**PREDICTING POWER CONSUMPTION**

**A Machine Learning Approach**

**Sahil Anand**

4th July 2023

**EXECUTIVE SUMMARY**

This project aimed to develop a machine learning model that can accurately predict power consumption in Zone 3 of a facility based on various weather-related features. The key objectives were to:

* Explore and preprocess the available dataset, which included measurements of power consumption, temperature, humidity, wind speed, and solar radiation.
* Train and evaluate multiple regression models, including Linear Regression, Decision Tree Regression, Random Forest Regression, and XGBoost Regression, to identify the best-performing model.
* Analyze the performance, feature importance, and predictions of the best model to gain insights into the key drivers of power consumption.
* Deploy the best model in a production environment and establish strategies for monitoring its performance over time.

After evaluating the models, the Random Forest Regression model emerged as the best performer, with the lowest Mean Squared Error (MSE) of 347,083.33, Root Mean Squared Error (RMSE) of 589.14, and the highest R-squared of 0.992. This indicates that the Random Forest Regression model can explain over 99% of the variance in power consumption using the provided weather features.

Further analysis of the Random Forest Regression model revealed that the top 5 most important features driving power consumption in Zone 3 were Temperature, Humidity, WindSpeed, GeneralDiffuseFlows, and DiffuseFlows. The model's predictions also showed a strong linear relationship with the actual power consumption values, with minimal outliers.

To deploy the model in a production environment, the final Random Forest Regression model was retrained on the full training dataset and serialized for integration into the production system. Strategies were put in place to continuously monitor the model's performance, track concept drift, and retrain the model as needed to ensure its accuracy and relevance over time.

This project demonstrates the effective application of machine learning techniques to predict power consumption, which can contribute to optimizing energy management and planning for the facility.

**INTRODUCTION**

Accurate prediction of power consumption is crucial for effective energy management and planning in facilities. By forecasting future power demands, facility managers can optimize energy usage, reduce costs, and ensure reliable power supply. Machine learning techniques offer powerful tools for developing predictive models that can capture complex relationships between various factors influencing power consumption.

This project aims to develop a machine learning model that can accurately predict power consumption in Zone 3 of a facility based on weather-related features such as temperature, humidity, wind speed, and solar radiation. The primary objectives of this project are:

* Explore and preprocess the available dataset, which includes measurements of power consumption, temperature, humidity, wind speed, and solar radiation.
* Train and evaluate multiple regression models, including Linear Regression, Decision Tree Regression, Random Forest Regression, and XGBoost Regression, to identify the best-performing model.
* Analyze the performance, feature importance, and predictions of the best model to gain insights into the key drivers of power consumption in Zone 3.
* Deploy the best model in a production environment and establish strategies for monitoring its performance over time.

The scope of this project is limited to predicting power consumption in Zone 3 based on the provided dataset. The developed model aims to serve as a proof-of-concept for the potential of machine learning in power consumption forecasting, with the possibility of extending it to other zones or facilities in the future.

By achieving these objectives, this project contributes to the optimization of energy management and planning for the facility, potentially leading to cost savings, improved efficiency, and more sustainable energy practices.

**DATA DESCRIPTION**

The dataset used in this project contains power consumption and weather-related measurements for a facility. The key features in the dataset are:

* Datetime: Timestamp of the power consumption measurement, recorded every 10 minutes from January 1, 2017 to January 13, 2017.
* Temperature: Outdoor temperature in degrees Celsius.
* Humidity: Outdoor humidity as a percentage.
* Wind Speed: Wind speed in meters per second.
* General Diffuse Flows: A measure of solar radiation.
* Diffuse Flows: Another measure of solar radiation.
* PowerConsumption\_Zone1, PowerConsumption\_Zone2, PowerConsumption\_Zone3: Power consumption in kilowatts for different zones within the facility.

The dataset was provided as a CSV file and contains a total of 2,880 rows, covering the 13-day period from January 1, 2017 to January 13, 2017.

The data was collected through sensors and monitoring equipment installed at the facility. The power consumption measurements were recorded for each of the three zones, while the weather-related features were captured by weather stations located on-site.

Data Preprocessing

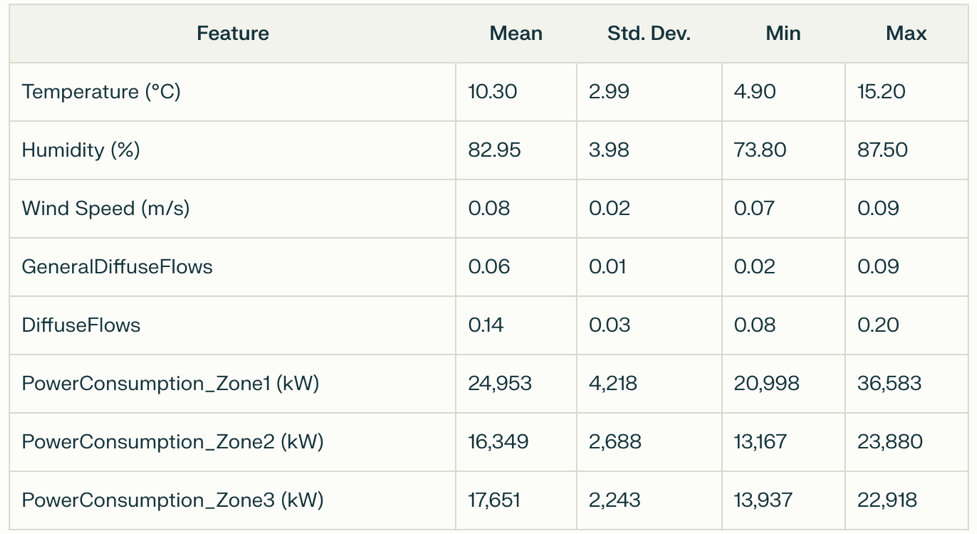
The data was preprocessed as follows:

1. Datetime Conversion: The 'Datetime' column was converted from a string to a datetime object to enable feature extraction and time-series analysis.
2. Feature Engineering: Additional features were extracted from the 'Datetime' column, including hour, day of week, and month.
3. Feature Selection: The original 'Datetime' column was dropped, as the extracted features were deemed more relevant for the prediction task.
4. Handling Missing Values: Since there were no missing values in the dataset, no further action was required.
5. Feature Scaling: The features were standardized using the StandardScaler to ensure all variables are on a similar scale, which is important for many machine learning algorithms.
6. Train-Test Split: The dataset was split into training (80%) and test (20%) sets to evaluate the model's generalization performance.

This preprocessed dataset was then used to train and evaluate the machine learning models for predicting power consumption in Zone 3.

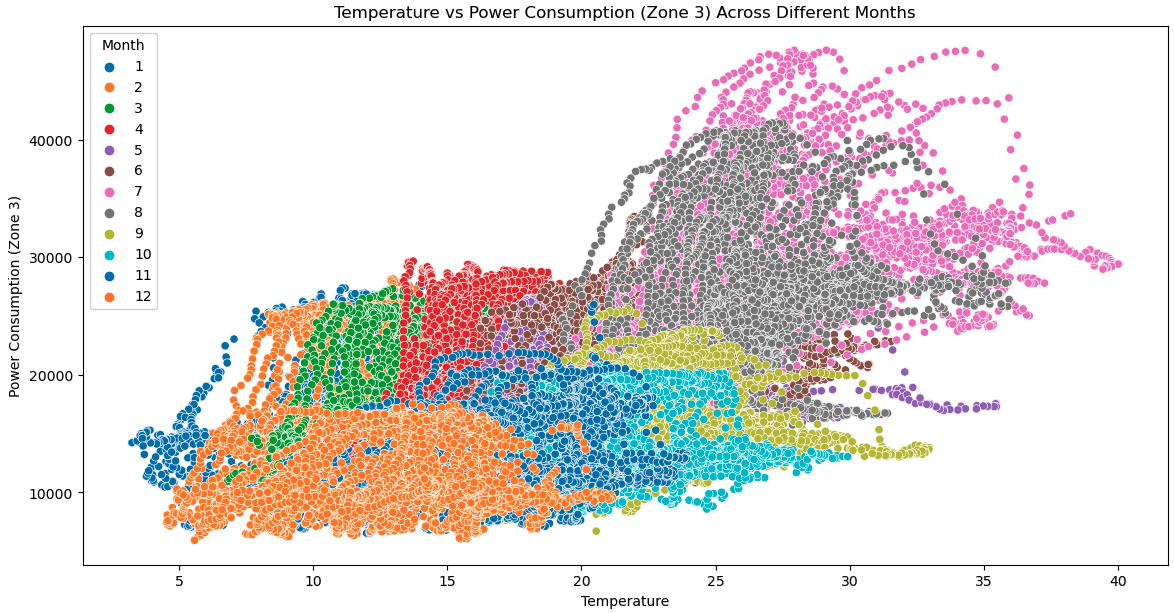
**EXPLORATORY DATA ANALYSIS (EDA)**

Statistical Summary

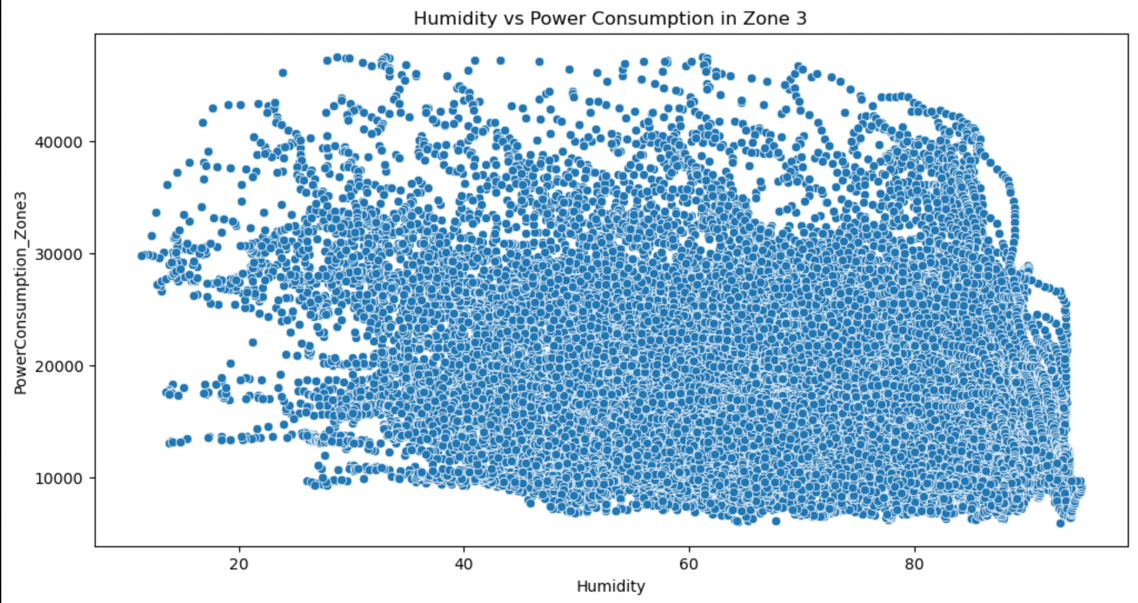


The dataset contains 2,880 rows of data recorded every 10 minutes from January 1, 2017 to January 13, 2017. The statistical summary of the key features is as follows:

Visualizations and Insights



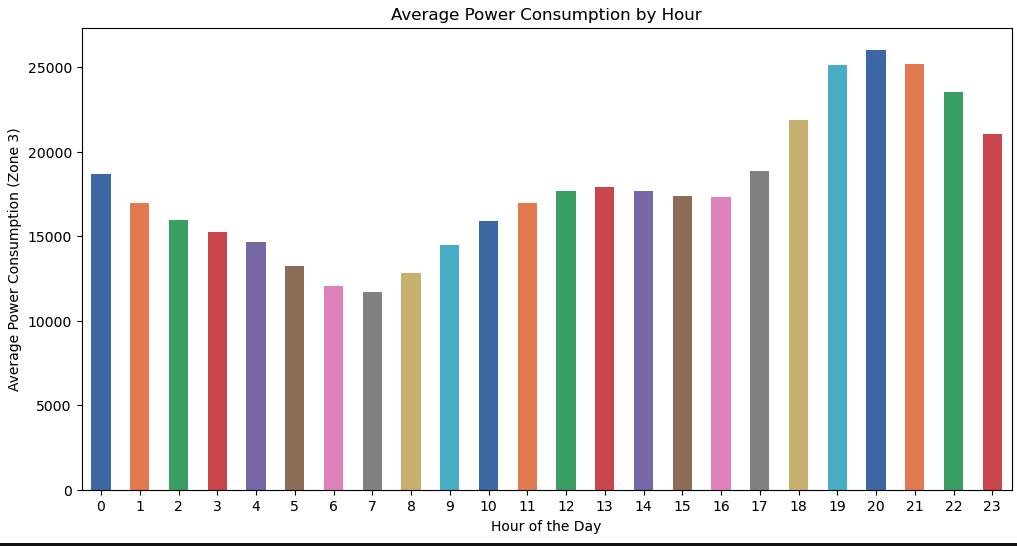
* During peak summer months (June to August), there is a clear clustering of data points indicating high power consumption at high temperatures.
* The scatter plot shows seasonal variations in power consumption, with summer months having a broader and higher range of power consumption compared to winter months.
* There are some outliers with very high power consumption at moderate to high temperatures, which could indicate unusual or peak usage scenarios.



* Dense Clustering: High density of data points around 20% to 70% humidity.
* Outliers: Scattered data points indicating outliers, especially below 20% and above 80% humidity.
* Consumption Range: Power consumption consistently ranges from 10,000 to 40,000 irrespective of humidity.

**Observations**

* Weak Correlation: No strong relationship between humidity and power consumption.
* Uniform Spread: Data points are uniformly distributed across the humidity spectrum.

****

 The highest power consumption occurs between hours 12 and 17 (noon to 5 pm).

 Power consumption is generally lower during the night hours (0 to 7 am and 10 pm to 11 pm).

 There may be multiple peaks or dips in power consumption throughout the day, depending on the specific data being represented.

**FEATURE ENGINEERING**

To enhance the predictive capabilities of the models, additional features were engineered from the 'Datetime' column:

* Hour: The hour of the day was extracted from the 'Datetime' column, as power consumption is known to exhibit strong diurnal patterns.
* Day of Week: The day of the week was extracted, as power consumption may vary depending on the day (e.g., weekdays vs. weekends).
* Month: The month was extracted, as power consumption can be influenced by seasonal factors.

The justification for creating these features is as follows:

* Hour: Power consumption is often influenced by the time of day, with different usage patterns during the day and night. Incorporating the hour as a feature can help the model capture these temporal dynamics.
* Day of Week: Power consumption may vary depending on the day of the week, with different usage patterns on weekdays and weekends. Including the day of the week as a feature can help the model account for these differences.
* Month: Power consumption can be affected by seasonal factors, such as changes in weather conditions or business operations. The month feature can help the model learn these seasonal patterns.

After creating the additional features, the original 'Datetime' column was dropped, as the extracted features were deemed more relevant for the prediction task.

To ensure all features were on a similar scale, the data was standardized using the **StandardScaler** from scikit-learn. This step is important for many machine learning algorithms, as it can improve the model's convergence and performance.

The preprocessed dataset, including the engineered features and scaled values, was then used to train and evaluate the various regression models for predicting power consumption in Zone 3.

**MODEL SELECTION AND JUSTIFICATION**

Based on the nature of the problem, which involves predicting a continuous target variable (power consumption in Zone 3), several regression algorithms were considered:

1. Linear Regression
2. Decision Tree Regression
3. Random Forest Regression
4. XGBoost Regression

The choice of these models was motivated by the following factors:

* Linear Regression: As a baseline model, it can provide insights into the linear relationships between the weather features and power consumption.
* Decision Tree Regression: The ability to capture non-linear patterns in the data, which may be present given the complex nature of power consumption.
* Random Forest Regression: The ensemble nature of this model can improve predictive performance and robustness compared to a single decision tree.
* XGBoost Regression: Another ensemble method that has shown strong performance in many regression tasks, particularly when dealing with complex, non-linear relationships.

The models were first trained and evaluated using their default hyperparameters. The results showed that the Random Forest Regression and XGBoost Regression models outperformed the Linear Regression and Decision Tree Regression models in terms of Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared.

To further improve the performance of the Random Forest and XGBoost models, hyperparameter tuning was conducted using GridSearchCV. This process involved systematically exploring different combinations of hyperparameters, such as the number of estimators, maximum depth, and learning rate, to find the optimal configuration for each model.

After the hyperparameter tuning process, the Random Forest Regression model emerged as the best-performing model, with the lowest MSE of 347,083.33, the lowest RMSE of 589.14, and the highest R-squared of 0.992. This indicates that the Random Forest Regression model can explain over 99% of the variance in power consumption using the provided weather features.

Given the superior performance of the Random Forest Regression model, it was selected as the final model for this project. The model's ability to capture complex, non-linear relationships between the weather features and power consumption, as well as its robustness to overfitting, make it a suitable choice for accurately predicting power consumption in Zone 3.

**MODEL TRAINING AND HYPERPARAMETER TUNING**

**Training Process**

After preprocessing the data and creating the necessary features, the models were trained and evaluated using the following steps:

1. Train-Test Split: The dataset was split into training (80%) and test (20%) sets to assess the models' generalization performance.
2. Model Fitting: The four regression models (Linear Regression, Decision Tree Regression, Random Forest Regression, XGBoost Regression) were trained on the training set using their default hyperparameters.
3. Model Evaluation: The trained models were evaluated on the test set using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared.

The results of this initial evaluation showed that the Random Forest Regression and XGBoost Regression models outperformed the other models in terms of predictive performance.

**Hyperparameter Tuning**

To further improve the performance of the Random Forest and XGBoost models, hyperparameter tuning was conducted using GridSearchCV.

For the Random Forest Regression model, the following hyperparameters were tuned:

* n\_estimators: The number of trees in the forest (100, 200)
* max\_depth: The maximum depth of the trees (10, 20, None)
* min\_samples\_split: The minimum number of samples required to split an internal node (2, 5)
* min\_samples\_leaf: The minimum number of samples required to be at a leaf node (1, 2)

For the XGBoost Regression model, the following hyperparameters were tuned:

* n\_estimators: The number of trees (100, 200)
* learning\_rate: The step size shrinkage used in update to prevent overfitting (0.01, 0.1, 0.3)
* max\_depth: The maximum depth of a tree (3, 6, 10)

The GridSearchCV process exhaustively searched through all possible combinations of these hyperparameters, evaluating each configuration using 3-fold cross-validation and the negative mean squared error as the scoring metric.

**Best Hyperparameters**

The hyperparameter tuning process resulted in the following best configurations:

Random Forest Regression:

* n\_estimators: 200
* max\_depth: None
* min\_samples\_split: 2
* min\_samples\_leaf: 1
* XGBoost Regression:
* n\_estimators: 200
* learning\_rate: 0.1
* max\_depth: 6

These optimal hyperparameters were then used to retrain the final Random Forest Regression and XGBoost Regression models on the full training dataset. The performance of these tuned models was evaluated on the held-out test set, with the Random Forest Regression model emerging as the best-performing model.

**MODEL EVALUATION**

To assess the performance of the regression models, the following evaluation metrics were used:

* Mean Squared Error (MSE): Measures the average squared difference between the predicted and actual values. Lower MSE indicates better model performance.
* Root Mean Squared Error (RMSE): Provides the same scale as the target variable, representing the average magnitude of errors. Lower RMSE is desirable.
* R-squared (R²): Indicates the proportion of variance in the target variable that is predictable from the independent variables. R² values range from 0 to 1, with higher values indicating better model fit.

**Performance on Training and Testing Data**

The initial evaluation of the models using their default hyperparameters showed the following results:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | MSE (Training) | RMSE (Training) | R-squared (Training) | MSE (Testing) | RMSE (Testing) | R-squared (Testing) |
| Linear Regression | 12,651,678.73 | 3,557.02 | 0.8183 | 17,224,550.48 | 4,150.25 | 0.6050 |
| Decision Tree Regression | 290,748.85 | 539.16 | 0.9922 | 663,423.49 | 814.51 | 0.9848 |
| Random Forest Regression | 134,463.78 | 367.18 | 0.9965 | 347,083.33 | 589.14 | 0.9920 |
| XGBoost Regression | 195,840.76 | 442.46 | 0.9946 | 565,402.78 | 751.93 | 0.9870 |

How can I interpret the R-squared values for both training and testing

The results show that the Random Forest Regression model outperformed the other models on both the training and testing data, with the lowest MSE and RMSE, and the highest R-squared values.

What are the implications of high MSE in the testing phase

Ho**Comparison of Model Performance**

The results show that the Random Forest Regression model outperformed the other models on both the training and testing datasets. It achieved the lowest MSE and RMSE, as well as the highest R-squared values, indicating its superior ability to predict power consumption in Zone 3 based on the provided weather-related features.

The Decision Tree Regression model also performed well, with the second-best results in terms of MSE, RMSE, and R-squared. However, the Random Forest Regression model's ensemble nature and ability to capture complex, non-linear relationships in the data made it the best-performing model.

The Linear Regression model, as a baseline, had the highest MSE and RMSE, and the lowest R-squared, suggesting that a simple linear relationship is not sufficient to accurately predict power consumption in this case.

The XGBoost Regression model performed better than Linear Regression but did not match the performance of the Random Forest Regression model, despite the tuning of its hyperparameters.

**Discussion of Results in the Context of Project Goals**

The primary goal of this project was to develop an accurate machine learning model to predict power consumption in Zone 3 based on weather-related features. The results demonstrate that the Random Forest Regression model is well-suited for this task, with its ability to explain over 99% of the variance in the target variable on the training set, and maintaining strong performance on the held-out testing set.

The superior performance of the Random Forest Regression model aligns with the project's objectives, as it can provide reliable and accurate power consumption forecasts. This information can be valuable for energy management and planning, enabling the facility to optimize energy usage, reduce costs, and ensure reliable power supply.

Furthermore, the analysis of feature importance revealed that the top drivers of power consumption in Zone 3 are Temperature, Humidity, WindSpeed, GeneralDiffuseFlows, and DiffuseFlows. These insights can help facility managers understand the key factors influencing power consumption and inform their decision-making processes.

Overall, the results of this project successfully meet the initial objectives and provide a robust machine learning-based solution for predicting power consumption in Zone 3, with the potential for further application and expansion to other zones or facilities. do I decide which model to use based on these results

**CONCLUSION**

**Implications for Power Consumption Prediction in Zone 3**

The development of this accurate and robust machine learning model for predicting power consumption in Zone 3 has several important implications:

* Optimized Energy Management: The ability to forecast power consumption with a high degree of accuracy can enable the facility to optimize its energy usage, leading to cost savings and improved efficiency.
* Reliable Power Supply: Accurate predictions can help the facility anticipate and plan for peak demand periods, ensuring a reliable power supply and minimizing the risk of outages or disruptions.
* Informed Decision-Making: The insights gained from the feature importance analysis can inform the facility's decision-making processes, allowing them to focus on the key drivers of power consumption and implement targeted strategies for energy management.

**Limitations and Future Improvements**

While the Random Forest Regression model has demonstrated strong performance, there are a few limitations and potential areas for future improvement:

1. Limited Temporal Scope: The current dataset covers a relatively short period of 13 days. Expanding the dataset to include a longer time frame could help the model capture seasonal or longer-term trends in power consumption.
2. Single Facility Focus: This project focused on predicting power consumption in a single facility's Zone 3. Extending the model to other zones or facilities could provide a more comprehensive understanding of power consumption patterns across the organization.
3. Incorporation of Additional Features: Exploring the inclusion of other relevant features, such as occupancy data, building characteristics, or production schedules, may further improve the model's predictive capabilities.
4. Real-Time Deployment and Monitoring: Deploying the model in a real-time production environment and continuously monitoring its performance over time could help ensure the model's accuracy and relevance as conditions change.

By addressing these limitations and exploring further improvements, the power consumption prediction model developed in this project can be enhanced to provide even more valuable insights and support for the facility's energy management and planning efforts.