processes

MDPI

Article

Explainable Machine Learning-Based Method for Fracturing Prediction of Horizontal Shale Oil Wells

Xinju Liu [1,2], Tianyang Zhang [3], Huanying Yang [1], Shihao Qian [3], Zhenzhen Dong [3], Weirong Li [3,*], Lu Zou [3], Zhaoxia Liu [4], Zhengbo Wang [4], Tao Zhang [2] and Keze Lin [2]

[1] Petrochina Changqing Oilfield Company, Xi'an 710021, China; liuxinju2022@163.com (X.L.); yanghy20232023@outlook.com (H.Y.)
[2] State Key Laboratory of Oil and Gas Resources and Exploration, China University of Petroleum (Beijing), Beijing 102249, China; zhangt20232023@outlook.com (T.Z.); keze.lin@hotmail.com (K.L.)
[3] Ptroleum Engineering Department, Xi'an Shiyou University, Xi'an 710065, China; zty16223334@gmail.com (T.Z.); ltxh990111@163.com (S.Q.); dongzz@xsyu.edu.cn (Z.D.); zoulu2409033593@gmail.com (L.Z.)
[4] Research Institute of Petroleum Exploration and Development, PetroChina, Beijing 100083, China; zhaoxliu@163.com (Z.L.); wangzhengbo@petrochina.com.cn (Z.W.)
* Correspondence: weirong.li@xsyu.edu.cn

Abstract: Hydraulic fracturing is a crucial method in shale oil development, and predicting production after hydraulic fracturing is one of the challenges in shale oil development. Conventional methods for predicting production include analytical methods and numerical simulation methods, but these methods involve many parameters, have high uncertainty, and are time-consuming and costly. With the development of shale oil development, there are more and more sample data on the geological parameters, engineering parameters, and development parameters of shale oil hydraulic fracturing, making it possible to use machine learning methods to predict production after hydraulic fracturing. This article first analyzes the impact of different parameters on initial production and recoverable reserves based on field data from Chang-7 shale oil in the Ordos Basin of China. Then, using the Particle Swarm Optimization (PSO) algorithm and the Gradient Boosting Decision Tree (GBDT) algorithm, machine learning models for initial production and recoverable reserves are established. The Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive exPlanations (SHAP) explanation methods are used to explain the models. The study found that initial production is highly correlated with parameters such as the number of fracturing stages and fracturing fluid volume, while recoverable reserves are significantly related to parameters such as well spacing, area, and reserver-controlled. The PSO-GBDT model established in this study has an accuracy of over 85% and can be used for production prediction and subsequent parameter optimization research. By comparing the LIME and SHAP local explanation methods, it is shown that different explanation methods can obtain reasonable and credible local explanation results. This article establishes a high-precision shale oil well production prediction model and two model interpretation methods, which could provide technical support for shale oil well production prediction and production analysis.

Keywords: shale oil; hydraulic fracturing horizontal wells; machine learning productivity prediction model; PSO-GBDT; explainable algorithm

Citation: Liu, X.; Zhang, T.; Yang, H.; Qian, S.; Dong, Z.; Li, W.; Zou, L.; Liu, Z.; Wang, Z.; Zhang, T.; et al. Explainable Machine Learning-Based Method for Fracturing Prediction of Horizontal Shale Oil Wells. *Processes* **2023**, *11*, 2520. https://doi.org/10.3390/pr11092520

Academic Editor: Qingbang Meng

Received: 11 July 2023
Revised: 8 August 2023
Accepted: 15 August 2023
Published: 22 August 2023

1. Introduction

With the further development of shale oil extraction in the United States, many countries around the world have followed the trend of shale oil development and have achieved corresponding results. China has relatively limited crude oil resources and high external dependence, so efficient development of unconventional energy sources, mainly shale oil,

is one of the most effective ways for China to alleviate its energy shortage and improve its energy structure [1].

Hydraulic fracturing is a very important production enhancement method in shale oil development [2]. The fracturing technology of horizontal wells in different development stages is a necessary condition for achieving efficient development of shale oil [3]. Reasonable fracturing parameters are one of the key factors determining the success or failure of unconventional oil reservoir development.

Forecasting production of fractured horizontal wells and optimizing fracturing parameters is one of the frontier topics in reservoir numerical simulation [4]. The fracturing parameter optimization model is generally based on the production equation to optimize the fracturing parameters. Fractured horizontal well production forecasting methods can be roughly divided into two categories: the first category is conventional forecasting methods, mainly including mathematical model-based analytical methods, and numerical simulation methods based on seepage theory. Analytical methods mainly use mathematical methods to solve established production formulas. Numerical simulation methods mainly use numerical simulation software to establish a model, and predict the production based on history fitting. However, in the process of completing traditional fracturing production evaluation and forecasting research using analytical and numerical simulation methods, the analysis of the factors affecting production is an indispensable and important step. Computer simulation can be used to simulate the design of reservoir physical properties, fracturing construction and transformation parameters, and the production dynamic parameters [5]. However, the relationships between these parameters are complex and not a simple function. Traditional reservoir evaluation simulate models are complex to calculate and take a long time, resulting in inaccurate parameter simulation. It is difficult to make a clear and concise functional relationship characterization of the above parameters, which leads to the inability of models established by traditional methods to complete the tasks of fracturing production prediction and parameter optimization [6]. The second category is production prediction methods based on machine learning. For the production data collected from the oilfield, machine learning algorithms are used to establish corresponding prediction models to achieve production prediction. Applying machine learning methods can efficiently solve the problem of data preprocessing, greatly improving the quality of on-site data analysis. Based on machine learning methods, it is possible to diagnose the main controlling factors of production and predict single-well production after fracturing, directly optimizing the design of on-site construction parameters, and improving single-well production.

With the widespread application of hydraulic fracturing in unconventional oil and gas reservoirs, it has become possible to obtain a large amount of geological parameters, fracturing engineering parameters, development parameters, and productivity parameters [7]. The application of machine learning-based unconventional hydraulic fracturing horizontal well productivity main control factor analysis, the productivity prediction model, and fracturing optimization model have been increasingly used [8].

For example, Fan Yilong et al. [9] collected data on construction conditions and production aspects from 800 wells in the Sudong gas field to establish a dataset, and then used the multiple linear regression algorithm to analyze various parameters and their standard errors and develop a parameter optimization plan based on this. Wang Hongliang et al. [10] used oilfield production history data, considering production indicators and their influencing factors and production changes, and constructed a production forecasting model using long short-term memory neural networks (LSTM) to predict production during the high water-cut period in the oilfield. Compared with fully connected neural network (FCNN) models and traditional water flooding methods, this model produced more accurate predictions. Costa et al. [11] trained an artificial neural network system and applied neural network models and genetic algorithms to mimic high-fidelity numerical models to successfully predict oil well production, solving the problem of historical fitting. Luo et al. [12] used random forests, Lasso regularization, and other methods to analyze the main factors affecting production, selecting the six most important variables from

geological parameters, reservoir modification parameters, and production parameters. They established an annual oil production prediction model based on a four-layer deep artificial neural network and completed parameter sensitivity evaluation through the model, optimizing production, and completing big data analysis of about 2000 hydraulic fracturing horizontal wells in the Bakken shale oil field. Liang and Zhao [13] established a production model based on the random forest method using data from 1069 hydraulic fracturing horizontal wells in the Eagle Ford shale gas reservoir.

Table 1 summarizes some examples of current domestic and foreign research on fracturing parameter optimization [12,14–19]. From a survey of the current domestic and foreign research status, it is found that machine learning methods related to fracturing parameter optimization are mainly focused on shale gas, with relatively few studies on shale oil, and even fewer studies on machine learning-based fracturing parameter optimization for Chinese shale oil reservoirs. This paper proposes a machine learning-based production forecasting method for shale oil development and conducts interpretability analysis on the model to clarify the main control factors of each well's production and the impact of each parameter on the production of different wells.

Table 1. Current status of research.

Time	Author	Data	Methods	Importation	Objectives	Research Block	Accuracy
2016 [14]	Esmaili	3700	Data-driven technology	Well locations, trajectories, static data, completion, hydraulic fracturing data, production rates, and operational constraints	Forecast well production	Marcellus Shale Oil and Gas, Southwestern Pennsylvania	97.18%
2019 [12]	Luo G, Tian Y	2061	Random Forest, Recursive Feature Elimination, Lasso Regularization Analysis	Formation pressure, porosity, reservoir thickness, TOC, thermal maturity, and brittleness	Forecast production	Bakken Shale Oil	60.00%
2020 [15]	Duplyakoz	5500	Hydraulic Fracturing Database	Formation parameters, well structure, field and layer IDs, all HF design parameters	Predicted production, fracturing design optimization	Data on 22 oil fields in Western Siberia, Russia	64.80%
2020 [16]	Wu Hua	137	RF, BP, XGBoost	Formation parameters, reservoir parameters, and fracturing parameters	Production, Fracturing parameter optimization	Weiyuan block	79.00%
2020 [17]	Li Juhua et al.	196	RF	Formation parameters, reservoir parameters	Predicted gas well production	Fuling Shale Gas Field	72.30%
2021 [18]	Yan Ziming	186	XGBoost, DNN, SVR	Formation parameters, reservoir parameters	Predicted recovery	Fuling Shale Gas Field	85.30%
2022 [19]	Ma Xianlin, Zhou Desheng, and Cai Wenbin	598	ANN, SVM, RF, GBDT SHAP Explanation	Formation parameters, reservoir parameters, and fracturing parameters	Horizontal well prediction, model interpretation	Surig Gas Field East	Train: 67.00% Test: 58.00%

This paper builds on the previous research and uses the parameters of the Chang 7 shale oil fracturing horizontal well and fracturing construction parameters as input for the machine learning model, with the production parameter as the model output, to carry out machine learning model training. The PSO combined with machine learning algorithm is used to establish and optimize the initial production and estimated ultimate recovery model. At the same time, we use two different interpretable methods (LIME, SHAP) to analyze and explain the production prediction model.

The main structure of this paper is as follows: Section 1 introduces the research status of machine learning-based unconventional hydraulic fracturing horizontal well productivity models. Section 2 introduces the machine learning algorithms, optimization algorithms, and model interpretation methods used in this study. Section 3 presents the workflow and research content of this paper. Section 4 combines the data from 89 wells in the Chang 7 shale oil formation to conduct correlation analysis, establish and optimize the initial productivity, and estimate ultimate recovery models, and uses the LIME and SHAP

methods to provide local and global explanations for the models. Section 5 concludes the paper.

2. Methodology

2.1. GBDT (Gradient Boosting Decision Tree)

GBDT is an iterative decision tree based on the boosting method, which combines gradient boosting and decision trees. By using an additive model (i.e., a linear combination of basis functions) and continually reducing the residual generated during training, it achieves the algorithm of classifying or regressing data.

The boosting tree is a machine learning algorithm that trains multiple weak learners using forward distribution algorithm, where each weak learner is constructed using the CART regression tree. These weak learners are then combined using an additive model to form a strong learner [20]. In GBDT, there is a connection between each weak classifier. The next weak classifier in GBDT is trained using the gradient of the loss function of the previous weak classifier, so that each iteration moves towards the direction of reducing the loss, ultimately resulting in an optimal solution [21].

The boosting tree model can be represented as an additive model with decision trees as base learners, with the specific formula:

$$f(x) = f_M(x) = \sum_{m=1}^{M} h_m(x; \alpha_m) \tag{1}$$

Among them, $h_m(x; \alpha_m)$ denotes the mth decision tree, M denotes the number of base learners, and α_m indicates the parameters of the mth learner, such as the number of leaf nodes, the depth of the tree, and so on.

First initialize: $f_0(x) = 0$

The model in step m is: $f_m(x) = f_{m-1}(x) + h_m(x; \alpha_m)$

Solved by minimizing the empirical risk, i.e., the loss function:

$$\alpha_m = argmin_{\alpha_m} \sum_{m=1}^{M} L(y_i; f_m(x_i)) \tag{2}$$

$L()$ is the loss function, and the commonly used loss functions for regression are MSE, absolute loss, Huber loss, and quantile loss. The loss functions commonly used for classification are exponential loss and logarithmic loss.

GBDT can be used for regression problems and compared to logistic regression, which can only be used for linear regression, GBDT can be used not only for linear regression and nonlinear regression, but also for dichotomous problems (set the threshold value, and greater than the threshold value is a positive case, and vice versa is a negative case) [11], which has strong applicability.

Advantages: high accuracy, both for classification and regression tasks; can handle nonlinear data; can handle both discrete and continuous values; and uses some robust loss functions that are insensitive to outliers. For example, the Huber loss function and quantile loss function.

Disadvantage: Since the next learner needs to fit the residuals of the previous learner, it must be executed in a walk-through fashion and cannot be parallelized, which also leads to a particularly slow speed when processing large data. Therefore, the later XGBoost and LightGBM are both based on GBDT but with improvements in parallelism.

2.2. PSO (Particle Swarm Optimization)

Particle Swarm Optimization (PSO), also known as the bird swarm algorithm, is an evolutionary algorithm [12]. Similar to the simulated annealing algorithm, it starts from a random solution and iteratively searches for the optimal solution by evaluating the quality of the solutions with fitness function. However, it is simpler than genetic algorithms and does not have the "crossover" and "mutation" operations. Instead, it seeks the global optimal by following the current best solution found. PSO is a parallel algorithm [13] that uses massless particles to simulate birds in a flock, with particles having only two properties: velocity and position. The velocity represents the speed of movement, while the particle represents the direction of movement [15]. This algorithm is highly valued for its ease of implementation, high precision, and fast convergence, and has demonstrated its superiority in solving practical problems.

The main process of the algorithm is as follows:

Step 1: Set the initial positions and velocities of the particles randomly, and also set the iteration times.

Step 2: Calculate the fitness value of each particle.

Step 3: For each particle, compare its fitness value with the fitness value of the best position it has ever experienced, and if it is better, set it as its current personal best position.

Step 4: For each particle, compare its fitness value with the fitness value of the global best position experienced by all particles, and if it is better, set it as the current global best position.

Step 5: Use the velocity and position formula to optimize the velocity and position of the particles, updating their positions.

Step 6: If the termination condition (usually the maximum number of iterations or the minimum error requirement) is not met, return to step 2.

The process is demonstrated more graphically through Figure 1.

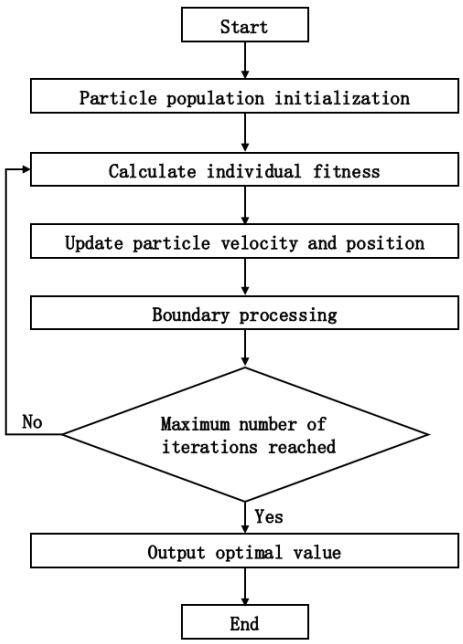


Figure 1. PSO flow chart.

Advantages: The PSO algorithm does not have crossover and mutation operations. It relies on particle velocity to perform the search, and only the optimal particle transmits information to other particles during the iterative evolution, resulting in a fast search speed. The PSO algorithm has memory, and the historical best position of the particle swarm can be remembered and transmitted to other particles. The number of parameters to be adjusted is small, the structure is simple, and it is easy to implement in engineering. The use of real

number encoding is determined directly by the solution of the problem, and the number of variables in the problem solution is directly used as the dimension of the particle [22].

Disadvantages: The PSO algorithm lacks dynamic adjustment of the particle speed, which can lead to falling into local optima and result in low convergence accuracy and difficulty in convergence. It cannot effectively solve discrete and combinatorial optimization problems. The algorithm requires parameter tuning, and how to choose suitable parameters for different problems to achieve optimal results is a challenge. PSO cannot effectively solve some problems described by non-rectangular coordinate systems [22].

2.3. Machine Learning Model Interpretability

Interpretability refers to the degree to which the reasons for the output decisions can be understood. Model interpretability refers to the understanding of the internal mechanisms of the model and the interpretation of the model results [23]. The higher the interpretability of a machine learning model, the easier it is to understand the reasons for the decisions or predictions obtained.

When solving machine learning problems, data analysts tend to focus on model performance indicators such as accuracy, precision, and recall. However, these metrics only explain a part of the model's predictive decisions. Over time, due to concept drift caused by various factors in the environment, the performance may change. Therefore, understanding what factors lead a model to make certain decisions is extremely important. Interpretability mainly refers to understanding the features, classification, and prediction indicators, and then understanding why a machine learning model makes certain decisions and which features play the most important role in the decision-making process. This helps us determine whether the model makes sense or not. Interpretability provides more transparency, explains why a model makes certain decisions, and can help us establish a certain level of trust in these machine learning models over time.

Interpretability mainly has the following characteristics:

(1) Importance: Understanding "why" can help us gain a deeper understanding of the problem, data, and reasons why the model may fail.

(2) Classification: Interpretability of data before modeling, interpretability of the model during the modeling stage, and interpretability of the results during the running stage.

(3) Scope: Global interpretability, local interpretability, model transparency, model fairness, and model reliability.

(4) Evaluation: Intrinsic or post-hoc, model-specific or model-agnostic, local or global.

(5) Features: Accuracy, fidelity, usability, reliability, robustness, and universality.

(6) Human-friendly interpretation: The degree to which humans can understand the reasons for the decision-making and the degree to which people can continuously predict the model results.

The relationship between the interpretability of a model and its predictive power is shown in Figure 2, where the weaker the predictive power of the model, the more likely it is to be interpreted.

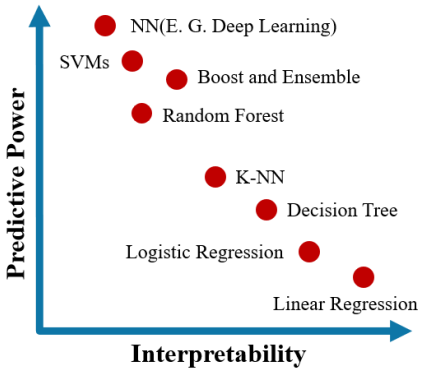


Figure 2. Classification of model interpretability.

Model interpretability can be divided into two categories: intrinsic interpretability and post-hoc interpretability. Intrinsic interpretability means that the model, such as linear models, parameter models, or tree-based models, is interpretable by its nature. Post-hoc interpretability means that interpretable methods, such as feature importance or partial dependence plots, are applied to black-box models (ensemble methods or neural networks) after they are trained. Depending on the scope of interpretability, it can be divided into global interpretability and local interpretability. Global interpretability refers to the interpretation between the entire model from input to output, which can obtain general rules or statistical inferences and understand the influence of each feature on the model. Local interpretability explains how the prediction result will change when the input value of a sample or a group of samples changes [24]. Different model interpretability methods have specific interpretability scopes, such as the LIME method that is mainly used for local interpretability, and the SHAP model, which can be used for both local and global interpretability, as well as for the interaction between different parameters.

2.3.1. LIME (Local Interpretable Model-Agnostic Explanations)

LIME was proposed by Marco Ribeiro and others in 2016 as a tool to help us understand how complex black box models make decisions. It is a model-agnostic machine learning interpretability method that can be applied to explain any type of model, including neural networks, XGBoost, random forests, and more. It can also be applied to various types of data, including tabular data, text data, image data, and so on [25].

LIME has three main characteristics:

(1) It only provides local explanations for the model, not global explanations, and explains each sample locally.

(2) It constructs simple interpretable models locally to predict and explain important features.

(3) It explains the relationship between the current input features and the predicted result, without including the abstract features generated during complex model training.

Based on the principle of the LIME algorithm, Figure 3 shows the specific flow schematic of the LIME model.

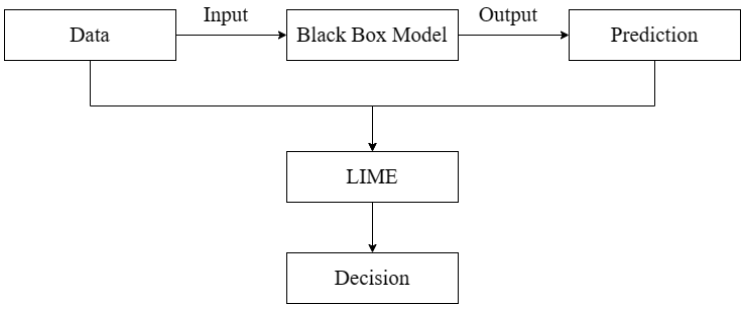


Figure 3. Schematic diagram of the LIME model.

For the black box model that needs to be explained, a sample point of interest is selected, and new samples are generated by perturbing it. Depending on the scope of the area around the sample point of interest, samples within the area are selected, and the black box model is used to predict their values. A new dataset is thus obtained, and a linear model is trained on it, which provides a good local approximation of the black box model. By using an interpretable model, we can understand the local decision-making behavior of the black box model [26].

Denote $x \in R^d$ the sample to be interpreted, first, select the more important d' dimensional features, and x becomes $x' \in R^{d'}$ after removing the unimportant components.

A new sample point, z', is generated by perturbation near x', and the new sample points forms a new dataset, Z'. The sample points are recovered into samples $z \in R^d$ after

adding the removed feature components. $\pi_x(z)$ is defined as the similarity of the samples before and after perturbation, which is calculated as follows:

$$\pi_x(z) = \exp\left(-\frac{D(x,z)^2}{\sigma^2}\right) \tag{3}$$

where $D(x,z)$ is the distance formula, which will be defined differently for different sample types, usually L2 parametric distance when it is image data, and cosine similarity when it is text data.

Denoting f as the complex model to be explained and g as the simple model, the objective function to measure the difference between the two models is shown below:

$$\xi(x) = \sum_{z,z'} \pi_x(z)\big(f(z) - g(z')\big)^2 2 + \Omega(g) \tag{4}$$

where $\Omega(g)$ is the complexity of model g. When g is a line regression model, the model complexity is the number of weight coefficients is not zero.

$E[f(x)]$ represents the expectation of all samples $f(x)$, which is the mean value of the model prediction on the incoming data set. $f(x)$ represents the 0th sample, and the magnitude of the $f(x)$ value is the prediction of the 0th sample. Assuming that the i-th sample is x_i, the j-th feature of the i-th sample is x_{i-j}, the model's predicted value for that sample is y_i, and the baseline (usually the mean of the target variable for all samples) of the entire model is y_{base}, then the SHAP value obeys the following equation:

$$y_i = y_{base} + f\ (x_{i1}) + f(x_{i2}) + I + f(x_{ik}) \tag{5}$$

where $f\ (x_{i-j})$ is the SHAP value of x_{i-j}. Intuitively, $f\ (x_i, 1)$ is the contribution value of the first feature in the i-th sample to the final prediction value y_i. When $f\ (x_{i1}) > 0$, the feature boosts the prediction value and also has a positive effect; conversely, the feature gives a lower prediction value and has a negative effect. That is the contribution of each feature to the first prediction.

2.3.2. SHAP (Shapley Additive Explainable)

The SHAP model can explain the model to determine whether known conditions have a positive or negative effect on the final prediction [27]. SHAP is an ex-post interpretation method, where the core idea is to calculate the marginal contribution of features to the model output and then interpret the "black box model" at both global and local levels. The relationship between the SHAP algorithm and the model is shown in Figure 4.

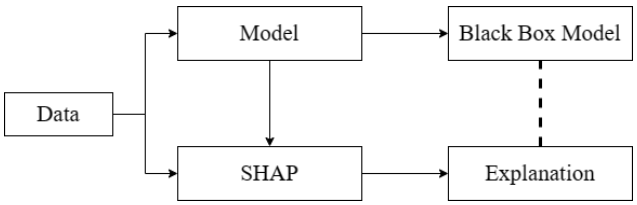


Figure 4. Schematic diagram of the Sharply model.

The SHAP model is mainly applied to calculate the feature SHAP values of individuals; the absolute value of the SHAP values of each feature of all individuals is summed or averaged to be the overall feature importance. The multicollinearity problem is solved by considering the effects of individual variables along with the effects of groups of variables and the possible synergistic effects between variables [27].

SHAP represents the SHAP value interpretation as an additive feature imputation method, and SHAP interprets the predicted value of the model as the sum of the imputed values of each input feature [27]:

$$g(z') = \varnothing_0 + \sum_{j=1}^{M} \varnothing_j z'_j \tag{6}$$

where g is the explanatory model, $z' \in \{0,1\}^M$ indicates whether the corresponding feature can be observed (1 or 0), this should be for unstructured data (e.g., text, images), and the features of each instance of structured data can be observed (including missing, which is also information about the feature being observed). M is the number of input features that, $\varnothing_j \in R$ is the imputed value (Shapley value) for each feature, and $\varnothing_0$ is the constant that explains the model (this value is actually $E_X(\hat{f}(X))$ from the previous introduction of Shapley values, i.e., the predicted mean of all training samples). The input to the tree model must be structured data, and for the instance x, z' should be a vector of all values of 1, i.e., all features that can be observed, so the formula simplifies to:

$$g(x') = \varnothing_0 + \sum_{j=1}^{M} \varnothing_j \tag{7}$$

The SHAP value is calculated by defining $f_x(S) = E(f(x)|x_s)$, where S is the set of possible subsets of the input features (the union v mentioned by the Shapley value) and $E(f(x)|x_s)$ is the conditional expectation value of the subset S of the input features (the val function mentioned by the Shapley value) [27]. The predicted values are obtained by calculating $E(f(x))$, as explained in Figure 5. It can be seen that the SHAP value assigns the imputed value of each feature as the expected change in the model prediction when adjusting the value, interpreting the prediction of the model f for the samples $\{x_1 = a_1, x_2 = a_2,\ x_3 = a_3, x_4 = a_4\}$ as the sum of the effects, $\varnothing_j$, of each feature that introduces the conditional expectation.

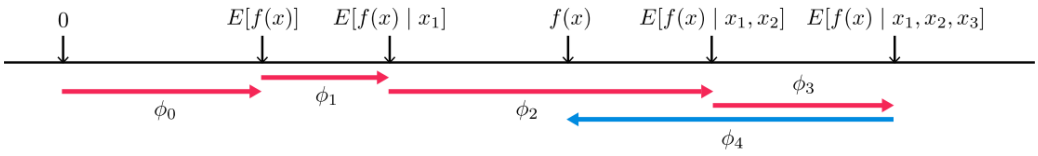


Figure 5. Four input variables to calculate the SHAP method.

3. Workflow

The main steps for applying machine learning to establish a production prediction model for shale gas hydraulic fracturing horizontal wells are shown in Figure 6.

(1) Data collection: The data includes the main factors affecting production and evaluation indicators for horizontal well production. The factors affecting production mainly include geology and hydraulic fracturing construction parameters, and production parameters can be initial production, decline rate, and estimated ultimate recovery.

(2) Data preprocessing: First, data filtering is performed to remove missing data, reduce dimensionality, and perform standardization processing. Then, the processed data is divided into an 80% training set and 20% test set or 70% training set and 30% test set.

(3) Machine learning modeling: A machine learning model is built using the training data, and the accuracy of different models [28] is compared to select the optimal method for establishing a production prediction model.

(4) Production prediction model optimization: The accuracy of the production prediction model is evaluated using the test set. Based on the evaluation indicators, the optimization model is used to further improve the prediction accuracy of the model and

the machine learning production prediction model with the highest prediction accuracy is selected [29].

(5) Model interpretation: Based on the optimal production prediction model, LIME and SHAP methods are used to perform global and local explanations of production predictions.

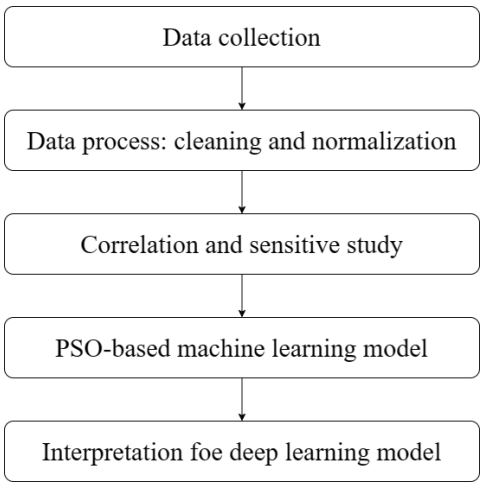


Figure 6. Hydraulic fracturing model interpretation and optimization process.

4. Results and Discussion

4.1. Work Zone Overview

The Ordos Basin is located at the junction of Shanxi, Gansu, and Ningxia, with its reservoir structure belonging to the combination of the east–west structures and part of the Paleozoic of the North China Basin. The study area (a shale oil reservoir of Chang 7 in the Ordos Basin) shown in Figure 7 is located in the secondary structural unit of the Yishan Slope, which is in the southwest of the Ordos Basin. The study area has abundant reserves and broad development prospects, but the overall development level is relatively low. As of the end of 2020, the predicted reserves of the block were 3.13 million tons, and the remaining untapped third-level reserves were 5.20 million tons (predicted to be 3.05 million tons).

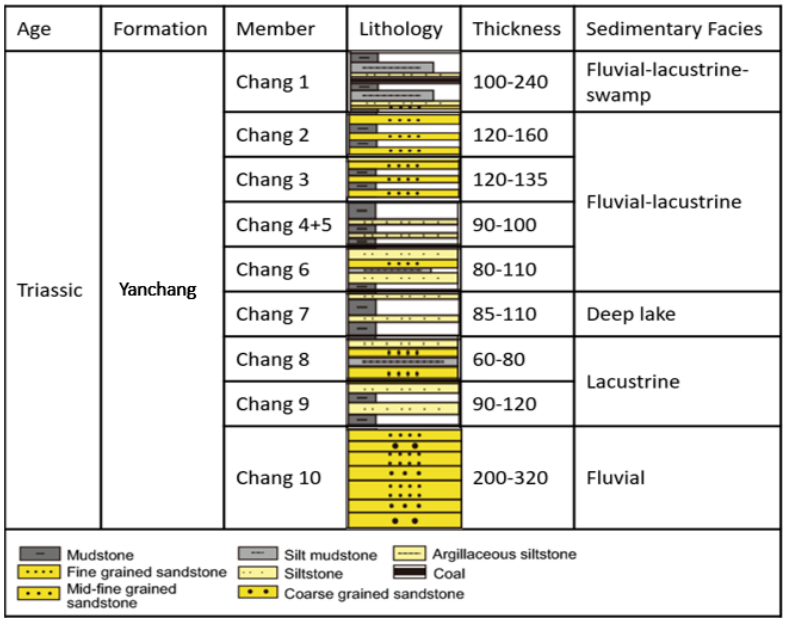


Figure 7. Ordos basin location, studied area, and regional stratigraphy [30].

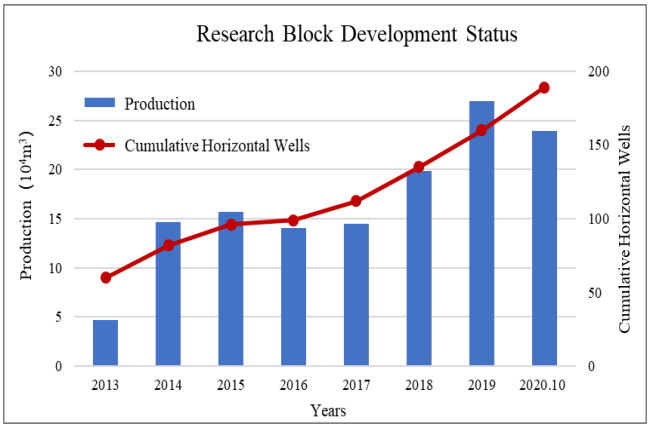
From 2011 to 2012, horizontal well pilot tests were carried out in the Chang 7 oil reservoir in Longdong, and two horizontal wells (NP1 and GP43-57) were put into production that year. The length of the horizontal section was 1021.5 m, and the horizontal well encounter rate of the oil layer was 92.7%. The initial daily oil production per well was 79.164 STB, with a comprehensive water content of 41.9%. In 2013, the Z230 area began to develop and build production using a combined deployment of horizontal wells and directional wells. Mainly using a five-point and seven-point well network, the average length of the horizontal section was 813 m, and the horizontal well encounter rate of the oil layer was 92.7%, with a planned production of 2,199,000 STB. Starting from 2014, the long horizontal section and natural energy development were deployed, with an average length of 1100 m, and a planned production of 901,590 STB.

Currently, the reservoir heterogeneity in the study area is strong, and the effectiveness of fracturing varies greatly. Low-liquid or high-water cut wells with low efficiency have appeared in some areas, which hinders the overall efficient development of the oilfield.

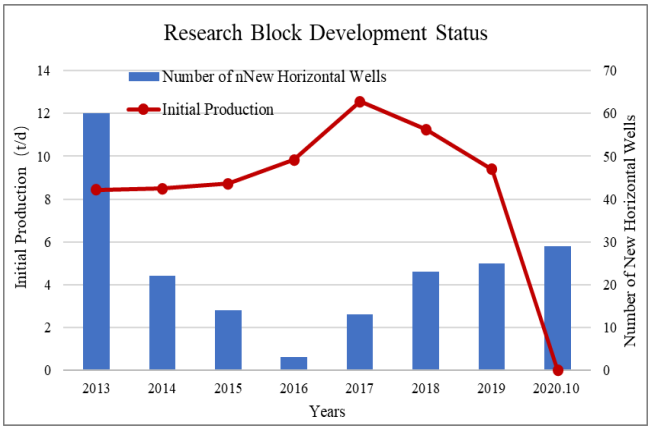
As shown in Figure 8, the number of horizontal wells has increased year by year from 2013 to 2020, and the production increased significantly between 2013 and 2015. The production was relatively low in 2016 and 2017, and increased significantly from 2018 to 2019, but decreased by October 2020.

As shown in Figure 8, the number of newly opened horizontal wells decreased year by year from 2013 to 2016, and the average initial production gradually increased. From 2016 to October 2020, the number of newly opened horizontal wells increased year by year, and the initial production reached its peak in 2017 before gradually decreasing.

As shown in Figure 9, the trend changes in the parameters of horizontal well fracturing in the study area from 2013 to 2018 are shown. The scale of fracturing has been increasing year by year in recent years.



(**a**)



(**b**)

Figure 8. Map of horizontal well data in the study block: (**a**) Plot of production and total number of horizontal wells over years; (**b**) Plot of initial production and total number of New horizontal wells over years.

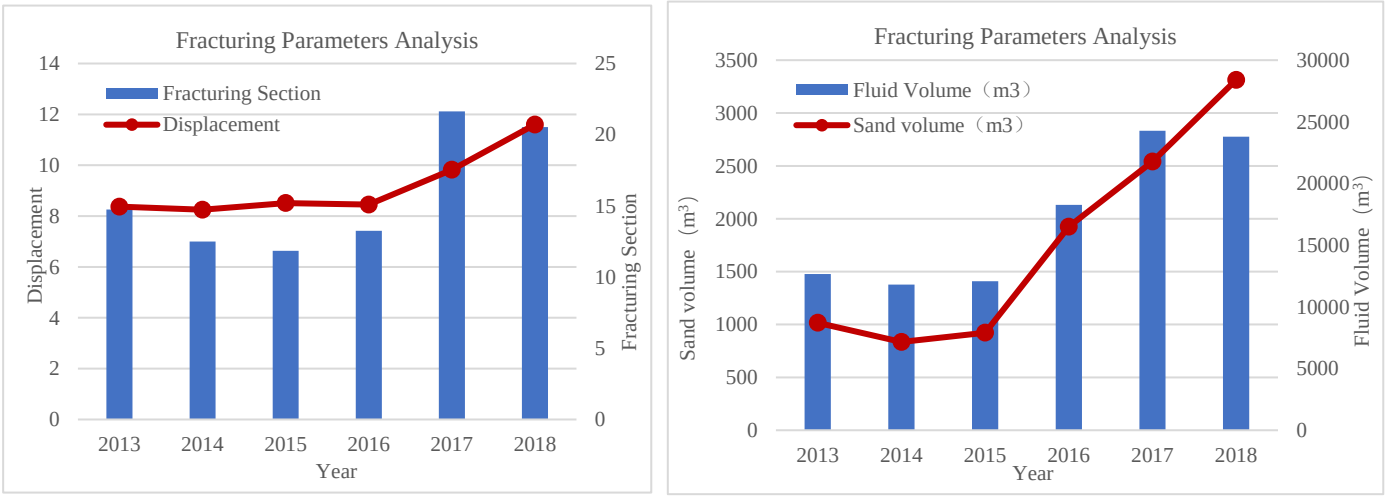


Figure 9. Trend of fracturing construction parameters: (**a**) Plot of displacement and fracturing section over years; (**b**) Plot of sand volume and fluid volume over years.

Based on the current situation, machine learning algorithms are used to analyze the factors affecting initial production and estimate ultimate recovery, and to determine the influencing parameters and their degree of influence. This can help to understand the trends in production and make more accurate predictions for future production in the study area.

4.2. Data Analysis

4.2.1. Data Collection and Analysis

The collected raw data mainly includes 89 segmented multi-cluster fractured horizontal wells in the Chang 7 shale oil reservoir in the Ordos Basin. Among them, 14 geological parameters and 8 engineering parameters were used as the input for production evaluation, and the output are initial production and estimated ultimate recovery (EUR). The specific parameters are shown in Table 2.

Table 2. Productivity impact parameters.

Type		Parameters
Input Parameters	Geological parameters	Well Distance, Row Spacing, Area, Reserves Abundance, Controlled Reserves, Oil Layer Length, Poor Oil Layer, Drilling Encounter Rate of Oil Layer, RT (Resistivity), AC (Acoustic time difference), SH (Shale volume), Φ (Porosity), K (Permeability), So (Oil saturation)
	Engineering Parameters	Horizontal Section, Fracturing Section, Single Well Ground Fluid Volume, Single Well Sand Proportion, Single-stage Sand Volume, Sand Ratio, Single-stage Volume
Output parameters	Production Dynamic Parameters	Initial Production, EUR (Estimated ultimate recovery)

Descriptive analysis of the data is an important step in the machine learning modeling process. Graphical descriptions were used to analyze the data from the screened 89 horizontal wells, and histograms of the distribution of capacity impact parameters and target parameters are shown in Figure 10. The distribution zones of specific geological and engineering parameters are shown in Tables 3 and 4.

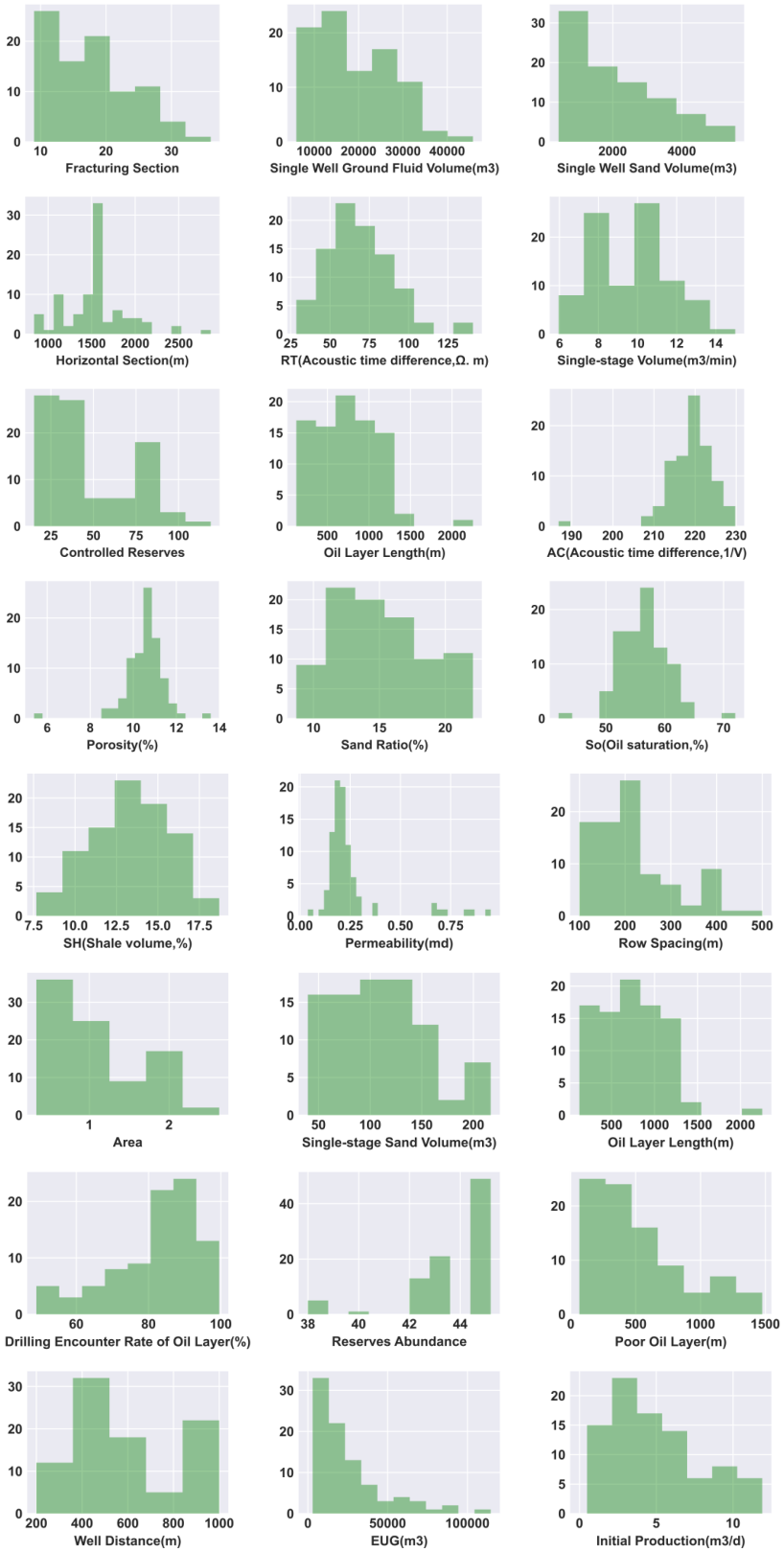


Figure 10. Field data distribution.

Table 3. Geological parameters distribution.

Parameters	Distributions		
	Percentage of Wells	Spilt Point	Percentage of Wells
RT	20%	50 Ω·m	80%
Controlled Reserves	60%	50×10^8 m^3	10%
Oil layer length	60%	1000 m	60%
AC value	80%	220 1/V	20%
Porosity	25%	0.1	75%
So	7%	0.5	93%
SH	35%	0.15	65%
Permeability	99%	0.5 md	1%
Area	99.5%	2	0.5%
Drilling encounter rate of oil layer	35%	80%	65%
Reserves Abundance	50%	44	50%
Poor oil layer	85%	1000	15%

Table 4. Engineering parameters distribution.

Parameters	Distributions		
	Percentage of Wells	Spilt Point	Percentage of Wells
Fracturing section	70%	20	30%
Single well ground fluid volume	60%	20,000 m^3	40%
Single well sand proportion of wells	58%	2000 m^3	42%
Horizontal section	90%	2000 m	10%
Single-stage volume	48%	10 m^3/min	52%
Sand ratio	77%	20%	23%
Row spacing	55%	200 m	45%

Target parameters: Recoverable reserves are mainly below 50,000 m^3, and 13% are larger than 50,000 m^3. Initial production is unevenly distributed between 0 and 10 m^3/d, with a few wells larger than 10 m^3/d.

4.2.2. Correlation Analysis

As shown in Figure 11, for the analysis of the linear relationship between the model input parameters and EUR, the linear regression R^2 refers to the correlation coefficient, which responds to the proportion of the total variation in the dependent variable that can be explained by the independent variable through the regression relationship; the larger the R^2, the higher the correlation, and when R^2 is 0, it indicates a non-linear correlation between the two. EUR was nonlinearly correlated with SH and showed linear correlations with other characteristic parameters in general, among which row spacing, permeability, porosity, and area were strongly correlated, and the drilling encounter rate of the oil layer and reserves abundance were weakly correlated.

This article applies the Pearson correlation coefficient method to reflect between different parameters and target parameters. The Pearson correlation coefficient is a statistical measure used to reflect the degree of linear correlation between two variables. The cor-

relation coefficient is represented by r, where n is the sample size, and the observations and means of the two variables are taken into account. The absolute value of r describes the degree of linear correlation between two variables, where a larger absolute value of r indicates a stronger correlation. The absolute value of r is defined as the quotient of the covariance and standard deviation between two feature variables. Its scope of application is as follows: the two features have a linear relationship and both are continuous data; the observations of the two features appear in pairs and are independent of each other. The correlation strength between different parameters changes with the range of the correlation coefficient, where 0.8–1.0 indicates an extremely strong correlation, 0.6–0.8 indicates a strong correlation, 0.4–0.6 indicates a moderate correlation, 0.2–0.4 indicates a weak correlation, and 0.0–0.2 indicates a very weak or no correlation.

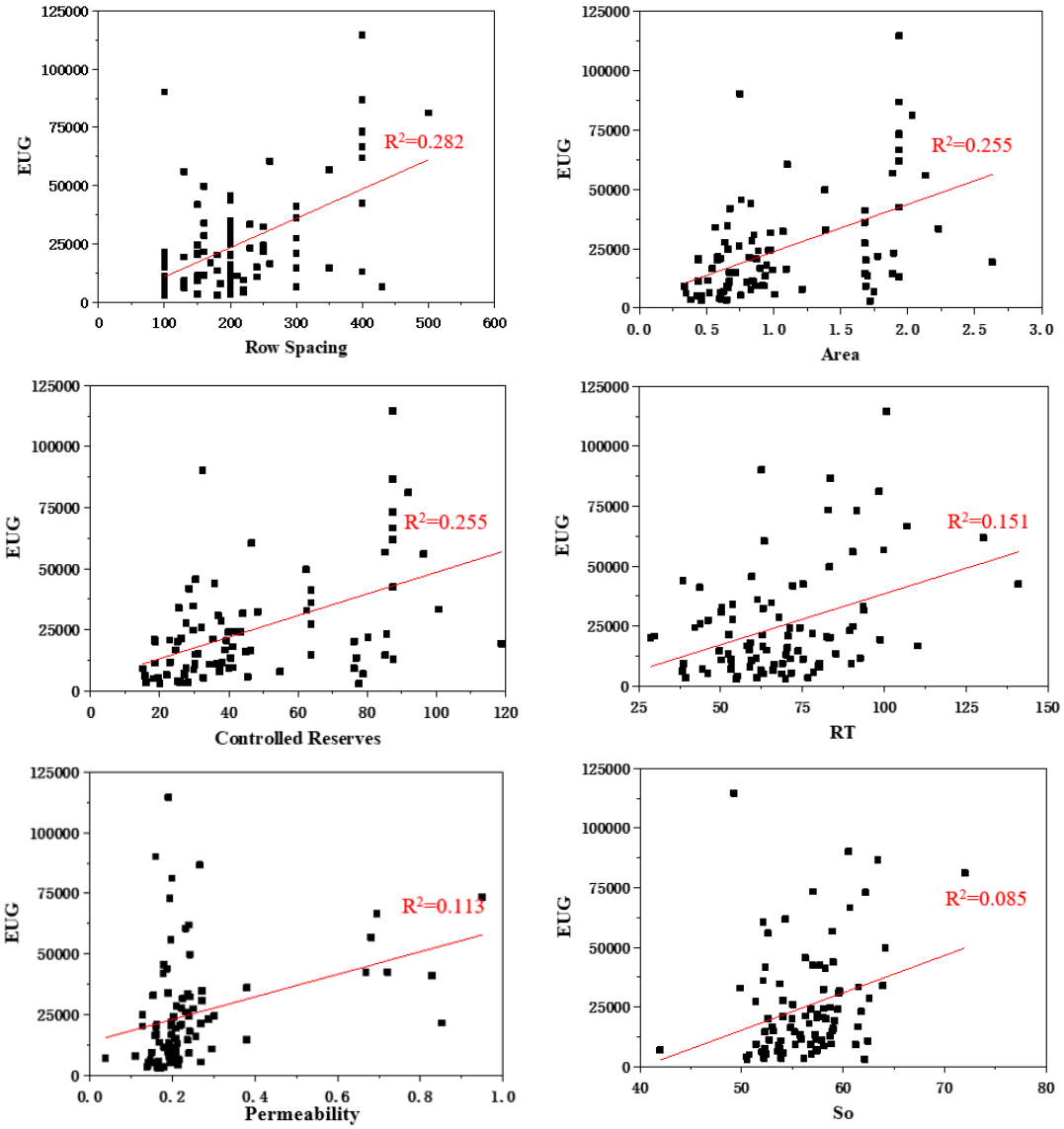


Figure 11. *Cont.*

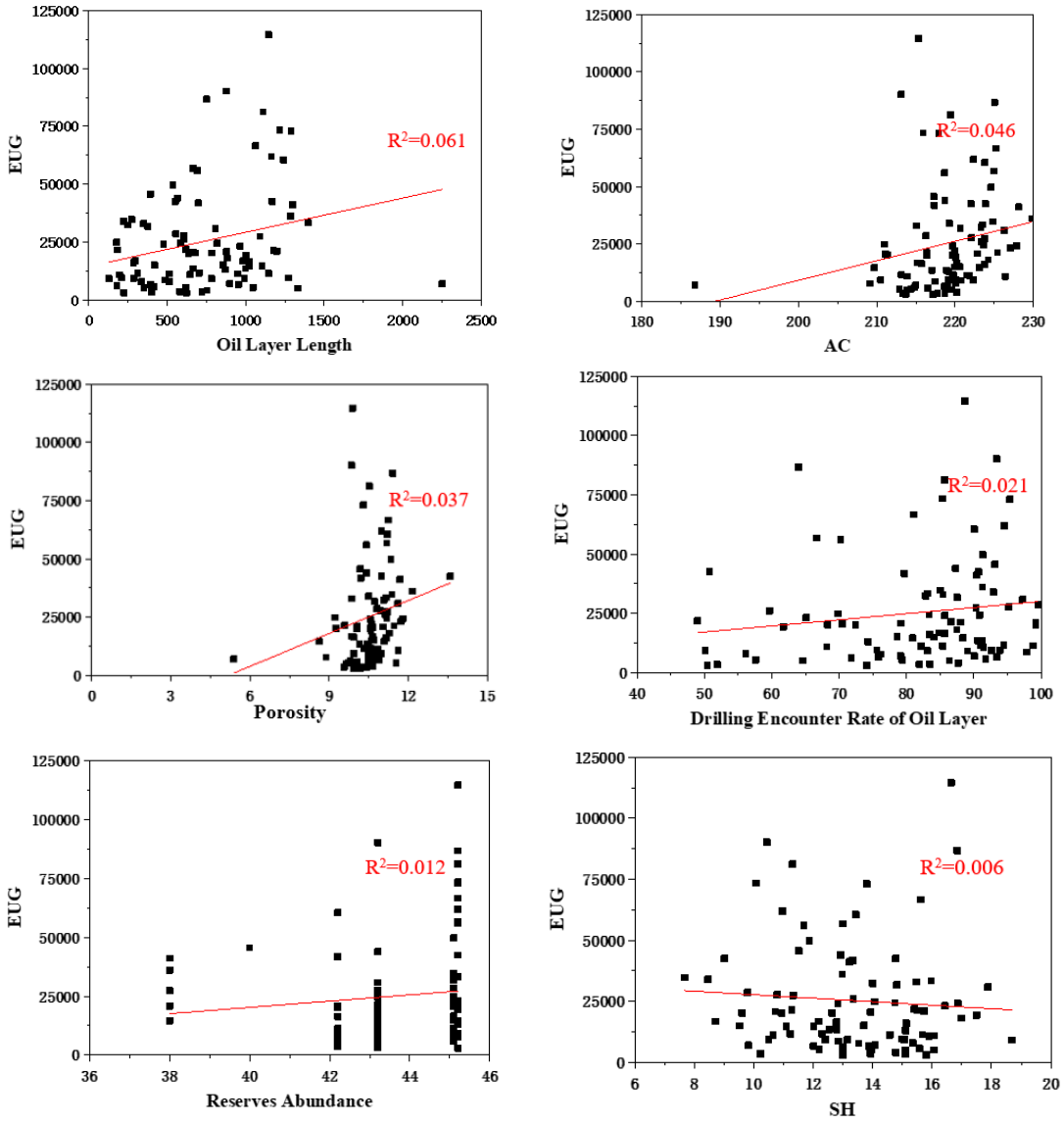


Figure 11. Correlation analysis between input parameters and EUR.

As shown in Figure 12, the geological parameters of initial production are moderately correlated with the horizontal section and single well ground fluid volume, weakly correlated with single well sand proportion, fracturing section, and single-stage volume, and very weakly correlated with the single-stage sand volume and sand ratio. The engineering parameters of estimated ultimate recovery are weakly correlated with resistivity and controlled reserves, and very weakly correlated with the oil layer length, acoustic time difference, porosity, So, SH, and permeability, indicating that the initial production is strongly correlated with engineering parameters.

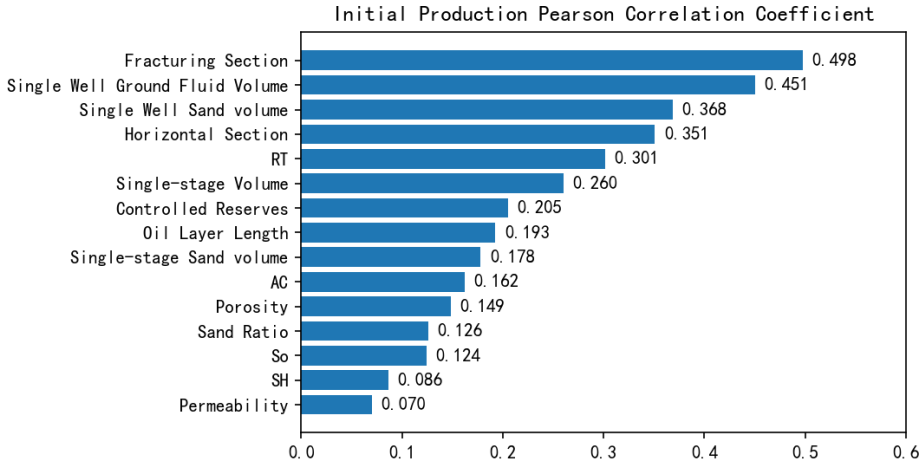


Figure 12. Initial production Pearson correlation coefficient.

As shown in Figure 13, the geological parameters of the estimated ultimate recovery are moderately correlated with the area and controlled reserves, weakly correlated with resistivity, permeability, So, oil layer length, and acoustic time difference, very weakly correlated with porosity, drilling encounter rate of oil layer, and reserves abundance, and negatively correlated with SH. The engineering parameter of estimated ultimate recovery is moderately correlated with row spacing, indicating that the estimated ultimate recovery is strongly correlated with geological parameters.

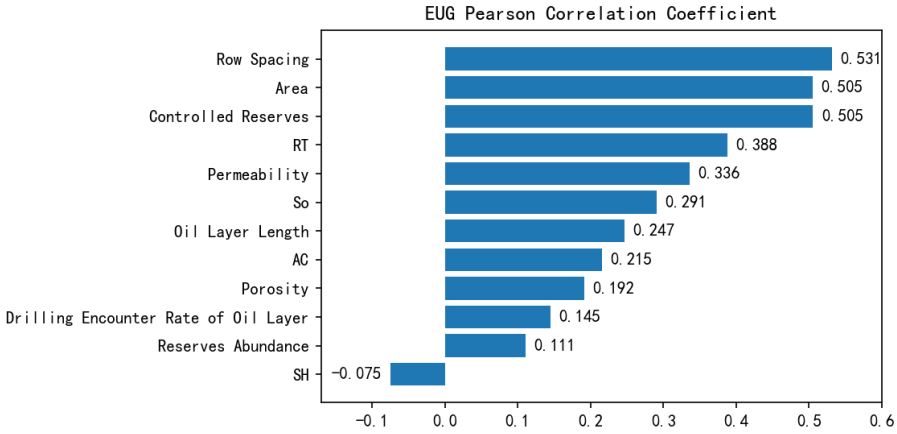


Figure 13. EUR Pearson correlation coefficient.

4.3. Machine Learning Model Building

After data preprocessing and correlation analysis, selected geological and engineering parameters used as input for the machine learning model, and the corresponding initial production and estimated ultimate recovery were used as output for model training.

To ensure the largest possible coverage and no overlap between the training and testing sets, the preprocessed data was randomly divided into a testing set (80%) and a training set (20%). Three different machine learning models, XGBoost, LightGBM, and gradient boosting decision trees, were established, and the PSO algorithm was used to optimize the hyperparameters of the models.

The fitness function represents the accuracy of the model, and the fitness value was optimized for EUR prediction after 100 iterations. The PSO algorithm mainly optimized four parameters: N-estimators, learning rate, maximum depth, and Alpha. The number of N-estimators refers to the number of decision trees, which is the number of data evaluations and has a monotonic effect on the accuracy of the model. The larger the N-estimators, the better the model, but after the N-estimators reach a certain level, the accuracy of the model

does not increase, and overfitting may occur. The value of the learning rate needs to be set within a certain range. A high learning rate can lead to unstable learning, while a too small learning rate can increase training time. The maximum depth of the decision tree can be applied to high-dimensional and low-sample-size data. Whether to increase the depth or not should be judged based on the results. Alpha is the weight of the L1 regularization term, which can be used to speed up the calculation in high-dimensional situations. Table 5 shows the range and optimal values of the four hyperparameters of the PSO algorithm used in the experiment.

Table 5. The hyperparameters of PSO used in the experiment.

		N_Estimators	Learning Rate	ax_Depth	Alpha
Default value		100	0.3	6	0.9
Value ranges		10–1000	0–1	1–100	0.5–0.95
Optimal values	XGB	122	0.11	22	0.80
	Light GBM	453	0.19	14	0.33
	GBDT	848	0.30	2	0.51

Figures 14 and 15 show the comparison of prediction accuracy of initial production and estimated ultimate recovery models established by GBDT, XGB, and Light GBM after PSO optimization. The GBDT method improved the accuracy of the initial production and estimated ultimate recovery models by 10.8% and 18% after PSO optimization. The XGB models improved by 23.8% and 7.7%. The Light GBM models improved by 20.7% and 31.1%. The accuracy of all three models improved to varying degrees, with the training set accuracy of the optimized initial production models being 99.7%, 96.1%, and 100%, and the testing set accuracy being 86%, 76.7%, and 71.7%. For the estimated ultimate recovery models, the training set accuracy was 98.7%, 99.7%, and 99.9%, and the testing set accuracy was 90.1%, 77.5%, and 57.9%. Overall, the PSO-GBDT production prediction model had the best accuracy after optimization.

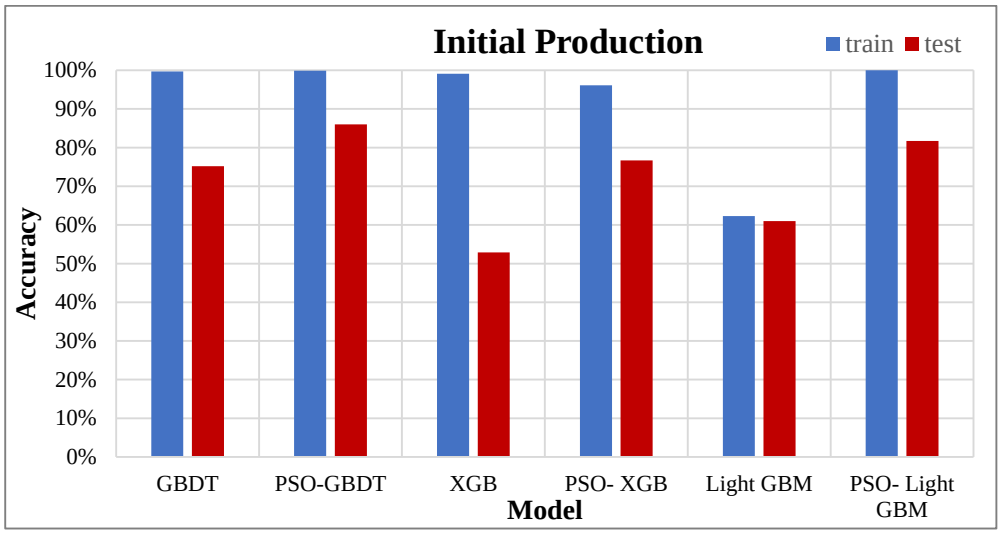


Figure 14. Initial production model PSO optimization comparison.

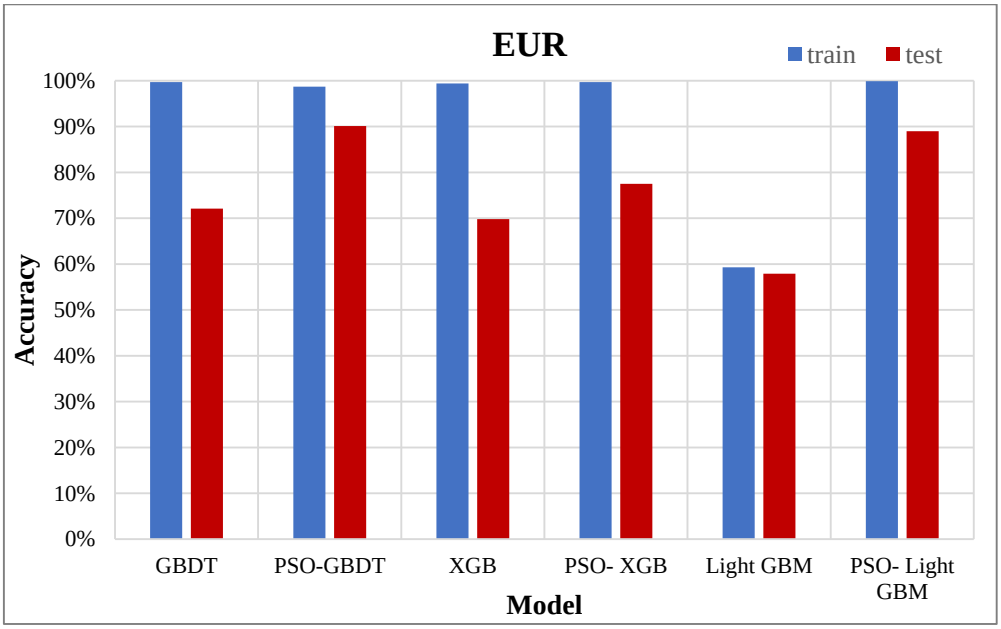


Figure 15. EUR model PSO optimization comparison.

According to Figure 16, the accuracy of the GBDT model before and after PSO optimization is shown. The horizontal axis represents the true value, and the vertical axis represents the predicted value. The accuracy of the initial production model training set is 99.7%, and the test set accuracy is 86.0%. The fitting accuracy of the estimated ultimate recovery model training set is 98.7%, and the test set accuracy is 90.1%.

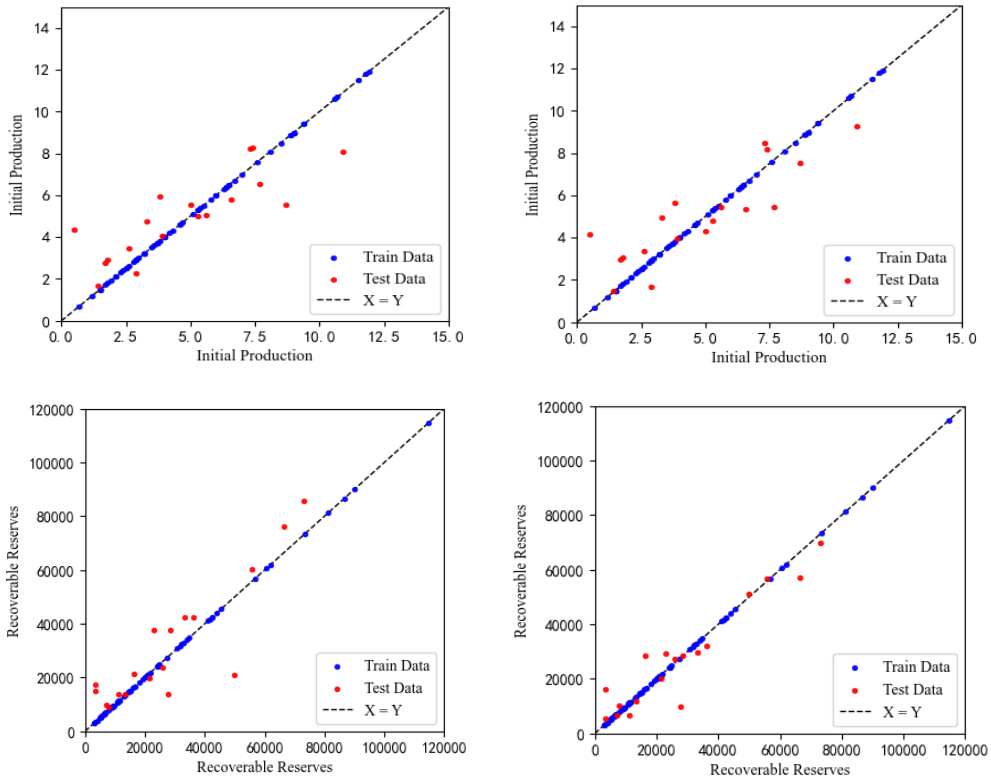


Figure 16. PSO-GBDT model training schematic.

4.4. Explanation of the Forecasting Model

4.4.1. Local Explanation

(1) LIME

The LIME model is a local interpretable model that trains on the production forecasting model dataset. This model is weighted based on the proximity between the generated and the real. Each feature is individually perturbed, and data is sampled from a normal distribution that is determined by the mean and variance of the feature.

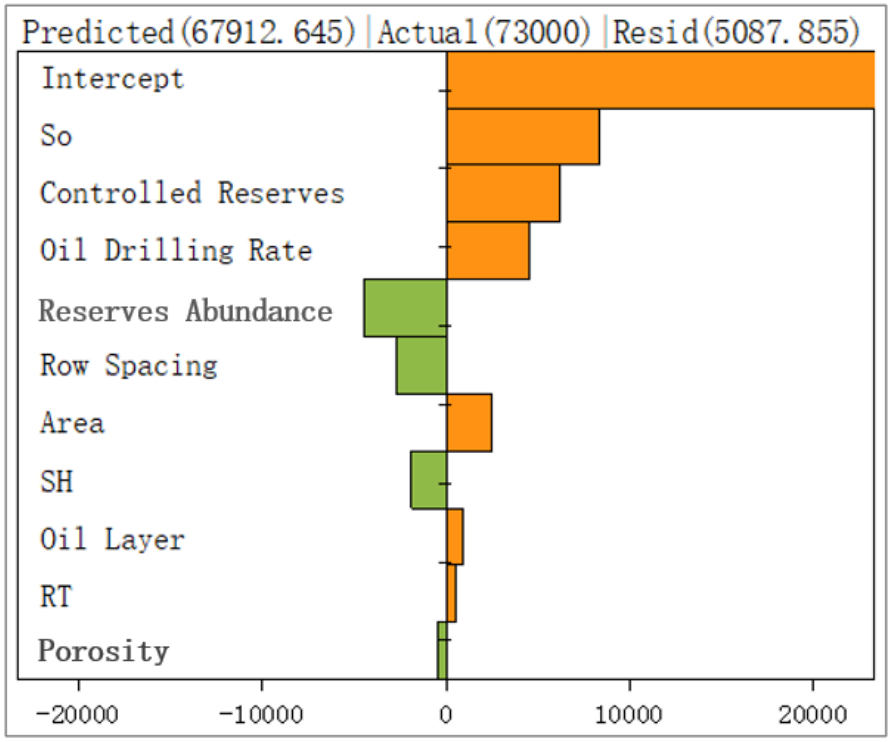
Based on the PSO-GBDT model of estimated ultimate recovery, an explanation of each feature parameter is obtained. The explanatory bar chart in Figure 17 shows the degree of influence of input parameters on different prediction results. Figure 17a shows the explanatory results of Horizontal Well 1. Controlled reserves, drilling encounter rate of the oil layer, area, reserve abundance, and RT have a positive impact on the prediction result of this feature, while reserve abundance, row spacing, SH, and porosity have a negative impact. The three parameters that have the greatest impact on this feature are So, controlled reserves, and drilling encounter rate of the oil layer. Figure 17b shows the explanatory results of Horizontal Well 54. Reserve abundance, row spacing, K, AC, and RT have a positive impact on the prediction result of this feature, while So, controlled reserve, area, SH, and porosity have a negative impact. The three parameters that have the greatest impact on this feature are reserve abundance, So, and controlled reserves.

The results show that different feature parameters have different degrees of influence on predicting EUR, and their correlations are uncertain and varied in size. This indicates that the local interpretation results of the model are random and variable.

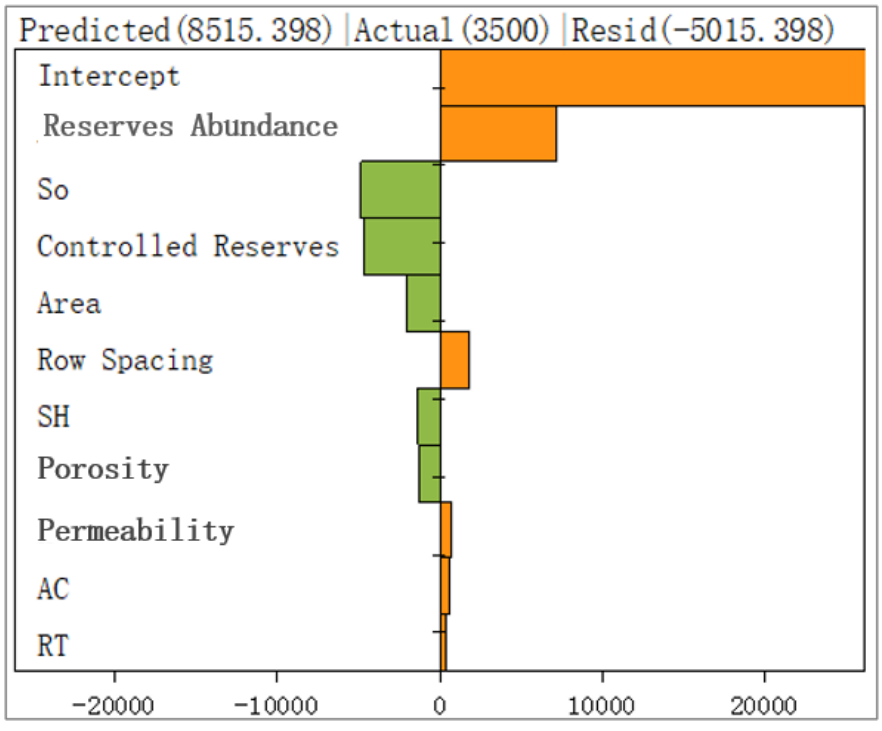
(2) SHAP

The local interpretability of SHAP focuses on explaining how a single prediction is generated by analyzing each training data point and explaining the degree to which each feature contributes to the final prediction.

Figure 18 shows the explanation of a specific prediction, where Figure 18a displays the explanation results for Horizontal Well 1. The top five parameters that have the highest impact on the prediction value are controlled reserves, area, reserve abundance, So, and RT. The red bars show to what extent an input feature increases the prediction value. Controlled reserves, oil layer length, and drilling encounter rate of the oil layer have a positive impact and increase the prediction value. For example, oil layer length has a positive impact of 2217.42 m^3 on the prediction value. The blue bars show to what extent an input feature decreases the prediction value. Area, reserve abundance, So, RT, SH, permeability, row spacing, and AC have a negative impact and decrease the prediction value to some extent. For instance, So has a negative impact of 4655.5 m^3 on the prediction value. After being influenced by multiple input features, the SHAP baseline value produces an output value, which is the mean prediction value by the model is 25,312.36 m^3.

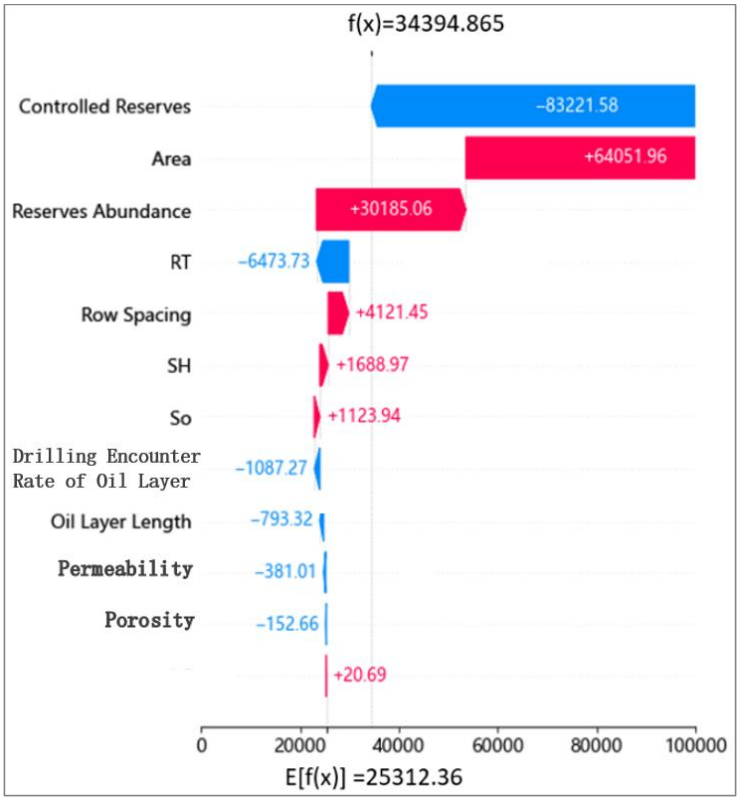


(**a**)

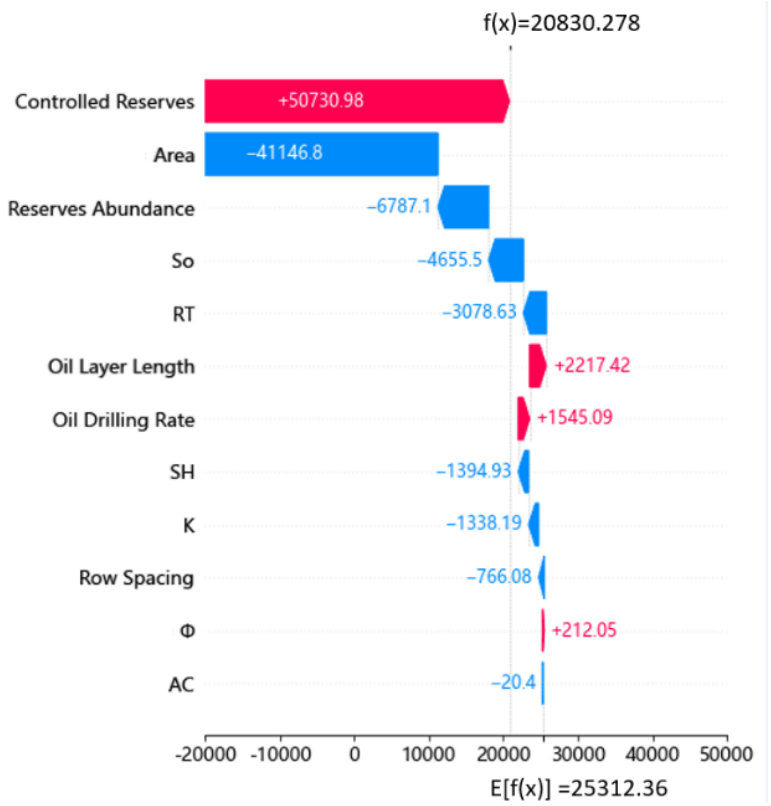


(**b**)

Figure 17. Local interpretation of recoverable reserve characteristics parameters: (**a**) Single-well LIME waterfall plot of well 1; (**b**) Single-well LIME waterfall plot of well 54.



(**a**)



(**b**)

Figure 18. Waterfall diagram for local interpretation of specific parameters: (**a**) Single-well SHAP waterfall plot of well 1; (**b**) Single-well SHAP waterfall plot of well 54.

Figure 18b shows the explanation results for Horizontal Well 54. The top five parameters that have the highest impact on the prediction value are controlled reserves, area, reserve abundance, So, and RT. Controlled reserves, oil layer length, drilling encounter rate of the oil layer, and porosity had positive effects on the predicted value and increased the predicted value. Area, reserves abundance, So, RT, SH, permeability, row spacing, and AC have a negative impact and decrease the prediction value to some extent. After being influenced by multiple input features, the SHAP baseline value produces an output value, which is the mean prediction value by the model is 25,312.36 m^3.

After comparing the results of the explanations for two Horizontal Wells, 1 and 54, using different methods, it was found that for Well 1, the top five features in terms of influence according to LIME were So, controlled reserves, reserve abundance, RT, and row spacing, while according to SHAP they were controlled reserves, area, reserve abundance, RT, and row spacing. The only different features in the local explanations between the two methods were So and area, while the other parameters were consistent. For Well 54, the top five features in terms of influence according to LIME were So, reserve abundance, controlled reserves, row spacing, and area, while according to SHAP they were controlled reserves, area, reserve abundance, So, and RT. The only different features in the local explanations between the two methods were row spacing and RT, while the other parameters were consistent.

In summary, different interpretable methods based on the same PSO-GBDT production prediction model can produce different interpretation results. In the process of computing interpretable models, the calculation method for interpreting the black box model can have different results. After comparing the results of LIME and SHAP explanations, it was found that only one feature in the top five local explanation results was different, indicating that the two methods have similar local analysis results and are credible. The difference in the influence level of local explanation results is due to the different emphasis points in local explanation. The LIME method focuses on the sorting and positive negative relationship of the influence level, while the SHAP explanation focuses on calculating the size of the SHAP value and analyzing how much impact it has on the prediction results, which can have a certain impact on the influence level sorting of the model. In this case, the fact that the influence parameters of the two methods are basically similar further demonstrates the feasibility of interpretable methods.

Figure 19 shows how the changes in the local independent relationships of feature parameters affect the model output and the distribution of the feature value. It represents the non-linear relationship between input feature parameters and estimated ultimate recovery, using porosity, permeability, and So as an example to analyze the local dependence graph.

When porosity is less than 9%, the expected model value is very low. With the increase of porosity, the predicted value rises rapidly until porosity reaches 12%, when the expected model value reaches its highest level and remains constant. When the porosity value is between 9% and 12%, the expected value fluctuates greatly, indicating that there are more data in this stage, and it has a greater impact on the expected value prediction. When permeability is less than 0.25, the predicted value is lower than the expected value. With the increase of permeability, the expected value increases in a stepwise manner. The first step is between 0.4 and 0.6, the second step is between 0.6 and 0.8, and the growth gradually slows down when permeability is greater than 0.8. When So is between 40% and 50% and greater than 65%, the expected predicted value is relatively high. This is a special case caused by the small number of horizontal wells in this range. When So is between 50% and 60%, the expected predicted value increases slowly. There is a fluctuation between 60% and 65%, indicating that the predicted value increases with the increase of oil saturation. However, due to the problem of the number of horizontal wells, there is a small-scale fluctuation.

4.4.2. Global Analysis

The SHAP method belongs to the class of post-factual explanation techniques, and its core idea is to calculate the marginal contribution of each feature to the model output, and

then explain the "black box model" from both global and local perspectives. Each predicted sample produces a prediction value, and the SHAP value is the number assigned to each feature in that sample.

Figure 20 shows the global explanation of the PSO-GBDT production prediction model, which is the ranking of the feature importance of input parameters on the model. The three parameters that have the largest impact on the estimated ultimate recovery model are controlled reserves, area, and reserves abundance, which are sorted by the absolute average value of their SHAP values. The top three parameters that have the highest impact according to the Pearson correlation coefficient are row spacing, area, and controlled reserves.

When comparing the results of the interpretable analysis using the SHAP method with those of the parameter sensitivity analysis based on the Pearson correlation coefficient for the PSO-GBDT model, it was found that the three parameters with the greatest impact on estimated ultimate recovery were the same, indicating that the SHAP interpretation results were correct. However, due to the use of the SHAP method to explain the production prediction model, the degree of influence of some parameters may change during the construction and optimization of the production model, resulting in some discrepancies between the ranking of SHAP global explanation and the results of the sensitivity analysis. Nevertheless, the fact that the SHAP global explanation can still produce results similar to those of the Pearson correlation coefficient analysis after model prediction further demonstrates the good trustworthiness of the SHAP global explanation.

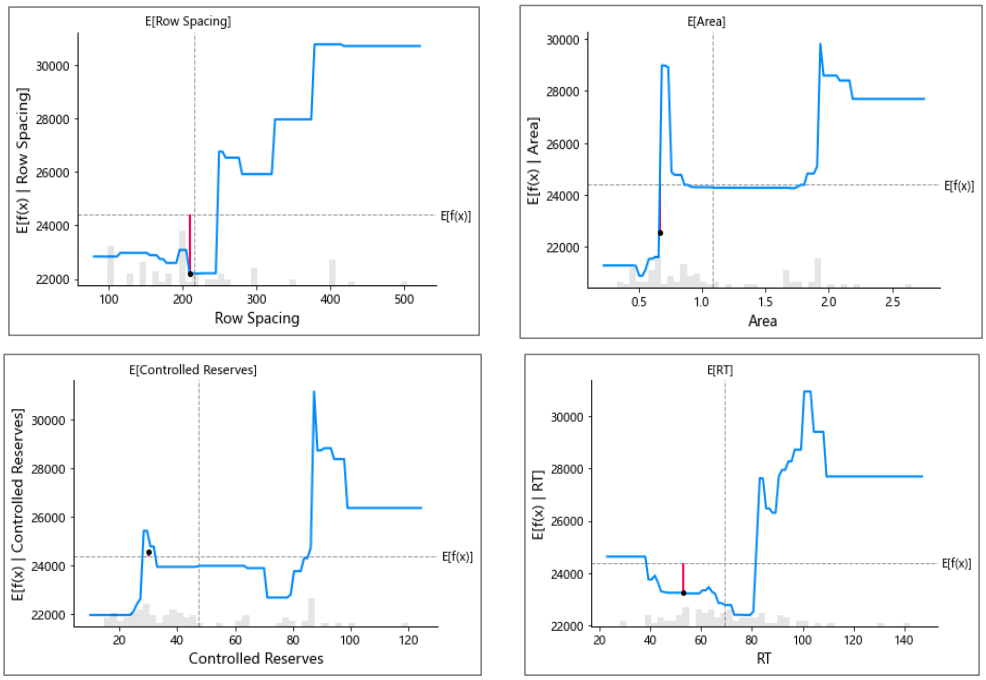


Figure 19. *Cont.*

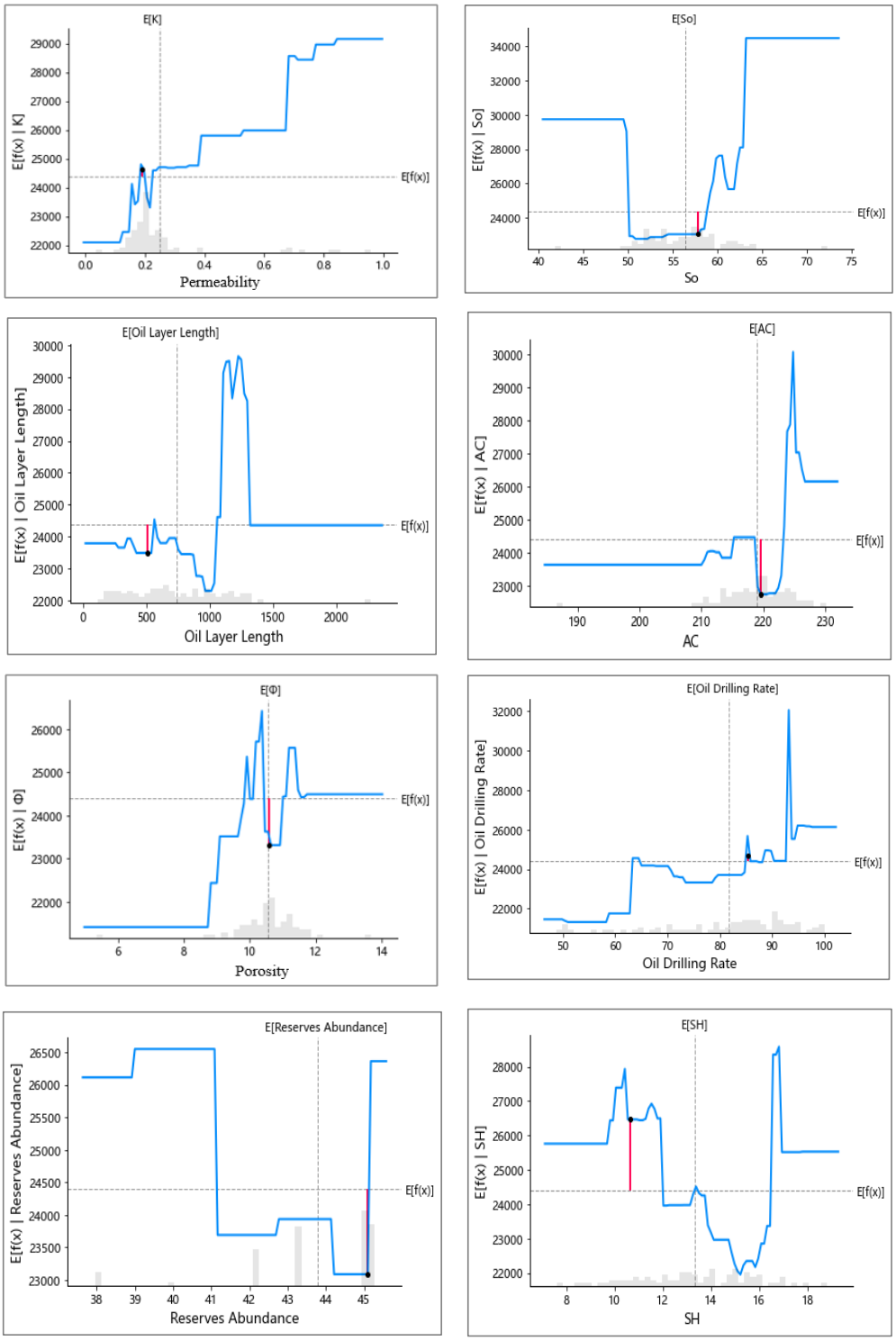


Figure 19. Partial dependence diagram of feature parameters.

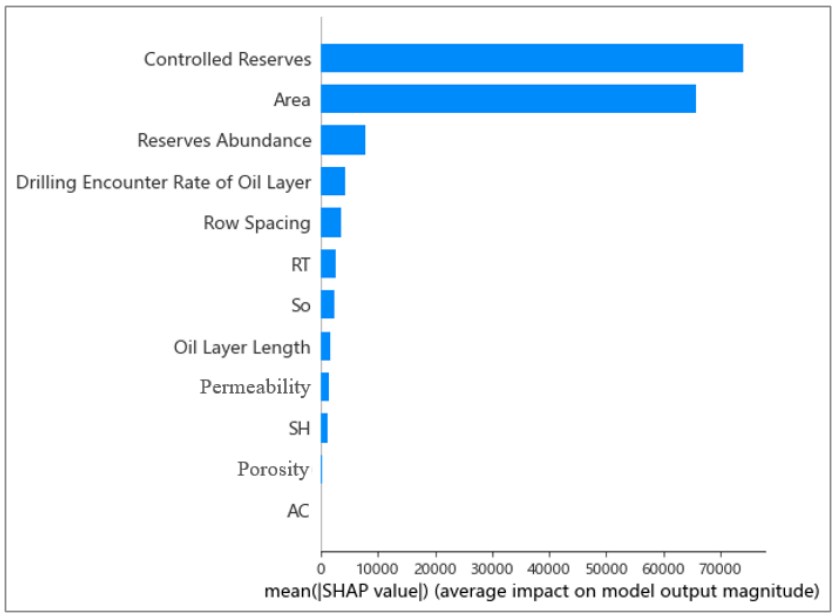


Figure 20. Ranking the importance of influencing factors.

Figure 21 shows the distribution of SHAP values for the input parameters. Each row represents a feature, and each point represents sample data, with the horizontal scale representing the SHAP value, and the features sorted according to the average absolute value of the SHAP value. Areas with more points indicate that there are a large number of samples gathered there. Each data point in the figure represents a fractured horizontal well, and the color represents the value of the variable, with blue to red indicating a low to high variable value. Red features make the prediction value larger (positive correlation), blue features make the prediction value smaller, and purple is close to the mean. Positive (negative) SHAP values indicate that the parameter has a positive (negative) correlation with estimated ultimate recovery, and the SHAP value characterizes the range of change in the influence of each parameter on estimated ultimate recovery. For example, as the controlled reserves of the reservoir increase, the SHAP value increases; as the area of the reservoir increases, the SHAP value decreases, and different permeability values result in a small range of variation in estimated ultimate recovery.

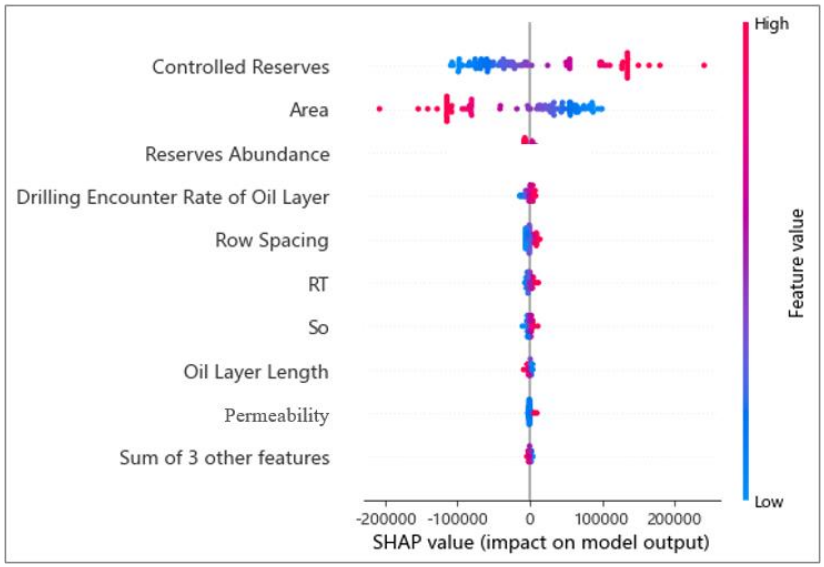


Figure 21. Distribution of the SHAP values of influencing factors.

As shown in Figure 22, the heatmap shows the distribution of clustered samples for all instances. The horizontal scale represents each instance, while the vertical scale represents the impact of each feature on that instance. The color describes the direction and strength of the impact of each feature on that instance. The samples are displayed in a regular order, such as those between 70 and 90 in the ranking, with the controlled reserves block being very red, indicating that these samples are greatly affected by controlled reserves in a positive way. Moreover, the sum of the sample's SHAP values, f(x), is also greater than the mean line. The area block gradually becomes blue, indicating that it is greatly affected by area in a negative way. In the case where controlled reserves are constant, the larger the oil-bearing area, the more difficult the reservoir development, the lower the exploitation degree, and the smaller the estimated ultimate recovery. Controlled reserves and area feature dimensions have significant SHAP sums and are high-quality samples. The horizontal scale below is the sample sequence, and the f(x) above the horizontal scale represents the deviation from the mean of the sum of each sample's SHAP values. The left side of the ordinates scale is the feature name, and the right side should be the feature importance (the sum of SHAP dimensions). The colored stripe in the middle represents the size of each sample's SHAP value for each feature.

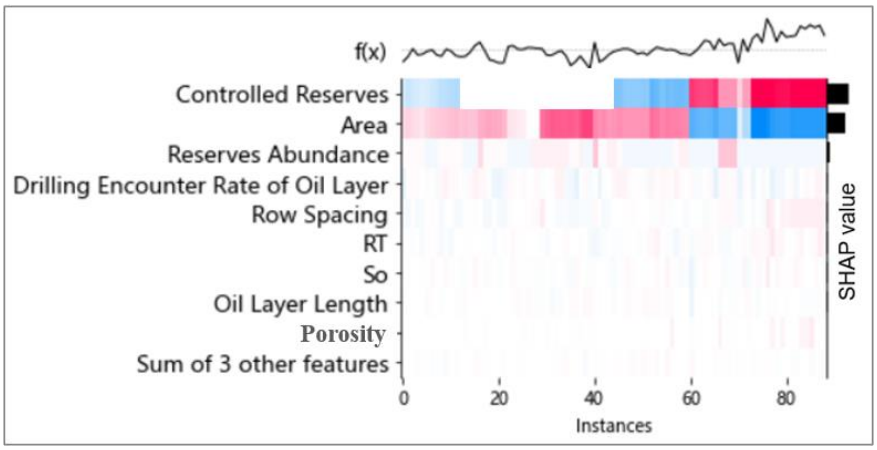


Figure 22. Heat map of feature distribution.

5. Discussion and Conclusions

5.1. Discussion

The shale oil prediction model constructed in the study can be an important tool used to assess and predict the potential reserves and production of subsurface shale oil resources. The model integrates the theories and skills of multiple disciplines, such as geology, geophysics, and engineering, to predict the production capacity of shale oil development through detailed studies of underground formation structure, mineral composition, pressure and temperature, and other factors.

Expanding to practical applications, the advantage of this research model is that it can provide a detailed depiction of underground shale oil reservoirs, which can help oil companies develop more effective exploration and development plans. In addition, this model is able to predict the difference in oil reserves between wells, thus optimizing the extraction strategy.

However, this model also has certain applicability and limitations. For example, due to the complexity and uncertainty of geological conditions, the prediction results may have certain errors. Therefore, when applying it, it is necessary to comprehensively consider the knowledge and skills of multiple disciplines and combine them with the actual situation to make reasonable corrections and improvements to the model. In addition, with the continuous progress in neural network modeling technology and the accumulation of data,

the shale oil prediction model will be continuously improved and refined to provide more accurate and reliable support for shale oil exploration and development.

At the same time, there are the environmental issues associated with extracting shale oil to ponder. In recent years, problems such as the triggering of earthquakes have led to a growing opposition to shale oil and gas extraction. Therefore, shale oil and gas development engineers should pay more attention to how to integrate with environmental engineering, balance the scale between formation development and sustainable development, and create an environmentally friendly reservoir development model.

5.2. Conclusions

Using wild data from Chang 7 formation in the Ordos Basin, this article built a machine learning model and explainable models to address the lack of shale oil well production prediction after hydraulic fracturing. Different machine learning algorithms were introduced and compared for their accuracy, and the best method was selected to establish production prediction model. The PSO optimization algorithm was used to improve the model's accuracy. On this basis, we used different methods analyze the interpretability of the model. The following conclusions were drawn:

(1) Based on the analysis of field data of shale oil hydraulic fracturing horizontal wells in the Ordos Basin from 2013 to 2018, the scale of horizontal well fracturing renovation in the study area continued to increase, showing a linear correlation with the EUR overall. The productivity of hydraulic fracturing horizontal wells is greatly affected by different parameters, and the influencing factors and their degree of impact need to be clarified.

(2) According to Pearson correlation analysis, the initial production is moderately correlated with the number of fracturing stages and fracturing injection fluid volume. It is weakly correlated with the amount of sand volume, the length of the horizontal section, the resistivity, and the controlled reserves. The estimated ultimate recovery is moderately correlated with the spacing, area, and controlled reserves, and weakly correlated with the resistivity, permeability, oil saturation, reservoir, and sonic transit time.

(3) We established different machine learning models based on the data analysis, and the GBDT model with the highest accuracy was selected. The test set accuracy of the prediction models for initial production and estimated ultimate recovery were 75.2% and 72.1%, respectively. PSO optimization algorithm was introduced to further optimize the model and improve its accuracy. The test set accuracy of the initial production and estimated ultimate recovery production prediction models based on GBDT were increased by 10.8% and 18%, respectively, after PSO optimization.

(4) LIME and SHAP were used to explanate the production for two horizontal wells, a and b. It was found that the explanation results are reasonable and can be used for other wells. Furthermore, the global explanation results of SHAP are basically consistent with those of the Pearson correlation analysis.

(5) The established model, which consisted of a high-precision shale oil well production prediction model and two model interpretation methods, could provide technical support for shale oil well production prediction and production analysis.

Author Contributions: Conceptualization, W.L.; Methodology, T.Z. (Tianyang Zhang), Z.D. and W.L.; Software, T.Z. (Tianyang Zhang) and Z.D.; Validation, T.Z. (Tianyang Zhang), L.Z. and Z.W.; Formal analysis, S.Q., W.L., L.Z. and T.Z. (Tao Zhang); Investigation, S.Q., L.Z. and Z.W.; Resources, T.Z. (Tao Zhang); Data curation, T.Z. (Tianyang Zhang) and H.Y.; Writing—original draft, T.Z. (Tianyang Zhang); Writing—review & editing, T.Z. (Tianyang Zhang) and Z.L.; Visualization, T.Z. (Tianyang Zhang) and K.L.; Supervision, X.L., H.Y. and W.L.; Project administration, X.L., Z.D. and W.L.; Funding acquisition, W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from Petrochina Changqing Oilfield Company and are available from the Correspondence author with the permission of Petrochina Changqing Oilfiled Company.

Acknowledgments: Special thanks to Instructor Professor for their careful guidance on the selection, collection, and writing of this thesis to its final draft, and to fellow lab members for their essential technical help. Thanks to Xi'an Shiyou University for cultivating and helping graduate students' innovation and practice ability, and thanks to the Petrochina Changqing Oilfield Company and State Key laboratory of oil and gas resources and exploration, China University of Petroleum (Beijing), for their support for this study in terms of data, technology, and funding.

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

References

1. Hu, S.; Zhap, W.; Hou, L.; Yang, Z.; Zhu, R.; Wu, S.; Bai, B.; Jin, X. Development potential and technical strategy of continental shale oil in China. *Pet. Explor. Dev.* **2020**, *47*, 819–828. [CrossRef]
2. Lu, Y. *Advances and Applications of Fracturing Technology in Shale Reservoirs*; Petroleum Industry Press: Beijing, China, 2016.
3. Li, Z.; Li, F.; Huang, Z. The key role of hydraulic fracturing in oil and gas field exploration and development. *Oil Gas Geol. Recovery* **2010**, *17*, 76–79+116.
4. Liang, Z. Dense cutting and multi-cluster volume fracturing technology for horizontal wells in well block 30 of Dongsheng gas field. *Pet. Geol. Eng.* **2022**, *36*, 98–101+108.
5. Ling, T. *Study on Optimization of Horizontal Well Fracturing in Shale Oil*; Northeastern Petroleum University: Daqing, China, 2022.
6. Kuangsheng, Z.H.A.N.G.; Meirong, T.A.N.G.; Liang, T.A.O.; Xianfei, D.U. Horizontal well volumetric fracturing technology integrating fracturing, energy enhancement, and imbibition for shale oil in Qingcheng Oilfield. *Pet. Drill. Tech.* **2022**, *50*, 9–15.
7. Zou, C.; Pan, S.; Jing, Z.; Gao, J.; Yang, Z.; Wu, S.; Zhao, Q. Shale oil and gas revolution and its impact. *Acta Pet. Sin.* **2020**, *41*, 1–12.
8. Jiang, T.; Wang, B.; Shan, W.; Li, A. A theoretical model for overall fracturing optimization scheme design. *J. Pet.* **2001**, *5*, 58–62+2.
9. Fan, Y.; Ma, X.; Lian, J. Research and comparison of filling methods for missing data in hydraulic fracturing. *Petrochem. Ind. Appl.* **2020**, *39*, 48–55.
10. Wang, H.; Mu, L.; Shi, F.; Dou, H. Production prediction at ultra-high water cut stage via Recurrent Neural Network. *Pet. Explor. Dev.* **2020**, *47*, 1009–1015. [CrossRef]
11. Costa, L.; Maschio, C.; Schiozer, D. Application of artificial neural networks in a history matching process. *J. Pet. Sci. Eng.* **2014**, *123*, 30–45. [CrossRef]
12. Luo, G.; Tian, Y.; Bychina, M.; Ehlig-Economides, C. Production-Strategy Insights Using Machine Learning: Application for Bakken Shale. *SPE Reserv. Eval. Eng.* **2019**, *22*, 800–816. [CrossRef]
13. Liang, Y.; Zhao, P. A Machine Learning Analysis Based on Big Data for Eagle Ford Shale Formation. In Proceedings of the SPE Annual Technical Conference and Exhibition, Calgary, AB, Canada, 30 September–2 October 2019.
14. Esmaili, S.; Mohaghegh, S.D. Full field reservoir modeling of shale assets using advanced data-driven analytics. *Geosci. Front.* **2016**, *7*, 11–20. [CrossRef]
15. Duplyakov, V.; Morozov, A.; Popkov, D.; Vainshtein, A.; Osiptsov, A.; Burnaev, E.; Shel, E.; Paderin, G.; Kabanova, P.; Fayzullin, I.; et al. Practical Aspects of Hydraulic Fracturing Design Optimization using Machine Learning on Field Data: Digital Database, Algorithms and Planning the Field Tests. In Proceedings of the SPE Symposium: Hydraulic Fracturing in Russia, Experience and Prospects, Virtual, 22 September 2020.
16. Wu, H. *Research on the Optimization Model of Shale Gas Well Fracturing Based on Machine Learning*; China University of Petroleum: Beijing, China, 2020. [CrossRef]
17. Li, J.; Chen, C.; Xiao, J. Yield prediction of shale gas muti-stage fracturing wells based on random forest algorithm. *J. Yangtze Univ. (Nat. Sci. Ed.)* **2020**, *17*, 34–38.
18. Yan, Z.; Wang, T.; Liu, Z.; Zhuang, Z. Machine-learning-based Prediction Methods on Shale Gas Recovery. *CHINese J. Solid Mech.* **2021**, *42*, 221–232. [CrossRef]
19. Ma, X.; Fan, Y. Productivity prediction model for vertical fractured well based on machine learning. *Math. Pract. Theory* **2021**, *51*, 186–196.
20. Kubota, L.K.; Reinert, D. Machine learning forecasts oil rate in mature onshore field jointly driven by water and steam injection. In Proceedings of the SPE Annual Technical Conference and Exhibition, Calgary, AB, Canada, 30 September–2 October 2019.
21. Bao, A.; Gildin, E.; Huang, J.; Coutinho, E.J. Data-driven end-to-end production prediction of oil reservoirs by En KF-enhanced recurrent neural networks. In Proceedings of the SPE Latin American and Caribbean Petroleum Engineering Conference, Virtual, 27–31 July 2020.
22. Perez, H.H.; Datta-Gupta, A.; Mishra, S. The Role of Electrofacies, Lithofacies, and Hydraulic Flow Units in Permeability Predictions from Well Logs: A Comparative Analysis Using Classification Trees. *Soc. Pet. Eng.* **2005**, *8*, 143–155. [CrossRef]
23. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

24. Molnar, C. Interpretable Machine Learning [M/OL]. Available online: https://christophm.github.io/interpretable-m-book (accessed on 29 March 2022).
25. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. Asurvey of methods for explaining black box models. *ACM Comput. Surv.* **2018**, *51*, 93–100. [CrossRef]
26. Baehrens, D.; Schroeter, T.; Harmeling, S.; Kawanabe, M.; Hansen, K.; Müller, K.R. How to explain individual classification decisions. *J. Mach. Learn. Res.* **2010**, *11*, 1803–1831. [CrossRef]
27. Feng, D.; Wu, G. Interpretable machine learning-based modeling approach for fundamental properties of concretcstructures. *J. Build. Struct.* **2022**, *43*, 228–238. [CrossRef]
28. Mai-Cao, L.; Truong-Khac, H. A Comparative Study on Different Machine Learning Algorithms for Petroleum Production Forecasting. *Improv. Oil Gas Recover* **2022**, *6*, 1–8. [CrossRef]
29. Doan, T.; Vo, M.V. Using Machine Learning Techniques for Enhancing Production Forecast in North Malay Basin. *Improv. Oil Gas Recovery* **2021**, *5*, 1–7. [CrossRef]
30. Hou, Y.; Wu, Y.; Hu, X.; Tang, M.; Liu, Y.; Zhang, J.; Niu, L.; Xu, W. Fracturing and Production Observations from the First Two Horizontal Wells for Shale Oil Exploration in Ordos Basin. In Proceedings of the Abu Dhabi International Petroleum Exhibition & Conference, Abu Dhabi, United Arab Emirates, 9–12 November 2020. [CrossRef]