symmetry

MDPI

*Article*

# Machine Learning-Based Research for Predicting Shale Gas Well Production

**Nijun Qi [1,2,3], Xizhe Li [1,2,3,*], Zhenkan Wu [3], Yujin Wan [1,2,3], Nan Wang [3], Guifu Duan [3], Longyi Wang [1,2,3], Jing Xiang [3], Yaqi Zhao [3] and Hongming Zhan [3,*]**

[1] School of Engineering Science, University of Chinese Academy of Sciences, Beijing 100049, China; qinijun22@mails.ucas.ac.cn (N.Q.); wanyj69@petrochina.com.cn (Y.W.); wanglongyi22@mails.ucas.ac.cn (L.W.)
[2] Institute of Porous Flow and Fluid Mechanics, Chinese Academy of Sciences, Langfang 065007, China
[3] Research Institute of Petroleum Exploration and Development, PetroChina, Beijing 100083, China; wuzhenkai@petrochina.com.cn (Z.W.); wn215@petrochina.com.cn (N.W.); duanguifu@petrochina.com.cn (G.D.); xiangjing4418@stu.cdut.edu.cn (J.X.); 202004010206@stu.cdut.edu.cn (Y.Z.)
* Correspondence: lxz69@petrochina.com.cn (X.L.); zhanhongming17@mails.ucas.edu.cn (H.Z.)

**Abstract:** The estimated ultimate recovery (EUR) of a single well must be predicted to achieve scale-effective shale gas extraction. Accurately forecasting EUR is difficult due to the impact of various geological, engineering, and production factors. Based on data from 200 wells in the Weiyuan block, this paper used Pearson correlation and mutual information to eliminate the factors with a high correlation among the 31 EUR influencing factors. The RF-RFE algorithm was then used to identify the six most important factors controlling the EUR of shale gas wells. XGBoost, RF, SVM, and MLR models were built and trained with the six dominating factors screened as features and EUR as labels. In this process, the model parameters were optimized, and finally the prediction accuracies of the models were compared. The results showed that the thickness of a high-quality reservoir was the dominating factor in geology; the high-quality reservoir length drilled, the fracturing fluid volume, the proppant volume, and the fluid volume per length were the dominating factors in engineering; and the 360−day flowback rate was the dominating factor in production. Compared to the SVM and MLR models, the XG Boost and the RF models based on integration better predicted EUR. The XGBoost model had a correlation coefficient of 0.9 between predicted and observed values, and its standard deviation was closest to the observed values' standard deviation, making it the best model for EUR prediction among the four types of models. Identifying the dominating factors of shale gas single-well EUR can provide significant guidance for development practice, and using the optimized XGBoost model to forecast the shale gas single-well EUR provides a novel idea for predicting shale gas well production.

**Keywords:** shale gas; EUR forecast; dominating factors; RF-RFR algorithm; XGBoost model; machine learning



**Citation:** Qi, N.; Li, X.; Wu, Z.; Wan, Y.; Wang, N.; Duan, G.; Wang, L.; Xiang, J.; Zhao, Y.; Zhan, H. Machine Learning-Based Research for Predicting Shale Gas Well Production. *Symmetry* **2024**, *16*, 600. https://doi.org/10.3390/sym16050600

Academic Editor: Vasilis K. Oikonomou

Received: 4 March 2024
Revised: 28 March 2024
Accepted: 8 April 2024
Published: 12 May 2024

**Correction Statement:** This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the backmatter of the website version of this article.



## 1. Introduction

Oil, natural gas, and coal are the three traditional energy sources dominating the global energy supply today. The proposed "double carbon" goal accelerates the transition of the global energy system [1]. As a clean and low-carbon fossil energy source, natural gas serves as a "bridge" and "pillar" in the energy transition process [2,3]. There are two types of natural gas: conventional and unconventional. Unconventional natural gas plays a vital role in exploration and development and will be the primary field in the present and future. Shale gas is a significant unconventional natural gas resource with extensive global reserves. The production of gas is also expected to increase as extraction technology advances.

The estimated ultimate recovery (EUR) is critical for accurately assessing the development potential of shale gas with abundant resources and, ultimately, achieving beneficial

development [4]. EUR is the foundation for developing shale gas reservoirs, linked to the design of gas well production and workover systems. It also forms the basis for gas fields' scientific and efficient development. Shale gas reservoirs have nanoscale pore characteristics, distinguishing them from conventional natural gas [5]. The production mode for gas wells is "non-constant pressure, non-constant production". The reservoirs' fluid flow characteristics are complex, with multiple flow phases and a fluctuating, declining production index. These factors raise the uncertainty of shale gas production forecasts and EUR valuations [6–8].

The empirical, numerical simulation, and analytical methods are the three most commonly used methods for evaluating EUR [9–12]. The empirical method is primarily an analytical method based on data fitting, which is simple to use and typically has broad applicability. Still, the human factor has a significant impact, and accuracy is difficult to ensure. The numerical simulation method uses basic seepage theory to create a detailed numerical gas reservoir model and estimate production. Nevertheless, the complexities of its modeling make it inaccessible. Finally, analytical methods typically rely on certain assumptions, such as the formation's homogeneity and the fluid's single-phase seepage nature. These methods have been widely used to guide shale gas development for many years. However, these methods have limited applicability, and the methods used at different stages vary greatly, as does the calculated EUR. Furthermore, due to the heterogeneity of shale reservoirs, the uncertainty of the transport mechanism, and the complexity of the fracture network, these methods have significant uncertainties in characterizing the reservoir and predicting the EUR [13,14].

Predicting shale gas production using data-driven machine learning techniques has become popular in the oil and gas industry as artificial intelligence technologies have been developed and improved [15]. Training machine learning models with data from older wells and using them to predict the production of new wells is now a viable approach using machine learning techniques. The oil and gas industry generates geological, engineering, and production data, the foundation for applying and popularizing machine learning models. In production prediction, Niu et al. [16] used multiple regression to predict the EUR of shale gas wells, which provided a new idea for EUR prediction. Hui et al. [17] conducted a shale gas production prediction study that combined geological and operational factors and used four models: linear regression, neural networks, gradient-boosting decision trees, and extra trees. Liu et al. [18] predicted the EUR of shale gas wells using a deep feed-forward network and integrating geological, engineering, and production factors. Han et al. [19] used a deep neural network based on a multilayer perceptron to predict natural gas production and achieved good results.

In this paper, the dominating factors influencing the production of a single well were screened from three perspectives—geology, engineering, and production—using Pearson correlation and mutual information value analysis, followed by combining the random forest algorithm (RF) with the recursive feature elimination (RFE) algorithm, the abbreviation of which is RF-RFE. Using the screened dominating factors as features and EUR as the label, we selected the multiple linear regression and support vector machine models in the single model, and the random forest and extreme gradient boosting models in the integrated model. From these, we could train and optimize the EUR prediction model, analyze the prediction effect, and finally select the best prediction model.

## 2. Methodology

This research is divided into three significant steps. First, the dominating factors were determined based on the feature set using Pearson's correlation and mutual information value analysis and combined with the RF-RFE algorithm. Second, prediction models were established using the dominant factors as inputs and EUR as the target. Third, the models' prediction performance was analyzed and evaluated to determine the best production prediction model.

### 2.1. Pearson Correlation

Correlation analysis analyzes the interdependence of two or more variables to determine the degree and direction of correlation and investigate the intrinsic relationship between variables [20]. The correlation coefficient measures the degree of correlation between two variables, whereas the Pearson correlation coefficient measures the degree of linear correlation [21]. In this case, we have two random variables, $X = (x_1, x_2, x_3, \ldots, x_n)$ and $Y = (y_1, y_2, y_3, \ldots, y_n)$. The Pearson correlation coefficient is thus denoted by [22]:

$$P(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \quad (1)$$

where: P (X, Y) is the Pearson correlation coefficient; x and y represent the corresponding sample values of the two variables; and $\overline{x}$ and $\overline{y}$ are the corresponding sample value means of the two variables.

P (X, Y) has values ranging $[-1, 1]$. The higher the absolute value, the stronger the correlation between the two variables. P (X, Y) > 0 indicates a linear positive correlation, while P (X, Y) < 0 indicates a linear negative correlation. When the value of P (X, Y) is close to 1 or $-1$, it indicates a strong linear correlation between the two variables. When the value of P (X, Y) is close to 0, it signifies a lack of linear correlation. Pearson's criteria for evaluating the degree of linear correlation are as follows: $|P(X,Y)| = 0$ indicates no linear correlation, $0 < |P(X,Y)| < 0.3$ indicates a low linear correlation, $0.3 \leq |P(X,Y)| < 0.8$ indicates a moderate linear correlation, $0.8 \leq |P(X,Y)| < 1.0$ indicates a high linear correlation, and $|P(X,Y)| = 1.0$ indicates a fully linear correlation.

### 2.2. Mutual Information

The mutual information value measures the correlation between two sets of events. This paper used the mutual information value to determine the degree of correlation between various features in the sample and EUR. Removing features from the chosen feature pairs with lower mutual information values can reduce the impact of redundant features.

Let P (x, y) be the joint distribution function of (X, Y), and P(x) and P(y) be the marginal distribution functions of X and Y, respectively. Thus, the mutual information value of X, Y, denoted as I (X, Y), is [23]:

$$I(X,Y) = \sum_{y \in Y}\sum_{y \in Y} P(x,y)\log\left(\frac{P(x,y)}{p(x)p(y)}\right) \quad (2)$$

I (X, Y) measures the information that X and Y share. It demonstrates how much uncertainty about one variable is reduced when the value of another is known. Mutual information can be used to determine the level of interdependence between two variables. If two variables are independent of one another and one does not reveal anything about the other, their mutual information value will be zero.

### 2.3. RF-RFE Algorithm

In the study of this paper, a large number of factors affecting EUR were selected, 31 in total, from which important factors need to be selected and their number determined. Commonly used methods, such as grey correlation analysis and the distance correlation coefficient, can analyze the correlation or distance correlation coefficient of each factor with EUR, but they cannot accurately give the number of important influencing factors, and they need to be selected by human beings, which increases the uncertainty of the inputs to the model [24,25].The RF-RFE algorithm is very suitable for solving the problem of feature selection, especially in the case of a large number of features and uncertainty about which ones are the most important, and it is able to give the number of important influencing

factors. Recursive feature elimination (RFE) is a feature selection algorithm that ranks feature variables [26]. Its main goal is to find the subset of features that contribute the most to the model's performance improvement by gradually removing features from it.

The RF-RFE algorithm analyzes the importance of variables using a random forest (RF) [27,28]. Then, it selects the important variables using the RFE method by ranking them in order of importance [29].

The basic steps are shown in Figure 1:

(1) Calculate and rank the importance of each feature in the initial variable training set data using RF.
(2) Remove the variable from the end of the feature importance degree ranking.
(3) Repeat steps (1) and (2) for the remaining variables, calculating the model's performance evaluation index each time, until all feature variables have been identified.

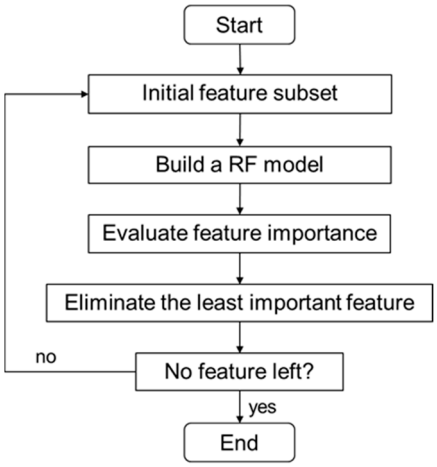



**Figure 1.** Flowchart of RF-RFE algorithm.

The variable importance measurement (VIM), which describes how much each variable contributes to the target variable, indicates the importance of each variable to the target variable [30]. RF can perform "variable importance measurement (VIM)" in the analysis process. The Gini index and out-of-band (OOB) index are commonly used as evaluation indexes. In this paper, the Gini index is used for evaluation with the following formulas [31].

The statistic $VIM_j^{(Gini)}$ denotes the average change in node-splitting impurity for the jth variable across all trees in RF. The Gini index is calculated as follows:

$$GI_m = \sum_{k=1}^{|k|} \hat{P}_{mk}\left(1 - \hat{P}_{mk}\right) \tag{3}$$

where K is the number of classes in the self-help sample set, and $\hat{P}_{mk}$ represents the probability estimate that the sample of node m belongs to the kth class when the sample is dichotomous (K = 2). The Gini index of node m is:

$$GI_m = 2\hat{P}_m\left(1 - \hat{P}_m\right) \tag{4}$$

where $\hat{P}_m$ is the probability estimate that the sample belongs to any class at node m.

The importance of variable $X_j$ at node m, or the amount of change in the Gini index before and after branching at node m, is as follows:

$$VIM_{jm}^{(Gini)} = GI_m - GI_l - GI_r \tag{5}$$

where $GI_l$ and $GI_r$ are the Gini indexes of two new nodes split by node m.

If a variable $X_j$ occurs M times in the ith tree, its importance in the ith tree is:

$$VIM_{ij}^{(Gini)} = \sum_{m=1}^{M} VIM_{jm}^{(Gini)} \quad (6)$$

The Gini importance of a variable $X_j$ in RF is defined as follows:

$$VIM_{j}^{(Gini)} = \frac{1}{n}\sum_{i=1}^{n} VIM_{ij}^{(Gini)} \quad (7)$$

where n is the number of classification trees in the RF.

After evaluation, each element has a positive value that adds up to 1.0. The higher an element's value, the more important the corresponding feature.

The advantages of the RF-RFE algorithm are: (1) Since random forest has good robustness to outliers and noise, the RF-RFE algorithm is able to resist the influence of these factors in feature selection. (2) Random forests can provide a quantitative assessment of feature importance, which helps to understand which features in the data are most critical for prediction. (3) The random forest model has good interpretability, which, combined with the feature selection process of RFE, makes the final set of selected features easy to understand. The limitations of the RF-RFE algorithm are: (1) The RFE algorithm has sequential dependency problems during feature elimination; the features that are eliminated first may affect the importance assessment of the subsequent features. (2) The performance of the algorithm may be affected by parameters such as the number of trees in the random forest, the maximum depth of the tree, etc., and careful tuning of the parameters is required to obtain optimal performance.

### *2.4. Prediction Models*

Linear regression is a statistical analysis method commonly used to determine the quantitative relationship between two or more variables in mathematical statistics [32]. Multiple linear regression (MLR) is a useful multivariate statistical analysis technique that determines the significance of each independent variable for the dependent variable [33]. The MLR model is expressed as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (8)$$

where y is the target variable for prediction, $\beta_0$ is the intercept, $x_1$, $x_2 \ldots$, $x_n$, are the feature variables to predict y, and $\beta_1$, $\beta_2$, …, $\beta_n$ are the weights of the feature variables. In this paper, y corresponds to EUR, and $x_1$, $x_2$, …, $x_n$ correspond to the screened dominant factors. The linear model is simple in form, easy to model, does not require very complex calculations, runs quickly even with large data volumes, and contains some critical machine learning fundamentals. However, it does not fit nonlinear data well.

Support vector machine (SVM) is a popular and effective supervised learning algorithm applied in various fields [34]. It is a binary classification model with the basic model being a maximally spaced linear classifier defined on the feature space. The basic idea behind SVM learning is to solve a separating hyperplane that correctly divides the training dataset while obtaining the maximum geometric spacing. A linearly divisible dataset has infinite hyperplanes (perceptual machines), but the hyperplane with the largest geometric interval is unique. SVM algorithms are intended to handle nonlinear data using a kernel function to map the original input data samples to a higher-dimensional space, making the samples linearly differentiable in the new feature space [35,36]. SVMs have the following advantages. First, they can solve high-dimensional problems with large feature spaces. Second, they can deal with nonlinear feature interactions. Third, they do not require the entire dataset. Fourth, their generalization ability is relatively strong. However, computational efficiency is low when there are many observation samples. There is no generalized solution for nonlinear problems, and finding a suitable kernel function can

be challenging. Therefore, SVM is commonly used to solve machine learning problems involving small samples.

Breiman et al. proposed RF in 2001 [27,28]. It is an extended variant of bagging that incorporates random attribute selection into the training process of decision trees, using the decision tree as the base learner to construct the bagging integration. It is trained by randomly selecting a subset of the original data and applying multiple models trained on that subset for regression and classification. For the regression problem, the RF predicts an average of all the decision trees' predicted results; for the classification problem, the RF determines the final result using the majority voting method. The advantages of RF are the following. First, it can handle high-dimensional (many features) data without having to be downscaled. Second, it can determine the importance of features as well as the interactions between features. Third, it is faster to train and easy to convert into a parallel method. Fourth, it is relatively simple to implement. However, RF has been shown to overfit in some noisy classification or regression problems.

Extreme gradient boosting (XGBoost), a machine learning algorithm, is based on the integration concept proposed by Chen et al. [37]. It systematically and efficiently implements gradient boosting, with a linear classifier or a tree as the base learner. Unlike traditional integrated learning, XGBoost boosts performance by reducing model bias. The main idea is to add another model based on the current model, resulting in a better combined model than the current one. Its advantages include the following. First, adding regularization simplifies the learned model and prevents overfitting. Second, for samples with missing feature values, XGBoost can automatically learn its split direction. Third, it supports parallel processing and has high computational efficiency. Fourth, it has a good processing speed and accuracy for low- and medium-dimensional data. However, it is unsuitable for processing high-dimensional feature data and does not perform well with unstructured data; the algorithm has too many parameters and complex tuning parameters, limiting its use to some extent.

### *2.5. Bayesian Optimization*

Bayesian optimization is a global algorithm based on Bayes' theorem that can produce an approximate optimal solution with minimal evaluation cost [38]. For a given optimized objective function, it first samples randomly in the parameter space to create a preliminary objective function distribution, then continuously searches for solutions that maximize or minimize the objective function based on historical information, iterating until the distribution fits through the sampling points and approximates the actual objective function. The relationship between the integrated learning model's numerous parameters and its performance exhibits a black-box characteristic with a complex structure, making it impossible to determine its internal structure. In Bayesian optimization, an agent model can describe the relationship between parameter selection and objective function. During the evaluation of the sample points, the entire search history is used to identify the most likely extreme points, which contributes to improving the probabilistic agent model. The method has the advantages of fast convergence and fewer optimization iterations, and it is beneficial for solving problems with multiple peaks, nonconvexity, a black box, and observation noise [39].

## 3. Development of Production Forecast Models

### *3.1. Experimental Data and Pre-Processing*

The Weiyuan shale gas field is in the southwestern part of the Sichuan Basin, specifically in southwest Sichuan's low-fold zone of the ancient central slope. The Weiyuan backslope tectonics has developed against the backdrop of ancient uplift, with the overall performance being a large-scale, wide, and slow monoclinic tectonics tilted northwest to southeast. The shale thickness ranges from 180 to 600 m, the burial depth is between 2000 and 4000 m, and the built-up area is 1520 $km^2$ [40]. The marine shale in the Weiyuan shale gas field has a low total organic carbon (TOC) value, low porosity, low gas saturation, a

thin layer of high-quality reservoir, a high degree of thermal evolution, complex formation conditions, and a significant horizontal stress difference [41,42]. By June 2023, 566 horizontal wells had been produced, with a total gas production of 236 × $10^8$ $m^3$. The drilled horizontal well has a lateral length of between 816 m and 3210 m, with an average of 1676 m and a well spacing of 300 m to 400 m. In terms of production, the wells exhibit low output and high variability.

The dataset for this paper was derived from the Weiyuan 202 and Weiyuan 204 well areas. In total, 31 features were identified for dominant-factor screening. There were thirteen geological features, and the features related to reservoir thickness were selected as high-quality reservoir thickness and type I reservoir thickness based on a comprehensive interpretation of well-logging data. The remaining features were Young's modulus, Poisson's ratio, formation fracture pressure, horizontal stress difference, permeability, vertical depth, TOC, porosity, pressure coefficient, total gas content, and average brittleness index, which were obtained from core sample testing and logging interpretations. Five drilling-related features—horizontal section length, high-quality reservoir length drilled, type I reservoir length drilled, drilling rate of high-quality reservoir, and drilling rate of type I reservoir—were obtained from drilling construction reports. Nine features regarding reservoir modification were obtained from fracturing construction summaries: fracturing section length, fracturing fluid volume, fluid volume per length, proppant volume, proppant volume per length, the number of fracturing stages, fracture clusters, proppant volume per cluster, and average pump rate. Four features regarding production were obtained from production dynamics summaries: 30−day flowback rate, 90−day flowback rate, 180−day flowback rate, and 360−day flowback rate.

In this study, 200 wells were collected as samples, with all data complete, with 80% serving as a training set and 20% as a test set.

To improve the model's prediction accuracy, eliminate the influence of magnitude, and improve training speed and classification effect, input and output data must be preprocessed. The data were more stable and there were no extreme maximum and minimum values. Therefore, this paper adopted the normalization processing method to normalize the data to the interval [0, 1]. The normalization formula is:

$$y_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{9}$$

where $x_i$ is the original data, $y_i$ is the normalized data, $x_{max}$ is the maximum value of the original data, and $x_{min}$ is the minimum value of the original data.

The model's performance was evaluated using three metrics: mean absolute error (MAE), mean square error (MSE), and coefficient of determination ($R^2$). These two metrics, MAE and MSE, are used to measure the difference between predicted and observed values, and their values range from 0 to positive infinity; the closer the value of MAE and MSE are to 0, the more accurate the prediction of the model is, and the better the model's performance is. $R^2$ is a statistic that measures the fit of a regression model and can take values from 0 to 1. The closer the value of $R^2$ is to 1, the closer the predicted values are to the observed values and the better the model performs. The equations are as follows [33]:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i| \tag{10}$$

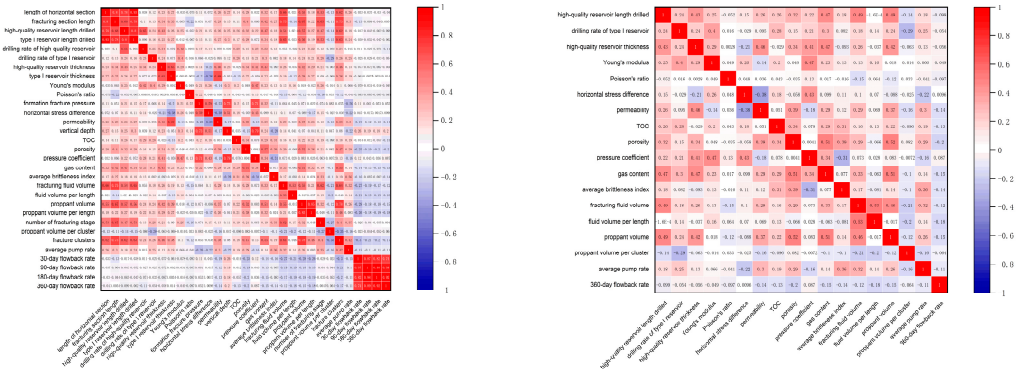
$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{11}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \overline{y}_i)^2} \tag{12}$$

where $y_i$ is the observed value of EUR, $\hat{y}_i$ is the predicted value of EUR, $\overline{y}_i$ is the average value of observed value of EUR, and m is the number of samples.

### *3.2. Screening of Dominating Factors*

The Pearson correlation analysis of the dataset's 31 features yielded the Pearson correlation coefficient mapping shown in Figure 2a.



(**a**) Before removing redundant features. (**b**) After removing redundant features.

**Figure 2.** Comparison of the Pearson correlation coefficient plots before and after feature screening.

Based on the evaluation of Pearson correlation coefficients, Figure 2a shows a strong correlation between some of the features, indicating that the features selected from the dataset contained redundant features. All features in the dataset that had a strong correlation were chosen as feature pairs with an absolute Pearson correlation coefficient greater than 0.6, and the mutual information values of these feature pairs were compared to eliminate redundant features.

Table 1 displays the mutual information values of each feature with EUR. The features with the highest mutual information values from each pair were high-quality reservoir length drilled, high-quality reservoir thickness, 360−day flowback rate, pressure coefficient, proppant volume, and fracturing fluid volume.

**Table 1.** Mutual information value between features and EUR.

| Features | Mutual Information Value | Features | Mutual Information Value |
|---|---|---|---|
| Length of horizontal section | 0.613 | 30−day flowback rate | 0.806 |
| Fracturing section length | 0.742 | 90−day flowback rate | 0.823 |
| High-quality reservoir length drilled | 0.825 | 180−day flowback rate | 0.826 |
| Drilling rate of high-quality reservoir | 0.714 | 360−day flowback rate | 0.854 |
| Type I reservoir length drilled | 0.652 | Number of fracturing stages | 0.816 |
| Formation fracture pressure | 0.796 | Proppant volume per length | 0.796 |
| Vertical depth | 0.812 | Proppant volume | 0.931 |
| Pressure coefficient | 0.846 | Fracturing fluid volume | 0.947 |
| Type I reservoir thickness | 0.699 | Fracturing clusters | 0.813 |
| High-quality reservoir thickness | 0.816 | | |

Figure 2b depicts a plot of the correlation coefficients after removing the redundant factors. As shown in the figure, the 18 features retained were high-quality reservoir length drilled, drilling rate of type I reservoir, high-quality reservoir thickness, Young's

modulus, Poisson's ratio, horizontal stress difference, permeability, TOC, porosity, pressure coefficient, total gas content, average brittleness index, fracturing fluid volume, fluid volume per length, proppant volume, proppant volume per cluster, average pump rate, and 360−day flowback rate.

The importance of the 18 initial features was assessed using the RF-RFE algorithm, and important features were chosen from them.

The initial features were arranged in descending order of importance, as illustrated in Figure 3. Each time, the features ranked last in importance were removed, and the $R^2$ value of the five-fold cross-validation was recalculated. $R^2$ explains the variance score in the regression model, reflecting the degree of regression fit. The closer it is to one, the closer the predicted value is to the observed value, which is used to evaluate the model's performance.

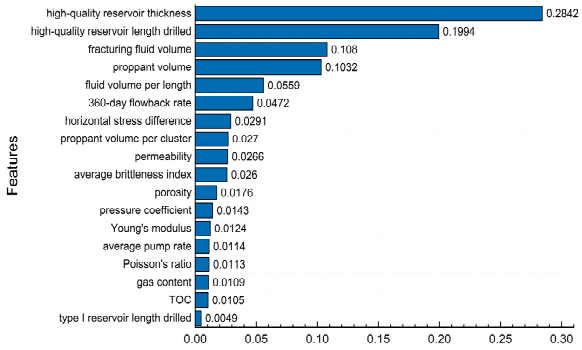



**Figure 3.** Distribution of feature importance.

Figure 4 depicts the five-fold cross-validation curve for the variation in $R^2$ value with the number of features. When the number of features was 18, the initial feature set, the $R^2$ value gradually increased as unimportant features were removed. That was because removing unimportant features could lessen the impact of redundant data on the algorithm. When the number of features was 6, the $R^2$ value was highest. However, as the number of features decreased, the $R^2$ value changed significantly due to deleting the most important features. As shown in Table 2, the six features with the highest $R^2$ scores were high-quality reservoir thickness, high-quality reservoir length drilled, fracturing fluid volume, proppant volume, fluid volume per length, and 360−day flowback rate. These features corresponded to the first six features in the importance ranking of the initial feature set. These six features were the most optimal feature set and were the primary factors for determining the EUR of a single well.

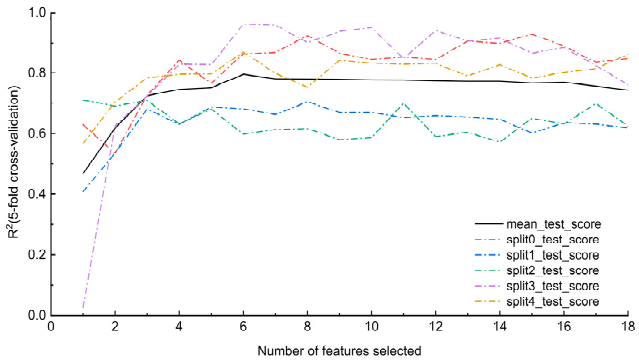



**Figure 4.** Recursive feature elimination curves.

**Table 2.** Distribution of dominating factors.

| Number of Features | Feature Types | Features |
|---|---|---|
| 6 | Geology<br>Drilling<br>Reservoir modification<br>Production<br>Geology | High-quality reservoir thickness,<br>high-quality reservoir length drilled,<br>fracturing fluid volume, proppant volume, fluid volume per length,<br>360−day flowback rate,<br>high-quality reservoir thickness |

### *3.3. EUR Forecast*

MLR and SVM models in single models and XGBoost and RF models in integrated learning were selected for EUR prediction. These four types of models were trained using EUR as the label. Six dominating factors were screened as inputs. Bayesian optimization algorithms were introduced in the model training process to perform super-parameters and improve the model (MLR models are excluded because the Bayesian algorithms are not used for parameter optimization in MLR models). Finally, the model simulation results and assessment indicators were compared and analyzed to determine the best model among the four.

#### 3.3.1. Comparative Analysis of Model Simulation Results

The six dominating factors screened out using the RF-RFE algorithm were used as inputs for the four models. The models were trained and parameters optimized, and the EUR values predicted by the four models were obtained on the test set. As shown in Figure 5, A–D are XGBoost, RF, SVM, and MLR models, respectively. The predicted values of EUR by two models, XGBoost and RF, were closer to the observed values. The predicted values of the SVM model slightly deviated from the observed values. The predicted values of the MLR model further deviated from the observed values. The XGBoost and RF models predicted values closer to the observed values than the SVM and MLR models. The predicted values differed significantly from the observed values in a few cases. However, on average, the predicted values were more similar to the observed values, and the predictions were more accurate.

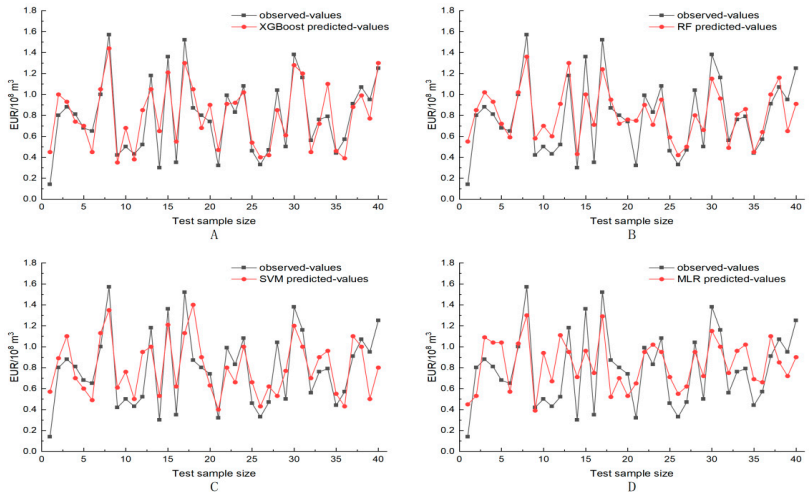



**Figure 5.** Comparison of model-predicted and observed values.

Figure 6 depicts scatter plots of predicted and observed values from the four models. The black line represents the linear regression line, and the red diagonal line serves as the reference line. Overall, the scatters of the XGBoost model exhibited a concentrated distribution along the diagonal line, indicating that the predicted and observed values were more consistent. In the RF model, when observed values were $<1 \times 10^8$ $m^3$, most

scatter points were distributed above the diagonal line, indicating an overestimating of the predicted value. When observed values were >1 × $10^8$ $m^3$, most scatter points were distributed below the diagonal line, indicating an underestimation of the predicted value. Compared to the XGBoost and RF models, the scatters of the SVM and MLR models were more dispersed. The scatters of the two models were distributed above the diagonal line when the observed values were <1 × $10^8$ $m^3$, indicating an overestimation of the predicted values. The scatters of the two models were distributed below the diagonal line when the observed values were >1 × $10^8$ $m^3$, indicating that the predicted value was generally underestimated.

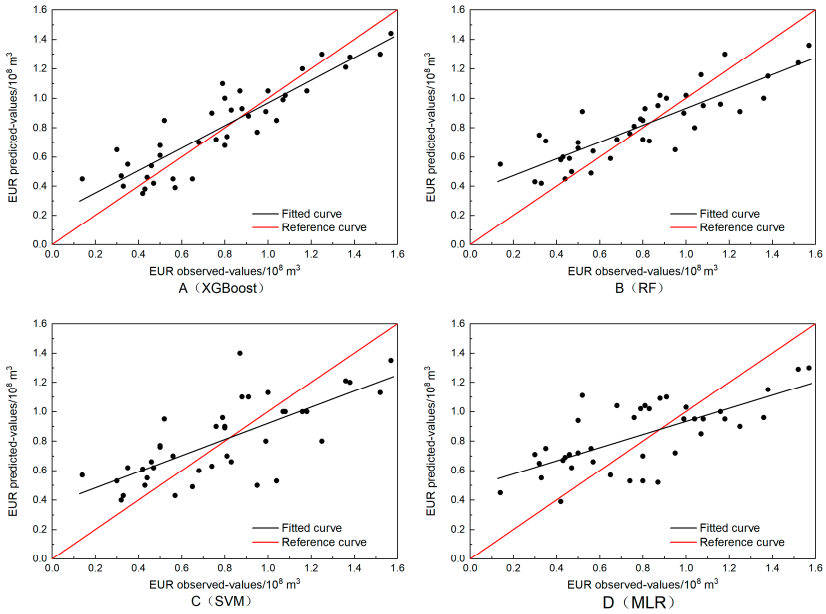



**Figure 6.** Scatterplot of model-predicted and observed values.

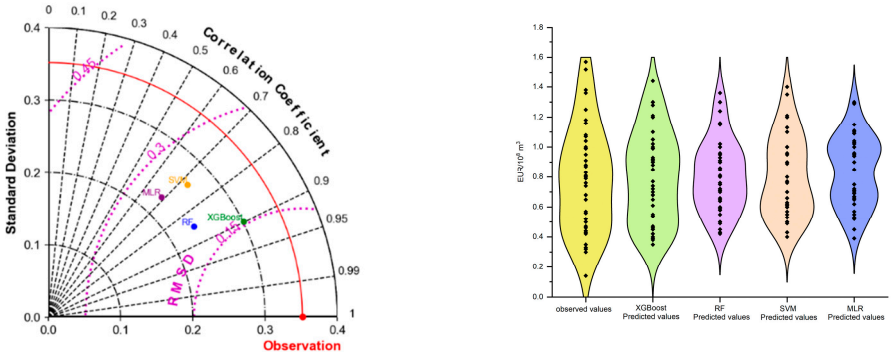
### 3.3.2. Comparative Analysis of Model Assessment Indicators

Table 3 shows the evaluation results for the four models. The MAE and MSE of the predicted and observed values of the XGBoost model were 0.130 and 0.024, respectively. These were correspondingly reduced by 18.2% and 38.5% compared to the RF model, by 37.5% and 59.3% compared to the SVM model, and by 44.4% and 64.7% compared to the MLR model. On the test set, the XGBoost model outperformed the RF, SVM, and MLR models (all with an $R^2$ lower than 0.70), with an $R^2$ of 0.804.

**Table 3.** Evaluation of the results of the four model predictions.

| Parameters | XGBoost | RF | SVM | MLR |
|---|---|---|---|---|
| MAE | 0.130 | 0.159 | 0.208 | 0.234 |
| MSE | 0.024 | 0.039 | 0.059 | 0.068 |
| $R^2$ | 0.804 | 0.688 | 0.522 | 0.459 |

As shown in Figure 7a, the Taylor diagrams of the prediction results of the four models show that the correlation coefficient between the XGBoost predicted and observed values was the highest, reaching 0.9. In contrast, the correlation coefficients between the RF, SVM, and MLR predicted and observed values were relatively low, at 0.85, 0.72, and 0.69, respectively. The standard deviations of the XGBoost predicted values were the closest to the observed values.




(**a**) (**b**)

**Figure 7.** (**a**) Taylor diagrams of the predicted results of the four models. (**b**) Comparison of the statistical distribution of predicted and observed values of the four models.

Statistical analyses of the observed and predicted values are shown in Figure 7b. The data distribution of the observed values is approximately symmetrical, with EUR = 0.7 × $10^8$ $m^3$ as the axis. Among the four types of models, the XGBoost model has the closest data distribution pattern between predicted and observed values and is also symmetrical, with EUR = 0.7 × $10^8$ $m^3$ as the axis, indicating that the XGBoost model is the closest to the observed values in terms of both the overall EUR prediction mean and the prediction of the high and low EUR values.

#### 3.3.3. Analysis of Model Application Performance

To test the efficacy of the dominating factor screened, the same method was used to train models and optimize the parameters using the 31 influencing factors that were not subjected to dominating factor screening as inputs, as well as to make predictions on the test set. The test set's prediction results were compared to those after the dominating factor screening, as shown in Table 4.

**Table 4.** Comparison of model performance under different methods before and after screening of the dominating factor.

| Method | Data Set | Number of Features | $R^2$ |
|---|---|---|---|
| XGBoost | Original dataset | 31 | 0.776 |
| | Dominating factors | 6 | 0.804 |
| RF | Original dataset | 31 | 0.662 |
| | Dominating factors | 6 | 0.688 |
| SVM | Original dataset | 31 | 0.501 |
| | Dominating factors | 6 | 0.522 |
| MLR | Original dataset | 31 | 0.421 |
| | Dominating factors | 6 | 0.459 |

Table 4 shows that using dominating factors as inputs improved the prediction effect of various machine learning methods, with the $R^2$ increasing by 3.6% and 3.9% for the XGBoost and RF methods and by 4.2% and 9% for the SVM and MLR methods, respectively.

## 4. Discussion

The six dominating factors identified by the RF-RFE algorithm cover parameters in each category of geology, drilling, reservoir modification, and production, providing comprehensive information coverage. The geological aspect is characterized by high-quality reservoir thickness. The greater the thickness of the high-quality reservoir, the greater the shale gas reserves and the EUR of a single well under a given level of fracturing. The drilling parameter is the length of the high-quality reservoir drilled. The length of

high-quality reservoir drilled is an essential factor in determining a single well's control range; the longer the length of high-quality reservoir drilled, the greater the control range of a single well. The fracturing fluid volume, proppant volume, and fluid volume per length reflect the degree to which the reservoir has been modified. The degree of modification determines the utilization of shale gas reserves. The production side is characterized by the 360−day flowback rate, reflecting the fracturing fluid flowback effect on EUR. The six dominating factors screened out by applying the method in this paper are basically consistent with the conclusions summarized by the production practice in reference, which verifies the accuracy of the high-quality reservoir thickness, the high-quality reservoir length drilled, and the degree of reservoir modification as the dominating factors of EUR, and also illustrates the reasonableness of the screened-out dominating factors for predicting the EUR [40]. Combining these dominating factors allows for a more comprehensive and accurate forecast of the EUR of a single shale gas well.

A comparison of the results of the four types of model simulations and the assessment indicators revealed that the XGBoost model performed best in EUR prediction, which was closer to the observed values. The RF model performed second best, while the SVM and MLR models performed poorly. The XGBoost model was the most accurate for EUR prediction among the four model classes. The optimized XGBoost model can complete the prediction of EUR by six features, and the data can be easily obtained and can achieve high accuracy, which provides a simple and convenient method for the prediction of EUR of shale gas wells. However, the sample size used in this study was 200 wells due to limited conditions. In general, the models based on machine learning methods have a strong relationship with the sample size of the dataset. The larger the sample size is, the higher the accuracy of the developed model is. Therefore, it would be helpful to develop a sample database with a larger capacity in the future.

Table 4 shows that prediction models based on production-dominating factors can improve prediction performance. It is not the case that the more features are input, the better the model performance will be. The cause of this phenomenon was a moderate or strong covariance between features with a lower importance ranking and other features, which was equivalent to adding "noise" to the training set. Therefore, a reasonable selection of the type and number of input features serves as the foundation for modeling.

In addition, the XGBoost and RF models based on the integration idea outperformed the SVM and MLR models in terms of prediction performance because the integration model was a combination of multiple base learners, which reduced the bias and variance of the individual models, thereby improving prediction accuracy. Furthermore, the XGBoost model outperformed the RF model in prediction accuracy. The reason could be that XGBoost fine-tuned the model by continuously building it to minimize the loss function, improving prediction accuracy.

## 5. Conclusions

This paper identified the dominating factors of EUR in a single well in the Weiyuan area using a mathematical method and the RF-RFE algorithm. On this basis, EUR was predicted using multiple models, and the best prediction model was chosen. The following conclusions were drawn:

1. The RF method was used to rank the factors' importance, and the importance of each factor to EUR was clarified. The geological, engineering, and production factors of 200 shale gas wells in the Weiyuan block were thoroughly analyzed using Pearson correlation and mutual information, combined with the RF-RFE algorithm. After removing redundant and unimportant factors, six factors were chosen as the dominating factors among 31 EUR influencing factors. The results showed that the dominating factor in geology was the thickness of high-quality reservoir. The dominating factors in engineering included high-quality reservoir length drilled, fracturing fluid volume, proppant volume, and the fluid volume per length. The dominating factor in production was the 360−day flowback rate.

2. With the six dominating factors screened as features and EUR as labels, XGBoost, RF, SVM, and MLR models were built and trained. The results showed that the XGBoost and RF models, based on the integration idea, outperformed the SVM and MLR models. Among the four models, the XGBoost model had an $R^2$ of 0.804, significantly higher than those of the RF, SVM, and MLR models. The MAE and MSE of the XGBoost model were 0.130 and 0.024, respectively, significantly lower than those of the other three models. The correlation coefficients between the predicted and observed values of the XGBoost model were around 0.9, and the standard deviation was closest to the observed values. Thus, the XGBoost model was the most effective of the four types of models for EUR prediction.
3. Identifying the dominant factors clarified the most important factors influencing the EUR of shale gas wells in the Weiyuan block, providing helpful guidance for development practice. Higher production could be achieved by selecting an area with a large thickness of high-quality reservoir for well deployment, increasing the length of high-quality reservoir drilled, improving the scale of fracturing, and reasonably controlling the flowback of fracturing fluids. Based on the dominating factors, the optimized XGBoost model was used to predict the EUR of shale gas single wells in a simple way that requires fewer data types, can significantly improve prediction accuracy, and provides a new idea for predicting shale gas well production.

**Author Contributions:** Conceptualization, N.Q. and H.Z.; methodology, Z.W. and N.Q.; software, N.Q. and Y.W.; validation, N.Q. and N.W.; formal analysis, Z.W. and H.Z.; investigation, L.W. and J.X.; resources, G.D. and Z.W.; data curation, H.Z. and Y.Z.; writing—original draft preparation, N.Q.; writing—review and editing, N.Q.; visualization, G.D. and L.W.; supervision, X.L.; project administration, N.Q. and X.L.; funding acquisition, N.Q. and J.X. All authors have read and agreed to the published version of the manuscript.

**Funding:** CYS-FW-2023-0188 Study on Technical Policy for Beneficial Development of Deep Shale Gas in Chongqing Shale Gas Company.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author. The data are not publicly available due to some data confidentiality restrictions.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Zou, C.; Xue, H.; Xiong, B.; Zhang, G.; Pan, S.; Jia, C.; Wang, Y.; Ma, F.; Sun, Q.; Guan, C.; et al. Connotation, Innovation and Vision of "Carbon Neutrality". *Nat. Gas Ind. B* **2021**, *8*, 523–537. [CrossRef]
2. Zou, C.; Zhao, Q.; Zhang, G.; Xiong, B. Energy Revolution: From a Fossil Energy Era to a New Energy Era. *Nat. Gas Ind. B* **2016**, *3*, 1–11. [CrossRef]
3. Bugaje, A.-A.B.; Dioha, M.O.; Abraham-Dukuma, M.C.; Wakil, M. Rethinking the Position of Natural Gas in a Low-Carbon Energy Transition. *Energy Res. Soc. Sci.* **2022**, *90*, 102604. [CrossRef]
4. Ibrahim, A.F.; Alarifi, S.A.; Elkatatny, S. Application of Machine Learning to Predict Estimated Ultimate Recovery for Multistage Hydraulically Fractured Wells in Niobrara Shale Formation. *Comput. Intell. Neurosci.* **2022**, *2022*, 1–10. [CrossRef] [PubMed]
5. Zhou, W.; Zhu, J.; Wang, H.; Kong, D. Transport Diffusion Behaviors and Mechanisms of $CO_2/CH_4$ in Shale Nanopores: Insights from Molecular Dynamics Simulations. *Energy Fuels* **2022**, *36*, 11903–11912. [CrossRef]
6. Wang, K.; Li, H.; Wang, J.; Jiang, B.; Bu, C.; Zhang, Q.; Luo, W. Predicting Production and Estimated Ultimate Recoveries for Shale Gas Wells: A New Methodology Approach. *Appl. Energy* **2017**, *206*, 1416–1431. [CrossRef]
7. Fang, X.; Yue, X.; An, W.; Feng, X. Experimental Study of Gas Flow Characteristics in Micro-/Nano-Pores in Tight and Shale Reservoirs Using Microtubes under High Pressure and Low Pressure Gradients. *Microfluid. Nanofluid.* **2019**, *23*, 5. [CrossRef]
8. Pang, W.; Du, J.; Zhang, T. Production Data Analysis of Shale Gas Wells with Abrupt Gas Rate or Pressure Changes. In Proceedings of the SPE Middle East Oil and Gas Show and Conference, Manama, Bahrain, 15–21 March 2019; p. D041S046R001. [CrossRef]
9. Niu, W.; Lu, J.; Sun, Y. An Improved Empirical Model for Rapid and Accurate Production Prediction of Shale Gas Wells. *J. Pet. Sci. Eng.* **2022**, *208*, 109800. [CrossRef]
10. Stalgorova, E.; Mattar, L. Analytical Model for History Matching and Forecasting Production in Multifrac Composite Systems. In Proceedings of the SPE Canada Unconventional Resources Conference, Calgary, AL, Canada, 30 October–1 November 2012; p. SPE-162516-MS. [CrossRef]

11. Nobakht, M.; Clarkson, C.R. A New Analytical Method for Analyzing Production Data from Shale Gas Reservoirs Exhibiting Linear Flow: Constant Pressure Production. In Proceedings of the SPE Unconventional Resources Conference/Gas Technology Symposium, The Woodlands, TX, USA, 14–16 June 2011; p. SPE-143989-MS. [CrossRef]
12. Zhan, J.; Lu, J.; Fogwill, A.; Ulovich, I.; Cao, J.P.; He, R.; Chen, Z. An Integrated Numerical Simulation Scheme to Predict Shale Gas Production of a Multi-Fractured Horizontal Well. In Proceedings of the Abu Dhabi International Petroleum Exhibition and Conference, Abu Dhabi, United Arab Emirates, 13–16 November 2017; p. D031S087R002. [CrossRef]
13. Du, F.; Huang, J.; Chai, Z.; Killough, J. Effect of Vertical Heterogeneity and Nano-Confinement on the Recovery Performance of Oil-Rich Shale Reservoir. *Fuel* **2020**, *267*, 117199. [CrossRef]
14. Huang, S.; Ding, G.; Wu, Y.; Huang, H.; Lan, X.; Zhang, J. A Semi-Analytical Model to Evaluate Productivity of Shale Gas Wells with Complex Fracture Networks. *J. Nat. Gas Sci. Eng.* **2018**, *50*, 374–383. [CrossRef]
15. Choubey, S.; Karmakar, G.P. Artificial Intelligence Techniques and Their Application in Oil and Gas Industry. *Artif. Intell. Rev.* **2021**, *54*, 3665–3683. [CrossRef]
16. Niu, W.; Lu, J.; Sun, Y. A Production Prediction Method for Shale Gas Wells Based on Multiple Regression. *Energies* **2021**, *14*, 1461. [CrossRef]
17. Hui, G.; Chen, S.; He, Y.; Wang, H.; Gu, F. Machine Learning-Based Production Forecast for Shale Gas in Unconventional Reservoirs via Integration of Geological and Operational Factors. *J. Nat. Gas Sci. Eng.* **2021**, *94*, 104045. [CrossRef]
18. Liu, Y.; Ma, X.; Zhang, X.; Guo, W.; Kang, L.; Yu, R.; Sun, Y.-P. A Deep-Learning-Based Prediction Method of the Estimated Ultimate Recovery (EUR) of Shale Gas Wells. *Pet. Sci.* **2021**, *18*, 1450–1464. [CrossRef]
19. Han, D.; Kwon, S. Application of Machine Learning Method of Data-Driven Deep Learning Model to Predict Well Production Rate in the Shale Gas Reservoirs. *Energies* **2021**, *14*, 3629. [CrossRef]
20. Zhou, Z.; Ding, Y.; Zhao, Y.; Chen, P.; Fu, Q.; Xue, P.; Liu, S.; Huang, S.; Shi, H. A New Perspective for Assessing Hydro-Meteorological Drought Relationships at Large Scale Based on Causality Analysis. *Environ. Res. Lett.* **2023**, *18*, 104046. [CrossRef]
21. Thao, N.X. A New Correlation Coefficient of the Pythagorean Fuzzy Sets and Its Applications. *Soft Comput.* **2020**, *24*, 9467–9478. [CrossRef]
22. Kumar, G.P.; Jena, P. Pearson's Correlation Coefficient for Islanding Detection Using Micro-PMU Measurements. *IEEE Syst. J.* **2021**, *15*, 5078–5089. [CrossRef]
23. Liu, H.; Li, Y.; Du, Q.; Jia, D.; Wang, S.; Qiao, M.; Qu, R. Prediction of production during high water-cut period based on multivariate time series model. *J. China Univ. Pet. (Ed. Nat. Sci.)* **2023**, *47*, 103–114. [CrossRef]
24. Xu, J.; Liu, Z.; Yin, L.; Liu, Y.; Tian, J.; Gu, Y.; Zheng, W.; Yang, B.; Liu, S. Grey Correlation Analysis of Haze Impact Factor PM2.5. *Atmosphere* **2021**, *12*, 1513. [CrossRef]
25. Niu, W.; Lu, J.; Sun, Y.; Guo, W.; Liu, Y.; Mu, Y. Development of Visual Prediction Model for Shale Gas Wells Production Based on Screening Main Controlling Factors. *Energy* **2022**, *250*, 123812. [CrossRef]
26. Gregorutti, B.; Michel, B.; Saint-Pierre, P. Correlation and Variable Importance in Random Forests. *Stat. Comput.* **2017**, *27*, 659–678. [CrossRef]
27. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
28. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P.; Feuston, B.P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958. [CrossRef] [PubMed]
29. Jiang, X.; Zhang, Y.; Li, Y.; Zhang, B. Forecast and Analysis of Aircraft Passenger Satisfaction Based on RF-RFE-LR Model. *Sci. Rep.* **2022**, *12*, 11174. [CrossRef] [PubMed]
30. Grömping, U. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *Am. Stat.* **2009**, *63*, 308–319. [CrossRef]
31. Davidson, R. Reliable Inference for the Gini Index. *J. Econom.* **2009**, *150*, 30–40. [CrossRef]
32. Zhou, Z. *Machine Learning*; Tsinghua University Press: Beijing, China, 2016; ISBN 978-7-302-42328-7.
33. Leng, J.; Gao, X.; Zhu, J. Application of Multivariate Linear Regression Statistical Prediction Model. *Stat. Decis.* **2016**, 82–85. [CrossRef]
34. Ding, S.; Shi, Z.; Tao, D.; An, B. Recent Advances in Support Vector Machines. *Neurocomputing* **2016**, *211*, 1–3. [CrossRef]
35. Chen, Y.; Zhou, X.; Huang, T.S. One-Class SVM for Learning in Image Retrieval. In Proceedings of the 2001 International Conference on Image Processing (Cat. No.01CH37205), Thessaloniki, Greece, 7–10 October 2001; Volume 1, pp. 34–37. [CrossRef]
36. Zhang, X.; Lu, X.; Shi, Q.; Xu, X.; Leung, H.E.; Harris, L.N.; Iglehart, J.D.; Miron, A.; Liu, J.S.; Wong, W.H. Recursive SVM Feature Selection and Sample Classification for Mass-Spectrometry and Microarray Data. *BMC Bioinform.* **2006**, *7*, 197. [CrossRef] [PubMed]
37. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
38. Zhu, J.; Chen, J.; Hu, W.; Zhang, B. Big Learning with Bayesian Methods. *Natl. Sci. Rev.* **2017**, *4*, 627–651. [CrossRef]
39. Binois, M.; Wycoff, N. A Survey on High-Dimensional Gaussian Process Modeling with Application to Bayesian Optimization. *ACM Trans. Evol. Learn. Optim.* **2022**, *2*, 1–26. [CrossRef]
40. Ma, X.; Li, X.; Liang, F.; Wan, Y.; Shi, Q.; Wang, Y.; Zhang, X.; Che, M.; Guo, W.; Guo, W. Dominating Factors on Well Productivity and Development Strategies Optimization in Weiyuan Shale Gas Play, Sichuan Basin, SW China. *Pet. Explor. Dev.* **2020**, *47*, 594–602. [CrossRef]

41. Zou, C.; Dong, D.; Wang, Y.; Li, X.; Huang, J.; Wang, S.; Guan, Q.; Zhang, C.; Wang, H.; Liu, H.; et al. Shale Gas in China: Characteristics, Challenges and Prospects (I). *Pet. Explor. Dev.* **2015**, *42*, 753–767. [CrossRef]
42. Dong, D.; Wang, Y.; Li, X.; Zou, C.; Guan, Q.; Zhang, C.; Huang, J.; Wang, S.; Wang, H.; Liu, H.; et al. Breakthrough and Prospect of Shale Gas Exploration and Development in China. *Nat. Gas Ind. B* **2016**, *3*, 12–26. [CrossRef]