*This is the main submission document**. Save and rename this document filename with your registered full name as Prefix before submission.***

Full Name	Koh Lianghao Eugene
Email Address	ekoh027@e.ntu.edu.sg

** : Delete and replace as appropriate.*

Declaration of Academic Integrity

By submitting this assignment for assessment, I declare that this submission is my own work, unless otherwise quoted, cited, referenced or credited. I have read and understood the Instructions to CBA.PDF provided and the Academic Integrity Policy.

I am aware that failure to act in accordance with the University's Academic Integrity Policy may lead to the imposition of penalties which may include the requirement to revise and resubmit an assignment, receiving a lower grade, or receiving an F grade for the assignment; suspension from the University or termination of my candidature.

I consent to the University copying and distributing any or all of my work in any form and using third parties to verify whether my work contains plagiarised material, and for quality assurance purposes.

Please insert an "X" within the square brackets below to indicate your selection.

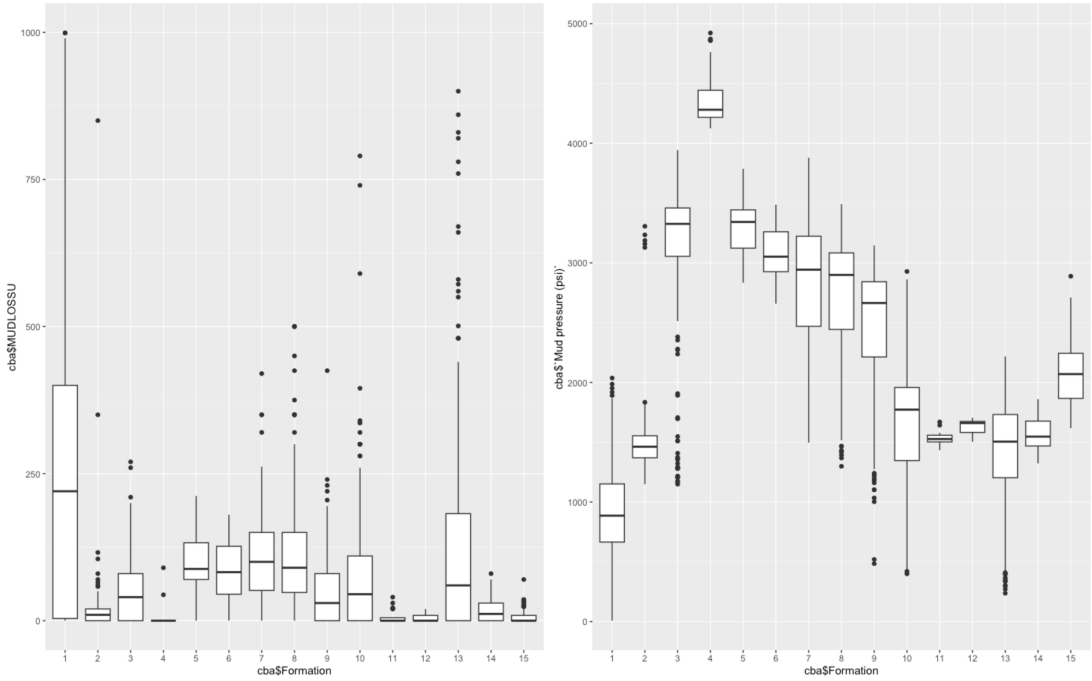
[X] I have read and accept the above.

Table of Contents

For each question, please start your answer in a new page.

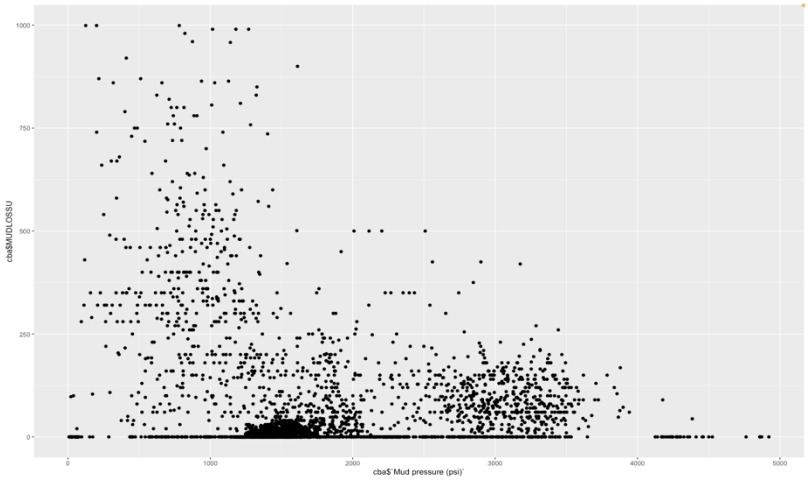
Answer to Q1:

Notable Findings 1:

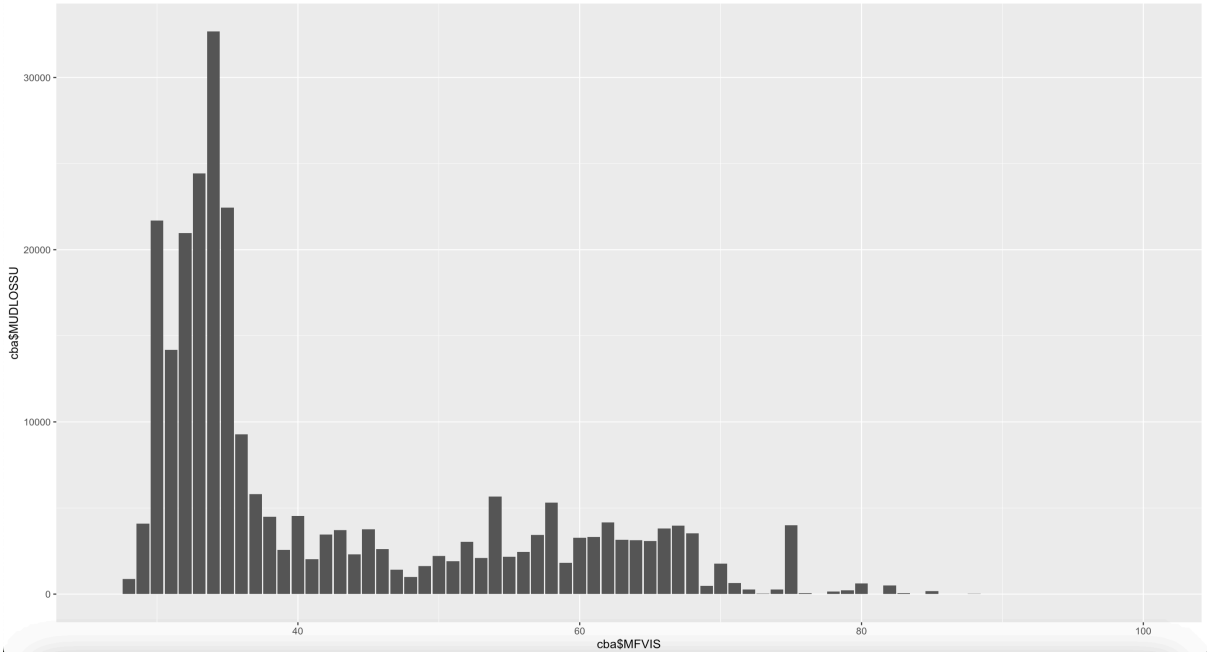


In this figure, one insight that I found was that Formation 1 requires the least mud pressure but has the highest mean of MUDLOSSU, which I assume is the loss of drilling fluid. This means that when drilling in Formation 1, the operators could use lower mud pressure to maintain the wellbore's stability when carrying out drilling operations effectively. Additionally, with a higher mean value of MUDLOSSU, it suggests that Formation 1 is prone to losing the drilling fluid. This means that drilling in Formation 1 is associated with the need for specific strategies to manage and reduce the loss of drilling fluid. On the other hand, Formation 4 seems to require a higher mud pressure to maintain the wellbore's stability. With the requirement of a higher mud pressure, the tendency of losing drilling fluid due to fracturing would be significantly lower.

Therefore, we can see that a higher mud pressure requirement reduces the chances of losing the drilling fluid (MUDLOSSU). This could be further supported by the bottom graph where the higher mud pressure results in a lower likelihood of losing drilling fluid (MUDLOSSU).



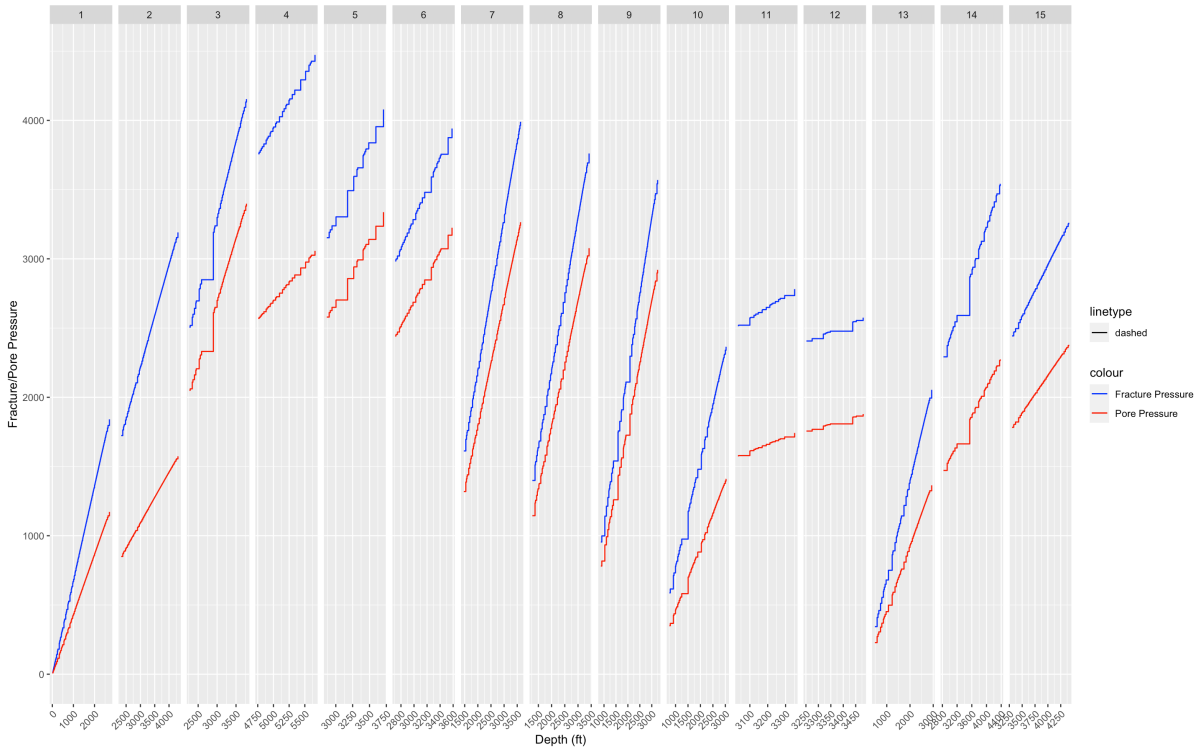
Notable Findings 2:



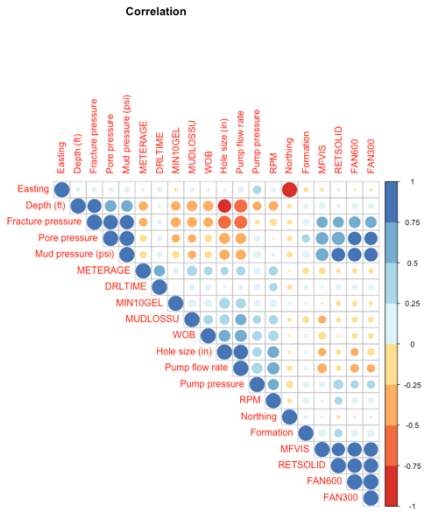
Another insight is that if the drilling mud has a lower viscosity, the pressure in the borehole might not be able to support the weight of the formation's weight. Thus, it could lead to wellbore instability and the formation could also collapse, causing immense amounts of loss of circulation. The figure shows that the amount of drilling fluid loss is significantly higher when the viscosity is lower than 40cp. However, one anomaly found was that even though the viscosity was low, the amount of loss circulation was found to be one of the lowest. There might be multiple reasons that resulted in the low loss of circulation. One speculation could be that there might have been unusual drilling operations that led to the unusual loss of drilling fluid.

In conclusion, mud engineers are required to ensure that the right amount of mud viscosity is placed during the drilling operations to minimise the amount of loss circulation.

Notable Finding 3:



From this step plot, depth is positively correlated to fracture pressure and pore pressure. The correlation between depth and fracture pressure is 0.84 meanwhile the correlation between depth and pore pressure is 0.58. This can be supported by the correlation heat map where the dark blue colour signifies the positive correlation. This would mean that the deeper the drill, it would increase the probability of fracturing the formation, causing a circulation loss. Additionally, different formations have different depths before fracture and pore pressure come into play. It is evident that Formation 1 is the weakest formation where there is a tendency to cause fractures even when the drill depth is less than 1000 feet. On the contrary, Formation 4 is the strongest formation where the likelihood of causing fracturing can only be seen at the lowest point of 4,750 feet.



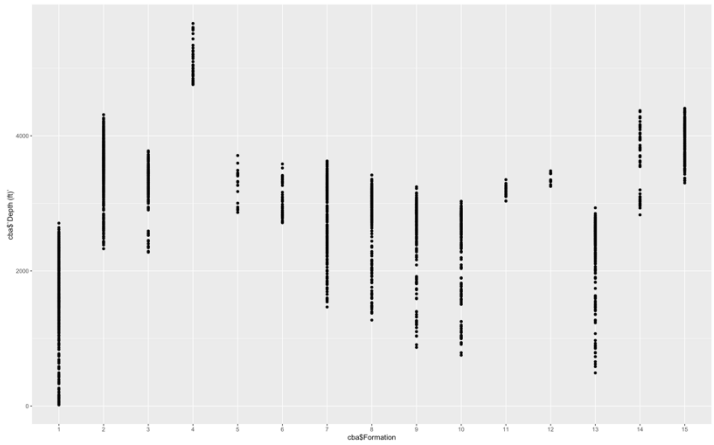
It is also noticeable that fracture pressure is generally higher than pore pressure. The discrepancy highlights the importance of careful monitoring and managing drilling operations to prevent any adverse effects.

Answer to Q2:

Part (1)

Formation was the categorical variable used. There is a total of 15 distinct formation types. In each formation type, formation affects various aspects of the production process.

From this simple example, different formation types possess different requirements. In this case, different formation types require different levels of depth. It is evident that Formation 1 requires a relatively lower depth as compared to the rest. It also allows for a variation of depth length as the number of points of depth is more scattered.



This figure compares the depth versus the formation type.

Part (2)

Missing values are present in this data set. The missing values were found in columns: FAN 600, FAN 300, MIN10GEL and MUDLOSSU. To handle missing values, I have taken the mean of the data column, assuming that the missing values are at random and taking the mean as a reasonable estimate to replace the missing values. It makes the dataset complete with no "NA" values. Removal of the entire row with missing values would not be ideal because retrieval of data is expensive for the company and removal of the data rows could be a waste.

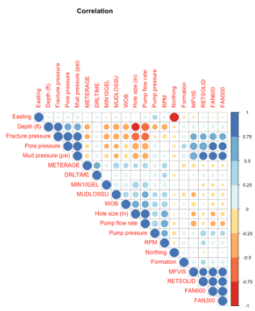
Answer to Q3:

Model	Complexity	Trainset RMSE	Testset RMSE
Linear Reg	Dependent variable count: 1 Independent variable count: 11	124.8089	124.3837
CART	Number of terminal nodes: 7	104.6169	113.231

Linear Regression

There are a total of 19 variables accompanied by 1 dependent variable. However, after much analysis has been done, there were several variables that multicollinearity was present in it. Therefore, those multicollinear variables such as Northing, Easting, Depth, Pore pressure, Fracture pressure, Mud pressure and both FAN600 and FAN300 were removed in the process of running the linear model.

Multicollinearity was determined by the vif() function. To validate the values, a correlation graph plot has also been used to cross-check the results.



Therefore, this leaves 11 independent variables for me to run the linear regression model. To compute RMSE for the linear regression model, the dataset was split into 70-30 train-test splits.

```
> train <- sample.split(Y = cba$MUDLOSSU, SplitRatio = 0.7)
> trainset <- subset(cba, train == T)
> testset <- subset(cba, train == F)
```

Afterwards, the train set values were predicted and compared against the actual values and RMSE was computed.

The following output shows the RMSE of the trainset and testset of the linear regression model, which are 124.8089 and 124.3837 respectively.

```
> m1 <- lm(MUDLOSSU ~ Formation
+          + `Hole size (in)` + METERAGE + DRLTIME + WOB + `Pump flow rate` + `Pump pressure` + MFVIS + RETSOLID + MIN10GEL +
+            RPM, data = trainset)
> trainset$prediction <- predict(m1, newdata = trainset)
> mean_squared_error <- mean((trainset$MUDLOSSU - trainset$prediction)^2)
> sqrt(mean_squared_error)
[1] 124.8089

> testset$prediction <- predict(m1, newdata = testset)
> # Calculate Mean Squared Error (MSE) - test set
> mean_squared_error <- mean((testset$MUDLOSSU - testset$prediction)^2)
> sqrt(mean_squared_error)
[1] 124.3837
```

Cart Model

The computation of CART RMSE is similar to the linear model where both train and test set predicted values are compared against the actual value.

The following output shows the RMSE of the trainset and testset of the CART model, which is 104.6169 and 113.231 respectively.

```
> set.seed(2)
> train <- sample.split(Y = cba$MUDLOSSU, SplitRatio = 0.7)
> trainset <- subset(cba, train == T)
> testset <- subset(cba, train == F)
> cart_model <- rpart(MUDLOSSU ~ ., data = trainset)
> prediction <- predict(cart_model, newdata = trainset)
> m3 <- rpart(MUDLOSSU ~ ., data = trainset, method = 'anova',
+             control = rpart.control(minsplit = 2, cp = 0))
> mean_squared_error <- mean((trainset$MUDLOSSU-prediction)^2)
> rmse <- sqrt(mean_squared_error)
> rmse
[1] 104.6169

> predictions <- predict(cart_model, newdata = testset)
> mean_squared_error <- mean((testset$MUDLOSSU - predictions)^2)
> rmse <- sqrt(mean_squared_error)
> rmse
[1] 113.231
```

To create a CART model, the dataset was split into 70-30 train-test splits. Subsequently, the trainset is used to grow the decision tree with the specific parameter of “rpart control(minsplit = 2, cp = 0)”. Oftentimes, presenting the fully grown tree is unmeaningful as there will be overfitting and it might be relatively tough to visualise the tree diagram. Overfitting may introduce unnecessary noise, which might be less likely to generalise and handle new and unseen data. Hence, pruning the tree is important and in question 4, I will discuss further how I prune the tree and present the optimal tree.

Answer to Q4:

Regression Model:

```
Call:
lm(formula = MUDLOSSU ~ Formation + `Pore pressure` + `Hole size
    METERAGE + DRLTIME + WOB + `Pump flow rate` + `Pump pressure
    MFVIS + RETSOLID + MIN10GEL + RPM, data = cba)

Residuals:
    Min      1Q  Median      3Q     Max
-734.85  -50.06   -5.23   32.56  871.23

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.190e+02  3.147e+01   6.958 4.32e-12 ***
Formation2       -1.390e+02  1.696e+01  -8.194 3.90e-16 ***
Formation3       -7.436e+00  2.405e+01  -0.309 0.757240
Formation4       -6.526e+01  3.109e+01  -2.099 0.035887 *
Formation5        1.679e+01  3.617e+01   0.464 0.642519
Formation6        1.583e+01  2.815e+01   0.562 0.573879
Formation7        7.630e+00  2.241e+01   0.340 0.733563
Formation8        7.730e+00  2.140e+01   0.361 0.717936
Formation9       -5.163e+01  1.973e+01  -2.616 0.008942 **
Formation10      -9.859e+01  1.258e+01  -7.835 6.72e-15 ***
Formation11      -9.525e+01  3.033e+01  -3.141 0.001705 **
Formation12      -9.039e+01  4.123e+01  -2.193 0.028421 *
Formation13      -9.594e+01  1.074e+01  -8.931  < 2e-16 ***
Formation14      -1.147e+02  2.464e+01  -4.656 3.39e-06 ***
Formation15      -1.065e+02  1.943e+01  -5.480 4.65e-08 ***
`Pore pressure`  -3.325e-02  8.942e-03  -3.718 0.000205 ***
`Hole size (in)` -8.487e+00  1.560e+00  -5.440 5.81e-08 ***
METERAGE          7.263e-01  8.551e-02   8.493  < 2e-16 ***
DRLTIME           5.625e-02  4.215e-01   0.133 0.893860
WOB               9.327e-01  3.452e-01   2.702 0.006939 **
`Pump flow rate`  1.742e-01  3.517e-02   4.954 7.75e-07 ***
`Pump pressure`   3.436e-03  4.601e-03   0.747 0.455268
MFVIS            -1.749e+00  4.208e-01  -4.157 3.32e-05 ***
RETSOLID         -6.240e-01  3.990e-01  -1.564 0.117969
MIN10GEL          2.146e+00  9.176e-01   2.338 0.019445 *
RPM               1.734e-01  8.363e-02   2.074 0.038191 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 124.6 on 2642 degrees of freedom
Multiple R-squared:  0.4066,    Adjusted R-squared:  0.401
F-statistic: 72.41 on 25 and 2642 DF,  p-value: < 2.2e-16
```

Having 19 independent variables and MUDLOSSU as a dependent variable, we can see that Formation, Hole size, Meterage, WOB, Pump Flow Rate, RETSOLID, MFVIS, FAN300, FAN600 and MIN10GEL are statistically significant. When determining the significance of the independent variables, the P-value is taken into consideration. The above variables identified are statistically significant because values are lower than the threshold of 0.05. In statistical analysis, it means that the observed data provides strong evidence that contradicts the null hypothesis (Mcleod S, 2023).

When analysing the impact of the independent variable (MUDLOSSU) based on the dependent variables, the coefficients are taken into consideration. Taking Hole Size as an example, a decrease in hole size by 6.304 inches will cause a unit increase in the loss of circulation. From a business perspective, to reduce the likelihood of circulation losses, the drilling operator should take into careful consideration the hole size.

In conclusion, the company should take into serious consideration the various factors that are statistically significant because changes in those factors could cause potential fractures, and time and resources would be wasted, which can be costly to the company.

Moving on to the r-squared error, it is computed to be 0.4192 while the adjusted r-squared is 0.4117. This means that approximately 41.92% of the dependent variables can be explained by the independent variables in the regression model. However, the r-squared error is not the best approach to compute the goodness of fit because the more variable input there is, the higher the r-squared error – which can result in inaccuracy. Therefore, the use of adjusted r-squared error is used to penalise the use of unnecessary independent variables. Hence, the adjusted r-squared error suggests that 41.17% of the dependent variables can be explained by the independent variable.

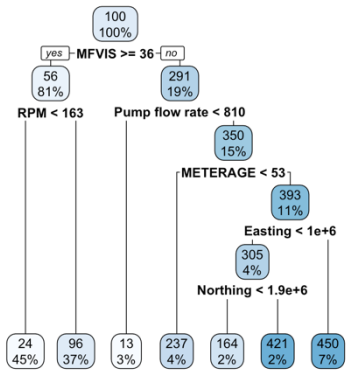
Cart Model:

```
> print(m.opt)
n= 1885

node), split, n, deviance, yval
      * denotes terminal node

 1) root 1885 50556080.0 100.30490
   2) MFVIS>=35.5 1533 11704970.0  56.45058
     4) RPM< 162.5 844  2211946.0  24.31012 *
     5) RPM>=162.5 689  7553165.0  95.82148 *
   3) MFVIS< 35.5 352 23062800.0 291.29550
     6) Pump flow rate< 810 61   114437.8  12.65574 *
     7) Pump flow rate>=810 291 17219540.0 349.70450
      14) METERAGE< 52.5 81  3494944.0 236.67900 *
      15) METERAGE>=52.5 210 12290720.0 393.30000
        30) Easting< 1025302 82  4355272.0 305.00000
          60) Northing< 1918365 37  1014094.0 164.21620 *
          61) Northing>=1918365 45  2004864.0 420.75560 *
        31) Easting>=1025302 128  6886523.0 449.86720 *
```

```
opt$variable.importance
          MFVIS      Pore pressure  Fracture pressure      FAN600 Mud pressure (psi)  Pump flow rate
     15788309.6          11878815.8         11583547.5          10181665.5          8527921.2         7476334.5
Hole size (in)          Formation            FAN300          Depth (ft)           Easting          Northing
      6562649.3          5950070.9          5696350.3           4862135.6         2385238.8         2385238.8
            RPM      Pump pressure           METERAGE                 WOB           DRLTIME
      2010665.1          1580171.4          1433874.8           855901.9          584171.2
```



In the CART analysis, MFVIS has the highest importance, followed by Pore Pressure. To achieve a pruned tree, the cost complexity (cp) of each node of the tree is calculated. The nodes are pruned based on the smallest increase in cost complexity from the 10-fold cross-validation error and standard error. Afterwards, the geometric mean of the two identified cp values in the optimal region would be computed to determine the cp value. Using the cp value, the optimal tree will then be produced.

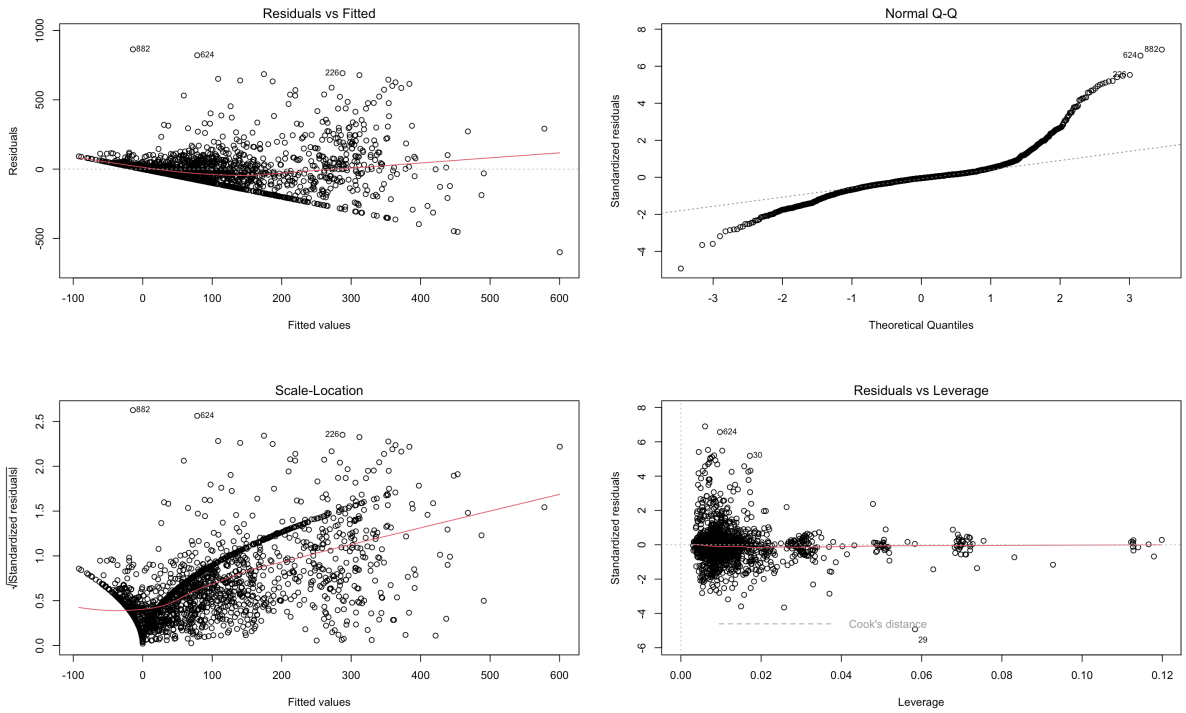
In this analysis, there are a total of 6 splits and 7 terminal nodes.

Using the first split based on the condition of MFVIS >= 35.5 as an example to do an analysis, it leads to two child nodes: left child – RPM < 162.5 and right child – RPM >= 162.5. There are a total of 844 observations and a predicted value of 24.3102 for the left child and 689 observations and a predicted value of 95.82148 for the right child. These predicted values provide insights into the behaviour of the target variable based on the influence of both MFVIS and RPM.

CART analysis would be useful for the business in guiding the drilling operators to optimise drilling parameters. For instance, when MFVIS is greater than 35.5, the CART model suggests that the RPM should be lower than 162.5 because the predicted value of loss of circulation is lower.

Conclusion

Both linear regression and CART models are good models because their RMSE is similar, ranging from 100-125. However, the trainset RMSE for the linear regression model is slightly higher than the test set while the trainset RMSE for the CART model is lower than the test set. This shows that there might be overfitting for the linear regression model and might not generalise well to new, unseen data. On the contrary, the CART model suggests that there might be underfitting and might not be able to capture the underlying patterns in the training data. In both cases, underfitting is usually better than overfitting and therefore the CART model is a better model in predictive accuracy.



Additionally, observing the key assumptions of a linear model, the model can be seen as a not optimal model because it does not follow the ideal path. For example, the plot to test the linear association between X and Y (Residuals v Fitted chart) shows some form of linearity but is not completely linear. The plot to test the error with a normal distribution with a mean of 0 (Normal Q-Q chart) shows that there is some deviation from the expected linearity, indicating potential issues with the assumption of normally distributed errors. Next, the plot to test for the constant standard deviation of errors (Scale-Location chart) shows that the points are not distributed evenly at each vertical slice. Therefore, it violates the standard deviation assumption of error. Lastly, the chart about influential outliers (Residuals vs Leverage chart) shows that the influential outliers are skewing the model's result and might potentially affect the model's fit. One recommendation to ensure a better fit is to perform a log transformation to meet the assumption of constant variance in the context of linear modelling (University of Virginia, n.d.).

Therefore, the CART model is still a better model as compared to the linear regression model.

Answer to Q5:

Sabah's CART model shows how outcome variable values can be predicted based on other values. There is a total of 37 terminal nodes and viscosity seems to be of the highest importance because it can be seen in the first split. From the model, the split is based on the "N" value, mean and standard deviation.

One example of using Sabah's CART model to conduct an analysis of loss circulation is when the drilling operator meets with this specific condition of viscosity lesser than 35.5, the optimal flow rate should be less than 810 because the predicted loss of circulation is much lower than when the flow rate is above 810.

Analysing the CART model, certain splits contain high standard deviations, which translates to high variances. For example, the split in drilling time produced a high degree of standard deviation, which shows that the decision tree was not pruned to its optimal cp. A high standard deviation or variance can lead to inconsistent or extreme predictions when presented with slightly different data points. Additionally, the CART model shows a relatively complex structure which signifies that the model is not pruned to produce the optimal tree. Therefore, it is important to prune the tree to its optimal tree to prevent the occurrence of overfitting.

Answer to Q6:

The 4 performance measures that Sabah used are Variance Accounted For (VAF), Root Mean Square Error (RMSE), Performance Index (PI) and R^2 to compare the prediction capability of the developed models. The results of VAF, PI and R^2 show that the model is overfitted because the train test result is higher than the test set result whereas the results for RMSE show that the model is underfitted. VAF is useful for understanding how well the model captures the variance in the target variable. R^2 is useful when trying to measure the goodness of fit. RMSE calculates the magnitude of error between predicted and actual errors. A lower error indicates better predictive accuracy. Lastly, PI leverages VAF and RMSE to calculate the performance index, omitting RMSE. Therefore, PI is determined by VAF as well as R^2.

Analysing the calculated values for each model, the GA-MLP model has the best predictive accuracy because the train test margin is the lowest among the rest of the models. However, the performance index for the GA-MLP model is the lowest among the rest, therefore the model is not the optimal model that should be utilised. On the other hand, the decision tree model has the highest difference between the train and test set but the model seems to be the best-performing model because it has the highest-performance index.

References

1. University of Virginia. (n.d.). Interpreting Log Transformations in a Linear Model. Retrieved from: https://library.virginia.edu/data/articles/interpreting-log-transformations-in-a-linear-model#:~:text=If%20we're%20performing%20a,non%2Dlinear%20relationship%20more%20linear.
2. Mcleod S. (2023). P-Value And Statistical Significance: What It Is & Why It Matters. Retrieved from: https://www.simplypsychology.org/p-value.html