Energy Consumptions Predictions
Capstone 2 Project Report
Angeles Olvera 5/15/2025

Table of Contents

Problem Identification

Problem Statement

The client, GreenWatts Energy, is a leading energy solutions provider aiming to optimize operational efficiency and meet growing energy demands. To achieve this, they want to predict energy consumption trends accurately and in advance. This will enable them to allocate resources more effectively, reduce wastage, and maintain an uninterrupted energy supply. A reliable prediction model must be developed and implemented by the end of the quarter to ensure readiness for the next operational cycle.

Context

GreenWatts Energy has operated for a decade, specializing in renewable energy solutions like solar, wind, and hydroelectric power. With a shift toward sustainable energy, the company has witnessed a steady rise in demand. However, predicting energy usage with precision is essential to balance production and distribution efficiently. Without this capability, they risk resource overproduction or underproduction, leading to inefficiencies and customer dissatisfaction. By harnessing data-driven insights, GreenWatts sees an opportunity to remain at the forefront of sustainable energy solutions.

Criteria for Success

The success of this analysis will be determined by the ability to accurately predict energy consumption trends. Specifically the predictive model should classify and forecast energy consumption with high precision, enabling GreenWatts Energy to optimize resource allocation.

Scope of Solution Space

The goal of this project is to analyze and predict energy consumption trends based on significant variables to optimize GreenWatts Energy's operational efficiency. The target variable for this predictive model is EnergyConsumption, and the predictors include

Square Footage, Month, Hour, Day of Week, Temperature, Humidity, and Occupancy. Different machine learning algorithms, such as Random Forest and Linear Regression, will be used to determine which predictors have the greatest impact on energy consumption.

Constraints Within Solution Space

Some things to consider during investigation are unpredictable influences like sudden changes in weather or energy consumption patterns may affect the reliability of the predictions.Ensuring that the insights generated from the analysis are practical and can be integrated into GreenWatts Energy's operations within the given timeframe. The dataset may include columns with inaccurate, incomplete, or missing entries that could affect the model's performance. Cleaning and preprocessing the data will be crucial.

Stakeholders

Energy Operations Manager - Daniel Clark
Database Manager - Amanda Johnson
Amanda Johnson will provide the dataset for the analysis

Key Data Source

A single CSV with 12 columns and 500 rows where the main columns of interest are Month, Hour, Day of Week, Temperature, Humidity, Square Footage, Occupancy, and EnergyConsumption.

Data Wrangling

Data Cleaning

In the Data Wrangling phase, I first imported essential libraries like NumPy and Pandas, along with tools for preprocessing, modeling, evaluation, and optimization. After examining the dataset, I found no missing values, and all column data types were correctly formatted, indicating a clean dataset.

Statistical Summary

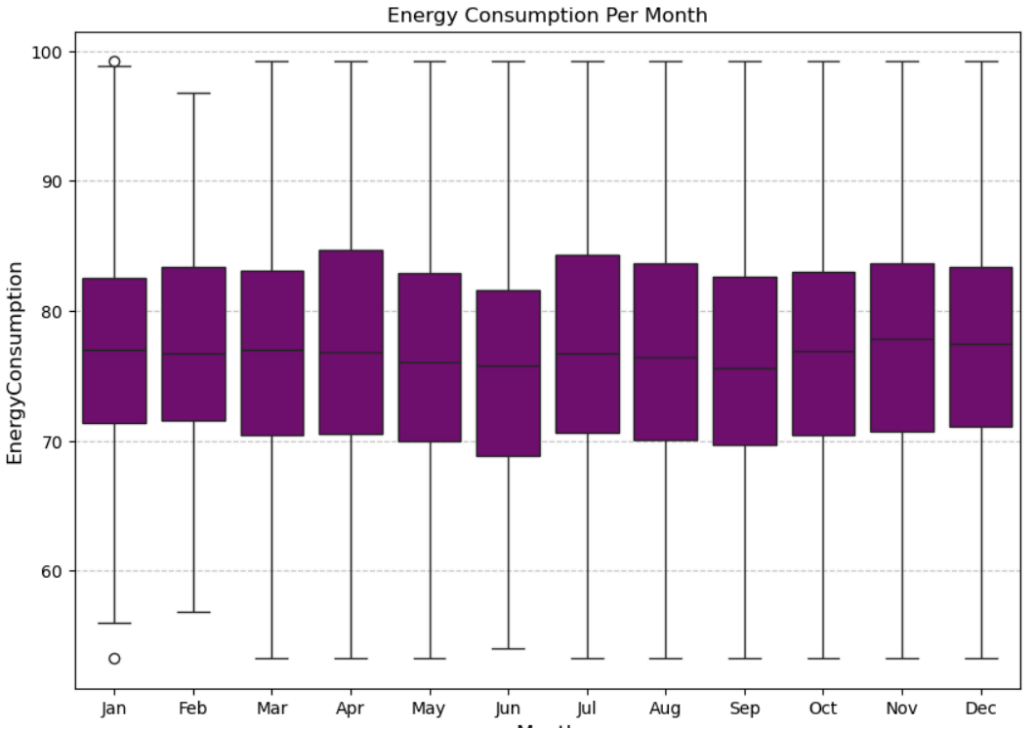
Additionally, I analyzed the float-type columns and performed a statistical summary to assess the data distribution. This summary provided insight into the spread of values, which is crucial for predictive accuracy. Ideally, for energy consumption prediction, the

standard deviation should be close to the mean across all columns. While most columns met this criterion, the square footage variable exhibited a significantly higher standard deviation compared to its mean, indicating substantial variation in home sizes which is normal.

	Temperature	Humidity	SquareFootage	RenewableEnergy	EnergyConsumption
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000
mean	24.946823	45.478146	1507.403201	15.201195	76.794919
std	3.041678	8.972690	293.147209	9.157038	9.231573
min	20.007565	30.015975	1000.512661	0.006642	53.263278
25%	22.453790	38.111104	1253.906598	7.477928	70.419588
50%	24.831846	45.793124	1513.581105	15.343830	76.696267
75%	27.427281	52.696651	1754.846457	22.889997	83.246274
max	29.998671	59.969085	1999.982252	29.965327	99.201120

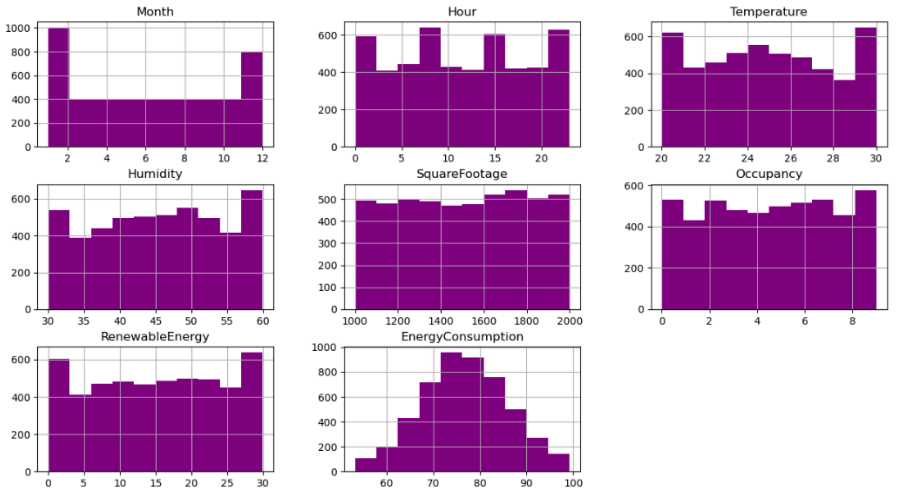
Outliers

When analyzing energy consumption predictions, the timeframe plays a crucial role. To explore monthly variations, I generated a boxplot spanning January to December, aiming to identify outliers that could indicate months with unusually high or low usage. The boxplots appeared fairly consistent across the months, with the only noticeable difference occurring in January. This month exhibited both low outliers, indicating reduced energy consumption, and exceptionally high consumption levels. This could mean that GreenWatts Energy, company needs to be ready for the month of January as it has both low and high extremes.



Distributions

I analyzed the distribution of float variables to understand data patterns. Energy consumption follows a normal distribution, meaning it is relatively predictable, with a standard deviation close to its mean. Month, temperature, humidity, and renewable energy show bimodal distributions, indicating two peaks at different stages, likely due to seasonal changes. Hour and occupancy are multimodal, suggesting energy usage and occupancy fluctuate across multiple time intervals. These trends highlight cyclical energy demand, requiring GreenWatts Energy to optimize resource allocation based on peak periods.



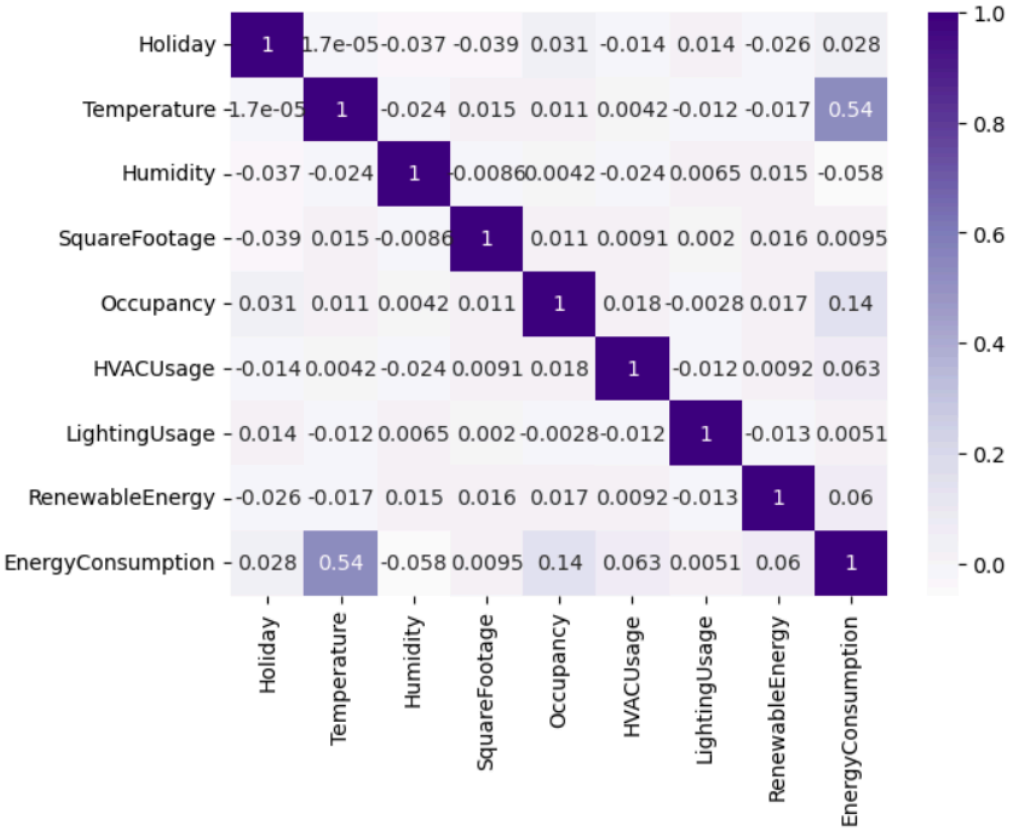
Exploratory Data Analysis

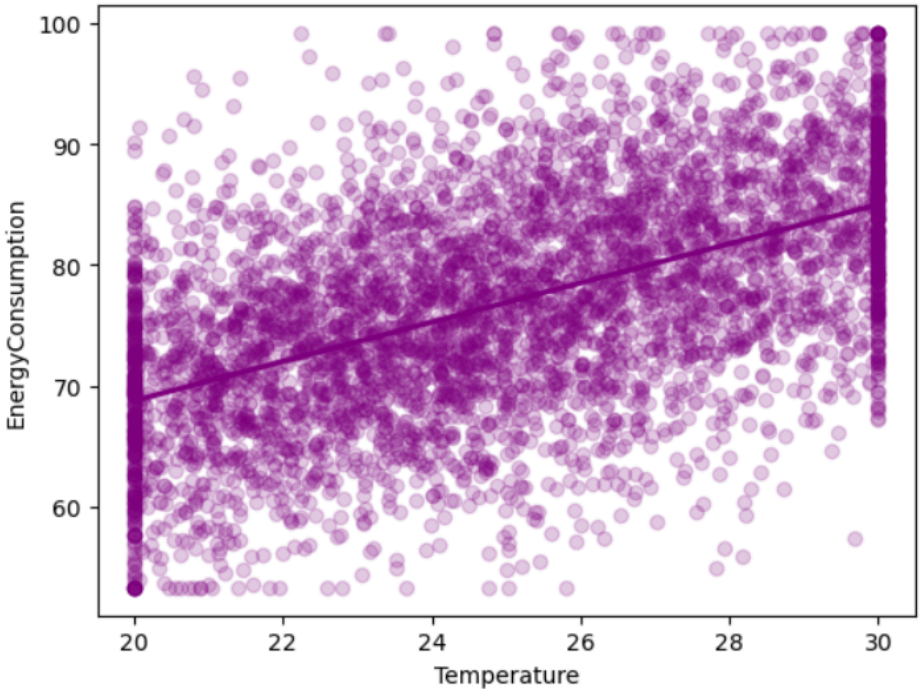
Feature Engineering

Moving on to feature engineering, all categorical data must be one-hot encoded to ensure compatibility with the predictive models. This transformation is essential, as the models I plan to use require properly encoded categorical features for optimal performance.

Correlations

I analyzed the correlation between variables to identify key factors influencing energy consumption. This insight is crucial for GreenWatts Energy, as it helps prioritize the most impactful variables for accurate predictions while minimizing less significant ones. The visualization reveals that temperature has the strongest correlation with energy consumption at 0.54, indicating a moderate positive relationship—suggesting that as temperature rises, energy usage tends to increase. Conversely, the holiday variable showed a negative correlation of -0.028, meaning energy consumption slightly decreases on holidays.


Holiday 1 1.7e-05 -0.037 -0.039 0.031 -0.014 0.014 -0.026 0.028
Temperature 1.7e-05 1 -0.024 0.015 0.011 0.0042 -0.012 -0.017 0.54
Humidity -0.037 -0.024 1 -0.0086 0.0042 -0.024 0.0065 0.015 -0.058
SquareFootage -0.039 0.015 -0.0086 1 0.011 0.0091 0.002 0.016 0.0095
Occupancy 0.031 0.011 0.0042 0.011 1 0.018 -0.0028 0.017 0.14
HVACUsage -0.014 0.0042 -0.024 0.0091 0.018 1 -0.012 0.0092 0.063
LightingUsage 0.014 -0.012 0.0065 0.002 -0.0028 -0.012 1 -0.013 0.0051
RenewableEnergy -0.026 -0.017 0.015 0.016 0.017 0.0092 -0.013 1 0.06
EnergyConsumption 0.028 0.54 -0.058 0.0095 0.14 0.063 0.0051 0.06 1
Holiday
Temperature
Humidity
SquareFootage
Occupancy
HVACUsage
LightingUsage
RenewableEnergy
EnergyConsumption
1.0
0.8
0.6
0.4
0.2
0.0


EnergyConsumption
100
90
80
70
60
20
22
24
26
28
30
Temperature

Hypothesis Testing

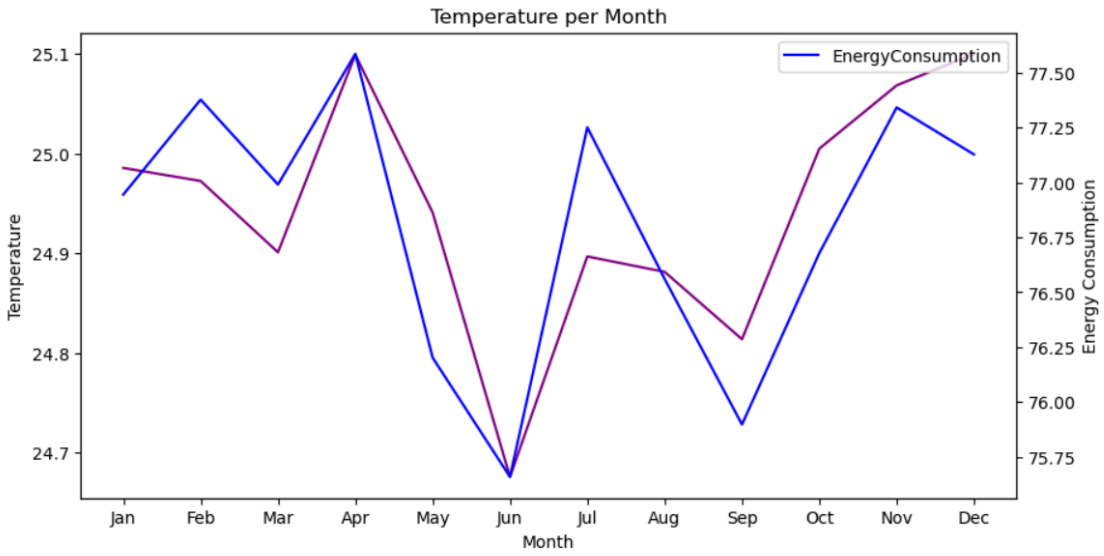
While temperature shows the highest positive correlation with energy consumption, correlation alone does not imply causation; it could simply be a random relationship. To verify this, I conducted hypothesis testing to determine whether the correlation was statistically significant.

The null hypothesis (H_0) states that there is no true correlation between temperature and energy consumption, meaning any observed relationship is due to chance. The alternative hypothesis (H_1) suggests a significant correlation, indicating a genuine connection between the two variables. Using the Pearson correlation test (pearsonr() method in SciPy), I calculated a p-value of 0.0, which allows us to reject the null hypothesis and accept that the correlation is statistically significant. This confirms that temperature plays a crucial role in energy consumption, validating its use as a key predictor variable in further analysis.

Additional Feature Exploration

I expanded my analysis to explore additional variables influencing energy consumption. I found that throughout the year, consumption ranged from 76.17 to 77.33, with Tuesdays and Saturdays showing the highest usage, while Mondays and Thursdays had the lowest. When comparing energy consumption on holiday versus non-holiday days, I discovered little variation and usage remained consistent with typical weekday patterns.

One of the most notable findings emerged when examining monthly temperature trends alongside energy consumption patterns. Generally, as temperature increases, energy consumption also rises, indicating a direct relationship. However, during colder months, energy consumption spiked even higher, suggesting increased heating demands in lower temperatures. This trend highlights the seasonal impact on energy usage and the importance of considering temperature fluctuations in predictive modeling.



Pre-processing & Training Data Development

Since the only remaining categorical columns were numerical categorical variables, I focused on **one-hot encoding** them to ensure compatibility with the predictive models. This step was essential because numerical categorical features, if left unencoded, could be misinterpreted as continuous variables rather than discrete categories. Proper encoding allows the model to recognize distinct groups within the data, improving its accuracy in capturing patterns and relationships.

Scaling

In the scaling step, I standardized all numerical features to ensure consistency across the dataset. Since numerical values had different ranges, failing to scale them could lead to models assigning disproportionate importance to larger valued features. To prevent this, I used StandardScaler() from sklearn, which transforms the data by centering it around a mean of 0 and standardizing deviations to 1. This improves model accuracy and helps algorithms converge faster during training.

Split the Data

Splitting the data is one of my favorite steps in building algorithms because I know I'm much closer to modeling. I used the train_test_split() method because my dataset is large, making it essential to divide it into distinct training and testing sets. This helps

prevent overfitting, ensuring the model learns patterns from training data while being able to make accurate predictions on unseen data. The training set allows the model to learn relationships, while the test set evaluates how well the model generalizes beyond the data it was trained on.

Modeling & Model evaluation

Cross- Validation, Modeling and Model Evaluation

To predict energy consumption, I trained four models: Linear Regression, Random Forest Regressor, Support Vector Regressor, and XGBoost Regressor. To ensure a robust evaluation, I applied cross-validation, allowing each model to be tested across different batches of data. My goal was to find a model that exhibited low error in energy consumption predictions.The ideal predictive model should have a low RMSE, a low MAE, a low MAPE and a high R^2 score. With Cross Validation the models gave the following scores.

Linear Regression Model Evaluation Scores {'Cross-Validated RMSE': 7.649073347969607, 'Cross-Validated MAE': 6.109298336476741, 'Cross-Validated R2-SCORE': 0.3160332814094657, 'Cross-Validated MAPE': 8.152251830124545}
Random Forest Regressor Model Evaluation Scores {'Cross-Validated RMSE': 7.662464749499648, 'Cross-Validated MAE': 6.123869157025364, 'Cross-Validated R2-SCORE': 0.3132864171483618, 'Cross-Validated MAPE': 8.169134536114072}
Support Vector Regressor Model Evaluation Scores {'Cross-Validated RMSE': 7.662316518822432, 'Cross-Validated MAE': 6.1279417938770875,

'Cross-Validated R2-SCORE': 0.3136910166781213, 'Cross-Validated MAPE': 8.194621999628904}
XGBoost Regressor Model Evaluation Scores {'Cross-Validated RMSE': 7.876765633358914, 'Cross-Validated MAE': 6.296055331004747, 'Cross-Validated R2-SCORE': 0.27451448056390654, 'Cross-Validated MAPE': 8.390630579555472}

The Random Forest Regressor proved to be the most effective model, achieving the lowest RMSE, MAE, and MAPE while securing the highest R^2 score among the four models. Its low MAPE score indicates that the model consistently produces accurate predictions with minimal percentage error. To enhance its predictive power and refine its performance, the next step involves hyperparameter tuning, allowing for fine-tuning of key parameters to optimize accuracy and improve the model's ability to explain variations in the data.

Model Optimization_

RandomizedSearchCV

Each model underwent hyperparameter optimization using RandomizedSearchCV, a technique that efficiently finds the best parameter values for a model. Instead of exhaustively testing all possible combinations, it randomly selects different sets of values and evaluates their performance, making the tuning process faster while still identifying effective configurations.

The Linear Regression model performed best with an n_jobs value of -1, achieving strong evaluation metrics. The Random Forest Regressor was fine-tuned across multiple parameters, ‘n_estimators’, ‘max_depth’, ‘min_samples_split’, ‘min_samples_leaf’, and ‘max_samples’ yielding the most effective parameter choices for optimal performance. For the Support Vector Regressor, the parameters ‘C’, ‘epsilon’, ‘kernel’, and ‘gamma’ were tested to enhance prediction accuracy. The

XGBoost Regressor underwent tuning for ‘n_estimators’, ‘learning_rate’, ‘random_state’, and ‘max_depth’ to improve model generalization. Below are the best-performing parameter values and the corresponding evaluation scores after the tuning process was completed.

Linear Regression

Best Parameters for RMSE: {'n_jobs': -1}
Best RMSE Score: 7.6490733479696065
Best Parameters for MAE: {'n_jobs': -1}
Best MAE Score: 6.109298336476738
Best Parameters for R^2: {'n_jobs': -1}
Best R^2 Score: 0.3160332814094661
Best Parameters for MAPE: {'n_jobs': -1}
Best MAPE Score: 8.152251830124543

Random Forest Regressor

Best Parameters for RMSE: {'n_estimators': 800, 'min_samples_split': 10, 'min_samples_leaf': 10, 'max_samples': 0.75, 'max_depth': 10}
Best RMSE Score: 7.620541791142202
Best Parameters for MAE: {'n_estimators': 800, 'min_samples_split': 10, 'min_samples_leaf': 10, 'max_samples': 0.75, 'max_depth': 10}
Best MAE Score: 6.084409995935495
Best Parameters for R^2: {'n_estimators': 800, 'min_samples_split': 10, 'min_samples_leaf': 10, 'max_samples': 0.75, 'max_depth': 10}
Best R^2 Score: 0.32082886610802797
Best Parameters for MAPE: {'n_estimators': 200, 'min_samples_split': 10, 'min_samples_leaf': 5, 'max_samples': 1.0, 'max_depth': 40}
Best MAPE Score: 8.173499811687648

Support Vector Regressor

Best Parameters for RMSE: {'kernel': 'rbf', 'gamma': 'auto', 'epsilon': 0.05, 'C': 8.0}
Best Scores for RMSE: 7.674991764225699
Best Parameters for MAE: {'kernel': 'rbf', 'gamma': 'auto', 'epsilon': 0.05, 'C': 8.0}
Best Score for MAE: 6.123428707909987

Best Parameters for R²: {'kernel': 'rbf', 'gamma': 'auto', 'epsilon': 0.05, 'C': 8.0}
Best Score for R²: 0.3112816560020096
Best Parameters for MAPE: {'kernel': 'poly', 'gamma': 'auto', 'epsilon': 0.5, 'C': 5.0}
Best MAPE Score: 9.414885742891526

XGBoost Regressor
Best Parameters for RMSE: {'random_state': 50, 'n_estimators': 100, 'max_depth': 40, 'learning_rate': 0.01}
Best Scores for RMSE: 8.276356200631842
Best Parameters for MAE: {'random_state': 50, 'n_estimators': 100, 'max_depth': 40, 'learning_rate': 0.01}
Best Score for MAE: 6.61367802949705
Best Parameters for R²: {'random_state': 50, 'n_estimators': 100, 'max_depth': 40, 'learning_rate': 0.01}
Best Score for R²: 0.19923273967388516
Best Parameters for MAPE: {'random_state': 1, 'n_estimators': 1500, 'max_depth': 10, 'learning_rate': 1.0}
Best MAPE Score: 10.531843849585117

After completing the hyperparameter tuning, I confirmed that the Random Forest Regressor remained the strongest model. It achieved the lowest RMSE and MAE, along with the highest R² score, outperforming the other models. While Linear Regression had the best MAPE score at 8.152, the Random Forest model followed closely with 8.173, a marginal difference. Given its superior performance across the other evaluation metrics, I confidently selected Random Forest Regressor as the best model overall.

Final Model

Now that I have my best model with its best parameters it is time to test it with data it has never seen before . Below is the model with its metric scores.

Model

```
final_model = RandomForestRegressor(n_jobs=-1, random_state=0, bootstrap=True,
n_estimators = 800, min_samples_split= 10, min_samples_leaf= 10,max_samples =
0.75, max_depth = 10)
final_model.fit(X_train, y_train)
y_pred = final_model.predict(X_test)
```

Evaluation Methods

```
rmse = mean_squared_error(y_test, y_pred) ** 0.5
mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
mape = mean_absolute_percentage_error(y_test, y_pred) * 100
```

Scores

RMSE for Final Model 7.844157484895022

MAE for Final Model 6.219433685792127

R2 for Final Model 0.2627934978203159

MAPE for Final Model 8.306582728385049

The final model's test results closely align with the scores obtained during hyperparameter tuning for the Random Forest Regressor. The RMSE and MAE remained low, indicating that the error rate stayed consistent even when applied to unseen data. Additionally, the MAPE below 10% confirms that, on average, the model's predictions deviate by less than 10% from actual values, reinforcing its reliability in making accurate forecasts.

Recommendations

GreenWatts Energy and its stakeholders can leverage this predictive model to forecast energy consumption trends with greater accuracy. By analyzing temperature data, they can anticipate fluctuations in energy demand and proactively adjust production levels. This capability is especially valuable in preventing power outages during peak periods, such as extreme heat events when energy usage surges. Preparing for high-demand

periods in advance ensures operational efficiency and cost savings, as GreenWatts Energy can optimize its energy production rather than relying on expensive external purchases from other suppliers. Ultimately, integrating this model into their strategic planning can enhance resource allocation, minimize unexpected shortages, and improve overall energy management.