PowerPulse: Household Energy Usage Forecast

Problem Statement:

- In the modern world, energy management is a critical issue for both households and energy providers. Predicting energy consumption accurately enables better planning, cost reduction, and optimization of resources. The goal of this project is to develop a machine learning model that can predict household energy consumption based on historical data. Using this model, consumers can gain insights into their usage patterns, while energy providers can forecast demand more effectively.

- By the end of this project, learners should provide actionable insights into energy usage trends and deliver a predictive model that can help optimize energy consumption for households or serve as a baseline for further research into energy management systems.

Business Use Cases:

1. Energy Management for Households:

Monitor energy usage, reduce bills, and promote energy-efficient habits.

2. Demand Forecasting for Energy Providers:

Predict demand for better load management and pricing strategies.

3. Anomaly Detection:

Identify irregular patterns indicating faults or unauthorized usage.

4. Smart Grid Integration:

Enable predictive analytics for real-time energy optimization.

5. Environmental Impact:

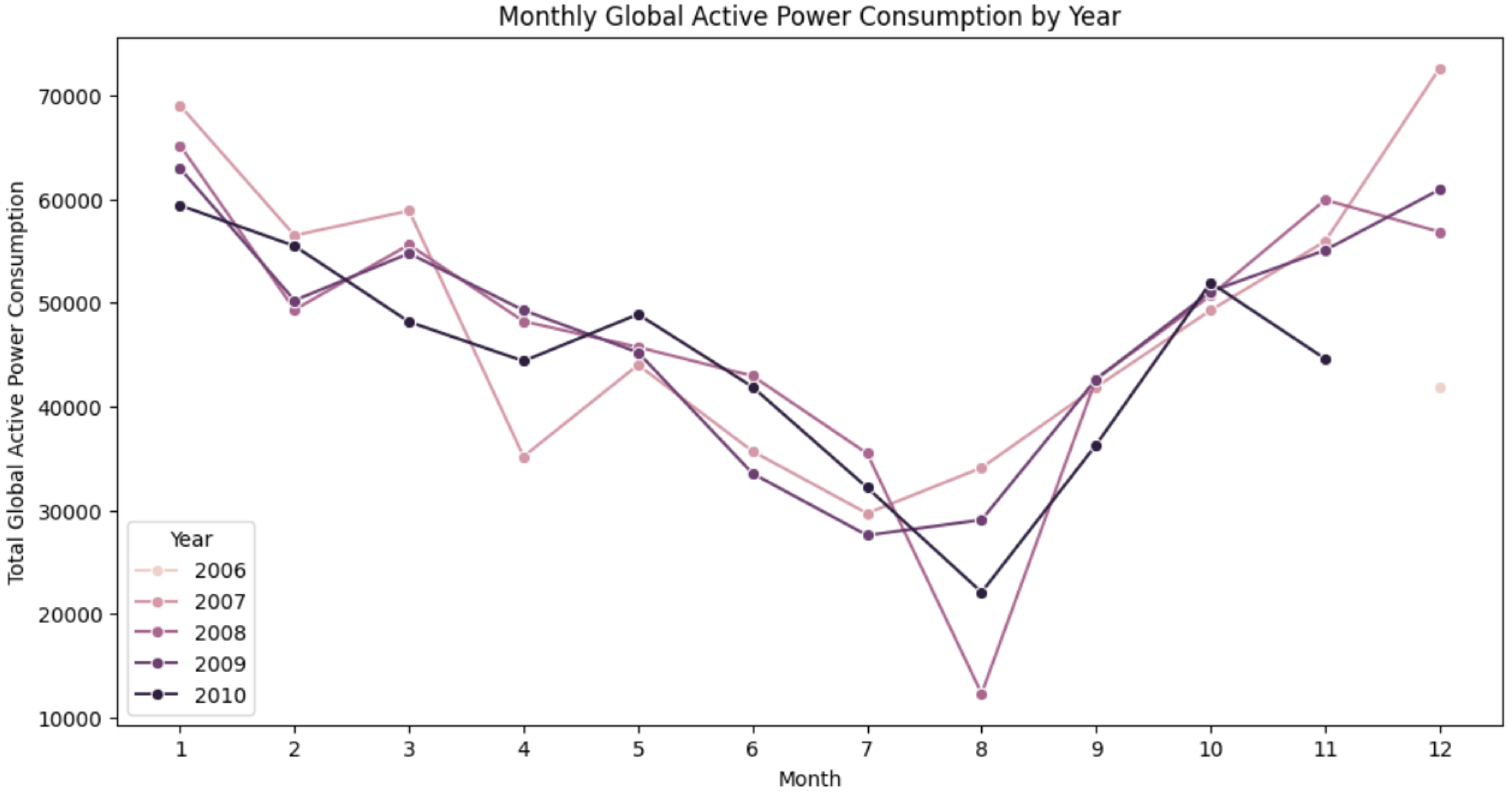
Reduce carbon footprints and support conservation initiatives.

Technical Tags:

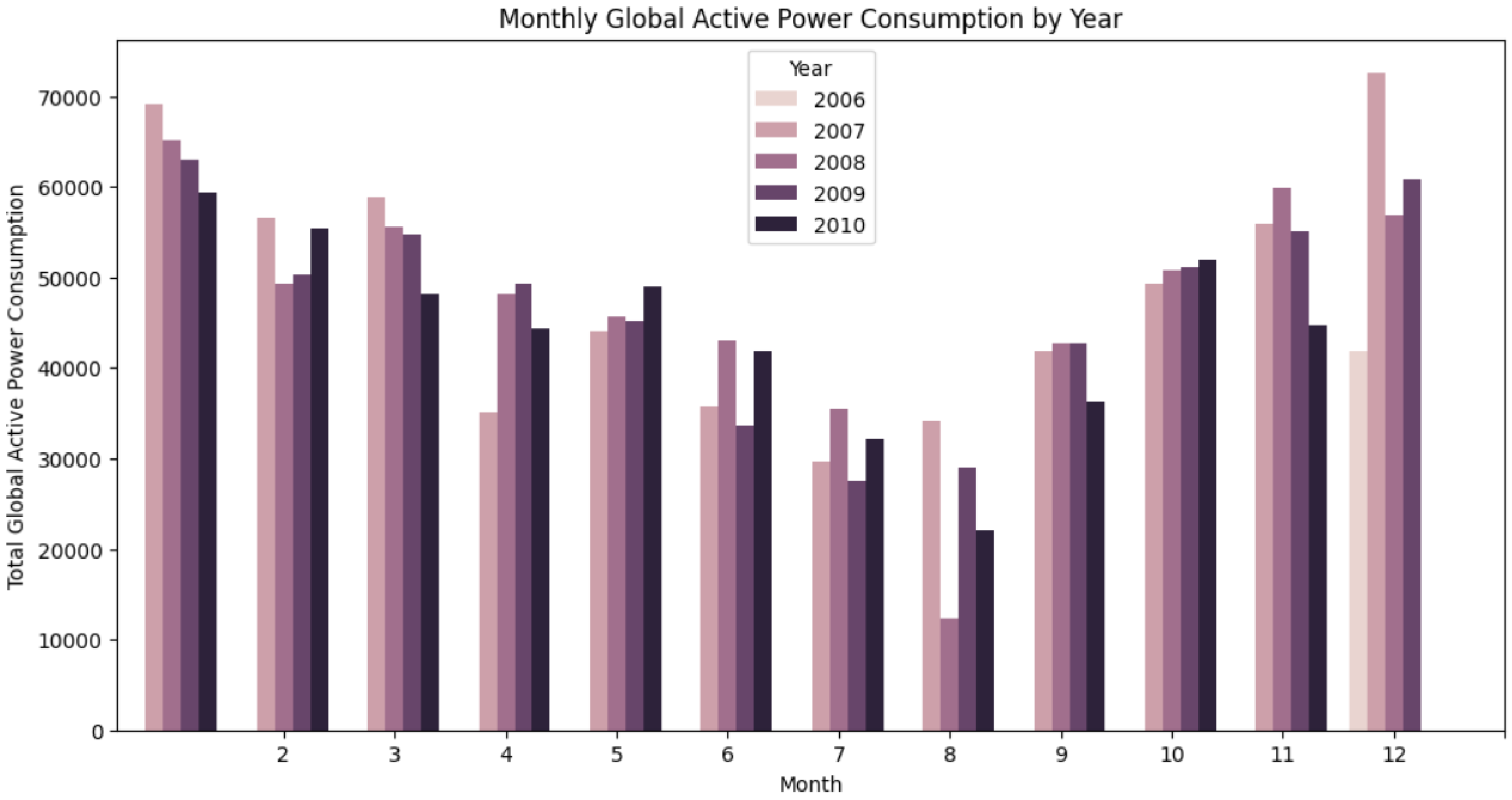
* Data Preprocessing
* Regression Modeling
* Feature Engineering
* Hyperparameter Tuning
* Visualization
* Python
* Scikit-learn
* Pandas
* Matplotlib/Seaborn

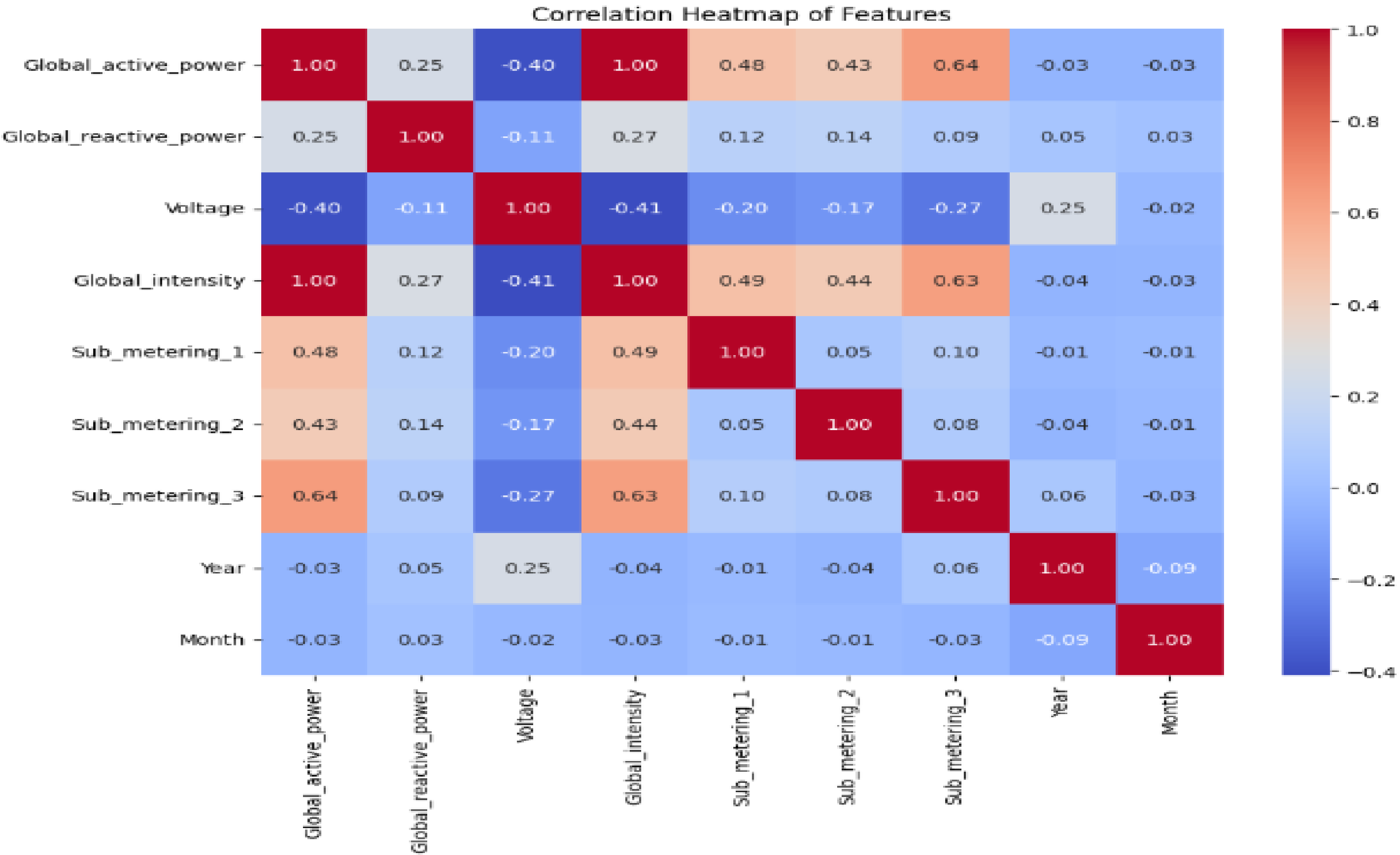
Dataset:

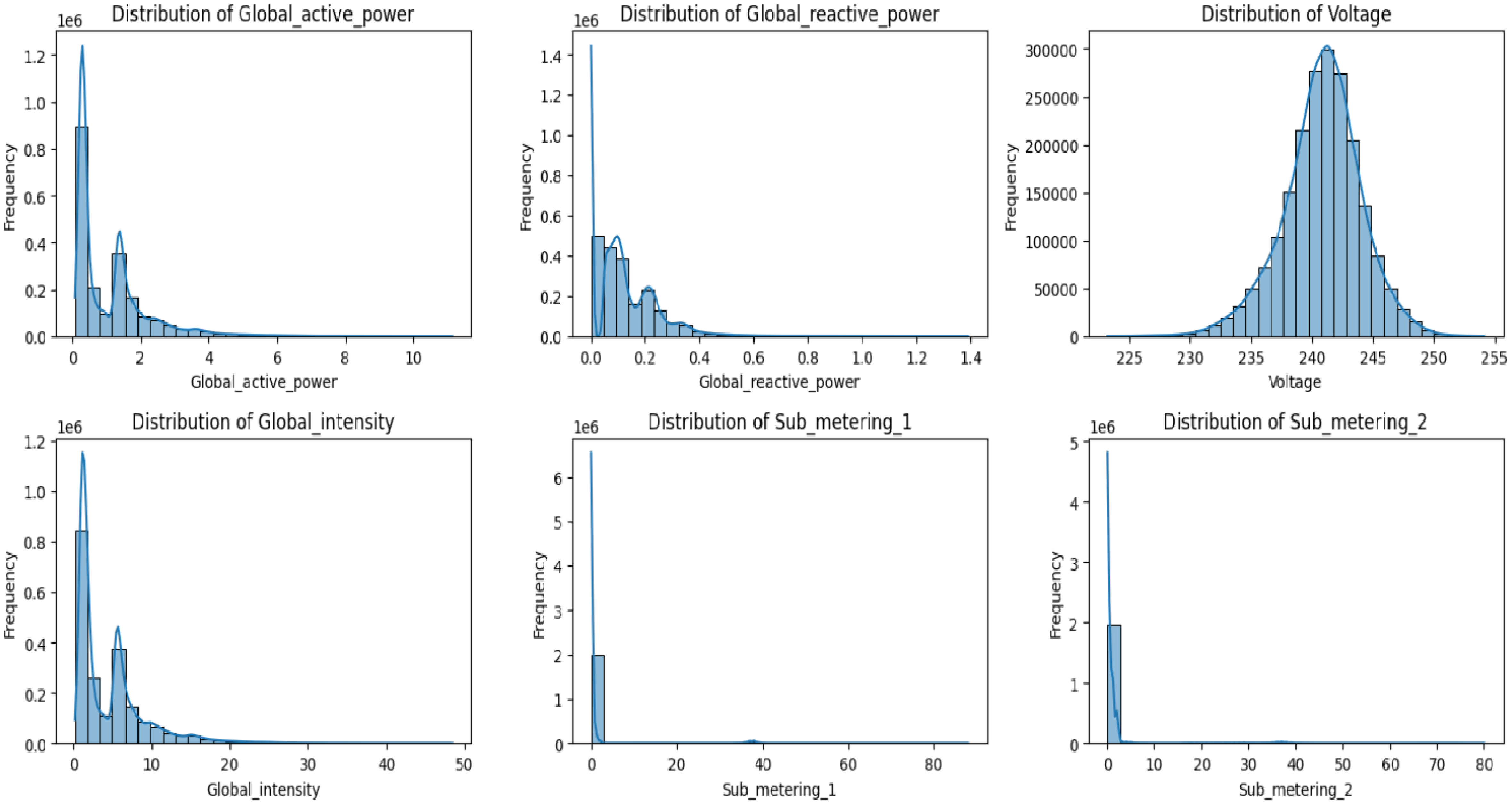
```
 #   Column                 Dtype
---  ------                 -----
 0   Date                   object
 1   Time                   object
 2   Global_active_power    float64
 3   Global_reactive_power  float64
 4   Voltage                float64
 5   Global_intensity       float64
 6   Sub_metering_1         float64
 7   Sub_metering_2         float64
 8   Sub_metering_3         float64
```



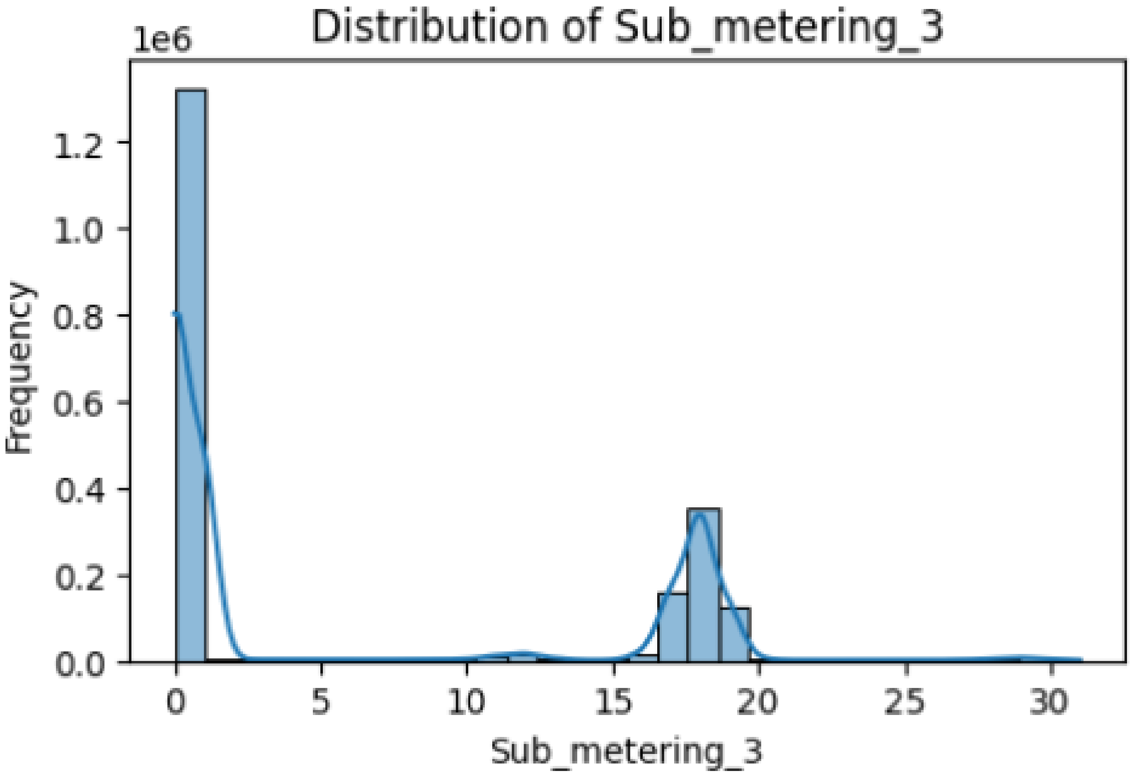
Monthly Global Active Power Consumption by Year
Total Global Active Power Consumption
Month
Year
2006
2007
2008
2009
2010
10000
20000
30000
40000
50000
60000
70000
1
2
3
4
5
6
7
8
9
10
11
12


Monthly Global Active Power Consumption by Year
Year
2006
2007
2008
2009
2010
Total Global Active Power Consumption
0
10000
20000
30000
40000
50000
60000
70000
2
3
4
5
6
7
8
9
10
11
12
Month


Correlation Heatmap of Features
Global_active_power 1.00 0.25 -0.40 1.00 0.48 0.43 0.64 -0.03 -0.03
Global_reactive_power 0.25 1.00 -0.11 0.27 0.12 0.14 0.09 0.05 0.03
Voltage -0.40 -0.11 1.00 -0.41 -0.20 -0.17 -0.27 0.25 -0.02
Global_intensity 1.00 0.27 -0.41 1.00 0.49 0.44 0.63 -0.04 -0.03
Sub_metering_1 0.48 0.12 -0.20 0.49 1.00 0.05 0.10 -0.01 -0.01
Sub_metering_2 0.43 0.14 -0.17 0.44 0.05 1.00 0.08 -0.04 -0.01
Sub_metering_3 0.64 0.09 -0.27 0.63 0.10 0.08 1.00 0.06 -0.03
Year -0.03 0.05 0.25 -0.04 -0.01 -0.04 0.06 1.00 -0.09
Month -0.03 0.03 -0.02 -0.03 -0.01 -0.01 -0.03 -0.09 1.00
Global_active_power
Global_reactive_power
Voltage
Global_intensity
Sub_metering_1
Sub_metering_2
Sub_metering_3
Year
Month
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4



Distribution of Global_active_power
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Frequency
0
2
4
6
8
10
Global_active_power
Distribution of Global_reactive_power
1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Global_reactive_power
Distribution of Voltage
300000
250000
200000
150000
100000
50000
0
Frequency
225
230
235
240
245
250
255
Voltage
Distribution of Global_intensity
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Frequency
0
10
20
30
40
50
Global_intensity
Distribution of Sub_metering_1
1e6
6
5
4
3
2
1
0
Frequency
0
20
40
60
80
Sub_metering_1
Distribution of Sub_metering_2
1e6
5
4
3
2
1
0
Frequency
0
10
20
30
40
50
60
70
80
Sub_metering_2


Distribution of Sub_metering_3
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Frequency
0
5
10
15
20
25
30
Sub_metering_3

Introduction

This report presents a comprehensive evaluation of multiple regression models applied to a dataset.
The goal is to identify the best-performing model based on various regression metrics.
The models evaluated include:

- Linear Regression (3 variations)
- Random Forest Regression (3 variations)
- Gradient Boosting Regression (3 variations)
- Neural Network Regression (3 variations)

Evaluation Metrics

Each model is assessed using the following metrics:

- **MAE (Mean Absolute Error):** Average of absolute differences between predicted and actual values.
- **MSE (Mean Squared Error):** Average of squared differences between predicted and actual values.
- **RMSE (Root Mean Squared Error):** Square root of MSE, representing error magnitude.
- **R^2 Score (Coefficient of Determination):** Indicates how well the model explains the variance in the target variable (1.0 is perfect).

Model	MAE	MSE	RMSE	R² Score
Linear Regression MODEL 1	0.0274	0.00179	0.0423	0.99840
Linear Regression MODEL 2	0.1078	0.02875	0.1695	0.99855
Linear Regression MODEL 3	0.1122	0.03055	0.1748	0.99846
Random Forest MODEL 1	**0.0156**	**0.00096**	**0.0309**	**0.99915**
Random Forest MODEL 2	0.0537	0.01557	0.1248	0.99921
Random Forest MODEL 3	0.0508	0.01450	0.1204	**0.99927**
Gradient Boosting MODEL 1	0.0209	0.00119	0.0345	0.99894
Gradient Boosting MODEL 2	0.0875	0.02060	0.1435	0.99896
Gradient Boosting MODEL 3	0.0874	0.02042	0.1429	0.99897
Neural Network MODEL 1	0.0602	0.00508	0.0713	0.99547
Neural Network MODEL 2	0.1514	0.03493	0.1869	0.99824
	0.0860	0.02033	0.1426	0.99897

Model 1:

X = df.drop(columns['Global_active_power','Sub_metering_3'])

y = df['Global_active_power']

Model 2:

X=df.drop(columns['Global_intensity','Sub_metering_3'])

y= df['Global_intensity']

Model 3:

X= df.drop(columns=['Global_intensity'])

y= df['Global_intensity']

Analysis

Linear Regression Models

- **MODEL 1** performs best among linear models with low MAE and RMSE.
- However, it is outperformed by tree-based models in all metrics.

Random Forest Models

- **MODEL 1** achieves the **lowest MAE (0.0156)** and **lowest RMSE (0.0309)**.
- It also has a **very high R^2 score (0.99915)**, making it the most reliable overall.
- **MODEL 3** has the highest R^2 (0.99927), but slightly higher error rates than MODEL 1.

Gradient Boosting Models

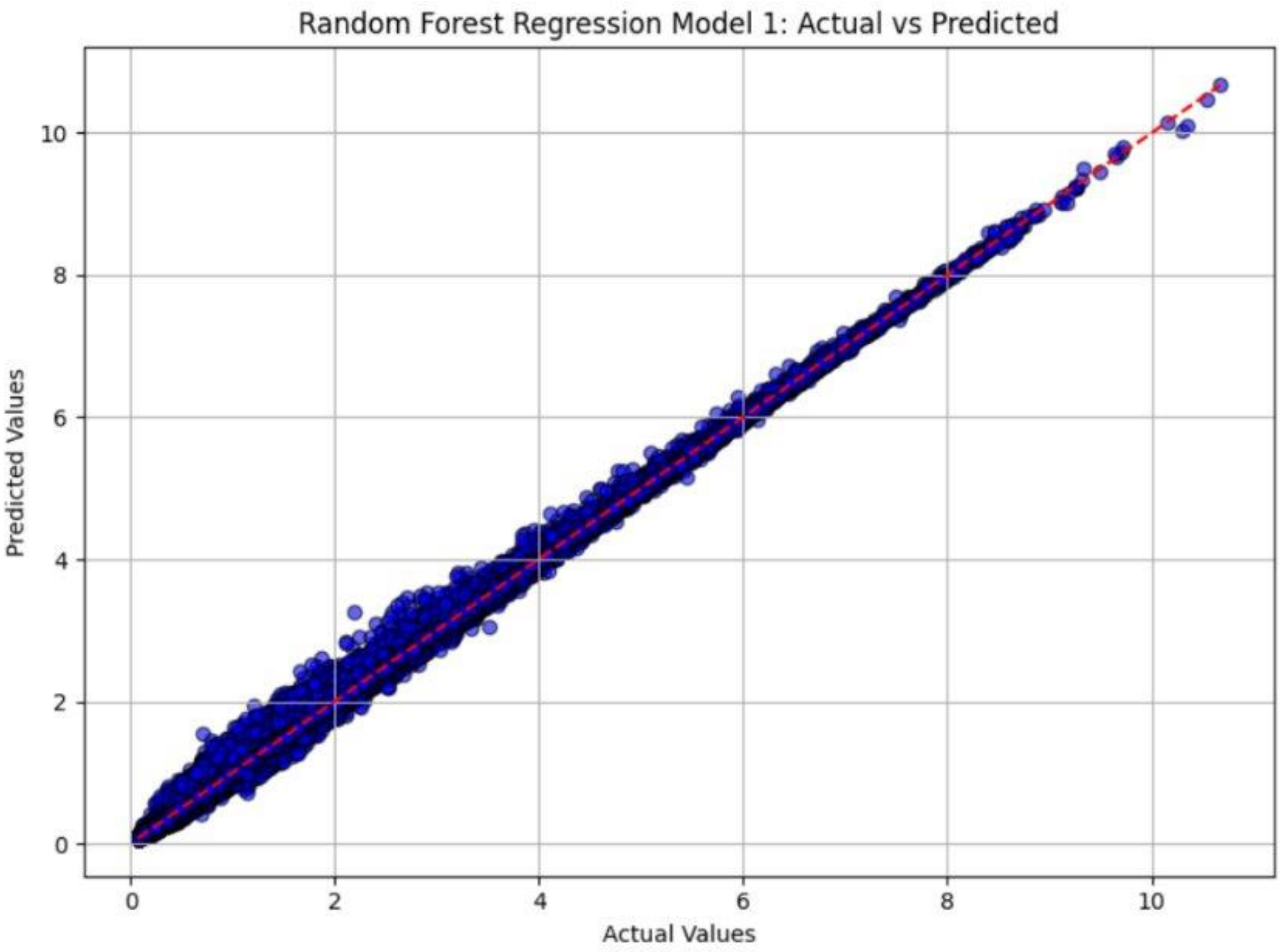
- These models also perform well, with MODEL 1 achieving competitive results (MAE: 0.0209).
- However, they are slightly less accurate than Random Forests.

Neural Network Models

- These show more variability.
- MODEL 3 performs the best among them (MAE: 0.0860), but it does not outperform the best tree-based models.

Final Thoughts

Tree-based models like **Random Forests** and **Gradient Boosting** consistently outperform linear and neural network models in this analysis. Particularly, **Random Forest MODEL 1** stands out due to its superior accuracy and generalization.



Random Forest Regression Model 1: Actual vs Predicted
Predicted Values
Actual Values
0
2
4
6
8
10

Conclusion

After evaluating multiple regression models using standard performance metrics (MAE, MSE, RMSE, and R^2 Score), it is evident that **Random Forest Regression MODEL 1** delivers the most accurate and reliable results. It achieves the **lowest error rates** and a **high R^2 Score of 0.99915**, indicating it explains nearly all the variance in the target variable. While other models such as Gradient Boosting and Neural Networks also performed well, they were slightly less consistent in terms of error minimization.

Therefore, **Random Forest Regression MODEL 1 is recommended as the optimal model** for this task due to its exceptional balance between prediction accuracy and model robustness. It is well-suited for real-world deployment where both precision and reliability are critical.