Smart Factory Energy Prediction — Final Report

Candidate: Abhiram
Company: Mechademy.
Role: Data Scientist Intern
Project: Forecasting equipment energy consumption using sensor and environmental data to support energy-efficient manufacturing operations.

📌 Objective

To develop a robust machine learning pipeline that accurately predicts equipment energy consumption (equipment_energy_consumption) based on multivariate sensor and environmental data from a factory. The final model should also offer insights into energy usage patterns and potential operational optimizations.

🔢 1. Data Overview

- **Source**: CSV file with 16,857 rows and 29 features.
- **Data Types**: A mix of environmental variables (temperature, humidity, pressure), operational readings, and sensor outputs across 9 factory zones.
- **Target**: equipment_energy_consumption (numeric).

⚠️ Issues Found:

- Missing values in several columns, including the target.
- Some numeric values were incorrectly typed as object.
- Presence of noise/outliers in several features (e.g., humidity values below 0 or extreme negatives).

📊 2. Exploratory Data Analysis (EDA)

- **Distribution**: Target variable is right-skewed; peak usage fluctuates during the day.
- **Correlation**: Top correlated features with energy usage include:
  - lighting_energy
  - zone1_temperature, zone2_humidity, outdoor_temperature
- **Random Variables**:
  - random_variable1 and random_variable2 had very high cardinality and weak correlation.
  - Scatterplots showed no meaningful trend — these were dropped.

📈 **Visuals Generated:**

Distribution of Equipment Energy Consumption

Time Series of Energy Usage

Correlation Heatmap

Random Variable vs Energy Scatterplots

🧹 3. Data Preprocessing

- Converted object-type numeric columns to float.
- Extracted time features: hour, and day_part (morning, afternoon, etc.).
- Applied one-hot encoding for categorical time-based variables.
- Imputed missing values using **median strategy**.
- Scaled features using StandardScaler.

🧠 4. Feature Selection

- Used RandomForestRegressor to evaluate feature importance.
- Top 15 most important features were selected for final modeling.
- Visualization: Top 20 Important Features Bar Chart.

🤖 5. Model Development & Tuning

Models Trained:

- Random Forest Regressor
- XGBoost Regressor

Evaluation Metrics:

Model	RMSE	MAE	R^2
Random Forest	161.81	68.34	0.02
XGBoost (Initial)	167.80	78.66	-0.048

Best Model (after tuning):

- Model: XGBoost
- Best Params: n_estimators=50, learning_rate=0.1, max_depth=4
- Cross-validated RMSE: 191.93
- Final RMSE: 158.08
- Final R^2: 0.07

📉 6. Model Insights

- **Errors**: Larger prediction errors occurred during peak hours.
- **Visuals**:
  - Distribution of Prediction Errors
  - Boxplot: Errors by Hour of Day
  - Scatterplot: Outdoor Temperature vs Energy Consumption
- **Conclusion**:
  - Energy usage increases during afternoon and evening time slots.
  - Outdoor temperature impacts indoor energy consumption, especially in climate-sensitive zones.

✅ 7. Key Takeaways & Recommendations

🔍 Insights:

- Lighting energy, zone temperature, and outdoor temperature are key energy drivers.
- Time-of-day effects are clear in prediction error patterns.
- Random_variable1 and 2 added noise and were safely excluded.

💡 Recommendations:

- **Shift energy-heavy operations** to low-demand hours (e.g., early morning).
- **Optimize lighting systems** in zones with high lighting_energy correlation.
- **Enhance insulation or temperature regulation** in high-impact zones to mitigate outdoor temperature influence.
- Consider using time-series specific models (e.g., LSTM or Prophet) in the future for improved temporal predictions.