Unlocking Tomorrow's Power: The Art of Energy Demand Forecasting

Data Science Nanodegree Capstone Project Report

1. Introduction

In our modern world, where electricity powers our homes, businesses, and industries, the ability to predict energy demand is more critical than ever. Energy demand forecasting is both an art and a science, playing a pivotal role in ensuring a stable and efficient energy supply. Accurate forecasts enable utilities, policymakers, and businesses to make informed decisions, allocate resources efficiently, and reduce environmental impacts. In this blog post, we'll delve into the fascinating world of energy demand forecasting, exploring its importance, methods, challenges, and the future of this essential practice.

1.1 Importance of Energy Demand Forecasting:

Energy demand forecasting serves as the backbone of the energy sector. Its significance can be distilled into several key points:

- **Resource Allocation:** Accurate forecasting helps utility companies allocate resources optimally. They can plan maintenance schedules, invest in infrastructure, and secure energy supplies accordingly.
- **Grid Management:** Predicting demand allows grid operators to maintain a stable and reliable power supply. This is especially crucial in preventing blackouts and brownouts.
- **Energy Pricing:** Forecasting enables energy providers to set prices that reflect the expected demand. This can contribute to more equitable pricing and help consumers manage their energy costs.
- **Environmental Impact:** By anticipating energy demand, we can minimize the environmental impact of energy generation. For instance, renewables can be integrated more effectively into the grid when we know when and where energy will be needed.

1.2 Factors Influencing Energy Demand:

A variety of factors affect the energy demand:

- **Economic Growth**: Higher economic activity often leads to increased energy consumption.

- **Population Growth**: A growing population generally results in higher energy demand.

- **Technological Advancements**: New technologies can either increase or decrease energy consumption, depending on their efficiency and adoption.

- **Weather Conditions**: Extreme temperatures can significantly affect energy usage for heating and cooling.

- **Government Policies**: Regulations and incentives can influence energy consumption patterns, such as energy efficiency programs.

1.3 Types of Forecasting:

In general, three types of forecasting are done depending on the time-period of the forecast:

- **Short-term Forecasting**: Predicts energy demand over hours to a few days, crucial for real-time grid management and supply chain logistics.

- **Medium-term Forecasting**: Focuses on weeks to months and helps with maintenance planning and resource allocation.

- **Long-term Forecasting**: Extends to several years or decades and is essential for infrastructure development and energy policy formulation.

1.4 Methods of Energy Demand Forecasting:

Energy demand forecasting relies on various methods, each suited to different applications and timeframes:

- **Time Series Analysis**: This method involves analyzing historical data to identify patterns, seasonality, and trends. It's useful for short-term and medium-term forecasting.

- **Machine Learning**: Advanced algorithms and machine learning models can handle large datasets and complex variables, making them valuable for both short-term and long-term forecasts.
- **Weather-Based Forecasting**: Weather conditions have a significant impact on energy demand. Meteorological data is integrated into forecasting models to predict energy usage more accurately.

- **Econometric Models**: These models combine economic indicators, such as GDP and population growth, with historical energy consumption data to make long-term forecasts.

Some of these methods will be explained in the next section.

1.5 Challenges in Energy Demand Forecasting:

Despite the advancements in forecasting techniques, challenges persist:

- **Uncertainty:** Energy demand can be influenced by unpredictable events, such as extreme weather, economic fluctuations, and technological disruptions.
- **Data Quality:** Forecasting accuracy relies on high-quality, up-to-date data. Inaccurate or incomplete data can lead to erroneous predictions.
- **Renewable Energy Integration:** As renewable sources like solar and wind become more prevalent, forecasting their intermittent output becomes increasingly challenging.
- **Regulatory Changes:** Changing regulations and policies can impact energy demand patterns, requiring forecasters to adapt quickly.

2. Methods for Energy Demand Forecasting

Different models can be employed to predict energy demand, and the choice of model depends on the specific requirements of the task and the available data. Here are some commonly used models for energy demand forecasting:

1. **Time Series Models**:
    - **ARIMA (AutoRegressive Integrated Moving Average)**: ARIMA models are well-suited for univariate time series data. They capture temporal patterns, seasonality, and trends in energy consumption. ARIMA models can be extended to SARIMA (Seasonal ARIMA) for handling seasonality.
    - **Exponential Smoothing Models**: Exponential smoothing methods, such as Holt-Winters, are used to forecast time series data by giving different weights to recent observations. They are effective for capturing seasonality and trends.
    - **Prophet**: Developed by Facebook, Prophet is a time series forecasting tool designed for forecasting with daily observations that display patterns on different time scales. It can handle missing data and outliers.
2. **Regression Models**:
    - **Linear Regression**: Simple linear regression can be used to predict energy demand based on one or more independent variables like temperature, time of day, and economic indicators. Multiple linear regression extends this to multiple predictors.
    - **Polynomial Regression**: In cases where the relationship between predictors and energy demand is nonlinear, polynomial regression can capture more complex patterns.
3. **Machine Learning Models**:

- **Random Forest**: Random forests are ensemble models that can capture complex relationships between predictors and energy demand. They are robust and handle both numerical and categorical variables.
- **Gradient Boosting (XGBoost, LightGBM, CatBoost)**: Gradient boosting algorithms like XGBoost, LightGBM, and CatBoost are powerful for regression tasks. They excel in handling large datasets and complex interactions between features.
- **Neural Networks**: Deep learning models, such as feedforward neural networks and recurrent neural networks (RNNs), can be used for energy demand forecasting. Long Short-Term Memory (LSTM) networks are particularly effective for time series data.

4. **Hybrid Models**:
    - **ARIMA with Exogenous Variables (ARIMAX)**: This combines the strength of ARIMA models with external factors or predictors (exogenous variables) like weather data to improve forecasting accuracy.
    - **Machine Learning and Time Series Hybrid**: Combining machine learning models like XGBoost with time series methods like ARIMA or LSTM can leverage both temporal patterns and external predictors.
5. **Statistical Models**:
    - **Gaussian Processes**: Gaussian processes are probabilistic models that can capture uncertainty in energy demand forecasting. They are suitable for situations where uncertainty estimation is crucial.
6. **Long Short-Term Memory (LSTM) Networks**:
    - **Deep learning-based LSTM networks can capture long-term dependencies and seasonality patterns in time series data. They are especially effective when dealing with sequences of data, such as hourly or daily energy consumption records.
7. **Hybrid Models with Domain Knowledge**:
    - Some forecasting tasks benefit from incorporating domain-specific knowledge. For example, a model that combines machine learning with physics-based models for HVAC systems can yield more accurate predictions for building energy demand.

The choice of model should be based on factors such as the nature of the data, the forecasting horizon (short-term or long-term), the availability of external variables, and the desired level of accuracy and interpretability. It's often beneficial to experiment with multiple models and select the one that performs best for a specific energy demand forecasting task. Additionally, ensemble methods and model combination techniques can further enhance forecasting accuracy by leveraging the strengths of multiple models.

In this work, we have used XGBoost, a popular ML algorithm, for energy demand forecasting.

3. Why XGBoost?

XGBoost, short for Extreme Gradient Boosting, is a robust and versatile machine learning algorithm that has gained widespread popularity in various domains due to its outstanding performance and versatility. Here are some reasons why XGBoost is an excellent choice for energy demand forecasting:

1. **Accuracy**: XGBoost is known for its accuracy in predicting both linear and nonlinear relationships within the data. This is crucial for capturing complex patterns in energy consumption data, which can be influenced by various factors such as weather, time of day, and economic conditions.
2. **Feature Importance**: XGBoost provides a built-in feature importance score, allowing us to understand which variables have the most significant impact on energy demand. This insight can be valuable for making informed decisions and optimizing energy management strategies.
3. **Handling Missing Data**: Energy consumption data can often have missing values or irregularities. XGBoost can handle missing data gracefully, reducing the need for extensive data preprocessing.
4. **Scalability**: XGBoost is highly scalable and can handle large datasets with millions of data points and numerous features, making it suitable for real-world energy demand forecasting applications.

4. XGBoost for Energy Demand Forecasting

Now, let's dive into the steps that we used to implement energy demand forecasting using XGBoost:

4.1 Data Collection and Pre-processing

We used the Hourly Energy Usage Data from PJM. PJM Interconnection LLC (PJM) is a regional transmission organization (RTO) located within the United States. It operates within the Eastern Interconnection grid, managing an electric transmission system that serves various areas, including Delaware, Illinois, Indiana, Kentucky, Maryland, Michigan, New Jersey, North Carolina, Ohio, Pennsylvania, Tennessee, Virginia, West Virginia, and the District of Columbia. The data provided here represents the hourly electricity consumption figures sourced from PJM's official website, measured in megawatts (MW). Please note that the covered regions have undergone changes over time, which may result in data availability being limited to specific dates for each region.

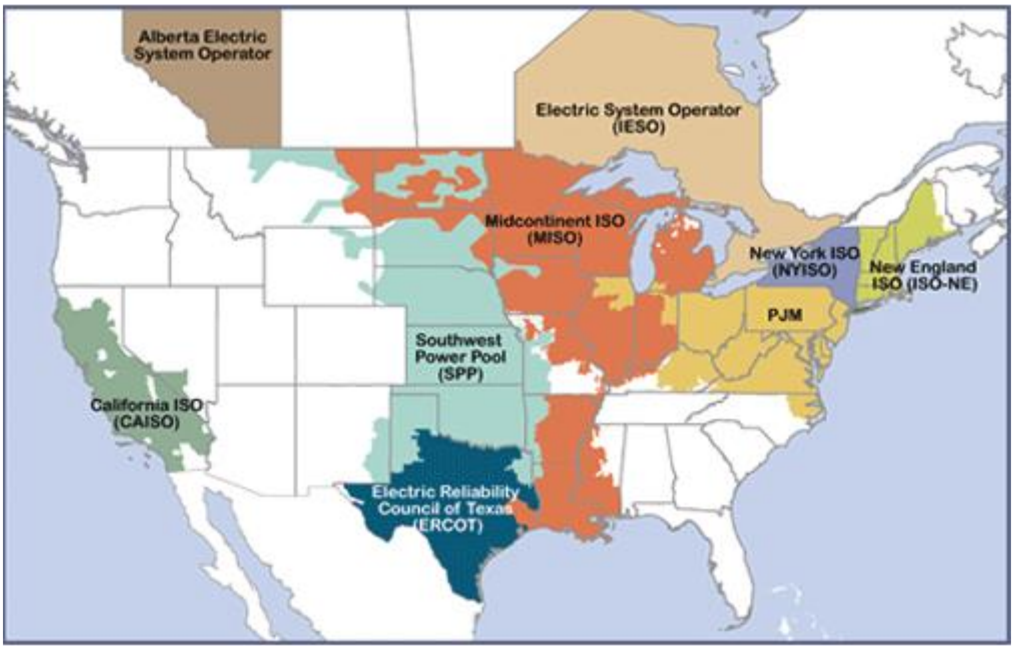

Figure 1. PJM Regions as of 2021 (Source: Wikipedia)

The dataset used can be found here: https://www.kaggle.com/datasets/robikscube/hourly-energy-consumption

No major data preprocessing was required because the dataset was clean and reliable.

4.2 Feature Engineering

Feature engineering is a crucial step in energy demand forecasting, as it involves creating and selecting relevant input variables to improve the accuracy of your predictive models. Here, we used time-based features where we extracted features like day of the week, month, hour etc to capture pattern in energy consumption.

4.3 Model Training

We used a baseline model for energy demand forecasting as a straightforward and minimalistic starting point. The baseline model that we used is a persistence model. A persistence model is a simple and baseline approach for energy demand forecasting that uses historical data to predict future values. In a persistence model, the forecast for the next time step is based on the actual value observed at the previous time step. This approach assumes that the future will be similar to the recent past.

As a ML model, we used the XGBoost library in Python to train your energy demand forecasting model.

4.4 Feature Importance Analysis

After training the model, we analyzed feature importance scores to gain insights into which variables most strongly affect energy demand. This information can guide decision-making and resource allocation.

4.5 Model Evaluation

We evaluated the model's performance on the test dataset using Root Mean Squared Error (RMSE).

5. Results

Now, let's look at some of the results.

5.1 Data Visualization

The entire dataset has the total load values from 2002 to early 2019. It is clear from the plot that there are trends and seasonality in the data.

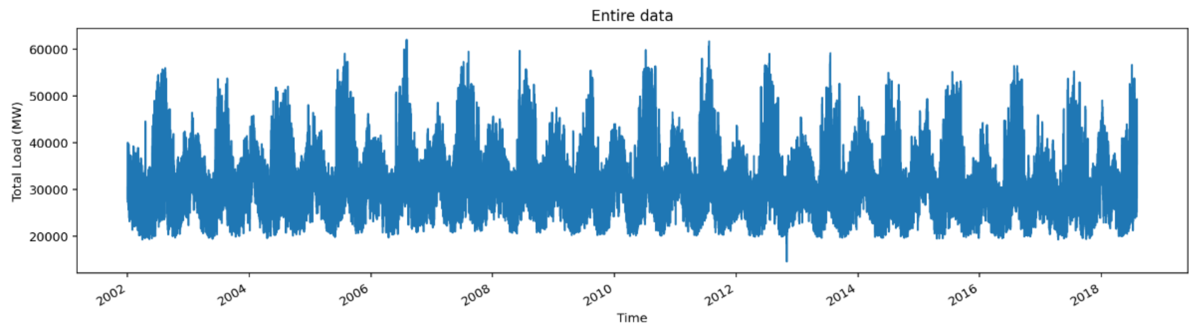

Figure 2. Timeseries plot of the entire dataset.

The same hold trues for a smaller timescale, for instance a week of data.

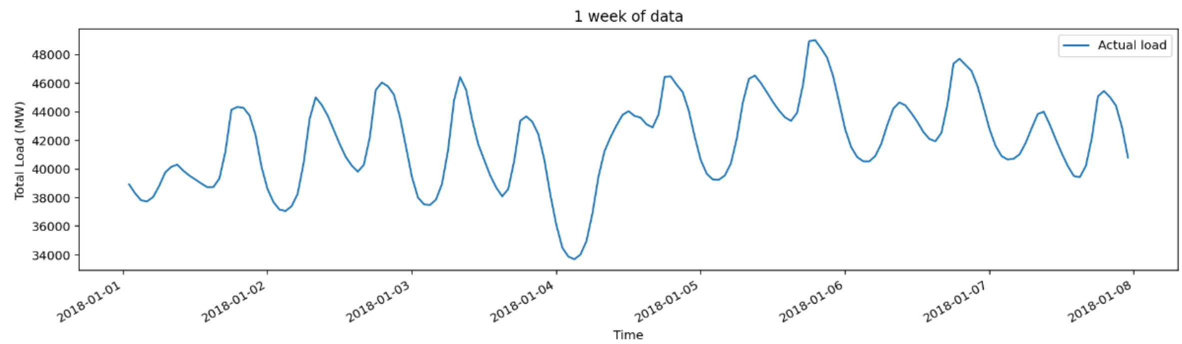

Figure 3. Timeseries plot for the actual load in the first week of 2018.

5.2 Data Decomposition

We decomposed the timeseries data to look for trends and seasonality. composing time series data involves breaking down a time series into its constituent components to better understand and model its underlying patterns. The typical components of a time series are:

1. **Trend Component:** The underlying long-term behavior or direction in the data. It represents the gradual increase or decrease in the time series over time.
2. **Seasonal Component:** The repeating patterns or cycles that occur at regular intervals, such as daily, weekly, or yearly. Seasonal patterns are often influenced by factors like the day of the week, month, or season.
3. **Residual Component (Noise):** The irregular or random fluctuations in the data that cannot be attributed to the trend or seasonal patterns. These are essentially the unpredictable elements of the time series.

To decompose time series data, we used the statsmodel library in Python. We found that decomposing the daily load data did not give any valuable insights due to the granularity. Hence, we resampled the data to weekly and monthly sums.

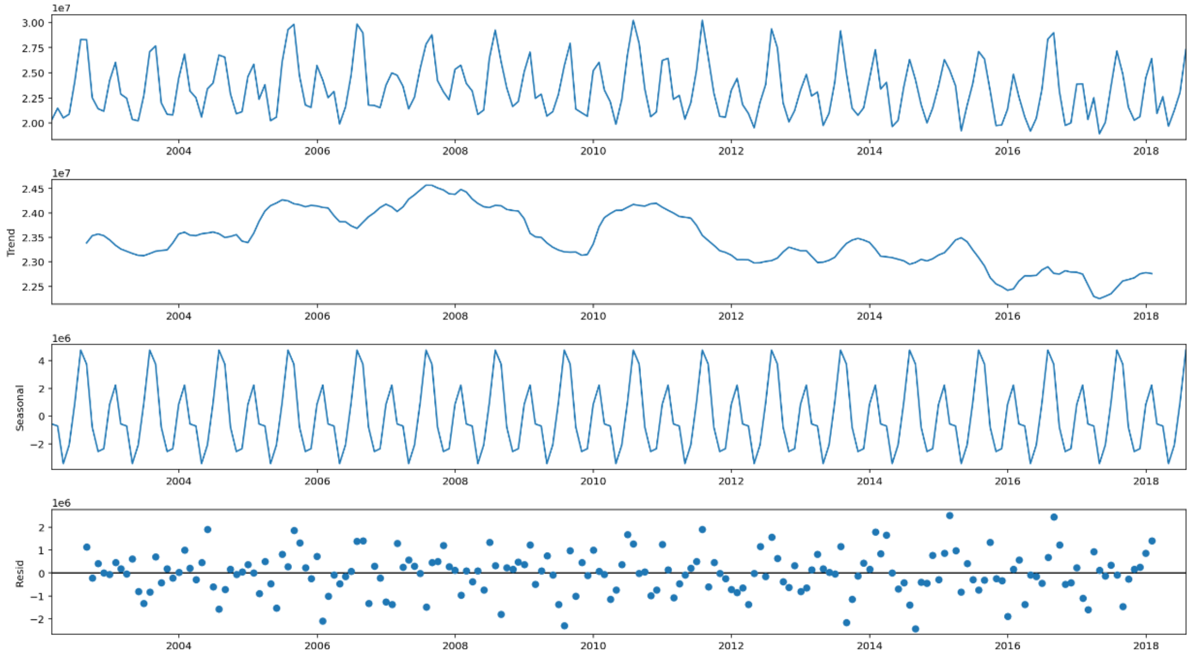

Figure 4. Decomposition of the dataset resampled monthly using statsmodel.

The decomposition of the electricity usage time series data has revealed several important observations that provide valuable insights into the underlying patterns and factors influencing energy consumption:

1. **Seasonal and Repetitive Nature:** The presence of distinct peaks and troughs in the electricity usage data strongly suggests the existence of seasonal and repetitive patterns. This could be attributed to various factors, including:
    - **Weather Patterns:** Weather conditions can have a significant impact on energy consumption. For instance, in colder months, heating systems

are often used more extensively, while in hotter months, air conditioning usage tends to rise.

- **Holidays:** Holidays can also influence electricity usage patterns. On weekends and certain holidays, when many offices and non-essential businesses are closed, there may be a decrease in electricity demand compared to regular workdays.

2. **Impact of Office Hours and Workdays:** The fluctuations in electricity usage tied to office hours and workdays highlight the importance of commercial and industrial activities in shaping energy consumption. Understanding these patterns can be instrumental in optimizing energy management strategies for businesses.

3. **Long-Term Decline:** The decrease in electricity usage observed in the latter years of the dataset is a significant finding. While the exact reasons behind this decline may require further investigation, it raises several hypotheses:

    - **Efficiency Improvements:** It's possible that businesses and organizations have implemented energy-efficient technologies and practices over time, leading to reduced energy consumption.

    - **Economic Factors:** Changes in the economy, such as shifts in industrial composition or economic downturns, could influence energy demand.

    - **Demographic Changes:** Population shifts or changes in occupancy patterns in the region covered by the dataset might affect electricity usage.

    - **Policy and Regulation:** Government policies and regulations aimed at energy conservation and sustainability could also play a role in reducing energy consumption.

In light of these observations, further analysis and exploration may be warranted. For example, conducting a regression analysis to identify the specific drivers behind the declining trend in electricity usage or exploring the impact of specific holidays and weather conditions on energy demand can provide deeper insights. Additionally, these insights can be used to develop more accurate energy demand forecasting models and to implement energy management strategies that are responsive to the identified patterns and factors.

5.3 Feature Importance

As mentioned earlier, we used time-based features where we extracted features like day of the week, month, hour etc. The results are shown below.

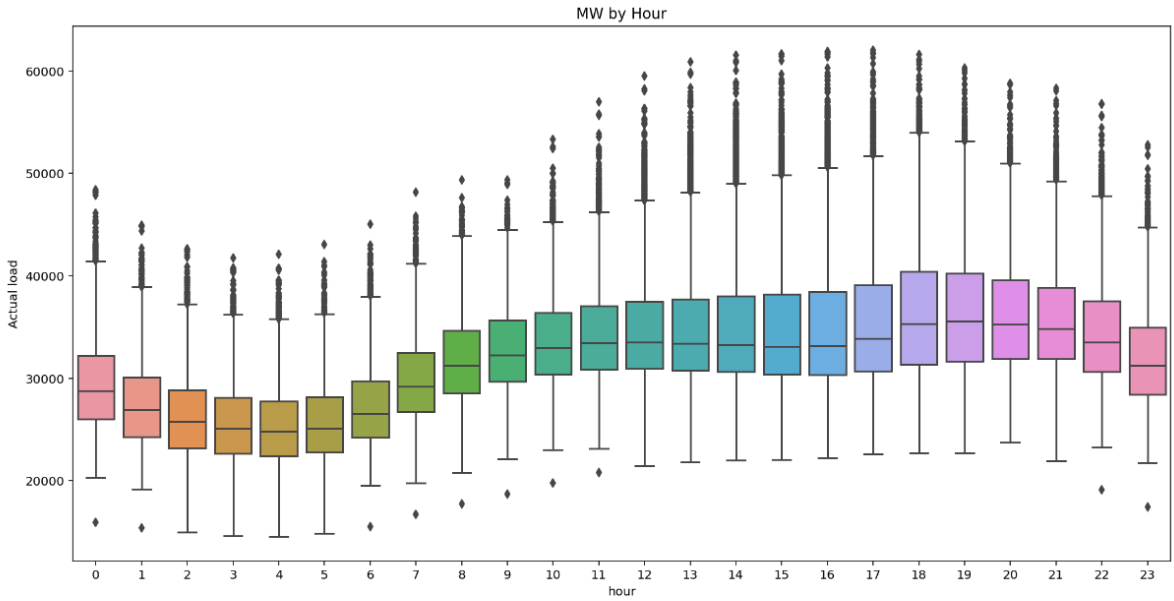

Figure 5. Relationship between the actual load and the hour of the day.

The analysis of the electricity load patterns reveals compelling insights into the temporal dynamics of energy consumption and its interaction with various external factors, with a particular focus on the daily load cycle:

1. **Minimum Load in the Early Hours:** The observation that the minimum load occurs during the early hours of the day aligns with common patterns of reduced activity. At this time, residential areas are often in a state of rest, with households sleeping, and many businesses and industrial facilities remaining closed. As a result, there is a decreased demand for electricity during these predawn hours.

2. **Maximum Load Post-Noon:** The spike in electricity demand occurring post-noon signifies the transition from the morning period of reduced activity to a period of heightened productivity and business operations. As the day progresses, various commercial and industrial activities resume, leading to a substantial surge in energy consumption. This trend can be attributed to factors such as:

    - **Business Hours:** Many offices, factories, and institutions typically open for the day around this time, requiring electricity for lighting, equipment, and HVAC systems.

    - **Lunchtime Activities:** The midday peak may also reflect increased energy use in cafeterias, restaurants, and food preparation facilities during lunch hours.

3. **Busiest Time from 11 AM to 9 PM:** The span from 11 AM to 9 PM represents the most active and energy-intensive period on the grid. This time frame encompasses the core operational hours of businesses and commercial establishments. It is during this interval that the grid experiences its highest levels

of electricity demand, and efficient management of resources and infrastructure is crucial to meet the requirements.

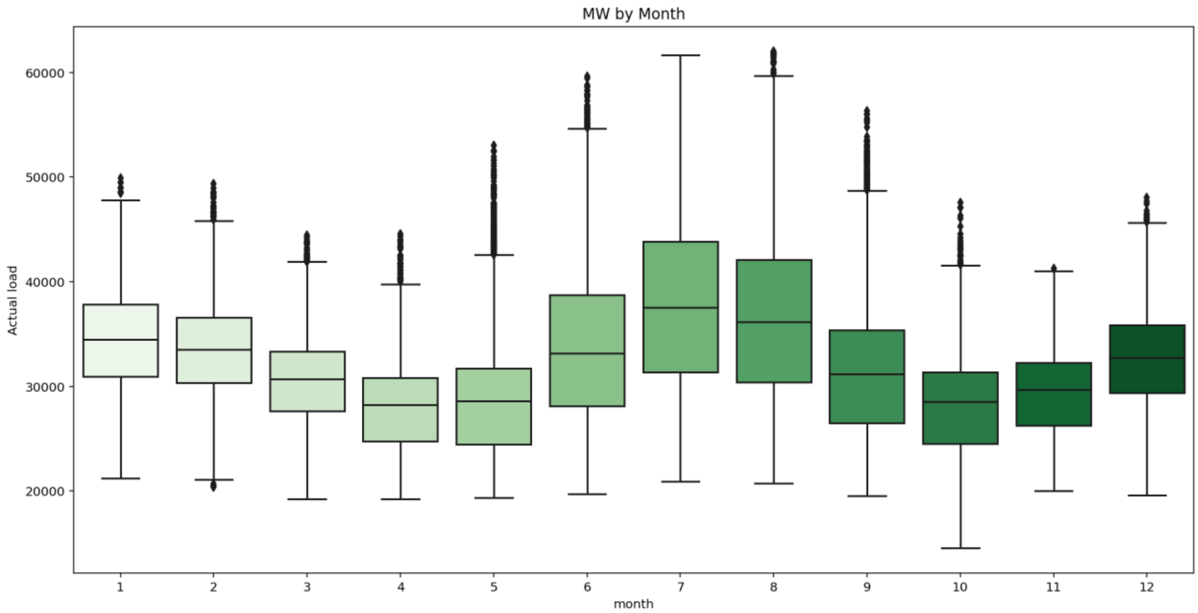

Figure 6. Relationship between the actual load and the month of the year.

The observation that the maximum and minimum electricity load occurs during the month of July and April, respectively, provides valuable insights into the seasonal variability of energy consumption in the region. Let's expand on this:

1. **Maximum Load in July - The Height of Summer:**

July being the month with the highest electricity demand aligns with the peak of summer in the region. Several factors contribute to this surge in energy consumption:

- **Increased Use of Air Conditioning:** One of the most significant drivers of high electricity demand in July is the soaring summer temperatures. As the weather gets hotter, residential, commercial, and industrial facilities rely heavily on air conditioning systems to maintain comfort. This heightened usage of air conditioning units leads to a substantial increase in electricity consumption.
- **Extended Daylight Hours:** Longer daylight hours during the summer season can also impact electricity usage. Although natural lighting reduces the need for artificial lighting during the day, the overall demand for electricity remains high due to cooling and refrigeration requirements.
- **Industrial Operations:** Industries that operate year-round, especially those involved in manufacturing and processing, may experience increased energy demands during the summer to power equipment and machinery.

- **Outdoor Activities and Entertainment:** People tend to engage in more outdoor activities during the summer months, but this does not necessarily translate to lower energy demand. Outdoor venues, such as amusement parks and stadiums, often require substantial energy inputs for lighting and equipment operation.

2. **Minimum Load in April - Transitional Season:**

The observation that the minimum electricity load occurs in April reflects the transitional nature of this season. Several factors contribute to the lower energy demand during this time:

- **Moderate Weather:** In April, the weather is often milder compared to the extremes of summer and winter. Consequently, there is reduced reliance on energy-intensive heating or cooling systems, resulting in lower overall electricity consumption.
- **Reduced Heating and Cooling Needs:** With temperatures in the transitional zone, both heating and cooling demands are relatively low. This leads to a balanced energy profile with less pronounced peaks and troughs.
- **Spring Holidays:** In some regions, April may coincide with spring holidays or breaks, during which businesses and schools might be closed or operating at reduced capacity, contributing to reduced electricity usage.

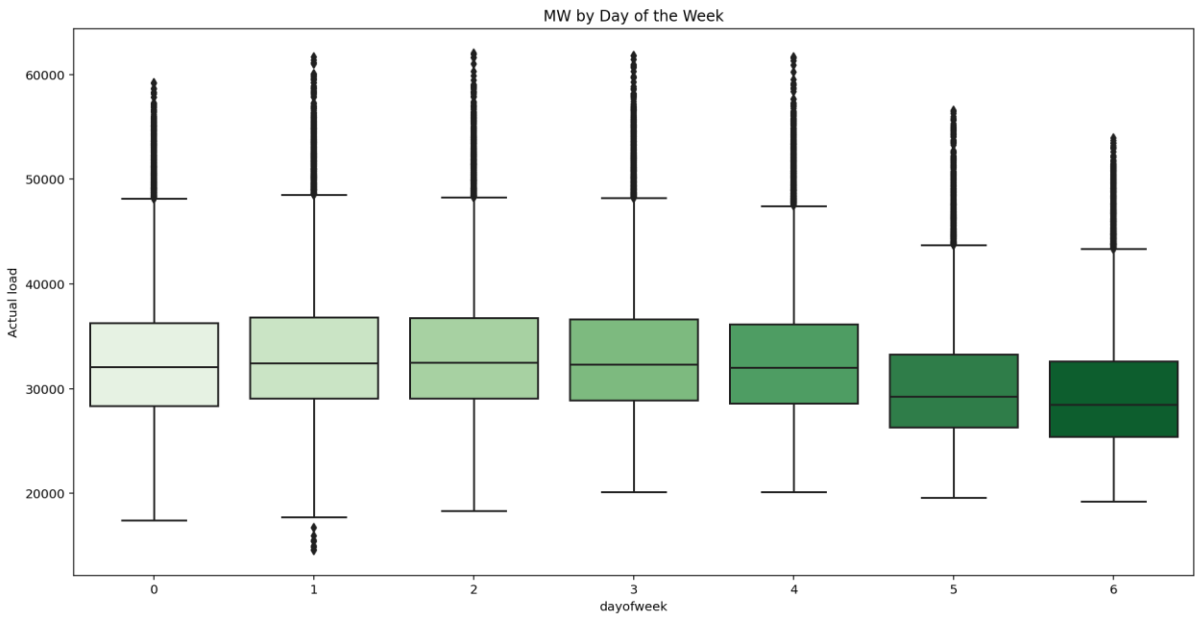

Figure 7. Relationship between the actual load and the day of the week.

The observation that the lowest electricity load occurs during the weekend when most businesses are closed is a common pattern in many regions. Let's expand on this observation:

1. **Weekend vs. Weekday Load Variation:**
    - **Weekend Lull:** Weekends, typically Saturdays and Sundays, often exhibit a noticeable decrease in electricity demand compared to weekdays. This phenomenon is primarily due to the closure or reduced activity of many businesses, offices, and industrial facilities during weekends.
    - **Commercial and Industrial Impact:** Businesses and industries are among the largest consumers of electricity. They operate various equipment, lighting, and HVAC systems during their regular workdays. However, over the weekend, a significant portion of these establishments shuts down or operates with minimal energy consumption.
2. **Factors Contributing to Weekend Load Reduction:**
    - **Business Closure:** Many offices and retail establishments close on weekends, leading to reduced lighting, heating, cooling, and equipment usage. This directly contributes to lower electricity demand.
    - **Industrial Pause:** Factories and manufacturing plants often curtail production or operate at a reduced capacity during weekends, leading to lower energy consumption for machinery and processes.
    - **Office Buildings:** Office buildings, which constitute a substantial share of commercial electricity usage, typically experience minimal occupancy during weekends. This results in lower energy requirements for lighting, climate control, and office equipment.
    - **Shift in Residential Demand:** While residential electricity usage may increase during weekends due to more people being at home, the reduction in commercial and industrial demand typically outweighs this increase, resulting in an overall decrease in total load.
3. **Implications for Grid Management:**
    - **Load Forecasting:** Grid operators and energy planners consider these load variations when forecasting electricity demand. Accurate load forecasting is essential for ensuring a stable and reliable power supply.
    - **Resource Optimization:** Understanding the weekend lull allows for better optimization of energy generation and distribution resources. It helps avoid overproduction and ensures that resources are allocated efficiently.
    - **Maintenance and Upkeep:** Weekends also offer an opportunity for maintenance and repair work on power infrastructure since there is typically lower demand. This can contribute to grid reliability.

5.4 Statistical Testing

Before embarking on model development, it was imperative to conduct statistical tests to assess the suitability of the data for traditional statistical methodologies.

5.4.1 Stationarity

Our initial focus was on evaluating stationarity, a term denoting the constancy of a dataset's characteristics over time. Non-stationarity implied that external factors, beyond the mere passage of time, influenced the data, causing trends and seasonality to evolve. Non-stationary datasets, also known as "Random Walk" data, were notably challenging to predict.

To assess stationarity, we employed the Augmented Dickey–Fuller (ADF) Test. In statistics, the Augmented Dickey–Fuller test (ADF) examined the null hypothesis of a unit root's presence within a time series sample. The alternative hypothesis varied depending on the specific version of the test employed but typically related to stationarity or trend-stationarity. [Source: Wikipedia (https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test)]. The ADF test showed that the data is stationary.

5.4.2 Autocorrelation

Autocorrelation measures the degree of relationship between two points in time, such as the extent to which today's value is related to last week's value.

We are checking the autocorrelation function (ACF) and the partial autocorrelation function (PACF). PACF represents a variation of ACF because it assesses residual correlation after eliminating the effects already explained by preceding time lags. This approach helps in avoiding a "cumulative" correlation effect.

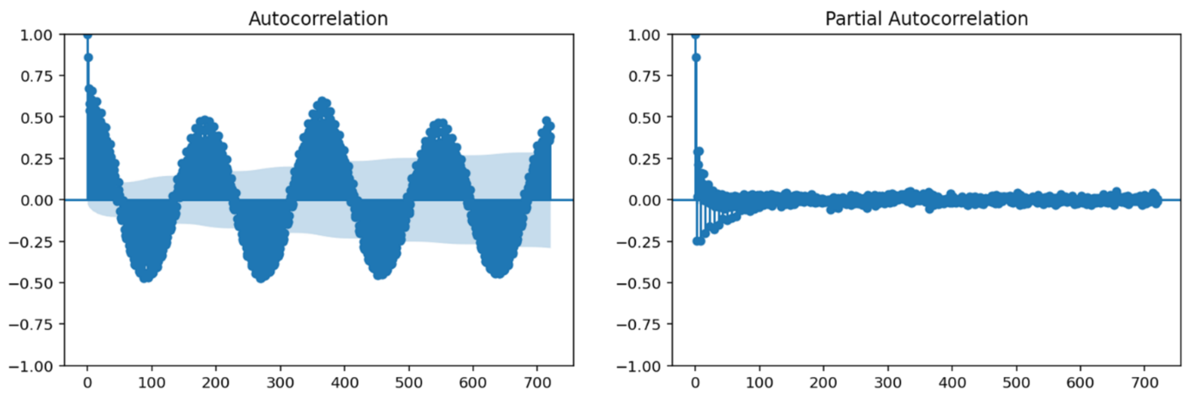

Figure 8. Plots of autocorrelation and partial autocorrelation of the data.

The graph presented above illustrates the correlation between the preceding data point and the current data point. The further the numeric value deviates from zero, the stronger the observed correlation.

In a general context, only data points falling above (for positive values) or below (for negative values) the shaded blue region, which represents the confidence interval, are considered statistically significant and deserving of attention. This pattern suggests a robust relationship between yesterday's values and today's values, indicative of a repeating seasonal trend. Notably, this cyclic pattern appears to recur every six months.

As previously discussed, this pattern aligns logically with weather conditions. During both winter and summer seasons, there is a heightened demand for electricity due to increased heating and cooling requirements. This phenomenon can be attributed to weather-related factors, as electricity consumption tends to be elevated during these climatic extremes.

The Partial Autocorrelation Function (PACF) plot reveals an intriguing pattern in the data. Specifically, it indicates that there is a notably stronger correlation among data points within the last 90 days. However, as you move further back in time, this correlation effect becomes less pronounced and apparent.

4.5 Modeling

We began by splitting the data into training and test sets. For our test data, we utilized the records from the year 2015 and onwards. The model was trained using the dataset spanning the years 2002 to 2015.

4.5.1 Baseline Models

Before we delved into machine learning, we established a 'baseline' using naive forecasting approaches. This entailed determining that if the ML models couldn't surpass these baseline projections, it would be more advantageous to stick with naive forecasting methods.

These forecasting approaches are referred to as 'naive' because of their simplicity, yet they often prove to be quite robust and effective. In fact, there are not many ML models that consistently outperform naive forecasts.

One such naive forecast we employed was the 'One-Year-Ago Persistent Forecast,' a common method for projecting multi-step forecasts, such as the subsequent 365 days. Essentially, this approach entails using the value from the same date one year ago to forecast the value for the current date. For instance, the value for January 01, 2018, would be forecasted using the value recorded on January 01, 2017.

We evaluated three types of persistence models:

1. **Previous Day Hour-by-Hour:** This approach involved forecasting each hour individually based on the corresponding hour from the previous day.
2. **Last 3-Day Average:** Here, we calculated the average value over the past three days to forecast the current value.

3. **Year Ago Day Hour-by-Hour:** Similar to the first method, this approach forecasted each hour separately, but it used data from the same date one year ago.

We utilized the root mean square error (RMSE) as the evaluation metric to directly compare the model performances with the energy values in the data. To generate forecasts, we employed the 'walk forward' method, which entails moving through the samples sequentially, generating a forecast at each step. After each forecast, the test value was appended to the training set and used for subsequent predictions.

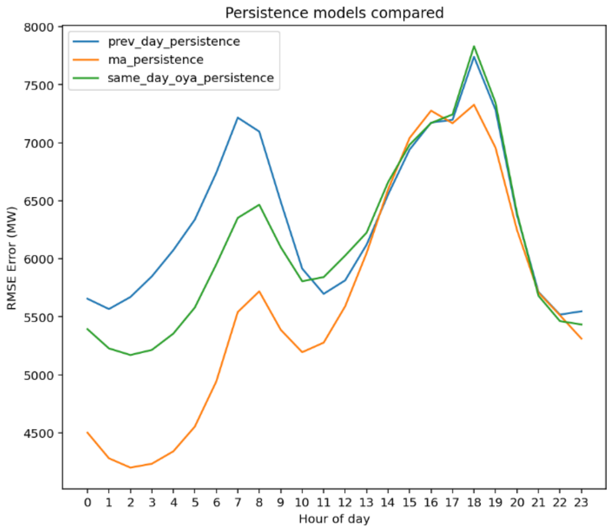

Figure 9. The errors values obtained for the 3 persistence models as a function of the hour of the day.

Among the three persistence models evaluated, it's evident that the moving average model stands out as the superior performer. However, it's important to note that there is still a significant amount of variation between the predictions, particularly during the hours from 0 to 12.

This observed variation in predictions during the early hours of the day suggests that the model may not capture the finer-grained patterns and nuances in the data during this time period as effectively as it does during other hours.

As we move forward with our analysis and modeling, the performance of the moving average model will serve as our baseline error benchmark. This means that we will use the moving average model's predictive performance as a reference point against which we will assess the performance of more advanced and sophisticated forecasting models. If these advanced models cannot consistently outperform the baseline established by the moving average model, it may indicate that the additional complexity of these models does not provide a significant improvement in predictive accuracy for the given dataset. Therefore, the moving average model's performance becomes a critical reference for gauging the effectiveness of our subsequent modeling efforts.

4.5.2 XGBoost Model

We used the XGBoost model for timeseries forecasting. XGBoost stands as a rapid gradient boosting implementation tailored to handling both classification and regression tasks with exceptional proficiency. This robust algorithm is characterized by its remarkable speed and efficiency, consistently demonstrating superior performance in an extensive array of predictive modeling endeavors. Furthermore, XGBoost extends its versatility to encompass the domain of time series forecasting, making it a valuable tool for a diverse range of data analysis and prediction tasks.

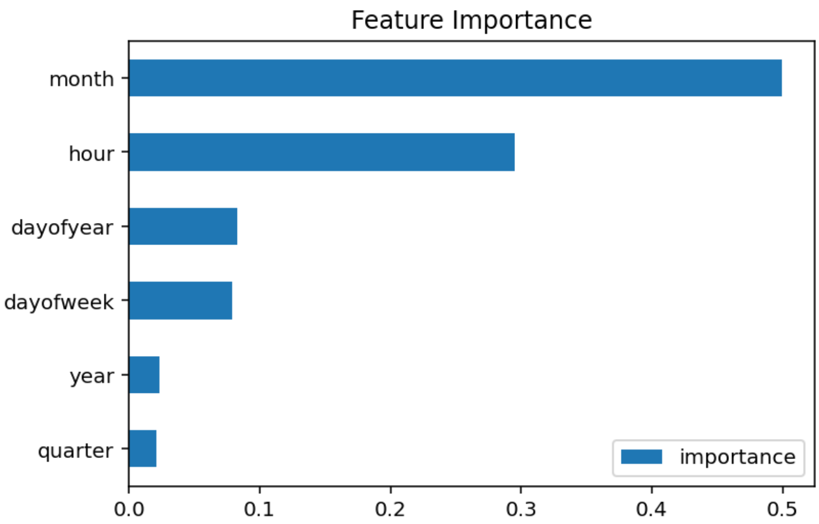

Figure 10. Feature importance on energy demand forecasting using XGBoost model.

The analysis of feature importance in energy demand forecasting has unveiled valuable insights into which factors significantly influence the predictive accuracy of the model. Specifically, it has indicated that the "month of the year" holds the highest degree of importance, followed closely by the "hour of the day."

1. **Month of the Year - Seasonal Influence:**

The prominence of the "month of the year" as a highly important feature underscores the significance of seasonality in energy demand forecasting. Different months exhibit distinct patterns in energy consumption due to various factors such as weather conditions, holidays, and cultural events. For example:

- **Weather Effects:** In regions with diverse climates, the energy demand typically surges during extreme weather conditions. This includes increased heating in winter and cooling in summer, significantly impacting energy consumption.
- **Holiday Season:** The holiday season often leads to variations in energy usage. Special occasions like Christmas and New Year's Eve can result in heightened electricity demand for lighting, decorations, and appliances.

- **Economic Activities:** Economic factors, such as industrial production and business operations, can also vary by month, contributing to fluctuations in energy demand.

The ability to capture these seasonality patterns is critical for accurate forecasting, as it allows models to adapt to the changing energy consumption dynamics throughout the year.

2. **Hour of the Day - Diurnal Trends:**

The "hour of the day" is identified as the second most important feature, highlighting the significance of diurnal (daily) trends in energy demand. Energy consumption typically follows distinct patterns throughout a 24-hour cycle:

- **Peak Hours:** There are specific hours during the day when energy demand reaches its peak. This often coincides with the typical working hours of businesses and institutions, leading to increased electricity usage for lighting, heating, cooling, and industrial processes.
- **Off-Peak Hours:** Conversely, during the night-time hours when many people are at home and businesses are closed, energy consumption tends to decrease, marking the off-peak hours.

Understanding and modeling these daily fluctuations in energy demand is vital for ensuring efficient energy resource allocation, grid management, and infrastructure planning.

In summary, the identification of the "month of the year" and the "hour of the day" as highly important features in energy demand forecasting underscores the critical role of seasonality and diurnal trends in shaping energy consumption patterns. Incorporating these features into forecasting models allows for more accurate predictions and better management of energy resources.

The plots below show the predictions from the XGBoost model for a week of data.

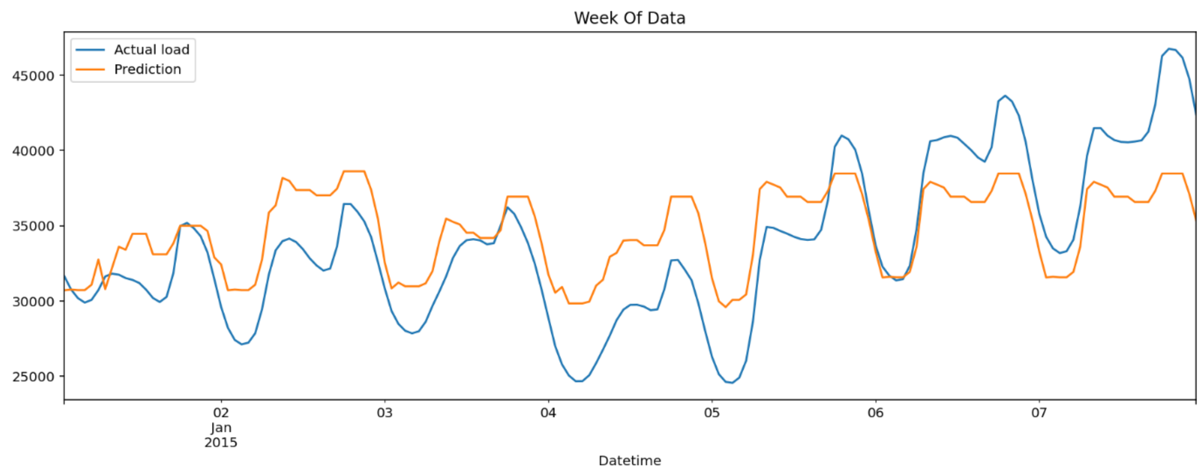

Figure 11. Predictions for a week of data obtained using the XGBoost model compared to the actual load values.

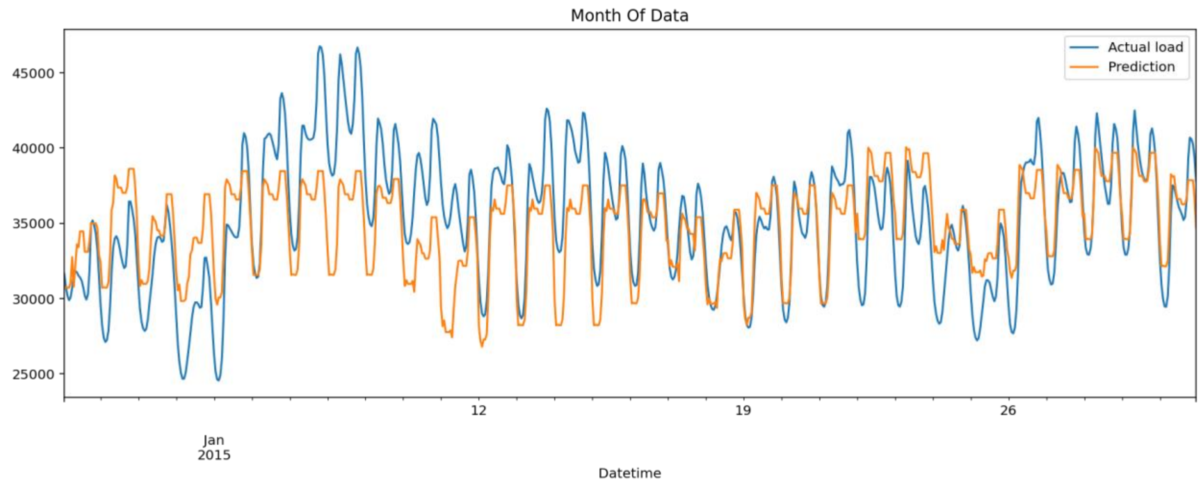

Figure 12. Predictions for a month of data obtained using the XGBoost model compared to the actual load values.

The XGBoost model demonstrates a commendable ability to provide reasonably accurate predictions of the actual energy load. However, the most noticeable disparities between the actual and predicted load values tend to occur during peak demand periods.

During these peak periods, which are typically characterized by exceptionally high energy consumption, the model encounters challenges in precisely forecasting the exact load values. Factors contributing to these discrepancies during peak periods may include sudden spikes in demand due to unforeseen events, variations in weather conditions, or other external influences that are difficult to capture accurately.

While the model excels in capturing general load trends and patterns, it may struggle to replicate the precise magnitude of load spikes during these peak moments. These disparities during peak periods highlight the complexities and intricacies of energy load forecasting, particularly when dealing with extreme and less predictable demand scenarios.

The error values generated by the XGBoost model are consistently lower when compared to those of the baseline model. This implies that the XGBoost model is delivering superior predictive accuracy in forecasting energy demand compared to the simpler baseline model. The reduction in error values signifies that the XGBoost model is better at capturing the underlying patterns and nuances in the data, resulting in more precise load predictions. This improvement in accuracy is a positive indication of the model's effectiveness in enhancing energy demand forecasting compared to the basic baseline approach.

6. Conclusion

The XGBoost model is able to provide reasonably accurate predictions. However, a few steps can be taken to improve the predictions. These are:

1. **Incorporate Weather Data:** Weather conditions have a significant impact on energy consumption. Include temperature, humidity, wind speed, and other relevant weather variables in the forecasting models to account for weather-related fluctuations.
2. **Advanced Machine Learning Models:** Explore advanced machine learning techniques such as Random Forests, Support Vector Machines, Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and Prophet. These models can capture complex patterns in time series data.
3. **Feature Engineering:** Carefully select and engineer features that have a direct or indirect influence on energy demand. Consider lag variables, holidays, special events, and economic indicators that may affect consumption.
4. **Cross-Validation:** Implement cross-validation techniques to assess model performance. This helps prevent overfitting and provides a more accurate estimate of a model's predictive capabilities.
5. **Ensemble Models:** Combine predictions from multiple models (ensemble learning) to reduce errors and increase robustness. Methods like stacking, bagging, and boosting can be effective.
6. **Hyperparameter Tuning:** Fine-tune model hyperparameters systematically to optimize model performance. Grid search or random search can help identify the best set of hyperparameters.

By combining these strategies, we can significantly enhance the accuracy and reliability of energy demand forecasting, enabling better resource allocation, grid stability, and energy management.