processes

MDPI

Article

CBM Gas Content Prediction Model Based on the Ensemble Tree Algorithm with Bayesian Hyper-Parameter Optimization Method: A Case Study of Zhengzhuang Block, Southern Qinshui Basin, North China

Chao Yang [1,2,*], Feng Qiu [1,2], Fan Xiao [1,2], Siyu Chen [1,2] and Yufeng Fang [3]

[1] School of Energy Resources, China University of Geosciences, Beijing 100083, China
[2] Coal Reservoir Laboratory of National Engineering Research Center of CBM Development & Utilization, China University of Geosciences, Beijing 100083, China
[3] The Tenth Oil Production Plant of PetroChina Changqing Oilfield Branch Company, Qingyang 745100, China
* Correspondence: 2006200031@cugb.edu.cn

Abstract: Gas content is an important parameter for evaluating coalbed methane reservoirs, so it is an important prerequisite for coalbed methane resource evaluation and favorable area optimization to predict the gas content accurately. To improve the accuracy of CBM gas content prediction, the Bayesian hyper-parameter optimization method (BO) is introduced into the random forest algorithm (RF) and gradient boosting decision tree algorithm (GBDT) to establish CBM gas content prediction models using well-logging data in the Zhengzhuang block, south of Qinshui Basin, China. As a result, the GBDT model based on the BO method (BO-GBDT model) and the RF model based on the BO method (BO-RF model) were proposed. The results show that the mean-square-error (MSE) of the BO-RF model and the BO-GBDT model can be reduced by 8.83% and 37.94% on average less than that of the RF and GBDT modes, indicating that the accuracy of the models optimized by the BO method is improved. The prediction effect of the BO-GBDT model is better than that of the BO-RF model, especially in low gas content wells, and the R-squared (RSQ) of the BO-GBDT model and the BO-RF model is 0.82 and 0.66. The accuracy order of different models was BO-GBDT > GBDT > BO-RF > RF. Compared with other models, the gas content curve predicted by the BO-GBDT model has the best fitness with the measured gas content. The rule of gas distribution predicted by all four models is consistent with the measured gas content distribution.

Keywords: gas content; random forests; gradient boosting decision tree; Bayesian optimization method

check for updates

Citation: Yang, C.; Qiu, F.; Xiao, F.; Chen, S.; Fang, Y. CBM Gas Content Prediction Model Based on the Ensemble Tree Algorithm with Bayesian Hyper-Parameter Optimization Method: A Case Study of Zhengzhuang Block, Southern Qinshui Basin, North China. *Processes* **2023**, *11*, 527. https://doi.org/10.3390/pr11020527

Academic Editors: Wu Li, Junjian Zhang, Tim A. Moore and Raymond Cecil Everson

Received: 26 November 2022
Revised: 26 December 2022
Accepted: 2 February 2023
Published: 9 February 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Coalbed methane (CBM) has become an important component for increasing natural gas storage and production after years of exploration and exploitation [1–4]. As gas content is an important parameter for characterizing coal reservoirs and an important basis for CBM resource and reserve estimation [5] the evaluation of gas content in coal is extremely significant for the exploration and production of coalbed methane [6–8]. The accuracy of gas content prediction is directly related to the estimation of coalbed methane resources and productivity prediction of CBM wells, so determining the gas content of CBM in the reservoirs is still an important problem to solve. The occurrence state of coalbed methane in coal reservoirs is complex due to the special pore-fracture system, which is different from the gas in conventional reservoirs [8,9]. At present, the theory most widely recognized is that CBM is endowed in coal reservoirs in three forms: water-soluble gas in the coalbed water, free gas between pore fractures, and adsorbed gas on the surface of pores [10]. Different occurrence states are transformed due to changes in coal reservoir properties and external geological conditions [11]. Adsorbed gas is the main component of total gas content in medium-high-rank coal seams [6,12,13].

The methods of gas content measurement include the direct method and the indirect method [5], but the gas content testing of coal is often lacking in the progress of CBM development [7], which means that the gas content of the coal seam cannot be determined. Therefore, various methods have been successively proposed based on different theories and methods to predict the CBM gas content, such as the isothermal adsorption method, gas content gradient method, regression method, and parameter calculation method [5,14]. With the rise of artificial intelligence science, machine learning and deep learning theories have been gradually applied to many scientific issues [15], and scholars have combined various algorithmic theories with coal reservoir gas content prediction to broaden the methods of gas content calculation. Artificial intelligence algorithms are mainly used to achieve gas content prediction by constructing mathematical models with nonlinear approximations. Supervised learning algorithms, which need explicit target labels, are widely applied to established models to predict the CBM gas content, such as support vector machines (SVM) [16], artificial neural networks (ANN) [17], random forests (RF) [18], gradient boosting decision trees (GBDT), and the Kernel Extreme Learning Machine [8]. Compared with supervised learning algorithms, unsupervised learning algorithms were employed less frequently. He et al. [19] applied the cluster analysis algorithm combined with the coal reservoir properties to predict gas content. Yu et al. [20] classified coal reservoirs by K-means clustering into different types and then used the random forest algorithm for gas content prediction in different reservoirs. The BP neural network was proved that its prediction effect was significantly better than that of the multiple regression method in the previous. [21,22]. Li et al. [23] predicted CBM gas content in the Linxing block by using four different algorithms, and the results showed that the prediction effect of GBDT was better than several other models. Although artificial intelligence algorithmic models outperformed conventional computational methods in most of the studied regions, the different artificial intelligence algorithmic models need to be adjusted to improve their applicability in different study regions.

Geophysical logging data, as a higher resolution data [24], has been widely used to evaluate the CBM gas content [8]. However, different features in the logging data need to be analyzed and chosen to establish the models because of the complex, nonlinear correlation between well logs and gas content [25]. Compared with the other algorithms, the ensembled tree algorithm adopts a classification and regression tree (CART) at the bottom layer and sets up an effective mechanism for selecting segmentation points, which can achieve the feature selection efficiently. The previous research has proved that the ensembled tree algorithm has the best effect and strong stability in predicting CBM gas content, which has more advantages than the SVM model and ANN model in terms of small sample size and dimension [5,23]. The ensemble method is a machine learning technique that combines several base ML models to produce one optimal predictive model, whose performance is better than the separate machine learning methods [25]; two of the most popular algorithms in these models are the "bagging" method [26] and the "boosting" method [27]. The RF model based on the "bagging" method and the GBDT model based on the "boosting" method were employed in the study to establish the CBM gas content prediction models.

Hyper-parameters are the parameters that must be set before training the machine learning models and cannot be directly estimated from data learning, which defines the architecture of models [28]. Although the RF model and GBDT model have been widely used in CBM reservoir evaluation and production prediction [18,23,25], numerous hyper-parameters in machine learning algorithms that control the accuracy and complexity of the model are often simply adjusted. The optimization problem usually relies on subjective judgment, experience, and trial-and-error methods with a certain degree of randomness and uncertainty [29]. Zhu et al. [30] introduced the genetic algorithm into the random forest CBM production prediction model to optimize the hyper-parameters, but only two hyper-parameters of the random forest were optimized. In this paper, Bayesian optimization is

employed to optimize almost all hyper-parameters of the RF model and GBDT model to make the model achieve optimal prediction.

2. Methodology

2.1. Random Forest Algorithm

Random forest (RF) is a classical parallel ensemble algorithm that incorporates the concept of "bagging" and the random subspace method [31]. In this study, each weak learner in the random forest is a decision tree [32] (Figure 1a), which is independent of each other, and the result of RF is the average aggregation of base learners. This method ensembles several decision trees with a controlled variation using the random selection of variables method [33], and each decision tree is first built using the bootstrap sampling method, which takes samples out randomly with put-back, and the process is carried out n times [32]. One of the advantages of using random forest is their ability to generate accurate outcomes based on a few computational resources [34]. The application of the bootstrap sampling method ensures the randomness of the random forest model. As a result, the random forest has better generalization ability due to multiple randomly built parallel decision trees for weak learners [35].

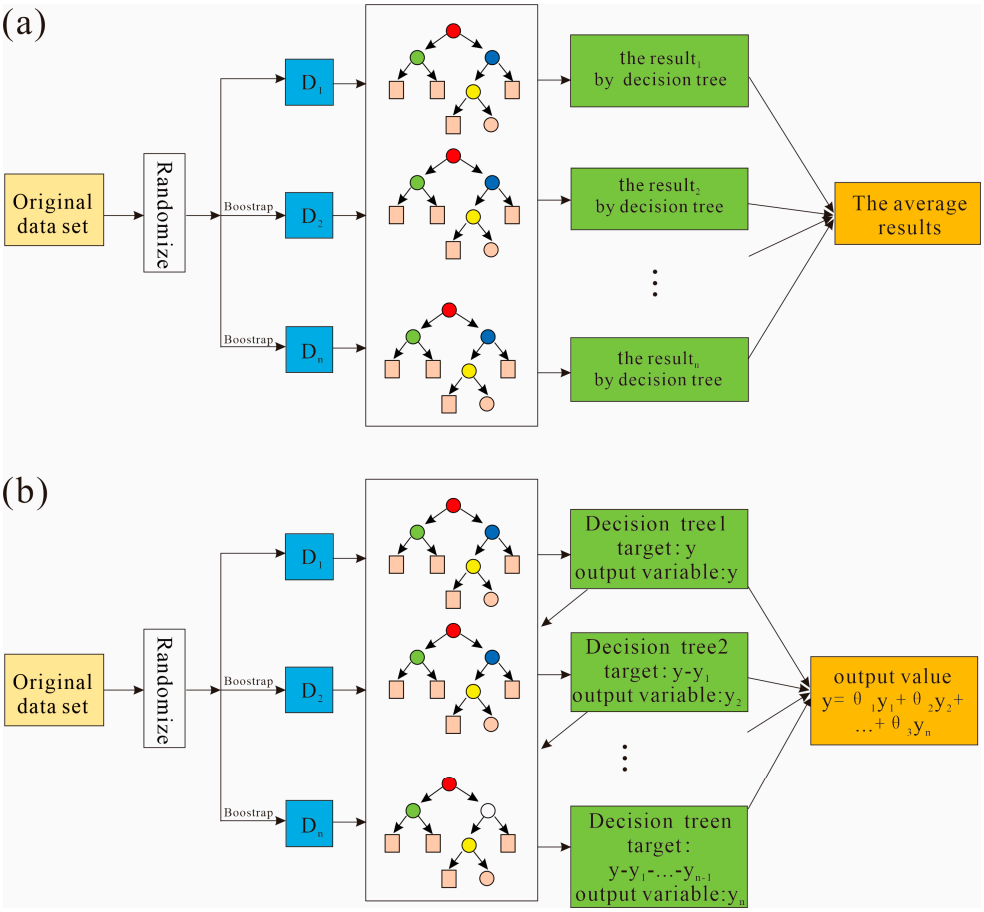


Figure 1. Principles of RF and GBDT algorithms: (**a**) RF model; (**b**) GBDT model.

In regression problems, random forest regression models are usually built using CART trees (classification and regression tree). In this study, the CART tree is built by using the Gini coefficient to select the best features for splitting. The Gini coefficient is calculated as follows:

$$Gini(A) = \sum_{i=1}^{I} p_i(1 - p_i) = 1 - \sum_{i=1}^{I} {p_i}^2 \quad (1)$$

2.2. Gradient Boosting Decision Tree Algorithm

Gradient Boosting Decision Tree (GBDT) is an ensemble machine learning algorithm based on the boosting method that combines the addition model with the forward distribution algorithm and uses the decision tree as the basis function. The weak learner in the GBDT is limited to CART [36]. The combination of the boosting and regression methods improves the model's accuracy and decreases the model's variance. While the bagging method is used in RF models, which means that the probability of each occurrence being selected in subsequent samples is equal, GBRT uses the boosting method, which the input data are weighted in subsequent trees [23]. The GBDT calculation process is based on residuals, and each iteration process proceeds in the direction of decreasing the loss function [32,37]. Each decision tree in the GBDT is not independent of the other (Figure 1b) because the establishment of the latter decision tree is dependent on the loss function calculated by the previous decision tree. So the final prediction result of the GBDT is the sum of the calculation results of all decision trees [38,39] (Figure 1b).

The negative gradient of the loss function at the current sub-model is regarded as the residual approximation in a boosting tree, which is employed by GBDT to fit a regression tree [39]. The negative gradient of the loss function can be calculated by:

$$r_{mi} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \quad \text{for } i = 1, 2, \ldots N \tag{2}$$

Given a training set T = {(x_1,y_1), (x_2,y_2), . . . , (x_n,y_n)}, The GBDT model construction process is as follows [39].

Initialization. Estimate the constant value that minimizes the loss function, and the $L(y_i, c)$ is the loss function:

$$f_0(x) = \underset{c}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, c) \tag{3}$$

To simplify the process of calculation, $f_0(x)$ is often defined as zero. In this study, $f_0(x)$ is the mean value of all data labels in the training set for improving the accuracy of the results.

Then, the negative gradient of the loss function is calculated by Equation (2) in the current model and used as an estimator of the residuals. Here, the fitting values r_{mj} of the j-th leaf node in the mth regression tree are optimized by minimizing the loss:

$$c_{mj} = \underset{c}{\operatorname{argmin}} \sum_{x_i \in R_{mj}} L(y_i, f_{m-1}(x_i) + c)(x \in R_{mj}), j = 1, 2, \ldots J \tag{4}$$

where R_{mj} denotes the region associated with the jth leaf node in the m-th regression tree. Summing up the contribution from the overall J nodes of the m-th regression tree results in an additive update to the model in the m-th iteration:

$$f_m(x) = f_{m-1}(x) + \sum_{J}^{j=1} c_{mj} I\left(x \in R_{mj}\right) \tag{5}$$

Obtain the final output model:

$$\hat{f}(x) = f_M(x) = \sum_{M}^{m=1} \sum_{J}^{j=1} c_{mj} I\left(x \in R_{mj}\right) \tag{6}$$

2.3. Bayesian Optimization Algorithm

The Bayesian optimization algorithm (BO) is an iterative algorithm whose main benefit is that it uses previous findings to choose future evaluation sites [40], which is composed of two key components, including a surrogate model and an acquisition function [41]. The

surrogate model aims to fit the current data to the objective function. After obtaining the predictive distribution of the probabilistic surrogate model, the acquisition function weighs the trade-off between exploration and exploitation to decide which points to use.

The two most commonly used methods in BO are the Gaussian process method (GP) and the Tree-structured Parzen estimator method (TPE), and since the TPE method outperforms the GP method in some complex problems, the TPE method is employed in this study for hyper-parameter optimization [40]. Two density functions are created for the generative model of all domain variables in the TPE method [29]. The expected improvement in the acquisition function is reflected based on the ratio between the two density functions, which is used to determine the new configuration for conducting the evaluation:

$$(x \mid y, D) = \begin{cases} l(x), & \text{if } y < y^* \\ g(x), & \text{if } y > y^* \end{cases} \tag{7}$$

where $l(x)$ is the density distribution of the observations associated with the loss value, $g(x)$ is the density distribution of the remaining observations, and y^* is the γ-quantile of the best observation y.

The first step to achieving the Bayesian optimization algorithm is that the probabilistic surrogate model of the objective function is built. Secondly, the optimal hyper-parameters are detected on the surrogate model and then applied to the real objective function to evaluate them. Thirdly, the surrogate model is updated with new results. Finally, the second and third steps are repeated until the maximum number of iterations is reached [42].

3. Establishment of the Model

3.1. Geological Background of the Study Area

Zhengzhuang Block, southern Qinshui Basin, China, is one of the important areas for the commercial development of coalbed methane. The Zhengzhuang Block is located on top of the Horseshoe Slope Tectonics and is west of the Sitou Fault. It has undergone multiple phases of tectonic deformation [43] and is developed as a near-SN and NNE-oriented broad and gentle fold [44,45]. Two major fracture structures, the Sitou Fault and the Houchengyao Fault, are developed in the study area (Figure 2). The main coal reservoir in the Zhengzhuang block is the No. 3 coal seam of the Carboniferous Shanxi Formation, which has good continuity and wide distribution. The buried depth of the coal seam ranges from 500 m to 1200 m, and the average thickness is 5.5 m [46]. The total gas-bearing area of coalbed methane is about 700 km^2, but the gas content of coal reservoirs varies widely, so a reasonable prediction model is needed to be proposed to evaluate the gas content of the coal seam accurately.

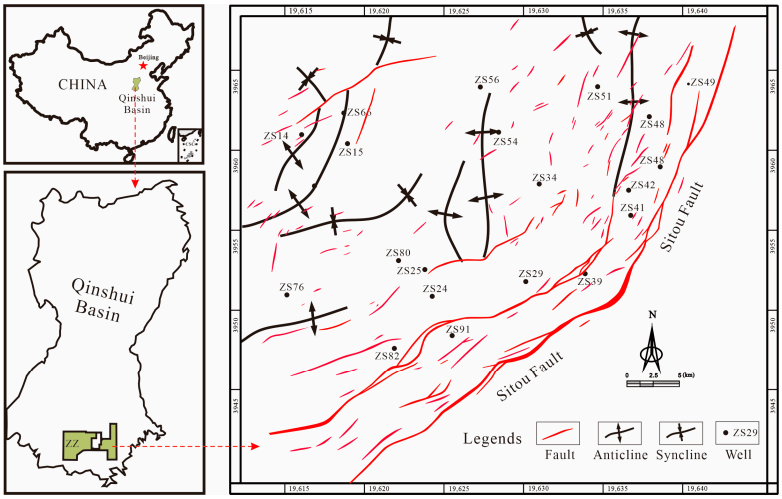


Figure 2. Regional geological characteristics of Zhengzhuang block.

3.2. Dataset Preparation

As shown in Table 1, the logging data and the coal gas content test data from 38 wells in the Zhengzhuang block were collected for the study. Since the logging instrument is subjected to tensile deformation by gravity during the measurement process, the logging depth was first corrected to ensure that the logging data accurately reflect the characteristics of the coal core [47]. The brittle and fragile characteristics of coal make the coal seam often appear as a phenomenon of diameter expansion [48], which leads to the distortion of logging value. So, the logging data collected from the strata with serious diameter expansion are corrected [49]. After data pre-processing, the data were standardized to eliminate the variability between different logging parameters. The standardization process was calculated as follows:

$$x_{std} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{8}$$

where x is the value of a logging parameter, $x_{\min}$ is the minimum value of the logging parameter, and $x_{\max}$ is the maximum value of the logging parameter.

Table 1. Part of logging data and measured gas content in Zhengzhuang Block.

Well	Depth (m)	AC (μm/s)	CAL (cm)	DEN (g/cm^3)	GR (API)	R_D (lg)	R_S (lg)	R_{Xo} (lg)	Gas Content (cm^3/t)
ZS86-1	425.15	399.74	29.66	1.28	63.87	3.09	3.01	3.15	13.01
ZS86-2	427.50	409.99	27.42	1.24	39.98	3.78	3.74	3.52	13.31
ZS73-1	895.53	372.73	22.85	1.35	66.68	3.25	3.30	3.34	22.81
ZS73-2	895.78	390.50	22.84	1.28	56.12	3.28	3.33	3.43	24.86
ZS73-3	897.20	413.05	23.09	1.16	27.31	3.95	3.91	3.76	25.34
ZS73-4	897.38	413.90	22.99	1.15	29.80	4.01	3.96	3.68	25.34
ZS73-5	898.88	396.00	26.77	1.19	50.39	3.35	3.46	2.18	22.42
ZS72-1	1108.40	417.78	22.76	1.30	32.94	3.56	3.61	2.91	26.24
ZS72-2	1111.15	405.65	22.40	1.36	44.76	3.17	3.30	3.37	25.13
ZS34-1	784.60	417.67	23.50	1.34	56.58	3.55	3.50	2.90	28.12
ZS34-2	785.05	426.89	23.59	1.36	45.61	3.44	3.39	3.19	27.13
ZS34-3	785.30	426.53	23.84	1.43	48.84	3.28	3.23	3.88	28.37
ZS93-1	674.35	397.72	21.85	1.24	79.45	2.55	2.61	2.38	11.86
ZS93-2	675.90	400.49	23.30	1.36	79.77	2.95	2.92	2.53	17.95
ZS78-1	702.50	399.70	23.73	1.28	62.66	2.79	2.83	2.61	21.14
ZS78-2	705.50	414.04	22.88	1.20	42.80	3.25	3.17	2.16	21.04
ZS98-1	1229.90	408.27	24.28	1.28	38.89	3.28	3.22	3.11	19.05

The standard data set for this study was established through data preprocessing and standardization, and then the data set was divided into two parts: the training set and the test set, of which the former is used to train the model and the latter is used to evaluate the model. According to previous practical experience [50], the data set is divided into the training set and the test set at 8:2. Therefore, the data from 30 wells were randomly selected from the 38 well data as the training data set, and the remaining 8 wells data were used as the validation data set.

3.3. Modeling Process

The flowchart of the CBM gas content prediction model is displayed in Figure 3, as follows:

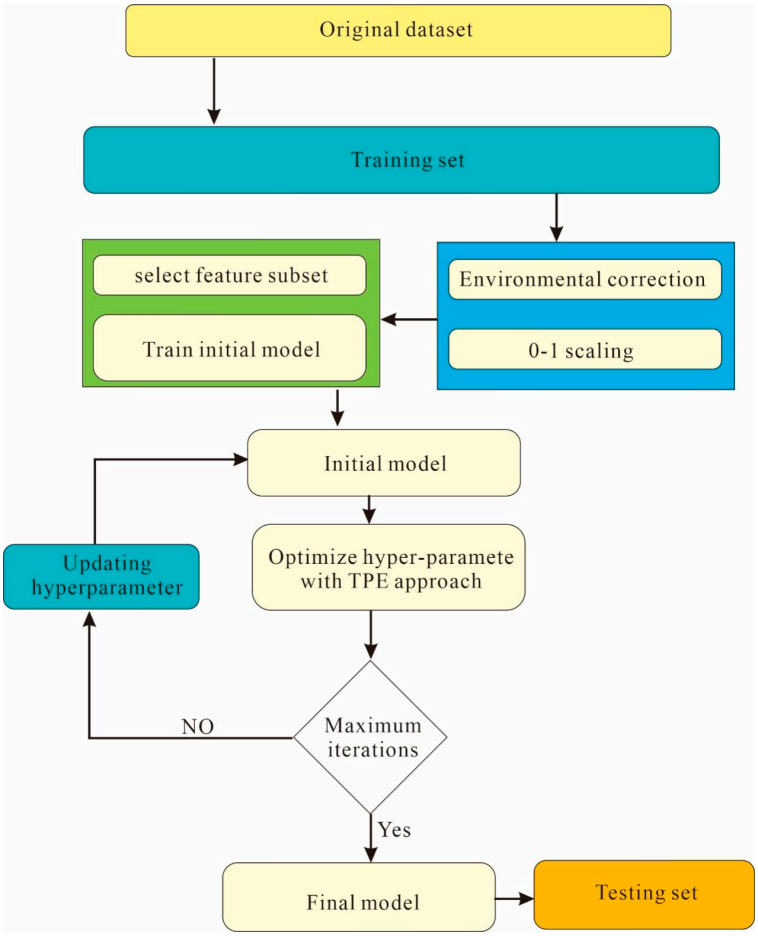


Figure 3. Establishment process of CBM gas content prediction model.

(1) Model initialization and parameter selection: The training data is brought into the ensemble tree model for initial training, and the hyper-parameters in the model are adjusted manually;

(2) Establishment of an ensemble tree model based on Bayesian hyper-parameter optimization: The reasonable search space is set for each hyper-parameter of the model, and the Bayesian optimization method is used to determine the hyper-parameters of the initialized ensemble tree model;

(3) Model validation: the test set is the input to the model for validating the established model, and the evaluation parameters are calculated for comparing the different models.

In this study, GBDT and RF models are operated with the Sklearn package on Python 3.9. Bayesian hyper-parameter optimization is performed using the hyperopt package for Python. All the experiments are performed on a desktop computer with a 3.0 GHz Intel i5 CPU, 8 GB of RAM, and the Windows 10 operating system.

4. Results and Discussion

4.1. Feature Selection

The changes in reservoir characteristics can be reflected by the difference in logging curve response, so the changes in CBM gas content can theoretically be characterized by changes in the logging response [8]. Considering the practicality of establishing the model, the logging curves that were available for all wells were selected. In this study, eight curves were selected to construct the models, but every curve has a different influence on the performance of the model. The data sets of different logging parameter combinations are adjusted and trained to find the best parameter combination. When choosing five logging parameters, the average RSQ of the RF and GBDT models is the highest, as

shown in Figure 4. According to the results of different logging parameter combinations, the parameters including depth, AC, R_s, R_d, and R_{xo} were chosen to establish the model.

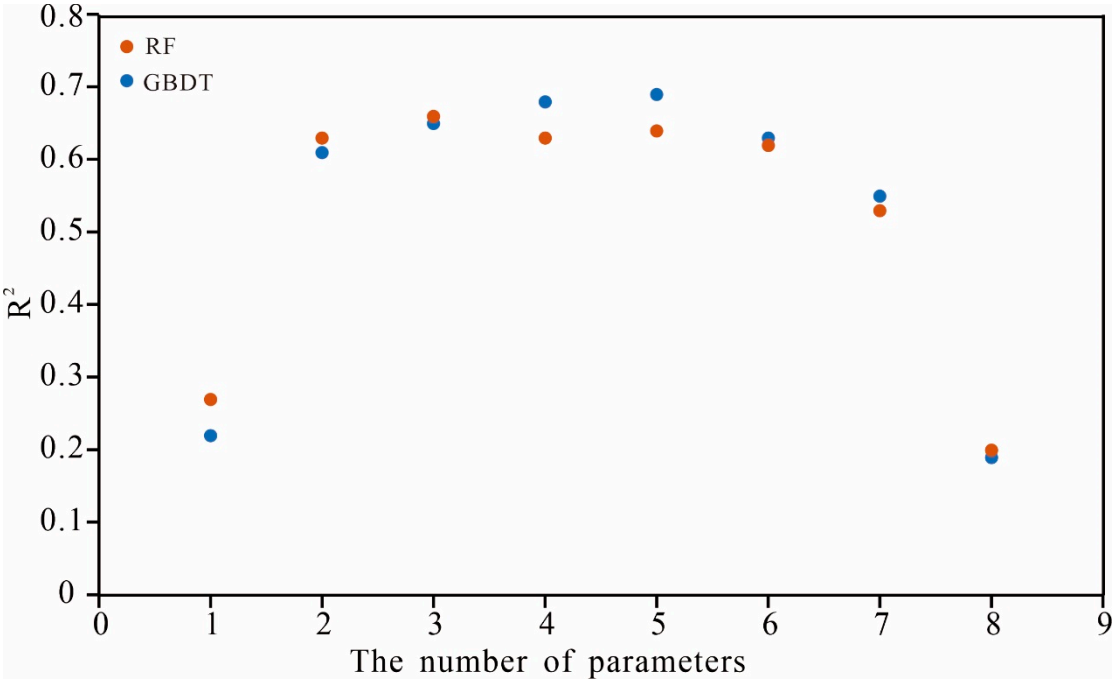


Figure 4. The relationship between RSQ and the number of parameters.

4.2. Ensemble Tree Models Based on Manual Research

There are numerous hyper-parameters in machine learning algorithms that have different influences on the accuracy of models (Table 2). A reasonable value of hyper-parameters can effectively avoid model overfitting. So hyper-parameter optimization is a key step to improving the accuracy of machine learning algorithms [40]. The most important hyper-parameters are often selected for model optimization manually and with experience. In the RF model, the n_estimators (n_{tree}) presented the number of trees, and the max_depth (n_{depth}) presented the maximum depth of the trees. These values have the greatest influence on the complexity of the model (Table 1), so the n_{tree} and n_{depth} in the RF were manually adjusted. As n_{tree} increased, the mean square error (MSE) of the RF model decreased rapidly and then gradually stabilized (Figure 5). After the value of n_{tree} reaches 300, the n_{tree} has no significant contribution to the decrease in MSE, but the complexity and operation time of the model increased, along with the increase in n_{tree}. Similarly, the increase in n_{depth} can reduce the model errors, but when the $n_{depth} > 5$, the model MSE tends to be stable. Therefore, the n_{tree} and n_{depth} are comprehensively analyzed to determine that the n_{tree} is 280 and the n_{depth} is 5. In the GBDT model, the n_{tree} and the learning_rate (n_{lr}) are the two most important hyper-parameters. The empirical parameters of the GBDT model (Figure 6) were analyzed to determine that n_{tree} is 100 and n_{lr} is 0.2.

Table 2. The importance of different hyper-parameter in the RF model and GBDT model.

Hyper-Parameter of GBDT	Importance	Hyper-Parameter of RF
n_estimators learning_rate max_feature	★★★★★	n_estimators Max_depth

Table 2. *Cont.*

Hyper-Parameter of GBDT	Importance	Hyper-Parameter of RF
Init Subsamples Loss function	★★★★	Min_samples_leaf
Max_depth Min_samples_split Min_impurity_decrease	★★★	Min_sample_split
Max_leaf_nodes criterion	★★	Max_feature
Random_state	★	Criterion

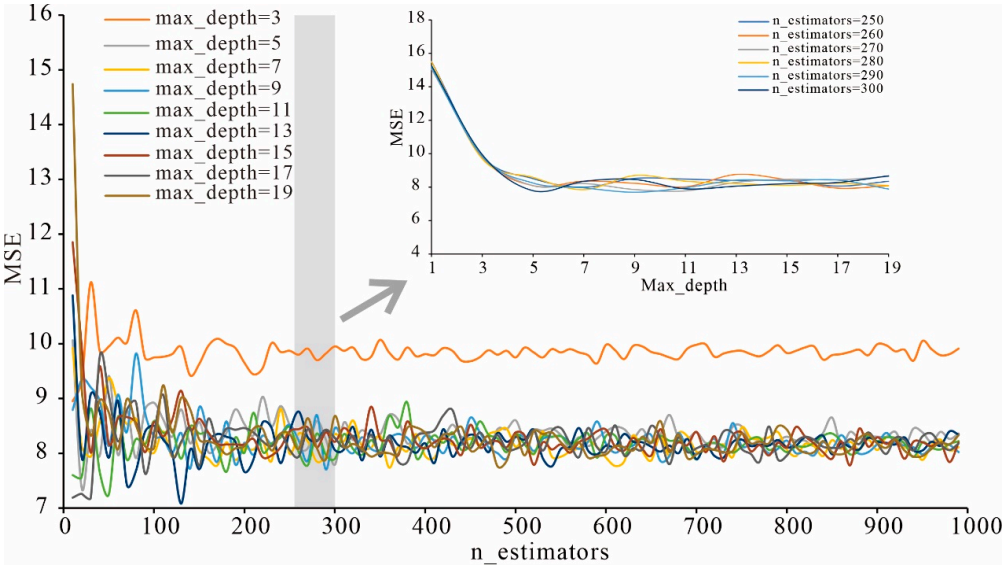


Figure 5. Manual tuning of RF model.

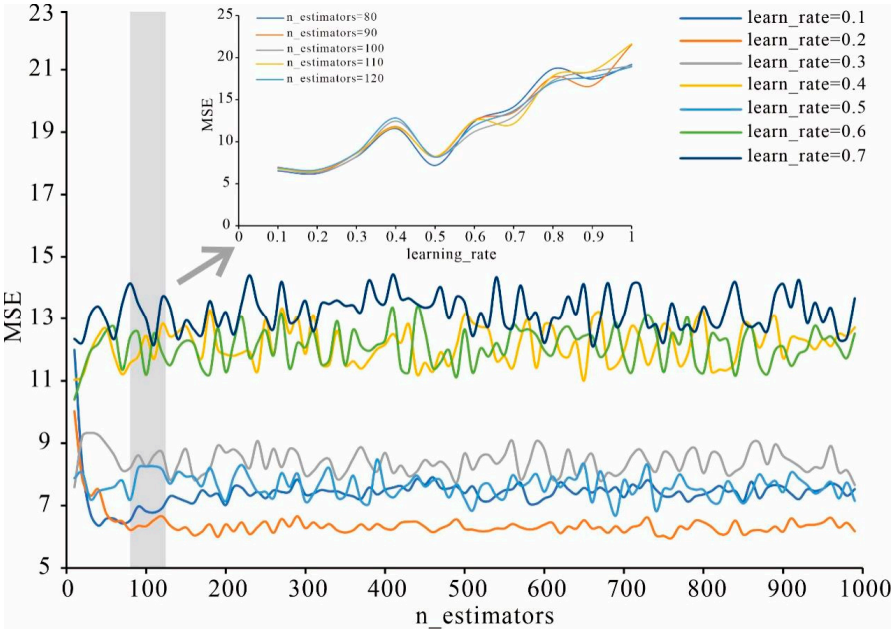


Figure 6. Manual tuning of GBDT model.

The models' effects can be reflected by the accuracy of the RF model and the GBDT model on the test set. As shown in Figure 7b,d, most of the data points from the two models are distributed within the 15% error line. The average relative error of the RF model is 10.42%, and that of the GBDT model is 9.54%. The R-squared (RSQ) of the GBDT model is larger than that of the RF model, so the accuracy of the GBDT model is higher than that of the RF model. No matter the RF model or the GBDT model, the ZS49 well has a significant abnormally high prediction value that seriously deviates from the measured gas content. However, compared with the GBDT model, the ZS49 well has a greater error in the RF model, indicating that the GBDT has a stronger error processing capability. The performance of the two models on the training set is illustrated in Figure 7a,c. As expected, the two models perform better on the training set than on the test set, and there is no obvious overfitting on the training set. So, the model has a certain generalization ability, indicating the prediction reliability of the models on the test set.

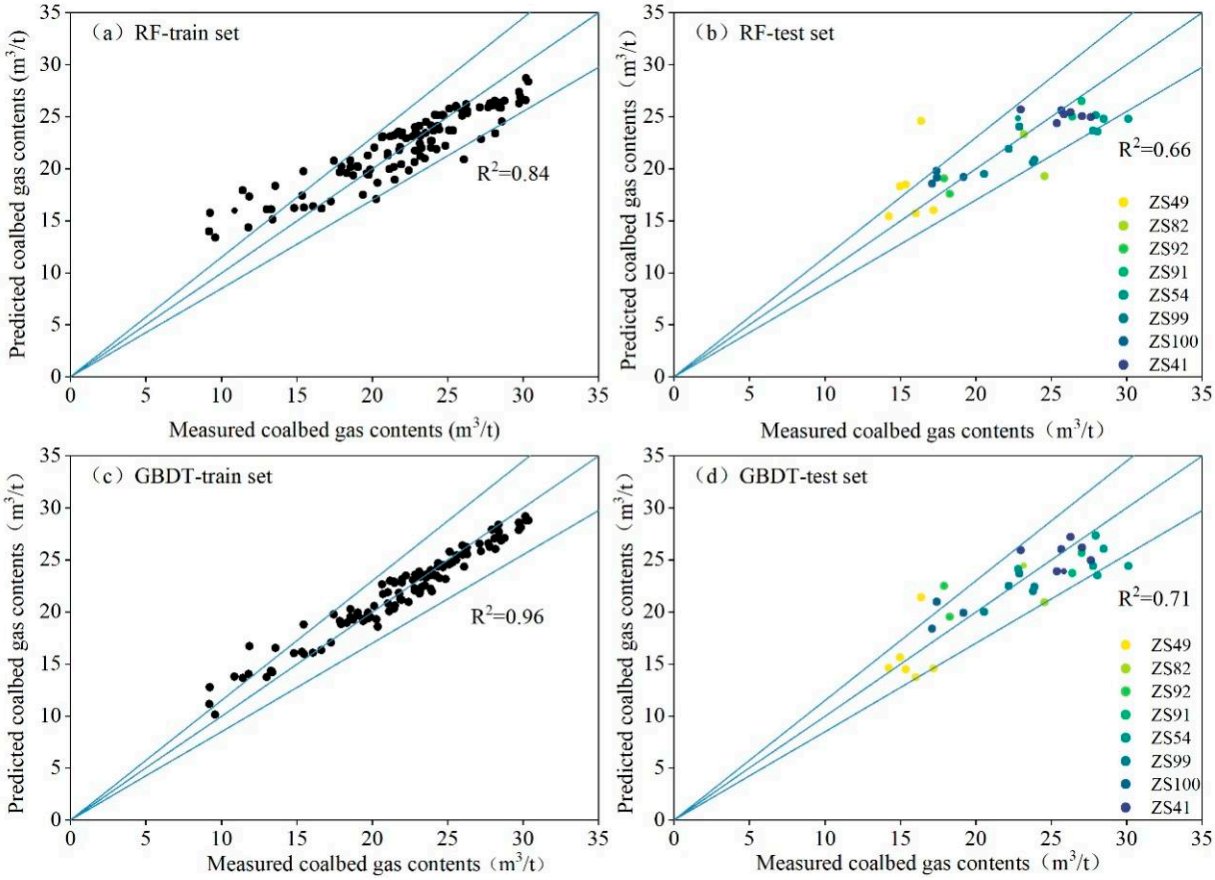


Figure 7. Performance of RF and GBDT models in training and test sets.

4.3. Ensemble Tree Models Based on the Bayesian Optimization Method

Although the ideal effect can be achieved by the algorithm design of the machine learning models, hyper-parameters can control the complexity and regularization of the models to affect their accuracy. Conventional parameter tuning methods such as manual parameter tuning, and grid search have limitations for models with numerous hyper-parameters. Bayesian optimization methods have their advantages in global optimization problems, especially in models with numerous hyper-parameters [40,51].

Based on the discussion in Section 4.2, the search space for Bayesian Hyper-parameter optimization is determined (Table 3). To obtain the best possible results, 1000 iterations of the model were carried out using the Bayesian optimization method, and the hyper-parameters determined by this method are shown in Table 3. As shown in Figure 8,

the prediction results of the gradient-boosting decision tree model based on Bayesian optimization (BO-GBDT) and random forest model based on Bayesian optimization (BO-RF) on the test set are basically within the 15% error line. The average relative error of the BO-RF model on the test set is 9.50%, and the RSQ is 0.66. The average relative error of the BO-GBDT model is 7.52%, and RSQ is 0.82. Therefore, the prediction effect of BO-GBDT on the test set is better than that of the BO-RF model. The two models also showed no obvious over-fitting in the training set. It is worth noting that ZS49 well still has an obvious abnormally high value on the BO-RF model, with a relative error of 52.24%, but the anomaly is not obvious in the BO-GBDT model. It shows that the BO-GBDT model has a stronger generalization ability and outlier processing ability.

Table 3. Searching space of Bayesian hyper-parameter optimization method.

RF Model	Search Space	Results	GBDT Model	Search Space	Results
n_estimators	[100,400]	210	n_estimators	[50,200]	170
Max_depth	[2,16]	6	learning_rate	[0.1,2.0]	0.25
Max_features	(log2,sqrt,auto)	auto	criterion	(friedman_mse,mse)	friedman_mse
Min_samples_leaf	[2,10]	4	Loss function	(ls,huber,quantile)	huber
Min_samples_split	[1,10]	9	max_depth	[2,30]	4
			max_features	['log2','sqrt','auto']	auto
			min_impurity_decrease	[0,5]	2

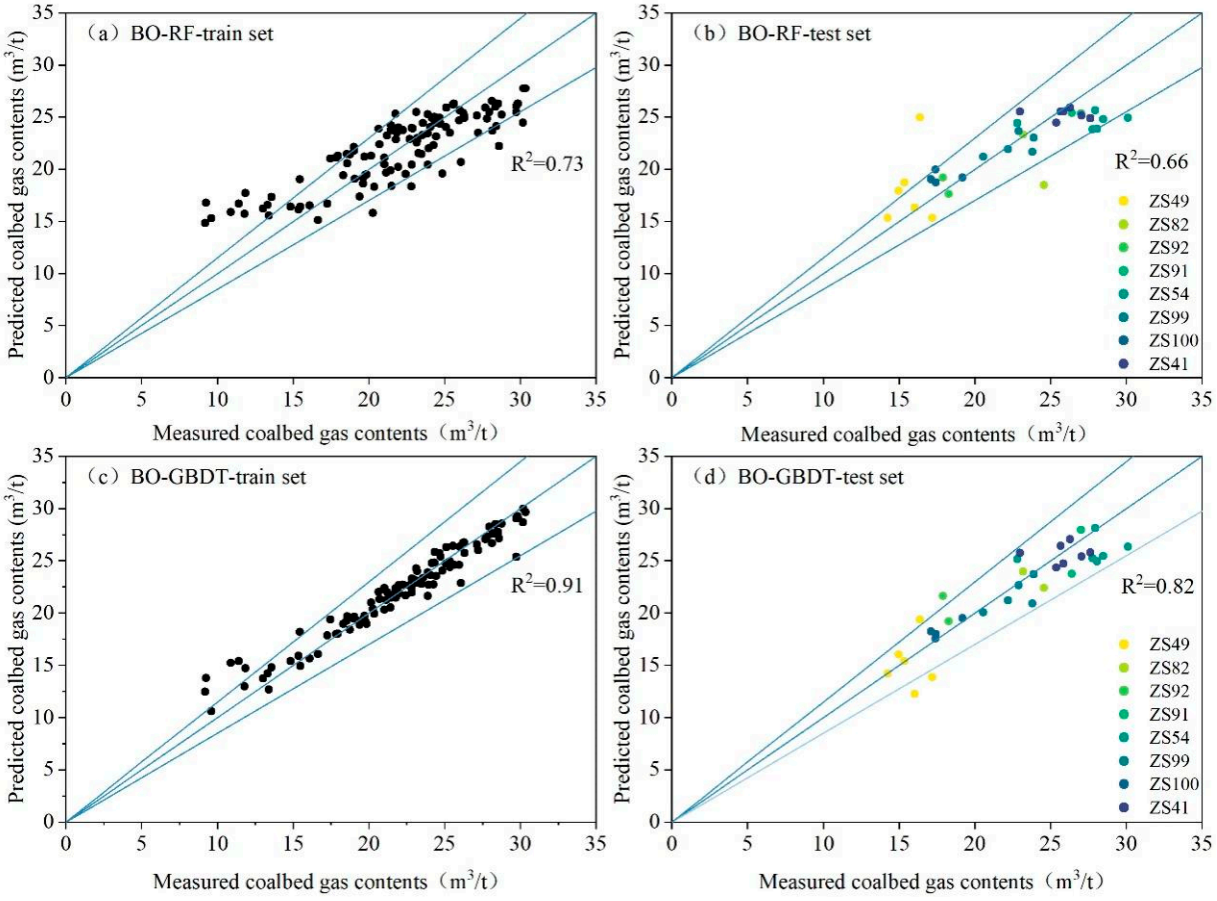


Figure 8. Performance of BO-RF and BO-GBDT models in the train set and test set.

4.4. Model Comparison

The different evaluation parameters of the models are shown in Table 4. Overall, the model prediction effect is BO-GBDT > GBDT > BO-RF > RF. For the same machine learning algorithm, the RSQ of the BO-GBDT model is 15.50% higher than that of the GBDT model, the MSE is 37.94% lower, and the relative error is 24.00% lower. The models based on the GBDT algorithm are significantly improved after being introduced to Bayesian optimization, showing that the model effect based on the GBDT algorithm is controlled by a variety of hyper-parameters.

Table 4. Error parameters of the models.

Models	RSQ	MSE	The Relative Error
RF model	0.66	8.38	10.42%
BO-RF model	0.66	7.63	9.50%
GBDT model	0.71	6.51	9.54%
BO-GBDT model	0.82	4.04	7.25%

The RSQ of the RF and BO-RF models on the test set is equal, but the relative error and MSE are reduced after being introduced to Bayesian optimization. Therefore, the effect of the Bayesian optimization method on the model based on the RF algorithm is not significant. The possible reason is that the hyper-parameters of the RF model are few, and the model is mainly affected by the number of trees and the maximum depth of trees. The remaining hyper-parameters with less influence will only realize the adjustment of model complexity but have no significant effect on its accuracy of the model. Therefore, the Bayesian optimization algorithm has a better optimization effect on machine learning algorithms with numerous hyper-parameters.

The RSQ of the training set on the BO-GBDT model is lower than that on the GBDT model, and the RSQ of the training set on the BO-RF model is also lower than that on the RF model, but the prediction ability of the BO-GBDT model and BO-RF model on the test set is improved, indicating that the models after introduction the Bayesian optimization method have stronger generalization ability and reduce time complexity. It is worth noting that both the BO-RF and RF models have data points that exceed the 15% error line on the training set (Figures 7a and 8a), especially on the data with low gas content, while the prediction results of the BO-GBDT model and the GBDT model on the training set are concentrated within the 15% error line.

The different accuracy for low gas content and high gas content on the RF model and BO-RF model is illustrated in Figure 9. In the RF model, the almost predicted gas content data is higher than the measured gas content data, while the measured gas content is less than 20 m^3/t, and the lower the measured gas content, the greater the deviation of the predicted gas content. While the measured gas content is higher than 20 m^3/t, the data points are concentrated within the 15% error line. As shown in Figure 9c,d, a similar characteristic also applies to the BO-RF model. On the two models, the MSE and relative error for gas content above 20 m^3/t are lower than for gas content below 20 m^3/t, so the accuracy of the RF model and BO-RF model at high gas content is higher. The prediction performance of ML algorithms is heavily affected by the number and quality of the dataset [38]. Because there are three times as many data with a gas content greater than 20 m^3/t as there are with less than 20 m^3/t, the RF model and BO-RF model are not fully trained when the gas content is lower than 20 m^3/t due to the bagging method.

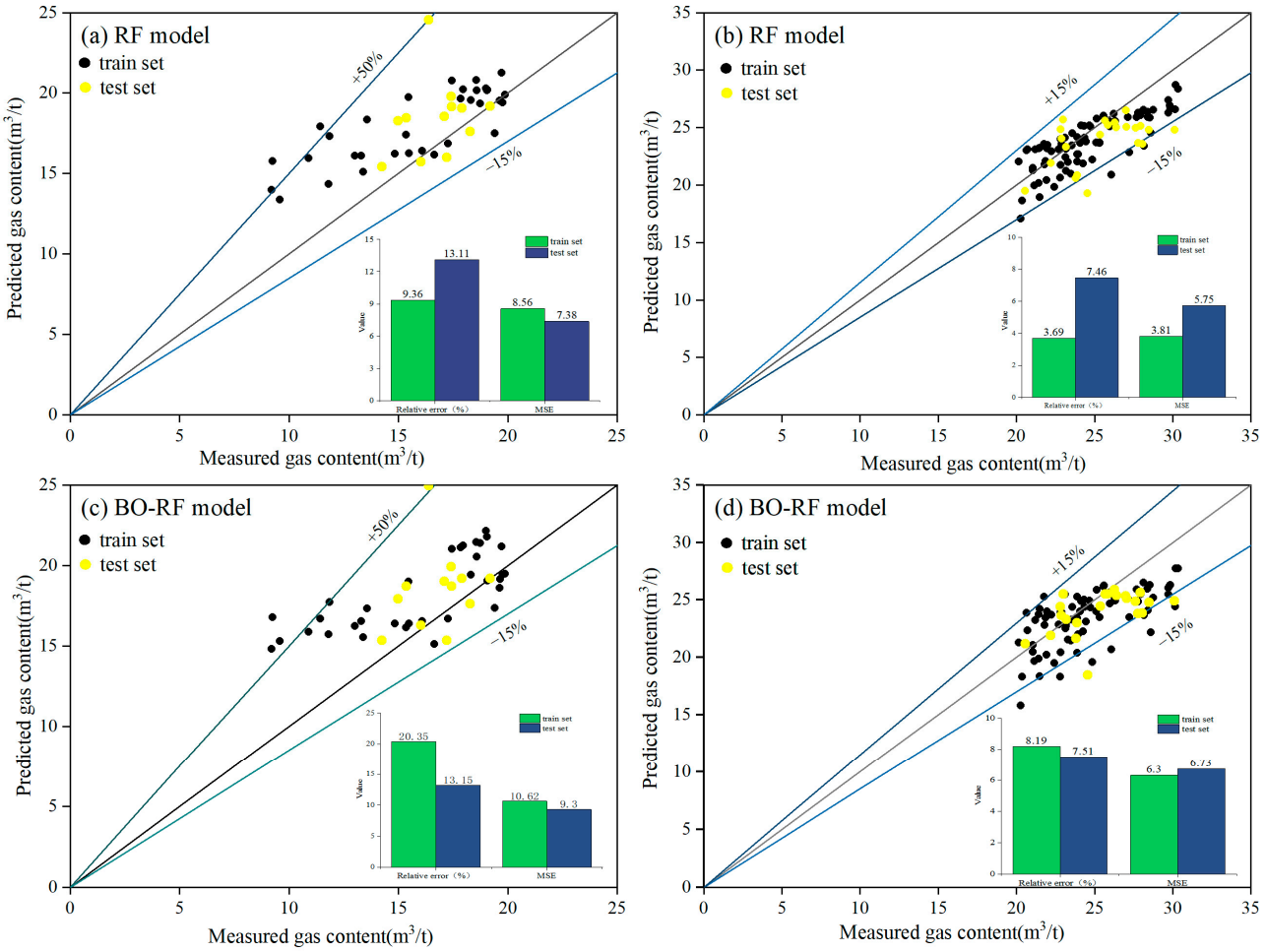


Figure 9. The different performance for low gas content and high gas content ((**a**) gas content below 20 m^3/t on RF model; (**b**) gas content above 20 m^3/t on RF model; (**c**) gas content below 20 m^3/t on BO-RF model; (**d**) gas content above 20 m^3/t on BO-RF model).

The prediction accuracy of each test well is illustrated in Figure 10. The accuracy of each test well is improved after the model based on the GBDT algorithm is optimized by the BO method. The relative error of 8 test wells is reduced by 23.82% on average, and the MSE is reduced by 31.75% on average. Among them, the accuracy improvement of the ZS100 well is the highest, the MSE is reduced by 93.82%, and the relative error is reduced by 76.48%. The accuracy of the models based on the random forest algorithm varies greatly among the test wells after being introduced to the BO method. Similarly, the accuracy improvement of the ZS100 well is the highest, the MSE is reduced by 74.79%, and the relative error is reduced by 57.08%. However, the relative error of the ZS82 and ZS54 wells increased after Bayesian optimization, especially in the ZS82 well, where the relative error growth almost doubled. It shows that the BO method has no obvious effect on improving the accuracy of the random forest model. The different performance in each test well also shows that the random forest model still has the problem of insufficient learning even after hyper-parameter optimization. In addition, due to the large variation range of logging data and the fluctuation of environmental factors such as logging instruments and formation characteristics, the ability to deal with the prediction errors caused by data outliers during the regression calculation of the random forest algorithm is weak. In contrast, the GBDT model has a stronger ability to deal with abnormal data through the calculation principle of gradual approximation of residuals.

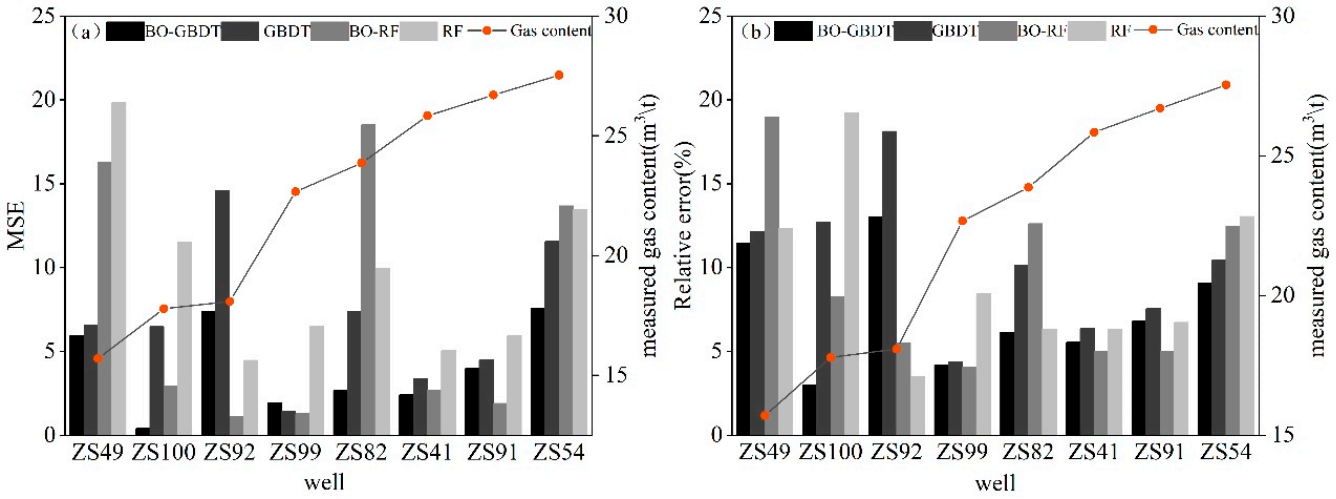


Figure 10. Comparison of prediction results of test wells:(**a**) MSE; (**b**) relative error.

As shown in Figure 10, as the gas content of the test well increases, the accuracy of the four models increases, and the RF model and the BO-RF model perform more significantly. When the gas content is less than 20 m^3/t, the test wells show a higher MSE value and relative error, while in the wells with gas content greater than 20 m^3/t, MSE value and relative error are relatively low. It shows that the random forest model has a better prediction effect in high gas-content coal reservoirs. Although the relative error and MSE in the test wells with gas content less than 20 m^3/t of the GBDT model and BO-GBDT models are higher than those in test wells with higher gas content when compared with the models based on random forest algorithm, the models based on gradient boosting decision tree algorithm have higher prediction accuracy in the test well with gas content below 20 m^3/t, indicating that the gradient boosting decision tree algorithm has better prediction effect in the whole gas content range.

4.5. Model Application on the Gas Content Prediction

The established model is used to predict the gas content of the coal seam in different depths of the single well (Figure 11). The gas content curve predicted by the BO-GBDT model is the closest to the measured gas content, and the gas content curve predicted by the RF model has the largest deviation from the measured gas content. In addition, the amplitude of the prediction curve of the GBDT model and RF model is small, while the amplitude of the prediction curve of the BO-GBDT model is large, which is sensitive to the change of the logging curve response, indicating that the model has stronger data analysis ability after being introduced to the BO method, so the BO-GBDT model has stronger generalization and credibility. The prediction curves of all models fluctuate greatly at the lithology interface, and one possible reason is that the logging curves at the lithology interface change greatly and models learn insufficiently or even lack learning at the lithology interface.

The measured gas content distribution of the No.3 coal seam in the Zhengzhuang block is shown in Figure 12. The distribution of gas content in the study area generally ranges from 9.96 to 28.83 m^3/t, with an overall trend of increasing from south to north. Due to the development of the tensile normal fault and poor preservation conditions, gas content is generally lower than 20 m^3/t in the south. In the middle region, the gas content varies widely, generally between 10 and 30 m^3/t, which is controlled by both faults and folds. In the northern region, except for the eastern edge, the gas content is generally higher, generally greater than 25 m^3/t.

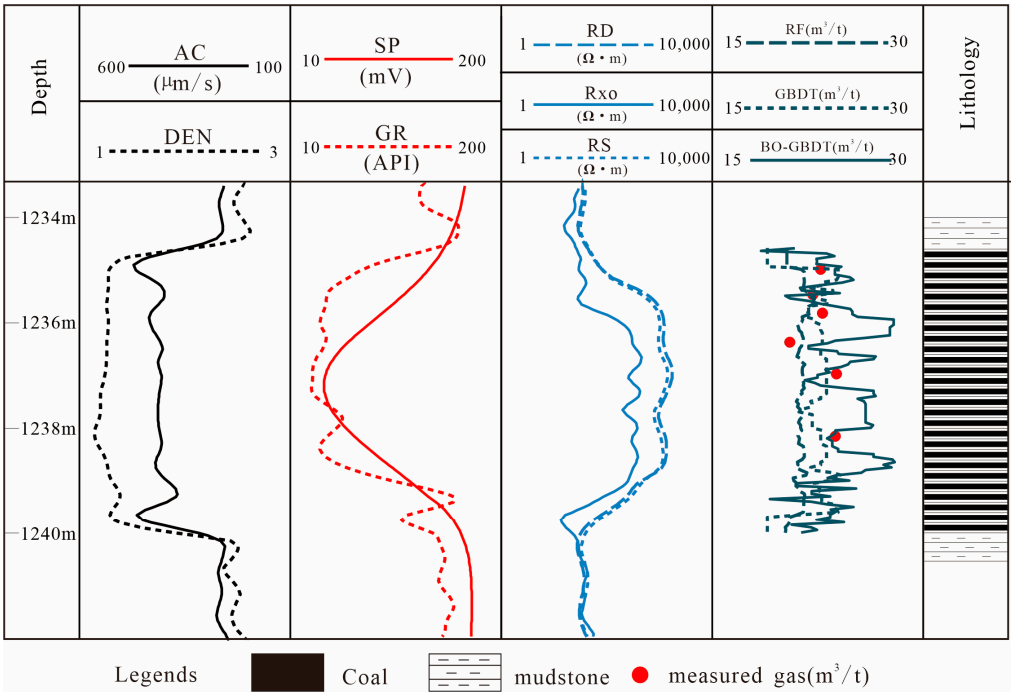


Figure 11. The gas content curve predicted by different models (taking ZS99 well as an example).

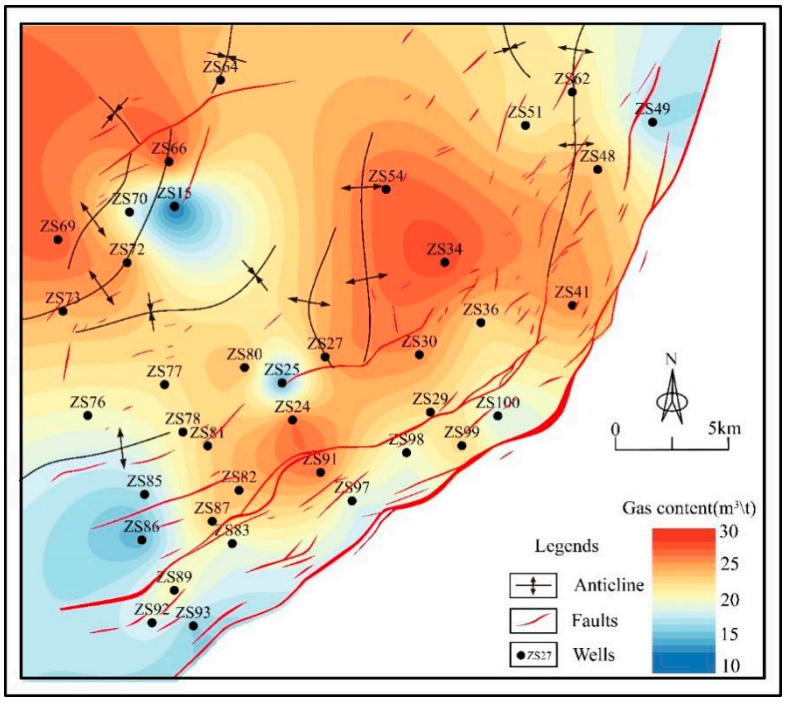


Figure 12. Gas content distribution of the No.3 coal seam in Zhengzhuang block.

The predicted distribution of gas content predicted by the four models (Figure 13) is consistent with the measured gas content distribution. The four models of high gas content centers and low gas content centers are in the same position. It shows that the four gas content prediction models have high reliability in predicting the plane characteristics of gas content; that is, the gas content prediction models based on random forest and gradient boosting decision tree algorithms combined with geophysical logging can be applied to CBM resource evaluation and favorable zones.

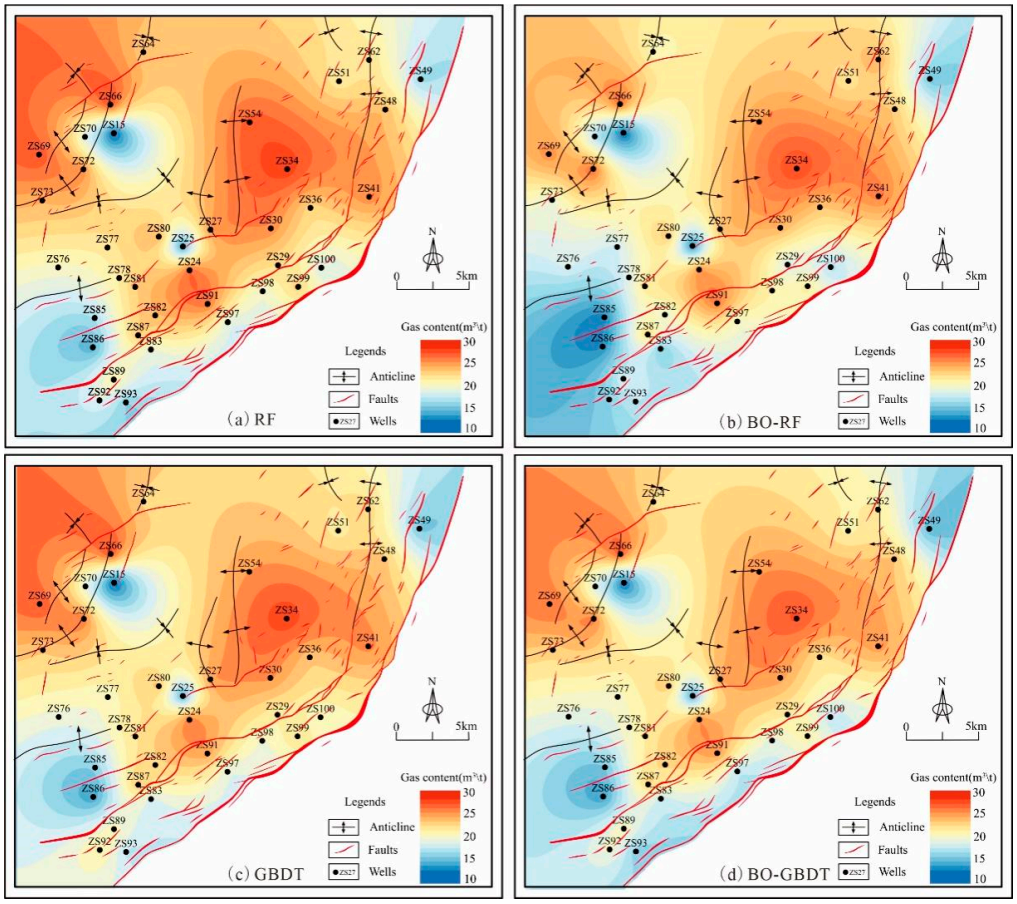


Figure 13. Gas content distribution predicted by different models ((**a**) RF model; (**b**) BO-RF model; (**c**) GBDT model; (**d**) BO-GBDT model).

5. Conclusions

Based on the logging data of the coal reservoir, the gas content prediction models of the coal reservoir are established by using ensemble tree algorithms, which are introduced in the Bayesian hyper-parameter optimization method. The results show that the established models can effectively evaluate the gas content of the coal seam in the Zhengzhuang block.

(1) The prediction errors of BO-GBDT, GBDT, BO-RF, and RF models are controlled within 15%, with the BO-GBDT model having the highest RSQ of 0.82, followed by the GBDT model with 0.72, and the BO-RF and RF models having an equal RSQ of 0.62. The accuracy order of different models was: BO-GBDT > GBDT > BO-RF > RF. The models based on the GBDT algorithm have stronger abnormal data processing abilities than the models based on the random forest algorithm.

(2) The RSQ of the BO-GBDT model is 15.50% higher than that of the GBDT model, the MSE is 37.94% lower, and the relative error is 24.00% lower. The accuracy of the BO-GBDT model is significantly higher than that of the GBDT model, while the accuracy of the BO-RF model has only slightly improved compared with the RF model. The application effect of Bayesian hyper-parameters optimization depends on the number of hyper-parameters included in the machine learning algorithm.

(3) In the longitudinal direction, the BO-GBDT model has the highest accuracy in predicting the gas content distribution curve, which is sensitive to logging curve response. The RF model has the maximum deviation between the predicted curve and measured gas content data.

(4) In the plane, the rule of predicted gas content distribution of the four models is consistent with the measured gas content distribution. The model in this paper can be used for coalbed methane resource evaluation and favorable zone selection.

Author Contributions: C.Y. and F.Q. conceived and designed the methods; C.Y. performed the methods and wrote the paper; C.Y and F.Q. analyzed the data; F.X. and S.C. revised the paper and provided language support; Y.F. and S.C. provided technical support. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Fund (Grant NOs. 41830427, 41772160, and 41922016).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jia, Q.F.; Liu, D.M.; Cai, Y.D.; Yao, Y.B.; Lu, Y.J.; Zhou, Y.F. Variation of adsorption effects in coals with different particle sizes induced by differences in microscopic adhesion. *Chem. Eng. J.* **2023**, *52*, 139511. [CrossRef]
2. Sun, Q.P.; Zhao, Q.; Jiang, X.C.; Mu, F.Y.; Kang, L.X.; Wang, W.Z.; Yang, Q.; Zhao, Y. Prospects and strategies of CBM exploration and development in China under the new situation. *J. China Coal Soc.* **2021**, *46*, 56–76. (In Chinese)
3. Li, Z.; Liu, D.; Wang, Y.; Si, G.; Cai, Y.; Wang, Y. Evaluation of multistage characteristics for coalbed methane desorption-diffusion and their geological controls: A case study of the northern Gujiao Block of Qinshui Basin, China. *J. Petrol. Sci. Eng.* **2021**, *204*, 108704. [CrossRef]
4. Lin, H.; Long, H.; Li, S.; Bai, Y.; Xiao, T.; Qin, A. CH4 Adsorption and Diffusion Characteristics in Stress-Loaded Coal Based on Molecular Simulation. *Fuel* **2023**, *333*, 126478. [CrossRef]
5. Fu, X.H.; Zhang, X.D.; Wei, C.T. Review of research on testing, simulation and prediction of coalbed methane content. *J. China U Min Technol.* **2021**, *50*, 13–31. (In Chinese)
6. Liu, D.; Yao, Y.; Chang, Y. Measurement of Adsorption Phase Densities with Respect to Different Pressure: Potential Application for Determination of Free and Adsorbed Methane in Coalbed Methane Reservoir. *Chem. Eng. J.* **2022**, *446*, 137103. [CrossRef]
7. Feng, R. A Method to Evaluated Gas Content with Coalbed Methane Reservoir Based on Adsorption Theory and Production Analysis. *Geofluids* **2022**, *2022*, 7341886. [CrossRef]
8. Guo, J.; Zhang, Z.; Guo, G.; Xiao, H.; Zhu, L.; Zhang, C.; Tang, X.; Zhou, X.; Zhang, Y.; Wang, C. Evaluation of Coalbed Methane Content by Using Kernel Extreme Learning Machine and Geophysical Logging Data. *Geofluids* **2022**, *2022*, 3424367. [CrossRef]
9. Pashin, J.C.; McIntyre-Redden, M.R.; Mann, S.D.; Kopaska-Merkel, D.C.; Varonka, M.; Orem, W. Relationships between Water and Gas Chemistry in Mature Coalbed Methane Reservoirs of the Black Warrior Basin. *Int. J. Coal Geol.* **2014**, *126*, 92–105. [CrossRef]
10. Shen, J.; Du, L.; Qin, Y.; Yu, P.; Fu, X.H.; Chen, G. Three-phase gas content model of deep low-rank coals and its implication for CBM exploration:a case study from the Jurassic coal in the Junggar Basin. *J. Nat. Gas Sci. Eng.* **2015**, *35*, 30–35. (In Chinese)
11. Crosdale, P.J.; Beamish, B.; Valix, M. Coalbed methane sorption related to coal composition. *Int. J. Coal Geol.* **1988**, *35*, 147–158. [CrossRef]
12. Moore, T. Coalbed methane: A review. *Int. J. Coal Geol.* **2012**, *101*, 36–81. [CrossRef]
13. Liu, D.M.; Jia, Q.F.; Cai, Y.D.; Gao, C.J.; Qiu, F.; Zhao, Z.; Chen, S.Y. A new insight into coalbed methane occurrence and accumulation in the Qinshui Basin, China. *Gondwana Res.* **2022**, *111*, 280–297. [CrossRef]
14. Cao, J.T.; Zhao, J.L.; Wang, Y.P.; Zhang, J.Y.; Xu, D.C. Review of influencing factors and prediction methods of gas content in coal seams and prospect of prediction methods. *J. Xi'an Shiyou Univ. (Nat. Sci. Ed.)* **2013**, *28*, 28–34+94. (In Chinese)
15. Yahya, S.I.; Rezaei, A.; Aghel, B. Forecasting of Water Thermal Conductivity Enhancement by Adding Nano-Sized Alumina Particles. *J. Therm. Anal. Calorim.* **2021**, *145*, 1791–1800. [CrossRef]
16. Liu, A.H.; Fu, X.H.; Wang, K.X.; Peng, L.; Zhou, B.Y. Prediction of coalbed gas content based on support vector machine regression. *J. Xian Univ. Sci. Technol.* **2010**, *30*, 309–313. (In Chinese)
17. Zhang, S.R.; Wang, B.T.; Li, X.E.; Chen, H. Research and application of improved gas concentration prediction model based on grey theory and BP neural network in digital mine. *Procedia Cirp* **2016**, *56*, 471–475. [CrossRef]
18. Guo, J.H.; Zhang, Z.S.; Zhang, C.M.; Zhou, X.Q.; Xiao, H.; Qin, R.B. The exploration of predicting CBM content by geophysical logging data: A case study based on slope correlation random forest method. *Geophys. Geochem. Explor.* **2021**, *45*, 18–28. (In Chinese)
19. He, H.J.; Zhao, Y.N.; Zhang, Z.M.; Gao, Y.N.; Yang, L.W. Prediction of coalbed methane content based on uncertainty clustering method. *Energ. Explor. Exploit.* **2016**, *34*, 273–281. [CrossRef]
20. Yu, J.; Zhu, L.Q.; Qin, R.B.; Zhang, Z.S.; Li, L.; Huang, T. Combining K-Means Clustering and Random Forest to Evaluate the Gas Content of Coalbed Bed Methane Reservoirs. *Geofluids* **2021**, *2021*, 9321565. [CrossRef]
21. Tang, Y.; Li, L.Z.; Jiang, S.X.; Zhong, M.H. Parameter selection and applicability of gas content logging interpretation methodology in coal seam. *Coal Geol. Explor.* **2015**, *43*, 94–98. (In Chinese)
22. Hou, J.; Zou, C.C.; Yang, Y.Q.; Zhang, G.H.; Wang, W.W. Comparison study on evaluation methods of coalbed methane gas content with logging interpretation. *Coal Sci. Technol.* **2015**, *43*, 157–161. (In Chinese)
23. Li, Z.C.; Du, W.F.; Hu, J.K.; Li, D. Interpretation method of gas content in logging of Linxing block in Ordos Basin. *J. China Coal Soc.* **2018**, *42*, 490–498. (In Chinese)

24. Zou, Y.; Chen, Y.; Deng, H. Gradient Boosting Decision Tree for Lithology Identification with Well Logs: A Case Study of Zhaoxian Gold Deposit, Shandong Peninsula, China. *Nat. Resour. Res.* **2021**, *30*, 3197–3217. [CrossRef]
25. Liu, Z.; Gilbert, G.; Cepeda, J.M.; Lysdahl, A.O.K.; Piciullo, L.; Hefre, H.; Lacasse, S. Modelling of Shallow Landslides with Machine Learning Algorithms. *Geosci. Front.* **2021**, *12*, 385–393. [CrossRef]
26. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
27. Ho, T.K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 832–844.
28. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: Berlin/Heidelberg, Germany, 2013.
29. Bergstra, J.; Bardenet, R.; Bengio, B. Algorithms for hyper-parameter optimization. In Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS'11), New York, NY, USA, 12 December 2011; pp. 2546–2554.
30. Zhu, J.; Zhao, Y.H.; Hu, Q.J.; Zhang, Y.; Shao, T.S.; Fan, B.; Jiang, Y.D.; Chen, Z.; Zhao, M. Coalbed Methane Production Model Based on Random Forests Optimized by a Genetic Algorithm. *ACS Omega* **2022**, *7*, 13083–13094. [CrossRef]
31. Han, T.; Jiang, D.X.; Zhao, Q.; Wang, L.; Yin, K. Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. *Trans. Inst. Meas. Control* **2017**, *40*, 1681–8693. [CrossRef]
32. Owusu, E.B.; Tetteh, G.M.; Asante-Okyere, S.; Tsegab, H. Error Correction of Vitrinite Reflectance in Matured Black Shales: A Machine Learning Approach. *Unconv. Resour.* **2022**, *2*, 41–50. [CrossRef]
33. Hariharan, S.; Mandal, D.; Tirodkar, S.; Kumar, V.; Bhattacharya, A.; Lopez-Sanchez, J.M. A Novel Phenology Based Feature Subset Selection Technique Using Random Forest for Multitemporal PolSAR Crop Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 4244–4258. [CrossRef]
34. Song, S.; He, R.; Shi, Z.; Zhang, W. Variable Importance Measure System Based on Advanced Random Forest. *CMES* **2021**, *128*, 65–85. [CrossRef]
35. Guo, G.S.; Guo, J.H.; Sun, L.C.; Liu, L.F.; Tian, Y.J. 3D fine modeling of coal seam gas content based on random forest algorithm. *China Offshore Oil Gas* **2022**, *34*, 156–163.
36. Liu, W.; Fan, H.; Xia, M. Step-wise multi-grained augmented gradient boosting decision trees for credit scoring. *Eng. Appl. Artif. Intell.* **2021**, *97*, 104036. [CrossRef]
37. Huan, J.; Li, H.; Li, M.; Chen, B. Prediction of dissolved oxygen in aquaculture based on gradient boosting decision tree and long short-term memory network: A study of chang Zhou fishery demonstration base, China. *Comput. Electron. Agric.* **2020**, *175*, 105530. [CrossRef]
38. Liang, W.; Luo, S.; Zhao, G.; Wu, H. Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms. *Mathematics* **2020**, *8*, 765. [CrossRef]
39. Friedman, J. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
40. Xia, Y.F.; Liu, C.Z.; Li, Y.Y.; Liu, N.N. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Syst. Appl.* **2017**, *78*, 225–241. [CrossRef]
41. Injadat, M.; Salo, F.; Nassif, A.B.; Essex, A.; Shami, A. Bayesian optimization with machine learning algorithms towards anomaly detection. In Proceedings of the 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 9–13 December 2018; pp. 1–6.
42. Snoek, J.; Larochelle, H.; Adams, R. Practical Bayesian optimization of machine learning algorithms. *Adv. Neural Inf. Process. Syst.* **2012**, *4*, 2951–2959.
43. Cai, Y.D.; Liu, D.M.; Yao, Y.B.; Li, J.Q.; Qiu, Y.K. Geological controls on prediction of coalbed methane of No. 3 coal seam in Southern Qinshui Basin, North China. *Int. J. Coal Geol.* **2011**, *800*, 101–112. [CrossRef]
44. Wang, H.; Yao, Y.B.; Huang, C.C.; Liu, D.M.; Cai, Y.D. Fault Development Characteristics and Their Effects on Current Gas Content and Productivity of No. 3 Coal Seam in the Zhengzhuang Field, Southern Qinshui Basin, North China. *Energ Fuel.* **2021**, *35*, 2268–2281. [CrossRef]
45. Liu, D.M.; Yao, Y.B.; Wang, H. Structural Compartmentalization and Its Relationships with Gas Accumulation and Gas Production in the Zhengzhuang Field, Southern Qinshui Basin. *Int. J. Coal Geol.* **2022**, *259*, 104055. [CrossRef]
46. Wang, M.; Zhu, Y.M.; Li, W.; Zhong, H.Q.; Wang, Y.H. Tectonic evolution and reservoir formation of coalbed methane in Zhengzhuang block of Qinshui basin. *J. China Univ. Min. Technol.* **2012**, *41*, 425–431. (In Chinese)
47. Fu, X.H.; Qin, Y.; Wang, G.G.X.; Rudolph, V. Evaluation of gas content of coalbed methane reservoirs with the aid of geophysical logging technology. *Fuel* **2009**, *88*, 2269–2277. [CrossRef]
48. Liu, Z.D.; Wang, J.; Yang, X.C.; Chen, C.H.; Zhang, J.K. Analyzing on applicability of expanding influence correction method of density logging in the coalbed methane reservoir. *Prog. Geophys.* **2014**, *29*, 2219–2223. (In Chinese)
49. Li, B.T. A New Correction Method for Acoustic Log. *Well Logging Technol.* **1990**, *14*, 305–310. (In Chinese)
50. Finlay, S. Multiple classifier architectures and their application to credit risk assessment. *Eur. J. Oper. Res.* **2011**, *210*, 368–378. [CrossRef]
51. Li, Y.; Shami, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing* **2020**, *415*, 295–316. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.