Appliances Energy Prediction

Nonlinear Regression with Ensemble Methods

Machine Learning Final Project | Mehdi Talebi | February 2026

1. Introduction

This project predicts household appliance energy consumption (Wh per 10-minute interval) using environmental sensor data from a low-energy house in Belgium. The dataset spans ~4.5 months with ~19,700 observations and 25 features including indoor/outdoor temperature, humidity, weather conditions, and time-derived variables.

Energy consumption exhibits inherently nonlinear patterns driven by occupancy thresholds, temperature comfort zones, and time-of-day effects. Linear models cannot capture these dependencies, motivating nonlinear and ensemble regression methods.

Dataset: UCI ML Repository (Candanedo et al., 2017). 19,735 observations, 29 columns, target: Appliances (Wh). No missing values.

2. Methods

EDA Summary

Target is strongly right-skewed (skewness ~3.6), most readings below 100 Wh. Individual feature correlations are weak (max |r| < 0.3). Clear temporal patterns with energy peaks during morning/evening hours.

Preprocessing

- Feature engineering: Extracted hour, day_of_week, month, is_weekend from timestamp
- Dropped: date (after extraction), rv1, rv2 (random noise variables)
- Outlier treatment: IQR-based capping on target variable
- Scaling: StandardScaler fitted on training data only (prevents data leakage)

Models and Hyperparameter Tuning

Six models trained with GridSearchCV (5-fold CV, neg_mean_squared_error scoring):

- Linear Regression (baseline, no tuning)
- Ridge Polynomial Regression (degree=2, alpha=0.1)
- Decision Tree Regressor (max_depth=20, min_samples_leaf=5)
- SVR with RBF kernel (C=100, gamma=0.1, epsilon=0.5)
- Random Forest Regressor (200 trees, min_samples_leaf=2)
- Gradient Boosting Regressor (200 estimators, max_depth=7, lr=0.1)

3. Results

Model Comparison Table

Model	Tr MAE	Te MAE	Tr RMSE	Te RMSE	Tr R2	Te R2
Random Forest	7.02	**14.66**	10.89	**22.11**	0.936	**0.734**
Gradient Boosting	10.70	15.71	14.68	23.16	0.883	0.708
SVR (RBF)	12.29	16.16	22.38	26.44	0.729	0.620
Decision Tree	9.53	17.13	15.62	27.66	0.868	0.584
Linear Regression	26.41	26.42	35.86	35.58	0.304	0.311

Model	Tr MAE	Te MAE	Tr RMSE	Te RMSE	Tr R2	Te R2
Polynomial Ridge	26.00	26.62	35.38	36.15	0.322	0.289

Key Findings

- **Random Forest** achieves best test performance: R2=0.734, MAE=14.66 Wh
- Ensemble methods consistently outperform individual models
- Nonlinear models improve R2 from ~0.31 (linear) to ~0.73 (ensemble)
- Decision Tree shows overfitting (Train R2=0.87 vs Test R2=0.58); Random Forest mitigates this via bagging

Feature Importance: Top predictive features are lights (occupancy proxy), indoor temperatures (T6, T3, T8), and temporal features (hour, day_of_week).

4. Discussion

Error Analysis

The best model struggles with high-consumption spikes. Large-error cases (~5% of test set) are biased toward under-prediction of peak consumption. Residuals show near-zero mean (unbiased) with mild heteroscedasticity at higher predicted values.

Limitations

- Single building: Results specific to one low-energy house in Belgium
- 4.5-month window: Incomplete seasonal coverage
- No explicit occupancy data: Model relies on indirect proxies
- Independence assumption: Ignores temporal autocorrelation

Future Work

- Time-series models (LSTM, GRU) for temporal dependencies
- Real-time occupancy sensors as additional predictors
- Weather forecast integration for anticipatory energy management
- Multi-building generalization with transfer learning

References

[1] Candanedo, L.M., Feldmann, A., and Degemmis, D. (2017). Data driven prediction models of energy use of appliances in a low-energy house. Energy and Buildings, 145, 13-25.
[2] UCI ML Repository: archive.ics.uci.edu/dataset/374
[3] Scikit-learn documentation: scikit-learn.org
[4] Pandas: pandas.pydata.org | Matplotlib: matplotlib.org

AI Usage Statement

Did you use any generative AI tools? No. All code, analysis, and written content were produced independently. External resources consulted: scikit-learn documentation, pandas documentation, and the original dataset paper (Candanedo et al., 2017).