Smart Factory Energy Prediction Challenge - Final Report

Problem Overview

SmartManufacture Inc. aims to reduce energy costs in its client's manufacturing facility by forecasting equipment energy consumption using environmental sensor data. The goal is to build a machine learning regression model that can accurately predict equipment_energy_consumption and offer actionable insights to optimize operations.

1. Data Preprocessing

Standard Deviation and Outlier Detection

- Plotted the standard deviation for each numerical feature.
- Identified outlier-prone features like zone humidity and outdoor pressure, requiring transformation or scaling.

Missing Value Treatment

- Replaced invalid entries ('???', 'error') with NaN.
- Dropped rows with missing target values.
- Imputed other missing values using median imputation for robustness against outliers.

Feature Selection and Cleaning

- Dropped irrelevant or low-signal columns after correlation and variance analysis.
- Removed random_variable1 and random_variable2 due to minimal correlation and no improvement in model performance.
- Extracted hour, dayofweek, month from timestamp to capture temporal patterns.

Final Features Retained:

- Zone-wise temperature and humidity
- Outdoor temperature, dew point, and pressure
- Time-based features

2. Data Evaluation

- After preprocessing, visualizations confirmed:

  - Smoother distributions
  - No extreme outliers
  - Well-normalized features using StandardScaler
- Dataset was split using an 80/20 train-test split, ensuring representative sampling.

3. Model Training

Problem Type:

- This is a regression problem: predicting continuous numerical values.

Model Selection Rationale:

- Chose Random Forest Regressor for its:
  - Non-linearity handling
  - Resistance to overfitting
  - Built-in feature importance estimation

Mathematical Intuition:

- Random Forest builds multiple decision trees using bootstrapped samples and averages their output.
- Reduces variance (compared to single decision trees) and works well with nonlinear relationships.

Training Setup:

- Applied 5-fold cross-validation for reliable evaluation.
- Compared multiple models; Random Forest gave best performance with default + tuned parameters.

4. Final Model Results

Metric	Value
MAE	~66.15
RMSE	~158.09
R Score	~0.07

Note: Classification metrics like precision, recall, F1-score, and confusion matrix are not applicable

here as the problem is regression-based.

Visualizations:

- Predicted vs Actual scatter plots show moderate prediction alignment.
- Residual plots indicate some variance yet to be explained.

5. Conclusion

We approached the problem with structured steps:

- Cleaned and standardized the dataset
- Identified and removed irrelevant features
- Extracted meaningful time-based features
- Trained and validated multiple regression models

While the Random Forest model achieved moderate performance (R 0.07), it established a strong baseline and highlighted key influencing features such as:

- Outdoor dew point and temperature
- Zone humidity levels

Recommendations

- Zone-Level Optimization: Improve HVAC efficiency in zones with high energy correlation.
- Smart Scheduling: Shift energy-intensive tasks to off-peak hours.
- Sensor Audit: Remove or replace low-signal sensors.
- Advanced Models: Explore XGBoost, LightGBM, or neural networks for enhanced accuracy.