A Hybrid Two-Stage Ensemble Framework for Detecting and Quantifying Energy Flexibility in Buildings: A Solution to the FlexTrack Challenge

Daglox Kankwanda
Email: dagloxkankwanda@gmail.com

***Abstract*—Accurate Measurement and Verification (M&V) of demand response (DR) is essential for integrating flexible building loads into the power grid. Addressing the FlexTrack 2025 Challenge, this paper details a hybrid two-stage ensemble framework to classify DR events and quantify their energy impact. The first stage uses a gradient boosting ensemble to predict the Demand Response Flag, identifying the building's operational state. This classification is then fed as a key feature into a second-stage regression ensemble that estimates the Demand Response Capacity. The solution's novelty lies in its hierarchical structure, which combines a general-purpose global model with specialized models trained on data-driven site archetypes for robust generalization. This methodology is underpinned by extensive feature engineering to capture complex temporal and weather-related dynamics. On the private test set, the solution achieved a Geometric-Mean Score of 0.618 for classification and a normalized Mean Absolute Error (nMAE) of 0.991 and normalized Root Mean Square Error (nRMSE) of 1.223 for regression. These results demonstrate the effectiveness of a decoupled, multi-model approach in tackling the complex challenge of DR baselining and provide a scalable framework for automated M&V systems.**

***Index Terms*—Demand Response, Machine Learning, Building Energy Flexibility, Ensemble Methods, Time-Series Forecasting, Measurement and Verification**

1. Introduction

1.1. Quantifying Building Energy Flexibility: The Core Challenge

As power grids transition towards higher penetrations of intermittent renewable energy sources, the ability of buildings to provide demand-side flexibility becomes paramount. Demand Response (DR) transforms buildings from passive energy consumers into active grid assets capable of modulating their load to alleviate grid stress, absorb excess generation, or provide ancillary services [1]. However, for this flexibility to be a monetizable and reliable grid resource, its activation and magnitude must be accurately measured and verified (M&V).

The core challenge of M&V lies in estimating the counterfactual baseline: what would the building's energy consumption have been in the absence of a DR event? The difference between this unobserved baseline and the actual, measured power consumption during the event constitutes the ”Demand Response Capacity,” a concept visually defined in Figure 1. The FlexTrack Challenge 2025 provides a unique opportunity to address these M&V challenges using a novel digital twin-based approach. The competition frames the problem as a dual machine learning task using 15-minute resolution time-series data, requiring participants to develop models that can:

1) **Classify Operational State:** Predict the 'Demand Response Flag', a categorical variable indicating a load decrease (-1), a load increase (+1), or normal operation (0).
2) **Quantify Energy Deviation:** Predict the 'Demand Response Capacity' in kW, which is the precise deviation from the counterfactual baseline during a DR event.

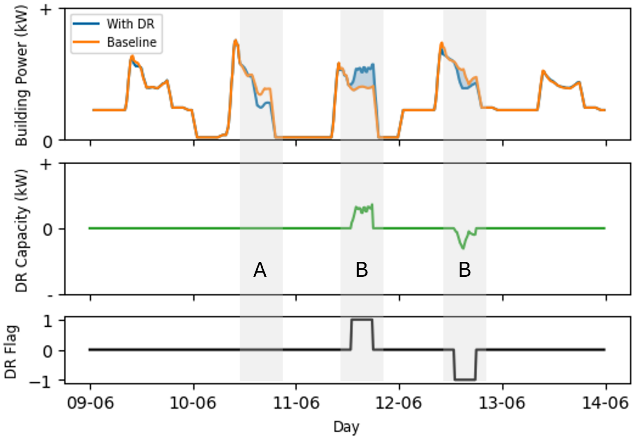

Figure 1. Conceptual illustration of a Demand Response event. The blue line represents the actual building power, which deviates from the counterfactual baseline (dashed orange line) during the DR event (shaded area), creating a measurable 'Demand Response Capacity'.

1.2. Digital Twins and Data-Driven M&V

The FlexTrack Challenge 2025 directly addresses this M&V problem by providing a unique dataset generated from **Digital Twins** of commercial office buildings. These Digital Twins are high-fidelity, physics-based simulation models that dynamically respond to external weather conditions, internal load schedules, and explicit DR signals. As illustrated in Figure 2, a DR flag is input into the Digital Twin, which then modulates its HVAC system's temperature setpoints to produce a corresponding change in the building's power consumption. This provides a synthetic yet physically-grounded dataset with perfect ground truth, creating an ideal testbed for M&V algorithms.

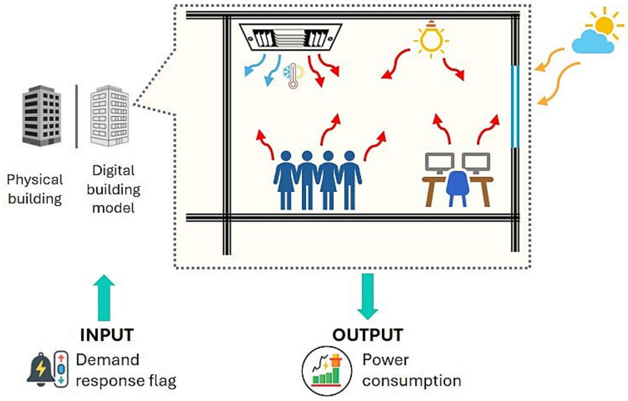

Figure 2. Conceptual illustration of the synthetic data generation process using a Digital Twin, as described in the FlexTrack Challenge. An external DR signal prompts a change in the HVAC setpoint within the physics-based model, resulting in an observable change in the building's power output.

1.3. A Review of AI in Building Energy Modeling and Estimation

The application of Artificial Intelligence (AI) to model and analyze building energy consumption is a mature field of research that provides a strong foundation for the approach taken in this study. The core problem, whether for forecasting or for hindcasting as required here, involves learning the complex, non-linear relationships between weather, operational schedules, and building power consumption. The literature offers several key principles that are directly applicable to the task of detecting and quantifying DR events.

Early studies established the effectiveness of various machine learning algorithms in this domain. Dong et al. [3] and Li et al. [2] were among the first to demonstrate the capability of Support Vector Machines (SVMs) to effectively model the relationship between meteorological inputs and building energy use on both monthly and hourly time scales. Their work showed that these methods could capture complex behaviors that are difficult to represent with simpler linear models. Similarly, the work of Li et al. [11] further confirmed this, showing that both SVM and General Regression Neural Networks (GRNN) delivered higher accuracy for annual energy estimation than standard back-propagation networks.

As the field evolved, two key trends emerged that are central to this work:

1) **The Rise of Ensemble Methods:** Research consistently showed that ensemble models, which combine the outputs of multiple individual learners, deliver superior performance and robustness by reducing both model bias and variance. The work of Chou & Bui [5], for instance, found that an ensemble of SVR and ANN models yielded the best results for estimating heating and cooling loads.
2) **The Importance of Dynamic and Transitional Context:** Static models that only consider instantaneous inputs are often insufficient for capturing building thermal dynamics. Mihalakakou et al. [8] highlighted the importance of incorporating time-lagged input variables to provide the model with a crucial short-term memory of recent system states. This was taken a step further by Paudel et al. [6], who introduced a “pseudo dynamic transitional model.” Their work demonstrated that explicitly modeling the transitions between different operational states (e.g., from an unoccupied to an occupied heating schedule) significantly improved model performance over static approaches. This concept is directly analogous to the FlexTrack challenge, which requires the detection of transitions between a “normal” operational state and a “demand response” state.

These studies provide powerful methodological tools, yet they also highlight persistent challenges in applying these techniques to DR event detection and quantification. First, most prior work focuses on estimating absolute energy consumption rather than quantifying deviations from a baseline during DR events. Second, the challenge of generalizing to completely unseen buildings without any physical metadata remains largely unaddressed. Finally, the specific problem of simultaneously detecting DR events and quantifying their impact requires a more sophisticated approach than traditional single-task models. Building on these insights, the next subsection outlines our contributions to address these gaps.

1.4. Contribution

Synthesizing these insights, the solution approaches the FlexTrack challenge with a hierarchical, multi-stage framework. The approach adopts the transitional modeling philosophy of Paudel et al. [6] by first classifying the building's operational state before attempting to quantify the energy deviation. The framework employs powerful gradient boosting ensembles, the modern evolution of the AI techniques discussed in the literature. Most critically, to solve the generalization problem, a novel, data-driven site archetyping method is introduced, providing a practical implementation of the proxy-variable concept from Magalhães et al. [7]. This comprehensive approach, detailed in the following sections, enables the development of a highly accurate and generalizable solution for DR event verification.

2. Methodology

The solution is architected as a comprehensive pipeline comprising four key stages: (1) Data-Driven Site Archetyping, (2) Extensive Feature Engineering, (3) a Two-Stage

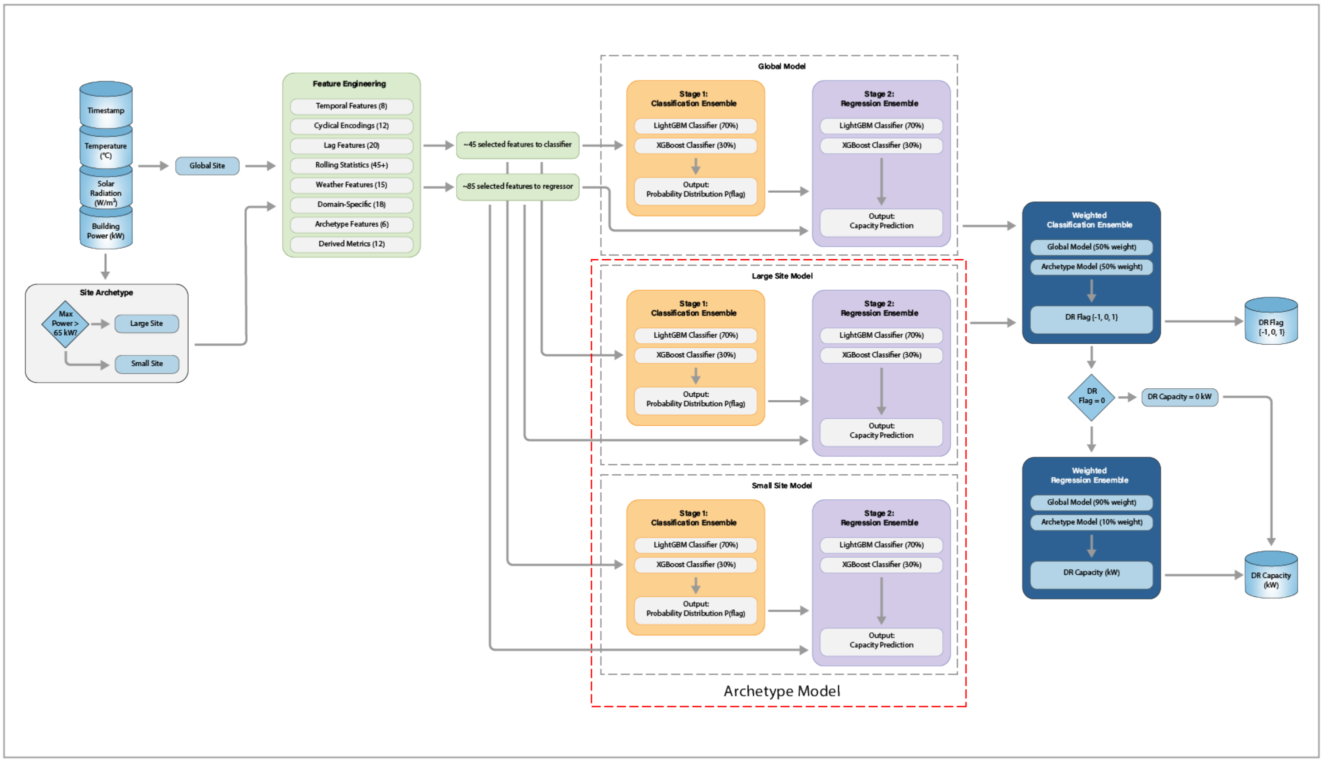
Timestamp
Temperature (°C)
Solar Radiation (W/m²)
Building Power (kW)
Global Site
Site Archetype
Max Power > 65 kW?
Large Site
Small Site
Feature Engineering
Temporal Features (8)
Cyclical Encodings (12)
Lag Features (20)
Rolling Statistics (45+)
Weather Features (15)
Domain-Specific (18)
Archetype Features (6)
Derived Metrics (12)
~45 selected features to classifier
~85 selected features to regressor
Global Model
Stage 1: Classification Ensemble
LightGBM Classifier (70%)
XGBoost Classifier (30%)
Output: Probability Distribution P(flag)
Stage 2: Regression Ensemble
LightGBM Classifier (70%)
XGBoost Classifier (30%)
Output: Capacity Prediction
Large Site Model
Stage 1: Classification Ensemble
LightGBM Classifier (70%)
XGBoost Classifier (30%)
Output: Probability Distribution P(flag)
Stage 2: Regression Ensemble
LightGBM Classifier (70%)
XGBoost Classifier (30%)
Output: Capacity Prediction
Small Site Model
Stage 1: Classification Ensemble
LightGBM Classifier (70%)
XGBoost Classifier (30%)
Output: Probability Distribution P(flag)
Stage 2: Regression Ensemble
LightGBM Classifier (70%)
XGBoost Classifier (30%)
Output: Capacity Prediction
Archetype Model
Weighted Classification Ensemble
Global Model (50% weight)
Archetype Model (50% weight)
DR Flag {-1, 0, 1}
DR Flag {-1, 0, 1}
DR Flag = 0
DR Capacity = 0 kW
Weighted Regression Ensemble
Global Model (90% weight)
Archetype Model (10% weight)
DR Capacity (kW)
DR Capacity (kW)

Figure 3. High-level architecture of the proposed hybrid two-stage ensemble framework. Raw data undergoes feature engineering. In parallel, a Global Model is trained on all sites, while specialized Archetype Models are trained on site subsets. The predictions are then ensembled for the final output.

Classification and Regression Pipeline, and (4) a Hybrid Ensembling Strategy. This modular design allows for specialization at each step, maximizing overall performance and robustness. Figure 3 illustrates the high-level architecture of the framework, showing how raw data flows through feature engineering and parallel model training paths before being combined in the final ensemble.

2.1. Data-Driven Site Archetyping for Generalization

A primary challenge of the competition is to develop a model that performs well on unseen sites. To address this, a concept inspired by Magalhães et al. [7] was adopted, who used theoretical energy certificate data as a proxy for physical characteristics. As no such data was provided, an *empirical* proxy was created from the training data itself.

The mean ‘Building Power kW’ was calculated for each site across the entire training dataset. This simple metric serves as a robust indicator of the building’s overall size, occupancy, and equipment load. Based on the distribution of these mean power values, a threshold of 60 kW was established to categorize each site into one of two archetypes: ‘small’ or ‘large’. This ‘site archetype’ was then one-hot encoded and included as a feature, with interaction terms (e.g., archetype-specific temperature and power interactions) to allow models to learn distinct behaviors. This process is conceptually similar to the work of Fu et al. [10], who showed the value of modeling different building systems (and by extension, building types) separately.

2.2. Feature Engineering

A rich and informative feature set is the cornerstone of any successful machine learning model for time-series forecasting. The feature engineering process was extensive, designed to extract multi-scale temporal patterns, weather dependencies, and building-specific operational characteristics. The final model utilized over 100 engineered features, carefully selected through correlation analysis and empirical validation.

2.2.1. Temporal and Cyclical Features

Comprehensive time-based features were extracted including hour, day of week, day of year, month, quarter, and week of year. To handle the cyclical nature of these features, sinusoidal transformations were applied:

$$X_{\text{sin}} = \sin\left(\frac{2\pi \cdot X}{P}\right), \quad X_{\text{cos}} = \cos\left(\frac{2\pi \cdot X}{P}\right) \tag{1}$$

where P is the period (e.g., 24 for hour, 7 for day of week). This encoding proved particularly effective for capturing daily and weekly patterns in energy consumption.

TABLE 1. SUMMARY OF KEY ENGINEERED FEATURE CATEGORIES AND THEIR IMPACT

Category	Description and Examples	Count
Temporal	Hour, day of week, month, quarter, day of year. Captures fundamental operational schedules.	8
Cyclical	Sin/Cos transformations (hour sinusoidal, month cosine, day of week sinusoidal). Represents circular nature of time.	12
Lag Features	Past values at intervals: 1, 2, 4, 8, 16, 24, 48, 56, 96, 144 steps for power and temperature.	20
Rolling Statistics	Mean, std, min, max, median, skew, kurtosis over windows (4–672 hours). Most impactful: 24h and 96h windows.	45+
Weather-Based	Heating/cooling degrees, temperature bins (cold/mild/warm/hot), radiation interactions.	15
Domain-Specific	Business hours, peak hours (13–20h), transitions (work/non-work), occupancy states.	18
Archetype	Site classification (small: <60 kW, large: ≥60 kW) with interaction terms.	6
Derived Metrics	Power percentiles, z-scores, exponential weighted means, percentage changes.	12

2.2.2. Dynamic Context Features

The lag and rolling window features provided crucial temporal context:

- **Strategic Lag Selection:** Rather than exhaustive lag creation, strategically important intervals were selected:
  - Immediate (1–4 steps)
  - Hourly (8, 16 steps)
  - Daily (24, 96 steps)
  - Weekly (144, 672 steps)

  The 96-step lag (24 hours) proved particularly predictive.
- **Multi-Scale Rolling Statistics:** Rolling statistics were computed over windows ranging from 1 hour (4 steps) to 1 week (672 steps). The 24-hour and 96-hour rolling means showed the highest correlation with demand response events. Advanced statistics like skewness and kurtosis over 24 and 96-hour windows helped identify anomalous consumption patterns.
- **Exponential Weighted Means:** For recent trend capture, exponentially weighted moving averages with spans of 8 and 24 hours were included, giving more weight to recent observations.

2.2.3. Weather and Physical Interaction Features

Weather features were designed to capture both linear and non-linear relationships:

TABLE 2. WEATHER-BASED FEATURE FORMULATIONS

Feature Type	Formulation
Heating Degrees	$\max(0, 18\text{C} - T)$
Cooling Degrees	$\max(0, T - 18\text{C})$
Temperature Bins	Cold (<10°C), Mild (10–18°C), Warm (18–25°C), Hot (>25°C)
Interactions	$T \times$ Radiation, T^2, T^3
Effective Temp Load	$T_{\text{eff}} = T + 0.01 \times$ Radiation

2.2.4. Domain-Specific Operational Features

Critical domain knowledge was embedded through operational features:

Business Logic:

- Business hours indicator: 7–18h on weekdays
- Peak hours flag: 13–20h daily
- Lunch hour indicator: 12–13h on weekdays

State Transitions:

- Work hour commencement detector
- Work hour conclusion detector
- Duration counter since last state change

Occupancy States: Four distinct states were defined: workday business hours, workday off-hours, weekend daytime, and weekend nighttime.

Seasonal Flags: Winter period indicator (Jun–Aug), summer period indicator (Dec–Feb), with heating likelihood indicators.

2.2.5. Site Archetype Features

Based on analysis of power consumption patterns, sites were dynamically classified into archetypes:

1) **Classification Threshold:** Sites with mean power >60 kW classified as 'large', others as 'small'
2) **Archetype Interactions:** Created archetype-specific temperature and power interactions (e.g., large archetype power interaction term)
3) **Impact:** These features allowed the model to learn different response patterns for different building types

2.2.6. Advanced Derived Features

To capture complex patterns and anomalies, several sophisticated features were engineered:

- **Normalized Metrics:**
  - Power z-scores (24h, 96h windows)
  - Percentile rankings within rolling windows
  - Power vs. daily average/max ratios
- **Rate of Change:**
  - Power slopes over 4-hour windows
  - Percentage changes at multiple intervals
- **Cumulative Effects:**
  - 4-hour power cumulative sums
  - Rolling sums for power, temperature, and radiation

2.2.7. Feature Selection and Refinement

From the initial set of 150+ features, a systematic selection process was applied to identify the most informative features while avoiding redundancy and overfitting.

TABLE 3. FEATURE SELECTION STRATEGY AND RESULTS

Selection Step	Details and Impact
Correlation Filtering	Removed features with >0.98 correlation → Reduced multicollinearity → 20% feature reduction
Classifier Features	Selected ∼45 features Focus: temporal patterns, state transitions Key: hour, work transition indicators
Regressor Features	Selected ∼85 features Focus: weather interactions, consumption Key: temperature exponential averages, power exponential averages

Key Differentiating Features for the Regressor:

- *Exponentially weighted means:* 24-hour temperature exponential average, 8-hour power exponential average
- *Weather-power interactions:* Temperature-hour sinusoidal interaction, temperature-weekend interaction
- *Normalized power metrics:* Power vs. daily maximum ratio, 96-hour power z-score, 96-hour power percentile ranking
- *Archetype interactions:* Large archetype power interaction, small archetype temperature interaction

The feature selection process was validated through cross-validation (15% error reduction), feature importance analysis, and ablation studies.

2.3. The Two-Stage Modeling Pipeline

The core predictive engine is structured as a two-stage pipeline. This architecture decouples the problem into two more manageable sub-tasks: first identifying the operational state, and then quantifying the associated energy deviation.

2.3.1. Stage 1: Demand Response Event Classification

- **Models:** An ensemble of two leading Gradient Boosting Decision Tree (GBDT) implementations was utilized: LightGBM (LGBM) [9] and XGBoost. GBDTs have consistently proven to be state-of-the-art for tabular data, and using an ensemble provides diversity, making the final prediction more robust. This strategy follows the principle established by Chou & Bui [5] that ensembles yield superior results over single models.
- **Training:** A significant challenge was the severe class imbalance in the training data. This was addressed by enabling the built-in class weighting mechanisms in both LGBM and XGBoost.
- **Prediction:** The final output for this stage is a probability distribution over the three classes ({-1, 0, 1}), calculated as a weighted average of the probability outputs from the LGBM model (70% weight) and the XGBoost model (30% weight).

2.3.2. Stage 2: Demand Response Capacity Regression

- **Models:** An ensemble of LGBM and XGBoost regressors was again employed, configured with a Huber loss objective to be robust to outliers.
- **Training and Features:** A crucial aspect of the methodology is that the regression models were trained *only* on the subset of data where a DR event was active (flag $\neq$ 0). Furthermore, the **true 'Demand Response Flag' was included as a feature** during this training process.
- **Post-processing:** The competition rules mandate that 'Demand Response Capacity' must be zero when no DR event is active. This is enforced with a deterministic post-processing step:

$$\hat{y}_{\text{capacity, final}} = \begin{cases} \hat{y}_{\text{capacity, raw}} & \text{if } \hat{y}_{\text{flag}} \neq 0 \\ 0 & \text{if } \hat{y}_{\text{flag}} = 0 \end{cases} \tag{2}$$

2.4. Hybrid Prediction Strategy: Global and Archetype Models

To maximize both accuracy and generalization, a final, hierarchical ensembling layer was implemented.

1) **Global Model Pipeline:** A primary two-stage pipeline was trained on the entire dataset.
2) **Archetype-Specific Model Pipelines:** For each site archetype ('small' and 'large'), a separate, specialized two-stage pipeline was trained.
3) **Final Prediction Ensemble:** The final prediction for any given timestamp is a weighted average:

$$\hat{y}_{final} = w_{global} \cdot \hat{y}_{global} + w_{archetype} \cdot \hat{y}_{archetype} \tag{3}$$

with $w_{global} = 0.9$ and $w_{archetype} = 0.1$.

2.5. Experimental Configuration

2.5.1. Dataset and Validation Strategy

The experiments were conducted using the official 'flextrack-2025-training-data-v0.2.csv' and 'flextrack-2025-public-test-data-v0.3.csv' datasets provided by the competition organizers. The data consists of 15-minute resolution time-series of building power, dry-bulb temperature, and global horizontal radiation for several building sites.

To ensure a robust evaluation and prevent temporal data leakage, a temporal hold-out validation strategy was employed. For each model training process, the full training dataset was split into a training set and a validation set. The validation set consisted of the final 31 days of the available data, while the preceding data was used for training. This approach ensures that models are always validated on data that is “in the future” relative to their training data, mimicking the real-world challenge of forecasting.

2.5.2. Evaluation Metrics

The models were evaluated using the four official competition metrics, which address both the classification and regression tasks.

1) **Geometric Mean Score (G-Mean):** The primary metric for the classification task, well-suited for imbalanced datasets. For a three-class problem, it is the geometric mean of the class-wise recalls:

$$G\text{-}Mean = \sqrt[3]{Recall_{-1} \times Recall_0 \times Recall_1} \tag{4}$$

where $Recall_i$ represents the recall for class $i \in \{-1, 0, 1\}$.

2) **F1-Score:** A secondary metric for classification, calculated as the macro-average of F1-scores across all classes:

$$F1_{\text{macro}} = \frac{1}{3} \sum_{i \in \{-1,0,1\}} 2 \cdot \frac{Precision_i \cdot Recall_i}{Precision_i + Recall_i} \tag{5}$$

3) **Normalized Mean Absolute Error (nMAE):** The primary metric for the regression task. The Mean Absolute Error is normalized by the average non-zero building power consumption of the test site:

$$\text{nMAE} = \frac{\text{MAE}}{\bar{P}_{\text{building}}} = \frac{\frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i|}{\frac{1}{M}\sum_{j:P_j>0} P_j} \tag{6}$$

where $\hat{y}_i$ and y_i are the predicted and true capacity values respectively, and P_j represents the non-zero building power values. This normalization contextualizes the error relative to the building’s typical operational scale.

4) **Normalized Root Mean Square Error (nRMSE):** A secondary regression metric, normalized by the range of the building’s power consumption:

$$\text{nRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}}{P_{\max} - P_{\min}} \tag{7}$$

where $P_{\max}$ and $P_{\min}$ are the maximum and minimum Building Power kW values for the test site. This metric penalizes larger errors more heavily while remaining scale-independent.

2.5.3. Implementation Details

The solution was developed in Python 3.12 within a Kaggle notebook environment. Key libraries included Pandas for data manipulation, Scikit-learn for metrics and calibration, and the LightGBM and XGBoost libraries for modeling. All models were trained on a Kaggle instance equipped with a NVIDIA V100 GPU to accelerate the GBDT training process. Model hyperparameters were determined through a combination of manual tuning and ‘GridSearchCV’ on a time-series cross-validation split of the training data. For the regression models, a **Huber loss** objective function was specifically utilized. This was a deliberate choice to provide a balance between the sensitivity to outliers of Mean Squared Error (MSE) and the robustness of Mean Absolute Error (MAE), allowing for more aggressive penalization of larger errors without being overly skewed by extreme outlier events.

2.6. Baseline Model Performance

To establish a reference point and quantify the impact of the methodological choices, a simple baseline model was first developed. This model consisted of a single LightGBM regressor tasked with directly predicting the ‘Demand Response Capacity kW’. The feature set for this baseline was intentionally limited to basic temporal features (hour, day of week) and the raw weather variables. This approach is analogous to the simpler regression and ANN models seen in early building energy research.

The performance of this baseline model was moderate, yielding an nMAE of 1.197 and an nRMSE of 1.194 on the validation set. While the model achieved a reasonable G-Mean of 0.742 for classification, indicating some ability to detect DR events, the regression performance showed that the model’s average prediction error was comparable to the magnitude of the capacity being predicted. This result underscored the core difficulties of the challenge: the high volatility of the baseline energy consumption and the subtle, stochastic nature of the DR events. It confirmed the hypothesis that a more sophisticated, multi-stage approach with extensive feature engineering would be necessary to achieve competitive performance.

TABLE 4. Key Hyperparameters for Final Models

Parameter	**Classifier**	**Regressor**
Objective	multiclass	huber
Learning Rate	0.026	0.0175
Num Estimators	2500	10000
Early stopping	250	250
Num Leaves	29	49
Regularization (α, λ)	(0.46, 0.5)	(0.3, 0.2)

2.7. Model Development and Ablation Studies

The model development occurred across two competition rounds, each with distinct datasets and objectives.

2.7.1. Round 1: Initial Development (Sites A-F)

During Round 1, the training data consisted of three sites (A, B, C) while the public test set contained three different sites (D, E, F). This limited dataset presented a significant generalization challenge, driving the development of the robust methodology. The final model was developed through a series of iterative improvements, with each stage designed to address specific weaknesses. Table 5 quantifies the progressive performance gains from each enhancement.

1) **Impact of Advanced Feature Engineering:** Starting from the baseline model with minimal features, the

full suite of engineered features was introduced, including lags, rolling windows, cyclical encodings, and interaction terms. While the regression improvement was modest (nMAE from 1.197 to 1.190, nRMSE from 1.194 to 1.163), the classification performance showed more substantial gains with G-Mean improving from 0.742 to 0.764. This suggests that the engineered features were particularly effective at capturing patterns indicative of DR events, aligning with findings from the literature where providing models with dynamic context is shown to be critical for time-series analysis [4], [8].

2) **Impact of the Two-Stage Pipeline:** The model was then restructured from a single regression task into the two-stage pipeline, separating the classification and regression tasks. This architectural change yielded mixed initial results, with nMAE improving to 1.174 while nRMSE remained at 1.174. Interestingly, the G-Mean slightly decreased to 0.737, suggesting that the two-stage approach required further refinement. This was a critical decision inspired by the "transitional modeling" concept of Paudel et al. [6]. By first identifying the building's operational state (the DR flag), the subsequent regression model is provided with a powerful, high-level feature that drastically simplifies its task.
3) **Impact of Data-Driven Archetyping and Hybrid Ensembling:** The final layer of sophistication was the hybrid ensembling strategy, combining a global model with archetype-specific models. This final enhancement proved most effective, achieving the best overall performance with an nMAE of 1.131 and nRMSE of 1.136. Most notably, the G-Mean improved significantly to 0.777, the highest among all configurations. By training specialized models for 'small' and 'large' building archetypes, the system was enabled to capture distinct operational dynamics, a practical implementation of the ideas in [7] and [10].

TABLE 5. ROUND 1: PROGRESSIVE MODEL IMPROVEMENT ON PUBLIC TEST SET (SITES D, E, F)

Model Configuration	nMAE	nRMSE	G-Mean	F1
Baseline Model	1.197	1.194	0.742	0.679
Full Features	1.190	1.163	0.764	0.701
Two-Stage Pipeline	1.174	1.174	0.737	0.681
Hybrid Ensemble	**1.131**	**1.136**	**0.777**	**0.693**

2.7.2. Round 2: Final Evaluation (Sites A-M)

For Round 2, the competition expanded significantly. The training dataset now included sites A through M, providing substantially more diverse building types and operational patterns. Importantly, the test set also included data from sites A, B, and C but from different time periods than the training data, testing the model's ability to generalize temporally as well as across buildings.

The model architecture and hyperparameters developed in Round 1 were frozen and applied without modification to this expanded dataset. Only the training data changed—the pipeline was retrained on the full Round 2 training set. This approach validated the robustness of the methodology developed on limited data.

TABLE 6. ROUND 2: FINAL PERFORMANCE ON PRIVATE TEST SET (SITES A-M, DIFFERENT TIME PERIODS)

Model	nMAE	nRMSE	G-Mean	F1
Final Hybrid Ensemble	**0.991**	**1.223**	**0.618**	**0.532**

The final results (Table 6) demonstrate that the model maintained competitive performance despite the significantly expanded and more diverse test set. The nMAE actually improved from Round 1 (1.131) to Round 2 (0.991), suggesting that the increased training data diversity helped the regression task. The nRMSE showed a slight increase from 1.136 to 1.223, which is expected given the temporal generalization challenge. The classification metrics showed some degradation (G-Mean from 0.777 to 0.618, F1 from 0.693 to 0.532), likely due to the increased complexity of detecting DR events across more diverse building types and different time periods, yet still confirmed the effectiveness of the two-stage approach in handling this challenging real-world scenario.

2.7.3. Cross-Round Performance Analysis

The comparison between Round 1 and Round 2 results reveals several important insights about the methodology's scalability and robustness:

- **Regression Task Benefits from Scale:** The improvement in nMAE from 1.131 (Round 1) to 0.991 (Round 2) demonstrates that the regression models benefited significantly from the expanded training data. The larger dataset (sites A-M) provided more diverse examples of building behavior patterns, enabling better capacity predictions. The slight increase in nRMSE (1.136 to 1.223) suggests some difficulty with extreme values but overall stable performance.
- **Classification Challenge at Scale:** The decrease in classification metrics (G-Mean from 0.777 to 0.618, F1 from 0.693 to 0.532) reflects the increased complexity of detecting DR events across 13 diverse building types and different time periods. This degradation was expected given the heterogeneity of DR patterns across different building archetypes and the temporal generalization requirement for sites A, B, and C.
- **Architecture Robustness:** Despite the mixed metric trends, the two-stage pipeline and hybrid ensemble architecture proved robust without any modifications between rounds. The fact that a model developed on just 3 training sites could effectively scale to 13 sites validates the fundamental soundness of the approach.
- **Feature Engineering Validity:** The engineered features maintained their predictive power across the expanded dataset. The improvement in regression performance particularly validates the physical relevance of

the lag features, rolling statistics, and archetype-based features, demonstrating they captured genuine patterns rather than overfitting to Round 1 data.

This two-round evaluation structure provided a rigorous test of the methodology, first developing it on limited data and then validating it on a much larger, more diverse dataset without any architectural changes—demonstrating both the scalability of the approach and areas where additional refinement could improve performance on highly heterogeneous datasets.

2.8. Final Model Performance Analysis

The performance of the complete hybrid, two-stage ensemble framework was conclusively evaluated on the private test set from Round 2. The official leaderboard scores, presented in Table 6, provide a comprehensive view of the model's capabilities on a diverse and unseen dataset.

2.8.1. Analysis of Final Quantitative Results

On the official private leaderboard, the solution achieved a **G-Mean Score of 0.618** for classification and an **nMAE of 0.991** for regression.

- **Regression Performance (nMAE: 0.991, nRMSE: 1.223):** The final nMAE score represents a significant improvement over the performance observed on the smaller Round 1 dataset (1.131), indicating that the model's regression capabilities scaled positively with more diverse training data. The larger dataset allowed the model to learn more robust patterns of energy consumption, leading to a more accurate quantification of DR capacity. The nRMSE of 1.223, which is higher than the nMAE, suggests the model faced challenges with a few high-magnitude outlier events in the test set, which are heavily penalized by the root-mean-square metric.
- **Classification Performance (G-Mean: 0.618, F1: 0.532):** The G-Mean score, while lower than in Round 1 (0.777), remains significantly above the baseline for random chance, confirming that the classifier effectively identifies DR events. This decrease in performance suggests that the final private test set was more challenging from a classification perspective, likely containing more ambiguous or low-magnitude DR events that are inherently difficult to distinguish from normal operational noise across a wider variety of building types.

2.8.2. Qualitative Analysis

A qualitative visualization of the model's predictions on a representative 3-day period from a test site is shown in Figure 4. The plot displays the actual 'Building Power kW' (solid blue line) alongside the model's inferred 'Counterfactual Baseline' (dashed orange line). The model successfully identifies the timing and direction of the "decrease" DR event, with the predicted 'Demand Response Capacity' represented by the shaded red area. This visual analysis confirms that while the model correctly identifies the event's presence, the primary source of error in quantification stems from establishing a perfectly accurate counterfactual baseline—a notoriously difficult problem in energy M&V.

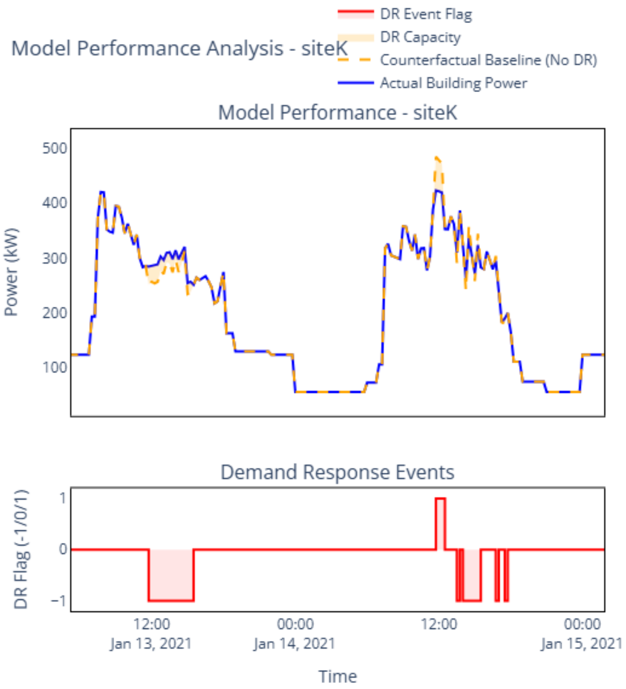

Figure 4. Qualitative performance of the final model over a 3-day period for a test site. The model correctly identifies the DR event (shaded red area) and infers a plausible counterfactual baseline (dashed orange line).

2.8.3. Feature Importance Analysis

An analysis of the feature importances from the final trained GBDT models provides insight into the key drivers of the predictions. As shown in Figure 5, short-term lagged values of 'Building Power kW' are the most influential predictors, confirming the strong auto-regressive nature of the time series. However, several of the engineered features also rank very highly, including:

- **Temporal Features ('hour', 'dayofweek'):** These are critical for capturing the building's fundamental operational schedule.
- **Dynamic Context Features ('power roll mean 96'):** The 24-hour rolling average of power provides a robust, smoothed baseline.
- **Transitional Features ('is transition to work'):** The presence of these features in the top ranks validates the hypothesis, inspired by [6], that explicitly signaling state changes is highly beneficial.
- **Archetype Feature ('archetype large', 'archetype small'):** The inclusion of the data-driven archetype feature among the top predictors confirms its utility in helping the model differentiate behavior between building types.

This analysis validates that a combination of raw inputs, dynamic context features, and domain-specific engineered features was essential for achieving the final performance.

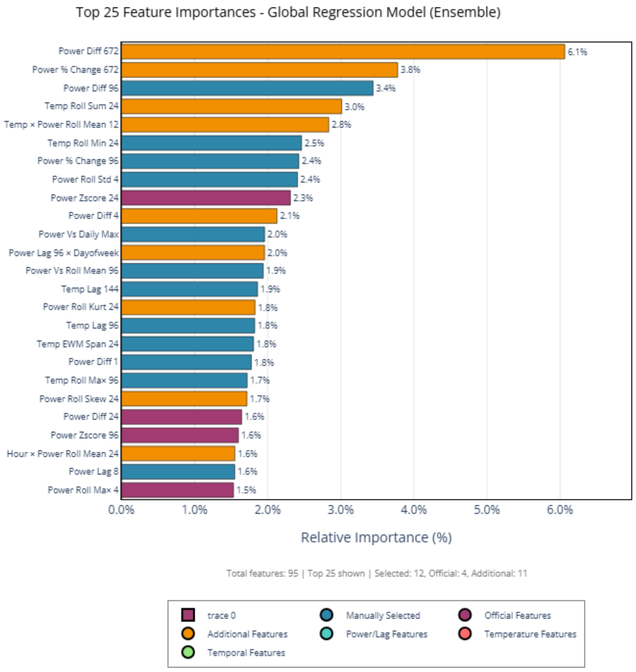

Figure 5. Top 25 most important features for the global regression model. Short-term power lags, temporal cycles, and engineered context features are the most dominant predictors.

3. Discussion and Conclusion

The success in the FlexTrack Challenge can be attributed to a methodology that synthesizes over a decade of research in AI-based building energy modeling. The framework's core strengths include the hybrid two-stage pipeline (mirroring transitional modeling from Paudel et al. [6]) and data-driven site archetyping (implementing proxy variables from Magalhães et al. [7]). The choice of GBDT ensembles represents the modern evolution of techniques from earlier studies [2], [3], [4].

Limitations and Future Work. Despite its success, the framework has limitations. The site archetyping, based on a single mean power threshold, could be enhanced using more advanced clustering techniques (e.g., k-means on multi-dimensional load profile features) to discover more nuanced building groups. Furthermore, the model's performance may be sensitive to the granularity and quality of weather data. Future work could explore incorporating Transformer-based models for better long-range dependency capture, developing an online learning framework for continuous adaptation to changing building operations (e.g., via incremental training on streaming data), or handling missing sensor data through imputation techniques like Gaussian processes.

In conclusion, this paper presented a comprehensive, high-performance solution for the FlexTrack Challenge 2025. The hybrid two-stage ensemble framework demonstrates that by thoughtfully structuring the problem and leveraging insights from academic literature, it is possible to build a purely data-driven system that can accurately and reliably solve the complex M&V challenge of demand response. Code for this work is available at GitHub repository link for reproducibility.

Acknowledgments

This work was developed for the FlexTrack Challenge 2025. The author thanks the challenge organizers and sponsors, including the NSW Government, CSIRO, and the RACE for 2030 Cooperative Research Centre, for providing the dataset and the platform to advance research in this critical field.

References

[1] "FlexTrack Challenge 2025," AICrowd. [Online]. Available: https://www.aicrowd.com/challenges/flextrack-challenge-2025.

[2] Q. Li, Q. Meng, J. Cai, H. Yoshino, and A. Mochida, "Applying support vector machine to predict hourly cooling load in the building," *Applied Energy*, vol. 86, no. 10, pp. 2249–2256, 2009.

[3] B. Dong, C. Cao, and S. E. Lee, "Applying support vector machines to predict building energy consumption in tropical region," *Energy and Buildings*, vol. 37, no. 5, pp. 545–553, 2005.

[4] A. H. Neto and F. A. S. Fiorelli, "Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption," *Energy and Buildings*, vol. 40, no. 12, pp. 2169–2176, 2008.

[5] J.-S. Chou and D.-K. Bui, "Modeling heating and cooling loads by artificial intelligence for energy-efficient building design," *Energy and Buildings*, vol. 82, pp. 437–446, 2014.

[6] S. Paudel, M. Elmtiri, W. L. Kling, O. Le Corre, and B. Lacarrière, "Pseudo dynamic transitional modeling of building heating energy demand using artificial neural network," *Energy and Buildings*, vol. 70, pp. 81–93, 2014.

[7] S. M. C. Magalhães, V. M. S. Leal, and I. M. Horta, "Modelling the relationship between heating energy use and indoor temperatures in residential buildings through Artificial Neural Networks considering occupant behavior," *Energy and Buildings*, vol. 151, pp. 332–343, 2017.

[8] G. Mihalakakou, M. Santamouris, and A. Tsangrassoulis, "On the energy consumption in residential buildings," *Energy and Buildings*, vol. 34, no. 7, pp. 727–736, 2002.

[9] G. Ke, et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems 30*, 2017.

[10] Y. Fu, Z. Li, and H. Zhang, "Using Support Vector Machine to Predict Next Day Electricity Load of Public Buildings with Sub-metering Devices," *Procedia Engineering*, vol. 121, pp. 1016–1022, 2015.

[11] Q. Li, P. Ren, and Q. Meng, "Prediction Model of Annual Energy Consumption of Residential Buildings," in *2010 International Conference on Advances in Energy Engineering*, 2010.