# Prediction of Downhole Pressure Using Machine Learning

### Project 3: Report

# Problem Statement

A machine learning model that can accurately predict the downhole pressure of an oil well from various well head measurements. Analysis will provide a deeper understanding of the relationships between six variables. (choke size, well head pressure, temperature, oil volume, gas volume, and water volume)

# Data-sets

Original file was in the format of a CSV. Data was sourced from Equinor’s website. Contained variables for five wells from 2008-2014.

# <https://www.equinor.com/en/how-and-why/digitalisation-in-our-dna.html>

# Data Wrangling

The data file was loaded to pandas data frame and analyzed. After discussions with the SME, some variables were dropped based on those correlations and the usability of the measure to predict downhole pressure. The resulting cleaned data had the choke size, well head pressure, temperature, oil volume, gas volume, and water volume measures for five wells. Null values and shutdown period rows were dropped. Categorical columns were also not taken into consideration.

# Exploratory Data Analysis

Matplotlib and Seaborn libraries were used to find out

1. Relationship between features and target variable.
2. Pairwise plot and heat map was used to find out correlation between the features
3. Histogram plots were used to plot frequency plots for each feature and target variable

Pandas data frame functions were used to explore data dimensionality, feature’s names, and feature types.

# Modeling

Models Used for our Analysis:

1. Multi-linear Regression
   * To explore the linear correlation between input variables with down hole pressure.
2. Lasso Regression with Grid Search CV Hyperparameter Tuning

* To explore the most important feature variables and improve the prediction accuracy.

1. Random Forest Regression with Randomized Search CV Hyperparameter Tuning
   * To explore the complex non-linear relationship between the input parameters and downhole pressure.
2. Neural Network
   * The use of deep learning to run modeling on the data to find the best model for predictions on test data.

All models’ MSE (mean squared error) and R-squared measures were evaluated on train and test data. Residual plots were generated to visualize these results. Tuning was done on each model to generate higher accuracy.

### Train-Test Split

The dataset included daily production datapoints from all the 5 Wells from 208-2016. The train and test data were split randomly across datapoints from all wells in 3:1 ratio. The assumption was that, downhole pressure was independent of the well location and the well performance. It was only dependent on the input variables used in our analysis.

### Multi-Linear Regression

#### Model Selection

Multi Linear Regression was performed to explore the linear relationship between the input variables and the target variables. The features with the highest coefficient values obtained from the model were, Average Wellhead Pressure, Average Wellhead Temperature and Bore Water Volume.

#### Model Scores

The following were the scores obtained from the base model:

Mean Squared Error (MSE)on Train Data: 0.4104764730468125

R-squared (R2) on Train Data: 0.5895235269531875

-----------------------------------

Mean Squared Error (MSE)on Test Data: 0.41490932151903154

R-squared (R2) on Test Data: 0.5781442928698933

The scores obtained were low compared to the other models.

* Lasso Regression with Grid Search CV Hyperparameter Tuning

#### Model Selection

Lasso Regression was performed to explore the linear relationship between the input variables and the target variable. This was used over the linear regression model to enhance the prediction accuracy and also to understand the variable selection process. The features with the highest coefficient value obtained from the model were, Average Wellhead Pressure, Average Wellhead Temperature and Bore Water Volume.

#### Model Scores

The following were the scores obtained from the base model and the model tuned with the best parameters obtained from Grid Search CV. The model yielded better performance, higher R2 score and low MSE, with the tuned parameters over the base model and also over the linear regression model.

Model Performance – Base Model

Mean Squared Error (MSE): 0.5299432795735386

R-squared (R2): 0.46118444356741584

Model Performance – Hyperparameters Tuned with Grid Search CV

Mean Squared Error (MSE): 0.4142533390099713

R-squared (R2): 0.5788112577966187

The following parameters were tuned and the optimized value of those parameters are as follows:

'alpha': 0.001, 'max\_iter': 5000, 'tol': 1e-05

* Random Forest Regression – With Randomized Search CV Hyperparameter Tuning

#### Model Selection

To explore the complex non-linear relationship between the input parameters and downhole pressure. With the best parameters obtained from Randomized Search Hyperparameter tuning on the model, it yielded an 87% R2 score with a low MSE of 0.12. The most important features as per the model were: Average Wellhead Pressure, Average Wellhead Temperature and Bore Water Volume.

#### Model Scores

The following were the scores obtained from the base model and the model tuned with the best parameters obtained from Randomized Search CV. The model yielded better performance, higher R2 score and low MSE, with the tuned parameters over the base model.

Model Performance – Base Model

Mean Squared Error (MSE): 0.17241049618066653

R-squared (R2): 0.8247030182000588

--------------------------------------------

Model Performance - Hyperparameters Tuned with Randomized Search CV

Mean Squared Error (MSE): 0.12480383126538175

R-squared (R2): 0.8731067108874555

The following were the best parameters that yielded the optimized result for the model:

'n\_estimators': 522,

'min\_samples\_split': 2,

'min\_samples\_leaf': 2,

'max\_features': 'sqrt',

'max\_depth': 110,

'bootstrap': False

R2: The score was 87%. This means that 87 percent of the variance of the target variable predicted through the model, could be explained by the variance of the input variables.

MSE: The mean squared error for the model was 0.12. It gives us an assessment on the quality of the predictor.

The high R2 and the low MSE helps us to evaluate the performance of the model. However, this was the second best, given the low score from the linear regression model and the better score from the Neural Network model.

### Neural Network

#### Model Selection

Deep learning with the use of neural networks. Keras sequential model used as data was relatively simple, numerical only, and from one input source. Tuning with done the hyperas. Created dropout layers to increase accuracy.

#### Model Scores

Model Performance – Base Model

Mean Squared Error (MSE): 0.14

R-squared (R2): 0.86

--------------------------------------------

Model Performance -with tuning

Model Score: 0.886

# Conclusion

1. The neural network model was the best model with the highest accuracy to predict downhole pressure given the following six input variables: Average Wellhead Pressure, Average Wellhead Temperature, Average Choke Size, Bore Water Volume, Bore Gas Volume, Bore Oil Volume.
2. The most important variables obtained from the models were Average Wellhead Pressure, Average Wellhead Temperature and Bore Water Volume.

# Further Exploration

1. Perform time series analysis on the dataset to explore if there is any effect of time on the downhole pressure prediction.
2. Explore the model performance considering other ratios including Gas Oil ratio and Water Oil Ratio.
3. Explore the model performance on other field data.

# References

Keras Metric Tuning

<https://github.com/keras-team/keras/issues/7947>

Journal pf Petroleum Exploration and Production Technology

<https://link.springer.com/article/10.1007/s13202-019-0728-4>