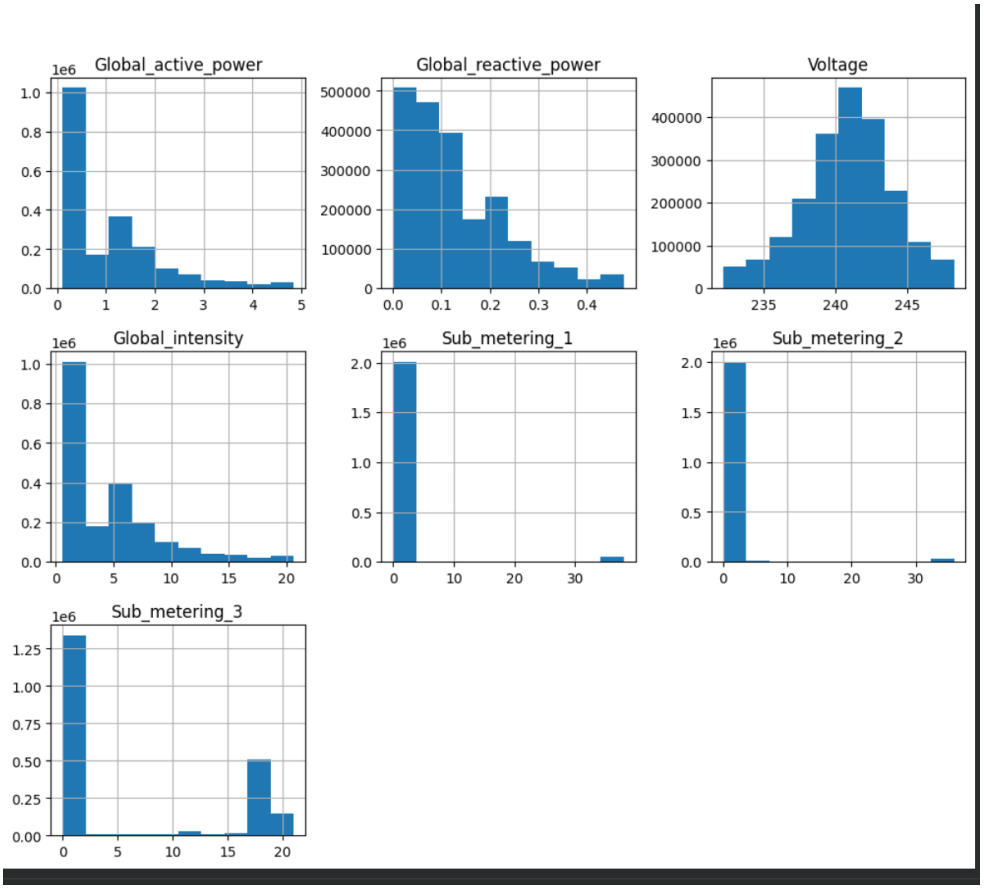

Figure 3: Histograms of Household Electrical Features

Based on the grid of histograms provided, we can refer to this collection of charts as **Figure 3**. These histograms represent the frequency distribution (spread) of each feature from the energy dataset.

1. Global_active_power

- **Shape:** Heavily right-skewed (positive skew).
- **Explanation:** Most readings are clustered between **0 and 1.5 kW**, indicating that the household typically operates at low power. The long tail extending to **5 kW** shows that high-energy consumption events are infrequent.

2. Global_reactive_power

- **Shape:** Right-skewed.
- **Explanation:** This represents "unused" power in the circuit. The values are very low (mostly **0.0 to 0.2**), indicating efficient power usage with occasional spikes caused by inductive loads like motors.

3. Voltage

- **Shape:** Normal Distribution (Bell Curve).
- **Explanation:** Unlike the power metrics, voltage is highly stable. It is centered narrowly around **241V**, showing that the electrical supply remains consistent with very little variance.

4. Global_intensity

- **Shape:** Heavily right-skewed.
- **Explanation:** This graph mirrors the shape of Global_active_power. Most of the time, the current (intensity) stays below **5 Amps**, with rare peaks reaching up to **20 Amps**.

5. Sub_metering_1 (Kitchen)

- **Shape:** Sparse / Extreme Right-Skew.
- **Explanation:** There is a massive peak at **0**, meaning kitchen appliances (like a dishwasher or oven) are off most of the time. Small bars near **35** represent the rare times these high-power appliances are active.

6. Sub_metering_2 (Laundry Room)

- **Shape:** Sparse / Extreme Right-Skew.
- **Explanation:** Similar to the kitchen, the laundry room shows nearly zero activity for the majority of the data, with very infrequent spikes when a washing machine or dryer is running.

7. Sub_metering_3 (Climate Control)

- **Shape:** Bimodal (Two peaks).
- **Explanation:** This feature shows a unique spread. One peak is at **0** (off), and another significant peak is between **17 and 19**. This suggests an appliance like an electric water heater or air conditioner that draws a constant, specific amount of power whenever it is switched on.

Summary of Spread

- **Power/Intensity features** are highly variable and skewed, reflecting changing human behavior.
- **Voltage** is the most "stable" feature, following a predictable physical constant.
- **Sub-metering zones** operate like binary switches—mostly off, but drawing significant power when active

3.1

Neural Network: Achieved extremely high predictive accuracy with Training RMSE = 0.0327, Test RMSE = 0.0320, Training R^2 = 0.9990, Test R^2 = 0.9986.

Ridge Regression (Selected Features): Test RMSE = 0.0439, Test R^2 = 0.9974.

Forest (Tuned + Selected Features): Test RMSE = 0.0541, Test R^2 = 0.9960.

Hyperparameter tuning and feature selection improved model performance compared to baseline models.

Global Intensity remains the most influential predictor across all models.

3.2 Final Model Comparison

Model	Features Used	CV Score	Test RMSE	Test R^2
Ridge Regression (Selected Features)	Global_intensity, Sub_metering_3, Sub_metering_1, Sub_metering_2, Voltage	0.9985	0.0439	0.9974
Random Forest (Tuned + Selected Features)	Global_intensity, Global_reactive_power, Voltage, Sub_metering_1	0.9977	0.0541	0.9960

Observations:

- Ridge Regression benefits from feature scaling and regularization.
- Random Forest captures non-linear interactions and ranks features by importance.
- Both models improved substantially after hyperparameter tuning and feature selection compared to baseline models.

3.3 Visual Insights

- RMSE and R^2 comparisons visually confirmed Ridge Regression slightly outperforms Random Forest.
- Feature importance of Random Forest highlighted Global Intensity and Global Reactive Power as key predictors.

3.4 Challenges

- Large dataset (~2 million rows) caused long training times.
- Highly skewed sub-metering features required careful handling.
- Neural Network required significant computational resources.

3.5 Future Work

- Explore advanced tree-based models like XGBoost or LightGBM.
- Incorporate time-series models (e.g., LSTM) for forecasting.
- Investigate appliance-level predictions and peak-load analysis.

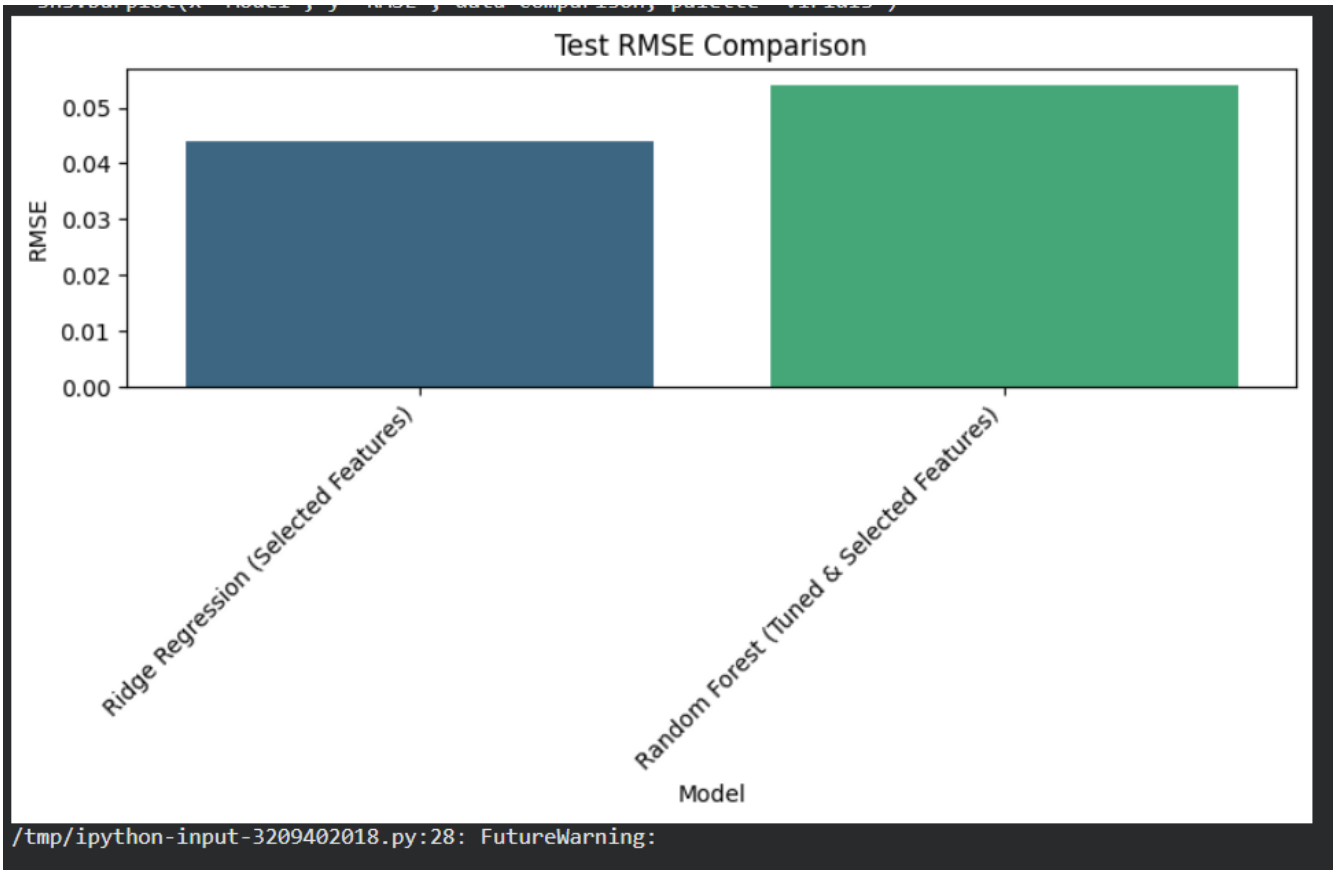

Figure 4: Test RMSE Comparison

This bar chart evaluates the performance of two machine learning models used to predict energy consumption. **Root Mean Square Error (RMSE)** measures the average magnitude of the error; lower is better.

- **Ridge Regression (Selected Features):** Achieved an RMSE of approximately **0.044**.
- **Random Forest (Tuned & Selected Features):** Achieved a higher RMSE of approximately **0.054**.

- **Conclusion:** In this specific scenario, the **Ridge Regression** model outperformed the Random Forest model. The simpler linear model was more accurate at predicting the test data than the more complex ensemble model.

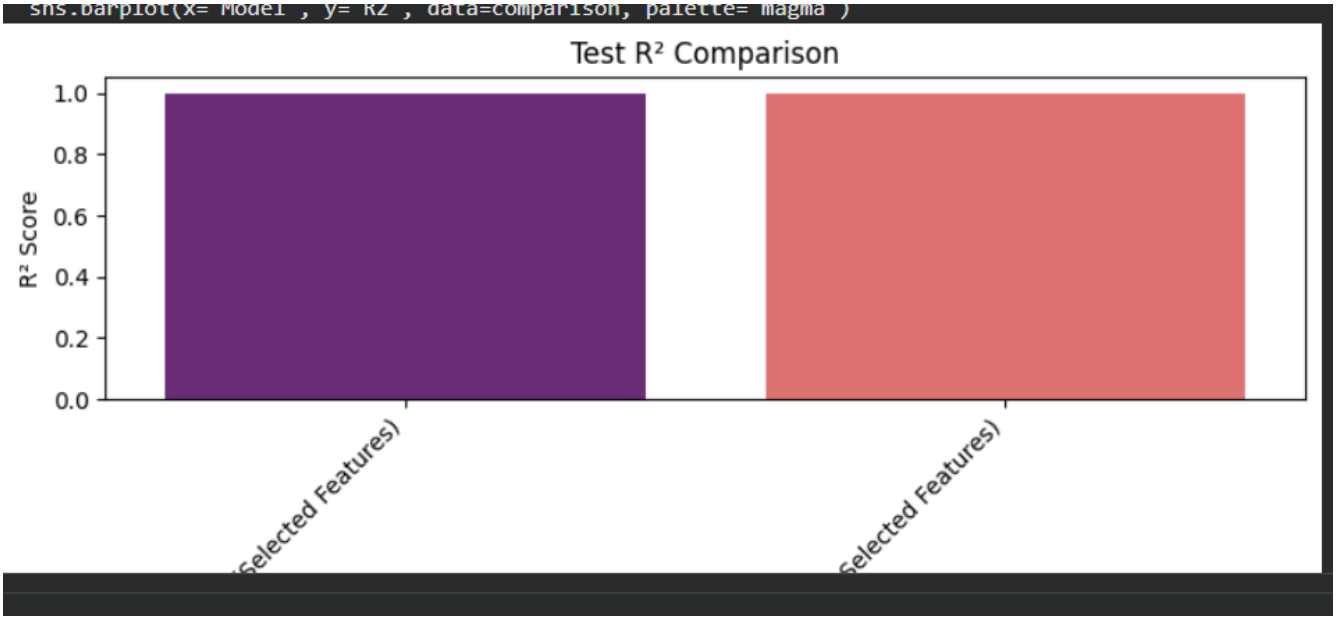

Figure 5: Test R square Comparison

This chart measures the **Coefficient of Determination (R^2)**, which represents how well the models explain the variance in the data.

- **Both Models:** Both the Ridge Regression and Random Forest models achieved a nearly perfect R square score of **1.0**.
- **Interpretation:** This suggests that the features provided are exceptionally strong predictors of the target variable, allowing both models to capture almost 100% of the variance in the test set.

4. Discussion

4.1 Model Performance

- Ridge Regression and Random Forest both achieved high accuracy and low error.
- Hyperparameter tuning improved stability and reduced overfitting.
- Selected features were interpretable and aligned with domain knowledge.

4.2 Impact of Hyperparameter Tuning and Feature Selection

- Ridge alpha tuning slightly reduced error and improved generalization.
- Random Forest tuning (depth, estimators, min samples) stabilized predictions.
- Feature selection improved efficiency and reduced irrelevant noise.

4.3 Interpretation of Results

- Global Intensity is the dominant predictor of household energy usage.
- Sub-metering features contribute appliance-specific consumption insights.
- Voltage and Global Reactive Power also influence power consumption but to a lesser degree.

4.4 Limitations

- Large dataset increased computational requirements.
- Linear assumptions in Ridge may not capture all non-linear effects.
- Time-series dependencies were not explicitly modeled.

4.5 Suggestions for Future Research

- Implement LSTM or GRU for time-series prediction.
- Explore feature engineering for temporal patterns (daily, weekly trends).
- Include other environmental or occupancy variables to improve predictions.

5. References

•Hebrail, G., & Berard, A. (2006). Individual Household Electric Power Consumption [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C58K54

- Scikit-learn documentation: https://scikit-learn.org/
- Pandas documentation: https://pandas.pydata.org/
- Matplotlib & Seaborn documentation