Smart Factory Energy Prediction Report

Prateek
SmartManufacture Inc.

Abstract

SmartManufacture Inc. collaborated with a manufacturing facility to develop a predictive pipeline for forecasting equipment energy consumption. Our optimized RandomForest model achieves an RMSE of 154.28, MAE of 62.24, and R^2 of 0.113 on held-out data, enabling actionable recommendations for energy reduction.

1 Results Overview

Model	RMSE	MAE	R^2
Baseline Linear & Ridge	162.00	71.70	0.02
Baseline RandomForest	160.92	69.67	0.035
Tuned RandomForest	154.28	62.24	0.113

2 Data Exploration and Cleaning

2.1 Data Overview

- **Records:** 16,857 from January 2016 onward
- **Features:** 29 total, including nine zone temperature/humidity pairs, lighting energy, and weather metrics
- **Missingness:** Approximately 5–6% per sensor; 5% for the target

2.2 Missing Value Handling

Columns with less than 10% missing were imputed using median values. Time-series sensor data were further refined with interpolation where appropriate.

2.3 Outlier Management

Applied IQR-based clipping at the 5th and 95th percentiles to mitigate the influence of extreme readings.

2.4 Key Observations

- Peak energy draws occurred daily between 14:00 and 17:00 with clear weekly cycles.
- Zone temperature variability correlated strongly (r 0.40) with energy use; outdoor humidity showed a moderate effect (r 0.30).

3 Feature Engineering

3.1 Temporal Features

Extracted hour, day of week, weekend flag, and month from timestamps to capture cyclical patterns.

3.2 Spatial Aggregates

Generated mean and standard deviation summaries across all nine zone temperatures and humidities, reducing dimensionality.

3.3 Interaction Term

Defined a temperature-humidity interaction to model combined environmental effects.

4 Modeling Pipeline

4.1 Preprocessing

Implemented a `ColumnTransformer` that performs median imputation and standard scaling on both raw and engineered features to ensure consistency.

4.2 Baseline Models

- **Linear Regression / Ridge:** RMSE 162, R^2 0.02
- **RandomForest:** RMSE 161, R^2 0.03

5 Hyperparameter Optimization

Conducted a randomized search over 25 hyperparameter sets and 5-fold CV for RandomForest, tuning n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features. The optimal configuration delivered an RMSE of 154.28 on the test set.

6 Interpretation

Using SHAP, we confirmed that zone temperature variability is the most influential predictor, followed by outdoor humidity and lighting energy. When SHAP is unavailable, RandomForest's built-in feature importances yield a consistent ranking.

7 GPU Acceleration

To accelerate training:

- Use XGBoost with `tree_method=gpu_hist` and LightGBM with GPU support.
- Leverage RAPIDS (cuDF and cuML) to perform end-to-end GPU-based data processing and modeling.

8 Deployment

Outlined a `predict.py` script for batch inference, which replicates preprocessing, loads the tuned model, and writes timestamped predictions to CSV, creating directories as needed.

9 Recommendations

1. Install localized temperature controls in high-variance zones to stabilize conditions.
2. Schedule dehumidification during peak humidity periods.
3. Shift high-energy tasks to off-peak time windows.
4. Retrain the model weekly to incorporate new data and seasonal effects.

10 Future Work

Potential expansions include integrating equipment utilization logs, exploring deep-learning architectures for sequence modeling, and adopting AutoML frameworks for automated pipeline optimization.