The PG&E Energy Analytics Challenge Report

Huynh Q. N. Vo *, H M Mohaimanul Islam*, Richard Reed *, Yash Patel *
*School of Industrial Engineering and Management Oklahoma State University, Stillwater, OK 74078 USA
Email: {lucius.vo, h_m_mohaimanul.islam, ricky.reed, yash.j.patel}@okstate.edu

I. Introduction

As the global energy landscape shifts towards renewable sources, accurate electricity load forecasting is increasingly vital for maintaining grid stability and optimizing resource allocation. The 2025 PG&E Energy Analytics Challenge addresses this imperative by tasking participants with predicting hourly electricity load for a full calendar year in an undisclosed California region, where solar energy significantly influences demand patterns. Participants are provided with historical datasets spanning two years, comprising hourly measurements of electricity load, temperature, and global horizontal irradiance (GHI) from five sites within the region. GHI, the total solar radiation on a horizontal surface, serves as a key predictor of photovoltaic generation potential and its impact on net load. The challenge mandates that forecasts rely solely on these datasets, with predictions for any given day using only data available up to that point, mirroring real-world forecasting constraints. Here, we detail our methodology—encompassing exploratory data analysis, model selection, and validation—to develop a robust forecasting tool that supports energy management and operational efficiency.

In this round of the Energy Analytics Challenge, people will use historical datasets that include hourly measurements of electricity load from the past two years, along with environmental variables like temperature and global horizontal irradiance, or GHI collected from different places in the target region. GHI, representing the total solar radiation received on a horizontal surface, is a crucial predictor reflecting the potential for photovoltaic energy generation and its impact on electricity demand patterns. Furthermore, to ensure realistic and operational applicability, the challenge necessitates the following stringent forecasting criteria: predictions must encompass an entire year and exclusively utilize the provided dataset without relying on external information. Thus, predictions for any given day must rely solely upon predictor data from that day or earlier, mimicking real-world forecasting constraints.

In this report, we described the methodologies applied in developing robust and accurate forecasting models, including exploratory data analysis, model selection, and rigorous validation procedures. Furthermore, we discuss the specific challenges encountered and the strategies implemented to overcome them. The ultimate objective of this effort is to deliver a practical and precise forecasting tool capable of supporting informed decision-making and enhancing operational efficiency in energy management.

II. Methodology

The fundamental principle of multivariate forecasting hinges on the premise that the future value of a variable can be more accurately predicted by taking into account the influence of other relevant variables in addition to its past values. This approach is particularly powerful in complex systems—where there is a significant interaction between multiple factors and thus affects the outcomes. The rationale behind multivariate forecasting is to capture with high fidelity these interactions and dependencies to improve forecast accuracy and reliability.

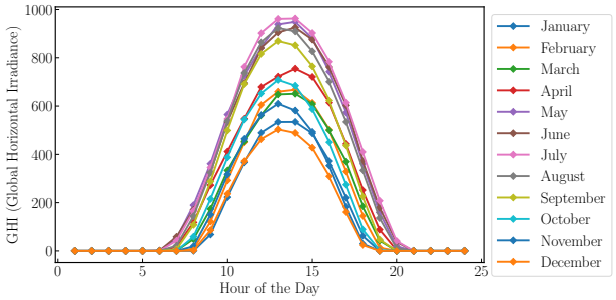
Our approach to multivariate forecasting leverages the interdependence of electricity load with exogenous variables—temperature, GHI, and temporal features—to enhance predictive accuracy. This methodology comprises: (1) Problem Statement: Forecasting hourly electricity load for a year using provided datasets; (2) Data Collection and Exploration: Analyzing distributions, trends, and correlations in the training data (see Section II.A); (3) Data Preprocessing: Standardizing features and engineering harmonic components (e.g., sine/cosine transformations of hour) and aggregated weather variables (e.g., mean temperature and GHI across sites) to capture cyclic and regional effects; (4) Model Selection: Evaluating multiple methods, with justification for selecting XGBoost (see Section III); (5) Training and Validation: Splitting the two-year training data chronologically (80% training, 20% validation) and optimizing hyperparameters via time series cross-validation; (6) Evaluation: Assessing performance with R^2, MAE, MSE, and MAPE; and (7) Forecast Generation: Producing the final year-long forecast with sanity checks.

A. Explanatory Data Analysis

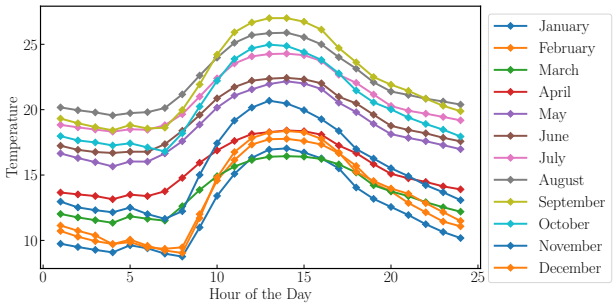
In this section, we present a brief overview of the provided datasets and discuss our EDA.

In this challenge, we received two sets of data: a training set contained in the file `training.xlsx` and a test set provided in the file `testing.xlsx`. Each set includes three primary variables: **Load**, representing the electricity load at the node of interest; **Site-X Temp**, indicating temperature measurements from various random locations within the node; and **Site-X GHI**, capturing the Global Horizontal Irradiance (GHI) at these locations. Data are collected from five different nodes, denoted by $\mathbf{X} = 1, \cdots, 5$ within the region of interest in California. Furthermore, the training set provides hourly observations spanning two years, whereas the test set includes hourly data for a subsequent third year, with the **Load** variable left blank, serving as the prediction target for the challenge.

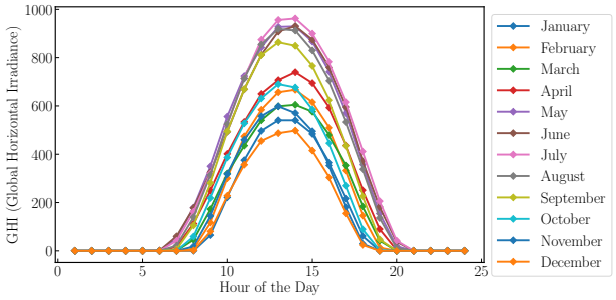
Based on the aforementioned information, we define the target (endogenous) variable as **Load**, and the explanatory (exogenous) variables as the time components—that is, **Year**, **Month**, and **Hour**— as well as **Temp** and **GHI** from multiple sites.



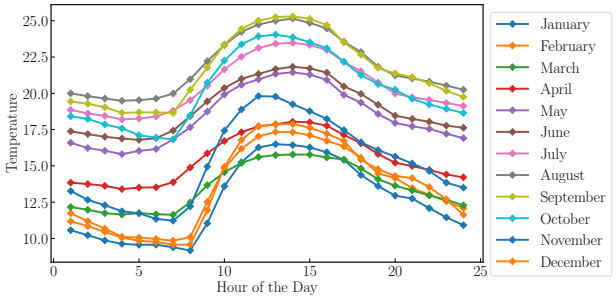
(a) **GHI** for Site 1 (Year 1).



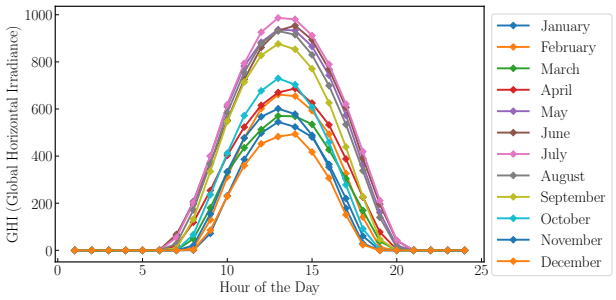
(b) **Temp** for Site 1.



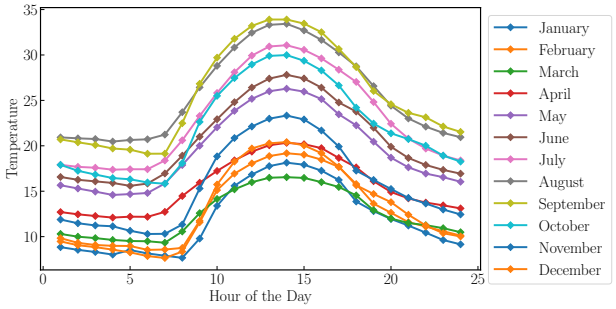
(c) **GHI** for Site 2 (Year 1).



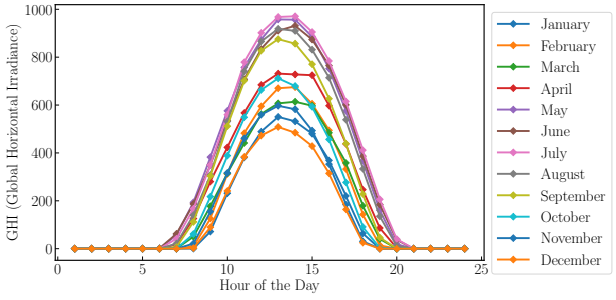
(d) **Temp** for Site 2.



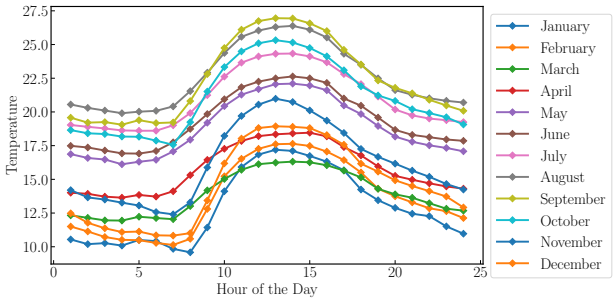
(e) **GHI** for Site 3 (Year 1).



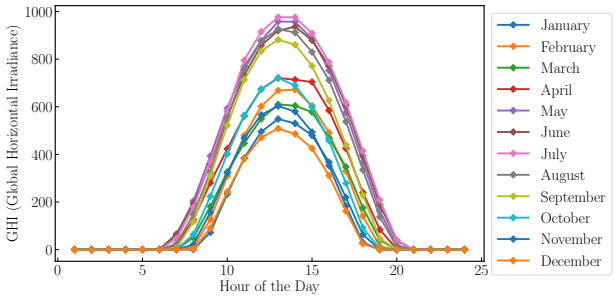
(f) **Temp** for Site 3.



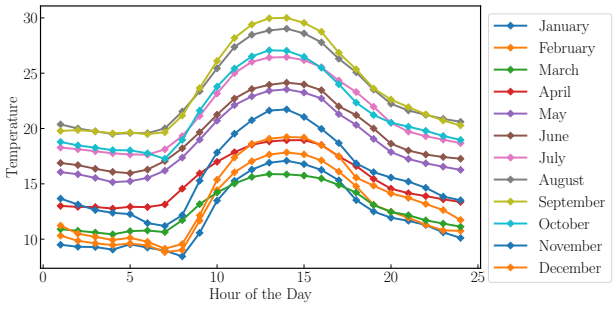
(g) **GHI** for Site 4 (Year 1).



(h) **Temp** for Site 4.

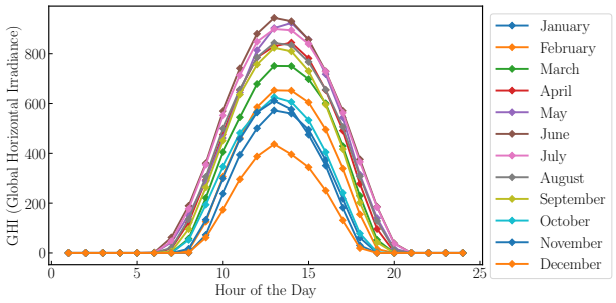


(i) **GHI** for Site 5 (Year 1).

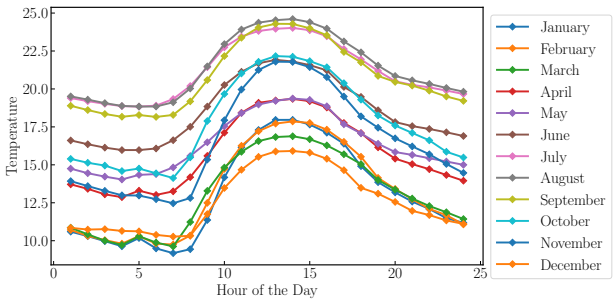


(j) **Temp** for Site 5.

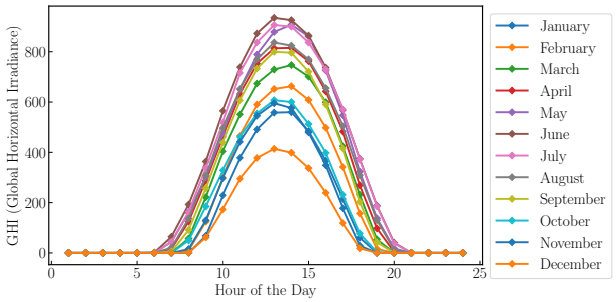
Fig. 1: Hourly variations of Global Horizontal Irradiance (GHI) and Temperature for Sites 1 to 5 (Year 1).



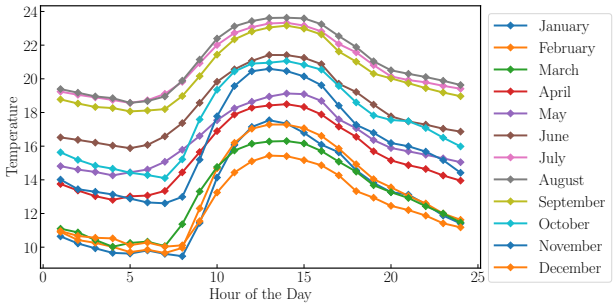
(a) **GHI** for Site 1 (Year 2).



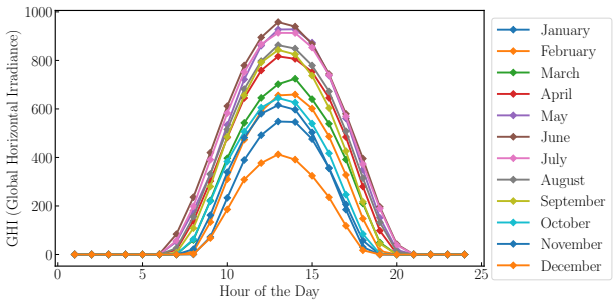
(b) **Temp** for Site 1 (Year 2).



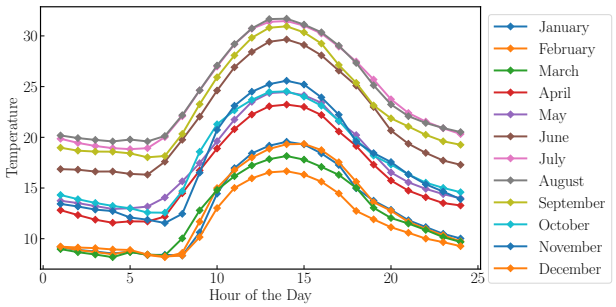
(c) **GHI** for Site 2 (Year 2).



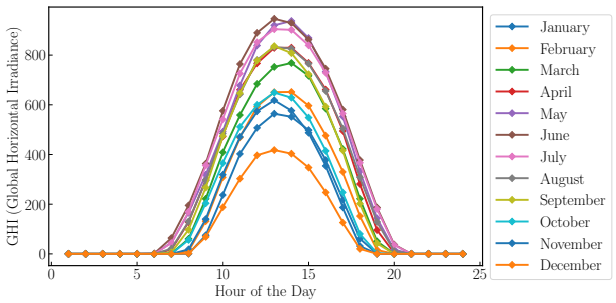
(d) **Temp** for Site 2 (Year 2).



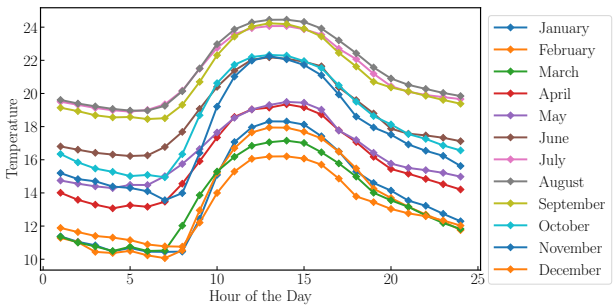
(e) **GHI** for Site 3 (Year 2).



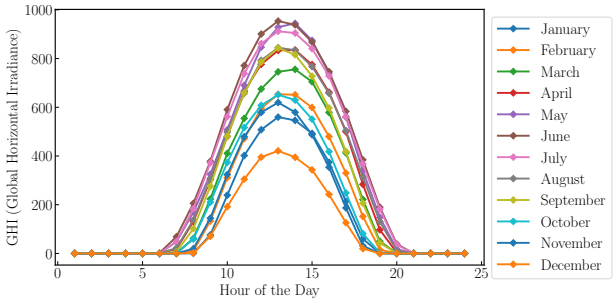
(f) **Temp** for Site 3 (Year 2).



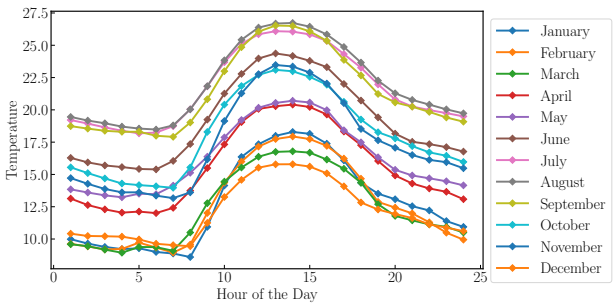
(g) **GHI** for Site 4 (Year 2).



(h) **Temp** for Site 4 (Year 2).



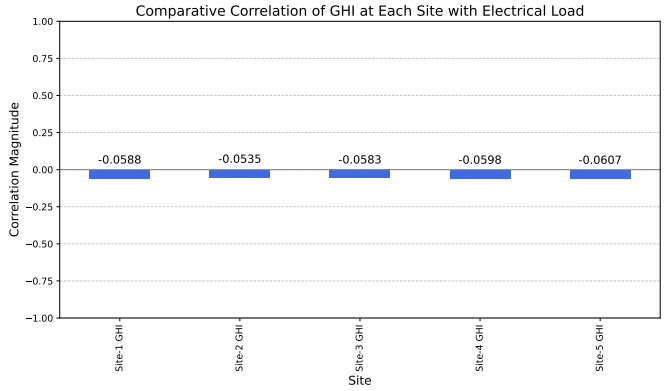
(i) **GHI** for Site 5 (Year 2).



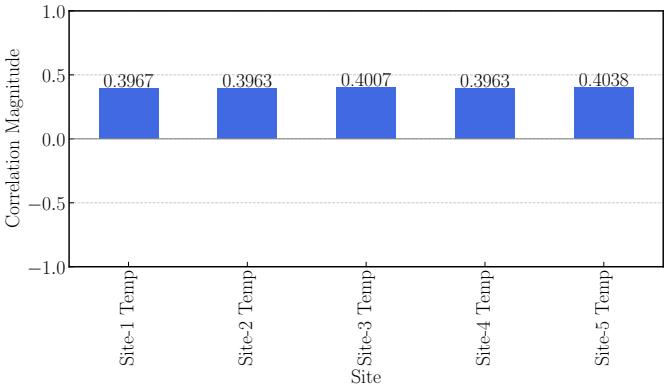
(j) **Temp** for Site 5 (Year 2).

Fig. 2: Hourly variations of Global Horizontal Irradiance (GHI) and Temperature for Sites 1 to 5 (Year 2).

Figures 1-2 demonstrates the hourly variations of GHI and temperature profiles over the first and the second year, respectively, across five different sites. We observed significant seasonal patterns and a robust interdependence between GHI and temperature, reflecting the inherent physical relationship between solar irradiance and temperature. This relationship manifests through the diurnal cycle, where increasing solar irradiance during daylight hours directly elevates temperature due to energy absorption by the Earth's surface. Moreover, pronounced seasonal trends highlight this connection further. For instance, higher irradiance during warmer seasons (e.g., summer) is associated with higher temperatures, and reduced irradiance aligns with cooler conditions in colder seasons (e.g., winter).



(a) Correlation of **GHI** with **Load**.

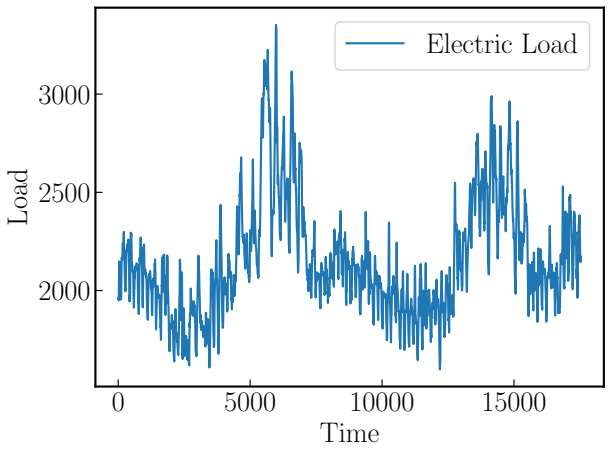


(b) Correlation of **Temperature** with **Load**.

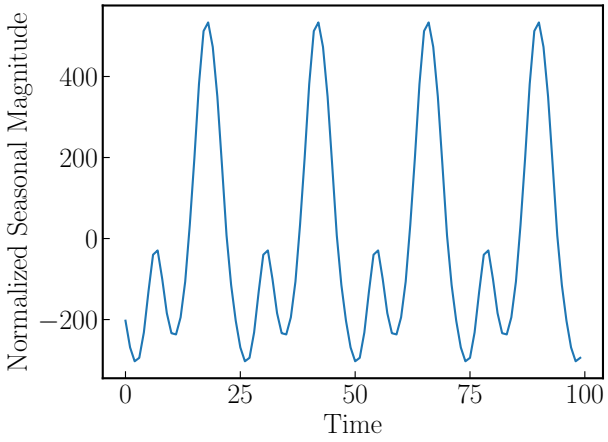
Fig. 3: Comparative correlations of GHI and temperature with electrical load across sites.

In solar energy systems, electricity generation directly correlates with solar irradiance; higher irradiance levels result in increased electricity production. Given that the **Load** variable represents the net electricity demand, calculated as the total demand minus electricity generated from combined solar and wind sources, we hypothesized a negative association between **Load** and **GHI** across all sites. Figure 3 demonstrates our observations regarding this relationship. Ideally, the net load decreases as solar generation increases, implying a negative correlation with solar irradiance. However, temperature exhibited a strong positive correlation with load, suggesting that as temperatures rise, electricity demand significantly increases due to presumably higher usage of cooling systems (e.g., air conditioning). Consequently, while increased GHI tends to reduce net load through enhanced solar generation, this effect is partly offset by elevated electricity consumption triggered by higher temperatures. As a result, GHI demonstrates only a slight negative correlation with net load.

Given that temperature and GHI are correlated with the load, as mentioned in our previous findings, we postulated that the electricity load's seasonality would match the other exogenous variables. Figure 4 corroborates our hypothesis: we see that the **Load** variable has the periodicity of $s = 24$.



(a) Trends of **Load** for a two-year period.



(b) Seasonal component with $s = 24$ for the first 100 hours. The component is normalized to have a mean of zero.

Fig. 4: Seasonal decomposition of **Load** for two years.

In conclusion, our EDA revealed several critical temporal interdependencies among **Load**, **GHI**, and **Temp** variables. Notably, the hourly and seasonal variations of **GHI** and **Temp** exhibit consistent diurnal and seasonal patterns, with the seasonality being monthly. The analysis also highlighted a nuanced interplay: while an increase in solar irradiance generally leads to reduced net electricity load through increased solar energy generation, higher temperatures simultaneously drive electricity consumption, particularly due to increased use of cooling systems. Moreover, the seasonality of these variables affects that of the electricity load. These results emphasize

the necessity for forecasting models, whether black-box or interpretable, to capture these temporal interdependencies with high fidelity for more reliable forecasted results.

III. Model Selection, Training, and Validation

During our EDA, we confirmed the absence of missing values in the provided datasets, eliminating the need for imputation. Using the insights from our EDA, we engineered the following features to enrich key patterns in the datasets:

- **Harmonic components**: To model the cyclic trends observed in the load data, we applied sine and cosine transformations to the **Hour** variable. These harmonic features effectively represent the diurnal periodicity evident in the time series.
- **Aggregated weather features**: To integrate multi-site weather effects, we calculated the average Temperature (Temp) and Global Horizontal Irradiance (GHI) across all sites. This aggregation succinctly captures the overall environmental influence on load.

We selected the eXtreme Gradient Boosting (XGBoost) algorithm for its demonstrated ability to model complex, non-linear relationships in multivariate datasets. Its ensemble-based structure and regularization properties make it well-suited for this forecasting task. The model was trained on hourly data spanning the first 49 months (approx. 80% of the data in the `training.xlsx` file) and validated on subsequent months. The model performance was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE). These metrics were selected to provide a comprehensive assessment of accuracy: the MAE offers a linear measure of error, the MSE provides insights on potential larger deviations, and MAPE facilitates relative error interpretation.

In the following sections, we provided a comprehensive breakdown of justifications for selecting XGBoost and how the model's parameters are fine-tuned.

A. XGBoost

The eXtreme Gradient Boosting (XGBoost) algorithm is an advanced implementation of gradient boosting that constructs an ensemble of decision trees iteratively. It optimizes a customizable loss function using gradients and Hessians, achieving faster and more precise convergence than traditional boosting methods. With L_1 and L_2 regularization to prevent overfitting and built-in handling of missing data, XGBoost is robust for real-world applications.

For electricity load forecasting, XGBoost excels at capturing non-linear relationships between the load and predictors such as temperature, solar irradiance, and temporal features. Its high accuracy, efficiency, and scalability suit it to large, complex datasets. Integration with the `scikit-learn` API further enables seamless preprocessing, hyperparameter optimization, and time series-aware cross-validation (e.g., the `TimeSeriesSplit` function), enhancing its utility in forecasting pipelines.

B. Hyperparameter Finetuning

In our implementation, the model is embedded within a pipeline that standardizes input features and then applies the XGBoost regressor configured with appropriate hyperparameters. We leverage `TimeSeriesSplit` to ensure that cross-validation respects the temporal ordering of data, preventing look-ahead bias. A grid search is performed over parameters such as the number of trees, maximum tree depth, and learning rate to identify the optimal model configuration, and the best model is subsequently used to generate forecasts on the test set.

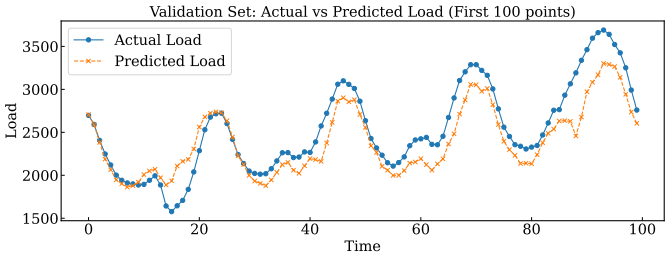


Fig. 5: Trends of the actual vs. the predicted load on the validation set in a particular time window.

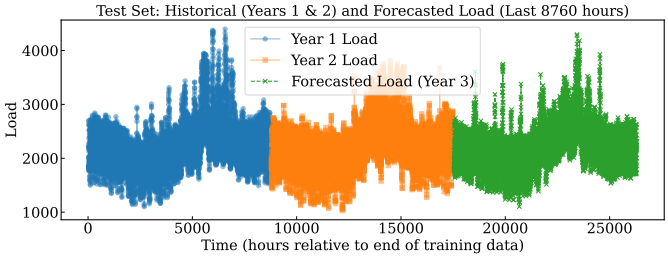


Fig. 6: Trends of the actual vs. the predicted load on the validation set in a particular time window.

IV. Model Evaluation

The optimized XGBoost model's performance on the validation set is summarized in Table I, indicating high explanatory power and accuracy. Moreover, Figure 5 compares actual versus predicted Load over a selected validation window, demonstrating the model's ability to capture trends and fluctuations.

	R^2	MAE	MSE	MAPE
Validation	0.9760	170.95	54911.12	0.0750%

TABLE I: Validation results of the optimal XGBoost model

V. Forecast Generation

Using the validated XGBoost model, we forecasted hourly Load for Year 3 using only exogenous data from `testing.xlsx`. To assess plausibility, we conducted a visual inspection (Figure 6) of the forecasted **Load** alongside historical **Loads** to confirm similar diurnal and seasonal patterns, supporting forecast reliability.

Supplementary Information

In this project, we implement the `ipywidgets` functions supported by Jupyter Notebook to provide interactive plots. These plots can be provided in the project's GitHub repository.