StackGBM: Stacked Gradient Boost Machine for Accurate Lost Circulation Prediction

Li Liang[1], Deng Hongmei[1], Yang Zhuo[1], Su Jianhua[1], Jiao Yang[1], Xie Yaorong[2,*], and Wu Chengyou[2]

[1]Exploration and Development Research Institute of PetroChina Changqing Oilfield Company, Xi'an, China, 710021
[2]Beijing KaiBoRui Petroleum Technology Co., Ltd., Beijing, China, 100083

*Corresponding author:bcpc@sina.vip.com

Abstract. Lost circulation leads to severe downhole accidents in some cases, and is common in oil or gas drilling. Lost circulation has become a serious threat to energy security and environmental protection and has thus attracted widespread attention. Recently, several studies introduce machine learning algorithms into lost circulation prediction, among which the Gradient Boosted Decision Tree (GBDT) methods take the lead. However, utilizing one single GBDT method can hardly generate optimal results. In this paper, we tackle this issue by the stacking technique. Besides, a lost circulation dataset is collected for further experiments. The proposed Stacked Gradient Boost Machine (StackGBM) adopts the two-stage paradigm to further enhance the original results that are produced by XGBoost, LightGBM and Catboost. In the second stage, a neural network system is employed due to its great prediction capability. Comprehensive experiments show that StackGBM achieves state-of-the-art performance in lost circulation prediction. In addition, we perform ablation studies on the variation of StackGBM architecture. The proposed StackGBM algorithm will benefit the development of drilling engineering in the long-term.

Keywords: Gradient boost machine; Lost circulation prediction; Machine learning; Energy security

1 Introduction

Lost circulation (LC) refers to that a considerable part of the drilling fluid enters into formations with high permeability, and it is common in major oil and gas production areas. Lost circulation may induce complex downhole accidents such as pipe sticking, overflow or blowout, and even cause the wellbore to be scrapped in severe cases [1]. Some studies reveal that about 25% of the oil and gas wells suffer leakage accidents every year, leading to enormous economic loss [14]. As a result, LC will cause significant energy security and environmental protection issues [2], which has drawn much attention and needs to be solved urgently.

R. Appleby et al. (eds.), *Proceedings of the 2nd International Conference on Intelligent Design and Innovative Technology (ICIDIT 2023)*, Atlantis Highlights in Intelligent Systems 10,
https://doi.org/10.2991/978-94-6463-266-8_25

The application of artificial intelligence methods in predicting and diagnosing lost circulation accidents has become a trend in the field of drilling engineering [3]. Recent studies mostly adopt neural networks, support vector machine and random forest, etc., to establish lost circulation prediction models based on machine learning algorithms, demonstrating excellent application prospects [4][5][6]. Generally speaking, these machine learning algorithms have achieved significant improvement in accuracy compared with human analysis.

The Gradient Boosted Decision Tree (GBDT) algorithm is an ensemble machine learning algorithm. GBDT combines several weak learners into one strong learner to improve prediction accuracy iteratively, and can effectively deal with complex engineering problems due to its low requirements for raw data processing. GBDT has been widely used in many areas such as urban traffic, electricity and drilling engineering.

However, previous works [13] show that utilizing just a single machine learning algorithm can hardly produce satisfactory results. In this paper, we aim to design a state-of-the-art algorithm by the stacking technique, which are capable of synergize the information generated by the single GBDT algorithms. Our contributions can be summarized as follows:

- We propose a novel machine learning algorithm termed as Stacked Gradient Boost Machine (StackGBM) for lost circulation prediction. StackGBM consists of several leading GBDT methods and neural networks, and is designed following the two-stage paradigm.
- We conduct thorough analysis on a lost circulation dataset collected from the petroleum company, and further provide visualization and preprocessing method, which are helpful for the following algorithm development.
- We perform comprehensive experiments and ablation studies on lost circulation prediction, and achieve significant improvement compared with the vanilla machine learning algorithms.

2 Dataset and Preprocessing

Our work utilizes dataset collected by Changqing Petroleum Company from March 14, 1988, to May 9, 2020. After data cleaning, 100,000 samples of the drilling data are employed, whose ratio of positive samples to negative samples is 44:1. The detailed description of the data field is presented in Table 1.

As shown in Table 1, most of the data fields are categorical features, while the others are numerical features. Specifically, the ‘‘Lithology’’ feature is deleted duo to its high proportion of missing values, and the drilling date is not taken into account in the experiment. We visualize the distribution of the accident wells in Figure 1 for better understanding.

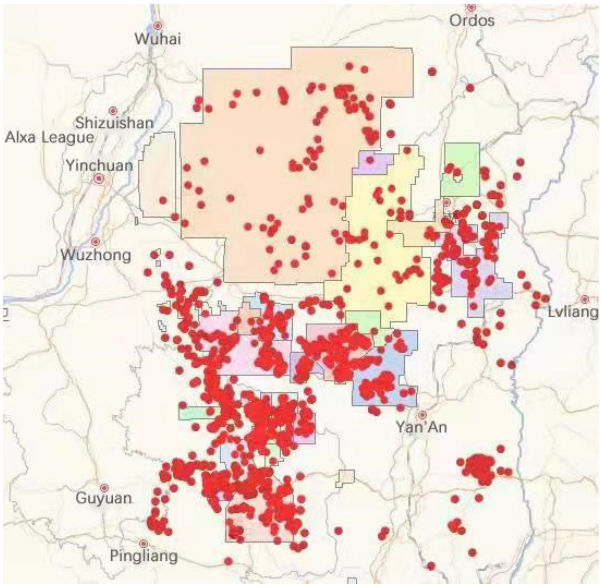

Fig. 1. Distribution of the Accident Wells

Table 1. Data field description

Feature Name	Feature Type	Category Number	Mean Value	Missing Value
Tectonic location	categorical	449		
Block	categorical	91		
Production factory	categorical	38		
Province	categorical	5		<1%
City	categorical	27		<1%
County	categorical	82		<1%
Geographical position	categorical	13564		
Well type 1	categorical	6		
Well type 2	categorical	26		
Well number	categorical	68658		
Drilling date	categorical	8807		
Drilling formation	categorical	221		
Lithology	categorical	7		>70%
Longitude	numerical		2.901×10^{7}	
Latitude	numerical		4.125×10^{6}	
Drilling depth	numerical		1995	

3 Methodology

3.1 XGBoost

As one of the most popular boosting models, XGBoost (Extreme Gradient Boosting Algorithm) improved the vanilla GBDT model by adding regularization [7]. In order to alleviate over fitting during training, the regularization term is constructed as a function of the number of leaf nodes and leaf weights of the decision tree, which can be represented as

$$\Omega(f_m) = \gamma J + \frac{1}{2}\lambda \sum_{j=1}^{J} \omega_j^2 \tag{1}$$

Based on the regularization term, we can conclude that the loss function is

$$L = \sum_{i=1}^{n} L(y_i, \hat{y}) + \sum_{m} \Omega(f_m) \tag{2}$$

In addition, XGBoost also performs second-order Taylor expansion on the loss function to improve the accuracy of the model.

3.2 LightGBM

LightGBM is another improved implementation of the GBDT, binging two merits into the boosting methods [8]. On the one hand, LightGBM can receive categorical variables as input features. On the other hand, it solves the issue of processing categorical features by one-vs-rest. Based on the conclusion of Fisher [9], LightGBM encodes category feature through many vs many strategy following three steps.

3.3 Catboost

The many-vs-many strategy utilized in LightGBM often leads to over fitting during training. Micci-Barreca [[10] deal with this issue by the target statistics (TS) method. TS transforms categorical features into numerical features by replacing the category with the tag average of each category, which is formulated as

$$x_k^i = \frac{\sum_{j=1}^{n} I(x_{j,k} = x_{i,k}) y_i + \alpha P}{\sum_{j=1}^{n} I(x_{j,k} = x_{i,k}) y_i + \alpha} \tag{3}$$

where $x_{i,k}$ represents the *k*-th category of the *i*-th sample, y_i represents the label of the *i*-th sample.

However, the above formulation is proved to be easily generating conditional offset [11]. Based on this prior, the Catboost [12] model is employed by integrating greedy-TS, which is formulated as

$$x_k^i = \frac{\sum_{x_j \in D_k} I(x_{j,k} = x_{i,k}) y_i + \alpha P}{\sum_{x_j \in D_k} I(x_{j,k} = x_{i,k}) y_i + \alpha} \tag{4}$$

where $D_k = D \setminus \{x_k\}$ represents the subset that exclude category x_k in the dataset.

3.4 Neural Network

Neural network technology first sprouted in the last century. After years of iterations, it has developed from the perceptron to the deep feedforward neural network. Neural network has achieved great success in many tasks such as natural language process and image classification [15]. For the lost circulation prediction task, we adopt the commonly used fully connected neural networks. Besides, all neurons in the system can be jointly optimized by the following cross entropy loss

$$L = -\sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{5}$$

where y_i represents the label of the *i*-th sample, and p_i represents the positive probability of the *i*-th sample.

3.5 StackGBM

Using each GBDT model alone does not provide optimal results for the lost circulation prediction. Previous works [13] show that combining several models together is likely to produce better results in the lost circulation prediction. Simply calculating the mean or maximum values of the output for the baseline models is intuitive, and we believe that stacking another machine learning model based on the output of the baseline models can achieve further improvement.

As shown in Figure 2, StackGBM utilizes all three GBDT models described above as the first stage, which deal with the input features respectively. Then, the output values of XGBoost, LightGBM and Catboost will be feed into another machine learning model in the second stage, e.g., neural network. Note that the above three boost models are running parallelly, and the structure of the neural network system employed in the second stage is 64×64×2 fully connected neural networks (dropout=0.2). Through the great processing capacity of the neural networks to the structural data, the output values of the single boost models can be improved. The pipeline is termed as stacked gradient boost machine (StackGBM). StackGBM produces accurate prediction values for lost circulation in drilling engineering.

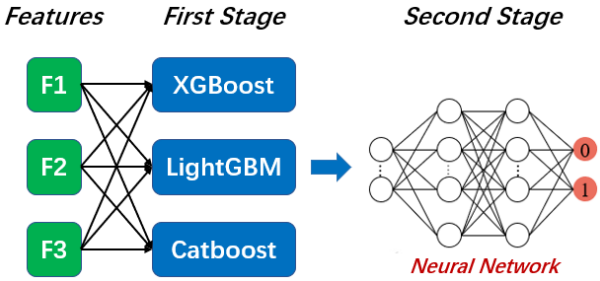

Fig. 2. The architecture of StackGBM.

4 Experiment

4.1 Implementation Details

We randomly split the dataset into 70% training set and 30% testing set to evaluate the effectiveness of the proposed StackGBM. As the baseline models, XGBoost, LightGBM and Catboost are also presented separately. We employ various machine learning methods to serve as the second stage model, including SVM (support vector machine), LR (logistic regression) and NN (neural network). It should be noted that the first stage models in the stack method are always XGBoost and LightGBM and Catboost. Moreover, AUC (Area Under ROC Curve) is adopted as the main metric of the experiments.

The detailed hyper-parameters of the GBDT models utilized in the proposed StackGBM are presented in Table 2. Note that these hyper-parameters are chosen by the grid-search algorithm, and we conduct the following experiments based on these hyper-parameters.

Table 2. The detailed hyper-parameters.

	XGBoost	LightGBM	Catboost
Learning rate	0.02	0.02	0.02
Iterations	298	153	693
Subsample	0.7	0.8	-
Max depth	8	9	9

4.2 Result Analysis

Our main results are presented in Table 3. Generally speaking, single models fail to achieve excellent results in the experiment, while the proposed stack methods produce satisfactory results. Specifically, combining boost models with the neural networks can generate the optimal result, and stacking SVM on the boost models performs poor. It is noticed that simply stacking LR can just do fine. Furthermore, the above boost models perform similarly in the experiment, demonstrating that utilizing only a single GBDT

model can hardly achieve optimal results. Consequently, we adopt ‘‘StackGBM (NN)’’ as our main model in the following experiment for its best performance.

Table 3. Results of various models. SVM, LR and NN indicate the stacked model in the second stage.

	Model	AUC
Single methods	XGBoost	93.55
	LightGBM	93.44
	Catboost	93.54
Stack methods	StackGBM (SVM)	90.12
	StackGBM (LR)	93.85
	StackGBM (NN)	**94.68**

Table 4 exhibits the ablation study on the model structure of the first stage. Note that the second stage are always set as neural network for its best performance. Results show that utilizing only two boost models in the first stage lags far behind the three-model architecture, demonstrating the effectiveness of the StackGBM model proposed in this paper (see Figure 2 for illustration). We believe that assembling the three GBDT models together helps improving the generalization ability and reducing the prediction variance of StackGBM.

Table 4. Ablation study on the model structure of the first stage. Neural network is utilized in the second stage.

First Stage	Second Stage	AUC
LightGBM + XGBoost	neural network	93.60
LightGBM + Catboost	neural network	93.62
XGBoost + Catboost	neural network	93.63
XGBoost + Catboost + LightGBM	neural network	**94.68**

The loss value during the training and validation phases are shown in Figure 3. It can be drawn that our StackGBM method achieves stabilized progress during the training phase, demonstrating that the proposed model is easy to train. And in the validation phase, the loss curve is similar to the curve in the training phase. Consequently, we can conclude that StackGBM suffers no overfitting during training.

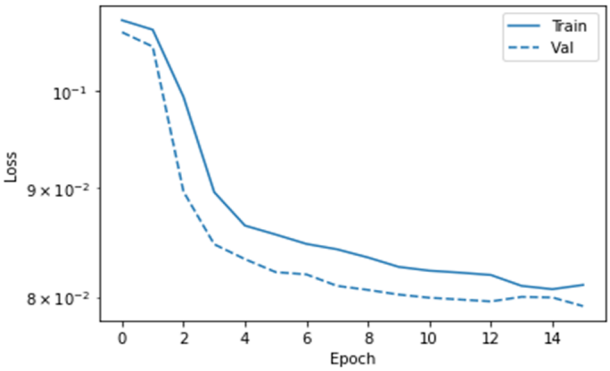

Fig. 3. Loss vs epoch.

5 Conclusions

In this paper, we propose a novel model termed Stacked Gradient Boost Machine (StackGBM) for accurate lost circulation prediction. Our contributions can be summarized as follows: (a) StackGBM adopts the two-stage paradigm to enhance the original results that are produced by three leading GBDT methods. And a neural network system is employed in the second stage. (b) Thorough analysis is conducted on a lost circulation dataset collected from the petroleum company, and visualization and preprocessing methods are further provided. (c) Comprehensive experiments and ablation studies show that StackGBM achieves state-of-the-art performance in the lost circulation prediction, promoting the development of drilling engineering.

References

1. Wang H M, Sweatman R, Engelman B, et al. Best practice in understanding and managing lost circulation challenges[J]. SPE Drilling & Completion, 2008, 23(02): 168-175.
2. Ho R, Moore D, Pirie D, et al. Drilling Deepwater Carbonates Using Managed Pressure Drilling on a Dynamically Positioned Drillship[C]//SPE/IADC Drilling Conference and Exhibition. SPE, 2014: SPE-167995-MS.
3. Alkinani H H, Al-Hameedi A T, Dunn-Norman S, et al. Applications of artificial neural networks in the petroleum industry: a review[C]//SPE middle east oil and gas show and conference. OnePetro, 2019.
4. Mohan R, Hussein A, Mawlod A, et al. Data driven and ai methods to enhance collaborative well planning and drilling risk prediction[C]//Abu Dhabi International Petroleum Exhibition and Conference. SPE, 2020: D012S116R120.
5. Lind Y B, Kabirova A R. Artificial neural networks in drilling troubles prediction[C]//SPE Russian Petroleum Technology Conference?. SPE, 2014: SPE-171274-MS.
6. Krishna S, Ridha S, Vasant P, et al. Conventional and intelligent models for detection and prediction of fluid loss events during drilling operations: A comprehensive review[J]. Journal of Petroleum Science and Engineering, 2020, 195: 107818.

7. Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.
8. Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[J]. Advances in neural information processing systems, 2017, 30.
9. Fisher W D. On grouping for maximum homogeneity[J]. Journal of the American statistical Association, 1958, 53(284): 789-798.
10. Micci-Barreca D. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems[J]. ACM SIGKDD Explorations Newsletter, 2001, 3(1): 27-32.
11. Zhang K, Schölkopf B, Muandet K, et al. Domain adaptation under target and conditional shift[C]//International conference on machine learning. PMLR, 2013: 819-827.
12. Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features[J]. Advances in neural information processing systems, 2018, 31.
13. Wang J, Li L, Ma J, et al. EGBM: an ensemble gradient boost machine for lost circulation prediction[C]//International Conference on Computational Modeling, Simulation, and Data Analysis (CMSDA 2021). SPIE, 2022, 12160: 1216002.
14. Arshad U, Jain B, Ramzan M, et al. Engineered solution to reduce the impact of lost circulation during drilling and cementing in Rumaila Field, Iraq[C]//International Petroleum Technology Conference. IPTC, 2015: D041S038R001.
15. Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. nature, 1986, 323(6088): 533-536.