energies

MDPI

Article

Well-Production Forecasting Using Machine Learning with Feature Selection and Automatic Hyperparameter Optimization

Ruibin Zhu [1,2], Ning Li [2,3], Yongqiang Duan [2,3], Gaofeng Li [2,3], Guohua Liu [2,3], Fengjiao Qu [2,3], Changjun Long [2,3], Xin Wang [2,3], Qinzhuo Liao [1,*] and Gensheng Li [1]

[1] College of Petroleum Engineering, China University of Petroleum-Beijing, Beijing 102249, China; cy1_zrb@petrochina.com.cn (R.Z.); ligensheng@cup.edu.cn (G.L.)
[2] Research Institute of Oil and Gas Technology, PetroChina Huabei Oilfield Company, Renqiu 062552, China; cyy_lin@petrochina.com.cn (N.L.); cy4_dyq@petrochina.com.cn (Y.D.); cyy_ligf@petrochina.com.cn (G.L.); cyy_liugh@petrochina.com.cn (G.L.); yjy_qfj@petrochina.com.cn (F.Q.); cyy_longcj@petrochina.com.cn (C.L.); cyy_wangxin@petrochina.com.cn (X.W.)
[3] Key Laboratory of Low-Permeability and Extra-Low-Permeability Reservoir Stimulation, Renqiu 062552, China
* Correspondence: liaoqz@cup.edu.cn

Abstract: Well-production forecasting plays a crucial role in oil and gas development. Traditional methods, such as numerical simulations, require substantial computational effort, while empirical models tend to exhibit poor accuracy. To address these issues, machine learning, a widely adopted artificial intelligence approach, is employed to develop production forecasting models in order to enhance the accuracy of oil and gas well-production predictions. This research focuses on the geological, engineering, and production data of 435 fracturing wells in the North China Oilfield. First, outliers were detected, and missing values were handled using the mean imputation and nearest neighbor methods. Subsequently, Pearson correlation coefficients were utilized to eliminate linearly irrelevant features and optimize the dataset. By calculating the gray correlation degrees, maximum mutual information, feature importance, and Shapley additive explanation (SHAP) values, an in-depth analysis of various dominant factors was conducted. To further assess the importance of these factors, the entropy weight method was employed. Ultimately, 19 features that were highly correlated with the target variable were successfully screened as inputs for subsequent models. Based on the AutoGluon framework, model training was conducted using 5-fold cross-validation combined with bagging and stacking techniques. The training results show that the model achieved an R^2 of 0.79 on the training set, indicating good fitting ability. This study offers a promising approach for the development of oil and gas production forecasting models.

Keywords: machine learning; production forecast; data preprocessing; principal component analysis; AutoGluon

Academic Editors: Manoj Khandelwal and Fernando Sánchez Lasheras

Received: 4 November 2024
Revised: 28 November 2024
Accepted: 26 December 2024
Published: 30 December 2024

Citation: Zhu, R.; Li, N.; Duan, Y.; Li, G.; Liu, G.; Qu, F.; Long, C.; Wang, X.; Liao, Q.; Li, G. Well-Production Forecasting Using Machine Learning with Feature Selection and Automatic Hyperparameter Optimization. *Energies* **2025**, *18*, 99. https://doi.org/10.3390/en18010099

1. Introduction

In the oil and gas industry, reservoir evaluation is a key parameter for petroleum engineers to make accurate decisions during extraction [1]. The well-production rate is one of the important parameters in reservoir evaluation, which can assist in reservoir modeling and numerical simulation and guide reservoir development strategies [2]. From current technological developments, traditional well-production prediction generally adopts three methods: Arps decline curve prediction, machine learning methods, and combination model prediction [3]. The Arps decline curve prediction method can achieve visualization

of prediction results through software [4]. Machine learning methods have excellent self-learning abilities and diverse algorithms, and their performance largely depends on the quality and quantity of data; however, they can efficiently process large-scale datasets [5]. The combination model prediction method has a wide range of applicability, can reduce systematic errors, and has strong model stability [6]. Arps decline curve prediction includes exponential decline, hyperbolic decline, and harmonic decline [7]. You et al. proposed a method for predicting the economic recoverable reserves of SAGD units through the decline constant method based on the Arps index decline and hyperbolic decline, as well as a new approach for predicting technical recoverable reserves using the intercept divided by slope method [8]. Chen et al. conducted a comparative analysis of three models, Arps, SPED, and MFF, and evaluated the predictive performance of the decline curve model and the combination of the decline curve and data-driven neural network model [9]. The prediction of the Arps decline curve requires a sufficiently long production time for the oil field to detect the trend of production decline, which is suitable for analyzing constant-pressure production situations. Due to its simplification of well assumptions, it cannot accurately estimate the actual well production [10].

Machine learning is an important branch of artificial intelligence (AI) that aims to enable computers to acquire knowledge and experience from data through the learning and automated reasoning of computer systems, and to use this knowledge and experience for pattern recognition, prediction, and decision-making [11]. Machine learning contains a large number of algorithms and models. For well-production prediction, machine learning methods establish a linear relationship between independent variables (geological parameters, production time, etc.) and dependent variables (production) [12]. Hou et al. applied machine learning methods to predict the porosity and permeability of reservoirs. Research has shown that logging parameters have a significant impact on the prediction results of porosity and permeability, and an optimal adaptive model can be selected [13]. In recent decades, some algorithms in machine learning have become increasingly frequent in predicting well production. Among the numerous algorithms in machine learning, ANN stands out because of its unique structure and learning ability, making it an effective means of solving complex problems [14]. Due to the strong nonlinear fitting ability of ANN, its prediction results are usually more accurate than those of traditional linear regression or time series analysis [15]. In production prediction, ANN can automatically learn patterns from historical data and predict future yields based on these patterns [16]. Amr et al. used machine learning methods to simultaneously train oil well productivity for multiple blocks. The monthly oil production is designated as the dependent variable of the model, which improves the accuracy of the prediction. They improved the robustness of the model by increasing the amount of data and studied the impact of the input variables on prediction accuracy [11]. ANN is a very suitable method for production prediction.

In terms of production prediction, in addition to the outstanding performance of ANN, other machine learning algorithms also play an important role. Chakra et al. constructed an oil and gas cumulative production prediction model based on high-order neural networks (HONN). HONN can effectively express the linear and nonlinear relationships between input variables. The model was trained using data from a sandstone reservoir in India, and the trained model was used to predict the cumulative production of oil and gas. Research shows that even in situations where on-site data are relatively insufficient, the model still exhibits good predictive ability and high accuracy [17]. Aizenberg et al. established an oil and gas production prediction model based on a complex neural network (MLMVN) with multi-valued neurons that can accurately achieve the dynamic prediction of oil and gas field production [18]. Shoeibi et al. used long short-term memory (LSTM) networks to predict oil and gas production capacity and validated the effectiveness of the model

through multiple numerical simulations and field data. The research results indicate that the model exhibits good predictive performance in both stable and dynamic production modes [19]. Lolon et al. used a multivariate statistical model to evaluate the relationship between well parameters and production in the Bakken and Three Forks formations. The study found that in tight reservoirs, the total fracturing fluid volume and proppant dose during hydraulic fracturing are the main engineering parameters that affect production [20]. Song et al. conducted a nonlinear analysis of multiple factors using the random forest method for a certain ultra-low-permeability reservoir, identifying the main influencing factors on the initial production capacity of the reservoir. They overcame the limitation of the gray wolf algorithm, which can only identify a single influencing factor using the random forest method. They combined random forest and gray wolf algorithms to predict and optimize the production capacity of oil wells, thereby achieving a fast training speed prediction model [21]. Zhang et al. proposed an unconventional oil and gas production prediction model based on local preserving projection (LPP). This model can accurately capture the nonlinear characteristics of various parameters. Through parameter testing and comparative experiments, research has shown that the LPP-based model has good adaptability and effectiveness [22]. In the same year, Cao et al. used machine learning algorithms to predict production capacity by combining geological data, historical production capacity, and pressure data. He used an artificial multi-layer perceptron to complete two tasks: to predict the future production capacity of existing oil wells based on historical data and to predict the production capacity of new wells based on the historical production of adjacent wells under similar geological conditions [15]. Wang et al. proposed a method that combines multi-layer perceptron (MLP) and long short-term memory (LSTM) networks to predict shale gas production based on geological and fracturing reservoir parameters, historical data, and other information. On the basis of reservoir numerical simulation, they constructed a dataset and trained the model to achieve high accuracy [23]. Rober S et al. developed a complete well-completion design optimization and rapid economic benefit evaluation process based on an ANN. In terms of data processing, they used a self-organizing feature map (SOM) clustering neural network to associate and classify data with reservoir types and completion features [24]. Dong et al. used the XGBoost algorithm to predict the initial production capacity of sandstone reservoirs and improved the model loss function, enhancing the physical constraints of data mining algorithms and significantly improving the prediction accuracy [25]. Khan et al. optimized the fracturing design for heterogeneous oil reservoirs and evaluated its accuracy using various machine learning methods. They found that the XGBoost model performed well in productivity prediction and significantly improved prediction performance [26].

On the basis of machine learning, the combination model prediction method combines the advantages of multiple prediction methods and can combine different machine learning algorithms to improve the accuracy and stability of predictions [11]. Berneti et al. combined the International Accounting Association (ICA) with an Artificial Intelligence Network (ANN) to develop a model for predicting oil well productivity. The model combines the local search function of the BP algorithm and the global search feature of the ICA. Based on the production data of 31 wells, the researchers use the ICA-ANN prediction model to forecast the production capacity and compare it with the ANN prediction results. The results show that the ICA-ANN model has good performance [27]. Aditya Vyas cleverly used machine learning to link the descent curve model with well-completion parameters. They proposed an evaluation criterion by combining the best descent curve and accurate EUR prediction with machine learning, providing a new method for predicting oil well productivity [28]. Noshi et al.'s joint decision-making based on multiple decision trees can effectively handle multidimensional data and has a strong ability to avoid overfitting [29].

Adesina et al. significantly improved the processing efficiency of nonlinear data by mapping data to a high-dimensional space using the kernel function method [30]. Considering that individual algorithms in machine learning have certain errors in yield prediction, the combination model prediction method provides a new approach for improving accuracy. AutoGluon is an open-source framework launched by Amazon with automated machine learning (AutoML) capabilities aimed at simplifying the training and deployment process of machine learning models [31]. AutoGluon utilizes the available computing resources to find the most powerful models in its allocated runtime, providing a user-friendly interface and documentation support. Nick et al. found that multi-layer combinations of many models make better use of allocated training time than finding the best model. Testing 50 classification and regression tasks on Kaggle and OpenML AutoML Benchmarks using AutoGluon showed that AutoGluon is faster, more robust, and more accurate [32].

However, there are significant limitations in the current research on the analysis of the main control factors of production capacity in intelligent production forecasting models. Most studies only analyze engineering or geological parameters separately and fail to achieve an organic combination of the two. This type of single-factor analysis method makes it difficult to fully reflect the production dynamics of oil and gas reservoirs, thereby affecting the accuracy of production capacity prediction. Therefore, there is an urgent need to develop an intelligent prediction model that comprehensively considers both geological and engineering factors to improve the accuracy of production capacity prediction. In addition, most of the current research on capacity prediction models is based on a single model. AutoGluon can automatically combine multiple models and conduct integrated learning, comprehensively utilize the advantages of each model, reduce the limitations of a single model, and enhance the stability of the trained model.

2. Methodology

2.1. Data Introduction

This study collected data from 435 fractured wells in the PetroChina North China Huabei Oilfield, encompassing over 100 characteristic parameters, ranging from quantitative to categorical data. From these parameters, 13 geological parameters and 14 engineering parameters were selected as input features due to their significance and relatively low rates of missing values. The cumulative oil production over 1800 days served as the output target [33]. However, there are issues with missing values, outliers, and non-standard on-site data that need to be addressed through preprocessing steps, such as data imputation and outlier identification.

2.2. Data Governance Process

The main problems with the data include missing data, presence of outliers, and confusion of labeled data. Data governance refers to filling in missing data, filtering out outliers, and standardizing labeled data according to certain processes and methods. The data governance process, also known as the data preprocessing process, can be divided into the following steps:

1. Data exploration and analysis: Data analysis is the process of examining, analyzing, and reviewing data to collect statistical information regarding its quality, starting with an investigation of existing data and its characteristics. Using histograms and box plots, the distribution of the data can be understood, including the shape of the data and the distribution of outliers and missing data.
2. Outlier screening: Based on the conclusions obtained from the data exploration and analysis, select data fields with outliers. Outlier recognition methods are used to

identify and reset outliers to null values, which can subsequently be filled using methods such as mean, mode, median, or quartile values.

3. Missing data processing: Visualize the degree of missing data, explore and analyze the data to understand the types of missing data, and use different filling methods (such as mean, median, or mode filling) based on the reasons and extent of missing data to fill all data, ensuring that there are no missing values in the overall dataset.

These steps are carried out sequentially, starting with data exploration and analysis to understand the distribution of the data. Subsequently, abnormal outlier data are screened out, and the empty outlier data can be processed in subsequent data filling. In addition, before filling in data, all data that needs to be cleaned up or left blank should be processed completely, and then multiple methods should be used to fill in the empty values in the data; After data filling, a dataset without null values was obtained.

2.3. Main Control Factor Analysis

According to the characteristics of the data and the requirements of machine learning algorithms, it is crucial to process the parameters appropriately to enhance the model's performance and effectiveness. Once data governance and target parameter selection are complete, analyzing the relationships between these parameters becomes essential [34]. This analysis aims to identify data fields that exhibit high correlation and strong representativeness with the target parameters. The selected fields are then utilized for tasks such as feature selection and variable screening in modeling, prediction, and classification.

Data correlation analysis refers to a method that calculates the correlation coefficient between two or more variables or employs graphical representations to illustrate the strength and direction of the relationships between these variables. Common techniques include the Pearson correlation coefficient, Spearman correlation coefficient, scatter plots, and heat maps. This type of analysis is invaluable, as it aids in uncovering hidden connections and influencing factors within the data, enabling causal inference, hypothesis testing, and other related work.

In this study, the Pearson correlation coefficient method was used to eliminate linearly irrelevant data, thereby ensuring a correlation between features. Additionally, the study calculated the gray correlation degree, maximum mutual information, AutoGluon feature importance, and SHAP importance. By combining these metrics with the entropy weight method for a comprehensive evaluation, the primary controlling factors were analyzed and ranked in depth.

2.4. AutoGluon

2.4.1. Introduction to AutoGluon

The research method chosen in this study is the combined model prediction method AutoGluon. It is capable of automating the entire process from data preprocessing to model training, hyperparameter optimization, and result evaluation. It is particularly suitable for tabular data, image data, and text data, allowing users to quickly build high-performance machine learning models with minimal coding. The AutoGluon framework can automatically train a variety of models, automatically select hyperparameters, and perform integrated learning on well-performing models. In this process, only hyperparameters need to be set, including the number of layers of the integrated training stack and the number of cross-validation. This makes AutoGluon suitable not only for data scientists but also for developers without an extensive background in machine learning, enabling them to rapidly apply machine learning techniques to real-world problems.

2.4.2. Core Features of AutoGluon

Automation: One of AutoGluon's core strengths lies in its fully automated machine learning workflow. It can automatically select the most suitable algorithms based on the characteristics of the data and perform hyperparameter tuning. This process not only reduces the workload for users but also lowers the requirement for domain expertise, allowing users without a deep background in machine learning to build high-performance models. The automation features of AutoGluon make machine learning more accessible and widely applicable.

Multi-task Support: AutoGluon supports multiple task types, including classification, regression, and time series forecasting. This extensive task support provides users with great flexibility, enabling AutoGluon to handle various data types and application scenarios. Whether dealing with structured tabular data or complex image and text data, AutoGluon offers effective solutions.

Simplicity and Efficiency: AutoGluon offers a simple and intuitive API, allowing users to perform model training and prediction using only a few lines of code. This simplification significantly lowers the entry barrier, enabling both beginners and experts to quickly start and work efficiently. Additionally, the AutoGluon framework utilizes parallel processing and GPU acceleration technologies to significantly enhance the speed of model training and prediction. This efficiency allows AutoGluon to handle large datasets in a relatively short amount of time, saving users' valuable time and computational resources.

2.4.3. Working Principles of AutoGluon

AutoGluon employs intelligent search strategies, such as Bayesian optimization and grid search, to automatically select the best model and optimize its hyperparameters. This intelligent search process ensures that the model's performance on a given dataset is maximized. AutoGluon can efficiently explore and evaluate multiple model configurations to determine the optimal combination, thereby maximizing both prediction accuracy and model stability. Additionally, by using ensemble learning strategies, this approach combines predictions from multiple models employing techniques like bagging and stacking. This not only enhances the model's stability and generalization capabilities, but also compensates for the shortcomings of individual models. Even if a particular model performs poorly under certain conditions, the predictions from other models can still maintain the overall accuracy and reliability of the predictions.

2.4.4. Steps for Using AutoGluon

The use of AutoGluon for model training involves a relatively simple process, making it accessible to users with varying backgrounds. First, users need to prepare the training data, which typically include identifying features and target variables, and configuring relevant parameters based on their project requirements, such as defining the problem type, specifying the model save path, setting training time limits, and selecting other hyperparameters. In practice, users can further enhance the model's generalization capability and stability by choosing a five-fold cross-validation to split the training and validation sets. This method involves dividing the dataset into five parts, with four parts used for training and one for validation in each iteration, thereby effectively reducing overfitting and improving the model robustness. Additionally, users can employ a layer of stacking ensemble learning to combine the predictions of multiple base models and further enhance the overall prediction performance through a meta-model. Figure 1 illustrates the AutoGluon framework process utilizing five-fold cross-validation. This approach leverages the strengths of different models, compensates for the shortcomings of individual models, and significantly improves the final model's accuracy and stability.

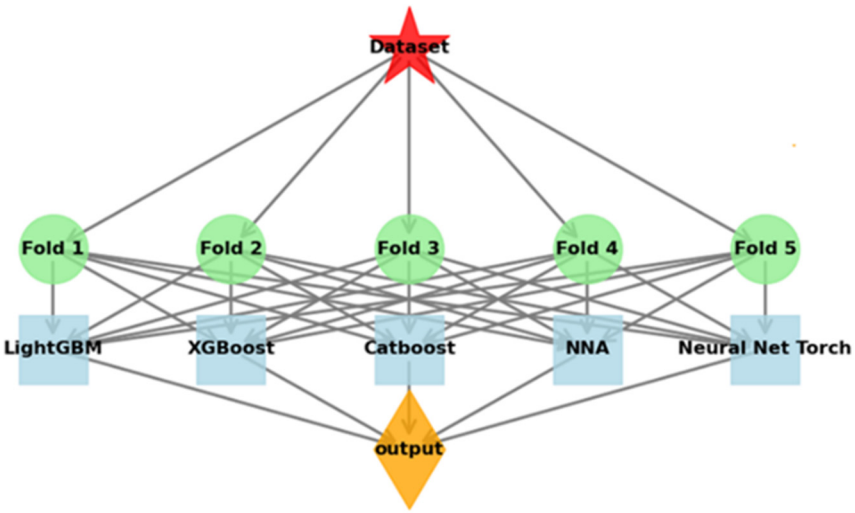

Figure 1. AutoGluon Framework Process Diagram.

Once the configuration is complete, users can initiate the training process. Notably, AutoGluon automatically handles complex operations, such as model selection and hyperparameter tuning, allowing users to focus on analyzing the results without delving into every detail. After the training is complete, users can provide a test set to evaluate the trained model. AutoGluon generates a detailed evaluation report, analyzing the model's performance across various metrics, such as RMSE and R^2. RMSE is the root mean square error, which measures the difference between the predicted value of the model and the true value. The smaller the value, the better the prediction effect of the model. R^2 is the coefficient of determination, a statistic that measures how well a regression model fits. It reflects the correlation between the predicted value of the model and the actual data and represents the proportion of variance that the model can explain. The higher the R^2, the more variability the model can explain in the data, and the better the model fits. This comprehensive evaluation helps users understand the strengths and weaknesses of the model and provides a basis for further improvement.

Through this process, AutoGluon not only simplifies the entire machine learning workflow but also makes the construction of high-performance models more efficient and convenient. In particular, the combination of five-fold cross-validation and stacking methods further enhances the model performance, ensuring good results across various datasets.

2.4.5. Advantages of AutoGluon

In the field of machine learning, as data volume and complexity continue to grow, choosing an appropriate tool for model training has become increasingly important. Compared to individual manual models, AutoGluon offers significant advantages. Firstly, AutoGluon features a high degree of automation, streamlining many machine learning processes. This enables users to easily obtain high-quality models without the need to delve into every detail. This simplified process allows non-expert users to start quickly, significantly reducing the time and cost of model development.

Additionally, by utilizing ensemble learning techniques, AutoGluon combines the strengths of multiple models, thereby significantly enhancing the performance and stability. Compared with a single training model, the AutoGluon framework can be used to integrate multiple models for learning, and cross-validation can improve the stability of the model. This method not only enhances the adaptability of the model to various data types, but also improves the stability of the model. Therefore, selecting AutoGluon for model training not only enhances efficiency but also provides more reliable results, making it my top choice among model training tools.

3. Case Study

This study employs ensemble techniques from hybrid machine learning methods to enhance oil well-production forecasting. Initially, various outlier detection and missing value imputation methods are applied to analyze and preprocess geological and engineering data from oilfield sites, removing outliers and filling in missing values. After data governance and target parameter selection, a relationship analysis is conducted among the parameters to filter out those with high correlations with the target parameters. The study utilizes gray relational analysis, the maximal mutual information method, AutoGluon feature importance evaluation, and SHAP value analysis, combined with the entropy weight method for comprehensive evaluation, to determine appropriate input parameters. Finally, the AutoGluon framework, which integrates multiple models, is used for target output prediction, with RMSE employed to validate the model performance, while MSE, MAE, and R^2 are used to assess the model's performance on both the training and testing datasets. All data have been desensitized for confidential consideration.

3.1. Abnormal Data Situation

By normalizing the on-site data and scaling them uniformly to the range [0, 1], the distribution of the data can be visually observed using a box plot. A box plot is a statistical chart designed to display the dispersion of a dataset, primarily reflecting the characteristics of the original data distribution and allowing for the comparison of distributional characteristics across multiple datasets. In the box plot, the middle line represents the median, while the upper and lower boundaries of the box represent the upper and lower quartiles (75% and 25%, respectively). The outer edges of the plot represent the outlier cutoff points. The distance from the box to the outer edges is typically 1.5 times the interquartile range. Data points within this range are considered inner limits, while data points outside this range are considered outer limits. Generally, data points beyond the outer limits are regarded as extreme outliers (IQR method). Perform normalization on the data and then construct a box plot, as shown in Figure 2.

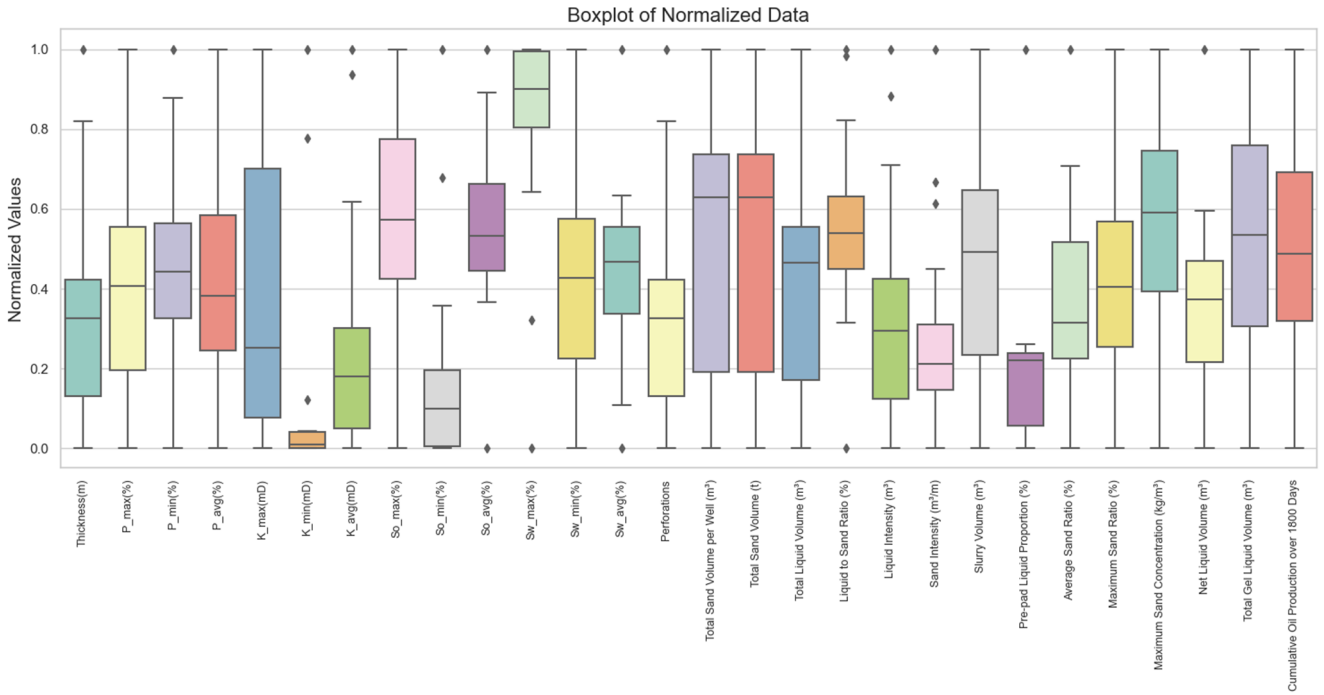

Figure 2. Box distribution graph after data normalization processing.

Due to recording errors, magnitude errors, calculation errors, deduction errors, and other reasons, there are many outliers in the on-site data, often reflected in the form of outlier data. The distribution of the data range is displayed in a histogram [1], as shown in Figure 3.

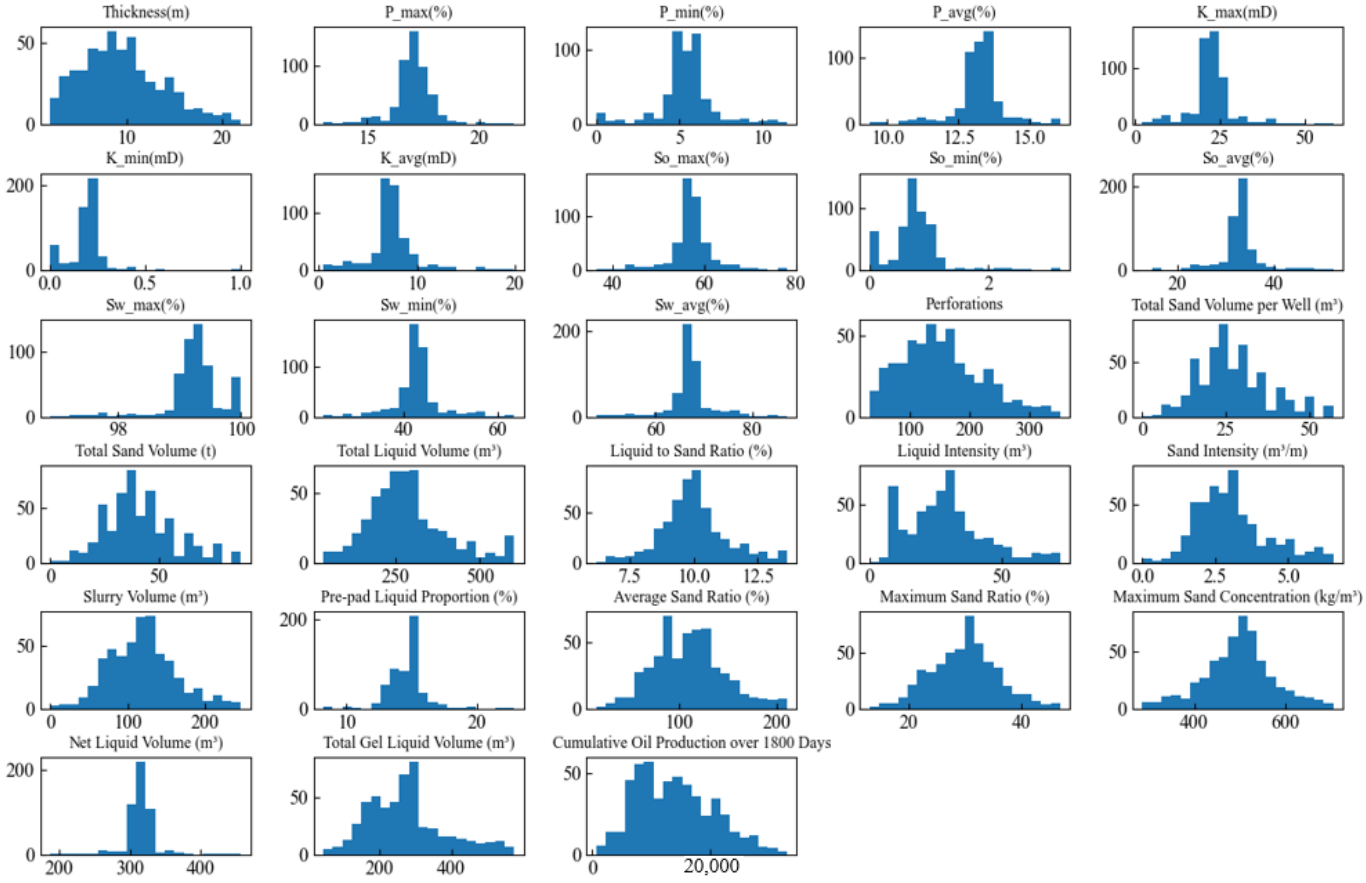

Figure 3. Histogram of the overall distribution of quantitative data.

As can be seen from the graph, the distribution of on-site data is mostly uneven, with many outliers. There are two reasons for determining outliers: one is that there are a large number of 0 s in sparse data, which affects the overall data distribution, and the second reason is abnormal values caused by erroneous construction records, as shown in Figure 4. For these two types of data, they need to be treated differently. Although sparse data are classified as outliers in the mathematical distribution, they do not affect model training, while outliers caused by recording errors can significantly affect the training process and the prediction accuracy of machine learning models. This part needs to be manually corrected.

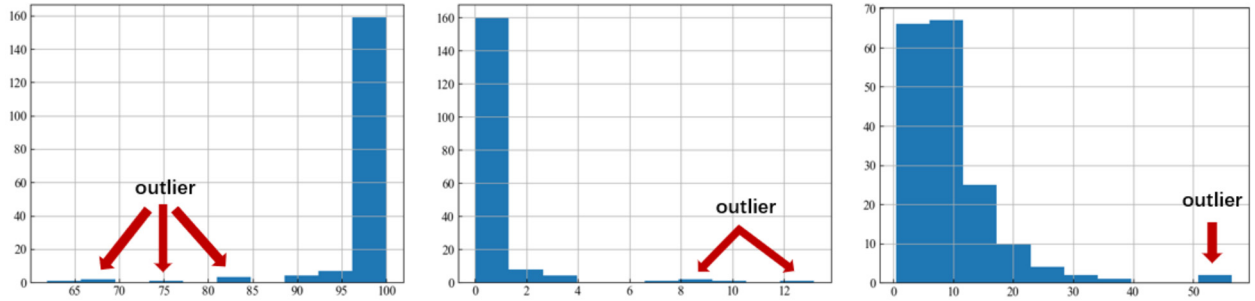

Figure 4. Abnormal value determination.

3.2. Data Preprocessing

3.2.1. Data Missing Situation

In addition to the issue of data anomalies, there are also serious missing data problems on-site. In traditional processing methods for big data, data fields with missing values exceeding 60% are usually deleted directly. However, due to the small amount of on-site data, these fields cannot be deleted. Therefore, special attention should be paid to data filling during data preprocessing.

As can be seen from Figure 5, only nearly 150 wells have complete data, while most wells lack ten to fifteen types of data, resulting in a high data loss rate.

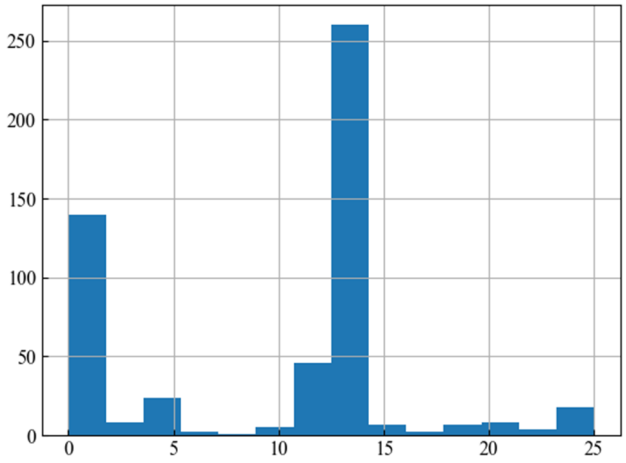

Figure 5. Distribution of missing values in each row of data.

As can be observed from Figure 6, the parameter data statistics are also incomplete, with 38.7% of parameters having a missing rate between 60% and 70%, and 3.2% of parameters having a missing rate exceeding 80%. These include important parameters such as porosity, permeability, and saturation.

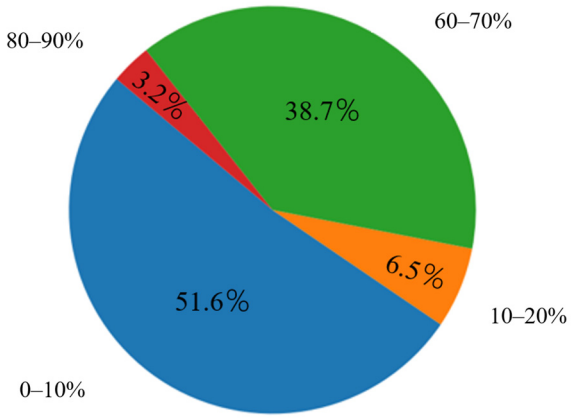

Figure 6. Distribution of missing rate of feature parameter data.

3.2.2. Outlier Handling

The primary foundation of outlier handling is to identify outlier data, and commonly used outlier identification methods include the IQR method, MAD method, and 3sigma method.

The IQR method is based on the lower quartile Q1 and upper quartile Q3 of the dataset for judgment. The size between the upper and lower quartiles is called IQR, while a value that exceeds 1.5 times IQR beyond the upper and lower quartiles is considered an outlier. The dataset is arranged from small to large, with the lower quartile Q1 located in the fourth

quarter and the upper quartile Q3 located in the third quarter. The difference between Q3 and Q1 is the IQR. By using box plots and bar charts, it is possible to visually observe the data points outside the whisker axis, namely, outlier points, as shown in Figure 7. The green dashed line represents the lower boundary of the IQR (Q1 − 1.5 × IQR), and any data below this value are considered outliers. The purple dashed line represents the upper boundary of IQR (Q3 + 1.5× IQR), and any data above this value are also considered outliers. The light blue Kernel Density Estimate (KDE) curve illustrates the smooth probability density distribution of the data, illustrating the central tendency and dispersion of the da-ta values. This curve, combined with the histogram, aids in visually understanding the data distribution and highlighting the location of outliers. The curves depicted in Figures 8 and 9 carry the same significance as the curves in Figure 7. The lower plot in Figure 7 shows the box plot distribution of the data, with points outside the two black lines identified as outliers. The combination of the histogram and box plot provides a better observation of both the quantity and distribution of outliers.

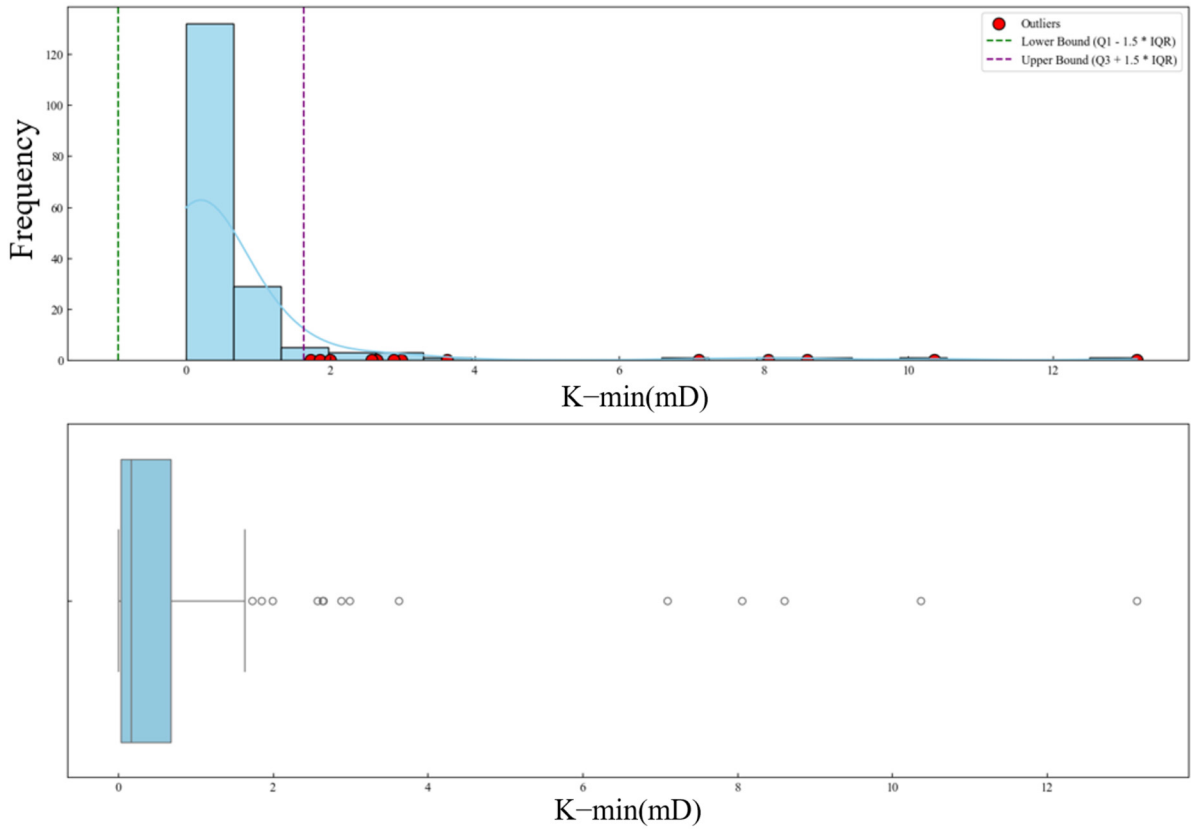

Figure 7. IQR method identifies outliers in the data.

The MAD method assumes that the data follow a normal distribution, and the values within the middle 50% area of the normal distribution are normal values, while the remaining 50% of the area on both sides are outliers. The outlier situation can be viewed in the histogram, and the judgment boundary of the MAD is shown in Figure 8.

The 3 σ method assumes that the data follow a normal distribution, with the mean as the center, and the probability within plus or minus 3 σ is 99.7%. Therefore, the probability of values outside the mean value of 3 σ occurring is 0.3%, which is a very rare and small probability event. Therefore, it can be identified as an outlier, and the outlier situation can be viewed in the histogram. The judgment boundary of 3 σ is shown in Figure 9.

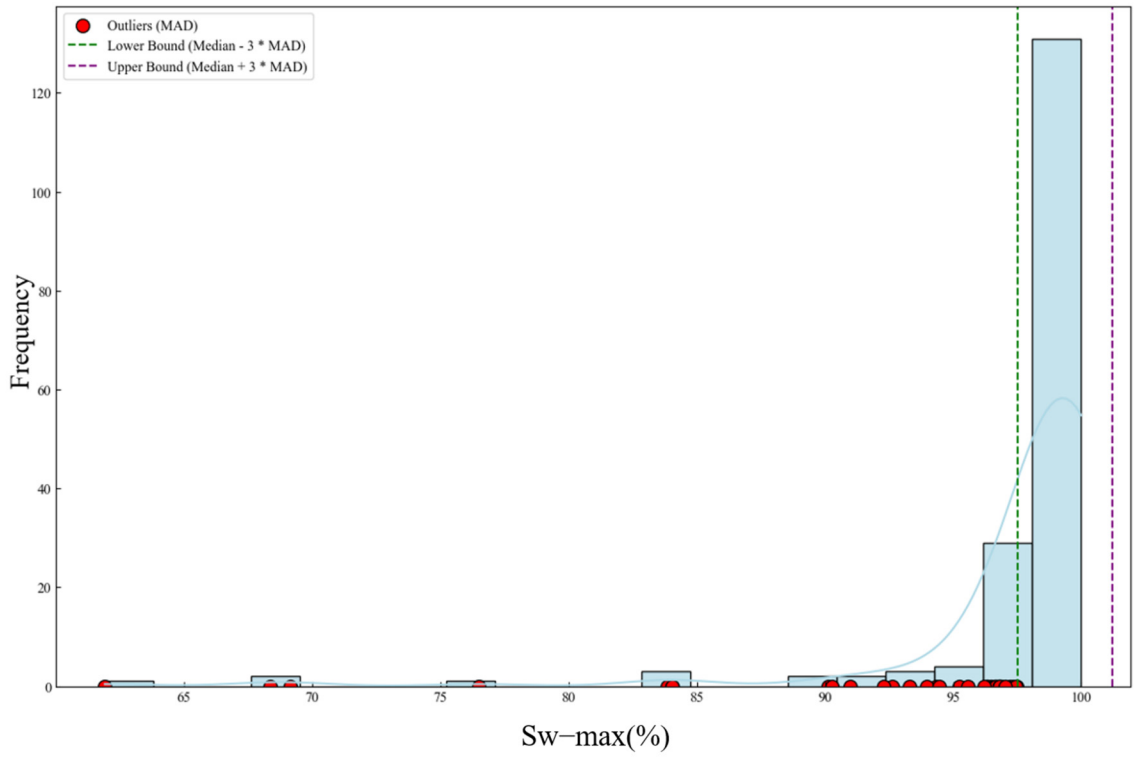

Figure 8. MAD method identifies outliers in the data.

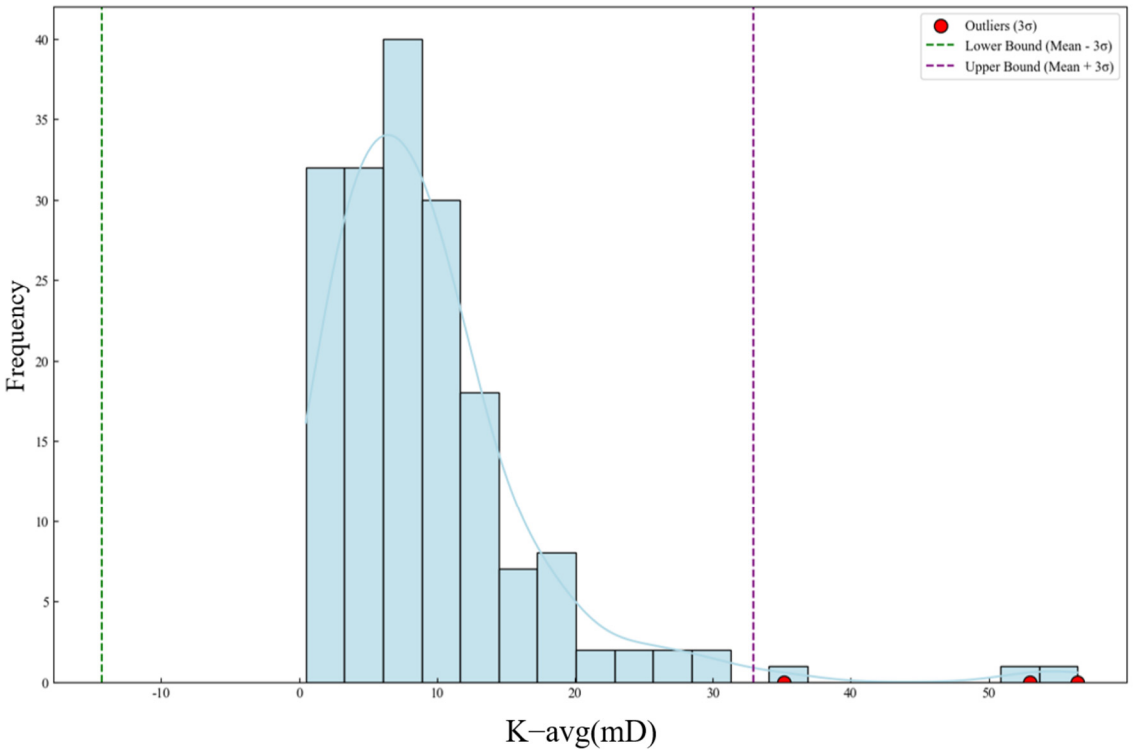

Figure 9. Identifying outliers in data using the triple standard deviation method.

This study adopts the density-based noise unsupervised spatial clustering (DBSCAN) clustering algorithm, which performs multidimensional clustering analysis based on Euclidean distance and automatically and accurately identifies outlier data points as outliers. DBSCAN only requires specifying two parameters: the search neighborhood radius ε and the minimum number of points contained within the search radius MinPts. In the clustering

process, any data point is determined to have neighboring points within its neighborhood radius ε, and points with a number of neighboring points greater than MinPts are identified as core points. Points with fewer neighboring points than MinPts, but adjacent to the core point, are identified as boundary points. The core and boundary points always belong to a certain category cluster, and points that do not belong to any cluster are identified as noise points. The core points and noise points are shown in Figure 10. In the figure, purple, yellow, and green dots represent core points, while black dots indicate noise points, which are also considered outliers.

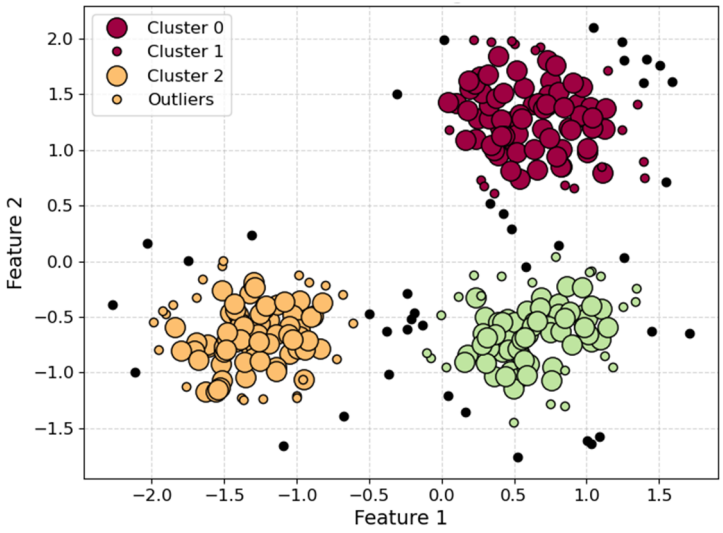

Figure 10. DBSCAN Abnormal value determination.

Choosing an appropriate method is crucial for the accuracy of results in the process of outlier detection. Compared with other commonly used outlier detection methods, such as IQR, MAD, and the triple standard deviation method, the DBSCAN method exhibits more reasonable characteristics in multiple aspects. Firstly, DBSCAN's density-based clustering method can automatically identify clusters and noise points in data without making assumptions about the distribution pattern of the data, while traditional IQR, MAD, and triple standard deviation methods typically assume that the data follows a normal distribution or a relatively uniform distribution, which may not always hold true in practical applications. Secondly, DBSCAN can effectively handle data with nonlinear structures and has stronger adaptability. In addition, DBSCAN does not require a pre-set number of clusters and can automatically determine the number and shape of clusters, which is particularly important for datasets with unknown or complex data features. Taking into account these advantages and data characteristics, the DBSCAN method is more reasonable and effective in outlier detection. Therefore, this method was chosen for this study.

3.2.3. Data Filling

In data analysis and machine learning projects, handling missing values is one of the key steps in ensuring data quality and model performance. If missing values are not handled properly, this may lead to analysis bias and even affect the predictive ability of the model. Therefore, in this project, we adopted a comprehensive strategy that combines mean imputation and the K-Nearest Neighbor (KNN) algorithm to ensure that missing values in the data are properly and effectively filled.

Firstly, we conducted preliminary processing on the geographical features in the dataset. For the missing values in these feature columns, we filled them with column

means. This method can reduce the information loss caused by missing values while preserving the overall trend of the data, providing a relatively complete data foundation for subsequent processing steps.

After completing the preliminary processing, we further used the K-Nearest Neighbor (KNN) algorithm to fill in the missing values in the data. KNN is an instance-based learning method that infers possible missing values by referencing several neighbors in the dataset that are most similar to the missing value samples. We set the number of reference neighbors to 50, which means that when filling in each missing value, the algorithm searches for the 50 most similar samples and infers the missing value based on the data of these neighbors. This method fully utilizes the similarity information in the data, providing more accurate and reasonable filling results, thereby effectively improving the integrity of the data.

After filling in using the above method, we updated the entire dataset to ensure that all missing values were properly processed. The situation before and after missing value processing is shown in Figure 11. In the figure, blue dots represent the original data distribution, while green dots show the data distribution after imputation.

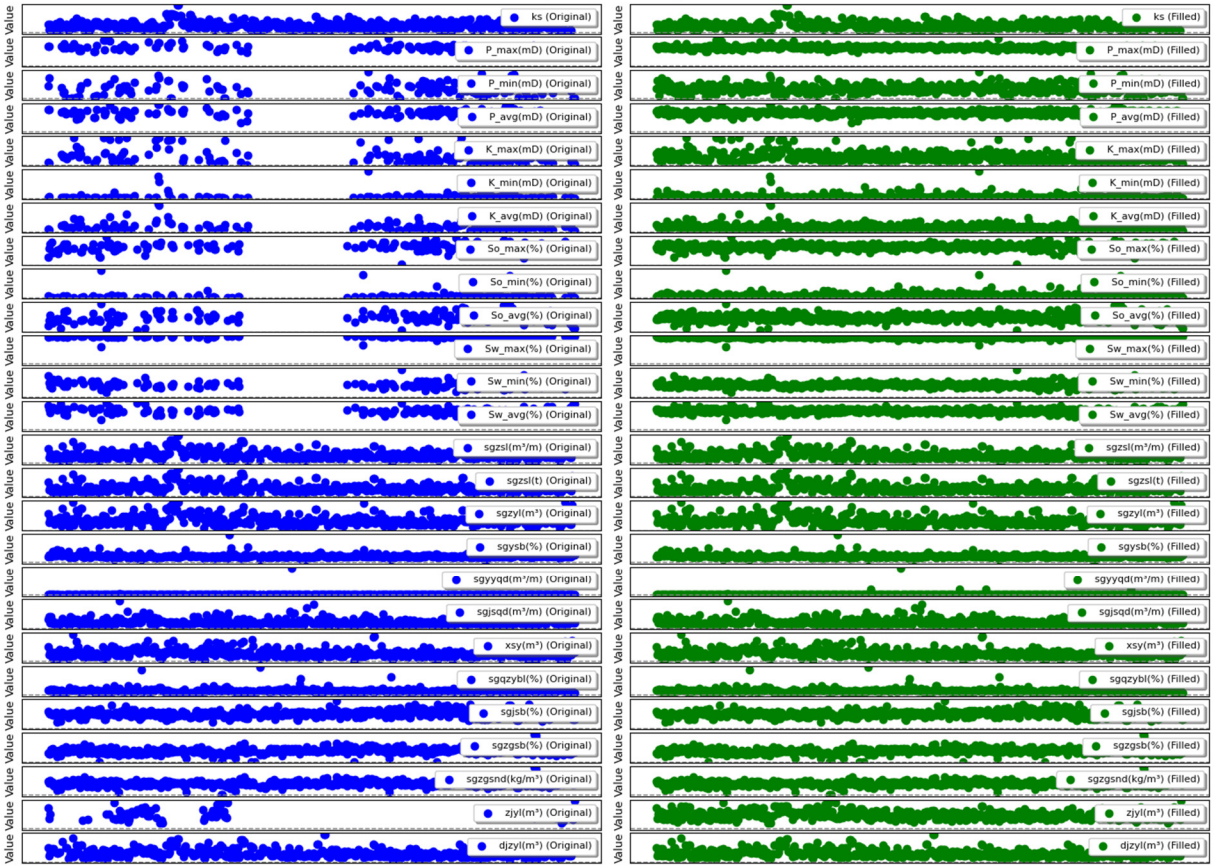

Figure 11. Comparison chart before and after missing value supplementation.

3.3. Feature Extraction

Pearson coefficient analysis was conducted on the extracted geological and engineering data. This analysis involved determining the coefficient value for each pair of parameters. Following the analysis of all parameters, Figure 12 was generated. The colors depicted in the heatmap correspond to the magnitude of the data. Specifically, Pearson coefficients closer to 1 are represented by redder hues in the heatmap, while coefficients nearer to −1 appear bluer [35]. This visual representation provides a clear and intuitive understanding of the correlations present within the dataset.

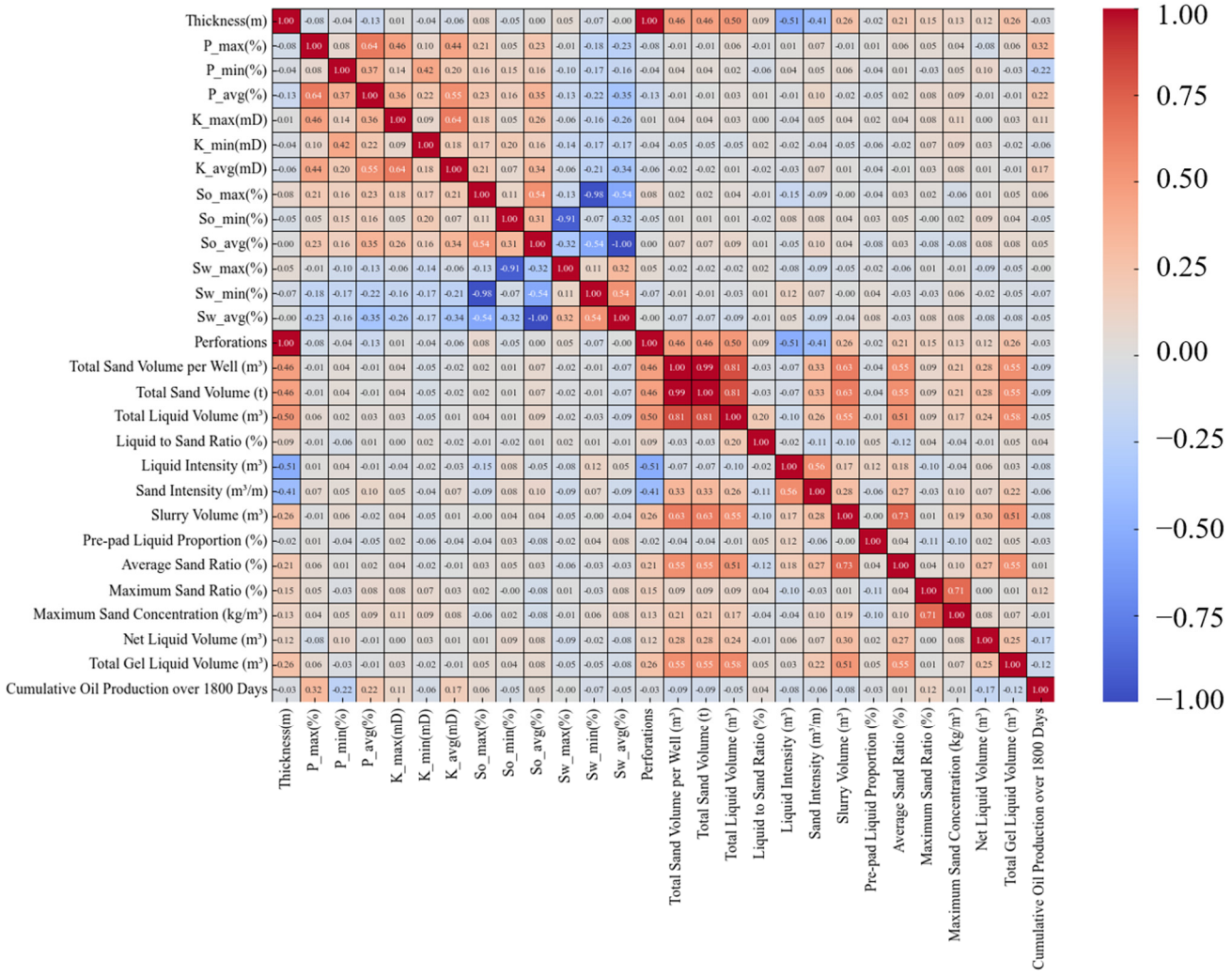

Figure 12. Pearson coefficient heatmap of production and construction data.

When the absolute value of the Pearson coefficient is greater than 0.4, a linear correlation is generally considered. By extracting parameters with Pearson correlation coefficient absolute values greater than 0.4, as shown in Figure 13, a linear relationship test was conducted between each parameter and 1800 days of cumulative oil production. The results show that although there is a certain linear correlation between the construction parameters, geological parameters, and cumulative oil production, this correlation is not strong. This indicates that there may be nonlinear relationships or other complex influencing factors that have not been fully identified. In order to comprehensively understand and evaluate the impact of these parameters on cumulative oil production, this article has decided to introduce more analytical methods. Subsequently, gray relational analysis, the maximum mutual information method, AutoGluon feature importance, and SHAP value analysis will be used, combined with the entropy weight method for comprehensive evaluation, to deeply analyze and determine key influencing factors. These comprehensive methods are expected to reveal more potential relationships and provide a more accurate basis for subsequent optimization decisions.

For the target variable of 1800 days of cumulative oil production, this study comprehensively utilized various methods, such as gray relational analysis, the maximum mutual information method, AutoGluon feature importance assessment, and SHAP value analysis, combined with the entropy weight method, for a comprehensive evaluation. Through this series of analytical methods, the impact of each parameter on the target variable was suc-

cessfully identified and quantified, and the final evaluation results of the main controlling factors are shown in Figure 14.

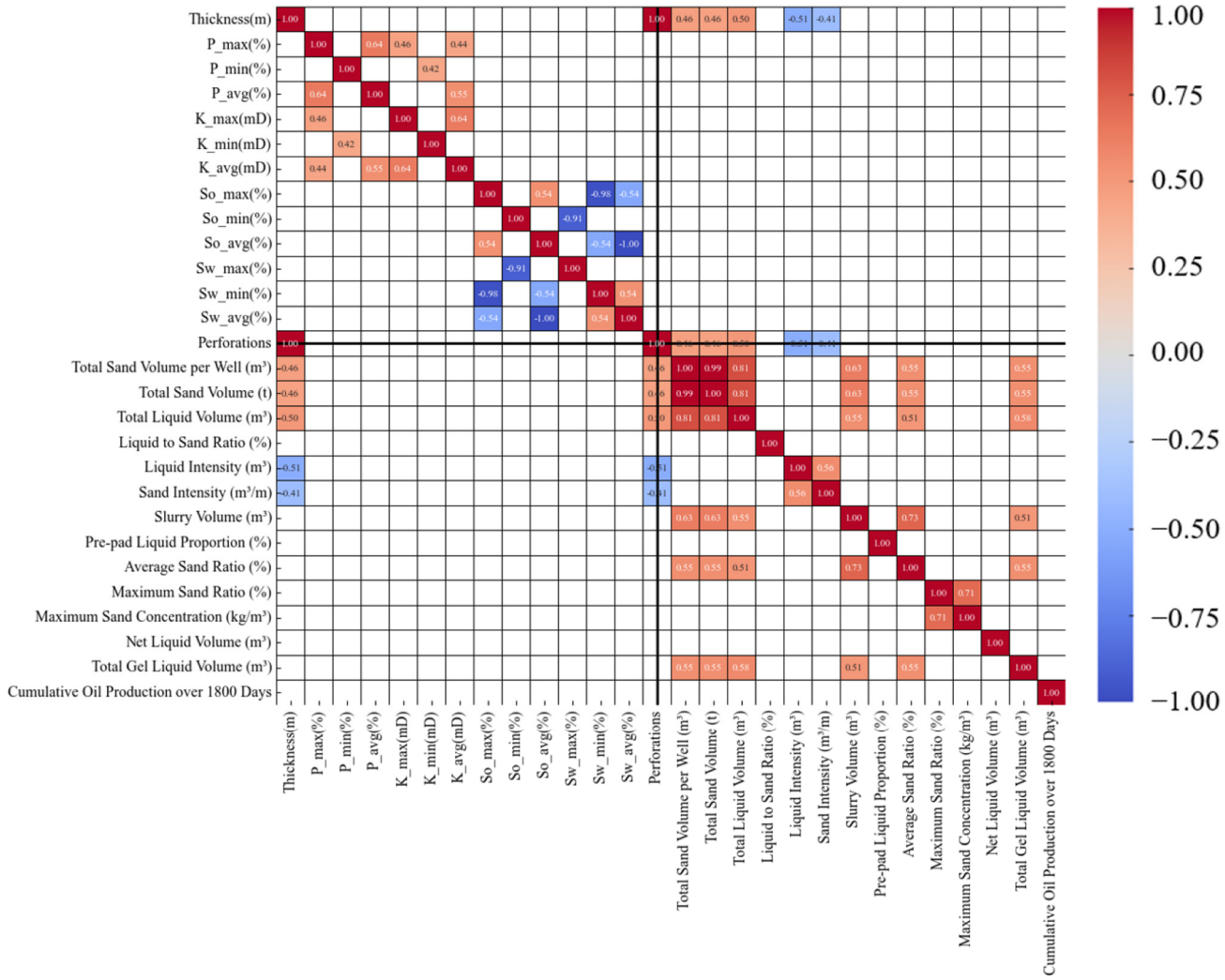

Figure 13. Combination of linear correlation above 0.4 in production and construction data.

In the feature extraction stage, this article selects features from the raw data that are highly correlated with the target parameters. The aim is to reduce the dimensionality of the data, enhance its nonlinearity, and increase its expressive power in order to more effectively adapt it to machine learning models and algorithms. To achieve this goal, this article primarily uses the principal component analysis (PCA) method for feature extraction. This method minimizes information loss by preserving the most representative information in the original data. After applying principal component analysis (PCA), this paper selected 19 geological and engineering parameters from an initial set of 29 variables, using the entropy weight method to rank them by their scores from high to low. These parameters were chosen to serve as input features for subsequent model training. The geological parameters include aspects such as the total gel liquid volume, liquid intensity, liquid-to-sand ratio, pre-pad liquid proportion, total sand volume, average sand ratio, sand intensity, total liquid volume, perforations, and net liquid volume. On the engineering side, the selected parameters include minimum permeability (K_min), average pressure (P_avg), minimum pressure (P_min), maximum oil saturation (So_max), maximum water saturation (Sw_max), average oil saturation (So_avg), minimum water saturation (Sw_min), thickness, and average water saturation (Sw_avg). Among the preferred parameters, Total Gel Liquid

Volume, Liquid intensity, and Total Liquid Volume are positively correlated to the yield, while other preferred parameters are negatively correlated.

Feature Columns	GRA	Maximal Mutual Information	AutoGluon Feature Importance	SHAP Importance	Entropy Weight Method Comprehensive Evaluation
djzyl(m³)	−0.321	0.274	1.865	1.530	1.650
sgyyqd(m³/m)	−0.285	0.388	0.507	0.959	0.474
sgzyl(m³)	0.016	0.339	0.280	0.589	0.281
dty(m³)	−0.371	−0.252	0.181	0.696	0.168
sgzsl(m³/m)	−0.140	0.133	0.006	−0.146	−0.014
sgqzy(m³)	−0.176	0.451	−0.075	0.666	−0.026
sgzgsb(%)	−0.338	0.066	−0.050	0.435	−0.038
P_max(mD)	−0.709	0.007	−0.108	0.422	−0.118
sgzgsnd(kg/m³)	−0.660	−0.356	−0.077	−0.026	−0.123
So_min(%)	1.409	−0.318	−0.270	−0.313	−0.140
K_avg(mD)	0.547	−0.305	−0.265	−0.250	−0.200
K_max(mD)	−0.561	0.423	−0.179	−0.402	−0.217
xsy(m³)	−0.147	0.890	−0.241	−0.535	−0.240
K_min(mD)	1.261	−0.040	−0.381	−0.491	−0.255
P_avg(mD)	−0.274	−0.690	−0.248	−0.491	−0.272
sgysb(%)	0.002	−0.305	−0.318	−0.160	−0.282
sgqzybl(%)	0.077	0.008	−0.365	−0.300	−0.321
sgzsl(t)	−0.140	0.133	−0.332	−0.542	−0.326
P_min(mD)	−0.470	0.079	−0.322	−0.303	−0.328
sgjsb(%)	−0.852	0.810	−0.296	−0.397	−0.334
sgjsqd(m³/m)	−0.220	−0.014	−0.355	−0.283	−0.335
So_max(%)	−0.457	−0.440	−0.379	−0.189	−0.373
Sw_max(%)	1.255	−0.058	−0.510	−0.650	−0.375
So_avg(%)	−0.193	−0.062	−0.430	−0.473	−0.410
Sw_min(%)	−0.579	−0.408	−0.438	−0.044	−0.422
ks	−0.457	0.120	−0.415	−0.729	−0.434
hd(m)	−0.457	0.120	−0.469	−0.547	−0.466
Sw_avg(%)	−0.419	−0.106	−0.537	−0.705	−0.534
zjyl(m³)	0.174	−0.683	−0.597	−0.720	−0.546
Objective Parameter	Cumulative Production over 1800 Days				

Figure 14. Using entropy weight method for comprehensive evaluation and analysis of main control factors.

3.4. Model Training

In this article, the AutoGluon machine learning framework was utilized for rapid model training and optimization, achieving higher accuracy predictions by integrating multiple models without the need for manual hyperparameter search. Specifically, various ensemble learning techniques from AutoGluon, including stacking, k-fold cross-validated bagging, and multi-level stacking, were employed to enhance the model's fitting and generalization performance.

Stacking technology trains multiple models independently on the same dataset and uses linear models to calculate the weighted average of all model predictions. Bagging through k-fold cross-validation effectively prevents model overfitting by performing k-fold cross-validation on all the models and obtaining the average output. Multi-layer stacking combines the original data with the results of single stacking to form a new linear weighted model, further improving the prediction accuracy.

In the specific implementation, the model was trained for regression problems with the key parameters set to optimize the training process. By specifying the model save path, controlling the training time, and utilizing GPU-accelerated computing, the training efficiency of the model was significantly improved. Additionally, 5-fold cross-validation was employed to enhance the model's generalization ability and ensure the effective application of stacking and bagging techniques through appropriate hierarchical control.

Through these methods and parameter settings, AutoGluon was able to automatically select and combine multiple models without human intervention, significantly improving the accuracy and efficiency of predictions.

In the data preprocessing stage, the DBSCAN algorithm is used to identify and remove outliers, and the mean and KNN methods are used to fill in missing values. Subsequently,

the data set was divided into a training set and a test set, with the test set accounting for 20% and not participating in model training. Based on the AutoGluon framework, bagging technology with 5-fold cross-validation and single-layer stacking integrated learning strategy were implemented. The RMSE was selected as the loss function, and the ability of the model to fit the actual data was evaluated using R^2. According to the results shown in Figure 15, the closeness between the predicted value of the model and the real value is measured by the distance between the predicted value and the red dotted line. The closer the distance, the better the prediction effect. The blue marks represent the prediction performance on the training set, the yellow marks represent the prediction performance on the test set, and the points falling in the green area indicate that the prediction accuracy is above 85%. It is worth noting that although the R^2 value on the training set reached 0.79, indicating good model fit, the R^2 value on the test set was only 0.23, which may indicate overfitting of the model.

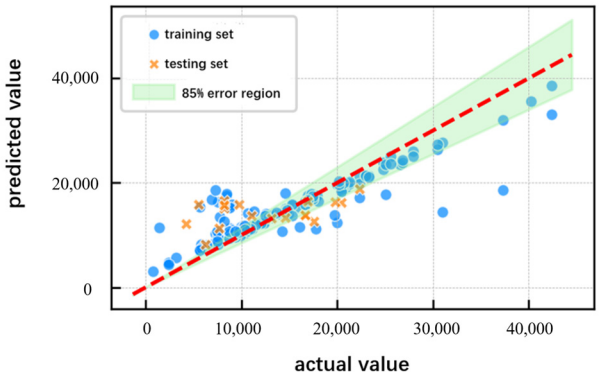

Figure 15. AutoGluon Prediction comparison chart.

4. Conclusions

This study aims to preliminarily construct a production capacity prediction model for a specific oilfield by employing regression analysis methods on a dataset containing 435 records and 19 features, with the goal of predicting the cumulative oil production over 1800 days. In the data preprocessing phase, we utilized a combination of graphical analysis and numerical processing methods. Specifically, we used the K-Nearest Neighbors (KNN) algorithm to impute missing values and employed the entropy weight method to conduct a comprehensive weighting of various key factors, extracting critical features to prepare for model training. Ultimately, we achieved automated model training using the AutoGluon framework, efficiently completing the prediction task.

During the model training process, we applied various algorithms, including Random Forest, CatBoost, Extra Trees, Neural Networks, XGBoost, NeuralNetTorch, and LightGBM. A comparative analysis revealed that the LightGBM model performed the best, with a root mean square error (RMSE) of 4175.08 on the training set and a coefficient of determination R^2 of 0.79, indicating that the model could explain 79% of the data variability. However, during the validation process, especially on the test set, the model's performance significantly declined, with the RMSE increasing to 11,113.33, and the R^2 dropping to 0.23, revealing a clear overfitting issue. This result emphasizes the importance of the model generalization capability. Future research will explore more advanced models and algorithms based on this foundation to further improve the prediction accuracy and reliability, providing more scientific guidance for the overall oil and gas development process.

Author Contributions: Conceptualization, R.Z.; Methodology, N.L.; Software, F.Q.; Validation, G.L. (Gaofeng Li); Formal analysis, X.W.; Resources, G.L. (Guohua Liu); Writing—original draft, C.L.; Visualization, Y.D.; Supervision, G.L. (Gensheng Li); Project administration, Q.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant number 52320105002 and 52374017), and PetroChina Huabei Oilfield Company (grant number HBYT-YQGYY- 2024-JS-28).

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy restrictions from the company.

Conflicts of Interest: Authors Ruibin Zhu, Ning Li, Yongqiang Duan, Gaofeng Li, Guohua Liu, Fengjiao Qu, Changjun Long and Xin Wang were employed by Research Institute of Oil and Gas Technology, PetroChina Huabei Oilfield Company. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Jin, M.; Liao, Q.; Patil, S.; Abdulraheem, A.; Al-Shehri, D.; Glatz, G. Hyperparameter tuning of artificial neural networks for well production estimation considering the uncertainty in initialized parameters. *ACS Omega* **2022**, *7*, 24145–24156. [CrossRef] [PubMed]
2. Ali, N.; Fu, X.; Chen, J.; Hussain, J.; Hussain, W.; Rahman, N.; Iqbal, S.M.; Altalbe, A. Advancing Reservoir Evaluation: Machine Learning Approaches for Predicting Porosity Curves. *Energies* **2024**, *17*, 3768. [CrossRef]
3. Zhang, Z. A well production prediction method based on blending heterogeneous ensemble learning optimized by OOA. *J. Phys. Conf. Ser.* **2024**, *2835*, 012002. [CrossRef]
4. Rushing, J.A.; Perego, A.D.; Sullivan, R.B.; Blasingame, T.A. Estimating reserves in tight gas sands at HP/HT reservoir conditions: Use and misuse of an Arps decline curve methodology. In Proceedings of the SPE Annual Technical Conference and Exhibition, Anaheim, CA, USA, 11–14 November 2007; p. SPE-109625.
5. Nasteski, V. An overview of the supervised machine learning methods. *Horizons* **2017**, *4*, 56. [CrossRef]
6. Li, G.; Tian, Z. A new method of network traffic prediction based on combination model. *Peer-Peer Netw. Appl.* **2024**, *17*, 1075–1090. [CrossRef]
7. Arps, J.J. Analysis of decline curves. *Trans. AIME* **1945**, *160*, 228–247. [CrossRef]
8. Yuanqian, C.; You, Z.; Xiuluan, L. A new method of using SAGD exploitation technique to predict the recoverable reserves of heavy oil reservoir. *Spec. Oil Gas Reserv.* **2015**, *22*, 85–89.
9. Chen, M.; Qu, Z.; Liu, W.; Tang, S.; Shang, Z.; Ren, Y.; Han, J. Production Prediction Model of Tight Gas Well Based on Neural Network Driven by Decline Curve and Data. *Processes* **2024**, *12*, 932. [CrossRef]
10. Wang, K.; Xie, M.; Liu, W.; Li, L.; Liu, S.; Huang, R.; Feng, S.; Liu, G.; Li, M. New Method for Capacity Evaluation of Offshore Low-Permeability Reservoirs with Natural Fractures. *Processes* **2024**, *12*, 347. [CrossRef]
11. Amr, S.; El Ashhab, H.; El-Saban, M.; Schietinger, P.; Caile, C.; Kaheel, A.; Rodriguez, L. A large-scale study for a multi-basin machine learning model predicting horizontal well production. In Proceedings of the SPE Annual Technical Conference and Exhibition, Dallas, TX, USA, 25 September 2018; p. D011S008R005.
12. Wu, T.; Fang, H.; Sun, H.; Zhang, F.; Wang, X.; Wang, Y.; Li, S. A data-driven approach to evaluate fracturing practice in tight sandstone in Changqing field. In Proceedings of the International Petroleum Technology Conference, IPTC, Virtual, 23 March–1 April 2021; p. D071S024R002.
13. Hou, X.M.; Wang, F.Y.; Zai, Y. Prediction of carbonate porosity and permeability based on machine learning and logging data. *J. Jilin Univ.* **2021**, *52*, 644.
14. Surguchev, L.; Li, L. IOR evaluation and applicability screening using artificial neural networks. In Proceedings of the SPE Improved Oil Recovery Symposium, Tulsa, OK, USA, 2–5 April 2000; p. SPE–59308.
15. Cao, Q.; Banerjee, R.; Gupta, S.; Li, J.; Zhou, W.; Jeyachandra, B. Data Driven Production Forecasting Using Machine Learning. In Proceedings of the SPE Argentina Exploration and Production of Unconventional Resources Symposium, Buenos Aires, Argentina, 1–3 June 2016.
16. Jia, X.; Zhang, F. Applying data-driven method to production decline analysis and forecasting. In Proceedings of the SPE Annual Technical Conference and Exhibition, Dubai, UAE, 26–28 September 2016; p. D021S020R007.
17. Chakra, N.C.; Song, K.Y.; Gupta, M.M.; Saraf, D.N. An innovative neural forecast of cumulative oil production from a petroleum reservoir employing higher-order neural networks (HONNs). *J. Pet. Sci. Eng.* **2013**, *106*, 18–33. [CrossRef]
18. Aizenberg, I.; Luchetta, A.; Manetti, S.; Piccirilli, M.C. System identification using FRA and a modified MLMVN with arbitrary complex-valued inputs. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 4404–4411.

19. Loh, K.; Omrani, P.S.; van der Linden, R. Deep learning history matching for real time production forecasting. In *First EAGE/PESGB Workshop Machine Learning*; European Association of Geoscientists Engineers: Utrecht, The Netherlands, 2018; Volume 2018, pp. 1–3.
20. Lolon, E.; Hamidieh, K.; Weijers, L.; Mayerhofer, M.; Melcher, H.; Oduba, O. Evaluating the relationship between well parameters and production using multivariate statistical models: A middle bakken and three forks case history. In Proceedings of the SPE Hydraulic Fracturing Technology Conference and Exhibition, Woodlands, TX, USA, 9–11 February 2016; p. D031S007R003.
21. Song, X.; Liu, Y.; Xue, L.; Wang, J.; Zhang, J.; Wang, J.; Jiang, L.; Cheng, Z. Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model. *J. Pet. Sci. Eng.* **2020**, *186*, 106682. [CrossRef]
22. Zhang, Y.; Hu, J.; Zhang, Q. Application of locality preserving projection-based unsupervised learning in predicting the oil production for low-permeability reservoirs. *Spe J.* **2021**, *26*, 1302–1313. [CrossRef]
23. Wang, T.; Wang, Q.; Shi, J.; Zhang, W.; Ren, W.; Wang, H.; Tian, S. Productivity prediction of fractured horizontal well in shale gas reservoirs with machine learning algorithms. *Appl. Sci.* **2021**, *11*, 12064. [CrossRef]
24. Shelley, R.; Oduba, O.; Melcher, H. Machine learning and artificial intelligence provides wolfcamp completion design insight. In Proceedings of the SPE Hydraulic Fracturing Technology Conference and Exhibition, Houston, TX, USA, 26–28 July 2021.
25. Dong, Y.; Qiu, L.; Lu, C.; Song, L.; Ding, Z.; Yu, Y.; Chen, G. A data-driven model for predicting initial productivity of offshore directional well based on the physical constrained eXtreme gradient boosting (XGBoost) trees. *J. Pet. Sci. Eng.* **2022**, *211*, 110176. [CrossRef]
26. Khan, A.M.; BinZiad, A.; Al Subaii, A.; Alqarni, T.; Jelassi, M.Y.; Najmi, A. Supervised Learning Predictive Models for Automated Fracturing Treatment Design: A Workflow Based on Algorithm Comparison and Multiphysics Model Validation. In Proceedings of the SPE International Hydraulic Fracturing Technology Conference and Exhibition, Muscat, Oman, 11–13 January 2022; p. D011S003R002.
27. Berneti, S.M.; Shahbazian, M. An imperialist competitive algorithm artificial neural network method to predict oil flow rate of the wells. *Int. J. Comput. Appl.* **2011**, *26*, 47–50.
28. Vyas, A.; Datta-Gupta, A.; Mishra, S. Modeling early time rate decline in unconventional reservoirs using machine learning techniques. In Proceedings of the Abu Dhabi International Petroleum Exhibition and Conference, Abu Dhabi, UAE, 31 October–3 November 2017; p. D041S113R002.
29. Noshi, C.I.; Eissa, M.R.; Abdalla, R.M. An intelligent data driven approach for production prediction. In Proceedings of the Offshore Technology Conference, Houston, TX, USA, 6–9 May 2019; p. D041S048R007.
30. Adesina, E.; Olusola, B. Application of Machine Learning Algorithm for Predicting Produced Water Under Various Operating Conditions in an Oilwell. In Proceedings of the SPE Nigeria Annual International Conference and Exhibition, Lagos, Nigeria, 1–3 August 2022; p. D021S009R004.
31. Qi, W.; Xu, C.; Xu, X. AutoGluon: A revolutionary framework for landslide hazard analysis. *Nat. Hazards Res.* **2021**, *1*, 103–108. [CrossRef]
32. Erickson, N.; Mueller, J.; Shirkov, A.; Zhang, H.; Larroy, P.; Li, M.; Smola, A. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv* **2020**, arXiv:2003.06505.
33. Liao, Q.; Zhang, D. Data assimilation for strongly nonlinear problems by transformed ensemble Kalman filter. *SPE J.* **2015**, *20*, 202–221. [CrossRef]
34. Liao, Q.; Zhang, D.; Tchelepi, H. Nested sparse grid collocation method with delay and transformation for subsurface flow and transport problems. *Adv. Water Resour.* **2017**, *104*, 158–173. [CrossRef]
35. Li, D.; You, S.; Liao, Q.; Sheng, M.; Tian, S. Prediction of Shale Gas Production by Hydraulic Fracturing in Changning Area Using Machine Learning Algorithms. *Transp. Porous Med.* **2023**, *149*, 373–388. [CrossRef]