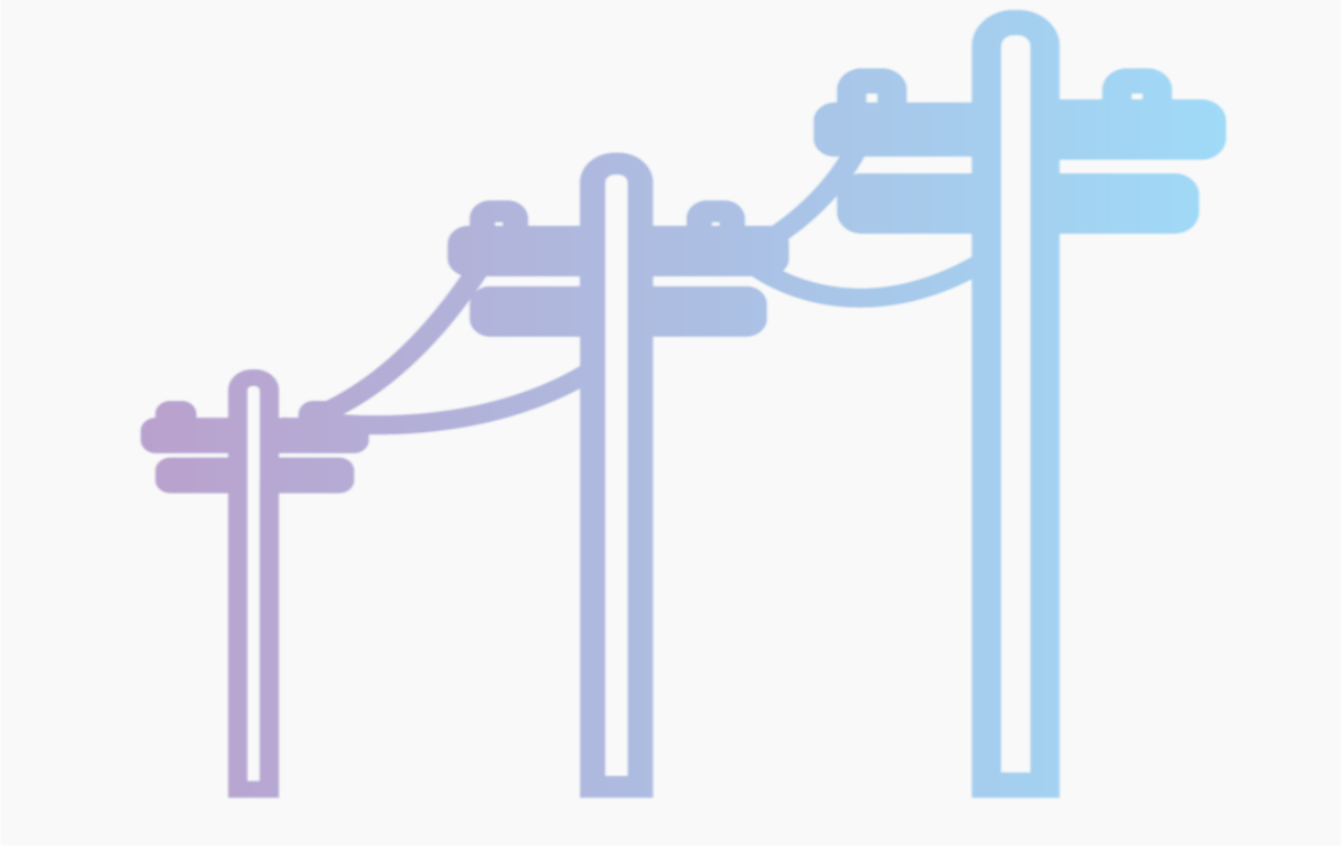
Power consumption prediction project



By
Naresh B Ghorpade

About me..

Mobile application developer with overall of 8 years experience in IT industry where in I have worked various projects and built end to end mobile applications.
I am now transitioning to AI/ML roles to align myself with current advance skillset and prove my expertise in the vast domain of AI world

ML Project Lifecycle

Developing a robust predictive model for power consumption involves a systematic approach, moving through distinct phases from initial understanding to continuous monitoring.

1. Understand Requirements

Clearly define the project's objectives, scope, and desired outcomes for power consumption optimization.

2. Collect Data

Gather all necessary environmental and meteorological datasets relevant to Wellington's Zone 1 power usage.

3. Understand Data

Explore the collected data's structure, identify data types, and discover initial relationships and patterns.

4. Clean Data & EDA

Preprocess data by handling missing values, inconsistencies, and conducting thorough Exploratory Data Analysis.

5. Feature Engineering

Create new, insightful features from existing data to enhance the predictive power of the model.

6. Model Building

Select and train appropriate machine learning algorithms to build the core predictive model.

7. Model Evaluation

Rigorously assess the model's performance using relevant metrics to ensure accuracy and reliability.

8. Deployment

Integrate the validated model into a production environment for real-time power consumption predictions.

9. Monitor Performance

Continuously track the model's performance in real-world scenarios and retrain as needed to maintain accuracy.

Objective & Problem Statement

Core Objective:

The primary objective is to optimize the power consumption of Zone 1 in Wellington, New Zealand. This will be achieved by developing a sophisticated machine learning model capable of accurately predicting power demand based on a comprehensive set of environmental and meteorological factors. This proactive approach will enable better resource allocation, reduce waste, and potentially lower energy costs for the region.

Problem Statement:

Current power consumption in Wellington's Zone 1 is influenced by a complex interplay of dynamic environmental and meteorological conditions. These reactive geological features are contributing more in ineffective usage of power which need to be observed and understand power consumption needs depending upon various geological conditions. The challenge lies in constructing a machine learning solution that can effectively capture these intricate relationships to predict Zone 1 power consumption based on various weather condition, thereby enabling a shift towards more intelligent and optimized energy distribution.

Dataset Overview: Core Data Characteristics

This base overview of data is built upon a power consumption of Zone 1 dataset, providing the foundational insights of predictive model. Understanding its structure and features helps in subsequent steps in our machine learning lifecycle.

Data Source

The raw data was initially provided in an **Excel sheet format**, from which it was extracted and prepared for analysis.

Total rows

The dataset comprises **52,583** individual records, representing a weather, power consumption details and environmental factors.

Total Columns

Given dataset contains 9 distinct columns, encompassing a blend of environmental observations and power consumption metrics.

Data Types

The dataset includes a mix of Integer, Float, and Object data types, accommodating numerical measurements and categorical information.

Feature Overview: Environmental & Meteorological Drivers

Dataset consist of several key environmental and meteorological factors, each playing a crucial role in influencing power consumption. Understanding these features is foundational to building an accurate and robust model.

Temperature

Ambient air temperature in degrees Celsius. Directly impacts heating and cooling demands, especially in residential and commercial sectors.

Humidity

Relative humidity percentage. High humidity can influence the perceived temperature and the efficiency of HVAC systems, indirectly affecting power usage.

Wind Speed

Wind speed in meters per second. Can influence heat loss from buildings (especially in colder conditions) and also impacts the operation of some industrial ventilation systems.

General Diffuse Flows

Total solar radiation received on a horizontal surface, often measured in W/m². Represents the combined direct and diffuse solar energy, impacting lighting and passive heating/cooling.

Diffuse Flows

Solar radiation that has been scattered by the atmosphere and clouds, measured in W/m². Higher diffuse flows often correspond to cloudy conditions, increasing demand for artificial lighting.

Air Quality Index

A numerical index indicating air pollution levels. While not directly a power consumption driver, poor air quality can lead to increased use of air purification systems or reduced outdoor activity, indirectly affecting indoor energy usage patterns.

Cloudiness

Percentage of sky covered by clouds. Directly influences the need for artificial lighting during daylight hours and impacts solar energy potential, thus affecting overall power demand.

Data Cleaning: Laying the Foundation for Accuracy

Before any model building, thorough data cleaning is essential to ensure the quality, consistency, and reliability of our dataset. This foundational step mitigates issues that could otherwise lead to biased or inaccurate predictions.

Column Renaming

All raw column names were standardized to snake_case for improved readability and consistency across the dataset, facilitating easier data manipulation and analysis.

Type Conversions

Critical features like 'Temperature' and 'Humidity' were converted from object (string) data types to float, enabling numerical operations and proper statistical analysis.

Null Value Imputation

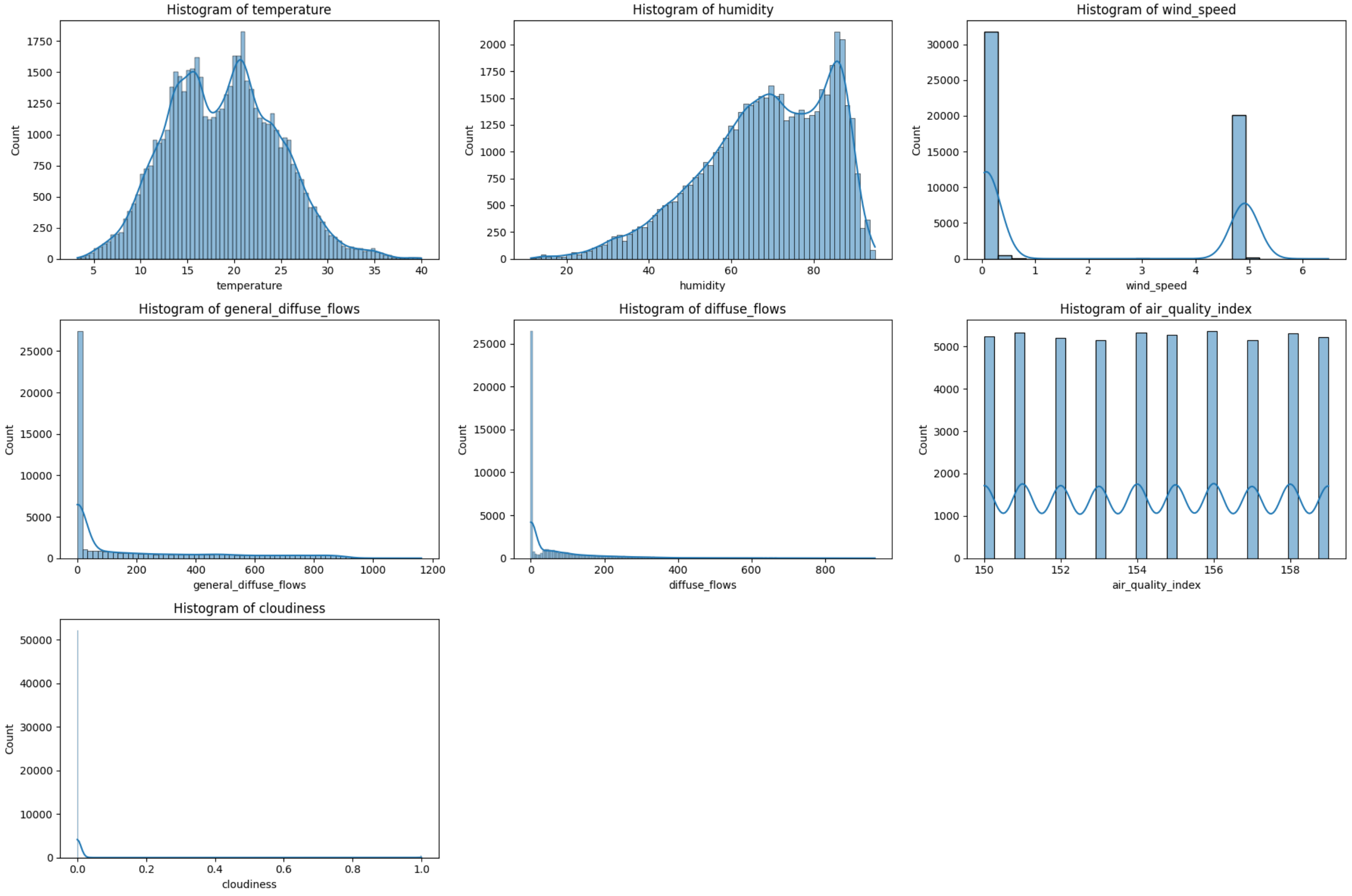
Missing values in key meteorological and environmental columns ('Temperature', 'Humidity', 'Wind Speed', 'general diffuse flows', 'diffuse flows', 'Air Quality Index (PM)') were imputed using respective mean and median strategies to preserve data integrity.

Outlier Capping

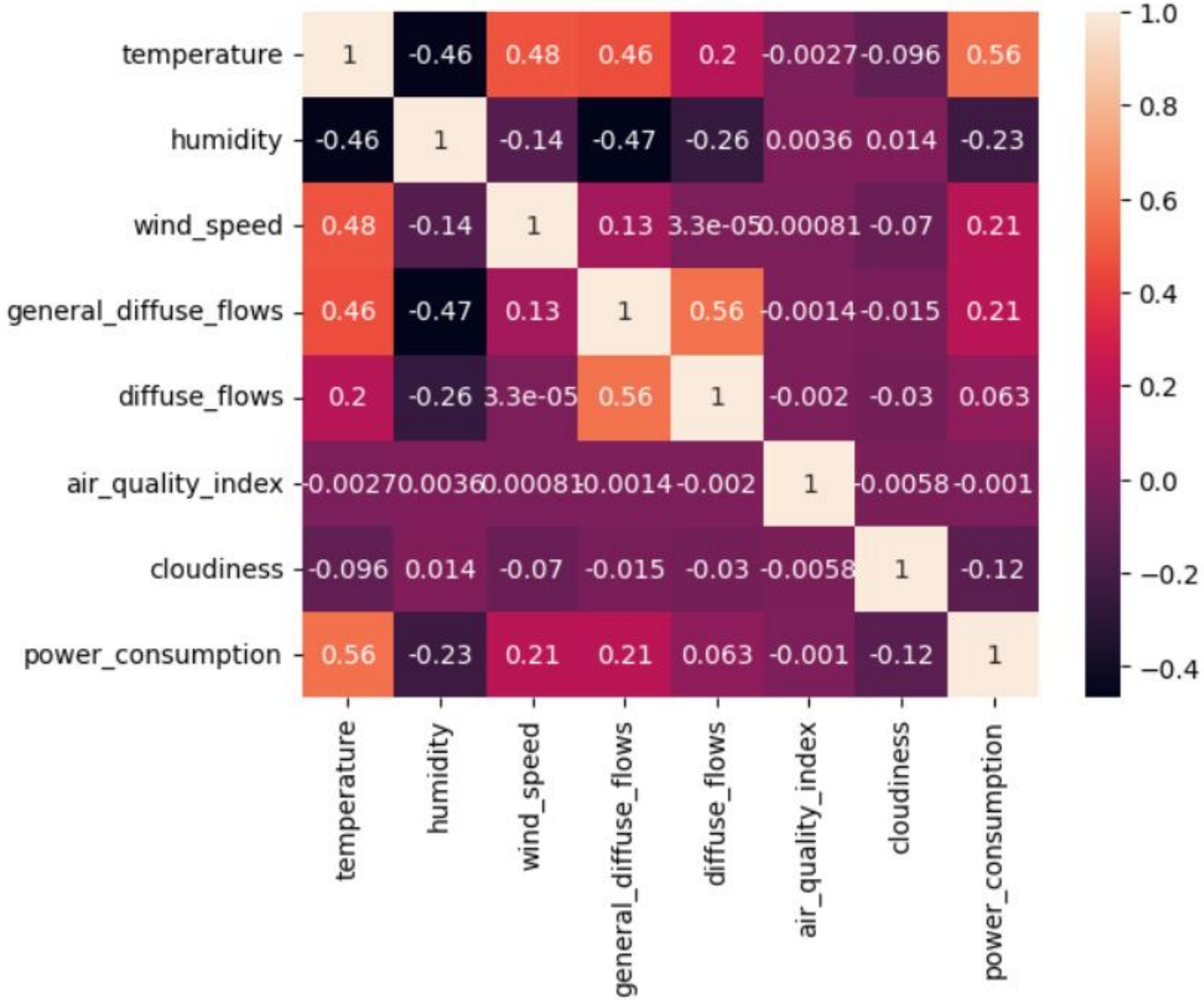
To prevent extreme values from skewing the model, upper and lower value capping techniques were applied to numerical features, ensuring robust data distribution.

Exploratory Data Analysis (EDA): Uncovering Insights

Data distribution

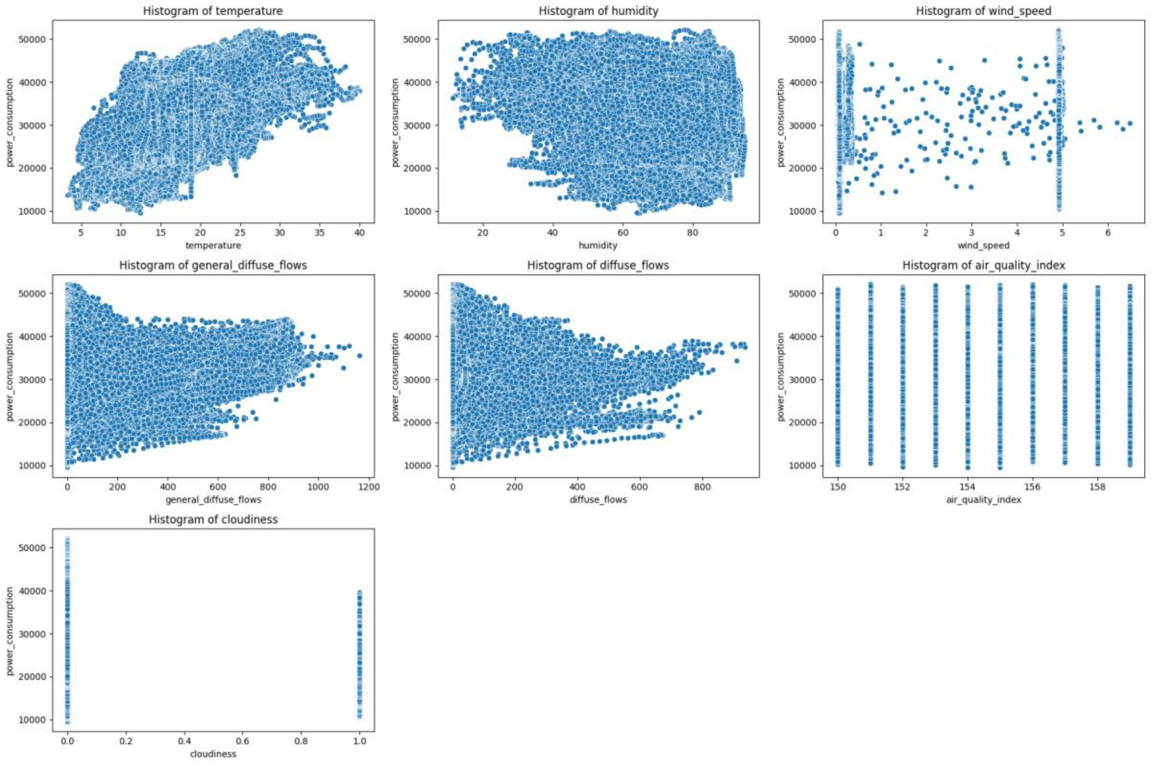


Correlation analysis

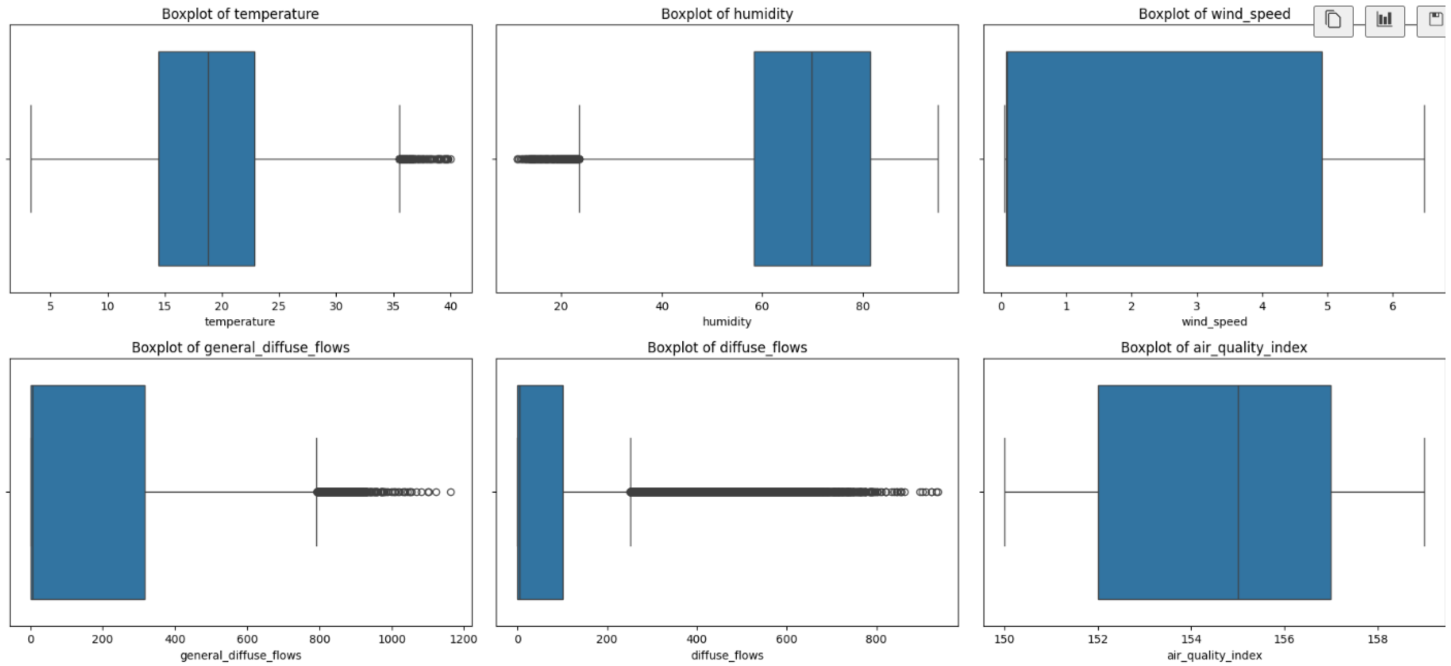


Exploratory Data Analysis (EDA): Uncovering Insights

Correlation pair plot



Outlier analysis



Exploratory Data Analysis (EDA): Overview

Exploratory Data Analysis shows critical patterns and characteristics within the dataset, forming the backbone for feature engineering and model selection. These insights highlight the unique behaviors of environmental drivers impacting power consumption.

Temperature & Humidity Patterns

Temperature exhibits a bimodal distribution (3-40°C), while humidity shows a left-skewed range (11-95%), indicating distinct environmental conditions.

Wind & Diffuse Flow Skewness

Wind speed and both general/diffuse flows are extremely right-skewed.

Cloudiness Imbalance

Cloudiness data is highly imbalanced (approx. 10:1 ratio), which could introduce bias and requires careful handling in model training.

Key Correlations Identified

Temperature positively correlates with power consumption, a vital predictor. Interestingly, humidity shows a slight, unexpected negative correlation.

Diffuse Flow Impact

Power consumption tends to be higher during periods of lower diffuse radiation, possibly due to increased direct sunlight or the need for artificial lighting.

Model Building & Evaluation

Develop & evaluate a machine learning model to accurately predict Zone 1 power consumption. This process involves careful model selection, training, and performance assessment.

Feature Engineering

- Creation of polynomial features (temperature, humidity square and cubical values) helped model to find bit more patterns in data.
- Finding interaction between feature contributed in model performance.

Data Splitting & feature scaling

- Splitting the dataset into training, validation, and testing sets using sklearn train_test_split with 80:20 ratio
- Used standard scaler to test with linear models.

Model Selection & Training

- Explored model with various regression algo models:
- Choose best model to evaluate performance with grid search with cross validation
- **Random Forest Regressor** gave best performance with available data.
- Training models on the training set and fine-tuning hyperparameters using the validation set (e.g., GridSearchCV).

Model Evaluation

- Evaluating the final model on the unseen test set to provide an unbiased estimate of performance.
- **Mean Absolute Error (MAE):** Average magnitude of errors.
- **Root Mean Squared Error (RMSE):** Penalizes larger errors more heavily, sensitive to outliers.
- **R-squared (R^2):** Proportion of variance in the dependent variable predictable from the independent variables.
- Visualizing actual vs. predicted values and residual plots to assess model fit and identify systematic errors.

Conclusion: Process Overview & Performance Metrics

This comprehensive project leveraged environmental and meteorological data to develop a robust machine learning model for optimizing power consumption in Wellington's Zone 1. The iterative process from data understanding to model evaluation ensures a reliable and impactful solution.

Illustrative Performance Metrics

Random forest regressor

Chosen model

Based on our final model, preliminary results on the test set indicate strong predictive capabilities:

3257.9

MAE

Mean Absolute Error, indicating the average absolute difference between predicted and actual power consumption.

4710.7

RMSE

Root Mean Squared Error, penalizing larger prediction errors more heavily, reflecting the model's overall prediction precision.

0.65

R-squared

R^2, indicating that approx. 66% of the variance in power consumption can be explained by our model's features.

Thank you