SPE-221583-MS

Application of Boosting Machine Learning for Mud Loss Prediction During Drilling Operations

M. I. Okai, Department of Petroleum and Gas Engineering, Nile University of Nigeria, Abuja, FCT, Nigeria; O. Ogolo, Petroleum Training Institute, Warri, Delta, Nigeria; P. Nzerem and K. S. Ibrahim, Department of Petroleum and Gas Engineering, Nile University of Nigeria, Abuja, FCT, Nigeria

 DOI 10.2118/221583-MS

This paper was prepared for presentation at the SPE Nigeria Annual International Conference and Exhibition held in Lagos, Nigeria 5 - 7 August 2024.

Abstract

Lost circulation during drilling operations is a persistent challenge in the oil and gas industry, leading to significant financial losses and increased non-productive time. The common use of lost circulation materials (LCMs) in drilling fluids helps mitigate mud loss only to an extent. However, predicting the extent of mud loss before drilling specific formations would greatly benefit engineers. This study aims to predict mud loss using advanced boosting machine learning frameworks, addressing the need for more accurate forecasting tools. We evaluated three ensemble boosting algorithms—Adaptive Boosting (AdaBoost), Light Gradient Boosting Machine (LightGBM), and Extreme Gradient Boosting (XGBoost)—and compared them to Random Forest, a baseline bagging algorithm. Utilizing a dataset of over 7,000 data points with 27 features from drilling operations in Well MXY at the Utah FORGE field, we found that XGBoost and Random Forest were the most accurate models, with R^2 scores of 0.935 and 0.934, respectively. These results indicate that while XGBoost is the top-performing framework, Random Forest remains a robust and reliable method for predicting lost circulation, providing valuable insights for drilling engineers.

Introduction

Lost circulation is the escape of drilling fluids into the reservoir formation during drilling operations. It can be characterized as a partial loss or complete loss of drilling fluid to the formation. Lost circulation is one of the biggest problems in drilling operations. It is attributed to many drilling problems such as pipe sticking and kicks that can cause a lot of lost time during drilling (Wang, et al., 2008). It is a frequent drilling problem majorly in fractured or vuggy and highly permeable formations (Nayberg & Petty, 1986). Lost circulation can range from the unconsolidated and shallow formation and extend to the highly consolidated deeper formation that has been fractured by the drilling fluid. (Moore, 1986).

In recent times, data analytics, artificial intelligence and machine learning are maturing and providing new ways to diagnose, predict and bring forth solutions to numerous drilling engineering 2 problems (Zhong, et al., 2020). The revolutionary data forward method has special advantages in identifying uncertainty in various drilling problems, pattern identification and unravelling unknown information (Noshi & Schubert,

2018). Presently machine learning applications in lost circulation problems involve pre-drilling predictions, well prediction/diagnosis based on identical properties and real-time prediction/diagnoses while drilling. The pre-drilling predictions are done most usually with seismic and well log data. Using the machine learning method, the seismic attribute which most strongly correlates with lost circulation can be filtered and used. The lost circulation risk model is then created based on the similarity between the data and selected attributes. Due to the resolution properties of seismic data, this method is most often selected for risk analysis before drilling operations commence. In comparison, mud-logging parameters have a real-time characteristic and information making it the fundamental data for lost circulation diagnosis and prediction (Pang, et al., 2021).

In the earlier stages, single/multiple regression techniques were used by scholars to predict drilling parameters of interest (Al-Hameedi, et al., 2018). Next, they selected different drilling parameters, mud and downhole variables, and geological variables according to a data classification technique and used singular or multiple machine learning algorithms to either diagnose or predict lost circulation (Pang, et al., 2021). Works conducted include changing the mud loss judgement to a binary problem (Li, et al., 2018), prediction using selected parameters (Abbas, et al., 2018) and mud loss classification based on geologic fracture (Alkinani, et al., 2019). Lost circulation classification and risk diagnoses/prediction is a very hot topic. At now, research into lost circulation involving machine learning tends to focus on predicting if it happens or not.

Boosting

Boosting is a powerful class of machine learning methods on the simple idea that an ensemble or combination of weak learners performs better than a single learner alone. A weak learner is an algorithm that produces an error slightly below that of random guessing or 0.5. On the other hand, a strong learner produces a probability of error that is very small or near 0. The ensemble is an algorithm that is a combination of multiple weak learning algorithms. The idea of boosting is that it may be better to train several weak learners and combine their results in some way than to train a single complex learner.

For instance, instead of training a sing decision tree-based model, we may train several smaller decision trees and combine their results to get a better prediction or classification output. Boosting can give good results even if the base classifiers have a performance that is only slightly better than random, and hence sometimes the base classifiers are known as weak learners. Originally designed for solving classification problems, boosting can also be extended to regression (Friedman, 2001).

The main difference between boosting and committee methods like bagging is that the base classifiers are trained in order, and each base weak learner is trained using a weighted form of the data set in which the weighting coefficient for each data point is dependent on the previous classifiers' performance.

Boosting is essentially the process of repeatedly applying the fundamental weak learning algorithm to multiple weighted versions of the training data, resulting in a series of weak classifiers that are then combined as in. At each round of the algorithm, the weighting of each instance in the training data is determined by the accuracy of the preceding classifiers, allowing the system to focus its attention on samples that are still improperly identified. The choice of base learners and criterion for updating the weights of the training samples differs among the various boosting methods.

This work aims to evaluate the application of ensemble boosting supervised machine learning methods to predict lost circulation using drilling parameters. The machine learning techniques namely Adaboost (Adaptive Boosting), LightGBM (Light Gradient Boosting) and XGBoost (Extreme Gradient Boosting) will be applied.

Methodology

Data acquisition

A collection of datasets must be available to develop a data-driven model. Dataset consisted of raw static drilling data from Well MXY, Salt lake City, Utah Forge (USA) for this project. The data was in comma-separated values format (CSV). Drilling reports and annotations were not included.

Table 1—Statistical summary of parameters obtained from Well MXY drilling data.

index	count	mean	std	min	25%	50%	75%	max
Depth(ft)	7311	3835.044	2147.504	85.18	1970.395	3851.85	5691.585	7536.25
ROP (ft)	7311	42.01032	75.92112	0	11.4	17.99	44.295	2977.91
weight on bit (klbs)	7311	23.1167	9.119495	0	18.31	23.83	29.68	47.05
Temp Out(degF)	7311	126.0595	12.25988	84.07	116.13	124.86	136.49	151.7
Temp In(degF)	7311	118.3156	11.93314	85	108.85	117.22	126.86	146.31
Pit Total (bbls)	7311	236.9292	18.26164	170.91	224.535	238.11	249.59	279.88
Pump Press (psi)	7311	1266.634	490.5546	19.94	665.59	1432.56	1669.68	2200.43
Hookload (klbs)	7311	81.28559	26.50383	27.27	54.72	80.14	105.63	148.93
Surface Torque (psi)	7311	130.9836	48.70665	0	117	140.31	157.28	273.71
Rotary Speed (rpm)	7311	54.94729	25.94765	0	38.09	50.38	75.965	271.58
Flow In (gal/min)	7311	716.2541	141.7842	0	620.26	700.21	824.61	3317.51
Flow Out %	7311	79.69283	11.9094	0.69	72.65	80.71	88.845	111.21
WH Pressure (psi)	7311	-35.7609	222.6695	-1231.83	2.92	5.94	8.26	17.41
H2S Floor	7311	-0.02737	0.042453	-0.1	-0.07	-0.01	0	0.78
H2S Cellar	7311	0.004303	0.025282	-0.08	-0.01	0	0.02	0.07
H2S Pits	7311	0.148833	0.11529	-0.06	0.06	0.14	0.22	0.72
Mud Loss (gal/min)	7311	144.0863	92.40076	0	84.9475	134.9043	169.0511	447.9308

Data Exploration and Visualization

The primary aim of exploratory data analysis (EDA) is to examine the data for distribution, outliers, and anomalies to direct specific testing of the hypothesis. After importing the data set using all necessary libraries. The shape of the data was seen to be 7311 rows against 27 columns.

From exploring the data, it was noticed that some of the features were in both imperial and S.I units and the features in the latter units were removed to avoid exact collinearity. On further exploration, it was found that the mud loss column wasn't available. However, a mud-return parameter and a mud loss percentage parameter were used to create a Mud loss column. This "Mud loss" column will be the target or label for the machine learning model, while the remaining parameters will be the input features for the machine learning models. In visualizing the data, various techniques were used. A log plot was implemented to measure the variability of the features against the depth parameter. A pair plot and a visual correlation matrix were used as a bivariate analysis tool to measure linear dependencies or collinearity in each of the data features. A kernel density plot was used to measure regions of high density for given parameters.

Data Pre-processing

Missing values/sections in all datasets were removed to allow for a cleaner dataset devoid of biases from using other methods of dealing with missing values. A larger amount of these missing values accounts for unlogged data starting from most of the well tops. These parts were dropped off from both the training and

validation datasets to allow the algorithms to work on the dataset properly. Missing and infinite values are seen as ambiguous for algorithms like random forest and extra tree regressors.

Data scaling/ Normalization

The data was passed through a data scaling or normalization process. Normalization is a scaling technique or a mapping technique or a pre-processing stage where we can find a new range from an existing range (Patro et al., 2015). Normalization is required when there are big differences in the ranges of different features. Data scaling was done to reduce the magnitude of the drilling properties passing through the algorithms to speed up training time. The StandardScaler module from SK-learn preprocessing package was used. It standardizes features by removing the mean and scaling to unit variance. The standard score of sample x is calculated as:

$$Z = \frac{X - U}{S} \tag{1}$$

Where U is the mean of the sample and S is the standard deviation of the training sample

Model Optimization

A bagging algorithm, a random forest and three boosting algorithms AdaBoost, XGBoost and LightGBM all imported to the Python sci-kit-learn library were used. 10-fold cross-validation (CV) was used with each algorithm to get a more generalized result and to estimate the overall performance of the models on unseen data. This is also a data resampling technique. In k-fold cross-validation, the available training set is partitioned into k disjoint subsets of approximately equal size (Berrar, 2018).

The model was then trained with k-1 of the folds and evaluated on the last one. Hyper-parameter tuning with GridSearchCV was done to obtain optimal hyper-parameters for the training of each model. It aims at finding a tuple of hyper-parameters that yields an optimal model that minimizes a predefined loss function on a given independent data (Ghawi & Pfeffe, 2019). Five-fold cross-validation was used to conduct a grid search on the hyper-parameters. The number of estimators, minimum split at each tree leaf node, minimum samples at each leaf node, and minimum impurity split were tuned to obtain each model's optimal hyper-parameters. This was done to prevent the overfitting of the models on the training dataset. Overfitting maps out noise as interesting relationships rather than the real data signature.

Model Performance measures.

The model performance was optimized and evaluated using the root mean square error (RMSE) and the R^2 score (coefficient of determination). The RMSE is the standard deviation of all the prediction errors. It measures the deviation of the predicted values from the actual values. A very high value indicates a more error-prone prediction while a lower value shows more prediction accuracy. R^2 score is a measure of how fit the predicted values are to the actual values. R^2 scores range between −1 to +1. An R^2 score of +1 indicates a perfect positive fit while −1 indicates a perfect negative fit. The equations describing the statistical parameters are shown below:

1. Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(w_{iexp} - w_{ipred}\right)^2}{\sum_{i=1}^{N}\left(w_{pred} - w\right)} \tag{2}$$

2. Root mean square error ($RMSE$):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(w_{iexp} - w_{ipred}\right)^2}{N}} \tag{3}$$

Results and Discussion

Table 2 shows the statistical results obtained for all the models used in estimating the mud loss before and after optimization. The table shows that model training improved accuracy in the models implemented. All the models developed gave good estimates of the target (Mud loss gal/min). XGBoost gave the best estimate RMSE and R^2 of 24.38 and 0.935 after optimization respectively. Achieving the highest R^2 and lowest RMSE of 24.38 and 0.935. The baseline Random Forest gave the next best estimate with RMSE and R^2 of 24.1 and 0.934. It had the best average performance before and after optimization indicating it as a powerful algorithm. LightGBM had an R^2 and RMSE of 0.917 and 27.31 respectively. AdaBoost had the least performance with RMSE and R^2 of 36.37 and 0.852 indicating relatively poor performance metrics. Comparing all the models XGBoost and Random Forest models had the best performance with the highest R^2 and lowest RMSE overall. Therefore, these two models can be used with high reliability (>90%) to predict mud loss occurrence.

Table 2—Model evaluation metrics for each model before and after optimization

Method	RMSE before optimization	RMSE after optimization	R^2 before optimization	R^2 after optimization
Random Forest	24.3	24.1	0.933	0.934
AdaBoost	37.17	36.37	0.845	0.852
LightGBM	26.17	27.31	0.923	0.917
XGBoost	28.66	24.38	0.908	0.935

Predicted Logs Comparison.

Comparative plots of the actual mud loss log and predictions based on the four models are presented in Figure 1. The predicted and actual logs were plotted against their corresponding depth and superimposed in the figure so variations in difference can be seen. A uniform scale was shown for all log types to properly display the variations and similarities among the curves produced. The XGBoost and Random Forest log showed a more distinct log signature compared to the other predicted logs. This is due to its better prediction accuracy hence showing more similarity to the actual log plot compared to the other two models. The AdaBoost log performed relatively poorly, mostly struggling to fit the actual log predictions.

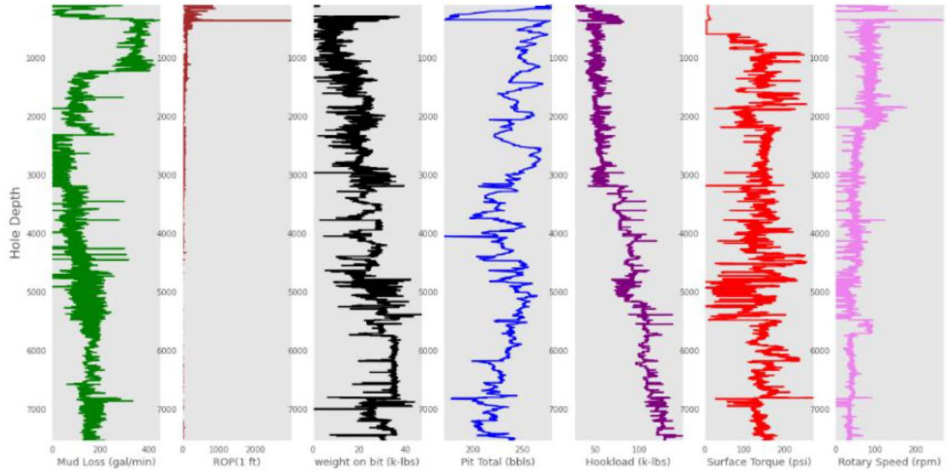

Figure 1—A visual log plot of selected parameters in the drilling data.

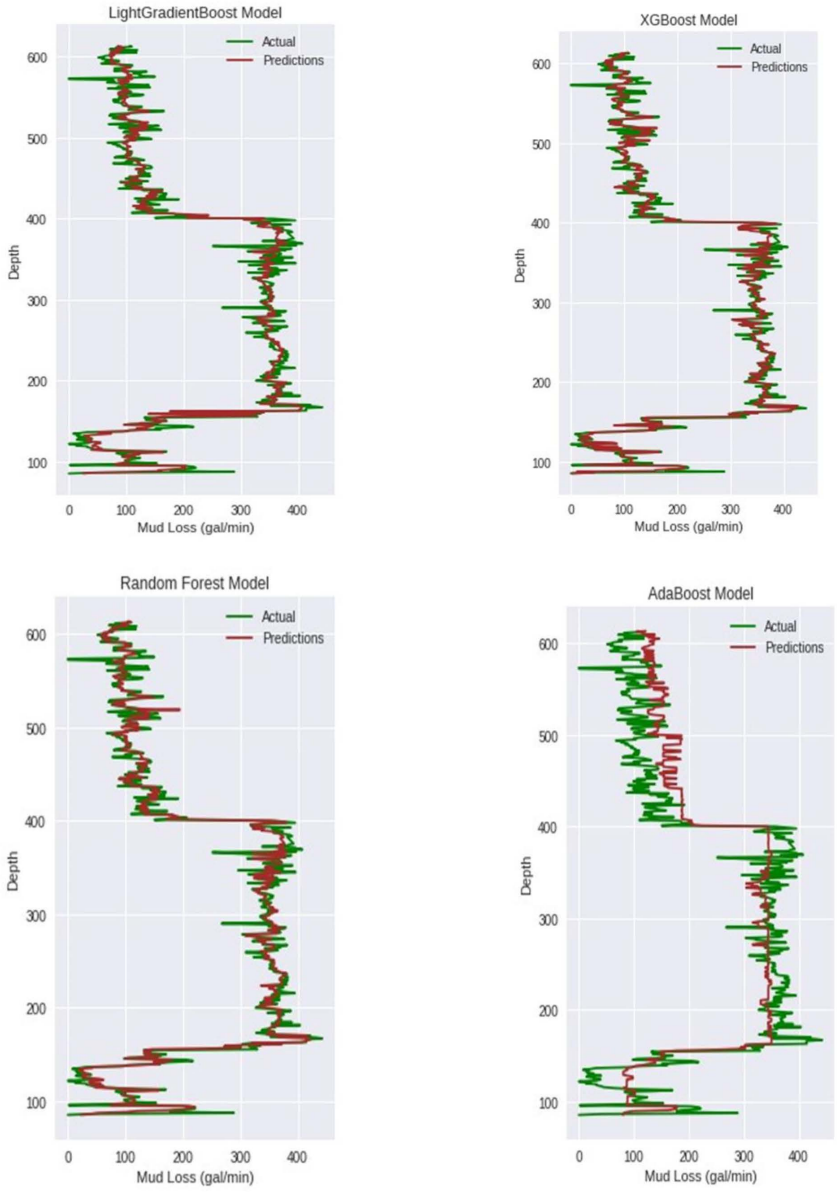

Figure 2—Predicted log comparisons for mud loss prediction in the four developed model.

Conclusion

1. Based on results gotten, drilling operational data alone is very sufficient in predicting lost circulation events with high accuracy.
2. XGBoost exhibited the highest performance in predicting lost circulation than other developed models with an R^2 score of 0.935, although followed very closely by Random Forest with R^2 of 0.934, next was the LightGBM model with R^2 of 0.917 and the worst performing AdaBoost with R^2 of 0.852. Moreover, XGBoost had the highest performance increase after optimization by tuning hyperparameters.
3. Boosting machine learning methods and particularly gradient boosting methods (XGBoost and LightGBM) are extremely effective and reliable in predicting mud losses with high accuracy rates (>90%) and minimal errors.
4. XGBoost and Random Forest had the best performance with an RMSE of 24.3 and 24.1 respectively. These results prove to show that although boosting methods are presumably an improvement to

bagging algorithms. The bagging methods are still robust and formidable with Random Forest having the lowest performance difference of 0.527% before and after optimization.

5. Based on field application, the proposed models are expected to perform well for real-time data. Although, these models are valid for datasets that fall within the range of datasets employed for the training process.

References

Abbas, A., Hammed, H., Al-bazzaz, W., & Abbas, H. (2018). Predicting the amount of lost circulation while drilling using artificial neural networks: an example of southern Iraq oil fields. Society of Petroleum Engineers - SPE Gas and Oil Technology Showcase and Conference 2019.

Al-Hameedi, A., Alkinani, H., Dunn-Norman, S., Flori, R., Hilgedick, S., Alkhamis, M., & Alsaba, M. (2018). Data Analysis of Lost Circulation Events in the Hartha Formation, Rumaila Field, Iraq. Society of Petroleum Engineers - SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition 2018. doi : https://doi.org/10.2118/192181-ms. SATS 2018

Alkinani, H., Al-Hameedi, A., Dunn-Norman, S., Flori, R., Alsaba, M., Amer, A., & Hilgedick, S. (2019). Using data mining to stop or mitigate lost circulation. *J. Petrol. Sci. Eng.* doi: https://doi.org/10.1016/j.petrol.2018.10.078.

Berrar, D. (2018). Cross-validation. *Encyclopedia of Bioinformatics and Computational biology*, **1**, 542545.

Feng, Y., & Gray, K. (2018). Modeling lost circulation through drilling-induced fractures. *SPE.J.*

Ghawi, R., & Pfeffe, J. (2019). *Efficient hyperparameter tuning with gridsearch for text categorization usingKNN*, 160–180.

Li, Z., Chen, M., Jin, Y., Lu, Y., Wang, H., Geng, Z., & Wei, S. (2018). Study on intelligent prediction for risk level of lost circulation while drilling based on machine learning. 52nd U.S. Rock Mechanics/Geomechanics Symposium.

Moore, P. (1986). *Drilling Practices manual* (2nd edition ed.). Tulsa: Penn and Well Publishing company

Nayberg, T., & Petty, B. (1986). *Laboratory study of lost circulation materials for use in oil-base drilling muds*. SPE Deep drilling and production symposium. Texas.

Noshi, C., & Schubert, J. (2018). The role of machine learning in drilling operations; a review. SPE Eastern Regional Meeting. doi: https://doi.org/10.2118/191823-18erm-ms

Pang, H., Meng, H., Wang, H., Nie, Z., & Jin, Y. (2021). Lost circulation prediction based on machine learning. *Journal of Petroleum Science and Engineering*, **208**(109364.). doi: https://doi.org/10.1016/J.PETROL.2021.109364

Patro, S., & Sahu, K. (2015). Normalization: A pre- processing stage. *Science journal*.

Wang, H., Sweatman, R., Engelman, B., Deeg, W., Soliman, M., & Towler, B. (2008). Best practice in understanding and managing lost circulation chalenges. *SPE drilling and completion*.

Zhong, R., Johnson, R., & Chen, Z. (2020). Using machine learning methods to identify coal pay zones from drilling and logging-while-drilling (LWD) data. *SPE J.* doi: https://doi.org/10.2118/198288-PA.