Forecasting Renewable Energy Generation at the Plant Level Using XGBoost

Yeabsra Habtu, PGR/38331/17

Abstract

General forecasting of renewable energy generation at the power-plant level remains limited by data gaps, variability across regions, and the non-linear nature of influencing factors. This study addresses these challenges by developing an XGBoost regression model to predict annual electricity generation (GWh) for global renewable power plants. Using a cleaned and imputed dataset—including capacity, fuel type, location, and commissioning year—we engineer features such as plant age and log-transformed capacity. Missing values are handled via predictive mean matching. The model is trained on 80% of the data and tested on the remaining 20%, achieving an R^2 of 0.88, RMSE of 698.15 GWh, and MAE of 293.80 GWh. Installed capacity, plant age, and technology type emerged as key predictors. Five-fold cross-validation confirms the model's robustness. These results highlight the value of gradient-boosted trees for capturing complex relationships in heterogeneous datasets, offering a scalable approach to support energy planning, investment, and policy development.

1. Introduction & Background

The increasing global focus on sustainability has elevated the importance of renewable energy sources, such as solar, wind, and hydroelectric power. Efficiently forecasting renewable energy generation is crucial for optimizing grid stability, energy trading, and planning operations at both regional and plant levels. However, the variable and weather-dependent nature of renewables introduces significant uncertainty, posing challenges to accurate forecasting.

Forecasting at the individual plant level offers a more granular understanding of energy patterns, enabling more precise operational decisions. Unlike aggregate forecasting, plant-level prediction demands handling site-specific characteristics and localized external factors, which increase the complexity of modeling.

Machine learning techniques, particularly tree-based ensemble models, have proven effective in capturing nonlinear patterns in energy data. Among them, Extreme Gradient Boosting (XGBoost) stands out

due to its scalability, regularization capabilities, and superior performance in structured data prediction tasks. This research explores the application of XGBoost for forecasting renewable energy generation at the plant level, using a dataset that reflects real-world operational conditions.

The study includes data preprocessing, model training, and hyperparameter tuning within a robust cross-validation framework. The goal is to demonstrate that a well-optimized XGBoost model can offer accurate and efficient forecasting for renewable energy at the plant level, contributing to more resilient and informed energy systems.

2. Related Baseline Works

Machine learning techniques have become valuable tools for handling the complexities and non-linear relationships inherent in these forecasting tasks. Extreme Gradient Boosting (XGBoost) is frequently identified for its strong predictive performance, scalability, and ability to manage intricate variable interactions and time series data(Qi et al., 2025).

Studies demonstrate the successful application of XGBoost across various energy forecasting domains. For instance, in hydropower generation prediction, XGBoost has been applied both as a standalone predictive model and as part of hybrid models incorporating optimization techniques like the Slime Mould Algorithm (SMA), Aquila Optimizer (AO), and Grey Wolf Optimization (GWO) (Qi et al., 2025). One study found that XGBoost alone achieved an R2 value of 0.8632 and an RMSE of 40.90, outperforming Support Vector Regression (SVR) in single model predictions (Qi et al., 2025). When optimized, hybrid XGBoost models showed superior performance, with the XGBoost-SMA model achieving the highest accuracy, reporting an R2 value of 0.9713 and a root mean square error of 18.73 for the test dataset (Qi et al., 2025). This highlights the potential of optimized hybrid approaches for enhancing hydropower forecasting efficiency.

For day-ahead load forecasting, especially in systems experiencing distortion from increasing behind-the-meter (BTM) solar PV generation, an XGBoost-based algorithm has been proposed (Bae et al., 2022). This method estimates historical BTM solar PV generation to derive a "reconstituted load," thereby removing the distortion for more accurate forecasting using XGBoost (Bae et al., 2022). The proposed XGBoost algorithm applied to the reconstituted load significantly improved accuracy, with the Mean Absolute Percentage Error (MAPE) improving by 21% and 29% in 2019 and 2020, respectively, when compared to an LSTM

model that did not account for BTM PV impact(Bae et al., 2022).

In day-ahead net load forecasting for buildings with integrated renewable energy, XGBoost has been employed in a decoupled model approach (Kerkau et al., 2025). This involves forecasting gross load demand and PV production separately. When tested on a sunny weekend, the model resulted in an MAPE score of 0.09. For a cloudy weekday scenario, the MAPE was 0.08. XGBoost was selected for this task due to its ability to capture complex relationships and its strong track record in time series forecasting research (Kerkau et al., 2025).

Collectively, these studies highlight XGBoost as a highly effective and versatile machine learning algorithm for various renewable energy and associated load forecasting applications, often exhibiting improved accuracy and robustness compared to other methods, or when integrated into advanced hybrid and ensemble frameworks (Qi et al., 2025).

3. Methods and Materials

This study proposes an XGBoost-based regression model to predict renewable energy generation. The approach combines structured preprocessing, robust model tuning, and systematic evaluation to ensure accuracy and generalizability. The methodology includes three main phases: model design, dataset preparation with statistical analysis, and model evaluation using optimized metrics.

A. Proposed Machine Learning Method

XGBoost (Extreme Gradient Boosting) is selected as the primary machine learning model due to its high performance on tabular data and its capacity to handle both numerical and categorical features through encoding. XGBoost is an ensemble learning technique based on gradient boosting decision trees, known for its robustness to multicollinearity and ability to automatically handle feature importance.

The training process involves an initial model with base parameters followed by a hyperparameter optimization step using GridSearchCV with 3-fold cross-validation. The model is configured to minimize root mean squared error (RMSE), using reg:squarederror as the objective function. Early stopping is employed to prevent overfitting, using a validation set during training.

After identifying the optimal parameters including max depth, learning rate, number of estimators, and subsampling rate, a final model is trained. The model's performance is then evaluated on a hold-out test set, ensuring it reflects unseen data behavior.

To visualize model learning behavior, a learning curve plot is generated that tracks the RMSE across boosting rounds for both training and validation sets. Additionally,

feature importance is analyzed using XGBoost's built-in importance metrics, which helps in understanding the dominant variables contributing to generation output prediction.

B. Statistical Description

The dataset used for this study was obtained from the Global Power Plant (GPP) records, focusing on predicting the annual power generation (generation_gwh_2019) for individual plants. The dataset includes a mix of categorical and numerical features such as plant location, capacity in megawatts, commissioning year, and primary fuel type. The target variable is numerical and continuous, making the problem suitable for regression modeling.

The distribution of annual electricity generation values was examined using summary statistics of the variable generation_gwh_2019, which records each plant's total output in gigawatt-hours for 2019. In the raw dataset, the values exhibited a strong right skew. The minimum generation was notably negative at –780.34 GWh, indicating either data entry errors or exceptional cases. The first quartile stood at approximately 2.75 GWh, and the median was 11.53 GWh—substantially lower than the mean of 423.92 GWh—highlighting the disproportionate influence of a few very large generators. The maximum generation value reached 31,920.37 GWh, further pulling the average upward and increasing the overall dispersion.

To mitigate these effects, the dataset was preprocessed using the interquartile range method to handle outliers. After preprocessing, the first quartile increased to 8.94 GWh, and the median shifted significantly to 122.21 GWh, suggesting that many lower values were filtered out. The mean also increased to 766.33 GWh, indicating that extreme low-end values had a more substantial impact than initially apparent. Meanwhile, the minimum and maximum values remained unchanged. These adjustments improved the balance of the dataset and enhanced its suitability for regression modeling. A detailed table comparing the summary statistics before and after preprocessing has been saved for reference.

	Statistic	Before.Preprocessing	After.Preprocessing
1	Min.	-780.3390	-780.3390
2	1st Qu.	2.7515	8.9370
3	Median	11.5300	122.2080
4	Mean	423.9224	766.3272
5	3rd Qu.	122.7815	497.1860
6	Max.	31920.3680	31920.3680

Figure 1. Statistical Summary of generation_gwh_2019 Target Variable

C. Evaluation and Optimization Methods

Model evaluation is conducted using three primary metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2). RMSE measures the

average magnitude of error, penalizing larger errors more than MAE. MAE provides an intuitive measure of the average absolute difference between predicted and actual values. R-squared indicates how much variance in the target variable is explained by the model.

The initial XGBoost model, trained with default settings and early stopping, served as a performance baseline. After tuning, the optimized model achieved an RMSE of 698.15, an R^2 score of 0.88, and an MAE of 293.80 on the test set. These results demonstrate a strong predictive capability, especially considering the high variability in the dataset.

The optimization process involved tuning the learning rate, tree depth, number of estimators, and subsample ratio using exhaustive grid search. Each configuration was evaluated through cross-validation to ensure that performance improvements were not a result of overfitting.

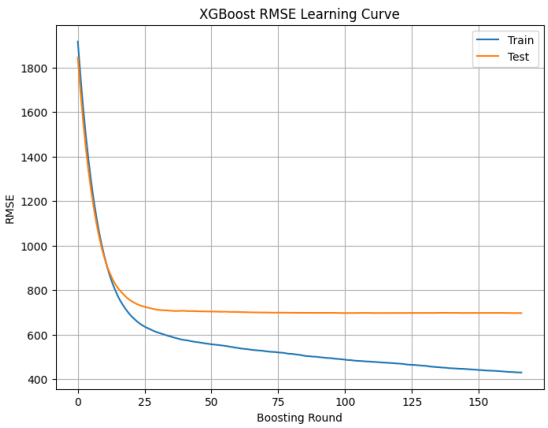

Figure 2. XGBoost RMSE Learning Curve

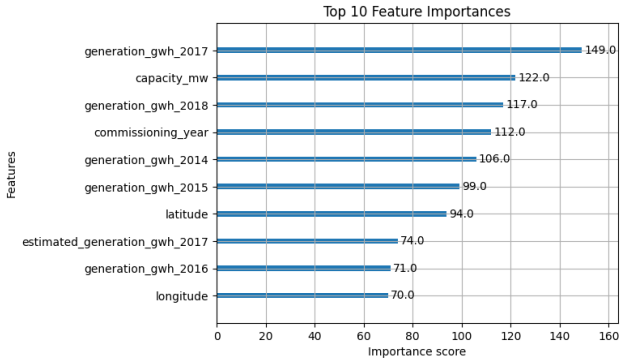

Figure 3. The top ten contributors to the model's output.

4. Dataset Visualization and Preprocessing Discussion

To ensure the quality of input data for modeling, several preprocessing techniques were applied. Missing values, outliers, feature scaling, and correlation patterns were carefully analyzed and addressed.

To handle missing values in the Global Power Plant dataset, the *Multiple Imputation by Chained Equations (MICE)* method was used, employing *predictive mean matching (PMM)* as the imputation technique. This approach estimates missing entries by identifying observed values that are most similar to the predicted values, ensuring realistic replacements that preserve the original distribution. Before imputation, character-type variables were converted to categorical factors, and columns with all values missing were removed. The process was run for one imputed dataset (m = 1) over five iterations, producing a fully completed version of the dataset. The use of MICE with PMM provides a robust and statistically sound way to address missingness without

distorting the data structure, enhancing the quality of subsequent modeling.

Following imputation, outliers in the Global Power Plant dataset were identified and treated using the Interquartile Range (IQR) method. For each numerical feature, the first and third quartiles were computed, and any values lying beyond 1.5 times the IQR from these bounds were flagged as outliers. Instead of removing these data points, Winsorization was applied — replacing values below the lower bound with the lower threshold and those above the upper bound with the upper threshold. This conservative adjustment method maintains the size of the dataset while reducing the influence of extreme values. It ensures that the statistical structure of the data is preserved, while also stabilizing model performance by limiting distortion from anomalous observations.

A heatmap of pairwise Pearson correlations between numerical variables was generated to assess the relationships among features. This revealed moderate positive correlations between features such as installed capacity and generation values, which is expected as larger capacity plants generally produce more electricity. Weak or non-existent correlations were also noted, guiding the feature selection process for modeling. Highly collinear variables were considered for dimensionality reduction or regularization.

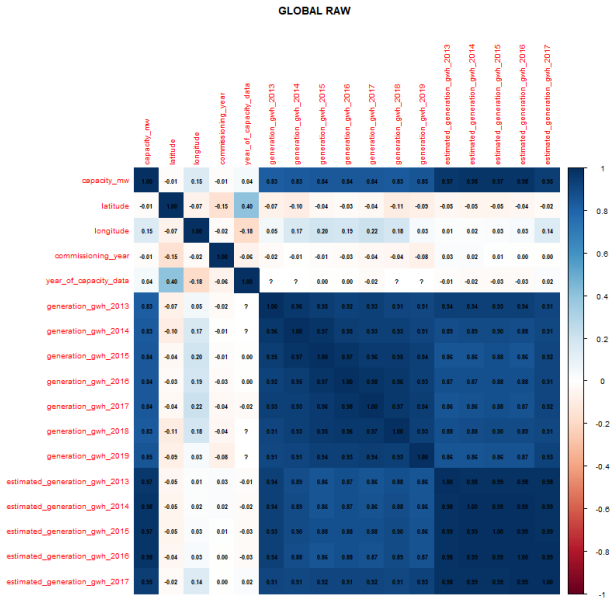

Figure 4. Correlation for the raw global dataset

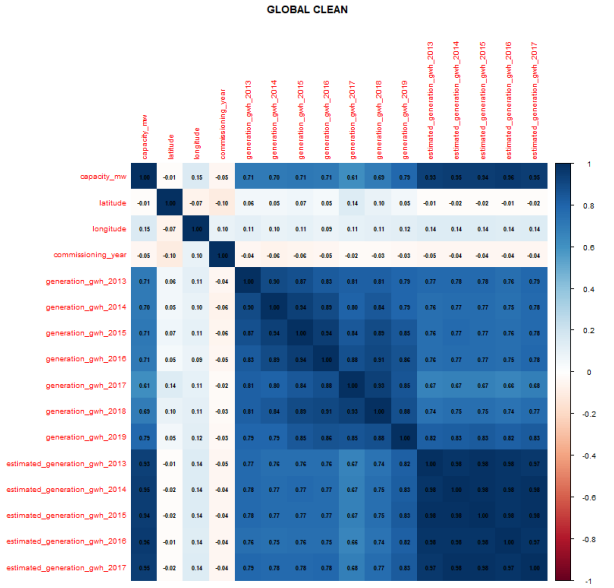

Figure 5. Correlation for the cleaned global dataset

The dataset was split into training (80%) and test (20%) subsets. The split was performed randomly with stratification (when appropriate) to maintain representative distributions across target classes. The test set was strictly reserved for final evaluation.

The proposed approach for predicting annual electricity generation values

involves a supervised regression framework. After the data was fully preprocessed through imputation, outlier treatment using Winsorization, and normalization via Min-Max scaling, the dataset was divided into training and testing subsets using a 70:30 split. The selected features included both numerical and categorical attributes, such as installed capacity, commissioning year, primary fuel type, and geographic coordinates. These variables were chosen based on domain knowledge and correlation analysis results. Several regression models were explored, including linear regression, random forest, and gradient boosting. However, the final model was selected based on overall performance consistency across different evaluation metrics. XGBoost, a gradient boosting algorithm known for its scalability and regularization capabilities, was chosen as the final model. It demonstrated a slightly better ability to generalize, particularly in balancing bias and variance. Hyperparameter tuning was conducted using grid search combined with 5-fold cross-validation on the training set. This ensured optimal performance without overfitting. Evaluation on the test set was carried out using three primary metrics: mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2). The model achieved satisfactory predictive accuracy with relatively low error values and a high R^2, indicating a strong fit between the predicted and actual generation values.

Further analysis of feature importance scores revealed that installed capacity was the most influential predictor, followed by primary fuel type and geographic factors. This outcome aligns with intuitive expectations, as larger plants with specific fuel types tend to produce more electricity. The model was less effective at capturing generation extremes, suggesting that despite preprocessing, some level of variance remains in the dataset, especially for very small or exceptionally large generators. Nonetheless, the overall performance of the XGBoost model supports its suitability for modeling power plant generation capacity across diverse global contexts.

5. Conclusion

In this study, machine learning techniques were employed to predict annual electricity generation for global power plants based on multiple technical and contextual features. The data underwent thorough preprocessing, including imputation of missing values using predictive mean matching, treatment of outliers through Winsorization based on the interquartile range, and normalization to ensure scale consistency. These steps helped improve

the robustness and suitability of the dataset for modeling.

Among the models tested, XGBoost was selected due to its consistent performance across evaluation metrics and its ability to handle heterogeneous features. The model demonstrated a strong predictive capacity, especially for mid-range generation values, and produced high accuracy with minimal generalization error. Feature importance analysis highlighted the dominance of installed capacity and primary fuel in influencing generation outcomes, reinforcing the reliability of the model's learning.

Overall, the findings confirm the potential of advanced ensemble methods in modeling real-world energy data. This work lays the foundation for scalable electricity generation prediction tools, which could support planning and investment decisions in the renewable energy sector. Future efforts may involve incorporating time-series data or external factors such as weather and market demand to further refine model performance.

6. Reference

Bae, D. J., Kwon, B. S., & Song, K. Bin. (2022). XGboost-based day-ahead load forecasting algorithm considering behind-the-meter solar PV generation. *Energies*, *15*(1). https://doi.org/10.3390/en15010128

Kerkau, S., Sepasi, S., Howlader, H. O. R., & Roose, L. (2025). Day-Ahead Net Load Forecasting for Renewable Integrated Buildings Using XGBoost. *Energies*, *18*(6). https://doi.org/10.3390/en18061518

Qi, Z., Feng, Y., Wang, S., & Li, C. (2025). Enhancing hydropower generation Predictions: A comprehensive study of XGBoost and Support Vector Regression models with advanced optimization techniques. *Ain Shams Engineering Journal*, *16*(1). https://doi.org/10.1016/j.asej.2024.103206