Oil Well Production Performance on Volve Dataset

Syed Sohaib Ali - 2209809, Shaik Misbah Bilal -2145022, Wajahat Ali Syed - 2049216, Mohammed Abdul Aziz -2215231
University of Houston
syedsoha@cougarnet.uh.edu, mshaik2@cougarnet.uh.edu, wsyed3@cougarnet.uh.edu, amoham70@cougarnet.uh.edu

Abstract

The report presents a comprehensive analysis of the Volve field production performance based on historical data. We determine the various factors that affect the performance of the wells and using machine learning models like K means Clustering, Decision Tree, and Linear Regression we differentiate between the wells that have significant oil production and the wells that are about to become dry. We also suggest various ways to enhance the performance of wells.

1 Introduction

1.1 Literature Review

Volve is an oil and gas field located in the North Sea that has been producing oil since 2008. The field has a complex reservoir structure with several layers and compartments, making it challenging to optimize production. The use of big data analytics can help in improving production efficiency, optimizing performance, and reducing costs in the Volve field.

The production data, which includes Oil production, gas production, water production, well temperature is one of the main sources of big data in the Volve field. Predictive analytics and machine learning algorithms can be used to analyze this data and find patterns and trends that can be leveraged to maximize output.

1.2 Business/ Analytics Problem and Question Framing

The core questions that this study aims to address include:

- What were the key factors driving production performance in the Volve field?
- Are there any insights from the Volve field that can be applied to other fields for production optimization and improved economic performance?

1.3 Objective

The primary objective of our project is to analyze the production performance of the Volve field and identify key factors, trends, and challenges that can be used to optimize production and enhance economic performance in similar fields. Additionally, suggest various methods to improve and enhance the performance of the well.

1.4 Impact and Values

The findings of our project offer operators, investors, and regulators in the oil and gas sector useful insights.

- By understanding the factors that contribute to production performance and management stakeholders can make informed decisions to optimize existing fields and develop new projects in a sustainable and profitable manner.
- The insights from the Volve field contribute to best practices for reservoir management and production optimization, ultimately enhancing the overall performance of the industry.
- Within the same reservoir if some wells are producing good and others are not, then there may be a case that reservoir is not uniform in terms of permeability which is always the case. In that event, the wells which do not have a good production require well stimulation techniques. Therefore, oil well

classification is important to know which wells may require well stimulation.

2 Data

2.1 Dataset background and quality

To provide a solution to our problem statement, a dataset is required which not only has oil production values with respect to the wells, but also with sufficient well details to help classify them. While there are several oil well production datasets in the public, very few provide details regarding the wells. Finally, the dataset is needed to contain production data, which is the main goal of this project. Therefore, the team decided to use the "Volve Well Production" dataset for this project.

The volve field dataset was collected and provided by the company Equinor. Equinor is major company in oil and gas and our chosen dataset comes from the Volve field on Norwegian continental shelf. Equinor specifically created this dataset for the purposes of research, study, and development, and has provided the license as well as the conditions of use of the dataset on their website. Altogether, this improves the credibility and quality of the provided dataset. The dataset comes with the following details:

Table 1 Dataset Information

Title	Total Rows	Total Columns
Daily Production Data	15634	24
Monthly Production Data	527	10

The daily production data holds several rows which can be used to determine the production performance of wells and has 15634 rows which is more beneficial for the training of the model. As such, this section was used for modeling. It has the following columns:

```
Index(['DATEPRD', 'WELL_BORE_CODE', 'NPD_WELL_BORE_CODE', 'NPD_WELL_BORE_NAME',
       'NPD_FIELD_CODE', 'NPD_FIELD_NAME', 'NPD_FACILITY_CODE',
       'NPD_FACILITY_NAME', 'ON_STREAM_HRS', 'AVG_DOWNHOLE_PRESSURE',
       'AVG_DOWNHOLE_TEMPERATURE', 'AVG_DP_TUBING', 'AVG_ANNULUS_PRESS',
       'AVG_CHOKE_SIZE_P', 'AVG_CHOKE_UOM', 'AVG_WHP_P', 'AVG_WHT_P',
       'DP_CHOKE_SIZE', 'BORE_OIL_VOL', 'BORE_GAS_VOL', 'BORE_WAT_VOL',
       'BORE_WI_VOL', 'FLOW_KIND', 'WELL_TYPE'],
      dtype='object')
```

Variables such as BORE_OIL_VOL, BORE_GAS_VOL, and BORE_WAT_VOL, are important indicators of well performance, as they provide information about the volume of different fluids produced by the well.

2.2 Data processing, EDA

Several steps were taken to ensure the dataset is clean and fit to use for the models. The dataset has the following features:

```
0   DATEPRD                  15634 non-null  datetime64[ns]
1   WELL_BORE_CODE           15634 non-null  object
2   NPD_WELL_BORE_CODE       15634 non-null  int64
3   NPD_WELL_BORE_NAME       15634 non-null  object
4   NPD_FIELD_CODE           15634 non-null  int64
5   NPD_FIELD_NAME           15634 non-null  object
6   NPD_FACILITY_CODE        15634 non-null  int64
7   NPD_FACILITY_NAME        15634 non-null  object
8   ON_STREAM_HRS            15349 non-null  float64
9   AVG_DOWNHOLE_PRESSURE    8980 non-null   float64
10  AVG_DOWNHOLE_TEMPERATURE 8980 non-null   float64
11  AVG_DP_TUBING            8980 non-null   float64
12  AVG_ANNULUS_PRESS        7890 non-null   float64
13  AVG_CHOKE_SIZE_P         8919 non-null   float64
14  AVG_CHOKE_UOM            9161 non-null   object
15  AVG_WHP_P                9155 non-null   float64
16  AVG_WHT_P                9146 non-null   float64
17  DP_CHOKE_SIZE            15340 non-null  float64
18  BORE_OIL_VOL             9161 non-null   float64
19  BORE_GAS_VOL             9161 non-null   float64
20  BORE_WAT_VOL             9161 non-null   float64
21  BORE_WI_VOL              5706 non-null   float64
22  FLOW_KIND                15634 non-null  object
23  WELL_TYPE                15634 non-null  object
```

After analyzing the entire dataset, we found that it contains missing values in the dataset, and replaced it.

checking for missing values in the dataset

```
In [12]: missing_values=df.isnull()

In [13]: missing_values_count = missing_values.sum()

In [14]: print(missing_values_count)
```

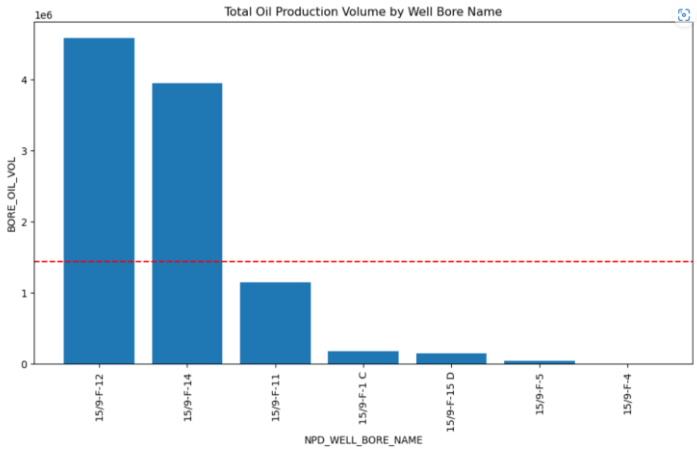
```
DATEPRD                        0
WELL_BORE_CODE                 0
NPD_WELL_BORE_CODE             0
NPD_WELL_BORE_NAME             0
NPD_FIELD_CODE                 0
NPD_FIELD_NAME                 0
NPD_FACILITY_CODE              0
NPD_FACILITY_NAME              0
ON_STREAM_HRS                285
AVG_DOWNHOLE_PRESSURE       6654
AVG_DOWNHOLE_TEMPERATURE    6654
AVG_DP_TUBING               6654
AVG_ANNULUS_PRESS           7744
AVG_CHOKE_SIZE_P            6715
AVG_CHOKE_UOM               6473
AVG_WHP_P                   6479
AVG_WHT_P                   6488
DP_CHOKE_SIZE                294
BORE_OIL_VOL                6473
BORE_GAS_VOL                6473
BORE_WAT_VOL                6473
BORE_WI_VOL                 9928
FLOW_KIND                      0
WELL_TYPE                      0
dtype: int64
```

We removed these missing values using interpolation methods,

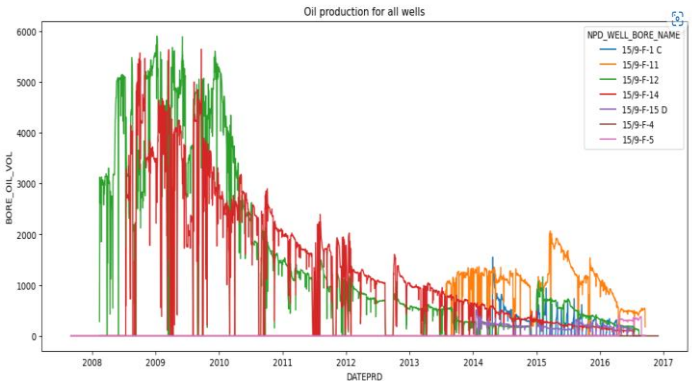
```
#replacing the missing values
df.interpolate(method='pad', inplace=True)
```

```
DATEPRD                        0
WELL_BORE_CODE                 0
NPD_WELL_BORE_CODE             0
NPD_WELL_BORE_NAME             0
NPD_FIELD_CODE                 0
NPD_FIELD_NAME                 0
NPD_FACILITY_CODE              0
NPD_FACILITY_NAME              0
ON_STREAM_HRS                  0
AVG_DOWNHOLE_PRESSURE          0
AVG_DOWNHOLE_TEMPERATURE       0
AVG_DP_TUBING                  0
AVG_ANNULUS_PRESS              0
AVG_CHOKE_SIZE_P               0
AVG_CHOKE_UOM                  0
AVG_WHP_P                      0
AVG_WHT_P                      0
DP_CHOKE_SIZE                  0
BORE_OIL_VOL                   0
BORE_GAS_VOL                   0
BORE_WAT_VOL                   0
BORE_WI_VOL                    0
FLOW_KIND                      0
WELL_TYPE                      0
```

Total Volume of Oil produced by the wells:



We observed that only 3 wells are significantly producing the oil while the others are at the verge of becoming dry.



From the Fig, we can interpret the volume of oil production of the 3 most significant wells has declined over the period of time.

3 Methodology

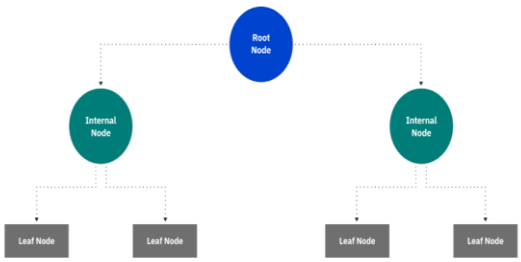
Methods and its Description:

Classification:

A classification machine learning model is a type of supervised learning algorithm that aims to categorize input data into one of several predefined classes or categories. Its fundamental objective is to accurately predict the class label of new, unseen instances based on the patterns learned from a training dataset.

i) Decision Tree

A Decision Tree is a machine learning algorithm used for both classification and regression tasks. It has a hierarchical, tree-like structure that divides the dataset recursively according to the most useful attribute at each node. The decision points, potential outcomes, and final class labels or predictions are represented by the tree's nodes, branches, and leaves, which are made up of nodes, branches, and leaves.



To select the best attribute at each node decision tree, make use of entropy and information gain.

ii) Random Forest

Random Forest is an ensemble learning method used for both classification and regression tasks. It combines multiple decision trees to improve the overall predictive performance, increase model robustness, and reduce the risk of overfitting. The main idea behind the Random Forest algorithm is to create a "forest" of decision trees, each trained on a random subset of the dataset, and aggregate their predictions to produce a final output.

Regression:

A regression machine learning model is a type of supervised learning algorithm that aims to predict a continuous target variable based on one or more input features. Finding the correlation between the input data and the target variable is the main objective of regression models, which enables the model to produce precise predictions for novel, unforeseen situations.

i) Linear Regression

Linear Regression is a machine learning algorithm used for predicting a continuous target variable based on one or more input features. It estimates the best-fitting straight line through the data points under the assumption that there is a linear connection between the input features and the target variable. Finding the ideal weights (coefficients) for the input features that minimize the sum of squared errors between the anticipated and actual target values is the aim of linear regression.

We determine the predictor and the target variables and build the model using it and then evaluate the model for its performance.

Clustering

Clustering is a type of unsupervised learning technique used in machine learning to group similar data points together based on their features or characteristics. Finding significant patterns, relationships, or structures within the data without any prior knowledge of the target labels or classes is the main objective of clustering. Many applications, including customer segmentation, anomaly detection, picture segmentation, and document grouping, can use clustering techniques.

i) K-means Clustering

K-Means is a partition-based clustering algorithm in machine learning that aims to partition the dataset into 'K' distinct, non-overlapping clusters based on the similarity between data points. The algorithm identifies 'K' cluster centers (centroids) in the feature space such that the sum of squared distances between data points and their corresponding centroids is minimized.

Here we determine the features required for model building and then number of clusters using elbow method and then train the model on the training data that is,
'WELL_BORE_CODE', 'BORE_OIL_VOL', so that our model gives the best outcomes.

Model Building and Evaluation :

Decision Tree:

Initially we defined a threshold for well that are producing oil significantly which acts as our target variable and build a decision tree model using the function Decision Tree Classifier() and fit it with 70% of the training data,

```
clf = DecisionTreeClassifier()

clf = clf.fit(X_train, y_train)
```

Once the model is trained, we make predictions on the test data to make predictions and determine the accuracy our model,

```
y_pred_test = clf.predict(X_test)

accuracy_test = metrics.accuracy_score(y_test, y_pred_test)
print("Accuracy on the test dataset: ", accuracy_test)
```

Similarly, we build a decision tree model to determine the amount of gas and water produced by the wells.

Random Forest

To compare the model performances, we also build a random forest machine learning model using the function RandomForestClassifier( ) ,

Now we fit our random forest model with the training data,

```
clf = RandomForestClassifier(random_state=1)

clf = clf.fit(X_train, y_train)
```

After the model is trained completely, we perform model evaluations using the 30% of the test data, And determine its accuracy and the confusion matrix,

```
y_pred = clf.predict(X_test)

print("Accuracy:", metrics.accuracy_score(y_test, y_pred))

conf_matrix = confusion_matrix(y_test, y_pred)
print("Confusion Matrix:")
print(conf_matrix)
```

Once the model is built, trained, and evaluated we determine the volume of gas and water produced by the well using the same technique.

Linear Regression.

We start build the model by splitting the training and testing data wherein ‘BORE_OIL_VOL’ is our target volume,

Moving further we create a Linear Regression model using the function Linear Regression() which is then fitted on the training data.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)


model = LinearRegression()


model.fit(X_train, y_train)
```

Once the model is trained, we make predictions using the test data and calculate the performance metrics of our model,

```
y_pred = model.predict(X_test)

r2 = r2_score(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
mae = mean_absolute_error(y_test, y_pred)
```

Lastly, we build a model to determine the production of gas and water respectively.

K-means Clustering

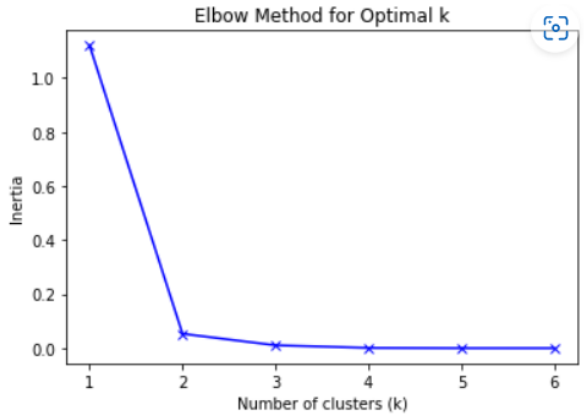
Initially, we extract the features that contribute to the production data,

```
production_data = df[['WELL_BORE_CODE', 'BORE_OIL_VOL']]

grouped_production = production_data.groupby('WELL_BORE_CODE').sum().reset_index()
```

Now, using the elbow method we determine the optimal number of clusters,

```
inertia = []
K = range(1, min(10, len(grouped_production)))
for k in K:
    kmeans = KMeans(n_clusters=k)
    kmeans.fit(grouped_production[['BORE_OIL_VOL']])
    inertia.append(kmeans.inertia_)
```



From the above fig, we can conclude that our model will give best results when k = 2

Once, the number of clusters are determined we apply the k-means clustering by utilizing the production data,

```
kmeans = KMeans(n_clusters=k)
grouped_production['Cluster'] = kmeans.fit
```

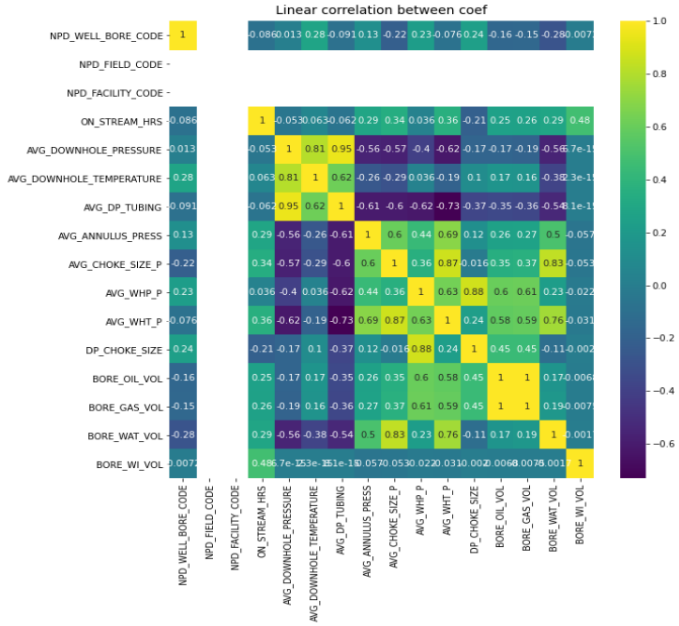
Once the model is trained, we determine the silhouette score and Inertia to determine the model performance.

```
silhouette = silhouette_score(grouped_production[['BORE_OIL_VOL']
print(f"Silhouette score: {silhouette}")
inertia = kmeans.inertia_
print(f"Inertia: {inertia}")
```

Then we repeat the same model for both gas and water production.

4 Results and Discussion

To identify the key factors that affect the production performance of the wells we tried to find the linear correlation between coefficient columns. So, we plot the heat map to determine the correlation among the features.



The production performance is judged based on BORE_OIL_VOL,BORE_WAT_VOL,BORE_GAS_VOL, which are the volumes of oil, gas, water produced by the wellbore respectively.

From the above Linear correlation between coefficients, we have derived the following conclusions:

The key factors that affect the production performance (BORE_OIL_VOL) of the wells are:

'AVG_DOWNHOLE_TEMPERATURE',
'AVG_ANNULUS_PRESS',
'AVG_CHOKE_SIZE_P', 'AVG_WHP_P'
,'DP_CHOKE_SIZE' , 'ON_STREAM_HRS'.

The key factors that affect the production performance (BORE_WAT_VOL) of the wells are:
'AVG_ANNULUS_PRESS',
'AVG_CHOKE_SIZE_P', 'AVG_WHP_P'
'ON_STREAM_HRS'

The key factors that affect the production performance (BORE_GAS_VOL) of the wells are:

'AVG_DOWNHOLE_TEMPERATURE',
'AVG_ANNULUS_PRESS',
'AVG_CHOKE_SIZE_P', 'AVG_WHP_P'
,'DP_CHOKE_SIZE' , 'ON_STREAM_HRS'.

Classification

i) Decision Tree

The accuracy and the confusion matrix on the test data

```
Accuracy on the test dataset:  0.9918993
[[4184   20]
 [  18  469]]
```

The classification of wells is given as

```
Wells with significant production:
['15/9-F-1 C', '15/9-F-11', '15/9-F-12', '15/9-F-14'

Wells at risk of becoming dry:
['15/9-F-15 D', '15/9-F-4', '15/9-F-5']
```

ii) Random Forest

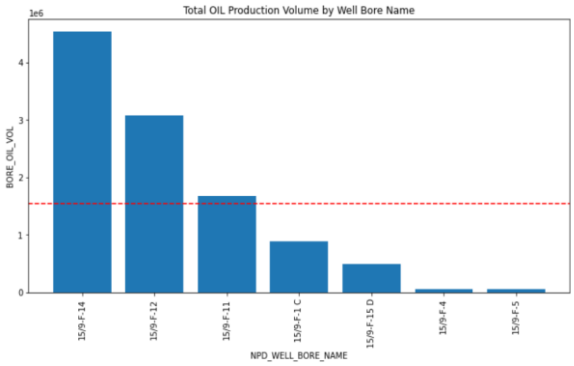
For random forest the accuracy and the confusion matrix on the test data is given as,

```
Accuracy: 0.996376039224046
Confusion Matrix:
[[4197    7]
 [  10  477]]
```

It could be observed that the results are like that of decision tree.

Regression

i) Linear Regression



From the above fig we can analyze that linear regression model correctly classify the significant wells and the tells us about the wells that are about to become extinct. Linear regression was run to predict the well production performance and see which wells are about to get dry and which have good production performance, the above figure is an example of oil production performance.

The below table summarizes the performance metrics of the linear regression model of oil, gas and water.

	R^2	*MSE*	*AMSE*	*MAE*
Oil	0.54	627739	729	453
Gas	0.55	124501	111580	64422
Water	0.73	651129	806	533

i. Clustering

K-means Clustering

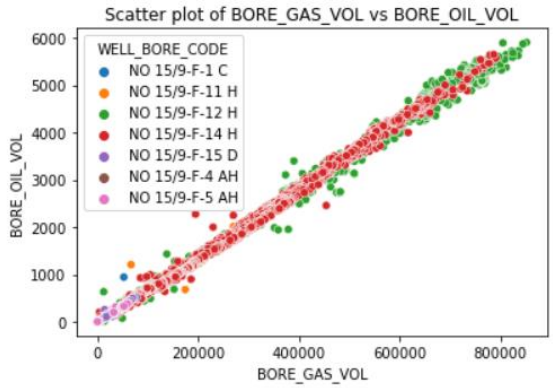
When k = 2, the model gives its best performance and correctly predicts the significant oil wells,

```
Cluster 0:
Significant wells:
   WELL_BORE_CODE  BORE_OIL_VOL
2  NO 15/9-F-12 H      1.000000
3  NO 15/9-F-14 H      0.860823
Cluster 1:
Significant wells:
   WELL_BORE_CODE  BORE_OIL_VOL
1  NO 15/9-F-11 H      0.250643
Other wells:
   WELL_BORE_CODE  BORE_OIL_VOL
0   NO 15/9-F-1 C      0.038804
4  NO 15/9-F-15 D      0.032430
5  NO 15/9-F-4 AH      0.000000
6  NO 15/9-F-5 AH      0.008988
```

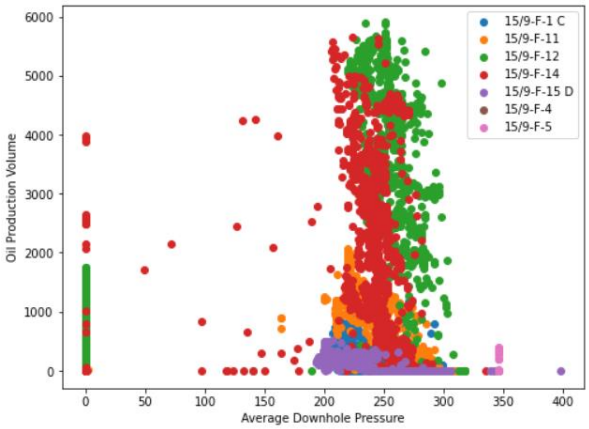
The below table summarizes the performance metrics of the k means clustering model of oil, gas and water.

	Silhouette score	*Inertia*
Oil	0.85	0.056
Gas	0.93	0.017
Water	0.936	0.017

4.1 Suggestions to enhance the wells production performance:

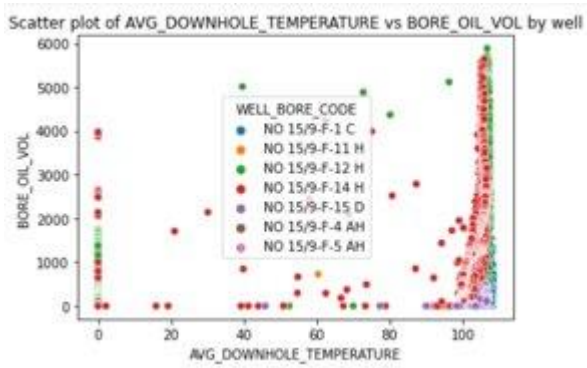


We can draw the following insights that can help improve the production performance of the wells in the Volve Field, from the above Fig Bore Gas Volume ('BORE_GAS_VOL') is the most significant factor affecting oil production performance. This insight suggests that optimizing gas production could have a substantial impact on oil production as well. The petroleum management team may consider evaluating gas lift systems, gas compression facilities, or other gas management strategies to optimize gas production and enhance oil recovery.

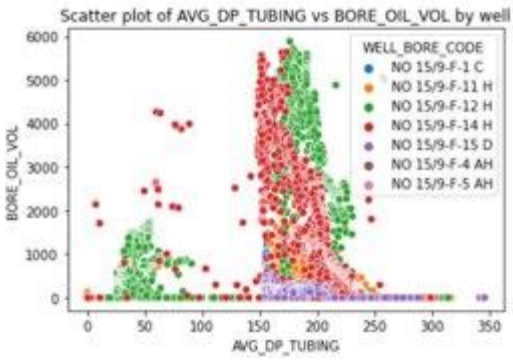


As we can see above plot when the down hole pressure is in between the range 200 to 300 wells are performing good, now to study any one well we can plot seperate graph and understand the range for that to improve the future production

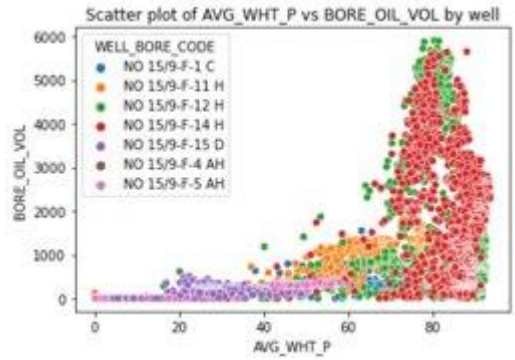
this was only one feature with which we were drawing insights below are plots for all the features from which we insights could be drawn



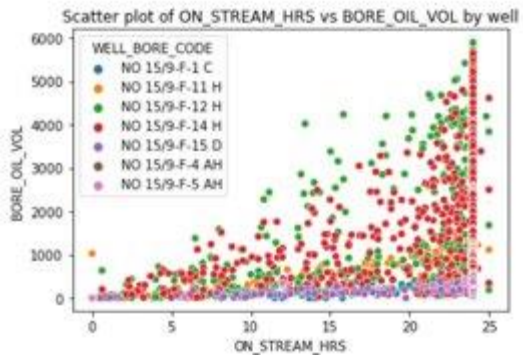
The above visualization provides the range of AVG_DOWNHOLE_TEMPERATURE which provides best production results



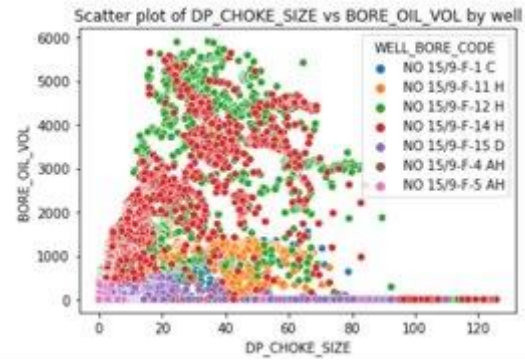
The above visualization provides the range of AVG_DP_TUBING which provides best production results



The above visualization provides the range of AVG_WHT_P which provides best production results.



The above visualization provides the range of ON_STREAM_HRS which provides best production results



The above visualization provides the range of DP_CHOKE_SIZE which provides best production results.

5 Conclusion and Recommendations

Based on the analysis of the Volve field production performance, we can draw several conclusions. Gas production optimization could boost the wells oil production. The production of oil is greatly hindered due to various parameters including the average temperature and pressure of the well. Also, production could be improved after understanding the ranges at which these parameters provide best production performance. These ranges or affects of parameters could be noted/understood using the above visualization and the parameters could be tuned accordingly to improve production results.

Of all the machine learning model applied li near regression and k-means clustering performs t he best with the r-square value of 0.55 and Silhou ette score of 0.93.

The future aspects of this project are large as more datasets with oil well quality data with its production can be requested from companies to make an even better and robust model.

6 Acknowledgements

We would like to express our gratitude to Equinor for sharing the Volve field data and related materials. We also appreciate the industry experts who provided valuable insights and guidance in various research papers.

7 Nomenclature

MSE - Mean Square Error.
MAE - Mean Absolute Error.
AMSE – Absolute Mean Square Error.

8 References

[1] H. Powers, W. Trainor-Guitton and G. Hoversten, "Classification of total oil production of wells in SEAM Life of Field from stochastic AVA inversion attributes via machine learning", SEG Technical Program Expanded Abstracts 2018, 2018.

[2] A. Sircar, K. Yadav, K. Rayavarapu, N. Bist, and H. Oza, "Application of machine learning and artificial intelligence in oil and gas industry", Petroleum Research, vol. 6, no. 4, pp. 379-391, 2021.

[3] T. Bikmukhametov and J. Jäschke, "Oil Production Monitoring using Gradient Boosting Machine Learning Algorithm", IFAC-PapersOnLine, vol. 52, no. 1, pp. 514-519, 2019.

[4] O. Innocent, "Application of Machine Learning in Predicting Crude Oil Production Volume", Day 2 Tue, August 03, 2021, 2021.

[5] W. Liu, W. Liu, and J. Gu, "Petroleum Production Forecasting Based on Machine Learning", Proceedings of the 2019 3rd International Conference on Advances in Image Processing, 2019.

[6] "Volve field data village download - data 2008-2016 - equinor.com", Equinor.com, 2022. [Online]. Available: https://www.equinor.com/en/what-we-

do/digitalisation-in-our-dna/volve-field-data-village-download.html. [Accessed: 19- Apr- 2022]