Smart Factory Energy Prediction Challenge - Final Report

1. Approach to the Problem

- Data Exploration and Understanding:
  - Loaded and explored the dataset structure and missing values.
  - Analyzed feature types and basic statistical properties.

- Data Preprocessing:
  - Replaced invalid entries ('error') with NaN.
  - Converted feature columns to appropriate numeric types.
  - Parsed the timestamp column into datetime format.
  - Dropped rows with missing values for clean modeling.

- Feature Engineering:
  - Extracted new time-based features (hour, dayofweek, month) from the timestamp.

- Feature Selection:
  - Performed correlation analysis to retain meaningful features.
  - Dropped random_variable1 and random_variable2.

- Model Development and Tuning:
  - Selected Random Forest Regressor.
  - Applied RandomizedSearchCV with 5-fold cross-validation.

- Model Evaluation:
  - Assessed model performance using MAE, RMSE, and R^2 score.

2. Key Insights from the Data

- Environmental Influence: Zone temperatures and outdoor temperatures are the most influential factors.
- Temporal Patterns: Consumption patterns vary by hour and day of week.
- Feature Relevance: Dropping irrelevant features improved performance.

3. Model Performance Evaluation

- Mean Absolute Error (MAE): ~0.1832
- Root Mean Squared Error (RMSE): ~0.2519
- R^2 Score: ~0.9247

The model achieved a high R^2 score (~92%), indicating strong predictive performance.

4. Recommendations for Reducing Equipment Energy Consumption

- Optimize HVAC Settings: Adjust temperature targets based on operational hours.
- Enhance Building Insulation: Stabilize internal temperatures.
- Leverage Weather Forecasting: Implement proactive energy control.
- Zone-Specific Energy Management: Focus on high-consumption zones.
- Implement Real-Time Monitoring: Detect and address anomalies promptly.

5. Model Limitations

- Random Forest may overfit smaller datasets.
- Model predictions may be less accurate under extreme conditions.
- Sensor anomalies can affect prediction accuracy.

6. Model Selection Justification

Random Forest was selected due to:

- Robustness to outliers and noise.
- Ability to model non-linear relationships.
- In-built feature importance ranking.
- Enhanced generalization after hyperparameter tuning.

Prepared by: Sharvari Sainath Kulkarni

Date: May 8, 2025