Appling Gaussian Process Regression to Predict Drilling Rate of Penetration with Hyperparameter Optimization using Bayes algorithm: A case study of FORGE site*

Kanghui Zeng[1][00009−0002−2449−8379], Second Author[2,3][1111−2222−3333−4444], and Third Author[3][2222−−3333−4444−5555]

[1] China University of Geoscience
[2] Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany
zkh@cug.edu.cn
http://www.springer.com/gp/computer-science/lncs
[3] ABC Institute, Rupert-Karls-University Heidelberg, Heidelberg, Germany
{abc,lncs}@uni-heidelberg.de

Abstract. The prediction of rate of penetration (ROP) is a key technology to optimize the drilling process and improve drilling efficiency. The existing calculation models are mainly based on physical experiments and theoretical analysis. The lack of application of measured data in drilling engineering makes the calculation accuracy difficult to meet the complexity on-site demand. In this study, we have used a public data set from Utah FORGE geothermal wells project. After cleaning and preprocessing the data, we identified 7 relevant characteristic parameters based on Spearman correlation coefficients to streamline the model and enhance efficiency. Subsequently, we selected 4 models and optimized their hyperparameters. Our results indicate that the Gaussian process regression (GPR) model, optimized using the Bayesian algorithm, outperformed the other models, achieving an R^2 value of 0.969, RMSE of 3.572, and MAE of 1.094 in predicting the ROP behavior in our field.

Keywords: rate of penetration (ROP) · predictive modeling · machine learning · Gaussian process regression (GPR) · Random Forest

1 Introduction

With the research and development of integrated drilling technology in geological engineering, the drilling rate equations obtained based on single-factor analysis in the early stage are difficult to meet the application requirements of massive actual drilling data on-site, severely affecting the adjustment and optimization of safe and efficient integrated geological drilling schemes[1]. There is an urgent need to establish new drilling rate equations. In recent years, with the application

* Supported by organization x.

of new technologies in the information field such as big data and machine learning to the petroleum industry, the integrated intelligent drilling rate equations that consider multiple factors have attracted increasing attention from domestic and foreign scholars[2]. Ahmed et al. [3] analyzed the prediction accuracy of models such as artificial neural networks, extreme learning machines, and support vector regression, verifying the feasibility of intelligent algorithms in drilling rate prediction. Jingning et al. [4] proposed an intelligent model combining layer analysis method with neural networks, using rock compressive strength, drill bit size, drilling parameters, and drilling fluid density to predict drilling rates, with prediction errors within 10%. Zhao Ying et al. [5] developed a mechanical drilling rate prediction model based on extreme learning machine regression algorithm, with prediction accuracy exceeding 90%. Although various intelligent prediction models for mechanical drilling rates have been established, a single intelligent algorithm is prone to falling into local optima, and the stability of prediction results is poor. Therefore, there is a need to develop mechanical drilling rate prediction models based on hybrid algorithms. Based on actual drilling engineering data, a new model for predicting mechanical drilling rates was established by combining the bayes optimization algorithm with Gaussian process regression. This model targets mechanical drilling rates, with input parameters including Depth(m), Hook load(kg), flow in (L/min), weight on bit (kg), rotary speed (rpm) and pump press (KPa). The results indicate that the Gaussian process regression optimized by the bayes optimization algorithm has higher prediction accuracy compared to other machine learning model. We share the developed code for public access as follows: https://github.com/vectorZeng/Predict-and-Optimization-of-Drilling-Rate-of-Penetration.

2 Materials

The source of the data used to generate this study is from the Utah FORGE project. This is a geothermal project for enhanced geothermal system (EGS) research. All the information processed in this study belongs to the well 58-32[6]. The well 58-32 is located in Milford, UT, 217 miles south of Salt Lake City. The well 58-32 reached a depth of 7,536 ft. drilling more than 4,500 ft. of granite. Most of the information generated by this project is public and can be accessed at Geothermal Data Repository (GDR) webpage (https://gdr.openei.org). The collected data contains observations of the drilling rig's performance from 27 parameters.

3 Methods

3.1 Machine Learning Algorithm

Supervised learning is a type of machine learning algorithm that requires both features (input variables) and the desired output. The algorithm tries to train and adjust its parameters to produce the desired output from the training set.

Then, the trained model is tested on unseen data, also known as the test set, for which we already know the output. Finally, the model predictions are compared with the known output values. These models are referred to as supervised learning algorithms because they have a "teacher" that provides supervision to the algorithms in the form of the known outputs for each example learned from the training data. Concerning the workflow for performing our data analysis and predictive modeling, the first step consists of the preprocessing of our data to prepare it for predictive modeling. The second step consists of the predictive modelling step. First, we train our model using the training set. Second, we validate it on an existing validation set to tune and optimize our model parameters. Lastly, the model is tested on unseen data[7]. The best model is selected according to the error. Fig. 1. summarizes the workflow of our predictive modeling.

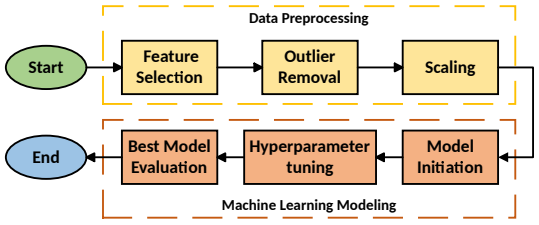

Fig. 1. Workflow required for machine algorithms

3.2 Data Preprocessing

Data preprocessing techniques are transformations applied to the training set to improve the performance of predictive models. Prediction is improved by transformations such as reducing data skewness and removing outliers. Feature selection is a simpler strategy that involves removing predictors based on their lack of information and is another effective technique for improving the performance of machine learning algorithms[8].

Feature Selection base on correlation Measurement Correlation is obtained using the Spearman correlation coefficient, which measures the linear relationship between two predictors.

Fig. 2 shows the correlation heatmap results.

Fig3 shows that the ROP is a strong function of Depth, Hook load, Flow In, weight On Bit, rpm and PumpPress where R's were -0.78, -0.74, 0.64, -0.60, 0.51 and -0.50, respectively. While the ROP a weak function of the other feature.

Therefore, this study selects the following features or predictors using domain knowledge for drilling engineering:

Target variable γ : ROP (m/hour).

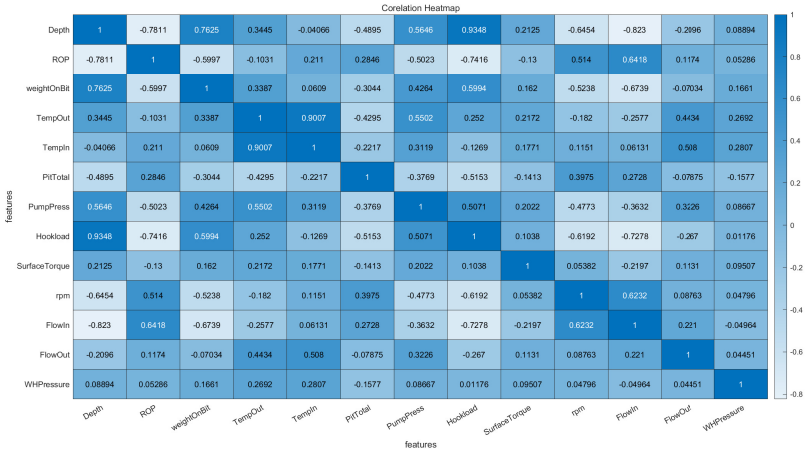

Fig. 2. correlation heatmap results

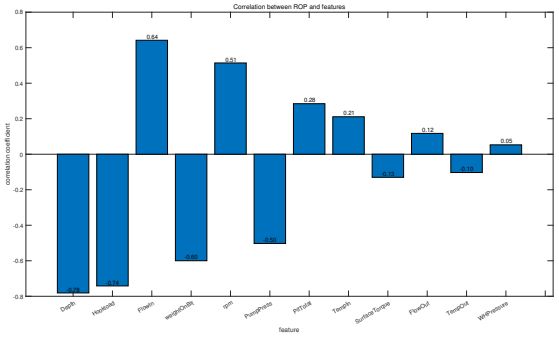

Fig. 3. The correlation coefficients between the input parameters and ROP.

Predictors X_i: Depth(m), Hook load(kg), flow in (L/min), weight on bit (kg), rotary speed (rpm) and pump press (KPa).

Table 1 summarizes the correlation results.

Table 1. Pearson correlation results.

Predictors	Spearman Correlation between ROP and Predictors
Depth (m)	-0.78
Hookload (kg)	-0.74
Flow In (liters/min)	0.64
weight on bit (kg)	-0.60
RPM	0.51
Pump Press (KPa)	-0.50

Outlier Removal A dataset can contain extreme values that are outside of the expected range and are unlike the other data. These outliers reduce the machine learning algorithm's ability to generalize. Removing these outliers improves the model's performance. There is no universal solution for outlier removal, but it is common is to rely on data visualization. We used both pair plot and boxplot for this purpose. Fig.4 shows the pair plot results[9].

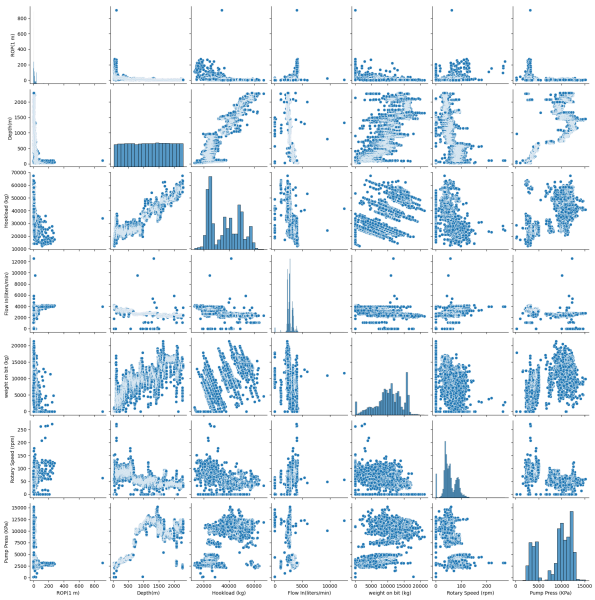

Fig. 4. Pair plot of the drilling dataset

The pair plot shows the distribution plot and the scatter plot associated with each pair of features. We found outliers in the ROP, rotary speed, flow in, and flow out features. The previous features were examined in-depth using the boxplot. The ROP boxplot below clearly demonstrates that ROP above 800 (m/h) is an outlier.Rotary speed above 200 (rpm) is considered an outlier.Flow in above 7000 (L/min) is considered an outlier.Finally, the flow out of less than 10% is considered an outlier (Fig.5).

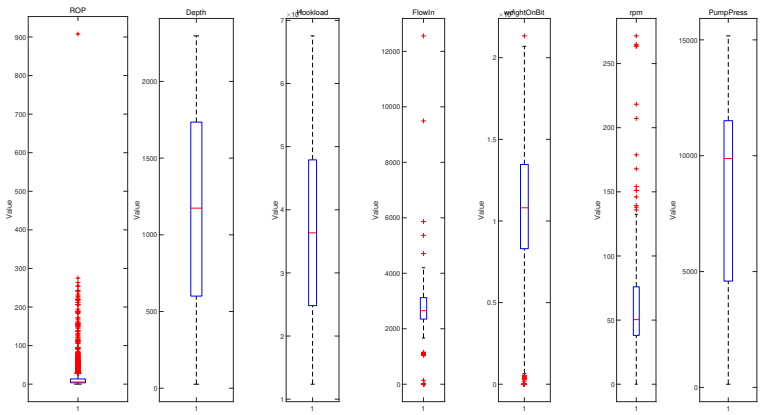

Fig. 5. boxplot of the drilling dataset

Data Scaling Machine learning algorithms rely on minimizing the cost function. Thus, data normalization is employed to speed up the calculation of gradient descent. This involves using a common scale for the values of the predictors [35]. Based on the probability distributions in Figure 7, standardization cannot be applied because our predictors do not follow a Gaussian distribution. Min-max normalization is used instead and described in Equation1

$$X_{MinMax} = \frac{X - X_{Min}}{X_{Max} - X_{Min}} \tag{1}$$

where X represents a predictor value, X_{Min} is the minimum value of the considered predictor, and X_{Min} is the maximum value of the predictor.

3.3 Accuracy Assessment of Regression Models

A commonly used metric to evaluate the performance of a predictive model on a given data set is the coefficient of determination R^2, root mean square error (RMSE). The coefficient of determination, or R^2, is a statistical measurement

that looks at how variations in one variable may be explained by the difference in another when predicting the result of an event. This measurement examines how strong the linear relationship is between two variables (predictors and response variables). The closeness of R^2 values to 1 indicates how well a statistical model predicts an outcome. RMSE is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far the data points are from the regression line, and RMSE is a measure of how dispersed these residuals are. In other words, it reveals the degree to which the data is centered on the line of best fit. The formulations are given as follows:

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y}_i)^2} \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_i^n (y_i - \hat{y}_i)^2} \tag{3}$$

$$MAE = \frac{1}{n} \sum_i^n |y_i - \hat{y}_i| \tag{4}$$

where y_i is the observed value of ROP; $\hat{y}_i$ is the prediction of ROP; and $\bar{y}_i$ is the average value of ROP.

3.4 Machine Learning Algorithms

In the scope of our study, the ROP of a geothermal well is assessed using ten distinct ML algorithms: (1) Random Forest model optimized with the Bayesian algorithm. (BO-RF), (2) Decision Tree, (3) Gaussian process regression model optimized with the Bayesian algorithm.(BO-GPR), (4) BP, Table 2 offers a succinct summary and comparison of the employed algorithms in this research.

4 Results

4.1 Training and Cross-Validation of Random Forest Regressor

After preprocessing and feature engineering, we trained the random forest with the 6 scaled predictors: Depth(m), Hook load(kg), flow in (L/min), weight on bit (kg), rotary speed (rpm) and pump press (KPa). Table.3 shows our new preprocessed dataset, which contains 7293 data samples following outlier removal.

First, we split our data into a 70% training set and a 30% test set. Then, the model's hyperparameters are optimized. The hyperparameter for the random forest method is the number of decision trees. Finally, the model evaluation is done by using a validation set. A total of 20% of the training set is used as the evaluation set for each fold, which corresponds to 14% of the total dataset. Additionally, we used the 5 folds cross-validation.

Table 2. A brief Summary and comparison of the used algorithms in this study.

Algorithm	Brief Description	Strengths	Weaknesses
Random Forest	Builds an ensemble of decision trees, each trained on a random subset of features andsamples.	Robust to noise and outliers. Handles nonlinear data. Can capture complex patterns. Performs well with highdimensional data.	Can be prone to overfitting. Less interpretable than individual decision trees.
Decision Tree	tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.	Easy to interpret and visualize. Can handle both numerical and categorical data. Nonlinear relationships can be captured.	Easy to interpret and visualize. Can handle both numerical and categorical data. Nonlinear relationships can be captured.
GPR	non-parametric approach for regression analysis that assumes a Gaussian process prior.	Can capture complex patterns in data. Provides uncertainty estimates for predictions. Can handle small datasets.	Computationally expensive for large datasets. Interpretability can be difficult.
BP	supervised learning algorithm used for training artificial neural networks.	Can learn complex patterns in data. Can handle large datasets.	Can be prone to overfitting. Requires a large amount of labeled data for training.

Table 3. Drilling dataset after pre-processing.

	count	mean	std	min	25%	50%	75%	max
ROP	7300	12.57	20.19	0.0	3.47	5.475	13.45	274.75
Depth	7300	1169.79	654.10	25.96	601.86	1174.75	1735.26	2296.94
Hook load	7300	36872.53	12021.41	12367.35	24816.33	36378.69	47908.16	67541.95
Flow In	7300	2707.047	514.85	0.0	2347.94	2649.86	3120.265	4709.92
weight on bit	7300	10491.04	4130.39	0.0	8307.26	10807.26	13460.32	21337.87
Rotary Speed	7300	54.789	25.36	0.0	38.09	50.38	75.92	178.86
Pump Press	7300	8737.21	3379.86	137.49	4593.17	9877.88	11512.44	15171.96

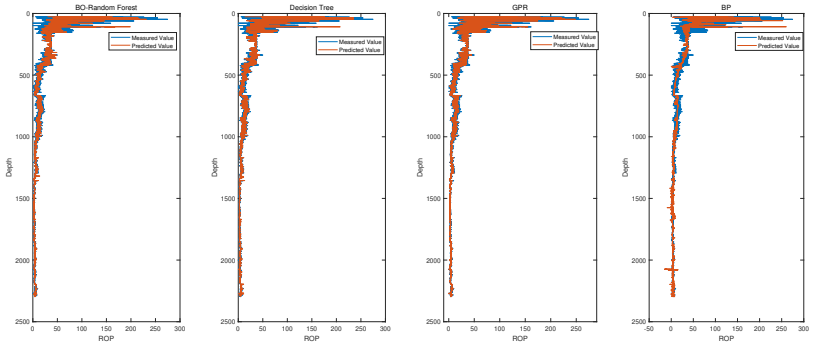

Fig. 6. Depth vs ROP profile using the tuned model of each algorithm where (a) is for BO-RF, (b) is for Desision Tree, (c) is for BO-GPR, (d) is for BP.

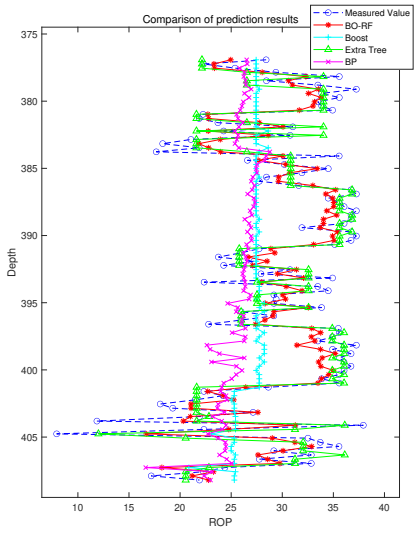

Fig. 7. Comparison of prediction results.

4.2 Model Comparison

Fig.6 illustrates the predicted and filed ROP values, along with the depth, for the model including BO-RF, decision tree, BO-GPR and BP.

Utilizing the best-tuned models among the 4 algorithms employed, which are shown in Table 4, the prediction performance on the test data was assessed through the measurement of 3 error metrics: MAE, RMSE, and R2. The analysis revealed negligible differences in the prediction performance across the models, as evidenced by the close alignment of the error metrics. Fig.8 visually depict the alignment between the actual test data and predictions generated by the tuned model for each algorithm.

Table 4. Ranges of each tuned hyperparameters and the hyperparameters of the best model for each algorithm.

ML- Model	Tuned hyperparameter	Best hyperparameter
BO-GPR	Sigma	22.288
	BasisFunction	Zero
	KernelScale	0.005057
BO-RF	EnsembleMethod	'Bag'
	NumLearners	160
	MinLeafSize	1
	NumberOfPredictorsToSample	4
Decision Tree	MinLeafSize	5
BP	Train function	Trainlgdx
	Transfer functions	tansig, purelin
	hidden layer	15

Table 5 summarizes the R^2, RMSE and MAE of each algorithm.

Table 5. Comparison of evaluation indexes of different models

ML-Model	R^2	RMSE	MAE
BO-RF	0.956	4.223	1.402
BO-GPR	**0.969**	**3.572**	**1.094**
Decision Tree	0.930	5.263	1.594
BP	0.838	8.045	3.182

As seen in Fig.6 and Fig.8, the predictive values of the BO-GPR model exhibit a consistent trend with the actual measured data. Furthermore, as shown in Table.5, the error analysis results of the BO-GPR model outperform those of the other three intelligent models. Therefore, it can be concluded that the proposed BO-GPR mechanical drilling speed prediction model demonstrates high predictive accuracy.

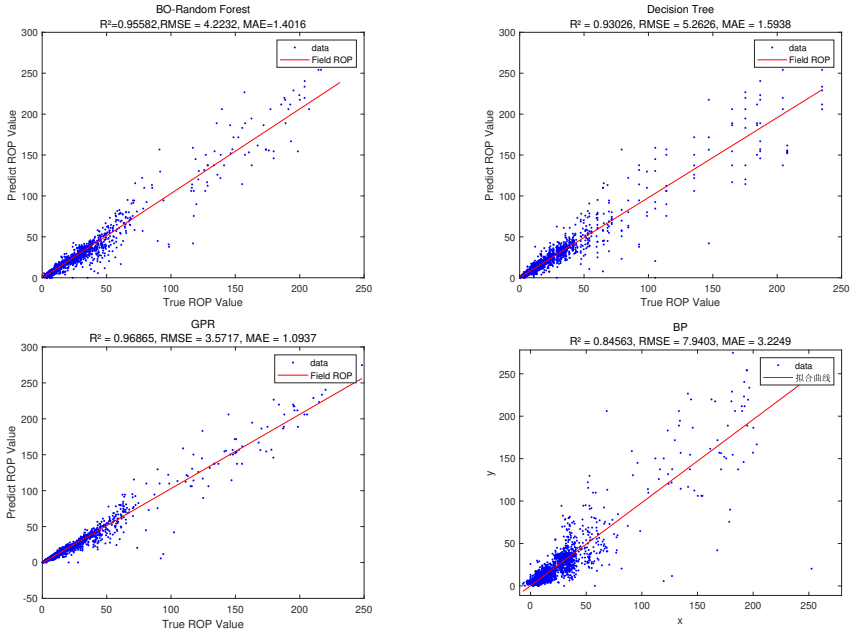

Fig. 8. Measured vs predicted ROP using the tuned model of each algorithm where (a) is for BO-RF, (b) is for Desision Tree, (c) is for GPR, (d) is for BP.

5 Conclusions

In this study, we proposed an intelligent prediction model for mechanical drilling rate based on Gaussian Process Regression optimized by bayes algorithm. Through Spearman Correlation analysis, 7 parameters including Depth, Hook load, Flow In, weight On Bit, rpm and Pump Press are selected as independent variables to establish the mechanical drilling speed prediction model. Compared to other models, this BO-GPR mechanical drilling speed prediction model demonstrates higher predictive accuracy. The application of hybrid algorithms in mechanical drilling speed prediction provides guidance for subsequent engineering applications. The intelligent prediction of mechanical drilling speed is discussed, and with the development of intelligent oilfields, the integration research of intelligent drilling speed prediction, intelligent drilling, and smart oilfields should be conducted to provide strong support for reducing drilling costs and improving drilling efficiency.

Acknowledgements Please place your acknowledgments at the end of the paper, preceded by an unnumbered run-in heading (i.e. 3rd-level heading).

References

1. Chao Gan, Weihua Cao, Min Wu, Xin Chen, and Suobang Zhang. Prediction of drilling rate of penetration (rop) using hybrid support vector regression: A case study on the shennongjia area, central china. *Journal of Petroleum Science and Engineering*, 181:106200–, 2019.
2. Rashidi, Behrad, Hareland, Geir, and Zebing. Performance, simulation and field application modeling of rollercone bits.
3. Mustafa M. Amer, Abdel Sattar Dahab, and Abdel Alim Hashem El-Sayed. An rop predictive model in nile delta area using artificial neural networks. 2017.
4. Jing Ning. Data mining technology-based research on the prediction method of deepwell rop. *China Petroleum Machinery*, 2012.
5. Zhao Ying, Sun Ting, Yang Jin, L. I. Yanjun, Huang Yi, and Yan Yulong. Extreme learning machine-based offshore drilling rop monitoring and real-time optimization. *China Offshore Oil and Gas*, 2019.
6. Moorea Nash, Greg and Joe. Utah forge: Logs and data from deep well 58-32 (mu-esw1). 04 2018.
7. Max Kuhn, Kjell Johnson, et al. *Applied predictive modeling*, volume 26. Springer, 2013.
8. Pang Ningtan, Michael Steinbach, and Vipin Kumar. Data mining introduction, 2006.
9. Mohamed Arbi Ben Aoun and Tamás Madarász. Applying machine learning to predict the rate of penetration for geothermal drilling located in the utah forge site. *Energies*, 15(12), 2022.