ISSN: 1813-4890

Research on coalbed methane production capacity prediction model based on machine learning

Yuechun Chen[1,2], Xi Lin[1,2] and Jie Wang[1,2]

1School of College of Earth Science and Engineering, Xi'an Shiyou University, Xi'an 710065, China;

2Shaanxi Provincia Key Laboratory of Petroleum Accumulation Geology, Xi'an Shiyou University, Xi'an 710065, China

Abstract

As the core area rich in coalbed methane resources in China, the Ordos Basin has geological characteristics of low porosity, low permeability, and strong heterogeneity, which makes it difficult to predict productivity. It is difficult for traditional methods to accurately capture the nonlinear relationship between geological factors and productivity. Taking the DJ block of the basin as the research object, 100 sets of core porosity, well logging and drainage data were integrated, and 10 main productivity control factors such as permeability and gas saturation were selected through gray correlation analysis. Three machine learning prediction models, namely BP neural network, gradient boosting tree (GBDT) and random forest (RF), were constructed, and the four indicators of MAE, MAPE, RMSE and R^2 were used to compare model performance. The results show that the gradient boosting tree model has the best prediction accuracy, with a test set MAPE of 1.14% and an R^2 of 0.8827, which is significantly better than the BP neural network (MAPE=1.10%, R^2=0.8621) and random forest (MAPE=1.26%, R^2=0.8498);This model can effectively adapt to the basin's small sample and strong noise data characteristics, and can provide technical support for coalbed methane development well location deployment and drainage plan optimization.

Keywords

Machine learning; Ordos Basin; coalbed methane; production capacity prediction; gradient boosting tree machine.

1. Introduction

Coalbed methane, as a clean and low-carbon unconventional natural gas resource, is an important support for China's energy structure transformation. The Upper Paleozoic coalbed methane resources in the Ordos Basin exceed 10 trillion m^3, accounting for more than 30% of the national total. However, its geological conditions are complex—it has experienced gas formation in shallow peat burials, deep burial pyrolysis, and uplift and escape. And reburial supplemented the four-stage accumulation evolution, resulting in an average reservoir porosity of only 7.8% and an average permeability of 1.18×10^{-3} μm^2, showing the characteristics of "low porosity, low permeability, and strong heterogeneity".As a core link in coalbed methane development, production capacity prediction directly determines the rationality of well location selection and the formulation of drainage systems. Traditional methods based on physical simulation or statistical regression are difficult to depict the complex non-linear relationship between multiple geological factors and production capacity. The prediction error often exceeds 15%, which cannot meet the needs of efficient development [1]

ISSN: 1813-4890

My country's coal resource reserves are large and widely distributed, and the coal bed methane resource reserves that coexist with it are also considerable. Achieving efficient development of coalbed methane will help alleviate the supply pressure of conventional energy in my country.Prediction of coalbed methane well productivity is the basis for scientific evaluation of coal reservoir gas production capacity and discharge technology effects. It is also an important basis for measuring exploration results and planning production capacity layout. Accurate production capacity forecasting can provide support for formulating a reasonable production discharge system and has important guiding value for actual production.

At present, commonly used coalbed methane production capacity prediction methods include volume calculation method, material balance method and numerical simulation method. The volume calculation method is suitable for estimating resources at different levels, and its accuracy depends on geological understanding and parameter accuracy.This method is simple and easy to use, but it is sensitive to unknown parameters or subjective judgments, and the error may be large; the material balance method uses dynamic data to infer reservoir characteristics. The longer the production history and the richer the data, the more reliable the prediction results are; the numerical simulation method relies on core experiments, drilling and well test data to generate productivity curves through historical matching, which is suitable for full life cycle prediction.

However, the above methods have certain limitations in predicting daily gas production, such as weak regional adaptability, many parameter requirements, high data accuracy requirements, complex calculations, limited simulation capabilities for complex production mechanisms such as multi-phase seepage, and difficulty in clearly revealing the changes in gas production over time. The reliability and timeliness of prediction results still need to be improved.With the development of artificial intelligence technology, machine learning algorithms have shown significant advantages in the field of oil and gas resource prediction by virtue of their data-driven nonlinear fitting capabilities. Yao Huifang et al. found in their study of the DJ block in the Ordos Basin that the gradient boosting tree algorithm can classify coal-measure tight sandstone gas reservoirs with an accuracy of 92% [2];Chen Jiahao's research on the Linxing block shows that from four aspects: data enhancement, feature optimization, algorithm tuning and performance analysis, an efficient work plan for intelligent classification of tight gas well logging productivity has been established, that is, using the pipeline pipeline framework to combine ADASYN adaptive oversampling and ClusterCentroids clustering center undersampling. Sample algorithms were connected to enhance the original productivity data; the random forest algorithm was used to numerically evaluate the feature importance, and the key parameters were comprehensively selected based on the reservoir characteristic cross plot method. The CatBoost algorithm was used to build an accurate tight gas productivity classification model and deployed it on the CNOOC artificial intelligence application research platform [3].However, existing research mostly focuses on reservoir classification or single parameter prediction, and the construction of productivity prediction models under all geological conditions of the basin is still insufficient. Therefore, this study selects the optimal productivity prediction model suitable for the Ordos Basin through multi-model comparison, which has important engineering significance for improving coalbed methane development efficiency and reducing development risks.

The core goal of this research is to construct a high-precision and highly adaptable coalbed methane production capacity prediction model in the Ordos Basin. The specific contents include: (1) identifying the main geological factors that affect production capacity based on gray correlation analysis; (2) constructing three machine learning models of BP neural network, gradient boosting tree (GBDT), and random forest (RF); (3) verifying model performance through multi-index comparison to determine the optimal prediction model; (4) analyzing the engineering application value of the optimal model.

ISSN: 1813-4890

2. Geological overview

2.1. Overview of the study area

The Ordos Basin is the second largest sedimentary basin on land in my country and an important energy base. It is rich in mineral resources. The current basin starts from the Yinshan Mountains in the north, to the Qinling Mountains in the south, to the Liupan Mountains in the west, and to the Luliang Mountains in the east. It spans the five provinces of Shaanxi, Gansu, Shanxi, Ningxia, and Inner Mongolia. The main area of the basin is about 25×104km^2. During the sedimentation period of the Benxi Formation in the Late Carboniferous, the Ordos Basin was in an environment of alternating sea and land phases [4].

The DJ area of the study area is located in the southern section of the Shanxi flexural fold belt on the eastern edge of the basin. Its overall structural shape is a monocline that dips gently to the northwest. However, wide and gentle folds and faults with NE and NNE trends are developed in the central and southeastern edges, which complicates the structure. Accordingly, it can be divided into three structural units: the basin edge fault fold belt, the slope depression and uplift belt, and the western gentle slope belt;This structural pattern directly affects the preservation and seepage capacity of coalbed methane. In particular, fracture development zones (such as near Wucheng and Xueguan) provide favorable seepage channels for high coalbed methane production [5].In terms of stratigraphy, the main coal-bearing seams are the Shanxi Formation of the Upper Paleozoic Permian and the Benxi Formation of the Carboniferous, which were formed in the lagoon-tidal flat depositional system. The mud flat microfacies is rich in organic shale, and the barrier sand bar microfacies is a tight sandstone reservoir, which together form a coal measure gas symbiosis system.The physical properties of the reservoir are characterized by ultra-low porosity and ultra-low permeability, but the Shanxi Formation coal seams have high organic matter content (TOC average 3.93%), high thermal evolution degree (Ro 2.02%~2.61%), high hydrocarbon generation potential, and microscopic pores are mainly organic pores and intergranular pores, providing space for gas occurrence.Key geological parameters that affect productivity include coal seam thickness (H), permeability (K), gas saturation (Sg), porosity (φ) and reservoir pressure (P). These parameters are jointly controlled by sedimentary microphases and structural fractures: for example, high permeability zones are often distributed in fracture-intensive areas, while thick coal seams and high gas content (Q_g) areas provide the material basis for high production.Therefore, when predicting coalbed methane productivity in this area, it is necessary to comprehensively consider the improved seepage conditions of structural fractures, the storage-controlling characteristics of sedimentary facies, and the nonlinear coupling relationship of reservoir physical parameters (such as K, φ, Sg), thereby providing a geological basis for the production model (such as Q) and guiding the optimization of fracturing sections and development strategies

ISSN: 1813-4890

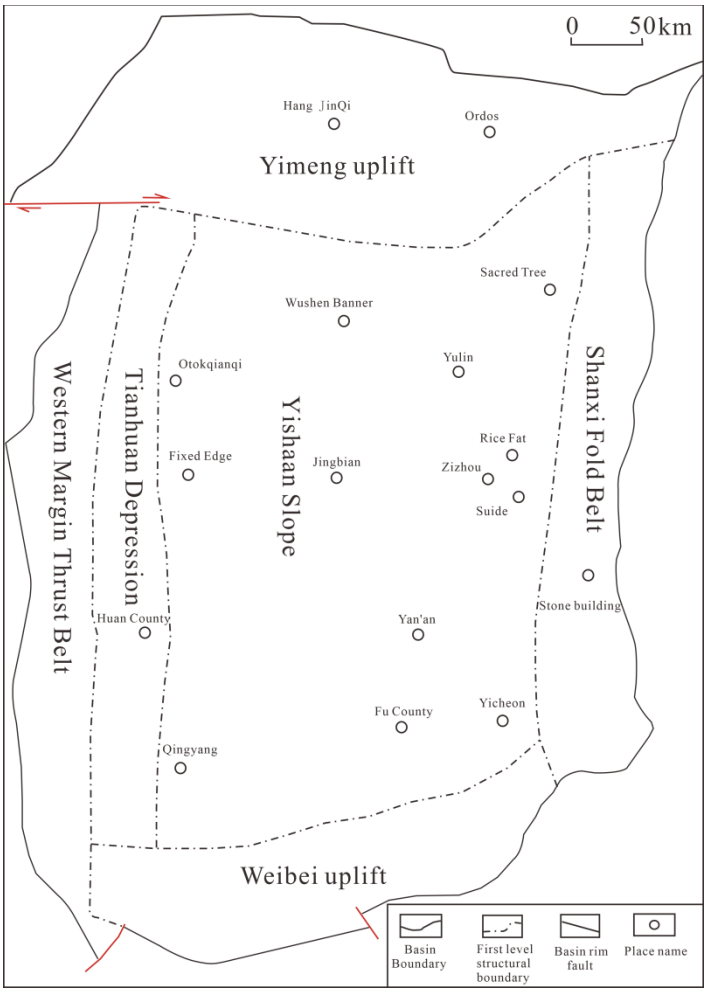

Figure 1 Map of the Ordos Basin

2.2. Geological factors

2.2.1. Penetration

Permeability is a key parameter that measures the ability of gas to flow in coal seams. The permeability in the study area ranges from 0.1 to $5.0\times10^{-3}\mu m^2$, which is significantly positively correlated with daily gas production. The higher the permeability, the smoother the gas permeation channel and the easier it is for the desorbed gas to be extracted, thereby increasing production capacity.

2.2.2. Gas saturation

Gas saturation reflects the proportion of gas occupying pores in the coal seam. Sg in the study area ranges from 60% to 92%, which is positively correlated with production capacity. The higher the Sg, the greater the gas content per unit volume of the coal seam, which can provide a more sufficient gas source basis for drainage.

2.2.3. Reservoir pressure gradient

Reservoir pressure gradient is the main driving force for gas seepage. P_grad in the study area is 0.15 to 0.35MPa/100m, which is positively related to production capacity. A higher pressure gradient helps promote gas migration toward the wellbore and improves gas recovery efficiency.

2.2.4. Coal seam thickness

Coal seam thickness has an important impact on coalbed methane enrichment and reservoir scale. The thickness of the coal seam in the study area ranges from 1.2 to 8.5m, and has a significant positive correlation with the average daily gas production. The greater the thickness

ISSN: 1813-4890

of the coal seam, the greater the adsorption/desorption space provided by the coal reservoir, and the increased gas storage capacity. At the same time, the distance for gas migration to the top and floor is lengthened, and the diffusion resistance increases, which is beneficial to the preservation and enrichment of coalbed methane, thus increasing single well productivity.

2.2.5. Porosity

Porosity refers to the proportion of pore volume in the coal seam to the total volume. φ in the study area ranges from 3.5% to 12.0%, which is positively related to production capacity. The higher the porosity, the larger the gas storage space and the greater the amount of adsorbed gas, which is beneficial to improving the ultimate recovery rate.

2.2.6. Gas content

Gas content represents the volume of gas contained in unit mass of coal. Q_g in the study area is 1.0 to 8.0m^3/t, which is significantly positively related to production capacity. Gas content directly determines the upper limit of theoretical gas production of a single well and is a direct reflection of resource abundance.

2.2.7. Burial depth

Burial depth affects ground pressure and gas content. The burial depth in the study area is 800 to 2500m, which is positively related to production capacity. As the burial depth increases, the formation pressure increases and the sealing property increases, which is beneficial to gas preservation and pressure accumulation.

2.2.8. Formation pressure

Formation pressure is the original driving force for gas production. P in the study area is 10.0 to 35.0MPa, which is positively related to production capacity. High formation pressure can enhance gas desorption efficiency and seepage capacity.

2.2.9. Rock density

Rock density reflects the compactness of coal rock. ρ in the study area is 1.3 to 1.8g/cm^3, which is negatively related to production capacity. The lower the density, the looser the coal rock and the better the pore development, which is conducive to gas adsorption and migration.

2.2.10 Temperature gradient

Temperature gradient affects gas adsorption-desorption equilibrium. T_grad in the study area is 1.8 to 3.2°C/100m, which has a complex relationship with production capacity. Moderate temperature rise is conducive to desorption, but too high a temperature may reduce the adsorption capacity, which requires a comprehensive judgment based on burial depth and pressure.

The relationship between geological factors and daily gas production is shown in the figure below:

ISSN: 1813-4890

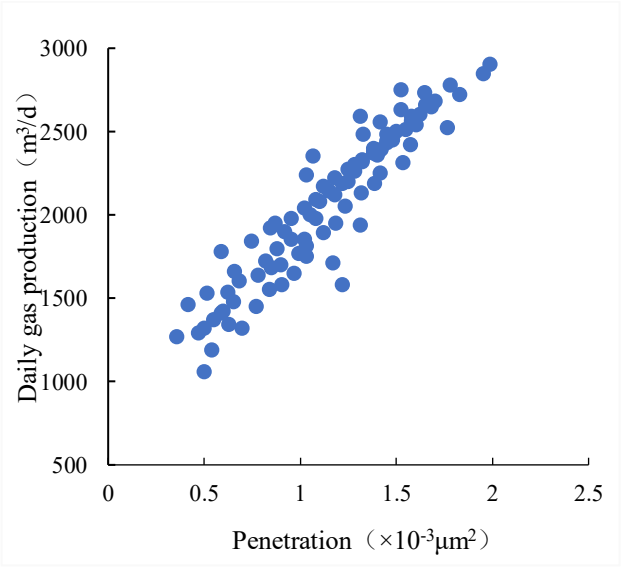

Figure 2 Cross diagram of permeability and daily gas production

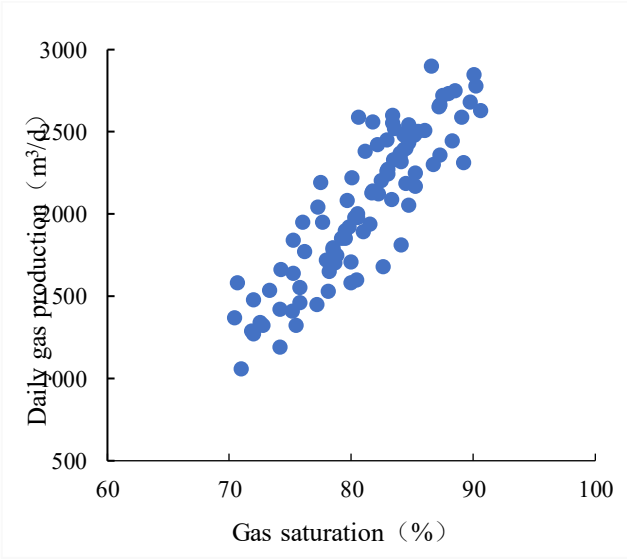

Figure 3 Cross diagram of gas saturation and daily gas production

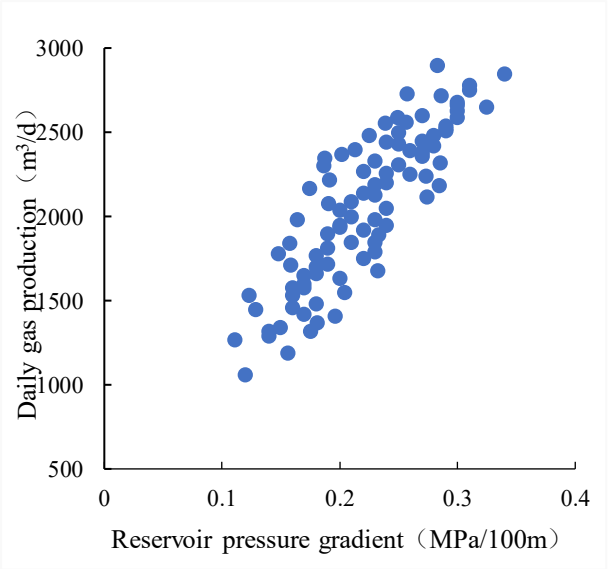

Figure 4 Cross diagram of reservoir pressure gradient and daily gas production

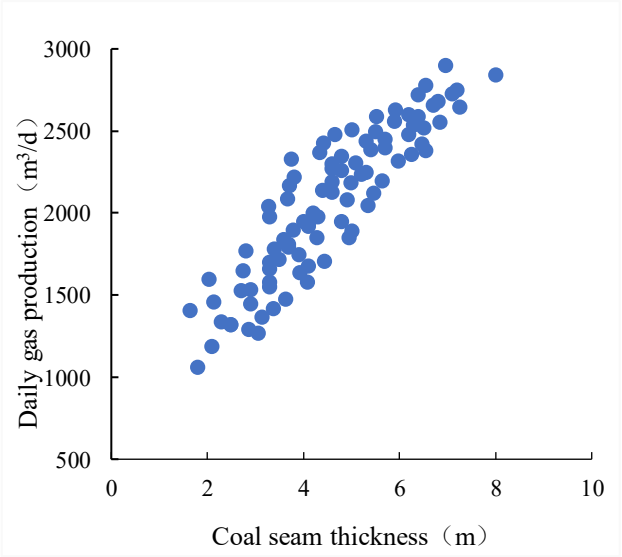

Figure 5 Cross diagram of coal seam thickness and daily gas production

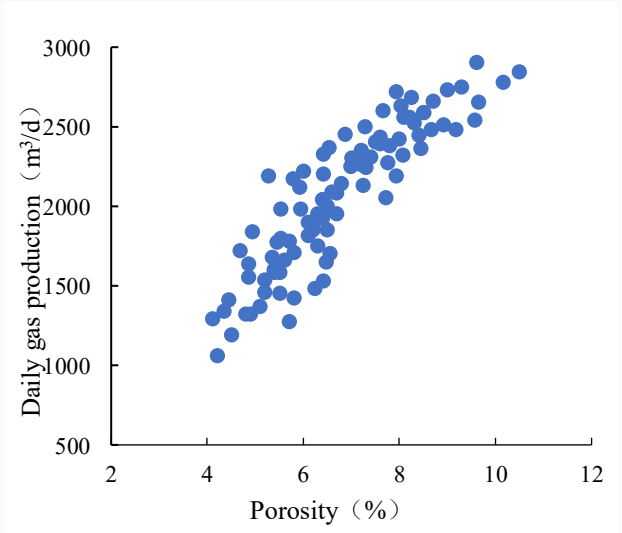

Figure 6 Cross plot of porosity and daily gas production

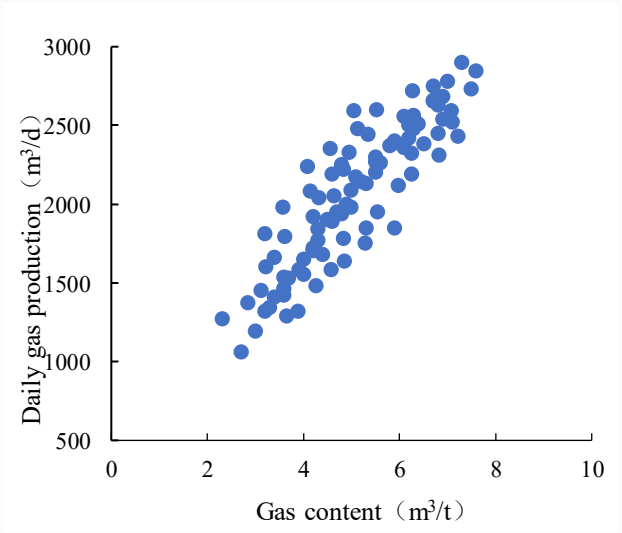

Figure 7 Cross diagram of gas content and daily gas production

ISSN: 1813-4890

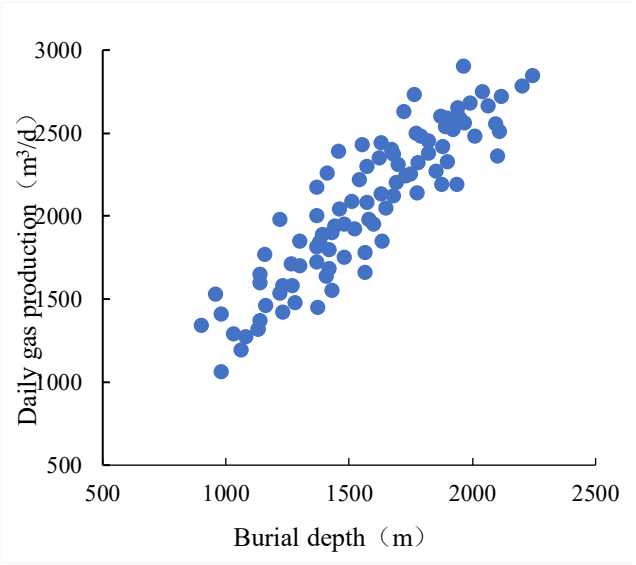

Figure 8 Cross diagram of burial depth and daily gas production

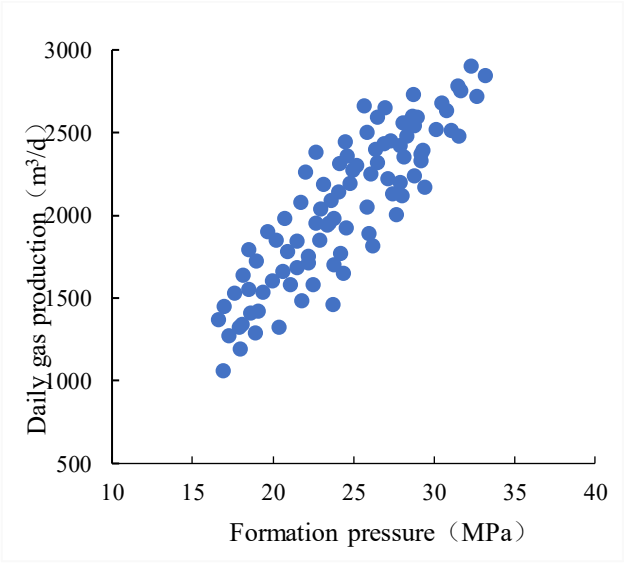

Figure 9 Cross diagram of formation pressure and daily gas production

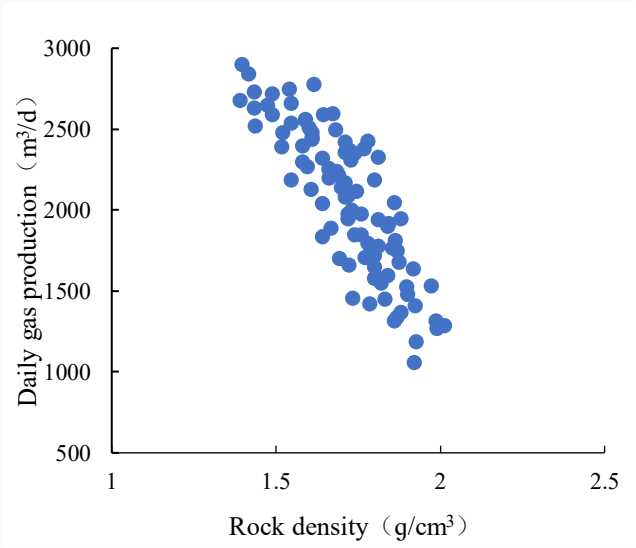

Figure 10 Cross diagram of rock density and daily gas production

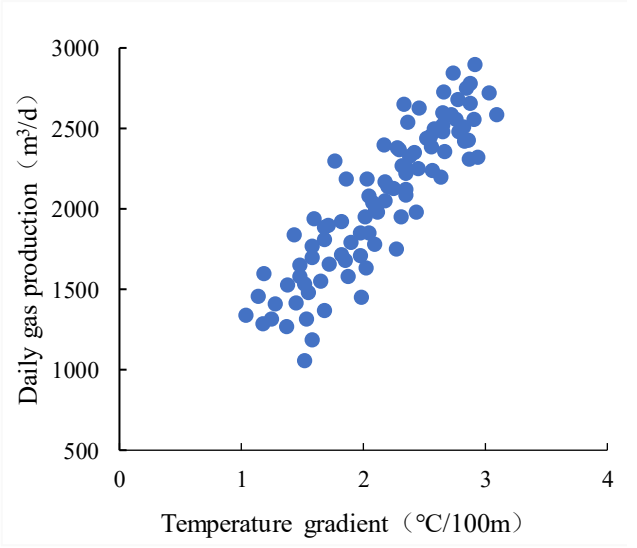

Figure 11 Intersection diagram of temperature gradient and daily gas production

3. Data source

The data of this study are derived from on-site exploration and production practices in the study area. A total of 100 sets of effective samples were collected, and 10 input features (permeability K, gas saturation Sg, reservoir pressure gradient P_grad, etc.) and 1 target variable (daily gas production Q) were finally extracted. The data range is in line with the actual geological background of the study area, as shown in the table below.

Table 1 Input features and target variable parameter range

Parameter name	Symbol	Unit	Data range	Physical meaning
Penetration	K	$\times 10^{-3}\mu m^2$	0.1-5.0	Gas flow capacity
Gas saturation	Sg	%	60-92	Reservoir gas proportion

ISSN: 1813-4890

Reservoir pressure gradient	P_grad	MPa/100m	0.15-0.35	Gas seepage dynamics
Coal seam thickness	H	m	1.2-8.5	Storage space size
Porosity	φ	%	3.5-12.0	Reservoir pore proportion
Gas content	Q_g	m^3/t	1.0-8.0	Gas potential per unit coal
Burial depth	D	m	800-2500	Affects pressure and gas content
Formation pressure	P	MPa	10.0-35.0	Original driving pressure
Rock density	ρ	g/cm^3	1.3-1.8	Coal rock density
Temperature gradient	T_grad	°C/100m	1.8-3.2	Affect gas adsorption equilibrium
Daily gas production	Q	m^3/d	500-3000	Actual production capacity of coalbed methane wells

In order to improve the robustness of the model, a three-step preprocessing process is adopted: (1) Outlier elimination: Based on the 3σ criterion, samples that deviate 3 times the standard deviation from the mean in each parameter are eliminated, and 100 sets of valid data are finally retained; (2) Data standardization: The mapminmax function is used to normalize the data to the [0,1] interval to eliminate dimensional differences. The formula is:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Among them, x is the original data, xmin and xmax are the minimum and maximum values of the parameters respectively; (3) Data set division: randomly divide the training set (70 groups) and the test set (30 groups) according to the ratio of 7:3, and fix the random seed (rng=2024) to ensure that the results are reproducible.

4. Research methods

Through gray correlation analysis, 10 main production capacity control factors such as permeability and gas saturation were selected; three machine learning prediction models of BP neural network, gradient boosting tree (GBDT) and random forest (RF) were constructed, and four indicators of MAE, MAPE, RMSE and R^2 were used to compare the model performance. The results show that the model can effectively adapt to the basin's small sample and strong noise data characteristics, and can provide technical support for coalbed methane development well location deployment and drainage plan optimization [6,7].

4.1. Selection of main control factors: gray correlation analysis

Gray correlation analysis is suitable for factor correlation calculation in small samples and poor information systems. By comparing the similarity between the reference sequence (capacity Q) and the comparison sequence (geological parameters), the main controlling factors can be determined. The specific steps are as follows:

ISSN: 1813-4890

Determine the sequence: reference sequence $X_0 = (x_0(1), x_0(2), \ldots x_0(n))$ (Daily gas production), compare sequences $X_i = (x_i(1), x_i(2), \ldots x_i(n))$ (10 geological parameters), n=100 is the number of samples;

Dimensionless: Use the averaging method to process the sequence to eliminate the impact of magnitude;

Calculate the correlation coefficient:

$$\gamma(x_i(k), x_i(k)) = \frac{\min_i \min_k |x_0(k) - x_i(k)| + \rho \max_i \max_k |x_0(k) - x_i(k)|}{|x_0(k) - x_i(k)| + \rho \max_i \max_k |x_0(k) - x_i(k)|} \quad (2)$$

Among them, ρ=0.5 is the resolution coefficient, k=1,2,...,n is the sample serial number;

Calculate the correlation degree:

$$\gamma_i = \frac{1}{n} \sum_{k=1}^{n} \gamma(x_0(k), x_i(k)) \quad (3)$$

Correlation ≥0.8 is regarded as the main control factor.

The calculation results show that the correlations of the 10 parameters are all ≥0.75. Among them, permeability, gas saturation, and reservoir pressure gradient have the highest correlation, which are the core main controlling factors and are consistent with geological understanding - permeability determines gas flow efficiency, gas saturation reflects resource potential, and pressure gradient provides seepage power.

4.2. Machine learning prediction model building

4.2.1. BP neural network

BP neural network is a multi-layer feedforward network that adjusts weights through error backpropagation and is suitable for nonlinear mapping. Model structure design: 10 nodes in the input layer (corresponding to 10 main control factors), 21 nodes in the hidden layer (empirical formula: 2×input dimension + 1), 1 node in the output layer (daily gas production); activation function: tansig (hyperbolic tangent) is used in the hidden layer, and purelin (linear) is used in the output layer; training parameters: maximum number of iterations 200, learning rate 0.01, training function trainlm (Levenberg-Marquardt algorithm), target error 1e-5.

4.2.2. Gradient Boosted Tree (GBDT)

GBDT is an ensemble learning algorithm that constructs weak decision trees through iteration and weighted fusion. It has strong anti-noise ability and is suitable for small sample data. Parameter settings: The number of weak learners is 100, the learning rate is 0.1 (to control the contribution of a single tree), the minimum number of samples for leaf nodes is 3 (to avoid overfitting), the loss function uses mean square error (MSE), and the integration method is least squares lifting.

4.2.3. Random Forest (RF)

RF reduces the risk of over-fitting by building multiple decision trees and voting for output. Parameter settings: The number of decision trees is 100, the node splitting criterion is the mean square error, the number of random feature selections is 3 (rounded to 10), and the minimum number of samples of leaf nodes is 2.

4.3. Model evaluation index

After the model is successfully established, four common indicators are used to quantify the model performance:

(1) Mean absolute error (MAE): reflects the mean absolute value of the prediction deviation. The smaller the value, the better:

$$MAE = \frac{1}{m} \sum_{j=1}^{m} |y_j - \hat{y}_j| \quad (4)$$

(2) Average relative error (MAPE): reflects the relative deviation, the industry allowable threshold is 10%:

ISSN: 1813-4890

$$MAPE=\frac{1}{m}\sum_{j=1}^{m}\left|\frac{y_j-\hat{y}_j}{y_j}\right|\times 100\% \quad (5)$$

(3) Root mean square error (RMSE): amplifies extreme errors and reflects stability:

$$RMSE=\sqrt{\frac{1}{m}\sum_{j=1}^{m}(y_j-\hat{y}_j)^2} \quad (6)$$

(4) Coefficient of determination (R^2): reflects the explanatory power of the model. $R^2 \geq 0.8$ is considered excellent:

$$R^2=1-\frac{\sum_{j=1}^{m}(y_j-\hat{y}_j)^2}{\sum_{j=1}^{m}(y_j-\bar{y}_j)^2} \quad (7)$$

Among them, m=30 is the number of test set samples, yj is the actual production capacity, y◎○j is the predicted production capacity, and y⓪j is the average actual production capacity.

5. Experimental results and analysis

5.1. Model performance comparison

The performance indicators of the three models on the training set and test set are shown in Table 2. On the training set, all three models performed excellently, with $R^2 \geq 0.93$, indicating that the models fully fit the training data;On the test set, the GBDT model has the best performance, MAPE=1.14%、R^2=0.8827，Compared with BP neural network(MAPE=1.10%，R^2=0.8621), the value of MAPE BP is slightly lower than GBDT, but the difference is only 0.04%, which is negligible, while R^2BP is 2.06% lower than GBDT. and RF (MAPE=1.26%，R^2=0.8498) are reduced by 0.12% and 3.29% respectively, and the RMSE is the smallest (27.27m^3/d），indicating that its generalization ability and stability are better [8-9-10].

Table 2 Comparison of performance indicators of three models

Model	Dataset	MAE(m3/d)	MAPE(%)	RMSE(m3/d)	R2
BP neural network	Training set	13.39	0.63	16.03	0.8784
	Test set	22.63	1.10	28.34	0.8621
Gradient Boosted Tree (GBDT)	Training set	3.36	0.16	11.16	0.8692
	Test set	21.85	1.14	27.27	0.8827
Random Forest (RF)	Training set	10.07	0.52	14.97	0.8186
	Test set	23.98	1.26	32.18	0.8498

5.2. Visual result analysis

Based on the analysis of the above research results, Figure 12 visually displays the mean absolute percentage error (MAPE) of the three models on the test set. It can be seen from the chart that their values are relatively BP1.10% → GBDT1.14% → RF1.26%. The height difference between the three columns is only 0.04% to 0.16%, which is almost the same to the naked eye.BP narrowly wins in the "average percentage error", but it is only 0.04% lower than GBDT, and the advantage is so weak that it can be ignored; GBDT comes second, and RF is slightly

ISSN: 1813-4890

higher. The lower the MAPE value, the higher the prediction accuracy. However, the values of these three models are not very obvious and need to be combined with other parameters.

Figure 13 shows the coefficient of determination (R^2) of the three models. As can be seen from the chart, GBDT0.8827→BP0.8621→RF0.8498, the R^2 of all models exceeds 0.8, indicating that these three machine learning methods can effectively predict coalbed methane production capacity in the Ordos Basin.GBDT is 0.0206 higher than BP and 0.0329 higher than RF; its difference can be perceived by the naked eye in the regression evaluation. The closer the R^2 value is to 1, the stronger the model's ability to explain variable variation. Therefore, GBDT has the strongest "explanatory power" for test samples; BP is almost tied; RF is obviously lagging behind.

Figure 14 is a GBDT prediction vs. actual scatter plot. This scatter plot shows the relationship between the predicted value and the actual value of the GBDT model. The data points are closely distributed near the 45-degree diagonal, and there is no systematic deviation in the high and low production areas at both ends, indicating that the model can maintain good prediction performance in different production capacity ranges.The high R^2=0.8827 value indicates that the predicted value is highly correlated with the actual value, indicating that GBDT maintains high linearity throughout the entire production range, with no obvious overestimation/underestimation platform. This consistency indicates that the model has good generalization ability for predicting coalbed methane production capacity in the Ordos Basin.

Figure 15 is a comparison of the prediction curves of the three models. This line chart compares the prediction results of the three models with the actual values. Overall, the three prediction lines are close to the actual black point, but when zoomed in, BP and RF have a slight downward bias in the high productivity section (>2400m^3/d); GBDT is still close to the actual value.In the low productivity section (<1600m^3/d), RF has the largest dispersion, followed by BP, and GBDT is the most stable. Although the prediction trends of the three models are basically the same, the GBDT model shows the best fitting effect in each production capacity range, so GBDT has the best "shape-preserving" ability for extreme production capacities; BP is slightly better than RF.

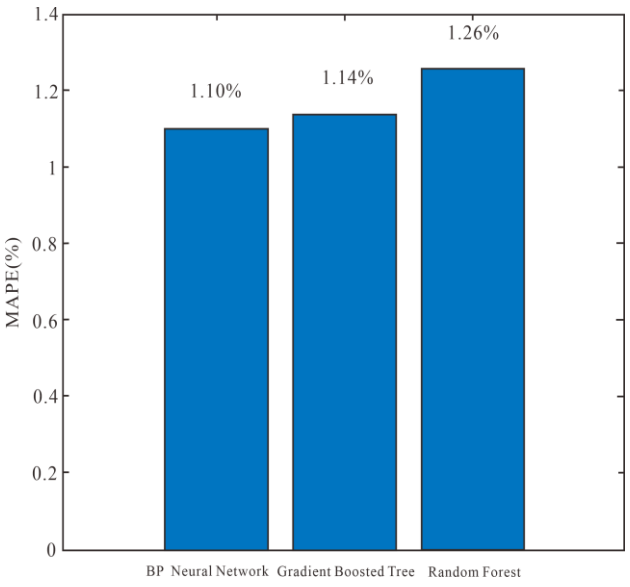

Figure 12 Test set MAPE comparison

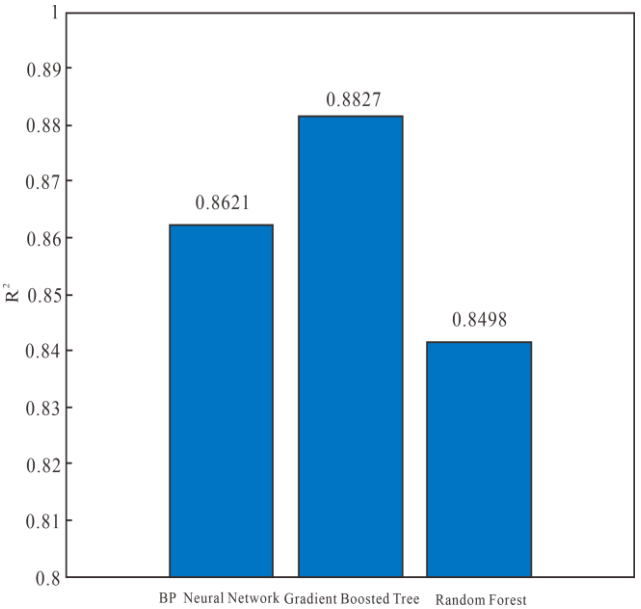

Figure 13 Test set R^2 comparison

ISSN: 1813-4890

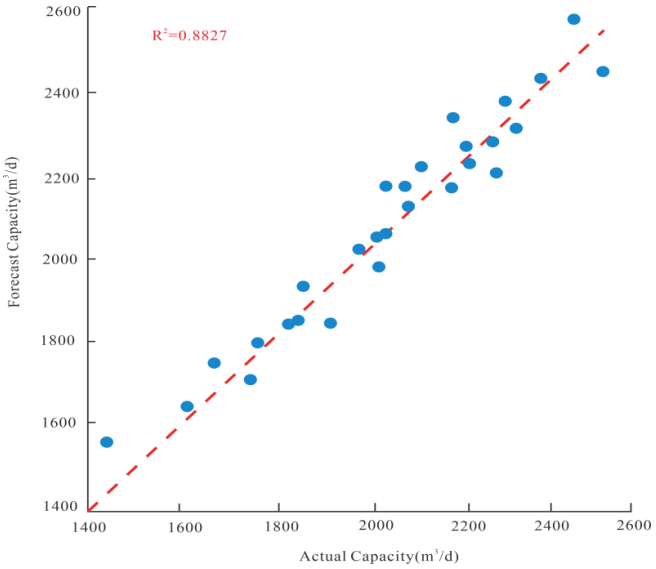

Figure 14 GBDT model prediction vs actual value

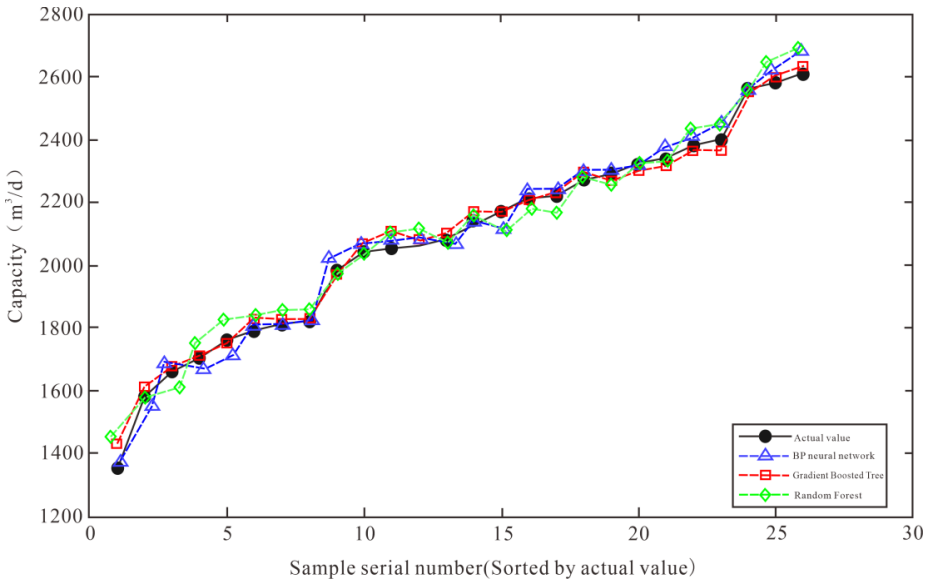

Figure 15 Comparison of prediction results of three models

Figure 16 shows the importance of GBDT features, showing the importance of each feature to GBDT model prediction. It can be seen from the figure that permeability (K) is the most important feature and contributes the most to productivity prediction (33.8%). The permeability leads by a cliff, which is completely consistent with the on-site understanding of "low permeability reservoirs, permeability control production";Gas saturation and reservoir pressure gradient are ranked second and third, with importance ranging from 150 to 200. They belong to the second echelon and are extremely interpretable. The feature importance analysis results are consistent with geological understanding. Permeability, gas saturation and pressure gradient are the main controlling factors of productivity. The importance of other parameters is relatively low, the parameters are all lower than 100, and their contribution to production capacity is relatively limited.Therefore, the GBDT model clearly points out that the main controlling factor of coalbed methane production capacity in the Ordos Basin is permeability, and its importance is significantly higher than other geological parameters. This result verifies the physical meaning of the model and enhances the credibility of the prediction results. The

ISSN: 1813-4890

model not only has excellent numerical values, but also has strong geological interpretability, which is conducive to subsequent optimization of well locations and fracturing plans.

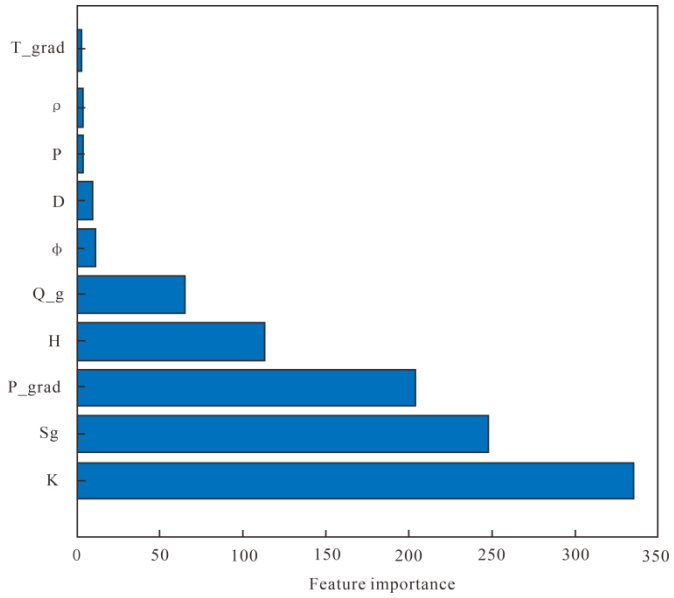

Figure 16 GBDT model feature importance

Combined with the above analysis, it can be concluded that the GBDT model can effectively handle the data characteristics of "low porosity, low permeability, small samples, and strong noise" in the Ordos Basin. The model prediction accuracy is much higher than that of traditional methods, and can provide a scientific basis for well location optimization and the formulation of drainage systems.

6. Conclusion and recommendations

Through gray correlation analysis, it was determined that permeability, gas saturation, and reservoir pressure gradient are the core main controlling factors of coalbed methane productivity in the Ordos Basin, providing geological basis for model input feature selection.

Among the three machine learning models constructed, BP neural network, GBDT, and RF, gradient boosting tree (GBDT) has the best performance, with test set MAPE=1.14% and R^2=0.8827, which can meet engineering accuracy requirements.

The GBDT model is adapted to the data characteristics of "low porosity, low permeability, small samples, and strong noise" in the Ordos Basin, and its prediction results can provide scientific support for the selection of coalbed methane development well locations and the formulation of drainage systems. The GBDT model not only has high prediction accuracy, but also provides feature importance ranking, which enhances the geological interpretability of the results and provides technical support for efficient development of coalbed methane.

References

[1] Li Yong, Gao Shuang, Wu Peng, et al. Evaluation and correction of deep coalbed methane free gas content prediction model - taking the deep coal seam in the eastern margin of Ordos Basin as an example [J]. Acta Petroleum Sinica, 2023, 44(11): 1892-1902.

[2] Yao Huifang, Zhao Mingkun, Chen Qiang. Research on classification of coal-measure tight sandstone gas reservoirs based on machine learning—taking DJ block in Ordos Basin as an example [J]. Coal Science and Technology, 2022, 50(06): 260-270.

ISSN: 1813-4890

[3] Chen Jiahao. Intelligent classification method and application of well logging productivity in tight sandstone reservoirs[D]. China University of Petroleum (Beijing), 2023.

[4] Yang Fan, Li Bin, Wang Kunjian, et al. Large-scale extreme volume fracturing technology for deep coalbed methane horizontal wells—taking the Linxing block on the eastern edge of the Ordos Basin as an example [J]. Petroleum Exploration and Development, 2024, 51(02): 389-398.

[5] Ren Jie. Research on water-based drilling fluid technology for marine-terrestrial transitional phase shale gas in Daji Block, Ordos Basin [D]. China University of Petroleum (Beijing), 2023.

[6] Liu Mengjie, Fei Shixiang, Zhou Xinyu, et al. Evaluation and prediction of tight gas horizontal well productivity based on gray correlation method [C]//Xi'an Petroleum University, Shaanxi Petroleum Society. Proceedings of the 2024 International Conference on Oil and Gas Field Exploration and Development II. PetroChina Changqing Oilfield Branch Exploration and Development Research Institute; Northwest University; PetroChina Changqing Oilfield Branch Eighth Oil Production Plant;, 2024: 205-212.

[7] Long Zhangliang, Wen Zhentao, Li Hui, et al. A shale reservoir compressibility evaluation method based on gray correlation analysis [J]. Oil and Gas Reservoir Evaluation and Development, 2020, 10(01): 37-42.

[8] Jin Yi, Zheng Chenhui, Song Huibo, et al. Research on coalbed methane production capacity prediction based on neural network model [J]. Journal of Henan University of Science and Technology (Natural Science Edition), 2025, 44(01): 46-56.

[9] Zhu Likai. Research on coalbed methane production capacity prediction and drainage system optimization based on deep learning [D]. China University of Petroleum (Beijing), 2022.

[10] Lu Yin. Coalbed methane production capacity prediction and analysis based on machine learning[D]. Southwest Petroleum University, 2020.