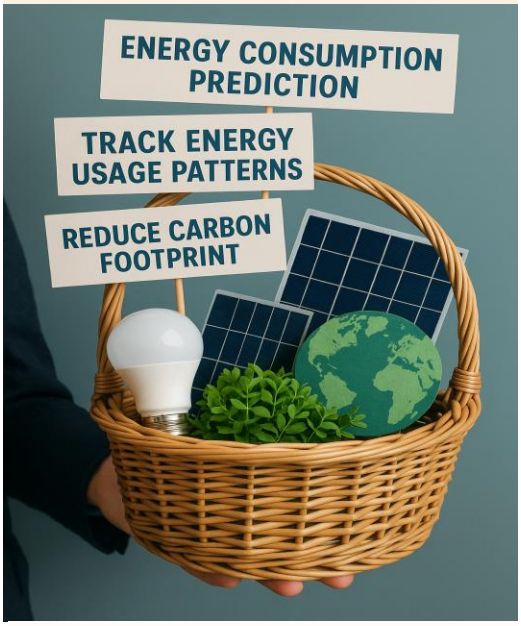
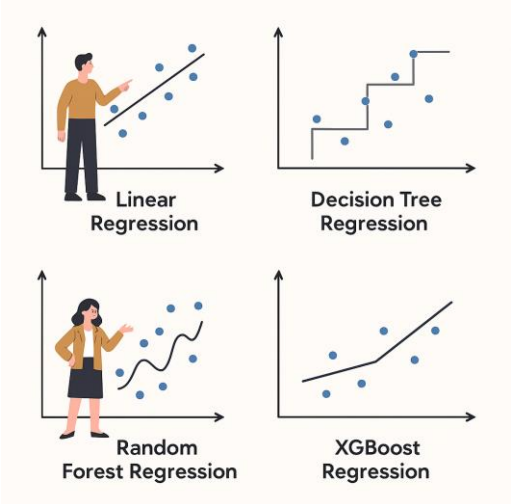
THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Energy Consumption TRACKER





AI-Driven Energy Consumption Prediction: Harnessing Data for a Smarter, Sustainable Future

A Project Report by:
- **Smarth Kaushal [102497023]**
- **Aniket Gupta [102317169]**
- **Aadar Mishra [102367002]**

Submitted to Mr. Abhishek Phull
In Partial Fulfilment of lab evaluation of the course
UCS411: ARTIFICIAL INTELLIGENCE, Jan – May 2025

Computer Science & Engineering Department
Thapar Institute of Engineering & Technology, Patiala
Dated: May 5th, 2025

In a world where every watt counts and sustainability is no longer a choice but a necessity, the imperative for intelligent, data-driven solutions has never been greater. Guided by the conviction that "energy misused today cannot be excused tomorrow," our group project addresses the pressing challenges of energy overconsumption and inefficiency by harnessing the transformative power of artificial intelligence. As Roy Cooper aptly stated, "A strong renewable energy industry is good for our environment and our economy." Yet, even the cleanest energy sources fall short without efficient consumption and management.

To tackle these challenges, we employ a suite of advanced regression models-including Linear Regression, Decision Tree Regression, Random Forest, and XGBoost-to build a robust energy consumption prediction system. By leveraging historical usage patterns, weather data, and sector-specific trends, our approach aspires not only to reduce waste and optimize distribution but also to support the global transition toward sustainability. This report details our journey-from identifying the urgent need for smarter energy management, through methodological innovation, to the realization of actionable insights-demonstrating the pivotal role of AI in shaping a smarter, greener future.

As you explore our methodology and findings, we encourage you to reflect on Barack Obama's vision: "To truly transform our economy, protect our security, and save our planet from the ravages of climate change, we need to ultimately make clean, renewable energy the profitable kind of energy." Read on to discover how our approach brings this vision closer to reality, paving the way for a future where efficiency and sustainability go hand in hand.

Abstract

As global energy demand surges, inefficient consumption patterns place tremendous strain on power grids, drive up costs, and exacerbate the climate crisis. The increasing pressure on energy systems underscores the need for intelligent solutions to optimize power distribution and enhance sustainable energy management. This project presents the development of an AI-driven energy forecasting system that leverages multiple regression models to address these challenges. The goal is to predict energy consumption with high accuracy, providing a powerful tool to guide decision-making for more efficient energy management.

A progressive modeling approach is employed in this study, which involves comparing a range of algorithms: Linear Regression, Decision Tree Regression, Random Forest, and XGBoost. Each model is evaluated based on its performance in terms of key metrics like Root Mean Squared Error (RMSE) and R^2. As illustrated in the accompanying visualizations, model performance improves systematically across this spectrum, with XGBoost achieving the highest accuracy, delivering an RMSE of 2815.23 and an R^2 of 0.97. These results highlight the increasing sophistication of machine learning models in addressing the complexities of energy forecasting.

Feature engineering plays a crucial role in this project, incorporating a wide variety of factors such as historical energy usage, weather conditions, temporal dynamics, and occupancy patterns. Data preprocessing, including rigorous cleaning and normalization, as well as hyperparameter tuning, contributed significantly to the robustness of the models. The comparative analysis also revealed that manual implementation of XGBoost outperformed automated approaches such as PyCaret AutoML, reinforcing the value of domain-specific customization for achieving optimal results.

This work not only bridges the gap between advanced machine learning techniques and sustainable energy management but also offers a scalable framework that can be applied to mitigate carbon emissions, optimize renewable energy integration, and enhance the resilience of energy systems. The findings underscore the critical role of intelligent prediction systems in balancing the increasing demands for energy with the urgent need to reduce the environmental footprint of energy consumption.
In addition to its contributions to energy efficiency, this research also highlights the potential of AI-driven solutions in transforming the way cities manage and consume energy. The scalable nature of the system makes it suitable for urban environments around the world, where power grids are under growing stress due to rising populations and energy consumption. By optimizing energy use in real-time, this project paves the way for a more sustainable, low-carbon future, offering the tools needed to navigate the challenges posed by the evolving energy landscape.

Keywords: Energy consumption prediction, Linear Regression, Decision Tree Regression, XGBoost, Random Forest, machine learning, artificial intelligence, smart grids, sustainability, PyCaret, feature engineering, energy efficiency, carbon footprint reduction, power distribution optimization, energy management systems.

ACKNOWLEDGEMENT

In the accomplishment of this report, we have been fortunate to receive blessings and unwavering support from many individuals. First and foremost, we express our deepest gratitude to the Lord Almighty for guiding us through every challenge and enabling us to complete this report timely.

The journey began with our keen interest in addressing real-world issues that impact society at large. After much discussion and exploration, we chose the topic of energy consumption prediction, recognizing its critical importance in today's world where sustainability and efficient resource management are paramount. This decision was inspired by our desire to contribute to the global pursuit of smarter, greener solutions through the application of cutting-edge technology.

From the very first day to the final stages of compiling and refining this report, we have encountered numerous challenges-technical hurdles, moments of uncertainty, and the complexities of teamwork. Yet, with perseverance, mutual support, and divine grace, we have transformed obstacles into valuable learning experiences. We are deeply thankful to our mentors, faculty, and peers for their encouragement and insightful feedback, which have been instrumental in shaping our work. Above all, we are grateful for the opportunity to collaborate, grow, and contribute to a project that aspires to make a meaningful impact.

We are also immensely thankful to our family members for their unconditional love, patience, and steadfast support during every phase of our research. Their encouragement has been a source of strength, motivating us to persevere through difficulties and achieve our goals.

Throughout this journey, we have learned and discovered many new things, both academically and personally. Our heartfelt gratitude goes out to everyone who has contributed-directly or indirectly-to the successful completion of this report.

Smarth Kaushal [102497023]
Aniket Gupta [102317169]
Aadar Mishra [102367002]

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Energy Consumption TRACKER

CONTENTS

Introduction

As cities continue to grow at an unprecedented pace, the demand for electricity has reached new heights. Imagine a future where, due to the depletion of energy resources and the inefficiencies embedded in urban systems, we may no longer have the luxury of switching on lights, fans, or even heating systems at will. With the global electricity demand rising exponentially, this scenario no longer seems like a distant possibility, but a looming reality. Empty skyscrapers that light up their entire structures at night, factories running inefficiently and overloading power grids during peak hours, and homes oblivious to their overconsumption of energy—all these factors exacerbate the energy crisis.

In the heart of this bustling city, where the hum of progress is deafening, Alex, a city planner, is troubled by the growing inefficiencies that lurk behind the shining lights of skyscrapers and the relentless hum of factories. He understands that it is not just about producing more energy, but about using it wisely. On the other side, Sophie, an environmental activist, is fighting against the rising tide of carbon emissions caused by poor energy management. She recognizes that renewable energy investments, though important, cannot address the systemic waste that permeates urban consumption patterns. The reality is that without a concerted effort to optimize consumption, society faces a future of blackouts, soaring costs, and ever-increasing carbon emissions.

It is within this context that this project takes form—an innovative solution designed to harness the power of artificial intelligence to predict and optimize energy consumption. Through machine learning models, specifically Linear Regression and Decision Tree Regression, the project aims to address the inefficiencies that currently plague urban energy systems. These models will serve as the foundation, providing a clear understanding of energy consumption patterns and trends. However, recognizing the limitations of these models, the project will continue to evolve, with plans to incorporate more advanced algorithms such as Random Forest and XGBoost as future enhancements. These sophisticated models hold the potential to refine predictions and optimize energy usage even further, bringing us closer to the vision of sustainable urban living.

By analyzing historical energy usage, factoring in weather trends, and accounting for sectoral demand, this project aims to transform raw data into actionable insights. These insights will empower cities to predict and manage energy consumption more effectively, reduce waste, and optimize the integration of renewable energy sources into the grid.

The goal is not just to predict energy demand but to create a system where energy consumption is intelligently optimized, enabling more sustainable urban living. In doing so, we hope to bridge the gap between renewable energy production and the wasteful patterns of consumption that continue to undermine our efforts toward a cleaner future. By employing advanced machine learning techniques, this project aims to provide a clear path toward a future where cities can thrive without sacrificing environmental sustainability or energy reliability. This is the vision—a future where intelligent energy management leads to more resilient, efficient, and sustainable cities.

Problem Statement

The accelerating depletion of finite energy resources has emerged as one of the most pressing challenges of the 21st century. This issue is not only a threat to global economic stability and environmental integrity, but it also raises questions of intergenerational equity. At the heart of this crisis is humanity's unsustainable consumption patterns, where the overutilization of fossil fuels and inefficient energy distribution systems threaten the very resources that power modern civilization. While advances in renewable energy infrastructure offer a glimmer of hope, the lack of intelligent demand forecasting and consumption optimization perpetuates a cycle of waste. This is evident in the needless illumination of empty skyscrapers, the inefficiencies of industrial complexes operating at peak capacity, and households unaware of their rising carbon footprints.

This inefficiency is not merely an economic concern—it represents a violation of the first law of thermodynamics, which states that energy within a closed system cannot be created or destroyed but only transformed. In today's energy landscape, mismanagement of resources effectively "destroys" energy's potential utility. This systemic waste results in a range of negative repercussions, including deteriorating air quality due to fossil fuel combustion, volatile energy markets that exacerbate geopolitical tensions, and aging power grids that buckle under avoidable strain. Traditional energy forecasting methods, which rely on static historical averages and simplistic regression models, fail to address the complexity of modern energy demand. They do not account for dynamic variables such as weather anomalies, temporal consumption fluctuations, and sector-specific usage behaviors, all of which significantly impact energy consumption patterns.

Current governance frameworks focus predominantly on renewable energy generation but remain disproportionately centered on supply-side solutions. While renewable energy investments are critical, they cannot address the underlying inefficiencies in demand-side management. Without an intelligent system for predicting energy consumption and optimizing its distribution, even the most advanced clean energy systems will fall short. This oversight creates a crucial gap—one that cannot be bridged by increasing supply alone. The growing demand for energy, paired with inefficient consumption practices, places tremendous strain on global energy systems and calls for a transformative approach to energy management.

Our project seeks to address this critical gap by leveraging machine learning models, including Linear Regression, Decision Tree Regression, Random Forest, and XGBoost, to forecast energy demand with unprecedented accuracy. By integrating temporal, meteorological, and sectoral data, we aim to develop a system that not only predicts future energy consumption but also identifies inefficiencies in real-world consumption patterns. The application of machine learning to energy forecasting represents a paradigm shift—one that transforms energy stewardship from a reactive process to a proactive, intelligent system. This system will allow for more efficient use of resources, mitigate carbon emissions, and enable smarter integration of renewable energy sources.

While XGBoost provides a powerful predictive tool with its ability to capture complex relationships in the data, the project also acknowledges the value of other models like Linear Regression, Decision Trees, and Random Forests in contributing to a comprehensive solution. Each model offers unique advantages, and by comparing their performance, we aim to identify the most effective combination of approaches for energy forecasting and optimization.

In conclusion, this project goes beyond technical innovation—it calls for a collective reimagining of energy management, where algorithmic intelligence bridges the gap between human behavior and the planet's limits. The implementation of advanced machine learning models provides a critical step toward mitigating the inefficiencies of current energy systems and ensuring a sustainable energy future for generations to come. This work not only demonstrates the potential of AI to optimize energy consumption but also sets the foundation for a smarter, more efficient energy landscape.

Objectives

1. **Develop a Robust Machine Learning Model to Predict Energy Consumption Accurately:**

   Our primary objective is to build an advanced machine learning system that can accurately predict energy consumption across different sectors. By using historical energy usage data, weather conditions, and temporal dynamics, we aim to create a predictive model that can foresee energy demand and consumption patterns with high precision. This system will leverage a range of algorithms, including Linear Regression, Decision Tree Regression, Random Forest, and XGBoost, to ensure a diverse approach to forecasting and optimizing energy usage.

2. **Reduce Electricity Waste in Residential, Commercial, and Industrial Sectors:**

   The project is designed to tackle inefficiencies in energy consumption, particularly in residential, commercial, and industrial sectors. By identifying and addressing areas where energy is wasted, such as in empty buildings, idle machinery, and excessive heating/cooling, we aim to reduce unnecessary electricity usage. This will not only lower costs but also optimize energy consumption patterns across these sectors, contributing to more sustainable energy usage.

3. **Help Businesses Optimize Power Usage and Lower Operational Costs:**

   Many businesses struggle with high operational costs due to inefficient energy use. This project aims to provide organizations with actionable insights on how to optimize their power consumption, particularly during peak demand hours. By leveraging predictive analytics, businesses will be able to adjust their energy usage, schedule operations more efficiently, and avoid high-cost consumption periods, leading to significant cost savings.

4. **Contribute to a Sustainable Future by Lowering Carbon Footprints and Minimizing Environmental Impact:**

   The ultimate goal of this project is to contribute to a more sustainable future by reducing the carbon footprints of individuals, businesses, and industries. By reducing energy waste, we can decrease reliance on fossil fuels, leading to a reduction in greenhouse gas emissions. Through more efficient energy usage, the project will help mitigate the environmental impact of global energy consumption, supporting international efforts to combat climate change.

5. **Enhance the Integration of Renewable Energy Sources into Power Systems:**

   As the world transitions toward renewable energy sources, one of the challenges is ensuring that these intermittent sources (such as solar and wind) are effectively integrated into existing power grids. The model developed in this project will help predict and balance renewable energy production with consumption patterns, enabling a more seamless integration of clean energy into the grid. This will help reduce dependence on non-renewable energy sources and increase the overall sustainability of energy systems.

6. **Improve Demand Response Capabilities for Energy Providers and Consumers:**

   Effective demand response is crucial to balancing energy supply and demand. By providing real-time consumption data and forecasts, this project will help energy providers manage load fluctuations more efficiently. It will also enable consumers to better understand their energy usage patterns and adjust their behavior accordingly. This two-way interaction will contribute to more balanced energy consumption, reducing the need for backup generation during peak times and helping avoid blackouts.

7. **Promote Public Awareness of Energy Efficiency and Sustainable Consumption:**

   In addition to technical goals, the project seeks to raise public awareness about the importance of energy efficiency and sustainable consumption. By providing users with clear and actionable insights into their energy usage, we can empower individuals to make informed decisions about their energy consumption. This, in turn, will encourage more widespread adoption of energy-saving technologies and behaviors.

8. **Provide Scalable Solutions for Global Adoption:**

   This project is designed with scalability in mind, meaning the solution can be adapted for use in cities, regions, and countries with diverse energy landscapes. The methodology and algorithms used in the project can be easily modified to accommodate different types of data, making it a versatile tool for addressing energy efficiency challenges worldwide.

9. **Facilitate Energy Equity and Access for Underserved Communities:**

   Ensuring that all communities have access to affordable, reliable, and sustainable energy is a critical global challenge. By optimizing energy consumption, this project can help reduce costs in low-income communities, improve access to electricity, and support equitable energy distribution. The system can be customized to meet the specific needs of underserved regions, ensuring that energy resources are used efficiently and equitably.

10. **Foster Innovation in the Energy Sector through Data-Driven Solutions:**

    This project seeks to catalyze innovation within the energy sector by demonstrating the power of machine learning and artificial intelligence in optimizing energy usage. By showcasing the potential of AI in this domain, we hope to inspire further research and development of smart energy solutions that can drive the future of energy management, consumption, and sustainability.

In summary, the project aims to provide a comprehensive, AI-driven solution to optimize energy consumption, reduce waste, lower operational costs, and contribute to a sustainable energy future. By addressing inefficiencies across various sectors and fostering smarter energy management, we can make a significant impact on global energy systems, improve environmental sustainability, and support efforts to mitigate climate change.

Methodology

"Regression is the art of learning from patterns in the past to predict the shape of tomorrow."

*Building on the foundational understanding of the problem and objectives, the **Methodology** section outlines the step-by-step approach taken to develop a robust energy consumption prediction model. This phase begins with comprehensive data preprocessing and feature engineering to ensure the dataset is clean, informative, and suitable for machine learning. Following this, different regression algorithms—Linear Regression and Decision Tree Regression—are implemented and evaluated based on key performance metrics. The methodology is designed not only to compare model effectiveness but also to provide a framework that can be extended with advanced techniques like Random Forest and XGBoost in future iterations.*

Phase 1: Data Acquisition & Exploratory Data Analysis

1. Data Collection and Loading:

Objective: Gather historical energy consumption and weather datasets, merge them, clean them, and prepare them for analysis.

Data Sources

For this project, we have utilized real-time energy consumption and weather data sourced from credible and authoritative platforms. The energy usage statistics were collected from:
https://iced.niti.gov.in/energy/electricity/distribution/national-level-consumption/load-curve,
Copyright © 2025, NITI Aayog. While the platform primarily relies on official sources, in certain instances, some assumptions have been made or data has been derived. These details are mentioned explicitly within the platform. Although we believe the data to be reliable and sufficiently comprehensive, NITI Aayog ICED does not guarantee absolute accuracy and accepts no liability for consequences arising from the use of this information.
Complementing this, the weather data — including variables such as temperature, humidity, and solar radiation — was obtained from https://www.visualcrossing.com/weather-query-builder/, ensuring both temporal precision and contextual relevance.
The combined dataset spans from January 2023 to April 2024, offering a robust 15-month window that captures seasonal variations and demand patterns across different weather conditions. By integrating these datasets, we aim to effectively capture the dynamic relationship between environmental factors and energy consumption — paving the way for smarter forecasting models and sustainability-driven insights.

2. Preliminary Data Inspection:

- ✓ Identifying Numerical and Categorical Features
- ✓ Summary Statistics of Numerical Features
- ✓ Checking Datatypes
- ✓ Checking Missing Values
- ✓ Checking Unique Values
- ✓ Checking Duplicate Records

3. Drawing Insights: Visualizing Data Trends:

Objective: Identify consumption patterns, seasonality, and anomalies.
Libraries Used: matplotlib , seaborn , plotly

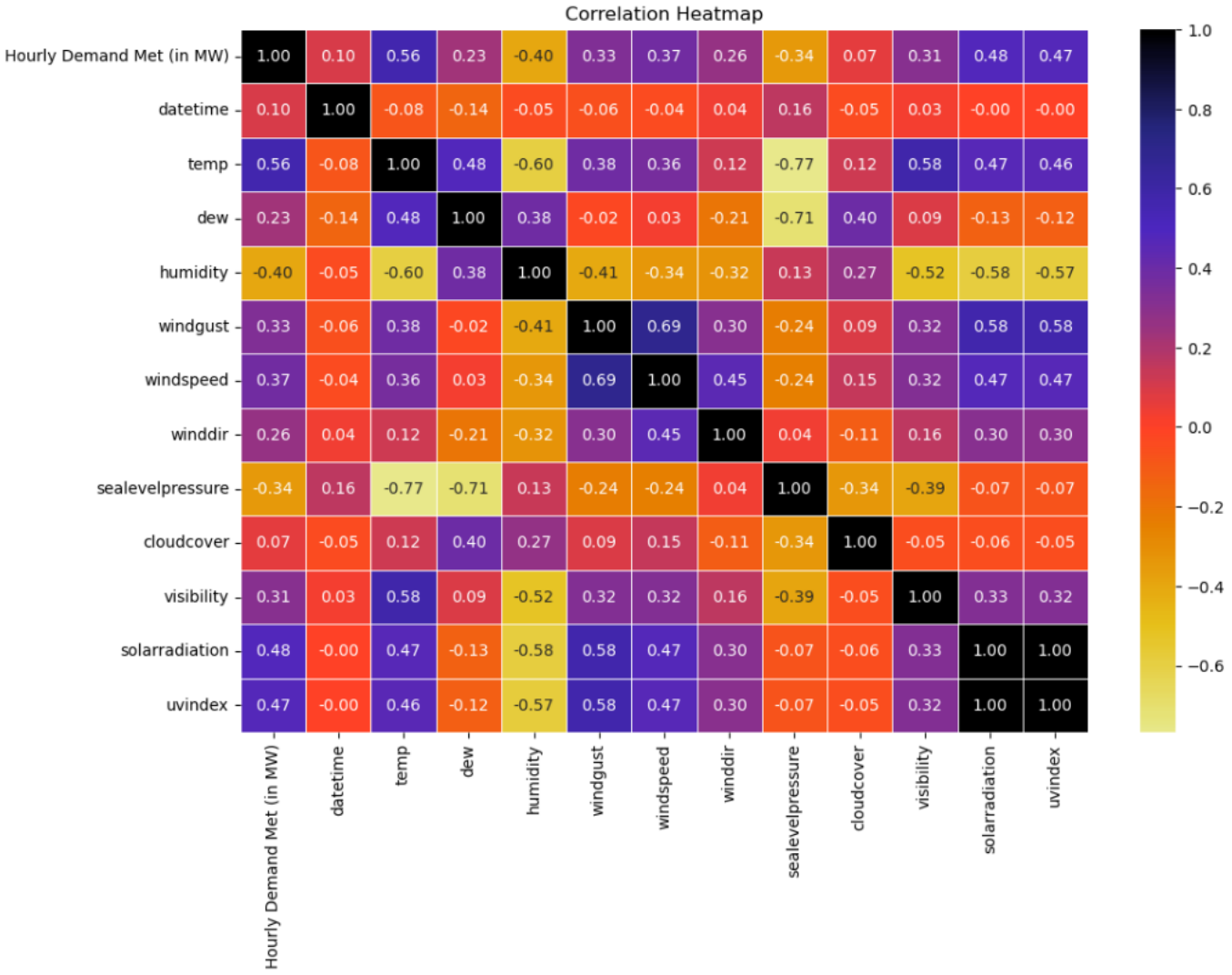
Investigation Tasks:

- ✓ Summary Statistics and Missing Value Overview
- ✓ Distribution Patterns (skewness, normality)
- ✓ Outlier Detection using Boxplots
- ✓ Correlation Heatmap
- ✓ Categorical Features Identification
- ✓ Feature vs Target (Hourly Demand Met) Trends
- ✓ Temporal Trends (Hourly, Daily, Monthly)

Mathematical Concept:

Pearson Correlation Coefficient:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$



Phase 2: Data Preprocessing & Feature Engineering

4. Handling Missing Values

Techniques: Mean/Median Imputation

5. Outlier Detection and Removal

Outliers have been detected using boxplot.

Removal Technique used: Interquartile Range (IQR) Method

$$\boldsymbol{IQR = Q3 - Q1}$$

Lower Bound: $Q1 - 1.5 * IQR$

Upper Bound: $Q3 + 1.5 * IQR$

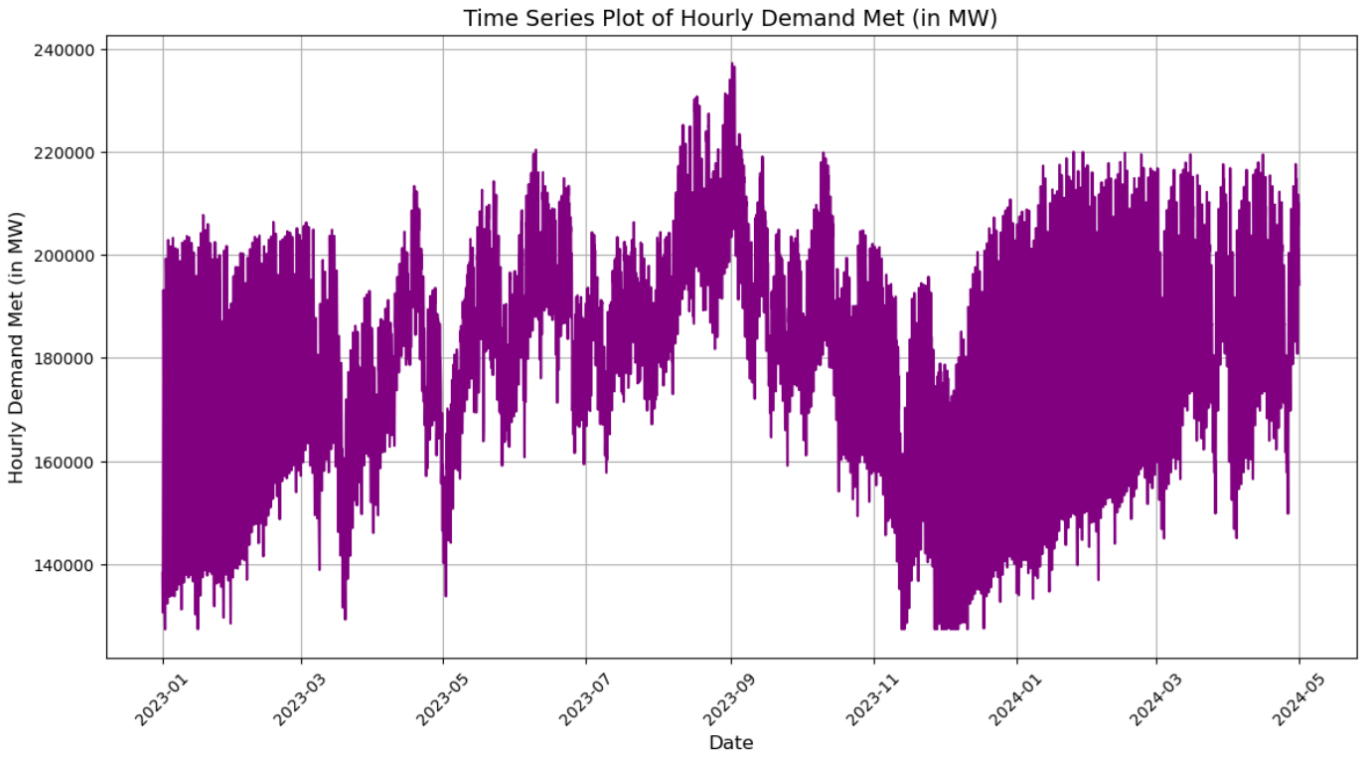
All those values lying outside this interval have been capped.

6. Feature Engineering

Extracting Time-Based Features:

Extracted the time-based features like Hour, Day, Month, Season, Quarter, Year, day of year, day of month, week of year etc. and performed EDA on the same.

All the graphs have been shown and presented in the jupyter notebook.



Lag Features: Previous 1-hour, 24-hour rolling mean, 7-day rolling mean.

Lag Features & Rolling Statistics: Lag features and rolling statistics are crucial for capturing temporal dependencies in time series data. By adding lag features, we are essentially creating new columns that represent the previous time steps (e.g., previous hour, previous day) in the dataset. Similarly, rolling statistics (like rolling mean, rolling standard deviation) are used to capture short-term fluctuations and smooth trends.

Lag Features:

A **lag feature** is a value of the target variable (or another feature) from a previous time step. Lag features are useful when past values have a strong relationship with the current or future target variable.

Given a time series y_t, a lag feature at time (t) with a lag of (n) is defined as:

$$\text{Lag}_n = y_{t-n}$$

Rolling Mean Smoothing:

Rolling Mean (or Moving Average) helps in reducing noise from the time series data by taking the average of data points over a defined window. It helps in identifying **trends** and **seasonal patterns** more clearly.

Let y_t be the original value at time t, and w be the window size. Then the Rolling Mean at time t is:

$$\text{RollingMean}_t = \frac{1}{w} \sum_{i=t-w+1}^{t} y_i$$

Rolling Statistics:

Rolling statistics help summarize the data over a moving window, giving a clearer view of trends and variations over time.

Rolling Standard Deviation at time (t) with window size (w):

$$\text{RollingStd}_t = \sqrt{\frac{1}{w} \sum_{i=t-w+1}^{t} (y_i - \text{RollingMean}_t)^2}$$

Phase 3: Model Training & Optimization

In this phase, the focus shifts to building robust machine learning models capable of accurately predicting energy consumption based on the features engineered and insights derived during the exploratory data analysis phase. Model training is a critical step where we take the cleaned and preprocessed data and feed it into various machine learning algorithms. The goal is to develop models that can capture the underlying patterns in energy usage while minimizing error. This process involves selecting the right model, tuning hyperparameters, evaluating model performance using appropriate metrics, and continuously optimizing the model to ensure the best generalization to unseen data. As we move forward, we will explore two essential regression models: Linear Regression and Decision Tree Regression. These models will serve as strong baseline models, allowing us to compare their performance and later fine-tune or switch to more advanced methods based on their results.

In this study, both **Linear Regression** and **Decision Tree Regressor** models were trained using an **80-20 train-test split**, where 80% of the dataset was allocated for training and the remaining 20% for testing the model's generalization ability. After training, the models were evaluated using standard regression performance metrics to assess their predictive capabilities. These metrics include **Mean Absolute Error (MAE)**, which measures the average absolute difference between actual and predicted values; **Mean Squared Error (MSE)**, which emphasizes larger errors by squaring them; **Root Mean Squared Error (RMSE)**, providing error in the same units as the target variable; and the **R^2 Score (Coefficient of Determination)**, which indicates how well the model explains the variance in the target variable. These metrics collectively provide a comprehensive view of model accuracy, bias, and consistency.

Formulae:

- Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{MSE}$$

- R^2 Score (Coefficient of Determination)

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Notes:

- Always check that you are evaluating on unseen validation/test data.
- Lower RMSE and MAE indicate better model performance.
- An R^2 Score close to 1 indicates a good fit.

Linear Regression

Goal: To establish a linear relationship between energy consumption (dependent variable) and the various features (independent variables) in the dataset.

Mathematical Concept:

Linear Regression aims to model the relationship between two or more variables by fitting a linear equation to the observed data. The equation for simple linear regression (with one feature) is:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Where:

- y is the dependent variable (energy consumption).
- β_0 is the intercept.
- β_1 is the coefficient of the feature.
- x is the independent variable (predictor).
- ε is the error term.

For multiple features, the equation extends to:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

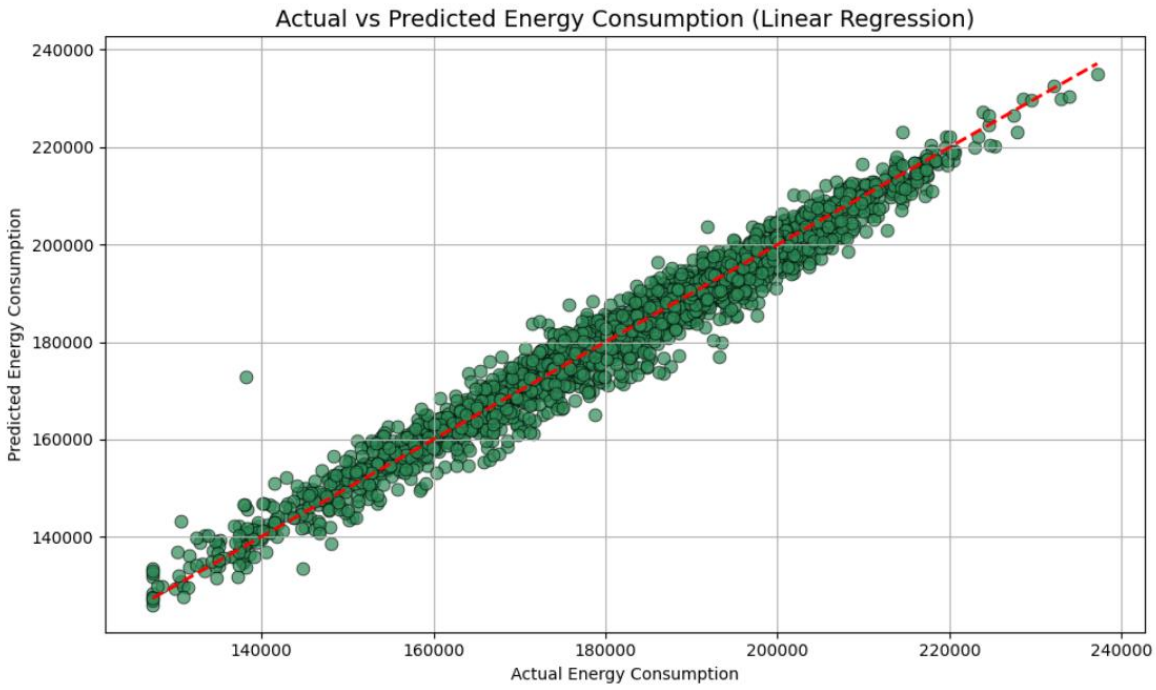
In this equation, each feature x_n has an associated coefficient β_n, which represents its weight in predicting energy consumption. The goal of Linear Regression is to find the values of $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ that minimize the **sum of squared residuals (errors)**.

Linear Regression Metrics:

MAE: 2789.2579

RMSE: 3668.5255

R^2 Score: 0.9664



Decision Tree Regression

Introduction: A **Decision Tree** is a non-linear machine learning algorithm used for both classification and regression tasks. In the case of regression, the model predicts continuous values by splitting the data at each decision node based on a feature value. It constructs a tree-like model of decisions and their possible consequences, where the leaf nodes represent the predicted continuous values for each segment of the data.

Goal: To model the relationship between features and energy consumption using a non-linear decision tree approach. Decision Trees split the data into smaller and more homogeneous subsets, where each leaf node represents a predicted value.

Key Components of a Decision Tree for Regression:

1. **Nodes:**
   - **Root Node**: The initial node that represents the entire dataset. It is the starting point of the decision tree.
   - **Leaf Nodes**: The terminal nodes of the tree where the predictions are made. Each leaf node contains the predicted continuous value.
   - **Decision Nodes**: Internal nodes that represent decisions or tests on a feature. Each decision node splits the data into two or more child nodes based on a chosen feature and threshold.

2. **Splitting:**
   The decision tree splits the dataset at each decision node based on the feature that minimizes the variance of the target variable. For regression, this process aims to reduce the **variance** (or minimize **mean squared error**) within the child nodes.

3. **Homogeneity:**
   The algorithm tries to make the data in each child node as homogeneous as possible, meaning the target variable values in each node should be close to each other.

4. **Stopping Criteria:**
   The tree-building process stops when a predefined criterion is met, such as:
   - Reaching a maximum tree depth.
   - Having fewer than a minimum number of samples in a node.
   - When further splits do not significantly reduce variance.

5. **Prediction:**
   The prediction for a new data point is determined by traversing the tree. Starting from the root node, the algorithm makes decisions at each internal node based on the feature values until it reaches a leaf node. The prediction is the mean of the target values of the samples in that leaf.

Mathematical Foundation:

1. **Variance Reduction (Best Split Selection):**
   The goal is to split the data at each node in a way that reduces the variance within the child nodes. Variance at a node t is calculated as:

$$V(t) = \frac{1}{n_t} \sum_{i=1}^{n_t} (y_i - \hat{y}_t)^2$$

   Where:

   - y_i is the target value of sample i.
   - $\hat{y}_t$ is the mean target value in the node t.
   - n_t is the number of samples in node t.

2. **Mean Squared Error (MSE) for Decision Making:**
   The decision tree algorithm evaluates splits based on the reduction in the **mean squared error** (MSE). The MSE for a node t is calculated as:

$$\text{MSE}_t = \frac{1}{n_t} \sum_{i=1}^{n_t} (y_i - \hat{y}_t)^2$$

   A **Decision Tree** recursively splits the data into subsets based on the values of the features. At each node, the decision tree chooses the feature that maximizes the **information gain** or minimizes the **mean squared error (MSE)** in the case of regression. This approach creates a tree structure where each node represents a decision based on one feature, and each branch represents a possible outcome of that decision.

The splitting criterion at each node is determined by minimizing the variance (or MSE) within the child nodes:

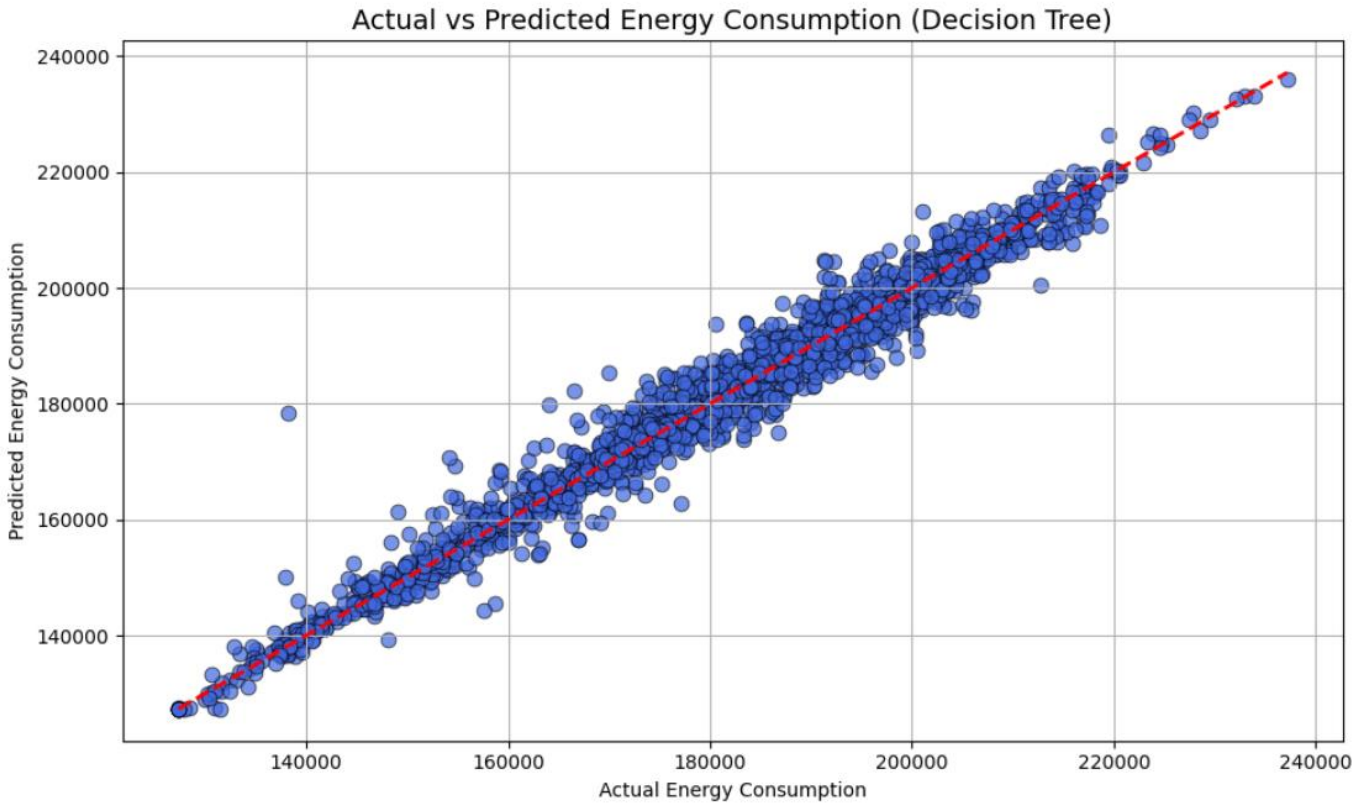
$$\text{MSE}_{\text{split}} = \frac{1}{N_{\text{left}}} \sum_{i=1}^{N_{\text{left}}} (y_i - \hat{y}_{\text{left}})^2 + \frac{1}{N_{\text{right}}} \sum_{i=1}^{N_{\text{right}}} (y_i - \hat{y}_{\text{right}})^2$$

Where:

- N_{left} and N_{right} are the number of data points in the left and right child nodes, respectively.
- $\hat{y}_{left}$ and $\hat{y}_{right}$ are the predicted values in the left and right child nodes.

3. **Prediction:**
   Once the tree is built, predicting a new sample involves following the path from the root to a leaf node based on the feature values of the sample. The predicted value for the sample is the mean of the target values of the samples in that leaf node.

   Decision Tree Metrics:
   MAE: 2286.5941
   RMSE: 3329.7746
   R^2 Score: 0.9723



Advantages of Decision Tree Regression:

1. **Interpretability**: Decision trees are easy to understand and visualize. The decision-making process can be directly traced.
2. **Non-Linearity**: Unlike linear models, decision trees can capture non-linear relationships between the features and the target.
3. **No Need for Feature Scaling**: Unlike other algorithms, decision trees do not require normalization or standardization of input features.
4. **Automatic Feature Selection**: Decision trees automatically select the most significant features for the splits.

Disadvantages of Decision Tree Regression:

1. **Overfitting**: Decision trees tend to overfit, especially with deep trees, as they learn the noise in the data. Overfitting can be mitigated by pruning the tree or using ensemble methods like Random Forest.
2. **Instability**: Small changes in the dataset can lead to a completely different tree, making decision trees sensitive to small data variations.
3. **Bias Towards Features with More Levels**: Decision trees can favor features with many categories when splitting.
4. **Poor Performance on Extrapolation**: Decision trees do not generalize well to unseen data that lies outside the range of the training data.

Hyperparameters for Decision Tree Regression:

1. **max_depth**: The maximum depth of the tree. Limiting the depth helps prevent overfitting.
2. **min_samples_split**: The minimum number of samples required to split an internal node.
3. **min_samples_leaf**: The minimum number of samples required to be at a leaf node.
4. **max_features**: The number of features to consider when looking for the best split.
5. **max_leaf_nodes**: Limits the number of leaf nodes in the tree.
6. **criterion**: The function to measure the quality of a split (e.g., "mse" for mean squared error).

Practical Considerations:

- **Pruning**: Decision trees can easily overfit, and pruning helps reduce the complexity of the tree by removing nodes that provide little predictive power. Post-pruning is a common technique where branches are removed after the tree has been fully grown.
- **Ensemble Methods**: To address the instability and overfitting of individual decision trees, ensemble methods like **Random Forest** and **Gradient Boosting Machines (GBM)** are often used. These methods aggregate multiple decision trees to create more robust models.

Decision Tree Regression is a flexible and interpretable model that works well with both linear and non-linear relationships. However, it is prone to overfitting, especially on noisy data. Techniques like pruning or using ensemble methods (e.g., Random Forest or XGBoost) are recommended for better performance and generalization.

XGBoost:

For implementing the **XGBoost** model, a **70-15-15 split** strategy was adopted, dividing the dataset into **70% training**, **15% validation**, and **15% testing**. Initially, model performance was assessed using standard evaluation metrics—**Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, **Root Mean Squared Error (RMSE)**, and **R^2 Score**—on the **validation set**. This informed the hyperparameter tuning process to optimize the model. Once the best hyperparameters were determined, the **final model was evaluated on the test set** to ensure unbiased performance measurement and validate its generalization capability.

XGBOOST

- Handles missing data internally.
- Ensemble of weak learners (trees).
- Regularization prevents overfitting.
- Fast and efficient gradient boosting.

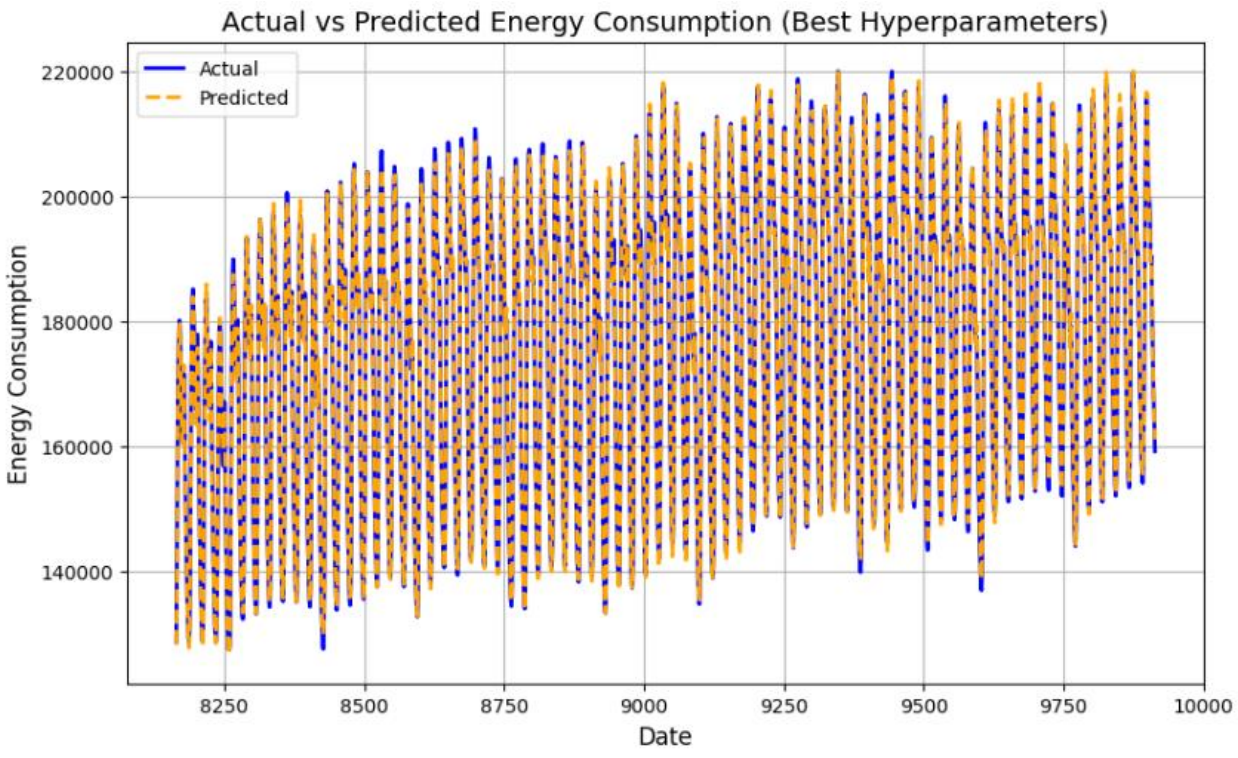
Key Terminologies:

Term	Explanation
Base Learner	A simple model (usually decision tree)
Boosting	Sequential learning from previous errors
Objective Function	Loss + Regularization
Regularization	Penalizes model complexity
Loss Function	Measures prediction error

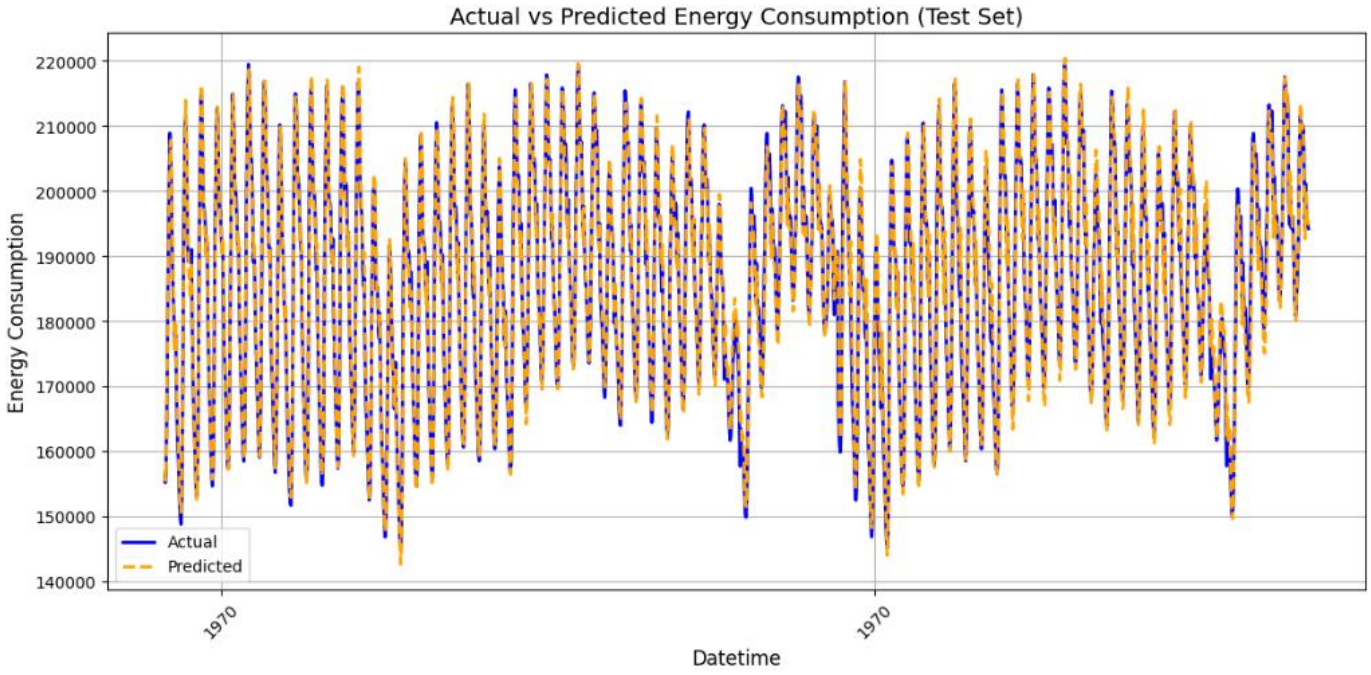
Important Hyperparameters:

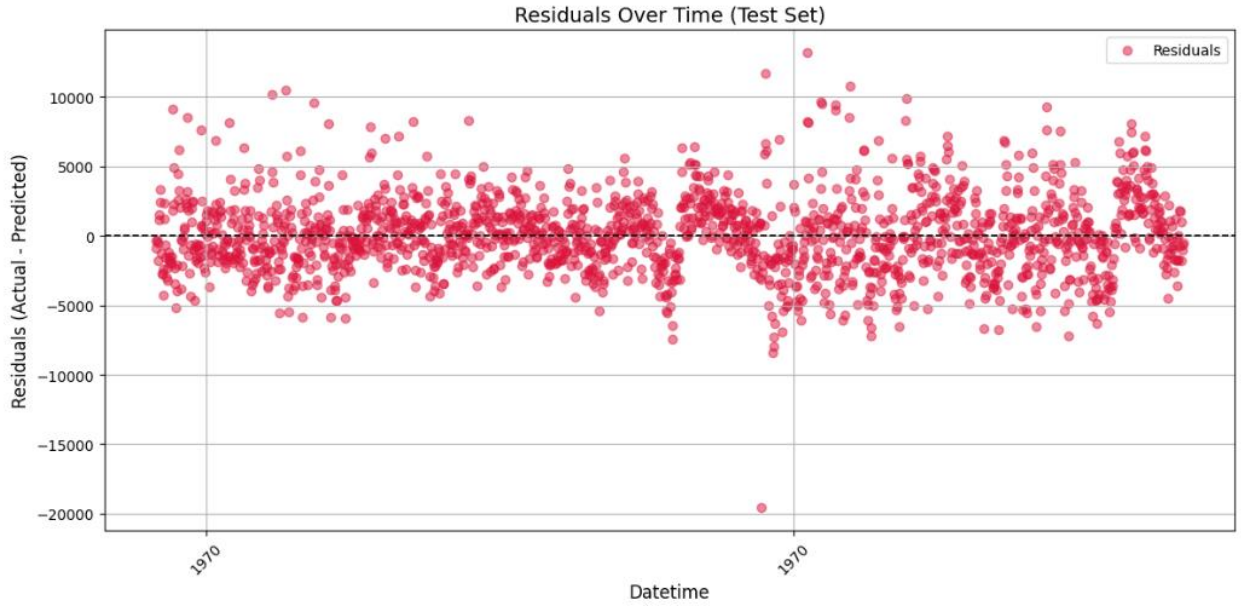
Hyperparameter	Meaning	Effect
`n_estimators`	Number of trees	More trees = better fit but slower
`learning_rate`	Step size shrinkage	Lower = slower but more accurate
`max_depth`	Tree depth	Larger = more complex patterns
`subsample`	Row sampling per tree	Controls overfitting
`colsample_bytree`	Feature sampling per tree	Controls overfitting

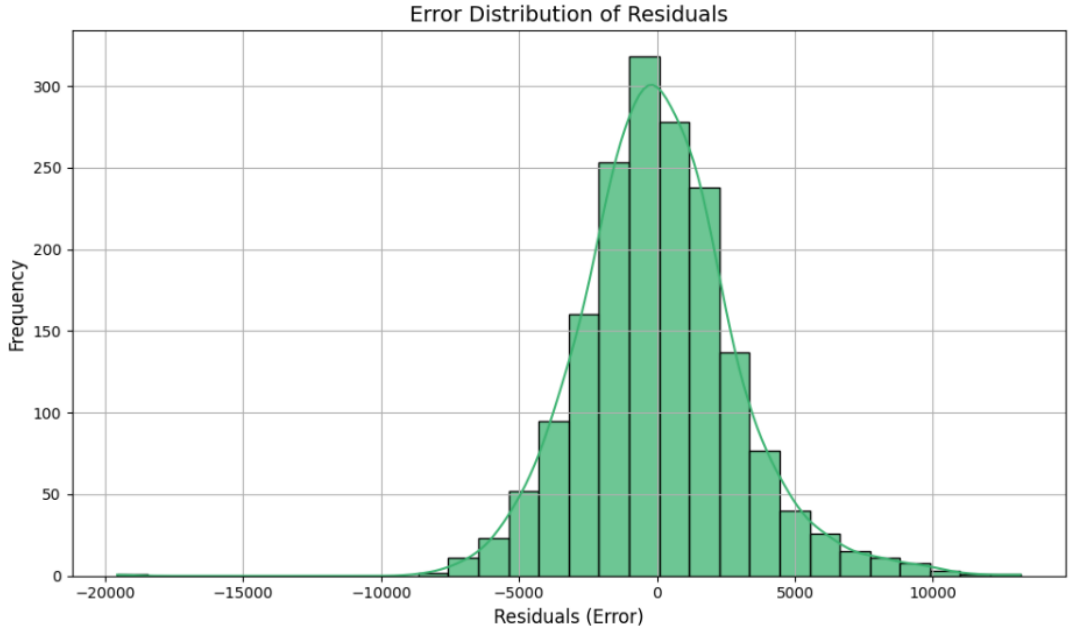
Final Hyperparameters of the XGBoost Model:{'colsample_bytree': 1.0,
'learning_rate': 0.0246,
'max_depth': 6,
'n_estimators': 1500,
'subsample': 0.8}



RMSE with best hyperparameters: 2351.5689933187964


Actual vs Predicted Energy Consumption (Test Set)
Energy Consumption
Datetime
Actual
Predicted
220000
210000
200000
190000
180000
170000
160000
150000
140000
1970
1970


Residuals Over Time (Test Set)
Residuals (Actual - Predicted)
Datetime
Residuals
10000
5000
0
−5000
−10000
−15000
−20000
1970
1970


Error Distribution of Residuals
Frequency
Residuals (Error)
300
250
200
150
100
50
0
−20000
−15000
−10000
−5000
0
5000
10000

Phase-4: Model Comparison (PyCaret vs XGBoost)

All the models implemented have been compared including XGBoost Regressor with PyCaret's Gradient Boosting Regressor (GBR).

Evaluation Metrics: RMSE, MAE, R^2 Score
Outcome: Whichever model shows lower error (MAE, RMSE) and higher (R^2) is selected for dashboard deployment.

FINAL RESULTS:

📊 Performance Comparison: Manual XGBoost vs PyCaret XGBoost

Metric	Manual XGBoost	PyCaret XGBoost
MAE	2125.12	3431.2131
RMSE	2815.2927	4406.6328
R^2 Score	0.9725	0.9513
MSE	—	19418414.0000
RMSLE	—	0.0257
MAPE	—	0.0196

Key Observations:

- The **manually trained XGBoost model** outperforms the PyCaret model in all common metrics (MAE, RMSE, and R^2).
- PyCaret's model shows slightly weaker predictive power, possibly due to:
  - Generic preprocessing pipeline
  - Broader and default hyperparameter search space
- The manual model benefits from **custom feature engineering and tailored tuning**, leading to significantly better performance.

Final Conclusion:

This comparison clearly demonstrates the value of **manual model development**—from curated data preprocessing to custom hyperparameter tuning—in delivering a more optimized and accurate XGBoost model for energy consumption forecasting.

While PyCaret offers a fast and automated baseline, **manual tuning unlocks superior performance** tailored to the problem domain. The additional effort in manual modeling directly translates to improved metrics and a deeper understanding of the data dynamics.

"Automation can accelerate the journey, but craftsmanship defines the destination."

This reinforces our project's approach: balancing the speed of automation with the precision of manual intervention to achieve the most reliable and actionable predictions.

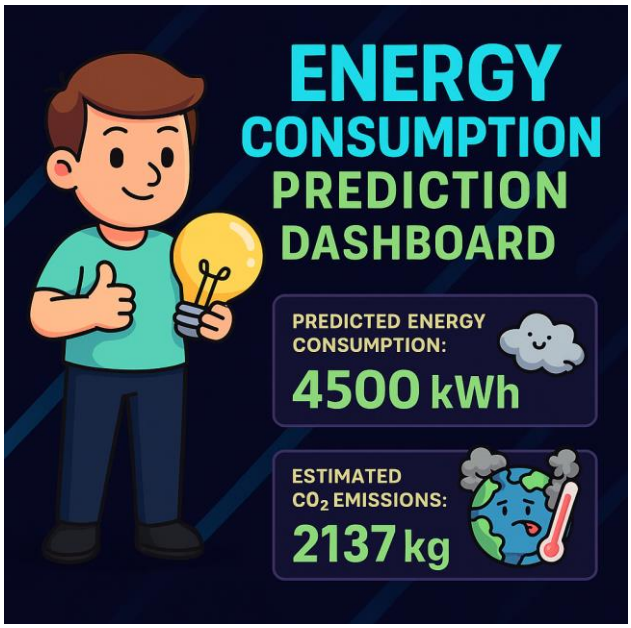
Concluding Remarks:

In this project, we explored and implemented multiple machine learning models—Linear Regression, Decision Tree Regressor, and XGBoost—to predict energy consumption effectively. While Random Forest was considered, it has been reserved for future enhancements to maintain scope and performance balance. Among all, XGBoost was selected as the final model due to its robust performance and widespread recognition in state-of-the-art predictive modeling tasks. This strategic choice allowed us to optimize both accuracy and computational efficiency. Each model was rigorously evaluated using appropriate metrics such as RMSE and R^2, ensuring robust validation across training, validation, and test sets.

From the outset, this project was driven by a vision. Alex imagined a smart city where lights switch off automatically in vacant buildings, industries dynamically manage power loads, and households benefit from intelligent energy-saving suggestions. Sophie, on the other hand, saw a greener planet—where AI-powered predictions play a crucial role in reducing carbon emissions, one insight at a time. This journey—from raw data to impactful predictions—embodies their vision and lays the foundation for a smarter, more sustainable future.

Phase-5: Dashboard Development

To truly experience the essence of our project, we invite you to explore the interactive dashboard—a seamless blend of innovation, technology, and purpose. Designed with user-friendly navigation and real-time visualizations, the dashboard brings our predictive model to life, allowing users to observe, analyze, and act on energy consumption patterns like never before. Whether you're a policymaker, a business leader, or a curious mind, this platform offers a window into the future of intelligent energy management. Let it not just be a tool you visit, but a step you take toward a smarter, greener, and more sustainable world. Come, be part of this vision—one click at a time.



The dashboard (deployed on Streamlit Community Cloud) allows users to **upload their own CSV files** for custom analysis or provide **manual input**, where the system smartly auto-fills current date, time, and other contextual features—while still offering full control for custom entries. It supports **real-time prediction of future energy consumption**, powered by our trained machine learning models, and provides actionable **suggestions to reduce carbon footprints** and optimize energy use. Users can instantly see **how much energy can be saved**, reinforcing eco-conscious decisions. With these features and a built-in **Carbon Footprint Calculator**, we invite you to explore how artificial intelligence can guide smarter energy consumption and help shape a greener tomorrow—one prediction at a time.

Visit https://energy-demand-forecasting-umd8asuuq5yyqm974d8cud.streamlit.app/

A glance at the dashboard below. Click on the link and see it in action. Experience the pulse!

References:

- https://github.com/archd3sai/Hourly-Energy-Consumption-Prediction/blob/master/PJME.ipynb
- https://github.com/arunsivakumar5/XGBoost-Decision-tree-for-Predicting-Energy-consumption
- https://github.com/MohamadNach/Machine-Learning-to-Predict-Energy-Consumption/blob/master/Machine_Learning_to_predict_Energy_Consumption.ipynb
- https://dimleve.medium.com/xgboost-mathematics-explained-58262530904a
- https://arxiv.org/pdf/1603.02754
- https://www.sciencedirect.com/science/article/abs/pii/S2210537923000185
- https://www.kaggle.com/code/songulerdem/energy-consumption-forecasting-with-pycaret
- https://www.dexma.com/blog-en/forecasting-energy-consumption-using-machine-learning-and-ai/
- https://onlinelibrary.wiley.com/doi/10.1155/2024/6812425
- https://www.youtube.com/watch?v=vV12dGe_Fho

🌙 Behind the Code: Our Journey

Every project has a story—and ours is no exception. What began as a shared vision turned into countless midnights filled with research papers, bugs, and breakthroughs. We juggled online meetings, cracked jokes between commits, and sometimes stared blankly at screens hoping for inspiration (or just fewer errors). And let's not forget the times when our laptops would hang, making us wonder if the universe was sending us a sign to take a break—or maybe just restart the system.

There were moments of frustration—when models wouldn't converge or when the dataset seemed to be working against us. But there were also moments of pure joy—when plots finally made sense, metrics improved, and our dashboard came to life. Some nights we worked to the beats of hip-hop playlists, turning debugging into a dance-off, and other times, we simply laughed off the chaos together.

We weren't just building a machine learning project—we were building memories, skills, and friendships. Through coffee-fueled brainstorming, shared Google Docs chaos, and that one weird bug that still haunts us, we realized this: the process mattered just as much as the product. This journey was driven by passion, perseverance, and a shared dream to make a difference—even if it meant a few sleepless nights, too many Git conflicts, and those endless laptop freezes. 😊

As we come to the close of this chapter, we're thrilled to share the outcome with you. ***Readers, enthusiasts, and fellow researchers are warmly invited to explore the codebase on Github Repository, visualizations, and the interactive dashboard to gain deeper insights into our approach.*** *While this marks a significant milestone in our journey, it is only the beginning. Future plans include extending this work using deep learning models for even more accurate and adaptive predictions. We also envision transforming this project into a comprehensive research paper to contribute meaningfully to the growing body of knowledge in energy analytics and sustainability. We hope you enjoy the journey as much as we did, and stay tuned as we continue to enhance and expand the project—perhaps with deep learning models and more insights to come!*