

**FA18-PETR6368**

**Predicting Rate of Penetration in Underbalanced Drilling Using Machine Learning**

**Nathan Tran, Dr. Robello Samuel, University of Houston.**

**SPE Members**

This paper was prepared as partial completion of the class Well Drilling and Completion I taught during the Fall 2018 semester at University of Houston. Team members include Nathan Tran. Contents of the paper have not been reviewed by the professor and are subject to correction by the authors. The material in this paper was compiled, developed, and/or synthesized by team members and does not necessarily reflect any position of the Department of Petroleum Engineering; its students, staff, or faculty; or University of Houston.

**Abstract**

Underbalance drilling has the advantage to minimize formation damage, lost circulation and differential sticking. However, its operation expense can be expensive. Predicting rate of penetration is important because it is a useful parameter for optimizing drilling cost. The main objective of this study is to predict the rate of penetration (ROP) for future underbalanced drilling processes in Ekofisk formation.

Data includes 13 operational parameters with over 3700 data samples collected from an underbalanced drilling process in Ekofisk formation. The study assumes that ECD at the bottom and bit equals 8.3 ppg, mud weight is as constant as 8.6 throughout the drilling process and bit size is 17.5 inches. Besides quality controlling the data, thoroughly examining and understanding the data in both quality, quantity and posible correlations is the differentiating part of this study. Exploring the data helps separate the data set into two clusters according gamma ray (GR) log (below 30 API and equal/above 30 API). In addition, sub-clusters (according to mud weight out (Mwout) for below 30 API group are necessary to develop efficient predictive models.

After clustering the data, random forest technique is applied for a major portion of each cluster to develop a predictive model forecasting ROP. The remaining portion of each data cluster is used to test the reliability of the model. To evaluate the model, mean square error (MSE), and variable importance have been calculated. The developed model for the group of GR < 30 API, MWout < 6.6 has MSE equals 1.93 with MWout and GR playing as the most important variables in the model. The developed model for the group of GR < 30 API, MWout > 6.6 has MSE equals 0.86 with MWout and GR playing as the most important variables in the model. The developed model for the group of GR > 30 API has MSE equals 1.65 with GR playing as the most important variables in the model.

**Introduction**

Underbalance drilling has the advantage to minimize formation damage, lost circulation and differential sticking. However, its operation expense can be expensive. Predicting rate of penetration is important because it is a useful parameter for optimizing drilling cost. The main objective of this study is to predict the rate of penetration (ROP) for future underbalanced drilling processes, in Ekofisk formation, using machine learning approach.

ML algorithms are mainly categorized into two group: Unsupervised learning and supervised learning (Table 1). Unsupervised learning is used on data with no labels and the goal is to find relationship in the data. On the other hand, supervised learning algorithms are trained with labeled data or the data comprised of examples of answered wanted.

|  |  |  |
| --- | --- | --- |
|  | Supervised Learning | Unsupervised Learning |
| Discrete | **Classification** | **Clustering** |
| Continuous | **Regression** | **Dimensionality Reduction** |

**Table 1. Types of Machine Learning Algorithms**

In situations where data is large enough and relevant to answer the question, machine learning algorithms can be used. In addition, depending on the desirable of the outcomes, some algorithms may be more favorable than others. Predicting ROP is a regression problem. There are many regression algorithms to use for prediction. They could be anything of Linear Regression, Random Forest, Gradient Boosting, Support Vector Machine, etc. The size of the training data set will determine the best combination of algorithms working well in our case. The predictive model of this study can be used to forecast ROP of newly drilled wells. From there, management team can make future plans for development of the field. The author also hopes to find applicability of this workflow in other reservoirs in the U.S.A and around the world.

**Methodology**

ROP depends on not only the drilling process but also the properties of the formation being drilled. Data includes 13 operational parameters with over 3700 data samples collected from an underbalanced drilling process in Ekofisk formation. The categorized parameters (operational and formation properties) are shown in Table 2.

|  |  |
| --- | --- |
| Parameters | Category |
| MWout | Operational |
| TempMudin | Operational |
| BlockPos | Operational |
| HookLoadA | Operational |
| WOBA | Operational |
| RPMA | Operational |
| TorqueAbs | Operational |
| Pressure | Operational |
| Flowin | Operational |
| HookLoad | Operational |
| Depth | Formation Property |
| Time | Operational |
| GR | Formation Property |

**Table 2. Parameters Used to Model ROP**

Crossplots between parameters are done to explore the data set. Because of the great number of data samples, random forest technique will be carried out to develop a regression model predicting ROP.

**Discussion**

**Data Exploration**

It is essential that understanding the data helps develop an appropriate model. Directly applying algorithms to regress the targeted variable will result in an inaccurate model. Figure 1 shows the result from directly applying random forest on the data set for a regression model predicting ROP. That most of the predictions are higher than the actual values. This indicates that the model is overpredicting.

**Figure 1. Actual ROP vs. Prediction from Modeling the Entire Data Set**

Crossplots between parameters have been carried out to distinguish the different clusters. Data points in each cluster will be modeled separately. Figure 2 shows two major clusters in terms of GR. Data points in the first cluster have GR < 30. Data points in the second cluster have GR>=30. Therefore, the data set is split into two group according to GR.

**Figure 2. ROP vs. GR**

Within each major group, parameters have been crossplotted again to identify sub-group for modeling. Using data in the group of GR < 30, a plot of Hookload versus MWout (Figure 3) indicates that there are two trends of operation during the drilling process. In the first trend, MWout is below 6.6 and the other is above 6.6. Therefore, the group GR<30 is sub-divided into two sub-group, MWout less than 6.6 and greater than 6.6.

**Figure 3. Hookload versus MWout**

**Modeling Data with GR < 30, MWout < 6.6**

Random forest (RF) has been applied to train 900 data samples. The generated predictions are mostly close to the actual value (Figure 4). The MSE is very low, 1.93 (Figure 5). The testing samples are used to assess the model. The red points in Figure 4 fall very close to the unity line. Therefore, the model is valid.

**Figure 4. Actual ROP vs. Prediction from Modeling Data Points with GR <30, MWout<6.6**

**Figure 5. Error Evolution During Modeling Data with GR<30, MWout<6.6**

The importance of the parameters is also evaluated and shown in Figure 6. In this case, GR and MWout are the most important parameter affecting ROP.

**Figure 6. Parameter's Importance in Modeling Group GR<30 and MWout<6.6**

**Modeling Data with GR < 30, MWout > 6.6**

This data set has over 800 data points. Similar methodology has been applied to model for ROP. Figure 7 shows that for the training data set (blue points), the regressed values are close to the actual values (MSE equals 0.8 in Figure 8). When the testing samples are used for prediction, it is important to notice the mismatches for actual ROP above 47. This is expected because around that region, there are not many points in the training data set to help the model learn. In the area where there are many training samples for the machine to learn, the prediction is quite excellent (ROP below 42).

**Figure 7. Actual ROP versus Prediction for GR < 30, MWout > 6.6**

**Figure 8. Error Evolution During Training Data GR <30, MWout > 6.6**

Similar with, the previous data set, GR and MWout are the most important parameters driving the value of ROP (Figure 9).

**Figure 9. Variables' Importance for GR < 30, MWout >6.6**

**Modeling Data with GR > 30**

There are not significant clusters in this data set. Therefore, RF is applied to the entire training set. Figure 10 show values of actual ROP versus the prediction. The regression converges to a MSE of 1.8 as the number of trees approaches 200 (Figure 11). The most important parameter in this regression model is GR. When the testing set is used to test the prediction, most of the predicted values is reasonable.

**Figure 10. Actual ROP versus Prediction for GR > 30**

**Figure 11. Error Evolution During Modeling Data Set GR > 30**

**Figure 12. Variables' Importance (GR >30)**

**Conclusion**

* Exploring the data is an important step before developing any model. In this study, reviewing the data helps identify three major clusters (GR < 30 and MWout < 6.6, GR < 30 and MWout > 6.6, and GR > 30). From there, modeling the cluster separately improve the accuracy for the predictions.
* RF is a suitable machine learning technique for data sets with numerous samples. Each cluster has at least 700 samples. Therefore, RF can learn from samples to generate accurate predictions.
* In this particular underbalanced drilling process, GR is the most important parameter for any cluster. When the formation is clean (GR<30), MWout is also an important parameter driving the value of ROP.
* The predictive model is developed under the assumption that the formation is Ekofisk, ECD at the bottom and bit equals 8.3 ppg, MWin is as constant as 8.6 throughout the drilling process and bit size is 17.5 inches. The entire data set is from a single well. Data from other wells may be necessary to address the variability of ECD, MWin and bit size. This could be the subject for future study.
* Besides the limitation stated in the assumption, other parameters affecting ROP are not desbribed in this data set. The drilling mud’s rhedology, bit types and rock’s mechanical properties may also the important parameters. Having these types of information may help improve the prediction in future study.
* The uniqueness of this study includes considering the importance of GR and incorporating its value in a proper manner to honor the formation’s properties. Another unique thought is the capture of different behavior of operational parameters (hookload and MWout) to differentiate the data set.

**Acknowledgement**

The author would like to kindly thank Dr. Samuel Robello for providing an excellent data set and his guidance to make this research posible. The author also respectfully thanks University of Houston’s Central of Advanced Computing and Data Science for providing training in machine learning.

**References**

1. Bell, Jason. *Machine Learning Hands-on for Developers and technical Professionals*. Indianapolis: John Wiley & Sons, 2015.
2. Robello, Samuel. *501 Solved Problems for Drilling Operations*, SigmaQuadrant.com, 2008