Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2022.020498

ARTICLE

A Data-Driven Oil Production Prediction Method Based on the Gradient Boosting Decision Tree Regression

Hongfei Ma[1,*], Wenqi Zhao[2], Yurong Zhao[1] and Yu He[1]

[1]School of Energy Resources, China University of Geosciences (Beijing), Beijing, 100083, China

[2]Research Institute of Petroleum Exploration and Development, China National Petroleum Corporation, Beijing, 100083, China

*Corresponding Author: Hongfei Ma. Email: 2106190066@cugb.edu.cn

Received: 27 November 2021 Accepted: 19 April 2022

ABSTRACT

Accurate prediction of monthly oil and gas production is essential for oil enterprises to make reasonable production plans, avoid blind investment and realize sustainable development. Traditional oil well production trend prediction methods are based on years of oil field production experience and expertise, and the application conditions are very demanding. With the rapid development of artificial intelligence technology, big data analysis methods are gradually applied in various sub-fields of the oil and gas reservoir development. Based on the data-driven artificial intelligence algorithm Gradient Boosting Decision Tree (GBDT), this paper predicts the initial single-layer production by considering geological data, fluid PVT data and well data. The results show that the GBDT algorithm prediction model has great accuracy, significantly improving efficiency and strong universal applicability. The GBDT method trained in this paper can predict production, which is helpful for well site optimization, perforation layer optimization and engineering parameter optimization and has guiding significance for oilfield development.

KEYWORDS

Gradient boosting decision tree; production prediction; data analysis

Nomenclature

GBDT	Gradient Boosting Decision Tree
RF	Random Forest

e.g.,

$\emptyset$	Porosity
s	Skin Factor
N	Sample Space
x	Input Variable
x_n	Input Vector of a Single Sample
$\boldsymbol{Y}$	Output Vector
$\{y_i, \boldsymbol{x}_i\}_1^N$	A Training Sample

$L(y,F(x))$	Loss Function
a	Residual
$F_0(\boldsymbol{x})$	The Initialization Model
M	Model Numbers
$g_m(\boldsymbol{x})$	Negative Gradient
$h_m(x)$	A Small Decision Tree
J_m	The Leaf of Each Decision
R_{1m}	Independent Regions
b_{jm}	The Constant Value of R_{jm} Region
ρ_{lm}	Step Size
R	Learning Rate
q_i	Oil Well Production Rate of i-th Layer
Q_t	Total Oil Production Rate
kh	Stratigraphic Coefficient
k	Formation Permeability
h	Formation Thickness
T	Conductivity
d	Top Layer Depth
NTG	Net to Gross Ratio
S_{oi}	Initial Oil Saturation
p_i	Initial Formation Pressure
k_{ro}	Relative Permeability of Crude Oil
μ	Viscosity of Crude Oil
Bo	Volume Coefficient of Crude Oil
C_t	Compressibility of Total
R_g	Gas-Oil Ratio
R^2	The Coefficient of Determination

1 Introduction

In the efficient development of oil and gas reservoirs, the most important thing is to determine the productivity of production wells, which is crucial for making development plans and reducing cost and improving efficiency. The conventional methods for predicting oil well production include: predicting oil well production under different flow pressures according to Inflow Performance Relationship (IPR) Curve. By modifying the IPR curve continuously to make the prediction effect more accurate [1–5]. The Arps decline curve, Fetkovich modern production decline curve, A-G production decline method, can be used to predict future production by establishing plates [6–11]. It can be seen that the traditional production prediction model usually needs to be simplified and ignore the influence of some factors, so the application conditions are very demanding. The application of traditional prediction methods requires researchers to have certain oilfield development knowledge and experience, which has low efficiency and high cost.

As an alternative, machine learning algorithms are gaining more and more attention from researchers because they can process large amounts of existing data quickly with the least cost. Nowadays, each reservoir has its own data system, which contains a large amount of geological and development data, but the potentials hidden in these massive data have not been fully explored, and it is urgent to make full use of them for further exploration and development. Scholars at home and abroad have used artificial intelligence algorithms to carry out a series of exploratory studies in geological

analysis, well logging interpretation, seismic interpretation, sweetness prediction, geological modeling, reservoir simulation and other aspects [12]. Among many algorithms, the integrated algorithm is the most effective [13]. Because it improves the prediction accuracy of the final algorithm by integrating multiple weak learners. These applications include: reservoir permeability prediction [14], generation of complex geological facies [15], identification of core CT scanning images [16]; Prediction of formation fluid parameters [17]; Predicting the gas-oil ratio of oil sand reservoirs [18]; Detection of faults and their dip angles. In terms of production prediction, Gupta et al. [19] used neural network and time series analysis to forecast production performance of an unconventional layer. The prediction algorithms are fast and have high accuracy within reasonable tolerance. Crnkovic et al. [20] collected more than 8 hundred well data and 2 million geological data points to predict estimated ultimate recovery based on deep learning algorithms. Anderson et al. [21] developed an integrated machine learning and statistical algorithm to predict the ultimate productivity of oil, gas, and water by calculating the importance weights of valuable data and convolving these weights with the actual attribute values of each wells. Liang et al. [22] collected a data set containing over 25 variable types from 4000 wells in the Eagle Ford and used random forest (RF) to predict the Oil and Gas EUR. Besides, the methods they used can rank the importance of relevant independent variables.

The above scholars mainly predict production based on some simple algorithms, such as support vector machines, neural networks or their variants. But for weak learner, which is the integrated algorithm of decision tree, only some scholars use it for production prediction. RF and Gradient Boosting Decision Tree (GBDT) are two typical integrated algorithms based on decision tree. Wang et al. [23] used RF to predict the cumulative production in Montney formation with relatively high accuracy. Xue et al. [24] established a mechanism model through numerical simulation, which generated a large amount of data and predicted the future production of shale gas based on multi-objective random forest regression. Wang et al. [25] predicted the first annual productivity of sub-wells based on RF, GBDT, Linear Regression and neural network, among which RF and GBDT had significantly better prediction effect than the latter. Alwated et al. [26] mainly studied the machine learning technology, including random forest, gradient boosting regression and decision tree to predict the migration of fluid in porous media. However, the GBDT algorithm has not been fully demonstrated its ability to predict single-layer production, and complex geological conditions have not been fully considered.

In this paper, GBDT algorithm will be employed to predict the production of oil wells in the early stage of single-layer. The GBDT method can construct a decision tree along the gradient descent direction, obtaining the final training model after weighting all the small decision trees, and the algorithm can be implemented in parallel on multi-core computers. These characteristics enable it to avoid over-fitting and deal effectively with large data sets with various features. Firstly, the detailed theoretical model of GBDT and production splitting method are introduced. Secondly, multiple features including overall geological data, fluid PVT data, and well data are considered. The data set of the training model is obtained through data preprocessing. Finally, optimal hyper parameters are obtained by multiple training and the final GBDT algorithm with good immunity and robustness is picked. The algorithm also predicts the prediction effect of single-layer oil well production, and obtains the ranking of each feature importance.

2 Equations and Mathematical Expressions

2.1 Loss Function

The input vector of a single sample is defined as Eq. (1):

$$\boldsymbol{x}_i = \left(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(p)}, \cdots, x_i^{(P)}\right), \tag{1}$$

where P is the dimension of single sample vector $\boldsymbol{x}_i$, equal to the features.

The data set for this article is defined as Eq. (2):

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(p)} & \cdots & x_1^{(p)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(p)} & \cdots & \vdots \\ \vdots & \vdots & \ddots & & & \vdots \\ x_n^{(1)} & x_n^{(2)} & & \ddots & & \vdots \\ \vdots & \vdots & & & \ddots & \vdots \\ x_N^{(1)} & x_N^{(2)} & \cdots & x_N^{(P)} & \cdots & x_N^{(p)} \end{pmatrix}, \tag{2}$$

where N is the sample space.

The final output vector is defined as Eq. (3):

$$\boldsymbol{Y} = (y_1, y_2, \cdots, y_n, \cdots, y_N)^T, \tag{3}$$

where y_n is the calculating result, which is corresponded to input vector X_n.

Based on a training data set $\{y_i, \boldsymbol{x}_i\}_1^{\mathrm{N}}$ with known $(y, \boldsymbol{x})$ value, the goal is to find a function $F*(\boldsymbol{x})$ that can match $\boldsymbol{x}$ to y. Thus, the expected value of some specified loss function $L(y, F(\boldsymbol{x}))$ is minimized on the joint distribution of all known $(y, \boldsymbol{x})$ values [27].

In the regression problem, the loss function is applied to measure the residual between the predicted value and the actual value. The smaller the residual, the better the model's prediction performance. The frequently used loss functions include square loss function, absolute loss function and Huber loss function. Huber loss function is a combination of square loss function and absolute loss function. In this study, Huber loss function will be employed as the loss function of the GBDT algorithm, which will help to improve the prediction accuracy of the model.

Variable a in Huber's loss function is the residual: $a = y - F(\boldsymbol{x})$, and the expression of the loss function is Eq. (4) [28]:

$$\mathrm{L}(y, F) = \begin{cases} \dfrac{1}{2}(y - F)^2 & |y - F| \le \delta \\ \delta\left(|y - F| - \delta/2\right) & |y - F| > \delta \end{cases}, \tag{4}$$

2.2 Gradient Boosting Decision Tree Algorithm

The idea of gradient boosting is derived from Breiman, which is a boosting optimization algorithm based on loss function [29]. Gradient boosting can be used as a machine learning technology to solve regression and classification problems. Friedman then developed the gradient boosting algorithm, which is an improvement to boosting algorithm: a new model is established in the gradient direction of residual reduction after each iteration. Finally, the loss function is minimized by moving towards negative gradient direction. The final model is obtained by weighting all decision trees. The advantage of this method is that it can deal with various types of data flexibly, including continuous value and discrete value [30]. The processes of GBDT algorithm is as follows:

The goal of the model is to find an approximate $F*(\boldsymbol{x})$ for $F(\boldsymbol{x})$. At the same time, the expected value of a given loss function $L(y, F(\boldsymbol{x}))$ on the training data set is the minimum. The minimum expected value can be defined as Eq. (5):

$$F^{*}(\mathbf{x}) = \arg\min_{F(x)} \sum_{i=1}^{N} L(y, F(x)), \tag{5}$$

where $F(\boldsymbol{x})$ is the prediction function, y is the observed value, and $L(y, F(\boldsymbol{x}))$ is the Huber loss function.

First, the Huber loss function is introduced to match a constant function $F_0(\boldsymbol{x})$ of model initialization, and the number of model numbers is M.

The initialization model is as Eq. (6):

$$F_0(\boldsymbol{x}) = \arg\min_{F(x)} \sum_{i=1}^{M} \mathrm{L}(y, F(x)). \tag{6}$$

The model iterates along the direction of gradient descent and continuously reduces the residual $L(y, F(\boldsymbol{x}))$. For the m-th iteration of the model, the current value of the negative gradient of the loss function is calculated and used as the estimated value of the residual. The negative gradient is calculated as Eq. (7):

$$g_m(\boldsymbol{x}) = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x_i)=F_{m-1}(x_i)}, \tag{7}$$

where $g_m(\boldsymbol{x})$ is the negative gradient value.

For residual, gradient boosting regression tree model will match a small decision tree $h_m(x)$ through CART algorithm [29]. Assuming that the leaf of each decision tree is J_m, the input space is divided into J_m independent regions: R_{1m}, R_{2m}, R_{3m}, ..., R_{jm}, and the constant output value from each region is determined. For example, b_{jm} is the constant value of R_{jm} region. The decision regression tree expression is Eqs. (8) and (9):

$$h_m(x) = \sum_{j=1}^{J_m} b_{jm} I(\boldsymbol{x} \in R_{jm}), \tag{8}$$

$$\text{where } I(\boldsymbol{x} \in R_{jm}) = \begin{cases} 1, & \text{if } \boldsymbol{x} \in R_{jm} \\ 0, & \text{if } \boldsymbol{x}/ \in R_{jm} \end{cases}, \tag{9}$$

The step size Eq. (10) of the model gradient descent is calculated by using a separate linear search method in each terminal node R_{lm}

$$\rho_{lm} = arg\min_{\rho} \sum_{xi\varepsilon Rlm} L(y_i, F_{m-1}(\boldsymbol{x}_i) + \rho h_{lm}), \tag{10}$$

In order to correct the residual of the previous decision regression tree, the model adds a new regression tree at each step. The model is determined as Eq. (11):

$$F_m = F_{m-1}(x) + \rho_m h_m(x), \tag{11}$$

In addition, the gradient boosting algorithm in this paper introduces Shrinkage coefficient. This method can control the step size each time and avoid fast approximation results causing overfitting. Therefore, GBDT model is updated as Eq. (12):

$$F_m = F_{m-1}(x) + R \times \rho_m h_m(x), 0 < R \leqslant 1, \tag{12}$$

The parameter R is called the "learning rate", which is the weight reduction factor for each base classifier.

The final GBDT algorithm training model of this paper is Eq. (13):

$$F^*(x) = F_M(x) = \sum_{m=1}^{M} R * F_m(x), \tag{13}$$

It can be seen that GBDT algorithm not only introduces Huber loss function as a method of fitting residual error, but also improves the prediction accuracy by constantly adjusting the weight of the weak learner of the model.

The determination coefficient (R^2) is employed to verify the generalization ability of the Regressor. The algorithm with a larger R^2 is always favored because the model is more reliable.

R^2 is defined as Eq. (14):

$$R^2 = 1 - \frac{\sum_i \left(F^{(i)} - y^{(i)}\right)^2}{\sum_i \left(\bar{F} - y^{(i)}\right)^2} = 1 - \frac{\left(\sum_{i=1}^{m} \left(F^{(i)} - y^{(i)}\right)^2\right)}{\left(\sum_{i=1}^{m} \left(F^{(i)} - \bar{y}\right)^2\right)}, \tag{14}$$

2.3 Data Generation

In order to establish a data-driven oil well production prediction model, it is necessary to collect a large number of geological data and production data. The features affecting oil production are used as the input data set, and the corresponding oil production is used as the output data set. The GBDT algorithm established in this paper is to establish the quantitative relationship between oil production and formation parameters, so that it can be used for accurate and rapid prediction, and replace or supplement the mathematical model. In order to ensure the practicability and reliability of the algorithm, it is necessary to collect the data set from the actual production.

In this paper, 50 oil wells identified in a geological unit A in North China are selected as data sources, the GBDT algorithm is constructed and its performance to predict single layer initial production is evaluated. Geological unit A is located in the southwest wing of the dome anticline structure, cut by faults in the north and east, and connected with the edge water in the southwest. It is a medium-high permeability sandstone reservoir of delta front deposits. The unit is characterized by a simple structure and gentle stratum, with a dip Angle of 2–5° to the southwest. The average porosity is 0.25, while average permeability is 2800 mD. The sedimentary unit division divides it into 3 sand groups and 27 small layers, including 12 sand layers.

All wells are multi-layer commingled production. In this study, the production will be divided according to the proportion of formation coefficient of each production layer Eq. (15):

$$q_i = \mathrm{Q_t} * \frac{k_i h_i}{\sum kh}. \tag{15}$$

where, q_i is oil well production rate of i-th layer, Q_t is the total oil production rate of all perforation zone, $k_i h_i$ is stratigraphic coefficient of i-th layer [31].

There are many complex factors affecting the initial production of single layer, which can be summarized into the following categories: geology parameters, fluid PVT, well parameters [31,32]. In this paper, 15 parameters are selected as the features of the training machine learning model. These characteristics are formation coefficient (kh), formation permeability (k), wellbore peripheral permeability (k_{local}), porosity (Φ), conductivity (T), top layer depth (d), formation thickness (h), net

to gross ratio (NTG), initial oil saturation (S_{oi}), initial formation pressure (p_i), relative permeability of crude oil (k_{ro}), viscosity of crude oil (μ), volume coefficient of crude oil (B_o), compressibility of total (C_t), and gas-oil ratio (R_g). The first nine are basic geological features, most of which are static parameters. Oil saturation is a dynamic parameter related to production time. The rest are fluid characteristics, all of which are dynamic parameters corresponding to the initial stage of perforation production and will change with the production process.

Data quality is the basis of prediction analysis, so sufficient data preprocessing must be carried out. In this study, the collected data were preprocessed as follows: (i) null values and single-value features were removed; (ii) few features with missing many values were eliminated as a whole; (iii) some samples with missing features were deleted; (iv) abnormal data was replaced by interpolation. Finally, 2512 valid data are obtained. The data sets are randomly divided in 8:1:1 ratio, with 80% as training data sets, 10% as validation data sets, and 10% as test data sets.

The format of the datasets is shown in Table 1. The features' scatter diagrams with weight greater than 2% can be seen in Appendix.

Table 1: The format of the datasets

k/mD	p_i/MPa	S_{oi}	k_{local}/mD	φ	T	d/m	h/m
1709.6	22.86	0.32	1550.60	0.28	17.28	2278.10	11.96
2156.3	22.58	0.40	1999.00	0.29	49.33	2259.90	5.41
2423.1	21.35	0.63	2781.70	0.31	141.06	2197.40	6.06
945.64	21.35	0.64	927.40	0.30	41.27	2182.80	8.64
2108.6	22.33	0.40	2276.90	0.29	18.96	2212.00	7.08
2312.0	21.58	0.59	2781.70	0.31	141.06	2197.40	6.06
847.60	21.59	0.61	927.40	0.30	41.27	2182.80	8.64
2151.1	8.88	0.51	2277.20	0.29	29.45	2247.00	4.32
1881.3	22.60	0.30	1999.00	0.29	49.33	2259.90	5.41
1611.0	22.55	0.50	1651.50	0.29	29.07	2253.70	3.63

NTG	Formation coefficient	Viscosity	Volume factor	Compressibility	Gas-oil ratio	Relative permeability	Production m^3/d
0.46	9362.54	4.00	1.30	0.0005	53.33	0.01	36.75
0.90	10541.15	4.00	1.30	0.0005	53.33	0.06	41.38
0.84	12311.80	4.00	1.30	0.0005	53.33	0.80	26.41
1.00	8170.24	4.00	1.30	0.0005	53.33	0.88	17.52
0.75	11188.76	4.00	1.30	0.0005	53.33	0.06	13.09
0.84	11747.30	4.00	1.30	0.0005	53.33	0.53	13.74
1.00	7323.18	4.00	1.30	0.0005	53.33	0.65	8.57
0.55	5108.10	4.14	1.16	0.0041	38.62	0.24	96.67
0.90	9196.80	4.00	1.30	0.0005	53.33	0.01	17.64
0.97	5700.02	4.00	1.30	0.0005	53.33	0.22	10.93

3 Results and Discussion

3.1 Hyper-Parameter Optimization

Hyper-parameter optimization is a significant step in improving the accuracy of the machine learning algorithm. This paper mainly considers three critical parameters to improve the accuracy of GBDT algorithm:

(1) N_estimators
(2) Learning rate
(3) Max depth

It is common that the final prediction model can effectively reduce the training error with the increase of the maximum number of N-estimators. However, its generalization ability will be greatly weakened if there are too many decision trees, which will even cause over-fitting phenomenon. It is necessary to optimize the number of decision trees in order to reduce training errors and prevent overfitting. The gradient boosting algorithm with a low learning speed ($R \leq 0.1$) can effectively increase the generalization capability of the algorithm. If learning speed is too low ($R < 0.01$), the number of decision trees should be increased to make the model converge. The max depth is also an essential factor affecting the prediction accuracy. The model with shallow depth has poor prediction performance, but deep depth will increase time cost. To achieve the best forecast effect and save trainin time, the specific optimal values of above three parameters need to be determined by continuous trial calculation based on the actual sample data set.

There are many kinds of hyper-parameter optimization methods, such as Ant colony optimization and particle swarm optimization [33,34]. But these hyper-parameter optimization algorithms are basically heuristic algorithms and can only get suboptimal solutions. To fully study the effect of hyper-parameters on prediction model, this paper will apply Grid Search method to determine the optimal values of the hyper-parameters. The value range and step size of the hyper-parameters are shown in Table 2.

Table 2: The hyper-parameters of GBDT

Algorithm parameters	Parameter value range	Optimal value combination
N_estimators	[10, 210] step size 25	60
Learning rate	[0.01, 0.1] step size 0.01	0.1
Max depth	[5, 15] step size 1	6.0

The optimization results of N_estimators are shown in Fig. 1a. It can be seen that when N_estimators increase at the initial stage (N_estimators < 60), R^2 increases rapidly until it reaches the maximum value of 0.977. At this point, the RMSE value of prediction model is 1.10 and the MAPE value is 0.58. This shows that the predictive performance of GBDT algorithm is pretty good. R^2 cannot increase any more when N_estimators is greater than 60. And as the number of weak learners increase, time cost of calculation will rise. Therefore, the optimal N_estimators are determined as 60. The relationship curve between tree depth and R^2 value is shown in Fig. 1b. The depth of tree determines the complexity of decision tree. The prediction accuracy of the model increases with the increase of depth in the early stage, but over-fitting phenomenon may occur in the later stage, and the prediction ability of the model decreases. The preset maximum depth range of weak learners is [5,15]. It can be found that the optimal maximum depth is 8 in this study. Based on the optimized

N_estimators and the maximum depth, the learning rate can be further determined. The curve of the relationship between learning rate and R^2 is shown in Fig. 1c. As can be seen from the figure, when the learning rate value is relatively small ($0.01 < R < 0.04$), the prediction performance of the model is general; when the learning rate is greater than 0.05, the training effect of the model is excellent. In order to achieve the optimal training model, learning rate is determined to be 0.1. The third column of Table 2 lists all the best hyper-parameters values.

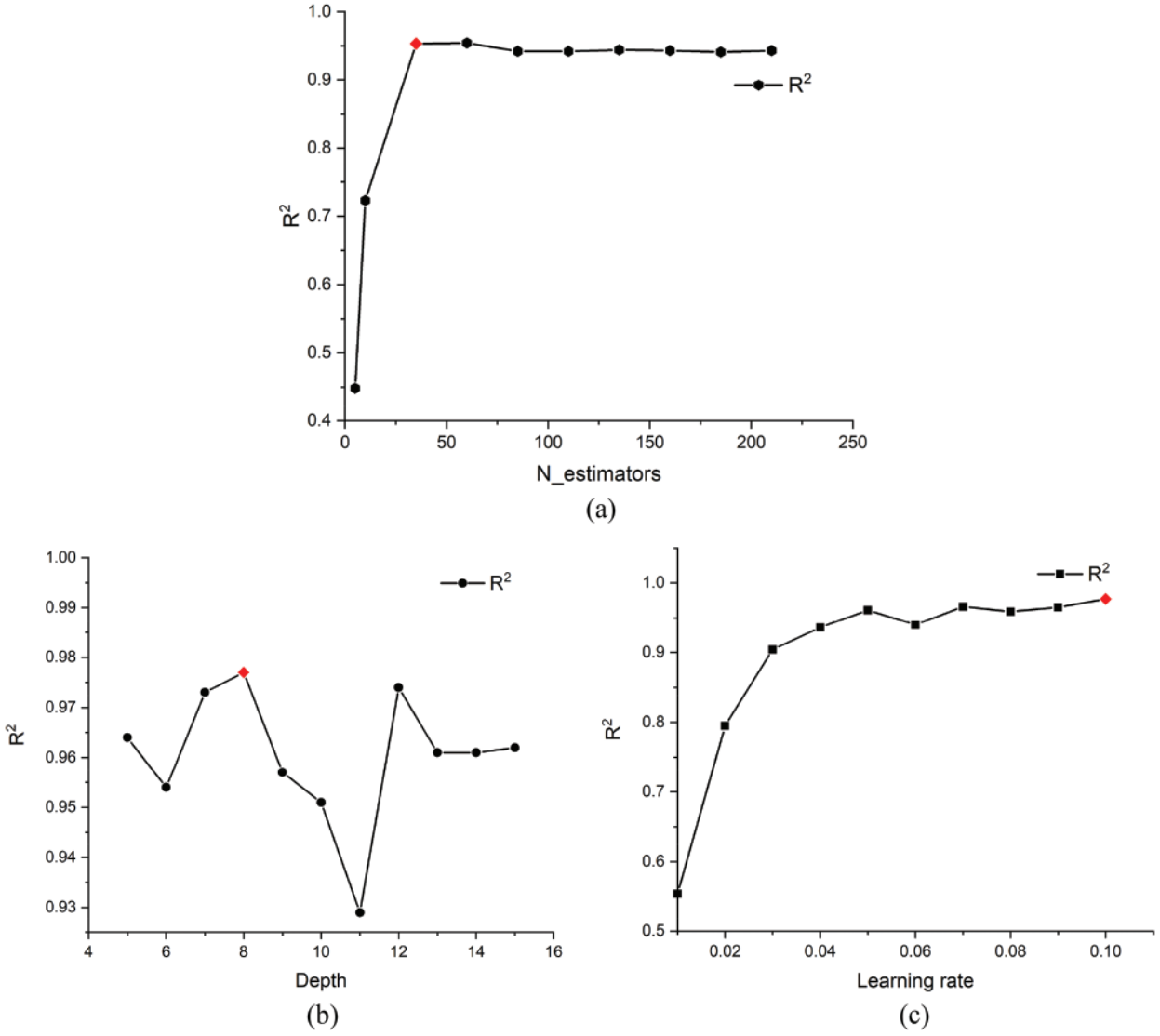

Figure 1: Determination of hyper-parameter for GBDT method

Fig. 2 shows a small decision tree of GBDT algorithm. The decision tree achieves prediction by constructing binary tree. Each internal node represents a test on a property, each branch represents a test output, and each leaf node represents a category. According to the feature parameters of the training data set, the binary tree is divided continuously and generated finally. GBDT prediction algorithm continuously optimizes the decision tree along the gradient descent direction.

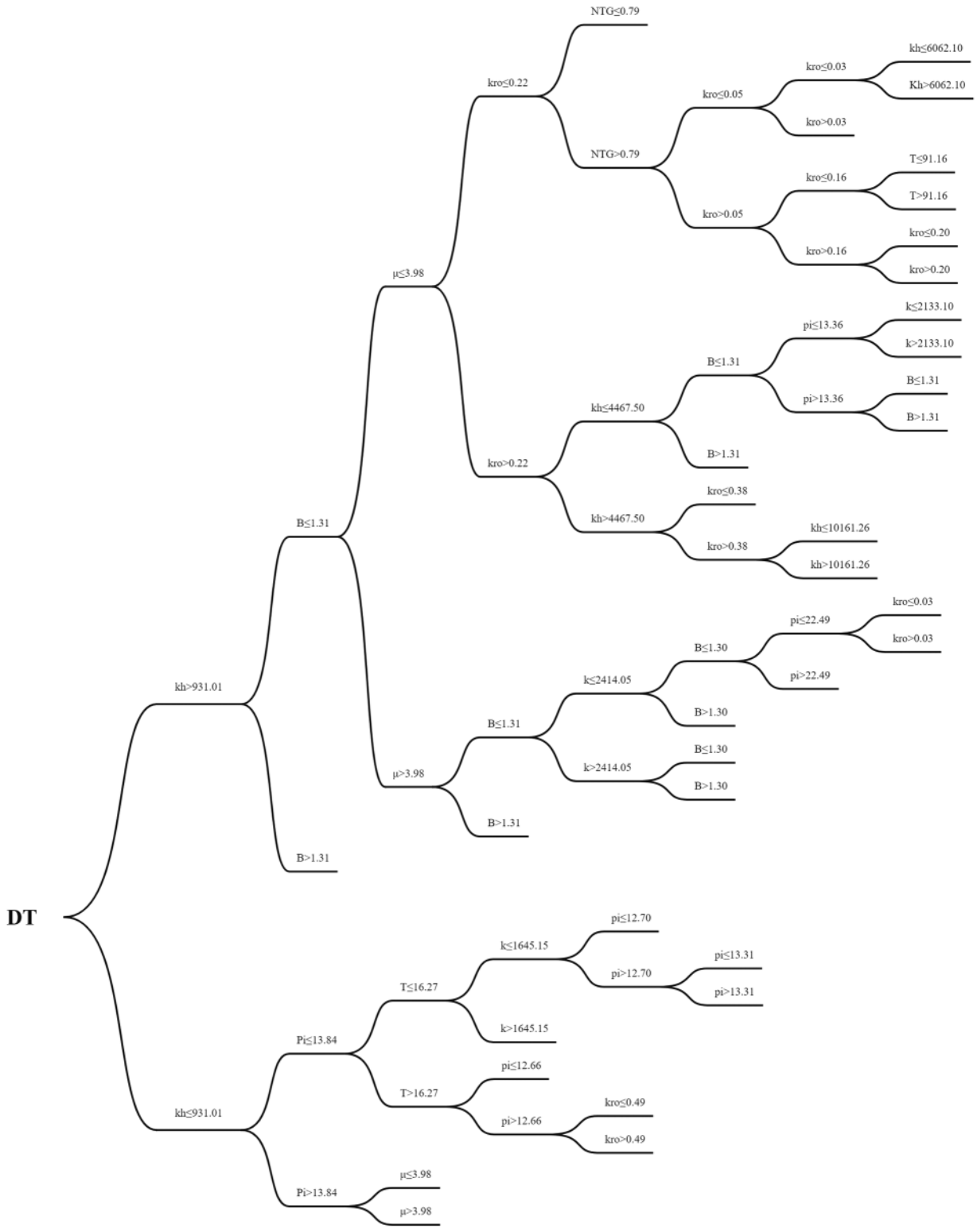

Figure 2: The diagram a decision tree of GBDT algorithm

3.2 Prediction Performance of GBDT Algorithm

To fully assess the prediction accuracy of the data-driven algorithm, this study used a test dataset from the same geological unit. It can solve such a difficult problem that how much the initial production will be if there are only fluid PVT data and geological parameters.

Fig. 3 shows the comparison between test dataset actual production and model predicted production. The curve represents relative error, the blue bar represents the initial daily production of the actual wells in a single layer, the orange band represents the predicted production of GBDT algorithm, and the red bar represents predicted production of RF algorithm. The initial single-layer daily production data is obtained by dividing the total production of three months by the number of production days. Random forest algorithm is also a common intelligent algorithm, which is used to verify GBDT in this paper. As a whole, the predicted productions of both algorithms fit very closely with the average single-layer daily production data, with an average relative error of 0.226 and a maximum relative error of 1.797. It shows that the GBDT and RF algorithms based on actual production data are very accurate and can be further used to predict single-layer production. And GBDT algorithm has better prediction performance than random forest.

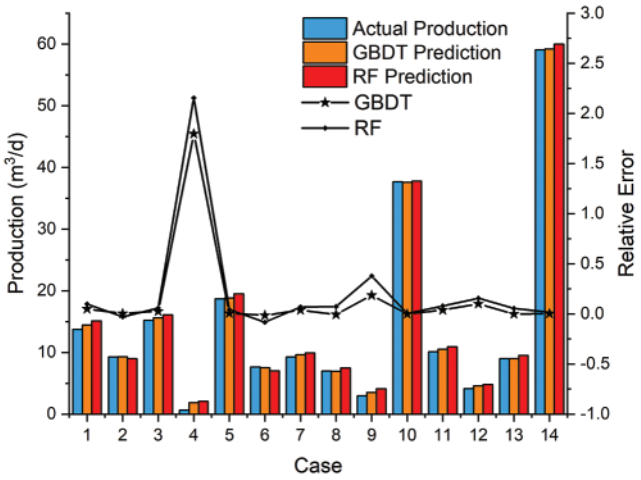

Figure 3: Comparing actual production and model predicted production

Fig. 4 shows the importance of each feature of the trained model. Features with great weight have a great influence on the predicted value. If the feature weight ratio is less than 2%, the feature can be ignored, which can make the model more concise and efficient. It can be seen that the first three features that have the greatest impact on the initial production data are formation coefficient, relative permeability of crude oil, and volume coefficient of crude oil. Besides, crude oil viscosity, initial formation pressure, average conductivity, net gross ratio and formation thickness are the secondary factors. The effect of compressibility and gas oil ratio on initial production can be ignored. The ranking of feature importance emphasizes that the physical properties of reservoir and fluid are significant to optimize the development layer and perforation location. If these features with great importance are considered before the drilling process, underground fluid can be efficiently extracted and development costs can be reduced.

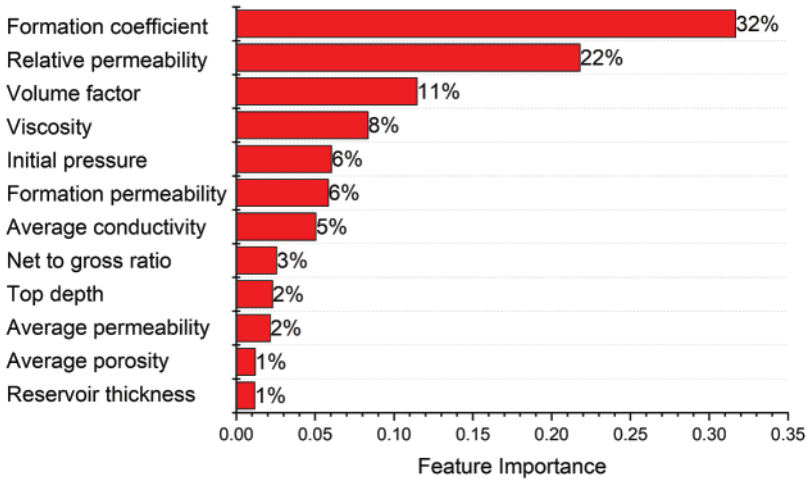

Figure 4: Feature importance of the 13 geological and PVT parameters

Fig. 5 shows the partial dependence of the model on the formation coefficients. By controlling other features unchanged and changing the values of formation coefficient *kh*, the partial dependence plots can be obtained. It is helpful to intuitively analyze the influence of formation coefficient on the initial production of target layer.

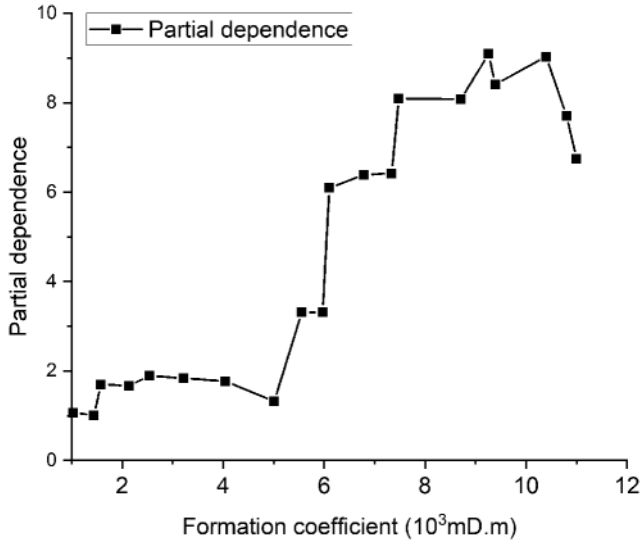

Figure 5: Partial dependence plots of the model on the formation coefficients

The Fig. 5 shows that when the formation coefficient is small ($kh < 7.48 * 10^3$ mD.m), the single-layer production increases rapidly with the rise of formation coefficient. However, when the formation coefficient is large ($kh > 7.48 * 10^3$ mD.m), single-layer production increases very slowly with the increase of formation coefficient. Because the production energy and geological reserves are limited, and they cannot be supplemented effectively during production.

4 Conclusions

In this paper, a new production prediction model is established based on actual oil reservoir production data and artificial intelligence algorithm. The algorithm can be used to predict the initial production of single-layer with very reasonable accuracy. The conclusions are listed as follows:

(1) The GBDT algorithm is trained to be an intelligent model for predicting initial production of single-layer of oil wells, with 15 features as input data, including geological data and fluid PVT data. Huber Loss Function and hyper-parameter optimization ensure the prediction accuracy of the model. The R^2 of the final model is 0.977 and RMSE is 1.10. The prediction results of test set data show that the prediction performance of this model is excellent because the average relative error is only 0.226%.

(2) The overfitting problem is mainly dealt with by controlling the number and depth of decision trees and limiting the value of learning rate. Because the datasets of major oil reservoirs are different, the model hyper-parameters can only be determined by continuous trial calculation of Grid Search method.

(3) Through weight analysis, it is found that formation parameters, such as formation permeability and relative permeability of crude oil, are the most important factors affecting production prediction. Further analysis of the partial dependence plots of the prediction model on formation coefficient *kh* shows that the initial single-layer production is positively correlated with formation coefficient, but it will be limited by formation energy and geological reserves.

Acknowledgement: This research was supported by China University of Geosciences (Beijing).

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Vogel, J. V. (1968). Inflow performance relationships for solution-gas drive wells. *Journal of Petroleum Technology, 20(1),* 83–92. DOI 10.2118/1476-PA.
2. Richardson, J. M., Shaw, A. H. (1982). Two-rate IPR testing-a practical production tool. *Journal of Canadian Petroleum Technology, 21(2),* 57–61. DOI 10.2118/82-02-01.
3. Standing, M. B. (1970). Inflow performance relationship for damaged wells producing by solution, gas drive. *Journal of Petroleum Technology, 22(11),* 1399–1400. DOI 10.2118/3237-PA.
4. Standing, M. B. (1971). Concerning the calculation of inflow performance of wells producing from solution gas drive reservoirs. *Journal of Petroleum Technology, 23(9),* 1141–1142. DOI 10.2118/3332-PA.
5. Arps, J. J. (1945). Analysis of decline curves. *Transactions of the AIME, 160(1),* 228–247.DOI 10.2118/945228-G.
6. Fetkovich, M. J. (1980). Decline curve analysis using type-curves. *Journal of Petroleum Technology, 32(6),* 1065–1077. DOI 10.2118/4629-PA.
7. Blasingame, T. A., McCray, T. L., Lee, W. J. (1991). Decline curve analysis for variable pressure drop/variable flowrate system. *SPE Gas Technology Symposium*, pp. 1–17. Houston. DOI 10.2118/21513-MS.
8. Agarwal, R. G., Gardner, D. C., Kleinsteiber, S. W. (1999). Analyzing well production data using combined type curve and decline curve concept. *SPE Reservoir Evaluation & Engineering, 2(5),* 478–486. DOI 10.2118/57916-PA.

9. Yu, Q. T. (2000). How to do it when arps decline index n<0 or n≥1. *Xinjiang Petroleum Geology, 21(5),* 408–411. DOI 10.3969/j.issn.1001-3873.2000.05.015.
10. Chen, Y. Q., Guo, E. P. (2008). Building and applying a new decline model of oilfield production. *China Offshore Oil and Gas, 20(6),* 379–391. DOI 10.3969/j.issn.1673-1506.2008.06.006.
11. Huang, J. C., Zhang, J. C. (2021). Overview of oil and gas production forecasting by machine learning. *Petroleum Reservoir Evaluation and Development, 11(4),* 613–620. DOI 10.13809/j.cnki.cn32-1825/te.2021.04.018.
12. Li, Y., Lian, P. Q., Xue, Z. J., Dai, C. (2020). Application status and prospect of big data and artificial intelligence in oil and gas field development. *Journal of China University of Petroleum, 44(4),* 1–11. DOI 10.3969/j.issn.1673-5005.2020.04.001.
13. Zhou, C. D., Wu, X. L., Chen, J. A. (1993). Determining reservoir properties in reservoir studies using a fuzzy neural network. *SPE Annual Technical Conference and Exhibition*, pp. 1–10. Houston. DOI 10.2118/26430-MS.
14. Saljooghi, B. S., Hezarkhani, A. (2015). A new approach to improve permeability prediction of petroleum reservoirs using neural network adaptive wavelet (wavenet). *Journal of Petroleum Science and Engineering, 133,* 851–861. DOI 10.1016/j.petrol.2015.04.002.
15. Zhang, T. F. (2019). Generating geologically realistic 3D reservoir facies models using deep learning of sedimentary architecture with generative adversarial networks. *Petroleum Science, 16(3),* 541–549. DOI 10.1007/s12182-019-0328-4.
16. Chen, G. J., Guo, W. H., Fan, P. Z. (2017). Study on rock image classification based on convolution neural network. *Journal of Xi'an Shiyou University (Natura Science Edition), 32(4),* 116–122. DOI 10.3969/j.issn.1673-064X.2017.04.020.
17. Chukwuma, O. (2018). Application of machine learning ideas to reservoir fluid properties estimation. *SPE Nigeria Annual International Conference and Exhibition*, pp. 1–20. Lagos. DOI 10.2118/208264-MS.
18. Akbilgic, O., Zhu, D., Gates, I. D., Bergerson, J. A. (2015). Prediction of steam-assisted gravity drainage steam to oil ratio from reservoir characteristics. *Energy, 93,* 1663–1670. DOI 10.1016/j.energy.2015.09.029.
19. Gupta, S., Franz, F., Benin, C. J. (2014). Production forecasting in unconventional resources using data mining and time series analysis. *SPE/CSUR Unconventional Resources Conference*, pp. 1–8. Canada, Calgary. DOI 10.2118/171588-MS.
20. Crnkovic, F. L., Erlandson, M. (2015). Geology driven EUR prediction using deep learning. *SPE Annual Technical Conference and Exhibition*, pp. 1–10. Houston. DOI 10.2118/174799-MS.
21. Anderson, R. N., Boyi, X., Leon, W., Arthur, A. K., Joseph, H. F. et al. (2016). Petroleum analytics learning machine to forecast production in the wet gas marcellus shale. *SPE/AAPG/SEG Unconventional Resources Technology Conference*, pp. 1–16. San Antonio. DOI 10.15530/URTEC-2016-2426612.
22. Liang, Y., Zhao, P. D. (2019). A machine learning analysis based on big data for eagle ford shale formation. *SPE Annual Technical Conference and Exhibition*, pp. 1–23. Calgary. DOI 10.2118/196158-MS.
23. Wang, S., Chen, S.,(2019). Insights to fracture stimulation design in unconventional reservoirs based on machine learning modeling. *Journal of Petroleum Science and Engineering, 174,* 682–695. DOI 10.1016/j.petrol.2018.11.076.
24. Xue, L., Liu, Y., Xiong, Y., Liu, Y., Cui, H. et al. (2021). A data-driven shale gas production forecasting method based on the multi-objective random forest regression. *Journal of Petroleum Science and Engineering, 196,* 1–13. DOI 10.1016/j.petrol.2020.107801.
25. Wang, H., Chen, Z., Chen, S., Hui, G., Kong, B.(2021). Production forecast and optimization for parent-child well pattern in unconventional reservoirs. *Journal of Petroleum Science and Engineering, 203,* 1–10. DOI 10.1016/j.petrol.2021.108899.
26. Alwated, B., El-Amin, M. F. (2021). Enhanced oil recovery by nanoparticles flooding: From numerical modeling improvement to machine learning prediction. *Advances in Geo-Energy Research, 5(3),* 297–317. DOI 10.46690/ager.

27. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29,* 1189–232. DOI 10.1214/aos/1013203451.
28. Huber, P. J. (1992). Robust estimation of a location parameter. In: Kotz, S., Johnson, N. L. (Eds.), *Breakthroughs in statistics*, pp. 492–518. New York, NY, USA: Springer. DOI 10.1007/978-1-4612-4380-9_35.
29. Breiman, L., Friedman, J. H., Olshen, R. A., Stone., C. J. (1986). Classification and regression trees. *Journal of the American Statistical Association, 81(393),* 253. DOI 10.2307/2288003.
30. Breiman, L.(1997). Prediction games and arcing algorithms. *Neural Computation, 11(7),* 1493–1517. DOI 10.1162/089976699300016106.
31. Wang, X. D., Liu, C. Q. (1999). Productivity analysis on commingled production wells. *Oil Drilling & Production Technology, 21(2),* 56–61.
32. Hou, R., Liu, Z. (2012). Resevoir evaluation and development strategies of daniudi tight sand gas field in the Ordos Basin. *Oil & Gas Geology, 33(1),* 118–128.
33. Helaleh, A. H., Alizadeh, M. (2016). Performance prediction model of miscible surfactant-CO_2 displacement in porous media using support vector machine regression with parameters selected by ant colony optimization. *Journal of Natural Gas Science and Engineering, 30,* 388–404. DOI 10.1016/j.jngse.2016.02.019.
34. Ebrahimi, A., Khamehchi, E. (2016). Developing a novel workflow for natural gas lift optimization using advanced support vector machine. *Journal of Natural Gas Science and Engineering, 28,* 626–638. DOI 10.1016/j.jngse.2015.12.031.

Appendix

In order to present the data set more clearly, this section shows the features' scatter diagrams with feature weight greater than 2%.

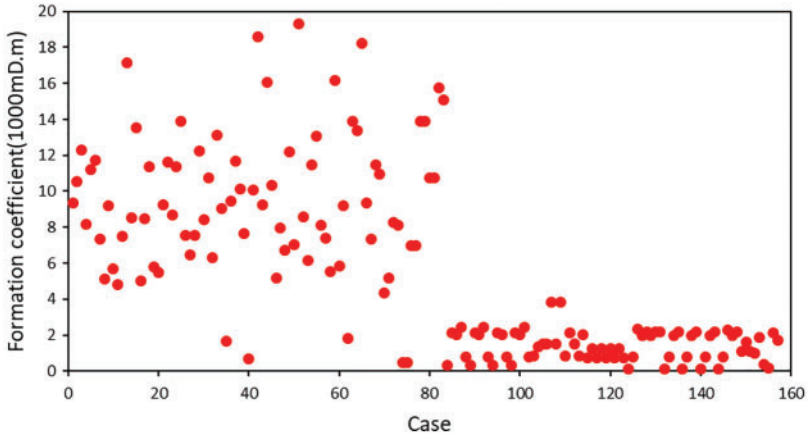

Figure 6: Scatter plot of formation coefficient

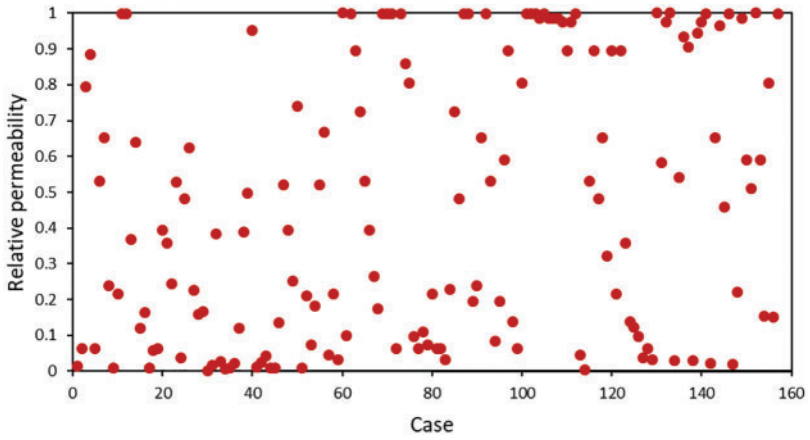

Figure 7: Scatter plot of relative permeability

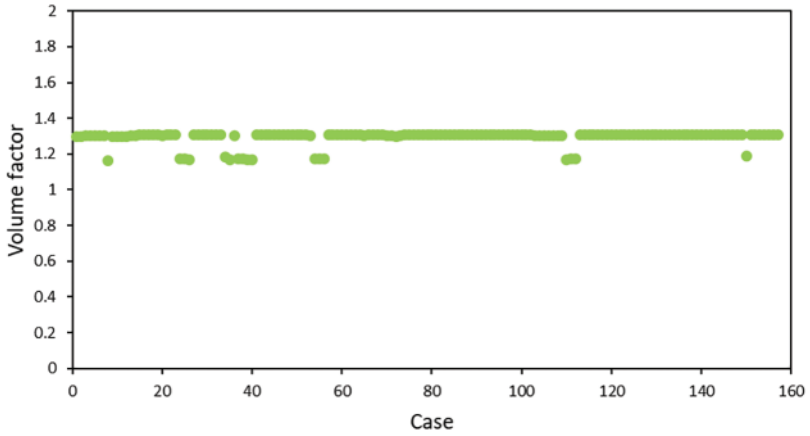

Figure 8: Scatter plot of volume factor

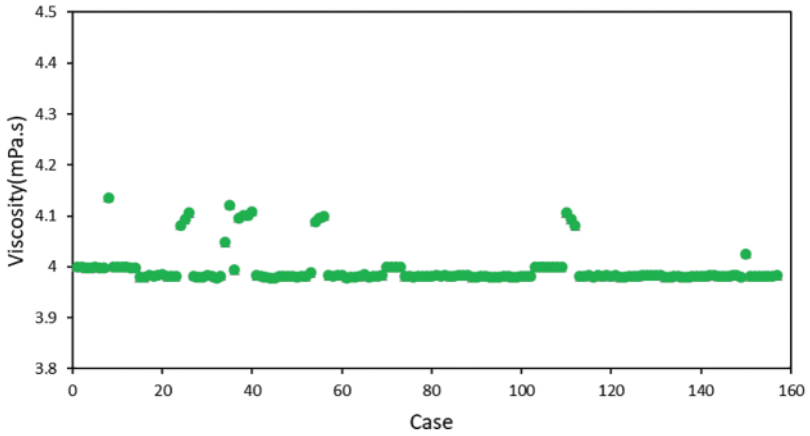

Figure 9: Scatter plot of viscosity

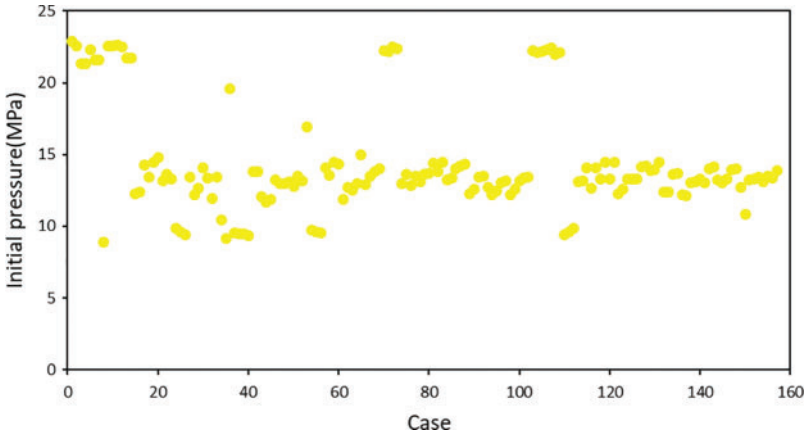

Figure 10: Scatter plot of initial pressur

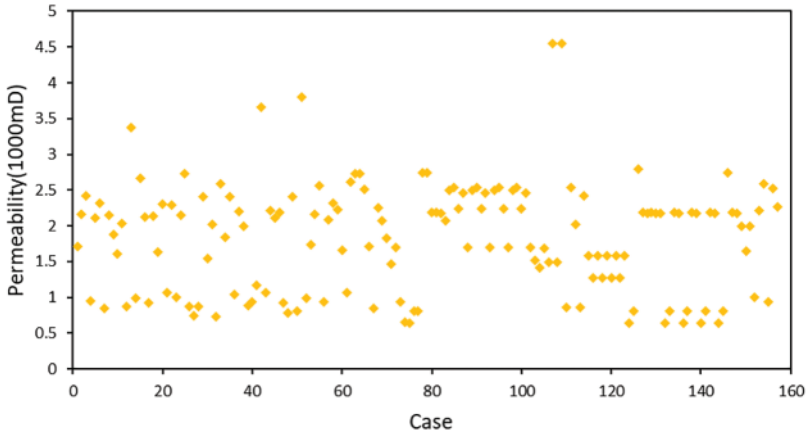

Figure 11: Scatter plot of formation permeability

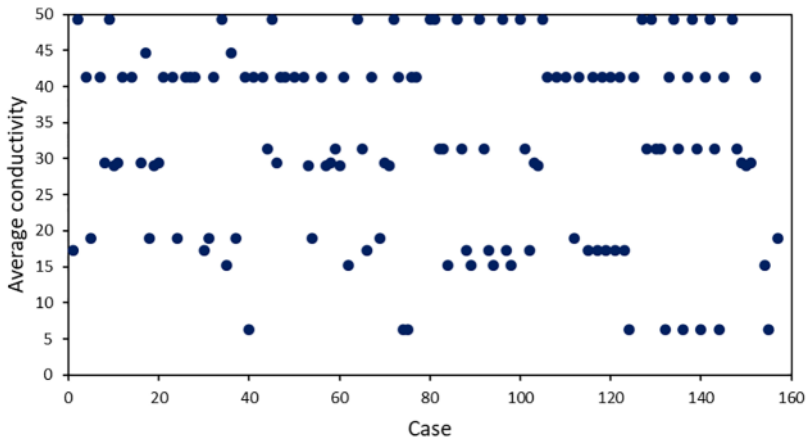

Figure 12: Scatter plot of average conductivity

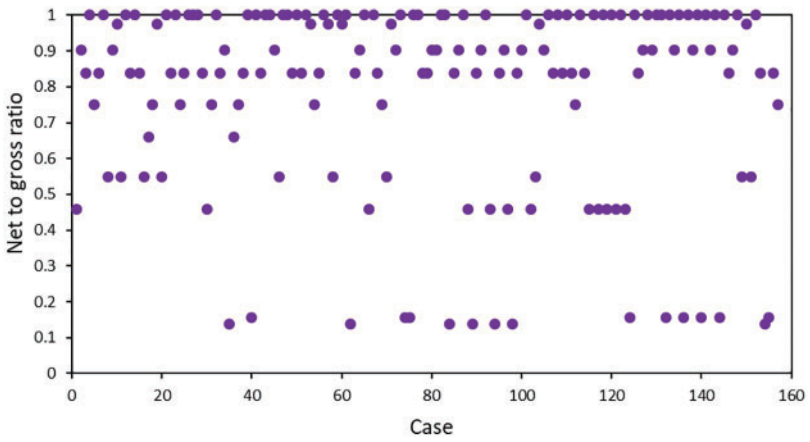

Figure 13: Scatter plot of net to gross ratio