Power Consumption Prediction – Zone 3

Final Project Report

Author: Rachit Patwa
Date: 11th October 2025

---

1. Introduction

Electricity consumption prediction is a critical component for efficient energy management in industrial and residential zones. Accurate forecasts help in load balancing, reducing operational costs, and planning for future energy requirements.

This project focuses on **predicting power consumption in Zone 3** using historical data and advanced machine learning and time series forecasting models. Both **Random Forest Regression** and **SARIMA** models were implemented to provide accurate and generalized predictions.

---

2. Objective

- Predict the **power consumption for Zone 3** using historical energy usage and weather-related features.
- Develop **generalized models** that can be applied to unseen data for forecasting.
- Compare performance between **machine learning models** and **time series models**.
- Identify the most important factors influencing power consumption.

---

3. Dataset Overview

Dataset Source: Internal/Provided CSV (`powerconsumption.csv`)

Dataset Size: 52,416 records, 9 columns

Key Columns:

Column	Description
Datetime	Timestamp of record
Temperature	Temperature (°C)
Humidity	Relative Humidity (%)
WindSpeed	Wind Speed (km/h)
PowerConsumption_Zone1	Energy consumed in Zone 1 (kWh)

Column	Description
PowerConsumption_Zone2	Energy consumed in Zone 2 (kWh)
PowerConsumption_Zone3	Energy consumed in Zone 3 (kWh)
DayOfWeek	Day of the week (0–6)
Is_Weekend	Weekend flag (0 = weekday, 1 = weekend)

Data Cleaning and Preprocessing:

- Converted `Datetime` to pandas datetime object.
- Extracted **time-based features**: Year, Month, Day, Hour, DayOfWeek, Is_Weekend.
- Detected outliers using **boxplots** and handled them with **winsorization**.
- Applied **log transformation** to `DiffuseFlows` and `GeneralDiffuseFlows` to reduce skewness.
- Standardized features using `StandardScaler` for ML models.

Missing Values and Duplicates:

- Dataset had **no missing values** after cleaning.
- **No duplicate records** found.

---

4. Exploratory Data Analysis (EDA)

4.1 Distribution of Power Consumption

- `PowerConsumption_Zone3` shows **right skewness**, indicating occasional high consumption spikes.
- Daily and hourly consumption trends reveal **peak usage hours** around early evening.

4.2 Correlation Analysis

- Strong positive correlation with **Zone 1 and Zone 2 consumption**.
- Moderate correlation with **Hour** and **Temperature**.

4.3 Visual Insights

1. **Hourly Trends**: Highest consumption in evening hours.
2. **Monthly Patterns**: Slight increase in summer months.
3. **Weekday vs Weekend**: Weekdays show slightly higher consumption.
4. **Feature Relationships**:
    - Temperature positively influences consumption.
    - WindSpeed has negligible effect.
5. **Feature Importance (Correlation-based)**:
    - Top features: `PowerConsumption_Zone1`, `PowerConsumption_Zone2`, `Hour`

Figures and charts: Add your plots here (line plots, heatmaps, scatterplots, etc.)

5. Machine Learning Models

Features Used:

```
Temperature, Humidity, WindSpeed, PowerConsumption_Zone1,
PowerConsumption_Zone2, Hour, DayOfWeek, Is_Weekend, Month
```

Train-Test Split: 80%-20% (time-aware, preserves chronological order)

5.1 Models Trained

- Linear Regression
- Ridge Regression
- Lasso Regression
- ElasticNet Regression
- Random Forest Regressor
- XGBoost Regressor

5.2 Model Performance

Model	RMSE (kWh)	MAE (kWh)	R^2
Random Forest v3	5742.20	5062.61	-2.0998
XGBoost v3	6190.19	5627.71	-2.6024
ElasticNet v3	7510.85	6756.50	-4.3035
Ridge Regression v3	7514.96	6760.15	-4.3093
Lasso Regression v3	7515.17	6760.33	-4.3096
Linear Regression v3	7515.17	6760.33	-4.3096

Observations:

- Random Forest provided the **lowest RMSE and MAE**, making it the best-performing ML model.
- Linear models performed poorly on high-variance data.
- Feature importance indicates **Zone 1 and Zone 2 consumption** as primary predictors.

6. Random Forest Generalization

- Random Forest retrained on **full dataset** for deployment.
- Features were **standardized** using `StandardScaler`.
- Saved both **model** and **scaler** using `joblib` for use in deployment (e.g., Streamlit).

File Outputs:

- `rf_generalized_model.pkl`
- `rf_scaler.pkl`

---

7. Time Series Forecasting – SARIMA

Approach:

- Aggregated daily average consumption for Zone 3.
- Applied **SARIMA (Seasonal ARIMA)** to capture trends and weekly seasonality.

Parameters:

- Non-seasonal: (1, 1, 1)
- Seasonal: (1, 1, 1, 7)

Train-Test Split: Last 20% of data as test set

7.1 SARIMA Performance

Metric	Value
RMSE	6462.85 kWh
MAE	4257.79 kWh
R^2	0.0605

Observations:

- SARIMA effectively captures **weekly patterns** in consumption.
- Slightly higher RMSE than Random Forest but provides **seasonality-aware forecasts**.
- Residual diagnostics indicate **well-behaved residuals** with no major bias.

Saved Model:

- `generalized_sarima_model.pkl`

Figures and charts: Include SARIMA forecast vs actual plots here.

---

8. Conclusion

- Random Forest is the **best ML model** for predictive accuracy.
- SARIMA provides a **time-series perspective**, capturing seasonal patterns.
- Combining ML and time series forecasts can enhance decision-making.
- Feature analysis shows **Zone 1 & Zone 2 consumption** as major influencers of Zone 3 usage.

9. Future Work

1. Include additional features such as **holiday effects, local events, and weather forecasts**.
2. Explore **hybrid models** combining Random Forest and SARIMA for enhanced predictions.
3. Deploy models in **real-time monitoring dashboards** using Streamlit or Flask.
4. Incorporate **explainable AI techniques** to provide interpretability.