Shale gas well productivity potential evaluation based on data-driven methods: case study in the WY block

Chaodong Tan1,2 · Hanwen Deng1,2 · Wenrong Song3 · Huizhao Niu3 · Chunqiu Wang1,2

Received: 12 September 2021 / Accepted: 3 January 2022 / Published online: 16 January 2022
© The Author(s) 2022

Abstract
Evaluating the productivity potential of shale gas well before fracturing reformation is imperative due to the complex fracturing mechanism and high operation investment. However, conventional single-factor analysis method has been unable to meet the demand of productivity potential evaluation due to the numerous and intricate influencing factors. In this paper, a data-driven-based approach is proposed based on the data of 282 shale gas wells in WY block. LightGBM is used to conduct feature ranking, K-means is utilized to classify wells and evaluate gas productivity according to geological features and fracturing operating parameters, and production optimization is realized through random forest. The experimental results show that shale gas productivity potential is basically determined by geological condition for the total influence weights of geologic properties take the proportion of 0.64 and that of engineering attributes is 0.36. The difference between each category of well is more obvious when the cluster number of well is four. Meanwhile, those low production wells with good geological conditions but unreasonable fracturing schemes have the greatest optimization space. The model constructed in this paper can classify shale gas wells according to their productivity differences, help providing suggestions for engineers on productivity evaluation and the design of fracturing operating parameters of shale gas well.

Keywords Shale gas · Productivity potential evaluation · LightGBM · K-means · Optimization

Abbreviations

\[ r(X,Y) \] Linear correlation coefficient
\[ \text{Cov}(X,Y) \] Covariance
\[ \sigma(X) \text{ and } \sigma(Y) \] Variances
\[ M \] Number of trees
\[ v_i \] Splitting feature
\[ T_j \] Corresponding empirical improvement in squared-error
\[ a(i) \] Mean distance between \( i \) and all other data points in the same cluster
\[ d(i,j) \] Distance between data points \( i \) and \( j \) in the cluster \( C_i \)
\[ b(i) \] Smallest mean distance of \( i \) to all points in any other cluster
\[ B_k \] Between group dispersion matrix
\[ W_k \] Within-cluster dispersion
\[ N \] Number of data points
\[ C_q \] Point set in cluster \( q \)
\[ c_q \] Center of cluster \( q \)
\[ c \] Center of matrix trace \( E \)
\[ n_q \] Number of points in cluster \( q \)
\[ b \] Individual bootstrap samples
\[ B \] Sum of trees
\[ t^*_h \] Individual learner
\[ y_i \] True value
\[ \hat{y}_i \] Predicted value

Introduction
The research area in this paper focused on Weiyuan (WY) shale gas block. WY block, which is located in Sichuan basin, is a representative block with the achievement of industrial production in the Lower Cambrian shale gas exploration in China (Cao et al. 2020). The whole block has
a monoclinic structure inclined to the southeast, and faults are not well developed in this block. The reservoir properties are generally high, with total organic carbon (TOC) value of 2.1–8.1% and average of 3.2%, porosity of 2.1–9.6% and average of 5.9%, brittle mineral content of 60–82% and average of 74%, gas content of 3.3–8.5 m³/t and average of 5.5 m³/t. The shale gas play is characterized by small thickness (1–7 m and an average of 3.8 m) of high-quality reservoir, large horizontal stress difference (7.7–18.7 MPa) (Ma et al. 2020; Liu et al. 2021).

Shale gas is an important unconventional natural gas resource with huge exploration potentials (Yao et al. 2019; Zhu et al. 2020). The success of the shale gas revolution in North America further highlights its resource potential and development prospects (Ma et al. 2018). Shale gas reservoirs are characterized by low porosity, low permeability and heterogeneity, and the flow law in the matrix of shale gas reservoirs no longer follows Darcy flow compared with conventional reservoirs (Yan et al. 2017). Due to the ultralow porosity and permeability, it requires the horizontal well drilling and the multi-stage hydraulic fracturing technology to produce the shale gas (Jiang et al. 2020). The productivity of shale gas wells is affected by the energy utilization rate of the fracturing fluid, the closure law of main fractures and induced fractures, etc. (Mohammadmoradi and Kantzas, 2018). The challenge in developing shale gas lies in the inherent problem of multi-parameter decision making, especially when substantial uncertainty may exist with underground conditions and major economic factors. Additionally, since the development cost of shale gas wells is generally higher than that of conventional oil and gas, especially the fracturing cost, it is important to develop wells with high production potential and design reasonable fracturing parameters to improve the economic benefits.

In petroleum geoscience, sweet spots are regarded as target regions within the best production potential (Hauge et al. 2017). Engineers have employed traditional approaches like core analysis, well logging and seismic test to identify the locations of sweet spots (Liu et al. 2019; Zhao et al. 2019). However, these methods are not persuasive, because there is no obvious equivalence between the measured parameters and the well’s productivity, and the fact that there is no uniform standard for the identification of the sweet spot location from one field to another makes conventional methods less applicable. Some scholars have also adopted data-driven methods, such as neural network, fuzzy logic and support vector machine, to conduct productivity evaluation. The usual practice is to use logging data, seismic test data and production data to predict geological parameters such as TOC that related to the location of sweet spots or perform other analyses (Tahmasebi et al. 2017; Qian et al. 2018; Kurison et al. 2019). But the above techniques remain uncertain as they still do not fully explain the relationship between geological parameters and gas productivity. Therefore, the evaluation of shale gas production potential should be established on the determination of the controlling factors of productivity. Furthermore, well optimization will be more reliable when based on the important factors influencing gas recovery.

With the rapid development of AI, machine learning methods are widely applied in reservoir flooding (Ma et al. 2019), heavy oil recovery (Lu et al. 2021), performance evaluation of shale gas (Syed et al. 2021) and other oil and gas fields. Compared with data-driven models, the mechanism-based reservoir simulation models for parameter sensitivity analysis are too time-consuming and computationally rigorous to conduct (Bettin et al. 2019; Mo et al. 2019). Selection of the principal parameters influencing gas production should consider both geologic and engineering parameters. Currently, the commonly used data-driven method to determine the main influencing factors of gas production is to establish a prediction model between the influencing factors and productivity parameters. He et al. (2019) established a prediction model of gas production in Marcellus shale gas area with 31 input parameters including well information, reservoir parameters, rock mechanics parameters, completion parameters, and fracturing parameters by using neural network algorithm, and optimized parameters such as perforation cluster number, fracturing stage and proppant volume through this model. Liu et al. (2021) applied deep learning method to predict estimated ultimate recovery (EUR) based on 17 parameters of the geological, engineering and production factors involved with the EUR. Liang et al. (2021) built a multi-objective random forest method to predict the shale gas production performance throughout the entire shale gas production process. However, there were no specific conclusions about the controlling factors of gas productivity. Vikara et al. (2020) used gradient boosted regression tree (GBRT)-based predictive model to estimate ultimate recovery and concluded that net thickness and porosity were the vital influencing factors. However, whether geologic attribute is the first principal controlling factor remains unclear. Meanwhile, it is still not explicit to know how important the engineering factors such as fracturing parameters are to the productivity. Therefore, it is necessary to rank the importance of the factors that affect the productivity of shale gas wells and determine the category of parameter attributes with the largest influence weight. If the weight of geological attribute ranks first, it is of great significance for shale gas wells to develop different fracturing optimization schemes according to different geological conditions to increase the output of single well and reduce production cost.

In view of the current research status, this study seeks to achieve the well productivity evaluation, dominating factors analysis and production optimization through machine learning methods. The model constructed in this paper can give
Data description and statistical analysis

The dataset obtained from the 282 fractured wells in the study area contains 12 different types of data items, including 7 geological features (vertical depth, TOC content, porosity, thickness of organic-rich gas reserves, gas saturation, formation pressure coefficient, brittle mineral content), 4 engineering features or fracturing operating features (average fracturing interval length, amount of fracturing fluid, amount of fracturing sand, average fracturing fluid displacement), and first-year daily gas production rate ($Q_g$) is taken as the evaluation index of gas productivity. Due to the limited field conditions, some parameters such as permeability cannot be obtained temporarily, but the method proposed in this paper is still feasible and can be extended to more data.

Table 1 is the descriptive statistics of all wells, serial number of geological features is 1–7, and that of fracturing operating features is 8–11, and No. 12 is the label of the dataset. Among them, geological factors are difficult to be changed artificially, while fracturing factors are more controllable to field engineers, and can be adjusted manually, both of which are related to $Q_g$. The productivity of these wells varies widely, with the maximum $Q_g$ value being more than 170 times of the minimum $Q_g$ value, and the variance of $Q_g$ is large, indicating that the productivity data is scattered. From a mechanical point of view, poor geological conditions or unreasonable fracturing parameters will lead to low fracturing productivity. Therefore, evaluating the productivity potential of shale gas wells prior to fracturing will greatly reduce the probability of low production caused by the failure of fracturing parameter design, thus improving the overall benefit.

This paper will calculate the influence weight of each feature through feature ranking, find the principal factors affecting productivity, evaluate productivity potential through feature reconstruction and clustering, and finally establish a random forest predictive model to optimize the comprehensive index of fracturing operating parameters of low production wells. A preliminary data analysis was conducted before the above work. The relationship between each feature and the response variable was independently measured by the univariate feature selection method, and the linear correlation coefficient $r$ between each parameter was calculated by Eq. (1). From Fig. 1, it can be seen that there is almost no strong linear relationship between geological and fracturing operation parameters with $Q_g$, only vertical depth is positively related to formation pressure coefficient. The correlation coefficients of X1 and X2 with X12 are all 0.5, indicating thickness of organic-rich gas reserves and gas saturation are moderately correlated with $Q_g$. Therefore, conventional single-factor analysis method has been unable to meet the demand of productivity potential evaluation due to the numerous and intricate influencing factors. Data-driven methods provide the possibility to solve the problem. And the scattered correlation matrix of these features is shown in Fig. 2.

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\sigma(X) \cdot \sigma(Y)}}$$

| No | Dataset                        | Unit       | Standard deviation | Average value | Minimum value | Maximum value |
|----|--------------------------------|------------|--------------------|---------------|---------------|---------------|
| 1  | Vertical depth                 | m          | 343                | 3015          | 2200          | 3800          |
| 2  | TOC content                    | %          | 0.75               | 5.38          | 3.30          | 7.80          |
| 3  | Porosity                       | %          | 0.94               | 7.15          | 5.70          | 8.90          |
| 4  | Thickness of organic-rich gas reserves | m | 1.37               | 5.19          | 2.30          | 7.60          |
| 5  | Gas saturation                 | %          | 3.17               | 75.79         | 60            | 83            |
| 6  | Formation pressure coefficient | a.u        | 0.18               | 1.74          | 1.35          | 2.05          |
| 7  | Brittle mineral content        | %          | 7.26               | 78.26         | 62.50         | 96            |
| 8  | Average hydraulic fracturing fluid displacement | m³/min | 1.30               | 11.82         | 6.19          | 15            |
| 9  | Average hydraulic fracturing interval length | m | 11.61              | 68.30         | 43.10         | 145.70        |
| 10 | Amount of hydraulic fracturing fluid | m³/m | 4.39               | 27.31         | 7.12          | 47.94         |
| 11 | Amount of hydraulic fracturing sand | t/m  | 0.39               | 1.64          | 0.29          | 3             |
| 12 | First-year daily gas production rate ($Q_g$) | $10^4$m³/d | 5.49               | 9.50          | 0.19          | 33.30         |

a.u. refers to arbitrary unit (dimensionless)
is the corresponding empirical improvement.

The greater the absolute value of \( r \), the stronger the correlation between two parameters, and vice versa. When \( r \) is equal to 0, it can only indicate that there is no linear correlation between the two parameters, but irrelevant.

**Methodology**

There are wide variations in gas productivity of these wells, in order to improve shale gas production, it is of great significance to further explore the reasons for the productivity differences among wells and find out the controlling factors that affect the productivity.

Specifically, the data from 282 shale gas wells will be used to: (1) calculate the importance weight of each influencing factors and conduct feature ranking by LightGBM. (2) Calculate comprehensive index of geological parameters and fracturing operating parameters through feature reconstruction, and combine new features with first-year daily gas production rate to form a new dataset. (3) Classify the new dataset so as to evaluate the gas productivity potential through K-means. (4) Using random forest predictive model to optimize wells with low \( Q_g \).

**Feature ranking by LightGBM**

The feature ranking in this paper is model-based, because some machine learning methods have a scoring mechanism for features. Therefore, LightGBM (Light gradient boosting machine learning) algorithm is selected to establish the prediction model between each feature and daily gas production, and its ranking mechanism is used to calculate the influence weight of geological features and fracturing parameters.

LightGBM is an improved variant of gradient boosting decision tree (GBDT) algorithm (Dj et al. 2020); it combines the predictions of multiple decision trees to make the final prediction generalizes well and contains two novel techniques: Gradient-based One-Side Sampling and Exclusive Feature Bundling to deal with large number of data samples and large number of features, respectively (Ke et al. 2017). LightGBM algorithm divides the training set into multiple models, and in each iteration, the selection of the first \( k \) features is determined by local voting, and then the selection of the first \( k \) features is determined by global voting. The default training decision tree in LightGBM algorithm uses gradient boosting algorithm, which does not need to find splits on the classified feature values, but stores continuous feature values in discrete bins which are used to construct feature histograms during the training process. This reduces the number of data instances and features, thus shortening training time (Qin et al. 2021).

LightGBM can be used for feature ranking since greedy feature selection already happens in the algorithm when selecting splitting features, for example, splitting features and splitting points of regression trees are found to minimize the squared-error loss for any given partition of the data. Moreover, as a byproduct, a sorted list of relative importance of features is automatically generated for each boosted trees model (Feng et al. 2009). Friedman (J.H, 2001) generalizes the global importance of feature \( x_j \) is measured by the average value of the importance of feature \( x_j \) in a single tree:

\[
\hat{T}_j^2 = \frac{1}{M} \sum_{m=1}^{M} \hat{T}_j^2(T_m)
\]  

(2)

\( M \) is the number of trees, and the relative importance of feature \( x_j \) in each tree is calculated as:

\[
\hat{T}_j^2(T) = \sum_{i=1}^{L-1} \hat{T}_j^2(v_i = j)
\]  

(3)

where the summation is over the internal nodes \( t \) of a \( L \)-terminal node tree \( T \), \( v_i \) is the splitting feature associated with node \( t \), and \( \hat{T}_j^2(T) \) is the corresponding empirical improvement in squared-error as a result of the split.

Through this method, the influence weight of each feature, the total influence weight of geological features and
fracturing operating features will be obtained. Thus, features have greater impact on gas production will be known.

**Feature reconstruction**

In order to facilitate the classification of shale gas wells, it is necessary to reduce the dimensionality of the original feature parameters, and feature reconstruction is used to obtain the comprehensive index of geologic features and fracturing operating features, respectively.

The commonly used method of feature dimension reduction is feature extraction. Feature extraction refers to the use of some specific feature extraction methods to transform the original data, such as PCA (Principal Component Analysis), LDA (Linear Discriminant Analysis), etc. (Martinez et al. 2001; Shlens, 2014). The dimension of the data is reduced after feature extraction. Although the new feature has statistical significance and retains the main information of the original data, the physical meaning is not clear. Feature construction refers to artificially...
constructing new features from original data, such as using mixed attributes or combined attributes to create new features, or decomposing and segmenting original features to create new features. Therefore, this paper will adopt feature construction, combined with domain knowledge to reconstruct new features to achieve feature reduction. The detailed calculation process is as follows.

A data set with \( N \) samples, \( M \) features and \( F \) attributes can be set as \( X_{NM} \). Suppose \( m_i \) is the number of features belonging to the \( i \) th attribute, thus \( M = \sum_i m_i \), \( X_{NM} = [x_{N1m_1}, x_{N2m_2}, \ldots, x_{NM_m}] \). Set the importance weight matrix as \( \mathbf{W} = [w_{1x_{m_1}}, w_{1x_{m_2}}, \ldots, w_{x_{m_1}m_1}]^T \), where \( w_{yym} = [y_1, y_2, \ldots, y_m] \) is the feature weight, \( \sum_i m_i = 1 \), and the new feature matrix can be represented as: \( C_NF = [c_1, c_2, \ldots, c_F] \), where:

\[
c_i = \frac{x_{N1m_1} \times w_{1x_{m_1}}}{\sum_i m_i y_j}
\]

In this study, there are 282 samples, 7 geological features and 4 fracturing operating features, thus the comprehensive index of geologic attribute \( c_1 = \frac{2x_{N1m_1}}{\sum_j y_j} \), and the comprehensive index of fracturing operating attributes \( c_2 = \frac{2x_{N2m_1}}{\sum_j y_j} \). Therefore, \( c_1, c_2 \) and \( Q_F \) will form a new data matrix of 282 rows and 3 columns.

**Productivity potential evaluation by K-means**

After constructing the new data set, the wells will be classified by clustering algorithm, and the productivity potential will be evaluated, so as to select those low productivity wells with good geological conditions but unreasonable fracturing parameters design to facilitate subsequent productivity optimization.

K-means algorithm is selected for well classification as it is one of the most commonly used and important clustering algorithms, the basic idea is to find the partition scheme of \( K \) clusters through iteration, and use the mean value of these \( K \) clusters to represent the corresponding sample types, so as to minimize the total error (Chu X et al. 2020). Suppose \( \mu_1, \mu_2, \ldots, \mu_k \) are \( K \) cluster centers, \( r_{nk} \in \{0, 1\} \) is used to indicate whether \( x_n \) belongs to cluster \( K \), the loss function is defined as:

\[
J(\mu, r) = \sum_{i=1}^K \sum_{k=1}^K r_{nk} \left| x_n - \mu_k \right|^2
\]  

(5)

The cost function of above formula can only be minimized by iterative methods. The K-means algorithm clusters samples into \( K \) clusters, where \( k \) is given by the user. The solution process of the algorithm is described as follows:

1. Input \( N \) samples \( \{x_1, x_2, \ldots, x_N\}, x_n \in \mathbb{R}^D \), and input the proposed number of clustering \( k \).
2. Randomly initialize \( K \) vectors of \( D \) dimension, and select \( K \) different sample points as the initial clustering center.
3. Repeat the following process until convergence:

   For each object \( x_n \), calculate the class it should belong to, as in the following formula:

   \[
c_k = \{ n : k = \arg\min_i \left| x_n - \mu_k \right|^2 \}
\]

   Recalculate the centroid of the class:

   \[
   \mu_k = \frac{1}{\left| c_k \right|} \sum_{m \in c_k} x_n
   \]

(6)

The iterative convergence definition is that the clustering center remains unchanged or the sum of distances from each sample to the corresponding clustering center no longer has a great change. Since the ground truth labels are not known, two internal evaluation methods, Silhouette coefficient \( (SC) \) and Calinski–Harabasz \( (CH) \) index, will be adopted as the evaluation indexes of clustering performance.

\( SC \) was first proposed by Peter (1999); it is based on the comparison of its tightness and separation, and can be used to evaluate the influence of different algorithms or different operation modes on clustering results based on the same original data. \( SC \) ranges from −1 to 1, the closer the value is to 1, the more reasonable the clustering result is. If the value is close to −1, it means that the sample should be classified into another cluster. And the closer the value is to −1, the more reasonable the sample clustering should be classified into another cluster. \( SC \) for a set of samples is given as the mean of the \( SC \) for each sample.

The \( SC \) of object \( i \) is:

\[
SC(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}
\]

(7)

\[
a(i) = \frac{1}{\left| C_i \right|} \sum_{j \in C_i, j \neq i} d(i, j)
\]

\[
b(i) = \min_{k \neq i} \frac{1}{\left| C_k \right|} \sum_{j \in C_k} d(i, j)
\]

\( a(i) \) is the mean distance between \( i \) and all other data points in the same cluster, \( d(i, j) \) is the distance between data points \( i \) and \( j \) in the cluster \( C_i \), \( b(i) \) is the smallest mean distance of \( i \) to all points in any other cluster, of which \( i \) is not a member.

The \( CH \) index, also known as the variance ratio criterion—can be used to evaluate the clustering model, where
a higher $CH$ score relates to a model with better defined clusters.

$$CH(k) = \frac{T_r(B_k)}{T_r(W_k)} \times \frac{N-k}{k-1}$$ \hspace{1cm} (10)

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$ \hspace{1cm} (11)

$$B_k = \sum_{q} n_q(c_q - c)(c_q - c)^T$$ \hspace{1cm} (12)

$B_k$ is between group dispersion matrix, $W_k$ is within-cluster dispersion, $N$ is the number of data points, $C_q$ is the point set in cluster $q$, $c_q$ is the center of cluster $q$, $c$ is the center of matrix trace $E$, and $n_q$ is the number of points in cluster $q$.

### Production optimization through random forest

Through the previous steps, wells with good geological conditions but low productivity will be selected for optimization, and random forest (RF) will be used to establish a regression model between new characteristic parameters and $Q_e$. When optimizing the gas productivity of a single well, fix the geological index of the well, change the fracturing operating index and iterate through the trained predictive model to search for the corresponding fracturing operating index when the output is maximum, so as to achieve productivity optimization.

RF is an ensemble of CART (Classification and regression tree) and uses a bootstrap method based on the classification and regression tree analysis to predict a continuous response variable (Di Fina et al. 2015; Ftpa B et al. 2021). Regression in RF is generated as an ensemble estimate from a number of decision trees based on bagging; the final output is the average of outputs of individual trees (Prasad et al. 2006; Adusumilli et al. 2013)

$$\hat{y}_B(x) = \frac{1}{B} \sum_{b=1}^{B} t_b^*(x)$$ \hspace{1cm} (13)

$b$ is the individual bootstrap samples, $B$ is the sum of trees, $t_b^*$ is the individual learner, and the individual decision tree:

$$t_b^*(x) = t(x; z_{b,1}, z_{b,2}, \ldots, z_{b,K})$$ \hspace{1cm} (14)

$z_{b,k}(k=1,2,\ldots,K)$ is the $k$th training sample with pairs of values for the target variable ($y$) and covariates ($x$): $z_{b,k} = (x_k, y_k)$.

### Results and discussion

LightGBM algorithm is selected as the feature selection algorithm to determine the importance weight of each feature, and normalization processing was conducted to eliminate any differences in the order of magnitude among the various input parameters. In the training process, 80% of the data set is divided into training set and 20% is divided into test set, and the grid search method is used to adjust the hyperparameters of the LightGBM predictive model, and the hyperparameters are shown in Table 2. In this paper, Mean square error (MSE) is used as the evaluation index of the regression model, it is calculated by Eq. (15), where $n$ represents the number of shale gas wells, $\hat{y}_i$ is true value, $y_i$ is predicted value, the smaller the value of MSE, the higher the accuracy of the prediction model. MSE value of training set is 0.006 and that of test set is 0.01, which is small enough to make the model feasible.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$ \hspace{1cm} (15)

The influence weight of each feature is shown in Fig. 3. It can be seen that thickness of organic-rich gas reserves and amount of hydraulic fracturing sand are vital parameters affecting gas productivity. The total weight of geological features is 0.64 and that of fracturing operating features is 0.36, indicating that geological condition is more deterministic to gas productivity. The result also proves that gas production of well in poor reservoir condition is difficult to be improved by adjusting engineering parameters.

New sample matrix $C_{282 \times 3}$ is constructed according to Eq. (4); the new data statistics is shown in Table 3. Through feature transformation, 11 features were compressed to 2-dimension according to 2 feature attributes, and a three-dimensional dataset, including comprehensive index of

| Hyperparameter       | Type/Value |
|----------------------|------------|
| boosting_type        | gbdt       |
| objective            | Regression |
| num_leaves           | 20         |
| max_depth            | 6          |
| learning_rate        | 0.09       |
| feature_depth        | 0.8        |
| bagging_fraction     | 0.8        |
| bagging_freq         | 1          |
| bagging_fraction     | 0.8        |
| bagging_seed         | 1          |
| lambda_11            | 0.001      |
| lambda_12            | 1          |
| verbosity             | -1         |
| n_estimators          | 50         |
geological features \((c_1)\), comprehensive index of fracturing operating features \((c_2)\) and \(Q_g\) was formed for further analysis.

Due to the large difference between the values of \(Q_g\), \(c_1\) and \(c_2\), data preprocessing is required before classification. This article uses the method of MinMaxScaler to compress these values to between 0 and 1. Then, clustering will be performed by K-means after the normalization of new dataset. In order to obtain the best cluster number, sensitivity analysis of cluster number \(k\) is carried out by using two internal evaluation methods: \(SC\) and \(CH\). From Fig. 4 and Fig. 5, it can be seen that when \(k\) is 2, the clustering effect is the best, while when \(k\) is 3 or 4, the clustering effect is also acceptable. In order to evaluate the productivity potential of the well more reasonably, the cluster diagrams when \(k\) is 2, 3, and 4 are drawn, as shown in Fig. 6. When the cluster number is 4, the difference between various types of wells is more obvious, which is more conducive to subsequent analysis. Therefore, this article divides the 282 shale gas wells in the study block into 4 categories, detailed feature description and data statistics of each cluster are shown in Table 4. The values of \(Q_g\) and \(c_1\) values in cluster 2 are very small, which demonstrate that this type of wells has the poorest geological conditions, low productivity potential, and low economic development benefits. The values of \(c_1\) in cluster 1 and cluster 3 is relatively large, indicating that these two types of

#### Table 3 Statistics of the reconstructed dataset

| Feature                                | Unit       | Standard deviation | Average value | Minimum value | Maximum value |
|----------------------------------------|------------|--------------------|---------------|---------------|---------------|
| \(Q_g\)                                | \(10^4\)m\(^3/d\) | 5.49               | 9.50          | 0.19          | 33.30         |
| Comprehensive index of geologic features \((c_1)\) | a.u       | 31.73              | 310.96        | 234.31        | 384.57        |
| Comprehensive index of fracturing operating features \((c_2)\) | a.u       | 2.11               | 24.85         | 19.37         | 38.42         |

Fig. 3 Histogram of importance weight of features

Fig. 4 Change curve of \(SC\) with cluster number \(k\)

Fig. 5 Change curve of \(CH\) index with cluster number \(k\)
Fig. 6 Clustering graph of normalized $Q_g$, comprehensive index of geologic features ($c_1$) and fracturing operation features ($c_2$)
wells have superior geological conditions, while the $Q_g$ value of cluster 1 is low, which proves that the fracturing operation parameter design of this type of wells is unreasonable and has large optimization space. If secondary fracturing is required, such kind of wells should be considered first. The fracturing parameter design of wells in cluster 3 is reasonable and has important significance. Table 5 summarizes the parameter range of wells in cluster 3, and the design of fracturing operating parameters can refer to the wells in cluster 3. Conclusively, the geological conditions of the well should be precisely evaluated to determine the productivity potential of the well before fracturing. For wells with good geological conditions, the investment can be increased as designing reasonable fracturing parameters for hydraulic fracturing will bring greater economic benefits, and for wells with poor geological conditions, the investment cost can be relatively reduced. By considering different reservoir characteristics, the fracturing scheme is customized to make it more specific and applicable.

After completing the above work, six wells are selected from cluster 1 for productivity optimization, because these wells have large optimization space. Taking $c_2$ as the optimization variable, the objective function is to maximize $Q_g$ by changing the value of $c_2$, and RF algorithm is used to establish a predictive model between $c_1$, $c_2$ and $Q_g$ of 282 wells. The ratio of the training set to the test set of the predictive model is 7:3. When the accuracy of the predictive model reaches the standard, it starts to iterate. During the optimization process of each well, only the value of $c_2$ is adjusted while $c_1$ remained unchanged in every iteration, and the value of $c_2$ varies in the range of 0 and 1. The results are shown in Fig. 7, the red line in the figure represents the original $Q_g$, and the blue dotted curve is the change trend of the optimized $Q_g$ with the number of

| Cluster | Number of wells | Normalized feature | Average value | Minimum value | Maximum value | Description                                      |
|---------|-----------------|--------------------|---------------|---------------|---------------|-------------------------------------------------|
| 1       | 77              | $Q_g$              | 0.24          | 0.024         | 0.45          | Good geological condition, poor fracturing measures and low productivity |
|         |                 | $c_1$             | 0.79          | 0.64          | 0.95          |                                                 |
|         |                 | $c_2$             | 0.41          | 0             | 0.70          |                                                 |
| 2       | 67              | $Q_g$              | 0.19          | 0.07          | 0.41          | Poor geological condition and the productivity is low regardless of the fracturing measures |
|         |                 | $c_1$             | 0.19          | 0             | 0.38          |                                                 |
|         |                 | $c_2$             | 0.50          | 0.31          | 0.71          |                                                 |
| 3       | 31              | $Q_g$              | 0.61          | 0.37          | 1             | Superior geological condition, good fracturing measures and high productivity |
|         |                 | $c_1$             | 0.83          | 0.60          | 1             |                                                 |
|         |                 | $c_2$             | 0.61          | 0.29          | 1             |                                                 |
| 4       | 107             | $Q_g$              | 0.27          | 0             | 0.58          | Medium geological condition, medium productivity |
|         |                 | $c_1$             | 0.49          | 0.30          | 0.64          |                                                 |
|         |                 | $c_2$             | 0.43          | 0.11          | 0.75          |                                                 |

| Feature                                         | Unit   | Mean value | Minimum value | Maximum value |
|-------------------------------------------------|--------|------------|---------------|---------------|
| Vertical depth                                   | m      | 2991.94    | 2600.00       | 3600.00       |
| TOC content                                      | %      | 5.41       | 4.85          | 6.40          |
| Porosity                                        | %      | 7.47       | 6.40          | 8.70          |
| Thickness of organic-rich gas reserves           | m      | 6.83       | 5.20          | 7.60          |
| Gas saturation                                   | %      | 80.16      | 73.50         | 83.00         |
| Formation pressure coefficient                   | a.u    | 1.73       | 1.52          | 1.90          |
| Brittle mineral content                          | %      | 82.39      | 73.80         | 96.00         |
| Average fracturing fluid displacement            | m³/min | 12.37      | 10.60         | 15.00         |
| Average fracturing interval length               | m      | 71.05      | 44.95         | 94.00         |
| Amount of fracturing fluid                       | m³/m   | 28.59      | 23.58         | 47.94         |
| Amount of fracturing sand                        | t/m    | 2.10       | 0.92          | 3.00          |
| $Q_g$                                            | $10^4$ m³/d | 20.54 | 12.61 | 33.30 |

Table 5: Statistical distribution of parameters in cluster 3
It can be seen from the figure that each well has achieved the largest $Q_g$ value within the range of 70 iterations. The statistics of the parameters before and after optimization of the 6 wells are shown in Table 6. The experimental result shows the gas production rate of these six wells has been greatly improved, indicating that the previous design of fracturing operating parameters such as the amount of hydraulic fracturing fluid and the amount of hydraulic fracturing sand is not reasonable. These wells with greater productivity potential should be taken seriously as their gas production has not been fully exploited.

In addition, the data obtained in this study is limited, with only 7 geological parameters and 4 engineering parameters. If more features can be obtained in the future, such as permeability, stimulated reservoir volume, drilling and completion parameters, they will also be added to the model to improve its generalization ability.

### Conclusions

This paper utilized machine learning methods to evaluate the productivity potential of 282 fractured shale gas wells in WY block, and the specific content of this study includes: featuring ranking, feature reconstruction, well classification, productivity potential evaluation and optimization. Several conclusions from this research are presented below:

1. Thickness of organic-rich gas reserves and amount of hydraulic fracturing sand are vital parameters affecting gas productivity, and geological condition is more deterministic to gas productivity.
2. The model established in this paper can accurately divide shale gas wells into four categories according to their geological conditions and productivity potential, mainly including wells with good geological condition,
poor fractoring measures and low gas production; wells with poor geological condition and the productivity is low regardless of the fractoring measures; wells with superior geological condition, good fractoring measures and high production; wells with medium geological condition and moderate gas production.

3. The model constructed in this paper can classify shale gas wells according to their productivity difference, it can select wells with good fractoring measures and good geological conditions as sample wells. Meanwhile, wells with high productivity potential but low production due to poor fractoring measures can also be selected as wells that can be refractured.

4. Six low production wells with good geological conditions but unreasonable design of fractoring operating parameters have been optimized successfully, and the gas production rate of those wells increased by an average of five times after optimization. Besides, this kind of wells should be highly valued as their productivity potential has not been fully exploited.

In the future, the number of samples and features will be increased to improve the generalization ability of the model.
Lu C, Jiang H, You C et al (2021) A novel method to determine the thief zones in heavy oil reservoirs based on convolutional neural network. J Petrol Sci Eng 201:108471

Xinhua Ma, Xie J . The progress and prospects of shale gas exploration and development in southern Sichuan Basin, SW China[J]. Petroleum Exploration and Development, 2018.

Ma H, Yu G, She Y et al. 2019 Waterflooding Optimization under Geological Uncertainties by Using Deep Reinforcement Learning Algorithms. Paper presented at the SPE Annual Technical Conference and Exhibition, Calgary, Alberta, Canada, 30–2 September. SPE-196190-MS. https://doi.org/10.2118/196190-MS.

Ma X, Li X, Liang F et al (2020) Dominating factors on well productivity and development strategies optimization in Weiyuan shale gas play, Sichuan Basin, SW China[J]. Petrol Explor Develop 47(3):594–602

Martinez AM, Kak AC (2001) PCA versus LDA. IEEE Trans Pattern Anal Mach Intell 23(2):228–233. https://doi.org/10.1109/34.908974

Mo S, Zhu Y, Zabaras N et al (2019) Deep convolutional encoder-decoder networks for uncertainty quantification of dynamic multiphase flow in heterogeneous media. Water Resources Research 55:703–728

Mohammadmoradi P, Kantzas A (2018) Wettability and Capillary Imbibition in Shales Analytical and Data-Driven Analysis. Soc Petrol Eng. https://doi.org/10.2118/182623-MS

Peter RJ (1999) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis [J]. J Comput Appl Math 20

Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9(2):181–199

Qian K, He Z, Lu X, Chen Y (2018) Intelligent prediction and integral analysis of shale oil and gas sweet spots. Petrol Sci 15:744–755

Qin X, Liu M, Zhang L et al (2021) Structural Protein Fold Recognition Based on Secondary Structure and Evolutionary Information Using Machine Learning Algorithms[J]. Comput Biol Chem 91:107456

Shlens J 2014 A Tutorial on Principal Component Analysis[J]. International Journal of Remote Sensing 51(2).

Syed FL, Muther T, Dahaghi AK et al (2021) AI/ML assisted shale gas production performance evaluation. J Petrol Explor Prod Technol 11:3509–3519. https://doi.org/10.1007/s13202-021-01253-8

Tahmasebi P, Javaphour F, Shamsi M (2017) Data mining and machine learning for identifying sweet spots in shale reservoirs. Expert Syst Appl 88(435):447

Vikara D, Remson D, Khann V (2020) Machine learning-informed ensemble framework for evaluating shale gas production potential: Case study in the Marcellus Shale[J]. J Nat Gas Sci Eng 84:103679

Yan B, Mi L, Wang Y, Tang H, An C, Killough JE (2017) Mechanistic Simulation Workflow in Shale Gas Reservoirs. Soc Petrol Eng. https://doi.org/10.2118/182623-MS

Yao M, Chang H, Li X et al (2019) An Integrated Approach for History Matching of Multiscale-Fractured Reservoirs. SPE J 24(4):1508–1525

Zhao X, Pu X, Jin F et al (2019) Geological characteristics and key exploration technologies of continental shale oil sweet spots: a case study of Member 2 of Kongdian Formation in the Cangdong sag in Huanghua depression, Bohai Bay Basin. Petrol Res 4:97–112

Zhu G, Kou J, Yao J et al (2020) A Phase-Field Moving Contact Line Model with Soluble Surfactants. J Comput Phys 405:109170. https://doi.org/10.1016/j.jcp.2019.109170

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.