ANALYSIS OF DRILLING DATA AND ROP OPTIMIZATION USING ARTIFICIAL INTELLIGENCE TECHNIQUES WITH STATISTICAL REGRESSION COUPLING

A Thesis

Presented to

The Faculty of the Department

Of Petroleum Engineering

University of Houston

In partial fulfillment

Of the requirements for the degree

Master of Science

In Petroleum Engineering

By

Bharat Mantha

August 2016

ANALYSIS OF DRILLING DATA AND ROP OPTIMIZATION USING ARTIFICIAL INTELLIGENCE TECHNIQUES WITH STATISTICAL REGRESSION COUPLING

_____
BHARAT MANTHA

Approved:

_____
Chair of the Committee
Dr. Konstantinos Kostarelos,
Associate Professor,
Petroleum Engineering

Committee Members:

_____
Dr. Robello Samuel,
Adjunct Faculty - Lecturer
Petroleum Engineering

_____
Dr. Thomas K. Holley,
Professor and Interim Chair,
Petroleum Engineering

_____
Dr. Suresh K. Khator,
Associate Dean,
Cullen College of Engineering

_____
Dr. Thomas K. Holley,
Professor and Interim Chair,
Petroleum Engineering

ACKNOWLEDGMENTS

I thank my supervisor, Dr. Robello Samuel, for taking me on as his student and for his advice and understanding on both technical and personal issues throughout the duration of my degree. He has been a pillar of support and motivation during my entire study.

I thank Dr. Konstantinos Kostarelos for agreeing to chair the committee and Dr. Thomas K. Holley for agreeing to be a member of my committee.

I thank Aimee Taylor and Avinash Wesley for providing me with all the data required and helping with several aspects all along, right from my internship which formed the basis of my understanding of the problem at hand. I am equally grateful to Halliburton for providing me the initial opportunity to learn and apply.

I am grateful to the instructors at Coursera for providing me with a financial waiver for all the coursework I have finished on Machine Learning and Data Analytics.

Thanks also go to my friends and colleagues, and the departmental faculty and staff, for helping me to adjust to this environment. Finally, thanks to my family for their support, understanding, and love.

DEDICATION

This thesis is dedicated to the almighty God, and my beloved family members without whose support none of this would have been possible, and without whom, I would not have been. It is also dedicated to my fiancé, Maitreyi for her incredible help and motivation.

ANALYSIS OF DRILLING DATA AND ROP OPTIMIZATION USING ARTIFICIAL INTELLIGENCE TECHNIQUES WITH STATISTICAL REGRESSION COUPLING

An Abstract

Presented to

The Faculty of the Department

Of Petroleum Engineering

University of Houston

In partial fulfillment

Of the requirements for the degree

Master of Science

in Petroleum Engineering

By

Bharat Mantha

August 2016

ABSTRACT

Predictive data-driven analytics has the potential to successfully predict the downhole environment in Drilling Engineering. In general, rate of penetration (ROP) optimization involves adjustment of the weight on bit (WOB) and rotary speed (RPM) for efficient drilling. ROP has a complex relationship with several other parameters, such as formation properties, mud properties, mud hydraulics, borehole deviation, as well as the size/type of bit. In this study, a new workflow based on statistical regression and artificial intelligence (AI) techniques was designed to forward predict ROP using field data gathered from the North Sea horizontal wells. Several machine-learning models such as such as step-wise regression, neural networks, support vector regression, classification-regression trees, random forests, and boosting, were applied for prediction. A web based prediction app was developed that could perform predictive analytics and uncertainty analysis on any data. The app was further tested on other wells and was shown to predict with significant accuracy.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Problem Statement

Wellbore complexities can result in increased well costs. Consequently, it is now more important than ever to optimize and achieve the best drilling rate of penetration (ROP). While several techniques exist to accomplish this, each has its own merits and limitations, and there is no acceptable universal model for all conditions, as the nature of the relationships among these parameters can be complex. Usually, conventional methods fail to predict ROP accurately owing to additional complexities of the downhole conditions. In general, ROP optimization involves the adjustment of weight on bit (WOB), mud flow rate (Flow) and rotary speed (RPM) for efficient drilling. However, ROP has a complex relationship with several other parameters, such as formation properties, compressive strength, pressure gradient, mud properties, mud hydraulics, borehole deviation, and size and type of bit used. In several instances, increasing WOB and RPM results in a decreased ROP, as there is an interaction of these inputs with the formation properties and flow, clearly highlighting the underlying complex relationships among these parameters. To account for these uncertainties, data –driven analytics can be used effectively to better understand ROP optimization. Traditional regression analysis models have limitations and failed to describe the dependence of one observed quantity on another observed quantity. On the other hand, the artificial intelligence methods failed to understand the physics behind the operations. To ensure the physical and technical feasibility of the prediction a coupling condition between the two have been developed for the ROP optimization. Therefore, we attempt to leverage the computational advances in predictive modeling and couple it with traditional approaches to help decipher the complex relationship ROP follows.

1.2 Background and Literature Review

The following literature review was performed before undertaking data analysis. Virtual Intelligence techniques in general and artificial neural networks (ANN) in particular have been used to solve problems in the various branches of petroleum engineering as shown in Table 1.

Table 1 - Summary of literature review.

Summary	Paper	Discipline
An Analytical Model coupled with Data Analytics to estimate PDC Bit wear- It discusses Warren drilling model to correlate rock strength from gamma ray, Abrasiveness from rock strength and Bit wear using above parameters.	Z. Liu et al. (2014)	Drilling
Analysis of Data from the Barnett Shale with conventional Statistical and Virtual Intelligence Techniques-This paper describes the application of ANN, both supervised and unsupervised and SOMs to predict water production.	Awoleke O. and Lane R. (2011)	Water production
Real Time Rate of Penetration Optimization using the Shuffled Frog Leaping algorithm- It provides details about ROP optimization using a heuristic function to seek a solution of optimization. This method is particularly useful for computing optimum drilling parameters in real time.	Ping Yi et al. (2014)	Drilling

Table 1 (continued)

Investigation of Various ROP Models and Optimization of Drilling Parameters for PDC and Roller-cone Bits in Shadegan Oil Field	Mahmood, B. et al. (2010)	Drilling
Data Analytics for Production Optimization in Unconventional Reservoirs	Schuetter, J et al. (2015)	Production optimization
Drilling Optimization Based on the ROP Model in One of the Iranian Oil Fields	Masood Mostofi et al. (2010)	Drilling
Enhancing Wellwork Efficiency with Data Mining and Predictive Analytics	Mohamed Sidahmed (2014)	Production optimization
Real-time Optimization of Rate of Penetration during Drilling Operation- This paper discusses the Bourgoyne and Young ROP model and introduces the moving-window method coupled with multiple regression, that computes coefficients from real time data for ROP calculation	Dan Sui et al. (2013)	Drilling
Data Mining and Predictive Analytics Transforms Data to Barrels	Richard Bailey et al. (2013)	Water flooding
Using Data-Driven Predictive Analytics to Estimate Downhole Temperatures while Drilling. – This paper discusses the usage of a machine learning technique called Support Vector Regression to estimate downhole temperatures.	Serkan Dursun et al. (2014)	Drilling

Table 1 (continued)

Real-Time Drilling Parameter Optimization System Increases ROP by Predicting/Managing Bit Wear	Yashodhan K. Gidh et al. (2011)	Drilling
Data Driven Analytics in Powder River Basin, WY	Mohammad Maysami et al. (2013)	Reservoir Management
Drilling Hydraulics Optimization Using Neural Networks	Yanfang Wang et al. (2015)	Drilling
Application of Neural Networks for Predictive Control in Drilling Dynamics	Dashevskiy D. et al. (1999)	Drilling
Predictive Analytics: Development and Deployment of Upstream Data Driven Models	Keith Richard Holdaway (2012)	Completion
Big Data Every Day: Predictive Analytics Used to Improve Production Surveillance	Scott Raphael (2015)	Production optimization
Efficient Use of Data Analytics in Optimization of Hydraulic Fracturing in Unconventional Reservoirs	C. Temizel (2015)	Hydraulic fracturing
Ensemble Learning - Boosting and Bagging. ListenData. http://www.listendata.com/2015/03/ensemble-learning-boosting-and-bagging.html (accessed 24 May 2016).	Bhalla, D. (2015)	Data analytics/ ML
Introduction to Data Mining. Boston: Addison-Wesley. Applications of SVR.	Tan, P-N, Steinbach, M., and Kumar, V. (2006)	Data analytics/ ML

Table 1 (continued)

How to Evaluate Machine Learning Algorithms. http://machinelearningmastery.com/how-to-evaluate-machine-learning-algorithms.	Brownlee, J. (2013)	Data analytics/ ML
Application of Neural Networks for Predictive Control in Drilling Dynamics	Dashevskiy, D. et al. (1999)	Drilling
Stuck Pipe Prediction and Avoidance: A Convolutional Neural Network Approach	Siruvuri, C. et al. (2006)	Drilling
Pipe Sticking Prediction and Avoidance Using Adaptive Fuzzy Logic Modeling	Murillo, A. et al. (2009)	Production optimization
Robust Well Cost Estimation Using Support Vector Machine Model	Buddharaju, P. et al. (2007)	Data analytics/ ML
Machine Learning in R for beginners. Datacamp. https://www.datacamp.com/community/tutorials/machine-learning-in-r	Datacamp (2015)	Data analytics/ ML
Modeling – Predicting the amount of rain. http://theanalyticalminds.blogspot.com/2015/04/part-4a-modelling-predicting-amount-of.html	Pedro M. (2015)	Data analytics/ ML
Model Selection for Support Vector Machines	Chapelle, O. and Vapnik, V. (1999)	Data analytics/ ML

1.3 Objectives

The objectives of this research work are as follows:

1. Analyze drilling data from the North Sea horizontal wells using statistical regression and machine learning techniques to understand intricate relationships among several variables (ROP, RPM, WOB, Flow, GR etc.)
2. Apply predictive modeling to build several models for ROP forward prediction using multivariate data and choose the best performing models.
3. Perform sensitivity and uncertainty analysis on the best models using Monte Carlo simulation to compute a range of ROP values (P10, P50 and P90). Calculate the most contributing parameters and their variation on ROP in order to maximize ROP for all the models.
4. Develop a web-based prediction application for drilling engineering ROP prediction, which can be later extended to production and reservoir engineering.

1.4 Thesis Outline

In this study, data from the North Sea horizontal wells was used and analyzed for the development of a model, based on statistical regression and machine learning methods. Figure 1 presents the workflow of the thesis project.

Chapter 2 of this study discusses the preliminary work completed pertaining to data preparation and performance of preprocessing operations such as extracting the relevant parameters for our analysis, and taking care of missing data and outliers. It also includes exploratory data analysis such as analyzing correlations between data sets, and outlier extraction. Several methods of data splitting were studied and tested to prepare test, validation and train datasets. Cross-validation was employed to prevent overtraining of the models. Error metrics that were used to differentiate our algorithms are also defined and explained in Chapter 2.

In chapter 3, various techniques of regression and machine learning were introduced. Influence of interactivity within parameters and relative importance of predictors was ascertained. A test harness of formations from Well 12a was used on several algorithms from different categories such as regression (linear, stepwise), neural networks (NN, SVR), and instance-based methods (KNN), trees and ensemble methods such as random forests (RF) and boosting. Tuning was performed to further enhance the performance of these models.

In chapter 4, the best algorithm was determined and then applied on the remaining five wells. These were further treated to Monte Carlo simulation using a pre-defined test set to check robustness and perform uncertainty analysis. Results and conclusions are presented within this chapter.

In chapter 5, the web-based prediction application designed using Shiny R was introduced and its applicability explained. The app enables any user to perform predictive analytics on ROP prediction, or other applications such as fracture design in production engineering.

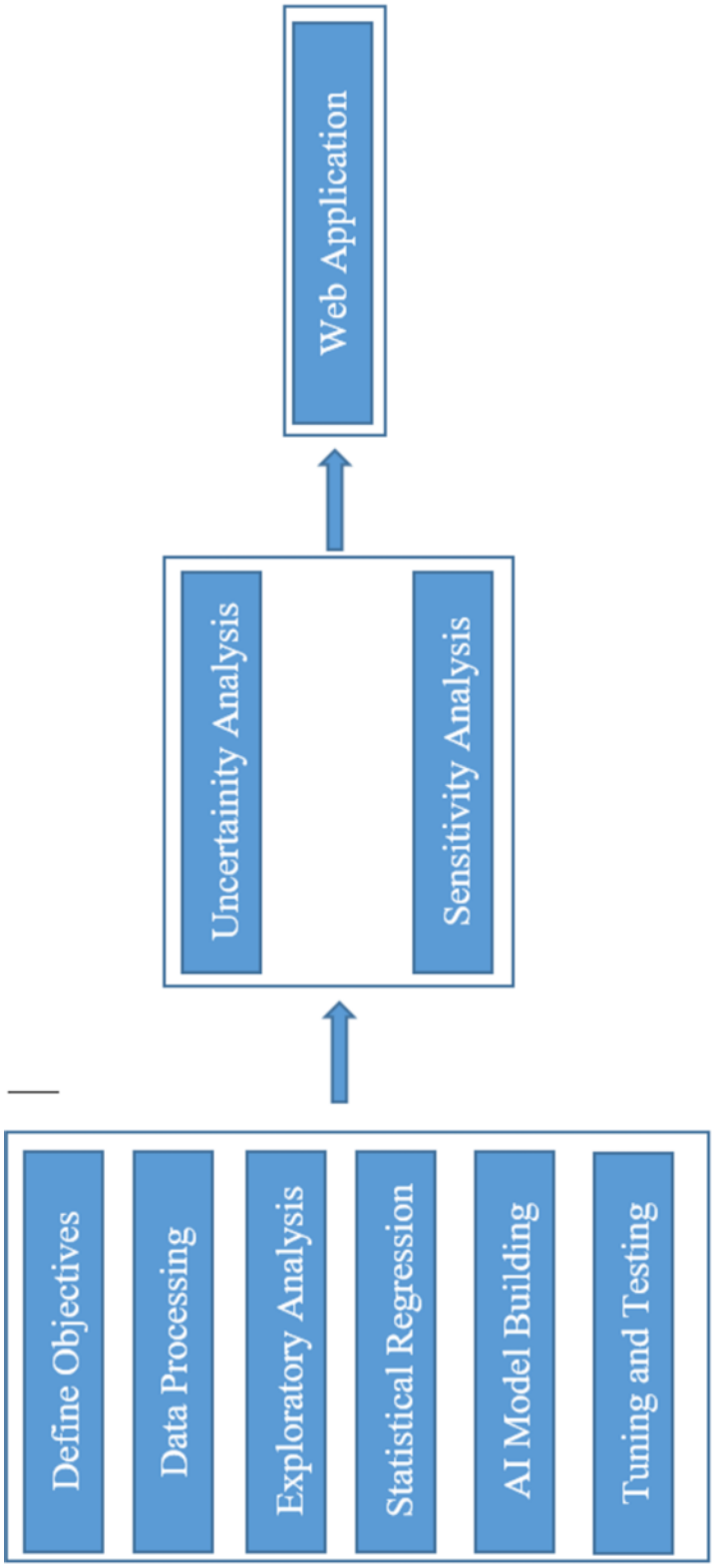


Figure 1- Workflow of the thesis project.

2. DATA PREPARATION

2.1 Data Acquisition and Description

Data preparation workflow is presented in Figure 2. Client drilling data from several wells in the North Sea was acquired (courtesy of Halliburton). All the wells have been code named to preserve the confidentiality of the data. All the data was recorded in real-time and was comprised of several parameters such as rate of penetration (ROP), weight on bit (WOB), rotary speed (RPM), flow rates (Flow), pressure, torque etc. Information about the formations, gamma ray logs (GR), and survey data were provided in separate files. Averaged values of RPM, WOB, Flow, GR, etc. were provided in *Parameter- Averaged* files, while real time values of the entire drilling operation were in *Time- Series* files. The time series fields include the NPT recordings as well. In general, ROP optimization involves the adjustment of WOB and RPM for efficient drilling. But ROP follows a complex relationship with several other parameters such as formation properties, mud properties, mud hydraulics, borehole deviation and size/type of bit. Hence, the presence of additional data such as GR is very helpful in optimizing the ROP and is an intrinsic part of the models built in the exercise. Survey data along with specific formation depths were also provided in separate files. This data is also extremely useful as separate models were built for each formation and this approach led to a better estimation of the ROP rather than the case when a single model was built for the entire dataset. More information about log data, and formation properties could have increased the accuracy of the models as it plays a vital role in influencing ROP.

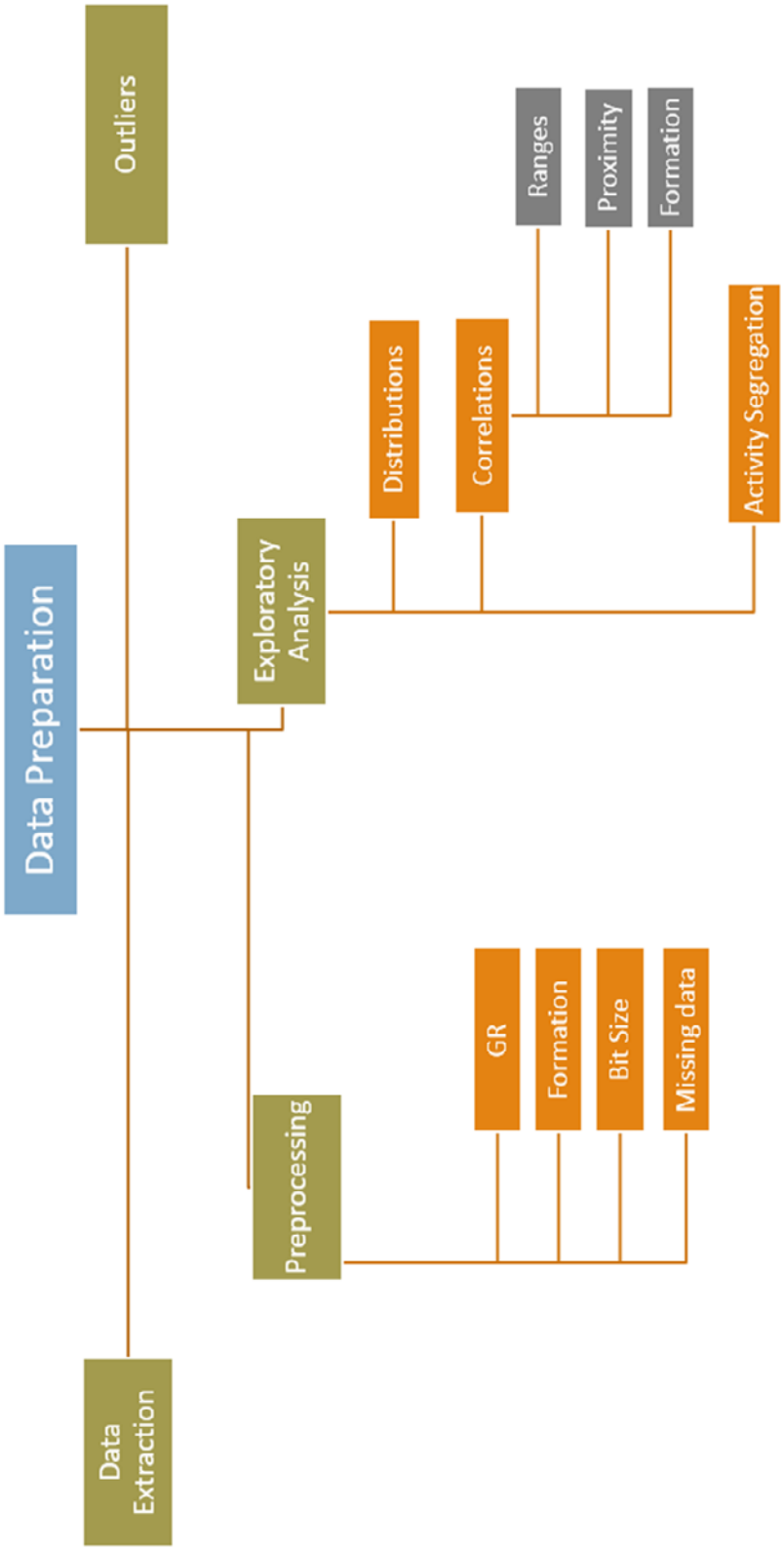


Figure 2- Data preparation workflow.

2.2 Data Preprocessing and Summary

The entire data set consisted of *Time- Series* data, *Parameter- Averaged* data, and *Survey* data for 7 wells in the North Sea. All the wells are horizontal in nature and are spread across about 13-14 formations. Table 2 lists the size of the data points of the above three data-frames for each well. *Survey* data included information about formation intervals, bit size and bit type while *Parameter- Averaged* data had values of GR.

Table 2-Time-series and Parameter-Averaged data of the wells.

Well ID	***Parameter- Averaged***	***Time-Series***
10	7005	510910
12a	11021	694762
B2a	8412	666918
B30y	5103	538476
13	9402	481921
E8	10333	941447

Since a comprehensive analysis for ROP prediction was being performed, effort was made to use as much information as possible. Drilling data such as ROP, WOB, Flow, mud details, and pressure was extracted from the *Time-Series* files; data containing formation depts., bit size and bit type was extracted from *Survey* files and GR values were extracted from the *Parameter-Averaged* files. All this data was then used to create a comprehensive master file containing all the relevant information for our analysis. This process was repeated for all the wells. As this data was recorded in real time, it also contained the NPT periods of tripping in/out, rotation, and sliding. Several parameters were then analyzed to segregate the NPT activities from drilling (as discussed in subsequent sections) and an additional “activity” column was annexed to this master

file. All the data taken from the master file was loaded into R (Statistical Programming Language) and then separate files (data frames) for each well were created. Figure 3 lists all the attributes of the master data-frame for Well 12a.

```
 [1] "TimeString"      "ROPA"            "MWin"
 [4] "MWout"           "TempMudin"       "ROP1"
 [7] "ROPI"            "BlockPos"        "TFAC"
[10] "DepthHole"       "HookLoadA"       "WOBA"
[13] "RPMA"            "TorqueRel"       "TorqueAbs"
[16] "Flowout"         "Pressure"        "Flowin"
[19] "ECDBottom"       "ECDBit"          "Overpull"
[22] "HookLoad"        "OnbtmStatus"     "InslipsStatus"
[25] "Depth"           "DateTime"        "Date"
[28] "Time"            "Wellid"          "Formation"
[31] "Activity"        "BitSize"         "HoleDepthRounded"
[34] "GR"
```

Figure 3-Attributes of the master data-frame for Well 12a.

The master data-frames of Wells 12a and E8 had about 1% of missing values (NAs) while those of wells B30y, B2a and 10 had about 5%, 9% and 14% NAs respectively. NAs were removed since there was a sizeable amount of data present without the risk of alienating important features. Well 13 had about 90% of NAs that belonged to the annexed GR column. Due to the presence of such a high number of NAs in GR, GR for Well 13 was not extracted and annexed. Well 12a was used for the entire course of the project, from model testing, algorithm tuning, and sensitivity and uncertainty analysis, while the remaining data of 5 wells was used at the end for the work flow validation purposes.

2.3 Exploratory Data Analysis

Exploratory analysis was performed on individual wells as well as on data clusters that were created using several approaches: grouping of several wells by a specific formation, grouping of several wells by proximity to each other (from survey data) and grouping of data by ranges for a

particular parameter (ROP, WOB etc.). This process gave several insights as to how data behaves. Grouping was followed by computing predictor importance to infer relative contributions and weights of input variables to these clusters.

2.3.1 Histograms and Box Plots

To undertake any model building, it is essential to perform a preliminary analysis of each parameter involved. Histograms and box plots were constructed for parameters ROP, WOB, RPM, Flow and GR, as shown in Figures 4 and 5. Histograms and boxplots are very essential to identify the distributions of existing data. They help understand what specific ranges of values are prevalent in a particular parameter, and indicate any physical phenomenon behind the occurrence.

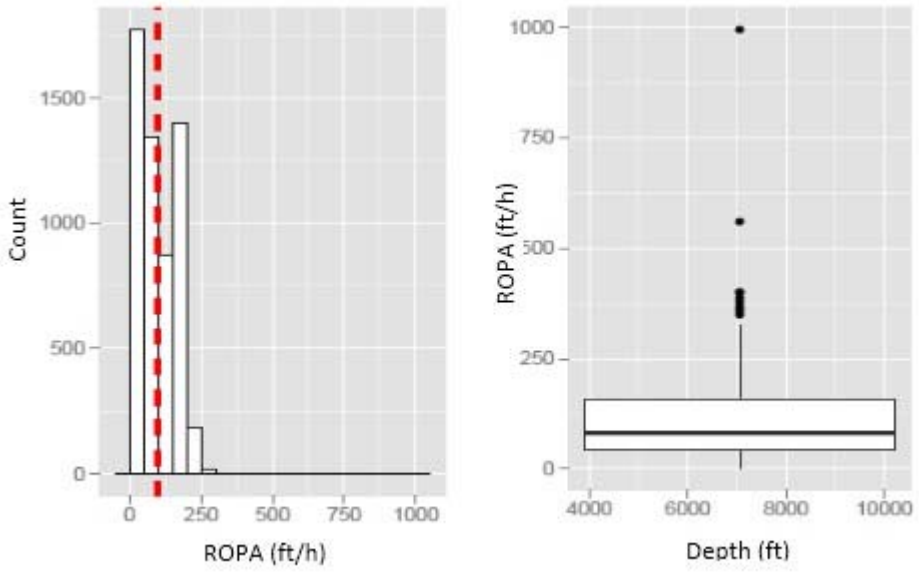


Figure 4-Histogram and boxplot of ROP for Well 10.

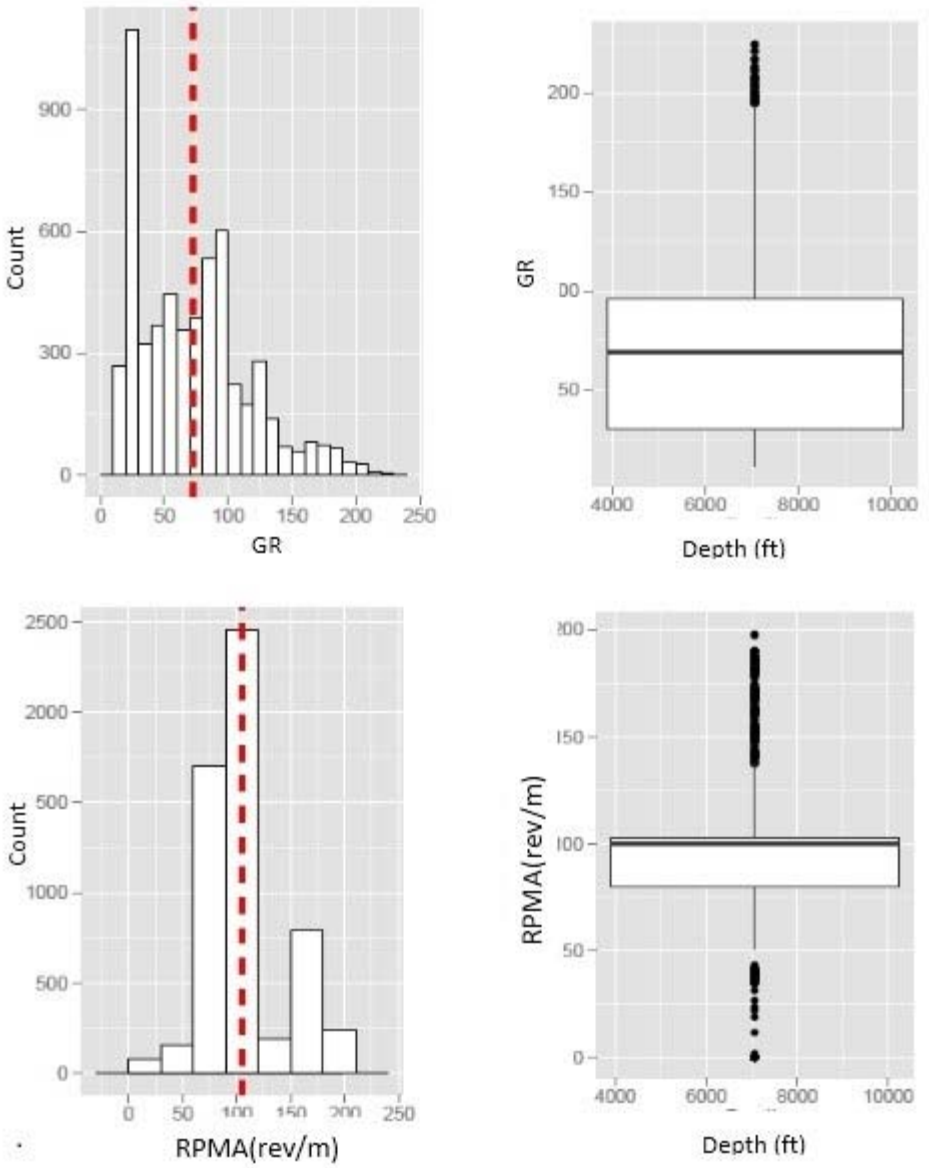


Figure 5-Histograms and boxplots of GR and RPM for Well 10.

The presence of outliers in the distributions made it clear that outliers could be detrimental to model performance. Domain expert advice was included in the isolation of such outliers, as explained in further chapters.

2.3.2 Correlations

Correlation matrices have been computed for different sets of parameters to identify if there is any high degree of correlation as this would influence the models later. Correlations were run on entire data initially for Well 12a as shown in Figure 6.

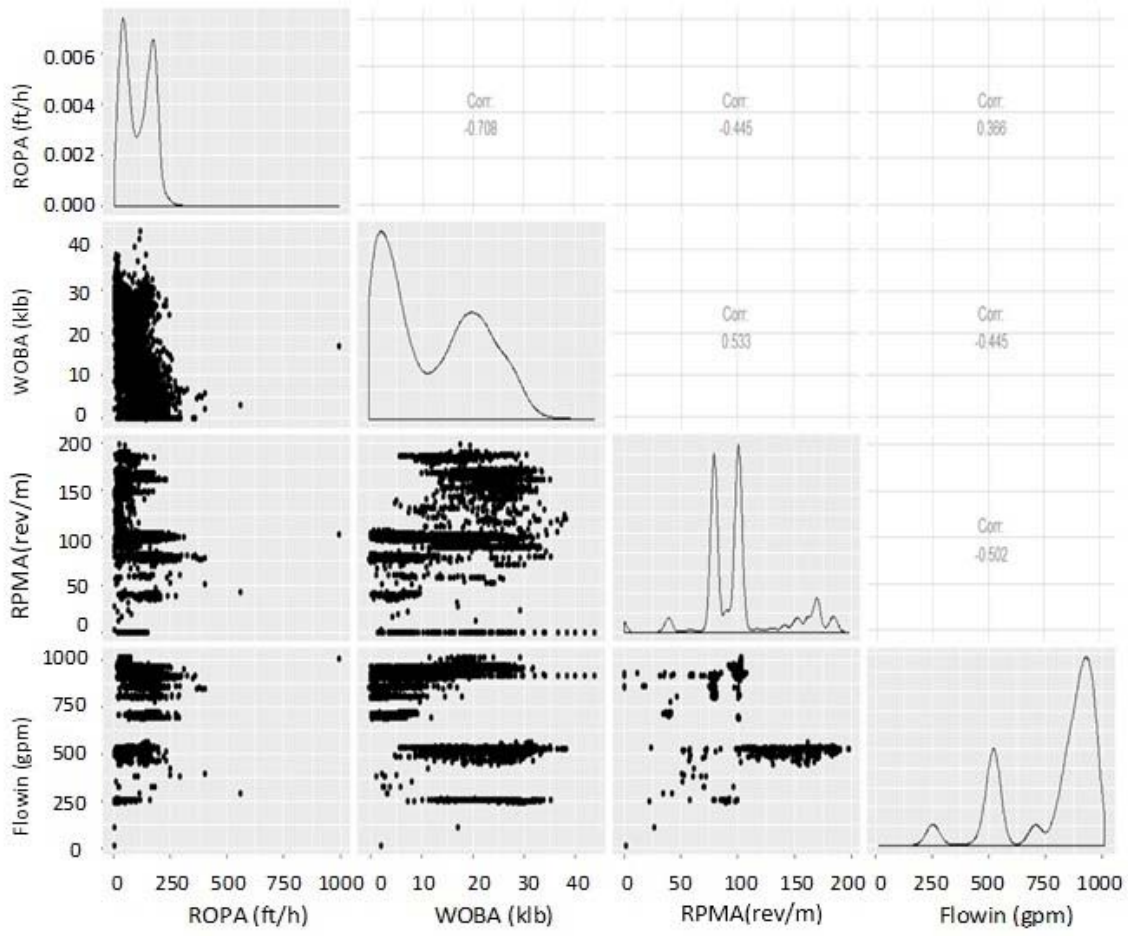


Figure 6- Correlation matrix with scatter plots for Well 12a.

There was no noticeable trend between various parameters with the ROP. More analysis was done later to see if grouping wells led to better correlations. Several wells were grouped by proximity to each other (using information on the location from the survey data), by formations and by trisecting the data into high, medium and low for a given parameter respectively.

2.3.2.1 Grouping by parameter ranges

Initially, a particular parameter (such as ROP) was chosen and the entire data for a well was segregated on the basis of a prefixed high, medium and low range for that parameter. Then, the correlation matrix was computed for these 3 groups of data branches as shown in Table 3, which shows the segregation by ranges (high and medium) for ROP in the Well 10.

Table 3- Correlation matrix of ROP vs. other variables for grouping by ranges of ROP.

ROP

ROP- High

	ROPA	WOBA	RPMA	Flowin
ROPA	1.00000000	0.1335727	-0.05450765	-0.16716567
WOBA	0.13357269	1.0000000	0.60829706	-0.25871104
RPMA	-0.05450765	0.6082971	1.00000000	0.02291333
Flowin	-0.16716567	-0.2587110	0.02291333	1.00000000

	ROPA	GR	Pressure	Temp
ROPA	1.00000000	-0.01373709	-0.12586529	0.03551352
GR	-0.01373709	1.00000000	0.04120238	0.31435978
Pressure	-0.12586529	0.04120238	1.00000000	0.71563163
Temp	0.03551352	0.31435978	0.71563163	1.00000000

	ROPA	HookLoadA	TorqueAbs	ECDA
ROPA	1.00000000	0.04764983	0.02287414	0.16374399
HookLoadA	0.04764983	1.00000000	0.76475295	0.13068562
TorqueAbs	0.02287414	0.76475295	1.00000000	0.01909608
ECDA	0.16374399	0.13068562	0.01909608	1.00000000

ROP- Medium

	ROPA	WOBA	RPMA	Flowin
ROPA	1.00000000	-0.2929078	-0.2648585	0.09190481
WOBA	-0.29290777	1.0000000	0.5653154	-0.48704140
RPMA	-0.26485847	0.5653154	1.0000000	-0.36100742
Flowin	0.09190481	-0.4870414	-0.3610074	1.00000000

	ROPA	GR	Pressure	Temp
ROPA	1.00000000	-0.03637899	-0.2565180	-0.40530824
GR	-0.03637899	1.00000000	-0.1990630	0.01442185
Pressure	-0.25651799	-0.19906300	1.0000000	0.84223850
Temp	-0.40530824	0.01442185	0.8422385	1.00000000

	ROPA	HookLoadA	TorqueAbs	ECDA
ROPA	1.0000000	-0.33127763	-0.2785656	-0.19982154
HookLoadA	-0.3312776	1.00000000	0.7882023	0.01895503
TorqueAbs	-0.2785656	0.78820231	1.0000000	0.17569040
ECDA	-0.1998215	0.01895503	0.1756904	1.00000000

Trends (correlation values of greater than 0.7 or lesser than -0.7) have been noticed but there has been no consistency when the same process was applied to a different well. Similarly, correlations were computed by taking WOB, RPM and GR ranges to find any meaningful relationships between the variables.

2.3.2.2 Grouping by proximity

Several wells were grouped by proximity (using latitude and longitude values, pad details etc.) from the given survey information and then the correlations were computed. Wells B28 and B30y, and B2a and E8 are two such groups which have been used in the analysis. Table 4 shows the results of both the groups. Better correlations were noticed even though the entire well data was utilized for the correlation computation, unlike segregation by ranges as done in the previous section.

Table 4-Correlation matrix of ROP vs. other variables using grouping by proximity.

##Wells B2a, E8

```
##            ROP_avg    WOB_avg    RPM_avg    Flow_in
## ROP_avg  1.0000000  0.8062987  0.9191268 -0.6818877
## WOB_avg  0.8062987  1.0000000  0.9544458 -0.7146659
## RPM_avg  0.9191268  0.9544458  1.0000000 -0.7435934
## Flow_in -0.6818877 -0.7146659 -0.7435934  1.0000000

##               ROP_avg         GR   Pressure    Temp_Mud
## ROP_avg    1.00000000  0.1685826  0.1718521 -0.09433698
## GR         0.16858261  1.0000000 -0.5487716  0.68487262
## Pressure   0.17185215 -0.5487716  1.0000000 -0.40327649
## Temp_Mud -0.09433698  0.6848726 -0.4032765  1.00000000

##                ROP_avg HookLoad_avg  Torque_avg     ECD_avg
## ROP_avg       1.0000000    0.5432922 -0.19470866  0.28446704
## HookLoad_avg  0.5432922    1.0000000  0.27467666  0.22781062
## Torque_avg   -0.1947087    0.2746767  1.00000000 -0.05980717
## ECD_avg       0.2844670    0.2278106 -0.05980717  1.00000000
```

##Wells B28, B30y

```
##            ROP_avg    WOB_avg    RPM_avg    Flow_in
## ROP_avg  1.0000000  0.9339527  0.9617455 -0.7282659
## WOB_avg  0.9339527  1.0000000  0.9830082 -0.7351787
## RPM_avg  0.9617455  0.9830082  1.0000000 -0.7429190
## Flow_in -0.7282659 -0.7351787 -0.7429190  1.0000000

##               ROP_avg         GR   Pressure    Temp_Mud
## ROP_avg    1.0000000 -0.16764140  0.3058432  0.30686716
## GR        -0.1676414  1.00000000  0.2209771 -0.02644867
## Pressure   0.3058432  0.22097713  1.0000000  0.66204800
## Temp_Mud   0.3068672 -0.02644867  0.6620480  1.00000000

##                ROP_avg HookLoad_avg  Torque_avg    ECD_avg
## ROP_avg       1.00000000    0.5291922 -0.04305868 0.4421812
## HookLoad_avg  0.52919220    1.0000000  0.36376873 0.6833833
## Torque_avg   -0.04305868    0.3637687  1.00000000 0.5592684
## ECD_avg       0.44218115    0.6833833  0.55926843 1.0000000
```

2.3.2.3 Grouping by formation

The groups of wells in the previous section were further analyzed by computing the correlation matrix for each formation separately. Table 5 shows the results for formations Valhall and Plenus of B28-B30y. The correlations were stronger and consistent when segregated by formation. This exercise highlighted the need to isolate each formation for model building while computing the ROP as it follows a complex relationship with changing formation characteristics.

Table 5-Correlation matrix of ROP vs. other variables for grouping by formation.

Valhall

```
##            ROP_avg    WOB_avg    RPM_avg    Flow_in
## ROP_avg  1.0000000  0.9851322  0.9843039 -0.9826936
## WOB_avg  0.9851322  1.0000000  0.9972796 -0.9974219
## RPM_avg  0.9843039  0.9972796  1.0000000 -0.9988267
## Flow_in -0.9826936 -0.9974219 -0.9988267  1.0000000

##               ROP_avg          GR   Pressure   Temp_Mud
## ROP_avg    1.00000000 -0.09376293 0.02874902 -0.1825401
## GR        -0.09376293  1.00000000 0.38650817  0.4226487
## Pressure   0.02874902  0.38650817 1.00000000  0.7534640
## Temp_Mud  -0.18254008  0.42264874 0.75346404  1.0000000

##                ROP_avg HookLoad_avg  Torque_avg    ECD_avg
## ROP_avg       1.00000000    0.9327051 -0.08885869  0.9765132
## HookLoad_avg  0.93270507    1.0000000 -0.18150366  0.9706150
## Torque_avg   -0.08885869   -0.1815037  1.00000000 -0.1874699
## ECD_avg       0.97651319    0.9706150 -0.18746994  1.0000000
```

Plenus

```
##            ROP_avg    WOB_avg    RPM_avg    Flow_in
## ROP_avg  1.0000000  0.9966539  0.9970954 -0.9943472
## WOB_avg  0.9966539  1.0000000  0.9996823 -0.9969033
## RPM_avg  0.9970954  0.9996823  1.0000000 -0.9974709
## Flow_in -0.9943472 -0.9969033 -0.9974709  1.0000000

##               ROP_avg          GR   Pressure    Temp_Mud
## ROP_avg    1.00000000 -0.01603629  0.2568598  0.06071687
## GR        -0.01603629  1.00000000 -0.2696636 -0.41499524
## Pressure   0.25685976 -0.26966356  1.0000000  0.78657548
## Temp_Mud   0.06071687 -0.41499524  0.7865755  1.00000000

##                ROP_avg HookLoad_avg Torque_avg     ECD_avg
## ROP_avg       1.0000000   -0.52909738  0.8340736  0.11046706
## HookLoad_avg -0.5290974    1.00000000 -0.6950546 -0.01085822
## Torque_avg    0.8340736   -0.69505457  1.0000000  0.18219407
## ECD_avg       0.1104671   -0.01085822  0.1821941  1.00000000
```

2.3.3 Segregation by Activity

As the given data consisted of all the activities such as tripping in/out, sliding, and rotation on bottom, there was a need to isolate the data for all these NPT activities from the actual drilling

data in order to effectively construct models. As shown in the Table 6, the following parameter ranges were used to isolate drilling data, which would later be further preprocessed before initiating model construction.

Table 6- Ranges for parameters applied for segregating data by activity.

	Drilling	**Trip in**	**Trip out**
ROP	> 0	0	0
RPM	≠ 0	~ 0	~ 0
WOB	≠ 0	0	0
Flow	≠ 0	0	0
Bit Depth	same range and increasing	Increasing	decreasing
Hole Depth		no change	no change

	Rotation Off Bottom	**Sliding**	**Back reaming**
ROP	0	> 0	0
RPM	≠ 0	bit ≠ 0, pipe = 0	≠ 0
WOB	0	≠ 0	< 0
Flow	≠ 0	≠ 0	≠ 0
Bit Depth	no change	same range and increasing	decreasing
Hole Depth	no change		no change

2.3.4 Outlier Identification

After the NPT activities were isolated, there was still a need for further preprocessing as the data contained a considerable number of outliers for each parameter. Therefore, outlier analysis was performed and the first layer of outliers were removed after deliberations with domain

experts about the noticed field ranges for the parameters involved in model construction such as WOB, ROP, GR, RPM and GR. This data is stored in separate data-frames for outlier analysis (as discussed in further chapters).

Figure 7 shows the boxplots by formation for ROP in Well 12a where the dots represent outliers. Figure 8 represents boxplots of ROP after the removal of outliers. Box plots present the distribution of data by quartiles. Outliers are those values beyond 1.5 times the interquartile range (Q3-Q1). The extremes in ROP values could be due to various reason: recording errors, bit wear, impending bit failure or due to change in formations. It is clearly evident that such high values of ROP need to be isolated for a better model performance. Separate analysis was conducted on the outliers file to see if they offer any meaningful insight as to what caused such spikes.

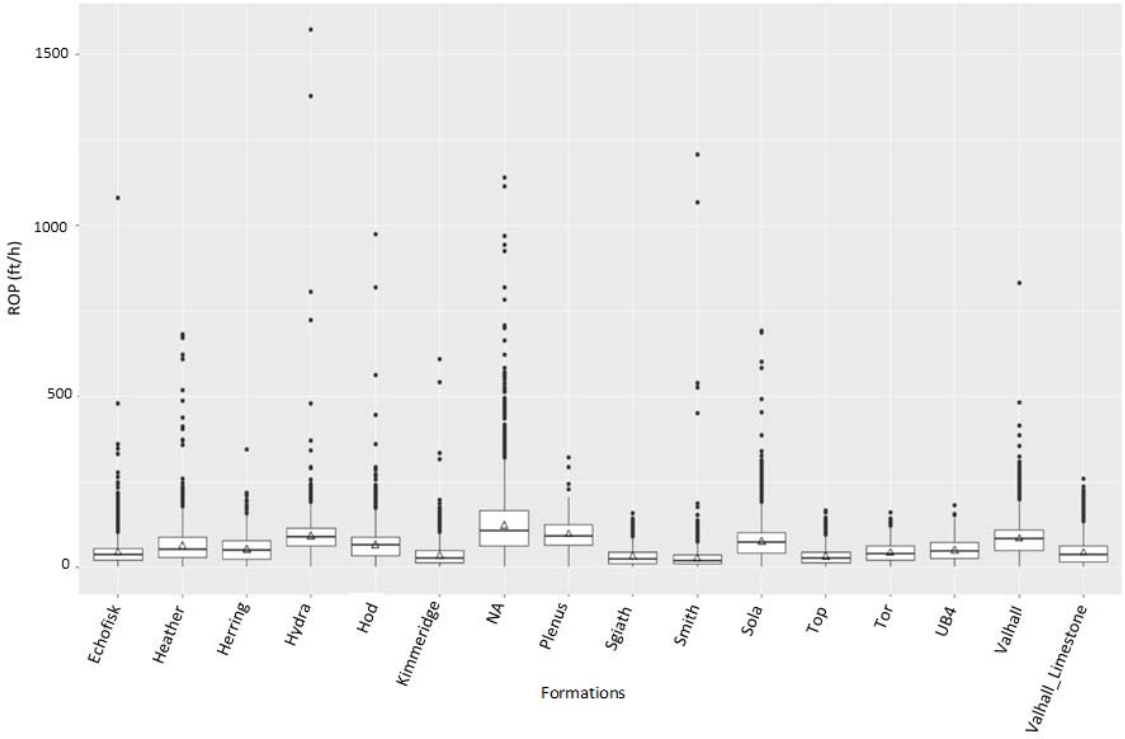


Figure 7-Boxplots by formation for ROP in Well 12a.

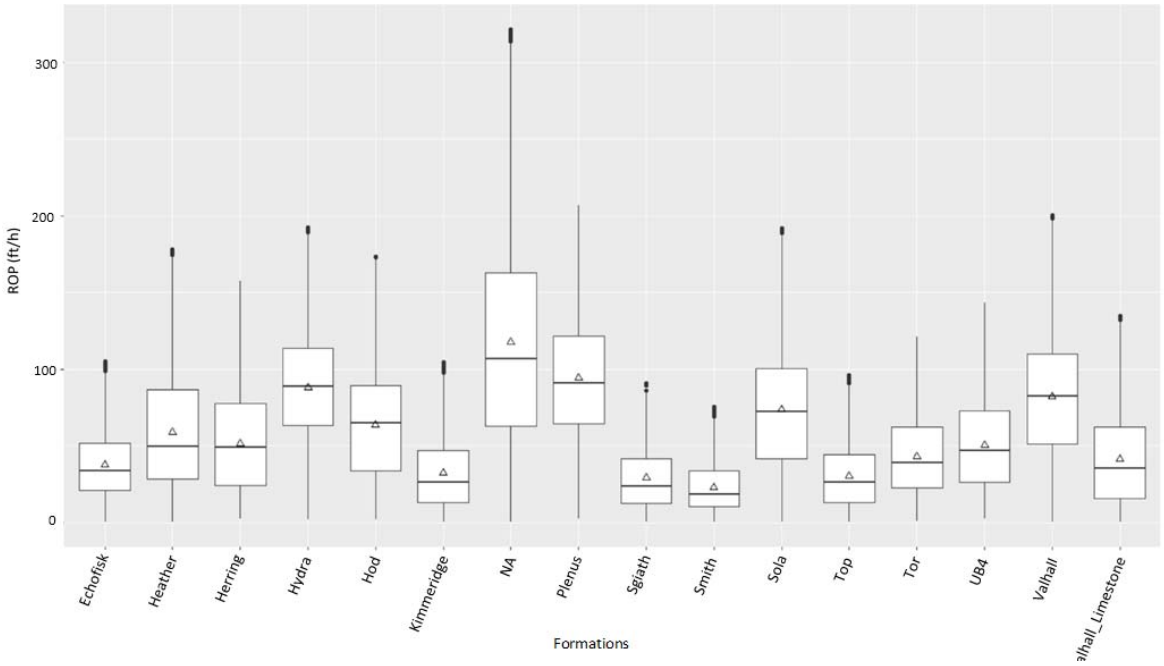


Figure 8-Boxplots by formation for ROP without outliers for Well 12a.

A similar process was followed to extract outliers from the other important parameters and the data was further cleaned. Therefore, after a series of operations such as data extraction, inclusion of data from formation and survey files, merging of relevant attributes, preprocessing, omission of irrelevant parameters, drilling data segregation, and outlier isolation, the processed data sets were finally ready for model building. Figure 9 lists the processed data by formation for Well 12a. This well has 15 formations. Data with missing formation depths are categorized as a "NA" formation. This data was also used as part of model building as it gives us an idea about the model performance in absence of formation intervals data.

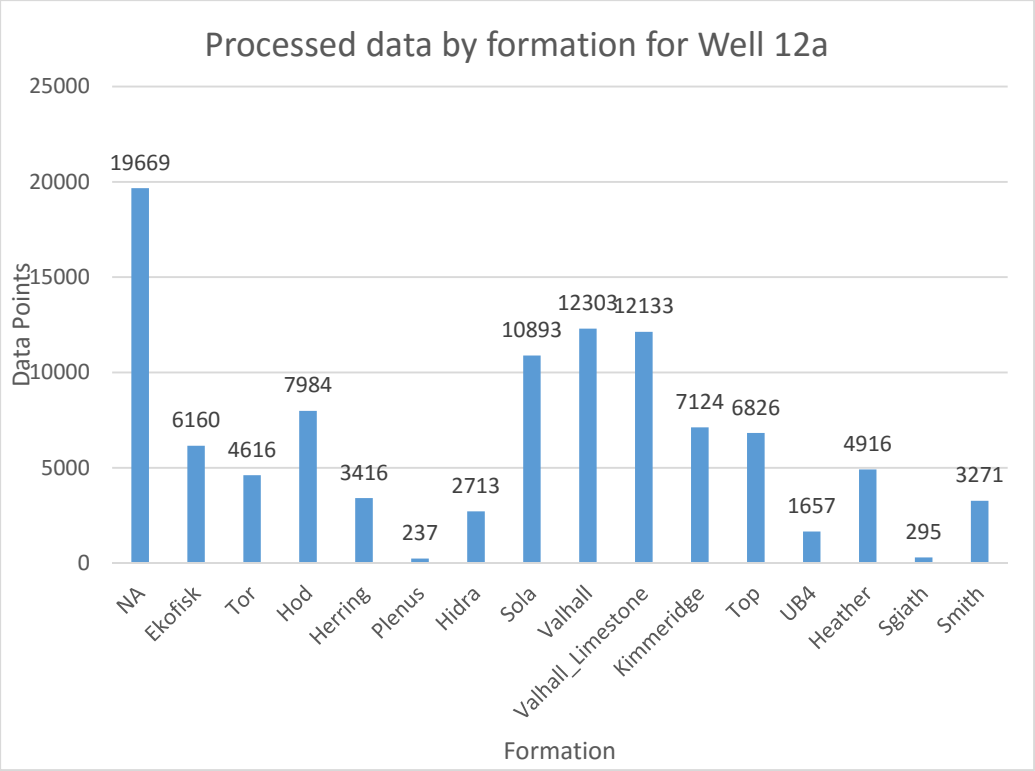


Figure 9-Processed data by formation for Well 12a.

As shown below, Figure 10 represents the ROP versus RPM and WOB while Figure 11 represents ROP versus GR and Flow during a specific time interval using the processed dataset. As the ROP follows a complex relationship with each of these parameters wherein there could also be interaction between the parameters themselves, all these are used in model building. Their relative weights and contributions are discussed in Chapter 3. These processed datasets were then used for cross validation, model building, refining and ultimately validation purposes. The Table 7 represents the notation used for final parameters in the subsequent analysis.

Table 7- Parameters used in analysis and their respective notations.

Parameter	**Notation**
Rate of penetration (ft/h)	*ROPA*
Weight on Bit (klb)	*WOBA*
String rotary speed (rev/min)	*RPMA*
Gamma Ray	*GR*
Mud Flow in (gpm)	*Flowin*

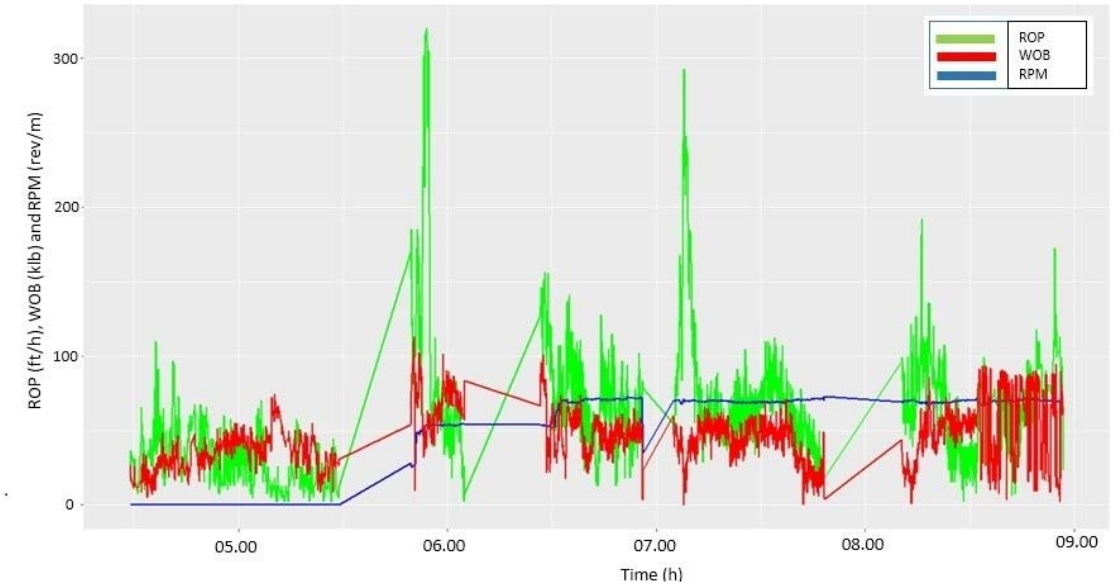


Figure 10-ROP versus RPM and WOB for Ekofisk formation in Well 12a.

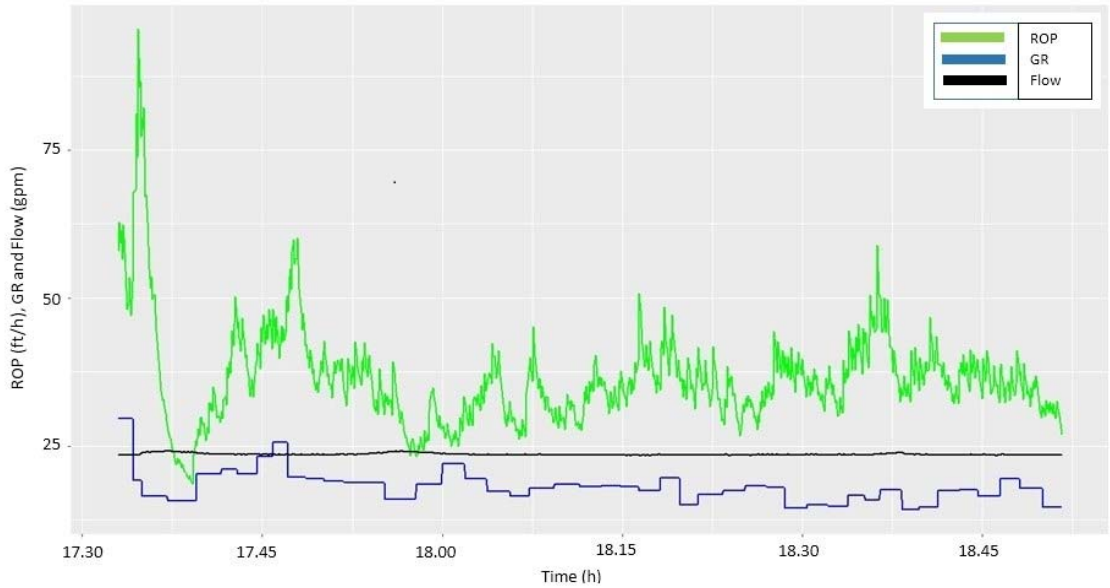


Figure 11-ROP versus GR and Flow for Ekofisk formation in Well 12a.

2.4 Data Splitting

Data splitting was performed to create train, validation and test data sets against which the algorithm is trained, modified and tested. Several error metrics such as root mean square error

(RMSE) and mean absolute error (MAE) were used in order to judge the performance of several algorithms. Therefore, it is extremely important to split data preserving integrity and similarity in all the three datasets so that there is no bias in evaluating different algorithms while choosing the best performing ones. Four methods were applied to see if there is an inherent advantage of using one over the other such as holdout, k-fold cross validation (CV), repeated k-fold cross validation and bootstrapping (Brownlee, J. 2013). The outcome of testing multiple algorithms against a sample dataset such as a single well data in this case is an estimation of how different algorithms perform on the problem against a chose performance measure. If all these varied algorithms fail to perform, it may be an indication of a lack of structure available for these algorithms to learn. Although this may occur due to an actual lack of learnable structure in the selected date, it also provides an opportunity to try different data transformations to interpret the structure to the learning algorithms.

2.4.1 Hold Out Method

The hold out method involves the concept of slicing the data into a training data that is usually used to prepare the model and an unseen test data that is then employed to evaluate the model's performance on unseen data. Usually 75% of the data is used for training purposes. In the case, where training, testing and validation sets are created, 60% is used for training, 25% for validation and the remaining for testing purposes. In this case study, 75% of the data in each formation was used for training and 25% for testing purposes directly. Well 12a was used for all the data analysis, model selection, and development of a unique algorithm. The other 5 wells are tested against the developed algorithm.

Figure 12 shows the histograms of train and test data sets for Ekofisk formation in Well 12a for RPM and WOB. Care should be taken that the distributions of test and train look similar or else it would lead to incoherency in model building and inaccurate results.

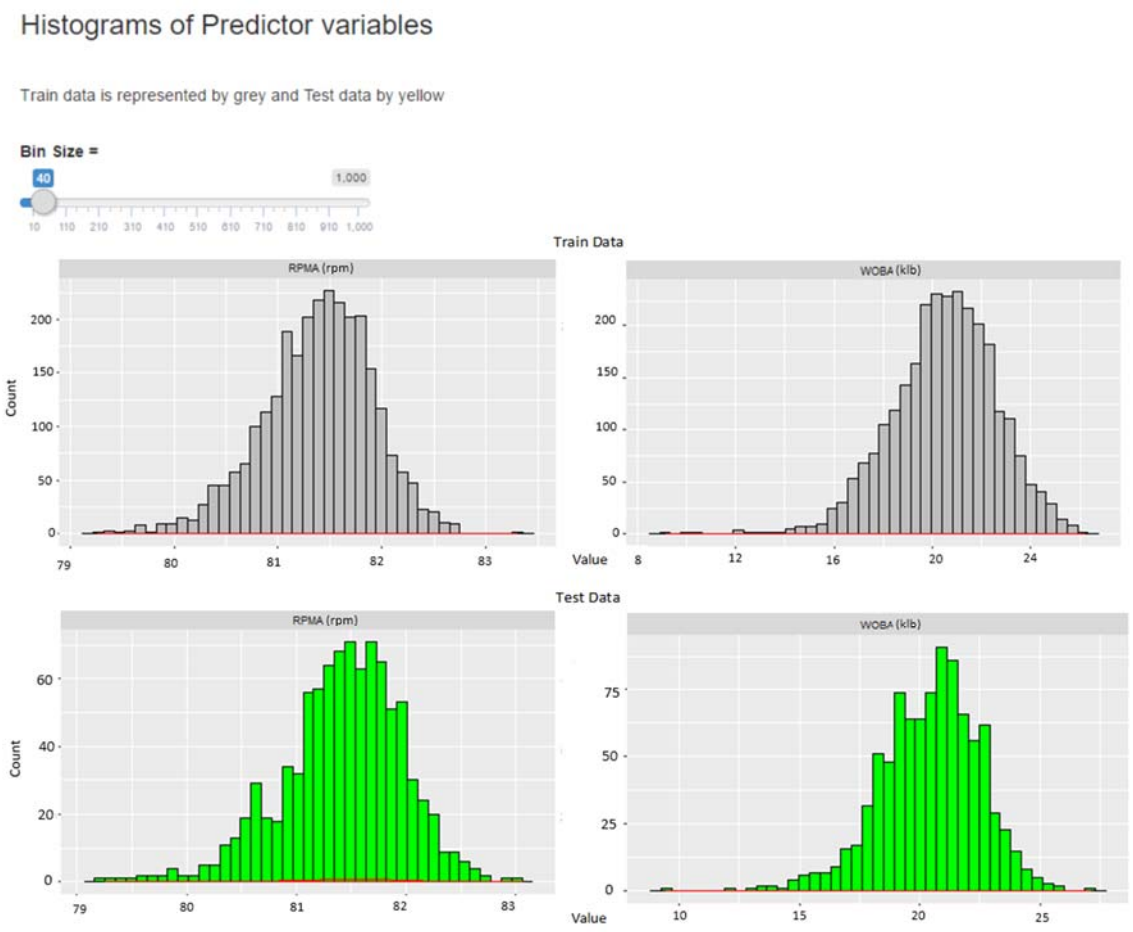


Figure 12-Histograms of train and test data for Ekofisk formation in Well 12a for RPM and WOB.

Table 8 shows the summary of the above mentioned datasets. The presence of outliers can greatly distort the distributions and hence their isolation can lead to better model building as long as relevant values are not discarded, even if they exceed ranges.

Table 8- Summary of test and train datasets for Ekofisk formation in Well 12a.

Summary of Test Data

	ROP1	RPMA	WOBA
1	Min. :20.31	Min. :79.26	Min. : 9.46
2	1st Qu.:30.37	1st Qu.:81.12	1st Qu.:19.07
3	Median :33.76	Median :81.47	Median :20.61
4	Mean :34.43	Mean :81.42	Mean :20.41
5	3rd Qu.:37.60	3rd Qu.:81.78	3rd Qu.:21.84
6	Max. :64.06	Max. :83.06	Max. :27.08

Summary of Train Data

	ROP1	RPMA	WOBA
1	Min. :21.43	Min. :79.27	Min. : 9.159
2	1st Qu.:30.36	1st Qu.:81.07	1st Qu.:19.206
3	Median :33.75	Median :81.43	Median :20.584
4	Mean :34.53	Mean :81.39	Mean :20.498
5	3rd Qu.:37.61	3rd Qu.:81.75	3rd Qu.:21.921
6	Max. :61.49	Max. :83.30	Max. :26.145

2.4.2 Cross Validation (CV)

CV and repeated CV involve isolating the dataset into a number of equally sized groups of instances (also referred to as folds). Every model is then trained on all folds except one that was left out. This model is then tested against the untouched fold. By repeating this procedure, every fold of data gets selected to be either a part of the training data set or an opportunity at being left out of the training data, and therefore acting as the test dataset. Finally, the performance measures are computed and averaged across all folds to estimate the capability of algorithms.

For example, a 3-fold CV would involve training and testing a model 3 times:

#1: Training on folds 1 and 2 while testing on fold 3

#2: Training on folds 1 and 3, while testing on fold 2

#3: Training on folds 2 and 3, while testing on fold 1

Figure 13 shows the workflow for a 10-fold CV. There are 10 runs for the same set of data, with the test folds being varied 10 times during the entire process. But the usual concern with CV is that it uses randomness to decide how to split the dataset into k folds.

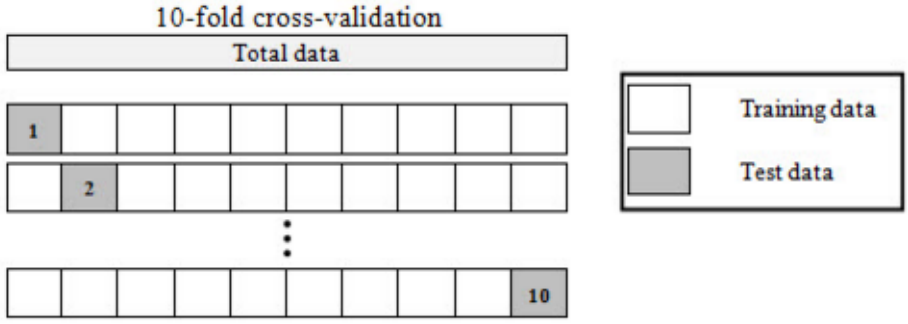


Figure 13- Ten-fold cross validation (CV) example.

2.4.3 Repeated K-fold Cross Validation

Repeated k-fold CV runs CV several times and computes mean of the accuracy. For example, six repeats of 10-fold CV would give 60 total iterations. Usually, if there is a smaller number of folds, error estimates are more biased (indicating higher error than there is in reality) and if there is a smaller number of folds, the variance decreases. On the extreme end, when using the Leave-One-Out-Cross-Validation technique (LOOCV), the error estimate is essentially unbiased but it could potentially have a high variance.

2.4.4 LOOCV

LOOCV is a special case of k-fold CV where k equals the number of instances in the data. So a model is constructed on other data points except one. This process is repeated for all data points. However, this method can be very time consuming and is only advisable when there are few data rows.

2.4.5 Boot Strapping

In bootstrapping, random samples are extracted from the dataset (with re-selection after replacement) and the model is applied to evaluate these samples. The model is first trained on the bootstrap sample and the data points not belonging to this boot strap sample are then predicted. These methods are very important when the number of data points is very low. All the above methods were applied using the linear regression model. Across all the methods, there has

not been much of difference when using a specific method, although the computation time in LOOCV is quite high. As all the methods discussed above delivered similar results, the holdout technique was used in regression methods for all subsequent analysis, as computation time was the least for holdout technique. CV was used in RF and Boosting and in some of the other machine learning algorithms where it is applied by default. In cases where there is no constraint on computation capabilities, it is advisable to apply CV or k-fold CV as it helps build the best models by splitting data accurately (preserving inherent data patterns) across test, train and validation data sets.

2.5 Error Metrics

When undertaking any model building exercises for prediction purposes, the primary goal should be to construct a model that accurately predicts an output for new data. To compute the accuracy, several error metrics can be analyzed. In this study, two error metrics: root mean squared error (RMSE) and mean absolute error(MAE) were used. Both the RMSE and MAE are used in predictive modeling quite often along with R-square and adjusted R-square. MAE is an average of the absolute errors, which is the difference between prediction and the true value. MSE computes the average of the squares of the deviations and RMSE is the square root of MSE.

Because of the square, large errors have relatively greater influence on RMSE than do the smaller errors. Therefore, MAE is more robust to outliers since it does not make use of squares of values. On the other hand, RMSE is more useful if we are concerned about large errors in which consequences are much bigger than equivalent smaller ones. So, the error metric chosen depends on the kind of target analysis that we are dealing with while considering these criteria. In the prediction app that was designed as a part of this project, the user can choose either of the two metrics (RMSE or MAE) to filter out the best models for each formation. This shall be discussed later in Chapter 5.

3. BUILDING MODELS

The different algorithms applied over the course of building models is mentioned in Figure 14. Using the master data-frames that included all the relevant parameters, several algorithms were applied including regression methods such as multivariate linear regression and stepwise regression, machine learning methods such as neural networks (NN), instance-based methods such as k-nearest neighbor (KNN) and support vector regression (SVR), classification and regression trees (CART). The model performance was further improved using ensemble methods such as random forest (RF) and boosting (GBM). The following chapter discusses the methodology employed in every algorithm using a train data set, its performance on a test dataset and tuning methods employed to further improve a particular model.

Preliminary analysis was performed using the regression methods to ascertain several variations in model building as shown in Figure 14. The findings of the preliminary analysis were then utilized in advanced machine learning models.

a. Effect of using the entire dataset for model building
b. Effect of clustering by formation and building models separately for each formation
c. Effect of variable interaction versus no-interaction
d. Variable importance

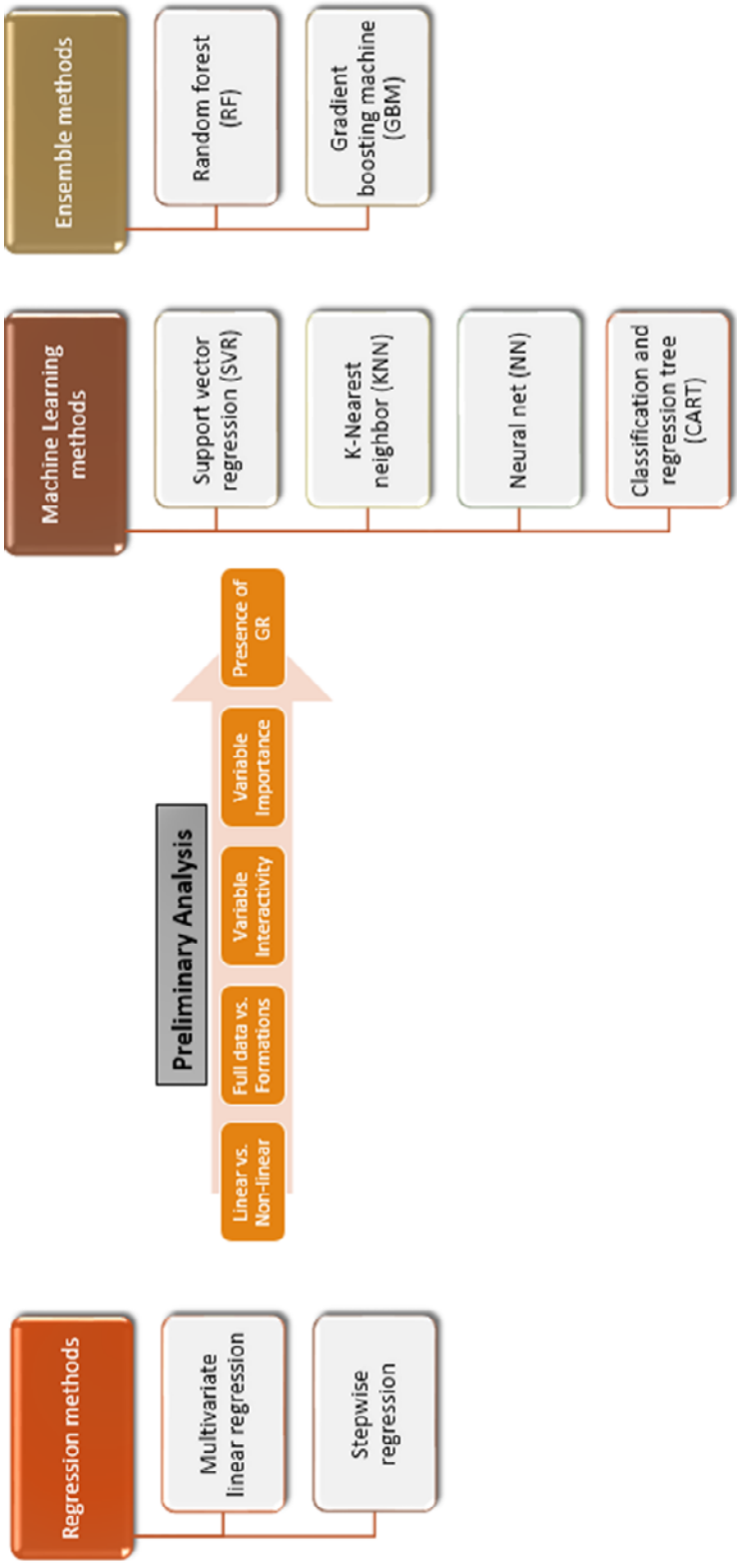


Figure 14- Workflow of the algorithm modeling analysis.

3.1 Regression Methods

3.1.1. Multivariate Linear Model (LM)

LM computes a relationship between the dependent variable (Y) and an independent variable (X) using a regression line, as shown in Figure 15. If there is more than one independent variable, it is referred to as multivariate regression.

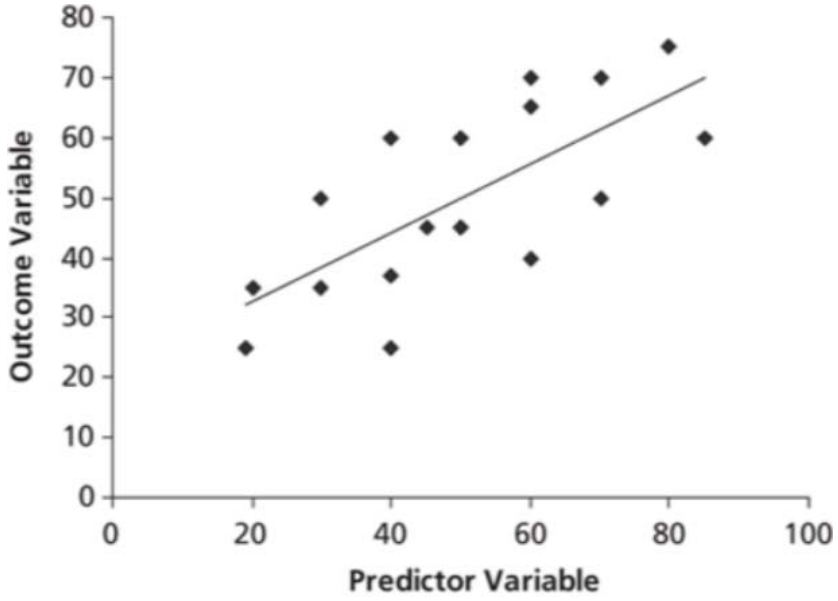


Figure 15- Example of Linear Regression

LM is usually denoted by an equation Y=a+b*X + e, where 'e' represents the error, 'a' is the intercept, 'b' is the slope. This equation can be applied to predict the value of the target variable based on the given predictors datasets.

There are some assumptions that must be considered for this model to be valid:

a. There must be linear relationship between independent and dependent variables.
b. Multiple regression suffers from multicollinearity, where there is a high degree of correlation between the predictors and from heteroscedasticity, in which the variability of a variable is unequal across the range of values of a second variable that predicts it.

c. Residuals have a mean of zero. All predictors are uncorrelated with the residuals and the residuals are not correlated with each other.

d. Residuals have a constant variance and are normally distributed.

Multivariate regression is linear regression considering multiple predictors. As mentioned in the chapter introduction, this section investigates several variations in model building.

3.1.1.1. Effect of using complete dataset

With ROP as the response variable, and RPM, WOB, and Flow as the input variables, regression was performed by considering the entire well data for 12a. This is important in order to analyze the significance of sub -setting the data set by formation versus using the entire data. The complete data set model was compared to models built after segregating by formation. Figure 16 show the linear model with all the coefficients for each predictor parameter in the full data set model.

```
Call:
lm(formula = ROP1 ~ RPMA + WOBA + Flowin, data = training)

Residuals:
     Min       1Q   Median       3Q      Max
-104.252  -32.784   -7.441   25.781  252.373

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.2666806  0.9542999  34.860  < 2e-16 ***
RPMA        -0.2760878  0.0060375 -45.729  < 2e-16 ***
WOBA        -0.2470047  0.0323988  -7.624  2.5e-14 ***
Flowin       0.0982381  0.0008052 122.010  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50.01 on 63956 degrees of freedom
Multiple R-squared:  0.2067, Adjusted R-squared:  0.2067
F-statistic:  5556 on 3 and 63956 DF,  p-value: < 2.2e-16
```

Figure 16- Linear model parameters using the complete dataset of Well 12a.

a. Building Model using Train Dataset

b. Interpreting Model Performance

c. Prediction using Test Dataset

The model results give information about the direction (by the sign of coefficient), magnitude (by the value of coefficient) and statistical significance (p-values) of the relationship between a predictor and response. Low p-values imply strong relationship while higher p-values negate the same. Usually the cut-off (alpha) is around 5%. The regression results show that all the three predictors are significant because of their low p-values. But the R-squared and adjusted R-squared values are low, around 0.2, which means that the predictors explain about only 20% of the variance in the predicted values.

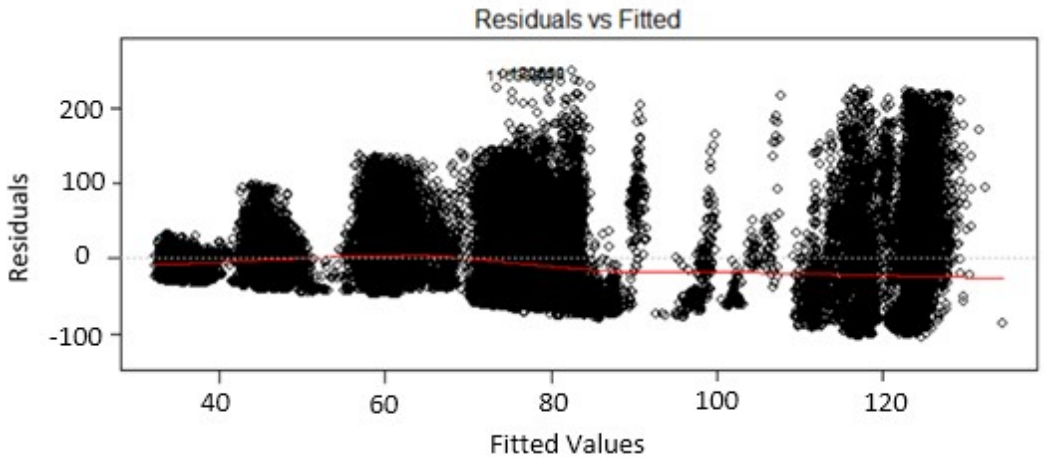


Figure 17- Residuals vs. Fitted plot using linear model on complete data of Well 12a.

Figure 17 is a standard residuals plot with which linearity and homoscedasticity can be evaluated. There should be a completely random, and equal distribution of points throughout the X axis and a flat red line failing which would result in heteroscedasticity. But there is no presence of equally spread residuals without distinct patterns, implying a non-linear relationship between the predictors and the response variable. As the fitted value increases, so does the spread which confirms the presence of heteroscedasticity and hence non-linearity.

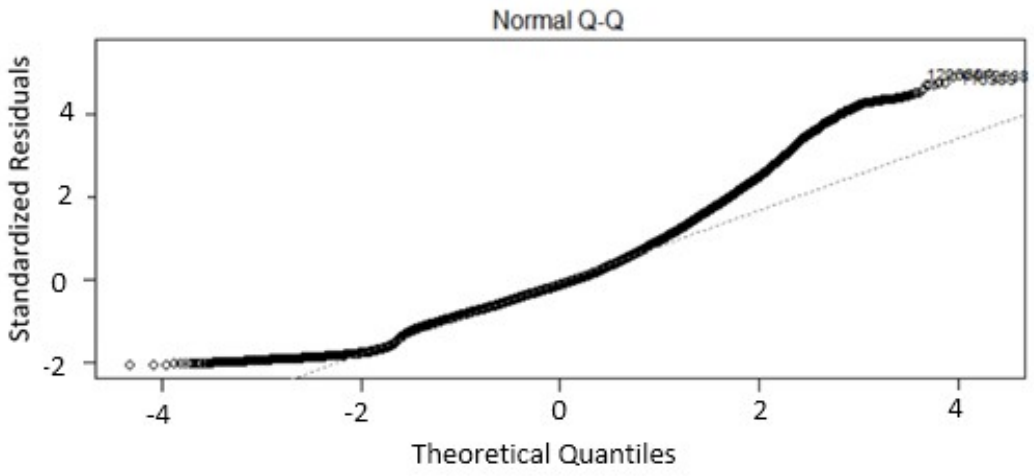


Figure 18- Quantile plot of complete data using the linear model on Well 12a.

Figure 18 is a normal quantile plot of the residuals, and ideally the residuals should be normally distributed. The residuals have to follow a normal distribution but in this figure, a lot of variation from the base line can be seen, which indicates the presence of non-linearity between predictors and response.

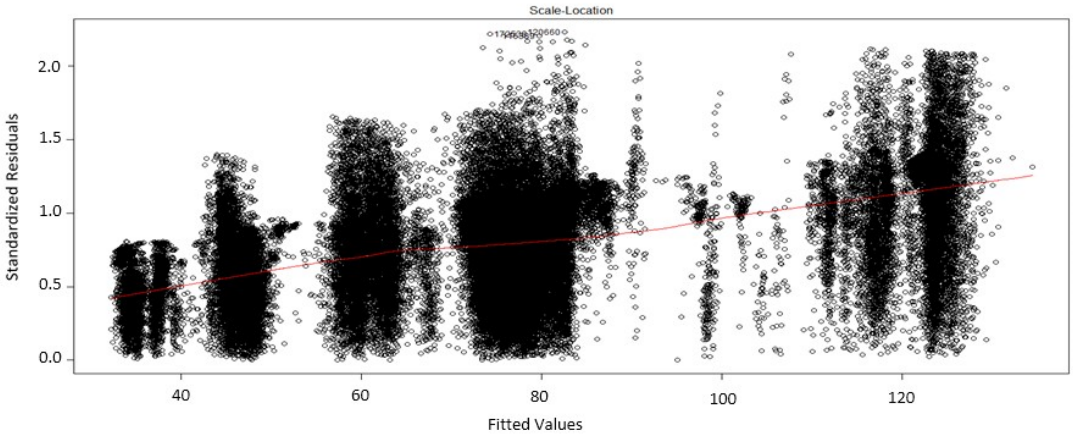


Figure 19-Spread -Location plot of complete data using the linear model on Well 12a.

Figure 19 is called Spread-Location plot. This plot illustrates whether residuals are spread equally along the ranges of predictors or not. It again confirms the presence of heteroscedasticity as the spread is increasing with fitted values.

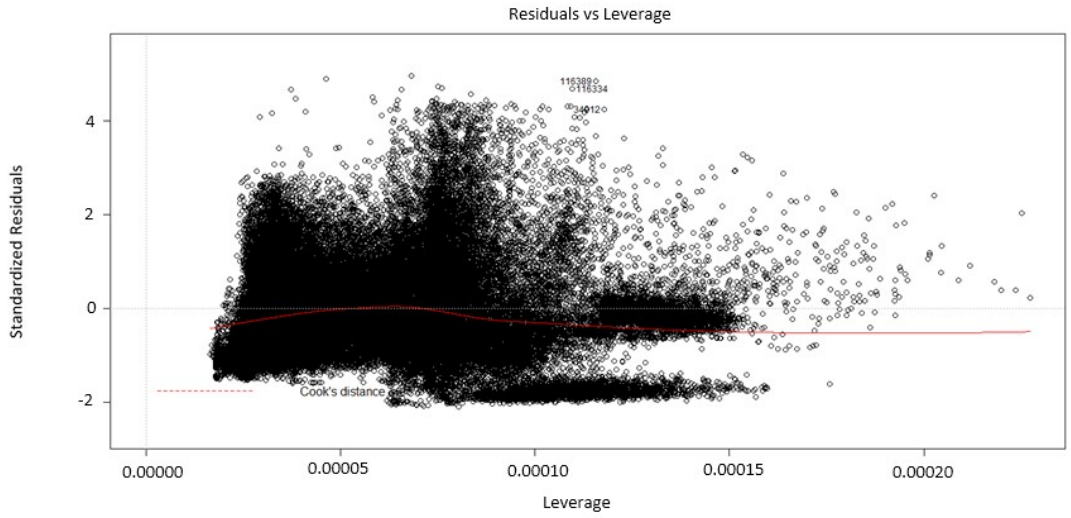


Figure 20- Residuals vs. Leverage plot of complete data using the linear model on Well 12a.

Figure 20 is called the Residuals vs. Leverage plot. This plot helps us to find influential data cases. The data can have extreme values and still not be influential to determine a regression line or can have reasonable value and still dominate.

The model above follows the typical scenario when there is no influential case, or cases. Cook's distance lines (represented by red dashed lines) are barely visible and all cases are inside of the Cook's distance lines. So there are not any specific cases influencing the model in a significant way.

The above four diagnostic plots show potential problematic cases with the row numbers in the data. These provide information about the model and data. Patterns in these plots tell that the current model might not be the best way to understand the data and perform predictive modeling. This would indicate perhaps there is a non-linear relationship between the predictors and the outcome. Polynomial regression or log transformations might better the model performance. And further employing machine learning methods such as SVR, KNN, trees, and RF is necessary to build better models.

```
[1] "All Data_ 3 Predictors_LM MODEL "

  County  RMSE   MAE MedAE Med % Mean % Adj. Rsqr
1      1 50.06 37.93 29.96 42.89  78.72      0.21
```

Figure 21- Prediction results of complete data using the linear model on Well 12a.

Figure 21 represents the RMSE and MAE after performing prediction on the entire data set. The RMSE and MAE errors were high, evident from the Percentage MAE and the percentage median of absolute errors as well. In order to improve the model and reduce the error metrics, investigation was done to see if clustering by formation would reveal any insight. Several models were built for each formation respectively as mentioned in the following sections.

3.1.1.2. Effect of Clustering by Formations

Data was segregated by formation and linear regression models were fitted for each formation separately. These models were then applied on the test datasets of each formation respectively and the error metrics were computed. Table 9 shows the summary of error metrics after prediction was performed. The overall adjusted R-squared has increased for most of the formations and is low for some bad performing formations. Combining the data as a whole prevented the capturing of such local formation specific trends which is essential for any good model. RMSE and MAE reduced from a global high of 50 and 38 in the previous case. The best formations defined by formations with RMSE < 10 are presented as well. For all the modeling techniques employed, clustering by formation was performed and individual models were built in order to achieve the best performance.

Table 9-Prediction results of clustering by formation using the linear model on Well 12a.

"TOTAL SUMMARY- LM MODEL "

County	RMSE	MAE	MedAE	Med %	Mean %	Adj. Rsqr	Algo
NA	64.597	51.475	43.8	33.2	60.5	0.18	4
Ekofisk	5.323	4.105	3.5	10.2	12.1	0.09	4
Tor	12.508	9.987	8.8	18.2	25.6	0.22	4
Hod	30.046	23.902	20.7	29.7	40.8	0.19	4
Herring	22.398	17.528	15.0	34.2	53.7	0.29	4
Plenus	56.001	42.690	35.9	32.2	33.8	0.29	4
Hidra	42.818	34.769	29.2	26.0	35.7	0.04	4
Sola	35.770	28.677	24.8	29.2	46.0	0.02	4
Valhall	33.676	26.722	22.5	23.6	34.1	0.31	4
Valhall_Limestone	18.009	14.183	12.2	33.5	46.1	0.01	4
Kimmeridge	8.047	6.209	5.2	15.6	20.3	0.44	4
Top	11.063	8.647	7.0	24.7	53.5	0.01	4
UB4	11.431	8.974	7.3	13.5	17.2	0.24	4
Heather	22.166	16.229	11.5	17.2	39.3	0.19	4
Sgiath	10.434	8.627	8.1	29.9	43.2	0.26	4
Smith	6.817	5.237	4.0	20.2	34.1	0.22	4

"BEST Formations - LM MODEL "

County	RMSE	MAE	MedAE	Med %	Mean %	Adj. Rsqr	Algo
Ekofisk	5.323	4.105	3.5	10.2	12.1	0.09	4
Tor	12.508	9.987	8.8	18.2	25.6	0.22	4
Kimmeridge	8.047	6.209	5.2	15.6	20.3	0.44	4
Top	11.063	8.647	7.0	24.7	53.5	0.01	4
UB4	11.431	8.974	7.3	13.5	17.2	0.24	4
Sgiath	10.434	8.627	8.1	29.9	43.2	0.26	4
Smith	6.817	5.237	4.0	20.2	34.1	0.22	4

3.1.1.3. Effect of Variable Interaction vs. No-interaction

Initially, two regression models were computed using ROP as the output and RPM, WOB and Flow as the inputs to check if there is an interactivity within predictors. Results of the Valhall formation are presented in Figure 22.

a. Interpreting Model Performance- No interactivity case

```
[1] "Valhall"

Call:
lm(formula = ROP1 ~ RPMA + WOBA + Flowin + GR, data = training)

Residuals:
     Min       1Q   Median       3Q      Max
-129.094  -23.416   -2.719   20.658  145.087

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -744.44695   32.23075 -23.097  < 2e-16 ***
RPMA           2.14127    0.11914  17.973  < 2e-16 ***
WOBA           3.72047    0.08230  45.208  < 2e-16 ***
Flowin         0.52124    0.03186  16.359  < 2e-16 ***
GR             0.25411    0.04446   5.716 1.13e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.61 on 7759 degrees of freedom
Multiple R-squared:  0.3133, Adjusted R-squared:  0.3129
F-statistic: 884.9 on 4 and 7759 DF,  p-value: < 2.2e-16
```

Figure 22-Linear model parameters considering no interactivity among predictors for Valhall formation in Well 12a.

The t-value is the ratio between the coefficient and its standard error. A large t-value suggests that the coefficient estimate is large and different from zero while a smaller t-value fails to show that the predictor has any influence on the response. The t-value is used in the computation of p-values. The p-values indicate that all the 3 predictors are very important.

b. Interpreting Model Performance- interactivity case

```
[1] "Valhall"

Call:
lm(formula = ROP1 ~ RPMA * WOBA * Flowin * GR, data = training)
Residuals:
     Min       1Q   Median       3Q      Max
-116.711  -23.308   -3.187   20.228  142.491

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          7.814e+04  2.927e+04   2.670  0.00761 **
RPMA                -5.376e+02  1.940e+02  -2.771  0.00560 **
WOBA                -9.734e+03  2.420e+03  -4.021 5.84e-05 ***
Flowin              -8.997e+01  3.343e+01  -2.691  0.00713 **
GR                  -9.962e+02  3.371e+02  -2.955  0.00314 **
RPMA:WOBA            6.572e+01  1.597e+01   4.115 3.92e-05 ***
RPMA:Flowin          6.191e-01  2.216e-01   2.794  0.00521 **
WOBA:Flowin          1.107e+01  2.769e+00   3.998 6.43e-05 ***
RPMA:GR              6.798e+00  2.235e+00   3.041  0.00237 **
WOBA:GR              1.228e+02  2.738e+01   4.487 7.34e-06 ***
Flowin:GR            1.143e+00  3.850e-01   2.969  0.00300 **
RPMA:WOBA:Flowin    -7.474e-02  1.827e-02  -4.090 4.35e-05 ***
RPMA:WOBA:GR        -8.273e-01  1.807e-01  -4.577 4.78e-06 ***
RPMA:Flowin:GR      -7.796e-03  2.553e-03  -3.054  0.00226 **
WOBA:Flowin:GR      -1.397e-01  3.132e-02  -4.461 8.27e-06 ***
RPMA:WOBA:Flowin:GR  9.413e-04  2.068e-04   4.552 5.40e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.32 on 7748 degrees of freedom
Multiple R-squared:  0.318,   Adjusted R-squared:  0.3167
F-statistic: 240.8 on 15 and 7748 DF,  p-value: < 2.2e-16
```

Figure 23-Linear Model parameters considering interactivity among predictors for Valhall formation in Well 12a.

Figure 23 shows the regression model considering interactivity. Here, the t-values and p-values for a combination of predictors give an insight about their relative contribution. The asterisk sign besides each variable represents the significance codes, therefore indicating the relative contribution of that particular predictor to the model, making it simple to interpret the model. So for Valhall formation, considering interactivity helped explain the model better, as there is considerable interaction between all the factors. In Valhall, both R square and adjusted R square went up in the case of the interactivity model but not by a large factor, indicating that

the models perform better when interaction between predicitors is considered. The p-values also indicate that the interactivity is better.

c. Investigation using ANOVA

ANOVA (Analysis of Variance) was performed to see if there truly is a statistical difference between both the models for all formations, as claimed in the previous sections.

```
[1] "Hod"
Analysis of Variance Table

Model 1: ROP1 ~ RPMA + WOBA + Flowin
Model 2: ROP1 ~ RPMA * WOBA * Flowin
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1   4966 4474235
2   4962 4468792  4    5442.2 1.5107 0.1962


[1] "Valhall"
Analysis of Variance Table

Model 1: ROP1 ~ RPMA + WOBA + Flowin
Model 2: ROP1 ~ RPMA * WOBA * Flowin
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1   7760 8720079
2   7756 8674414  4     45666 10.208 3.059e-08 ***
```

Figure 24- ANOVA results for Hod (above) and Valhall formations (below) in Well 12a.

Figure 24 lists the results of ANOVA for 2 formations HOD and Valhall. The F statistic is a ratio of 2 different measures of variance for the data. F statistic is 1 when null hypothesis is true implying that above models are both estimates of the same thing.

Valhall has a good F value and a low p-value suggesting that it rejects the null hypothesis and considering Interactivity is statistically relevant. But HOD formation has a high p-value with an F-value close to 1, suggesting that it accepts the null hypothesis and hence, interactivity did not

improve the model by a considerable degree. Therefore, whether interactivity exists and is statistically relevant varies from formation to formation as explained.

d. Prediction using Test Datasets

Finally, prediction was perfomed using the 2 models on test datasets and the results are summarized in Tables 10 and 11.

Table 10-Prediction results without considering interactivity among predictors in Well 12a.

"TOTAL SUMMARY- LM MODEL "

County	RMSE	MAE	MedAE	Med %	Mean %	Adj. Rsqr	Algo
NA	66.384	52.736	44.9	34.6	62.2	0.19	lm
Ekofisk	5.509	4.200	3.4	10.2	12.1	0.09	lm
Tor	12.752	10.175	8.4	17.9	25.9	0.23	lm
Hod	29.493	22.963	19.0	27.5	40.4	0.21	lm
Herring	21.601	16.904	13.5	33.2	50.9	0.28	lm
Plenus	43.676	35.600	30.2	24.7	31.4	0.29	lm
Hidra	39.814	32.112	27.7	23.7	33.4	0.07	lm
Sola	34.850	27.858	23.8	28.9	43.6	0.05	lm
Valhall	33.146	26.435	22.8	23.6	33.8	0.31	lm
Valhall_Limestone	17.122	13.269	10.8	29.8	42.4	0.11	lm
Kimmeridge	8.500	6.546	5.2	16.1	21.4	0.46	lm
Top	10.664	8.354	6.9	24.4	48.1	0.03	lm
UB4	11.711	9.108	7.2	14.0	17.3	0.25	lm
Heather	22.456	16.452	12.3	17.5	40.3	0.20	lm
Sgiath	9.948	8.059	6.4	23.9	38.5	0.27	lm
Smith	7.068	5.517	4.2	22.1	35.5	0.22	lm

Table 11- Prediction results considering interactivity among predictors in Well 12a.

"TOTAL SUMMARY- LM MODEL "

County	RMSE	MAE	MedAE	Med %	Mean %	Adj. Rsqr	Algo
NA	64.971	51.239	42.3	33.3	60.3	0.21	lm
Ekofisk	5.513	4.173	3.3	10.0	12.1	0.13	lm
Tor	12.379	9.843	8.3	17.9	25.6	0.25	lm
Hod	29.906	23.470	19.8	28.9	40.6	0.21	lm
Herring	22.452	16.995	13.4	32.8	51.4	0.32	lm
Plenus	40.455	31.199	22.6	21.5	26.3	0.51	lm
Hidra	38.193	29.791	24.6	21.3	30.8	0.19	lm
Sola	33.923	26.765	22.6	26.8	42.4	0.13	lm
Valhall	33.288	26.071	22.0	22.6	33.6	0.32	lm
Valhall_Limestone	16.870	13.214	10.9	30.1	42.6	0.12	lm
Kimmeridge	7.915	6.170	5.0	15.9	19.8	0.48	lm
Top	9.743	7.493	5.9	21.3	41.5	0.19	lm
UB4	11.203	8.703	7.1	12.8	16.7	0.28	lm
Heather	21.557	15.656	11.0	16.1	35.3	0.26	lm
Sgiath	15.017	10.489	7.6	28.1	43.2	0.31	lm
Smith	6.883	5.320	4.0	22.0	31.8	0.25	lm

The error rates as shown indicate that interactivity improved the error rates in some formations while it did not in others, supporting the results of ANOVA analysis performed above. In conclusion, the exercise implies that the ROP follows a complex relationship with the predictors that also varied with each formation.

One drawback considering interactivity is the computation time involved when considering interactivity. Another factor is that there was no considerable difference noticed when interactivity was considered. Since the inclusion of interactivity yielded mixed results, models without interaction were considered for further usage in algorithm construction.

3.1.2. Stepwise Regression

Stepwise regression is used to determine the variable significance, the last step of the investigation process as described in the beginning of this chapter. In this technique, the selection of independent variables is achieved by automatically observing statistical values like R-square, t-stats etc. to discern significant variables. Stepwise regression fits the regression model by adding/dropping predictors one at a time based on a specified criterion. The aim of this modeling technique is to maximize the prediction power with a minimum number of predictor variables.

3.1.2.1. Interpreting Model Performance

Predictors RPM, WOB, Flow and GR are considered in stepwise regression. The technique was applied on all formations and the model details are presented for two formations – Ekofisk and Kimmeridge in Table 12. Figure 25 illustrates the relative contribution of each predictor for formations, Ekofisk and Kimmeridge, respectively.

Table 12- Stepwise model parameters for Ekofisk (left) and Kimmeridge (right) formations.

```
[1] "Ekofisk"
Start:  AIC=9533.12
ROP1 ~ RPMA + WOBA + Flowin + GR

          Df Sum of Sq   RSS    AIC
- GR       1       1.5 85163 9531.2
<none>                 85162 9533.1
- WOBA     1     108.7 85270 9534.7
- RPMA     1     516.7 85678 9548.0
- Flowin   1    4004.6 89166 9659.0

Step:  AIC=9531.17
ROP1 ~ RPMA + WOBA + Flowin

          Df Sum of Sq   RSS    AIC
<none>                 85163 9531.2
- WOBA     1     119.5 85283 9533.1
- RPMA     1     648.9 85812 9550.3
- Flowin   1    4309.0 89472 9666.6
```

```
[1] "Kimmeridge"
Start:  AIC=21026.95
ROP1 ~ RPMA + WOBA + Flowin + GR

          Df Sum of Sq    RSS    AIC
- Flowin   1         0 333708 21025
- RPMA     1        14 333722 21025
<none>                333708 21027
- GR       1      1073 334781 21041
- WOBA     1    101977 435685 22359

Step:  AIC=21024.95
ROP1 ~ RPMA + WOBA + GR

          Df Sum of Sq    RSS    AIC
- RPMA     1        16 333724 21023
<none>                333708 21025
- GR       1      1149 334858 21040
- WOBA     1    227519 561227 23624

Step:  AIC=21023.18
ROP1 ~ WOBA + GR

          Df Sum of Sq    RSS    AIC
<none>                333724 21023
- GR       1      1251 334975 21040
- WOBA     1    230131 563855 23646
```

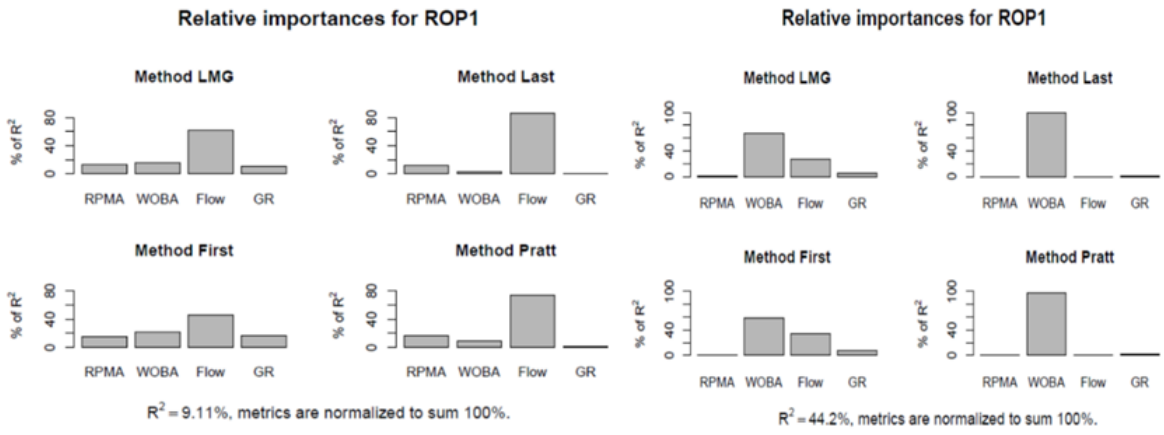


Figure 25-Relative importance plots for Ekofisk (left) and Kimmeridge (right) formations.

For Kimmeridge, the stepwise regression chosen predictors are GR and WOB while for Ekofisk, WOB, Flow and RPM were chosen. The process suggested that the chosen predictors were different for each formation. Next, prediction was done using both models to see if there was a change in error metrics.

3.1.2.2. Prediction using Test data

Both the models – one with inclusion of important predictors only (Stepwise regression) and the other with all predictors (linear regression) were computed. Figure 26 summarizes the results of linear vs. stepwise regression RMSE metric for all formations.

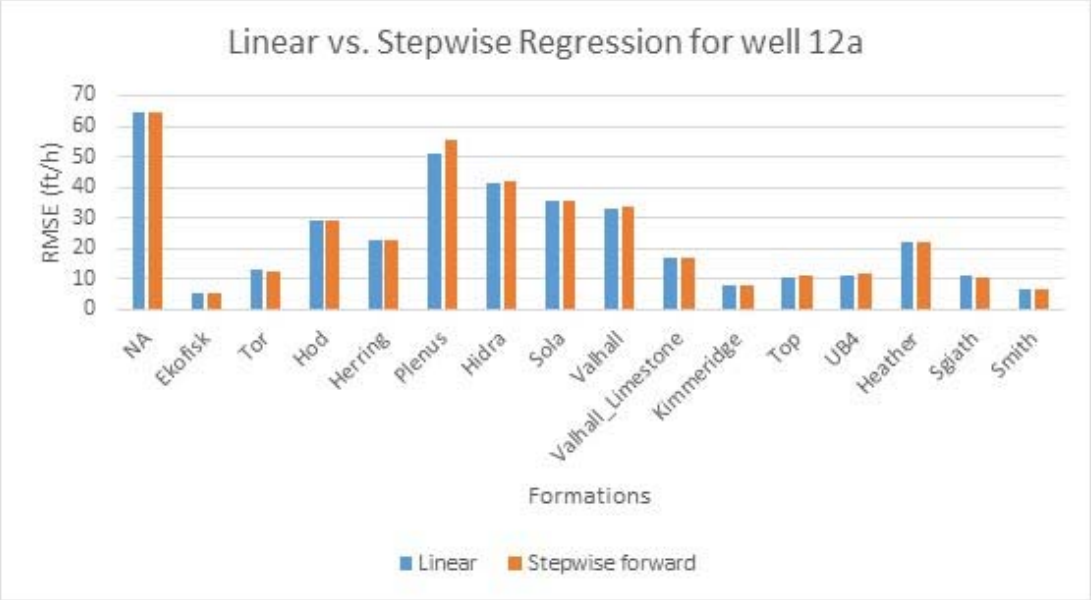


Figure 26-Linear vs. stepwise regression results of all the formations in Well 12a.

There is no significant variation in using only the important variables as stepwise performed almost the same or worse than linear regression models. These results also varied for each formation and hence, it once again emphasizes the importance of building separate models for each formation. Although error rates remained the same, knowledge of the important predictors would help optimize ROP on a real time basis if the driller can get knowledge of the most relevant factors affecting ROP.

3.1.2.3. Investigation using ANOVA

Another analysis was performed with and without considering GR data. GR was chosen for this purpose because usually when drilling data is avaliable it includes RPM, WOB and Flow while

GR data is available only with formation data. Another reason is that most of the numerical models use only WOB, RPM and Flow. So this process would determine if the presence of GR would help to create a better model. ANOVA was performed to see if there truly is a statistical difference between changing predictors – including and excluding GR.

```
[1] "Ekofisk"
Analysis of Variance Table

Model 1: ROP1 ~ RPMA + WOBA + Flowin
Model 2: ROP1 ~ RPMA + WOBA + Flowin + GR
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1   2780 85163
2   2779 85162  1    1.4638 0.0478  0.827
```

```
[1] "Kimmeridge"
Analysis of Variance Table

Model 1: ROP1 ~ RPMA + WOBA + Flowin
Model 2: ROP1 ~ RPMA + WOBA + Flowin + GR
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   5000 334781
2   4999 333708  1    1073.1 16.075 6.176e-05 ***
```

Figure 27- ANOVA results for Ekofisk and Kimmeridge formations with and without GR

Figure 27 presents the ANOVA details with and without GR. For Ekofisk, the F- value is low and P- value high suggesting that there was no change when GR was included while for Kimmeridge, the opposite results was observed. Upon repeating the process for all other formations, it was determined that about 50% of the time GR had influence on making the model better.

Hence, GR was considered because the given data already had formation values and in case where it's missing, the other 3 parameters were used for modeling.

3.1.3. Conclusions of regression methods and preliminary analysis

1. Sub-setting by formation from entire dataset helped build better predictive models and explain variation in the data compared to using complete well data for a single model construction.
2. The analysis of diagnostic plots has revealed that following a linear approach modeling would be error prone with the case study data. Hence, there emerged a need to employ machine learning models and other advanced ensemble methods.

3. Models built with and without interaction within the variables revealed mixed results as far as reducing error rates were concerned. There was little variation in RMSE/ MAE between both the methods and this also varied across formations. So for the sake of achieving better computation speeds, non-interactivity would be employed across further modeling.
4. Relative importance revealed that among ROP, RPM, Flow and GR would be relevant in model building in some formations and not so much in others as they had varied results for each formation. The reduction in error metrics was not very significant as well. So all the predictors will be considered if GR data is available for a given formation/well.

The following chapters discuss the application of advanced modeling algorithms such as Neural Networks, Decision trees, Instance based methods and Ensemble Models. As concluded from above, non-interactivity would be followed across all models. WOB, RPM and flow shall be used as default predictors in all cases as they are input at the surface real time. It is therefore logical to include these parameters even though relative importance gave weightage to one or more among these. GR shall be used if available for a particular well/formation.

3.2 Machine Learning Methods

3.2.1. Support Vector Regression (SVR)

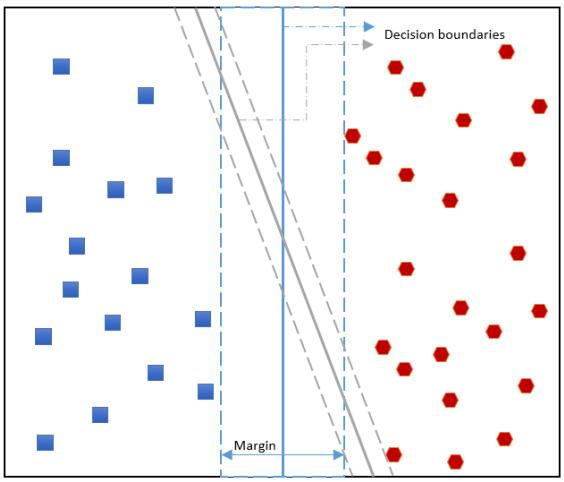


Figure 28- Hyper planes of SVM

SVR is used to construct a hyperplanes or a set of hyperplanes in a high-dimensional or infinite dimensional space, which can be used for regression, classification etc. In a sample space as shown in Figure 28, there are infinitely many possible hyperplanes and to find the best combination, SVM relates the margin of each possible hyperplane to its generalization error using a statistical principle called structural risk minimization (SRM). SRM defines an upper boundary to the generalization error and hence the maximum margin for the hyperplane in terms of its training error, number of training samples and the model complexity. (Tan et al. 2006).

To be able to use SVM, the package e1071 needs to be installed in R. In this section, a support vector machine algorithm was applied to the Well 12a. The Holdout method was used for splitting data into training and testing. Later, how to better the algorithm through tuning will be discussed.

Since the process of extraction, processing, data splitting and model building has already been described in the preceding chapters, the prediction results for all models will be directly presented.

3.2.1.1. Base Model

The four predictors: RPM, Flow, WOB, and GR were used to predict ROP. Table 13 lists the results of the base model (using default values for all parameters without tuning). SVR was more efficient with lesser error metric values and a better prediction. Tuning was then performed to see if the models for formation can further be improved.

Table 13-Prediction results of SVR by formation for Well 12a.

County	RMSE	MAE	MedAE	Med %	Mean %
NA	56.61	42.17	32.45	26.62	42.71
Ekofisk	5.39	3.98	3.06	9.20	11.31
Tor	12.22	9.44	7.58	17.20	22.76
Hod	28.47	21.46	16.87	25.00	35.08
Herring	21.71	15.60	11.71	28.72	39.70
Plenus	42.63	30.92	23.10	17.10	23.50
Hidra	37.94	29.49	23.92	22.24	28.61
Sola	30.32	23.15	18.22	23.35	34.24
Valhall	30.97	23.78	18.82	20.41	29.26
Valhall_Limestone	16.53	12.25	9.00	26.38	36.19
Kimmeridge	7.84	5.81	4.26	13.93	17.84
Top	8.55	6.21	4.58	18.04	31.34
UB4	10.91	8.50	7.13	13.69	15.90
Heather	20.55	14.17	9.74	14.53	34.65
Sgiath	8.59	5.98	4.31	17.77	21.82
Smith	6.48	4.75	3.54	19.98	25.56

3.2.1.2. Tuned Model

The performance of the support vector regression can be further improved by a process called hyper parameter optimization, or model selection. Through this process, the best parameters for the model can be selected- epsilon (єє) and a cost parameter to avoid overfitting. Usually grid search is used for this purpose. The four predictors- RPM, Flow, WOB, and GR were used to predict ROP using SVR. The darkest area in Figure 29 (right) represents the best combination. The results improved for most of the formations over the linear regression models as SVR captures non-linearity better for most of the formations in well 1. The drawback with SVR was mainly computation time.

```
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  epsilon cost
##      0.7    4
##
## - best performance: 88.74364
```

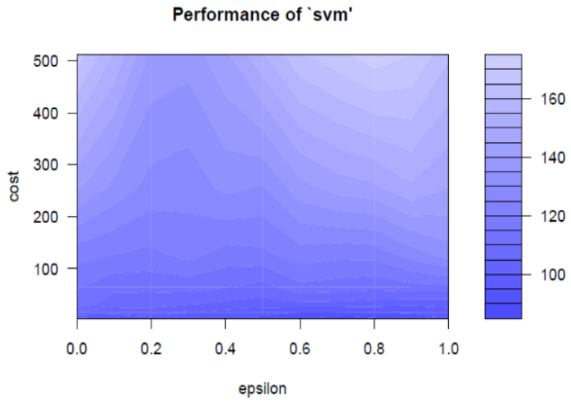


Figure 29- Tuning Parameters: cost and epsilon of SVR for Sgiath Formation in Well 12a.

Since the process of tuning was time consuming (close to 4-5 hours for a couple of formations), this process was repeated on a select group of formations. Figure 30 provides a comparison between SVR (tuned models) and Linear Regression.

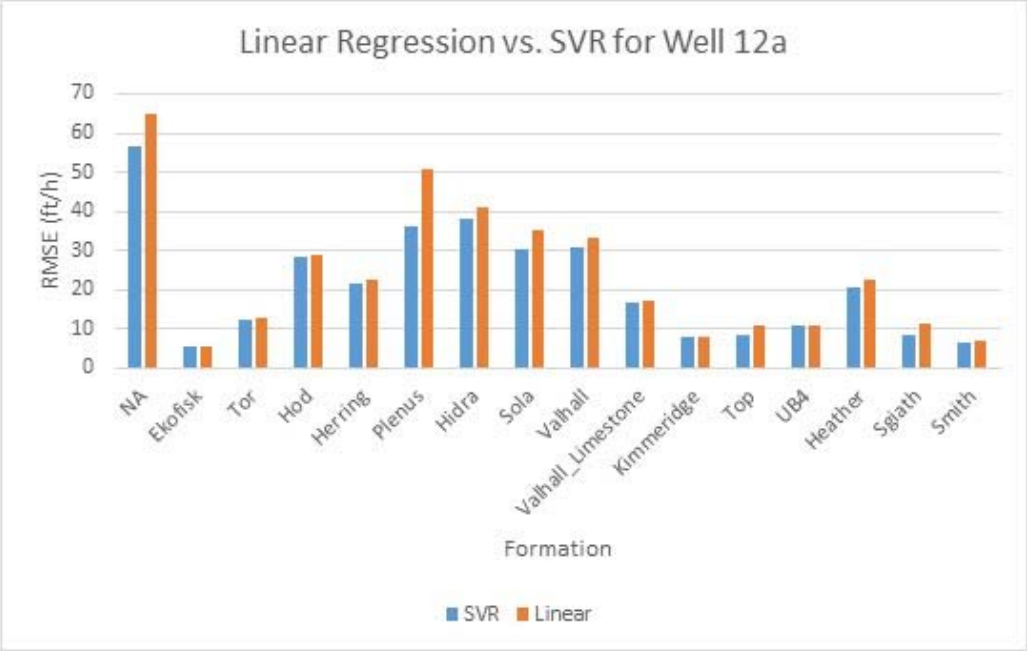


Figure 30-Linear Regression vs. SVR for all the formations in Well 12a.

The results improved for most of the formations over the linear model as SVR captures non-linearity better. For most of the formations, SVR seemed to perform better although the improvement in error rates was minimal.

3.2.1.3. Prediction Plot

Figure 31 shows the plot between actual and predicted ROP for SVR model for the formation Sgiath while Figure 32 shows the same for a linear model.

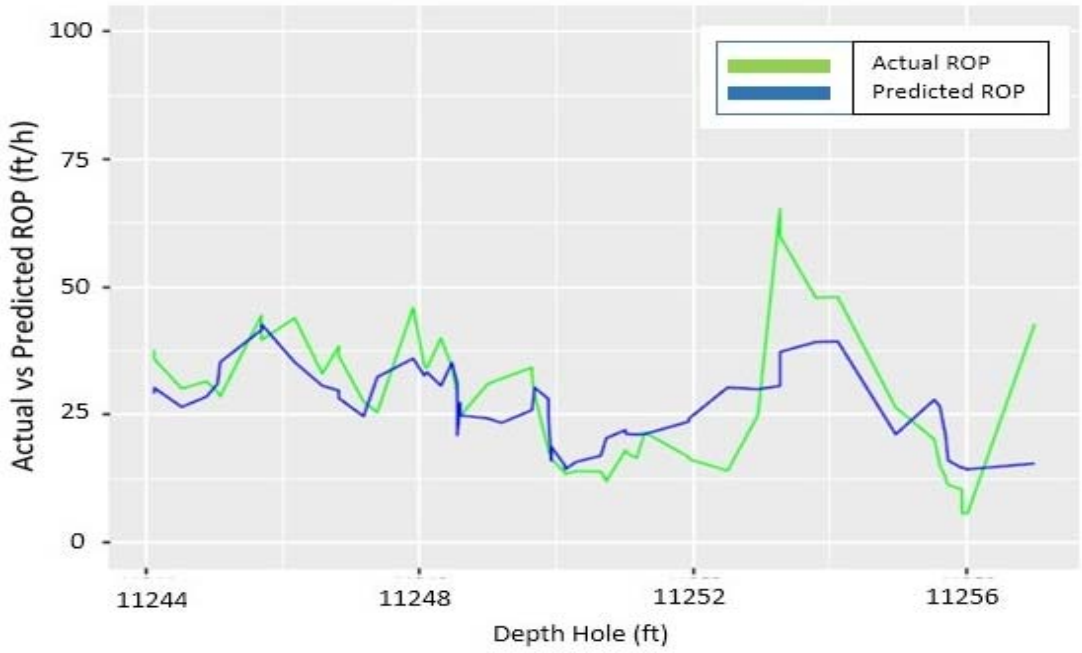


Figure 31- Actual vs. Predicted ROP for Sgiath formation using SVR in Well 12a.

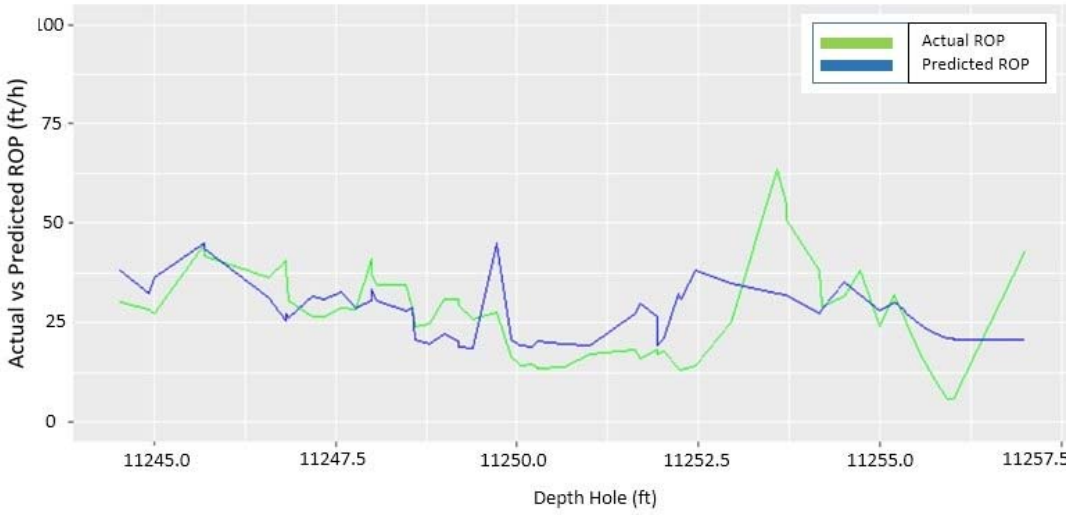


Figure 32- Actual vs. Predicted ROP for Sgiath formation using Linear Model in Well 12a.

As seen from the above figures, SVR captures non-linearity in the data better than the linear regression model. The next section discusses the application of Neural Network and KNN algorithms.

3.2.2. K-Nearest Neighbor (KNN)

One of the simplest machine learning algorithms is the KNN algorithm. It is one of the methods of instance-based learning, wherein new data points are classified on the basis of stored, labeled instances. Some kind of similarity measure, typically expressed by a distance measure such as Euclidean distance, cosine similarity or the Manhattan distance, is used to compute the distance between the store data and the new instance (data point). The KNN algorithm calculates and adds the distance of a new point to all stored data points, and then the distance values are sorted and the k-nearest neighbors are computed. After gathering the labels of these neighbors, a vote is taken. For regression, mean of k-nearest neighbors is assigned to new data point. One of the uses of KNN is when outliers are involved. It is insensitive to outliers that makes it resilient

to errors in the classification process. The efficacy of the model is greatly dependent on the number of nearest neighbors chosen i.e. determining the value of k.

3.2.2.1. Base Model

The four predictors- RPM, Flow, WOB, and GR were used to predict ROP. Table 14 lists the results of the base model (using default values for all parameters without tuning). The package caret is used to undertake the KNN model building and the function knnreg () is applied. KNN also performed better than the linear model and its results are closer to SVM.

Table 14-Prediction results of KNN for all formations in Well 12a.

County	RMSE	MAE	MedAE	Med %	Mean %
NA	49.042	35.765	25.8	21.4	37.1
Ekofisk	5.807	4.396	3.5	10.2	12.8
Tor	11.708	9.144	7.4	17.1	22.0
Hod	30.900	23.961	19.1	29.0	40.2
Herring	22.941	16.749	12.0	31.1	46.1
Plenus	45.852	32.015	19.6	18.5	25.8
Hidra	41.660	32.064	25.4	22.2	33.3
Sola	30.856	23.448	18.1	22.7	35.6
Valhall	32.518	25.012	19.7	20.7	31.7
Valhall_Limestone	17.943	13.945	11.4	31.3	44.1
Kimmeridge	7.545	5.697	4.5	13.9	18.1
Top	8.941	6.559	4.9	18.5	32.2
UB4	10.444	8.203	6.8	12.7	15.5
Heather	21.062	15.615	12.0	18.0	32.9
Sgiath	11.155	8.161	5.9	22.8	36.0
Smith	6.849	5.279	4.2	22.0	31.5

3.2.2.2. Tuned Model

The efficacy of the model is determined by k i.e. number of nearest neighbors. A large k value helps reduce the variance due to the noisy data. However, the side effect of a large k value includes the development of a bias due to which model might ignore smaller patterns in data containing useful information. Tuning was performed by iterating over several values of k to see

improve the efficacy of the KNN algorithm. So in the cases where there are constraints on time and computational power, a value of k equal to square root of the observations can be used directly. RMSE vs. k values are presented in Figure 33 for formation Ekofisk, where the tuning was done upon k values from 1 to 300.

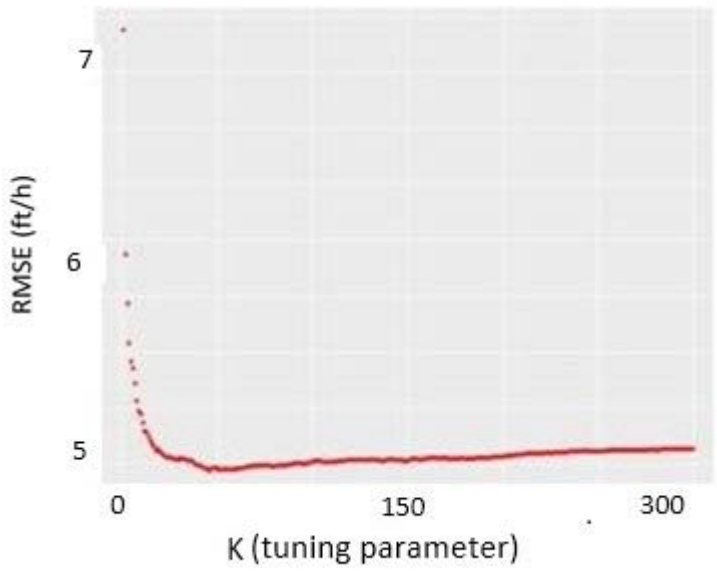


Figure 33- RMSE vs. k-min (tuning parameter) for KNN formation Ekofisk in Well 12a.

The error rate approaches a minimum at around 50 k, which is almost equal to square root of the number of observations in Ekofisk. So in the cases where there could be constraints on time and computational power, a value equal to the square root of the observations for k could be used directly.

Table 15 includes the final output using the KNN algorithm in which tuning was applied on each formation separately. The last column gives the values of the most optimum value for k. As compared with the previous section, the RMSE and MAE metrics have decreased for most of the formations.

Table 15-Prediction results of KNN by formation after applying tuning in Well 12a.

County	RMSE	MAE	MedAE	Med %	Mean %	K-min
NA	47.662	35.092	26.0	21.5	36.4	16
Ekofisk	5.201	4.030	3.4	9.9	11.8	38
Tor	11.197	8.790	7.2	16.3	21.6	23
Hod	29.489	23.029	19.5	28.2	38.3	100
Herring	20.376	15.462	12.2	32.1	44.5	76
Plenus	51.025	38.391	27.2	22.2	30.2	2
Hidra	40.635	32.286	27.5	24.7	32.8	31
Sola	30.150	23.000	18.5	22.6	35.2	19
Valhall	31.744	24.665	19.5	20.9	31.4	38
Valhall_Limestone	17.623	13.718	11.3	30.9	44.3	36
Kimmeridge	7.209	5.491	4.4	13.5	17.4	15
Top	8.421	6.159	4.7	17.4	33.3	17
UB4	10.984	8.651	7.2	13.5	16.5	37
Heather	20.083	14.816	11.2	17.2	33.5	11
Sgiath	9.245	6.938	4.7	20.9	32.7	7
Smith	5.937	4.608	3.6	18.8	28.0	16

3.2.2.3. Prediction Plot

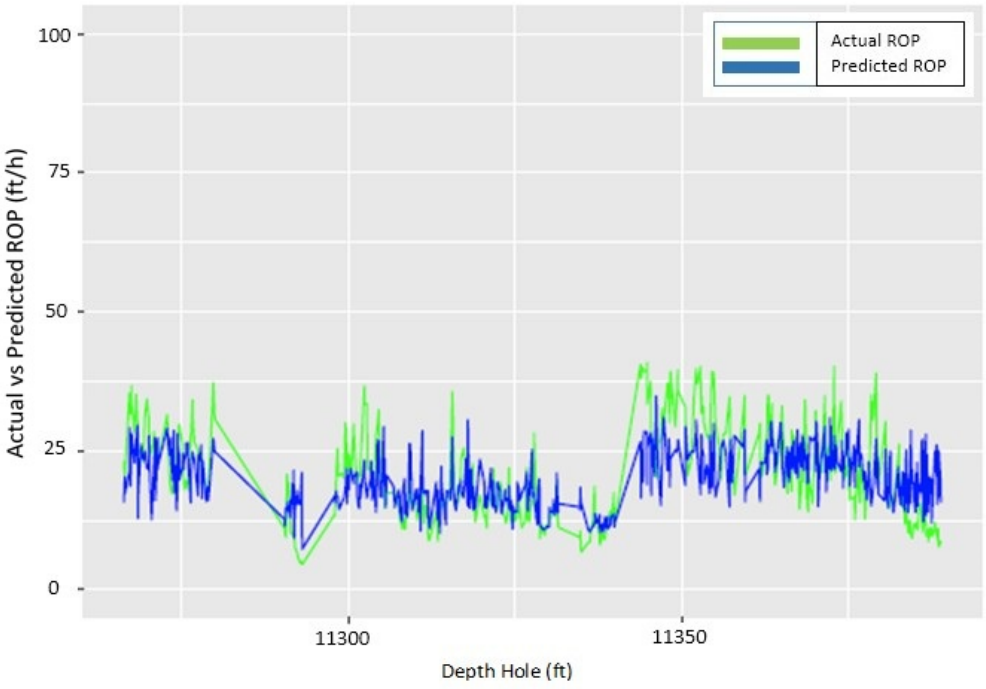


Figure 34-Actual vs. Predicted ROP for Smith formation using KNN in Well 12a.

Figure 34 shows the plot between actual and predicted ROP for the KNN model for the Smith formation. The plot captures the non-linear trend better than the linear model and similar to SVR. The main advantage of KNN over SVR is that KNN is simpler and faster.

3.2.3. Neural Networks

Neural networks are quite popular because of their hidden complexity and resemblance to the working of neurons and human brain structure. A neural network (Bayesian Regularized Neural Network or BRNN) is a graph of computational units that receive inputs and transfer the result into an output that is passed. The features of an input vector are connected to the features of an output vector by ordering the units into layers, as shown in Figure 35. It is possible to design and train neural networks to the model the primary relationship in the data with training such as the Back-Propagation algorithm. Neural networks were also applied on the training dataset. The following packages nnet and brnn in R were used to undertake neural network model building and prediction. Initially base models with the default number of neutrons were run and then tuning was incorporated into model performance.

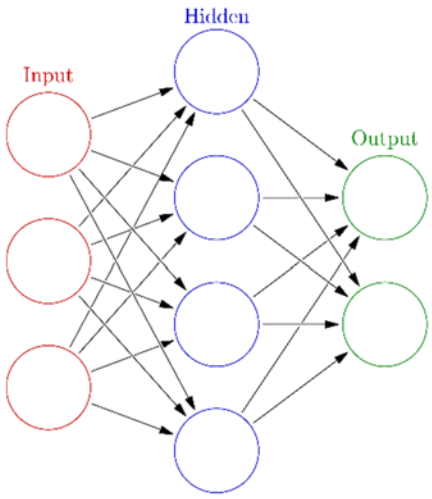


Figure 35- Pictorial description of a simple neural network.

3.2.3.1. Tuned Model

Tuning is usually done either by employing a tune grid search or manually altering the value of neurons. In this study, a loop for different values of neurons was run and then the best neurons with the least RMSE, as shown in Figure 36. As compared with KNN and SVR, NN did not perform better in either the prediction results or in the actual vs. predicted ROP plots. However, NN performed better compared to regression models: linear and stepwise.

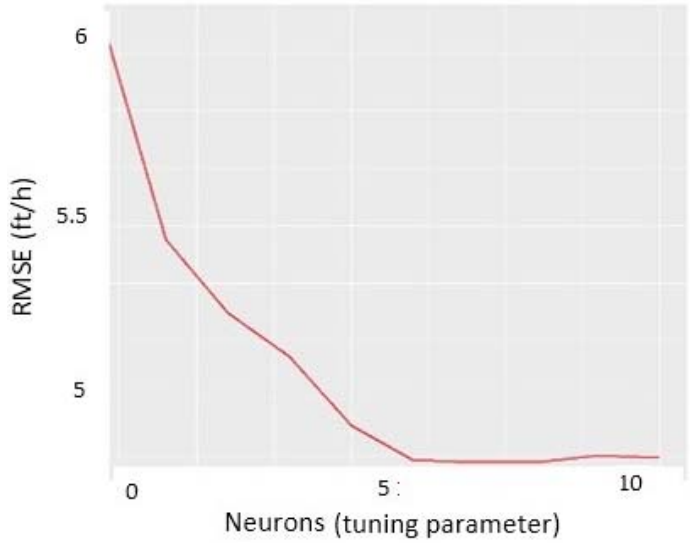


Figure 36-RMSE vs. Neurons (tuning parameter) using NN for Smith formation in Well 12a.

Table 16 shows the results of prediction and Figure 37 shows the predicted vs. actual ROP plot. As compared with KNN and SVR, NN did not perform better in either the prediction results or in the actual vs. predicted ROP plots. However, NN performed better compared to regression models.

Table 16-Prediction results of NN for all formations in Well 12a

"TOTAL SUMMARY"

County	RMSE	MAE	MedAE	Med %	Mean %
NA	54.422	41.819	32.9	26.5	46.2
Ekofisk	5.292	4.097	3.3	9.9	12.1
Tor	11.716	9.431	8.2	18.2	24.2
Hod	29.695	23.344	19.3	28.2	39.8
Herring	21.125	15.690	12.3	30.5	44.7
Plenus	36.932	29.611	24.0	18.2	24.4
Hidra	40.554	31.760	25.0	23.2	31.1
Sola	31.599	24.345	19.5	23.5	38.1
Valhall	32.349	25.076	20.5	21.2	32.6
Valhall_Limestone	16.863	13.440	11.7	31.6	44.7
Kimmeridge	7.577	5.728	4.5	14.0	18.0
Top	9.333	6.958	5.3	19.4	40.4
UB4	10.689	8.411	7.3	13.9	16.3
Heather	20.838	14.812	10.3	15.4	32.8
Sgiath	9.558	7.867	7.4	27.9	35.8
Smith	6.254	4.870	3.8	20.6	28.8

3.2.3.2. Prediction Plot

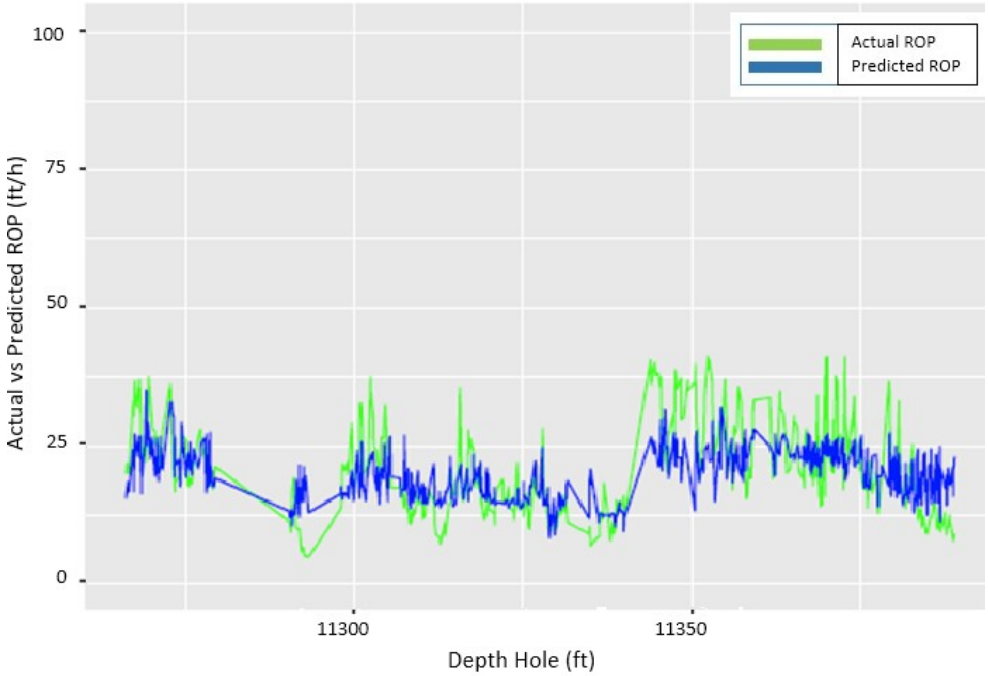


Figure 37-Actual vs. predicted ROP for Smith formation using NN model in Well 12a.

3.2.4. Classification and Regression Tree (CART)

A decision tree or regression tree is a simple yet popular machine learning algorithm. To build the tree, a root node containing all the training data is taken and split into two new nodes on the basis on the most important variable which separates the outcomes into two categories in the best possible manner. Each new node is again split based on the variable that provides the better splits for that particular node (which is not necessarily the same variable as the one that the analysis was started with). This process of nodes splitting is continued until stop criteria is achieved. Decision trees are popular for certain reasons. They do not make any assumptions about linearity and consider both linear and non-linear models. Their interpretability is quite straight forward. But the downside is they tend to over fit. Usually fitting is done through cross validation technique.

3.2.4.1. Tuned Model

Since testing was completed at the same time the tree was grown, an error measurement was used to find the optimal number of splits. The original tree was then pruned, and only the optimal number of splits were retained. This helped optimize the computational time, especially when large datasets with a number of predictors were involved. Figures 38 and 39 shows the tree construction of the formation Sgiath before pruning and after pruning respectively.

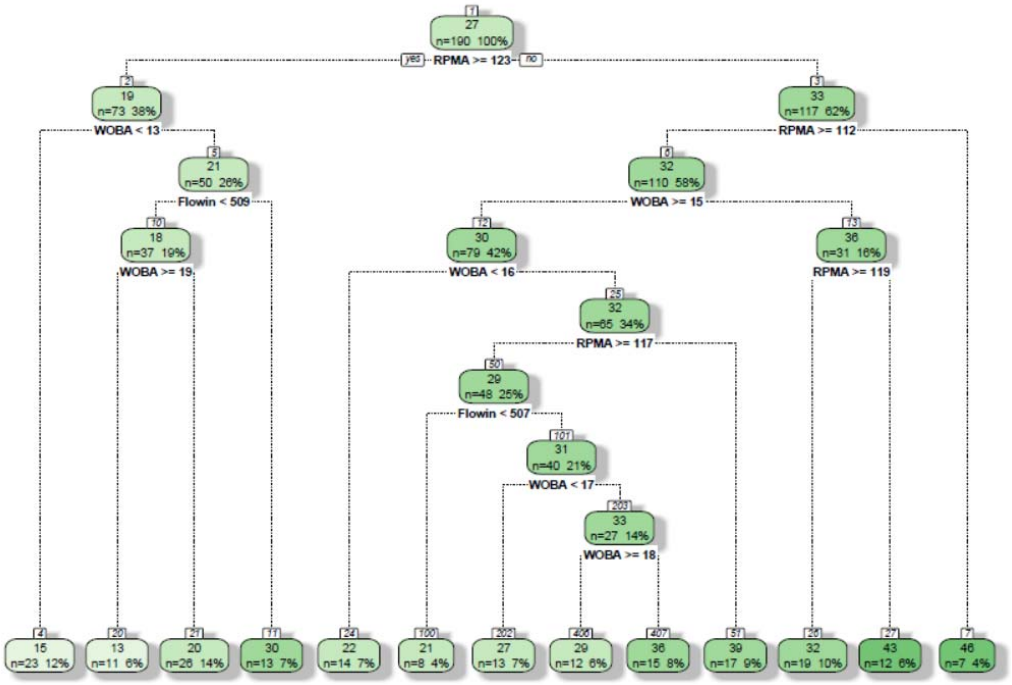


Figure 38- Decision tree representation for formation Sgiath before applying tuning in Well 12a.

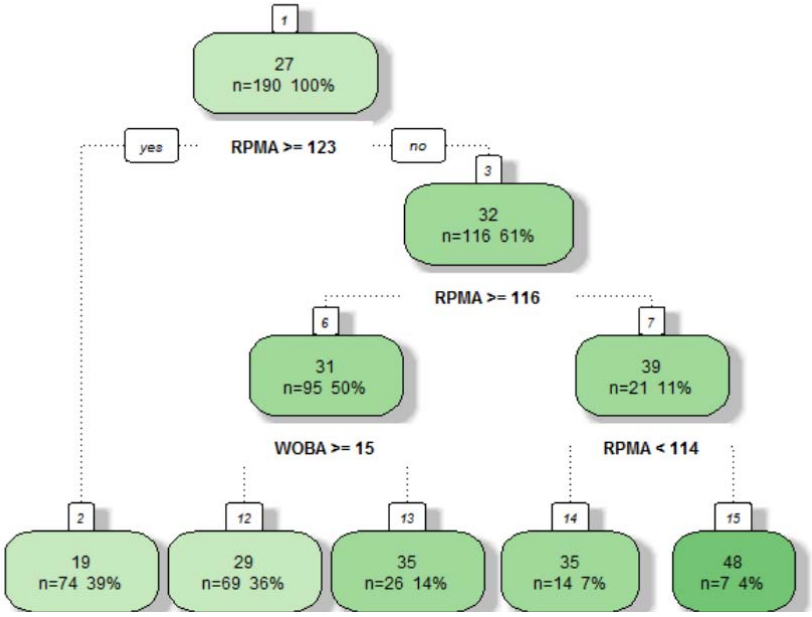


Figure 39-Decision tree representation for formation Sgiath after applying tuning in Well 12a.

Tuning is extremely important as it helps bring down the computation time, as the number of nodes decreases without compromising the error metrics and model performance.

Table 17 lists the summary of prediction results applied using the tuned tree models for each formation and Figure 40 presents the actual vs. predicted ROP plot.

Table 17-Prediction results of CART for all formations in Well 12a.

"TOTAL SUMMARY"

County	RMSE	MAE	MedAE	Med %	Mean %
NA	55.199	42.418	33.3	26.9	47.7
Ekofisk	5.543	4.175	3.4	10.1	12.1
Tor	11.673	9.281	8.0	17.4	24.1
Hod	29.349	23.306	20.1	28.1	41.3
Herring	21.842	16.374	12.8	33.1	46.9
Plenus	48.446	34.666	24.5	19.7	26.0
Hidra	40.892	32.466	27.6	23.8	33.7
Sola	33.661	26.306	21.7	26.9	40.9
Valhall	34.439	27.135	22.4	24.5	34.9
Valhall_Limestone	17.807	14.082	12.1	33.2	46.8
Kimmeridge	8.200	6.357	5.1	15.5	21.2
Top	9.954	7.494	5.5	20.5	45.7
UB4	10.800	8.636	7.6	13.6	16.5
Heather	21.740	15.828	11.5	16.8	39.1
Sgiath	11.692	8.783	5.3	30.3	41.3
Smith	7.144	5.492	4.3	22.1	33.6

3.2.4.2. Prediction Plot

As compared with KNN and SVR, RPART also did not perform better in either the Prediction results or in the Actual ROP vs. Predicted plots. However, it performed better compared to the regression models. Finally, ensemble methods were employed to achieve the best possible modeling algorithms for the given data, results and procedures of which are discussed in the following sections.

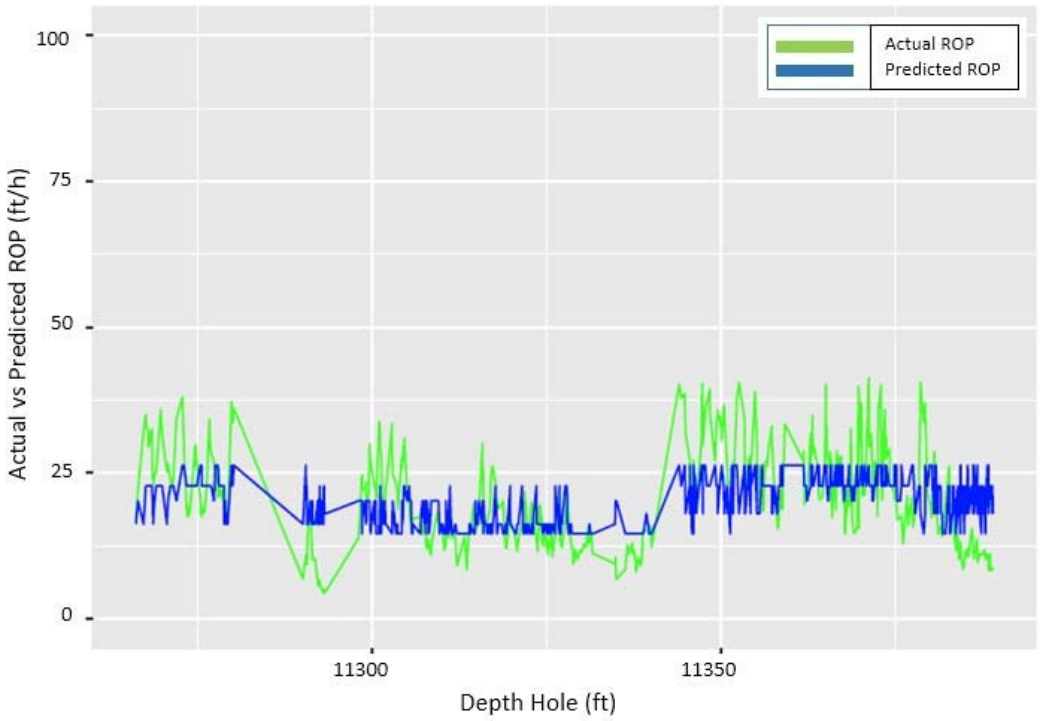


Figure 40-Actual vs. predicted ROP for Smith formation using CART in Well 12a.

3.3 Ensemble Methods

Ensemble learning is an advanced concept that trains multiple models simultaneously. It has several advantages such as reduced variance (due to the presence of multiple models) and reduced bias.

3.3.1. Random Forests (RF)

RF is an advanced ensemble method when compared to decision trees. RF grows several trees instead of one single tree and infuses randomness into each tree so that a 'forest' of such individual tree models is created. In regression, such as in ROP prediction, the predicted value (ROP in this case) is a weighted average of the value predicted by each individual tree. One of the most significant bagging ensemble learning algorithm is RF. Bagging algorithm (or Bootstrap Aggregating) generates 'x' new training data sets. Each new training data set picks of a sample of observations with replacement (also referred to as bootstrap sample) from the original data set.

Repetitive observations may occur in each new training dataset by sampling with replacement. The x models are put together using x bootstrap samples generate above and then combining them by averaging the output for regression. RF employs approximately two-third of the total training data used for growing each tree. The remaining data cases are not applied in tree construction. RF models have advantages over decision trees. They do not over fit (due to the presence of several trees which average out errors and minimize overfitting). There is no necessity for CV as the out-of-bag data are used for error estimation. The disadvantages of RF are in its interpretability. Random Forest applied to the Ekofisk formation is discussed here. Later the most relevant variables for each formation are ascertained. Relative parameter ranking is similar to the variable selector process that was discussed earlier in relation to Stepwise regression. This process is quite important as it helps physically control (or not control!) the most relevant variables in order to achieve the best ROP. The number of optimum trees required can be tuned. Figure 41 shows the error rate versus trees. The optimum trees are around 100 for Ekofisk. Model building is continued using the optimum number of trees required and feature selection using the RFE function will be discussed.

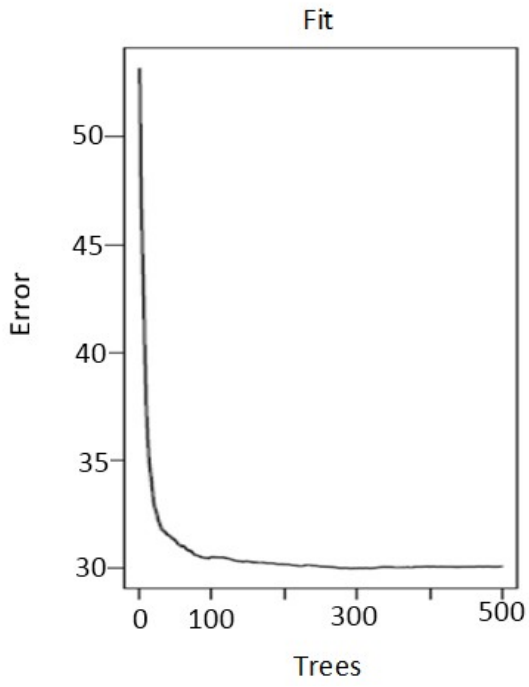


Figure 41- Error Rate vs. number of trees for Smith formation using RF in Well 12a.

3.3.1.1. Variable Importance (Recursive Feature Elimination)

RF is useful in finding the relative importance of the variables indicated by the %IncMSE. Higher the %Inc. MSE values, greater is the relative importance of that particular variable as shown in Table 18. Figure 42 is an illustrative way of looking at the most relevant parameters.

Table 18-%Inc MSE and IncNodePurity and RMSE values in RF for Smith in Well 12a.

	%IncMSE	IncNodePurity
RPMA	45.75628	18786.71
WOBA	43.29991	20053.56
Flowin	58.30745	21597.40
GR	41.02100	23133.07

Variables	RMSE	Rsquared	RMSESD	RsquaredSD	Selected
1	6.407	0.02379	0.4002	0.01602	
2	5.504	0.13318	0.4364	0.12390	
3	5.041	0.25295	0.3411	0.09434	
4	4.760	0.34046	0.3756	0.05404	*

The top 4 variables (out of 4):
Flowin, WOBA, GR, RPMA

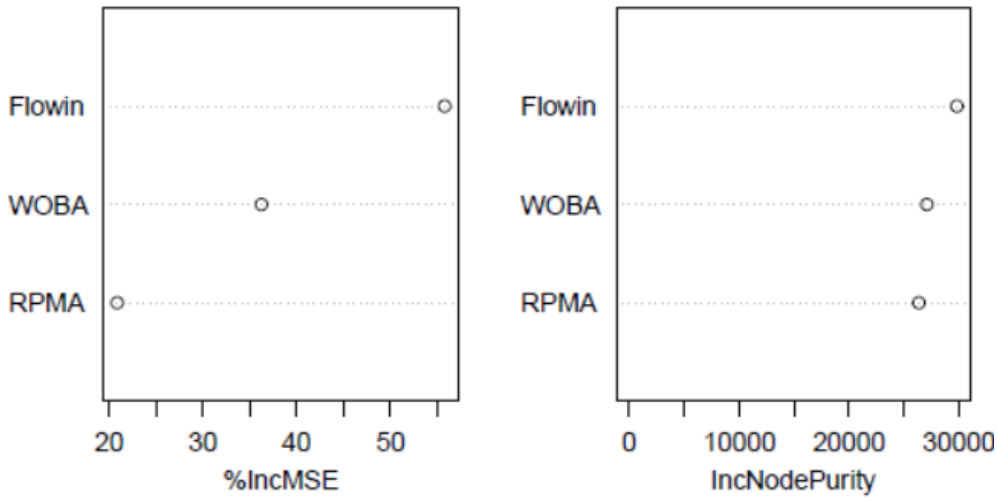


Figure 42- Parameter selection using RFE for Smith formation using RF in Well 12a.

3.3.1.2. Tuned Model

The summary of prediction results for all formations is shown in Table 19.

Table 19-Prediction results of RF for all formations in Well 12a.

[1] "TOTAL SUMMARY"

County	RMSE	MAE	MedAE	Med %	Mean %
NA	45.814	33.165	23.7	19.8	35.0
Ekofisk	5.282	4.081	3.4	9.9	11.9
Tor	11.488	9.055	7.4	16.1	22.2
Hod	30.026	23.695	19.0	28.6	42.1
Herring	21.733	15.857	11.5	29.3	44.3
Plenus	43.104	32.704	25.5	21.2	27.0
Hidra	37.963	30.255	26.1	22.0	30.9
Sola	29.639	22.761	17.7	22.2	34.0
Valhall	32.000	24.932	20.3	21.5	32.0
Valhall_Limestone	15.956	12.363	10.3	28.0	39.5
Kimmeridge	7.272	5.527	4.3	13.5	17.2
Top	8.565	6.421	4.9	18.5	33.7
UB4	10.749	8.572	7.2	14.2	16.5
Heather	21.153	15.388	10.9	16.2	38.5
Sgiath	10.628	7.664	4.5	17.1	32.3
Smith	6.340	4.844	3.8	20.5	29.0

RF performs better than regression and NN while it is on par with SVR and KNN as far as prediction results are concerned. The actual vs. predicted ROP plot in Figure 43 indicates that it performed decently. Due to inherent cross-validation and testing, the tree building exercise contributed to the success of RF. The error metrics are also significantly low in RF and it does a good job in mapping the actual ROP.

3.3.1.3. Prediction plot

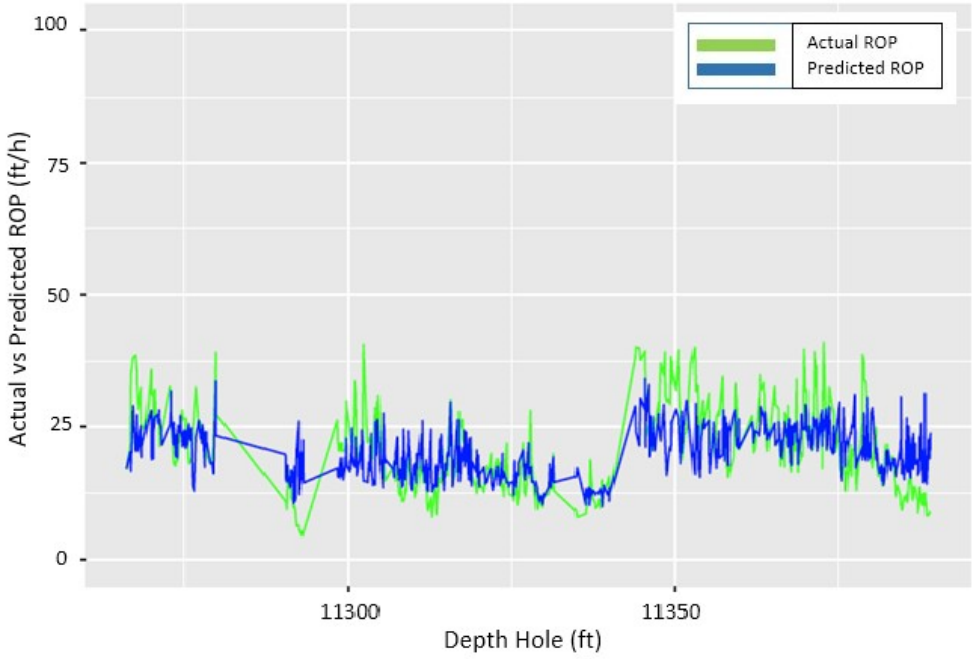


Figure 43-Actual vs. predicted ROP for Smith formation using RF in Well 12a.

3.3.2. Gradient Boosting Machine (GBM)

The last method employed in the study was gradient boosting machine (GBM). In GBM, many simple decision trees are created, where each tree is built for the prediction errors of the previous trees (Bhalla, D. 2015). Once the first tree is created, weighted trees are created following weight determination and subsequent iterations. The final prediction is the weighted sum of the decisions made by trees as a whole. So the main idea here is to combine several simple decision trees such that each tree complements the previous ones and keeps track of the errors of the

previous trees. For example, from the Figure 44, we start with the first box. The one vertical line which is seen becomes the first week learner. Now in total there are 3/10 misclassified observations. Now higher weights are given to three plus misclassified observations. Hence, the vertical line towards right edge. The process is repeated and then each of the learner are combined in appropriate weights. Models are built independently in bagging, whereas in boosting, models are improved and built upon previous ones. This helps reduce variance and bias but also leads to overfitting.

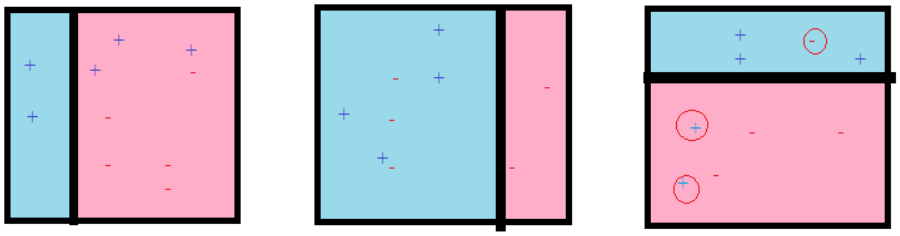


Figure 44- Boosting model explained through learners.

3.3.2.1. Tuned Model

GBM has a large number of hyper-parameters to tune, out of which the most important are the following: Number of trees or GBM iterations, Interaction depth or number of splits to be performed starting from a single node, as shown in Figure 45, and shrinkage used for lessening the impact of each additional fitted base-learner. With the tuning parameter shrinkage held at a constant value of 0.1 and a minobsinnode (minimum observations) of 20, the best model (least RMSE) for n.trees (number of trees) = 950 and interaction.depth = 5 is obtained.

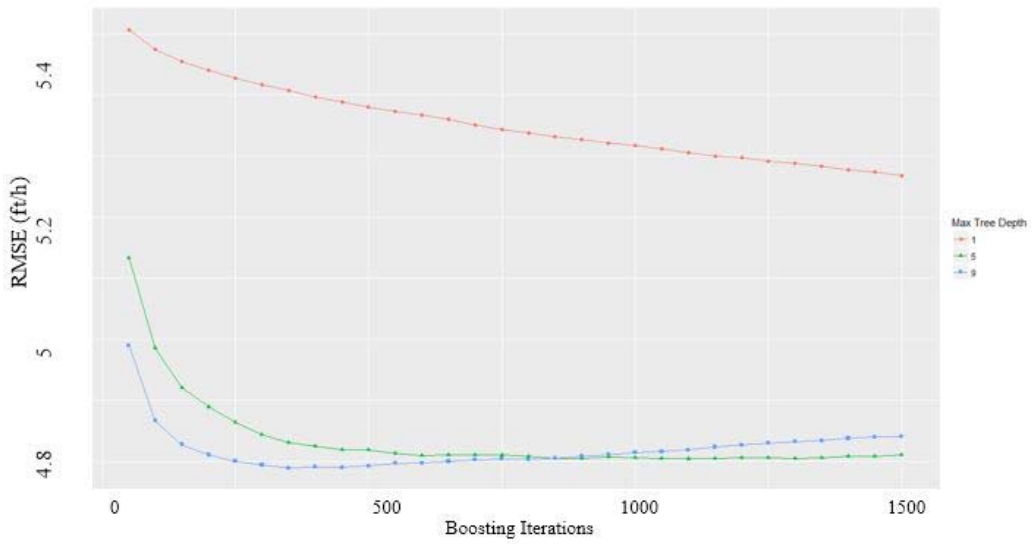


Figure 45- Tuning parameters vs. RMSE of GBM for Ekofisk formation in Well 12a.

3.3.2.2. Variable Importance

Similar to the linear models and RF, variable importance can be calculated in GBM. In Linear Models: the t-statistic for each parameter is used to find relative variable ranking. And in random forests, only the out-of-bag observations are used but GBM computes using the entire training dataset (not the out-of-bag observations). At GBM, relative variable ranking for parameters is normalized to sum upto 100, as shown in Figure 46.

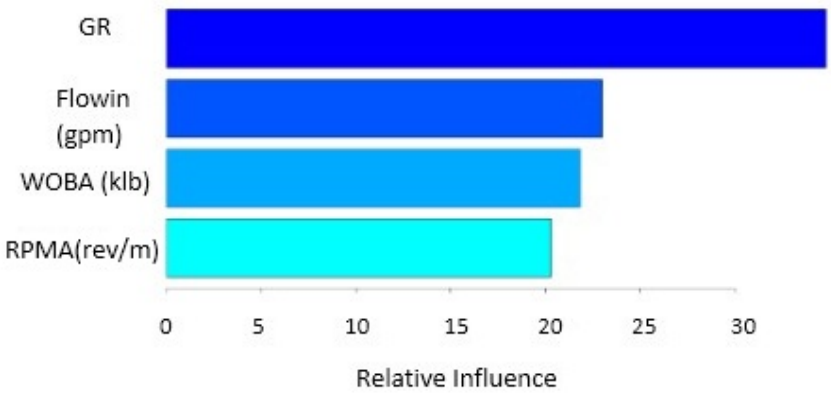


Figure 46-Parameter selection in GBM for Ekofisk formation in Well 12a.

Table 20 provides the results of prediction after tuning is performed. Figure 47 shows the predicted vs actual ROP values for the Smith formation in well 12a. GBM does a good job of prediction. As shown in Figure 45, errors have considerably reduced and it is the best modeling technique for almost all formations. Visual inspection of the attached actual vs. predicted ROP plots revealed that GBM, RF, KNN and SVR performed decently.

Table 20-Prediction results of GBM for all formations in Well 12a.

County	RMSE	MAE	MedAE	Med %	Mean %
NA	39.544	28.711	21.1	18.2	27.8
Ekofisk	4.689	3.527	2.8	8.3	10.1
Tor	9.270	7.211	5.9	13.6	17.2
Hod	27.389	20.865	16.3	24.8	34.9
Herring	19.995	14.796	11.7	28.5	40.8
Plenus	42.494	30.417	19.7	18.2	23.4
Hidra	34.559	27.151	23.5	20.5	27.5
Sola	25.027	19.105	15.1	19.7	26.4
Valhall	27.714	21.144	16.6	18.5	24.6
Valhall_Limestone	11.891	8.629	6.5	18.1	25.1
Kimmeridge	5.422	4.141	3.3	10.6	13.0
Top	6.032	4.482	3.5	13.0	22.0
UB4	9.546	7.395	6.0	11.3	14.2
Heather	16.113	12.084	9.2	13.9	22.0
Sgiath	9.165	6.931	6.0	22.1	30.8
Smith	4.483	3.389	2.6	14.1	17.9

3.3.2.3. Prediction plot

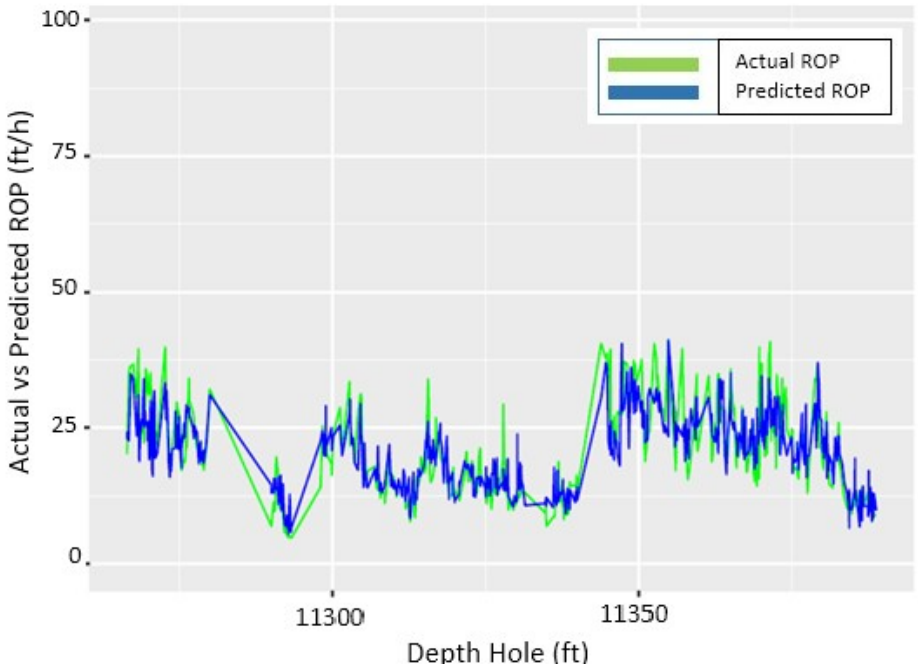


Figure 47-Actual vs. predicted ROP for Smith formation using GBM in Well 12a.

3.4 Comparison of all algorithms

Table 21- Advantages and disadvantages of all algorithms

Algorithm	Advantages	Disadvantages
Linear	Fast and Interpretable. Less prone to overfitting.	Assumes linear relationship. Difficulty modeling nonlinear relationships
Stepwise	Computes most important predictors. Not computation intensive.	Assumes linear relationship.
SVR	Can model complex relationships. Robust to noisy data.	Large computation time and processing power required.
KNN	Simple, powerful and fast.	Fails on high dimensional data.

Table 21 (continued)

NN	Powerful. Can model complex relationships.	Overfitting, computation time and processing power. Black box model.
CART	Treats both linear and non-linear data without assumptions. Accurate and easier to interpret.	Overfitting.
RF	Data splitting not required as algorithm has CV inbuilt. One of the best performing algorithms. Reduced variance.	Interpretation is tricky. Many hyper-parameters to tune.
GBM	Reduced variance and bias. One of the best performing algorithms.	Tuning requires many hyper-parameters. Overfitting.

Table 21 lists the advantages and disadvantages of all algorithms. The prediction errors for all models are discussed in the follow chapter for all algorithms in Well 12a and the remaining five wells.

The next chapter summarizes the results of all the algorithms by comparing the RMSE and MAE for the remaining wells. A selection of best modeling techniques is prepared which was then applied to these wells.

4. RESULTS AND CONCLUSION

4.1 Model Evaluation on Test Wells

Different types of algorithms belonging to regression, instance–based, trees, neural networks and ensemble methods were run on Well 12a in order to determine a bag of best performing models for each formation. Results of all models for the Smith formation are presented in Figure 46 and the results for missing formation data, represented as "NA" are presented in Figure 48.

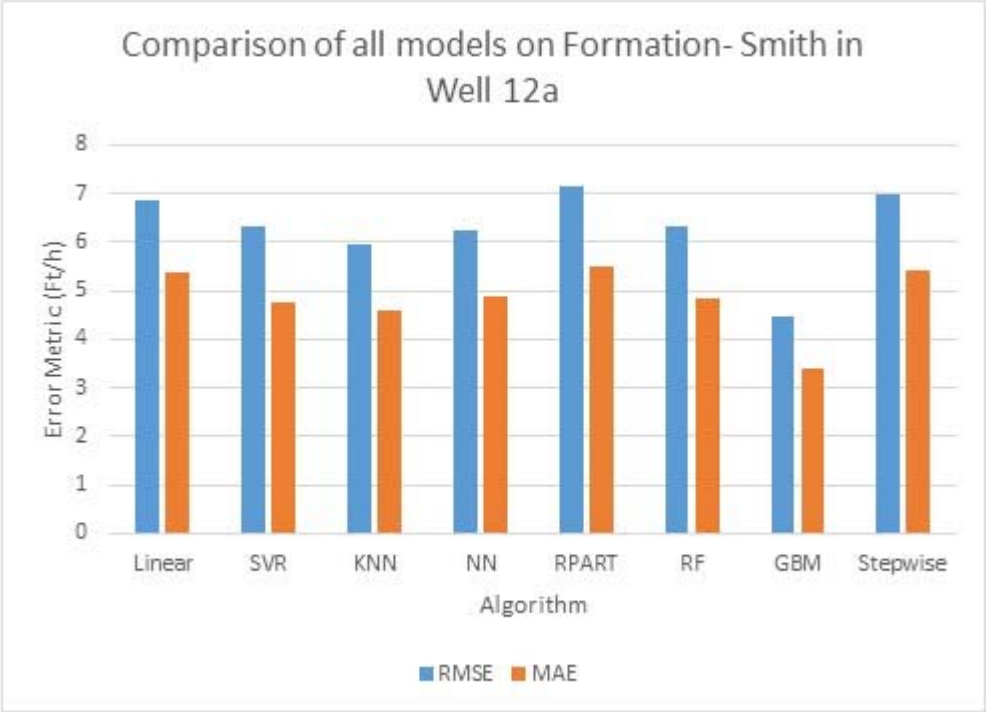


Figure 48-Comparison of all models on Smith formation in Well 12a.

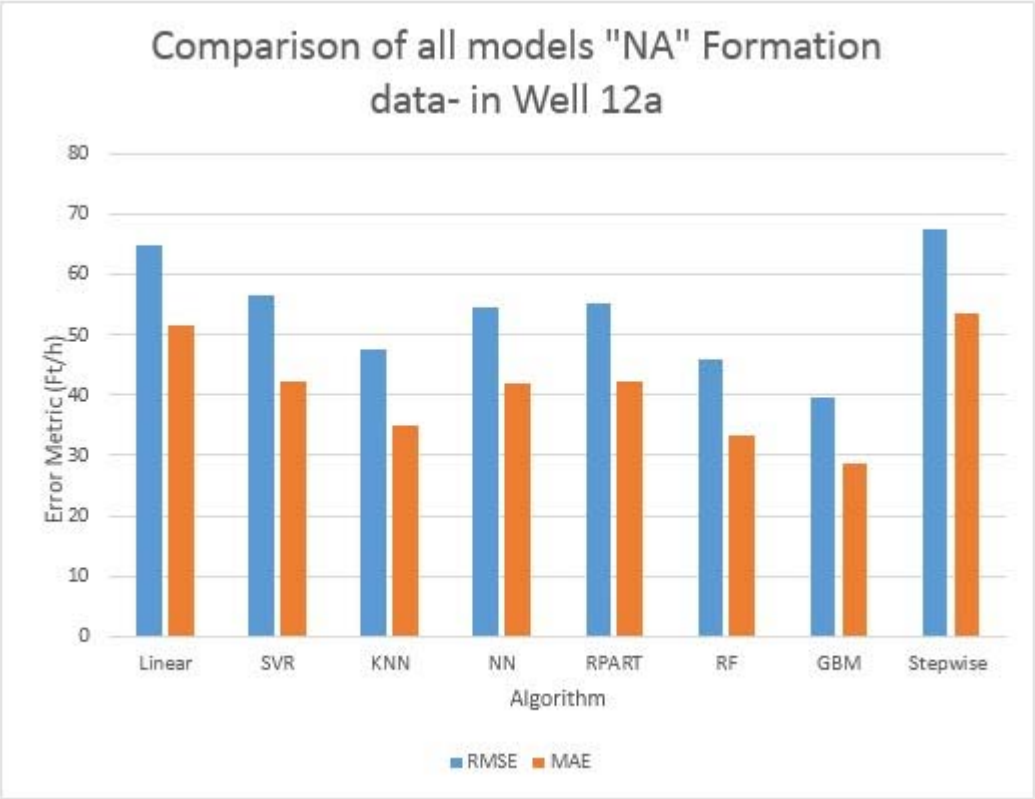


Figure 49-Comparison of all models for missing formation data (NA) in Well 12a.

GBM gave the best results for most of the formations, followed by KNN. But the other algorithms such as SVR and RF also performed very well. Significant differences can be noticed when models were run on data with missing formation (NA) information as shown in Figure 49. RF, SVR and GBM did particularly well. Hence, the same analysis was run on all the formations. Results of the best models and the second best models by formation for Well 12a are shown in Figure 49.

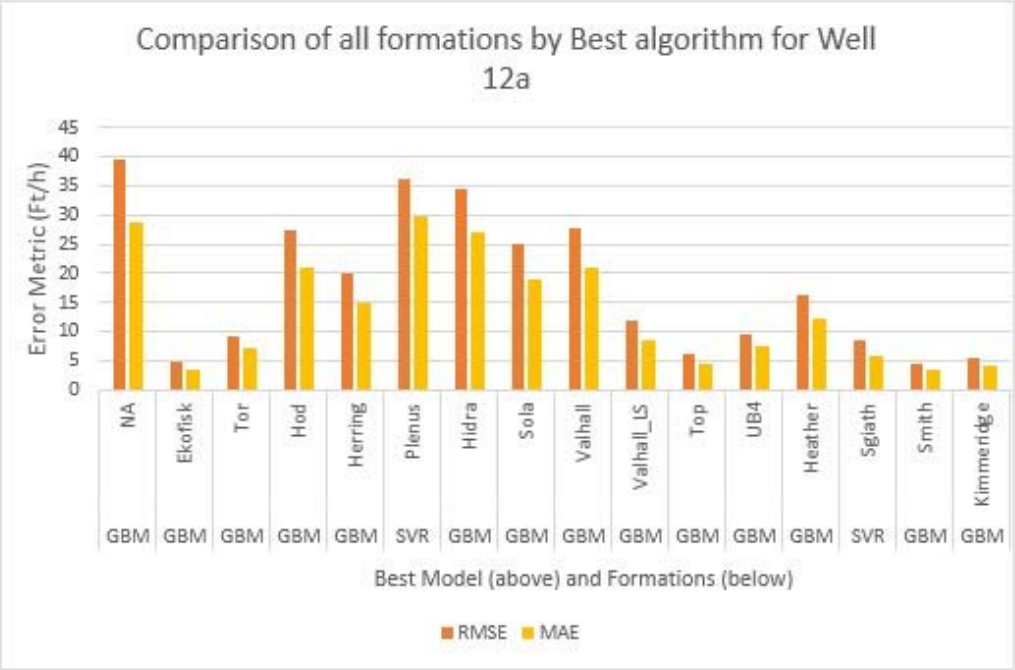


Figure 50-Comparison of all formations using the best algorithm for Well 12a.

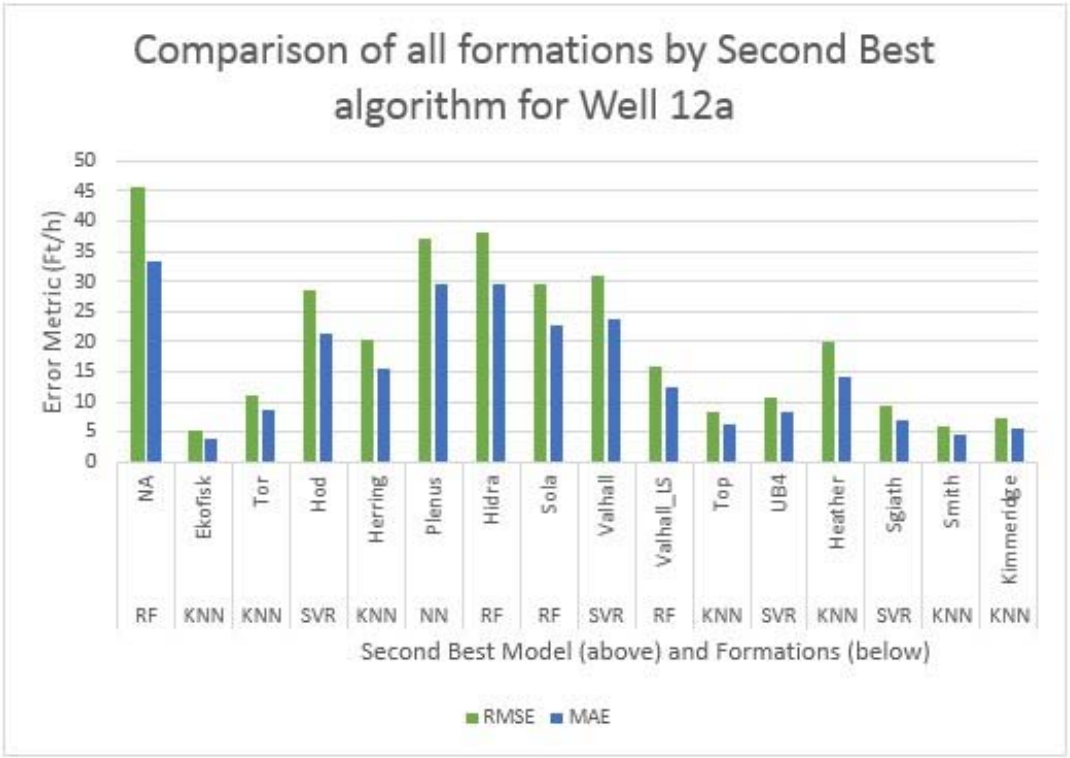


Figure 51-Comparison of all formations using the second best algorithm for Well 12a.

Figures 50 and 51 indicate that there are four algorithms which delivered good results across all formations. Although, SVR was performing well on few formations, KNN, RF, and GBM were applied due to less computation time taken. The developed workflow consisting of the top three-Boosting, RF, SVR and KNN are chosen to be used for testing on the other Wells-B2a, E8, 10, 13, and B30y to see how they performed.

Well B30y: There were about 36000 rows of data for Well B30y. The best modeling algorithm was found to be RF followed by KNN, as shown in Figure 52.

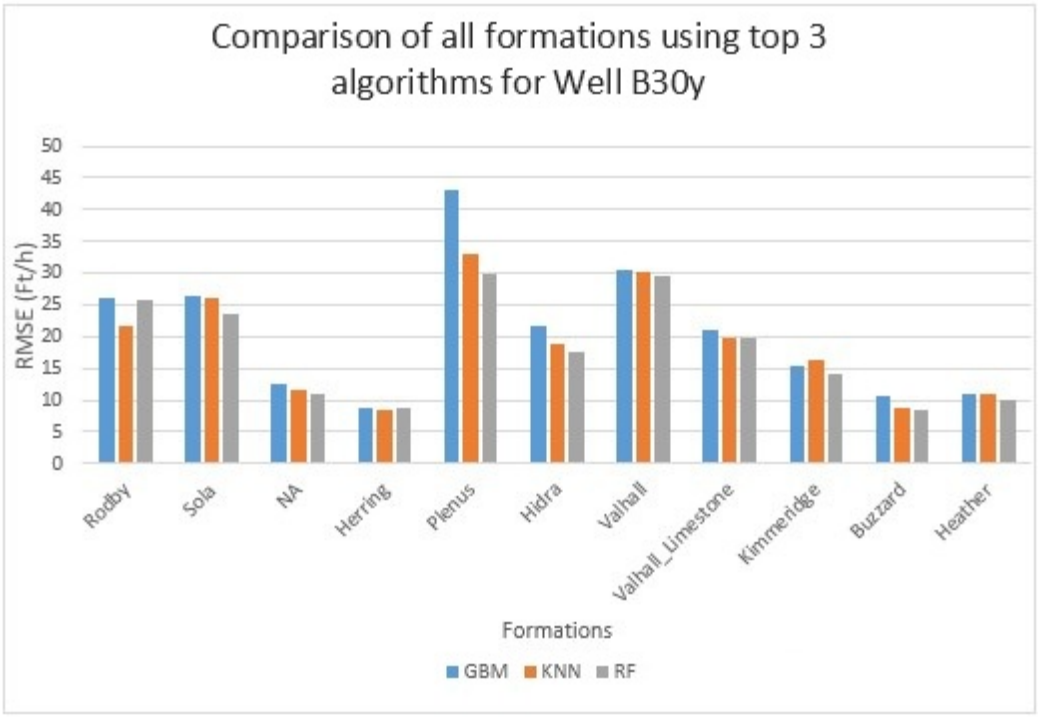


Figure 52-Comparison of all formations using the top 3 algorithms for Well B30y.

Well E8: There were about 121,000 rows of data for Well E8. The best modeling algorithm was found to be RF followed by KNN, as shown in Figure 53.

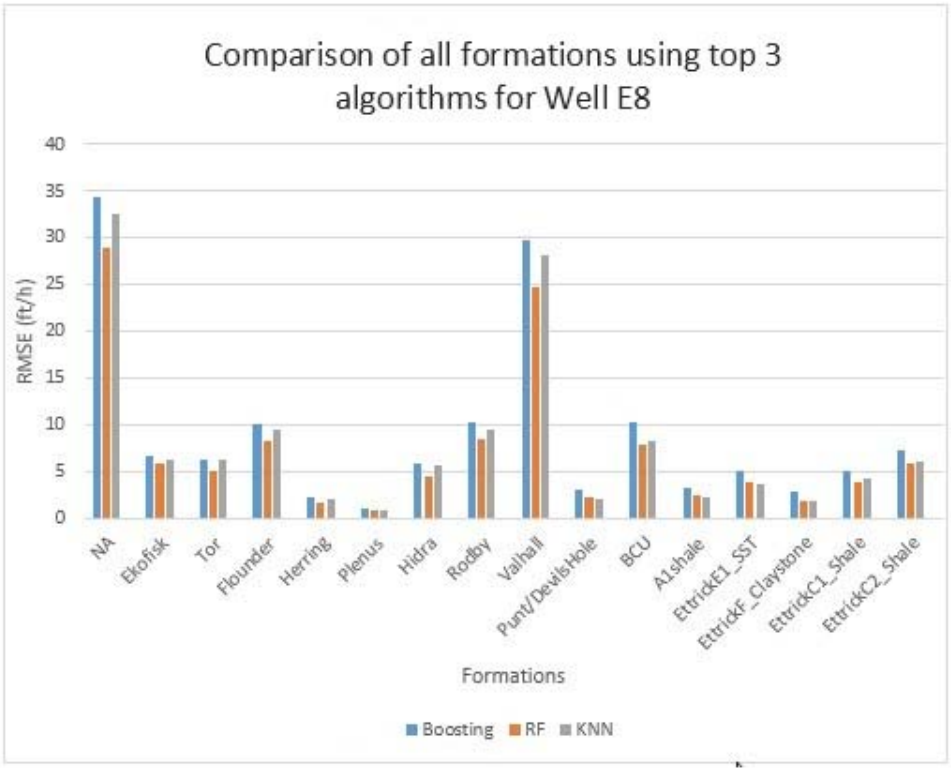


Figure 53-Comparison of all formations using the top 3 algorithms for Well E8.

Well 10: There were about 50,000 rows of data for Well E8. GBM and RF performed well in some formations and not as good in others as shown in Figure 54.

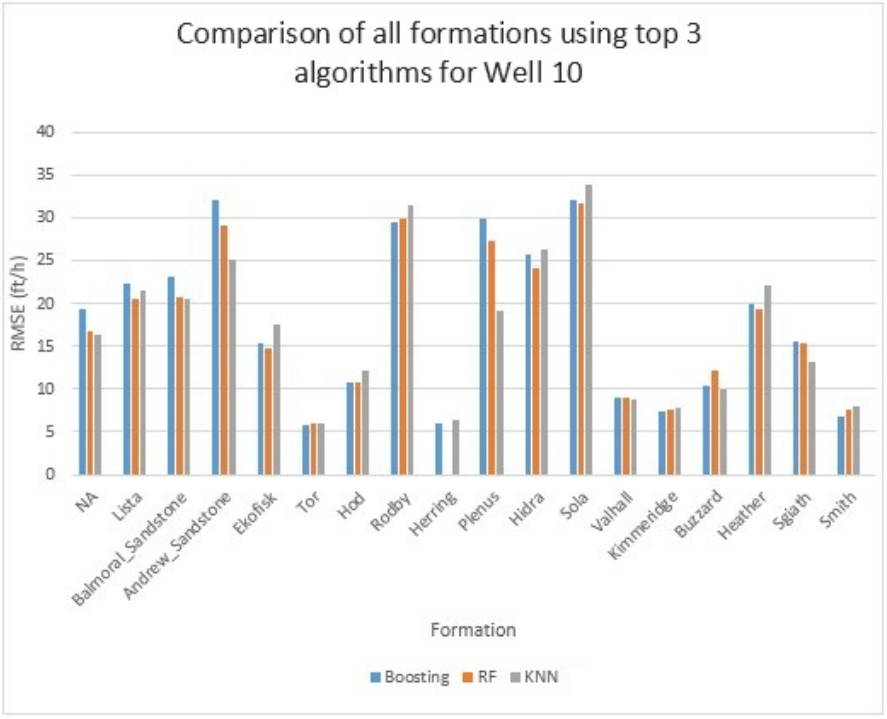


Figure 54-Comparison of all formations using the top 3 algorithms for Well 10.

Well B2a: There were about 67,000 rows of data for Well E8. The best modeling algorithm was found to be GBM followed by RF as shown in Figure 55.

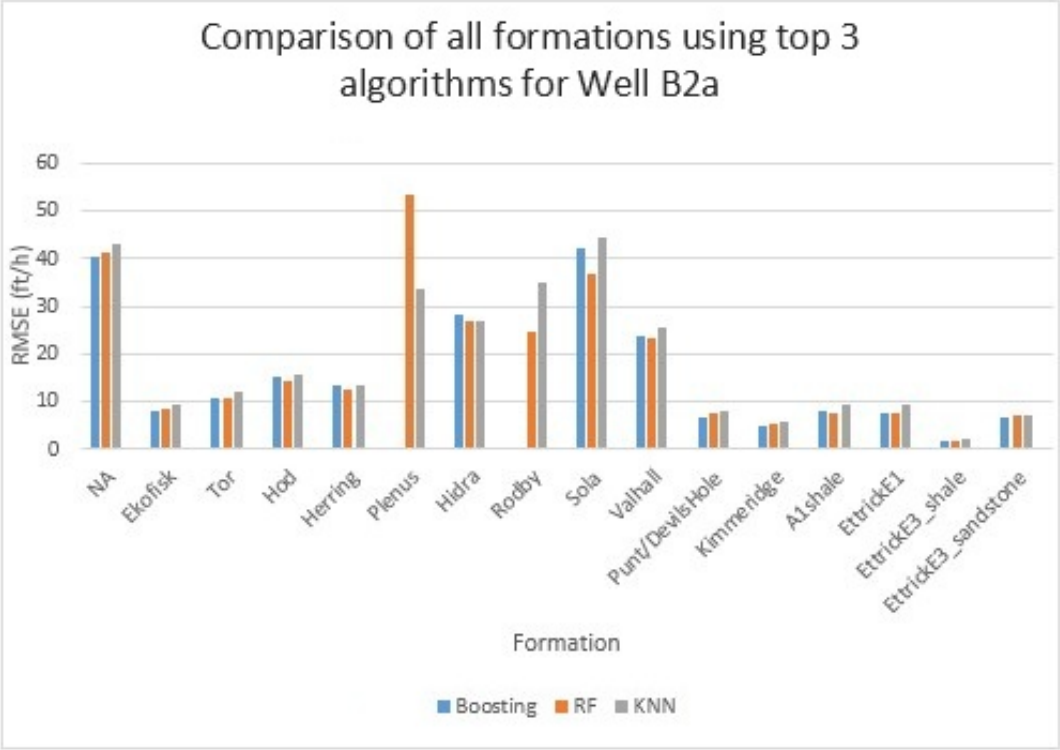


Figure 55-Comparison of all formations using the top 3 algorithms for Well B2a.

Well 13: There were about 59,000 rows of data for Well E8. Boosting and Random Forest performed good in some formations and failed in others as shown in Figure 56.

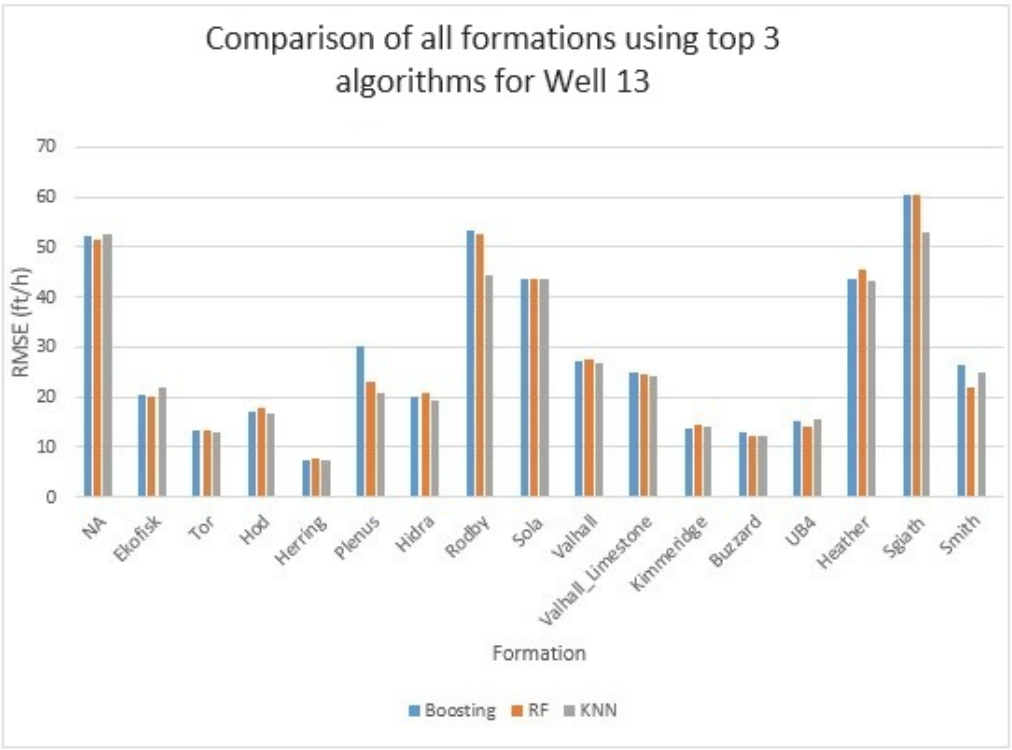


Figure 56-Comparison of all formations using the top 3 algorithms for Well 13.

4.2 Uncertainty Analysis

The usefulness of any workflow depends on the accuracy of the output it generates. Input data are rarely captured accurately. Consequently, this imperfection gets transferred to the output parameters leading to inconsistent, inaccurate and sometimes irrelevant answers.

Drilling engineering data is also subject to errors as there could be several errors while capturing data. The RPM values that are noticed at the surface are usually different from those that are present underground. Same is the issue with WOB.

In uncertainty analysis, attempts were made to understand and predict a range of outputs while taking errors in the predictors into consideration. To be able to predict outputs more accurately, a range of ROP predictions are needed.

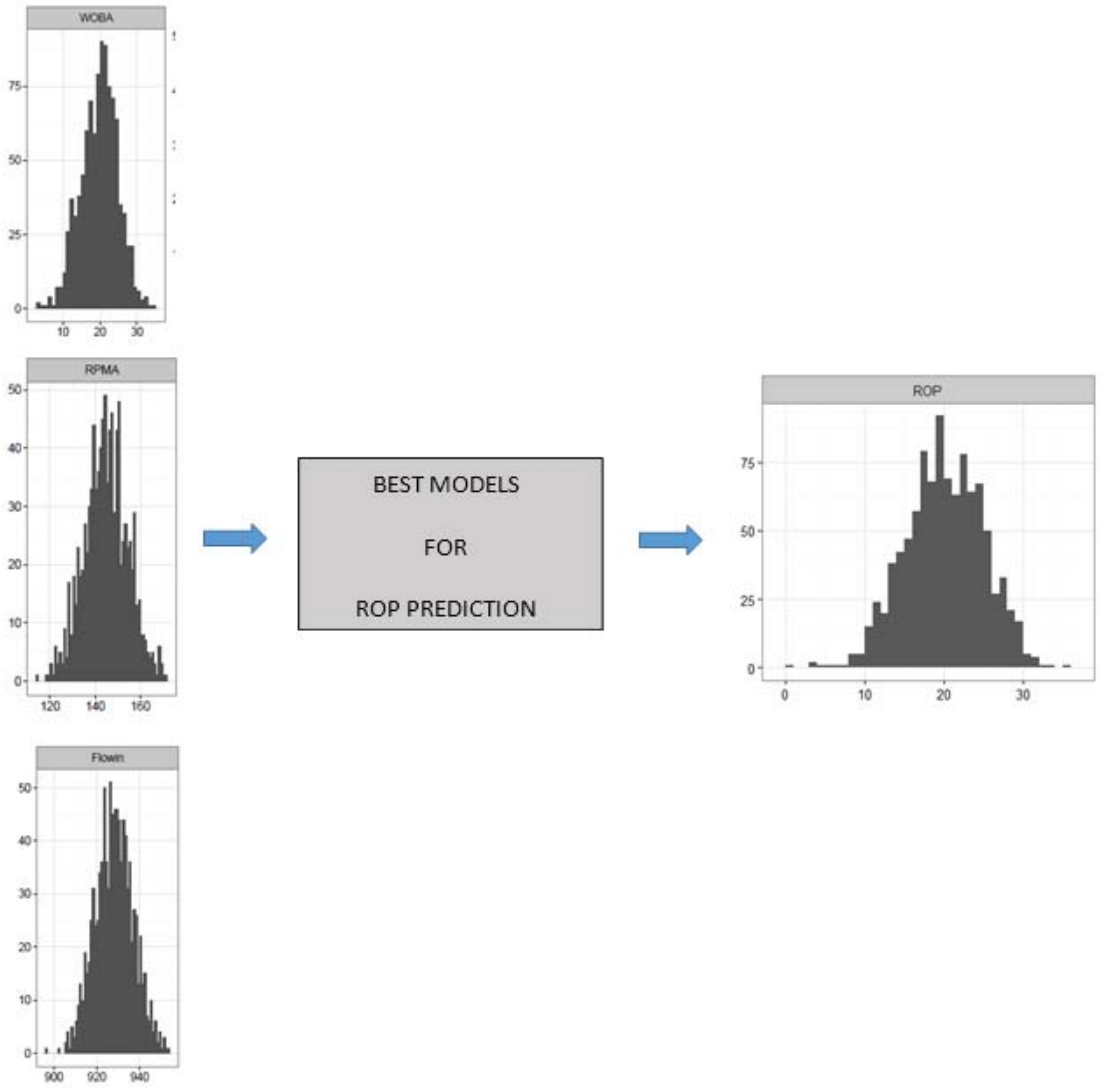


Figure 57-Model input distributions resulting in a range of Output distributions

Monte Carlo simulation was used to undertake uncertainty analysis. The predictors were assumed to follow normal distribution (which usually is the case with Petroleum data). New predictor data, which incorporated errors for each of the parameters, was simulated. Initially, the best model from all the algorithms applied was chosen (on the basis of RMSE and/or MAE).

Random sampling was used to create normal distributions for the predictors, by fixing a mean and standard deviation (usually averaged from the test data). After choosing 1000 generated data sets (ranges similar to the actual field data), ROP for each of the data points was predicted for a chosen number of simulation runs (usually 10000) and the distributions of the ROP along with P10, P50 and P90 values were captured, as shown in Figure 57.

Using the mean and standard deviation of Ekofisk formation test data, normal distributions of three predictors (WOB, ROM and Flow) containing 1000 rows of data was simulated. Figure 58 represents the distributions of one such simulation run. Using this data, and the best performing model (linear model is used as an example), ROP values (vector of 1000 values) were predicted. The same process was repeated with a different set of randomly sampled data for 10 runs (usually 1000 preferred) and generated P10, P50 and P90 values.

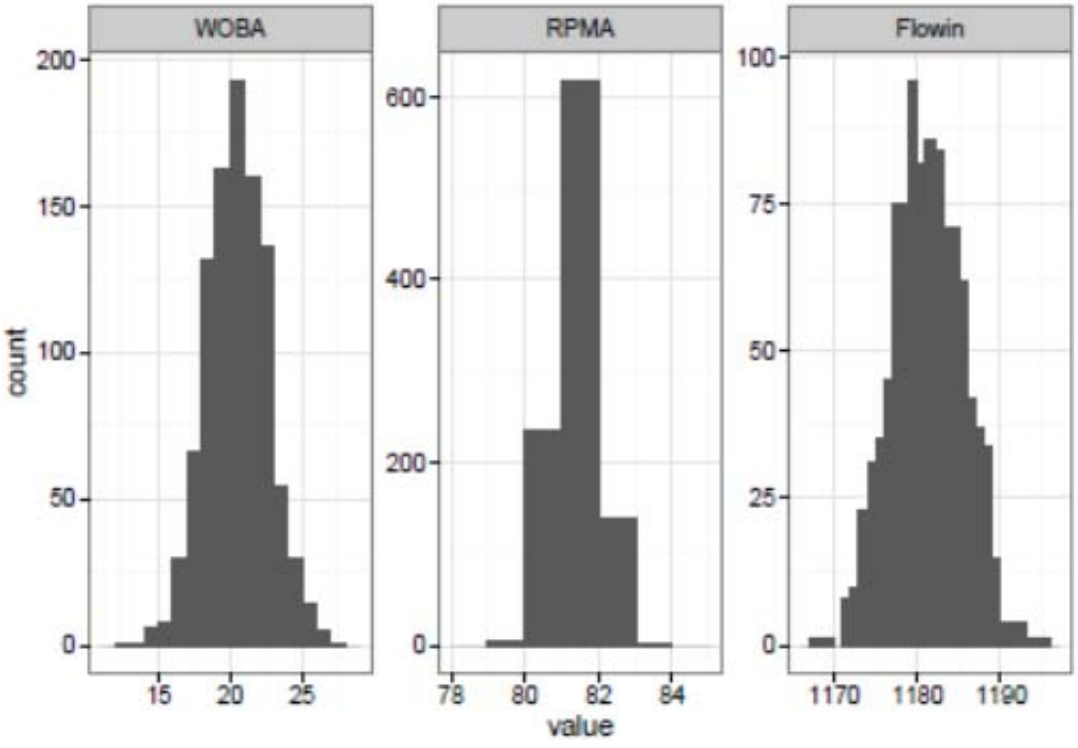


Figure 58- Simulated distributions (normal) of predictors for Ekofisk Formation in Well 12a

Table 22 represents the range of predicted values for the Ekofisk formation. Ten simulation runs were performed on 1000 simulated data points and the distributions were plotted at the end of each run. Then, the mean (median can also be considered) of P10s P50s and P90s values were taken to determine a range for each formation and to develop a probabilistic estimate rather than a single pinpointed value.

Table 22-P10, P50 and P90 values of predicted ROP distribution using Regression model for Ekofisk Formation in Well 12a

	Run	P90	P50	P10
1	1	32.51	34.42	36.32
2	2	32.63	34.47	36.30
3	3	32.54	34.45	36.37
4	4	32.60	34.53	36.46
5	5	32.49	34.43	36.38
6	6	32.54	34.49	36.44
7	7	32.47	34.44	36.41
8	8	32.53	34.47	36.41
9	9	32.44	34.41	36.38
10	10	32.58	34.50	36.42

Table 23 gives the summary of averaged P10, P50 and P90 values after the Monte Carlo simulation (10 runs- 1000 data points) was conducted for each formation of Well 12a. Hence, this approach helps to better understand the range of ROP predictions for each formation and once again emphasizes the importance of sub setting data by formations for model building and analysis.

Table 23- Summary of P10, P50, P90 distributions using regression for all formations in Well 12a.

	Formation	P90	P50	P10
1	NA	84.140	133.519	182.898
2	Ekofisk	32.533	34.461	36.389
3	Tor	37.077	46.140	55.202
4	Hod	52.693	72.875	93.063
5	Herring	27.958	45.691	63.429
6	Plenus	104.517	136.558	168.597
7	Hidra	106.037	117.022	128.008
8	Sola	76.189	83.975	91.762
9	Valhall	72.121	98.140	124.159
10	Valhall_Limestone	35.689	38.310	40.931
11	Kimmeridge	25.075	34.067	43.058
12	Top	25.433	27.403	29.376

4.3 Sensitivity Analysis

Sensitivity analysis was performed using models built from the best performing algorithms. Four predictors- GR, RPM, WOB and Flow were used to create simulated data. The simulated data was constructed using normal distributions for each of the predictors. Three of the predictors were fixed at their mean values, while the fourth predictor had simulated data values that followed a normal distribution (with a pre specified mean and standard deviation- which represented the errors in capturing data). This data was then tested against the best performing models to capture a range of ROP values. By repeating this process for each predictor, ranges of ROP were obtained and compared. The main feature that was noticed was that the most sensitive parameter varied for each formation, highlighting the importance of segregating data by formation for building models. This process of identifying the most relevant parameter was also done using algorithms directly. RF and GBM have good inbuilt techniques for achieving the same (as discussed in the previous sections). Figure 59 presents the most important parameter using

Recursive feature elimination in the RF algorithm. As stated earlier, the relative ranking of each predictor provided an idea as to what parameters to focus on while optimizing ROP. For example, in Figure 59, the emphasis was to alter Flow and WOB as they were the most sensitive parameters for Smith formation. The other important insight revealed was direction of dependency of the output on a particular predictor.

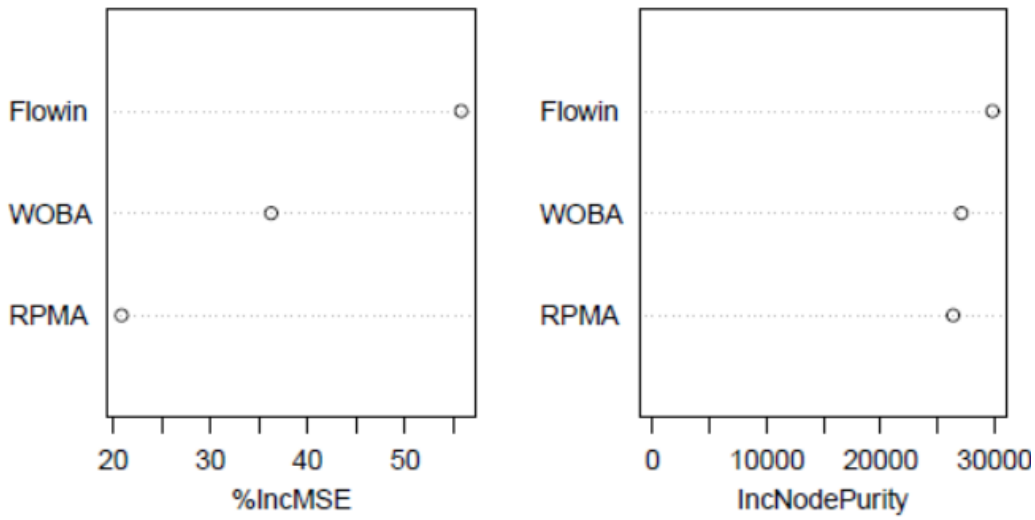


Figure 59- Recursive feature elimination using RF for relative predictor ranking in Smith formation of Well 12a.

RF also helps to understand the direction of influence of each parameter on the output. Figure 60 lists the influence for each of the four parameters versus ROP. Flow has a well-defined negative trend with ROP and it is also the highest contributor from the previous graphs.

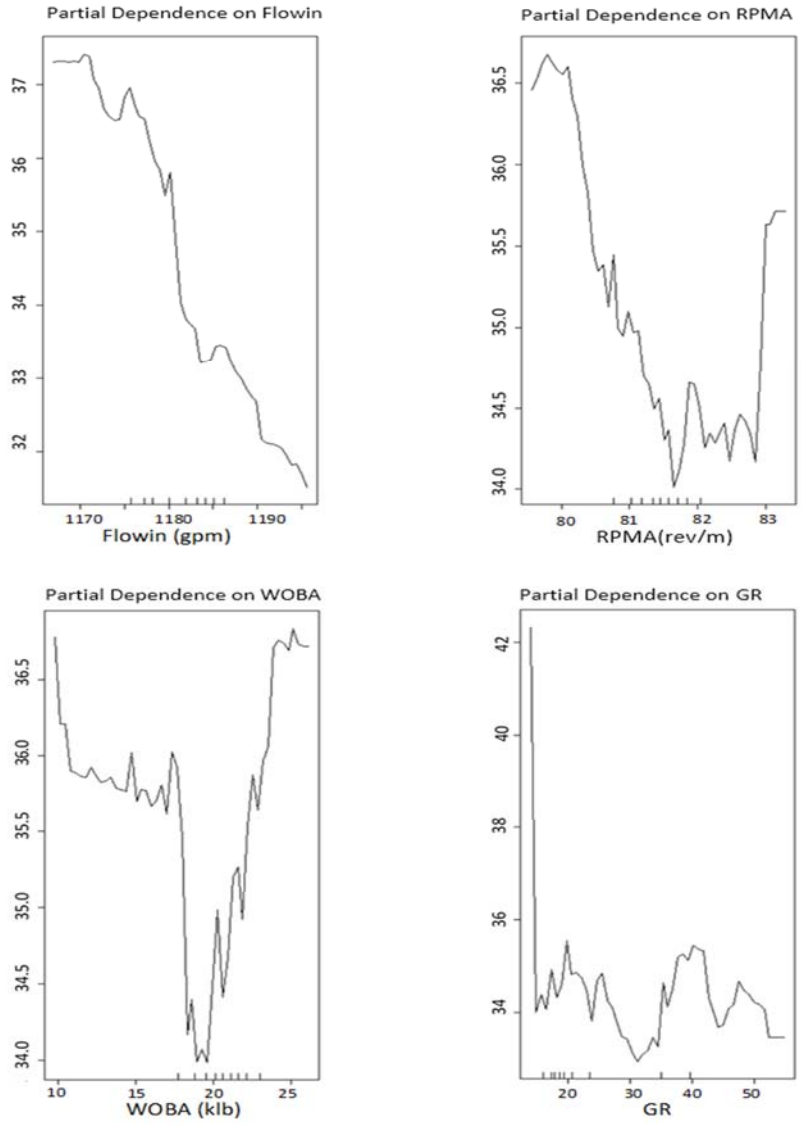


Figure 60-Partial dependency plots of WOB, RPM, Flow, and GR on ROP for Smith in Well 12a.

Figure 61 shows the partial dependency plots for each predictor. WOB was found to decrease around 15 klbs, Flow followed a negative pattern with ROP while RPM was highest around 80 rev/min. By following the pattern of ranges of each predictor as illustrated above, simulated data was recreated and tested against the best algorithms. ROP was found to increase when predictor values as determined from the partial dependency were used as art of the simulated data. This exercise confirmed the importance of undertaking sensitivity analysis in order to optimize the

ROP. The analysis not only helped identify the most relevant parameters, but also suggested a possible range of values for each predictor in order to achieve the best possible values for ROP.

In the prediction application developed as part of this project (that will be discussed in chapter 5), there is an option for the user to assign a distribution or to provide a test dataset. Within the app, a single row of predictors can be used to generate data points using a pre-specified mean and standard deviation (or the error in measurement). This data set would then be used against the best model(s) computed from chapter 4 to predict a range of outputs and also perform sensitivity analysis.

4.4 Conclusions

Extensive simulations were performed to test the robustness of the best models developed. The study presents several case studies (wells) in which the hybrid models are successfully used to estimate and optimize the drilling parameters using the prediction app. The accuracy of the models was increased with a hybrid intelligent system based on the evolutionary computation system that combines the statistical regression models. The new system provides better predictive accuracy and performance as compared to the traditional models.

The following conclusions were reached from this study:

1. ROP follows a complex relationship which cannot be comprehensively explained by traditional models alone. Application of data-driven analytics using several machine learning algorithms coupled with regression analysis can better predict ROP, irrespective of the formation/location/region.
2. A multivariate analysis comprising of data from drilling, formation and survey data can better model the complex relationship and uncover innate relationships among predictors, rather than using only drilling data such as RPM, WOB and Flow for modeling.

3. Sub-setting the data by formation or bit size yielded better prediction results than using the entire data for a single model. This finding highlights the importance of additional data other than just drilling while predicting ROP and model construction.
4. An ensemble of methods: GBM and RF helped achieve the best prediction with the least error metrics for most of the formations across all the five wells of data. Algorithms such as KNN and SVR also performed better and can be used if there is a constraint on computing capabilities.
5. Outliers should not be discarded but analyzed carefully to uncover any trend/ anomaly. They can give a picture as to what caused the extreme values to be present. And may help predict tool failure or stuck pipe etc.
6. Uncertainty analysis account for the inclusion of errors in predictor data and provides a realistic range of predictions (P10, P50 and P90) using Monte Carlo simulation. This analysis presents a holistic picture to the driller when compared to models that predict only one value.
7. Sensitivity analysis helped determine the most contributing predictor for each model built as well as its trend with the response variable (ROP). This analysis helps the driller to optimize and achieve the best possible ROP by varying only the most contributing parameters according to the trend they follow with ROP.

4.5 Limitations

1. Artificial intelligence and statistical regression techniques require very good quality data for model building purposes. Bad data can severely hamper the model performance while the presence of missing values can be detrimental as difficulties in learning patterns arise.

2. ROP optimization can sometimes provide irrelevant/ impossible values for predictors, and hence it is extremely important to crosscheck these results against domain knowledge and field-reported values.
3. Lack of multivariate data (drilling, formation, survey, logs etc.) can handicap the predictive capabilities of the model as a good ROP optimization model utilizes as many parameters as possible to account for intricate relationships among the variables.
4. The usage of advanced machine learning ensemble methods such as random forests and boosting requires good computation capabilities and hardware support.
5. The data used in the analysis consisted only of horizontal wells. So the best performing algorithms might vary if vertical wells are analyzed. The range of ROP predictions can also be different from the results observed in this project, so caution should be exercised while any comparison of the observed trends is done.

A web-based prediction Application has been developed that directly enables a user to apply the above methodology and use it to predict ROP. The app has several steps ranging from exploratory analysis to applying inbuilt algorithms featuring regression, neural networks, trees, instance-based methods and ensemble techniques. It also allows the user to perform Monte Carlo simulation to test the robustness of the best models built. The app also can be used to apply predictive analytics in production and reservoir engineering and can be easily modified for use in the other industries as well. Most importantly, the any software knowledge is not required to use the app.

5. DATA PRODUCTS: Prediction_APP

5.1 About

A predictive analytics app has been developed as part of this thesis project. Coding was done using R and the app is deployed using Shiny R. The user does not need any prior knowledge of coding in R or how each algorithm functions to access the app. The app has a wide range of features and runs an automated process by which the user can upload a dataset, perform data splitting, build model using several algorithms such as regression, CART, ensemble methods and then proceed to testing the models against test dataset. There is also an option to choose the best model by all the existing algorithms and the app returns the best performing model for each formation. The app can be accessed using a mobile phone/ tablet or a PC and does not require the installation of any software. The details are discussed in the following section.

5.2 Description

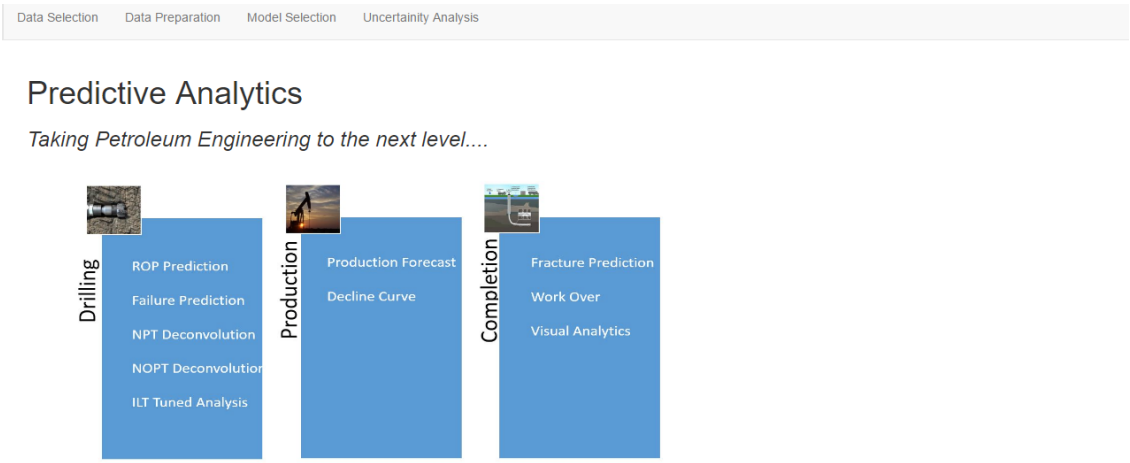
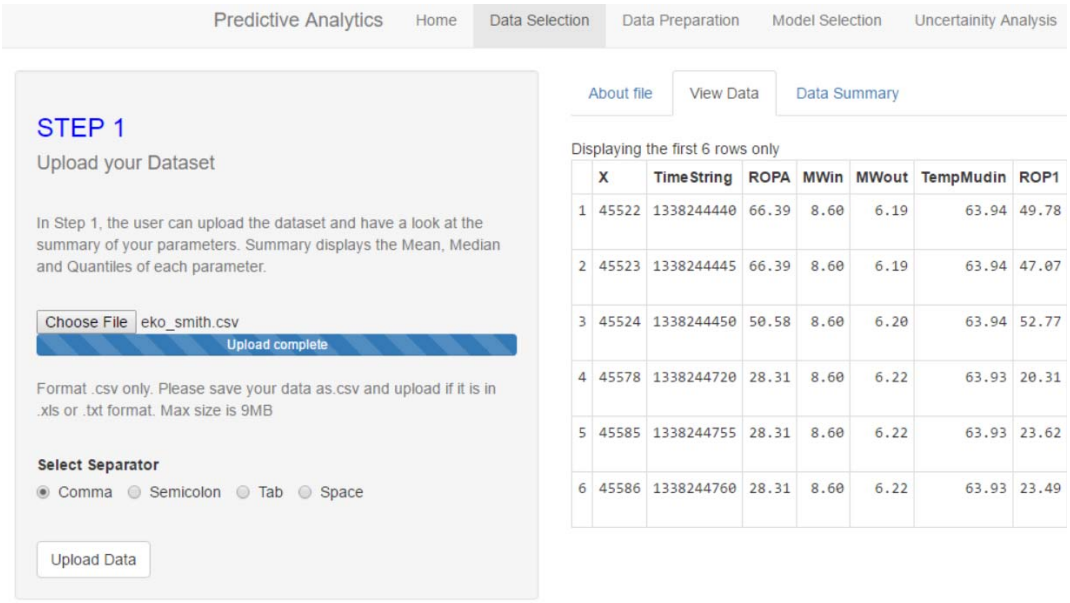


Figure 61- Homepage of the web-based prediction app developed.

Design and usability were one of the most important factors considered while designing the app. This software can accommodate several analyses pertaining to other operations in

petroleum engineering. Figure 61 shows the introduction page. The following tabs explain how to use the software for prediction purposes.

5.2.1. Data Selection



	X	TimeString	ROPA	MWin	MWout	TempMudin	ROP1
1	45522	1338244440	66.39	8.60	6.19	63.94	49.78
2	45523	1338244445	66.39	8.60	6.19	63.94	47.07
3	45524	1338244450	50.58	8.60	6.20	63.94	52.77
4	45578	1338244720	28.31	8.60	6.22	63.93	20.31
5	45585	1338244755	28.31	8.60	6.22	63.93	23.62
6	45586	1338244760	28.31	8.60	6.22	63.93	23.49

Figure 62- Data Selection of the Prediction app displaying uploaded data in Step 1.

In *Step 1***,** Data can be uploaded (using a .csv format) to the App. Once the data is uploaded, *About File, View Data and Data Summary* tabs can be used in the main panel to view the data and summary of each variable present in the uploaded data. The Data *Summary* tab displays mean, median, and quantiles information as shown in Figure 62. The code uses a reactive format, so any changes in the file would automatically change the displayed data and its summary, as in Figure 63.

About file | Data | Summary

	X	TimeString	ROPA	MWin	MWout	TempMudin	ROP1	ROPI
1	Min. : 45522	Min. :1.338e+09	Min. : 4.37	Min. : 8.60	Min. : 6.192	Min. : 63.90	Min. : 4.375	Min. : 0.1023
2	1st Qu.: 47524	1st Qu.:1.338e+09	1st Qu.:19.70	1st Qu.: 8.60	1st Qu.: 6.896	1st Qu.: 64.75	1st Qu.:20.711	1st Qu.: 13.0406
3	Median : 49507	Median :1.338e+09	Median :29.73	Median : 8.60	Median : 7.475	Median : 65.62	Median :29.938	Median : 26.0286
4	Mean :225217	Mean :1.339e+09	Mean :27.92	Mean :10.07	Mean : 9.128	Mean : 84.26	Mean :28.290	Mean : 30.4355
5	3rd Qu.:461090	3rd Qu.:1.340e+09	3rd Qu.:34.45	3rd Qu.:11.99	3rd Qu.:11.997	3rd Qu.:109.61	3rd Qu.:35.139	3rd Qu.: 41.6175
6	Max. :464116	Max. :1.340e+09	Max. :68.16	Max. :12.32	Max. :12.050	Max. :111.71	Max. :64.058	Max. :136.6170

Figure 63- Summary tab displaying mean, median, and quantiles of the uploaded data.

5.2.2. Data Preparation

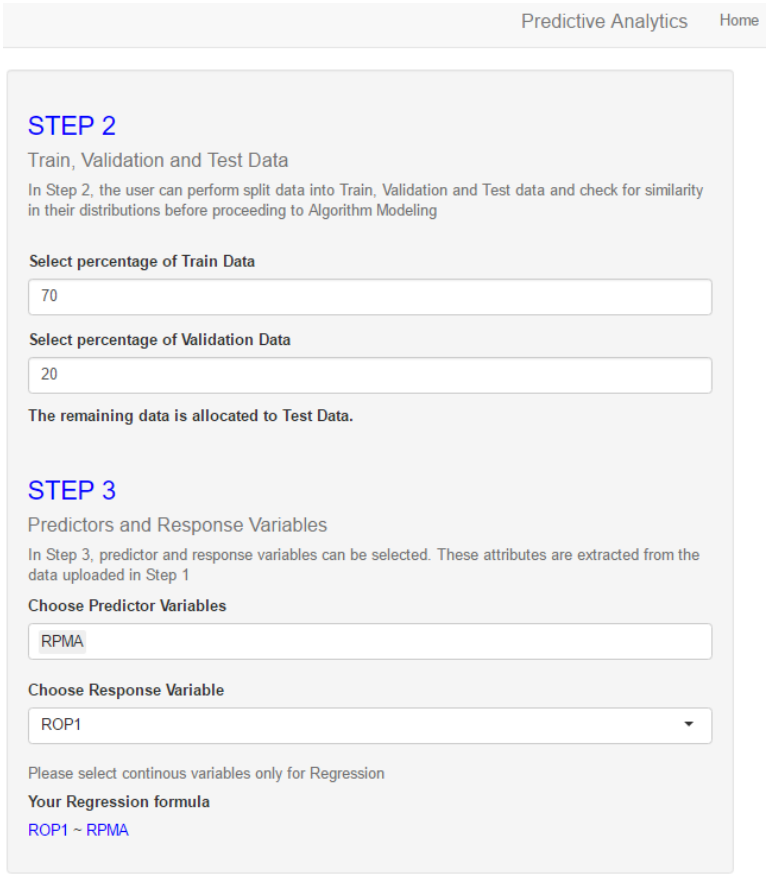


Figure 64- Step 2 of the prediction app showing data splitting options.

Data Preparation consists of two steps. *Step 2* presents options to perform data splitting by selecting a percentage of train, test and validation data. The distributions and summaries of train, test and validation data can also be analyzed as shown in Figure 64. The *Train Data*, *Validation and Test Data* display the split data and the *Summary* tab displays their summaries respectively. It is important to ensure that both train and test data follow similar distributions for an accurate model building and *Histograms* tab provides the means to ensure accurate modeling as shown in Figure 65. The tab also shows the same for the response variable.

In *Step 3*, predictors and response variables can be chosen from the list which is populated by automatically selecting column names from the uploaded data. This feature is important as it lets the user decide the parameters depending on whether he is doing a drilling analysis such as ROP optimization or production analysis such as fracture geometry prediction.

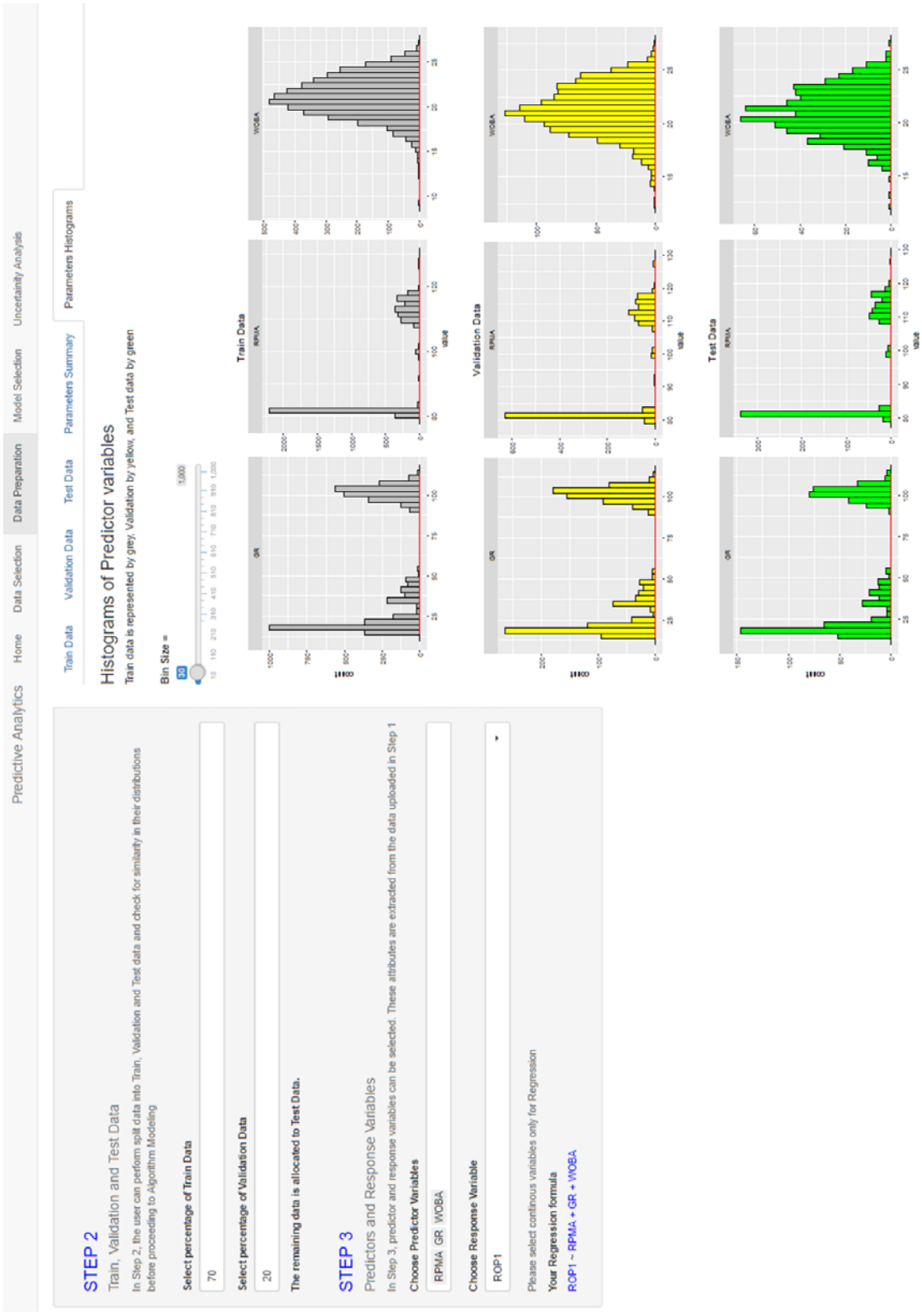


Figure 65- Histograms of train, test and validation data for predictors in Step 2.

5.2.3. Model Selection

Model Selection also included two steps. In this Step 4, the user can perform model building using various combinations of algorithms and grouping parameters, as shown in Figure 66.

1. Algorithms: Regression methods, trees, RF, GBM, KNN, and SVR.
2. Grouping parameters: Formation, Bit size, K-means clusters and Entire data.

The user can select the algorithms from a drop-down list present in the sidebar panel and compute to view several results. Train data can be selected to compute the model by looping over a grouping parameter selected (e.g. when *formation/bit size* is selected, the app builds one model for every formation/bit size level present in the data) and building separate models for each formation. If none of these attributes (formation, bit size) are present, the user can perform artificial clustering using the k-Means Algorithm on Train data. Or, the user can simply select the entire data and the app will build one model for validation and testing purposes.

The models are initially tested against the validation data (train, validation and test data are created from *Step 2*) and the prediction results, model details and prediction plots are then displayed. The *Validation Data Results* tab displays the summary of RMSE and MAE error metrics for each formation. As shown in Figure 67, the tab also has additional data such as mean and median of absolute errors and their percentage variation with respect to actual values (Med % and Mean %).

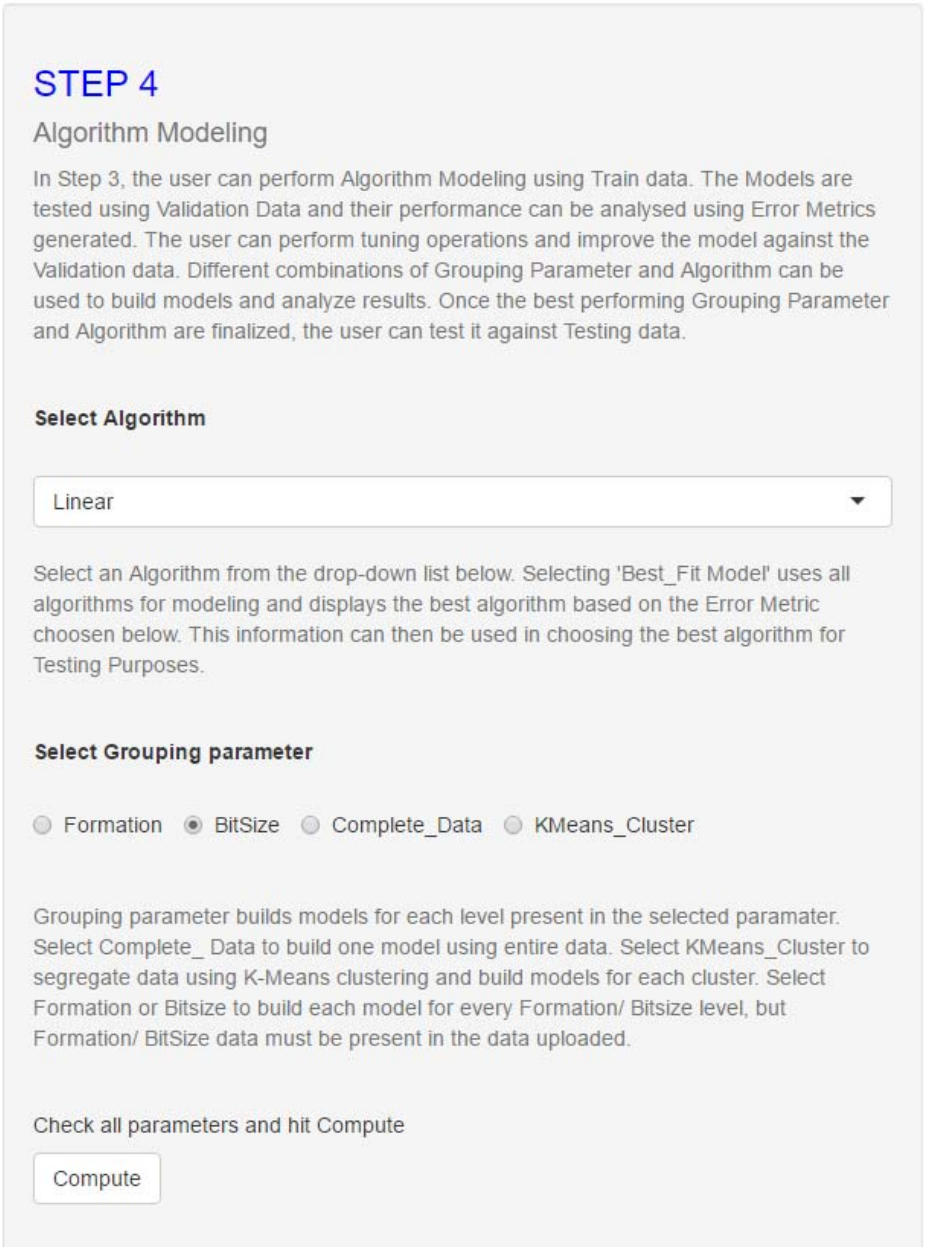


Figure 66- Details of Step 4 in the prediction app showing algorithm selection.

Data Selection | Data Preparation | Model Selection | Uncertainity Analysis

Validation Data Results | Validation Data Plots | Validation Data Models

Please donot refresh web page during Model building or Simulations. The results will be automatically displayed upon completion.

	BitSize	RMSE	MAE	MedAE	Med %	Mean %	Adj. Rsqr	Data	Algorithm
1	17.5	5.67	4.31	3.60	10.70	12.40	0.03	3332	Linear
2	8.5	7.75	6.13	5.10	28.00	39.80	0.05	2519	Linear

Figure 67- Summary of results computed using linear regression applied on grouping parameter-Bit size.

The *Validation Data Plots* tab presents the actual vs. predicted ROP plot and the *Validation Data Model* displays the details of the selected modeling technique as shown in Figure 68. Figure 69 shows the actual vs. predicted ROP plot for models built using linear regression and looped over the *Bit Size* Grouping parameter.

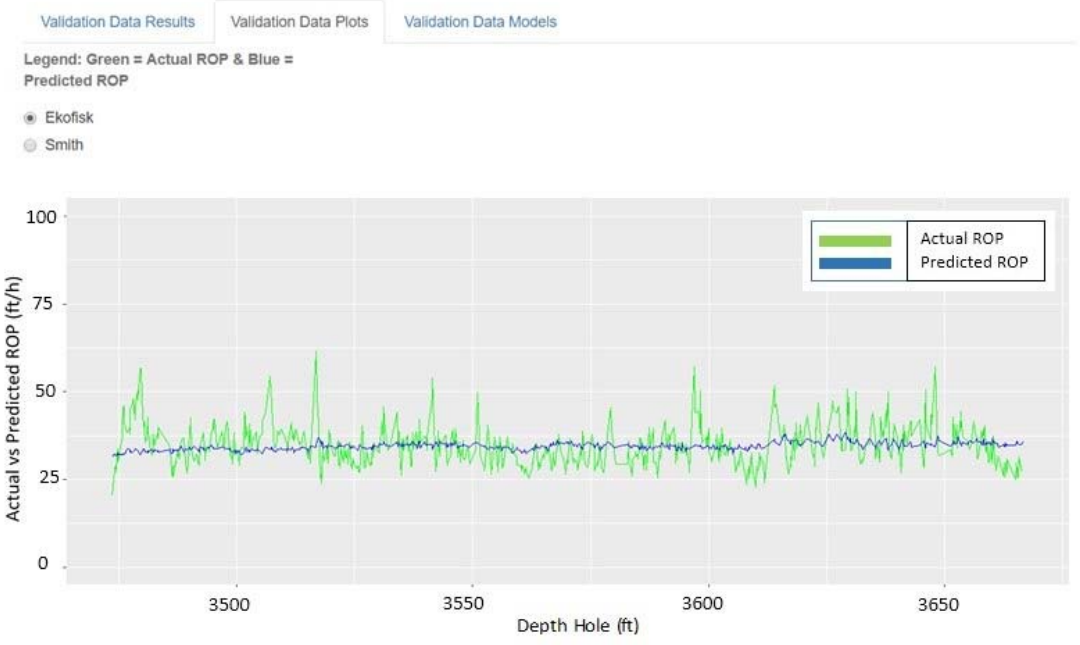


Figure 68-Prediction plot of actual vs. predicted ROP for Ekofisk formation using linear regression applied on grouping parameter-Formation.

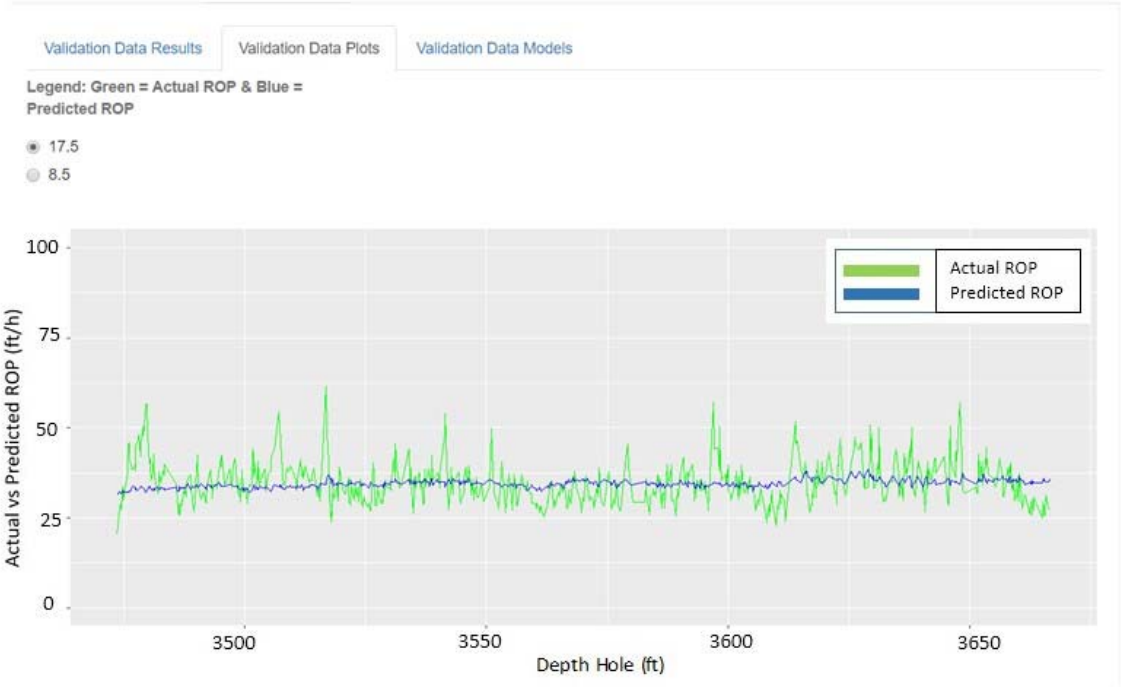


Figure 69-Prediction plot of actual vs. predicted ROP for 17.5 inch Bit size using linear regression applied on grouping parameter-Bit size.

Similar results can be achieved using the other modeling techniques such as linear, CART and GBM. Additionally, there is another option in the dropdown menu called the *Best_Fit_Model.* This option automatically uses all the algorithms present to build models, use them against test data and then chooses the best performing model (or the one with least error metric) for each formation as shown in Figures 70 (all models) and 71 (best models) respectively. The error metrics are visible in the sidebar panel.

All Algorithms | Best Algorithms

Total summary of all algorithms applied on each Clustering Parameter Level

	Formation	RMSE	MAE	MedAE	Med %	Mean %	Algorithm	Data
1	Ekofisk	5.67	4.31	3.60	10.70	12.40	Linear	3332
2	Smith	7.75	6.13	5.10	28.00	39.80	Linear	2519
3	Ekofisk	5.60	4.26	3.50	10.00	12.30	CART	3332
4	Smith	7.29	5.75	4.80	24.40	37.10	CART	2519
5	Ekofisk	6.61	5.05	4.20	12.30	14.70	Random_Forest	3332
6	Smith	8.46	6.55	5.00	26.40	41.20	Random_Forest	2519
7	Ekofisk	5.60	4.25	3.40	10.20	12.20	GBM	3332
8	Smith	7.21	5.70	4.90	24.10	36.60	GBM	2519

Figure 70-Summary of all algorithms computed for Ekofisk and Smith formations in Well 12a.

All Algorithms | Best Algorithms

The following Algorithms have the least chosen error metric and are the best for

	Formation	RMSE	MAE	MedAE	Med %	Mean %	Algorithm	Data
7	Ekofisk	5.60	4.25	3.40	10.20	12.20	GBM	3332
8	Smith	7.21	5.70	4.90	24.10	36.60	GBM	2519

Figure 71-Summary of best algorithms for Ekofisk and Smith formations in Well 12a.

The user can also tune the models using validation data and can try several combinations of algorithms and grouping parameters. After trying several combinations and achieving the best possible Grouping Parameter and Formation, the user can test these models against test data, as shown in Figure 72 in *Step 5*. The results will automatically be displayed in the main panel upon choosing test data for testing purposes.

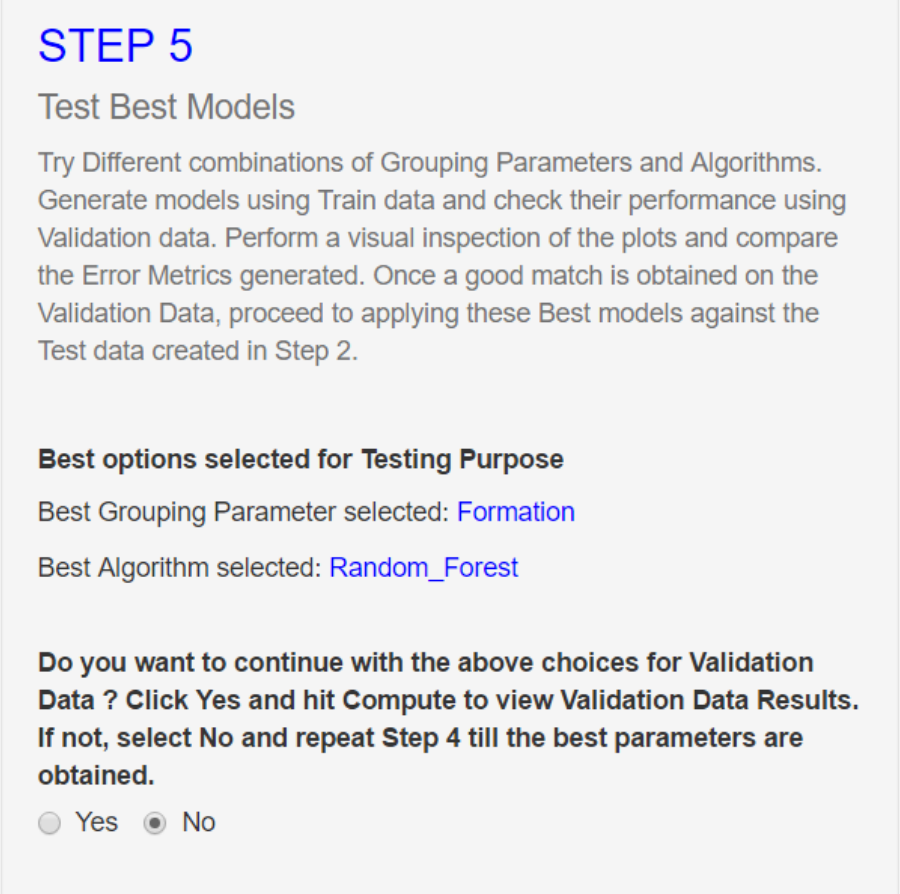


Figure 72- Description of Step 5 in the prediction app explaining the testing phase.

5.2.4. Uncertainty Analysis

In *Step 6,* the best models, based on the Grouping Parameter and Algorithm chosen in Step 5, will be used to perform both uncertainty and sensitivity analysis. The user can check model robustness and generate a range of outputs: P10, P50 and P90 on the Best Models selected in Step 5 by using simulated data, as shown in Figure 73. The input data is simulated by adjusting means and standard deviations (SDs) for predictor variables that can then be selected from the input bars in the sidebar panel.

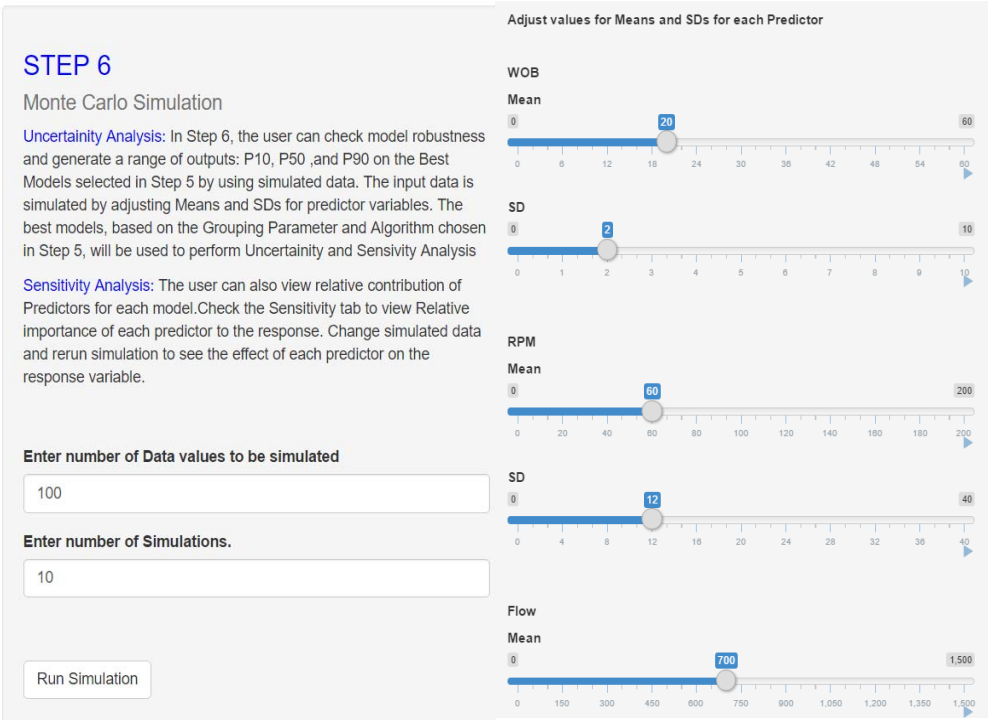


Figure 73-Description of Step 6 in the Prediction app demonstrating Monte Carlo simulation.

The SD represents the errors noticed while recording real time values in the field and the user can select a value for each predictor after consulting with domain experts. Using the simulated test data (explained above) and the best performing model from *Step 3*, Monte Carlo simulation can be performed. The user can specify the number of simulations and number of data points for each simulation. This is then used to predict a distribution of outputs for which the P10, P50 and P90 values are captured, as shown in Figure 74.

	Formation	P10	P50	P90
1	Ekofisk	36.93	36.93	36.93
2	Smith	24.39	24.39	24.39

Figure 74- Results of Monte Carlo simulation- P10, P50, P90 values for formations Ekofisk and Smith in Well 12a.

For sensitivity analysis, the user can also view the relative contribution of predictors for each model, as shown in Figures 75 and 76. Further, the user can change simulated data and rerun simulations to see the effect of each predictor on the response variable. Therefore, *Step 6* enables forward prediction and displays the observed ranges for a response variable instead of a single output value.

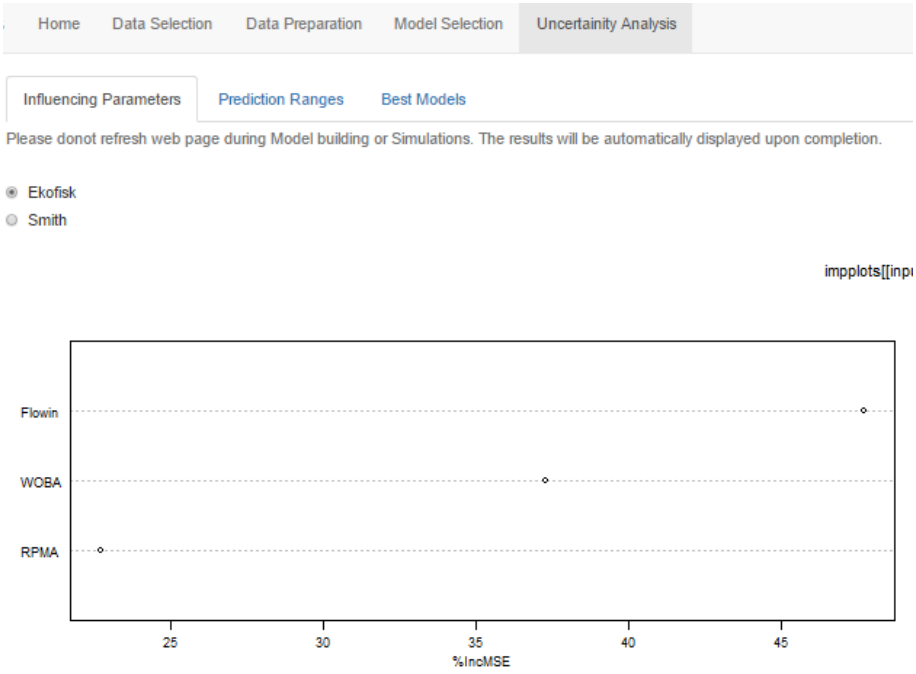


Figure 75- Sensitivity analysis indicating the relative ranking of predictors using Step 6.

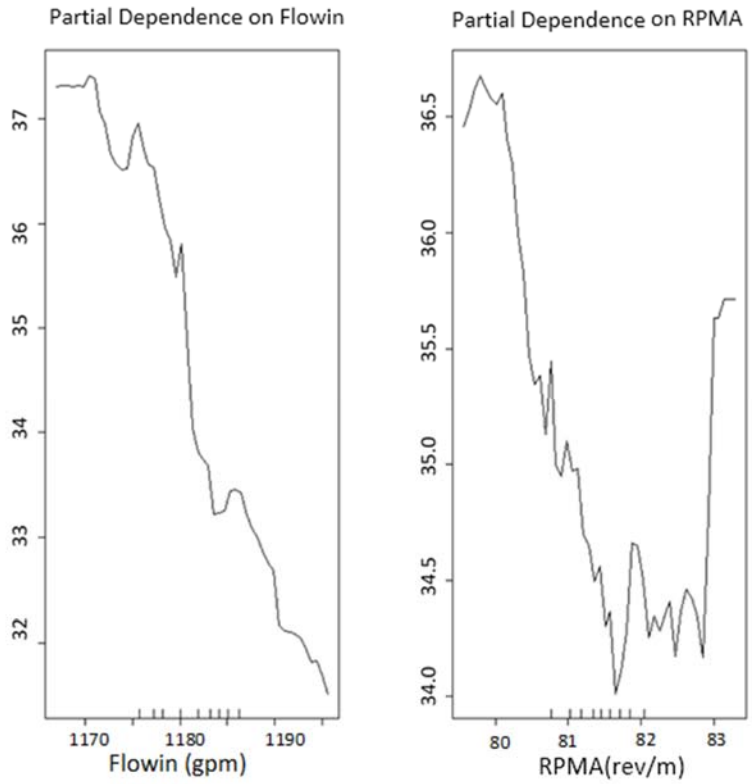


Figure 76- Partial dependency plots of Flow and RPM for Ekofisk formation in Well 12a.

Due to the vast range of features available and an easily navigable design, the predictive analytics app enables a user to apply the above methodology and use it to accurately and quickly predict ROP. The user is not expected to be proficient in R, predictive modeling or machine learning. The app also has the potential to be extended to other applications in petroleum engineering such as in production and reservoir engineering, and can be extended to other industries as well. More algorithms can be easily added to the app to assist in better and faster modeling. Also, the app can be accessed anywhere using a smart device.

6. NOMENCLATURE

ANOVA = analysis of variance

BRNN = Bayesian regularized neural network

CART = classification and regression tree

CV = cross validation

Flow = rate of flow [gpm]

GBM = gradient boosting machine

GR = gamma ray

k = tuning parameter

KNN = K-nearest neighbor

LM = multivariate linear model

LOOCV = leave-one-out cross validation technique

MAE = mean absolute error

NN = neural networks

RF = random forests

RMSE = root mean square error

ROP = rate of penetration [ft/h]

RPM = string rotary speed/revolutions per minute [rev/m]

SRM = structural risk minimization

SVR = support vector regression

WOB = weight on bit [klb]

REFERENCES

1. Liu, Z., Marland, C. and Li, D., 2014. “An Analytical Model Coupled with Data Analytics to Estimate PDC Bit Wear”. Presented at SPE Latin America and Caribbean Petroleum Engineering Conference, Maracaibo, Venezuela, 21-23 May. SPE-169451-MS. http://dx.doi.org/10.2118/169451-MS
2. Awoleke, O.O. and Lane, R., 2011. “Analysis of Data from the Barnett Shale with conventional Statistical and Virtual Intelligence Techniques”. MS Thesis, Texas A & M University, College Station, Texas, USA (October 2011). SPE-127919-PA. http://dx.doi.org/10.2118/127919-PA
3. Yi, P., Kumar, A. and Samuel, R., 2014. “Real Time Rate of Penetration Optimization using the Shuffled Frog Leaping algorithm (SFLA)”. Presented at SPE Intelligent Energy Conference & Exhibition, Utrecht, The Netherlands, 1-3 April. SPE-167824-MS. http://dx.doi.org/10.2118/167824-MS
4. Mahmood, B., Mohammadreza, K. and Ashenax, R., 2010. “Investigation of Various ROP Models and Optimization of Drilling Parameters for PDC and Roller-cone Bits in Shadegan Oil Field”. Presented at International Oil and Gas Conference and Exhibition in China, Beijing, China, 8-10 June. SPE-130932-MS. http://dx.doi.org/10.2118/130932-MS
5. Schuetter, J., Mishra, S., Zhong, M. and RaFollette, R., 2015. “Data Analytics for Production Optimization in Unconventional Reservoirs”. Presented at Unconventional Resources Technology Conference, San Antonio, Texas, USA, 20-22 July. SPE-178653-MS. http://dx.doi.org/10.2118/178653-MS
6. Mostofi, M., Shahbazi, K., Rahimzadeh, H. and Rastegar, M., 2010. “Drilling Optimization Based on the ROP Model in One of the Iranian Oil Fields”. Presented at International Oil and Gas Conference and Exhibition in China, Beijing, China, 8-10 June. SPE-131349-MS. http://dx.doi.org/10.2118/131349-MS

7. Sidahmed, M., Ziegel, E., Shirzadi, S., Stevens, D. and Marcano, M., 2014. “Enhancing Wellwork Efficiency with Data Mining and Predictive Analytics”. Presented at SPE Intelligent Energy Conference & Exhibition, Utrecht, The Netherlands, 1-3 April. SPE-167869-MS. http://dx.doi.org/10.2118/167869-MS
8. Sui, D., Nybø, R. and Azizi, V., 2013. “Real-time optimization of rate of penetration during drilling operation”. Presented at 10th IEEE International Conference on Control and Automation (ICCA), Hangzhou, 12-14 June. 10.1109/ICCA.2013.6564893
9. Bailey, R., Shirzadi, S. and Ziegel, E., 2013. “Data Mining and Predictive Analytics Transforms Data to Barrels”. Presented at SPE Digital Energy Conference, The Woodlands, Texas, USA, 5-7 March. SPE-163731-MS. http://dx.doi.org/10.2118/163731-MS
10. Dursun, S., Kumar, A. and Samuel, R., 2014. “Using Data-Driven Predictive Analytics to Estimate Downhole Temperatures while Drilling”. Presented at SPE Annual Technical Conference and Exhibition, Amsterdam, The Netherlands, 27-29 October. SPE-170982-MS. http://dx.doi.org/10.2118/170982-MS
11. Gidh, Y.K., Ibrahim, H., and Purwanto, A., 2011. “Real-Time Drilling Parameter Optimization System Increases ROP by Predicting/Managing Bit Wear”. Presented at SPE Digital Energy Conference and Exhibition, The Woodlands, Texas, USA,19-21 April. SPE-142880-MS. http://dx.doi.org/10.2118/142880-MS
12. Maysami, M., Gaskari, R., and Mohaghegh, S.D., 2013. “Data Driven Analytics in Powder River Basin, WY”. Presented at SPE Annual Technical Conference and Exhibition, New Orleans, Louisiana, USA, 30 September-2 October. SPE-166111-MS. http://dx.doi.org/10.2118/166111-MS.

13. Wang, Y. and Salehi, S., 2015. “Drilling Hydraulics Optimization Using Neural Networks”. Presented at SPE Digital Energy Conference and Exhibition, The Woodlands, Texas, USA, 3-5 March. SPE-173420-MS. http://dx.doi.org/10.2118/173420-MS
14. Dashevskiy, D., Dubinsky, V., and Macpherson, J.D., 1999. “Application of Neural Networks for Predictive Control in Drilling Dynamics”. Presented at SPE Annual Technical Conference and Exhibition, Houston, Texas, 3-6 October. SPE-56442-MS. http://dx.doi.org/10.2118/56442-MS
15. Holdaway, K.R., 2012. “Predictive Analytics: Development and Deployment of Upstream Data Driven Models”. Presented at SPE Latin America and Caribbean Petroleum Engineering Conference, Mexico City, Mexico, 16-18 April. SPE-153454-MS. http://dx.doi.org/10.2118/153454-MS
16. Raphael, S., Fuge, C.P., Gutierrez, S., Kuzma, H.A., and Arora, N.S., 2015. “Big Data Every Day: Predictive Analytics Used to Improve Production Surveillance”. Presented at SPE Digital Energy Conference and Exhibition, The Woodlands, Texas, USA, 3-5 March. SPE-173444-MS. http://dx.doi.org/10.2118/173444-MS
17. Temizel, C., Purwar, S., Abdullayev, A., Urrutia, K. and Tiwari, A., 2015. “Efficient Use of Data Analytics in Optimization of Hydraulic Fracturing in Unconventional Reservoirs”. Presented at Abu Dhabi International Petroleum Exhibition and Conference, Abu Dhabi, UAE, 9-12 November. SPE-177549-MS. http://dx.doi.org/10.2118/177549-MS
18. Siruvuri, C., Nagarakanti, S., and Robello, S., 2006. Stuck Pipe Prediction and Avoidance: A Convolutional Neural Network Approach. Presented at IADC/SPE Drilling Conference, Miami, Florida, 21-23 February. SPE-98378-MS. http://dx.doi.org/10.2118/98378-MS

19. Murillo, A., Neuman, J. and Robello, S., 2009. Pipe Sticking Prediction and Avoidance Using Adaptive Fuzzy Logic Modeling. Presented at SPE Production and Operations Symposium, Oklahoma City, Oklahoma, 4-8 April. SPE-120128-MS. http://dx.doi.org/10.2118/120128-MS
20. Buddharaju, P., Laskar, S.A., Robello, S., 2007. Robust Well Cost Estimation Using Support Vector Machine Model. Presented at Digital Energy Conference and Exhibition, Houston, Texas, U.S.A., 11-12 April. SPE-106577-MS. http://dx.doi.org/10.2118/106577-MS
21. Bhalla, D. 2015. "Ensemble Learning - Boosting and Bagging". ListenData, 24 Mar 2015, http://www.listendata.com/2015/03/ensemble-learning-boosting-and-bagging.html
22. Tan, P-N., Steinbach, M., and Kumar, V. 2006. Introduction to Data Mining. Boston: Addison-Wesley.
23. Brownlee, J. 2013. "How to Evaluate Machine Learning Algorithms". MachineLearningMastery, 27 December 2013, http://machinelearningmastery.com/how-to-evaluate-machine-learning-algorithms.
24. DataCamp. 2015. "Machine Learning in R for beginners". Datacamp, 25 March 2015, www.datacamp.com/community/tutorials/machine-learning-in-r
25. Pedro, M. 2015. "Modeling-predicting amount of rain", 24 Mar 2015, http://theanalyticalminds.blogspot.com/2015/04/part-4a-modelling-predicting-amount-of.html
26. Chapelle, O. and Vapnik, V., 1999. "Model Selection for Support Vector Machines". *Advances in Neural Information Processing Systems*, Vol 12 (1999).