ISSN No:-2456-2165

Investigating and Ranking the Rate of Penetration (ROP) Features for Petroleum Drilling Monitoring and Optimization

Ijegwa David Acheme, Osemengbe Oyaimare Uddin*Ayodeji Samuel Makindes
Department of Computer Science, Edo State University Uzairue,
Nigeria

Abstract:- The drilling phase has been reported to be the most expensive phase of oil exploration and production, hence several research efforts have been targeted at improving its efficiency. The rate of penetration (ROP) has also been identified as the most important metric for improving drilling performance, hence, several research efforts have reported different methods of predicting ROP optimal values. Recently, artificial intelligence (AI) and machine learning (ML) models have been reported for the prediction of ROP. However, the ROP is influenced by several factors, and the interactions among these factors introduces a kind of complexity that affects its accurate prediction. This research work sets out to achieve two important objectives, firstly, to investigate and rank the most important factors for the prediction of the ROP, and secondly, to carry out a comparative study and ranking of selected machine learning algorithms for the prediction of ROP. In order to achieve this, the open source volve dataset which is a complete set of data from the North Sea oil field was utilized. Eighteen (18) machine learning models were built using this dataset and their performances compared. The result showed the random forest regressor with an RMSE value of 0.0010 and R2 score of 0.891 as the most efficient algorithm among the eighteen chosen for this work. Further experimentation also revealed the most influential factors for predicting the rate of penetration, these features in order of importance are; measured depth, bit rotation per minute, formation porosity, shale volume, water saturation, log permeability. The output of this study work offers a blueprint for choosing algorithms and features when implementing ML solutions for optimizing oil drilling, and this is helful in the development of real-time ROP prediction models and hybridization.

Keywords:- *Rate of Penetration Prediction, oil drilling, machine learning, feature selections*

I. INTRODUCTION

Several researchers have noted that the drilling phase remains the most expensive phase of oil exploration and production (Cao, *et* al, 2021, Sircar, *et al*, 2021; Darwesh *et al.*, 2019; Ameloko *et al.*, 2019; and Lashari *et al.*, 2019). Therefore, ongoing research projects aimed at drilling process optimization to achieve a decrease in the overall expenses connected with the drilling process have been reported. Although equipment, products, and processes are always being improved, machine learning methods have only just started to play a substantial role in oil drill optimization. This has largely been made possible by the current accessibility of enormous datasets. (Braga, 2019). Machine learning (ML) models holds promising results in this sector, as this will lead to the efficient processing of the massive amounts of data, which are produced by several internets of thing (IoT) sensors at oil rigs to aid decision making. Major oil firms have already invested hundreds of dollars in the IT infrastructure to establish Real-Time Operation Centers (RTOC), which read drilling data from rigs in real-time. With the help of these readings, specialists can instantly assess data in the centers, enabling quicker decision-making, a decrease in stuck pipe incidents, hole cleaning problems, and fluid loss occurrences, as well as an increase in the number of wells that can be monitored with the same amount of staff. (Al-khudiri et al., 2015). Additionally, the accessibility of this data has provided the essential groundwork for the application of artificial intelligence and machine learning techniques for the creation of smart models for more precise and reliable real-time drilling performance monitoring and optimization.

As a result of the enormous amounts of instrumentation that modern drilling rigs possess for the collection of parameters from almost every piece of equipment installed in the drilling rig, using sensors to measure their states, and enabling remote and safe operations, there has been an exponential increase in the amount of data generated at oil rigs. This has prepared the way for the creation of predictive analytics machine learning models and decision support systems.

As researchers continue to study these datasets created at oil rigs, choosing the appropriate machine learning algorithms and features for the precise prediction of ROP poses a challenge because the ROP is influenced by a number of variables that have complex relationships, and the extent of their influence also varies as some are more relevant than others. In addition to implementing and contrasting various ML techniques, the goal of this research is to investigate and rank the factors that have been published in the literature for ROP prediction. A machine learning model built with many of the lowly influential factors or with a less efficient algorithm is not likely to give satisfactory results. By focusing on the most crucial elements, these models will perform better in terms of prediction, computation, and training time, and will be easier to understand. (Acheme, *et al.*, 2022).

ISSN No:-2456-2165

II. RELATED WORKS

The speed at which a wellbore is being drilled can be used to define the rate of penetration (ROP). By monitoring the depth at regular periods of time in feet or meters per hour, one can manually calculate this. High ROP values suggest quick drilling, which translates to higher drilling productivity. Reducing this time in order to attain a greater ROP is a crucial optimization approach for oil firms because ROP is such a direct measurement of the overall time necessary to drill an oil well. This section presents many approaches that have been used to optimize ROP. These methodologies, which can be broadly categorized into traditional and data-driven models, have focused on modeling and predicting the ROP using specific drilling parameters that can be manipulated on the surface, such as weight-on-bit (WOB), rotary speed (RPM), etc. Data-driven models refer to machine learning methods for the prediction of ROP, while traditional models refer to mathematical equations that have been developed by tests and field experience.

A. Traditional ROP Models

One of the early mathematical models for ROP prediction was developed by Maurer in 1962, who used a rock cratering technique to develop a formula using the parameters bit diameter, rock strength, weight on bit (WOB), and rotations per minute (RPM). This is according to Alsaihati et al (2022). Another early mathematical equation-based ROP prediction model is the Bingham model, which is described in Hegde et al. (2018). It uses similar input parameters along with an extra empirical constant, "k," which stands for a parameter that was dependent on formation. In Eckel (1967), Eckel presented a further early conventional model that examined the impact of mud on ROP. The Bourgoyne and Young (BY) model (Bourgoyne & Young, 1974) is the earliest model that has garnered the most attention and media coverage. The formation strength, undercompaction, normal compaction trend, differential pressure, bit diameter and weight, rotational speed, tooth wear, and bit hydraulics were other geological and physical aspects that were taken into consideration.

B. Data Science Models

The goal of leveraging data gathered during drilling to create predictive models of ROP is the application of data science and machine learning techniques for the prediction and optimization of ROP. In order to forecast the rate of penetration, such models use surface-measured characteristics as input variables, such as weight on bit, rotations per minute, and flow rate. Oil drilling typically entails extensive data collection from both surface and subsurface areas employing IOT sensors. These sensors can gather a lot of information on the condition of the bit underneath. Plotting, analyzing, and controlling bit performance, in this case the ROP, are done using the obtained data. Due to the fact that ML models the relationship between input factors in order to predict an output (target) variable, the availability of these datasets has created the groundwork for the construction of models for the prediction of ROP. The research that have proposed ML models for ROP prediction are reviewed in this section.

The majority of research has been reported on using neural networks as a machine learning method. For ROP prediction, these neural network models have used a variety of input parameters (Jahanbakhshi, 2012). A hybrid neural network model was proposed by Ashrafi et al. (2011) that made use of the Savitzky-Golay (SG) smoothing filter to remove noise from retrieved data in order to estimate the rate of penetration.

A feedforward neural network model for predicting penetration rate was published by Lashari et al. in 2019. The work made use of a few elements, including differential pressures, mud flow, bit weight, and bit rotations per minute. The input variables were made up of these attributes. Datasets used for the creation of their model came from both an oil field and lab simulations. By comparing the projected values with the actual measured value, the anticipated ROP values are then utilized to detect bit failure or malfunction. Any detected variance suggests that the bit is performing below par, and this can be a red signal.

An artificial neural network (ANN) model was used in the study by Wang and Salehi (2015) to forecast hydraulics pump pressure and to provide early warnings. The model was implemented using MATLAB's fitting tool, and the sensitivity of the chosen input parameters was examined using the forward regression method. Data sets were gathered from chosen well samples and used to verify the model. In similar formations, the model predicted pump pressure vs well depth. While powerful tools, neural networks have proven to be particularly effective at handling high-dimensional modeling. (Hinton et al., 2012; Schmidhuber, 2015; Hegde et al., 2015) contend that when applied to low dimensional issues, they typically underperform when compared to simpler machine learning models like random forest, which have reported greater prediction accuracies. ROP is typically monitored in real-time by equipment that uses measurement-while-drilling (MWD) techniques. The optimization of the rate of penetration is required since greater ROP values indicate that drilling distance is being covered more quickly. Oil drilling businesses want to cover greater distances more quickly in order to save time and money. WOB and RPM are two factors that can be directly regulated and have an impact on the rate ROP. The soil formation affects the other factors (PHIF, VSH, SW, and KLOGH). ROP first rises until a point called the founder point or the sweet spot (optimum point), after which it starts to fall. This has been observed through studies. As a result, to retain the best performance moving forward, the values of the external variables must be raised. Regrettably, ROP does not always rise proportionately to changes in these variables' values.

ISSN No:-2456-2165

III. PROPOSED METHODOLOGY

The process used in this work to create machine learning and data science models is conventional. Six (5) phases make up the methodology depicted in figure 1, and they are as follows:

- understanding the business issue, that is the problem
- Data preparation and cleanup
- Data modeling
- Model Assessment
- Model Implementation

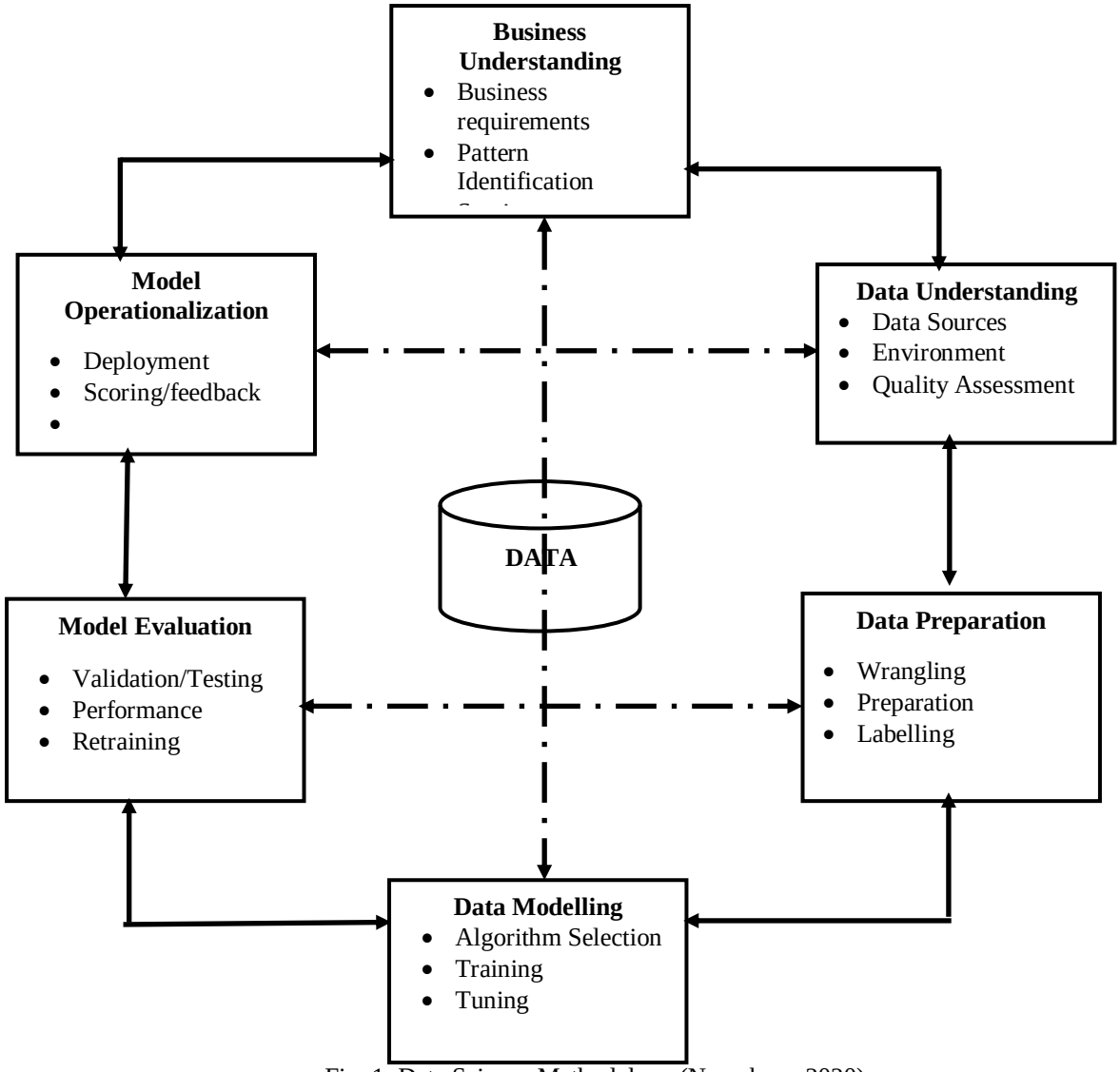

Fig. 1: Data Science Methodology (Nwankwo, 2020)

A. Understanding the Problem and Data Collection

The open source volve data was the dataset used in this study. This comprehensive set of North Sea oil field data is made up of real-time drilling data and Computed Petrophysical Output (CPO) log data from well number 15/9-F-15 in the Volve Oil Field in the North Sea (Equinor 2018). It is available for research, study, and development purposes. Seven (7) input variables and one (1) target variable make up this dataset. which are:

- Height (measured height)
- WOB (Weight on bit)
- SURF_RPM (surface rotation per minute)
- PHIF (formation porosity)
- Shale Volume (VSH)
- Water saturation (SW)
- Log permeability (KLOGH).
- TARGET VARIABLE: ROP_AVG (rate of penetration average)

ISSN No:-2456-2165

Table 1: Snapshot of the Dataset

	Depth	WOB	SURF_RPM	ROP_AVG	PHIF	VSH	SW	KLOGH
0	3305	26217.864	1.314720	0.004088	0.086711	0.071719	1.000000	0.001000
1	3310	83492.293	1.328674	0.005159	0.095208	0.116548	1.000000	0.001000
2	3315	97087.882	1.420116	0.005971	0.061636	0.104283	1.000000	0.001000
3	3320	54793.206	1.593931	0.005419	0.043498	0.110040	1.000000	0.001000
4	3325	50301.579	1.653262	0.005435	0.035252	0.120808	1.000000	0.001000
...	...	...	...	...	...	...	...	...
146	4065	71081.752	2.104258	0.008808	0.087738	0.291586	1.000000	0.162925
147	4070	72756.626	2.333038	0.008824	0.019424	0.503175	1.000000	-0.001124
148	4075	83526.789	2.333326	0.008799	0.054683	0.689640	1.000098	0.002261
149	4080	84496.549	2.334673	0.008375	0.022857	0.640100	1.000000	0.001000
150	4085	86658.559	2.331339	0.008454	0.022857	0.640100	1.000000	0.001000

There were a total of 150 entries in the dataset, each with eight (8) features.

B. METHODOLOGY

The dataset listed in Table 1 was used to construct the chosen machine learning algorithms in order to meet the goals of this study. Figure 2 displays the many steps of the complete procedure.

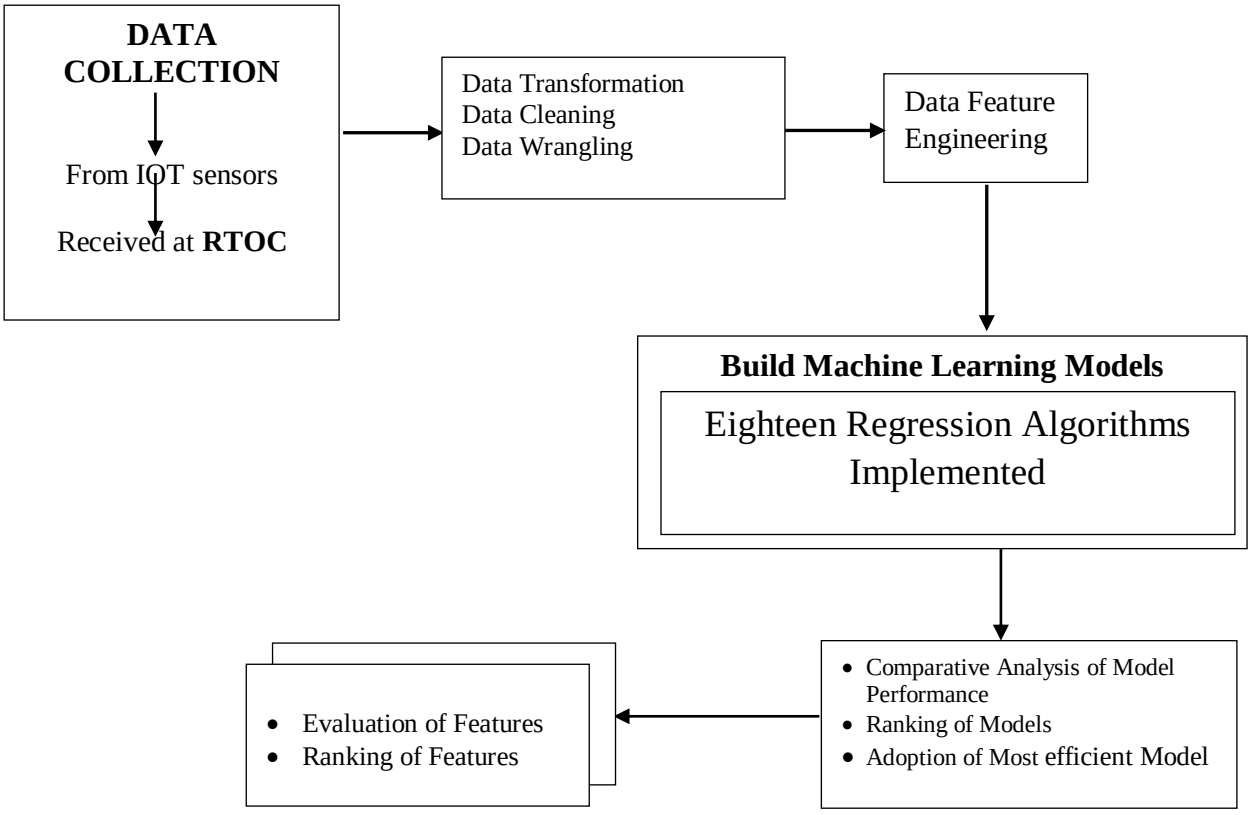

Fig. 2: Proposed Architecture

C. Data Modelling and Evaluation

In order to learn more from the data and find hidden trends, exploratory data analysis was next done. The data is then split into training and testing portions in a 70:30 ratio using the chosen features. Then, machine learning algorithms receive this. The machine learning algorithms employed and their performance comparison are shown in Table 1.

Cooks distance outlier detection was performed to estimate outliers in the dataset (figure 3). An estimation of a data point's influence is called the Cook's Distance. It takes each observation's leverage and residual into account. When the ith observation is taken out of a regression model, the change in the model is calculated as Cook's Distance.

ISSN No:-2456-2165

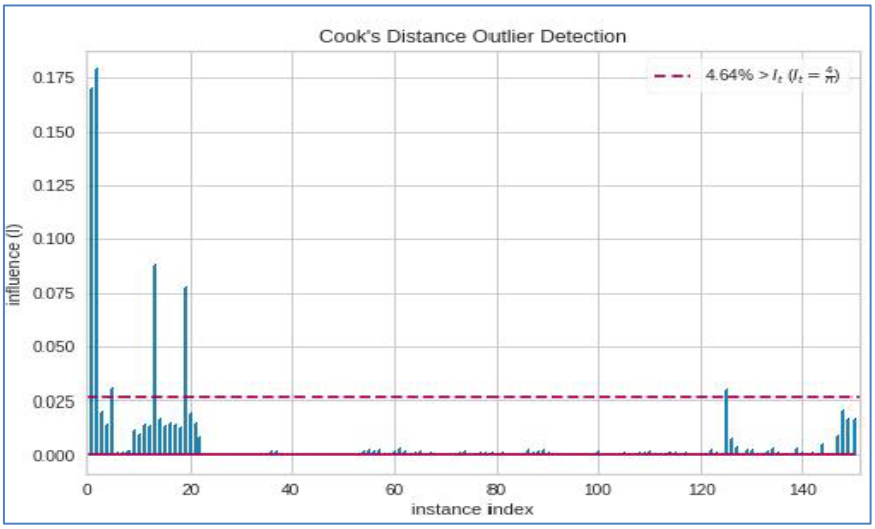

Fig. 3: Cook's Distance Outlier Detection

Table 2: Selected Regression Models and their Performances

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
rf	Random Forest Regressor	0.0006	0.0000	0.0010	-0.0891	0.0010	0.1082	0.407
gbr	Gradient Boosting Regressor	0.0006	0.0000	0.0010	-0.1076	0.0010	0.1133	0.045
et	Extra Trees Regressor	0.0006	0.0000	0.0009	-0.1805	0.0009	0.1064	0.363
huber	Huber Regressor	0.0007	0.0000	0.0010	-0.1856	0.0010	0.1213	0.029
dt	Decision Tree Regressor	0.0007	0.0000	0.0012	-0.3617	0.0012	0.1298	0.014
knn	K Neighbors Regressor	0.0009	0.0000	0.0013	-0.5364	0.0012	0.1572	0.059
ridge	Ridge Regression	0.0007	0.0000	0.0010	-0.5516	0.0010	0.1219	0.012
br	Bayesian Ridge	0.0007	0.0000	0.0010	-0.5530	0.0010	0.1224	0.014
en	Elastic Net	0.0009	0.0000	0.0012	-0.5729	0.0012	0.1531	0.013
lightgbm	Light Gradient Boosting Machine	0.0007	0.0000	0.0011	-0.5771	0.0011	0.1281	0.046
lr	Linear Regression	0.0007	0.0000	0.0010	-0.5861	0.0010	0.1227	0.304
lar	Least Angle Regression	0.0007	0.0000	0.0010	-0.5861	0.0010	0.1227	0.013
lasso	Lasso Regression	0.0009	0.0000	0.0012	-0.6040	0.0012	0.1532	0.014
llar	Lasso Least Angle Regression	0.0009	0.0000	0.0012	-0.6929	0.0012	0.1531	0.014
dummy	Dummy Regressor	0.0009	0.0000	0.0012	-0.6929	0.0012	0.1531	0.013
omp	Orthogonal Matching Pursuit	0.0007	0.0000	0.0011	-0.7197	0.0011	0.1262	0.012
ada	AdaBoost Regressor	0.0007	0.0000	0.0011	-0.7845	0.0011	0.1267	0.067
par	Passive Aggressive Regressor	0.0079	0.0001	0.0080	-138.4456	0.0080	1.0000	0.013

To examine the ROP_AVG target variable's prediction accuracy using the chosen features, the models provided in Table 2 were put into practice. Regression model evaluation standards like MAE, MSE, RMSE, R2, and others are used as the comparison measures. According to our findings, the random forest regressor model performed better than all the others and is ranked number 1, whereas the passive aggressive regressor performed poorly and is ranked number 18.

D. Evaluation of the Random Forest Regressor

Further evaluation analysis of the algorithm, including the residual plot, error plot, learning and validation curves, was conducted after it was determined that the random forest (rf) algorithm was the most effective among the selected eighteen (18) machine learning algorithms tested with the dataset. Additionally, feature priority ranking was done to determine which features were most crucial for predicting ROP. Figures 4 show these results.

ISSN No:-2456-2165

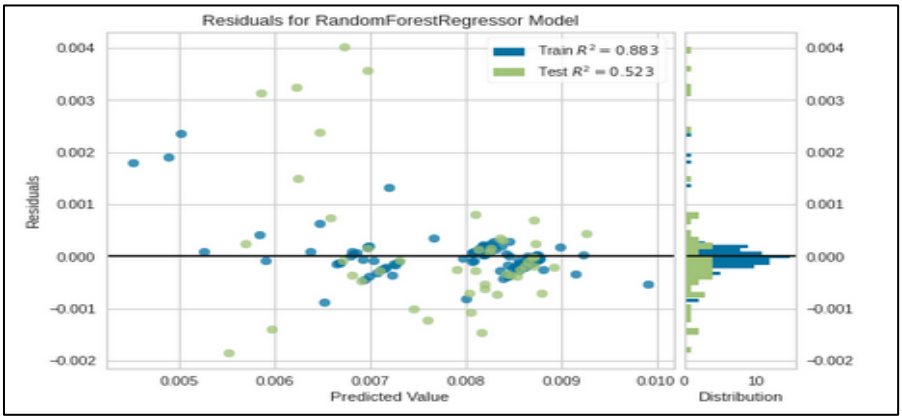

Fig. 4: Residuals for the random forest algorithm

The model's fit was verified using the residual plot under the assumptions of constant variance, normality, and error independence. The discrepancy between the observational and fitted values can be seen on the plot. Figure 4's plot displays erratically spaced points that retain an approximately constant width around the line of identity; this is a sign of a sound model because it is close to a null residual plot.

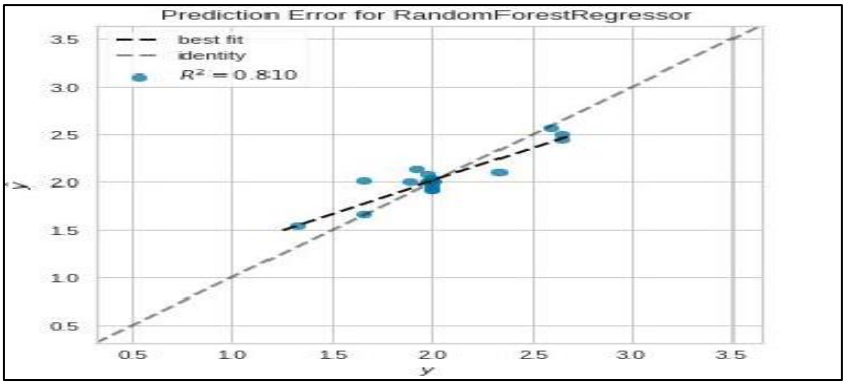

Fig. 5: Prediction error for the random forest algorithm

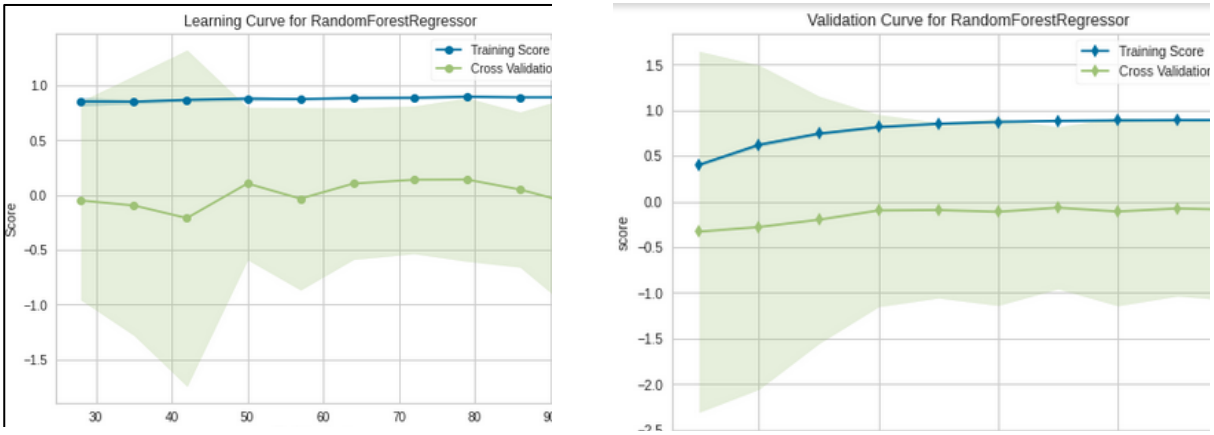

Fig. 6: Learning and validation curve for the random forest regressor

With the use of the learning and validation curves (Figure 6), the performance of the model was further examined. These diagrams display a model's performance with time or as the training data set grows. They are helpful for models created using incremental datasets. The validation curve demonstrates how effectively the model generalizes with values that have not previously been observed, while the training curves demonstrate how well the model learns.

E. Feature Importance and Ranking

Calculating the relevance of a feature involves weighing the decrease in node impurity by the likelihood of reaching that node. The node probability can be computed by dividing the total number of samples by the number of samples that reach the node. The values of the more significant traits are higher.

ISSN No:-2456-2165

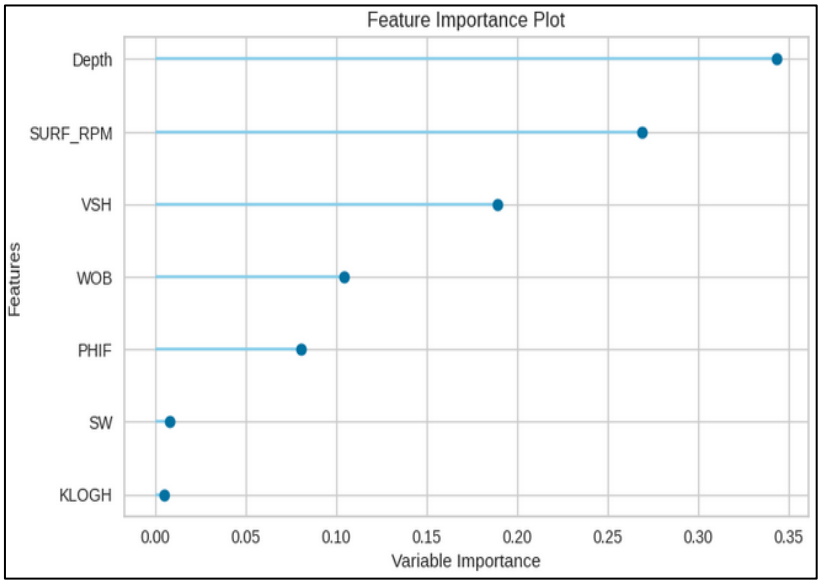

Fig. 7: Feature importance

As seen in figure 7, the features are measured depth, rotations per minute on the surface, shale volume, weight on bit, formation porosity, water saturation, and log permeability, in that order of importance.

IV. SUMMARY AND CONCLUSION

In comparison to other stated traditional methodologies, the creation and implementation of efficient machine learning applications for ROP prediction offers superior outcomes. This is because there are more datasets available that are produced at oil rigs, but choosing the best machine learning features and algorithms presents a real difficulty. It's unlikely that a model created using a lot of insignificant factors or a less effective algorithm can produce adequate results. Because of this, we evaluated 18 machine learning methods in this research effort by creating these models from the Volve drilling dataset of the North sea in order to compare and rate their performance. The end result of this work offers a blueprint for choosing algorithms and features for developing ML solutions for optimizing oil drilling. Hybridization and the creation of real-time ROP prediction algorithms can both benefit from this.

REFERENCES

[1.] Acheme, I. D., Vincent, O. R., & Olayiwola, O. M. (2022). Data Science Models for Short-Term Forecast of COVID-19 Spread in Nigeria. In Decision Sciences for COVID-19 (pp. 343-363). Springer, Cham. https://doi.org/10.1007/978-3-030-87019-5_20

[2.] Al-khudiri, M. M., Al-sanie, F. S., Paracha, S. A., Miyajan, R. A., Awan, M. W., Aramco, S., Kashif, M., and Ashraf, H. M. (2015). Application Suite for 24 / 7 Real-Time Operation Centers 2.Operation Centers ' Systems.

[3.] Alsaihati, A., Elkatatny, S., & Gamal, H. (2022). Rate of penetration prediction while drilling vertical complex lithology using an ensemble learning model. Journal of Petroleum Science and Engineering, 208, 109335.

[4.] Ameloko A.A., Uhegbu G.C. and Bolujo E. (2019) Evaluation of Seismic and petrophysical parameters for hydrocarbon prospecting of G-field, Niger Delta, Nigeria Journal of Petroleum Exploration and Production Technology (2019) 9:2531–2542.