Name: Oyindamola Obisesan

Course: DSA 5900

Credit Hours: 4 units

Introduction:

Rate of Penetration (ROP) is a singular most important parameter during the drilling of an oil and gas well. Drilling days per thousand feet (DDPTF) is a measure of drilling performance. The DDPTF is highly influenced by the ROP which is why the ROP is a very important drilling variable. The aim of a drilling operation is to reduce drilling time i.e., maximize ROP and reduce drilling risk. There are approaches to optimize ROP, optimum drilling variables section prior to a run and real time optimization of ROP by varying Weight on Bit (WOB), Revolution Per Minute (RPM), Torque.

There have been efforts (theoretical and experimental) to model ROP as a mathematical function of some variables, but this is not a trivial problem as it is a highly non-linear problem. The most common traditional ROP model (physics-based model) is the Bourgoyne and Young equation based on analytical equations. The physics made model attempts to understand the physical nature of well drilling, describing the drilling activities with analytical equations. This work aims to use machine learning algorithms to understand the relationship between these parameters and use this understanding to optimize the ROP.

Objective

Technical Objectives:

The plan of this work is however to predict and optimize the ROP and reduce drilling time by adjusting surface drilling parameters such as WOB, torque, surface RPM, flowrate, MSE, HSI. The importance to this is to understand the changes in ROP based on the drilling parameters. The objective is to access how some important variables (RPM, WOB, flowrates) affect drilling parameter; ROP. In this way, the ROP can be increased by tweaking the other drilling parameters.

Individual Objectives:

- Learn a program better in Python
- Understand dive deeper into data preprocessing and preparation.
- Understand Machine learning models better with these hands-on projects.
- Unite Machine Learning with practical experience in the engineering field.
- Add more knowledge to my petroleum engineering domain and solve a critical problem.

Data Exploration and Analysis

Data Description

The data includes the drilling data such as WOB, ROP, RPM, Torque, SPPA and Formation Evaluation Data such as Gamma Ray, Resistivity and other formation evaluation logs from a field in Niger Delta Nigeria. The data collected would be cleaned and processed to remove missing data and other data inconsistency. The data was from downhole tools run while the wells were being drilled. The data is provided in excel sheets. It has columns for: Depth ('Depth'), Equivalent Circulating Density ('ecd'), Gamma Ray ('GR'), Resistivity ('Res B', C, D), Rate of Penetration ('Rop'), Revolution Per Minute ('Rpm'), Stand Pipe Pressure (' Sppa'), ' flow','torque', 'Weight on Bit ('wob'), Mechanical Specific Energy ('Mse') etc. These are raw features measured directly from the formation. Some feature engineering would be done to convert these features to more useful ones for the model.

Data Exploration:

The data is an excel file that contains 22 columns.

The table 1 below shows the output of the unfiltered data:

Res A	Depth	Drhb	ecd	GR	Res B	Res C	Res D	Pef	Robb	Rop
18485	18485	11203	18485	18485	18485	18485	18485	18485	18485	18485
-92.4949	9355.535	-150.828	-119.677	-23.1116	-87.7917	-87.7963	-88.6225	-88.7709	-90.1481	262.6342
965.7218	1597.174	1218.854	1135.501	1149.314	969.5374	969.454	965.402	951.9744	951.8388	933.9549
-9999	6631.5	-9999	-9999	-9999	-9999	-9999	-9999	-9999	-9999	-9999
1.2153	8042	-0.001	10.6442	101.7405	1.2104	1.2099	1.2315	0	0	208.7198
1.417	9197.5	0.0038	10.8079	109.3697	1.5236	1.5782	1.6783	4.1893	2.1421	277.0063
1.6115	10433	0.0182	11.1007	117.278	2.0005	2.2688	2.5061	4.6334	2.2954	366.5508
59.8834	12743.5	0.3502	47.4156	833.9749	2468.162	3000	2815.48	42.8781	2.7306	3494.236

Table 1: Description of Raw Data

The minimum values of some of the columns show -9999 and this is an impossible value. -9999 is usually a placeholder for missing value in this case, possibilities include that the tool failed or the measurements from the tool were wrong. For the initial exploration of the data, these values are substituted to zero. The percentage of the columns that composes of zero were checked. This is shown in Table 2.

Columns	% Zeros
Res A	0.048093
Depth	0.000000
Drhb	0.050906
ecd	0.013038
GR	0.013092
Res B	0.048093
Res C	0.048093
Res D	0.048039
Pef	0.441764
Robb	0.441764
Rop	0.007574
Rpm	0.036354
Sppa	0.009143
tor	0.008656
wob	0.008656
flow	0.674385
pb	0.606059
vn	0.606059
JIF	0.606059
HSI	0.606059
neutron	0.442521
Mse	0.000000

Table 2: Percentage of Zeros per Feature

The columns with high percentages of zeros are removed from the analysis.

The number of colomns were reduced to 15. The new data set is explored to understand the data distribution better (Figure 1).

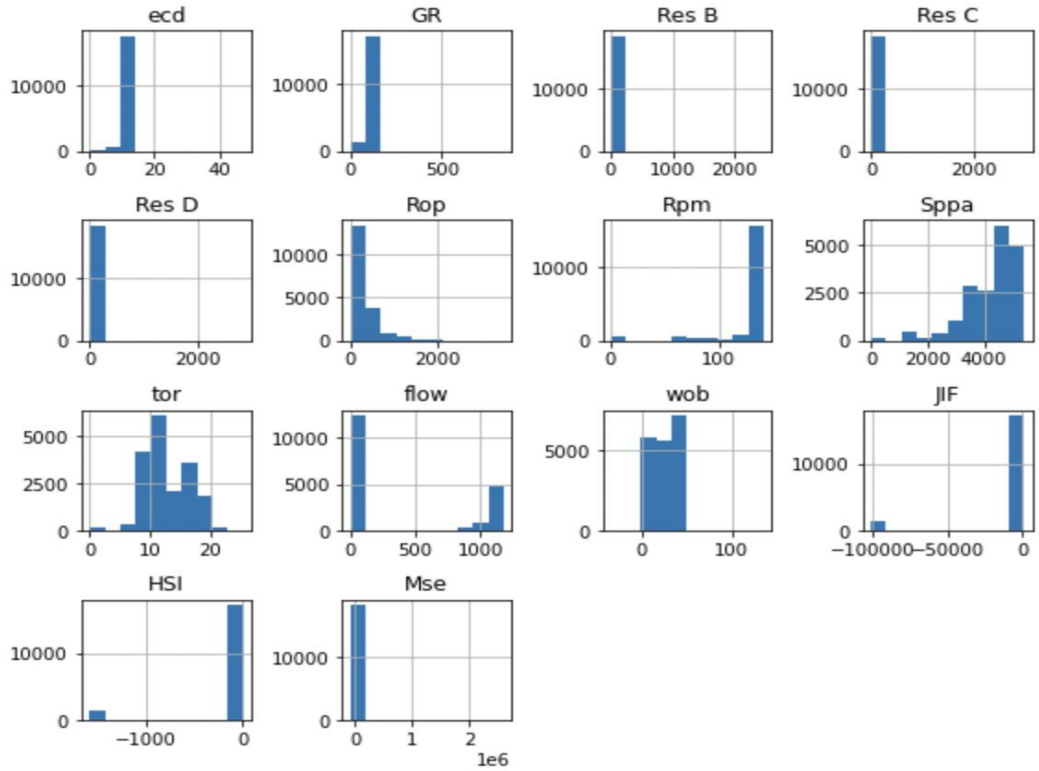


Figure 1: Histogram for Raw Input Features

The MSE, HIS and JIF can still be seen to have negative values. In the real case scenario, this is not feasible. This can be ascribed to error from the tools and they have to be removed from the data to ensure that they do not interfere with the results of the prediction models. The only data that would be considered for these features are those above zero.

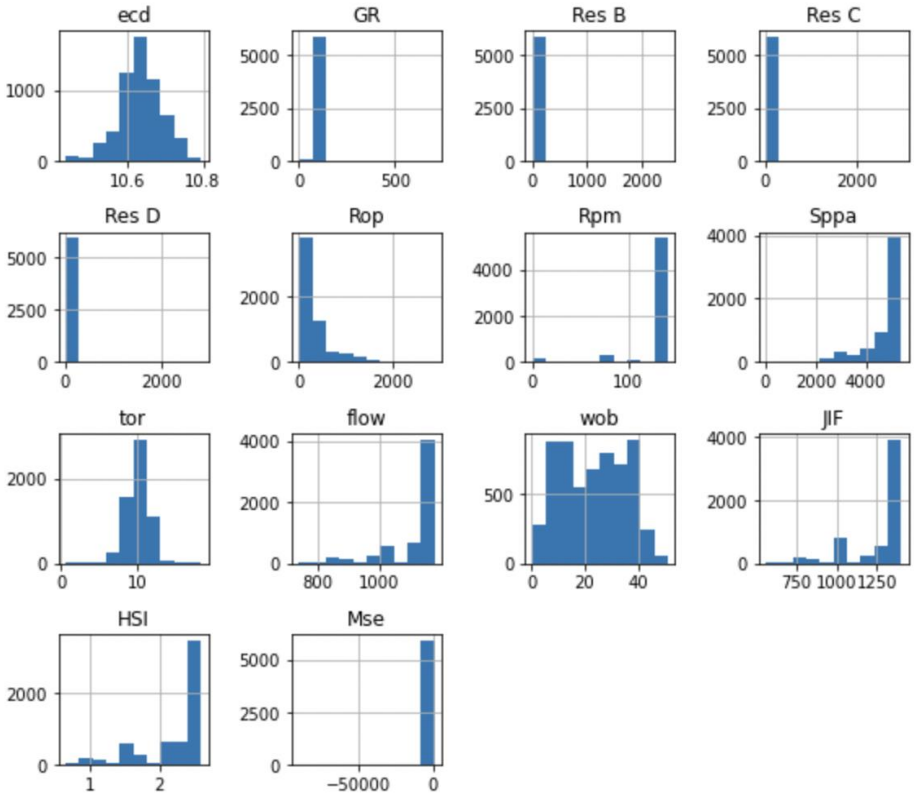


Figure 2: Histogram for Cleaned Input Features

This the new histogram with the values zero and below being removed. A very observable behavior of the some of the histogram is the skewed distribution of the some features. GR, Res B and other features have a non-normal distribution. This is ensure the normaility of the features. Many Machine learning models work best with data that is normally distributed.

```
Lambda value used for Transformation: -0.616028847038434
```

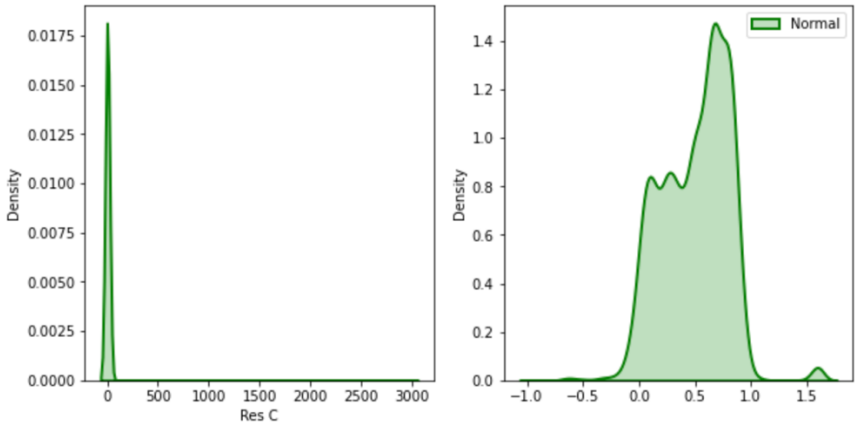


Figure 3: BoxCox Transformation of Res C

```
Lambda value used for Transformation: -0.37408201288441184
```

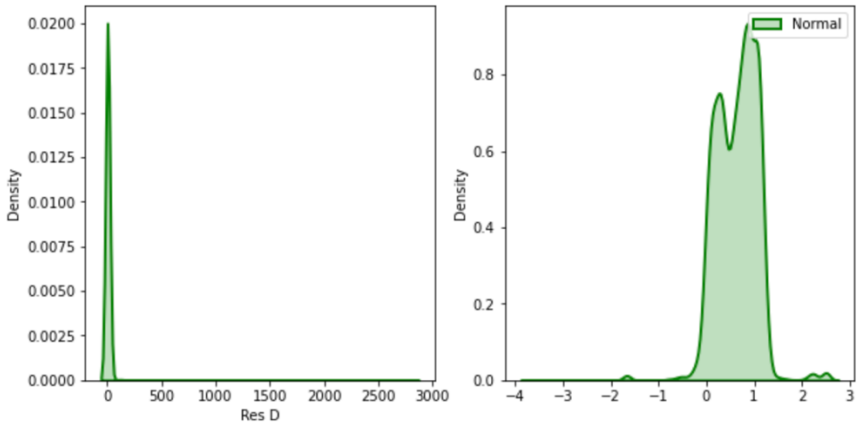


Figure 4: BoxCox Transformation of Res D

```
Lambda value used for Transformation: 0.12569530229995518
```

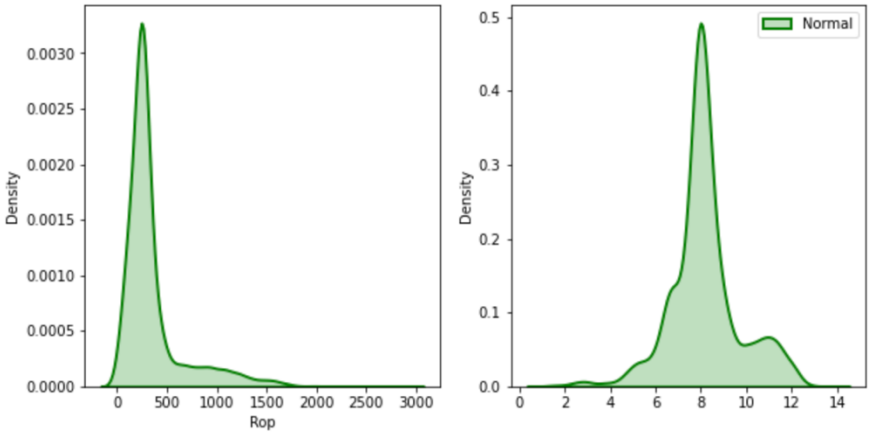


Figure 5: BoxCox Transformation of Rop

Each of the figure show the density plots before and after the transformation. These transformed features are used in further analysis in this work.

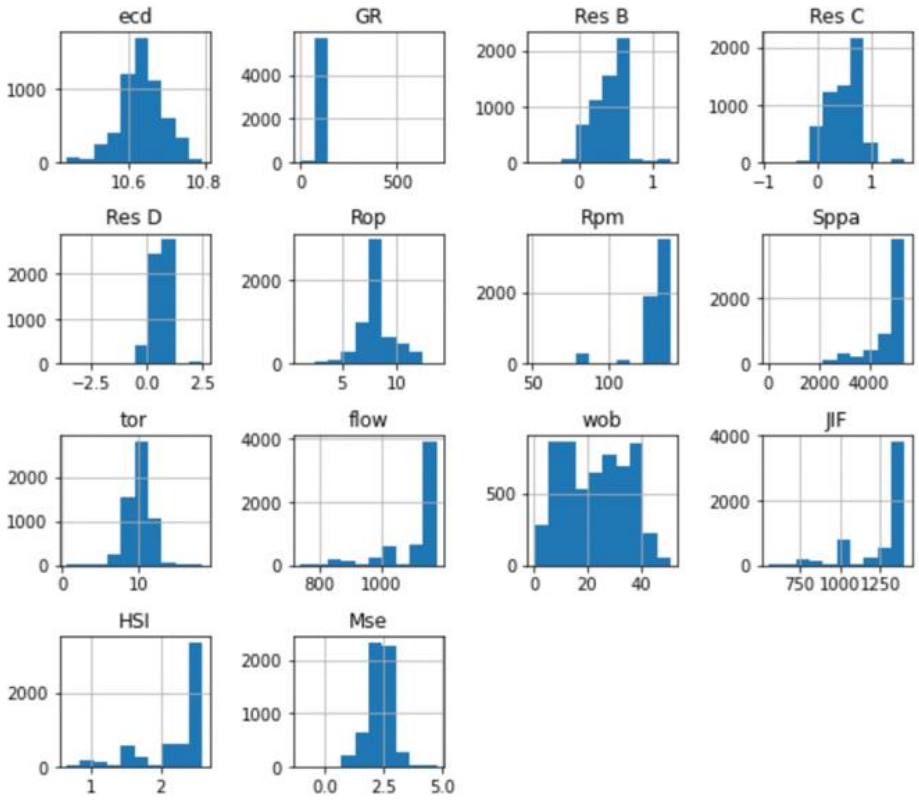


Figure 6: Histogram for Normalized Input Features

The features have a more normal distribution.

Data Preparation

Feature Correlation:

Machine learning models are as good as the data that are put into them. The individual features of the data must be studied to ensure that only important features are put into model.

Data correlation is a way to determine the relationship between multiple features and attributes. Investigating the correlation of a data would help to determine if one data depends on another or if one feature is the cause of another feature. The features can be positively correlated or negatively correlated. It is positively correlated if an increase (decrease) in feature A then there an increase (decrease) in feature B. The fetaures are negatively correlated if an increase in feature A then there is a decrease in feature B. The values for the correlation can range from 0 to 1 or negative in case of negative correlation. The values from 0.5 to 0.8 can show highly correlated features. The correlation in the range of 0.9 to 1.0 shows perfectly correlated features.

If the data contains fetaures that are perfectly correlated postively or negatively, the model would mostly encounter a problem called Multicollinearity. This occurs when one feature in a multiple regression model can linearly predict another one with a high degree of accuracy. This

can cause very wrong and skewed results. However, there are models that re immune to multicollinearity such as decision trees and boosted trees that use tree which picks only one from these perfectly correlated features. The other models used in this work like linear regression do not have this capability.

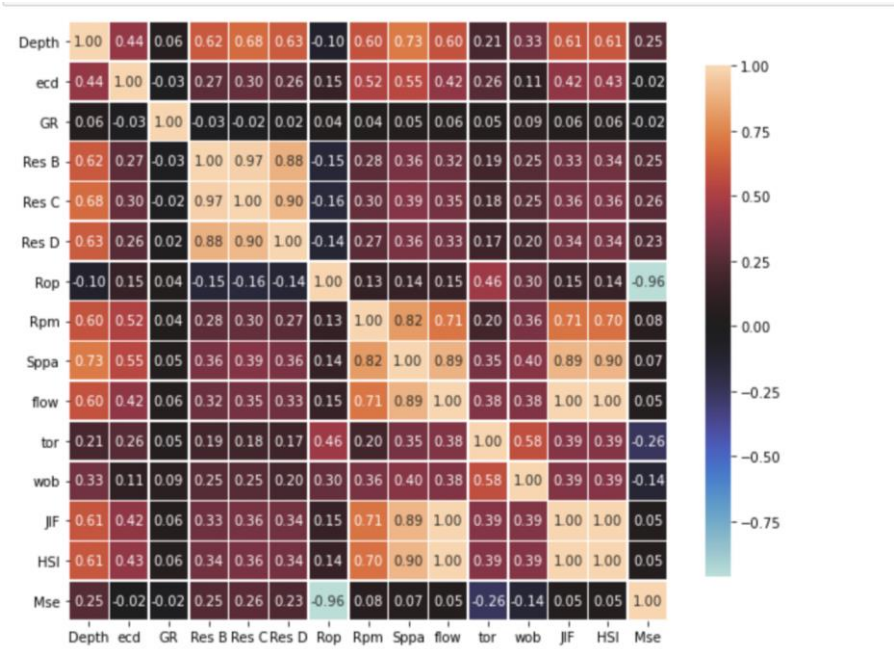


Figure 7: Heat Map of Features

From the heatmap in Figure 7, we can see highly correlated features. The most critical features are those that are perfectly correlated such as Res B, Res C and Res D. Another group of perfectly correlated features are HSI and SPPA, HSI and JIF. In the practial sense, the correlation makes sense as the RES B, C and D are measured from the same tool and they measure the same property of the formation which is the resistivity. Also, HSI and SPPA are calculated from each other with a linear equation same as HSI and JIF. In this case , some of these features have to removed. The Res B would be used and the HIS would also be used. The other perfectly correlated features would be removed.

Outliers:

The zscore is used to check for the outlier. If the score is greater than 3 then the data point is an outlier and it is taken out of the data. This was how outlier removal was done in this analysis.

The MinMax scaler is appled on the data to ensure that each of the features have equal weights in the model. The range of all the features is between 0 and 1. In this way, features with high values are not unfairly favored in the model. This would ensure the regression model coefficient for all the features are of the same unit.

Methodology

Data Clustering

The data is split into two different groups, the first is the drilling data which contains features such as 'Rop','Rpm','tor','wob','HSI','flow' and the formation data which contains features such as 'Res B','GR','Mse'. The formation data is the data that shows the type of formation were are drilling in. During the drilling of the 12.25in section, we encoutered various types of formation and this data would show if we pass through various formation types while drilling. The 12.25 in hole is usually the longest section and it os highly likely that the section is layered with multiple formations.

The plan for the work is to cluster the data based on the type of formation. Clustering would be done on the data to determine if there is enough reason to split the 12.25 in section based on formations or consider it as one formation type.

The formation data is fitted to a kmeans clustering algorithm, This would determine the **Within sum of squares** (SSW) values to check the variance between the clusters. Higher SSW values mean that there is more variance between the clusters and smaller values mean that the clusters are more or less same.

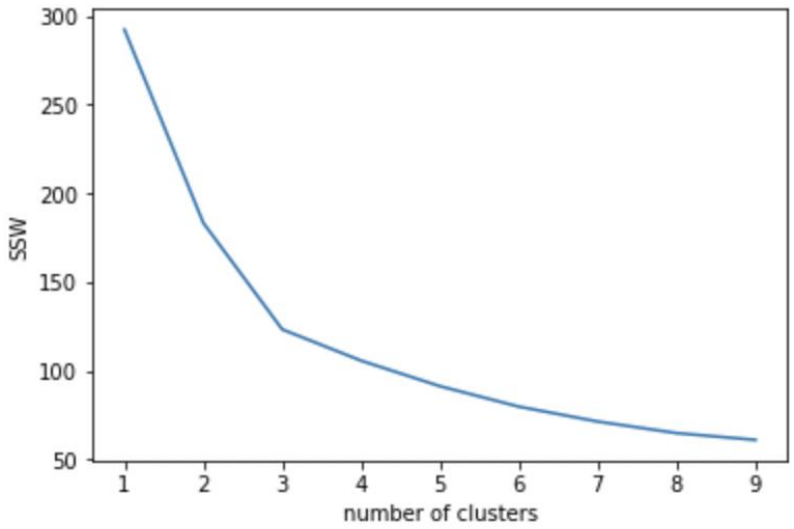


Figure 8: SSW values for Clusters

The plot shows various numbers of clusters being analyzed. After 3 clusters, the SSW value reduces significantly which means that clustering the data into more than 3 clusters do not really conserve any variance or explain data behavior. Therefore, we would pick three as the number of the clusters.

The properties of each of the clusters are studied. From the Figure 9, each of the clusters have differing properties which goes further the need to prove that this bit section contains various

formation types with different behaviors. This is the reason behind the need to split the data into three portion and build models for each section. This gives a model that is suited to each formation present.

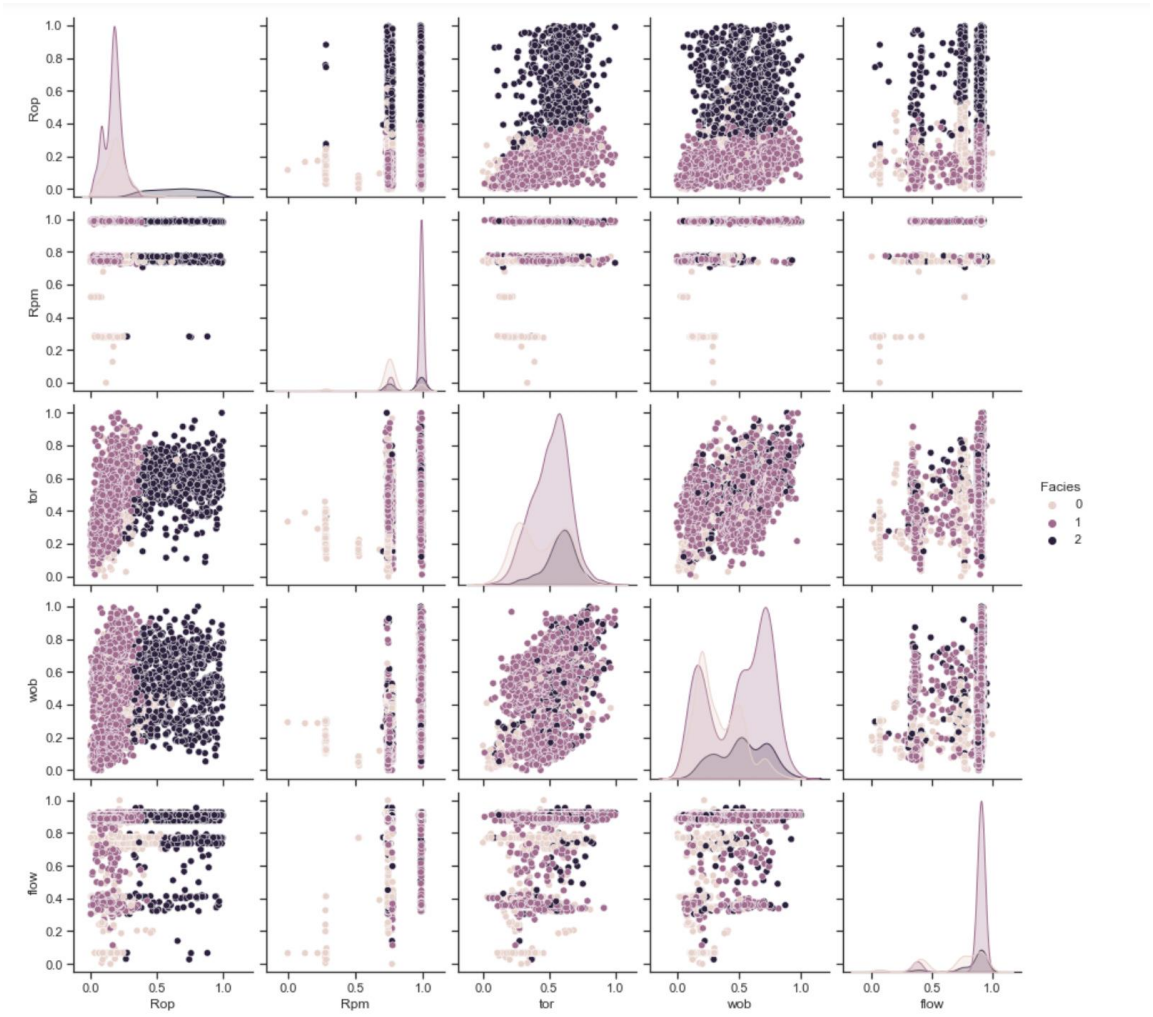


Figure 9: Pairplot for the three data clusters/Facies

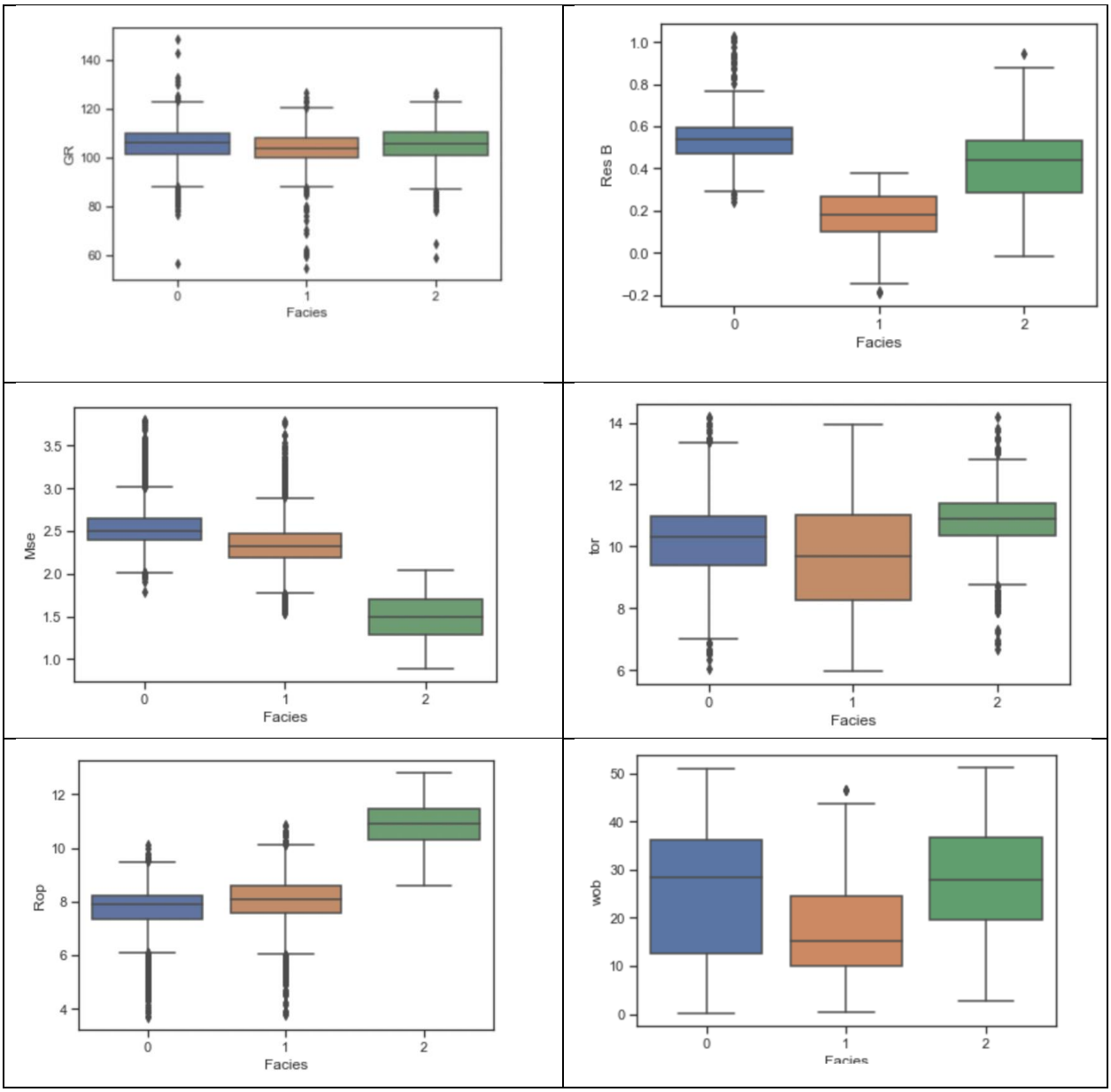


Figure 9: Boxplots Showing Properties for Each Cluster

Machine Learning Algorithms:

The algorithms that are considered in this work include:

- Support Vector Machine (SVM)
- Random Forest
- Multiple linear regression

The best model was selected based on the least error.

The data would be split into training and testing with the cross-validation technique. Hyperparameter optimization and feature importance would be done to determine the most efficient parameter combination. A major part of this work would be to determine the most efficient parameter combination for the optimization of the ROP. Sensitivity analysis would be done on the data to determine the effect of increase and decrease of certain controllable drilling parameters on the ROP.

Model Evaluation Methods:

- R-square
- RMSE
- MSE
- Residual analysis would be done on the results of the model the model to ensure the error of the model is normally distributed. If a pattern is observed in the residual analysis, it means there is a problem of heteroscedastic and it must be resolved in the model until the same error is observed across all the data points. This is to ensure that all the patterns are captured in the model i.e., the model accurately represents the data.

Residual Analysis

The residual analysis was done on the data using a linear regression model. The residual analysis was done on the linear regression model. The distribution of the residuals is not distributed uniformly randomly around the zero x-axes. We can see clusters and patterns in the data. This means that the data cannot be fitted to a linear model.

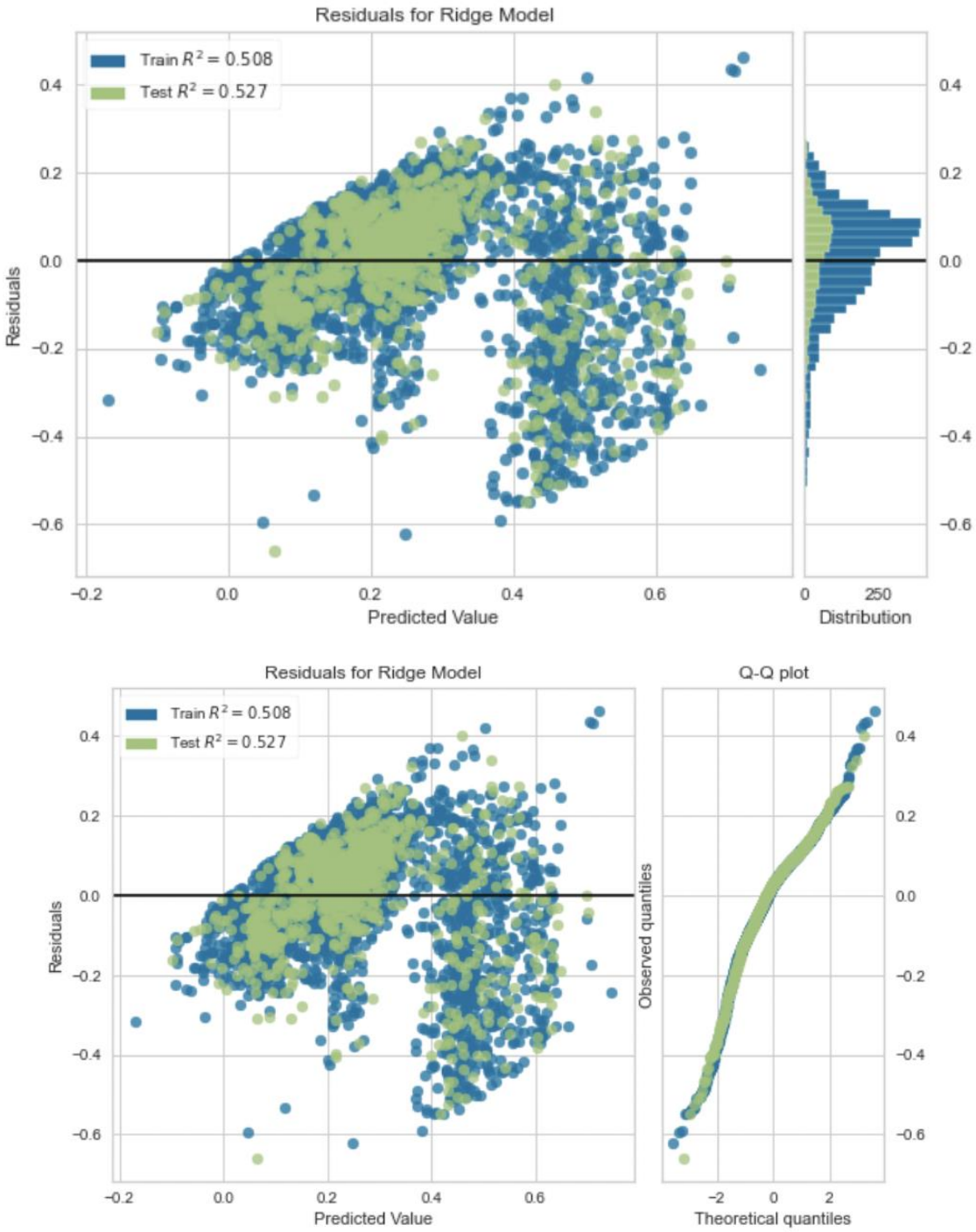


Figure 9: Residual Analysis

The models that were considered were the SVM and Random Forest.

To justify splitting the data into columns, a model was run with all the data in one cluster, inputting the facies as input parameters and writing a model for each cluster. Each of the techniques are compared based on the r2 value and the error analysis.

Modelling

- SVM Model
- Random Forest

The modelling is considered in two cases, where we use all the data as one Facie/Cluster and when we cluster the data and include the Facies in the model estimation.

SVM Model

Method	One Cluster	Data + Facies
R2	0.12	0.78
Root Mean Squared Error	0.18	0.09

Table 2: Model Evaluation results for SVM

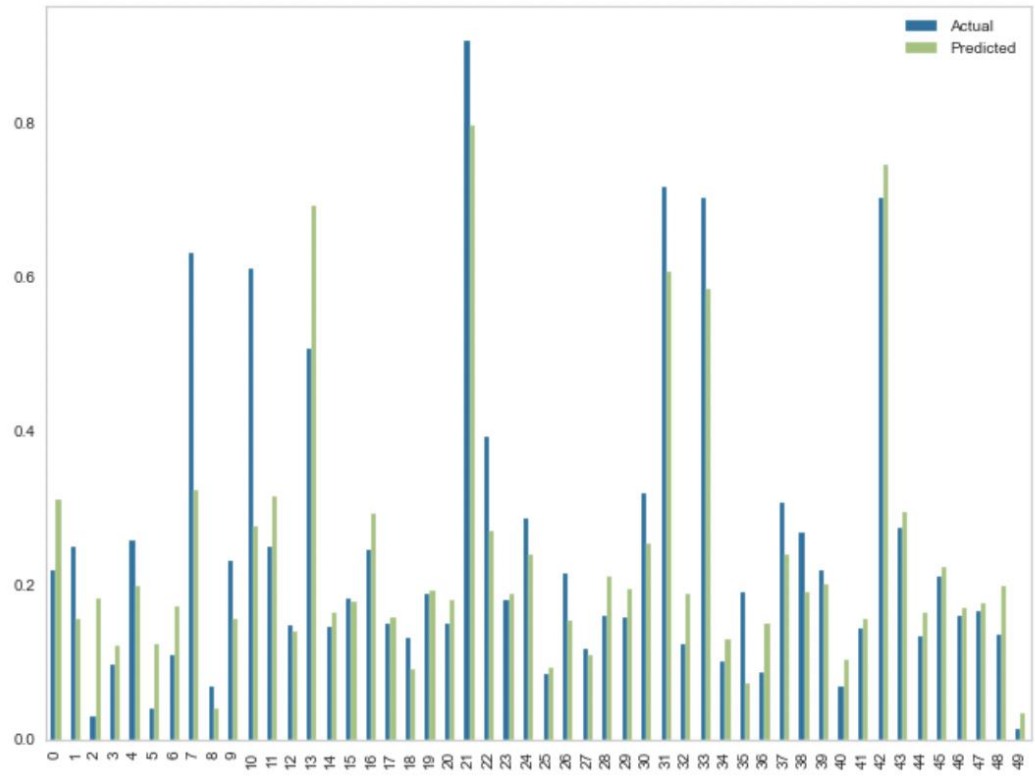


Figure 10: Bar Chart of predicted Vs Actual ROP

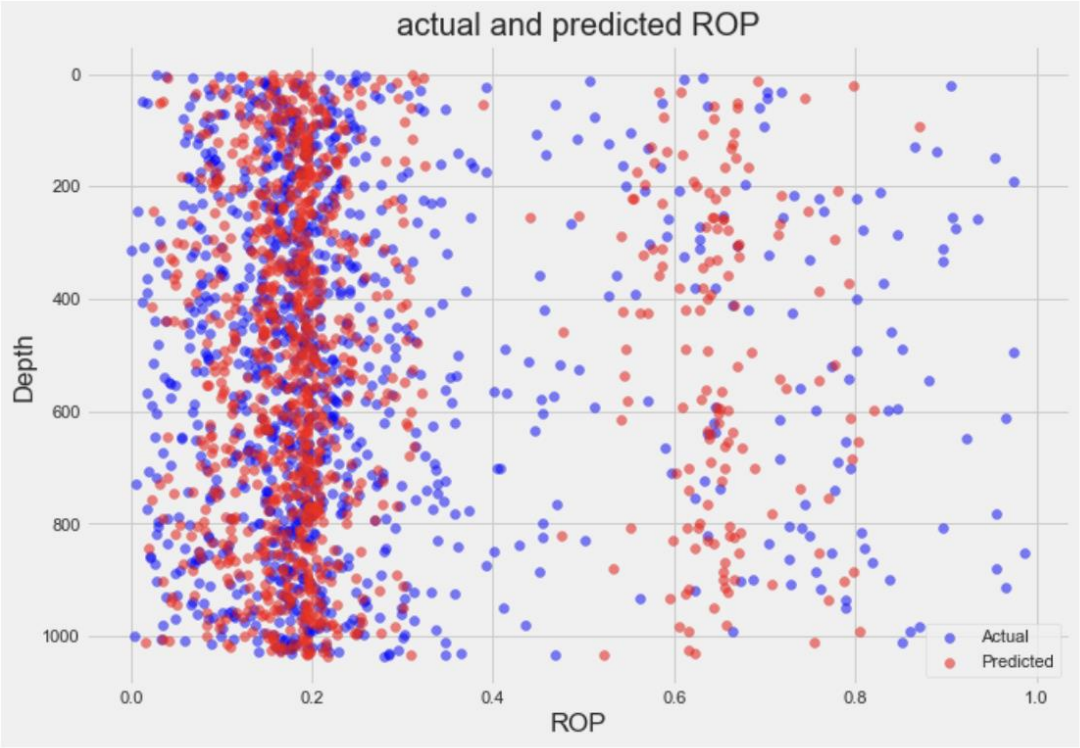


Figure 11: Scatter Chart of predicted Vs Actual ROP

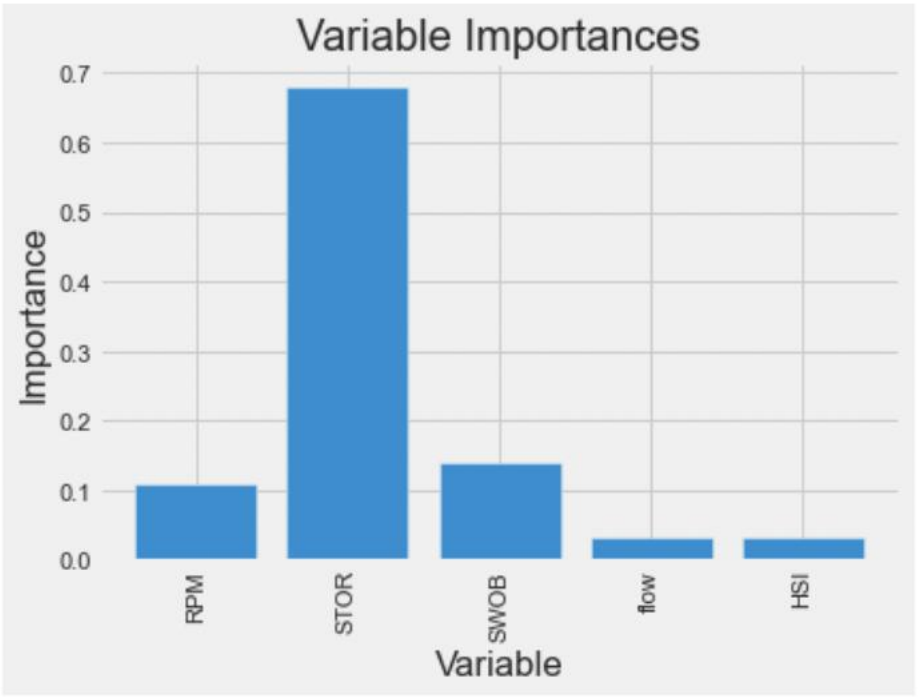
Random Forest

Through the process of feature optimization and cross validation:

Only Data:

Best parameters: {'max_depth': 5, 'min_samples_leaf': 4, 'n_estimators': 1000}

```
Variable: STOR                 Importance: 0.68
Variable: SWOB                 Importance: 0.14
Variable: RPM                  Importance: 0.11
Variable: flow                 Importance: 0.03
Variable: HSI                  Importance: 0.03
```



The feature importance shows STOR as the most important feature that affects the model prediction. The STOR is an uncontrollable feature as it is a response of the formation to the RPM. This feature importance plot is for the case when the model only used the columns but not the facies. The r2 is very low for this case and the mean square error is higher. This model did not pass the 0.5 benchmark for model accuracy.

Data + Facies:

Best parameters: {'max_depth': 6, 'min_samples_leaf': 5, 'n_estimators': 500}

Method	No Facies Identification	Data + Facies
R2	Score on train set: 0.280390 Score on test set: 0.224883	Score on train set: 0.821236 Score on test set: 0.786496
Root Mean Squared Error	0.17406024284826363	0.09229494948723312

Table 3: Model Evaluation results for Random Forest

The feature importance plot was gotten from the Random Forest Model.

```
Variable: Facies               Importance: 0.86
Variable: STOR                 Importance: 0.08
Variable: RPM                  Importance: 0.02
Variable: SWOB                 Importance: 0.02
Variable: flow                 Importance: 0.01
Variable: HSI                  Importance: 0.01
```

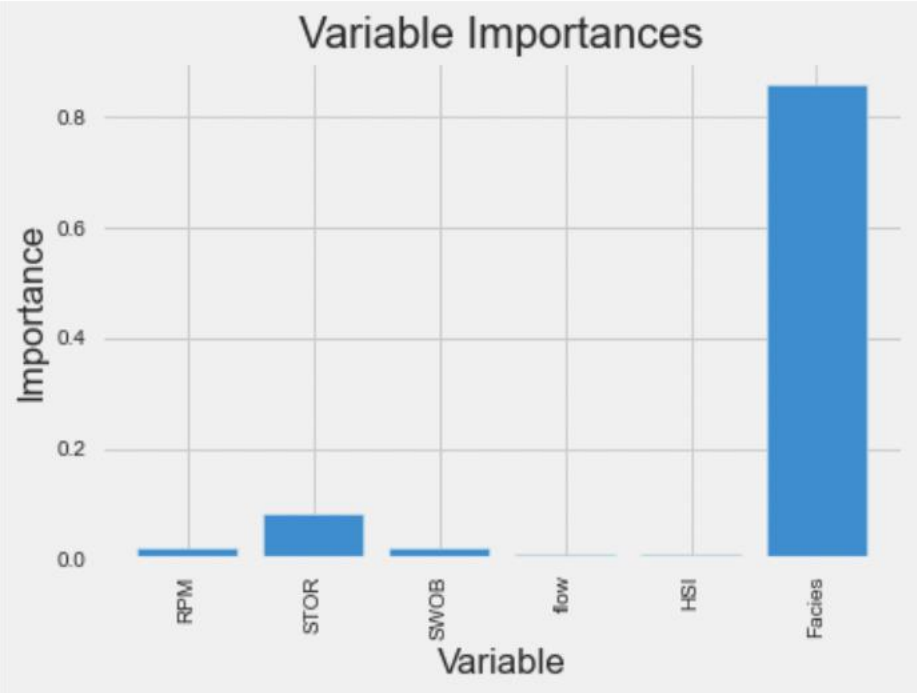
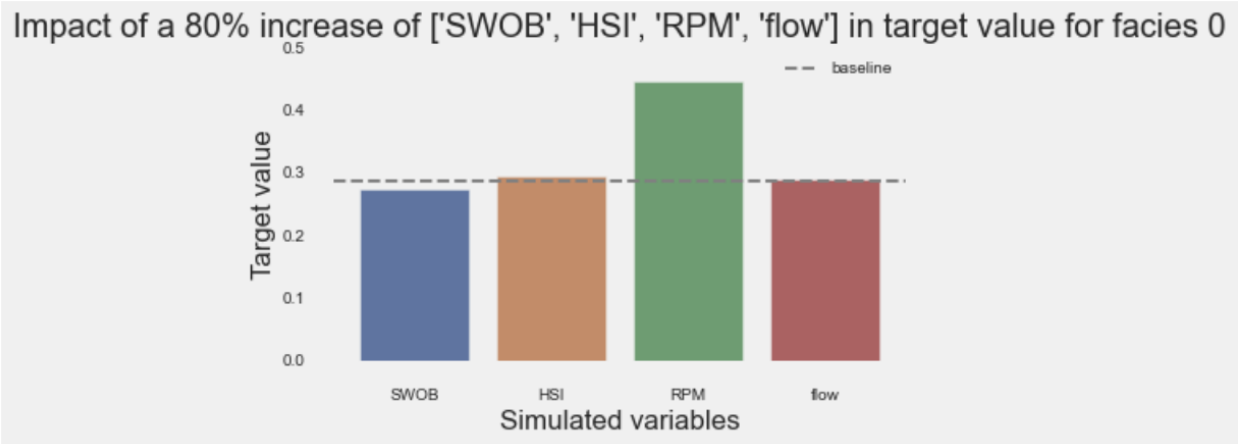


Figure 12: Feature Importance

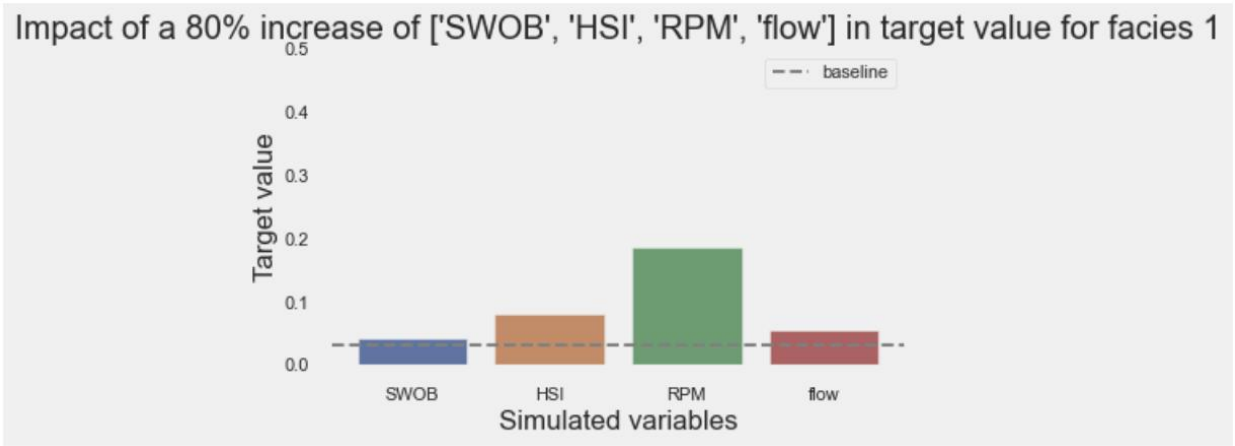
From the variable importance plot, the Facie type is the most important input into the model. The weight it carries is 86%. This means that in the practical sense, the type of formation being drilled is the most important feature to determine the rate of penetration. The formation to be drilled is a parameter that we cannot control. The next most important parameters are those

that we can control such as STOR, RPM and SWOB (torque, revolutions per minute and weight on bit). These are the parameters we can concentrate on optimizing.
The optimization is split for each of the three groups. This plot for each of them is shown below for an increase in 80% of each of the controllable parameters.
The plots show that increase in these parameters can cause an increase or a decrease in the ROP. The formation type (Facie) is an important feature to ensure the right optimization is done in the right Facie.



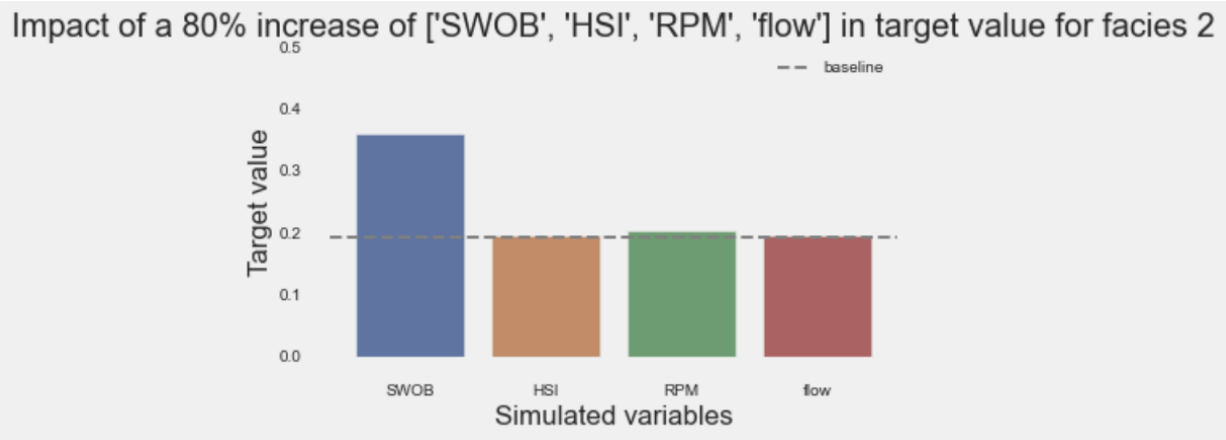
	test	simulated	baseline	perc_change_%
0	SWOB	0.272182	0.285303	-4.599254
1	HSI	0.293864	0.285303	3.000537
2	RPM	0.445698	0.285303	56.219010
3	flow	0.290035	0.285303	1.658419

Figure 13: Feature Optimization in Facies 0



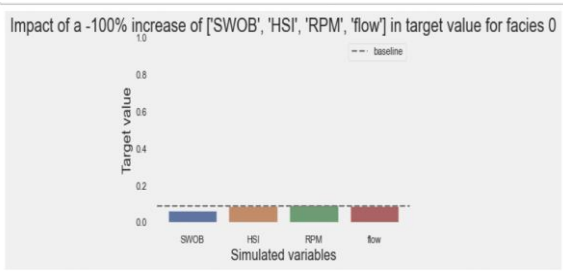
	test	simulated	baseline	perc_change_%
0	SWOB	0.042016	0.032325	29.978989
1	HSI	0.081722	0.032325	152.810056
2	RPM	0.185585	0.032325	474.115597
3	flow	0.056003	0.032325	73.245998

Figure 14: Feature Optimization in Facies 1



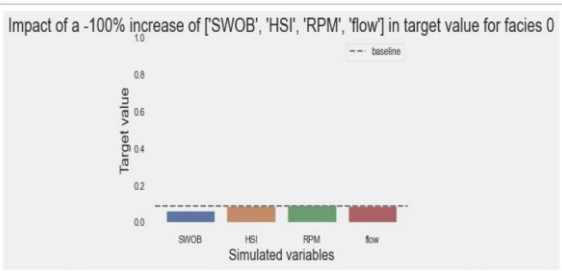
	test	simulated	baseline	perc_change_%
0	SWOB	0.358535	0.193042	85.729165
1	HSI	0.195390	0.193042	1.216515
2	RPM	0.203503	0.193042	5.419355
3	flow	0.194002	0.193042	0.497741

Figure 15: Feature Optimization in Facies 2



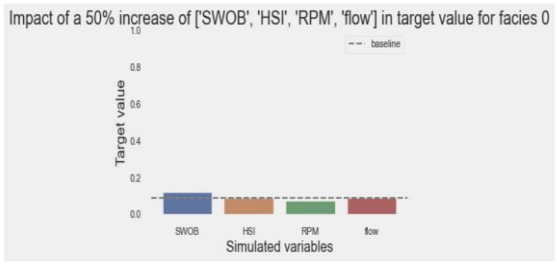
	test	-100	baseline	perc_change_%
0	SWOB	0.062718	0.088764	-29.342543
1	HSI	0.089123	0.088764	0.405438
2	RPM	0.092762	0.088764	4.504230
3	flow	0.089097	0.088764	0.375934

(a)



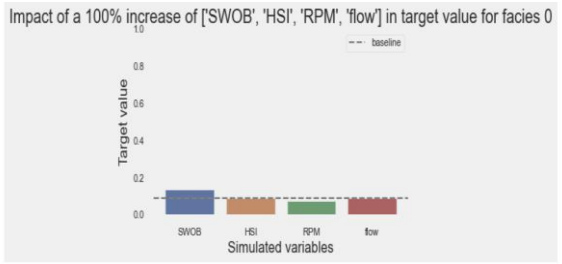
	test	-100	baseline	perc_change_%
0	SWOB	0.062718	0.088764	-29.342543
1	HSI	0.089123	0.088764	0.405438
2	RPM	0.092762	0.088764	4.504230
3	flow	0.089097	0.088764	0.375934

(b)



	test	50	baseline	perc_change_%
0	SWOB	0.117411	0.088764	32.273362
1	HSI	0.088469	0.088764	-0.332376
2	RPM	0.075953	0.088764	-14.432207
3	flow	0.088379	0.088764	-0.432743

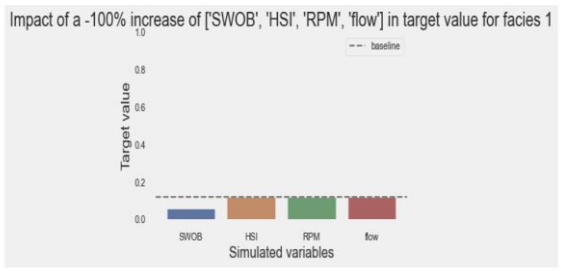
(c)



	test	100	baseline	perc_change_%
0	SWOB	0.134977	0.088764	52.063062
1	HSI	0.088367	0.088764	-0.446250
2	RPM	0.075953	0.088764	-14.432207
3	flow	0.088435	0.088764	-0.370159

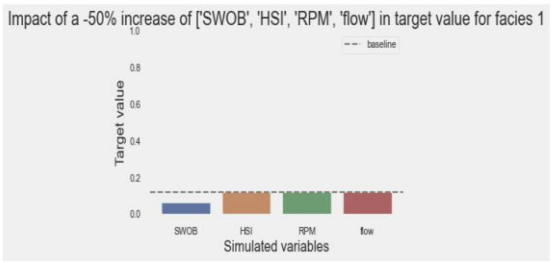
(d)

Table 4: Feature Optimization from range (-100, 100) in Facie 0



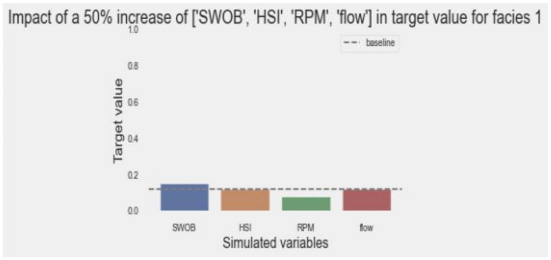
	test	-100	baseline	perc_change_%
0	SWOB	0.056037	0.118457	-52.694212
1	HSI	0.118191	0.118457	-0.224276
2	RPM	0.118011	0.118457	-0.376279
3	flow	0.118402	0.118457	-0.046279

(a)



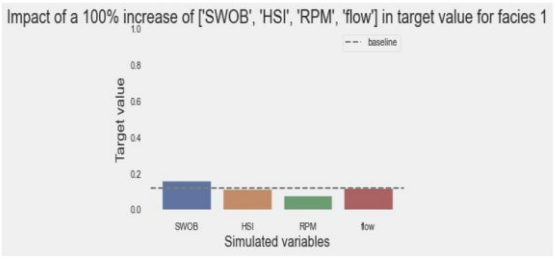
	test	-50	baseline	perc_change_%
0	SWOB	0.062246	0.118457	-47.452911
1	HSI	0.118191	0.118457	-0.224276
2	RPM	0.118011	0.118457	-0.376279
3	flow	0.118402	0.118457	-0.046279

(b)



	test	50	baseline	perc_change_%
0	SWOB	0.148599	0.118457	25.445807
1	HSI	0.118337	0.118457	-0.100620
2	RPM	0.076862	0.118457	-35.114112
3	flow	0.118457	0.118457	0.000000

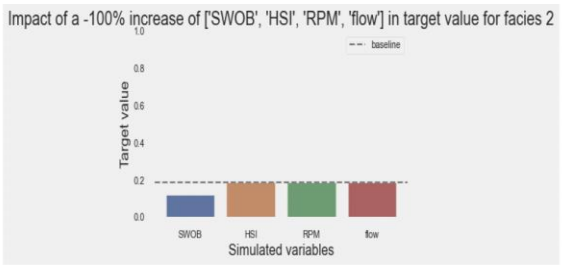
(c)



	test	100	baseline	perc_change_%
0	SWOB	0.161411	0.118457	36.261596
1	HSI	0.116894	0.118457	-1.319387
2	RPM	0.076862	0.118457	-35.114112
3	flow	0.118339	0.118457	-0.099293

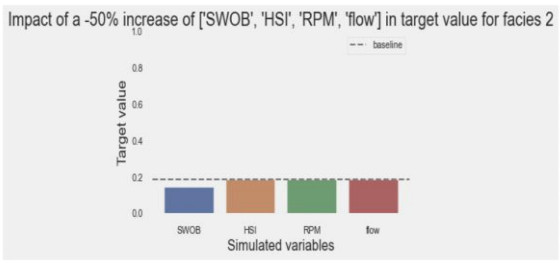
(d)

Table 5: Feature Optimization from range (-100, 100) in Facie 1



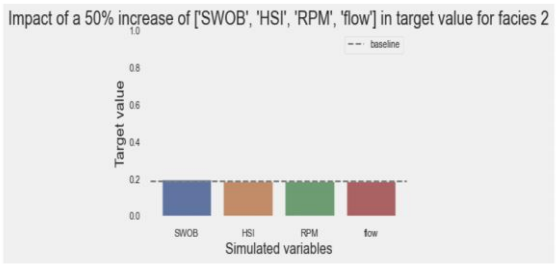
	test	-100	baseline	perc_change_%
0	SWOB	0.122181	0.188044	-35.02521
1	HSI	0.188044	0.188044	0.00000
2	RPM	0.188044	0.188044	0.00000
3	flow	0.188044	0.188044	0.00000

(a)



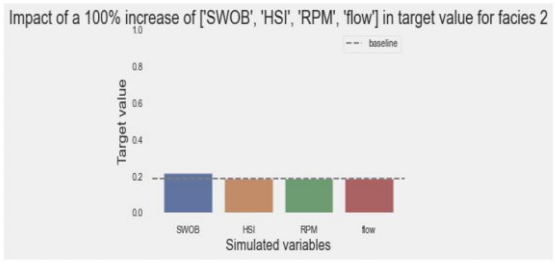
	test	-50	baseline	perc_change_%
0	SWOB	0.146990	0.188044	-21.832247
1	HSI	0.188044	0.188044	0.000000
2	RPM	0.188044	0.188044	0.000000
3	flow	0.188044	0.188044	0.000000

(b)



	test	50	baseline	perc_change_%
0	SWOB	0.199579	0.188044	6.134527
1	HSI	0.188044	0.188044	0.000000
2	RPM	0.188044	0.188044	0.000000
3	flow	0.188044	0.188044	0.000000

(c)



	test	100	baseline	perc_change_%
0	SWOB	0.219859	0.188044	16.918768
1	HSI	0.188044	0.188044	0.000000
2	RPM	0.188044	0.188044	0.000000
3	flow	0.188044	0.188044	0.000000

(d)

Table 6: Feature Optimization from range (-100, 100) in Facie 2

This project is very important in the practical oil field as well as looking at this from the business perspective. If the ROP can be maximized with this model, then the days for drilling an hydrocarbon well is decreased and thousands of dollars is saved on the drilling wells.

For every data point, the clustering system would identify the right cluster and the model for that cluster would specify the feature combination to give a higher ROP. Models have a r2 value of above 0.75 on the training dataset. This project can be modelled into drilling systems, and this can significantly improve the drilling efficiency.

Self-Assessment:

- Learn a program better in Python: I was able to learn the use of python for data analysis.
- Understand dive deeper into data preprocessing and preparation – this project helped me to be a be data engineer/data scientists. I used libraries such as numpy, pandas, math, stats etc.
- Understand Machine learning models better with these hands-on projects – since I worked with various machine learning models, I was able to learn about them more.
- Unite Machine Learning with practical experience in the engineering field – this project was a perfect fir to learn about how to apply data science to domain/practical knowledge.
- Add more knowledge to my petroleum engineering domain and solve a critical problem- the project solved an important problem in the oil field.

References

1. The data is from a Niger Delta Field in Nigeria.
2. Drilling in the digital Age: An approach to optimizing ROP Using Machine Learning; Peter Batruny, Hanif Yahya, Norazan Kadir, Amir Omar, Zahid Zakaria, Saravanan Batamale, and NoreffendyJayah, PETRONAS
3. Looking ahead of the bit using surface drilling and petrophysical data: machine-learning-based real-time geosteering in volve field. Ishank Gupta, Ngoc Tran, Deepak Devegowda, Vikram Jayaram, Chandra Rai, Carl Sondergeld, and Hamidreza Karami.
4. Application of Data Science and Machine Learning Algorithms for ROP Prediction: Turning Data into Knowledge. Christine Ikram Noshi and Jerome Jacob Schubert, Texas A&M University.
5. Use of machine learning and data analytics to increase drilling efficiency for nearby wells. Chiranth Hegde, K.E. Gray.
6. Machine learning methods applied to drilling rate of penetration prediction and optimization - A review. Luis Felipe F.M. Barbosa, Andreas Nascimento, Mauro Hugo Mathias, Joao Andrade de Carvalho Jr.