Research Article

Yuchen Xie, Jingfu Zhang*, Jiancheng Wang, Hujia Zhu, and Shuai Xie

A physical model for calculating cementing quality based on the XGboost algorithm

https://doi.org/10.1515/phys-2022-0024
received February 13, 2022; accepted March 21, 2022

Abstract: A physical model can be used to judge cementing quality to help drilling engineering. This article reports a physical model based on the XGboost algorithm to solve the cementing quality prediction problem of oil and gas wells. Through the physical model, the nonlinear, time-varying, and uncertain influencing factors, the high latitude of the data set, the lack of data, data imbalance and other characteristics are comprehensively analyzed. Finally, through numerical example verification, the physical model we reported can effectively predict the key factors affecting quality, improve process quality and reduce unit cost.

Keywords: physical model, XGboost algorithm, quality prediction, artificial intelligence

1 Introduction

In recent years, artificial intelligence has spread to various fields [1], such as autonomous vehicles in the field of transportation [2], intelligent medical imaging in the field of medicine [3], robot escort in the field of elderly care [4], and so on. In these fields, data mining, data modeling, and data analysis play a key role in solving problems based on artificial intelligence.

In petroleum engineering, experts have proposed intelligent drilling to improve drilling efficiency [5]. The wellbore pressure and temperature gradient can be predicted by establishing a physical model [6]. Gradually, scholars developed a logging driller intelligent interpretation system based on intelligent drilling [7]. It has laid the foundation for promoting artificial intelligence drilling and completion engineering technology. However, there is still a lack of good physical models for field guidance in drilling and completion engineering. Particularly, in cementing engineering technology, due to many factors affecting cementing quality, there are nonlinear, time-varying, and uncertain factors [8].

Cementing is an important process in the drilling and completion of oil and gas wells [9]. High-quality cementing quality will improve the production life of oil and gas wells [10]. How to improve the quality of cementing technology has become a hotspot in the field of cementing research [11]. In the research on cementing quality prediction of oil and gas wells, experts and scholars from all over the world have used different methods to analyze the main factors affecting the cementing quality. On this basis [12], an evaluation method for cementing quality is initially established [13]. Li and Shi organically combined different logging evaluation methods to form a cementing quality evaluation system for different requirements [14]. Yang *et al.* established a casing-hole sound field model and improved the cementing quality evaluation standard and method by using the influence of cement slurry density [15]. Zhan and Zhu simulated different downhole environments through numerical simulations and physical simulations to realize the evaluation of cementing quality of oil and gas wells [16]. Yang *et al.* established a multifactor statistical model of cementing quality using a combination of grey correlation and fuzzy evaluation [17].

With the development of the artificial neural network, many scholars have begun to apply neural network theory to the prediction of cementing quality of oil and gas wells. Ai *et al.* established a multifactor cementing quality evaluation model for the first time using an orthogonal wavelet neural network [18]. Bu *et al.* used a neural network algorithm to establish a mathematical model for cementing quality prediction [19]. Lu *et al.* combined an immune optimization algorithm with BP neural network and proposed a cementing quality prediction model established by immune neural network [20]. Pan *et al.* used the previous research data to use database statistics to analyze and summarize the influence of various factors on cementing quality.

* **Corresponding author: Jingfu Zhang,** Key Laboratory of Enhanced Oil Recovery, Northeast Petroleum University, Daqing, Heilongjiang 163318, China, e-mail: nepu_zjf@163.com
Yuchen Xie, Jiancheng Wang, Hujia Zhu, Shuai Xie: Key Laboratory of Enhanced Oil Recovery, Northeast Petroleum University, Daqing, Heilongjiang 163318, China

With the advancement of science, solving problems through physical models has become an inevitable choice [21]. Oil and gas well cementing quality prediction mainly relies on neural network theory to establish related models. Using the neural network model must rely on field experience, *etc.*, and artificially select several representative data identified as the most important and substitute it into the model for prediction. However, there are a large number of characteristic parameters that affect the cementing quality. To ensure accurate data, we need an updated physical model to help determine cementing quality. Therefore, the establishment of a set of methods and models for intelligent analysis of cementing quality based on big data mining and intelligent calculation is of great significance for deepening the research and development of cementing engineering technology [22].

2 Physical model building

2.1 Physical model influencing factors

Cementing quality prediction models for oil and gas wells are based on factors that affect cementing quality. Whether the cementing quality is high or not depends on the influence of the formation conditions, wellbore conditions, cementing equipment, cementing design, and other factors on the cementing process. In this article, combined with cementing construction experience and research literature, the following four major influencing factors are summarized to collect data and establish physical models: (1) formation and wellbore factors, (2) cement slurry factors, (3) drilling fluid factors, and (4) construction operation factors.

Formation and wellbore factors include the following: location, depth, formation pressure, minimum formation fracture pressure, bottom hole mixing, well type, wellbore quality, annular gap, *etc.*; cement slurry factors include the following: cement slurry segment density, fluidity, thickening time, filtration loss, cement dosage, displacement selection, *etc.*; drilling fluid factors include the following: drilling fluid segment density, drilling fluid water loss, drilling fluid shear force, total drilling fluid, displacement selection, *etc.*; construction operation factors including the following: displacement selection, cement slurry technology, slurry mixing equipment and accessories, casing treatment, *etc.*

2.2 Data processing

According to the above theory, as a basis, the model characteristic data are collected. In this study, the field data of an oil field are taken as an example to collect data. At the same time, the establishment of a database for data collection, storage, and invocation lays the foundation for subsequent modeling. The collected data are given in Table S1.

After analyzing the collected data, data preprocessing is performed, including data screening, missing value processing, one-hot coding, data balance processing, and normalization processing.

The samples whose missing value feature of a single sample is greater than 30% of the total number of features are directly deleted, and the samples with less than 30% of the missing value are taken as a separate feature for processing; the category label features are processed by one-hot encoding; the query is positive. After the proportion of negative samples, it is found that the proportion of negative samples is relatively small, and oversampling and repeated sampling are adopted; the order of magnitude difference of each feature data is calculated according to the statistical analysis of the database, and it is judged whether standardization and normalization are required.

When the order of magnitude difference between the variables of cementing quality data is too large, after standardizing and normalizing the data, the process of finding the optimal solution will become smoother, and it will be easier to converge to obtain the optimal solution. In addition, the processed data reduce the influence of abnormal data in training, making each feature data more comparable. The data of each dimension in the data set are brought into the model so that the variance is 1 and the mean is 0. The standardized and normalized models are as follows:

$$x^{\star} = \frac{x - \mu}{\sigma}, \tag{1}$$

where x is the cementing quality data, μ is the mean value of the cementing quality data, and σ is the standard deviation of the cementing quality data.

The processed features are represented by fn.

2.3 Physical model based on the XGboost algorithm

Extreme gradient boosting (XGboost), an implementation of extreme gradient trees, is very important in most regression and classification issues. The algorithm is based

on the traditional gradient improvement of the gradient boosting decision tree algorithm. The traditional gradient improvement algorithm is a step-of-order development of the previous round of loss functions, while XGboost uses Taylor's fifth-order development to fit. It is mainly manifested in the approximation of the loss function through Taylor's second-order expansion and the use of regularization to reduce overfitting, which belongs to ensemble learning. The purpose of integrated learning is to combine the prediction results of multiple base learners with improving single learning, generalization ability, and robustness of the device. Its advantages are fast speed, good effect, ability to process large-scale data, support for multiple languages, support for custom loss functions, and so on. Therefore, the accuracy of the physical model based on the XGboost algorithm is higher, the same training effect is satisfied, and the number of iterations is fewer, which makes the model easier, avoiding the fit.

A general way to build the optimal model is to minimize the loss function of training data; assume that there are K decision trees in the model:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in D, \tag{2}$$

where f_k is a function in the function space D, $\hat{y}_i$ is the predictive value, x_i is the input of the text of i sample, and D is all possible collections of all possible Classification and Regression Tree.

The XGboost algorithm adds each time a new tree, and it is assumed that the predicted value of the term is $\hat{y}^{(t)}i$, the derived process is as follows:

$$\begin{cases} \hat{y}_i^{(0)} = 0, \\ \hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i), \\ \hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i), \\ \hat{y}_i^{(3)} = f_1(x_i) + f_2(x_i) + f_3(x_i) = \hat{y}_i^{(2)} + f_3(x_i), \\ \vdots \\ \hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i). \end{cases} \tag{3}$$

As shown in Eq. (4), the target function of the XGboost algorithm is the loss function

$$\begin{aligned} \text{Obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) \\ + \Omega(f_t) + \text{con}, \end{aligned} \tag{4}$$

where $\Omega(f_t)$ is the regular item, which reduces complexity and prevents the effect of predation. Take $\hat{y}_i^{t-1} + f_t(x_i)$ in the target function as a whole, and this overall value is taken at $\hat{y}_i^{(t-1)}$

$$\begin{cases} \text{Obj} \cong \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_i(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] \\ \qquad + \sum_{i=1}^{t-1} \Omega(f_i) + \Omega(f_t), \\ g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}), \\ h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}), \end{cases} \tag{5s}$$

where g_i, h_i is one order and two steps. All constant items are removed in the function:

$$\begin{aligned} \text{Obj} \cong \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_i(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] \\ + \sum_{i=1}^{t-1} \Omega(f_i) + \Omega(f_t). \end{aligned} \tag{6}$$

The regular item for the tree in XGboost is defined as follows:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2, \tag{7}$$

where w_j is the fractional value on the j left leaf node in tree f and T is the total number of leaves nodes in tree f. γ and λ are custom parameters for XGboost, γ is the L1 regular item, and λ is the L2 regular item.

Define the sample set on each leaf node j as $I_j = \{i | q(x_i) = j\}$, After substituting the function f, the regular term and the leaf node sample set into the objective function formula (6), We can get a new formula (8). The final objective function formula (8) is obtained as follows

$$\begin{aligned} \text{Obj} &\cong \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i \right) w_j^2 \right] + \gamma T + \lambda \frac{1}{2} \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \\ &= \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T. \end{aligned} \tag{8}$$

When the derivative of formula (8) is 0, we can get the optimal value of X. Bringing the optimal value of X into the objective function, the final loss can be obtained as:

$$\begin{cases} w_j^{\star} = -\dfrac{G_j}{H_j + \lambda}, \\ \text{Obj}^{\star} = -\dfrac{1}{2} \sum_{j=1}^{T} \dfrac{G_j^2}{H_j + \lambda} + \lambda T. \end{cases} \tag{9}$$

As can be seen from Eq. (9), the result is best when w_j^* decreases to the minimum.

3 Results

Through oilfield on-site case verification, the predictive model established herein, the data set is the data set in the field database, which belongs to the high latitude and less sample quantity. The data set is the final quality of the well, where "f0," "f1," and "f124" indicate the quality characteristics of the sample, "*Y*" indicates the quality result, "1" indicates that qualification, and "0" means unqualified.

In the process of establishing a model, in order to verify the XGboost algorithm more accurately, three algorithms are used to simulate experiments: XGboost, random forest and linear regression. According to the score of training and test values in the simulation experiment, XGboost is shown in Figure 1, for which the score difference ranges between 0.05 and 0.2; and random forest and linear regression are shown in Figures 2 and 3, respectively. The score difference between the training value and the test value is between 0.15 and 0.34 and between 0.2 and 0.8, respectively.

Comparing the training value and the test value of the XGboost algorithm model, the score difference is in the optimal range in three algorithms; the accuracy of the XGboost test set reaches 85%, while that of the random forest reaches only 70%, indicating that these two algorithms fit well. While the degree is high, the linear regression algorithm is almost unqualified. The verification results have proved that the XGboost algorithm is more accurate, fits better, and is more suitable for predicting the quality of oil and gas wells.

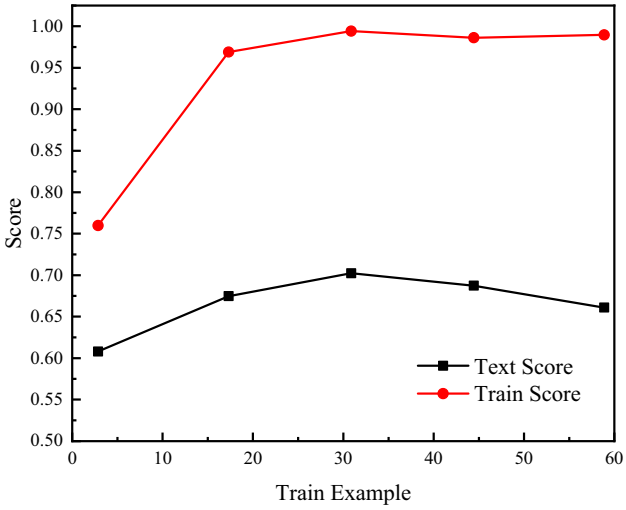

Figure 2: Random forest algorithm fitting graph.

To further improve the accuracy of the model prediction, the built-in parameters of XGboost are optimized, the main parameters are as follows: n_estimators, max_depth, min_child_weight, subsamples, colsample_bytree, *etc.* where the n_estimators parameter determines the number of model iterations, as shown in Figure 4. When the value of n_estimators is 900, the loss value reaches the lowest point of 0.1915, and the fit is optimal.

The max_depth parameter is the maximum depth of the tree; adjusting this value to avoid the effect of

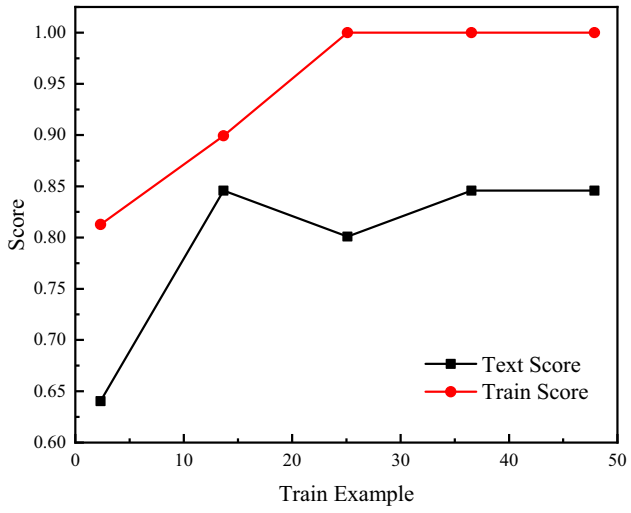

Figure 1: XGboost algorithm fitting graph.

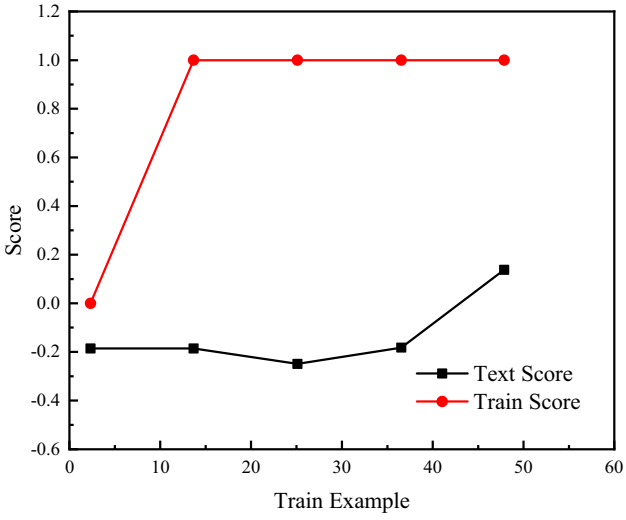

Figure 3: Linear regression algorithm fitting graph.

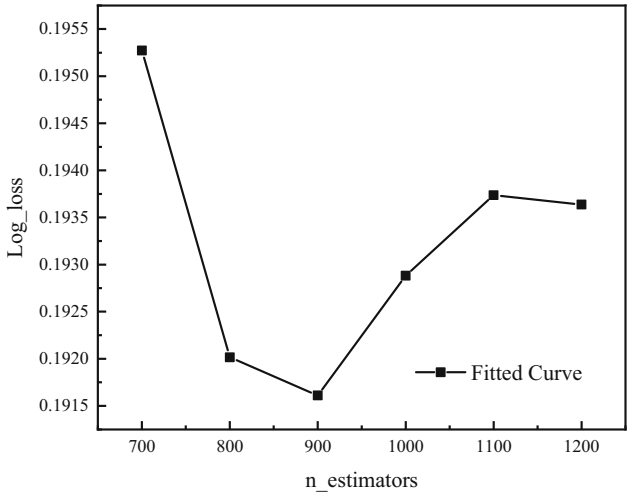

Figure 4: n_estimators parameter tuning map.

predation makes the prediction effect of the test set more accurate. As shown in Figure 5, when the value of max_-depth is 3, the loss value reaches the lowest value of 0.19, and the fit is optimal.

The min_child_weight parameter determines the minimum leaf node sample weight and adjusts the parameter to avoid the fit. As shown in Figure 6, when the value is 1, the loss value is 0.19 at the lowest, and the fit is optimal.

The subsample parameter controls for each tree randomly sampled, and it is used to optimize the fitting effect. As shown in Figure 7, when the value is 1, the loss value reaches 0.19, and the fit is optimal.

The colsample_bytree parameter is used to control the proportion of each tree randomly sampled, which is used to for debugging fit effects. As shown in Figure 8, when the value is 0.9, the loss value reaches a minimum of 0.187, and the fit is optimal.

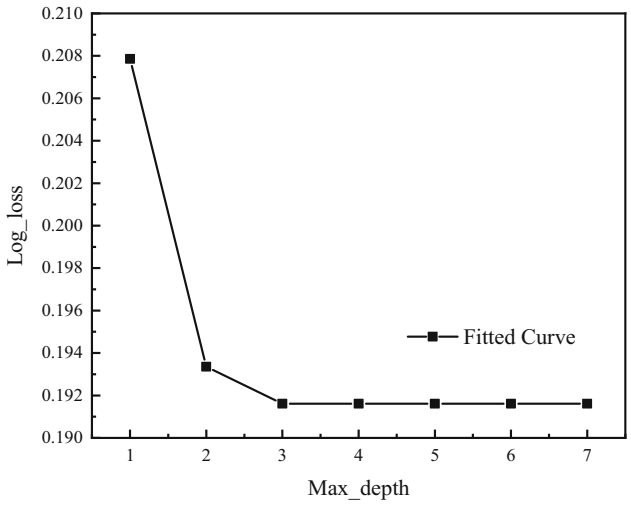

Figure 5: Max_depth parameter tuning map.

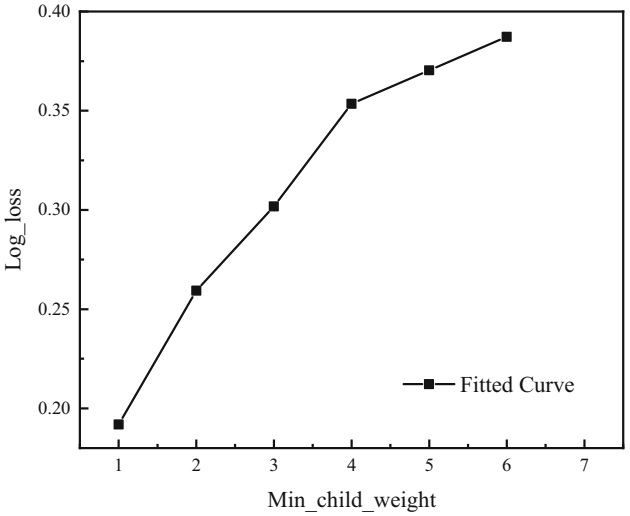

Figure 6: Min_child_weight parameter tuning map.

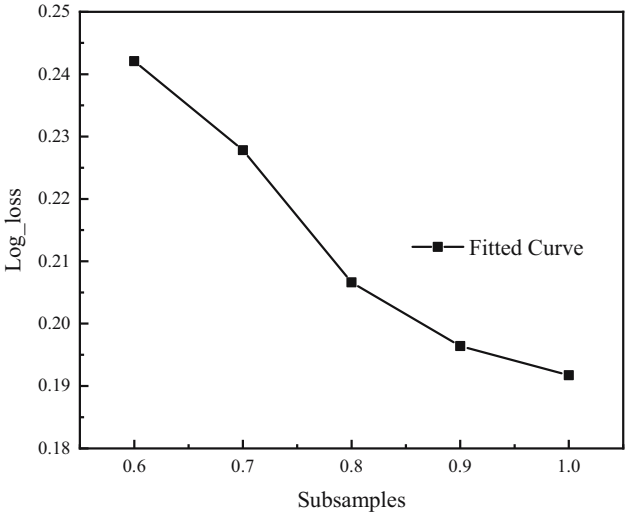

Figure 7: Subsample parameter tuning map.

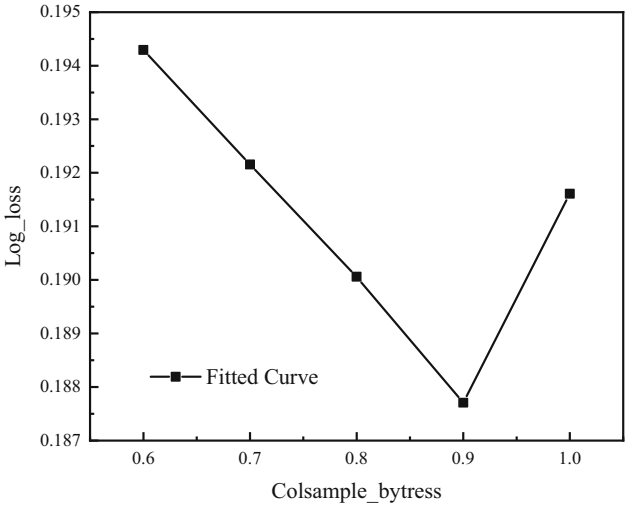

Figure 8: Colsample_bytree parameter tuning map.

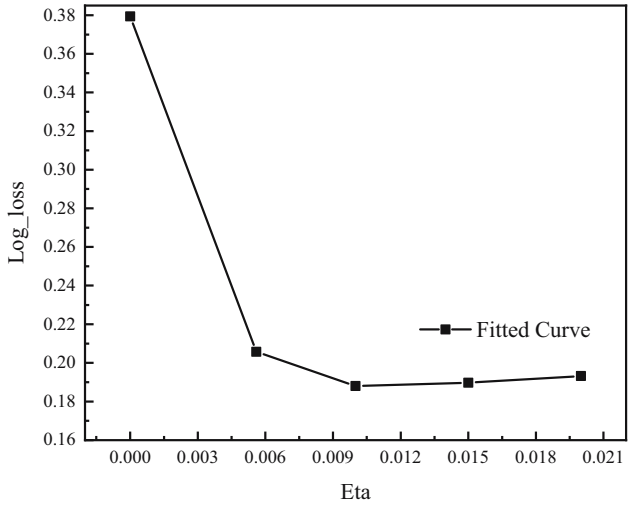

Figure 9: Eta parameter tuning map.

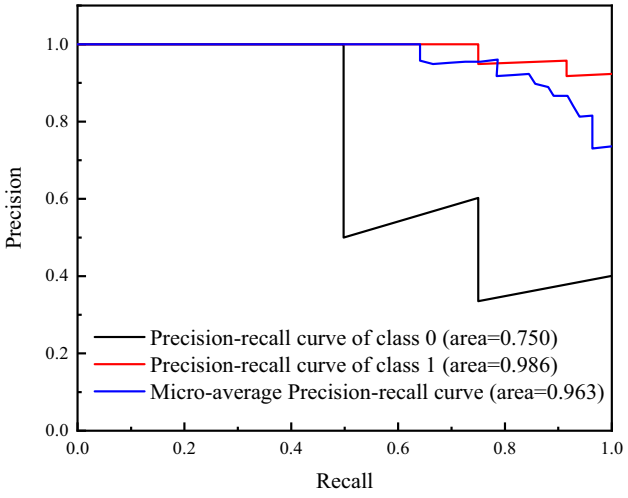

Figure 11: P–R graph.

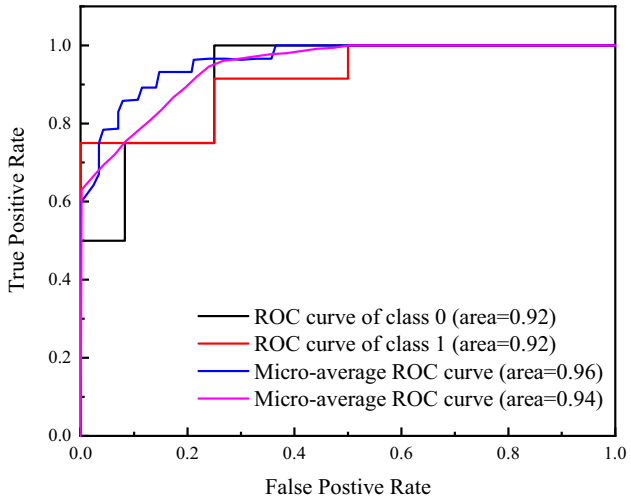

Figure 10: ROC evaluation indicator map.

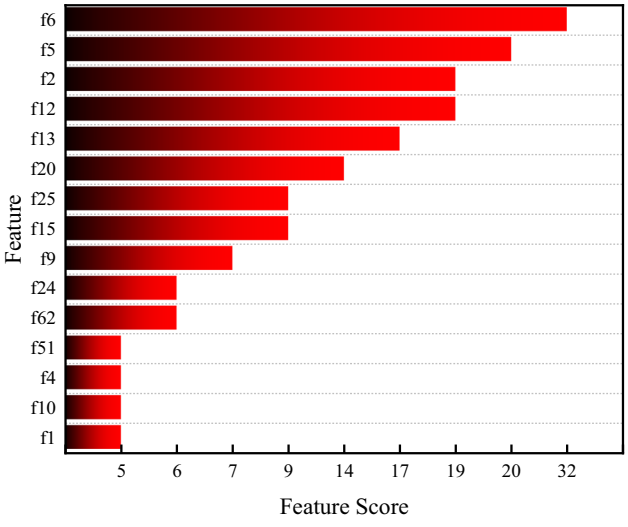

Figure 12: Feature parameter weight score diagram.

The eta parameter refers to the learning rate and improves the robustness of the model by reducing the weight of each step. As shown in Figure 9, when the value is 0.01, the loss value reaches a minimum of 0.187, and the fit is optimal.

After parameter tuning, the test set score increased from 0.85 to 0.89; that is, the prediction accuracy reached 89%. During the training iteration, the prediction accuracy of the model and the extensive ability of model forecasts are changed each time, but they all meet the 95% confidence interval. Finally, the receiver operating characteristic (ROC) curve of the model and the P–R curve are shown in Figures 10 and 11.

The area value surrounded by the curve is basically more than 90%; with a large predictive value, the model has a good application value.

According to model training, the weight score of each feature variable is shown in Figures 3–12, the characteristic parameters of the most critical key of the well quality can be obtained as f6, f5, f2, f12, f13, f20 and f25. These represent 1. Cement slurry segment density #2, 2. Cement slurry segment density #1, 3. Cement slurry minimum displacement, 4. Displacement minimum displacement, 5. Cement slurry 50BC thickening time, 6. Drilling Fluid segment density #1, 7. Drilling fluid segment density #2; The secondary key specific parameters are f15, f9, f24, f62, f51,

Table 1: Key feature mapping table

fn	Key characteristic	fn	Second key signature
f6	Cement slurry second segment density	f15	Total grout
f5	Cement slurry first segment density	f9	Cement slurry injection time
f2	Cement slurry minimum displacement	f24	Rushing density
f12	Rigid minimum displacement	f62	Well diameter expansion rate
f13	Cement 50bc thickening time	f51	Cement back
f20	Drilling liquid first segment density	f4	Rolling liquid
f25	Drilling liquid second segment density	f10	Flushing liquid injection time
		f1	Maximum displacement of cement slurry

f4, f10, f1, respectively 1. total amount of cement slurry, 2. cement slurry injection time, 3. flushing fluid density, 4. well diameter expansion rate %, 5. Cement return height, 6. Total amount of flushing fluid, 7. Flushing fluid injection time, 8. Maximum displacement of cement slurry

This score diagram is obtained by averaging the superposition of the feature gain. In the XGboost prediction model, the tree is branched with the greed method, and the gain means the relative contribution of the model calculated by the contribution of each feature through each tree in the model. In other words, a high gain value means that it is more important for generating prediction models (Table 1).

Through the characteristics of high research weight scores, these parameters play a key role in the top replacement efficiency, and in the model of conventional assessment of solid wells, the top replacement efficiency has also the most critical influence on the quality of the well. The research evaluation plays a major role. The side of this result verifies that the results predicted by this research model have greater absolute reliability.

4 Conclusion

The combination of artificial intelligence big data and cementing engineering technology is a new development of cementing engineering technology in the new era. This article reports a new physical model that can effectively process, analyze, and evaluate a large number of complex cementing quality influencing parameters and their validity.

The results of this research have been modeled through XGboost, random forest, linear regression, and other machine learning algorithms. It is not difficult to see the effect of the XGboost algorithm on each model. The prediction accuracy is 89%. Through the physical model, the maximum impact in cement slurry density, drilling fluid density, replacement displacement, cement pulp performance, *etc*., can be determined. In the cementing design of oil and gas wells in the future, the cementing design calculation can be carried out based on this model. Ultimately, this model can further improve the efficiency of cementing design work.

Funding information: The authors state no funding involved.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: The authors state no conflict of interest.

References

[1] Taniguchi M, Minami S, Ono C, Hamajima R, Tomono K. Combining machine learning and nanopore construction creates an artificial intelligence nanopore for coronavirus detection. Nat Commun. 2021;12(1):3726.
[2] Bonnefon JF, Shari FF, Rahwan I. The social dilemma of autonomous vehicles. Science. 2016;352(6293):1573–6.
[3] Zhou Y, Zhou T, Zhou T, Fu H, Shao L. Contrast-attentive thoracic disease recognition with dual-weighting graph reasoning. IEEE T Med Imaging. 2021;99:1.
[4] Qiu X, Feng Z, Xu T, Yang X, Zhang X. Research on intention flexible mapping algorithm for elderly escort robo. Sci Program. 2021;8:1–14.
[5] Wang MS, Guang XJ. Status and development trends of intelligent drilling technology. Acta Petrolei Sin. 2020;41(4):505–12.
[6] Carpenter C. Intelligent drilling advisory system optimizes performance. J Prerol Technol. 2020;72(2):65–7.
[7] Tewari S, Dwivedi UD, Biswas S. Intelligent drilling of oil and gas wells using response surface methodology and artificial bee colony. Sustainability. 2021;13(4):1–27.
[8] Egorova EV, Minchenko YS, Dolgova UV, Selivanov SV, Salavatov TS. Study of dispersed-reinforced expanding

plugging materials to improve the quality of well cementing. Earth Environ Sci. 2021;745(1):12019.
[9] Zheng S, Li W, Cao C, Wang C. Prediction of the wellhead uplift caused by HT–HP oil and gas production in deep-water wells. Energy Rep. 2021;7:740–9.
[10] Deryugina OP, Trapeznikov EA. The issue of "oil shrinkage" during the compounding of oils in the processes of production, collection, preparation and transportation of hydrocarbon raw materials. Oil Gas Stud. 2021;2:104–13.
[11] Xi Y, Lian W, Fan L, Tao Q, Guo X. Research and engineering application of pre-stressed cementing technology for preventing micro-annulus caused by cyclic loading-unloading in deep shale gas horizontal wells. J Pet Sci Eng. 2021;200(2):108359.
[12] Zheng S, Zhang C. Influence of cement return height on the wellhead uplift in deep-water high-pressure–high-temperature wells. ACS Omega. 2021;6:2990–8.
[13] Xu BC, Zhou JL, Liu W, Fu JS. Data driven prediction method for gas cut in drilling process. Acta Pet Sin. 2019;40(10):1263–9.
[14] Li DW, Shi GR. Optimization of common data mining algorithms for petroleum exploration and development. Acta Pet Sin. 2018;39(2):240–6.
[15] Yang JH, Qiu MX, Hao HN, Zhao X, Guo XX. Intelligence-oil and gas industrial development trend. Pet Sci Technol Forum. 2016;35(6):36–42.
[16] Zhan XD, Zhu ZX. Study of intelligent drilling technology. Oil Drill Pro Technol. 2010;32(1):1–4 + 16.
[17] Yang CS, Li CS, Sun XD, Huang LM, Zhang HL. Research method and practice of artificial intelligence drilling technology. Pet Drill Technol. 2021;49(5):7–13.
[18] Ai C, Bu ZD, Zhao WC, Li Q. Cementation quality prediction using wavelet neural network based on orthogonal scaling function. Pet Drill Technol. 2008;36(6):56–8.
[19] Bu YH, Song WY, He YJ, Shen ZC. Dicussion of a method for evaluating cementing quality with low-density cement slurries. Pet Drill Technol. 2015;43(5):49–55.
[20] Lv HY. Applications of neural network in prediction of cementing quality. Pet Drill Technol. 2002;30(3):24–6.
[21] Sohail M, Ali U, Zohra T, Al-Kouz W, Thounthong P. Utilization of updated version of heat flux model for the radiative flow of a non-Newtonian material under Joule heating: OHAM application. Open Phys. 2021;19(1):100–10.
[22] Elmaboud YA, Abdelsalam SI. DC/AC magnetohydrodynamic-micropump of a generalized Burger's fluid in an annulus. Phys Scrip. 2019;94(11):115209 (13pp).