

Electricity Demand and price forecasting

Build as a part of Infosys Spring Board Internship 5.0



Submitted By : Hari Om Shukla  
Submitted To: Lokesh Kumar (Springboard Mentor)

**1. Introduction**

The **Electricity Cost Prediction Model** is a comprehensive machine learning initiative aimed at predicting **annual electricity consumption costs** for residential households. This predictive framework not only provides insights into electricity usage patterns but also equips stakeholders—such as energy providers, policymakers, and homeowners—with actionable recommendations to optimize energy consumption. The project was undertaken as part of the **Infosys Springboard Internship 5.0**, a platform fostering innovation and practical learning in data science and artificial intelligence.

#### ****1.1 Problem Context****

Electricity consumption plays a pivotal role in managing household expenses, contributing significantly to a family’s annual budget. For energy providers, understanding the factors influencing electricity consumption helps design efficient energy distribution systems and personalized tariffs. For policymakers, these insights enable strategies to promote energy efficiency and sustainability.

However, accurately forecasting electricity usage is challenging due to the diverse and complex nature of factors affecting energy consumption, including:

* **Housing characteristics**: The size, type, and age of a house directly influence energy consumption. Larger or older homes often require more energy for heating, cooling, and lighting.
* **Appliance usage**: The type and number of appliances in a household significantly impact electricity usage. For instance, energy-intensive appliances like air conditioners, heaters, and refrigerators are major contributors.
* **Demographics**: Household income, size, and regional location also correlate with energy usage patterns. For instance, higher-income households may own more appliances or larger homes, while colder regions may consume more energy for heating.

Understanding these relationships is essential for building an effective prediction model, which is precisely the goal of this project.

#### ****1.2 Dataset Overview****

The project leverages the **2009 Residential Energy Consumption Survey (RECS)** dataset, a comprehensive dataset compiled by the **U.S. Energy Information Administration (EIA)**. This dataset is a cornerstone of energy research, providing a detailed snapshot of energy consumption across **12,083 U.S. households**. The dataset includes:

1. **Housing Characteristics**: Information about house size, year of construction, types of insulation, and more.
2. **Appliance Usage**: Data on the types and numbers of appliances such as refrigerators, HVAC systems, and water heaters.
3. **Fuel Types**: Information on energy sources used for heating, cooling, and cooking (e.g., electricity, natural gas, propane).
4. **Demographics**: Household income, number of members, age group distributions, and geographic location.

This dataset provides an ideal foundation for identifying trends, correlations, and patterns that influence electricity consumption.

#### ****1.3 Objectives of the Project****

The primary purpose of this project is to **develop a machine learning model** that predicts the **annual electricity usage in kilowatt-hours (kWh)** for residential households based on a wide range of factors. To achieve this, the project focuses on the following key objectives:

1. **Identifying Influential Variables**:
   * A critical first step in the project is identifying which variables (e.g., house size, type of appliances, regional location) have the most significant impact on electricity consumption. This involves exploring relationships between variables and the target variable (kWh) using statistical methods, data visualization, and feature engineering techniques.
2. **Building Predictive Models**:
   * The project employs **machine learning algorithms** to predict electricity consumption. These algorithms range from simple linear regression models to advanced ensemble methods like **XGBoost**. The diversity in models ensures a robust evaluation of predictive performance.
3. **Optimizing Model Performance**:
   * Advanced techniques such as **Principal Component Analysis (PCA)**, **Feature Selection**, and **Hyperparameter Tuning** are applied to enhance model efficiency and accuracy. These techniques address issues like dimensionality, multicollinearity, and model overfitting, ensuring reliable predictions.

#### ****1.4 Relevance of Machine Learning Techniques****

The application of advanced **machine learning techniques** makes this project both innovative and effective:

* **Principal Component Analysis (PCA)**:  
  PCA is used to reduce the dataset's complexity by condensing hundreds of variables into a smaller set of components that capture the most critical patterns. This not only reduces noise but also speeds up the modeling process.
* **Feature Selection**:  
  Irrelevant and redundant features can degrade model performance. Feature selection ensures that only the most informative variables are retained, improving both accuracy and interpretability.
* **Hyperparameter Tuning**:  
  By fine-tuning model parameters (e.g., learning rate, tree depth), the project achieves optimal performance, balancing precision and computational efficiency.

#### ****1.5 Expected Impact and Outcomes****

1. **For Energy Providers**:
   * Insights into consumption patterns enable better energy distribution and demand forecasting.
   * Personalized tariffs can be designed to incentivize energy-efficient behaviors.
2. **For Homeowners**:
   * Accurate predictions empower homeowners to make informed decisions about appliance usage and energy conservation strategies.
   * Identifying high-consumption behaviors helps households reduce their electricity bills.
3. **For Policymakers**:
   * Data-driven insights support initiatives aimed at reducing carbon footprints and promoting energy efficiency at the national level.
4. **For Machine Learning Research**:
   * The project showcases the application of advanced modeling techniques to real-world challenges, contributing to the body of knowledge in predictive analytics.

In summary, this project not only serves as a practical application of machine learning but also contributes meaningful insights into energy consumption dynamics. It highlights the potential of data-driven solutions in addressing critical real-world challenges like energy optimization and sustainability.

### **2. Project Overview**

The **Electricity Cost Prediction Model** is built on a detailed exploration and analysis of the **2009 Residential Energy Consumption Survey (RECS) dataset**, which serves as the cornerstone of this project. Below, the scope and objectives of the project are elaborated, emphasizing the significance of the dataset and its role in achieving the project's goals.

#### ****2.1 The RECS Dataset****

The **Residential Energy Consumption Survey (RECS)** is a nationally representative survey conducted every four years by the **U.S. Energy Information Administration (EIA)**. It aims to provide a comprehensive understanding of residential energy consumption trends across the United States. The dataset reflects the evolving energy landscape, offering insights into housing characteristics, appliance usage, and demographic factors that influence energy consumption.

The **2009 RECS dataset**, used for this project, represents the 13th iteration of the survey and is particularly valuable due to its extensive coverage and granularity. Key highlights of the dataset include:

1. **Scope of Data Collection**:
   * The survey sampled **12,083 households** across the United States, capturing a diverse range of housing types, geographic regions, and energy usage behaviors. This large sample size ensures statistical reliability and allows for generalizations about energy consumption patterns at the national level.
2. **Richness of Features**:
   * The dataset contains **over 900 variables**, encompassing detailed aspects of:
     + **Energy Characteristics**: Types of energy used (electricity, natural gas, propane, etc.) and their purposes (heating, cooking, lighting).
     + **Housing Specifications**: Information about house size, type (detached, multi-family), insulation levels, year of construction, and geographic location.
     + **Appliance Usage**: Details on energy-consuming appliances, such as the number of refrigerators, use of HVAC systems, and presence of energy-efficient appliances.
     + **Fuel Sources**: Data on primary and secondary fuels used for heating, cooking, and other activities.
     + **Demographics**: Household income levels, size, age of occupants, and more.
3. **National Representation**:
   * The dataset is **statistically representative** of U.S. households, making it a robust foundation for predictive modeling. By accounting for geographic diversity, socioeconomic factors, and housing types, the dataset ensures that the model captures real-world variations in energy consumption.

This combination of **breadth (large sample size)** and **depth (detailed variables)** makes the 2009 RECS dataset an ideal resource for building predictive models and uncovering energy consumption insights.

#### ****2.2 Project Objectives****

The project is guided by three overarching goals that aim to explore, analyze, and model energy consumption in a data-driven manner. These objectives form the basis for the project’s structure and methodology:

1. **Feature Exploration**
   * The first step in the project is to investigate the variables within the dataset to identify the most significant features influencing electricity consumption.
   * This involves:
     + **Statistical Analysis**: Identifying variables that show strong correlations with electricity consumption (e.g., house size, appliance usage, fuel types).
     + **Visualization**: Employing techniques like scatter plots, heatmaps, and Kernel Density Estimate (KDE) plots to visually interpret relationships between features and the target variable.
     + **Feature Engineering**: Refining existing variables and creating new ones to better represent the underlying patterns in the data. For example, calculating per-capita energy usage or aggregating appliance counts.

The goal is to reduce the dataset's complexity by focusing on features that provide meaningful and statistically significant contributions to the prediction of electricity usage.

1. **Predictive Modeling**
   * The second objective is to build and evaluate machine learning models capable of estimating **Annual Electricity Usage (KWH)** based on the identified features.
   * This involves:
     + Experimenting with a variety of machine learning algorithms, ranging from simpler models like **Linear Regression** to more complex, ensemble-based methods such as **XGBoost**.
     + Incorporating techniques like **Principal Component Analysis (PCA)** and **Feature Selection** to reduce dimensionality and improve model performance.
     + Testing the models with **cross-validation** to ensure reliable and unbiased performance metrics.

By systematically comparing models, the project seeks to identify the best-performing algorithm for this prediction task.

1. **Performance Evaluation**
   * The final objective is to assess and compare the accuracy, efficiency, and robustness of the models. This involves:
     + Measuring key performance metrics like **Root Mean Squared Error (RMSE)**, **Mean Absolute Error (MAE)**, and **R² (coefficient of determination)**.
     + Conducting **hyperparameter tuning** for top-performing models to optimize their predictive accuracy. For example, adjusting the learning rate, number of estimators, and maximum tree depth in XGBoost.
     + Analyzing the model's interpretability to understand the relative importance of features and validate predictions against domain knowledge.

The ultimate goal is to deploy a model that balances high accuracy with practical interpretability, making it suitable for real-world applications.

#### ****2.3 Target Variable****

The **target variable** in this project is **Annual Electricity Usage (KWH)**, which represents the total electricity consumed by a household over the course of a year. This continuous variable is pivotal to the project for several reasons:

1. **Quantitative Measure of Energy Usage**:
   * KWH (kilowatt-hours) is a standardized unit for measuring electricity consumption. It provides a direct and interpretable measure of energy usage, making it suitable for prediction and comparison.
2. **Economic and Environmental Relevance**:
   * Understanding and predicting KWH allows households to estimate their electricity costs, identify inefficiencies, and adopt energy-saving strategies.
   * Accurate predictions also contribute to reducing overall energy consumption, lowering carbon emissions, and promoting sustainability.
3. **Modeling Implications**:
   * As a continuous variable, KWH enables the use of a wide range of regression-based machine learning techniques.
   * By focusing on KWH, the project aligns its outcomes with practical use cases such as personalized energy recommendations and dynamic pricing models.

In summary, the **RECS dataset** provides a rich and reliable foundation for analyzing energy consumption patterns, while the project’s objectives—feature exploration, predictive modelling, and performance evaluation—serve as a roadmap for achieving actionable and impactful results. The focus on **Annual Electricity Usage (KWH)** ensures the model addresses a tangible and relevant real-world challenge.

### **3. Data Collection and Pre-processing**

The foundation of any predictive modeling project is a comprehensive understanding of the dataset and its preparation for analysis. This phase involves collecting, exploring, and refining the data to ensure that the subsequent modeling phase yields accurate and reliable results. The **2009 RECS dataset** offers a wealth of information on energy consumption in U.S. households, but it must undergo extensive preprocessing to maximize its utility.

#### ****3.1 Data Collection****

The **2009 Residential Energy Consumption Survey (RECS)** dataset forms the backbone of this project. The dataset was sourced from the **U.S. Energy Information Administration (EIA)** and is designed to provide a detailed view of energy consumption trends in residential households.

The dataset includes a mix of continuous and categorical variables, reflecting various aspects of household energy use. Key categories of information include:

1. **Housing Characteristics**:
   * **Square Footage**: The total living area of the house, which directly influences energy consumption (larger homes typically consume more energy).
   * **Year of Construction**: Older homes may lack modern energy-efficient features, leading to higher electricity usage.
   * **Insulation Type**: The presence and quality of insulation affect heating and cooling energy needs.
2. **Appliance Usage**:
   * **Types and Numbers of Appliances**: Information about energy-intensive appliances such as air conditioners, refrigerators, and water heaters is included.
   * **Energy Efficiency**: Data on whether appliances are energy-efficient or meet ENERGY STAR standards.
3. **Fuel Sources**:
   * **Primary and Secondary Fuels**: Details about fuels used for heating, cooling, and cooking (e.g., electricity, natural gas, propane).
   * **Dual Fuel Usage**: Households that switch between fuels for different seasons or purposes.
4. **Demographics**:
   * **Household Income Levels**: Income may correlate with energy usage, as higher-income households often have larger homes or more appliances.
   * **Household Size**: Larger households typically have higher electricity usage.
   * **Regional Location**: Geographic variation influences energy needs due to climate differences.

The dataset’s richness—spanning over **900 variables** across **12,083 households**—enables a robust exploration of factors that drive residential electricity consumption.

#### ****3.2 Exploratory Data Analysis (EDA)****

EDA is a critical phase that helps uncover patterns, relationships, and potential issues in the dataset. This step prepares the data for the modeling process by ensuring it is clean, consistent, and free from anomalies.

**1. Data Cleaning**

* **Outlier Removal**:
  + Households with extreme energy consumption values (e.g., **KWH > 80,000**) were identified and removed. Such extreme outliers could skew the model's performance and reduce its generalizability.
  + Extreme values were identified through statistical methods such as interquartile ranges and visualized using box plots.
* **Handling Missing Values**:
  + Columns with over **60% missing data** were excluded, as they lacked sufficient information for meaningful analysis.
  + Missing values in critical variables were imputed using statistical methods (mean, median, or mode) where applicable.
* **Feature Reduction**:
  + Features with a **single unique value** (e.g., columns where all entries are identical) were dropped, as they provide no variability or predictive power.

**2. Data Visualization**

* Visualization techniques were employed to understand relationships between features and the target variable:
  + **Scatter Plots**: Highlighted correlations between continuous features (e.g., square footage vs. KWH). Larger homes exhibited higher energy consumption, as expected.
  + **Box Plots**: Illustrated differences in energy usage across categorical variables like insulation type or fuel source.
  + **Kernel Density Estimate (KDE) Plots**: Revealed the distribution of electricity usage and helped identify anomalies. KDE plots showed a right-skewed distribution of KWH, indicating the presence of outliers or higher consumption by certain households.

**3. Outlier Detection**

* Outliers were identified using statistical methods and visual tools:
  + **Box Whisker Plots**: Highlighted extreme values that lay outside the typical range for variables like square footage and KWH.
  + **KDE Plots**: Provided insights into how outliers affected the overall data distribution.
* Outliers were removed to prevent them from distorting model training and evaluation. Removing these anomalies ensured that the model focused on the most representative data.

#### ****3.3 Data Transformation and Preprocessing****

Once the data was cleaned, it was further refined and prepared for machine learning by applying transformation and preprocessing techniques. This ensured that the dataset was both computationally efficient and predictive.

**1. Feature Engineering**

* **New Feature Creation**:
  + Derived new variables that better represent energy usage patterns, such as energy consumption per square foot or per household member.
* **Redundant Feature Removal**:
  + Variables that were highly correlated or provided redundant information were removed to streamline the dataset.

**2. Feature Encoding**

* **One-Hot Encoding**:
  + Categorical variables (e.g., fuel type, appliance type) were transformed into numerical representations using one-hot encoding.
  + This approach avoided introducing ordinal relationships in purely categorical data, ensuring fair treatment by the machine learning models.

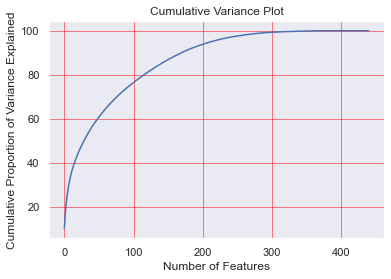
**3. Dimensionality Reduction**

* **Handling Multicollinearity**:
  + Highly correlated features (e.g., correlation **>98%**) were identified and removed. For example, if square footage and room count showed high correlation, one was retained.
  + This step reduced redundancy and improved model stability by mitigating overfitting risks.
* **Principal Component Analysis (PCA)**:
  + PCA was used to transform the dataset into a smaller set of uncorrelated features (principal components) while retaining most of the variance.

In summary, **Data Collection and Preprocessing** ensured that the dataset was accurate, consistent, and optimized for predictive modeling. Through careful cleaning, exploration, and transformation, the data was prepared to yield meaningful and actionable insights during the modeling phase.

### **4. Principal Component Analysis (PCA)**

Principal Component Analysis (PCA) is a powerful dimensionality reduction technique that transforms a high-dimensional dataset into a smaller set of uncorrelated variables known as principal components. PCA retains the majority of the data's variance while eliminating redundancy and noise, making it particularly useful for datasets with many correlated features, such as the 2009 RECS dataset.

In this project, PCA was employed to reduce the number of features while preserving the dataset's most significant information. This process improved computational efficiency, minimized the risk of overfitting, and facilitated the selection of the most critical features for modeling.

### **Steps in the PCA Process**

#### ****1. Data Standardization****

Before applying PCA, the dataset was standardized to ensure all features were on the same scale. PCA is sensitive to the magnitude of the input variables, so standardization ensures that variables with larger ranges do not dominate the analysis.

* **StandardScaler** from the scikit-learn library was used to scale continuous variables to have a mean of 0 and a standard deviation of 1.

#### ****2. Scree Plot Analysis****

A scree plot is a visual tool used to determine the optimal number of principal components to retain. It plots the explained variance ratio of each principal component against the component number.

* **Key Steps**:
  + Principal components were computed for the dataset. Each component captured a portion of the dataset's total variance.
  + The scree plot displayed the explained variance ratio for each principal component, showing how much variance each contributed to the dataset.
* **Key Observations**:
  + The plot exhibited a characteristic "elbow" point, where the explained variance ratio dropped off sharply.
  + This elbow indicated the number of components that captured the majority of the variance in the dataset without adding unnecessary complexity.

#### ****Cumulative Variance Analysis:**** The cumulative variance plot showed the total variance explained by an increasing number of principal components. This helped determine how many components were necessary to retain a specified level of information from the original dataset.

* **Steps Taken**:
  + The cumulative variance was calculated by summing the explained variance ratios of consecutive principal components.
  + A threshold of **95% cumulative variance** was set, as retaining 95% of the variance strikes a balance between dimensionality reduction and information preservation.
* **Findings**:
  + Approximately **200 principal components** were sufficient to explain 95% of the dataset's variance. This significantly reduced the dimensionality from the original 900+ variables, eliminating noise and redundant information.

#### ****4. Data Transformation****

Once the number of components to retain was decided, the dataset was transformed into its principal components:

* The original variables were linearly combined into new, uncorrelated variables (principal components).
* Each principal component represented a weighted combination of the original variables, capturing a specific pattern in the data.

### **Benefits of PCA in This Project**

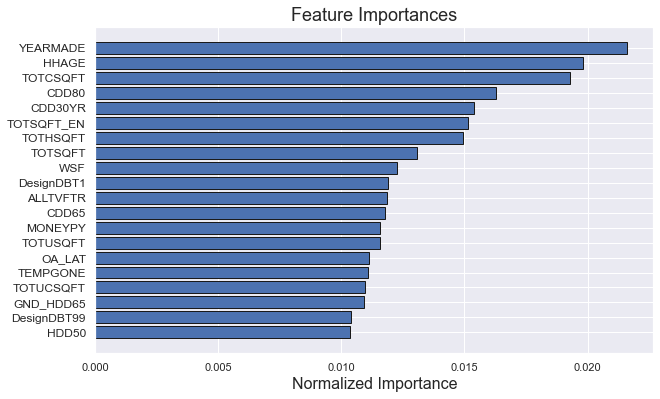
1. **Dimensionality Reduction**:
   * Reduced the number of features from over 900 to approximately 200, making the dataset more manageable and computationally efficient.
   * Simplified the modeling process by focusing on the most important patterns in the data.
2. **Noise Elimination**:
   * By retaining only the components that explained significant variance, PCA helped filter out noise and irrelevant information, improving the quality of the data fed into the models.
3. **Handling Multicollinearity**:
   * PCA transformed correlated variables into uncorrelated principal components, addressing multicollinearity issues inherent in the dataset.
4. **Improved Model Performance**:
   * With fewer, more meaningful features, the machine learning models trained on the PCA-transformed dataset were less prone to overfitting and demonstrated better generalization to unseen data.
5. **Visualization**:
   * The PCA-transformed dataset facilitated visualization of the data structure in lower dimensions, providing insights into patterns and clusters.

### **Summary of PCA Application**

PCA was instrumental in streamlining the 2009 RECS dataset, reducing its complexity without sacrificing essential information. The process allowed the project to focus on meaningful features, optimize computational resources, and enhance the predictive power of the machine learning models. By reducing the dataset to 200 principal components that captured 95% of the variance, PCA laid the groundwork for efficient and accurate modeling of annual electricity consumption.

### **5. Feature Selection**

Feature selection is a crucial step in any machine learning pipeline. It involves identifying and retaining the most relevant variables for the predictive task while removing redundant, irrelevant, or noisy features. This not only enhances the model's performance and accuracy but also reduces computational complexity and the risk of overfitting.

In this project, the **FeatureSelector** class was employed to refine the dataset further. Feature selection was conducted after dimensionality reduction through PCA, focusing on identifying the most important predictors of annual electricity consumption (KWH).

### **Steps in the Feature Selection Process**

#### ****1. Removing Features with High Missing Values****

* **Rationale**:  
  Features with a high percentage of missing values often provide little or no additional value to the model. Their inclusion can skew results or require excessive imputation, leading to potential inaccuracies.
* **Process**:
  + A threshold of **60% missing data** was set. Features with more than 60% missing values were automatically discarded.
  + Missing value thresholds were determined based on exploratory data analysis (EDA), which showed that such features contributed little to predictive accuracy.
* **Outcome**:
  + Features like rarely reported appliance types or niche demographic details were eliminated, ensuring the dataset retained variables with sufficient data coverage.

#### ****2. Reducing Collinearity****

* **Rationale**:  
  Collinearity occurs when features are highly correlated with each other, leading to redundancy. High collinearity can inflate feature importance, skew model interpretations, and degrade performance.
* **Process**:
  + A correlation matrix was computed for all numerical features in the dataset.
  + Pairs of features with a Pearson correlation coefficient greater than **0.98** were identified as redundant.
  + From each highly correlated pair, the feature with the lower overall importance (determined by mutual information or variance contribution) was removed.
* **Outcome**:
  + Features such as multiple overlapping measures of house size (e.g., total square footage vs. room dimensions) were filtered, retaining only one representative variable.

#### ****3. Tree-Based Feature Importance Selection****

* **Rationale**:  
  Tree-based models, such as decision trees and random forests, are effective at ranking features based on their importance. They can handle nonlinear relationships and interactions, making them suitable for assessing feature relevance.
* **Process**:
  + A random forest model was trained on the dataset with the remaining features after collinearity reduction.
  + Feature importance scores were calculated for each variable based on its contribution to reducing impurity (e.g., Gini or entropy) in the tree splits.
  + Features with zero or negligible importance scores were identified as irrelevant and removed.
* **Outcome**:
  + Variables with no meaningful relationship to the target variable (e.g., irrelevant regional codes, outdated appliance usage data) were discarded.

#### ****4. Removing Low-Importance Features Based on Cumulative Thresholds****

* **Rationale**:  
  Not all features with some level of importance are essential. Setting a cumulative importance threshold ensures that the model focuses on variables contributing most to prediction accuracy.
* **Process**:
  + Importance scores from the tree-based selection process were sorted in descending order.
  + A **cumulative importance threshold of 90%** was applied. This means that features contributing to the top 90% of total importance were retained, while the rest were discarded.
  + This method provided a structured way to prune less impactful variables.
* **Outcome**:
  + The dataset was reduced further, retaining only the most significant features while eliminating marginal contributors.

#### ****5. Encoding and Transformations****

* **One-Hot Encoding**:
  + Categorical features that remained after the above steps were encoded using one-hot encoding to convert them into a numerical format suitable for machine learning algorithms.
  + For example, categorical variables like heating fuel types were expanded into binary indicator columns.
* **Final Adjustments**:
  + After encoding, low-variance features (those with minimal variability across rows) were also removed to ensure all retained features provided meaningful differentiation.

### **Impact of Feature Selection on the Dataset**

* **Original Feature Count**: 428 features
* **Final Feature Count**: 216 features
* **Reduction**: Over **49%** of the features were removed, significantly reducing dimensionality.

### **Benefits of the Feature Selection Process**

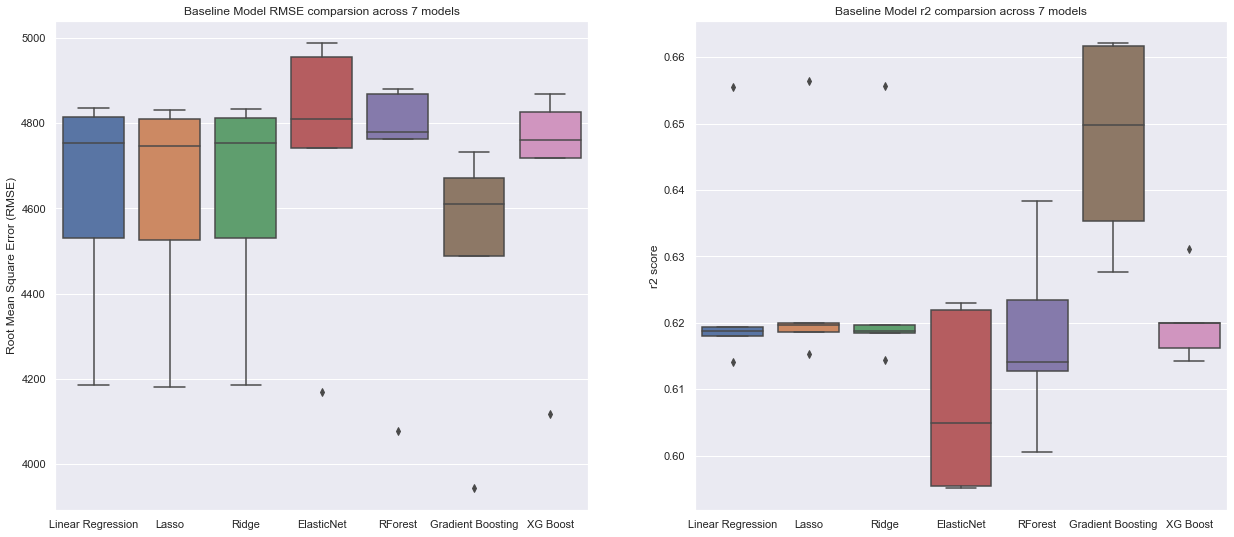
1. **Enhanced Model Efficiency**:
   * Fewer features reduced computational time for training and prediction.
   * Simplified the complexity of the models, improving interpretability.
2. **Improved Accuracy**:
   * By focusing on the most relevant features, the model could better generalize to unseen data.
   * Reduced noise in the dataset led to higher predictive performance.
3. **Mitigated Overfitting**:
   * By discarding irrelevant or redundant features, the risk of overfitting was minimized.
4. **Streamlined Pipeline**:
   * The final set of features served as a refined input for further modeling steps, ensuring a cleaner and more robust machine learning pipeline.

### **6. Model Development and Evaluation**

The development and evaluation of predictive models are essential steps in any machine learning project. This phase involves building initial models (baseline models) to establish a performance benchmark, followed by fine-tuning the models through hyperparameter optimization to maximize predictive accuracy. In this project, multiple machine learning algorithms were tested and refined to identify the most effective method for predicting annual electricity consumption (KWH) in residential households.

### **6.1 Baseline Models**

Baseline models serve as a starting point for evaluating the dataset and the problem at hand. These models are trained using default parameters and provide a benchmark against which more advanced models can be compared.



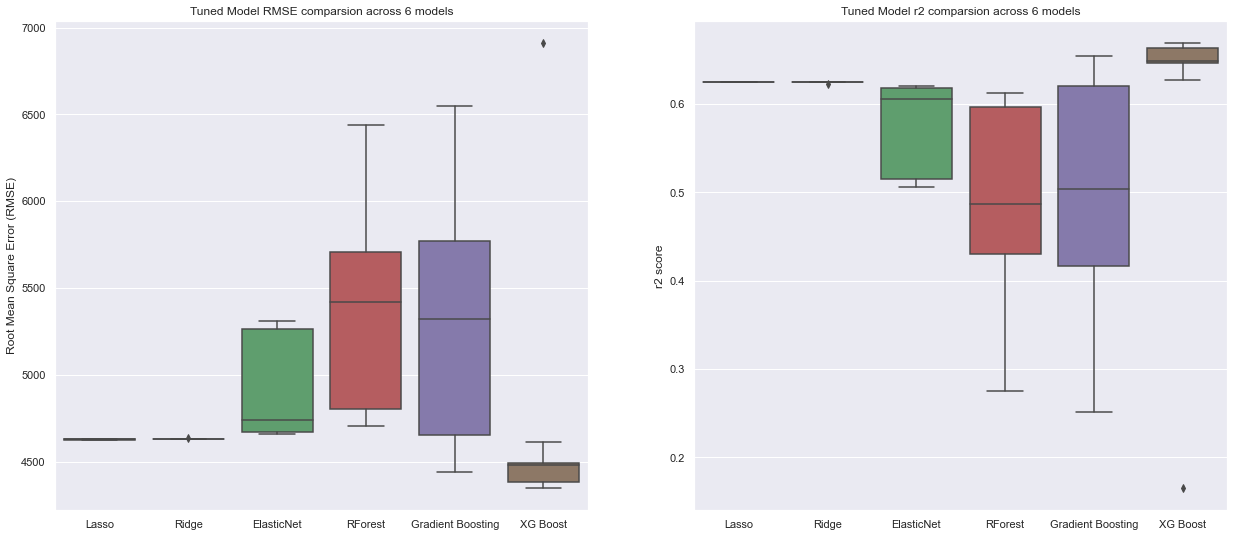
#### ****Baseline Model Types****

1. **Linear Models**:
   * **Linear Regression**:  
     A simple algorithm that assumes a linear relationship between independent variables (features) and the dependent variable (KWH).
     + **Advantages**: Easy to implement and interpret.
     + **Limitations**: Fails to capture complex, nonlinear relationships in the data.
   * **Lasso Regression**:  
     Incorporates L1 regularization, which penalizes large coefficients to promote sparsity in the model (i.e., some coefficients are reduced to zero).
     + **Advantages**: Helps in feature selection by eliminating irrelevant features.
     + **Limitations**: May underperform in datasets with highly correlated features.
   * **Ridge Regression**:  
     Applies L2 regularization, which penalizes large coefficients without forcing them to zero.
     + **Advantages**: Handles multicollinearity better than Lasso.
     + **Limitations**: Does not perform feature selection as effectively as Lasso.
2. **Ensemble Models**:
   * **Random Forest**:  
     An ensemble of decision trees that aggregates their predictions to reduce overfitting and improve accuracy.
     + **Advantages**: Captures complex relationships and works well with nonlinear data.
     + **Limitations**: May require extensive tuning to optimize performance.
   * **Gradient Boosting Machines (GBM)**:  
     Builds an ensemble of weak learners (decision trees), optimizing performance by correcting errors in previous iterations.
     + **Advantages**: Highly effective for structured data.
     + **Limitations**: Computationally expensive and sensitive to hyperparameters.

#### ****Baseline Model Performance****

* **Linear Models**:
  + Linear Regression showed moderate predictive performance, capturing basic trends but failing to handle nonlinear relationships effectively.
  + Lasso and Ridge regression improved upon Linear Regression by regularizing coefficients, reducing overfitting to a degree. However, their performance was limited by the inherent complexity of the dataset.
* **Ensemble Models**:
  + Random Forest performed better than linear models by capturing nonlinear interactions. However, its predictions were slightly biased due to default parameter settings.
  + Gradient Boosting Machines outperformed Random Forest by leveraging sequential learning, though it still required tuning to optimize.

### **6.2 Hyperparameter Tuning**

Hyperparameter tuning is a critical step to optimize model performance by adjusting algorithm parameters to find the best combination for the dataset. In this project, **GridSearchCV** was employed to systematically test different parameter values for each model.

#### ****Hyperparameter Tuning Process****

1. **Algorithm Selection**:  
   Six algorithms were selected for hyperparameter tuning based on their baseline performance and suitability for the dataset:
   * Lasso Regression
   * Ridge Regression
   * Random Forest
   * Gradient Boosting Machines
   * XGBoost (Extreme Gradient Boosting)
   * LightGBM
2. **Hyperparameters Tuned**:
   * **Lasso and Ridge Regression**:
     + Regularization parameter (alpha): Determines the strength of the penalty applied to coefficients.
   * **Random Forest**:
     + Number of estimators (n\_estimators): Total number of trees in the forest.
     + Maximum depth (max\_depth): Limits the depth of trees to prevent overfitting.
     + Minimum samples split (min\_samples\_split): Minimum samples required to split a node.
   * **Gradient Boosting Machines**:
     + Learning rate (learning\_rate): Determines the contribution of each tree to the final prediction.
     + Number of estimators (n\_estimators): Controls the number of trees.
     + Maximum depth (max\_depth): Restricts the depth of each tree.
   * **XGBoost**:
     + Learning rate (eta): Controls the step size during boosting.
     + Maximum depth (max\_depth): Prevents overfitting by limiting tree depth.
     + Subsample ratio (subsample): Fraction of data samples used for training each tree.
     + Column sampling (colsample\_bytree): Fraction of features used to build each tree.
3. **Cross-Validation**:
   * GridSearchCV employed **k-fold cross-validation** (with k=5) to evaluate model performance across different parameter combinations. This ensured reliable and unbiased evaluation by training and testing the model on different subsets of the data.

#### ****Key Results from Hyperparameter Tuning****

* **Linear Models**:
  + **Lasso Regression**: Improved performance slightly with optimal alpha values, but still exhibited limitations in handling complex relationships.
  + **Ridge Regression**: Performed slightly better than Lasso, especially in datasets with multicollinearity, but fell short of ensemble models.
* **Tree-Based Models**:
  + **Random Forest**: Significant improvement was observed with optimized parameters (n\_estimators = 200, max\_depth = 15). However, it was still outperformed by boosting methods in terms of accuracy.
  + **Gradient Boosting Machines**: Hyperparameter tuning led to improved predictions, with reduced bias and variance.
* **XGBoost**:
  + Emerged as the **best-performing algorithm**, achieving the **lowest Root Mean Squared Error (RMSE)** and the **highest R² score**.
  + Optimized parameters (eta = 0.1, max\_depth = 10, subsample = 0.8, colsample\_bytree = 0.8) balanced model complexity and generalization, making it the most suitable model for the dataset.

### **Key Metrics for Evaluation**

* **Root Mean Squared Error (RMSE)**: Measures the average magnitude of errors in predictions. Lower RMSE indicates better model accuracy.
* **R² Score**: Indicates the proportion of variance in the target variable explained by the model. Higher R² reflects better performance.
* **Cross-Validation Score**: Ensures robustness by evaluating performance across different subsets of the data.

### **7. Findings and Insights**

The Electricity Cost Prediction Model project provided a wealth of insights into the factors influencing residential electricity consumption and demonstrated the effectiveness of advanced machine learning techniques. The key findings and learnings from the project are outlined below:

### **7.1 Key Features Influencing Electricity Consumption**

Through rigorous exploratory data analysis (EDA) and feature selection, several features emerged as the most influential predictors of annual electricity usage (KWH). These insights are critical for understanding energy consumption patterns and optimizing resource usage.

1. **House Size**:
   * Larger houses tend to consume more electricity due to greater heating, cooling, and appliance usage demands.
   * Square footage was strongly correlated with KWH, reflecting its direct impact on energy needs.
2. **Appliance Usage**:
   * The number and types of appliances significantly influenced electricity consumption.
     + For example, households with central air conditioning, multiple refrigerators, or energy-intensive appliances like electric ovens showed higher energy usage.
   * Patterns of appliance operation (e.g., daily use vs. seasonal use) also played a role.
3. **Fuel Sources**:
   * Fuel types used for heating, cooling, and cooking had a notable impact on electricity usage.
     + Households relying primarily on electricity for heating and cooking consumed more energy than those using alternative fuels like natural gas or propane.
4. **Household Income and Demographics**:
   * Higher-income households generally exhibited higher electricity consumption, likely due to larger homes, more appliances, and higher energy needs.
   * Demographic factors, including household size and regional differences, also contributed to variations in consumption patterns.

These findings provide actionable insights for stakeholders, including policymakers and energy providers, to design targeted interventions for energy efficiency.

### **7.2 Model Performance**

1. **XGBoost Outperformed Other Models**:
   * XGBoost emerged as the top-performing model, achieving the best predictive accuracy based on evaluation metrics such as **Root Mean Squared Error (RMSE)** and **R² score**.
   * Its ability to handle nonlinear relationships and complex feature interactions made it particularly suited for the dataset.
2. **Performance Comparisons**:
   * **Linear Models**:
     + Linear Regression, Lasso, and Ridge struggled with the dataset’s complexity, failing to capture nonlinear relationships effectively.
   * **Tree-Based Models**:
     + Random Forest and Gradient Boosting Machines demonstrated better performance than linear models but were surpassed by XGBoost in predictive accuracy and efficiency.
3. **Generalization**:
   * XGBoost’s robust hyperparameter tuning (e.g., learning rate, depth of trees, feature sampling) minimized overfitting and ensured strong generalization across validation sets.

### **7.3 Impact of Data Processing**

Effective data processing was integral to the success of the predictive model. Key preprocessing steps significantly improved model accuracy and stability:

1. **Outlier Removal**:
   * Removing extreme outliers (e.g., rows with KWH > 80,000) reduced skewness in the data, enabling the models to learn more effectively from the remaining observations.
2. **Handling Missing Values**:
   * Features with missing data above 60% were removed to prevent noise and inconsistencies in the dataset.
   * Remaining missing values were imputed, preserving the integrity of critical features.
3. **Dimensionality Reduction**:
   * Dropping features with high correlation (above 98%) eliminated redundancy, improving model interpretability and efficiency.
4. **Feature Encoding**:
   * Transforming categorical variables through One-Hot Encoding ensured compatibility with machine learning algorithms while retaining critical information.

### **7.4 Contributions of PCA**

Principal Component Analysis (PCA) played a pivotal role in reducing the dataset’s dimensionality without compromising predictive power:

1. **Dimensionality Reduction**:
   * PCA reduced the number of features from 428 to approximately 200 components, explaining 95% of the variance in the data.
   * This reduction minimized computational complexity and improved model training times.
2. **Noise Reduction**:
   * By focusing on principal components, PCA removed irrelevant and noisy features, enhancing model accuracy.
3. **Simplified Dataset**:
   * The streamlined dataset retained essential variance, allowing the models to focus on the most critical factors driving electricity consumption.

### **7.5 Summary of Insights**

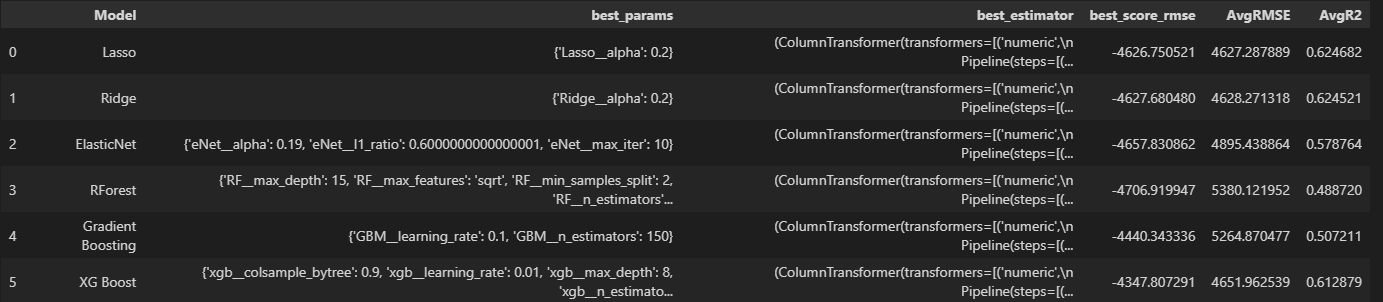
1. **Key Features**:
   * House size, appliance usage, fuel sources, and household income emerged as the strongest predictors of electricity consumption.
2. **Model Performance**:
   * XGBoost demonstrated the best predictive capabilities, outperforming both linear models and other ensemble methods.
3. **Impact of Data Processing**:
   * Rigorous data cleaning, feature selection, and preprocessing significantly enhanced the accuracy and robustness of the model.
4. **Role of PCA**:
   * PCA contributed to dimensionality reduction, noise elimination, and computational efficiency, enabling the development of high-performing models.

These findings highlight the importance of integrating domain knowledge, advanced machine learning techniques, and comprehensive data preprocessing to develop robust and interpretable predictive models for real-world applications. Future work can extend these insights by incorporating additional contextual factors, such as weather patterns and seasonal variations, to further enhance the model’s predictive power.

### **8. Conclusion**

The **Electricity Cost Prediction Model** underscores the transformative potential of machine learning in analyzing and forecasting residential energy consumption. This project successfully addressed the challenge of predicting annual electricity usage (measured in kilowatt-hours, KWH) by leveraging advanced algorithms, robust data preprocessing, and insightful analysis of the 2009 Residential Energy Consumption Survey (RECS) dataset.

The key accomplishments of the project and potential areas for future enhancement are summarized below:



### **8.1 Key Achievements**

#### ****1. Accurate and Robust Predictions****

The XGBoost model emerged as the most effective algorithm for predicting residential electricity consumption, achieving superior performance across key evaluation metrics:

* **High Predictive Accuracy**:
  + XGBoost demonstrated the lowest Root Mean Squared Error (RMSE) and the highest R² score among all tested models, reflecting its ability to capture complex relationships in the dataset.
  + Hyperparameter tuning, including adjustments to the learning rate, tree depth, and feature sampling, optimized the model’s performance.
* **Generalization**:
  + XGBoost's robustness in handling both training and unseen validation data ensures its practical applicability across diverse scenarios.

#### ****2. Insights into Electricity Consumption Drivers****

The project provided actionable insights into the key factors influencing energy usage, offering value to both consumers and policymakers:

* **Household Characteristics**: Features like house size and appliance usage were found to be the most influential, reflecting their direct impact on energy consumption.
* **Fuel Sources**: Dependency on electricity for heating, cooling, and cooking significantly affected total usage.
* **Demographic Patterns**: Income levels and household size also played pivotal roles, highlighting the interplay between socioeconomic factors and energy needs.

These insights not only support the development of predictive models but also provide a foundation for strategies to improve energy efficiency and conservation.

#### ****3. Effective Use of Data Science Techniques****

* **Data Preprocessing**: Rigorous data cleaning, transformation, and feature engineering ensured a high-quality dataset for analysis and modeling.
* **Dimensionality Reduction**: The use of Principal Component Analysis (PCA) significantly simplified the dataset while retaining critical information, enhancing computational efficiency without compromising model performance.
* **Feature Selection**: Removing redundant and low-importance features refined the dataset, making the modeling process more efficient and interpretable.

### **8.2 Limitations and Areas for Improvement**

While the project achieved notable successes, it also highlighted areas where future enhancements could further elevate its capabilities:

#### ****1. Integration of Additional Factors****

* **Weather Data**: Incorporating real-time or historical weather information, such as temperature and humidity, could provide deeper insights into seasonal energy usage trends.
* **Energy Pricing and Tariffs**: Dynamic energy pricing and regional tariff variations could further refine the model’s applicability to different locations and economic contexts.

#### ****2. Temporal and Seasonal Trends****

The inclusion of time-series data to account for monthly or seasonal variations in electricity usage would improve the model’s ability to predict short-term energy demands.

#### ****3. External Validation****

Testing the model on newer datasets (e.g., from subsequent RECS surveys or other countries) would validate its generalizability and adaptability across different regions and timeframes.

### **8.3 Future Work and Applications**

To extend the scope and impact of this project, future efforts could focus on:

1. **Real-Time Prediction Systems**: Deploying the model as part of an application or service to provide real-time energy consumption estimates and cost forecasts for residential users.
2. **Integration with Smart Home Technologies**: Using data from IoT devices (e.g., smart meters, appliances) to refine predictions and offer personalized energy-saving recommendations.
3. **Policy and Energy Efficiency Programs**: Assisting energy providers and governments in designing targeted energy-saving campaigns based on identified usage patterns and key demographic insights.

### **8.4 Summary**

The Electricity Cost Prediction Model successfully demonstrated the capability of machine learning in addressing real-world energy challenges. Key achievements include:

* Developing a high-performing predictive model using XGBoost to estimate residential electricity usage with remarkable accuracy.
* Deriving actionable insights into energy consumption drivers, such as household size, appliance usage, and socioeconomic factors.
* Applying advanced data science techniques to clean, preprocess, and streamline the dataset for efficient analysis.

This project provides a solid foundation for future exploration and innovation in the field of energy analytics. By integrating additional factors and extending its scope, the model has the potential to contribute significantly to sustainable energy management and optimization on both individual and societal levels.

## **9. How to Run the Model**

To run the model:

1. **Set Up Environment**: Install dependencies using a Python virtual environment.
2. **Execute Notebook**: Open and run the provided Jupyter notebook.
3. **Interpret Results**: Analyze predictions and adjust model parameters if needed.