A DATA-DRIVEN WORKFLOW FOR PREDICTION OF FRACTURING PARAMETERS WITH MACHINE LEARNING

by

Zhihua ZHU[a], Maoya HSU[b], Kun DING[a], Tianyu WANG[b*], Xiaodong HE[a], and Shouceng TIAN[b]

[a] Research Institute of Engineering Technology,
PetroChina Xinjiang Oilfield Company, Karamay, China
[b] National Key Laboratory of Petroleum Resources and Prospecting,
China University of Petroleum, Beijing, China

Original scientific paper
https://doi.org/10.2298/TSCI230718029Z

In the realm of unconventional reservoir hydraulic fracturing design, the conventional optimization of mechanistic model parameters is a time-consuming process that impedes its responsiveness to the swift demands of on-site development. This study, rooted in Xinjiang oilfield data, delves into the utilization of machine learning methods for extensive field data. The research systematically elucidates the training and optimization procedures of a production forecasting model, achieving effective optimization of hydraulic fracturing design parameters. By employing polynomial feature cross-construction generate composite features, feature filtering is performed using the maximal information coefficient. Subsequently, wrapper-style feature selection techniques, including ridge regression and decision trees, are applied to ascertain the optimal combinations of model input parameters. The integration of stacking during model training enhances performance, while stratified K-fold cross-validation is implemented to mitigate the risk of overfitting. The ultimate optimization of hydraulic fracturing design parameters is realized through a competitive learning particle swarm algorithm. Results indicate that the accuracy of the data-driven production forecasting model can reach 85%. This model proficiently learns patterns from mature blocks and effectively applies them to optimize new blocks. Furthermore, expert validation confirms that the optimization results align closely with actual field conditions.

Key words: *hydraulic fracturing, machine learning, controlling factors, production forecasting*

Introduction

In the hydraulic fracturing design for unconventional oil reservoirs, significant time is traditionally spent using mechanistic models for parameter optimization, which may not meet the requirement for rapid on-site development. With the accumulation of oilfield data and advancements in machine learning methods, data-driven production prediction models can efficiently achieve fracturing design parameter optimization. In actual oilfield development, researchers face difficulties in acquiring a large number of accurate sample data due to block restrictions, incomplete records, and improper operations. Common data cleaning methods in big data preprocessing, such as deleting samples or features with missing values, removing outliers, and mean (median) filling, do not yield satisfactory results in the presence of data with

* Corresponding author, e-mail: wangty@cup.edu.cn

substantial missing values and noisy data jumps [1]. Therefore, achieving accurate production forecasting results from small sample data has become a focal point for many scholars in recent years, with the primary focus on analysis and prediction of production [2]. Factors influencing oil and gas field production trends include geological parameters such as stress sensitivity, porosity, fracture pressure, and brittleness index, as well as production factors such as pressure, liquid intensity, sand addition intensity, displacement, fracturing segment length, inter-cluster spacing, and nozzle size [3]. Traditional analytical methods have limitations in high dimensional data correlation analysis, while machine learning models demonstrate powerful non-linear fitting capabilities, enabling them to learn complex non-linear relationships [4, 5]. Researchers have employed various machine learning models, including ANN [6], imperialist competitive algorithm (ICA) [7], higher order neural networks (HONN) [8], non-linear autoregressive neural Network (NARX) [9], and multi-valued neuron complex neural network (MLMVN) [10]. Neural network models are mainly applied in the prediction of time series data, such as in the prediction of production data. For instance, the construction of a fusion model of multi-layer perceptron (MLP) networks and long short-term memory (LSTM) networks, utilizing geological and fracturing reservoir parameters, historical data, and other indicators to predict production [11, 12].

This study, based on data from the Xinjiang Oilfield, discusses the application of machine learning methods to large-scale field data. It provides a detailed demonstration of the training and optimization process for a production prediction model. Additionally, the study utilizes the production prediction model to optimize fracturing design parameters.

Material and method

Data collection

The study collected geological, engineering, and production data for hydraulic fractured horizontal wells in the study block from the Xinjiang Oilfield from 2017 to 2022. To meet the optimization requirements, 20 geological and engineering parameters were extracted as feature parameters. The specific parameters are outlined in tab 1. The target parameter is the Production Rate over 330 days [tone per meter], and after preliminary processing, a total of 112 wells were obtained.

Table 1. Geological-engineering parameters

Geological parameters	Engineering parameters
Horizontal section length, [m]	Actual average section spacing, [m]
Type I oil layer, [m]	Actual average cluster spacing, [m]
Type II oil layer, [m]	Construction sand intensity, [m^3m^{-1}]
Type III oil layer, [m]	Construction liquid intensity, [m^3m^{-1}]
$Porosity_{min/avg/max}$, [%]	Construction proppant ratio, [%]
$K_{min/avg/max}$, [mD]	Construction pre-flush liquid ratio, [%]
$So_{min/avg/max}$, [%]	Maximum actual construction discharge, [m^3 per minute]

The probability density distribution of the primary data in the study area is illustrated in fig. 1. For the majority of the data within the block, a symmetrical distribution is observed, including parameters such as horizontal section length, thickness of Type I oil reservoir, permeability, oil saturation, and liquid injection intensity. However, there are some outliers present. A small portion of the data exhibits a skewed distribution, as seen in Type II oil reservoir, porosity,

and thickness with sanding strength. The remaining portion displays a bimodal distribution, such as maximum construction displacement, construction sand ratio, and pre-flush liquid ratio.

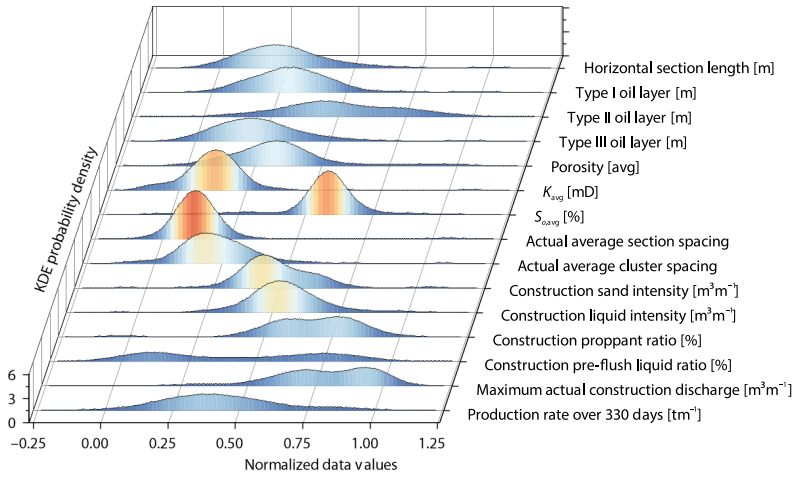

Figure 1. The kernel density distribution plots of key parameters

The distribution pattern of the data indicates a significant presence of artificial interference traces in the samples from the study area, resulting in a distinctive personalized distribution of the data. This may impact the generalization performance of the trained prediction model, rendering it less applicable to other blocks.

Feature engineering

To more effectively capture non-linearity and correlations among feature parameters, we introduce a feature cross method, augmenting the model's comprehensive understanding of data patterns. This polynomial feature cross entails combining two or more features to generate novel features 210 cross-features were generated through the combination of original features. Initially, a filter approach was employed for feature selection, involving the calculation of maximum mutual information between feature variables and the target variable. Features with a mutual information below 0.2, indicating weak correlation, were filtered out. Subsequently, a wrapper approach was implemented for feature evaluation.Various feature subsets were assessed for their importance using linear regression, ridge regression, L1 regularization (Lasso regression), Elastic Net with both L1 and L2 regularization, and decision tree regression models as base models. Different quantities of features were selected as inputs for the final predictive model.

Model training

This study utilized the AutoGluon machine learning framework for rapid model training and optimization, achieving higher accuracy and faster predictions without the need for hyperparameter tuning. The AutoGluon integrates multiple models using fusion techniques such as stacking, K-fold cross-bagging, and multi-layer stacking. Stacking involves independently training multiple models and calculating weighted results through a linear model. K-fold cross-bagging performs cross-validation on all models and averages the outputs, while multi-layer stacking merges data with the results of a single stacking process to form a new linear weighted model. These techniques contribute to improved fitting performance and prevent overfitting. The specific framework of AutoGluon is illustrated in fig. 2.

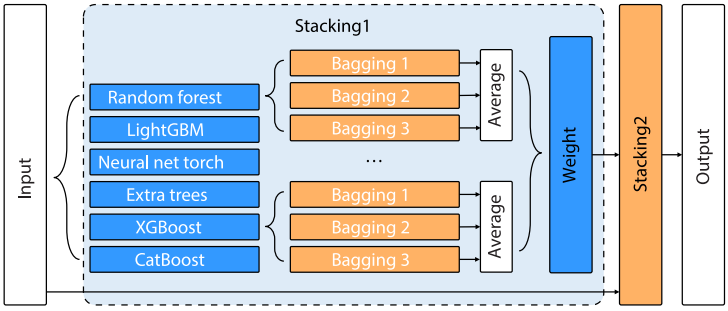

Figure 2. Flowchart of the AutoGluon framework

Result and discussion

The impact of feature combinations on model error was observed, as illustrated in fig. 3. The results indicate that using decision tree regression as the base learner for feature selection is the most stable model. When the top seven features are selected, the overall prediction model achieves the minimum error.

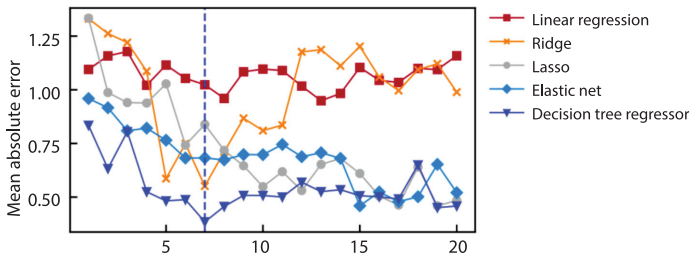

Figure 3. The impact of feature quantities on the predictive performance of the model

The inclusion of cross-features contributes to the reduction of prediction model errors, leading to 4% increase in prediction accuracy, as illustrated in fig. 4.

This study employed simple learners, including KNeighborsUnif, LightGBMXT, LightGBM, RandomForestMSE, CatBoost, ExtraTreesMSE, XGBoost, NeuralNetTorch, and LightGBMLarge. Utilizing stacking technique, the outcomes of simple learners were combined through a weighted layer to form the integrated model named WeightedEnsemble. The average absolute errors of each model in the test and validation sets are compared, as illustrated in fig. 5.

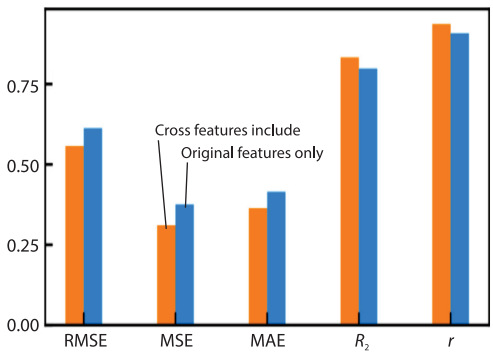

Figure 4. The increase in prediction accuracy by feature cross

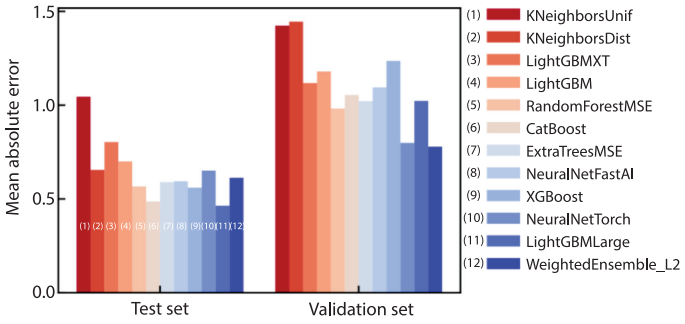

Figure 5. The average absolute errors of different models

The weighted ensemble model, while sacrificing a certain degree of fitting accuracy, achieves optimal predictive performance. This enhances the model's robustness and reduces overfitting.

Conclusion

This study optimized predictive models for hydraulic fractured horizontal wells in Xinjiang Oilfield, utilizing geological, engineering, and production data from 2017 to 2022. Analysis of diverse distribution patterns highlighted potential interference traces, impacting model generalization. Feature engineering, introducing 210 cross-features and employing decision tree regression, improved non-linearity capture. Model training with the AutoGluon framework, featuring stacking techniques, demonstrated high accuracy and rapid predictions without hyperparameter tuning. The inclusion of cross-features significantly reduced prediction errors, leading to a 4% increase in accuracy. Despite sacrificing fitting accuracy, the weighted ensemble model, named WeightedEnsemble, achieved optimal predictive performance, enhancing robustness and reducing overfitting. In summary, the study's comprehensive approach, incorporating advanced techniques in feature engineering, machine learning frameworks, and hybrid optimization, effectively optimized predictive models for hydraulic fractured wells, resulting in improved accuracy and robustness.

Acknowledgment

The authors would like to thank the financial project supported by the National Natural Science Foundation of China (Grant/Award No. 5231001009).

References

[1] Duplyakov, V., *et al*., Practical Aspects of Hydraulic Fracturing Design Optimization Using Machine Learning on Field Data: Digital Database, Algorithms and Planning the Field Tests, *Proceedings*, SPE Symposium: Hydraulic Fracturing, SPE, Virtual, Russia, 2020
[2] Cao, Q., *et al*., Data-Driven Production Forecasting Using Machine Learning, *Proceedings*, SPE Argentina Exploration and Production of Unconventional Resources Symposium, Buenos Aires, Argentina, 2016
[3] Zhou, Q., *et al.*, Evaluating Gas Production Performances in Marcellus Using Data Mining Technologies, *Journal of Natural Gas Science and Engineering*, *20* (2014), 2, pp. 109-120
[4] Lolon, E., *et al.*, Evaluating the Relationship Between Well Parameters and Production Using Multivariate Statistical Models: A Middle Bakken and Three Forks Case History, *Proceedings*, SPE Hydraulic Fracturing Technology Conference and Exhibition, SPE, Woodlands, Tex., USA, 2016
[5] Pankaj, P., *et al*., Application of Data Science and Machine Learning for Well Completion Optimization, *Proceedings*, Offshore Technology Conference, Houston, Tex., USA, 2018
[6] Clar, F. H., *et al*.,Data-Driven Approach to Optimize Stimulation Design in Eagle Ford Formation, *Proceedings*, Unconventional Resources Tech. Conf., Denver, Cal., USA, 2019, pp. 4317-4336
[7] Berneti, S. M., *et al*., An Imperialist Competitive Algorithm Artificial Neural Network Method to Predict Oil Flow Rate of The Wells, *International Journal of Computer Applications*, *26* (2011), 10, pp. 47-50

[8] Chakra, N. C., *et al.*, An Innovative Neural Forecast of Cumulative Oil Production from a Petroleum Reservoir Employing Higher-Order Neural Networks (HONN), *Journal of Petroleum Science and Engineering*, *106* (2013), 2, pp. 18-33
[9] Sheremetov, L., *et al*., Data-Driven Forecasting of Naturally Fractured Reservoirs Based on Non-Linear Autoregressive Neural Networks with Exogenous Input, *Journal of Petroleum Science and Engineering*, *123* (2014), 3, pp. 106-119
[10] Aizenberg, I., *et al*., System Identification Using FRA and a Modified MLMVN with Arbitrary complex-Valued Inputs, *Proceedings*, Int. Joint Conf. on Neural Networks, Vancouver, Canada, 2016, pp. 4404-4411
[11] Wang, T., *et al*., Productivity Prediction of Fractured Horizontal Well in Shale Gas Reservoirs with Machine Learning Algorithms, *Applied Sciences*, *11* (2021), 2, ID12064
[12] Shi, Y., *et al*., Productivity Prediction of a Multilateral-Well Geothermal System Based On a Long Short-Term Memory and Multi-Layer Perceptron Combinational Neural Network, *Applied Energy*, *282* (2021), 2, ID116046

Paper submitted: July 18, 2023
Paper revised: August 28, 2023
Paper accepted: November 21, 2023