Xgboost-Based Prediction of Net Electrical Energy Output of Combined Cycle Power Plants and Model Interpretability Study

Yibo Shang

School of Computer Science and Engineering, Northeastern University, Shenyang, China, 110169

20235769@stu.neu.edu.cn

Abstract. Combined cycle power plants occupy an important position in modern power systems due to their energy efficient and environmentally friendly characteristics. Accurate prediction of electrical energy output faces significant challenges due to complex changes in energy demand and environmental conditions. Existing methods are difficult to effectively capture the nonlinear relationship between the input variables and the output, and the interpretability of the model is poor. In this study, a prediction model based on the XGBoost algorithm is constructed and optimized for the prediction of net electrical energy output of combined cycle power plants. In addition, this study also compares two models, Random Forest and Support Vector Machine, and shows through experimental results that XGBoost is better than the other two models in terms of prediction accuracy, robustness and computational efficiency. Specific experimental results show that the root mean square error (RMSE) of the XGBoost model is 0.17, the explainable variance value (EVS) is 0.97, the mean absolute error (MAE) is 0.13, the mean square error (MSE) is 0.03, and the goodness-of-fit (R^2) is 0.97. XGBoost performs better in all these metrics compared to the random forest and support vector machine models. To solve the opacity problem in the model prediction process, the SHAP framework is introduced to perform global and local interpretability analysis of the model, which effectively reveals the effects of ambient temperature (AT), ambient pressure (AP), relative humidity (RH) and exhaust vacuum (V) on the power output. The XGBoost model combined with the SHAP analysis not only improves the accuracy of the prediction, but also enhances the model's transparency and practicality, which provides strong decision support for the operation optimization of the power plant.

Keywords: XGBoost, combined cycle power plant, net electrical energy output prediction, SHAP model interpretability analysis.

1. Introduction

Combined-cycle power plants play a vital role in meeting global energy demand and reducing carbon emissions due to their high efficiency and environmental benefits. Their share of the electricity supply is increasing as energy demand grows and environmental protection requirements increase. However, accurately predicting the net electrical energy output of power plants faces a significant challenge as it is subject to complex fluctuations in energy demand and changes in environmental conditions. This challenge is not only related to the operational efficiency and economy of power plants, but also important for optimizing scheduling strategies, reducing operational costs, and achieving intelligent and refined power management.

In the domestic and international scientific research field, machine learning applied to electrostatic energy output prediction has been explored by many scholars and achieved remarkable results. For example, scholars Zhou Yu et al [1] use PSO-BiLSTM prediction model optimized based on particle swarm algorithm to predict the output power of combined cycle power plant, and the experimental results show that the root mean square error, average absolute percentage error and other indexes of PSO-BiLSTM model are better than the typical algorithms listed as well as optimization combination algorithms, and the model has higher accuracy in predicting power load data. accuracy. Scholar Shao Kejia [2] used the LSTM_Adaboost model to predict the output power of combined cycle power plants, and the results show that the LSTM_Adaboost model can effectively reduce the prediction error and further improve the accuracy of short-term prediction of the output power of circulating power plants, which is of guiding significance for the formulation of operation plans and increasing

economic benefits of the power plants. Scholars Ceyhun Yildiz and Hakan Acikgoz [3] proposed a neural network based on kernel-limit machine learning for predicting the short-term power output of grid-connected photovoltaic and nuclear power plants, and through the comparison of the Extreme Learning Machine (ELM) and the Kernel-Limit Learning Machine (KELM), it was confirmed that the KEL has a better generalization ability compared to the ELM, and the prediction system of the KELM has better generalization ability in the winter and spring improved the R-values of ELM and LM by an average of 3.94% to 35.38% and 2.90% to 9.81%. Scholar Joao Gari da Silva Fonseca Jr et al [4] used support vector regression and numerical prediction clouds to predict the power generation of photovoltaic power plants in Kitakyushu, Japan, and the predicted results had a root mean square error of 0.0948 MWh and an average absolute error of 0.058 MWh. Scholars Anastasia G. Rusina et al [5] used decision tree integration algorithm to predict solar power generation in Mongolian power system and the results of the study had a normalized absolute percentage error of 6.5 to 8.4%. In the present time, the research on the prediction of electrostatic energy output of power plants using machine learning methods still needs to be further improved, and avoiding overfitting and improving the generalization ability of the model during the model training process are still the focus of the research. The dataset used in this study contains 9569 sets of samples, which effectively solves the overfitting problem caused by small samples, and verifies the prediction accuracy and practicability of the model through field research and data collection, which improves the effectiveness of the machine learning model in practical applications. This study applies the more advanced XGBoost model [6-8] to the prediction of complex phenomena and compares it with the random forest model [9-11] and the support vector machine model [12-14], with the expectation that by improving the accuracy and speed of prediction, it will further promote the application and development of the machine learning technology in various subject areas.

In this study, a prediction model is constructed based on the XGBoost algorithm, and the SHAP [15] (Shapley Additive Explanations) framework is introduced to analyze the interpretability of the model at both global and local levels. The study firstly organizes and analyzes the historical data of combined cycle power plants, constructs the XGBoost model with multi-factor inputs, and improves the model performance through cross-validation and parameter tuning. The experimental results show that the XGBoost model performs excellently in predicting the net power output, with a root mean square error (RMSE) of 0.17, an explainable variance value (EVS) of 0.97, a mean absolute error (MAE) of 0.13, a mean square error (MSE) of 0.03, and a goodness-of-fit (R^2) of 0.97, which is superior in all of these metrics compared to the Random Forest [1] and Support Vector Machine models. performs better in all these metrics. Combined with SHAP analysis, this study effectively reveals the effects of ambient temperature (AT), ambient pressure (AP), relative humidity (RH), and exhaust vacuum (V) on the electrical energy output, improves the transparency and practicability of the prediction model, and provides an important basis for the operation optimization and intelligent management of power plants.

This study is of great significance for improving the operation efficiency and realizing intelligent management of combined cycle power plants. Through the in-depth analysis in this study, we not only improve the accuracy of the prediction model, but also significantly enhance the interpretability of the model by introducing the SHAP framework, which provides powerful decision support for power plant managers.

Figure 1 illustrates the flowchart of the problem analysis in this study, describing in detail the various steps from data preprocessing to model evaluation. This flowchart demonstrates the methodological rigor and systematicity of this study, and provides a clear roadmap for subsequent research.

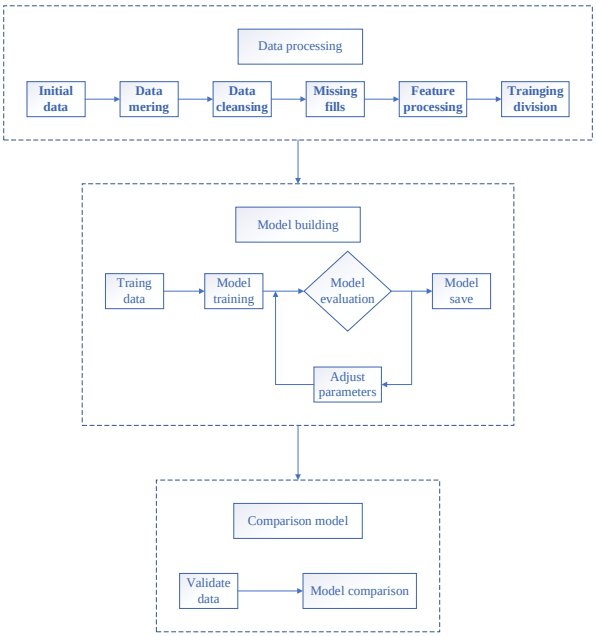

Figure 1. Problem analysis flowchart

2. Analysis of data sets and research methods

2.1. Data sources and analysis

The data for this study were obtained from https://paperswithcode.com/datasets. For the purpose of this study, several key environmental and operational parameters relevant to the operation of a combined cycle power plant were carefully selected:

(1) Ambient Temperature (AT): the ambient temperature around the power plant, which is an important factor affecting the thermal efficiency of the plant and the performance of the power generation. When subdividing the statistical characteristics, this study found the mean to be 19.65123; the median to be 20.345; the maximum value to be 37.11; and the minimum value to be 1.81.

(2) Ambient Pressure (AP): Changes in atmospheric pressure may affect the amount of air intake into the power plant, which in turn affects combustion efficiency and energy conversion. When subdividing the statistical characteristics, this study found the mean to be 54.3058; the median to be 52.08; the maximum value to be 81.56; and the minimum value to be 25.36.

(3) Relative Humidity (RH): Humidity levels have a direct impact on fuel combustion efficiency and heat exchange processes. When subdividing the statistical characteristics, this study found the mean to be 1013.259; the median to be 1012.94; the maximum value to be 1033.3; and the minimum value to be 992.89.

(4) Exhaust Vacuum (V): Vacuum in the exhaust system can reflect the energy loss of a power plant, and is a key indicator for optimizing the efficiency of power generation. When subdividing the statistical characteristics, this study found the mean to be 73.30898; the median to be 74.975; the maximum value to be 100.16; and the minimum value to be 25.56.

(5) Net Electrical Energy Output (PE): The actual electrical energy output of the power plant for a given period of time, which is the target variable of our prediction model. When subdividing the statistical characteristics, this study found the mean to be 454.365; the median to be 451.55; the maximum value to be 495.76; and the minimum value to be 420.26.

In order to ensure the accuracy of the data and the validity of the analysis, the collected data were thoroughly preprocessed in this study, including missing value processing, outlier detection and data normalization. Missing values were supplemented by interpolation methods, while outliers were identified and corrected by statistical analysis methods. The data standardization process is to eliminate the influence of different scales and ensure the balanced weight of each feature during the

model training process. After completing the preprocessing, this study used statistical analysis methods to conduct a descriptive analysis of the data and plotted Figure 2 by visual means to help understand the basic features of the data, which provides an important basis for the subsequent model construction and feature selection.

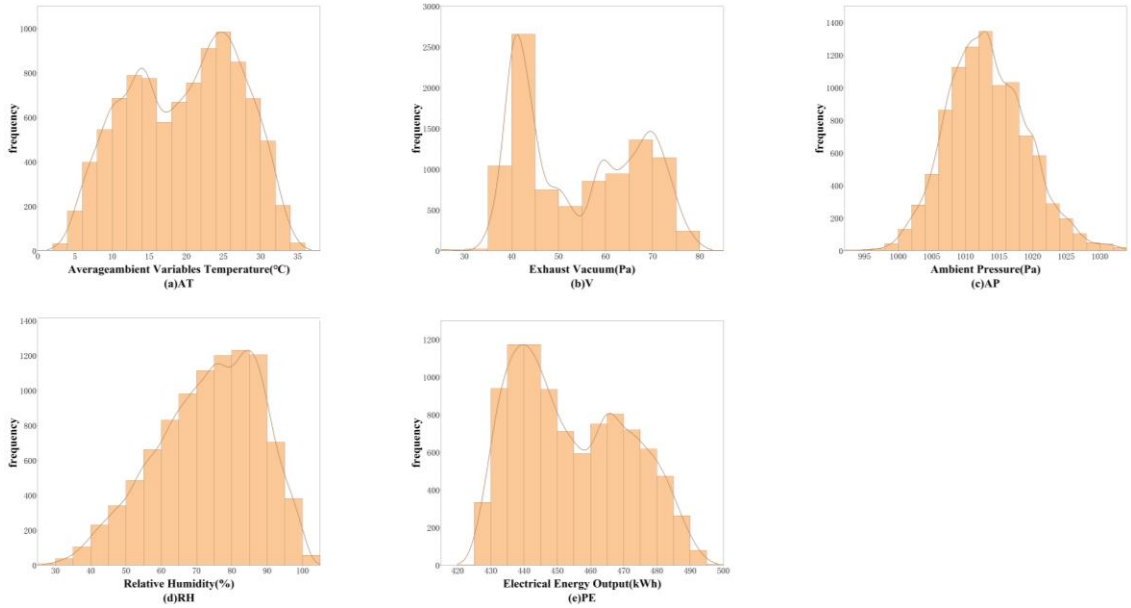

Figure 2. Data distribution map

2.2. Model analysis

2.2.1 Random Forests

Random Forest is an integrated learning method based on decision trees, which improves the predictive performance and robustness of the model by constructing multiple decision trees and integrating them. Random Forest performs well in handling high-dimensional data, nonlinear relationships and feature selection, and has strong generalization ability. Its underlying mathematical logic is based on the decision tree construction process: (1) Decision tree: the decision tree is constructed by recursively partitioning the dataset, with each partition based on the principle of maximizing information gain (or Gini impurity). (2) Randomness: Random Forest introduces randomness in the construction of each tree, including random selection of samples (bootstrap sampling) and random selection of feature subsets. (3) Integration: the final prediction is derived from all decision trees by majority voting (classification) or averaging (regression). It reduces the risk of overfitting by introducing randomness into the construction of each tree, making each tree differentiated in feature selection and sample selection. The information gain is calculated as follows:

$$Information\ Gain(D, A) = Entropy(D) - \sum_{v \in values(A)} \frac{|D_v|}{|D|} Entropy(D_v) \quad (1)$$

Where, D is the dataset, A is the attribute and D_v is the subset of data that has value v under attribute A.

Before model training, hyperparameter optimization was performed using a grid search to determine the best model configuration, using cross-validation (CV=3) and negative mean square error as scoring criteria. The optimal hyperparameter combination was obtained through grid search:

Table 1. Optimal hyperparameter combinations for random forest models

Hyperparameter name	optimum value
n_estimators	300
max_depth	10
min_samples_split	2
min_samples_leaf	1
bootstrap	True
max_features	'log2'

2.2.2 Support vector machines

Support Vector Machine (SVM) is a widely used machine learning algorithm for classification and regression tasks. In regression problems, SVM finds an optimal segmentation plane by maximizing the intervals between data points, a method known as Support Vector Regression (SVR):

(1) Interval Maximization : SVM tries to find a hyperplane that maximizes the interval between different categories.

(2) Regularization: In order to prevent overfitting, SVM introduces a regularization parameter to balance the interval maximization and classification error.

(3) Kernel trick: SVM maps the data to a high-dimensional space by means of a kernel function, which simplifies the computation by eliminating the need to compute the mapping explicitly.

Before performing SVM modeling, we first perform data preprocessing on the historical data of the power plant, including missing value processing, outlier detection, and data normalization. These steps ensure the consistency and completeness of the input variables (e.g., ambient temperature (AT), ambient pressure (AP), relative humidity (RH), exhaust vacuum (V)) and the target variable (net electrical energy output (PE)). 80% and 20% of the data are divided into training and test sets for model training and model evaluation, respectively.

Mathematically, it can be achieved by solving the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \tag{2}$$

$$s.t. y_i - (w \cdot x_i + b) \le \varepsilon + \xi_i, \xi_i \ge 0 \tag{3}$$

Where w and b are hyperplane parameters, ξ_i is a slack variable, C is a regularization parameter, and ε is the epsilon in the loss function.

In order to find the best model configuration, we optimized the hyperparameters of the SVR model using the Grid Search method with cross-validation (CV=3) and negative mean square error as scoring criteria. Through Grid Search, we obtained the optimal combination of hyperparameters:

Table 2. Optimal hyperparameter combinations for support vector machine models

Hyperparameter name	optimum value
C	10
epsilon	0.2
kernel	‘rbf’
gamma	‘scale’
degree	2
shrinking	True
tol	0.001

2.2.3XGBoost

XGBoost (eXtreme Gradient Boosting) is an integrated learning algorithm based on gradient boosting, which performs well in many machine learning tasks, especially in the prediction of tabular data.XGBoost improves the performance of a model by constructing multiple decision trees and optimizing the predictions for the next round using the predictions from the previous round of

iterations; gradient Boost: XGBoost minimizes the gradient of the loss function by adding new trees at each step:

$$w_t = -\eta \cdot grad(L(w_{t-1})) \tag{4}$$

Where w_t is the model parameters of round t, η is the learning rate, and $grad(L(w_{t-1}))$ is the gradient of the loss function with respect to the current model parameters.

Before performing XGBoost modeling, we first performed data preprocessing on the historical data of the power plant, including missing value processing, outlier detection and data normalization. These steps ensure the consistency and completeness of the input and target variables. In order to find the best model configuration, we optimized the hyperparameters of the XGBoost model using the Grid Search method with cross-validation (CV=3) as a scoring criterion. Through Grid Search, we obtained the best combination of hyperparameters:

Table 3. Optimal hyperparameter combinations for the XGBoost model

Hyperparameter name	optimum value
learning_rate	0.1
max_depth	6
n_estimators	275
subsample	0.9
colsample_bytree	0.8
gamma	0
reg_lambda	1.5

3. Research process and results

3.1. Research process

In this study, historical data from combined cycle power plants were first thoroughly preprocessed, including missing value processing, outlier detection, and data normalization to ensure data consistency and completeness. Three different machine learning models; Random Forest, Support Vector Machine, and XGBoost, were then employed to predict the net electrical energy output.

To optimize the performance of each model, this paper implements grid search and cross-validation to determine the best combination of hyperparameters. Through this process, optimal configurations were found for each model and the models were trained on a normalized training set using these configurations.

3.2. Simulation results

The performance of the models was assessed using several metrics:

(1) Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{\mathrm{n}}\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2} \tag{5}$$

Where y_i is the true value, $\widehat{y_i}$ is the predicted value, and n is the sample size.

(2) Explainable Variance Value (EVS):

$$EVS = 1 - \frac{Var\left(y - y\right)}{Var\left(y\right)} \tag{6}$$

Where $Var\left(y\right)$ is the variance of the true value and $Var\left(y - y\right)$ is the variance of the prediction error.

(3) Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y_i}\right| \tag{7}$$

(4) Mean Square Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2 \tag{8}$$

(5) Goodness of Fit (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2} \tag{9}$$

Where $\overline{y}$ is the mean of the true values.

Data for these specific metrics across the three models can be found in Table 7.

Table 4. Performance metrics data for each model

Model name	RMSE	EVS	MAE	MSE	R^2
random forest	0.20	0.96	0.15	0.04	0.96
support vector machine	0.23	0.95	0.18	0.05	0.95
XGBoost	0.17	0.97	0.13	0.03	0.97

In order to present a comprehensive and intuitive picture of the prediction performance of each model, this paper draws three charts, including scatter plots, data fitting plots and error plots, which are used to analyze and compare in depth the performance of Random Forest, Support Vector Machine and XGBoost models. These graphs not only show the correspondence between model predictions and actual values, but also reveal the ability and accuracy of each model in approximating the real data.

The scatterplot in Figure 3 visually depicts the comparison between model predictions and actual observations. In an ideal predictive model, we would expect to see all data points tightly distributed around the contour line, which indicates that the model's predicted values are highly consistent with the actual values, thus demonstrating the model's predictive accuracy. With this intuitive visualization method, we can easily identify any deviations or trends in the model's predictions and thus assess the reliability of the model in real-world applications.

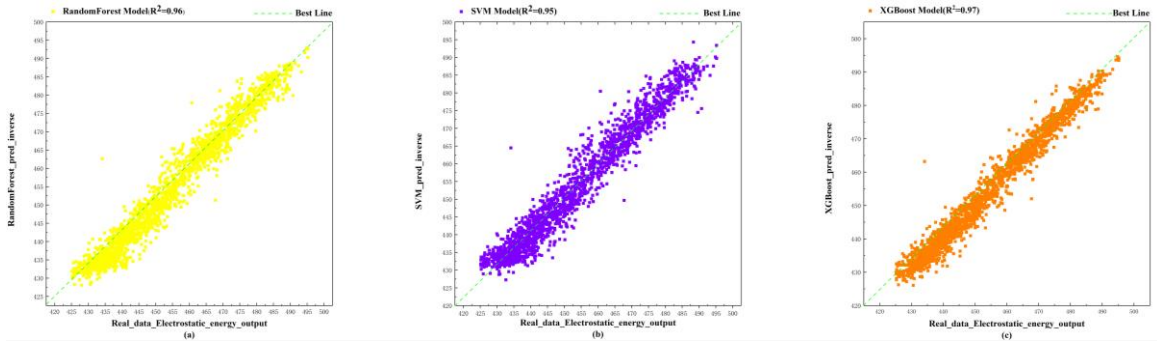

Figure 3. Scatterplot of predicted values for each model

Figure 4 vividly demonstrates how well the model fits the training data, where the close correspondence between the model and the data points highlights its ability to sensitively capture data patterns and trends. The good fit not only proves the superior performance of the model on the training set, but also hints at its strong generalization potential on unknown data.

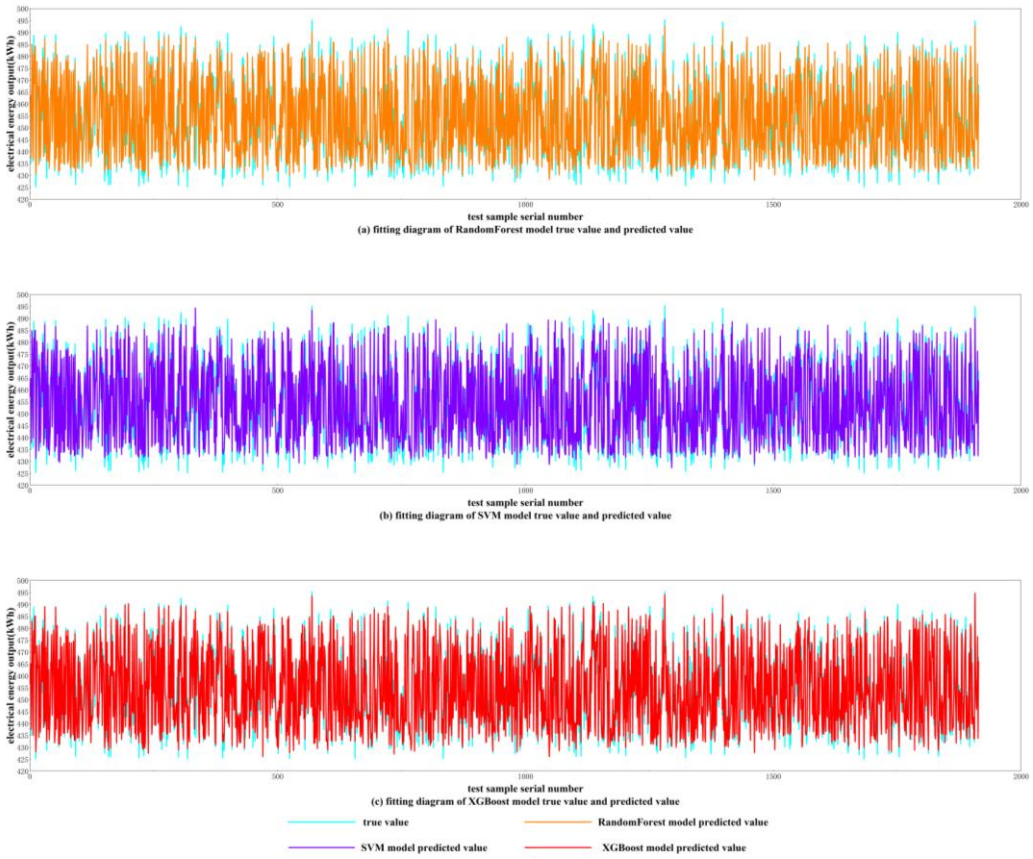

Figure 4. Data fit of predicted values to true values for each model

Figure 5 visually reveals the difference between the model predicted values and the actual observed values, and through the representation of error plots, we can clearly see an intuitive measure of the model's prediction accuracy. In Figure 5, each deviation is quantified and presented, thus providing us with a direct view of assessing the accuracy of the model. The smaller error represents the closer the model's prediction results are to the actual situation, thus reflecting the model's efficient prediction performance in real-world applications. This high degree of accuracy is an important indicator of the model's reliability and demonstrates the model's superior ability to process data and predict future trends.

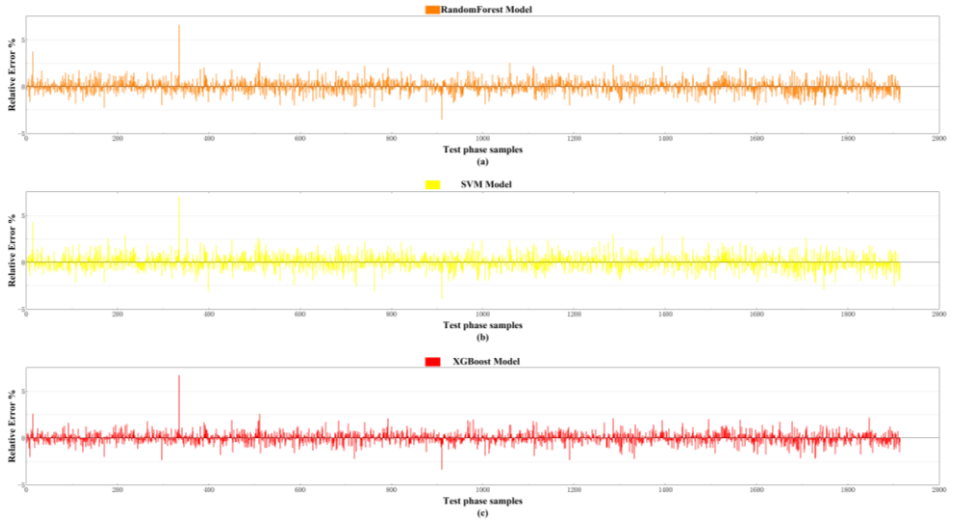

Figure 5. Error plots between predicted and true values for each model

4. SHAP model interpretability analysis

4.1. SHAP model analysis

In order to enhance the interpretability of the model and reveal its inner workings, this study introduces the state-of-the-art SHAP (Shapley Additive Explanations) framework. SHAP analysis is a powerful tool for model prediction interpretation, which quantifies the specific contribution of each feature to the model's predicted outcomes by assigning a numerical value (SHAP value) to that feature.

The SHAP values are computed based on cooperative game theory, where each feature is considered a player contributing to the overall prediction. The calculation involves the following steps:

(1) Coalition Formation: For each prediction, consider all possible combinations of features (coalitions) and compute the model output with and without each feature.

(2) Marginal Contribution: For each feature, calculate its marginal contribution, defined as:

$$\Phi_i(f)=\sum_{S\subseteq N\setminus\{i\}}\frac{|S|!(|N|-|S|-1)!}{|N|!}\left(f\left(S\cup\{i\}\right)-f\left(S\right)\right) \tag{10}$$

Where $\Phi_i(f)$ is the SHAP value for feature i, N is the set of all features, and S is a subset of features not including i.

(3) Shapley Value Calculation: Aggregate these marginal contributions across all possible coalitions to derive the SHAP value for each feature.

These values reflect the relative importance of each feature in the model's decision-making process, allowing us to visualize which factors play a key role in driving model output.

4.2. Interpretability analysis

4.2.1 Scatter plot analysis of SHAP value distribution

Fig. 6 shows the distribution of SHAP values for each feature on the output prediction. In Fig. 6, the horizontal axis represents the SHAP values and the vertical axis lists the features (AT, V, AP, RH) used in the model. Figure 6 also shows the SHAP values for each feature at different data points. In this case, the gradient of the color indicates the range of feature values: more skewed red represents higher feature values, while more skewed blue represents lower feature values. From the distribution of SHAP values of each feature in Fig. 6, we can understand that: the SHAP values of ambient temperature (AT) show a wider distribution with both positive and negative effects: when the SHAP values are shown as negative (blue area), it implies that lower ambient temperatures are favorable to increase the net electrical energy output, while the positive values (red area) indicate that, under certain conditions, an increase in ambient temperatures may have an adverse effect on the net electrical energy output adversely. On the other hand, the SHAP values for exhaust vacuum (V) are mostly positive, implying that an increase in exhaust vacuum is usually associated with an increase in net electrical energy output. Compared to AT and V, the SHAP values of ambient pressure (AP) and relative humidity (RH) are more concentrated and less fluctuating, suggesting that they have a more limited impact on the model predictions.

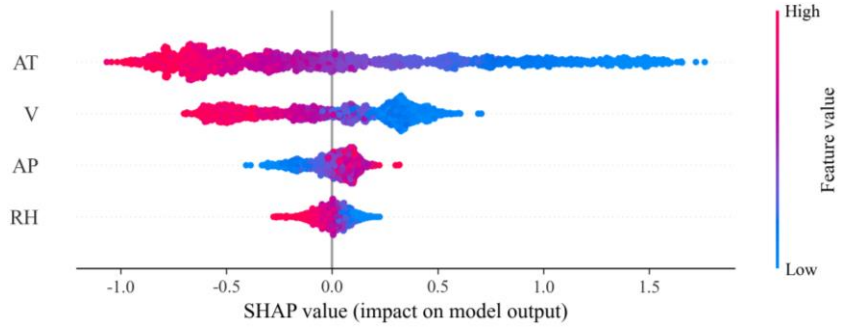

Figure 6. Scatter plot of SHAP value distribution

4.2.2 Influence diagram analysis of SHAP value characteristics

Figure 7 illustrates the magnitude of the average effect of each feature on the model predictions. In Fig. 7, the horizontal axis represents the features considered in the model, including ambient temperature (AT), exhaust vacuum (V), ambient pressure (AP), and relative humidity (RH), and the vertical axis represents the average absolute SHAP value of these features on the model predictions, i.e., it reflects the magnitude of their average contribution to the model output. By analyzing Figure 7, we can observe that ambient temperature (AT) has the highest average absolute SHAP value, indicating that it plays the most critical role in model prediction. This is closely followed by exhaust vacuum (V), implying that it also has a significant effect on the predictions.

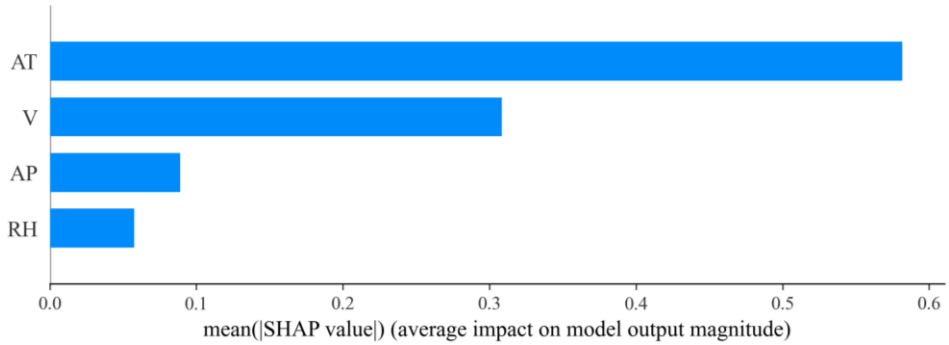

Figure 7. Characteristic impact map of SHAP values

4.2.3 SHAP value interaction plot analysis

Figure 8 shows the interaction effects of the features in the model on the prediction of the net electrical energy output of the combined cycle power plant using interactive visualization charts, which provides insight into the complex relationships that may exist between different features and their combined effects on the model predictions. In Fig. 8, the horizontal and vertical axes represent two interacting features, and the change in color shade indicates the intensity of the interaction effect of the feature on the model prediction results. The analysis of Fig. 8 shows that the interaction between ambient temperature (AT) and exhaust vacuum (V) is particularly significant, and the strong change in color indicates that the combination of these two features has a significant effect on the net electrical energy output prediction. This interaction may imply that at certain specific ambient temperatures, the adjustment of exhaust vacuum is particularly critical for enhancing net electrical energy output, and vice versa.

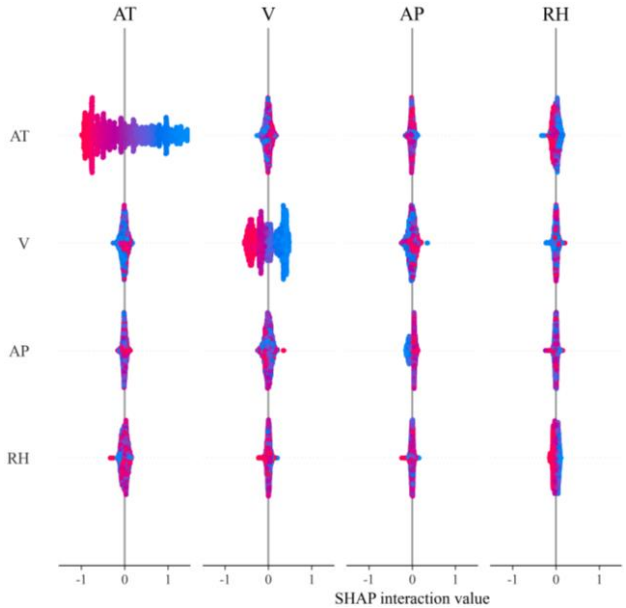

Figure 8. Interaction diagram of SHAP values

5. Conclusions and expectation

5.1. Conclusions

The XGBoost model performs best in all simulation results. The scatterplot reveals XGBoost's strength in prediction accuracy, with almost all points closely surrounding the contour line, indicating that its predicted values are highly consistent with the actual values. The data fitting plot further demonstrates XGBoost's superior ability to capture and model data trends. The error plot also shows that XGBoost has the smallest prediction error, which again demonstrates its significant advantage in prediction accuracy.

The SHAP (Shapley Additive Explanations) framework is introduced in this study, which significantly improves the interpretability of the XGBoost model. Through in-depth analysis of the SHAP values of the features, we are able to identify features that have a significant impact on power generation efficiency, such as ambient temperature (AT) and exhaust vacuum (V), and how they affect the model's prediction results as the feature values change. This intuitive visual analysis not only enhances the transparency of the model, but also provides plant managers with powerful information to optimize key operating conditions, helping to maximize power generation efficiency.

5.2. Expectation

Although this study has achieved some results, there is still room for further improvement and expansion. For example, consider introducing more features related to power plant operations, such as fuel cost, equipment maintenance status, etc., to further improve the accuracy and applicability of the predictions. Explore integrated learning methods that incorporate multiple machine learning algorithms with a view to obtaining better prediction performance. Conduct long-term real-world operational tests to evaluate the model's performance in real-world environments and continuously optimize it based on feedback.

References

[1] Shao Kejia, Zhou Yu, Song Hao, et al. Output power prediction of combined cycle power plant based on PSO optimized BiLSTM [J]. Power Science and Engineering, 2022, 38 (2): 9-17.

[2] Shao Kejia. Based on LSTM___Adaboo... Study on output power prediction of combined cycle power plants_Shao Kejia [J]. 2022.

[3] Yildiz C, Acikgoz H. A kernel extreme learning machine-based neural network to forecast very short-term power output of an on-grid photovoltaic power plant [J]. Energy sources. part A, Recovery, utilization, and environmental effects, 2021, 43 (4): 395-412.

[4] Da Silva Fonseca Jr J G, Oozeki T, Takashima T, et al. Use of support vector regression and numerically predicted cloudiness to forecast power output of a photovoltaic power plant in Kitakyushu, Japan [J]. Progress in photovoltaics, 2012, 20 (7): 874-882.

[5] Rusina A G, Osgonbaatar T, Stepanova A I, et al. Ensemble Machine Learning Model for Day Ahead Solar Power Forecasting for Mongolia Power System [C] //. IEEE, 2023.

[6] Euryong Liu, Xue-De Huang. Influence factor analysis of heart disease based on XGBoost and SHAP [J]. Information and Computer (Theoretical Edition), 2024, 36 (06): 68-70.

[7] Li Jinxia, Bian Huaxing, Wen Fuguo, et al. Performance risk prediction of grid material suppliers based on XGBoost [J]. Computer Science, 2024, 51 (z1): 1174-1182.

[8] Wang Dili. Research on the efficiency of pumping well system based on XGBoost decision tree [J]. Chemical Engineering and Equipment, 2024 (07): 116-119.

[9] Y. Zhang, Y. Chen Tang. Calcium ion concentration prediction model for seawater circulation cooling system based on random forest algorithm [J]. Salt Science and Chemical Engineering, 2024, 53 (7): 19-22, 26.

[10] JIA Haoxin, ZOU Menghua, TANG Xin, et al. Evaluation of construction suitability of geothermal power plant based on random forest method [J]. China Geology: 1-28.

[11] CHENG Zhiyuan, ZHANG Boliang, CAI Yu-chen, et al. Fusion of random forest and SHAP for heart disease prediction and its characterization [J]. Intelligent Computer and Applications, 2023, 13 (11): 172-179.

[12] Yang S. Deflection prediction of main girder of rigid bridge based on optimized support vector machine [J]. Hunan Transportation Science and Technology, 2024, 50 (2): 105-109.

[13] Luan Bing. Comprehensive evaluation of energy saving and emission reduction effect in industrial parks based on improved support vector machine [J]. Low Carbon World, 2024, 14 (3): 37-39.

[14] HE Ninghui, WU Xutao, ZHANG Pei, et al. Error pattern recognition of transformer online monitoring data based on multi-classification support vector machine [J]. High Voltage Apparatus, 2024, 60 (7): 173-181.

[15] Wu Y. A malicious encrypted traffic prediction model fusing random forest and SHAP [J]. Journal of Harbin University of Commerce (Natural Science Edition), 2024, 40 (2): 167-178.