UK Energy Usage 2024: Machine Learning and Data Analysis

Summary report

Date: 2026-01-04

Executive summary

This report presents an end-to-end supervised learning workflow to predict hourly UK energy usage in 2024 using time, region, and basic weather features. Exploratory analysis shows strong annual seasonality and a stable intraday cycle. After feature engineering (time decompositions, cyclic encodings, and lag/rolling usage predictors), a Linear Regression baseline generalises extremely well on a chronological hold-out test set (RMSE ≈ 20.78 kWh, MAE ≈ 16.64 kWh, R^2 ≈ 0.973). A Random Forest model is competitive but slightly worse under the chosen configuration (RMSE ≈ 21.26 kWh). Results indicate that demand is primarily driven by time-of-day, time-of-year, and recent usage, while weather and region add minimal incremental signal in this dataset.

1. Problem statement and objectives

Goal: forecast hourly energy usage (kWh) from observed covariates. The analysis follows a standard machine-learning project pipeline—problem definition, EDA, preprocessing, feature engineering, model selection, training, and evaluation with regression metrics (RMSE, MAE, R^2).

2. Dataset and context

The dataset contains 70,272 hourly observations across eight UK regions for the 2024 calendar year (8 regions × 24 hours × 366 days). Fields include timestamp, region, energy_usage_kWh, temperature_C, humidity, and weather_condition.

3. Methodology

3.1 Preprocessing and data representation

Timestamps were parsed into datetime format and used as the index. Calendar/clock features were extracted (hour, weekday, month, day_of_year, is_weekend). Categorical variables (region, weather_condition) were one-hot encoded. Rows affected by lag construction were dropped to ensure complete feature vectors.

3.2 Feature engineering for periodicity and autocorrelation

To reflect known structure in demand time series, features were engineered to capture periodic behaviour and short-term persistence. Lagged usage variables (energy_lag_1h, energy_lag_24h) and a 24-hour rolling mean (energy_roll_avg_24h) model autoregressive effects. Cyclic variables were encoded with sine/cosine transforms (hour_sin/hour_cos; month_sin/month_cos) to avoid discontinuities at cycle boundaries.

3.3 Train-test split, scaling, and models

A chronological split was used (first 80% train, last 20% test) to mimic forecasting and reduce leakage. Standardisation (StandardScaler) was applied for Linear Regression only; Random Forest was trained on unscaled features. Models compared: (i) Linear Regression baseline (interpretable, low-variance) and (ii) Random Forest Regressor (non-linear, higher capacity).

3.4 Metrics

Evaluation uses RMSE (penalises large errors; kWh units), MAE (average magnitude of error; kWh), and R^2 (variance explained relative to a mean baseline).

4. Exploratory findings

EDA highlights two dominant patterns: annual seasonality (higher winter demand, lower late-summer demand) and a consistent intraday cycle. Weather variables show weak correlation with energy usage in this dataset, motivating an emphasis on time-based and lagged predictors.

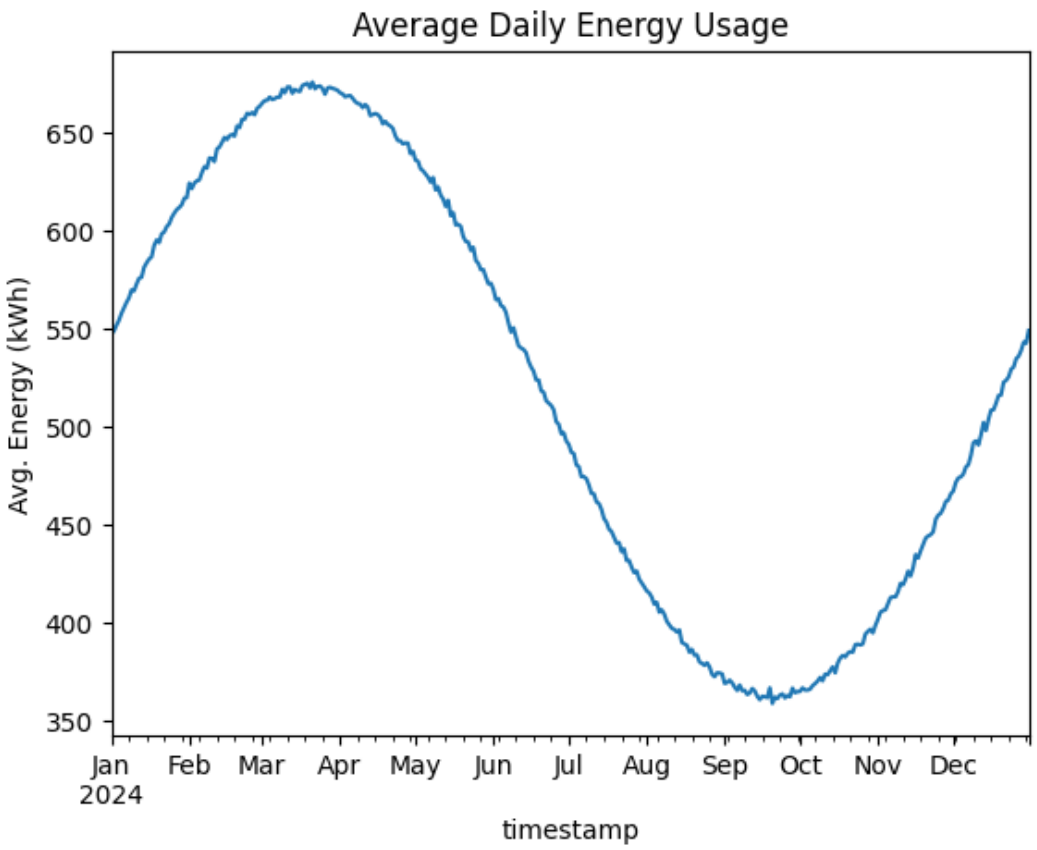

Figure: Average daily energy usage across 2024 (seasonality).

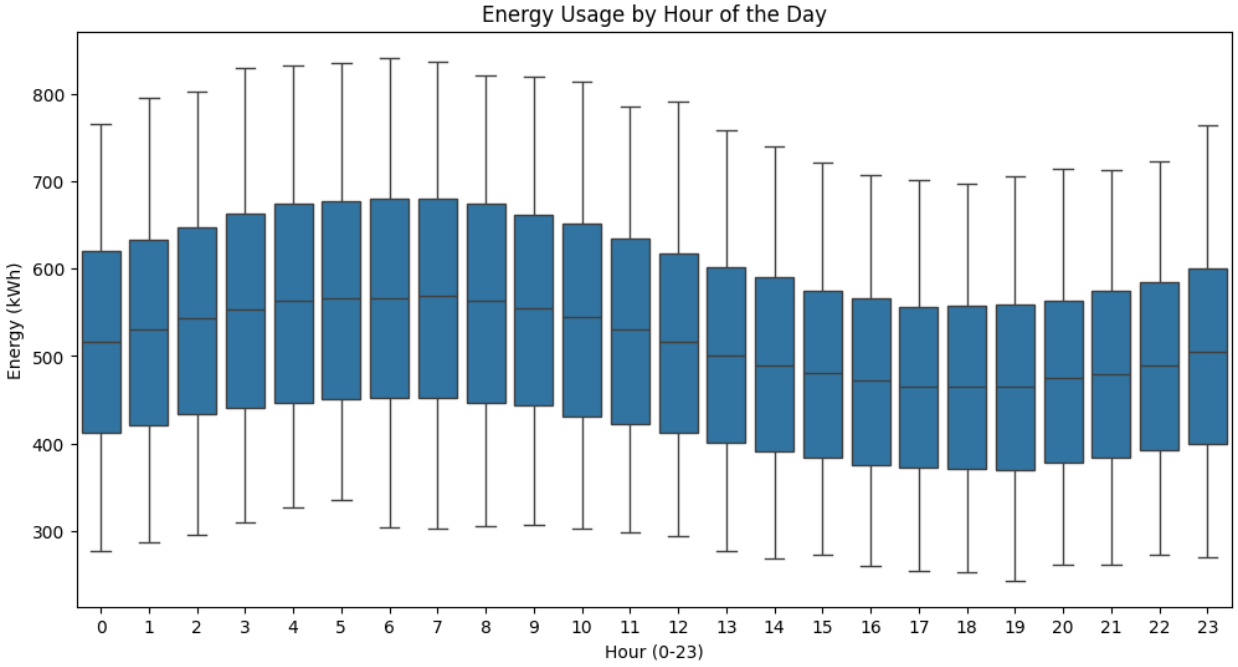

Figure: Energy usage distribution by hour-of-day (intraday cycle).

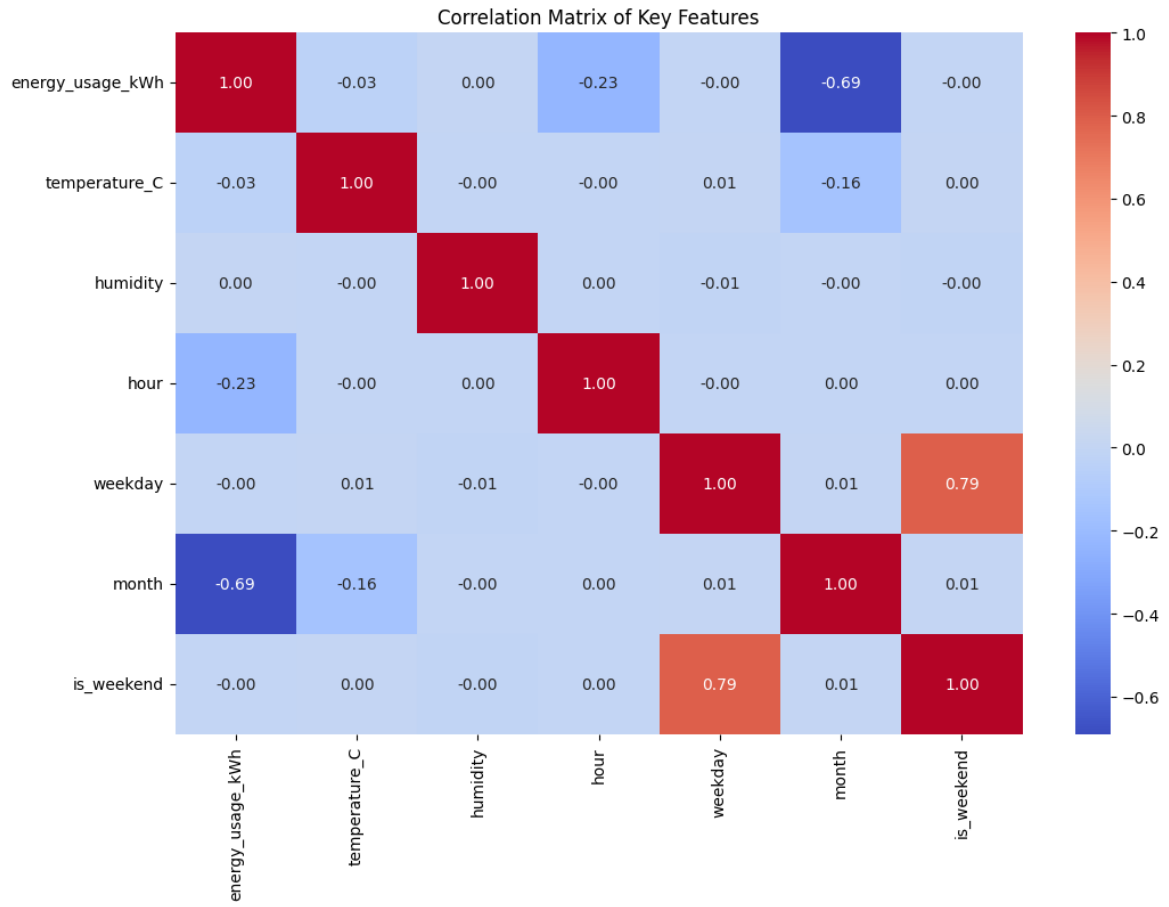

Figure: Correlation matrix for key numeric variables; month shows strong association with energy usage (seasonality).

5. Results

Both models achieve strong predictive performance on the chronological test window. Linear Regression slightly outperforms Random Forest, suggesting that engineered features capture most predictable structure and that additional non-linear capacity provides limited incremental benefit here.

Model	RMSE (kWh)	MAE (kWh)	R^2
Linear Regression	20.7849	16.6418	0.9728
Random Forest	21.2562	17.0101	0.9716

Table 1: Test-set performance comparison.

5.1 Drivers of demand (interpretability)

Linear Regression coefficients indicate that demand is driven primarily by recent usage and time-based cyclic structure. The 24-hour rolling average and lagged usage features dominate, followed by hour_sin/hour_cos and month_sin/month_cos. Weather and region indicators contribute comparatively little signal.

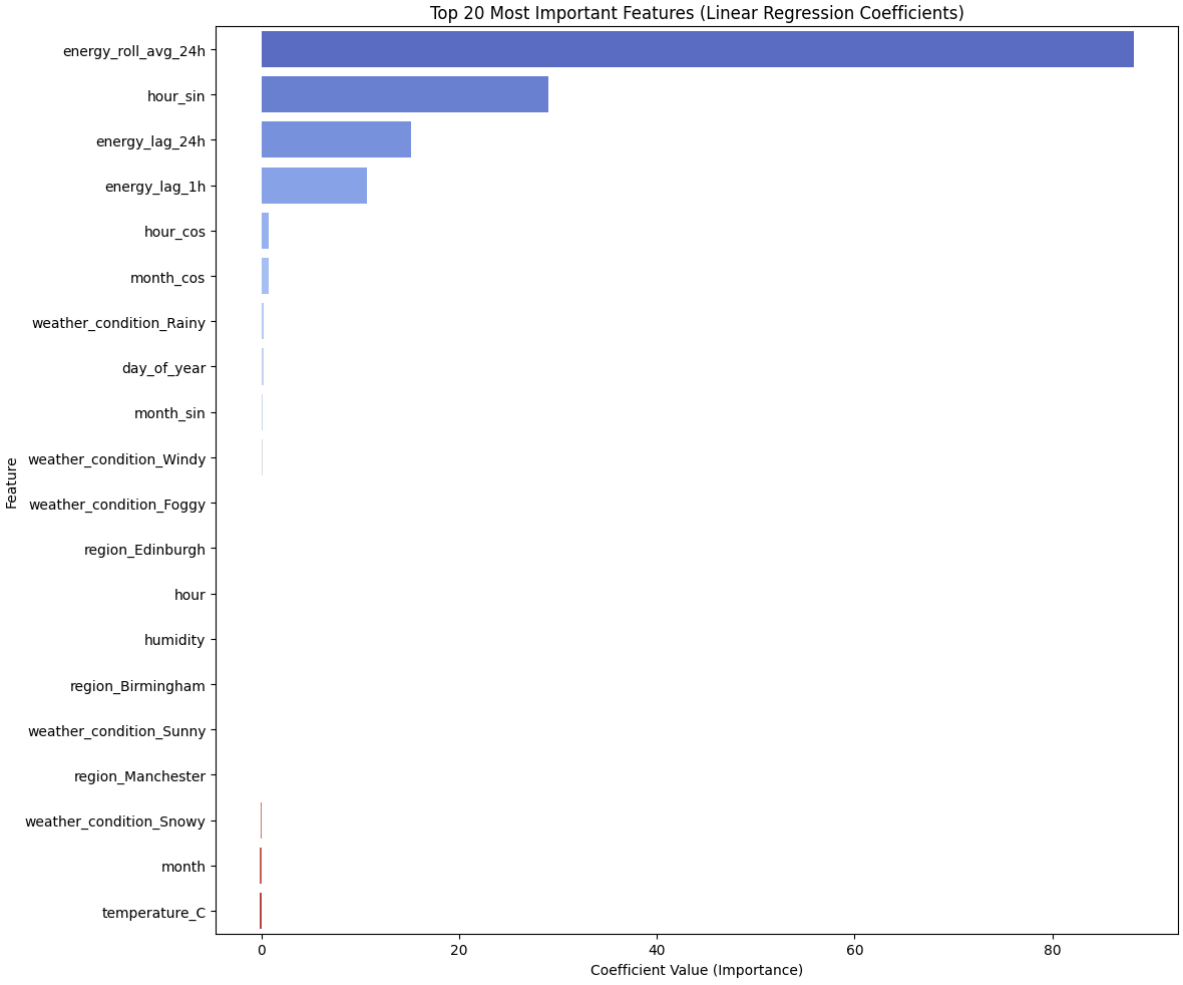

Figure: Top Linear Regression coefficients (magnitude): rolling/lag and cyclic time features dominate.

5.2 Visual validation

A one-week overlay of actual vs predicted values shows that the model captures the primary oscillations while smoothing short-lived spikes, consistent with a squared-error objective and the use of rolling statistics.

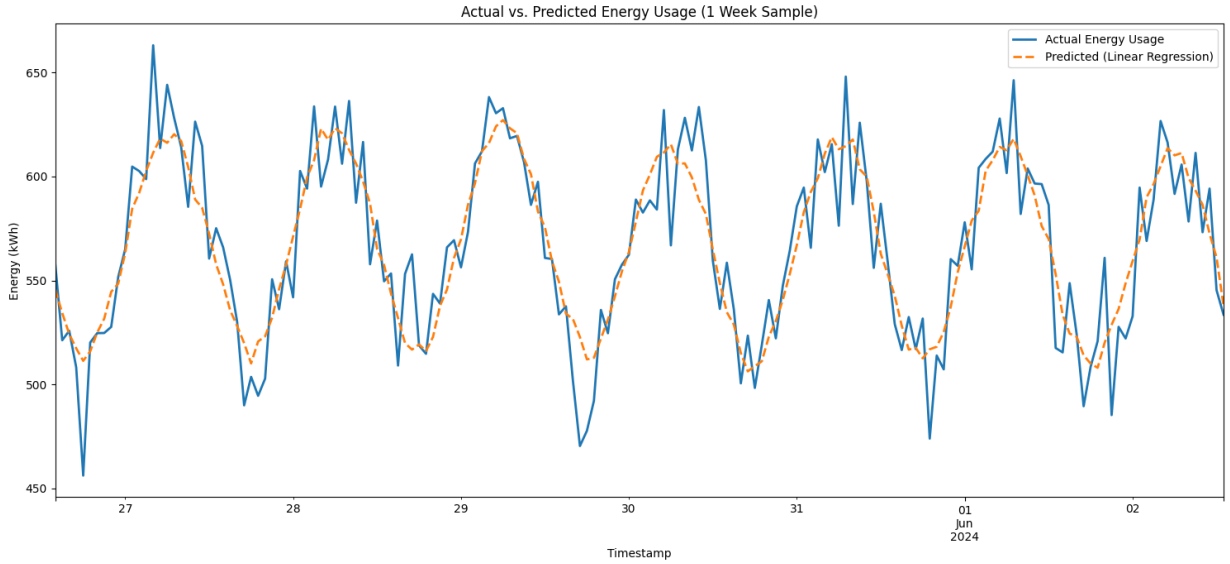

Figure: Actual vs predicted energy usage for a one-week sample in the test window.

6. Discussion and recommendations

The project demonstrates that strong baselines are achievable when feature engineering encodes domain structure (seasonality, periodicity, and autocorrelation). Given the modest uplift from Random Forest, future modelling should focus on (i) richer exogenous variables (holidays, tariffs, daylight hours), (ii) rolling-window evaluation for time series, and (iii) boosted tree models (e.g., XGBoost/LightGBM) with early stopping, which often outperform Random Forests on tabular data.

7. Conclusion

Energy usage in this dataset is highly predictable from time-of-day, time-of-year, and recent usage. A well-featured Linear Regression model achieves RMSE ≈ 20.78 kWh and R^2 ≈ 0.973 on a chronological hold-out, providing an accurate and interpretable forecasting baseline.