Energy Consumption Prediction using ML

Agenda

1. Introduction
2. The data
3. Method 1: Linear Regression
4. Method 2: Random Forest Regressor
5. Method 3: Neural Network Regression
6. Results
7. Conclusion

Introduction

This project is about predicting energy use.

We use historical data about energy consumption and try three different methods.

We employ data analysis, feature engineering, and multiple machine learning models, including Linear Regression, Random Forest, and Neural Networks, to evaluate their performance.

The ultimate goal is to determine the best model for accurate energy forecasting.

Data Exploration and Visualization

Data: Hourly energy consumption (PJME_MW) over time from Kaggle, .

Visualization: Explored trends and seasonality in energy usage with time-series plots.

Observations: Clear patterns influenced by time-related factors such as hours, days, and season

To begin our analysis, we import essential Python libraries for data manipulation and visualization. This includes Pandas for data handling, NumPy for numerical operations, Seaborn for enhanced visualizations, and Matplotlib for customizable plotting. Additionally, we incorporate the XGBoost model for predictive analytics and the mean squared error metric for evaluating model performance.

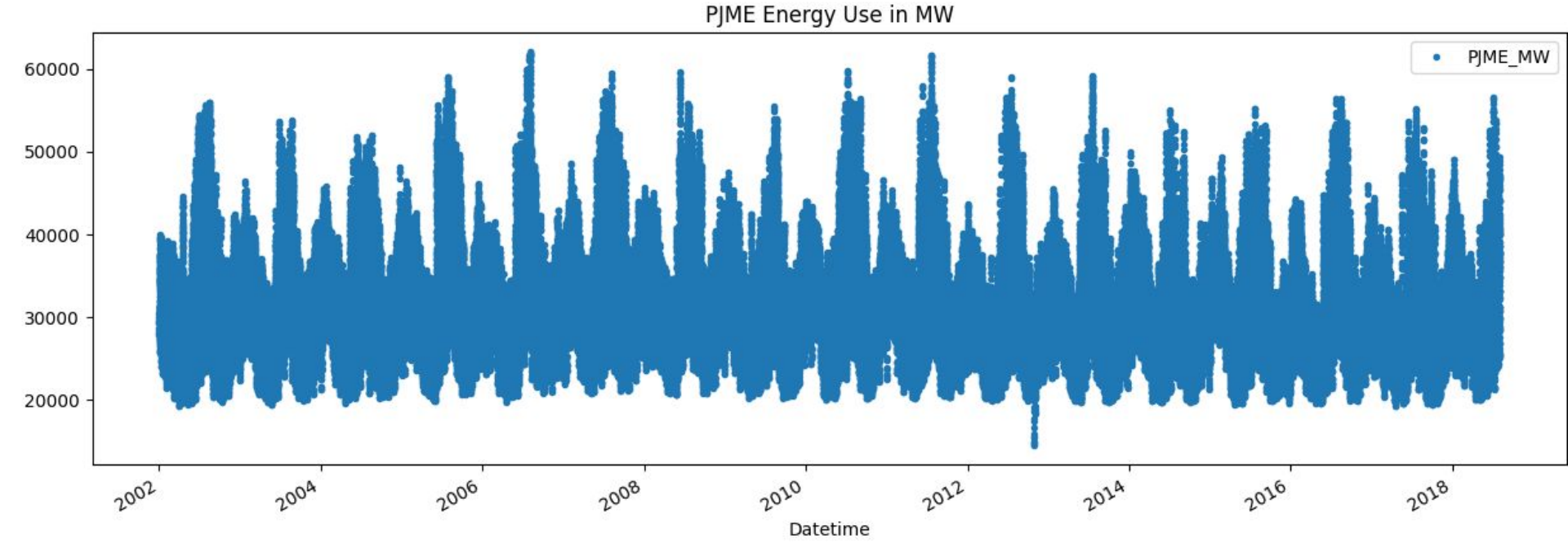
(The data consists of sequential observations recorded over time.)

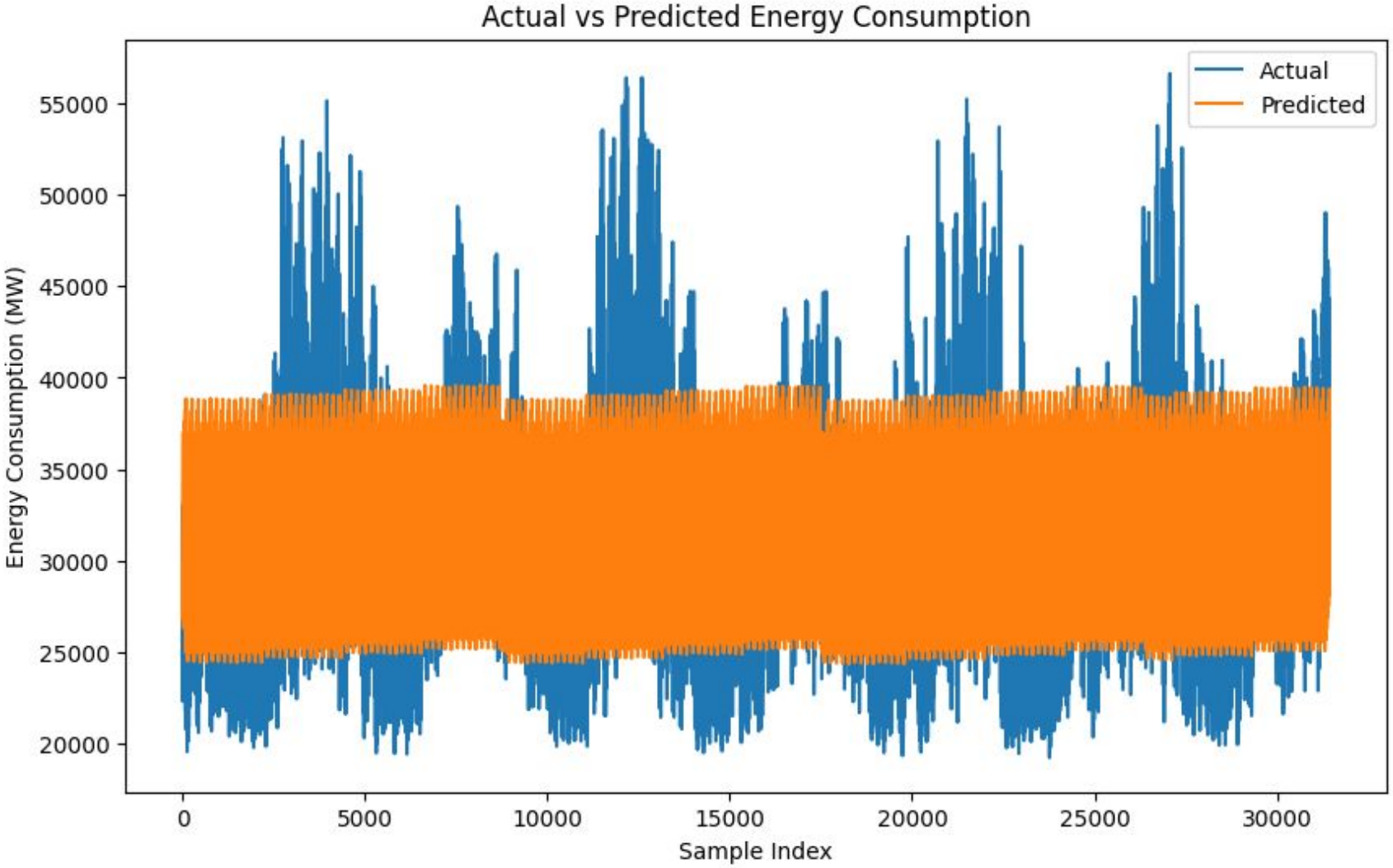
We start by loading the dataset using Pandas. The dataset contains hourly energy consumption records with a `Datetime` column representing the timestamp.

Next, we set the `Datetime` column as the index and convert it to a `datetime` type for easier manipulation.

Visualizing the data helps us understand its trends and seasonality. We can plot the entire time series to observe the patterns over time.

The plot shows how energy consumption fluctuates over time, with noticeable seasonal patterns corresponding to different times of the year.


PJME Energy Use in MW
PJME_MW
60000
50000
40000
30000
20000
2002
2004
2006
2008
2010
2012
2014
2016
2018
Datetime


Actual vs Predicted Energy Consumption
Actual
Predicted
Energy Consumption (MW)
55000
50000
45000
40000
35000
30000
25000
20000
0
5000
10000
15000
20000
25000
30000
Sample Index

The plot shows the comparison between actual and predicted energy consumption for a time series forecasting model using linear regression.

Key Observations:

1. **Performance Gap**: The predicted values appear to be much smoother and do not capture the fluctuations and peaks seen in the actual energy consumption data, indicating that the model struggles to predict the variability present in the time series.

The Root Mean Square Error (RMSE) calculated provides a quantitative measure of the model's prediction accuracy. While the prediction provides a rough approximation, it highlights the limitations of linear regression in capturing the complex patterns in time series data.

RMSE: 5683

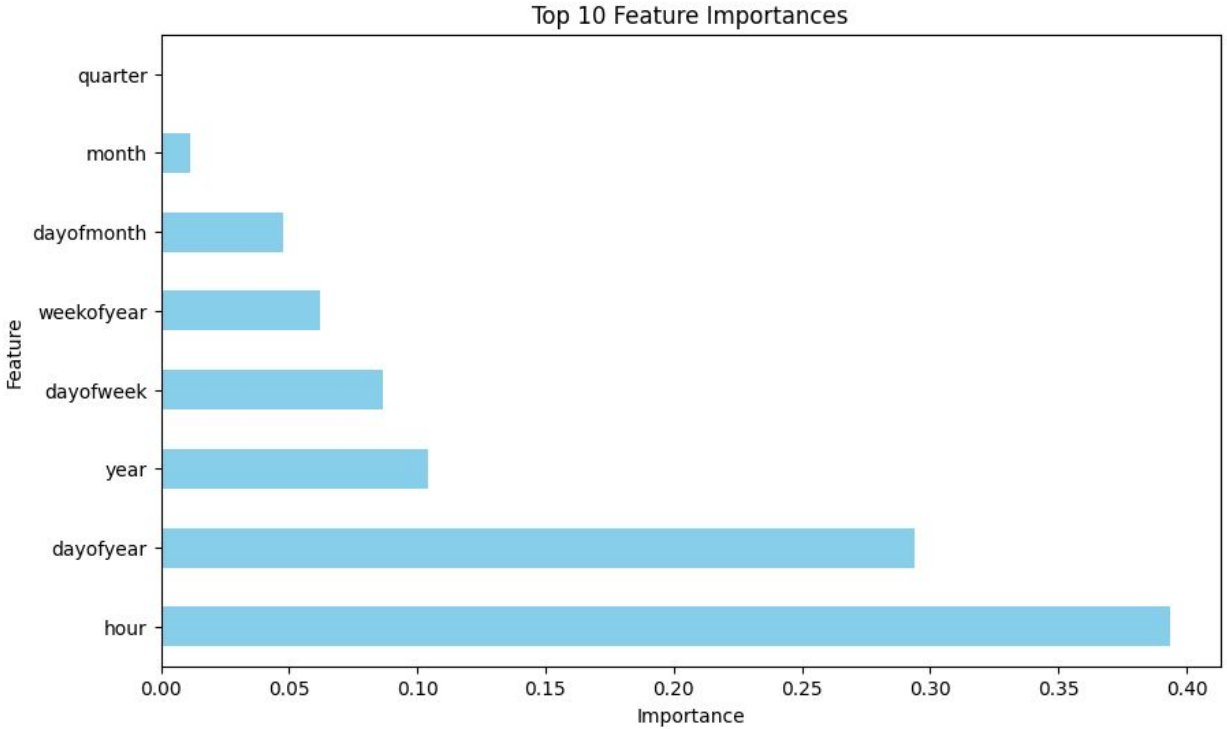
Comparison with Linear Regression

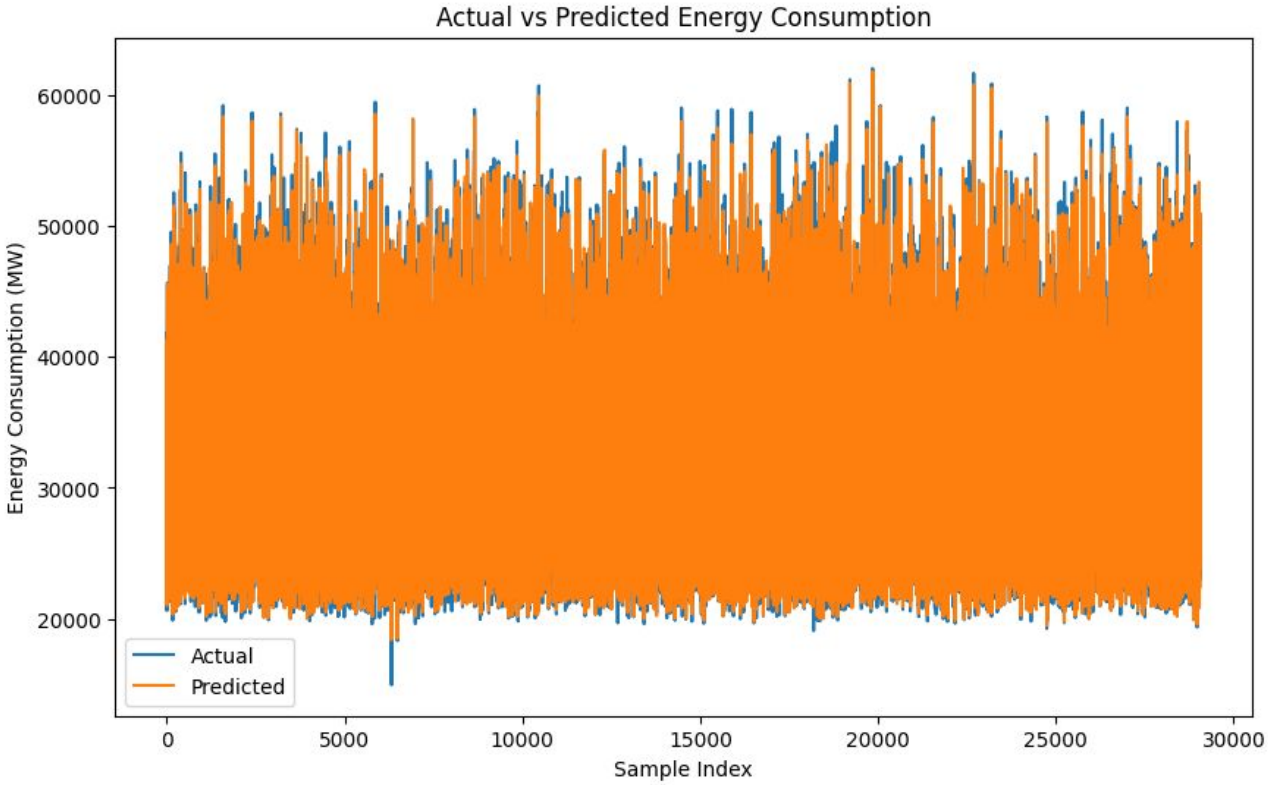
The stark difference in RMSE scores illustrates the benefits of using ensemble techniques, as random forests systematically outperform linear models in this application.

Feature Importance Analysis

Random forests provide insights into the importance of various features, guiding understanding of which factors most significantly influence energy consumption.

Method 2: Random Forest Regressor




Actual vs Predicted Energy Consumption
Energy Consumption (MW)
60000
50000
40000
30000
20000
0
5000
10000
15000
20000
25000
30000
Sample Index
Actual
Predicted

Key Insights:

1. **Model Performance:**
   - Root Mean Squared Error (RMSE): `1381.90`
   - The Random Forest model performed significantly better than linear regression in capturing energy consumption trends.
2. **Prediction Accuracy:**
   - The **Predicted** values closely track the **Actual** energy consumption values (blue line).
   - There are still some mismatches during peak fluctuations, but the Random Forest model's predictions align more consistently.
3. **Feature Importance:**
   - Top contributing features include:
     - **Hour of the Day** (most significant)
     - **Day of the Year**
     - **Year**
   - Temporal variables play a critical role in determining energy consumption patterns.

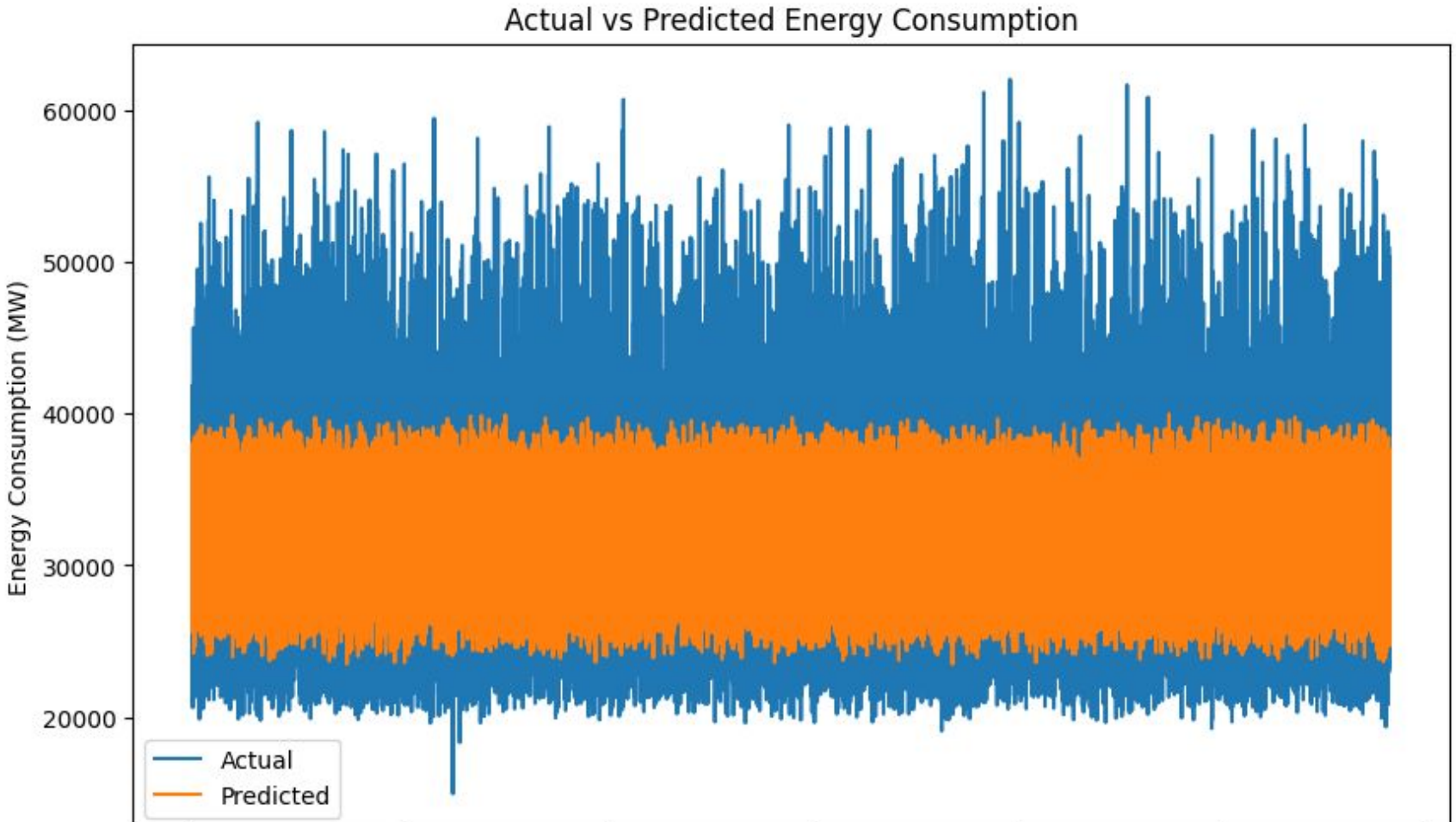
Method 3: Neural Network Regression

Key Insights:

Model Training Progress:

- **Epochs:** 14 (out of 50)
- **Loss (Training):** Reduced from ~407,729,152 in the first epoch to ~45,380,592 by the 14th epoch.
- **Validation Loss:** Stabilized around ~31,000,000 after several epochs, indicating convergence but potential overfitting.

2. **Performance Metrics:**
   - **Final RMSE on Test Set:** 5,504.23 MW
   - The neural network achieved better granularity in predictions compared to simpler models (e.g., linear regression) but still exhibits room for improvement.
3. **Observations:**
   - While the RMSE value is competitive, the validation loss plateau suggests a need for fine-tuning:
     - **Potential Enhancements:**
       - Introduce regularization techniques (e.g., dropout or L2 regularization).
       - Experiment with additional epochs, but monitor for overfitting.
       - Incorporate external features (e.g., weather or seasonal effects).


Actual vs Predicted Energy Consumption
Energy Consumption (MW)
60000
50000
40000
30000
20000
Actual
Predicted

Model Comparison (RMSE):

- **Linear Regression:** 5,683.95
- **Random Forest: 1,381.90** (best performer).
- **Neural Network:** 5,504.23 (potential for improvement with hyperparameter tuning).

Conclusion:

- Random Forest is the most accurate model.
- Neural Network can improve with fine-tuning and more complex architectures.