UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

USING DATA ANALYTICS ON DIMENSIONLESS NUMBERS TO PREDICT THE ULTIMATE RECOVERY FACTORS FOR DIFFERENT DRIVE MECHANISMS OF GULF OF MEXICO OIL FIELDS

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

GOWTHAM TALLURU
Norman, Oklahoma
2017

USING DATA ANALYTICS ON DIMENSIONLESS NUMBERS TO PREDICT THE
ULTIMATE RECOVERY FACTORS FOR DIFFERENT DRIVE MECHANISMS OF
GULF OF MEXICO OIL FIELDS

A THESIS APPROVED FOR THE
MEWBOURNE SCHOOL OF PETROLEUM AND GEOLOGICAL ENGINEERING

BY

Dr. Xingru Wu, Chair

Dr. Deepak Devegowda

Dr. Mashhad Fahes

Acknowledgements

I would like to thank Dr. Xingru Wu for continuously guiding me in the right direction of research. His hard working nature motivated me a lot during my research. Dr. Xingru Wu's down to earth nature inspite of being so knowledgeable always fascinates me and enhances my respect towards him. I would also thank Dr. Deepak Devegowda and Dr. Mashad for their valuable suggestions to my research. I would also like to thank Dr. Rouzbeh G. Moghanloo for being my advisor during first year of my Masters. I enjoyed working with Dr. Sidharth Misra as a teaching assistant for well testing classes. Many thanks to Priyank Srivastava for gathering GoM dataset which made my work easy. I would also like to ackowlege the support of MPGE staff during two years of my Masters degree.

I express my deepest gratitude to my parents for their love and efforts to make me a successful individual. Last but not the least, I would like to thank my friends Adithya, Eiman, Japneet, Pratiksha, Pritesh, Saket, Saurabh, Shipra for making my two years of Masters memorable.

Table of Contents

List of Tables

List of Figures

Abstract

The ultimate recovery factor is strongly affected by petrophysical parameters, engineering data, structures, and drive mechanisms. The knowledge of the recovery factor is needed for multiple decision makings and it should be known in the whole development process. This study is to estimate the recovery factor from a different perspective with traditional methods.

This study capitalizes on existing database from the same basin, but explores parametric relationships between different reservoirs using data analytics. Given that there are hundreds of attributes to characterize a reservoir, and some of the records in a database may not be accurate or contradictory to each other, the propose is to use dimensionless quantity first to categorize them based on similarity theorems. Using independent dimensionless variables not only reduces the number of variables for data analytics, but also they have particular physical meanings. This research presents a comparative study of different data mining techniques and statistical significance of various geological, reservoir and engineering parameters. A public dataset related to oil fields in Gulf of Mexico is used for this study.

This dataset consists of 4000 oil reservoirs and each reservoir has 80 attributes. Initial data cleaning was carried out on this dataset to remove reservoirs with erroneous data entries. Dimensionless reservoir parameters are defined and used for the study to make the models consistent to other reservoirs. In the model development, 80% dataset was used to train the model and the rest dataset was used to evaluate the trained models. A few models based on their intrinsic design predicted the ultimate recovery factor with an accuracy of 8-9%, and a few other models predicted the same with an accuracy of 10-

12%. Ensemble of a few models predicted oil recovery factor with and accuracy of 6%. In addition to predicting ultimate recovery factor, relative importance of various dimensionless parameters, and sensitivity of ultimate recovery factor to reservoir and engineering parameters is studied. This kind of study uses already available reservoir data to provide a quick means to evaluate new oil reservoirs even with limited data.

1. Introduction

1.1 Purpose and Significance of the Study

Data Analytics has been used in upstream oil industry from a long time. Data analytics techniques such as linear regression were used even before the advent of computers and numerical simulation. But recently significant development of data analytics algorithms, increasing computational power and decreasing data storing and data handling costs enabled effective simulation of complex relationships in E&P data.

This study uses analogous reservoirs data and data analytics for predicting ultimate recovery of oil reservoirs in Gulf of Mexico. Dimensionless reservoir parameters representing various reservoir characteristics were defined along with geological and engineering parameters. Predictive power and importance of various parameters were measured and predictors having significant predictive power were used for further study.

A comparative study of different data analytics techniques such as multi linear regression, robust regression, least absolute shrinkage and selection operator, decision trees, k nearest neighbours, random forests and artificial neural networks has been carried out. Based on the intrinsic methodology of these algorithms, a few methods predicted recovery factor with a mean error of 10-12% whereas a few predicted the same with an error of 9-10%. But Ensemble model due to its capability to combine multiple models predicted ORF with an accuracy of 6%.

The code is made in such a way that it can be upscaled to other reservoirs and fields. Inclusion of different kinds of development strategies and EOR techniques used can help in predicting the best combination of production strategies and EOR techniques for improving the ultimate recovery factor. This kind of analytical models provides methodology for predicting ball point ultimate recovery factor and helps in reducing the number of reservoir simulation cases that has to be run. These models can also be used with partially available reservoir data.

1.2 Objectives

The objectives of the study were as follows

1. Defining dimensionless parameters to characterize development strategy and heterogeneity of the reservoir based on the available data. These numbers will be used along with conventional dimensionless parameters and qualitative reservoir parameters for making predictive models.
2. Building data analytical models based on the data related to already exploited reservoirs to predict ultimate recovery factor of new reservoirs. Using dimensionless parameters to reduce the number of parameters and to make the models easily scalable to new reservoirs.
3. Evaluating the developed models not only based on the error metrics but also on their ability to capture the natural trend in the data. Also evaluating the models based on their ability to reproduce natural trends exhibited by oil reservoirs.

1.3 Outlay of the thesis

This thesis is divided into five chapters. Chapter 2 discusses about previous works related to this study, new dimensionless numbers defined and methodology of various data analytical models. Previous data analytical models developed by Arps et al (1956), Isehunwa & Nwankwo (1994), Gulstad (1995), Sharma et al (2010), Darhim et al (2016) and Srivastava et al (2016) are discussed in first section of this chapter. Second section of this chapter describes the logic behind development number and heterogeneity number. Final section explains modelling techniques likes Multiple linear regression, Robust and Penalized regression, regression trees, Random forests, k nearest neighbors and Artificial neural networks.

Chapter 3 discusses about sequence of data filtering and processing applied to the available dataset such that it can be used for building various analytical models. Section two of this chapter discusses various data processing and transformation techniques like identification and removal of outliers, Skewness and Box cox transformation, centering and scaling the predictors and converting categorical predictors into dummy variables. Final section in this chapter explains concept of bias and variance and requirement of test data to evaluate data analytical models.

Chapter 4 discusses various steps in each modelling techniques used in this study. It also explains the logic behind various error metrics and visualizations used to evaluate the relationship between predictors and ultimate oil recovery factor. Chapter 5 summarizes the work done in this thesis followed by conclusions and recommendations for future work.

2. Review of Literature

In this chapter, previous works carried out by the different authors on use of data analytics for predicting the performance of oil and gas reservoirs is reviewed. Use of data analytics for predicting the performance of oil and gas reservoirs dates back to 1945. Many researchers, Craze and Buckley (1945), Vietti et al (1945), Muskat and Taylor (1946), Guthrie and Greenberger (1955), Arps (1955), Arps et al (1955), Isehunwa and Nwankwo (1994), Gulstad (1995), Oseh and Omotara (2014), Srivastava et al (2015), and Priyank et al (2016) used different data analytics techniques to develop relationship between oil recovery factor and reservoir properties. In recent times, due to the improvements in the computational power and analytical algorithms, researchers are using sophisticated algorithms such as self-organizing maps(SOM), Decision Trees and Random Forests, Artificial Neural Networks and fuzzy logic for predicting recovery factor using reservoir and fluid parameters. This chapter describes different stages of data analytical models developed by various researchers and data analytical techniques.

2.1 Previous Data Analytics models

Guthrie and Greenberger (1955) model was based on 73 sandstone reservoirs with water drive mechanism. Reservoir and fluid properties such as permeability, porosity, oil viscosity, formation thickness, connate water saturation, oil formation volume factor, depth of the reservoir, well spacing and area were used as predictor variables and recovery factor (RF) is the target variable. The developed correlation is as follows

$$E_R = 0.2719 \log k + 0.25569\, S_{wi} - 0.1355 \log \mu_o - 1.5380\phi - 0.00003488H + 0.11403 \qquad (2 - 1)$$

This correlation predicted 50 percent of the time within 6.2% of the actual value and 75 percent of the time within 9.0% of the actual value. Arps et al. (1972) model was based on data from 312 sandstone reservoirs. It had different equations for water drive reservoirs and depletion drive reservoirs. The relationship for recovery factor for water drive reservoirs is as follows

$$E_{R_{water\ drive}} = 0.54898(\phi \frac{(1-S_{wi})}{B_o})^A (\frac{K\mu wi}{\mu_{oi}})^B S_{wi}^C\ (\frac{P_i}{P_a})^D \qquad (2 - 2)$$

Where

A=0.0422, B=0.0770, C=-0.1903, D=-0.2159

The correlation coefficient for predicted and actual recovery factors for the above model is 0.958. The relationship for solution gas drive reservoirs is as follows

$$E_{R_{sol\ gas\ drive}} = 41.815(\phi \frac{(1-S_{wi})}{B_o})^A (\frac{K*1000}{\mu_{ob}})^B S_{wi}^C\ (\frac{P_i}{P_a})^D \qquad (2 - 3)$$

Where

A=0.611, B=0.0979, C=0.3722, D=0.1741

The limitation of the study is selection of small number of reservoirs.

2.1.1 Arps and Isehunwa & Nwankwo (1994)

Arps et al. *(1956)* developed a residual oil saturation model for water drive reservoirs and depletion drive reservoirs. The equations developed are as follows

$$RF_{waterdrive} = \frac{1-S_{wi}-S_{or}}{1-S_{wi}} \qquad (2 - 4)$$

$$RF_{depletion} = 1 - \frac{1-S_w-S_g}{1-S_w} \qquad (2 - 5)$$

Isehunwa & Nwankwo (1994) further developed Arps model based on data from 12 reservoirs in Niger Delta. They induced a constant C into Arps model.

$$RF = C * \frac{1-S_{wi}-S_{or}}{1-S_{wi}} \quad (2 - 6)$$

Where C is 0.8447862 for the set of reservoirs considered in the study. This model is based on only a few reservoirs and may not be generalized.

Gulstad (1995) developed a model for predicting the recovery factor using multi-linear regression on water drive and solution gas drive reservoirs with including the heterogeneity of the reservoir. The equations developed by Gulstad are as follows

$$REC_{waterdrive} = -279 + 0.44(OOIP) - 56.70\ln(\mu_{oa}) - 119.45\ln(S_w) + 0.04(P_{ep}) - 4.73(\mu_{oi}) + 4.38(\mu_{oa}) + 0.24(OOIP)_{calc} - 0.88(T) \quad (2 - 7)$$

$$REC_{sol.\ drive} = -264.034(OOIP) + 29.37\ln(R_{si}) - 0.06\lambda_o - 12.64\ln(h) \quad (2 - 8)$$

It can be observed from the solution drive equation that formation thickness has negative impact on recovery factor. It can also be observed that a few important factors such as number of wells, area of the reservoir were not included in the equation.

2.1.2 Sharma et al

Sharma et al. (2010) used TORIS and GASIS data sets for building statistical models to predict ultimate recovery factor. TORIS data set was developed by the National Petroleum Council (NPC) for assessing the EOR potential (Sharma et al., 2010). This database consists of over 1300 oil reservoirs with 29 variables each. Whereas the Gas Information System (GASIS) is a similar data base for gas fields (Sharma et al., 2010). Sharma et al. used various data analytical models such as multiple linear regression and Principal component analysis to model the effect of various reservoir and fluid parameters on ultimate recovery factor. Oil recovery factor (ORF) is split into different categories and likelihood of recovery factor being in each bin is also estimated. One of the limitations of this study is poor cross validation. Even though the model is built on more than 1300 reservoirs, it is cross validated only on 6 reservoirs.

Darhim et al. (2016) used artificial neural networks to predict oil recovery factor. Predictors related to asset economics, technology, facilities, start of production, number of wells, reservoir architecture, rock and fluid properties and others were used. Two Artificial neural networks were developed with different levels of complexity. Both these Artificial neural networks predicted ultimate recovery factor with an accuracy of 9.5% and 8.0%; respectively. In addition to regular predictors, this model quantified the type of technologies used in the field such as 4D-Seismic, 3D-Seismic, VSP, type of tertiary recovery techniques used, type of secondary recovery techniques, type of artificial lift, Asset remoteness and facilities etc. This model cannot be interpreted openly due to the use of artificial neural networks.

Srivastava et al. (2016) used dimensionless numbers with data mining techniques to predict the recovery factor of oil fields having water drive in Gulf of Mexico. The reservoirs considered for the study are clustered into different groups using k-means clustering. The predictions on each cluster is evaluated based on the correlation between predicted and actural ORF. This study concentrates only on water drive reservoirs in Gulf of mexico. Training and test data were not defined for more reliable evaluation of the model. Additionally, Srivastava's model does not include the reservoir heterogeneity effect on recovery factor.

2.2 Dimensionless parameters

It can be observed that most the predictive models discussed in the previous section are reservoir dependent and can’t be scaled to apply them for predicting RF in different fields. This calls for a model that can be independent on field scale. This work continues based on the dimensionless numbers described in Srivsatavas model (Srivastava et al. 2016) with two additional dimensionless parameters discussed as follows.

2.2.1 Development number

This number is defined to represent the extent of development in the field and is defined as follows

$$Dev\ number = \frac{k * Number\ of\ wells}{Area\ of\ the\ reservoir} \tag{2 - 9}$$

It can be observed that :

i. For the same area and number of wells, higher permeability leads to higher development number and vice versa.

ii. For same permeability and area of the reservoir, higher number of wells leads to higher development number and vice versa.

iii. Similarly, for same number of wells and permeability, reservoirs with low area will have high development number and reservoirs with high area will have low development number.

It can also be observed that development number defined above is dimensionless.

$$Dimension\ of\ Permeability = [L]^2 \qquad Dimension\ of\ Area = [L]^2$$

$$Dev\ number = \frac{[L]^2 * Dimensionless}{[L]^2} \tag{2 - 10}$$

Development number can be further improved by incorporating average number of wells produced over the reservoir life. Time factor is not incorporated in this study due to non availability of time data in the dataset.

$$Average\ wells = \frac{(w_1t_1+w_2t_2+w_3t_3+\cdots w_nt_n)}{t} \quad (2 - 11)$$

$$Dev\ number = \frac{k*Average\ wells}{Area\ of\ the\ reservoir} \quad (2 - 12)$$

2.2.2 Heterogeneity number

The heterogeneity of the reservoir plays very crucial role in oil recovery and should be included in recovery factor estimation model. There are different ways to characterize heterogeneity of a reservoir, and in this study the heterogeneity number is defines as follows.

$$Hetro\ number = \frac{1}{(NTG)(\frac{Oil\ Area}{Total\ Area})} \quad (2 - 13)$$

It can be observed that if NTG=1 and oil area is equal to the total area of the reservoir, the heterogeneity number will be equal to 1. The lower the product of NTG and $\frac{Oil\ Area}{Total\ Area}$ the higher will be the Hetro number. Reservoirs having low NTG or low $\frac{Oil\ Area}{Total\ Area}$ or both will have high heterogeneity number. With the available data in the data set, heterogeneity of the reservoir is incorporated into the model with this heterogeneity number.

2.3 Data Analytics Models

2.3.1 Multiple Linear Regression

A linear regression models can be expressed as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots\ldots.\beta_n x_n + e_i \qquad (2 - 14)$$

The error metric sum of squared errors (SSE) for ordinary least squares regression is as follows

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \qquad (2 - 15)$$

The objective of least squares regression is to obtain combination of beta coefficients which minimizes the sum of squared errors (SSE). Where $\widehat{y_i}$ represents the numeric outcome, β_0 represents the intercept, β_i represents the coefficient of i^th predictor x_i. e_i represents the random error. The predictors x_1, x_2, x_3 ……. can be independent or can be combination of multiple predictors to depict non linear interactions between them. The beta parameters ($\beta_0, \beta_1, \beta_2,$ ……) are estimated in such a way that the error metric SSE is minimized. Each variant of linear regression such as robust regression and penalized regression has a different definition for error metric to attain optimum balance between bias and variance. The beta parameter estimation by ordinary least squares regression tends to have minimum bias whereas parameter estimates by other regression techniques such as robust regression and penalized regression tends to balance between bias and variance. This tradeoff between bias and variance characterizes their predictions (Graybill & Franklin 1976).

The advantage of linear regression is easy interpretability of the model. Predictors with negative beta coefficients have negative impact on the target variable where as predictors with positive beta coefficients have positive impact on the target variable. It can also be understood that predictors having higher beta coefficients have higher weightage on target variable than predictors having lower beta coefficients. The main limitations of linear regression is it cannot model non linear interactions between predictors implicitly. All non linear interactions should be tried and interactions having statistical significance should be entered into the model explicitly.

The beta coefficients for ordinary least squares regression (OLS) can be computed using the matrix multiplication shown below (Graybill 1976).

$$(X^T X)^{-1} X^T Y$$

Where X is the matrix of predictor parameters and Y is the matrix of target variable. Matrix $(X^T X)$ is invertible only if

a) None of the predictors can be expressed as a linear combination of others

b) Number of observations is more than number of predictors.

The determinant of $(X^T X)$ will tend to zero if highly correlated predictors are present in the data. In that case beta coefficients will get inflated and loose their meaning. This makes it necessary to remove highly correlated predictors before modelling.

2.3.2 Robust and Penalized Regression

The balance between bias and variance of multiple linear regression can be manipulated by changing the objective function. With increasing degee of error term in the objective function, the model will become more sensitive to outliers and will have more variance. **Table 1** shows the objective function of Least absolute value (L1) regression which is less sensitive to outliers than OLS regression. **Figure 1** shows the relationship between residuals and their contribution to objective function. It can be seen that higher residuals get more weightage in OLS regression than L1 regression. Multiple linear regression based on Huber loss function (Huber 1964) uses squared residuals for residuals with magnitude less than k and absolute value for residuals more than k. L1 regression and linear regression based on Huber loss function are two variants of robust regression methods.

Models with near zero $(X^T X)$ determinant will have high variance due to inflated beta coefficients. Penalized regression controls variance of these models by adding penalty for higher beta coefficients in objective function. **Table 1** shows objective functions of ridge regression, Least absolute shrinkage and selection operator and Elastic net regression which are variants of penalized regression (Hoerl 1970; Tibshirani 1996; Zou & Hastie 2005). $\sum(y_i - \hat{y}_i)^2$ term represents the bias of the model and $\sum {\beta_j}^2$, $\sum |\beta_j|$ represents the variance of the model. The trade off between bias and variance can be changed by changing λ value. Using $|\beta_j|$ in objective function shrinks the coefficients of predictors to zero which provides LASSO additional feature selection ability (Tibshirani 1996). The number of predictors and model accuracy of LASSO can be optimized by changing λ. As the value of λ increases, more important predictors will remain in the model and less important predictors will get discarded. Additional advantage of LASSO over MLR is automatic feature selection and elimination of highly correlated predictors.

Model	Contribution of each observation to loss function
Ordinary least squares regression	$\sum (y_i - \hat{y}_i)^2$
Least absolute value (L1) regression.	$\sum \lvert y_i - \hat{y}_i \rvert$
Huber loss function	$\frac{1}{2}(y_i - \hat{y}_i)^2 \;\; if \;\; \lvert y_i - \hat{y}_i \rvert \;\le\; k$ $k\lvert y_i - \hat{y}_i \rvert - \frac{1}{2}k^2 \;\; if \;\; (y_i - \hat{y}_i) \;>\; k$
Ridge regression	$\sum (y_i - \hat{y}_i)^2 + \lambda \sum {\beta_j}^2$
Least absolute shrinkage operator (LASSO)	$\sum (y_i - \hat{y}_i)^2 + \lambda \sum \lvert \beta_j \rvert$
Elastic net regression	$\sum (y_i - \hat{y}_i)^2 + \lambda_1 \sum \lvert \beta_j \rvert + \lambda_2 \sum {\beta_j}^2$

Table 1: Loss function for different variants of multiple linear regression

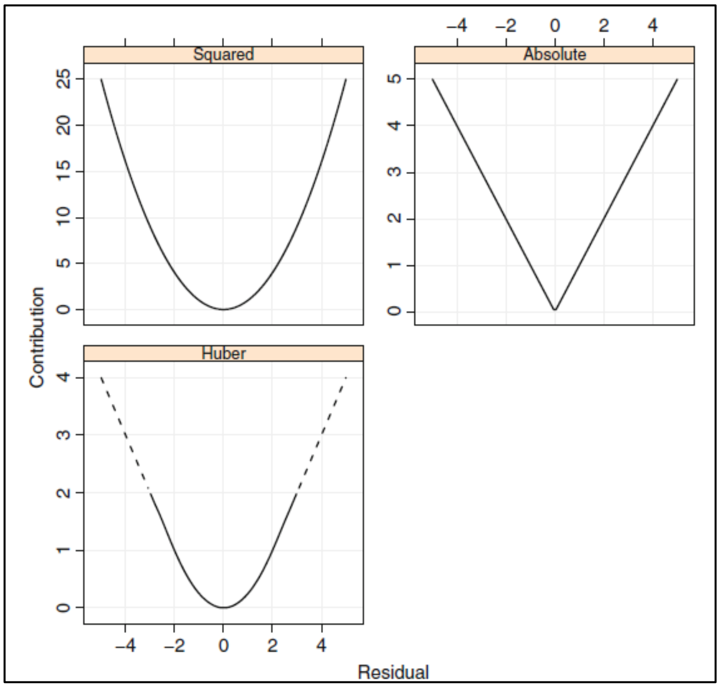

Figure 1: Residuals and their contribution to objective function (*Kuhn & Johnson 2013*)

2.3.3 Regression Trees

Classification and regression trees (CART) is the one of the oldest modelling techniques. Based on their approach classification and regression trees can model non linear interactions between predictors. CART models splits the data using if-else conditions on various predictors such that sum of with in sum of squared error (SSE) in each group is minimized (Breiman et al. 1984). **Figure 2** shows the schematic of how regression tree splits the observations based on predictor variables. The groups of observation at the end of the regression tree are called as leaf nodes. The objective function of a regression tree having n leaf nodes is defined as follows

$$SSE_n = \sum_{j=1}^{n} \sum_{i \in S_j} \left(y_i - \overline{y_j} \right) \qquad (2 - 16)$$

S_j and $\overline{y_j}$ represents the observations and mean of jth leaf node respectively. Based on the above definition, large size trees will always produce smaller SSE_n Complexity parameter is used to penalize large sized trees as shown below (Breiman et al. 1984).

$$SSE_{cp} = SSE + C_p * (\,No.of\ leaf\ nodes) \qquad (2 - 17)$$

Figure 3 shows relationship between C_p, crossvalidated error and size of regression tree. As C_p decreases, size of the tree increases and optimum size of the tree is where minimum cross validated error is observed. Standard error bars in Figure 3 represent the tolerance for selecting the size of the tree. Tree size having relative error within one standard error of best tree size can be selected for reducing the complexity.

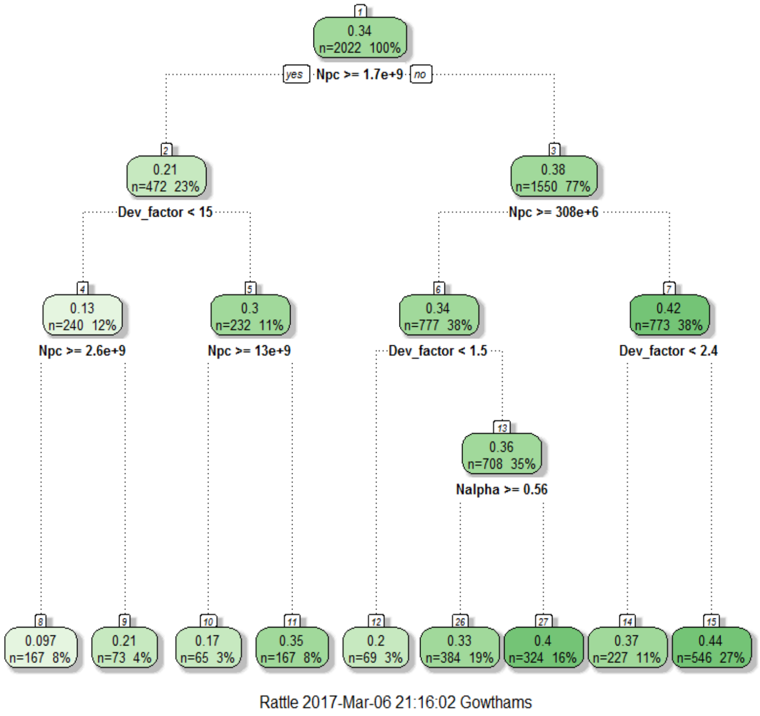

Figure 2: Decision Tree model

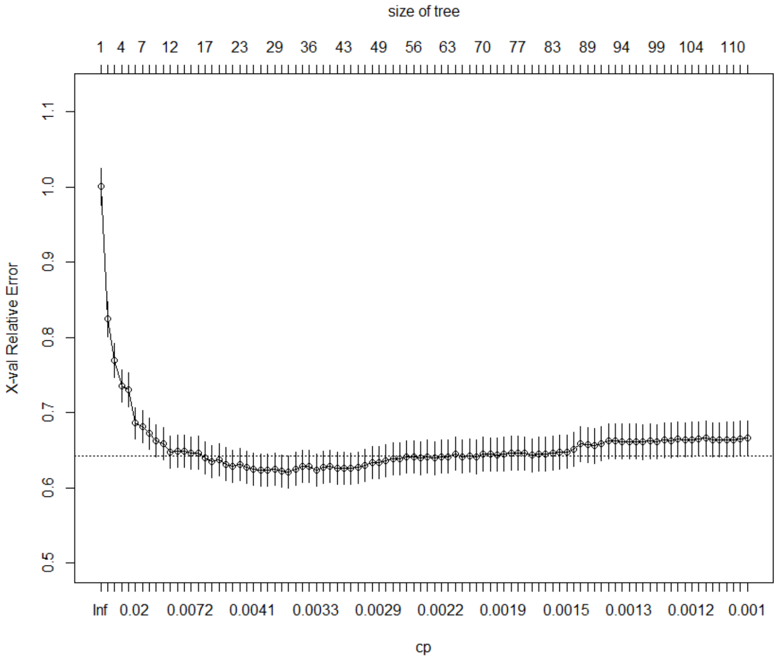

Figure 3: Change in C_p with regression tree size

Advantages of regression trees include

- Automatic feature selection
- Robust to outliers
- Doesn’t require normalization of predictors
- Easy to interpret

One of the major disadvantages of regression trees is instability (Breiman 1996). Regression trees are highly unstable and addition of new data will lead to change in the structure of the tree and decision rules (high variance). Other disadvantages include finite number of outcome levels, ability to make only linear splits and selection bias for predictors with higher number of factor levels (Loh & Shih 1997).

2.3.3 Random Forests

Random forests model is an ensemble of many regression trees and each node in these trees is split based on randomly selected ‘m’ predictors (Breiman 2001) . The main objective of Random forests is to use large number of decorrelated trees for prediction. Decorrealtion is achieved by making splits based on randomly selected predictors. The final outcome of random forest is average prediction of all the trees in the model. This kind of decorrelated ensemble models brings out the signal with suppressing the noise (Breiman 2001).

The variance of Random forests model increases with increasing ‘m’. As ‘m’ approaches total number of predictors in the model, the random forests model will tend to ack like a single regression tree. The number of trees in the random forest can be as many as possible but should be within the limits of computational power. The performance of the random forest model improves with increasing number of trees and reaches plateau at optimum point. Advantages of random forests model include stable predictions, automatic feature selection, robust to outliers, doest require normalization of predictors and resistant to overfitting. Even though individual trees in the random forest model can be interpreted, the actual model in whole cannot be interpreted.

2.3.4 K-Nearest Neighbors

The K-Nearest neighbors model identifies k nearest neighbors in training data to any instance of new data based on distance metric calculated on predictor variables (Cover & Hart 1967). The outcome of the prediction can be any one of the summary statistics such as mean, median or mode of the target variable of k nearest neighbors. **Table 2** shows a few distance metrics used for identifying k nearest neighbors.

Method	**Distance Expression**
Euclidean distance	$\sqrt{\sum_{j=1}^{P}(x_{aj} - x_{bj})^2}$
Murkowski distance	$\sqrt[q]{\sum_{j=1}^{P}\lvert x_{aj} - x_{bj}\rvert^q}$
Manhattan distance	$\sum_{j=1}^{P}\lvert x_{aj} - x_{bj}\rvert$

Table 2: Distance Metrics for k nearest neighbors

The scale of the predictors highly effects the weightage of the predictor in calculating distance metric. It is important to center and scale the predictors before calculating distance metrics. In case of high dimensional datasets, number of predictors in the model can be selected based on statistical significance of predictors.

2.3.5 Artificial Neural Networks

Artificial neural networks (ANN) are one of those modelling techniques which can implicitly model nonlinear relationships between predictors (Bishop 1995). ANNs loosely mimic the way biological brain works using large clusters of neurons connected by axons. Biological brain is much faster than computers in tasks such as pattern recognition. By mimicking biological brain, ANNs have extra advantage in tasks where human brain is good at.

The outcome of ANNs is a linear combination of inputs from hidden nodes which in turn are linear combination of inputs from other hidden nodes or predictor variables. Artificial neural networks can have one or more than one hidden layers. The linear combination at each node can be transformed using appropriate nonlinear, linear, exponential, or logical functions. An example of typical calculation in neural network is as follows

$$h_k(x) = g\left(\beta_{0k} + \sum_{i=1}^{P} x_j \, \beta_{jk} \right) \qquad (2 - 18)$$

$$g(u) = \frac{1}{(1+ e^{-u})} \qquad (2 - 19)$$

$$f(x) = \gamma_0 + \sum_{k=1}^{H} \gamma_k \, h_k \qquad (2 - 20)$$

Figure 4 shows a typical sequence of modelling in ANN. β_{jk} depicts the effect of the j_{th} predictor on k_{th} hidden unit. γ_k is the contribution of each hidden unit to the outcome. For a model with P predictors and H hidden units, the total number of parameters that has to be estimated is equal to $H(\, P + 1) + H + 1$. The objective of ANN is to calculate beta parameters such that the squared error is minimized. Beta parameters are initialized to random values and are corrected using gradient descent techniques such as back propagation algorithm (Rumelhart et al. 1986). Though ANNs can model non linear interactions between predictors implicitly, these models are difficult to interpret.

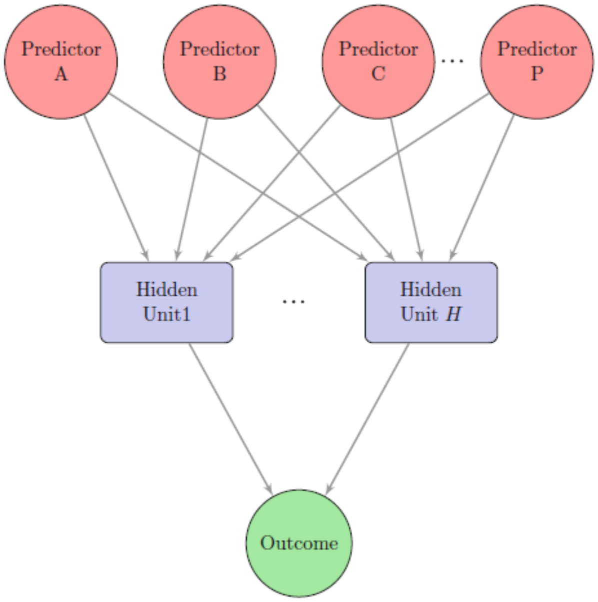

Figure 4: Schematic of basic ANN *(Kuhn & Johnson 2013)*.

3. Data preparation

The data set used in this study has 13289 reservoirs and each reservoir has 82 attributes (BOEM). Only one Reservoir having missing data is removed and this left the data with 13288 instances. The following techniques have been applied to clean the data before analysis.

3.1 Data Filtering

Table **3** shows the sequence of steps followed for data filtering. Reseroirs with play type (B1, R1, X2) are removed from the data set as the number of reservoir instances with this play types are less than the number of predictors. Similarly reservoir instances with structure (F, I, G, H) were also removed from the dataset. These type of reservoir instances would result in singular matrices which cannot be solved for estimating beta coefficients. Categorical predictors in the dataset are converted into dummy variables to incorporate them into regression models. As further described in section 3.3.6, a categorical predictor with n levels will require n dummy variables to represent it in regression models. It is necessary to minimize the number of levels in categorical predictors to minimize total number of predictors in the model. **Table 4** shows the approach followed for re organizing factor variables.

Table 5 shows dimensionless reservoir parameters that are calculated and merged with original data set. The definitions for dimensionless reservoir parameters provided by Shook et al., 1992 well be used in this study. In addition to these, end-point mobility ratio is also calculated and merged with the dataset. 80% and 40% of effective permeability is considered as relative permeability of oil and water respectively.

Filter No	**Reason**	**Conditions used**	**Remaining Reservoirs**
1	Missing Data	Sub setting complete cases	13288
2	Selecting oil reservoirs	SD_TYPE==” O” | SD_TYPE==” B” | SD_TYPE==0	5019
3	Deselecting reservoirs having GOR>30	oilRes$GOR<30	4313
4	Dropping reservoirs having play type B1, R1, X2 due to very few observations	PLAY_TYPE != ‘B1’ & PLAY_TYPE != ‘R1’ & PLAY_TYPE != ‘X2’	4244
5	Dropping reservoirs having structure F, G, H, I due to very few observations	FSTRU != F & FSTRU !=I & FSTRU != G & FSTRU != H	4011
6	Dropping reservoirs having drive O, GCP, SLG, UNK	DRIVE != O & DRIVE !=GCP & DRIVE != SLG & DRIVE != UNK	3724
7	Dropping reservoirs having zero OIP, ORF, BHCOMP, PERMEABILITY	OIP !=0 & BHCOMP !=0 & PERMEABILITY !=0 & ORF !=0	3342
8	Considering reservoirs which produced more than 80% of estimated recoverable oil	P_CUMOIL> 0.8* P_RECOIL	3038

Table 3: Steps followed for data filtering

Predictor	Original factors levels	Releveled factor level
CHRONOZONE	MML, MUL, MLM, Mmmm, MUM, MMM, MLL, MUU, KUL, MLU	MIOCENE
	PL	PLIO_LWR
	PU, PU-PL	PLIO_UPR
	PLU-LL, PLM, PL, PLU, PLL	PLEISTOCENE
DRIVE	DRIVE==’0’	UNK
	DRIVE==’GCP’	COM (Merging Gas cap with combination drive reservoirs)

Table 4: Re organizing factor variables

Dimensionless number	Formula
Capillary Number (N_{pc})	$N_{Pc} = \frac{\lambda^{o}_{r2}\sigma}{LU_t}\sqrt{\phi K_x}$
Gravity Number (N_g)	$N_g = \frac{K_z\lambda^{o}_{r2}\Delta\rho g cos\alpha}{u_t}\frac{H}{L}$
Aspect Ratio (R_l)	$R_l = \frac{L}{H}\sqrt{\frac{K_z}{K_x}}$
Density Number (D_n)	$N_\rho = \frac{\rho_o}{\Delta\rho}$
Development factor	$Dev_{factor} = \frac{k * No\ of\ wells}{Area}$
Heterogeneity factor	$Hetro_{factor} = \frac{1}{(\frac{oil\ Area}{toal\ Area})(NTG\ ratio)}$

Table 5: Dimensionless paramters used in the study

3.2 Identifcation and Removal of extreme predictors

Different descriptive statistics and visualizations were used to identify and remove nonphysical data entries. Reservoir instances with any one of the predictors having nonphysical entries was removed from the dataset.

a) End Point Mobility Ratio

Table 6 shows the distribution of End Point Mobility Ratio in the original data set. It can be observed that the 3rd quartile of the data is below 1 and maximum value is 27. Also **Figure 5** shows skewed distribution of End point mobility ratio in the original dataset. It indicates that reservoir instances with very high end point mobility ratio would be outliers. In this study, reservoirs having End Point Mobility Ratio more than 10 were removed from the dataset

Min	1st Quartile	Median	Mean	3rd Quartile	Max
0.249	0.576	0.701	0.854	0.875	**27**

Table 6: Summary statistics of End point mobility ratio

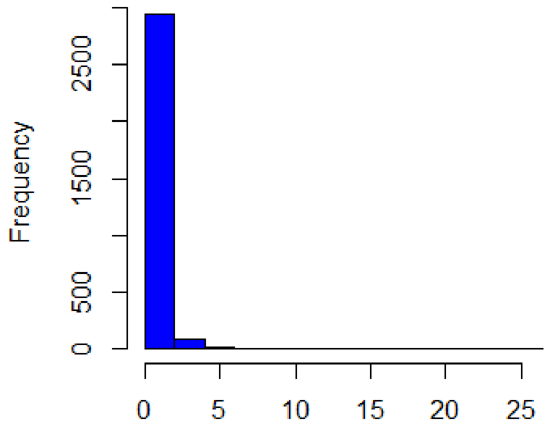

Figure 5: Skewed distribution of End point Mobility Ratio

b) Density Number (N_ρ)

Table 7 shows the the distribution of N_ρ in original dataset. Most of the reservoirs in the dataset are having N_ρ between 3.628 and 6.738 and the maximum value of N_ρ is 141.5. Also **Figure 6** shows the skewed distribution of density number. Reservoir instances having N_ρ more than 15 are considered as outliers and removed from the data set. Similarly summary statistics of each predictor is verified and reservoirs having extreme values in any of their predictors were removed from the dataset.

Min	1st Quartile	Median	Mean	3rd Quartile	Max
3.628	5.421	5.896	6.312	6.738	**141.5**

Table 7: Summary statistics of N_ρ

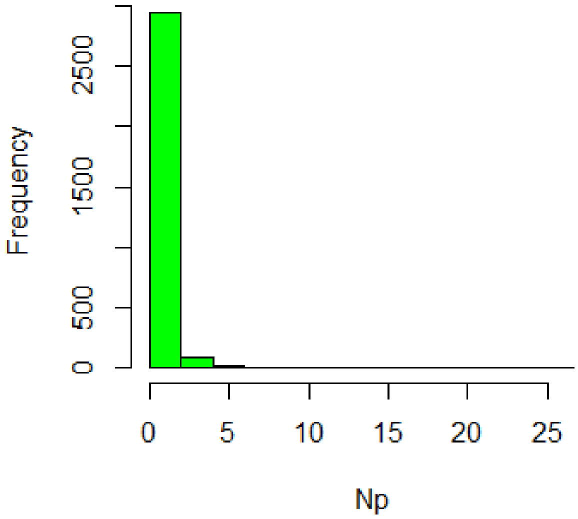

Figure 6: Skewed distribution of N_ρ

3.3 Data transformation

Figure 7 shows the initial distribution of the predictors in the data set. It can be observed that these predictors are in different scales. To have uniform weightage for all the predictors in modelling techniques like LASSO, k nearest neighbours and ANN they have to be normalized, scaled and centered. Therefore predictors number of wells, density number (N_{α}), capillary number (N_{pc}), development factor, gravity number (N_{g}) were normalized as described in the following sections. Heterogenity number is not normalized to preserve its physical meaning

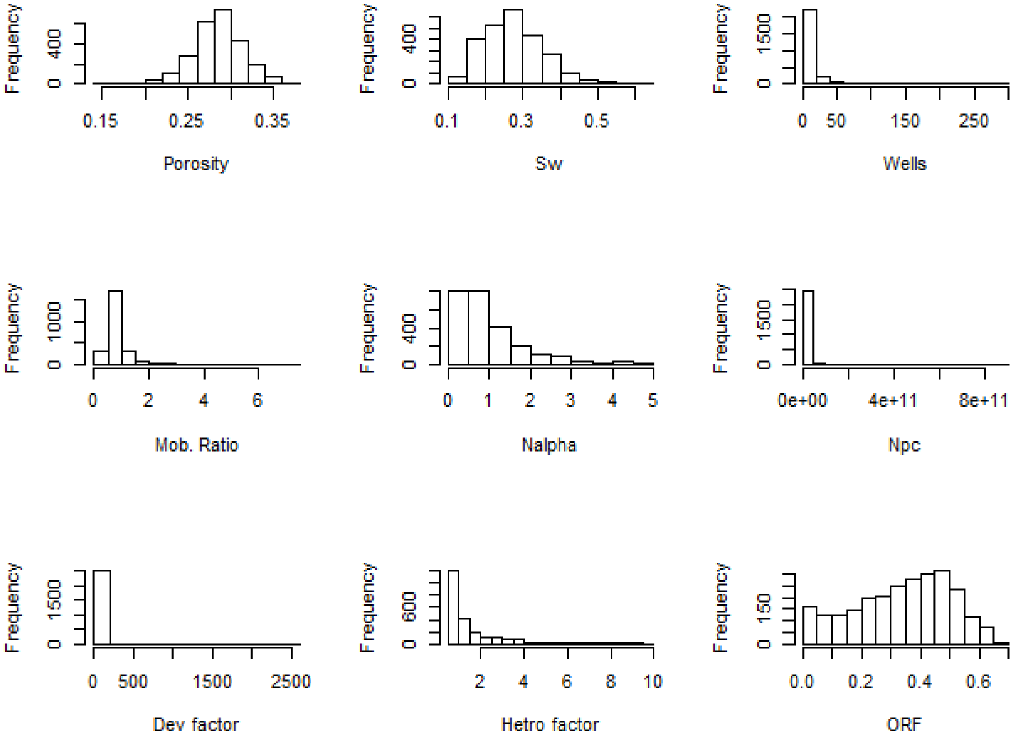

Figure 7: Distribution of predictors in original dataset

3.3.1 Skewness and Box-Cox transformation

An unskewed distribution is roughly symmetric about the mean, whereas right skewed distribution has more percentage of values left of the mean in a histogram plot. Similarly left skewed distribution has more percentage of values on right side of the mean. Skewness statistic is defined as follows

$$skewness = \frac{\sum(x_i - \overline{x})^3}{(n-1)v^{3/2}} \tag{3 - 1}$$

$$where \quad v = \frac{\sum(x_i - \overline{x})^2}{(n-1)} \tag{3 - 2}$$

Box-Cox transformation can be used to transform skewed data into normally distributed data (Box, 1964). It is defined as follows

$$x' = \frac{x^{\lambda\_boxcox} - 1}{\lambda\_boxcox} \tag{3 - 3}$$

Where λ_boxcox is the transformation coefficient. λ_boxcox is calculated using trail and error method, by plotting transformed data in a normal quantile plot. λ_boxcox having highest correlation is considered as the best suitable λ_boxcox for the transformation. If the λ_boxcox is near to zero, applying log transformation will be suitable for normalizing the data. **Table 8** shows the skewness of each predictor used in the analysis.

Predictor	**Skewness**
Porosity	-0.33
Sw	0.59
Permeability	2.5
No. of wells	5.2
Npc	15.9
Ng	22.3
Nalpha	1.8
Dev_factor	22.6
Heterogenity factor	2.16

Table 8: Skewness of each predictor in the original dataset

3.3.2 Centering and Scaling

Certain data analysis techniques such as k-means clustering and PCA require all the predictors in common scale. In general predictors such as porosity will be in the order of 0.1 and predictors such as N_{pc} will be in the order of 10^5. It is necessary to center and scale these predictors such that all of them will have similar influence on the predictive model. Predictor variables are centered and scaled using the following transformation.

$$x_i'' = \frac{x_i - mean(x_i)}{sd(x_i)} \qquad (3 - 4)$$

Figure 8 shows the distribution of predictors number of wells, development factor and capillary number before and after transformation. It can be observed that these parameters which are not having any variance before transformation exhibits significant variance after transformation.

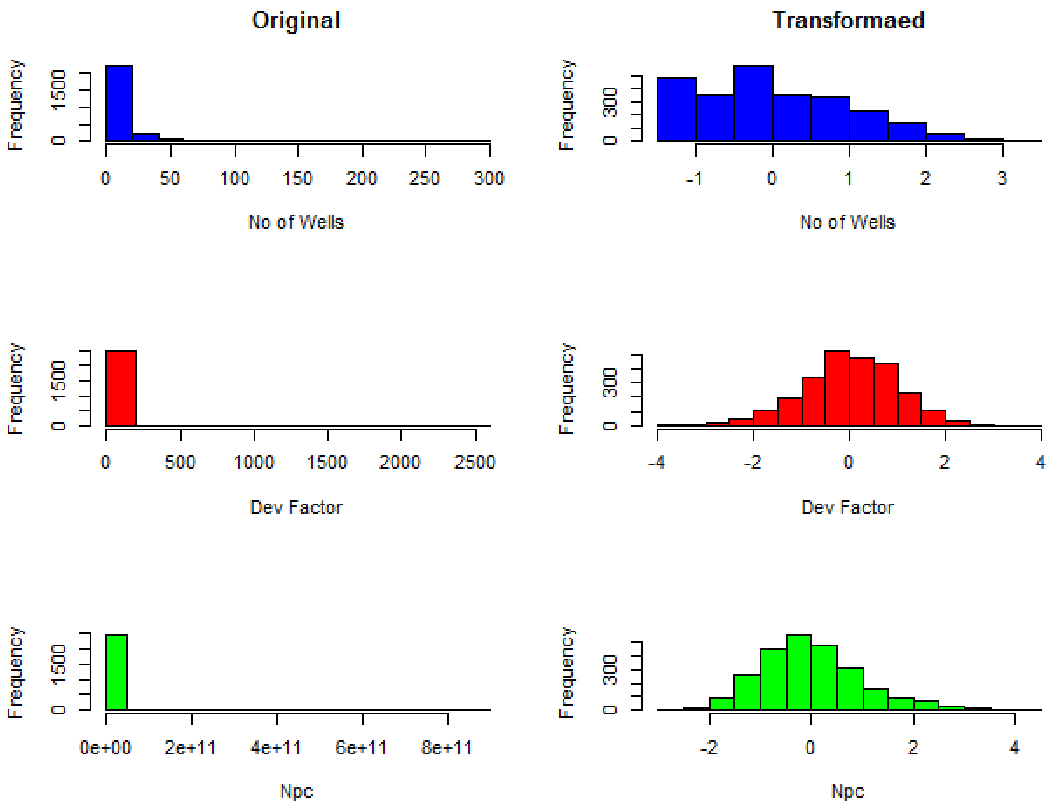

Figure 8: Predictor variables before and after transformation

3.3.3 Removal of outliers

Outliers are those data instances with extreme values which represent unusual circumstances. Sometimes they might have entered by mistake. If outliers are not removed, the derived model without correction to outliers has a poor capability in predicting the general trend. Ordinary least squares regression is sensitive to outliers whereas penalized regression models such as L1 regression, regression based on Huber loss function are somewhat resistant to outliers.

There are various methods available to identify and remove outliers. One of them is based on number of standard deviations. In this study an observation is considered as outlier if it is 3 standard deviations away from the mean and is removed from the data.

Figure 9 shows the distribution of porosity before and after transformation. It can be observed that porosity is converted into standard deviation units and outliers were removed.

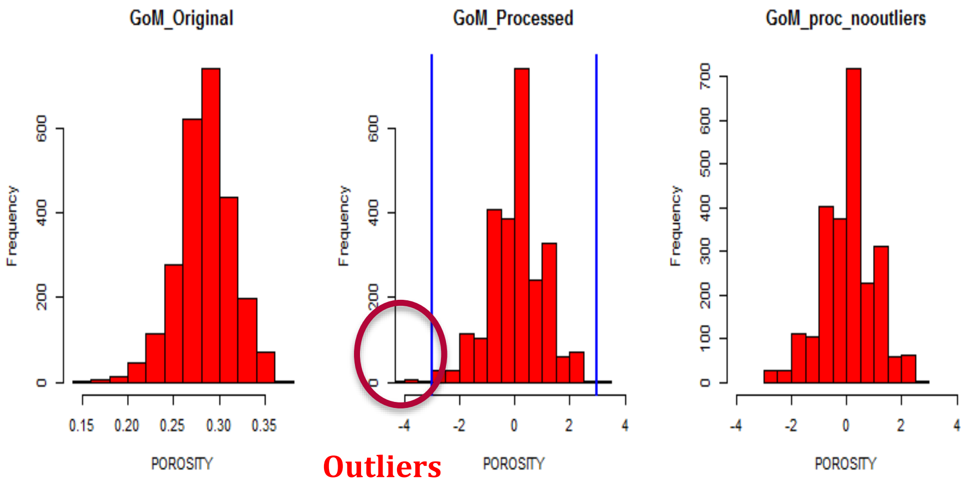

Figure 9: Distribution of porosity in different data sets

3.3.4 Error metrics

In this study, Root mean square error (RMSE) and Mean absolute error(MAE) are used to evaluate the accuracy of different data analytical models. The RMSE and MAE are defined as follows

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2} \qquad (3 - 5)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{f}(x_i)\right| \qquad (3 - 6)$$

$$where\ y_i\ is\ the\ original\ ORF\ and\ \hat{f}(x_i)\ is\ the\ predicted\ ORF$$

3.3.5 Bias and Variance

Bias of an estimator is the difference between the estimated value and original value. Variance of an estimator of the sensitivity of the model to small changes in predictors. Models which are less complex may have bias on the training data but will perform better on new data. Whereas models which are more complex will have low bias on the training data but may perform poor on new data. **Figure 10** shows behavior of models with different variance (complexity). It can be seen that as the complexity of the model increases, MSE on the training data reduces where as MSE on new data decreases upto some extent and then increases. This point is considered as the optimum complexity of the model.

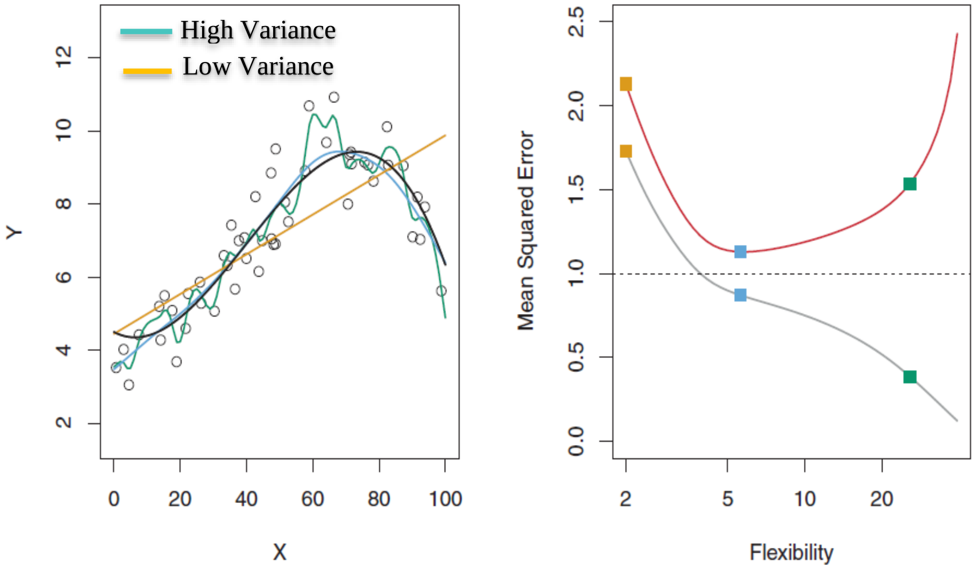

Figure 10: Balancing Bias and Variance (***Kuhn & Johnson 2013*** **)**

3.3.6 Dummy Variables

Dummy variables act like a proxy for categorical predictors in the dataset. They take only 0 or 1 to indicate the presence or absence of a particular category for any reservoir instance. **Table 9** shows the process of converting categorical predictors into dummy predictors. Any categorical predictor originally having 4 levels will require 4 dummy predictors to represent the original predictor. For example structure of the field (FSTRU) having 4 levels (A, B, C, D) is split into four dummy variables FSTRU.A, FSTRU.B, FSTRU.C, FSTRU.D. Structures having positive impact on the recovery will have positive coefficients for their dummy variables and viceversa. To reduce the total number of predictors in the model , categorical predictors were releveled to fewer number of levels.

Reservoir	**FSTRU**
Res-1	A
Res-2	B
Res-3	C
Res-4	D

Reservoir	**FSTRU.A**	**FSTRU.B**	**FSTRU.C**	**FSTRU.D**
Res-1	1	0	0	0
Res-2	0	1	0	0
Res-3	0	0	1	0
Res-4	0	0	0	1
Res-5	0	0	0	0
Res-6	0	0	0	0

Table 9: Example showing converting categorical variables into dummy variables

3.4 Training and Test data sets.

Once the original dataset has been processed, six different datasets were made for various modelling techniques. **Table 10** shows the training and test datasets defined for various modelling techniques. “GoM_original” is the original dataset without nonphysical entries and is further split into training and test data. This dataset is used for Regression tree and Random forest models. GoM_processed dataset is obtained after normal transforming, scaling, and centering the original dataset. This dataset is used for Multi linear regression, Robust regression, LASSO, prediction using kNN and Artificial neural network models. GoM_proc_nooutliers is obtained from GoM_processed after removing outliers.

Data Set	Modelling techniques	Comments
GoMTrain_original	Regression Tree Random Forest	Original dataset
GoMTest_original		
GoMTrain_processed	Multi linear regression kNN Artificial Neural Network	Removed extreme values Normal transformation Centering and Scaling
GoMTest_processed		
GoMTrain_proc_nooutliers	Multi linear regression kNN Artificial Neural Network	Removed outliers
GoMTest_proc_nooutliers		

Table 10: Data Sets used in the Study

To evaluate different models, training and test data sets are separated. **Figure 11** shows the distribution of target variable ORF in training and test dataset. It can be observed that similar distributin of ORF is maintained in training and test data sets. This ensures unbiased error metrics for evaluating predictive models.

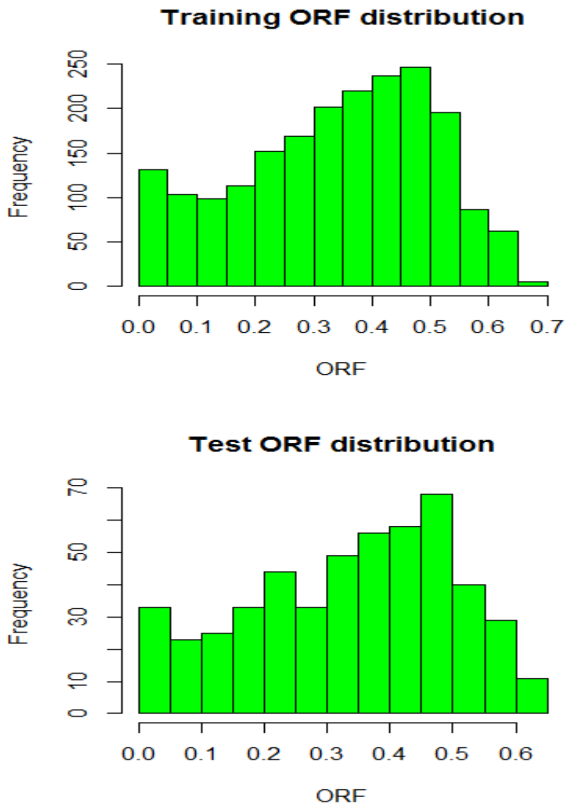

Figure 11: Distribution of ORF in training and test data

4. Evaluation of Modelling Techniques

With the processed data various modelling techniques were used to predict ultimate recovery factor using dimensionless predictors. A few models are easy to interpret but have limitations in modelling whereas few other modelling techniques are difficult to interpret but have better modelling capabilities. It is always essential to balance between complexity and interpretability.

4.1 Multiple Linear Regression

4.1.1 Multiple linear regression without nonlinear terms

Multiple linear regression model is fit on 20 dummy variables and 10 numeric predictors to predict target variable ultimate recovery factor using dataset “GoMTrain_processed”. **Table 11** and **Table 12** shows the coefficients of categorical and numeric predictors respectively. The asterisks in the table depicts the statistical significance of the predictor. Predictors with “***” has a p value of less than 0.001. **Table 13** shows other model statistics. The units of residuals in residuals plot is same as the Oil recovery factor (ORF). The residuals are standardized and converted into standard deviation units to make them consistent. The formula for converting residuals into standardized residuals is shown as follows where s_i is the standardized residual corresponding to residual r_i

$$Standardized\ Residual \quad s_i = \frac{r_i}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n} r_i}} \tag{4 - 1}$$

Standardized residual is measured in units of standard deviations. Standardized residuals of more than 2.5 standard deviations are considered as outliers.

Predictor Variable	**Estimate**	**Std. Error**	**t value**	**Significance**
Intercept)	0.365	0.010	37.901	***
FSTRU.A	-0.019	0.009	-2.098	*
FSTRU.B	-0.020	0.011	-1.869	.
FSTRU.C	0.008	0.009	0.913	
FSTRU.D	-0.040	0.010	-3.827	***
FSTRU.E	-0.040	0.012	-3.34	***
PLAY_TYPE.A1	0.031	0.008	3.909	***
PLAY_TYPE.F1	-0.036	0.007	-5.235	***
PLAY_TYPE.F2	-0.043	0.011	-3.926	***
CHRONOZONE2.MIOCENE	-0.004	0.006	-0.719	
CHRONOZONE2.PLEISTOCENE	-0.004	0.006	-0.689	
DRIVE.COM	-0.009	0.008	-1.084	
DRIVE.DEP	-0.034	0.009	-3.693	***
DRIVE.PAR	-0.001	0.005	-0.122	
RES_TYPE.N	0.030	0.013	2.314	*
RES_TYPE.S	0.002	0.008	0.279	

Table 11: Coefficients of Categorical Predictors

Predictor Variable	**Estimate**	**Std. Error**	**t value**	**Significance**
POROSITY	-0.010	0.003	-2.974	**
SW	0.005	0.006	0.744	
BHCOMP	-0.084	0.006	-14.016	***
Mobility_Ratio_endpoint	0.024	0.003	6.937	***
Nalpha	-0.074	0.007	-10.597	***
Np	0.004	0.004	1.071	
Ng	-0.172	0.015	-11.61	***
Npc	-0.081	0.014	-5.988	***
Dev_factor	0.151	0.007	22.532	***
Hetro_factor	-0.016	0.004	-4.622	***

Table 12: Coefficients of Numeric Predictors

Redidual standard error on 1997 degrees of freedom	0.099
R^2	0.644
Adjusted R^2	0.639
RMSE on test data	**9%**
MAE on test data	**7%**

Table 13: Multiple Liner Regression model Statistics

Figure 12 depics the relationship between standardized residuals and predicted values. It shows that reservoir instances “1261_EI88_K4”, “0961_EI238_c09”, “1361_EI188_N0” have very high standardized residuals. This could be due to some unusual reservoir management techniques or errors in data entry. The area marked with red circle in Figure 12 shows negative ORF predicted by MLR model.

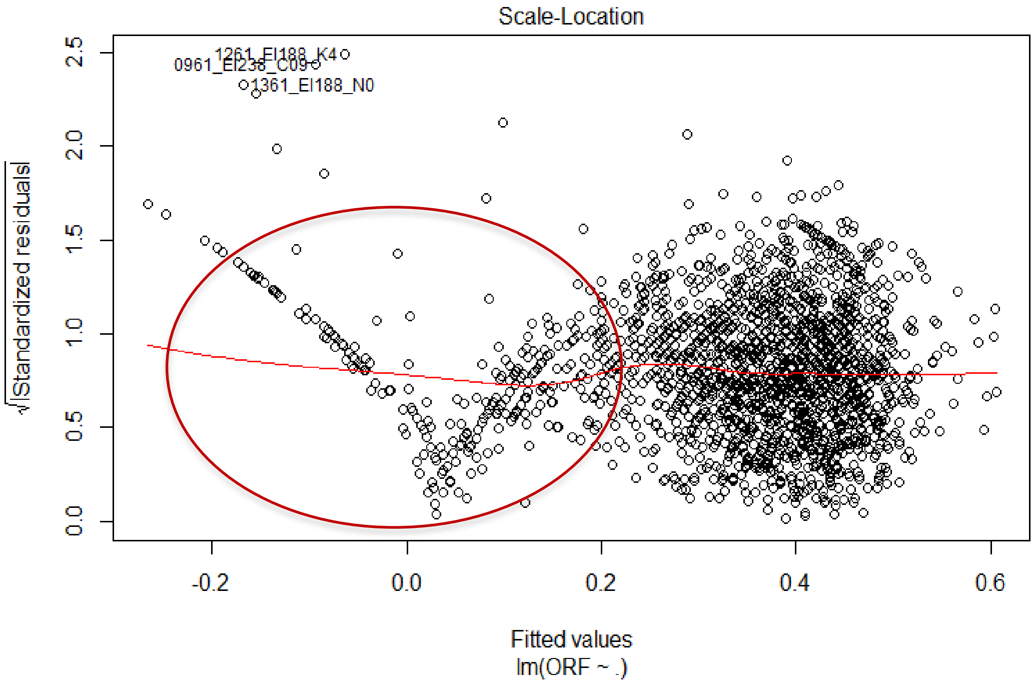

Figure 12: Non linear trend in standardized resicuals vs predicted ORF plot

Figure 13 shows the normal Q-Q plot of residuals in which quantile values of residuals are plotted along with quantile values of a normal distribution. These points will fall on a straight line if residuals are normally distributed which is one of the charecteristics of MLR model. It can be observed that the residuals deviated from straight line at extremes. **Figure 14** depicts the relationship between predicted ORF and error vs original ORF. **Figure 15** shows the distribution of absolute error with MLR without non linear interactions. Figure 12 and Figure 14 indicate the presence of non linear interactions between predictors and indicate that independent predictors are not sufficient to model the trend. Dimensionless numbers like gravity number, capillary number and endpoint mobility ratio has different kind of weightage in reservoirs with different geometries and heterogeneities. It is necessary to add non linear terms to the model to capture these kind of interaction between the predictors.

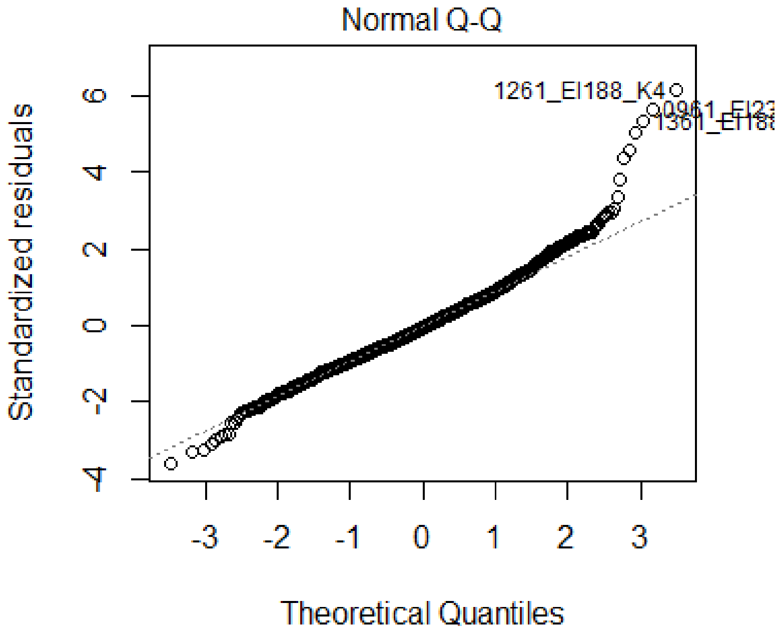

Figure 13: Normal Q-Q plot of residuals (MLR)

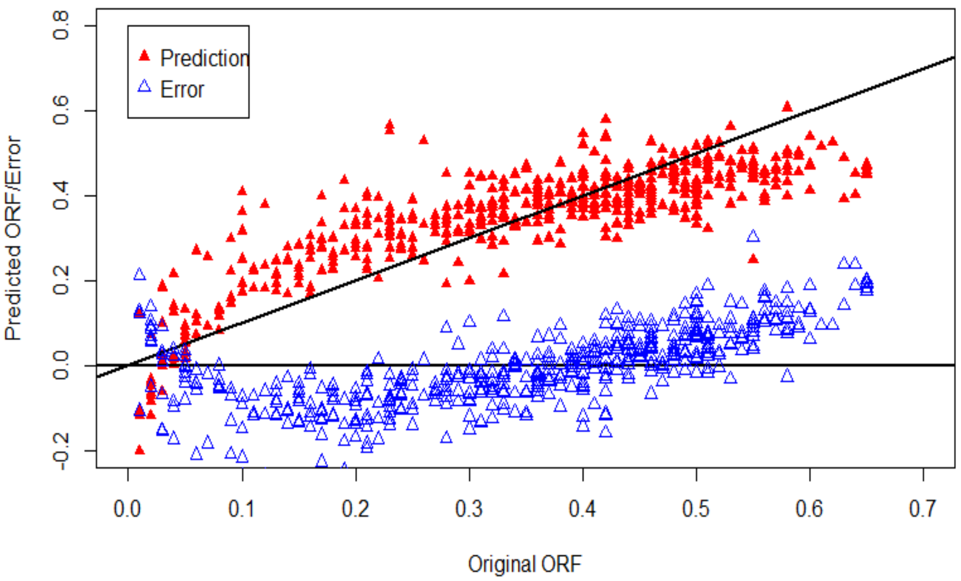

Figure 14: MLR- Prediction and Error vs Original ORF

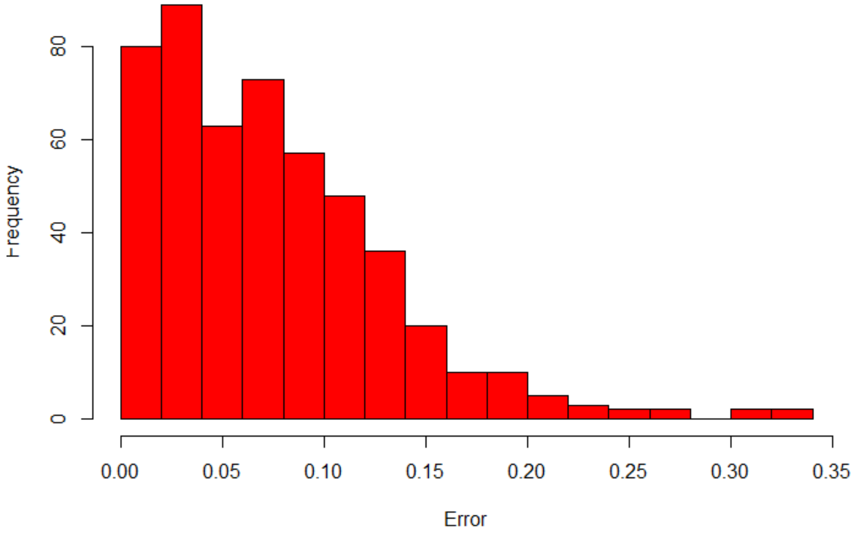

Figure 15: Distribution of absolute error on Test data (MLR)

Table 14 shows predictors related to one of the reservoirs in test data set which has high error of 0.53. It can been observed that high error may be due to extreme predictors which are not sufficient to charectrize the actual reservoir. It may also be due to inadequateness of the dimensionless parameters defined in the study

Predictor	Original value	Transformed Normalized value
Ng	9.7e+14	2.9
Npc	9.2e+10	2.9
Hetro_factor	7.1	2.6
Original ORF		0.36
Predicted ORF		-0.16

Table 14: Extreme Predictors of Reservoir "1361_EI188_N0"

4.1.2 Multiple linear regression with nonlinear terms

Figure 14 has indicated the presence of non linear interactions between the predictors. To capture nonlinear interactions between the predictors, multiple linear regression model with all possible interactions between the predictors is generated. **Figure 16** shows the relationship between predicted ORF and error vs original ORF with all possible non linear interactions. It can be observed that MLR with all possible non linear interactions can model the trend in the data. The trend in the residuals indicates that the model didn't left out signal in the dta. The RMSE and MAE of MLR with all possible interactions were 9% and 7.3% respectively.

With 30 predictors in initial data set, MLR with all possible interactions had $C_2^{30} + 30 = 245$ predictors. Among all possible interactions, predictors FSTRU: Mobility_Ratio_endpoint, FSTRU: Play_type, , FSTRU: CHRONOZONE2, FSTRU: N_ρ had statistical significance. Only these interactions were used to simplify the model. This is because only a few dimensionless numbers may have signigicance in a partciular geometry. For example, gravity number and buoyancy number will have significance in reservoirs with a significant dip. **Figure 17** shows relationship between predicted ORF and error vs original ORF for MLR with limited non linear interactions. Non linear trend in the plot indicates requirement of more number of predictors and feature selection methods.

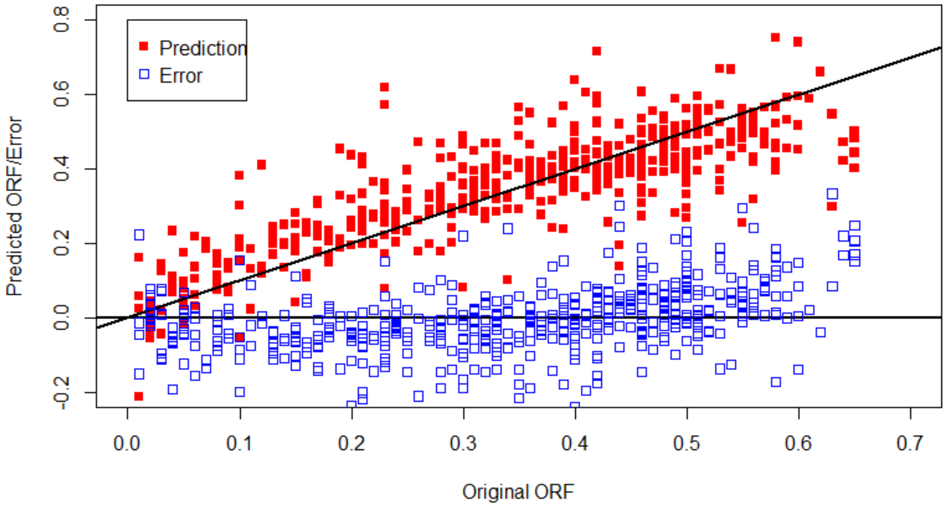

Figure 16: MLR (All Non linear predictors) – Diagnostics

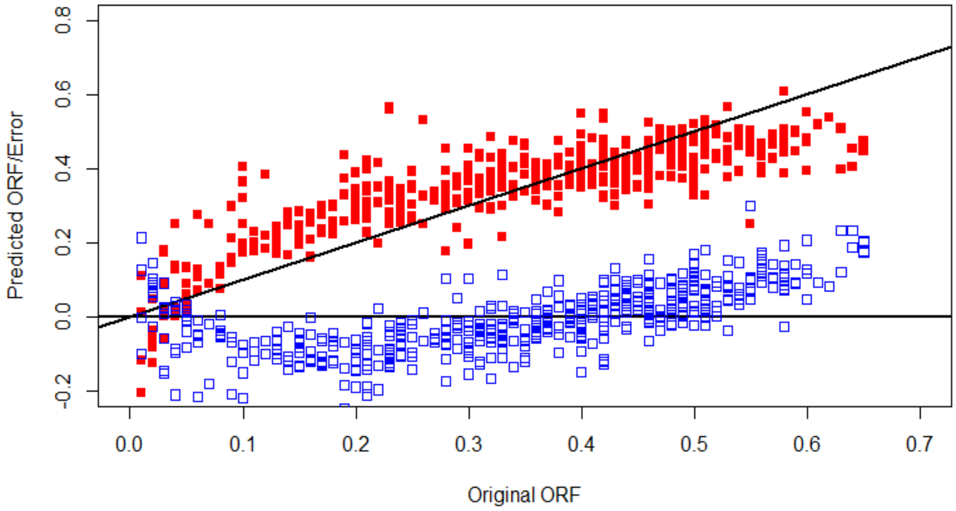

Figure 17: MLR (limited non linear predictors) – Diagnostics

4.2 Robust regression

As discussed in chapter-2.3 robust regression is relatively less sensitive to outliers due to its loss function. Depending on the tradeoff between bias and variance, loss function and tuning parameters can be selected. In this analysis Tukey Bi square M (Anderson 2008) estimation model is used.

Coefficients	Value	Std.Error	t value
Intercept	0.3583	0.0091	39.5004
FSTRU.A	-0.0108	0.0086	-1.2528
FSTRU.B	-0.0287	0.0101	-2.8319
FSTRU.C	0.0059	0.0084	0.7
FSTRU.D	-0.0299	0.0099	-3.0301
FSTRU.E	-0.0339	0.0112	-3.0283
PLAY_TYPE.A1	0.0333	0.0073	4.5922
PLAY_TYPE.F1	-0.0353	0.0065	-5.431
PLAY_TYPE.F2	-0.0472	0.0104	-4.5488
CHRONOZONE2.MIOCENE	-0.0056	0.0057	-0.9711
CHRONOZONE2.PLEISTOCENE	-0.0002	0.0056	-0.0321
DRIVE.COM	-0.0165	0.0078	-2.1276
DRIVE.DEP	-0.0348	0.0085	-4.1124
DRIVE.PAR	0.0001	0.0048	0.0233
RES_TYPE.N	0.0212	0.0119	1.7765
RES_TYPE.S	0.0055	0.0074	0.7372
POROSITY	-0.0119	0.0031	-3.8792
SW	0.0084	0.0054	1.5507
BHCOMP	-0.092	0.0054	-16.9382
Mobility_Ratio_endpoint	0.032	0.0036	8.9234
Nalpha	-0.0751	0.0062	-12.1536
Np	0.0019	0.0035	0.5453
Ng	-0.1845	0.0132	-14.0248
Npc	-0.0889	0.012	-7.3804
Dev_factor	0.1675	0.006	28.1273
Hetro_factor	-0.0197	0.0034	-5.8667

Table 15: Coefficients of predictors for Robust Regression

Table 15 and **Table 16** shows the coefficients of predictors and model statistics of Robust regression respectively. **Figure 18** shows the distribution of absolute error of test data using robust regression model. It can be observed that the spread of the error is wide with respect to MLR model due to higher bias than MLR model.

Rediddual standard error on 1996 degrees of freedom	0.085
RMSE on test data	9.3%
MAE on test data	7.2%

Table 16: Robust Regression Model Statistics

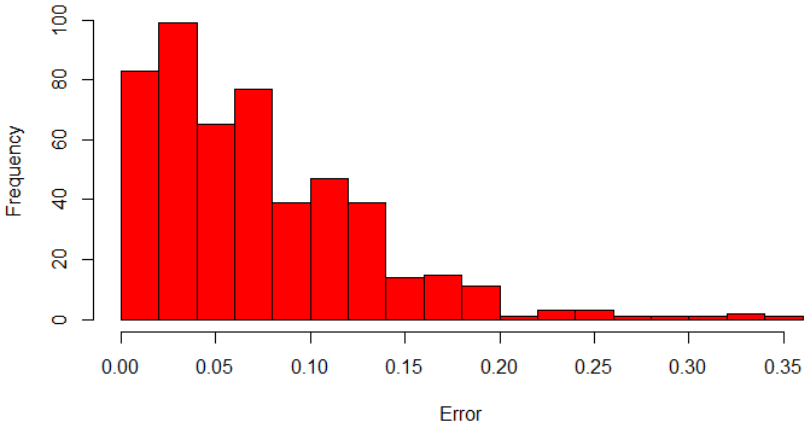

Figure 18: Distribution of Absolute Error (Robust Regression)

4.3 Penalized Regression model (LASSO)

As described in capter-2.3 Least absolute shrinkage and selection operator(LASSO) is capable of automatically discarding highly correlated predictors. **Figure 19** shows the relationship between λ, number of predictors and model accuracy. It can be observed that model accuracy with 23 predictors is similar to model accuracy with all the predictors. **Figure 20** shows the path of coefficients with increasing λ value. The λ value can be changed to select optimum number of predictors and model accuracy. **Table 17** and **Table 18** shows the predictor coefficients and model statistics for λ values of 0.0098 and 0.014 respectively. It can be observed that interpretability of the model increased with small compormize of 0.5 % (MAE) in model accuracy.

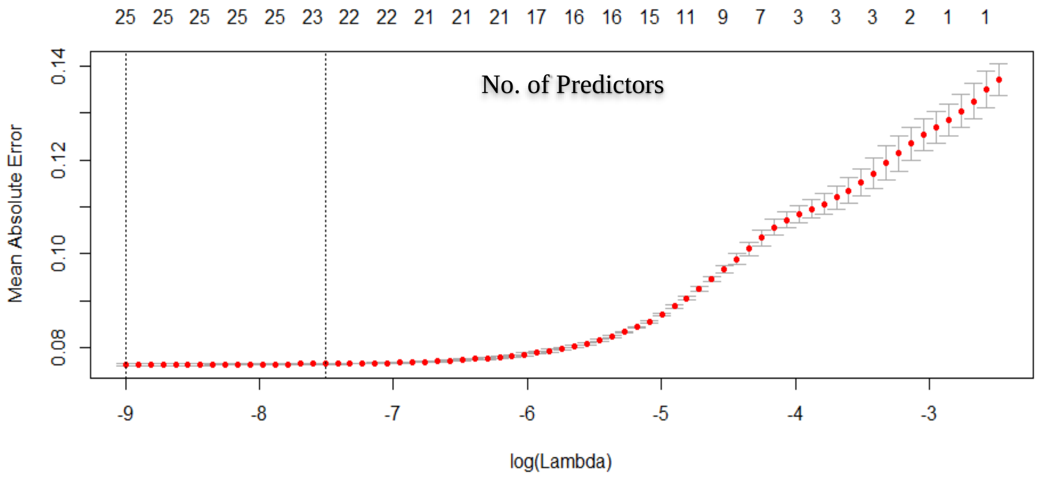

Figure 19: Lambda vs No. of Predictors and Model Accuaracy

Figure 21 shows the relationship between predicted ORF, Error and original ORF. Similar to MLR model non linear trend can be observed which indicates non linear effect of predictors on target variable. This is because except feature selection LASSO acts similar to MLR and cannot model non linear interactions between predictors implicitly.

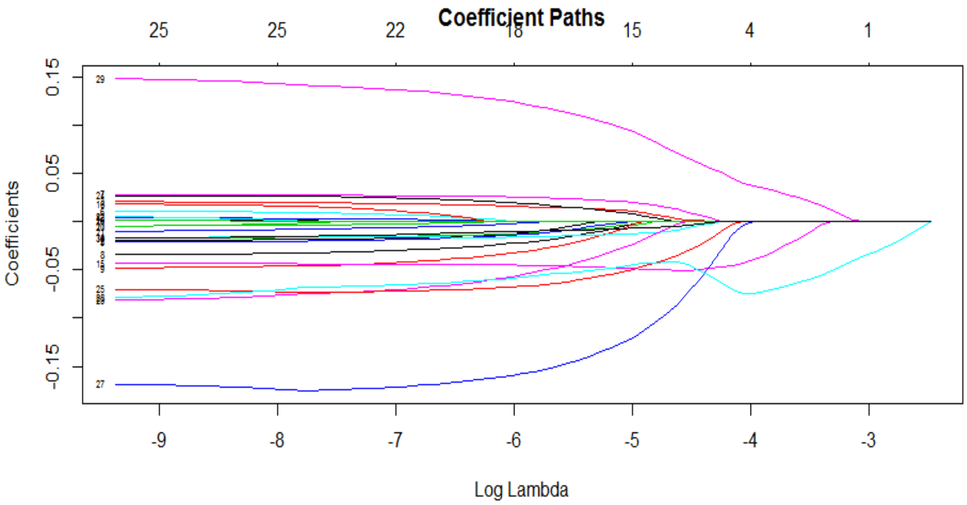

Figure 20: Coefficient path vs value of λ

Coefficient	Value
(Intercept)	0.346
Dev_factor	0.071
PLAY_TYPE.A1	0.013
FSTRU.C	0.003
Mobility_Ratio_endpoint	0.000
BHCOMP	-0.002
Hetro_factor	-0.005
DRIVE.COM	-0.008
Nalpha	-0.036
Npc	-0.041
DRIVE.DEP	-0.051
Ng	-0.086
No.of Predictors	11
λ	0.0098
MAE	9.5%
RMSE	11.0%

Table 17: Model Parameters λ=0.0098 (11 Predictors)

Predictor	**Value**
(Intercept)	0.347
Dev_factor	0.051
PLAY_TYPE.A1	0.000
FSTRU.C	0.000
Nalpha	-0.011
Ng	-0.030
DRIVE.DEP	-0.047
Npc	-0.063
No. of Predictors	7
λ	0.014
MAE	10%
RMSE	12.1%

Table 18: Model Parameters λ=0.014 (7 Predictors)

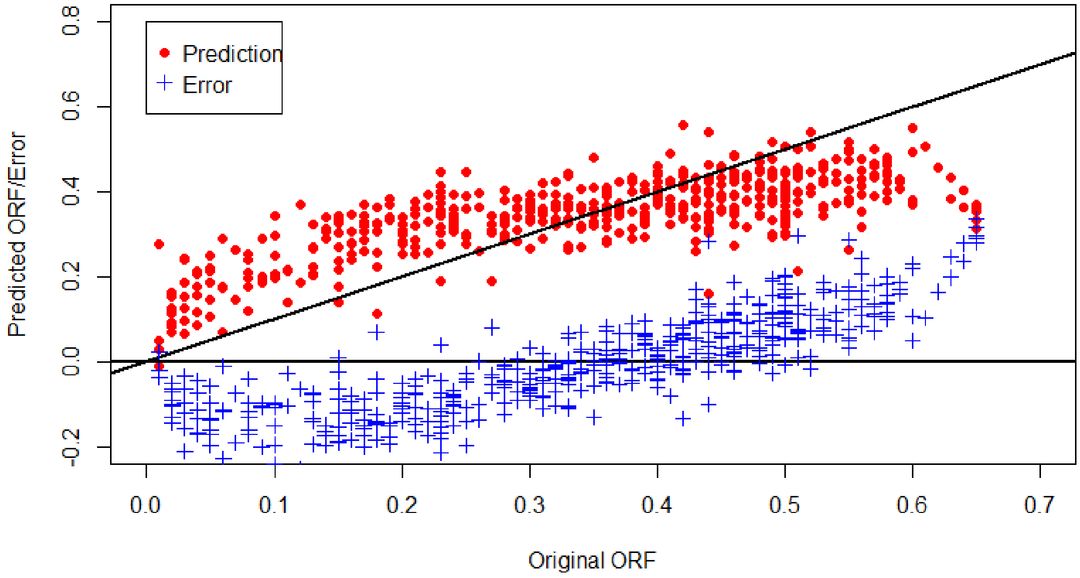

Figure 21: LASSO (11 Predictors) Prediction and Error vs Original ORF

4.4 K Nearest Neighbors

In k nearest neighbours model, k nearest reservoir instances are selected from training data based on distance metric (Eucledian distance in this study). ORF of new reservoir is predicted using the mean ORF of k nearest neighbors in the training data. **Figure 22** shows relationship between number of neighbours used for prediction and error on test dataset. It can be observed that optimum error rate is achieved by using 25 predictors for prediction. The RMSE and MAE for kNN model with 25 neighbours were 13.2% and 10.6% respectively.

Figure 23 shows that kNN model could not predict the trend. Except in the range of (0.3, 0.45) kNN model has significant error. **Figure 24** shows the histogram of absolute error of Knn. More than 50% of the predictions on test data had an error of greater than 10%. This could be due to giving equal importance to all the predictors in the data.

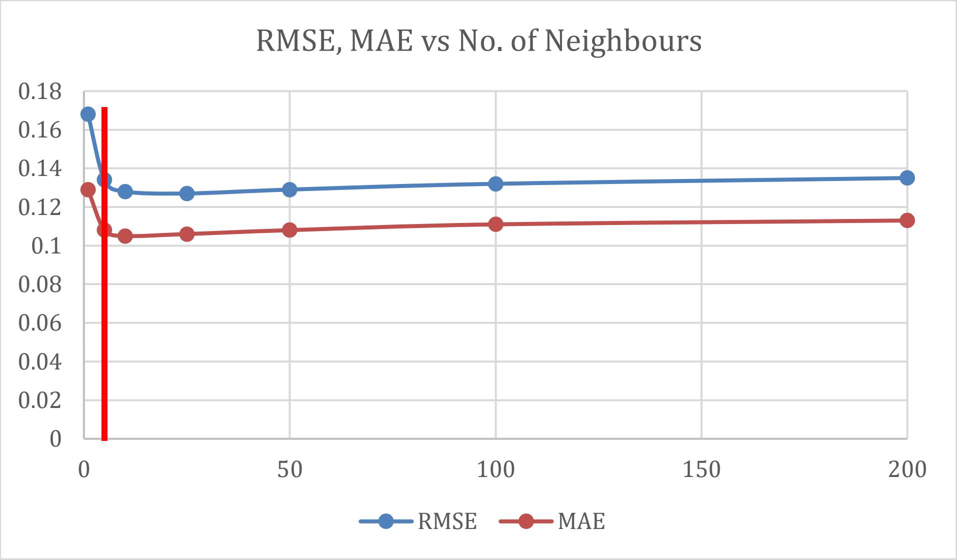

Figure 22: No. of Neighbours vs Accuracy plot

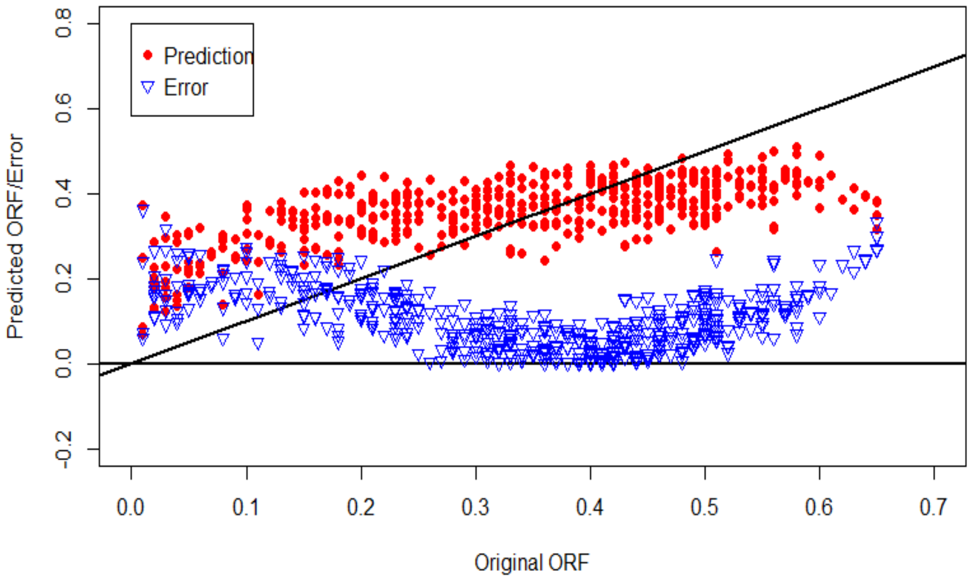

Figure 23: kNN (25 Neighbours) Prediction and Error vs Original ORF

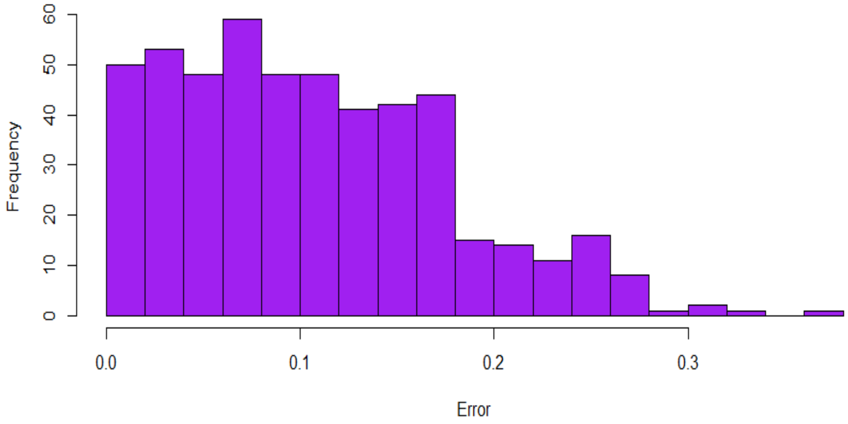

Figure 24: Distribution of Absolute Error (kNN-25 Neighbours)

4.5 Regression trees

As discussed in chapter 2.3.3, an extensive regression tree is built initially using a low cost-parameter (Cp). **Figure 25** shows the relationship between cross valided error, complexity parameter and size of the tree. The optimum tree size is 7 with a Complexity Parameter is 0.01. **Figure 26** and **Figure 27** shows regression trees of sizes 7 ($C_p = 0.01$) and 24 ($C_p = 0.005$) respectively. Even though these two trees are of different sizes, they have similar accuracy.

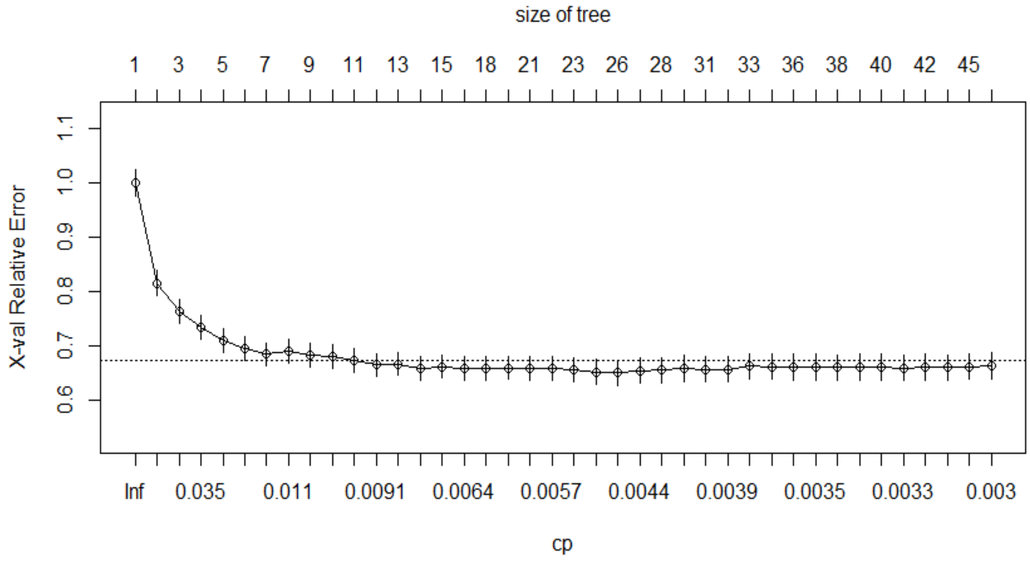

Figure 25: Change in C_p and Cross validated relative error with Tree Size

Figure 26 and Figure 27 depicts the way, regression tress make decisions to predict target variable. Various measures of central tendency or local regression model can be used to predict target variable from filtered ORF values after transversing through the tree.

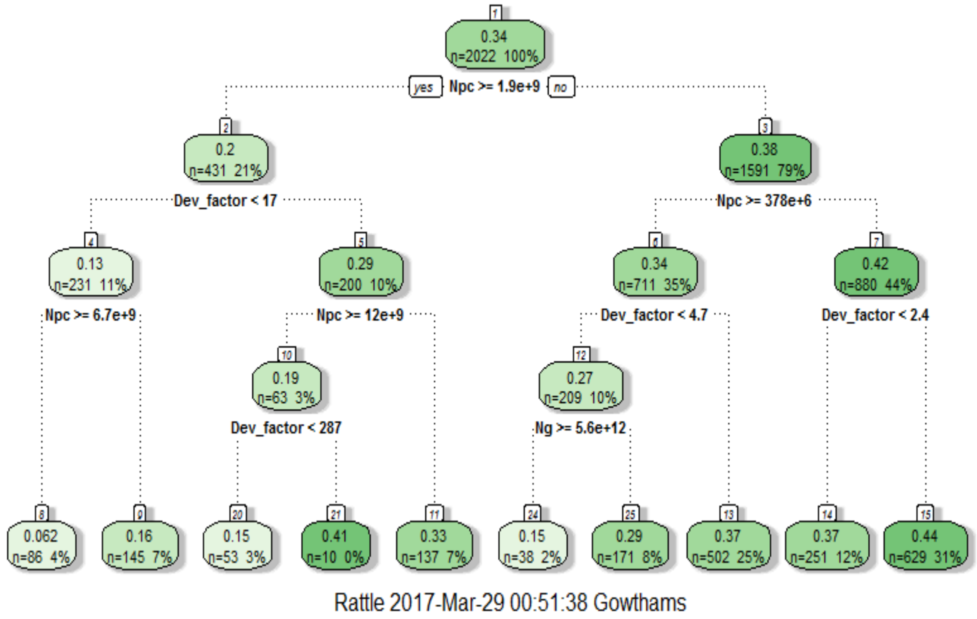

Figure 26: Regression tree of size 7, Complexity Parameter 0.01

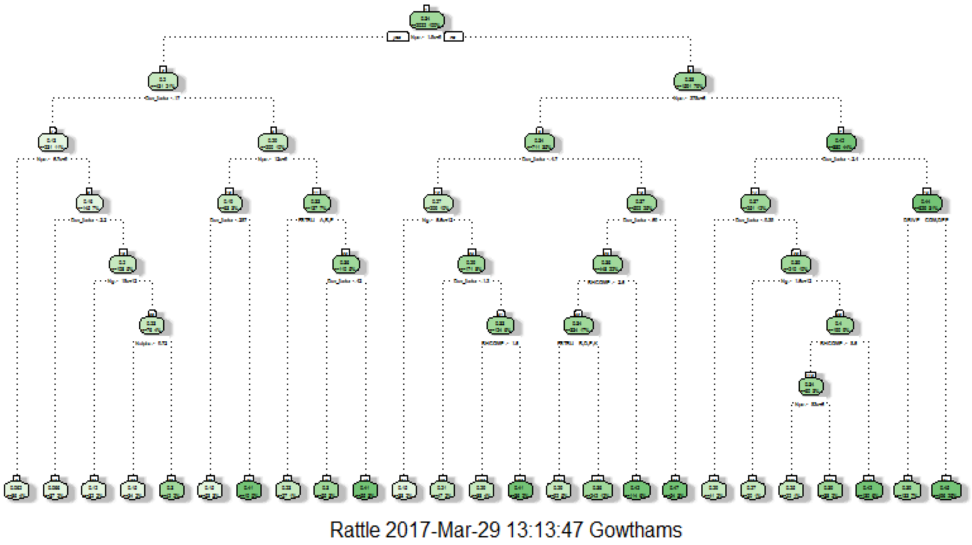

Figure 27: Regression Tree of size 24, Complexity Parameter 0.005

The RMSE and MAE for a tree size of 9 (No. of splits in the tree) is 13% and 10% respectively. Whereas, RMSE and MAE for a tree size of 24 is also 13% and 10% respectively. Figure 28 and Figure 29 shows the relationship between predicted ORF and error vs original ORF for tree size of 9 and 24 respectively. It can be observed that the number of levels in the prediction increases with increase in the size of regression tree and tends to follow unit slope line. **Figure 30** shows that around more than 50% of predictions on test data had an error of more than 10% ORF.

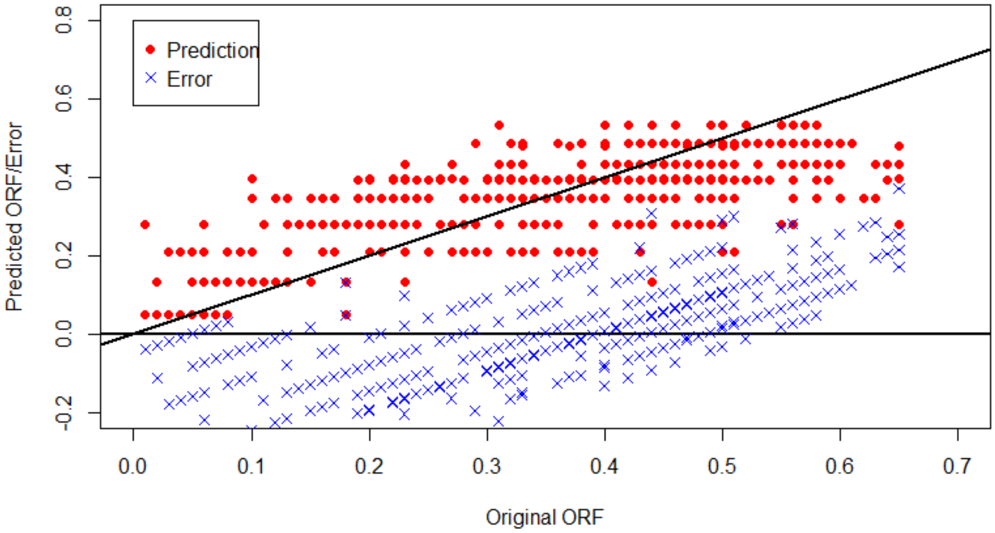

Figure 28: Predicted ORF and error vs original ORF (Tree size-9)

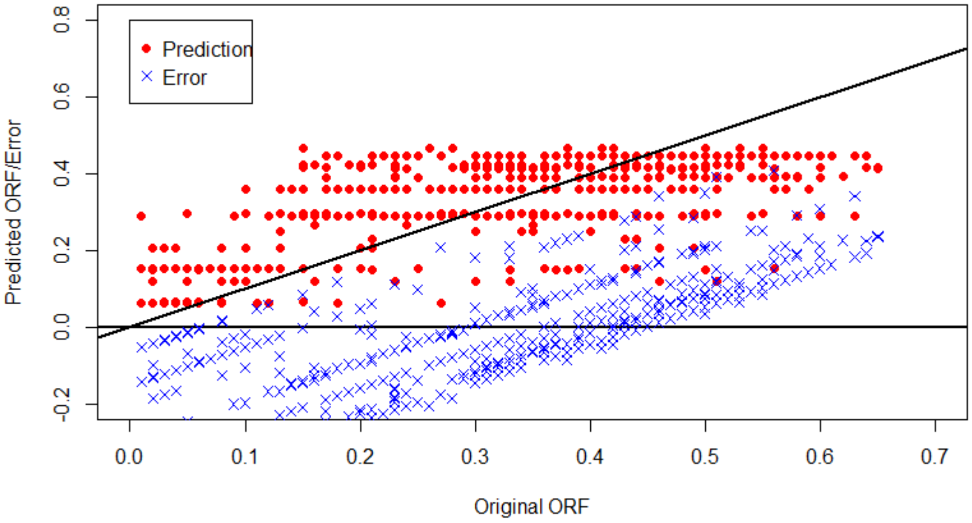

Figure 29: Predicted ORF and error vs original ORF (Tree size-24)

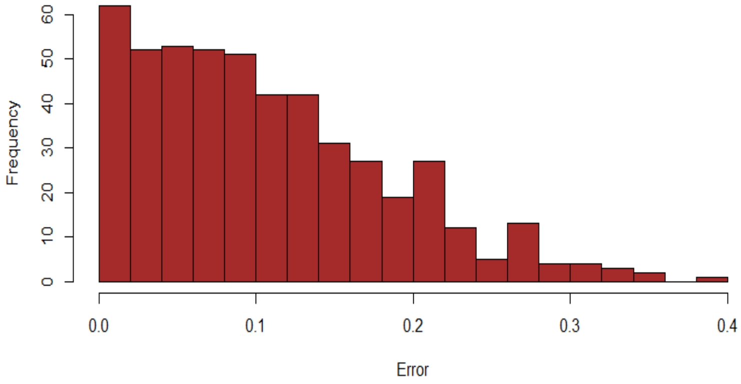

Figure 30: Distribution of absolute Error using regression tree (Size 9)

In addition to prediction, the influence of each predictor on the target variable can be estimated depending on the level of appearance, cleanness of the splits and number of splits connected to each predictor. If there are any duplicate variables, then they will share the variable importance. **Figure 31** shows the variable importance of various predictors used in the model. It can be observed that dimensionless parameters such as N_{pc}, N_g, R_l, N_α have high importance than other predictors like Porosity, No. of wells etc. It can also be observed that the dimensionless parameter defined for this study "Dev_factor" has 3rd highest importance after N_{pc} and N_g,

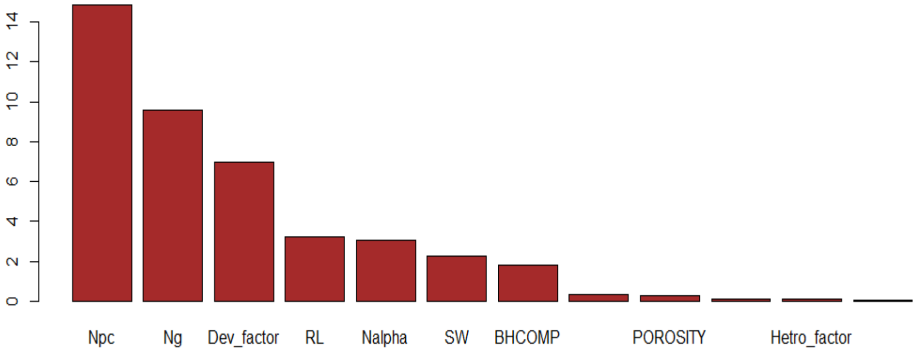

Figure 31: Predictor importance plot based on regression tree splits

4.6 Random Forest Model

As seen in the previous section, regression trees are rigid and its output space is limited. A small change or addition in the data may change the tree model. The predictions may change with the addition of new data. To address this issue, Random forests model uses large number of trees with splits based on randomly selected predictors. **Figure 32** shows the trend of RMSE using Random forests model with increasing number of trees. The RMSE over test data flattens off at a optimum number of trees and wont change with further increase of trees indicaing the robustness of random forests to overfitting. In this research, random forests model with 1000 trees is generated with each split in each tree is based on randomly selected 3 predictors. **Table 19** shows that model statistics of random forest model is similar to MLR. **Figure 33** shows relationship between predicted ORF and error vs original ORF. Even though there is no nonlinear trend in the residuals, random forest model has left away some signal. **Figure 34** shows that around 40% of test data has residuals of more than 10% ORF.

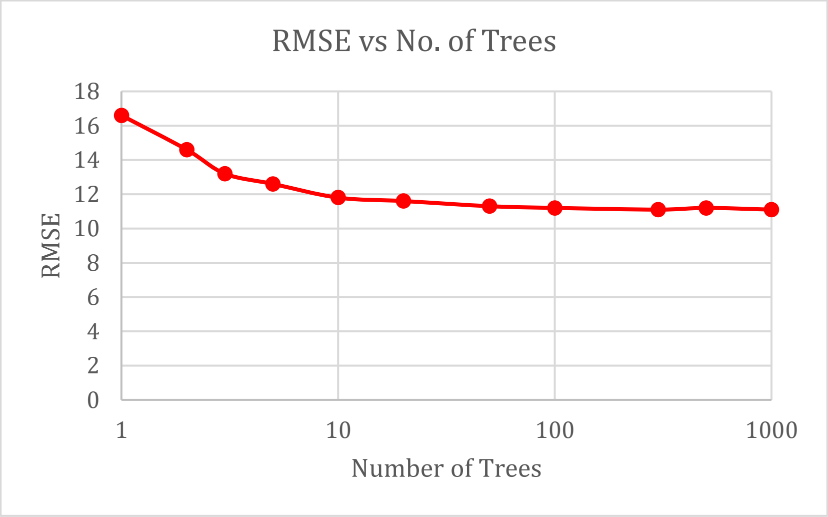

Figure 32: Decrease in RMSE with No. of trees in Random Forest model

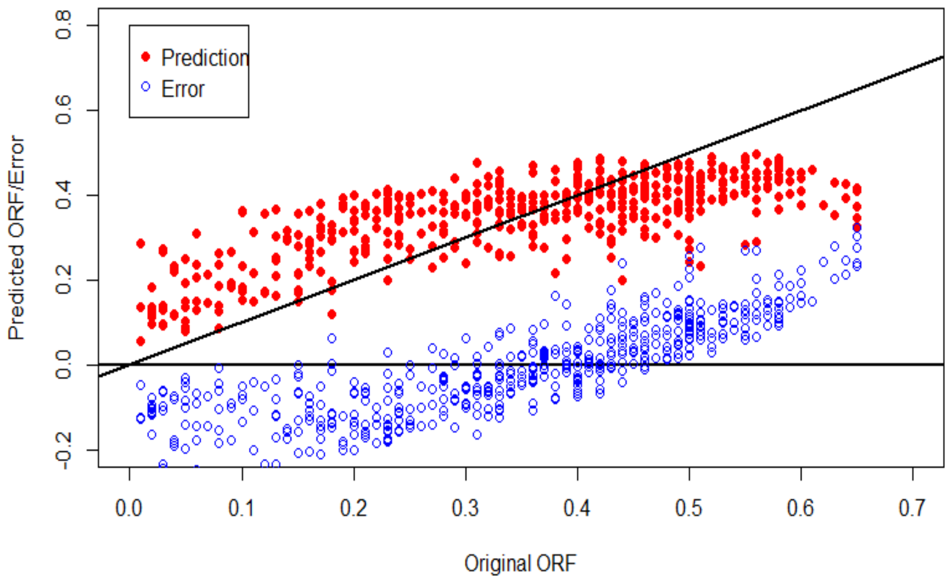

Figure 33: Random Forest Model - Prediction and Error vs Original ORF

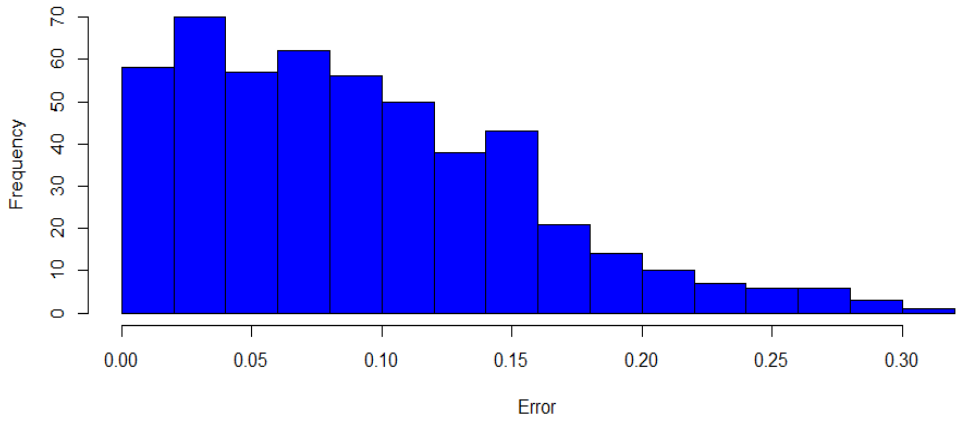

Figure 34: Distribution of Absolute Error (Random Forest)

No of Trees	1000
No of random variables at each split	3
RMSE on test data	11.1%
MAE on test data	9.1%

Table 19: Model Statistics (Random Forest)

4.7 Artificial Neural Networks

Predictive models discussed in previous chapters except linear regression with non linear interactions upto some extent have limitations in modelling nonlinear interactions between the predictors. Figure 29 and Figure 33 shows that random forests and decision tree models were not able depict non linear trend even though they can allow nonlinear interactions between predictors up to some extent. In this section artificial neural networks were used to predict ORF by incorporating non linear interactions between predictor variables. Artificial neural networks of various dimensions were analyzed and the best possible combination of hidden layers and nodes (1 hidden layer with 3 nodes) is selected. Simple summation is used as the transformation function at hidden nodes. **Figure 35** shows the relationship between predicted ORF and error vs original ORF. The residuals are more close to zero line than other models discussed so far indicating better performance of ANN.

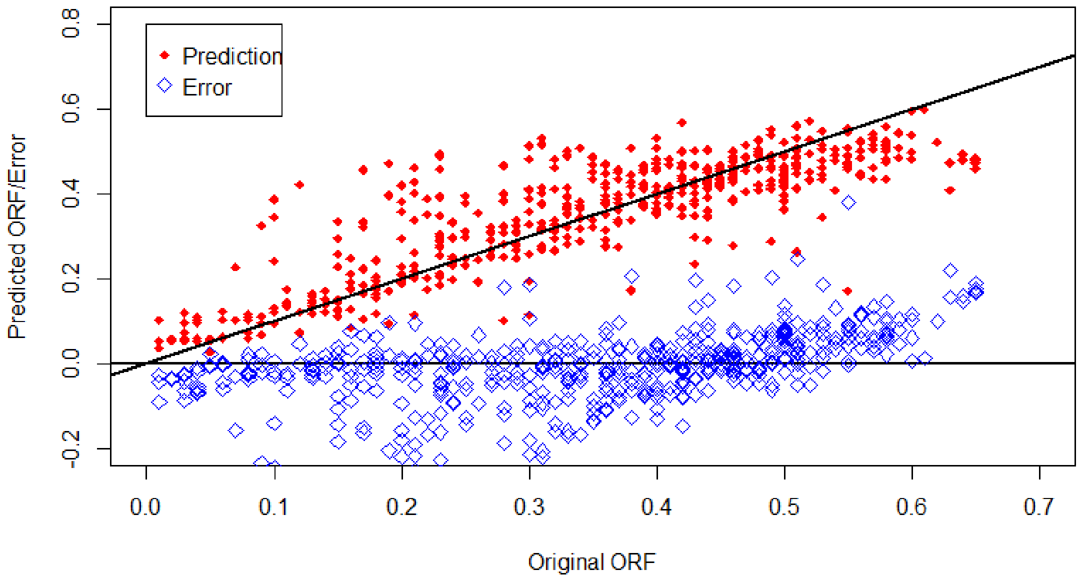

Figure 35: Artificial Neural Network - Prediction and Error vs Original ORF

Models with non linear interactions like Regression trees and Random forests have left out some signal. But ANN was able to model the data satisfactorily leaving the noise away. **Figure 36** shows that around 80% of the predictions on test data had an error of less than 8%. **Table 20** shows the model statistics of ANN model. It can be observed that ANN model has smaller error than models discussed so far.

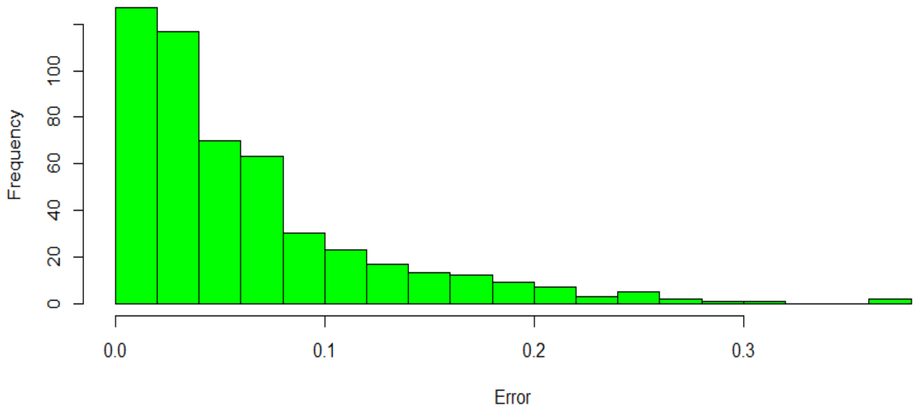

Figure 36: Distribution of Absolute error (ANN)

No. of Hidden Layers	1
No. of Nodes in hidden layers	3
RMSE	8.5%
MAE	6.0%

Table 20: ANN model Statistics

4.8 Ensemble Modelling

Few model predictions are good in a specific range of ORF whereas a few other models prediction is good in different range of ORF. **Figure 37** shows the distribution of error vs original ORF for different modelling techniques. ANN predicted better than Random forest at low ORF levels and random forests predicted better than ANN at high ORF levels. MLR had high error at low ORF and similar erro as ANN and random forest in high ORF level. Ensemble models combine various predictive models to leverage the strength of each model. There are different types of ensembling techniques based on the target variable to provide better and stable predictions. The aggregate of all the models will be less noisy than a single model.

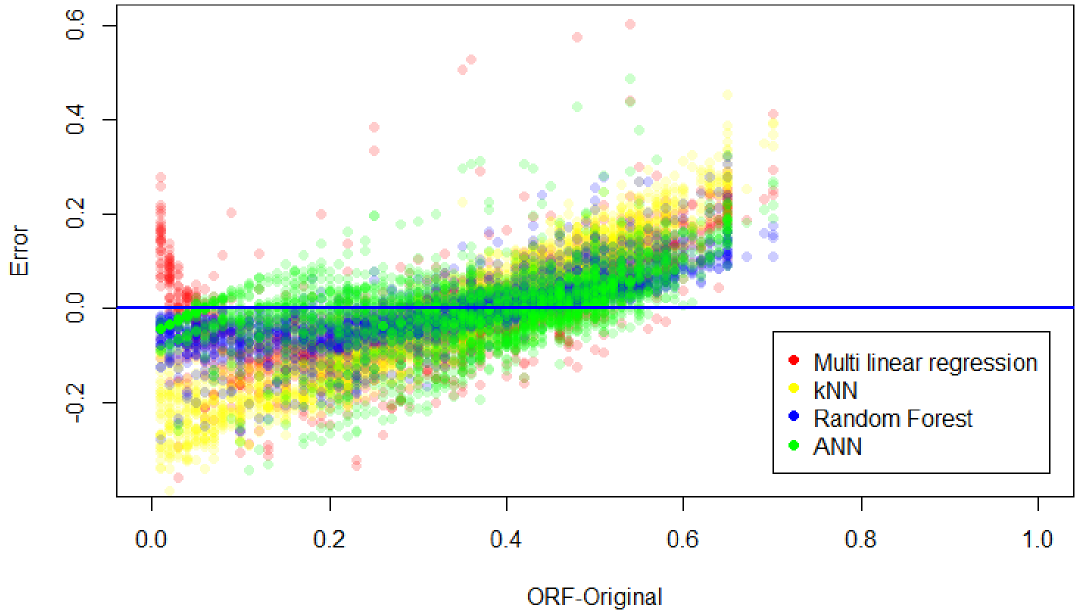

Figure 37: Distribution of Error vs Original ORF for different models

The models used in the analysis Simple linear regression, Random forest and Artificial neural network are combined to model more robust model. kNN is not selected because of its high error rate. Robust regression and LASSO are not selected because they are similar to simple linear regression. One of the simplest ensemble model is averaging the predictions from all the models. Sometimes weighted average can also be selected. **Table 21** shows the error metrics of average ensemble model based on Multiple linear regression, Random Forest and Artificial neural network. Even though the error is slightly higher than the ANN model, the model will be robust and more accurate on new data than a single model.

Models taken for averaging	MLR Random Forest ANN
RMSE	8.8%
MAE	6.8%

Table 21: Average Ensemble model statistics

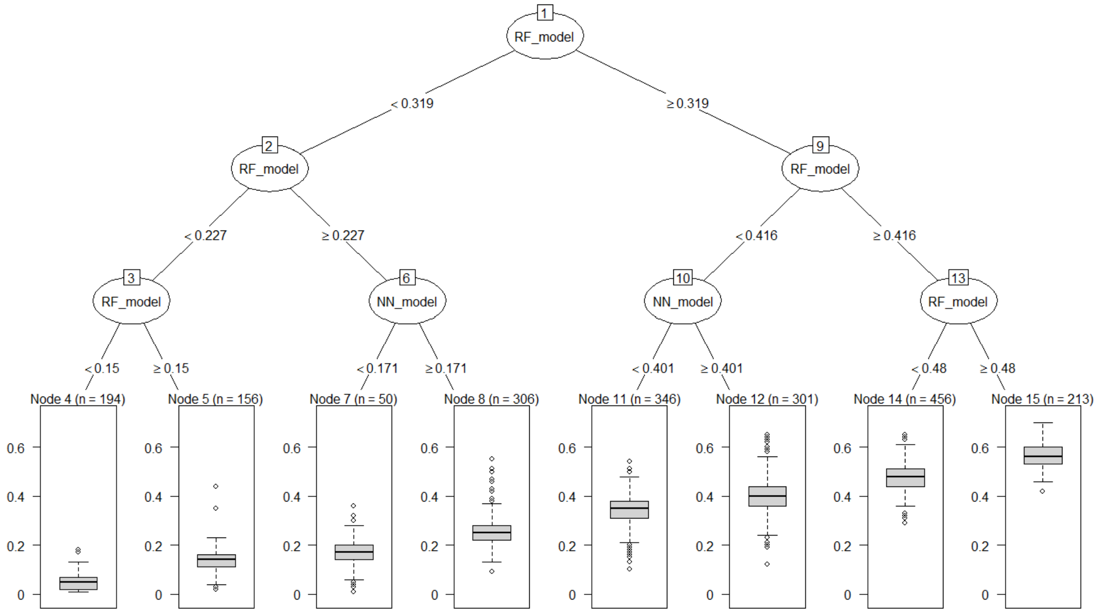

Figure 38: Decision tree Ensemble model for selecting model based on predicted ORF

Another kind of ensemble is using decision trees. As discussed earlier a few models work better in a specific range of ORF whereas other models work better in a different range of ORF. Decision trees as an ensemble technique will help in selecting different models at different ranges of ORF. **Figure 38** depicts the way decision tree selects predicted ORF from various models based on their predicted ORF. It can be observed that ANN model is selected between the predicted ORF range of (0.15, 0.45) and Random Forests is selected in the extremes. MLR model is not selected because of the better performance of Random forest and ANN. **Figure 39** shows the relationship between predicted ORF and error vs original ORF of ensemble model. Even thought the number of output levels is limited, they followed the unit slope line and residuals followed unitslope line.**Table 22** shows that ensemble model based on decision tree has a minimum MAE of 4.6%. **Figure 40** shows that more than 70% of the predictins on test data had an error of less than 5%.

Models taken for ensemble	MLR, Random Forest, ANN
RMSE	6.3%
MAE	4.6%

Table 22: Decision tree Ensemble model Statistics

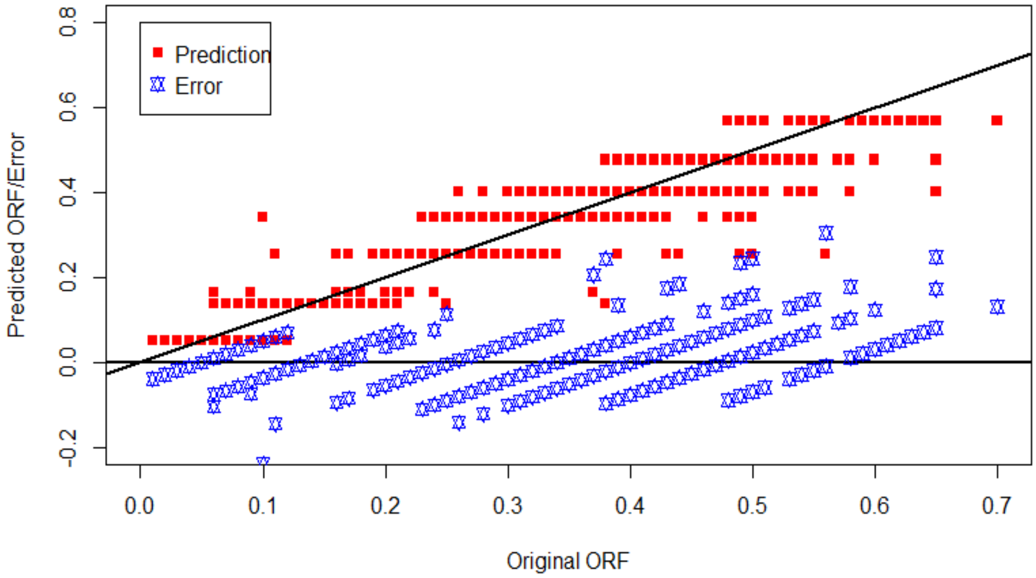

Figure 39: Final Ensemble model - Prediction and Error vs Original ORF

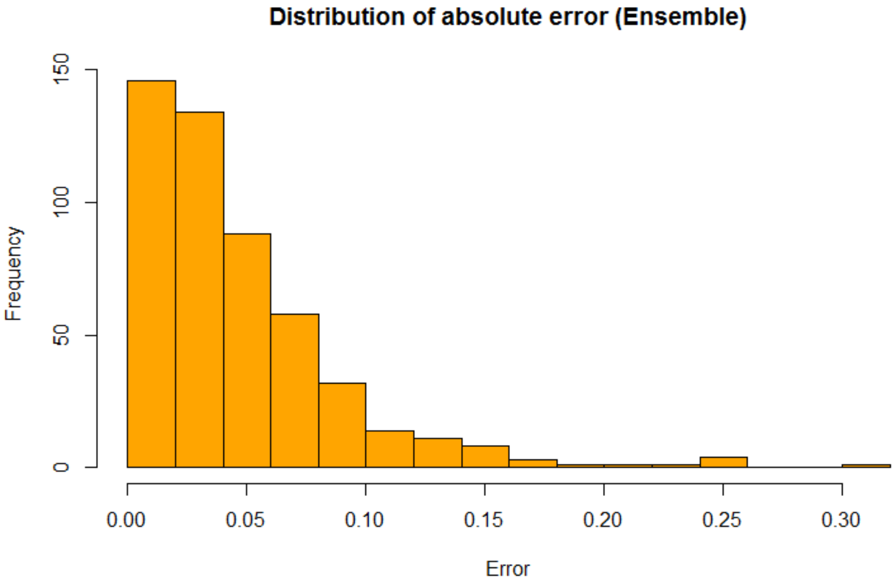

Figure 40: Distribution of Absolute Error (Ensemble model)

4.9 Sensitivity Analysis

The response of the data analytical model to changes in various predictors such as porosity, permeability and number of wells was analyzed using Random Forest model. Random forest model is selected because it does not need any transformation of the data. This model was able to successfully capture the natural tendency of the reservoirs up to some extent. **Figure 41** shows the effect of change in number of wells on ultimate recovery factor for a few sands considered in the study. It can be observed that in all the reservoirs, ultimate recovery factor increased up to some extent and flattened after that. This indicates that the model has captured the natural tendency of the reservoirs.

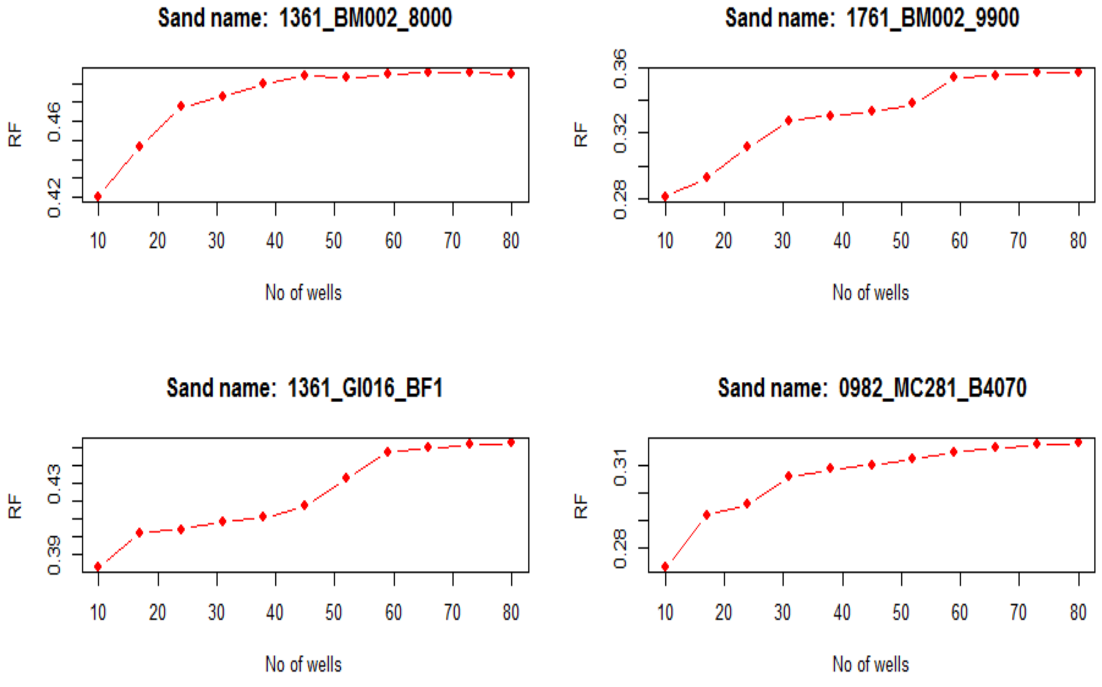

Figure 41: Effect of No. of wells on ORF

Figure 42 shows the change in ultimate recovery factor with change in porosity and keeping all other parameters same. It can be observed that ultimate recovery factor remained constant upto some extent and increased after a certain threshold. It can also be observed that ultimate recovery factor decreased after a certain extent of porosity which may be due presence of unconsolidated reservoirs in the data set. This is also possible if reservoirs with high porosity in the dataset had low ultimate recovery factor due to any other reason. But on overall, random forest model captured the dependence of ultimate recovery factor on Porosity of the reservoirs. **Figure 43** shows the effect of change in

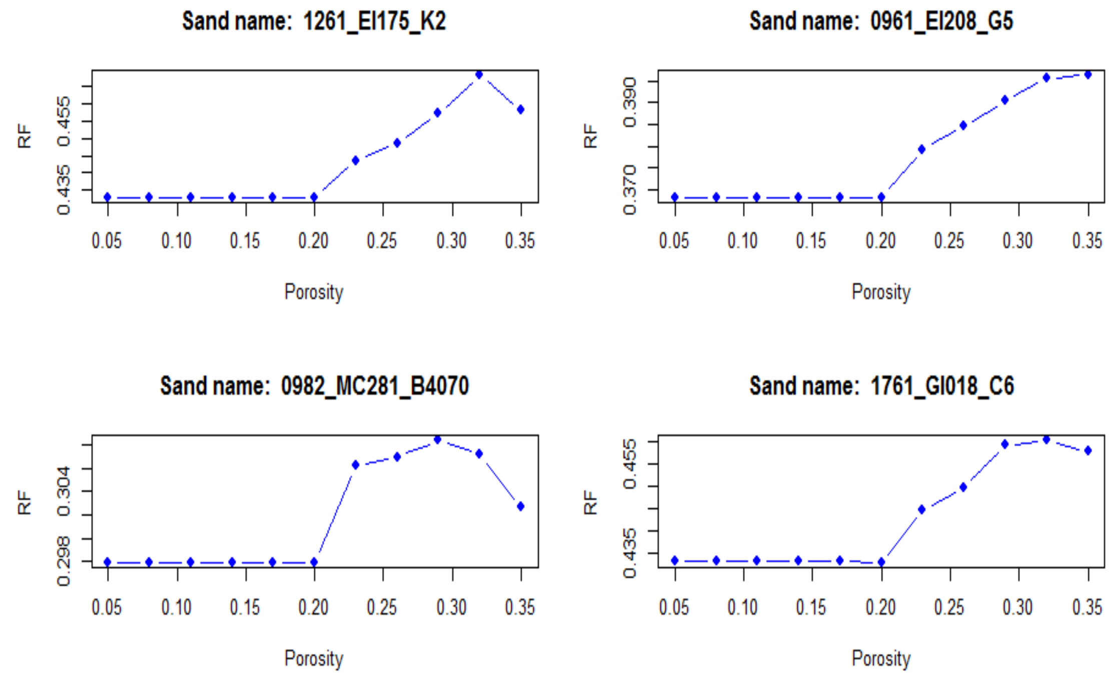

Figure 42: Effect of Porosity on ORF

permeability on ultimate recovery factor. It can be observed that ORF increases upto some extent and flattens out similar to the trend shown with number of wells. This kind of trend indicated the effectiveness of random forest model in capturing the relationship between premiability and ultimate oil recovery.

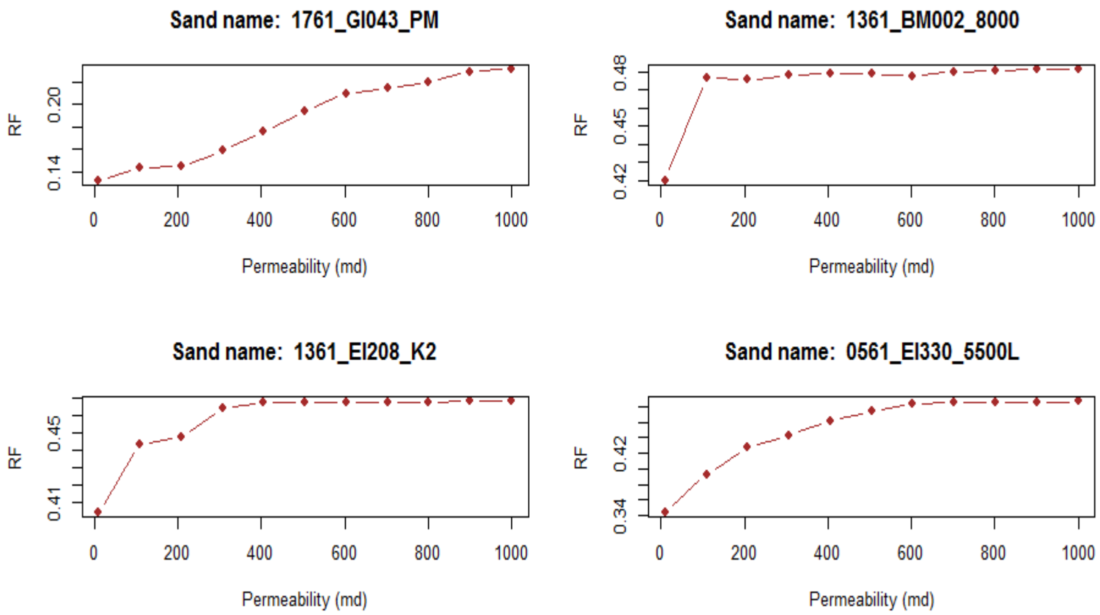

Figure 43: Effect of Permeability on ORF

5. Conclusions and Recommendations

5.1 Summary of the work

The main objective of this study is to predict the ultimate recovery factor of oil reservoirs using various data analytics techniques. In addition to that the sensitivity of the models to changes in predictor variables such as porosity, permeability and number of wells is also studied. Various data analytical models were used to predict ultimate recovery factor of oil reservoirs in Gulf of Mexico. Error metrics such as Root mean square error (RMSE) and Mean absolute error (MAE) were used on test data set to evaluate the analytical models used in the study. **Figure 44** shows the distribution of error for various models. It can be observed that Random forest model and Artificial neural network model have better predictions than other models. An Ensemble model is made using Multiple linear regression, Random forest model and Artificial neural network to take advantage of strengths of these models. This resulted in better RMSE and MAE on prediction over test data. **Table 23** shows the RMSE and MAE for various data analytical models used in this study.

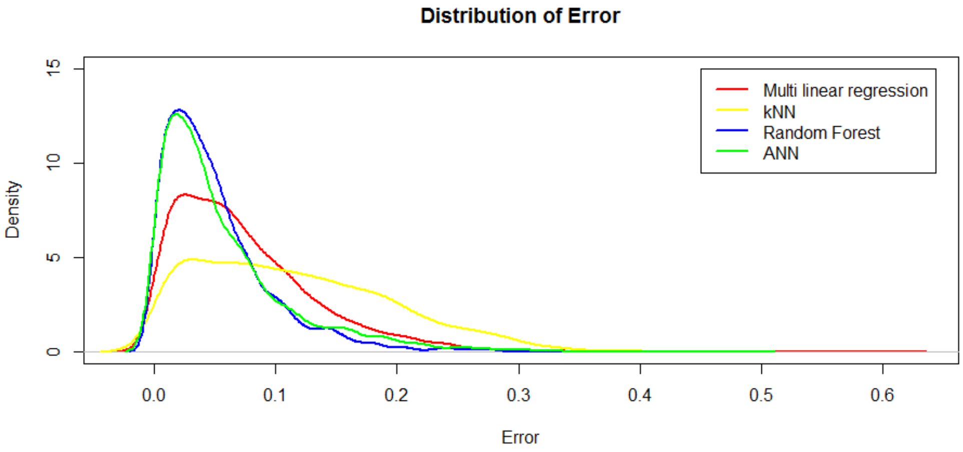

Figure 44: Distribution of absolute error for various models

Model	RMSE	MAE
Multiple linear Regression	9.0%	7.0%
Robust linear Regression	9.3%	7.2%
LASSO	11.0%	9.5%
K nearest neighbors	12.7%	10.6%
Decision Tree	13.2%	10.6%
Random Forest	11.1%	9.0%
Artificial Neural Network	8.5%	6.0%
Ensemble Model	6.3%	4.6%

Table 23: Summary Statistics of various models

5.2 Conclusions

Conclusions of the following study were as follows

1. Two new dimensionless numbers defined in this study characterized the field development and heterogeneity of the reservoirs. These two along with other dimensionless numbers can be used to reduce the number of predictors required to scale reservoirs for predicting ultimate recovery. Each of these dimensionless numbers have physical meaning which helps the model to conncest physical processes in oil reservoirs with statistical modelling techniques.
2. The trend of residuals in MLR model with idependent predictors indicated presence of non linear interactions between the predictors. Various models which can model non linear interaction were tried to predict ORF. MLR with all possible non linear predictors successfully captured the trend but it required 245 predicters as input. Inspite of number of predictors, MLR with non linear interactions has an advantage of easy interpretability.

3. Random forests model and ANN performed better than MLR in modelling the non linear interactions between predictors. ANN had better RMSE of 8.5% which is better than other individual models. In addition to prediction Random forest model captured the natural relationship between ultimate recovery factor and predictors like number of wells, porosity and permeability. Inspite of their ability to implicitly model non linear interactions between predictors these models are difficult to interpret.
4. Ensemble model used in the study selects models between MLR, Random forest and ANN at different ranges of ORF based on their prediction accuracy. This model further boosted the RMSE of prediction to 6.3%. Similar to Random Forest and ANN, this model is also difficult to interpret.
5. It is necessary to make trade off between between interpretability and modelling capability of data analytical models. Based on the task and requirement of interpretability and model accuracy, these models can be selected and used to predict the ultimate recovery factor of new reservoirs.

5.2 Scope of future work

It has been observed that a few models predicted negative values of ORF due to extreme values in predictors. One more limitation of these models is they will not replicate real conditions in case of extreme inputs. For example, these models may still predict positive ORF in cases with zero number of wells, zero permeability, zero porosity etc. Mathematical models with constraints can be used to model relationship between dimensionless parameters and ultimate recovery factor. The parameters of mathematical models can be estimated using maximum likelihood estimation. Simple mathematical model for predicting ultimate recovery factor can be defined as follows.

$$ORF = (b_1x_1 + b_2x_2 + b_3x_3 + \cdots)\,(1 - \ e^{a\,\Phi})(1 - e^{bBHCOMP})(1 - e^{c\,Dev_{factor}}) \qquad (5 - 1)$$

x_1, x_2, x_3 ... are predictors which does not have extreme effect on the target variable. Whereas, ORF should be zero when any of the parameters Φ, BHCOMP and Dev_factor is zero.

Therefore, the following work would be recommended to do in future.

1. Developing mathematical models as shown in Eqn. (5-1) and using the already available data to estimate the model parameters. This work is similar to relating decline curve parameters to petrophysical, fluid and development related properties of the reservoir using data analytics.
2. The dimensionless parameters in this study were defined based on the available data in the dataset. Development number needs to be further improved to account for development pace, stimulation techniques and artificial lift. Reservoir simulation studies can be used to further improve this expression to capture various development phenomenon.

References

Abou-Sayed, Ahmed. 2012. "Data mining applications in the oil and gas industry." Journal of Petroleum Technology 64 (10):88-95.

Arps, JJ, and TG Roberts. 1955. "The effect of the relative permeability ratio, the oil gravity, and the solution gas-oil ratio on the primary recovery from a depletion type reservoir." Trans. AIME 204:120-127.

Alabboodi, M. J., & Mohaghegh, S. D. (2016, September). Conditioning the Estimating Ultimate Recovery of Shale Wells to Reservoir and Completion Parameters. In SPE Eastern Regional Meeting. Society of Petroleum Engineers.

Andersen, R. (2008). Modern methods for robust regression (No. 152). Sage.

Asadollahi, R. (2014). Predict the flow of well fluids: a big data approach (Master's thesis, University of Stavanger, Norway).

Bishop C (2006). Pattern Recognition and Machine Learning. Springer.

Breiman L (2001). “Random Forests.” Machine learning, 45, 5-32.

Breiman L, Friedman J, Olshen R, Stone C (1984). Classification and Regression Trees. Chapman and Hall, New York.

Box G, Tidwell P (1962). “Transformation of the Independent Variables.” Technometrics, 4(4), 531–550.

Cervantes Bravo, R., Fernández, E., Jiménez Nieves, E., & Suma Tairo, G. (2015, October). A Data Mining Approach to Model Portfolios Oil Assets at High Risk. In OTC Brasil. Offshore Technology Conference.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

David, R. M. (2016, September). Approach towards Establishing Unified Petroleum Data Analytics Environment to Enable Data Driven Operations Decisions. In SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers.

Dimensionless Numbers with Data Mining Techniques. In SPE Intelligent Energy International Conference and Exhibition. Society of Petroleum Engineers.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1). Springer, Berlin: Springer series in statistics.

Graybill, Franklin A. "Theory and applications of the linear model." (1976).

Guthrie, R. K., & Greenberger, M. H. (1955, January). The use of multiple-correlation analyses for interpreting petroleum-engineering data. In Drilling and Production Practice. American Petroleum Institute.

Gulstad, R. L. (1995). The determination of hydrocarbon reservoir recovery factors by using modern multiple linear regression techniques (Doctoral dissertation, Texas Tech University).

Gowtham, T., Rouzbeh, G. M., Vamsi, K. B., & Srikanth, P. (2016, May). Possible Misinterpretations in Well Test Analysis Due to Unfiltered Tidal Signal. In SPE Western Regional Meeting. Society of Petroleum Engineers.

Grujic, O., Da Silva, C., & Caers, J. (2015, September). Functional approach to data mining, forecasting, and uncertainty quantification in unconventional reservoirs. In SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers.

Geertsma, J., Croes, G. A., & Schwarz, N. (1956). Theory of dimensionally scaled models of petroleum reservoirs. Trans. AIME, 207, 118-127.

Hoerl A (1970). “Ridge Regression: Biased Estimation for Nonorthogonal Problems.” Technometrics, 12(1), 55–67.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, *35*(1), 73-101.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26). New York: Springer.

Lake, L. W. (1989). Enhanced oil recovery.
Lee, B. H. (2015). Analyzing databases using data analytics (Doctoral dissertation).

Liu, Y. (2013). Interpreting Pressure and Flow Rate Data from Permanent Downhole Gauges Using Data Mining Approaches (Doctoral dissertation, STANFORD UNIVERSITY).

Li, D., & Lake, L. W. (1995). Scaling fluid flow through heterogeneous permeable media. SPE Advanced Technology Series, 3(01), 188-197.

Loh WY, Shih YS (1997). “Split Selection Methods for Classification Trees.”Statistica Sinica, 7, 815–840.

Mohaghegh, S. D., & Abdulla, F. A. S. (2014, October). Production Management Decision Analysis Using AI-Based Proxy Modeling of Reservoir Simulations–A Look-Back Case Study. In SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers.

Mohaghegh, S. D., Abdulla, F., Abdou, M., Gaskari, R., & Maysami, M. (2015, November). Smart Proxy: An Innovative Reservoir Management Tool; Case Study of a Giant Mature Oilfield in the UAE. In Abu Dhabi International Petroleum Exhibition and Conference. Society of Petroleum Engineers.

Novakovic, D. (2002). Numerical reservoir characterization using dimensionless scale numbers with application in upscaling (Doctoral dissertation, Louisiana State University).

Noureldien, D. M., & El-Banbi, A. H. (2015, September). Using Artificial Intelligence in Estimating Oil Recovery Factor. In SPE North Africa Technical Conference and Exhibition. Society of Petroleum Engineers.

Novakovic, Djuro. 2002. "Numerical reservoir characterization using dimensionless scale numbers with application in upscaling." Louisiana State University.
Ripley B (1995). “Statistical Ideas for Selecting Network Architectures.” Neural Networks: Artificial Intelligence and Industrial Applications,pp. 183–190.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation* (No. ICS-8506). California Univ San Diego La Jolla Inst for Cognitive Science.

Sakia, R. M. (1992). The Box-Cox transformation technique: a review. *The statistician*, 169-178.

Sharma, A., Srinivasan, S., & Lake, L. W. (2010, January). Classification of oil and gas reservoirs based on recovery factor: a data-mining approach. In SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers.

Shook, M., Li, D., & Lake, L. W. (1992). Scaling immiscible flow through permeable media by inspectional analysis. *IN SITU-NEW YORK-*, *16*, 311-311.

Srivastava, P., Wu, X., Amirlatifi, A., & Devegowda, D. (2016, September). Recovery Factor Prediction for Deepwater Gulf of Mexico Oilfields by Integration of

Soto, B., Wu, C. H., & Bubela, A. M. (1999, January). Infill Drilling Recovery Models for Carbonate Reservoirs-A Multiple Statistical, Non-Parametric Regression, and Neural Network Approach. In SPE Eastern Regional Conference and Exhibition. Society of Petroleum Engineers.

Strobl C, Boulesteix A, Zeileis A, Hothorn T (2007). “Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution.”BMC Bioinformatics, 8(1), 25.

Temizel, C., Energy, A., Aktas, S., Kirmaci, H., Susuz, O., Zhu, Y., ... & Tahir, S. (2016, September). Turning Data into Knowledge: Data-Driven Surveillance and Optimization in Mature Fields. In SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers.

Tibshirani R (1996). “Regression Shrinkage and Selection via the lasso.”Journal of the Royal Statistical Society Series B (Methodological), 58(1), 267–288.

Titterington M (2010). “Neural Networks.” Wiley Interdisciplinary Reviews: Computational Statistics, 2(1), 1–8.

Wood, D. J., Lake, L. W., Johns, R. T., & Nunez, V. (2006, January). A screening model for CO 2 flooding and storage in Gulf Coast reservoirs based on dimensionless groups. In SPE/DOE Symposium on Improved Oil Recovery. Society of Petroleum Engineers.

Appendix A: Nomenclature

API	Oil API gravity
ANN	Artificial Neural Network
BHCOMP	No. of completion in each sand (Wells + Additional completions using workover)
BOEM	Bureau of Ocean and Energy Management
CART	Classification and regression trees
Cp	Complexity Parameter
CHRONOZONE	PLU-LL Upper Pleistocene, PLM Middle Pleistocene, PLL Lower Pleistocene, PU Upper Pliocene, PL Lower Pliocene, MLU & MUU Upper Miocene, MUM &MMM Middle Miocene, MLM & MUL & MML Lower Miocene
Dev_factor	Dimensionless Development factor
DRIVE	Dominant drive mechanism DEP- Depletion, GCP- Gas cap drive, WTR- Water Drive COM-Combination drive, PAR- Partial water drive, UNK- Unknown
g	Acceleration due to gravity
GOR	Gas oil Ratio
H	Reservoir thickness
Hetro_factor	Dimensionless Heterogenity factor
FSTRU	Field Structure Code A- Anticline, B- Fault, C- Shallow Salt Diapir 4000 SS, D-Intermediate Salt Diapir 4-10,000 SS, K- Rollover into growth fault
K_x	Average horizontal permeability, md
K_Z	Vertical permeability of reservoir , md
K_{rw}	Relative permeability to water
L	Length of reservoir
LASSO	Least absolute shrinkage and selection operator
PLAY_TYPE	Type of Reservoir play
P_CUMOIL	Cumulative oil produced (Bbl)
P_RECOIL	Ultimate reserves (Bbl)
R^2	Coefficient of determination in multi linear regression

MAE	Mean Absolute Error
MLR	Multiple linear regression
N_g	Dimensionless gravity number
N_{pc}	Dimensionless capillary number
NTG	Net to Gross ratio
ORF	Ultimate oil recovery factor
OLS	Ordinary least squares
R_l	Dimensionless aspect ratio
RMSE	Root Mean Square Error
RES_TYPE	Reservoir Type U- Under saturated oil, S- Saturated Oil
SPGR	Specific gravity of oil
SD_TYPE	Type of reservoir 'O' – Oil, 'G'- Gas, 'B'- Both, Zero - Unknown
SSE	Sum of squared errors
S_w	Initial water saturation
U_t	Subsurface fluid velocity (oil+water) , ft./day
$\hat{y}_i$	Predicted value using model
λ_{r2}^{o}	Relative mobility of residual phase-2
ρ_l	Density of non-wetting liquid phase
N_ρ	Dimensionless density number
σ	Interfacial tension of hydrocarbon-water system
$\Delta\rho$	Density difference between oil-water
α	Dip angle
ϕ	Porosity

Appendix B: R packages used in the study

Package	**Purpose**
EnvStats	Box cox transformation
Rpart	Regression tree modelling
Partykit	Visualizing regression trees
Rattle	Visualizing regression trees
Glmnet	Least absolute shrinkage and selection operator
Flexclust	Adjusted box plot
randomforest	Random forst model
neuralnet	Artificial neural network
Nnet	
Readxl	Reading data from excel file
Class	K nearest neighbours
MASS	Robust regression
Hmisc	Combined histogram of dataset
E1071	For measuring skewness
Caret	Unified approach for various data models
Corrplot	Visualizing correlation between predictors

Appendix C: Datasets

Dataset	Dimensions	Discription
Rawdata	13289 x 82	Initial data set containing oil and gas reservoirs of Gulf of Mexico
rawDataComlCases	13288 x 82	Initial data set with only complete cases
oilRes	5019 x 82	Oil Reservoirs in the dataset which produced more than 80% of ulitimate reserves
oilResClean	3038 x 82	Oil Reservoirs after removing reservoirs with erroneous data
rCleanSelect, GoM_original	2524 x 18	Oil Reservoirs selected for the study with dimensionless parameters
GoMTrain_original	2022 x 18	Training dataset with original parameters
GoMTest_original	502 x 18	Test data set with original paramters
GoM_Processed	2524 x 31	Oil Resevoirs selected for the study with normalized paramters and dummy variables
GoMTrain_processed	2022 x 31	Training data with processed parameters and dummy variables
GoMTest_processed	502 x 31	Test data with processed parameters and dummy variables
GoM_proc_nooutliers	2423 x 31	Oil Reservoirs selected with normalized paramters and dummy variables without ouliers
GoMTrain_proc_nooutliers	1940 x 31	Training data with processed parameters, dummy variables and no outliers
GoMTest_proc_nooutliers	483 x 31	Test data with processed parameters, dummy variables and no ouliers