Smart Factory Energy Prediction Challenge

REZIG Wassila

May 10, 2025

1. Approach to the Problem

The goal of this challenge was to build a regression model to predict energy consumption in a smart factory based on timestamped sensor data. The pipeline I followed includes:

- Data cleaning: handling missing values, removing invalid entries (e.g., 'error', 'unknown'), and eliminating outliers.
- Feature engineering: extracting features from timestamps (e.g., hour, day, month), and converting data types appropriately.
- Normalization: applied Z-score normalization to numerical features.
- Model training: used Random Forest Regressor as a baseline model.

2. Key Insights from the Data

- Several columns contained invalid strings like "error" or "check" in numerical fields.
- Some negative energy readings were likely anomalies or sensor errors and were removed.
- Timestamp was a strong feature; extracting 'hour' and 'month' revealed potential patterns in energy usage.

3. Model Performance Evaluation

- **Model used:** Random Forest Regressor
- **Metrics:**
  - MAE: 68.85
  - MSE: 25,084.24
  - R^2: 0.04
- These scores indicate the baseline model has poor performance, likely due to high variance in the data or missing important predictive features. Future iterations will aim to improve this.