Article

Optimization and Application of XGBoost Logging Prediction Model for Porosity and Permeability Based on K-means Method

Jianting Zhang 1,*, Ruifei Wang 1, Ailin Jia 2 and Naichao Feng 2

1 Institute of Petroleum Engineering, Xi’an Shiyou University, Xi’an 710016, China
2 PetroChina Research Institute of Petroleum Exploration and Development, Beijing 100083, China; fengnaichao@petrochina.com.cn (N.F.)
* Correspondence: 22110101013@stumail.xsyu.edu.cn

Abstract: The prediction and distribution of reservoir porosity and permeability are of paramount importance for the exploration and development of regional oil and gas resources. In order to optimize the prediction methods of porosity and permeability and better guide gas field development, it is necessary to identify the most effective approaches. Therefore, based on the extreme gradient boosting (XGBoost) algorithm, laboratory test data of the porosity and permeability of cores from the southern margin of the Ordos Basin were selected as the target labels, conventional logging curves were used as the input feature variables, and the mean absolute error (MAE) and the coefficient of determination ($R^2$) were used as the evaluation indicators. Following the selection of the optimal feature variables and optimization of the hyper-parameters, an XGBoost porosity and permeability prediction model was established. Subsequently, the innovative application of homogeneous clustering (K-means) data preprocessing was applied to enhance the XGBoost model’s performance. The results show that logarithmically preprocessed (LOG(PERM)) target labels enhanced the performance of the XGBoost permeability prediction model, with an increase of 0.26 in its test set $R^2$. Furthermore, the application of K-means improved the performance of the XGBoost prediction model, with an increase of 0.15 in the $R^2$ of the model and a decrease of 0.017 in the MAE. Finally, the POR_0/POR_1 grouped porosity model was selected as the final predictive model for porosity in the study area, and the Arctan(PERM)_0/Arcatan(PERM)_1 grouped model was selected as the final predictive model for permeability, which has better prediction accuracy than logging curves. The combination of K-means and the XGBoost modeling method provides a new approach and reference for the efficient and relatively accurate evaluation of porosity and permeability in the study area.

Keywords: XGBoost; K-means; porosity and permeability prediction model

1. Introduction

In the oil and gas industry, the porosity (POR) parameter characterizes the reservoir’s storage capacity, and the permeability (PERM) characterizes the reservoir’s ability to pass fluids. Porosity and permeability are important indicators for the evaluation of the reservoir during the development and production of oil and gas fields [1,2]. Consequently, the prediction and distribution of reservoir porosity and permeability are crucial for the exploration and development of regional oil and gas resources. The determination methods for porosity and permeability of unconventional oil and gas reservoirs can be broadly classified into two categories: laboratory measurement and indirect interpretation.

The porosity experimental measurement method is selected based on the main distribution range of porosity in regional core samples. In sandstone reservoirs with a predominant distribution of micropores (<2 nm), some scholars have applied the low-temperature carbon dioxide (CO$_2$) adsorption method to measure porosity. This approach is based on the results of carbon dioxide condensation in the capillary under critical conditions, which enables an estimation of the pore size distribution range of the core sample. This method...
is more accurate in characterizing micropores than other techniques [3]. In sandstone reservoirs with predominantly medium porosity distribution (2–50 nm), the use of the low-temperature nitrogen (N\textsubscript{2}) adsorption method is a more common approach for porosity determination [4–7]. Some scholars have applied constant-rate mercury injection technology to measure porosity in tight porous sandstone reservoirs. The constant-rate mercury injection method measures porosity with the unique advantage of an extremely low and constant injection rate of mercury [8,9]. Based on the variation of mercury injection rate with pressure, further research is conducted on the pore structure parameters of reservoir rocks [10]. Cheng et al. (2020) and Andrew et al. (2012) [11,12] applied nuclear magnetic resonance radiography to characterize pore diameters from 2 nm to 1000 nm. The processing and interpreting relaxation time data and its indirect conversion to porosity values had the advantages of not damaging the core and having high speed, high accuracy, and strong experimental operability, making it one of the important tools for reservoir evaluation and exploration engineering in oil and gas reservoirs. Therefore, the range of porosity measurement varies between different methods, and a single method cannot fully measure the pore characteristics of the reservoir. Multiple methods must be combined to quantitatively characterize the reservoir porosity. In contrast, permeability laboratory measurements are mainly conducted through gas measurement. According to Darcy’s law of fluids, when the gas reaches a steady state, the permeability parameters of the rock core are calculated, and the test results are more accurate [13,14]. In oil and gas production sites, permeability testing is mainly applied through methods, such as well testing and oil and gas production, in order to provide true formation permeability. The cost of testing methods is relatively higher [15,16]. In summary, the above experimental methods for determining porosity and permeability mentioned above are expensive, with a limited number of test points, which is not conducive to analyzing the porosity and permeability characteristics of regional and vertical reservoirs.

Indirect porosity interpretation requires a combination of multiple logging curves, such as acoustic time-difference logging (AC) and neutron logging (CNL), corrected with laboratory porosity test data to obtain a more accurate representation of logging porosity; then, permeability is calculated from the indirect porosity by analyzing pore-permeability relationships in the core to obtain a regional longitudinal and planar porosity and permeability logging evaluation [17–20]. In addition, many scholars have analyzed the linear or nonlinear relationship between experimentally measured porosity and permeability data and geophysical logging parameters to realize multi-angle and multi-method quantitative prediction of formation porosity and permeability parameters, which can then be applied to the actual production process of oilfields [21–24]. Among them, machine learning, with its sophisticated algorithms, feature engineering, and unique advantage of processing large-scale data to discover trends in the data as well as patterns of correlation between the data, has achieved numerous applications in the oil and gas industry. Andrei Erofeev et al. (2018) comparatively investigated the applicability of support vector machine, linear regression, and neural network methods in predicting rock properties, which show good predictive and generalization abilities [25]. Yile Ao et al. (2019) compared the predictive ability of eight algorithms in predicting formation property parameters, such as porosity and permeability, and found that linear random forest showed good superiority [26]. Daniel Asante Otchere et al. (2021) concluded that when the dataset is limited, SVM has better reservoir prediction ability than artificial neural networks [27]. In view of this, the application of the random forest regression, support vector machine algorithm, and integrated learning algorithm, combined with a variety of hyper-parameter optimization methods as well as neural network-based and deep learning methods to predict the porosity and permeability have achieved significant prediction results [28–31]. Compared to traditional linear relationship models, machine learning methods can better solve complex multidimensional nonlinear problems, have strong fault tolerance and reliability, and also provide new ideas for predicting porosity and permeability. Therefore, numerous optimization methods and model applications are also an indispensable part of modeling.
The XGBoost algorithm has demonstrated superior performance in prediction and optimization. It has been utilized across numerous applications, where it consistently outperformed other existing algorithms [32–34]. This paper combined logging and core porosity and permeability data from Qingyang Gas Field in the Ordos Basin, learned experimental test data, established the XGBoost porosity and permeability prediction model, and analyzed the applicability for the study area. Then, the innovative application of homogeneous clustering (K-means) method was used to optimize the XGBoost method and to comprehensively determine the XGBoost prediction model for porosity and permeability that is suitable for the stratigraphy in the study area. Finally, the final prediction model of porosity and permeability was applied in the plane to analyze the distribution of porosity and permeability, which provides a reference basis for the exploration and development of favorable zones.

2. Methodology

2.1. Overview of Regional Geology

The Ordos Basin is rich in tight gas resources, and the porosity and permeability of the reservoirs vary widely in different regions of the basin [35]. The Qingyang Gas Field is in the southwest edge of the basin and spans two first-level regional tectonic units of the Yishan Slope and the Tianhuan Depression. The tectonic unit is subdivided as in Figure 1a, which is located in the southern part of the Ordos Basin, and is influenced by the western margin of the alluvial fault zone, and the overall performance is a westward-tilted monoclinic. The reservoir is dominated by deltaic deposits characterized by deep burial, thin sands, rapid vertical changes, and dense low permeability. Therefore, further analysis of the porosity and permeability of the deep, thin, low-permeability sandstone reservoirs is an important reference for the exploration and development of the Qingyang Gas Field and the evaluation of favorable zones.

![Figure 1. (a) Tectonic unit division, (b) well distribution, (c) stratigraphic composite histogram, and (d) single-well lithologic profile in Qingyang Gas Field, Ordos Basin.](image)

The Qingyang Gas Field is characterized by the bottom-up development of sedimentary strata in the Middle and Late Proterozoic, Paleozoic, Mesozoic, and Cenozoic, with...
the Upper Paleozoic Permian Shihezi Formation and Shanxi Formation being the main
target strata for gas exploration and development. The region has a single purposeful strati-
graphic system vertically and a large distribution area in plan. After uplift and denudation
in the Upper Early Paleozoic, the study area began to subside in the Carboniferous, and the
Upper Paleozoic stratigraphy consists of the Carboniferous Benxi Formation, the Permian
Taiyuan Formation, the Shanxi Formation, the Lower Shihezi Formation, the Upper Shi-
hezi Formation, and the Shiqianfeng Formation, from the bottom to the top, as shown in
Figure 1c [36,37]. Among them, the H8 section of the Lower Shihezi Formation and the S1
section of the Shanxi Formation are the main gas-bearing segments in the study area. The
sandstone at the bottom boundary of H8 is “camel neck” shaped and topped with high
gamma mudstone. H8 section III stratigraphic sequence can be divided into two IV strati-
graphic sequences according to the sequence cycle; the sandstone in the lower section of
H8 is dominated by grayish-white, light-gray coarse sandstone, and conglomerate-bearing
coarse sandstone, and with the development of horizontal bedding, part of the sandstone
develops fissures. The type of pores is dominated by intergranular dissolution pores and
lithic clastic dissolution pore. The sedimentary system of braided river delta is mainly
developed in the He8 stage. The S1 section is dominated by gray medium-coarse sandstone
and conglomerate-bearing coarse sandstone, with the development of parallel bedding,
wedge bedding, and oblique bedding; the pore type is dominated by intergranular solution
pores; and the S1 stage is mainly dominated by the Meandering-delta sedimentary system.
The single-well lithological profile is shown in Figure 1d [38,39].

2.2. Method Process

Applying the porosity and permeability laboratory test data of cores and geophysical
logging data, based on the extreme gradient boosting tree (XGBoost) regression prediction
method in machine learning, an XGBoost prediction model of porosity and permeability
of geophysical logging was established, followed by optimization with K-means method,
which was applied to predict the porosity and permeability in the hole section of the study
area, both in the longitudinal direction and in the planar direction. The main processes
included were as follows, and the flowchart is shown in Figure 2:
1. Data preparation;
2. XGBoost model establishment: (1) Logging features and label extraction: feature
variables suitable for machine learning methods were selected and extracted from
the raw data to establish the dataset; (2) Dataset division: the dataset was randomly
divided with 80% as the training set and the remaining 20% as the test set; (3) Feature
combination optimization: the exhaustive feature variable combination method was
used to weaken the problem of multicollinearity between feature variables, and the
feature variable combination method of XGBoost method was preferred to provide a
better feature variable representation for the model; (4) Hyper-parameter optimization:
Random search was used to ensure that the hyper-parameters were searched in a wide
range, and the hyper-parameters were manually adjusted in combination with the grid
search so as to achieve a fine search of the parameters and to improve the stability and
performance of the model; (5) Establishment of machine learning models: porosity
and permeability XGBoost models were established through feature combination
and hyper-parameter optimization steps; (6) Model evaluation: evaluated the model
performance based on MAE and \( R^2 \);
3. K-means optimization model: Based on K-means clustering, similar input data were
grouped to minimize intra-cluster differences and maximize inter-cluster differences,
and the performance of the grouped porosity and permeability XGBoost prediction
model was established and evaluated;
4. Determine the model: The performance of the models was compared in steps (2) and
(3), and the porosity and permeability prediction model that fits the study area was
comprehensively determined;
5. Model application: The final model from step (4) was applied to the layer S13 section in the study area to analyze the planar porosity and permeability distribution.

![Flow chart of XGBoost prediction model for porosity and permeability.](image)

Figure 2. Flow chart of XGBoost prediction model for porosity and permeability.

### 2.3. Data Description

With the core porosity and permeability experimental test data as the target label, the core layer is the section of H8 and S1 in the study area, there are 94 wells and 3389 experimental points, and the logging depth is 3441.43–5119.06 m. We counted the average values of the curves at the same depth of the single-well test points and logging curves and removed some outliers. The distributions of RLLD, RLLS, and RT were more discrete, as shown in Table 1. The distribution of wells is shown in Figure 1b. The laboratory-determined porosity distribution of H8 section was 5.0–10.0% with an average value of 7.0%, the laboratory-determined permeability distribution of H8 section was $0.1 \times 10^{-3} \mu m^2$ with an average value of $0.6 \times 10^{-3} \mu m^2$, the laboratory-determined porosity distribution of S1 section was 4.4–10.0% with an average value of 6.4%, and the laboratory-determined permeability distribution of S1 section was $0.1 \times 1.2 \times 10^{-3} \mu m^2$ with an average value of $0.51 \times 10^{-3} \mu m^2$.

<table>
<thead>
<tr>
<th>Logging Curves</th>
<th>Mean</th>
<th>First Quartile</th>
<th>Median</th>
<th>Third Quartile</th>
<th>Min</th>
<th>Max</th>
<th>Std Deviation</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC (µs/m)</td>
<td>216.20</td>
<td>203.68</td>
<td>212.68</td>
<td>225.65</td>
<td>181.86</td>
<td>379.57</td>
<td>19.17</td>
<td>367.71</td>
</tr>
<tr>
<td>CNL (%)</td>
<td>11.12</td>
<td>6.17</td>
<td>9.27</td>
<td>13.51</td>
<td>0.13</td>
<td>50.59</td>
<td>7.05</td>
<td>49.73</td>
</tr>
<tr>
<td>DEN (g/cm³)</td>
<td>2.53</td>
<td>2.49</td>
<td>2.56</td>
<td>2.62</td>
<td>1.29</td>
<td>2.85</td>
<td>0.16</td>
<td>0.03</td>
</tr>
<tr>
<td>GR (API)</td>
<td>66.96</td>
<td>38.78</td>
<td>59.17</td>
<td>85.31</td>
<td>15.64</td>
<td>437.66</td>
<td>35.80</td>
<td>1282.35</td>
</tr>
<tr>
<td>PE</td>
<td>2.91</td>
<td>2.21</td>
<td>2.55</td>
<td>3.00</td>
<td>0.79</td>
<td>13.70</td>
<td>1.45</td>
<td>2.11</td>
</tr>
<tr>
<td>RLLD (Ω·m)</td>
<td>67.07</td>
<td>28.25</td>
<td>45.88</td>
<td>73.25</td>
<td>5.37</td>
<td>679.66</td>
<td>75.83</td>
<td>5751.95</td>
</tr>
<tr>
<td>RLLS (Ω·m)</td>
<td>60.68</td>
<td>25.65</td>
<td>42.45</td>
<td>68.13</td>
<td>5.51</td>
<td>666.82</td>
<td>66.98</td>
<td>4487.29</td>
</tr>
<tr>
<td>RT (Ω·m)</td>
<td>66.59</td>
<td>28.16</td>
<td>45.96</td>
<td>72.95</td>
<td>0.00</td>
<td>676.38</td>
<td>75.80</td>
<td>5747.06</td>
</tr>
<tr>
<td>SP (MV)</td>
<td>56.52</td>
<td>47.81</td>
<td>57.45</td>
<td>65.93</td>
<td>–7.18</td>
<td>113.31</td>
<td>14.33</td>
<td>205.47</td>
</tr>
<tr>
<td>K (%)</td>
<td>1.32</td>
<td>0.69</td>
<td>1.18</td>
<td>1.79</td>
<td>0.11</td>
<td>4.32</td>
<td>0.77</td>
<td>0.60</td>
</tr>
<tr>
<td>TH (mg/L)</td>
<td>8.91</td>
<td>4.94</td>
<td>7.60</td>
<td>11.62</td>
<td>1.51</td>
<td>42.88</td>
<td>5.15</td>
<td>26.53</td>
</tr>
<tr>
<td>U (mg/L)</td>
<td>2.36</td>
<td>1.36</td>
<td>1.88</td>
<td>2.88</td>
<td>0.22</td>
<td>15.78</td>
<td>1.52</td>
<td>2.31</td>
</tr>
<tr>
<td>POR_L (%)</td>
<td>4.55</td>
<td>2.16</td>
<td>4.69</td>
<td>6.74</td>
<td>0.00</td>
<td>15.20</td>
<td>3.16</td>
<td>9.97</td>
</tr>
<tr>
<td>PERM_L ($10^{-3} \mu m^2$)</td>
<td>0.33</td>
<td>0.01</td>
<td>0.11</td>
<td>0.40</td>
<td>0.00</td>
<td>11.37</td>
<td>0.72</td>
<td>0.52</td>
</tr>
</tbody>
</table>

The maximum–minimum normalization method was applied to the input feature variables and target label values to unify the data range from 0 to 1. This eliminates the
influence of the units of measure of different logs and improves the performance of the model operation, as shown in Equation (1).

\[ x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \]  

(1)

where:

- \( x_i \) is the input feature variable data;
- \( x_i^* \) is the normalized feature variable;
- \( x_{\max} \) is the maximum value of feature variable data;
- \( x_{\min} \) is the minimum value of feature variable data.

The order of magnitude difference is large, and the data transformation using a logarithmic function effectively compresses the data magnitude and improves the stability of the modeling.

\[ x_i^* = \log x_i \]  

(2)

Normalization of the data can also be achieved using the inverse tangent function; when using this method, it should be noted that if the interval you want to map is \([0, 1]\), the data should all be greater than or equal to 0, and data less than 0 will be mapped to the interval \([-1, 0]\).

\[ x_i^* = 2 \arctan x_i / \pi \]  

(3)

Target Labels and Feature Variables

By organizing the logging and recording data in the study area, the target label was the porosity and permeability data tested in the laboratory, and the characteristic variables were derived from the average of the logging series data at the same depth location of the target label. By applying max–min normalization to feature variables and target labels, we found that among them, the medians of the normalized data for acoustic time difference logging (AC), compensated neutron logging (CNL), natural gamma logging (GR), resistivity shallow logging (RLLS), resistivity deep logging (RLLD), resistivity logging (RT), thorium logging (Th), uranium logging (U), measured permeability logging (PERM_L), anyway tangent preprocessing (Arctan(PERM)), and measured permeability (PERM) are distributed between 0 and 0.2; the median distribution of normalized data for caliper logging (CAL), potassium logging (K), measured porosity (POR), and logarithmic preprocessing (LOG(PERM)) ranged from 0.2 to 0.4; and the median distribution of normalized data for photoelectric factor logging (PE), density logging (DEN), spontaneous potential logging (SP), and density logging (DEN) ranged from 0.4 to 0.8, as shown in Figure 3. An amount of 80% (2711) of the feature variable and target label dataset was used as the training set and 20% (678) as the test set for training the predictive model.

The Pearson correlation coefficients between the analyzed feature variables and the target labels are shown in Figure 4. It can be observed that the CNL, DEN, GR, PE, RLLS, RLLD, RT, SP, K, Th, and U logging curves are negatively correlated with the target labels POR and PERM, whereas the AC, CAL, POR_L, and PERM_L logging curves are positively correlated with the target labels POR and PERM. When there is a strong correlation between feature variables, such as feature variables RLLS, RLLD, and RT, considering multicollinearity, only one of them can be selected as the feature variable. When the feature variables are strongly correlated with the target label and can better represent the target label, they are selected as feature variables. The most suitable combination of XGBoost algorithm feature variables is selected using feature variable combination.
The Pearson correlation coefficients between the analyzed feature variables and the target labels are shown in Figure 4. It can be observed that the CNL, DEN, GR, PE, RLLS, RLLD, RT, SP, K, Th, and U logging curves are negatively correlated with the target labels POR and PERM, whereas the AC, CAL, POR_L, and PERM_L logging curves are positively correlated with the target labels POR and PERM. When there is a strong correlation between feature variables, such as feature variables RLLS, RLLD, and RT, considering multicollinearity, only one of them can be selected as the feature variable. When the feature variables are strongly correlated with the target label and can better represent the target label, they are selected as feature variables. The most suitable combination of XGBoost algorithm feature variables is selected using feature variable combination.

Figure 3. Normalized distribution of feature variables and target labels.

Figure 4. Plot of Pearson’s correlation coefficient of feature variables.

2.4. Evaluation Indicators

This paper employed the mean absolute error (MAE) and R-squared score ($R^2$) to assess the performance of porosity and permeability model based on XGBoost algorithm. The objective was to provide guidance on the selection of the combination of feature variables and the hyper-parameter. Additionally, the impact of the K-means method on the XGBoost porosity and permeability model was evaluated.

The mean absolute error (MAE): The average value of the absolute difference between the predicted value and the true value is calculated as shown in Equation (4). This metric measures the average difference between the predicted value and the true value, with smaller values indicating a more accurate prediction.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|$$  (4)
The R-squared score ($R^2$) is also known as the coefficient of determination and is used to measure the extent to which a model explains the variability of the dependent variable, as shown in Equation (5). $R^2$ take values between 0 and 1. The closer $R^2$ is to 1, the better the fit between the predicted and true values, and the closer $R^2$ is to 0, the worse the fit.

$$R^2 = 1 - \frac{\sum(\hat{y}_i - \overline{y}_i)^2}{\sum(y_i - \overline{y}_i)^2}$$  \hspace{1cm} (5)$$

In the above equation, the variables are defined as follows:

- $y_i$ is actual data value.
- $\overline{y}_i$ is actual data average value.
- $\hat{y}_i$ is regression prediction value.

2.5. Principle of Extreme Gradient Boosting Tree

Chen Tianqi designed the extreme gradient boosting tree (XGBoost), and the core of the XGBoost algorithm is the integration of multiple weak learners to build a single strong learner by progressively optimizing the loss function [40,41]. The input samples of each decision tree and its predecessor tree are trained to correlate with the prediction results, and finally, the prediction results of all the decision trees are accumulated as the final prediction results. Assuming that the model itself consists of $N$ decision trees, the training dataset samples are $D = \{(x_i, y_i)\} (i = 1, 2, \ldots, n, x_i \in R^J, y_i \in R)$, which contain $n$ samples and $J$ features, where $x_i$ = $x_{i1}, x_{i2}, \ldots, x_{in}$ denotes the input feature values, $y_i$ denotes the labels, and $\hat{y}_i$ is the predicted value for the $i$th sample, and the model objective function is Equation (6):

$$obj(t) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, g^{(t)}) + \Omega \sum_{j=1}^{t} (f_j), \hspace{1cm} (6)$$

The first term of the objective function represents the loss function of the predicted value of the entire strong learner with respect to the target value. The second term represents the complexity of the $j$ weak learners in the strong learner.

The optimization objective of the model is to minimize the loss function until the predefined stopping condition is reached. This objective is expressed in Equation (7):

$$\arg\min_c \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{y}_i^{(t)}) + \Omega \sum_{j=1}^{t} (f_j), \hspace{1cm} (7)$$

If the previous $t - 1$ regression tree has been trained and the objective function has been added $t$ times, then Equation (8) is obtained:

$$obj(t) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega \sum_{j=1}^{t} (f_j) + \Omega(f_t), \hspace{1cm} (8)$$

Based on the second-order derivative of Taylor’s formula Equation (9), the objective function can be transformed into Equation (10):

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2$$

$$obj(t) \approx \frac{1}{n} \sum_{i=1}^{n} \left[ L(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega \sum_{j=1}^{t} (f_j) + \Omega(f_t)$$ 

$$\hspace{1cm} (10)$$
where \( g_i \) and \( h_i \) are the first-order derivative Equation (11) and the second-order derivative Equation (12) of the loss function, respectively:

\[
g_i = \frac{\partial L(y_i, \hat{y}^{(t-1)})}{\partial \hat{y}^{(t-1)}} \tag{11}
\]

\[
h_i = \frac{\partial^2 L(y_i, \hat{y}^{(t-1)})}{\partial \hat{y}^{(t-1)}} \tag{12}
\]

The complexity of each weak learner of XGBoost is determined by two factors: \( \gamma T \) and \( \lambda \|w\|^2 \), as indicated in Equation (13). \( T \) in \( \gamma T \) denotes the number of leaf nodes in a tree, and in general, the higher the number of leaf nodes, the more complex the tree model is; \( w \) in \( \lambda \|w\|^2 \) denotes the value of a leaf node, and it is the complexity of the predicted value of the weak learner, where \( \gamma \) and \( \lambda \) are hyper-parameters in the model:

\[
\Omega \sum_{j=1}^{t} (f_j) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{t} \|w_j\|^2 \tag{13}
\]

The objective function is Equation (14):

\[
obj(t) = \sum_{i=1}^{n} \left[ g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i) \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{t} \|w_j\|^2 + C \tag{14}
\]

The final objective function is Equation (15):

\[
obj^{(t)} = \sum_{j=1}^{t} \left[ w_j G_j + \frac{1}{2} w_j^2 (H_i + \lambda) \right] + \gamma T \tag{15}
\]

where \( G_i \) and \( H_i \) are Equations (16) and (17):

\[
G_i = \sum_{i \in I_j} g_i \tag{16}
\]

\[
H_i = \sum_{i \in I_j} h_i \tag{17}
\]

From the quadratic Equation of the leaf node \( w \), the optimal value of \( w \) can be obtained as shown in Equation (18):

\[
w_j = -\frac{G_i}{H_i + \lambda} \tag{18}
\]

In the above equation, the variables are defined as follows:
- \( t \) is the \( t \)-th regression tree;
- \( \Omega \) is the complexity of the regression tree to prevent overfitting;
- \( L \) is the loss function.

2.6. Principles of the K-means Method

The K-means algorithm is a clustering algorithm proposed by MacQueen in 1967. The goal is to minimize the variance of the means of the data objects within a group by dividing the \( n \) data objects into \( K \) distinct groups. The K-means clustering method splits the data into different parts with the aim of dividing the data points into clusters, minimizing the intra-cluster data point differences, and maximizing the inter-cluster differences [42]. The K-means optimization objective is given by Equation (19):

\[
E = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - u_i||^2 \tag{19}
\]
where \( u_i \) is Equation (20):

\[
\sum_{x \in C_i} \frac{1}{|C_i|} \cdot x
\]

(20)

3. Results

In order to construct an XGBoost model structure optimized for the K-means method, Python 3.8 software and the Scikit-Learn library were utilized in this article. The train_test_split module was imported to partition the training set and validation set. The GridSearchCV and RandomizedSearchCV modules were employed to achieve 5-fold cross validation and hyper-parameter optimization. The XGBRegression model was utilized to implement the XGBoost algorithm.

3.1. XGBoost Model Evaluation

The optimal combination of feature variables can provide a more accurate representation of the input features. Based on the correlation between feature variables and target labels, AC, CAL, CNL, DEN, SP, RLLS, PE, GR, RLLD, and RT were selected as input feature variables. An exhaustive search method was applied to identify all possible combinations. In this paper, the number of logging feature combinations was pre-set to 1, and 10 combinations of feature variables were searched. Subsequently, the combination of feature variables was input into the XGBoost model, and the minimum absolute error (MAE) and maximum coefficient of determination (\( R^2 \)) were used as the evaluation indices to optimize the feature combination method.

The results of the 1024 feature combination scores are shown in Table 2. The optimal feature combination for the POR model is \([AC, CAL, CNL, DEN, SP, RLLS, PE, GR]\), with a MAE of 0.0075 for the training set and 0.0195 for the test set and an \( R^2 \) of 0.97 for the training set and 0.68 for the test set. The optimal feature combination for PERM is \([AC, CAL, SP, RLLS, PE, GR, RLLD]\), the training set MAE is 0.0003, the test set MAE is 0.0006, the training set \( R^2 \) is 0.99, and the test set \( R^2 \) is 0.31. The optimal feature combination for LOG(PERM) regression is \([AC, CAL, CNL, DEN, SP, RLLS, PE, GR, RLLD]\), the training set MAE is 0.3757, the test set MAE is 0.0920, the training set \( R^2 \) is 0.97 and the test set \( R^2 \) is 0.57. The optimal feature combination for Arctan(PERM) regression is \([CAL, CNL, DEN, SP, PE, GR, RLLD]\), the training set MAE is 0.0932, test set MAE is 0.0244, training set \( R^2 \) is 0.96, and test set \( R^2 \) is 0.50.

<table>
<thead>
<tr>
<th>Model</th>
<th>Feature Combination</th>
<th>MAE of Training Set</th>
<th>MAE of Test Set</th>
<th>( R^2 ) of Training Set</th>
<th>( R^2 ) of Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>POR</td>
<td>[AC, CAL, CNL, DEN, SP, RLLS, PE, GR]</td>
<td>0.0075</td>
<td>0.0195</td>
<td>0.97</td>
<td>0.68</td>
</tr>
<tr>
<td>PERM</td>
<td>[AC, CAL, SP, RLLS, PE, GR, RLLD]</td>
<td>0.0003</td>
<td>0.0006</td>
<td>0.99</td>
<td>0.31</td>
</tr>
<tr>
<td>LOG(PERM)</td>
<td>[AC, CAL, CNL, DEN, SP, RLLS, PE, GR, RLLD]</td>
<td>0.3757</td>
<td>0.0920</td>
<td>0.97</td>
<td>0.57</td>
</tr>
<tr>
<td>Arctan(PERM)</td>
<td>[CAL, CNL, DEN, SP, PE, GR, RLLD]</td>
<td>0.0932</td>
<td>0.0244</td>
<td>0.96</td>
<td>0.50</td>
</tr>
</tbody>
</table>

The RandomizedSearchCV function was used for large-scale and efficient searches based on random search results. Subsequently, the GridSearchCV function was utilized for more precise searches, ensuring a wide search range and accurate search results while reducing computational model runtime. Additionally, the functions were combined with 5-fold cross-validation to evaluate different parameter configurations with the objective of maximizing the model performance. The main hyper-parameters that must be configured to build the regression prediction model using the XGBoost algorithm are shown in Table 3.
Table 3. Description of hyper-parameter optimization of XGBoost model.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Hyper-Parameter Optimization</th>
<th>Hyper-Parameter Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost</td>
<td>learning_rate</td>
<td>The smaller the learning rate, the smaller the impact of each tree, and the more stable the model training</td>
</tr>
<tr>
<td></td>
<td>max_depth</td>
<td>Control the maximum depth of each tree, with small values, making it difficult for the model to overfit</td>
</tr>
<tr>
<td></td>
<td>min_child_weight</td>
<td>Preventing model overfitting on the training set</td>
</tr>
<tr>
<td></td>
<td>n_estimators</td>
<td>The number of decision trees, the more decision trees there are, the better the model performance</td>
</tr>
</tbody>
</table>

With the combination of random search and grid search and based on the optimal combination of features, the results of the hyper-parameter selection are obtained as Table 4. The results show that the optimal feature combination of the POR model is \{AC, CAL, CNL, DEN, SP, RLLS, PE, GR\}. Following the hyper-parameter optimization step, the MAE of the training set is 0.0008, and that of the test set is 0.0671; the $R^2$ of the training set is 0.99, and that of the test set is 0.72. The optimal feature combination of the PERM model is \{AC, CAL, SP, RLLS, PE, GR, RLLD\}, and after the hyper-parameter optimization step, the MAE of the training set is 0.0017, the MAE of the test set is 0.0084, the $R^2$ of the training set is 0.99, and the $R^2$ of the test set is 0.32. The optimal feature combination of the LOG(PERM) model is \{AC, CAL, CNL, DEN, SP, RLLS, PE, GR, RLLD\}, and after the hyper-parameter optimization step, the MAE of the training set is 0.0205, that of the test set is 0.3543, that of the training set $R^2$ is 0.99, and that of the test set $R^2$ is 0.62. The optimal feature combination of the Arctan(PERM) model is \{CAL, CNL, DEN, SP, PE, GR, RLLD\}, and after the hyper-parameter optimization step, the MAE of the training set is 0.0225, and that of the test set is 0.0893; the $R^2$ of the training set is 0.96, and that of the test set is 0.52.

The effect of fitting the model test and prediction sets is shown in Figure 5, where the POR and LOG(PERM) models are insignificantly overfitted and the PERM and Arctan(PERM) models are significantly overfitted.

Table 4. XGBoost model hyper-parameter preference values.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Feature Combination</th>
<th>Hyper-Parameter Values</th>
<th>MAE of Training Set</th>
<th>MAE of Test Set</th>
<th>$R^2$ of Training Set</th>
<th>$R^2$ of Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>POR</td>
<td>{AC, CAL, CNL, DEN, SP, RLLS, PE, GR}</td>
<td>learning_rate: 0.1 max_depth: 50 min_child_weight: 9 n_estimators: 300</td>
<td>0.0008</td>
<td>0.0671</td>
<td>0.98</td>
<td>0.72</td>
</tr>
<tr>
<td>PERM</td>
<td>{AC, CAL, SP, RLLS, PE, GR, RLLD}</td>
<td>n_estimators: 100</td>
<td>0.0017</td>
<td>0.0084</td>
<td>0.99</td>
<td>0.32</td>
</tr>
<tr>
<td>LOG(PERM)</td>
<td>{AC, CAL, CNL, DEN, SP, RLLS, PE, GR, RLLD}</td>
<td>learning_rate: 0.02 max_depth: 90 min_child_weight: 9 n_estimators: 600</td>
<td>0.0205</td>
<td>0.3543</td>
<td>0.97</td>
<td>0.62</td>
</tr>
<tr>
<td>Arctan(PERM)</td>
<td>{CAL, CNL, DEN, SP, PE, GR, RLLD}</td>
<td>learning_rate: 0.02 max_depth: 50 min_child_weight: 9 n_estimators: 300</td>
<td>0.0225</td>
<td>0.0893</td>
<td>0.96</td>
<td>0.52</td>
</tr>
</tbody>
</table>
3.2. K-means Optimized XGBoost Model Evaluation

The K-means clustering method was applied to group the experimental test porosity and permeability data into groups 0 and 1 based on the values of logging feature variables GR and PE in the study area. The distribution of data points was shown in Figure 6, and the feature combination preference and hyper-parameter optimization steps were carried out for different combinations, respectively. The results of the model evaluation were obtained in Tables 5 and 6. The optimal feature combination for the POR_0 group is [AC, CAL, CNL, DEN, SP, RLLS, PE, GR], the training set MAE is 0.0246, the test set MAE is 0.0915, the training set R² is 0.96, and the test set R² is 0.54. The optimal feature combination for the Arctan(PERM)_0 group is [CAL, CNL, DEN, SP, RLLS, PE, GR, RT], with the MAE of the training set being 0.0246, the MAE of the test set being 0.96, and the test set R² being 0.54. The optimal feature combination for the Arctan(PERM)_1 group is [CNL, SP, PE, GR, RT], the training set MAE is 0.0017, the test set MAE is 0.0084, the training set R² is 0.99, and the test set R² is 0.32.

For the permeability model, the inverse tangent permeability model (Arctan(PERM)) is chosen to perform better, and the optimal feature combination of the Arctan(PERM)_0 group is [CAL, CNL, DEN, SP, RLLS, PE, GR], the training set MAE is 0.0744, the test set MAE is 0.0563, the training set R² is 0.99, and the test set R² is 0.84. For the permeability model, the inverse tangent permeability model (Arctan(PERM)) is chosen to perform better, and the optimal feature combination of the Arctan(PERM)_0 group is [CAL, CNL, DEN, SP, RLLS, PE, GR, RT], the training set MAE is 0.0225, the test set MAE is 0.0915, the training set R² is 0.96, and the test set R² is 0.54. The optimal feature combination for the Arctan(PERM)_1 group is [CNL, SP, PE, GR, RT], the training set MAE is 0.0010, the test set MAE is 0.0530, the training set R² is 0.99, and the test set R² is 0.32. For the grouped model, the inverse tangent permeability model (Arctan(PERM)) is chosen to perform better, and the optimal feature combination of the Arctan(PERM)_0 group is [CAL, CNL, DEN, SP, RLLS, PE, GR], the training set MAE is 0.0744, the test set MAE is 0.0563, the training set R² is 0.99, and the test set R² is 0.84. For the permeability model, the inverse tangent permeability model (Arctan(PERM)) is chosen to perform better, and the optimal feature combination of the Arctan(PERM)_0 group is [CAL, CNL, DEN, SP, RLLS, PE, GR, RT], the training set MAE is 0.0225, the test set MAE is 0.0915, the training set R² is 0.96, and the test set R² is 0.54. The optimal feature combination for the Arctan(PERM)_1 group is [CNL, SP, PE, GR, RT], the training set MAE is 0.0010, the test set MAE is 0.0530, the training set R² is 0.99, and the test set R² is 0.32. In comparison to the ungrouped model, the grouped model improves the model’s coefficient of determination and reduces the mean absolute error.

Figure 5. Figure of model test set and prediction set fitting effect.

Figure 6. Data grouping diagram of K-means method.
Table 5. Grouped machine learning feature variable combination preference results.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Group</th>
<th>Feature Combination</th>
<th>MAE of Training Set</th>
<th>MAE of Test Set</th>
<th>R^2 of Training Set</th>
<th>R^2 of Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>POR</td>
<td>0</td>
<td>[AC, CAL, CNL, DEN, SP, RLLS, PE, GR]</td>
<td>0.0744</td>
<td>0.0544</td>
<td>0.98</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>[CAL, CNL, SP, RLLS, PE, GR, RT]</td>
<td>0.0007</td>
<td>0.0563</td>
<td>0.99</td>
<td>0.84</td>
</tr>
<tr>
<td>PERM</td>
<td>0</td>
<td>[AC, CNL, RLLS, RLLD]</td>
<td>0.0023</td>
<td>0.0085</td>
<td>0.98</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>[AC, CAL, DEN, RT]</td>
<td>0.0008</td>
<td>0.0131</td>
<td>0.99</td>
<td>0.79</td>
</tr>
<tr>
<td>LOG(PERM)</td>
<td>0</td>
<td>[CAL, CNL, DEN, SP, RLLS, PE, GR, RLLD]</td>
<td>0.0969</td>
<td>0.3761</td>
<td>0.96</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>[CNL, SP, RLLS, RLLD, RT]</td>
<td>0.0010</td>
<td>0.2493</td>
<td>0.99</td>
<td>0.83</td>
</tr>
<tr>
<td>Arctan(PERM)</td>
<td>0</td>
<td>[CAL, CNL, DEN, SP, RLLS, PE, GR, RT]</td>
<td>0.0246</td>
<td>0.0915</td>
<td>0.96</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>[CNL, SP, PE, GR, RT]</td>
<td>0.0009</td>
<td>0.0530</td>
<td>0.99</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Table 6. Preferred values of hyper-parameters for the grouped XGBoost model.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Group</th>
<th>Feature Combination</th>
<th>Hyper-parameter Values</th>
<th>MAE of Training Set</th>
<th>MAE of Test Set</th>
<th>R^2 of Training Set</th>
<th>R^2 of Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>POR</td>
<td>0</td>
<td>[AC, CAL, CNL, DEN, SP, RLLS, PE, GR]</td>
<td>learning_rate: 0.07 max_depth: 100 min_child_weight: 20 n_estimators: 500</td>
<td>0.0006</td>
<td>0.0685</td>
<td>0.99</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>[CAL, CNL, SP, RLLS, PE, GR, RT]</td>
<td>n_estimators: 17</td>
<td>0.0135</td>
<td>0.0552</td>
<td>0.99</td>
<td>0.85</td>
</tr>
<tr>
<td>PERM</td>
<td>0</td>
<td>[AC, CNL, RLLS, RLLD]</td>
<td>n_estimators: 100</td>
<td>0.0023</td>
<td>0.0085</td>
<td>0.98</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>[AC, CAL, DEN, RT]</td>
<td>n_estimators: 500</td>
<td>0.0008</td>
<td>0.0131</td>
<td>0.99</td>
<td>0.79</td>
</tr>
<tr>
<td>LOG(PERM)</td>
<td>0</td>
<td>[CAL, CNL, DEN, SP, RLLS, PE, GR, RLLD]</td>
<td>learning_rate: 0.01 max_depth: 40 min_child_weight: 9 n_estimators: 800</td>
<td>0.0570</td>
<td>0.3520</td>
<td>0.98</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>[CNL, SP, RLLS, RLLD, RT]</td>
<td>n_estimators: 31</td>
<td>0.0482</td>
<td>0.2494</td>
<td>0.99</td>
<td>0.83</td>
</tr>
<tr>
<td>Arctan(PERM)</td>
<td>0</td>
<td>[CAL, CNL, DEN, SP, RLLS, PE, GR, RT]</td>
<td>learning_rate: 0.1 max_depth: 80 min_child_weight: 15 n_estimators: 100</td>
<td>0.0211</td>
<td>0.0842</td>
<td>0.96</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>[CNL, SP, PE, GR, RT]</td>
<td>n_estimators: 37</td>
<td>0.0009</td>
<td>0.0530</td>
<td>0.99</td>
<td>0.85</td>
</tr>
</tbody>
</table>

4. Discussion

1. Data preprocessing methods for XGBoost prediction model of permeability

In this paper, two experimental permeability data preprocessing methods were explored, and three XGBoost permeability prediction models were established: permeability (PERM), logarithmic preprocessed permeability (LOG(PERM)), and arctangent preprocessed (Arctan(PERM)). Among the ungrouped models, the final model of LOG(PERM) had a training set R^2 of 0.98 and a test set R^2 of 0.68, although there was a certain degree of overfitting. However, the model performance was better than that of PERM and Arctan(PERM). Among the grouped models, the penetration data model constructed using the inverse tangent preprocessing method demonstrated the most favorable performance.

2. Optimization effect of K-means method for XGBoost model

The K-means method of data grouping proved beneficial in improving the performance of the XGBoost model. The experimental porosity and permeability data were grouped by using K-means, and the porosity and permeability grouped XGBoost model reduced the MAE by 0.017 and improved R^2 by 0.15 relative to the ungrouped model, as shown in Figure 7. It is worth noting that the data volume is reduced to some extent by K-means, which increases the risk of model overfitting. However, the overall prediction effect is improved.
3. Effectiveness of XGBoost porosity and permeability modeling application

This paper took the M42 well as an example to analyze the vertical application effect of the established XGBoost porosity and permeability model. Comparing the grouped porosity model and logging curve to calculate the porosity and comparing the grouped permeability model and logging curve to calculate the permeability, the grouped XGBoost porosity and permeability model is closer to the real value, and it can realize the prediction of the whole well section, as shown in Figure 8. Through the above analysis, the grouped porosity model was chosen as the final prediction model for porosity in the study area, and the grouped model was chosen as the final prediction model for permeability. Next, the grouping model predicts single-well porosity and permeability and applies the polynomial interpolation method to obtain the distribution of porosity and permeability in the plane of layer S13 of the established model, as shown in Figure 9.
4. Model Limitations

K-means-based XGBoost porosity and permeability model data volume should be as large as possible to prevent model overfitting and improve the generalization ability of the model. Based on Python 3.8 software and Scikit-Learn library, the XGBoost algorithm supports the advantage of parallel operation, running on eight threads simultaneously, and the model building and running time is about 10 h. The model built has good applicability in the study area, and the ideas and methods are useful in other regions.

5. Conclusions

In this study, the application of logarithmic and inverse tangent methods to preprocess the permeability data improved the accuracy of the permeability XGBoost prediction model. The LOG(PERM) demonstrated an improvement in the $R^2$ of the test set by 0.26 over PERM. Similarly, the Arctan(PERM) improved the $R^2$ of the test set by 0.19 over PERM. Additionally, the degree of training set fitting was reduced by 0.02 and 0.03, respectively, which attenuates overfitting to some extent.

The K-means method has positive implications for the optimization of porosity and permeability XGBoost logging prediction models. The K-means data grouping method was applied to the data input to the machine learning model, which was divided into two groups. The $R^2$ was improved by 0.15 on average, and the MAE was reduced by 0.017 on average when compared with the ungrouped model. This resulted in an improvement in the model prediction ability and a closer alignment of the prediction results with the laboratory test data.

The $R^2$ of the K-means grouped porosity model POR_0/POR_1 test set was 0.73 and 0.85, and the optimal feature combinations were [AC, CAL, CNL, DEN, SP, RLLS, PE, GR] and [CAL, CNL, SP, RLLS, PE, GR, RT]; the $R^2$ of the K-means grouped permeability model Arctan(PERM)_0/Arctan(PERM)_1 test set was 0.58 and 0.85, and the optimal feature combinations were [CAL, CNL, DEN, SP, RLLS, PE, GR, RT] and [CNL, SP, PE, GR, RT]. The accuracy of the single-well application was good relative to the porosity and permeability logging curves, and a full-well section prediction can be realized. Consequently, the K-means optimized porosity and permeability XGBoost model offers a novel approach and a reference point for the efficient and relatively accurate evaluation of porosity and permeability in the study area at a reduced cost. Furthermore, it provides a reference for K-means to combine with other machine learning methods to predict important parameters in the process of oil and gas field development, thereby expanding the scope of the applicability of the method.
Author Contributions: Conceptualization, R.W.; methodology, J.Z.; software, J.Z.; validation, N.F.; formal analysis, N.F.; investigation, J.Z.; resources, A.J.; data curation, A.J.; writing—original draft preparation, J.Z.; writing—review and editing, R.W.; visualization, J.Z.; supervision, A.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Key R & D Plan of Shaanxi Province (key industrial innovation chain (Group)) “Research, development and industrialization promotion of small-molecular recyclable self-cleaning fracturing fluid (Project No.2022ZDLSF07-04”).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

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

Abbreviations

The following abbreviations are used in this manuscript:

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>Acoustic time difference logging</td>
</tr>
<tr>
<td>CAL</td>
<td>Caliper logging</td>
</tr>
<tr>
<td>CNL</td>
<td>Compensated neutron logging</td>
</tr>
<tr>
<td>DEN</td>
<td>Density logging</td>
</tr>
<tr>
<td>GR</td>
<td>Gamma ray logging</td>
</tr>
<tr>
<td>PE</td>
<td>Photoelectric factor logging</td>
</tr>
<tr>
<td>RLLD</td>
<td>Resistivity deep logging</td>
</tr>
<tr>
<td>RLS</td>
<td>Resistivity shallow logging</td>
</tr>
<tr>
<td>RT</td>
<td>Resistivity logging</td>
</tr>
<tr>
<td>SP</td>
<td>Spontaneous potential logging</td>
</tr>
<tr>
<td>K</td>
<td>Potassium logging</td>
</tr>
<tr>
<td>TH</td>
<td>Thorium logging</td>
</tr>
<tr>
<td>U</td>
<td>Uranium logging</td>
</tr>
<tr>
<td>POR_L</td>
<td>Porosity logging</td>
</tr>
<tr>
<td>PERM_L</td>
<td>Permeability logging</td>
</tr>
<tr>
<td>PERM</td>
<td>Measured Permeability</td>
</tr>
<tr>
<td>SH</td>
<td>Shale logging</td>
</tr>
<tr>
<td>SAND</td>
<td>Sand logging</td>
</tr>
<tr>
<td>XGBoost</td>
<td>Extreme gradient boosting</td>
</tr>
</tbody>
</table>

References


23. AlHomadhi, E.S. New correlations of permeability and porosity versus confining pressure, cementation, and grain size and new quantitatively correlation relates permeability to porosity. *Arab. J. Geosci.* 2013, 7, 2871–2879. [CrossRef]
35. Shengli, G.; Jinxia, Y. Exploration of tight oil resources based on stratigraphic paleo-tectonics during hydrocarbon generation in the Ordos Basin, China. *Arab. J. Geosci.* 2019, 12, 387. [CrossRef]
37. Wei, Q.; Cui, G.; Liu, M. Reservoir characteristics and control factors of the Lower Member of the Permian He-8 in the southwestern part of the Ordos Basin. Lithol. Reserv. 2021, 33, 17–25.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.