Feature investigation on the ROP machine learning model using realtime drilling data

Jie Cao¹²³*, Jiaxuan Gao¹², Tong Jiao¹², Ruiyi Xiang¹², Yanyan Pang¹², Tiantai Li¹²
¹College of Petroleum Engineering, Xi’an Shiyou University, Xi’an, Shan’Xi,710065,China
²Shaanxi cooperative innovation center of unconventional oil and gas exploration and development (Xi’an Shiyou University), Xi’an, Shan’Xi 710065, China
³State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum (Beijing), Beijing 102249, China
*Corresponding author’s e-mail:jie.cao@xsyu.edu.cn

Abstract. To meet the demand for reducing drilling costs in petroleum engineering, improving drilling efficiency is one of the main objectives in field operations. Highly accurate prediction of ROP is an essential basis for improving drilling efficiency and reducing development cycle time. However, realtime prediction of the Rate of Penetration (ROP) is not straightforward, affected by a number of operational and mechanical parameters. The interactions between these parameters also complicated the analysis and modeling ROP. In the presented study, we apply the Feature Engineering approach to analyze the features affecting ROP according to their relevance and relative importance. In addition, the input features are reduced from 14, manually selected based on physical relevance, to optimized 8. Then the model is retrained again for comparing the accuracy of the two prediction models. As a result, it is concluded that by reducing the input of low impact features, the model is substantially simplified, while there are only insignificant accuracy changes.

1. Introduction
In the practical exploration and development process of oil fields, the cost of drilling and completion operation occupies most of the entire investment, as much as to 80%. Therefore, to reduce the cost of drilling operations, the best way is to improve drilling and completion efficiency. Improving ROP is the key means to reduce drilling cycle time, reduce development costs, and improve drilling and completion efficiency. Unfortunately, conventional drilling efficiency improvement methods have almost reached their limits, and hence further improvement, especially the substantial ones, is hard to achieve. The use of machine learning methods for effective ROP prediction and optimization is one of the latest means to break through the bottleneck of drilling speed, which can be an essential guide in the actual drilling process[1-3]. Drilling optimization can be conducted in two aspects with machine learning methods. On the one hand, it is possible to improve the efficiency of use by reasonably deploying materials, equipment, personnel, etc., in advance. On the other hand, it calls for optimizing the drilling parameters to increase the ROP to achieve optimal and fast drilling[4,5].

In general, neural networks have different architectures, unit activation function types, and training algorithms from each other. Artificial Network model structures can be divided into several types. For
example, one is the feedforward neural networks with multilayer perceptrons (MLP), and another one is the feedforward networks with single hidden layer units radial basis functions (RBF), in which the activation function of the RBF is not a logical function but a radial basis function \[6,7\]. The loss functions adopted in the training phase differ in order to reduce losses in training, leading to differences in the mutual adjustment weights and biases (or centers for RBF). The most commonly used one is backpropagation (BP) neural networks, showing the error between the actual output and the predicted output through training \[8\]. The loss functions adopted in the training phase differ in order to reduce losses in training, leading to differences in the mutual adjustment weights and biases (or centers for RBF). In addition to this, there are ANNs with weights and biases randomly to draw, also known as non-iterative algorithms \[9\]. As early as the 1990s, Arehart used artificial neural networks to predict drill wear parameters in oil drilling \[10\]. In 1997, Bilgesu et al. published the first paper on artificial neural networks for predicting mechanical drilling speed \[11\]. More than that, the role of artificial neural networks has gradually expanded in recent years, being used by Wang and Salehi in 2015 to predict the pumping pressure of drilling fluid \[12\], by Khoosravian et al. in 2016 to predict bit weight \[13\], by Ahmadi to predict drilling fluid density \[14\], etc. Figure 1 shows the 70 research works using machine learning models to predict mechanical drilling speeds mentioned in the related articles.

In the drilling process, a large amount of static and dynamic data is generated in realtime. Much helpful information can be obtained by thoroughly analyzing and mining these data, including the prediction of downhole ROP. However, when constructing an ROP prediction model, features have complex relationships, but there may also be features that are not relevant to predicting mechanical drilling speed and even contain factors that can reduce the accuracy of the prediction \[15\]. In the actual drilling process, many low-value or even non-influential feature factors can make the model computationally heavy, reduce the run rate, and degrade the prediction performance. Therefore, in order to improve the prediction performance of the model, enhance the computation speed, reduce the training time, and enable the model to be better understood and maintained \[16\], it is necessary to use feature engineering to filter all the data and extract the features suitable for model building \[17\].

Figure 1.Machine learning methods for mechanical drilling speed prediction (Barbosa et al., 2019)

2. Dataset pre-processing

2.1. Filtering
The dynamic drilling data is mainly obtained from the well site using the drilling equipment and downhole sensors. However, it is common to have the missing section of data points, unusual data points such as outliers, which will lead to inaccuracy when the model is trained. Therefore, it is necessary to pre-process the obtained data first, i.e., data cleaning, divided into three main ways.

2.1.1. Check and compensate missing data. Missing data may occur during data collection, storage, and transmission. Missing data can lead to biased prediction results in practice. Therefore, data elimination is the simplest and most effective way in the case of a small number of missing data. However, for a
large amount of missing data, the interpolation and substitution methods can effectively ensure the integrity of the data while supplementing the missing data.

2.1.2. Check and eliminate outlier data. There are usually two types of data outliers. First, isolated data points, also referred to as “outliers”, are caused by manual errors, equipment affected during information collection, or data processing errors. Second, there is also invalid data, such as invalid values of -999.25. Matching performance of the model is significantly influenced by anomalous data, typically dealt with the same as missing data.

2.1.3. Check and delete overlapping data. Overlapping data collection due to equipment malfunction or different information collection methods must be checked and cleared before modeling.

2.1.4. Data resampling. Different instruments have varying sampling periods, causing that sampling frequency for each parameter is different. Hence, it is necessary to resample all the data to the same frequency. In typical cases, set resampling step/frequency, use interpolation or other methods to achieve the final resampled datasets sampled at the same depth.

2.2. Normalization
Even though the data has been cleaned, the data cannot be put together for comparison because of the large and complex types of data and the different specifications of each feature. First, therefore, these data need to be normalized so that data with different specifications can be transformed into a uniform specification Standard. Second, since the range of values taken for each feature differs significantly, the data should be scaled to reduce the data magnitude errors.

2.2.1. Min-max normalization. This method uses the maximum (Max) and minimum (Min) values of each selected feature and scales the value range of this feature (x) to the range (0,1) by using the difference of the maximum value minus the minimum value as a divisor.

\[ x' = \frac{x - Min}{Max - Min} \]

(1)

2.2.2. Z-score normalization. This method uses the mean \( \bar{x} \) and standard deviation \( \delta \) of the original data to normalize the data.

\[ x' = \frac{x - \bar{x}}{\delta} \]

(2)

2.2.3. Fractional calibration normalization. The value interval of the feature is mapped between [-1, 1] by shifting the number of decimal places of the feature value, and the number of decimal places shifted depends on the maximum value of the absolute value of the feature.

\[ x' = \frac{x}{10^k} \]

(3)

2.2.4. Sigmoid function normalization. The Sigmoid function is centrosymmetric at (0, 0.5) and has a relatively large slope around (0, 0.5). Thus, the image is “S” shaped, and when the data tends to positive and negative infinity, the mapped values tend to 1 and 0 infinitely.

\[ x' = \frac{1}{1 + e^x} \]

(4)
Normalization of data can improve the convergence speed of the model and improve the accuracy of the prediction model, and the appropriate way of normalization can be selected in practical application.

3. Model training
This research uses the simple backpropagation Neural Network (BP) model, which consists of three parts: the output layer, one or more hidden layers, and the output layer. However, it is fairly to use more sophisticated machine learning models, while the feature investigation process described below remains the same. A type of network that learns by being guided by constant errors. Two ways of forward propagation and error backpropagation are used for learning. First, the data enters from the input layer to each implicit layer to calculate the result. The result is transmitted to the output layer to compare with the expected correct result. When the deviation is significant, the output result is back-propagated into the input layer in a specific way and based on the feedback. The weights are re-assigned and re-computed until the error is reduced to a specific range or until a pre-defined number of learning times is reached. All the processed features are input into the BP model, and 70% of the data are randomly selected to train and learn the model, 15% of the data are validated, and then the remaining 15% of the data are tested against the trained model. The final fit of this model to all data is shown in Figure 2 below.

4. Feature investigation
Then, the model is constructed by reducing one feature at a time and is retrained. We define the following equation to quantify the decreased accuracy (relative importance) by omitting the corresponding feature. Here,

$$\eta = R_{\text{full}} - R_{\text{omitted}}$$

and the degree of influence of this feature ($\eta$) on the prediction result of drilling ROP can be obtained. Furthermore, since the training data are randomly selected, the average value is used for five training sessions to reduce errors. The effect of each feature on the prediction results of drilling speed is as follows.

| Omitted feature | $R^2$ (%) | Affected $R^2$ (%) |
|----------------|----------|--------------------|
| None           | 98.203   | 0                  |
| Depth          | 97.5975  | 0.6055             |
| WOH            | 97.652   | 0.551              |
| WOB            | 97.9635  | 0.2395             |
| RPM            | 97.902   | 0.301              |
| Torque         | 97.154   | 1.049              |
| Pump flow      | 97.876   | 0.327              |
| SPP            | 97.8225  | 0.3805             |
| Pump           | 97.918   | 0.285              |

Figure 2. Results of model training, validation, and testing
Given the decreased accuracy of prediction models, we can determine the relative importance of the features and rank them in the following.

| Feature      | Value 1 | Value 2 |
|--------------|---------|---------|
| MW out       | 97.7255 | 0.4775  |
| Bit Time     | 97.5585 | 0.6445  |
| Bit Run      | 97.329  | 0.874   |
| Pump time    | 97.457  | 0.746   |
| Flow Out     | 97.5335 | 0.6695  |
| DEXPONENT    | 96.2975 | 1.9055  |

Figure 3. Feature ranking based on the relative importance

Based on the feature importance, we select the most critical features for retraining and compare the accuracy with the original model. As shown in Figure 5, the total accuracy with all data is 0.9813, only slightly less than that from the original model. The parameters in use in only around half of the original model.

Figure 4. Model retraining with optimized parameters

5. Conclusion
This research demonstrated that screening the features before model construction and deployment reduced the complexity of the model to a certain extent. This can facilitate the understanding and maintenance of the personnel at a later stage and fully capture the key influencing the drilling speed, resulting in less overfitting and lower cost in conducting data information collection. In addition, the amount of data to be processed is significantly reduced, while there is only a slight impact on the accuracy of the predictions.

Acknowledgments
The project is supported by the National Natural Science Foundation (No. 52004214), the Foundation
of State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum, Beijing (PRP/open-2010), China Postdoctoral Science Foundation (2021M693493), Natural Science Foundation of Shaanxi Province (No.2021JQ-598), the Youth Innovation Team of Xi’an Shiyou University (No. 2019QNKYCTXD04), and the Graduate Student Innovation and Practical Ability Training Program of Xi’an Shiyou University [Grant No. YCS20212079].

References
[1] Pollock J, Stoecker-Sylvia Z, Veedu V, et al. (2018) Machine learning for improved directional drilling In: Offshore Technology Conference. Houston. pp. OTC-28633-MS.
[2] Hegde C, Gray K. (2018) Evaluation of coupled machine learning models for drilling optimization. Journal of Natural Gas Science and Engineering., 56: 397-407.
[3] Hegde C, Gray K E. (2017) Use of machine learning and data analytics to increase drilling efficiency for nearby wells. Journal of Natural Gas Science and Engineering., 40: 327-335.
[4] Eren, T. (2010) Real-time-optimization of Drilling Parameters during Drilling Operations. In: Paper presented at the SPE Oil and Gas India Conference and Exhibition. Mumbai, India. pp. SPE-129126-MS.
[5] Payette, G.S., Spivey, B.J., Wang, L., Bailey, J.R., Sanderson, D., Kong, R., Pawson, M., Eddy, A. (2017) Realtime well-site based surveillance and optimization platform for drilling: technology, basic workflows and field results. In: SPE/IADC Drilling Conference and Exhibition. Society of Petroleum Engineers. Hague, The Netherlands. pp. IPTC-18736-MS.
[6] Abramson N., Braverman D J., Sebestyen G S. (2006) Pattern Recognition and Machine Learning. Publications of the American Statistical Association., 103(4):886-887.
[7] Broomhead, D.S., Lowe, D.(1988) Multivariable functional interpolation and adaptive networks. Complex Systems., 2: 321-355.
[8] Barbosa L F F M, Nascimento A, Mathias M H, et al. (2019) Machine learning methods applied to drilling rate of penetration prediction and optimization-A review. Journal of Petroleum Science and Engineering., 183: 106332.
[9] Cao, W., Wang, X., Ming, Z., Gao, J. (2018) A review on neural networks with random weights. Neurocomputing., 275: 278-287.
[10] Arehart, R. (1990) Drill-bit diagnosis with neural networks. SPE Comp App 2.,24-28.
[11] Bilgesu, H., Tetrick, L., Altmiss, U., Mohaghegh, S., Ameri, S. (1997) A new approach for the prediction of rate of penetration (ROP) values. In: SPE Eastern Regional Meeting. Lexington, Kentucky. pp. SPE-39231-MS.
[12] Wang, Y., Salehi, S. (2015) Application of realtime field data to optimize drilling hydraulics using neural network approach., Energy Resour. Technol., 137(6): 62903.
[13] Khosravanian, R., Sabah, M., Wood, D.A., Shahryari, A. (2016) Weight on drill bit prediction models: Sugeno-type and Mamdani-type fuzzy inference systems compared., Journal of Natural Gas Science and Engineering., 36: 280-297.
[14] Ahmadi, M.A. (2016) Toward reliable model for prediction Drilling Fluid Density at wellbore conditions: a LSSVM model. Neurocomputing.,211: 143-149.
[15] Soares, C., Gray, K.(2019) Realtime predictive capabilities of analytical and machine learning rate of penetration (ROP) models. Journal of Petroleum Science and Engineering.,172: 934-959.
[16] James, G., Witten, D., Hastie, T., Tibshirani, R.(2013) An Introduction to Statistical Learning. Springer, New York.
[17] Jorden, J.R. and Shirley, O.J. (1966) Application of Drilling Performance Data to Overpressure Detection. J. Pet. Technol., 18 (11): 1387–1394.