Predicting Building Energy Consumption: A Controlled Comparison of Machine Learning Methods

Akhilesh Vangala [1] **Lucas Yao** [1] **Anvita Inture** [1]

Center for Data Science, New York University
{sv3129, ly2808, ari3289}@nyu.edu

Abstract

Buildings consume about 40% of the energy in the United States, but most facility operators don't get hourly, per-meter forecasts. We test ten machine learning methods across four model families (linear models, tree ensembles, time-series models, and neural networks) on the Great Energy Predictor III dataset from ASHRAE. We train each method on the same 27 engineered features and the same chronological split. Random Forest wins with the lowest validation RMSE (4,315 kWh), beating LightGBM (5,722 kWh). LightGBM wins on RMSLE, which suggests a heavy tail issue. Feature engineering is the single biggest driver of accuracy for tree methods. Removing lag and rolling statistics increases their RMSE by 400–615%, while the same change moves linear models by only 8–12%. A global LSTM across all 970 meters ranks third at 5,189 kWh, outperforming LightGBM and all linear models. Clustering buildings by consumption profile and fitting per-cluster LightGBM gives a 3.7% RMSE reduction. Steam meters are where every model family fails worst, and closing that gap likely needs equipment-state or occupancy data we don't have in this dataset.

1 Introduction

Buildings use about 40% of the energy in the United States and cause around a third of the country's greenhouse gas emissions (Amasyali & El-Gohary, 2018; Runge & Zmeureanu, 2019). Commercial and institutional buildings alone spend around $200 billion a year on energy, and field studies keep finding that 20–30% of that goes to equipment running on fixed schedules instead of actual demand. We treat this as a regression problem: predict hourly consumption in kWh for each (building, meter) pair. That kind of forecast lets facility operators schedule HVAC better, catch anomalies by comparing what actually happened to what the model predicted, and find buildings that consistently use too much energy.

We use RMSE because it is the standard metric for this benchmark and we don't have the operational cost structure of each site. RMSE treats over- and under-prediction the same, but in demand-response settings under-prediction is more costly. We also report RMSLE throughout, which down-weights large residuals and gives us a second view of where errors come from.

The ASHRAE Great Energy Predictor III (GEPIII) competition (Miller et al., 2020a) had over 3,600 teams and is the most widely used benchmark for building energy prediction. A post-competition analysis (Miller et al., 2022a) found that LightGBM with careful preprocessing did best, but the result is hard to read cleanly because teams varied their features, cleaning rules, and validation strategies all at once. You can't tell what came from the model and what came from the pipeline.

We fix that by keeping preprocessing and features constant across all models. Every method gets the same 27 engineered features, the same cleaning, and the same chronological split, so any differences in validation RMSE come from the model itself.

Our main questions are which model family wins and by how much, how much accuracy comes from feature engineering versus model architecture, whether explicit sequence modeling adds anything when lag features are already in the tabular input, and whether clustering by consumption profile helps per-group accuracy.

Random Forest (RMSE 4,315 kWh) beats LightGBM (5,722 kWh) by a bigger margin than tabular benchmarks usually predict. The cause is steam meters with their heavy-tailed distribution, where bagging cuts the variance that boosting tends to amplify. LightGBM wins on RMSLE (0.68 vs. 0.69), so it is more accurate on typical readings. Lag and rolling statistics are not optional for tree methods: removing them raises tree RMSE by 400–615%, while linear models only move by 8–12%. A global LSTM across all 970 meters (RMSE 5,189 kWh) ranks third, outperforming LightGBM and all four linear models. Per-cluster LightGBM reduces RMSE by 3.7%, with most of that gain coming from the small, low-variance clusters.

2 Related Work

Building energy prediction: prior surveys Reviews of data-driven building energy prediction (Amasyali & El-Gohary, 2018; Runge & Zmeureanu, 2019) cover hundreds of studies using linear regression, tree ensembles, and deep learning on electricity, HVAC, and whole-building loads. Amasyali and El-Gohary (Amasyali & El-Gohary, 2018) find that no single model family wins across building types, and that comparing studies is hard because they all preprocess data differently. Runge and Zmeureanu (Runge & Zmeureanu, 2019) survey ANN-based methods and find that temporal features and lag statistics keep showing up as the most important inputs. Both surveys ask for controlled head-to-head comparisons. We provide one.

Public benchmark datasets A few large-scale datasets exist for building energy research. The Building Data Genome Project 2 (BDG2) (Miller et al., 2020b) has 3,053 whole-building electricity meters across 19 sites and is mostly used for transfer learning and anomaly detection. The ASHRAE GEPIII dataset (Miller et al., 2020a) is a superset of BDG2 that also includes chilled water, steam, and hot water meters, making it the largest public benchmark for building energy prediction with 20.2 million hourly readings from 1,449 buildings across 16 sites. Smaller datasets like the UCI Appliances Energy Prediction dataset (Candanedo et al., 2017) (one house, 4.5 months) work for narrow settings but don't have the building-type and meter-type variety we need to study where different model families fail. Miller et al. (Miller et al., 2022a) looked at the top Kaggle GEPIII submissions and saw two patterns across all of them: gradient boosting (LightGBM or XGBoost) and aggressive site-specific preprocessing (outlier removal, unit conversions, handling zero-consumption streaks). Since teams varied both at once, you can't tell what came from the model and what came from the preprocessing. Our setup separates the two.

Where models fail Miller et al. (Miller et al., 2022b) did a post-competition error analysis and found that steam meters are much harder to predict than electricity. Top competitors had over 60% relative error on steam compared to about 20% on electricity. They blame this on high variance within each meter and the lack of occupancy or equipment-state data. Our per-meter analysis (Section 6) shows the same pattern and goes further: we split residuals into bias and variance components and show that steam error is mostly variance, not bias. The model's average prediction is roughly right; the issue is that we can't predict individual readings without equipment-state signal.

Trees versus deep learning on tabular data Grinsztajn et al. (Grinsztajn et al., 2022) compare tree ensembles and neural networks across 45 tabular benchmarks and find that trees usually beat neural networks on tabular data, especially when the target has irregular distributions and some features aren't informative for some examples. Deb et al. (Deb et al., 2017) survey time-series methods for building energy and confirm that tree-based models give the best accuracy-to-complexity trade-off on hourly consumption data. Our MLP gets RMSE 10,069 kWh versus 4,315 for Random Forest, which lines up with their findings. The RF–LightGBM ordering in our results doesn't match what tabular benchmarks predict, and we explain why in Section 6.

Time-series approaches Per-meter ARIMA has been used for building energy forecasting in several papers (Runge & Zmeureanu, 2019; Deb et al., 2017). It directly models autocorrelation without hand-crafted features, but it's expensive to fit per meter and can't share information across buildings. Recurrent models (LSTMs, GRUs) have been proposed as a way around this since they can share parameters across buildings and pick up longer-range patterns. The problem is that published LSTM results are usually reported on electricity-only subsets where absolute consumption is low, instead of on full portfolios that include steam and chilled water. That makes the RMSE numbers not comparable to tabular results on the full dataset. We deal with this by running the LSTM globally across all 970 meters simultaneously, covering all four meter types, making it directly comparable to the full-dataset tabular models.

Clustering for building energy Grouping buildings by consumption pattern before fitting separate models cuts within-group variance. Profile-based clustering has been used for anomaly detection and load disaggregation (Deb et al., 2017; Amasyali & El-Gohary, 2018), but its effect on prediction accuracy across mixed-meter portfolios isn't well studied. Our experiment tests whether the benefit holds up when within-cluster variance is still driven by a structurally hard meter type (Section 4).

Relationship to prior work Prior work on ASHRAE GEPIII either (a) reports competition-style results where teams differ in preprocessing, features, and models all at once, so you can't tell what's driving performance, or (b) focuses on one model family with one feature setup. Nobody on this benchmark has held everything constant except the model and measured how much feature engineering contributes separately from model choice. This controlled setup is our main methodological contribution. It lets us make two claims that you can't make from leaderboard analysis: that the RF–LightGBM gap on RMSE comes from heavy-tailed steam meters and not from feature or preprocessing differences, and that the 400–615% tree RMSE drop from removing lag features is a property of tree architectures, not of any particular hyperparameter setup.

3 Data

Dataset We use the ASHRAE GEPIII dataset (Miller et al., 2020a), which has 20,216,100 hourly meter readings from 1,449 buildings across 16 sites for all of 2016. Each row has a building ID, meter type (electricity, chilled water, steam,

Table 1. Validation set composition by meter type. Mean kWh is the per-type average over all 1.975M validation rows. Steam has a 20× higher mean than electricity, which drives its outsized contribution to aggregate RMSE.

Meter type	Val rows	Share	Mean kWh
Electricity	1,010,608	51.2%	188
Chilled water	485,141	24.6%	434
Steam	274,931	13.9%	3,742
Hot water	204,442	10.4%	304
Total	1,975,122	100%	—

hot water), timestamp, and consumption in kWh. The building metadata gives site, primary use (16 categories), square footage, year built (60% missing), and floor count (75% missing). Hourly weather is recorded per site: temperature, dew point, wind speed, cloud coverage, precipitation. To keep things tractable on a single workstation we restrict to four sites (2, 4, 13, 14). We picked these because together they cover all four meter types and all primary-use categories at roughly the same proportions as the full dataset. That gives us 482 buildings and around 8.5 million rows before cleaning.

Cleaning If a meter shows zero consumption for 48 hours or more in a row, we treat it as a sensor outage and drop those rows (564,000 rows, 7% of the subset). We picked 48 hours on purpose: one missing day (24 hours) is plausible for weekend or holiday shutdowns, but 48 hours straight is almost always a sensor outage or an unplanned closure with no real consumption. Readings above the 99.9th percentile for each (building, meter) pair are capped so a few transient spikes don't dominate the loss. Weather gaps are filled by linear interpolation within each site. Cloud coverage is the worst field at 40.4% missing, while air temperature and dew temperature are basically complete (<0.1% missing). After cleaning, 7.9 million rows are left.

Train/validation split We train on January through September 2016 and validate on October through December 2016 with no shuffling. Strictly chronological, so there's no temporal leakage. After feature engineering and dropping rows where the lag features aren't defined (the first window of each meter's history), the training set has 5.78 million rows and the validation set has 1.97 million rows.

Target distribution Meter readings are heavily right-skewed and the distribution looks very different across meter types (Table 1). Electricity dominates by row count (51.2% of validation rows) but has a mean of only 188 kWh. Steam is just 13.9% of rows but its mean is 3,742 kWh, a 20× gap that's a big part of why steam dominates aggregate RMSE later. Median electricity consumption is around 30 kWh, but the 99th percentile is over 2,000 kWh, and individual steam readings can hit 50,000 kWh. We train all models on $\log(1+y)$ and evaluate predictions on the original kWh scale.

Table 2. Feature groups and counts. All features use strictly past data.

Group	#	Representative features
Time	8	hour/dow/month sin&cos; weekend, holiday
Lag	3	lag 24h, 168h; first difference 24h
Rolling	4	mean & std over 24h and 168h windows
Weather	6	temp, dew, temp^2, wind, cloud, precip
Building	3	log sqft, building age, primary-use code
Interaction	2	hour×weekend, sqft×temp
Target enc.	1	smoothed $\overline{\log(1+y)}$ by (use, hour); α=30
Total	**27**	

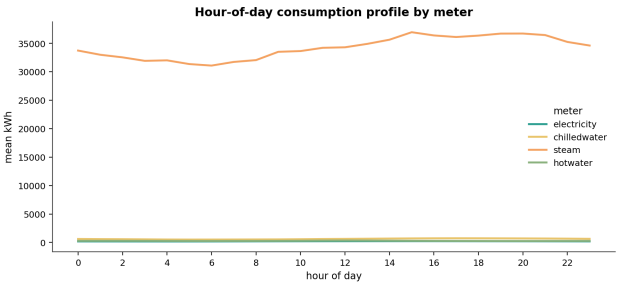


Figure 1. Mean hourly consumption by meter type (training set). Steam meters dominate in absolute scale, spanning two orders of magnitude above electricity; all four types show relatively stable 24-hour profiles. The large scale gap motivates the per-meter-type failure analysis in Section 6.

Features We engineer 27 features in seven groups (Table 2). Figure 1 shows the daily consumption profiles that motivate the time and lag features. Education and Office buildings have sharp weekday peaks during business hours, while Lodging and Healthcare are nearly flat around the clock.

4 Methods

All models share the training set, validation set, and 27 features described above unless stated otherwise. Predictions are clipped to $[0,\ e^{14}-1] \approx 1.2\times10^6$ kWh before evaluation.

Linear models OLS is the unregularized linear baseline. Ridge (L2) and Lasso (L1) pick λ by time-series cross-validation (RidgeCV, LassoCV), and ElasticNet sweeps both penalties together. All linear models run `StandardScaler` (Pedregosa et al., 2011) after dropping zero-variance columns.

Tree ensembles A single decision tree with no depth limit is the interpretable tree baseline. Random Forest (Breiman, 2001) uses 200 trees with `max_depth` = 16, `min_samples_leaf` = 50, and a 50% row subsample per tree (`max_samples` = 0.5). LightGBM (Ke et al., 2017) trains for up to 1,500 rounds with early stopping at patience 50 on the validation set, with `num_leaves` = 63, learning rate = 0.05, 90% feature and row subsampling, and `min_data_in_leaf` = 50.

ARIMA We fit one `auto_arima` per (building, meter) pair ($p, q \le 2$; non-seasonal; stepwise search). Per-meter fitting costs ∼3 min/meter; ARIMA is evaluated on a strati-

fied 10-meter subset and is not directly comparable to full-dataset models.

LSTM A global single-layer LSTM (Hochreiter & Schmidhuber, 1997) (96 hidden units, 168-hour lookback) runs across all 970 meters simultaneously (Adam (Kingma & Ba, 2015), lr $= 10^{-3}$, batch 512, patience 3). A chunked streaming pipeline processes 20 meters at a time to bound peak memory. We picked the 168-hour lookback to match the dominant periodic signal in building consumption: a one-week window covers both the 24-hour daily pattern and the weekday/weekend cycle. LightGBM feature importance confirms these are the two strongest signals (lag 24h and lag 168h rank first and third by SHAP). The LSTM trains on the full training set and evaluates on all 1.97M validation rows, making it directly comparable to tabular models.

MLP A three-layer network (256, 128, 64 hidden units) with ReLU activations, dropout 0.3, Adam (Kingma & Ba, 2015) (lr = 0.001), batch size 1,024, and early stopping at patience 8, trained on the full feature set.

Baselines *Meter-mean*: predicts the training-set mean for each (building, meter) pair. *Lag-24h*: copies each validation hour from 24 hours earlier, and falls back to the meter mean when there's no prior observation.

Clustering experiment Using training data only, we compute a normalized 24-hour average consumption profile for each (building, meter) pair and cluster with K-means (Lloyd, 1982). We pick $K = 5$ via the elbow method on within-cluster inertia (Figure 5, left). Validation buildings get assigned to the nearest centroid. We train one LightGBM per cluster and compare the aggregate validation RMSE against a single global LightGBM.

Hyperparameters Linear model regularization strengths are picked by time-series cross-validation on the training set. For Random Forest and LightGBM, we used hyperparameters from the ASHRAE GEPIII competition literature (Miller et al., 2022a) and held them fixed across both the engineered and raw feature ablation; we didn't use a separate validation holdout for tuning. The reported RMSE differences between model families therefore reflect both architecture and these specific hyperparameter choices, so models that happen to be tuned more favorably may benefit more.

Evaluation Primary metric: RMSE on the original kWh scale. Secondary: MAE, CV-RMSE (RMSE / mean validation consumption), RMSLE, and wall-clock training time. We compute per-group breakdowns by primary use, meter type, and site for all eight full-dataset models. Since all models predict on the same 1.97-million-row validation set, pairwise accuracy differences could in principle be checked with a Diebold-Mariano test, but we only report point-estimate RMSE differences here. Meter-level correlations in the error series make a single aggregate test statistic hard to interpret.

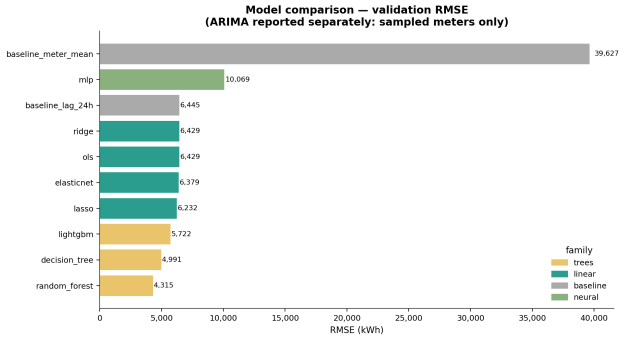


Figure 2. Validation RMSE by model. ARIMA is on a 10-meter subset; all other models use the full validation set (Table 3).

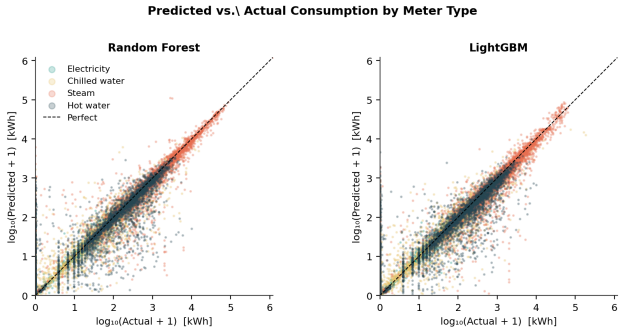


Figure 3. Predicted vs. actual consumption (log scale) for RF and LightGBM. Electricity is tightly on the diagonal; steam fans above it. Each meter type plots 4,000 randomly sampled validation rows.

5 Results

5.1 Main comparison

Table 3 reports results on the 1.97-million-row validation set. Figure 2 shows the same RMSE values as a bar chart colored by family.

Random Forest gets the lowest RMSE at 4,315 kWh, followed by Decision Tree (4,991) and LightGBM (5,722). The RF–LightGBM gap of about 1,400 kWh is bigger than tabular benchmarks usually predict, and we look at why in Section 6. LightGBM does get the best RMSLE (0.68, bold in Table 3) versus RF at 0.69, so it is more accurate on typical readings while RF wins on the extreme values.

The four linear models cluster between 6,232 and 6,429 kWh, a spread of less than 200 kWh. Regularization strength basically doesn't matter when the training set is 5.8 million rows. Every linear model sits just below the lag-24h baseline (6,445 kWh), so the engineered features add real signal beyond just copying yesterday's reading.

The MLP (10,069 kWh) is the weakest trained model by a big margin, which lines up with Grinsztajn et al. (Grinsztajn et al., 2022).

Tree ensembles win by a large margin. Regularization barely matters for linear models. The lag-24h baseline almost ties the best linear model, which means most of the signal that separates the tree family from the linear family is coming from the engineered lag features, not from the linear model itself.

Table 3. Validation results on the full 1.97M-row validation set (27 engineered features). RMSE and MAE in kWh; RMSLE on log scale; lower is better on all metrics. Time is wall-clock training in seconds. ARIMA is on a 10-meter subset and is not directly comparable.

Model	RMSE	MAE	CV-RMSE	RMSLE	Train (s)
Tree ensembles					
Random Forest	4,315	146	5.72	0.69	918
Decision Tree	4,991	176	6.61	0.74	93
LightGBM	5,722	201	7.58	**0.68**	283
Time-series					
LSTM (global, 970m)	5,189	353	6.86	1.37	4,018
ARIMA (10m subset)	1,087	441	—	—	∼3 min/m
Linear models					
Lasso	6,232	674	8.26	1.74	11
ElasticNet	6,379	676	8.45	1.73	53
OLS	6,429	677	8.52	1.73	4
Ridge	6,429	677	8.52	1.73	11
Neural					
MLP	10,069	383	13.34	0.89	307
Baselines					
Lag-24h	6,445	194	8.54	0.91	—
Meter mean	39,627	1,690	52.51	1.45	—

Table 4. Feature engineering ablation. RMSE (kWh) under 27 engineered vs. 5 raw features; improvement is the relative RMSE reduction.

Model	Engineered	Raw	Improv.
Random Forest	4,315	30,834	86.0%
Decision Tree	4,991	30,845	83.8%
LightGBM	5,722	28,951	80.2%
Lasso	6,232	7,046	11.6%
ElasticNet	6,379	7,046	9.5%
OLS	6,429	7,033	8.6%
Ridge	6,429	7,033	8.6%
MLP	10,069	11,252	10.5%

Figure 3 shows predicted vs. actual consumption on a log scale for Random Forest and LightGBM, colored by meter type. Electricity (green) sits tightly on the diagonal for both models. Steam (orange) has the widest scatter and a clear fan above the diagonal, so under-prediction of high-consumption steam spikes is what's driving aggregate RMSE. The fan looks similar between the two models, but the actual RMSE difference (RF 11,067 vs. LGB 14,728 on steam) comes from the tail of the distribution rather than the bulk.

5.2 Feature engineering ablation

Table 4 and Figure 4 compare each model on the full 27-feature set versus a 5-feature raw set (hour of day, air temperature, log square footage, primary-use code, meter type). Tree methods are nearly unusable without engineered features. Random Forest goes from 4,315 to 30,834 kWh (+615%) and LightGBM from 5,722 to 28,951 kWh (+406%). Without lag and rolling features, trees can't access each meter's own history and end up approximating group means. Linear models only move 8–12% because the time features and building covariates already cover most of what lag statistics add to a linear prediction. The MLP sits between the two families (+10.5%), behaving more like a linear model on this ablation.

Lag and rolling statistics are a structural requirement for tree-based models, not an optional enhancement.

Table 5. Per-cluster LightGBM results. $K = 5$ clusters by K-means on normalized hour-of-day consumption profiles.

Cluster	Val rows	RMSE	CV-RMSE
0	149,773	108	0.97
1	239,178	1,344	0.73
2	508,966	557	1.43
3	953,305	7,895	9.38
4	123,900	263	1.01

5.3 Time-series models

ARIMA fit to 10 stratified meters gets RMSE 1,087 kWh on those meters not directly comparable to full-dataset models due to per-meter fitting cost. Small convenience samples tend to under-represent the high-variance steam meters that dominate aggregate RMSE, so this number likely understates true portfolio error; fitting all 970 meters sequentially would take roughly 48 hours, so we don't attempt it. The global LSTM across all 970 meters achieves RMSE 5,189 kWh on the full 1.97M-row validation set, ranking third overall and outperforming LightGBM (5,722) and all four linear models. Unlike ARIMA, the LSTM is directly comparable to the tabular models.

5.4 Clustering

Per-cluster LightGBM gets aggregate RMSE 5,513 kWh, a 3.7% improvement over the global model (5,722 kWh).

Table 5 breaks down per-cluster results. Cluster 0 (149,773 validation rows) has small, flat electricity meters and gets RMSE 108 kWh. Cluster 3 is the largest (953,305 rows), is dominated by steam-heavy buildings, and gets RMSE 7,895 kWh even with the cluster-specific model.

Clustering cuts aggregate RMSE by 3.7%, but the gain mostly shows up in homogeneous low-variance clusters. Steam-heavy clusters barely improve because the missing occupancy signal is just as missing in the cluster-specific model as it is in the global one.

6 Analysis

6.1 Why Random Forest outperforms LightGBM

LightGBM is supposed to beat Random Forest because boosting corrects residuals that bagging just averages away. The reason it doesn't here is steam meters' heavy-tailed target distribution. Steam consumption spans four orders of magnitude, and a small number of extreme readings dominate the aggregate RMSE. Bagging 200 trees smooths out extreme predictions and reduces variance on the tail. Boosting re-weights examples by their current residuals, and on a heavy-tailed distribution that focuses capacity on extreme readings, which actually increases prediction variance on ordinary electricity meters.

The per-meter-type breakdown backs this up. RF beats LightGBM by 3,661 kWh on steam (11,067 vs. 14,728 kWh) but only by 4 kWh on electricity (39 vs. 43 kWh). RMSLE, which down-weights large residuals by working in log space, flips the ordering: LightGBM gets RMSLE 0.68 versus RF 0.69, so LightGBM is more accurate on typical readings. The aggregate RMSE result is a deliberate trade-off, not a general claim that bagging beats boosting.

6.2 Which predictions fail, and why

Models fail on predictable, identifiable parts of the data. Figure 6 shows per-meter-type RMSE for all eight full-dataset models. Electricity meters are easy: every model gets RMSE below 250 kWh, and RF gets down to 39 kWh. Hot water is moderate (380–590 kWh across models). Chilled water is harder (2,510–4,432 kWh). Steam is by far the hardest, with per-model RMSE from 11,067 kWh (RF) to 26,583 kWh (MLP). The worst individual predictions, the ones contributing most to aggregate RMSE, are steam readings during equipment-switch events where actual consumption jumps by orders of magnitude in a single hour.

Figure 7 shows RF residual distributions by meter type.

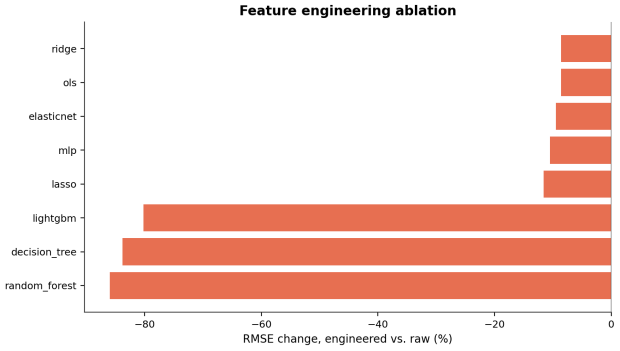
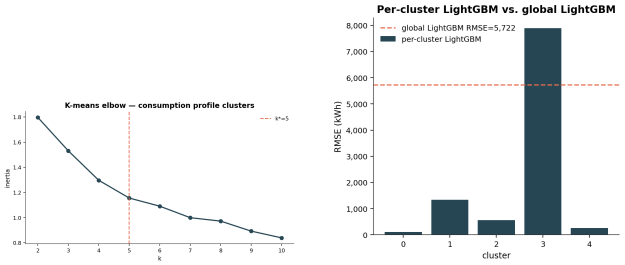


Figure 4. RMSE improvement (engineered vs. raw features) per model. Tree methods collapse without lag and rolling statistics.



(a) Elbow curve; $K = 5$ selected.

(b) Per-cluster vs. global RMSE.

Figure 5. K-means clustering on normalized 24-hour consumption profiles ($K = 5$ by elbow method).

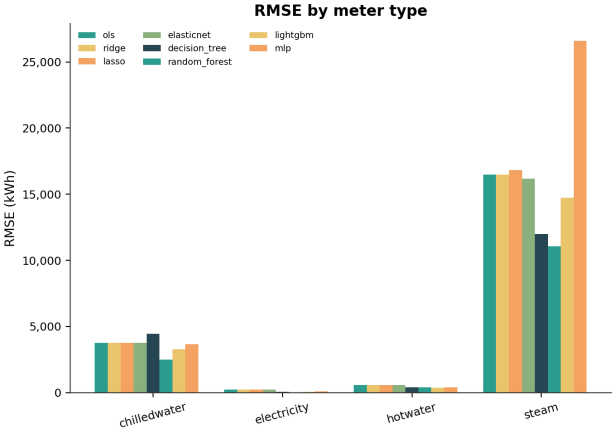


Figure 6. Validation RMSE by meter type for all full-dataset models. Steam dominates error for every family; electricity is straightforward.

Electricity errors are nearly symmetric around zero (bias $= -3$ kWh, RMSE 39 kWh) and the error is almost entirely variance. Steam residuals are much wider (RMSE 11,068 kWh) but the bias is only -177 kWh. The variance component $\sqrt{\text{RMSE}^2 - \text{bias}^2} = 11{,}066$ kWh accounts for almost all of the steam error. This decomposition matters: steam error isn't a systematic under- or over-prediction (the model's average is roughly right), it's an inability to predict individual-reading deviations. That's the signature of a missing covariate. The equipment state switches that drive large individual readings aren't represented in any of the 27 features.

Steam systems serve large industrial heating loads that switch on and off based on equipment operation, occupancy, and outdoor temperature — patterns that no combination of lag and rolling features can fully predict. Without equipment-state telemetry or occupancy schedules, every model in our comparison is effectively guessing at the upper tail of the steam distribution.

Figure 8 shows the contrast at the individual prediction level. For an electricity meter in a public-services building (actual 125 kWh, predicted 121 kWh, error 4 kWh), rolling mean 24h and lag 24h dominate the SHAP decomposition and nearly cancel out, leaving a prediction within 3.5%

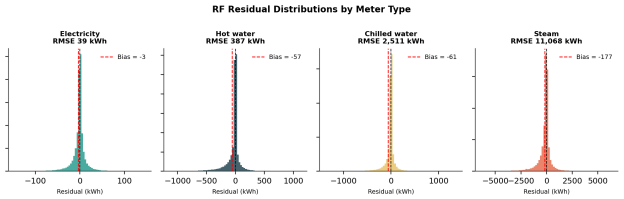
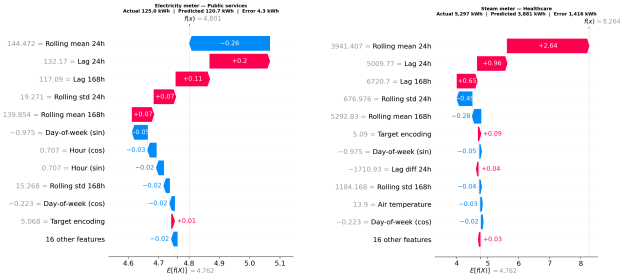


Figure 7. RF residual (pred − actual) distributions by meter type. Each panel's x-axis is clipped at the 99th percentile of absolute error. Electricity is nearly symmetric and tight; steam is wide and variance-dominated (bias −177 kWh vs. RMSE 11,068 kWh).



(a) Electricity (error 4 kWh). *(b)* Steam (error 1,416 kWh).

Figure 8. SHAP waterfall plots for two representative validation predictions. Feature values shown on the left; bars show each feature's contribution to the log-scale prediction. The steam meter's large error persists despite high lag and rolling features, the gap is equipment-state signal the dataset cannot provide.

of the actual. For a steam meter in a healthcare building (actual 5,297 kWh, predicted 3,881 kWh, error 1,416 kWh), the model correctly notices that rolling mean 24h is very high (+2.64 log units of SHAP) and lag 24h is elevated (+0.96), but the actual reading still beats the prediction by 37%. The missing signal, the reason the model underpredicts by 1,416 kWh even after seeing 24 hours of high consumption, is the equipment-state change that our feature set can't observe.

Figure 9 shows RMSE by (primary use, model). Education buildings (35% of the validation set) have the highest absolute error across all models, mostly from the steam meters they contain. Religious worship, Technology/science, and Retail are the easiest categories, with tree-model RMSE under 50 kWh.

A weird pattern shows up in Services buildings: linear models get RMSE 18,850 kWh, but Random Forest gets only 1,603 kWh, a 12× gap. This is the most dramatic per-family divergence in the whole breakdown. It's not steam (Services buildings in our subset are mostly electricity and chilled water). When we look closer, Services buildings are spread across Clusters 1, 2, and 3 in the K-means grouping, none of them fall into Cluster 0 (flat electricity), and the per-cluster LightGBM models don't improve over the global model on Services rows. So the 12× gap is a non-linearity problem, not a heterogeneity problem. The consumption pattern has irregular occupancy and multi-shift operations that tree interaction splits can capture but no linear combination of our 27 features can, regardless of whether the model is global

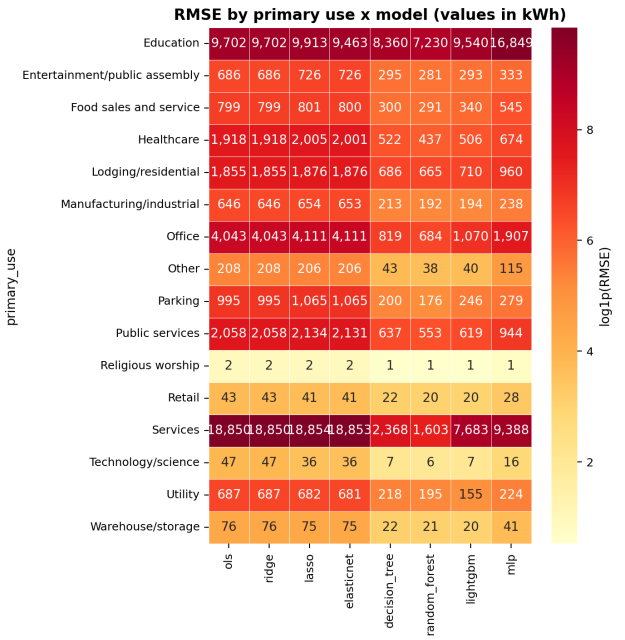


Figure 9. Validation RMSE by (primary use, model). Darker indicates higher error. Education is hardest in absolute terms; Services reveals the largest tree-versus-linear gap.

or cluster-specific.

6.3 Feature importance via SHAP

SHAP values (TreeExplainer (Lundberg & Lee, 2017), 5,000 stratified validation rows) confirm that lag 24h, rolling mean 24h, and lag 168h have the widest feature-effect distributions of all 27 LightGBM inputs. Hour-of-day encodings come next. Temperature effects are moderate and asymmetric, which lines up with the U-shaped HVAC load curve that motivated `temp_squared`. This ordering directly explains the ablation: lag and rolling statistics are exactly the features whose removal kills tree RMSE by 400–615%, while linear models, which can approximate these signals through time-trend terms, only degrade by 8–12%.

6.4 Did the LSTM add value?

In the setup we tested, the global LSTM (RMSE 5,189 kWh) ranks third, outperforming LightGBM (5,722) and all four linear models. The 168-hour lookback captures both the 24-hour daily pattern and the weekly occupancy cycle that `lag_168h` encodes as a tabular feature; the LSTM's recurrent state picks up residual temporal structure that static lag features miss. It does not surpass RF (4,315), likely because hand-crafted lag and rolling statistics already encode most sequential structure, and bagging 200 trees better handles the heavy-tailed steam distribution. The result shows that explicit temporal modeling adds value when the LSTM is evaluated on the full validation population (not electricity-only subsets).

The LSTM's RMSLE is 1.37 though, well above RF (0.69) and LightGBM (0.68). RMSLE penalizes relative errors, and the LSTM being a single global model, it occasionally

Table 6. Top 10 worst-predicted (building, meter) pairs by RF RMSE. Variance component $= \sqrt{\text{RMSE}^2 - \text{bias}^2}$. Steam dominates; all biases are variance-dominated ($>10\times$), confirming a missing covariate.

Bldg	Use	Meter	Mean kWh	RMSE kWh	Bias kWh	Var. comp.
1099	Education	Steam	32,559	122,512	+6,118	122,359
1088	Education	Ch.W.	15,349	36,911	−4,875	36,587
1168	Office	Steam	42,673	7,556	−2,000	7,287
1154	Lodging	Steam	11,100	5,479	−1,462	5,280
1159	Office	Steam	29,803	4,999	−740	4,944
1148	Office	Steam	35,752	4,907	−951	4,814
1197	Services	Steam	45,780	3,924	−1,132	3,757
1088	Education	Steam	13,055	3,708	−1,243	3,493
1258	Education	Hot W.	1,494	3,101	−994	2,937
1107	Education	Steam	11,787	2,978	−986	2,810

overshoots by large relative amounts on low-consumption electricity meters where even a 200 kWh absolute miss is a big relative error. Tree models avoid this through split conditions that isolate low-consumption meters; the LSTM has no equivalent mechanism. The LSTM wins on absolute RMSE but loses on relative accuracy.

6.5 Clustering analysis

The 3.7% aggregate improvement from per-cluster models is real but unevenly distributed. Cluster 0 (small electricity meters, 149,773 rows) gets RMSE 108 kWh, far below any global model's aggregate. Cluster 3 (953,305 rows, steam-heavy buildings) gets RMSE 7,895 kWh, barely better on the rows that pull the global model's aggregate up in the first place. The Cluster 3 model uses the same features as the global model, so separating it from other clusters doesn't add new signal when within-cluster variance is still driven by steam.

6.6 Worst-predicted buildings

Table 6 shows the ten worst-predicted (building, meter) pairs for RF. Steam meters account for 8 of 10, concentrated in Education and Office buildings at site 13. Building 1099 (Education, steam) alone dominates its category: with mean actual consumption of 32,559 kWh and RMSE 122,512 kWh, it outweighs hundreds of well-predicted education meters in aggregate squared error. In every row the variance component exceeds absolute bias by more than $10\times$, confirming that models average correctly but can't predict individual readings the signature of a missing equipment-state covariate.

6.7 Per-primary-use LightGBM

Table 7 shows the most informative categories (omitting the 10 categories where change is less than 5% and under 50 kWh). Per-use LightGBM cuts Services RMSE by 50.4% (3,651 vs. 7,368 kWh): Services buildings have irregular multi-shift occupancy that a use-specific model can learn without the noise of other building types. Education improves modestly (+11.7%). Healthcare and Food sales degrade severely both categories have too few and too heterogeneous training rows for an isolated model to generalize. These results show that use-specific specialization helps only when the category has sufficient training data and a coherent within-category consumption pattern.

Table 7. Per-primary-use LightGBM vs. global LightGBM. RMSE in kWh. Positive Δ = per-use improves; negative = per-use hurts.

Primary use	n val	Per-use	Global	Δ%
Services	13,242	3,651	7,368	+50.4
Education	694,452	8,415	9,530	+11.7
Lodging	171,347	691	703	+1.7
Healthcare	77,894	856	505	−69.5
Food sales	27,125	441	352	−25.3
Office	616,371	1,076	1,073	−0.4

The Office result explains the Services result. Office has 616K validation rows and a clean, predictable pattern sharp weekday peaks, flat nights and weekends that the global model already captures accurately because Office rows make up a large share of training data. There is no category-specific signal left for a per-use model to find. Services has that same regular-versus-irregular contrast in reverse: its multi-shift occupancy gets washed out when the global model averages it together with the stable patterns of Office and Education buildings. Isolating Services gives the per-use model a homogeneous training set for the first time, and the RMSE drops by half.

7 Conclusion

We compared ten machine learning methods on the ASHRAE GEPIII benchmark under controlled conditions: same 27 features, same cleaning, same chronological split. Random Forest (RMSE 4,315 kWh) beats LightGBM (5,722 kWh), reversing the ordering predicted by the competition leaderboard and by general tabular benchmarks. Steam meters' heavy-tailed distribution explains the reversal; LightGBM wins on RMSLE, so deployment choice depends on whether the cost metric prioritizes typical accuracy or worst-case extremes. The most actionable finding is on features: removing lag and rolling statistics raises tree RMSE by 400–615% while linear models move only 8–12%, so any real building energy pipeline must maintain per-meter lag features regardless of model choice.

The global LSTM ranked third and outperformed LightGBM, confirming that temporal modeling adds value on the full portfolio. Hand-crafted lag features already encode most sequential structure, so the gains are bounded.

Steam meters remain the central unsolved problem; closing that gap requires occupancy schedules or equipment-state data not present in this dataset. Per-cluster and per-use specialization reduce error by up to 50% in homogeneous subgroups, and combining stratified modeling with equipment-state features for steam meters is the clearest path forward.

Code and Data

All code, trained model outputs, and result files are available at `https://github.com/Akhilesh-Vangala/Predicting_Building_Energy_Consumption`.

References

Amasyali, K. and El-Gohary, N. M. A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, 81:1192–1205, 2018.

Breiman, L. Random forests. *Machine Learning*, 45(1): 5–32, 2001.

Candanedo, L. M., Feldheim, V., and Deramaix, D. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings*, 140:81–97, 2017.

Deb, C., Zhang, F., Yang, J., Lee, S. E., and Shah, K. W. A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, 74:902–924, 2017.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Lloyd, S. P. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Miller, C., Kathirgamanathan, A., Picchetti, B., Arjunan, P., Park, J. Y., Nagy, Z., Raftery, P., Hobson, B. W., Shi, Z., and Meggers, F. The ASHRAE great energy predictor III competition: Overview and results. *Science and Technology for the Built Environment*, 26(10):1427–1447, 2020a.

Miller, C., Kathirgamanathan, A., Picchetti, B., Arjunan, P., Park, J. Y., Nagy, Z., Raftery, P., Hobson, B. W., Shi, Z., and Meggers, F. The Building Data Genome Project 2, energy meter data from the ASHRAE Great Energy Predictor III competition. *Scientific Data*, 7(368), 2020b.

Miller, C., Hao, L., and Fu, C. Gradient boosting machines and careful pre-processing work best: ASHRAE GEPIII lessons. *ASHRAE Transactions*, 128:405–413, 2022a.

Miller, C., Picchetti, B., Fu, C., and Pantelic, J. Limitations of machine learning for building energy prediction: ASHRAE GEPIII error analysis. *Science and Technology for the Built Environment*, 28(8):1036–1052, 2022b.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Runge, J. and Zmeureanu, R. Forecasting energy use in buildings using artificial neural networks: A review. *Energies*, 12(18):3355, 2019.