DATA PREPROCESSING IN OIL AND GAS INDUSTRY: PREDICTION OF ROP AS A CASE STUDY

ABSTRACT

Machine learning and Artificial Intelligence is the key to optimizing processes in the Oil industry and saving cost, in a time whereby the growth of the industry is threatened by climate change and global crisis, Oil Industries are using new methods to make sure their processes earn profits instead of cost incurring. Big Oil Giant are investing a lot into Research and Development, Artificial Intelligence and Cloud Computing. They are also doing a lot to learn from the data they obtain from drilling processes, well logging and wellheads so they can improve processes and maximize profit.

In this project, the Supervised Machine Learning Technique is used in the prediction of the Rate of Penetration using variables like the weight of bit, rotation speed, and standpipe pressure. Before building the machine learning model in this case which is the Random Forest Regressor, rigorous data preprocessing is employed using various statistical measures and domain knowledge of the variables that are used to predict the ROP. From the statistical measures carried out, we begin to see the skewness of each variable, the cardinality of each variable and the frequency. Using domain knowledge of how these variables are generated on the field, we are able to come up with a comprehensive approach of categorizing each variable as a continuous or categorical variable and also applying various statistical measures to prepare the data for the machine learning model. These go a long way in improving the accuracy of the model and the predictive power of the model as it helps the Machine Learning Algorithm to avoid any form of generalization and reduces biases.

The results gotten from running the Model with ordinary data cleaning processes are compared to results obtained using extensive data cleaning processes and it can be clearly seen that using rigorous data preprocessing techniques with expert domain knowledge on the variables improves the model by a far higher accuracy.

1.0 INTRODUCTION

Machine learning is an automated process that extracts patterns from data, i.e. learns from the data. The applications of Machine learning are diverse but is still very limited to applications of Artificial intelligence in the Oil Industry and its potential has not fully been exploited. Data preprocessing is a very important process in Machine learning as it helps us understand patterns and variables that will be used in building the model. Because of how widespread and how implementable all this algorithms are, it is easy for analysts to jump into building the model without going through proper and rigorous data cleaning steps that will give insight about the data and help improve the accuracy of the model. Machine learning is basically divided into three namely : Supervised learning, Unsupervised Learning and Reinforced learning, for this particular project a Supervised learning algorithm is used in the prediction of the Rate of Penetration using variables such as: Weight on Bit, Mode of Operation, Stand Pipe Pressure, Torque, Flow rate and Rotary Speed. The Random Forest Regressor algorithm is used in constructing the predictive model. Supervised Machine learning techniques learns the model of the relationship between the descriptive features as mentioned above and the target variable which is the Rate of Penetration based on a set of historical examples or instances. We can then use this mode to make predictions for new instances.

In machine learning one has to be very careful when building the model and take note of significant features in the dataset, the truth about machine learning algorithms is that they have been programmed to generalize In the absent of an obvious prediction, so if care in not taken when developing the model, the algorithm can over fit the model and hence give wrong results.

A better way of preventing errors such as this and knowing if the model is over fitted is to go through the 4-step process:

- Data understanding
- Data preparation
- Modelling
- Evaluation

DATA UNDERSTANDING:

The analyst has to understand the goal, the different data sources and also where the data is being derived from. For this project the data is gotten directly from drilling operations and there are a lot of features that comes with it, using domain knowledge from previous research work, the analyst can narrow down and extract features which are of high performance and those that probable have a form of relationship with the target variable. It is also important that the analyst know where the data is coming from and whether it is recorded from sensors or gotten manually from gauges.

Different acquisition techniques are used while drilling ranging from mugging while drilling (MWD), Logging while drilling (LWD), Formation evaluation while drilling (FEWD), and all these parameters helps the drillers while drilling.

Some of the measurements done at the surface include:

- Hook displacement
- Hook load
- Rotations per minute
- Torque
- Stand pipe pressure
- Pump stroke rate
- Inflow rate
- Outflow rate
- Mud tank level
- Density in and out
- Cuttings flow
- Mud temperature in and out
- Mud rheology properties at atmospheric and ambient temperature

With all these information, certain parameters can be computed such as total bit depth, weight on bit, rate of penetration and many others.

The table below shows how the features used in the development of the model was recorded

Measurement	How	Where
Stand pipe pressure	Pressure sensor	Stand pipe manifold
Mud flow in	Flow meter Stroke pump rate	After the pumps Stroke pump rate
Input mud temperature	Temperature sensor	In the mud tanks
Hook load	Pressure on the dead line anchor or tension on the deadline	Rig floor
RPM	Number of rotations count	Rig floor

Torque	Electrical engine intensity	Rig floor
Cutting flow	Mass flow of cuttings discarded at the shale shakers	Shale shakers

TABLE 1.1

DATA PREPARATION:

Building a predictive model requires the data to be organized and ready for the machine learning model, data preparation ranges from data extraction, data clean-up, feature engineering and data preprocessing which is one of the strong holds that many analyst don't pay attention to. In the next chapter, rigorous data preprocessing techniques will be applied to the development of the Rate of penetration model and the effect of each parameters chosen i.e. the descriptive features which will be used in the development of the model: Weight on bit, Mode of Operation, RPM, SPP, Torque, Flow rate will all be analyzed against the target feature (Rate of penetration).

MODELLING:

There are various machine learning algorithm that can be used in the development of the model, the common principle is to try all using cross validation and see which model performs better. In this project the random forest algorithm, an ensemble learning technique is used for the modelling because of its ability to perform better with categorical features and it is known to reduce biasness.

EVALUATION:

Using various statistical metric, we are able to draw conclusion to the results gotten from the model. In order to test the accuracy of the model and to know whether the model is overfitted, the data is divided into a training dataset and a test dataset. The training dataset is used in the preparation of the model , the model is then used to predict the test data and evaluated, the results gotten from the model are used to judge whether the model is overfitted or not. A scenario where you have an overfitted model is one in which you get an high accuracy for the predictions made on the training data set and a low accuracy on the test data with a huge gap.

2.0 DATA PREPARATION

Data preparation is one integral aspect of machine learning that plays a huge role in improving the accuracy of the model. It can be divide into two categories:

- Data Exploration
- Data Preprocessing

The importance of these two processes is equal to that of building the model and most times more work is spent on preparing the data compared to building the machine learning model, either ways it's very important that thorough data preparation be done before implementation of the machine learning model so as to get insights about the features that will be used in building the model.

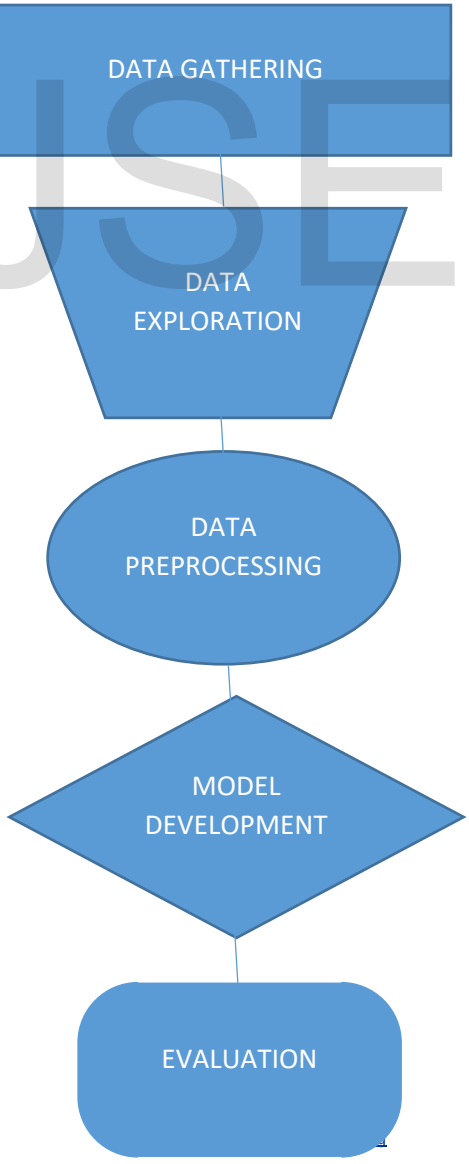


FIGURE 1 - FLOWCHART

2.1 DATA EXPLORATION

Data exploration is a key part of both data understanding and data preparation. There are two goals in Data Exploration. The first goal is to fully understand the characteristics of the features in the dataset. It is important that for each feature we understand characteristics such as the types of values a feature can take, the ranges into which the values in a feature fall, and how the values in a dataset for a feature are distributed across the range that they can take. We refer to this as getting to know the data.

The second goal of data exploration is to determine whether or not the data suffers from any data quality issues that could adversely affect the models that we build. Examples of typical data quality issues include an instance that is missing values for one or more descriptive features, an instance that has an extremely high value for a feature, or an instance that has an inappropriate level for a feature. Some data quality issues arise due to invalid data and will be corrected as soon as we discover them. Others, however, arise because of perfectly valid data that may cause difficulty to some machine learning techniques. We note these types of data quality issues during exploration for potential handling when we reach the modeling phase of a project.

In Data Exploration, we take note of various instances that gives us more insights about the data and take not of data points that have different variations compared to others, by doing these we are able to gain more insight about the data and it helps in building a more accurate model.

For this project we will divide the Data Exploration into three phases:

- Detecting Missing values
- Irregular Cardinality
- Outliers

2.1.1 Detecting Missing Values

If features have missing values, we must first determine why the values are missing. Often missing values arise from errors in data integration or in the process of generating values for derived fields. If this is the case, these missing values are due to invalid data, so the data integration errors can be corrected, and the ABT can be regenerated to populate the missing values. Missing values can also arise for legitimate reasons and that is why proper analysis must be done before removing the missing values.

The amateur approach is to get rid of the missing values which is wrong as this might affect the model we are trying to build.

Below is a diagram showing the features that have missing values present, the missing data points are identified by the yellow lines and as can be seen varies for each feature and are absent in some in
at

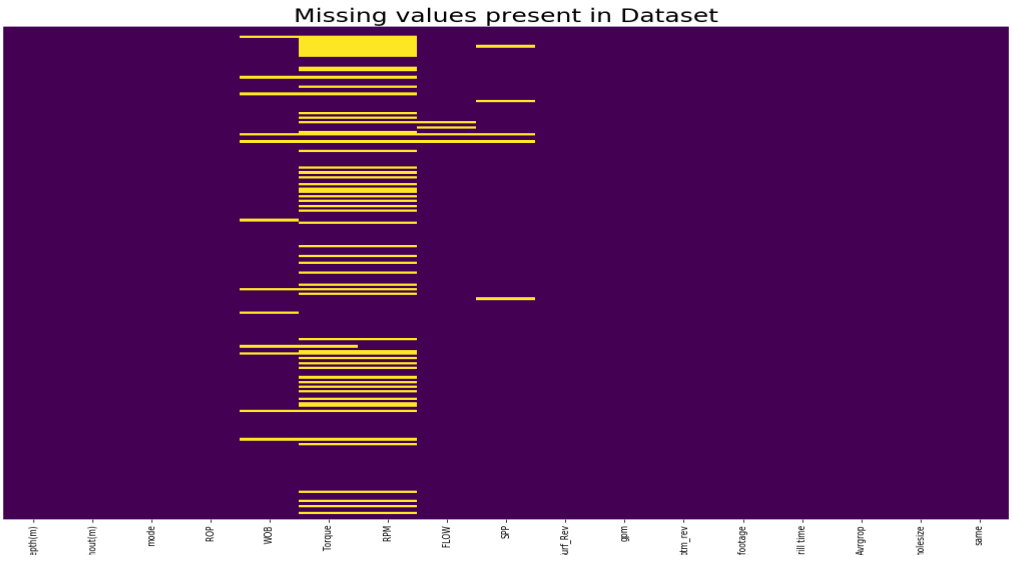


FIGURE 2 – Missing values of each Feature

The following diagram shows the missing values present in the data and it is reinforced by the data below:

Features	Number of Missing Values
Depth(m)	0
Depthout (m)	0
mode	0
ROP	0
WOB	12
Torque	60
RPM	59
FLOW	4
SPP	5
Surf_Rev	0
gpm	0
btm_rev	0
total_footage	0
total_drill time	0
Avrgrop	0
holesize	0
same	0

TABLE 2.1

CORRECTIONS MADE TO REPLACE THE MISSING VALUES

Instead of getting rid of all the cells that contain the missing values, the following corrections were made which gave more insights about the data and helped to improve the accuracy of the model.

Weight on bit Missing Values

After careful analysis, it was observed that the Weight on Bit had a lot repeated values and the values are in a small range thereby allowing us to impute values that most likely fall into that range. Using the python skLearn preprocessing module 'Imputer', we are able to replace the null values of the Weight on bit with the most frequent value.

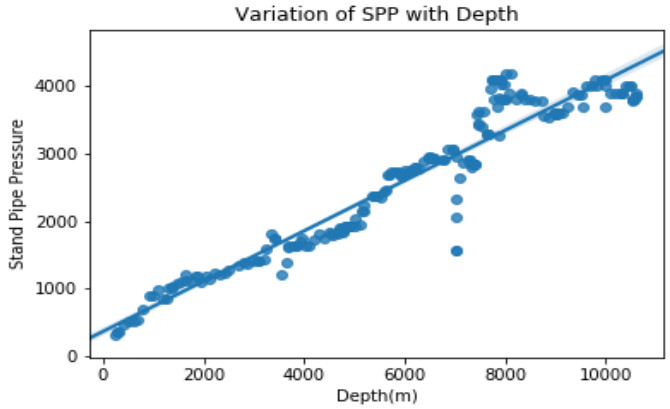
Torque and RPM Missing Values

From table 2.1 above we can see that the torque and RPM have the same number of missing values, this called for further investigation and after careful analysis, it was discovered that the RPM and torque had missing values where the mode of operation (which is another important feature) indicated sliding mode instead of rotation mode. The mode of operation tells us that the drilling occurred using the conventional rotary system and also a mud motor, whenever the mud motor is being operated, the rotary table is shut off thereby recording missing values for the torque and RPM.

To make the model work better, the missing values present in Torque and RPM columns are replaced with **0**, this allows the model to understand the relationship between the Torque, RPM and mode of operation better

Stand Pipe Pressure

The stand pipe pressure showed a significant increase with depth and hence a linear relationship can be seen to exist between the depth and Stand Pipe Pressure, using these linear relationship, we can use the interpolate method to replace with the missing values. The interpolate method works where there is linear relationship between the two variables, in this case it is the Depth and the Stand pipe pressure as can be seen in the diagram below



FFFF

FIGURE 3 – Variation of SPP with Depth

2.1.2 IRRGEULAR DATA CARDINALITY

The Cardinality shows the number of distinct values present for a feature within a dataset. A data quality issue arises when the cardinality for a feature does not match what we expect, a mismatch called an irregular cardinality. The first thing is to check for features with a cardinality of 1. This indicates a feature that has the same value for every instance and contains no information useful for building predictive models. Features with a cardinality of 1 should first be investigated to ensure that the issue is not due to any data processing error. If this is the case, then the error should be corrected. If the generation process proves to be error-free, then features with a cardinality of 1, although valid, should be removed from the dataset because they will not be of any value in building predictive models.

The second thing is to check for is the cardinality of categorical features incorrectly labeled as continuous. Continuous features will usually have a cardinality value close to the number of instances in the dataset. If the cardinality of a continuous feature is significantly less than the number of instances in the dataset, then it should be investigated. Sometimes a feature is actually continuous but in practice can assume only a small range of values—for example, the number of children a person has. In this case there is nothing wrong, and the feature should be left alone. In other cases, however, a categorical feature will have been developed to use numbers to indicate categories and might be mistakenly identified as a continuous feature in a data quality report. Checking for features with a low cardinality will highlight these features. For example, a feature might record gender using 1 for female and 0 for male. If treated as a continuous feature in a data quality report, this would have a cardinality of 2. Once identified, these features should be recoded as categorical features.

Using python analysis tool, we can generate the cardinality of each feature and the no of instances they appear in the dataset. From the table below we can see the cardinality of each feature:

Features	cardinality	total no of occurrences
Depth(m)	201	201
mode	2	201
RPM	8	201
Surf_Rev	113	201
gpm	16	201
btm_rev	172	201
total_footage	81	201
total_drill time	113	201
Avrgrop	191	201
holesize	2	201
Torque	21	201

WOB	16	201
SPP	98	201

TABLE 2.2

From the table above the cardinality of the mode of operation, RPM, Torque, holesize, gpm and WOB are smaller compared to the number of occurrences, this calls for a more in-depth view into these features so as to classify them into categorical and continuous variables.

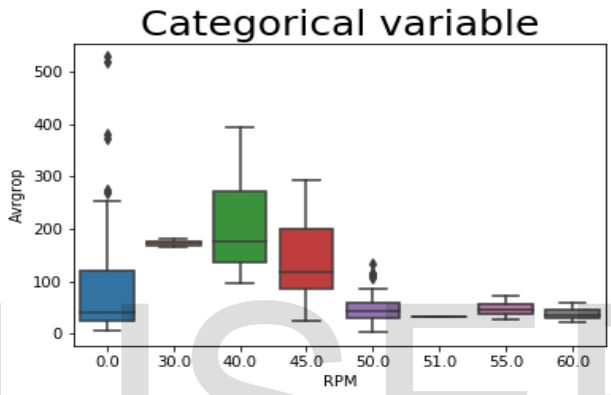


FIGURE 4 – Variation of RPM with ROP

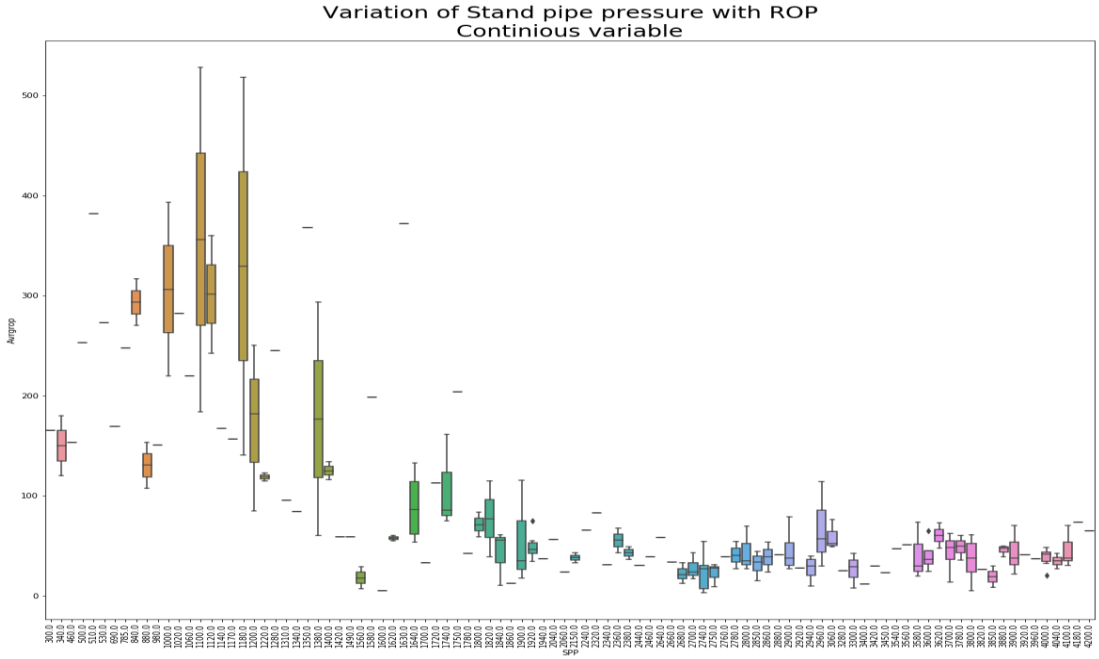


FIGURE 5 – Variation of stand pipe pressure with ROP

From the two diagrams above, we can see the difference between the variation of the RPM and SPP with the Rate of Penetration, this helps us to classify the RPM as a categorical variable because of low cardinality and the SPP has a continuous variable because of high cardinality.

After careful analysis, features are grouped into categorical and continuous features based on their cardinality, features whose cardinality is far lower than the number of instances are converted to categorical variables, this allows the model to better understand the relationship between the variables and reduce generalization.

Continuous features	**Categorical features**
Rate of penetration	Mode of Operation
Stand Pipe Pressure	Hole size
Total footage	RPM
	Torque
	GPM

TABLE 2.3

2.1.3 OUTLIER DETECTION:

Outliers are values that lie far away from the central tendency of a feature. There are two kinds of outliers that might occur in a dataset: invalid outliers and valid outliers. Invalid outliers are values that have been included in a sample through error and are often referred to as noise in the data. Invalid outliers can arise for all sorts of different reasons. For example, during a manual data entry process, a fat fingered 5 analyst may have entered 100,000 instead of 1,000. Valid outliers are correct values that are simply very different from the rest of the values for a feature, for example, a billionaire who has a massive salary compared to everyone else in a sample. There are two main ways that the data quality report can be used to identify outliers within a dataset. The first is to examine the minimum and maximum values for each feature and use domain knowledge to determine whether these are plausible values. Outliers identified in this way are likely to be invalid outliers and should immediately be either corrected, if data sources allow this, or removed and marked as missing values if correction is not possible. In some cases we might even remove a complete instance from a dataset based on the presence of an outlier. The second approach to identifying outliers is to compare the gaps between the median, minimum, maximum, 1st quartile, and 3rd quartile values. If the gap between the 3rd quartile and the maximum value is noticeably larger than the gap between the median and the 3rd quartile, this suggests that the maximum value is unusual and is likely to be an outlier. Similarly, if the gap between the 1st quartile and the minimum value is noticeably larger than the gap between the median and the 1st quartile, this suggests that the minimum value is unusual and is likely to be an outlier. The outliers shown in box plots also help to make this comparison. Exponential or skewed distributions in histograms are also good indicators of the presence of outliers. It is likely that outliers found using the second

approach are valid outliers, so they are a data quality issue due to valid data. Some machine learning techniques do not perform well in the presence of outliers, so we should note these in the data quality plan for possible handling later in the project.

For this project the following measures were taken for the detection of outliers:

1) ROP values lesser than 0 were removed as it is not possible to obtain a negative ROP
2) Using the mean and standard deviation of the ROP, an interquartile range was used in the removal of outliers
3) All null values from the other features had already been reviewed and since most of the features are categorical features, they have bi/multi modal form of distribution
4) The Random forest regressor performs accurately even in the presence of outliers and measures were taken to reduce the outliers present in the dataset.

2.2 DATA PREPROCESSING

Instead of explicitly handling problems like noise within the data in an ABT, some data preparation techniques change the way data is represented just to make it more compatible with certain machine learning algorithms. This section describes two of the most common such techniques: binning and normalization. Both techniques focus on transforming an individual feature in some way.

2.2.1 NORMALIZATION

Having continuous features in your dataset that cover very different ranges can cause difficulty for some machine learning algorithms. For example, a feature representing customer ages might cover the range [16, 96], whereas a feature representing customer salaries might cover the range [10,000, 100,000]. Normalization techniques can be used to change a continuous feature to fall within a specified range while maintaining the relative differences between the values for the feature. The simplest approach to normalization is range normalization, which performs a linear scaling of the original values of the continuous feature into a given range. We use range normalization to convert a feature value into the range [low, high] as follows:

$$\acute{a}_i = \frac{a_i - \min(a)}{\max(a) - \min(a)} * (high - low) + low$$

Where $\acute{a}_i$ is the normalized feature value, ai is the original value, min (a) is the minimum value of feature a, max (a) is the maximum value of feature a, and low and high are the minimum and maximum values of the desired range. Typical ranges used for normalizing feature values are [0, 1] and [−1, 1].

The Table below shows the range of maximum and minimum values for each features and it can be seen that for the continuous features, the range is large compared to the categorical values that just have [0&1], hence there is need to use normalization so as to come up with a more equivalent range.

FEATURES	MAX	MIN
Depth(m)	10613	256
Surf_Rev	10250	0
gpm	900	460
Torque	16	0
SPP	4200	300
S	1	0
16	1	0
30	1	0
40	1	0
45	1	0
50	1	0
51	1	0
55	1	0
60	1	0
Avrgrop	253.0435	2.903226

TABLE 2.4

2.2.2 TRANSFORMATION

The transformation used in the project is dummyfication and this involves transforming the categorical variables into binary digits of [1&0], during exploratory data analysis, the features where divided into categorical and continuous variables based on how their cardinality and how they affect the response variable (ROP).

RPM, hole size, Torque, gpm and mode of operation have been identified as the categorical variables and hence are converted to dummy variables [0&1], this will allow the machine learning model predict the response variable effectively because using the categorical variables as continuous variables will only reduce the accuracy of the model.

For a feature that has 6 unique values in the entire set of instances, we can create 6 unique columns of [0&1] where the value is 0 when that unique value is not present and 1 when the value is present.

Converting all these variables is very important especially when the machine learning algorithm used is an ensemble form of learning algorithm in this case is the Random forest regressor. The algorithm is

able to split the data easily and see the relationship clearer compared to taking the variables as continuous variables and spiting it using uncertain ranges and conditions.

From the table above, some features have already been converted into categorical variables and hence their value ranges between [0&1].

The accuracy gotten as shown in the results supports the assumptions that the categorical variables be converted to dummy variables.

3.0 RESULTS

3.1 EVALUATION

The basic process for evaluating the effectiveness of predictive models is simple. We take a dataset for which we know the predictions that we expect the model to make, referred to as a test set, present the instances in this dataset to a trained model, and record the predictions that the model makes. These predictions can then be compared to the predictions we expected the model to make. Based on this comparison, a performance measure can be used to capture, numerically, how well the predictions made by the model match those that were expected. There are different ways in which a test set can be constructed from a dataset, but the simplest is to use what is referred to as a hold-out test set. A hold-out test set is created by randomly sampling a portion of the data we created in the Data Preparation phase. This random sample is never used in the training process but reserved until after the model has been trained, when we would like to evaluate its performance.

The process is known as the train, test and split which in simple words means dividing the dataset into a training and test set.

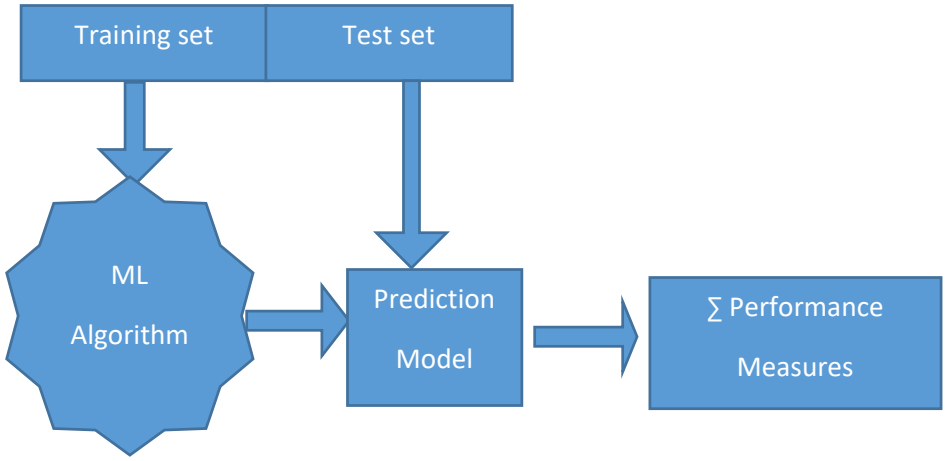


FIG 6

Using a hold-out test set avoids the issue of peeking, which arises when the performance of a model is evaluated on the same data used to train it; because the data was used in the training process, the model has already seen this data, so it is probable that it will perform very well when evaluated on this data. An extreme case of this problem happens when k nearest neighbor models are used. If the model is asked to make a prediction about an instance that was used to train it, the model will find as the

nearest neighbor, for this instance, the instance itself. Therefore, if the entire training set is presented to this model, its performance will appear to be perfect. Using a hold-out test set avoids this problem, because none of the instances in the test set will have been used in the training process. Consequently, the performance of the model on the test set is a better measure of how the model is likely to perform when actually deployed and shows how well the model can generalize beyond the instances used to train it. The most important rule in evaluating models is not to use the same data sample both to evaluate the performance of a predictive model and to train it.

3.2 PERFORMANCE MEASURES

When evaluating the performance of prediction models built for continuous targets, there are fewer options to choose from. In this section we describe the performance measures used for the prediction of ROP. The basic process is the same as for categorical targets. We have a test set containing instances for which we know the correct target values, and we have a set of predictions made by a model. We would like to measure how accurately the predicted values match the correct target values.

For this project, the training data and test data were split by 80% and 20% respectively, the training data was used in the development of the model and the test data was used for evaluation.

The following are the various metrics used in the evaluation of the model

- **Mean Absolute Error –** Mean Absolute Error is the average of the absolute difference between the Original Values and the Predicted Values of data. It gives us the measure of how far the predictions were from the actual output i.e., the magnitude of the error. However, they don't give us any idea of the direction of the error i.e. whether we are under predicting the data or over predicting the data. Below is the formula :

$$MAE = \frac{1}{N}\sum_{i=1}^{N} y_i - \hat{y}_i$$

Where $y_i = actual\ value\ and\ \hat{y}_i = predicted\ value$

MAE value of 0 indicates no error or perfect predictions.

- **Mean Squared Error – Mean** Squared Error is much like Mean Absolute Error except that it finds the average squared error between the predicted and actual values. It also provides a rough idea of the magnitude of the error.

An MSE of zero means that the estimator predicts observations of the parameter with perfect accuracy, is ideal but is generally not possible. The smaller the means squared error, the closer you are to finding the line of best fit. Hence, MSE is a measure of the quality of an estimator and is always non-negative, and values closer to zero are better.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

- **Root Mean Squared Error –** Root Mean Squared Error (RMSE) measures the average magnitude of the error by taking the square root of the average of squared differences between prediction and actual observation. It tells us how concentrated the data is around the line of best fit. The RMSE is the square root of the variance of the residuals. Lower values of RMSE indicate a better fit. RMSE is a good measure of how accurately the model predicts the response.

The RMSE will always be larger or equal to the MAE; the greater the difference between them, the greater the variance in the individual errors in the sample. If the RMSE=MAE, then all the errors are of the same magnitude.

$$RMSE = \sqrt{MSE}$$

- **R Squared** – The r2_score or commonly known as the R^2 (R-squared) is the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It is known as the coefficient of determination. It is a statistical measure of how close the data are to the fitted regression line or indicates the goodness of fit of a set of predictions to the actual values. The value of R^2 lies between 0 and 1 where 0 means no-fit and 1 means perfectly-fit.

RESULTS OBTAINED WITH NORMAL PREPRROCESSING TECHNIQUES

METRICS	Training set	Test set
MSE	0.06078	0.14369
MAE	0.16358	0.27093
R Squared	0.93757	0.86978

Table 3.1

RESULTS OBTAINED WITH RIGOROUS PREPRROCESSING TECHNIQUES

METRICS	Training set	Test set
MSE	0.04827	0.08029
MAE	0.15112	0.23078
R Squared	0.95248	0.90615

Table 3.2

Emphasis is given to the test data set and it can be seen from the above that with we are able to get a better result in our test set with extra preprocessing techniques applied.

4.0 CONCLUSION

This research focuses on the preprocessing techniques that should be done before running the machine learning model and not the Machine learning model, preparing the data is a very important aspect when it comes to model development and the processes above have showed that by using various statistical measures, we are able to gain insight about our data this helps us in preparing the data in such a way that we can account for any misappropriations.

There are still more advanced techniques and statistical measures that can be used to gain more insight about the data and that can also help in feature engineering and evaluation such as correlation methods, displaying scatter plot diagrams that shows the relationship between the features and how they affect the Target feature, also by using cross validation an extension of evaluation metrics which involves splitting the training set into different folds so as to reduce overfitting before evaluating the performance on the test set. All these are various methods that can be used in improving the accuracy of the model but this research focuses on how one can use statistical measures and domain knowledge on the field to understand the data and prepare it for modelling.

For this project we were able:

1) Replace null values of certain features instead of getting rid of them by carefully examining them, from the analysis these null values were generated due to the difference mode of operation during the drilling process and gave more insights about how the data should be structured.
2) Distribute the features into categorical and continuous features using their individual cardinality, numerical features are mostly categorized as continuous variables but after careful analysis, we were able to get insights about the data and it was revealed that certain numerical features are actually categorical and have a defined range, these features where therefore converted to categorical features and it helped improve the accuracy of the model
3) Using various statistical measures such as the mode and interpolation method, we were able to replace missing values of certain features which showed a dependent relationship with other features allowing the missing values to be imputed based on the relationship with the other features.
4) Outliers were removed using domain knowledge of the variables and setting a defined range so as to improve the accuracy of the model
5) Normalization of the ranges of the continuous features allowed all data points to fall within a particular range of [-1&1] so as help the model understand the relationship better.

It is not just enough to build the Machine Learning model, there are certain things we can do that will help us understand how to develop the model better and that is the mantra of this research, understanding the data so as to develop a better Machine learning model to make better decisions.