The Thesis Committee for Chiranth Manjunath Hegde
Certifies that this is the approved version of the following thesis:

Application of Statistical Learning Models to Predict and Optimize Rate of Penetration of Drilling

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Kenneth E. Gray

Hugh C. Daigle

Harry R. Millwater Jr

Application of Statistical Learning Models to Predict and Optimize Rate of Penetration of Drilling

by

Chiranth Manjunath Hegde, B.Tech.

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Engineering

The University of Texas at Austin
December 2016

Dedication

To my family, lindt, friends, upperclassmen, professors and everyone else who has helped me along the way.

Acknowledgements

The opportunity for me to come to Austin TX from Bangalore IN to attend the best program in Petroleum Engineering has been wonderful. I am extremely grateful to my research advisor, Dr. Kenneth E. Gray, who has been a source of advice, mentorship and guidance since our first email correspondence. Dr. Gray's knowledge about rock mechanics, drilling, completions, etc. stretches over the duration of the field's development, and his overview has been paramount in shaping my research and work during my Master's degree. I would also like to thank Dr. Hugh C. Daigle and Dr. Harry R. Millwater for being readers of my thesis.

I would also like to thank sponsors of the Wider Windows Industrial Affiliate Program: British Petroleum; BHP Billiton; Chevron; ConocoPhillips; Halliburton; Marathon Oil; National Oilwell Varco; Occidental Oil and Gas; and Shell Oil Company. Research guidance and constructive critiques at program review meetings were very helpful. Additionally, field data sets provided by Chevron, Marathon, and National Oilwell Varco for use in model validations are gratefully acknowledged.

Finally, I would like to thank the other students of the Wider Windows IAP. Our meetings have been very useful, and I wish you all the best in your academic and professional careers.

Abstract

Application of Statistical Learning Models to Predict and Optimize Rate of Penetration of Drilling

Chiranth Manjunath Hegde, M.S.E

The University of Texas at Austin, 2016

Supervisor: Kenneth E. Gray

Modeling the rate of penetration of the drill bit has been essential to optimizing drilling operations. Optimization of drilling – a cost intensive operation in the oil and gas industry– is essential, especially during downturns in the oil and gas industry. This thesis evaluates the use of statistical learning models to predict and optimize ROP in drilling operations.

Statistical Learning Models can range from simple models (linear regression) to complex models (random forests). A range of statistical learning models have been evaluated in this thesis in order to determine an optimum method for prediction of rate of penetration (ROP) in drilling.

Linear techniques such as regression have been used to predict ROP. Special linear regression models such as lasso and ridge regression have been evaluated. Dimension reduction techniques like principal components regression are evaluated for ROP prediction. Non-linear algorithms like trees have been introduced to address the low accuracy of linear models. Trees suffer from low accuracy and high variance. Trees are bootstrapped and averaged to create the random forests algorithm. Random forests algorithm is a powerful algorithm which predicts ROP with high accuracy.

A parametric study was used to determine the ideal training sets for ROP prediction. It was conclude that data within a formation forms the best training set for ROP prediction. Parametric analysis of the length of the training set revealed that 20% of the formation interval depth was enough to train an accurate predictor for ROP.

The ROP model built using statistical learning models were then used as an equation to optimize ROP. An optimization algorithm was used to compute ideal values of input feature to improve ROP in the test set. Surface controllable input features were varied in an effort to improve ROP. ROP was improved to save a predicted total of 22 hours of active drilling time using this method.

This thesis introduces statistical learning techniques for predicting and optimizing ROP during drilling. These methods use input data to model ROP. Input features (surface parameters which are controllable on the rig) are then changed to optimize ROP. This methodology can be utilized for reducing nonproductive time (NPT) in drilling, and applied to optimize drilling procedures.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1: Motivation

Drilling costs occupy a huge portion of the oil and gas budget (Kitchel et al., 1997). This emphasizes the need for drilling efficiently. The rate of penetration (ROP) of drilling, correlates well with drilling efficiency, although it is not the only factor (Hegde et al., 2015). High ROP will save time which translates to operational costs savings. As a result the Wider Windows IAP on drilling performance and rate of penetration (ROP) modeling looked into traditional models (or physics-based modeling) of ROP, torque and MSE. The traditional models were problematic given their low accuracy. This was attributed to the presence of empirical constants. However a better method for prediction was required which would not rely on empirical constants. This problem was solved using data analytics and statistical learning. This thesis makes use of data analytics and statistical learning to build data-driven models for the accurate prediction of ROP during drilling. An introduction is given to statistical learning algorithms and model building. ROP models are then built using statistical learning algorithms. These models are compared with traditional models to evaluate their accuracy and goodness of fit. A parametric study has been conducted to evaluate the data requirements for these statistical learning models.

1.2: THESIS ORGANIZATION

This thesis has been divided into nine chapters. The second chapter will provide a literature review on the topic of statistical learning and its application in drilling engineering. It will also include a literature review of the several traditional ROP models in drilling. The third chapter will introduce the data set used for validation of these models. It will also describe methods adapted ensure that there is no overfitting of the data. Cross validation and bootstrapping are covered which will be actively used in the algorithms in following chapters. The fourth chapter will introduce regression and its application in ROP prediction. The applications of regularized regression and principal components regression (PCR) have been discussed. The fifth chapter will cover nonlinear prediction techniques such as trees and random forests. The sixth conducts a parametric study of these models to determine the ideal size, volume and type of data required for accurate ROP predictions. The seventh chapter explores the use of these models to predict ROP and optimize ROP. The predicted ROP can be optimized based on changing input parameters on the surface of the rig based on model recommendations. The eighth chapter presents future work and continuation of this thesis. The ninth and the last chapter will provide a summary of the content of this thesis.

Chapter 2: Literature Review and Background

Optimizing drilling operations is extremely important to the success of an oil and gas project. Nonproductive time (NPT) accounts for a significant portion of drilling budget, and reducing NPT is important to keep drilling costs low. ROP is directly proportional to the cost of drilling, since increase of ROP reduces operational costs. This section looks at a review of ROP models.

2.1: Traditional ROP Models

The speed of drilling generally has high correlation to the rate of penetration of a well as long as the bit is intact. As a result improving the ROP as well as predicting it has been a subject to a great deal of research in the past. These models have been improved over decades based on their requirement and technical advances in drilling. However, most of these models are still empirical in nature i.e. they contain empirical constants. These empirical contestants need to be determined and adjusted for each formation or lithology. These adjustments are made based on the data acquired during drilling the well (or pad wells). Some of these models have bounds on the empirical constants. But for the most case this range must be determined by engineering judgement or data.

One of the earliest ROP models was developed by Maurer (1962) where the author applied a rock cratering approach to develop a ROP formula for roller-cone bits. The parameters included weight-on-bit (WOB), rotary speed (RPM), bit diameter and strength of rock. Despite theoretical backing for this model, an empirical coefficient was adopted. An important concept introduced by Maurer was rock floundering: beyond a certain WOB there was no improvement in ROP because of the reduction in hole cleaning ability. The cuttings would accumulate around the bit, making it harder to clean at the bit. This would subsequently reduce the ROP. A model for prediction of ROP was introduced by Bingham (1965), using parameters of weight on bit (WOB), rotations per minute (RPM), and bit diameter. An empirical constant 'k' was used, which was formation dependent. This paper stressed on the importance of hole cleaning ability and its relation to ROP. A model introduced by Eckel (1967) incorporated the effects of drilling mud on ROP. A Reynolds number function was used to correlate ROP with mud properties. It was showed that an increase in the Reynolds number function correlated well with high ROP measurements. Based on this paper it was concluded that a mud with a low kinematic viscosity would be recommended for easier drilling or higher ROP yield.

Bourgoyne and Young (1974) introduced a more sophisticated model with additional parameters in order to include more physical and geological aspects involved in drilling. This model is perhaps the most comprehensive model to date which describes ROP. The model contained eight parameters namely: formation strength, normal compaction trend, under compaction, differential pressure, bit diameter and bit weight, rotary speed, tooth wear, and bit hydraulics.

Walker et al. (1986) introduced a model which utilized triaxial rock strength tests and the Mohr-Coulomb failure criterion to develop a roller cone ROP equation dependent on WOB, borehole pressure, rock porosity, average grain size, and in-situ formation compressive strength. Warren (1987) developed a model which separated the effects of drilling into physical breakage of the rock and hole cleaning. This model has been shown by Soares (2015) to work well in low differential pressure but fails in cases of higher differential pressures. Winters et al. (1987) added a fourth term to the Warren (1987) equation: rock ductility.

Hareland and Rampersad (1994) introduced a drag bit model which was later modified by Motahari et al. (2010). The original drag bit model contained three empirical parameters to model lithology and other eccentric factors. Optimization of drilling was reported with some success as authors began to use well logs along with drilling simulators. Reports by Gjelstad et al. (1998) and Nygaard et al. (2002) have shown good cost reduction in North Sea drilling operations using ROP models. Motahari et al. (2010) discussed a PDC (polycrystalline diamond compact) bit model where the effect of PDMs (positive displacement motors) were accounted for. This model is useful given the prolific use of PDC bits for drilling in the present day scenario. The paper emphasizes the importance of torque at the drill bit.

All the models covered so far are not predictive in nature and cannot adapt to new lithology while drilling. Entering a new lithology, or change in wellbore trajectory or change in rock type, for example, would require re-determination of all the empirical constants.

2.2: STATISTICAL LEARNING MODELS

Bilgesu et al. (1997) used neural networks to predict ROP, however, this paper failed to adequately address the issue of data quality, data volume, algorithmic development, and ROP prediction between multiple formations. The authors also included some empirical variables in their ROP formulation - bit-wear, tooth-wear, and formation drill ability – which defeat the purpose of using these models as compared to traditional ROP models. Jahanbakhshi and Keshavarzi (2012) explored technique using different input parameters. The authors included empirical input feature in their paper which determined by the neural network algorithm. They employed a 75% training set, which would not be practically applicable in drilling. Dunlop et. al (2011) created a model with two input parameters to optimize ROP: RPM and WOB. The paper employed analytical methods to solve for the best ROP. Their paper failed to include other effects on the bit such as bit cleaning, bit wear, and pressure in the annulus which affect the ROP. ROP was optimized purely based on WOB and RPM. The work of Hegde et al. (2015a) has been insightful in introducing statistical learning methods to predict ROP without the inclusion of empirical parameters. The authors used machine learning and ensemble learning techniques to predict the ROP during drilling. The paper predicted ROP with a good accuracy using the random forests algorithm. Training and test-sets were labelled clearly which indicated their usability in drilling. Other work includes the use of statistical methods by Hegde et al. (2015b) to infer rather than predict which can be used to make decisions. Wallace et al. (2015) developed a method to determine incorporate this statistical model into real time drilling operations. This paper laid out the blueprint to use statistical learning techniques and incorporate these techniques in the drilling workflow so that they may be used on a rig for real-time drilling analysis. In contrast with the traditional models, statistical learning models utilize surface measured parameters such as weight on bit, rotations per minute, and flow rate to predict ROP. Machine learning can be used for accurate ROP prediction during drilling within a given facies or even for multiple facies in succession (with adequate training data). Machine Learning (ML) methods are advantageous since they do not contain any empirical constants or bit specifications and are not bound to a borehole assembly (BHA).

Chapter 3: Model Validation

The wider windows statistical learning model (WWSLM) uses input parameters (input features), such as weight on bit (WOB), rotary speed of the bit (RPM), and flow-rate. Drilling data measured on the surface include other parameters such as block height, differential pressure, hook load, rock strength, and torque. These input parameters (WOB, RPM and flow-rate) are then utilized to 'train' an ROP predictor. Training a model is where the model is built (or formed) based on the training data. The input parameters are user-selected; the accuracy of the resulting model will depend on input parameters, data quality and model algorithm. In the examples shown in this thesis, only surface measurements were used for ROP prediction. Other variables such as mud properties, drill string and bottom-hole assemblies were not included, but they could be. Basic requirements for WWSLM are minimal making it user-friendly and rig adaptable (Hegde et al., 2015a).

3.1: Data Exploration

Since this project is based on the use of data-driven models for ROP prediction, analyzing the data is important. This thesis utilizes drilling data (measured at the surface) from one vertical well drilled by Marathon Oil in the Williston Basin, North Dakota. The data includes measured ROP with depth. Other drilling parameters like weight on bit (WOB), rotations per minute (RPM), flow rate, differential pressure, strength of the rock, and torque were measured among several other entities. The well was drilled through 18 formations consisting of three types of rocks: sandstone, shale and limestone. Figure 1 shows a plot of the ROP against depth for the data set used in this thesis. All the formations have been identified separately.

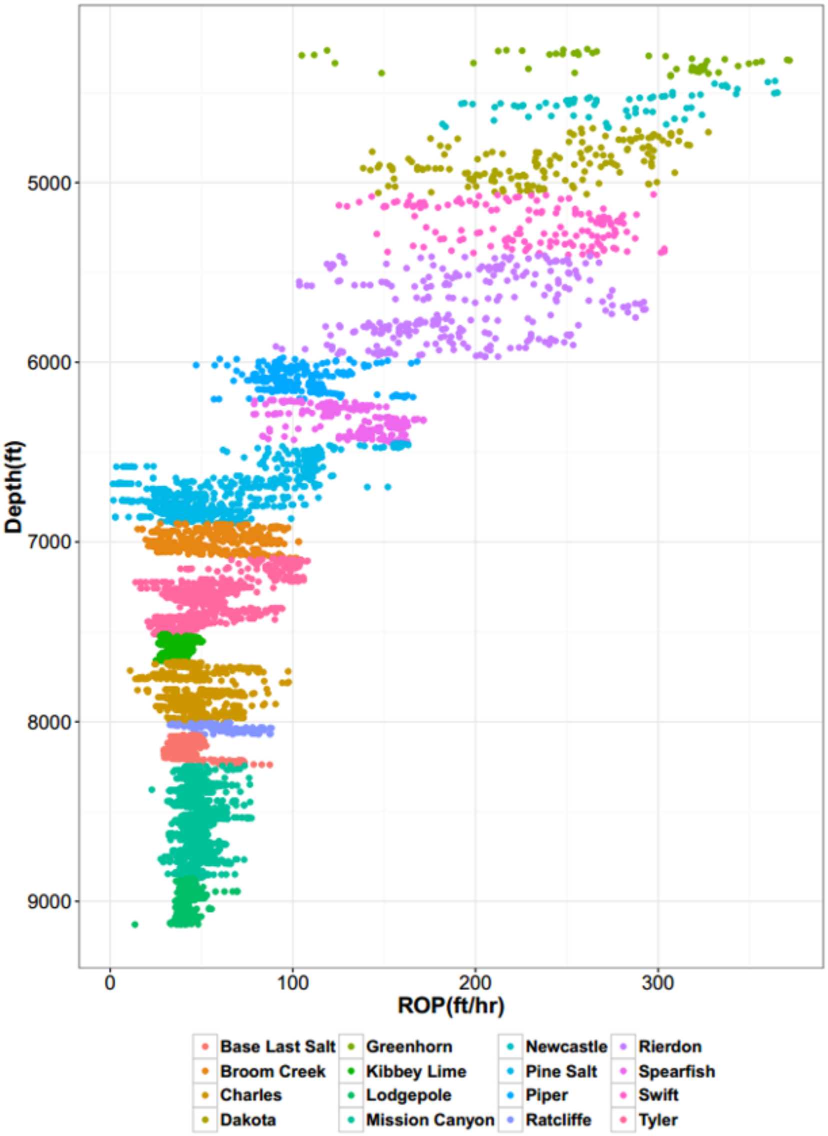


Figure 1: ROP versus Depth of Drilling for the Vertical Well Dataset

Based on figure 1, some outlier data points can be omitted from the analysis. The ROP spikes at depths of 7200 ft, 8550 ft and 9150 ft can be attributed to errors in measurement.

The data was provided in the form of .csv files which contain a data row for each 0.5 ft of the well's progress. There were over twenty different columns of data measured. The file contains Measured Depth, Bit Position, Bit Weight, Flow Weight, and recorded ROP, among others. Also included were data from several sensors. While the data from these extra sensors were interesting to look at, they were not used in this study as one of the main objectives of this work is to come up with a valid method to predict ROP using surface measured parameters that are always available (independent of the drilling or LWD contractor used). Figures 2, 3 and 4 below show plots of individual rock types drilled in the vertical section of the hole.

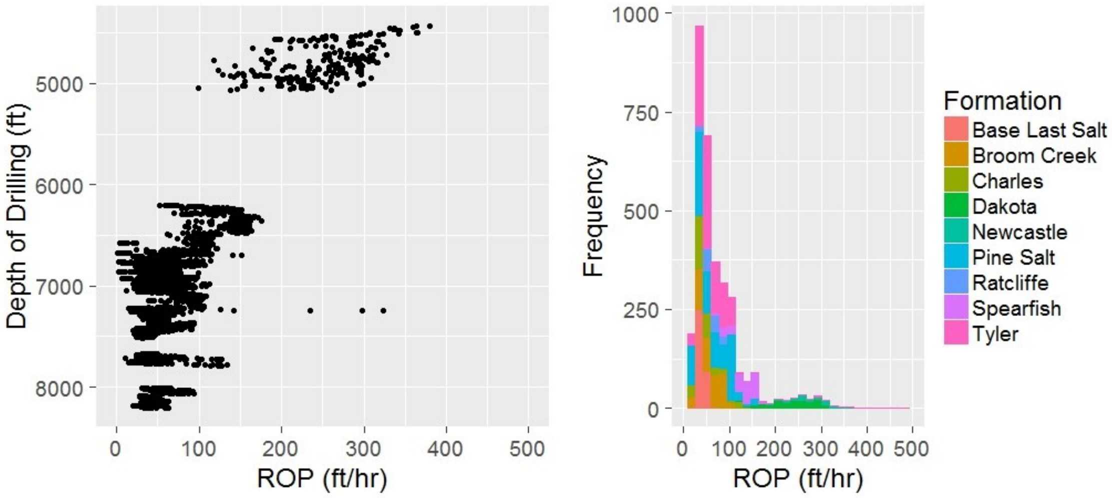


Figure 2: ROP vs Depth for sand rock in vertical well (left) and histogram of the ROP color coded by formation (right) for Sandstone

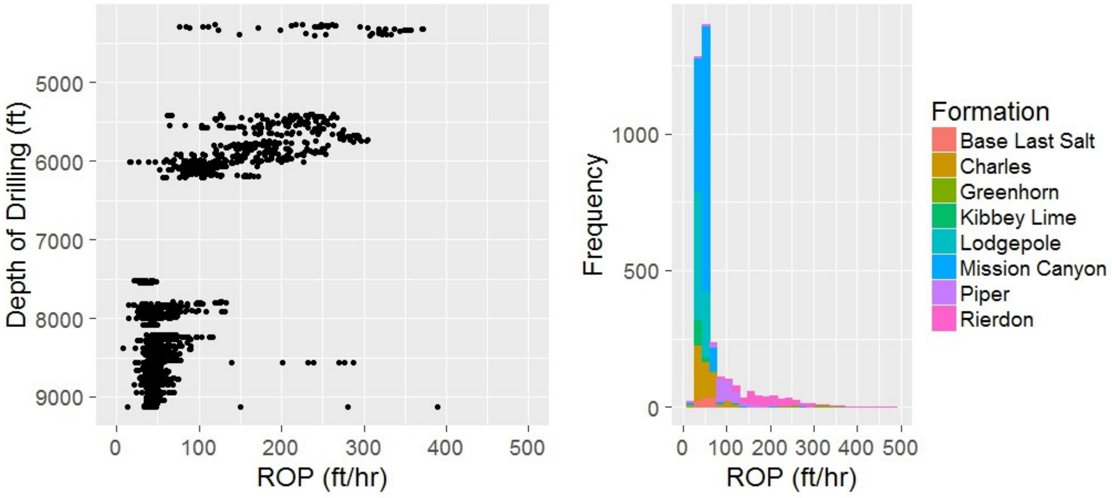


Figure 3: ROP vs Depth for sand rock in vertical well (left) and histogram of the ROP color coded by formation (right) for Limestone

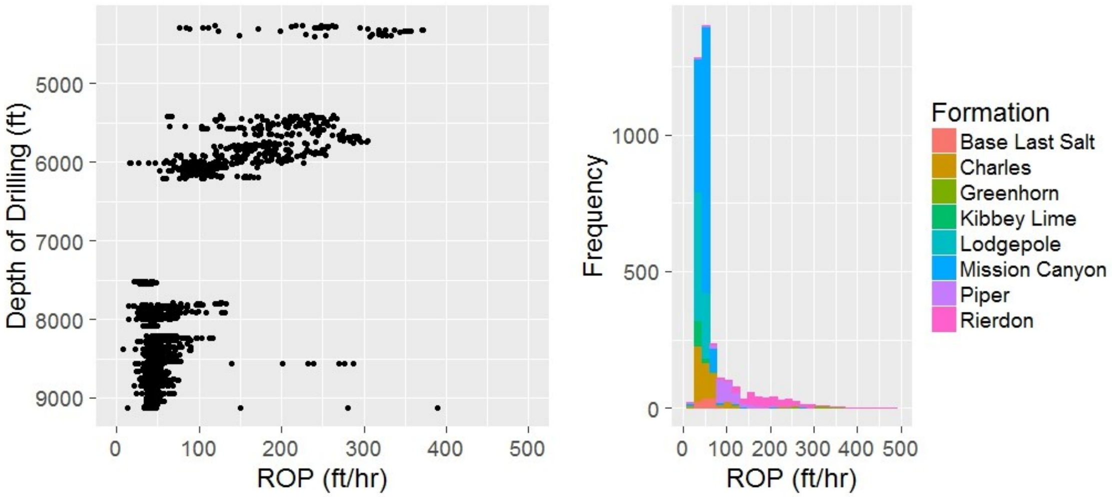


Figure 4: ROP vs Depth for sand rock in vertical well (left) and histogram of the ROP color coded by formation (right) for Shale

Figures 2, 3 and 4 can be used to draw basic conclusions about the data. Figure 2 shows data for sandstone; the ROP has no clear correlation to depth. This is also true for limestone and shale (Figures 3 and 4). The histograms in Figures 2, 3 and 4 can be used to infer the modal ROP in each rock. This is around 50-60 ft/hr for limestone and shale, a bit higher for sand which is about 80ft/hr. Sandstone formations have a clearer-cut demarcation of ROP by formation, as can be deduced by looking at the formation color-coded histogram in Figure 2.

A pairs plots can be used to determine correlations between different parameters in the data. Figure 5 below shows a pairs plot for data collected in sandstones. The pairs plot allows the study of interaction of multiple features on one plot.

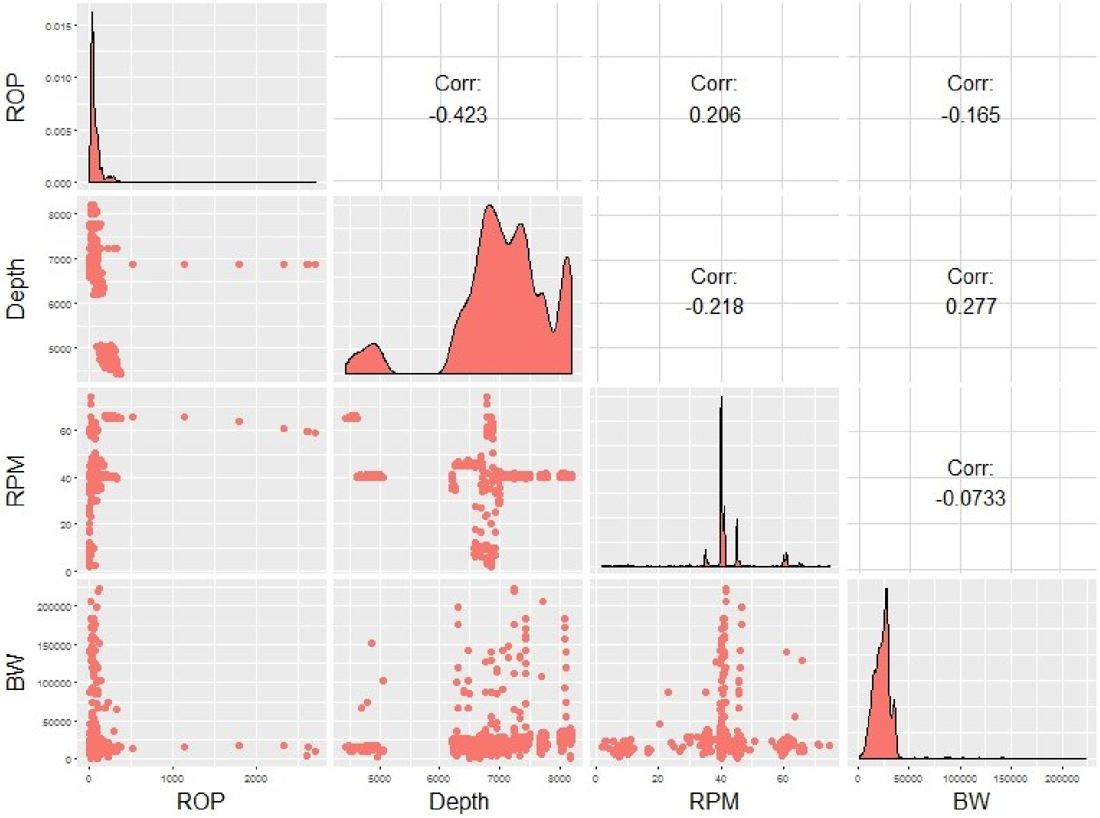


Figure 5: Pairs plot for a subset of the sandstone rock data

Figure 5 shows a generalized pairs plot for a subset of sandstone data. ROP has been compared with bit weight, depth, and RPM. The correlations between the variables are visible on the plot in Figure 5; this enables us to look at different variables and their pairwise correlation. The correlation coefficient between different variables can be used to select input parameters for the data-driven model. Variables with low correlation to the data may be discarded at this stage.

3.2: Data Management

Data-driven models are prone to overfitting of the data. This can lead to errors in the prediction stage. To prevent such errors, techniques have to be adapted to avoid overfitting of models. The data are split into three partitions. The partitions include training set, test set and validation set. Training sets consisted of 60% of the data, which was used to train the data-driven models. The validation set was used to fine tune parameters in the algorithms used to build the ROP models. The test set was a blind set (held out set) which is used to assess model accuracy. The behavior of the model on the test set is considered to be an ideal representation of model performance on new data.

3.2.1: Training Set

This constitutes the largest portion of the data set. The training set is the portion of the data set that will be used to "train" or build the model. A large percentage is used for training since the accuracy of the model generally depends on the volume of data sued in the training set. However, this set cannot be used to assess the model since the model was built using this data. This would drive the assessed error close to zero, making the result a false deduction of accuracy. This thesis has used about 60% of the data set for training.

3.2.2: Test Set

This is the blind set which is held out form the model during training. It used to test the asses the quality of the model. This thesis has used 20% of the data set towards testing of the model.

3.2.3: Validation Set

The validation set is used to fine tune the models. Algorithms have parameters which are determined based on the data. In these cases validations sets maybe used. The remainder 20% of the data set is used as the validation set in this thesis.

3.3: Model Assessment

The advantages of a model is evaluated based on error rate to determine the accuracy of the model. Assessment of model accuracy utilizes the root mean squared error (RMSE) of the model on the test data. The mathematical definition of RMSE is shown in Equation 1.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(Actual\ ROP - Predicted\ ROP)^2} \qquad \text{(Equation 1)}$$

where ‘n’ is the total number of points being evaluated. The advantages of using RMSE to measure the error is the consistency of units. RMSE has the same units as the entity measured (ft/hr).

The goodness of fit of a model (R^2) can also be used to assess a model, also called the coefficient of correlation. This can be important in forecasting trends of ROP. Sometimes the RMSE may be high due to inaccurate predictions of a few points. Evaluation of model fit (R^2) are useful in these situations since it shows the trends of the models.

3.4: Cross Validation

Overfitting - a common phenomenon associated with statistical and machine learning models – is when the model performs well on training data but fails to replicate similar results on a test set. To avoid overfitting, common practice dictates the use of a separate test set (as mentioned in a previous section). Cross validation is often used to avoid overtraining. Cross validation splits the training set into *K* parts (called K-fold cross validation). Let us assume a case with *K*=N: the training set is in a 1:N split with the validation set. An ROP model is built based on the current training set. This process is repeated 'N' times, creating 'N' ROP models until all parts of the data are used effectively for training and validation. The 'N' models are averaged to yield one final ROP prediction model. This randomization (of validation and training sets) and averaging (of models) helps improve the accuracy of the model by reducing the variance associated with prediction.

In summary the following procedure is followed for each of the *K* 'folds':

- A model is trained using *K-1* of the folds as training data;
- The resulting model is validated on the remaining part of the data

This process is computationally intensive, but it helps increase the model accuracy on test or blind data sets. Another cross-validation method employed is: leave one out cross validation (LOOCV), where all but 1 data points are used for training and it is validated on the remaining single left out data point. This process is repeated until all points have been tested. The number of splits in this case will be equal to the number of data points. This will be more computationally intensive, however, the models tend to have lower variance since they've been averaged more times making it more accurate.

3.5: The Bootstrap

The "bootstrap" is one of the most powerful computational statistical measures which can be used to assign accuracy to statistical estimates (Efron and Tibshirani, 1993). The concept behind the bootstrap is to draw multiple samples from a distribution with replacement, thereby creating multiple samples from the original sample. Each bootstrapped sample will contain the same number of samples as the original sample set. Since some draws are repeated, each sample set will be unique in its identity. Bootstrap has numerous applications as summarized by Davison and Hinkley (1997), and will be very pivotal in the development of the random forest algorithm. The bootstrap helps create multiple pseudo training sets for building ROP models. Simulated data can be resampled using many resampling procedures such as the Monte Carlo method. However, resampling of an unknown population is not possible. In this case, drilling data have been measured, which can be assumed to be a sample of a larger population. Re-creation of this population is possible to a certain extent by randomly drawing samples from the provided sample (i.e. drilling data). Resampling drilling data randomly with repetition will create a pseudo population of the drilling data. This will enable building multiple ROP models on this sample, and then averaging them to decrease the variance of prediction of these models. Therein lies the power of the bootstrap. This concept will be used in Chapter 5 where the random forests overcome the shortcomings of trees will be overcome by random forests by using the bootstrap.

Chapter 4: Linear Prediction Methods[1]

Regression is the most widely used algorithm for prediction of linear data. It can be a very powerful (yet simple) technique for prediction of linearly related data. For non-linear data regression methods can be used for inferential analysis as outlined by Hegde et al. (2015b).

4.1: Least squares regression

Simple linear regression is the most widely adopted algorithm used to predict a response given input data. It assumes a linear relationship between the input and output variables. Mathematically it can be described as shown in equation 2:

$$\text{ROP} = \sum_{n=1}^{N} f_n x_n, \qquad \text{(Equation 2)}$$

where x_n are the input variables (or input features) of the model. ROP is the target variable predicted as the linear sum of input features. The coefficients f_n are calculated by minimizing the sum of squares of the errors. Estimation of regression coefficients is described in more detail by Hastie et al. (2013).

[1] Hegde,C.M., Wallace S.P. and Gray, K.E. (2015b). Use of regression and bootstrapping in drilling: inference and prediction. Presented at SPE Middle East Intelligent Oil & Gas Conference & Exhibition, Abu Dhabi, United Arab Emirates, 15-16 September. SPE-176791.
The author of this thesis was the primary author of the paper

Data from the Tyler sandstone formation were used to predict ROP using linear regression. The data were partitioned into training and test sets; the ROP model was built on the training set using a linear regression algorithm. ROP was predicted on the test set and the results have been plotted in Figure 6.

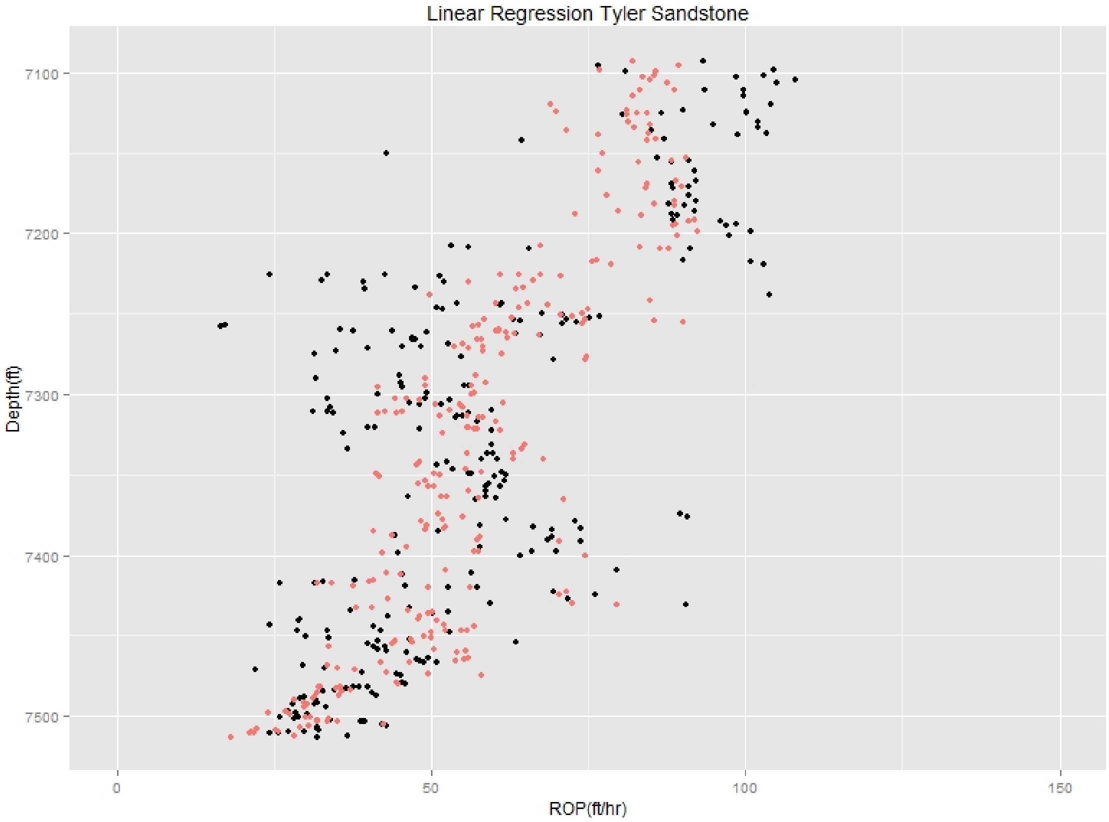


Figure 6: Linear Regression Model used to predict ROP while drilling in Tyler Sandstone. Pink represents the model whereas the black points on the plot represent the actual data.

Overall the fit seems satisfactory where the predicted values seem to lie in the same neighborhood as the actual data. The RMSE for the model was 19.23 ft/hr (33.9 % of the mean ROP in the formation); this makes using linear regression as a prediction algorithm infeasible from a practical point of view. The model's R^2 was 0.45. The low R^2 and high error rate of the model indicate the need to improve this method for ROP prediction. Table 1 summarizes the input variables their importance (t-value).

Input Variable	f_n	t-value
Depth	-0.054	-3.47
Hook Load	0.001634	5.96
RPM	-11.17	-5.127
Standpipe Pressure	-7.034	0.017695
Block Height	3.08E-02	0.336812
Bit weight	-2.99E-04	-4.711
Pump Pressure	-0.534	-0.599
Rock strength	-9.15E-04	-2.828
Intercept	19970	2.5

Table 1: Linear Model Feature Analysis

A t-test is used to determine the importance of features, higher the t-value of an input parameters, the higher is its importance. Each feature is associated with a physical meaning. The intercept should be the value of ROP when all other input features are set to zero. The intercept should be dropped since it does not makes sense from an engineering point of view (if RPM is zero as is the bit weight, it's impossible to have a non-zero ROP). When the intercept is dropped and a new linear ROP model is built, the R^2 term increases to 0.9052. Figure 7 shows the ROP predictions of the improved ROP model.

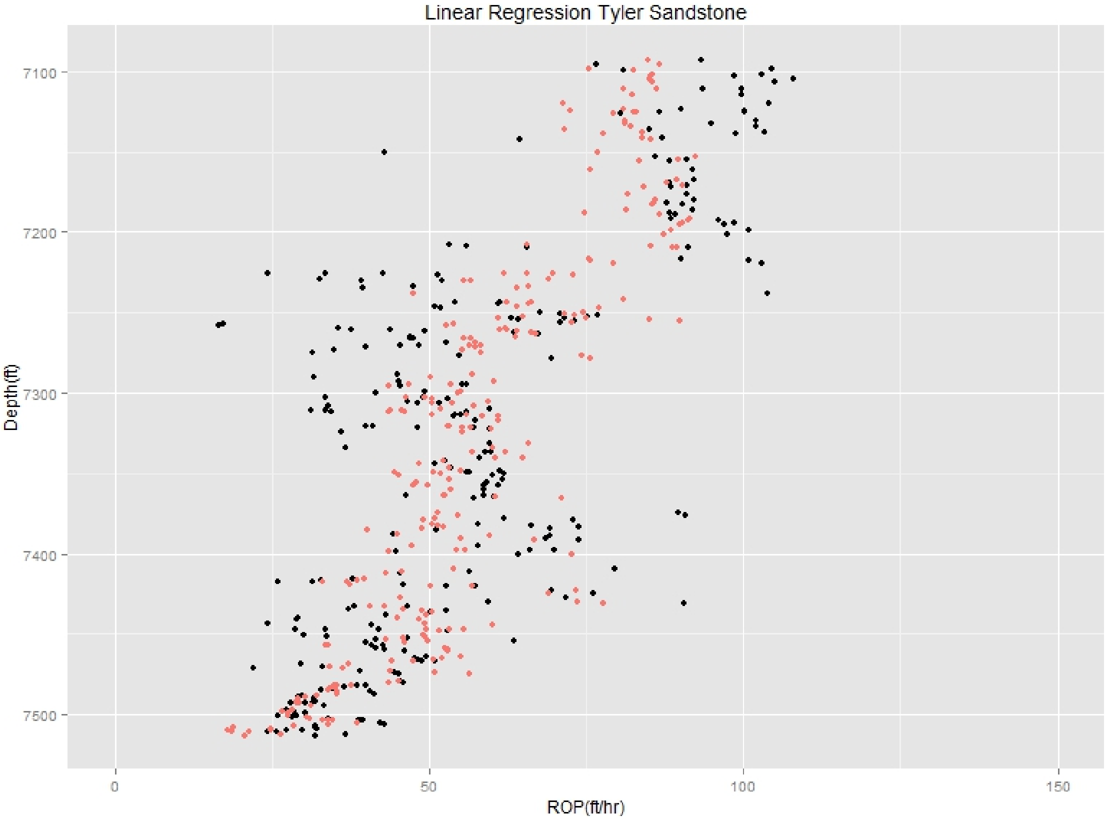


Figure 7: Improved regression model predictions in Tyler sandstone

The success of the model largely depends on the input features used in the model. Adding more features or dropping features may increase or decrease the accuracy. Forward selection or backward selection can be used to select the optimum number of features in the model. The Bayesian information criterion (BIC) or the adjusted R^2 (Hastie et al, 2013) can be used as tests to evaluate the quality of the resultant model.

The forward model starts with a null model and adds features. Measuring the RMSE and R^2 of each additional feature can help determine the best model. The backward model removes one feature repeatedly from the model, until the best model has been determined. Figure 8 shows this modeling procedure where BIC was measured for a model built using a different number of input variables. A higher BIC indicates a better model. A feature is colored black in Figure 8 if it is included in the ROP model; a white block would indicate exclusion. The best model as determined by feature selection includes hook load, bit weight, strength of rock and RPM as input features. Features which can be directly controlled on the surface should be used as input variables for the model since they can be controlled in an effort to improve ROP.

Figures 9, 10, 11 and 12 show plots of ROP predictions using ordinary linear regression (OLS) algorithm for ROP prediction in different formations. For Rierdon Limestone in Figure 9 the prediction is fairly good until a depth of 6700 ft after which the predictions are skewed to the right. However, the R^2 is really high. Figure 10 barely has enough data points to make concluding remarks, which is a possible realistic scenario, i.e. in cases of thin formations. Figure 10 has a lower error but the R^2 is low as well. Several conclusions can be drawn from Figures 9 - 12. R^2 needs to be high to ensure that the model will continue to follow the same trend as the actual data. Error percentage (error normalized to the mean) needs to be low to ensure useful predictions.

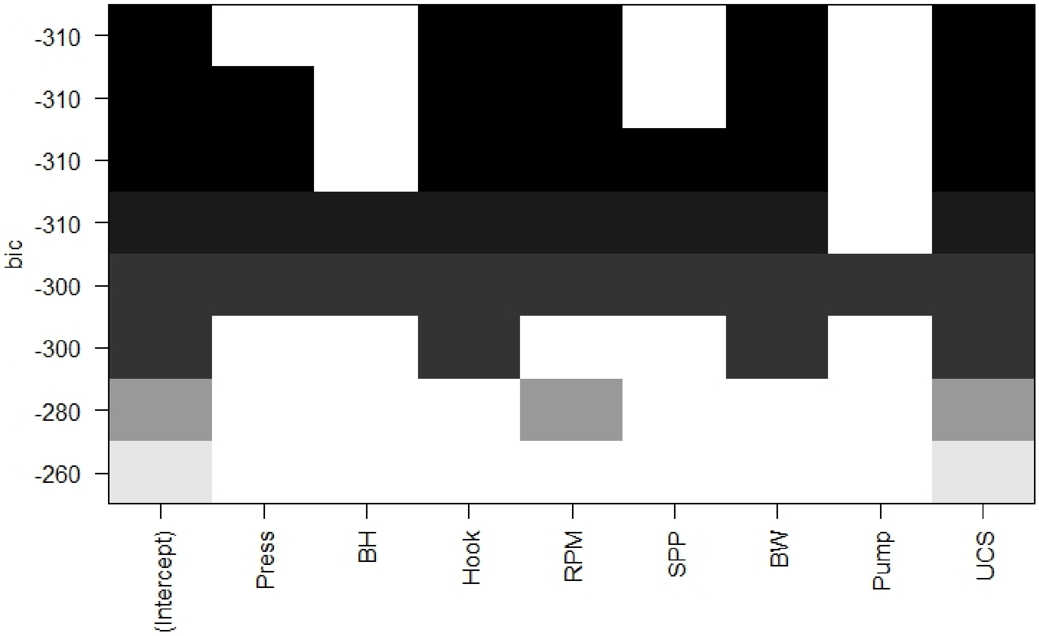


Figure 8: BIC vs feature selection for ROP prediction in Tyler Sandstone

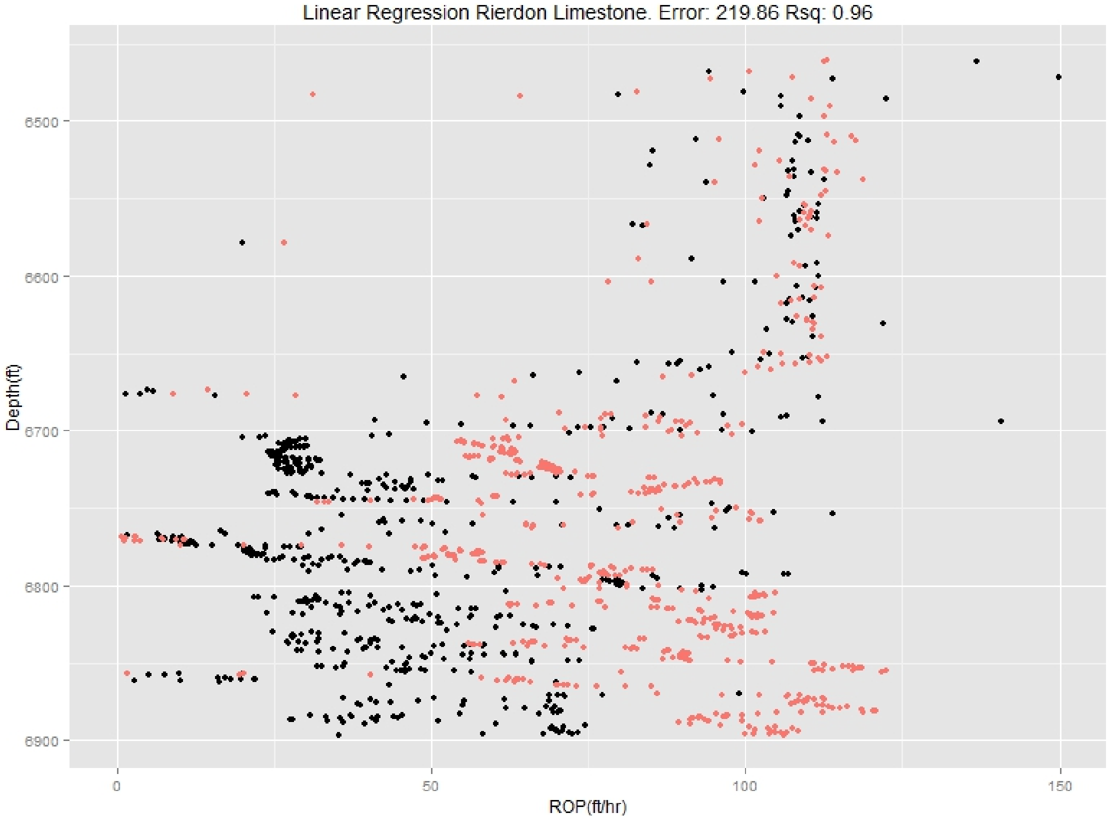


Figure 9: Prediction of ROP in Rierdon Limestone using OLS Regression

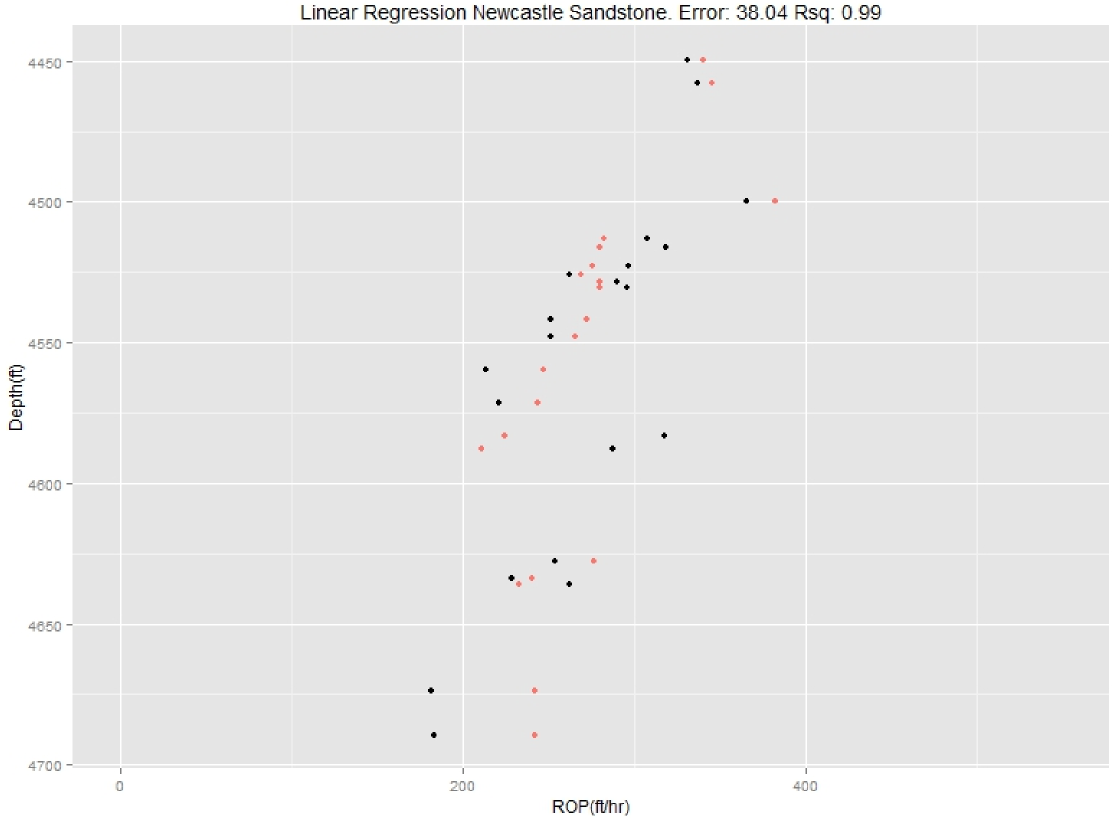


Figure 10: Prediction of ROP in Newcastle Sandstone using OLS Regression

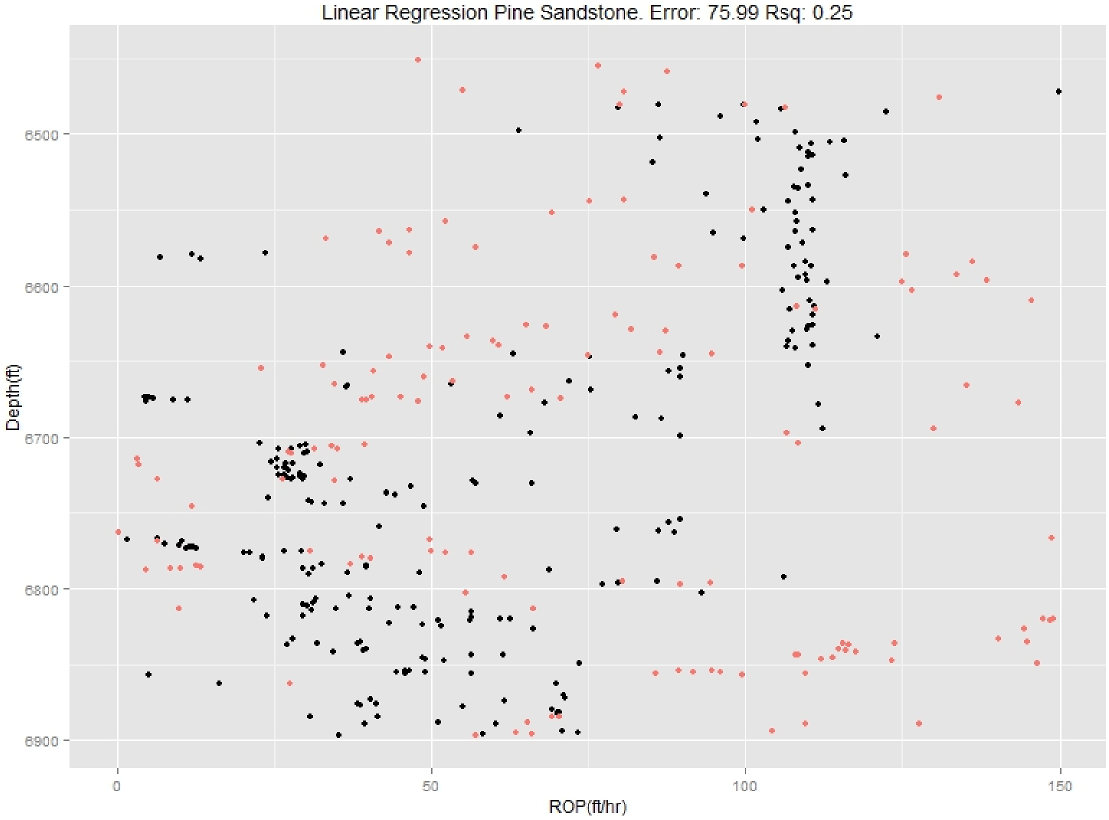


Figure 11: Prediction of ROP in Pine Sandstone using OLS Regression

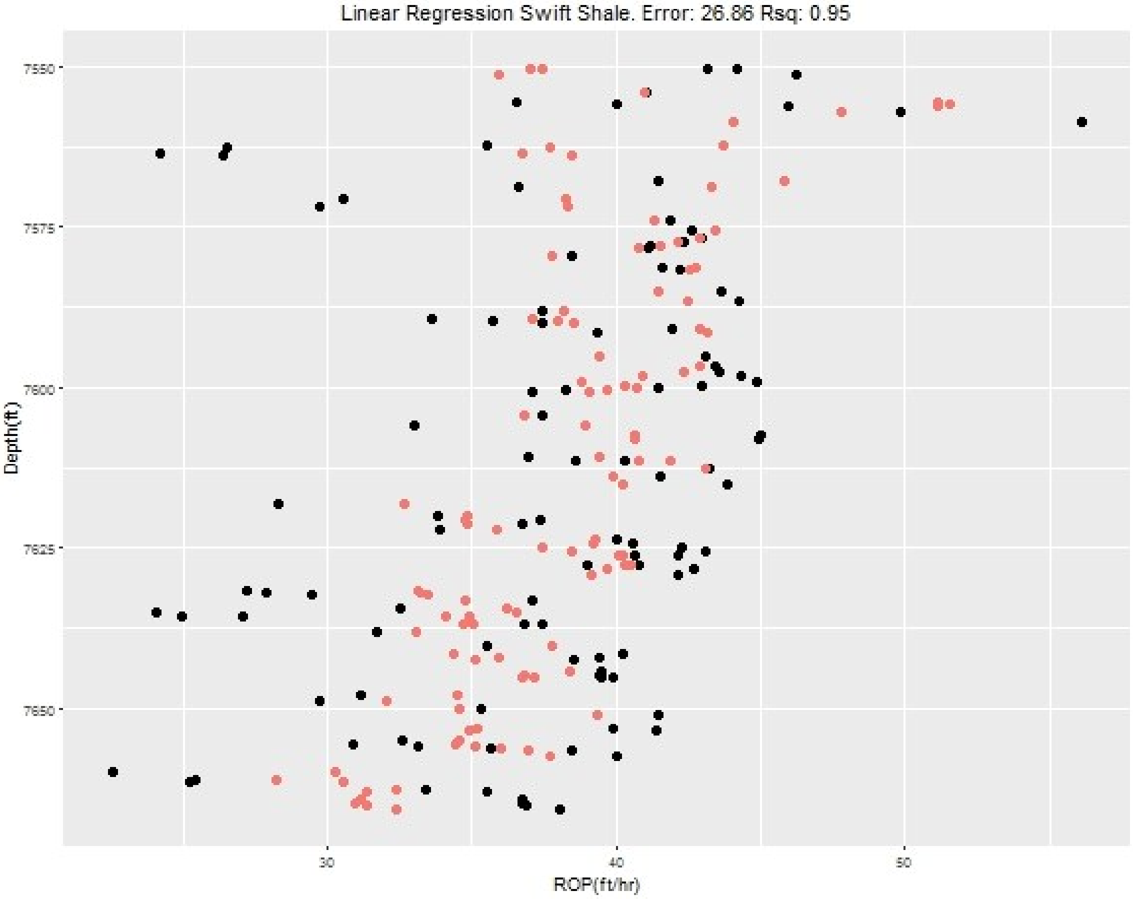


Figure 12: Prediction of ROP in Swift Shale using OLS Regression

Models tend to change with a change in formation, and will have to be adapted by changing some tuning parameters. However the model can be applied to any formation without changing the inputs, i.e. using the same input in all models. Wallace et al. (2015) have demonstrated application of similar models to horizontal wells while drilling in unconventional reservoirs.

In conclusion regression offers a simple model which can be used to predict ROP. The RMSE for the ROP models used in this section were high. Other nonlinear algorithms can be used instead to improve the accuracy of the ROP prediction as described later in this thesis.

4.2: REGULARIZED REGRESSION

Least square regression minimizes the sum of squares to find the coefficients of input features in regression. However, in certain situations this might not be the best course of action especially when the number of input features are large.

Although not a specific form of regression, this form deviates from least squares by imposing a penalty on the size/value of the coefficients in the model. This makes the regression model coefficients impervious to collinearity. While ridge regression enforces penalties on the values of the coefficients, the lasso forces the coefficients to zero using the l_2 norm (Hastie et al., 2007). Cross validation maybe used to determine the regularization parameters. The equations for ridge and lasso regression are described as Equation 3 and 4:

$$\text{ROP} = \sum_{n=1}^{N} f_n x_n + \lambda \sum_{n=1}^{N} f_n^2,$$

(Equation 3)

$$\text{ROP} = \sum_{n=1}^{N} f_n x_n + \lambda \sum_{n=1}^{N} |f_n|.$$

(Equation 4)

Equation 3 represents the ridge regression formulation, where penalties are induced on large coefficients. λ is the tuning parameter, which is determined using cross validation to reduce the error of the model. This is particularly effective when the number of features are large or in cases where features have a high degree of collinearity. Thus, when the value of λ is zero this model behaves like a least squares regression model. The main difference between ridge and lasso technique is the range of values of λ.

Equation 4 represents the Lasso regression equation, where the penalty (λ) can shrink coefficients of the regression model to zero. These methods are useful in cases where the number of predictors outweigh the samples and there is a high correlation between input features.

The penalty (λ) is varied the between10^{10} to 10^{-2}, encompassing all of the regression models (a model with just the intercept to the model containing all of the parameters); the model with the lowest squared error is chosen as the best model and ROP prediction results have been plotted. Figures 13,14,15,16 and 17 show the results of using ridge regression for ROP predictions in varying formations.

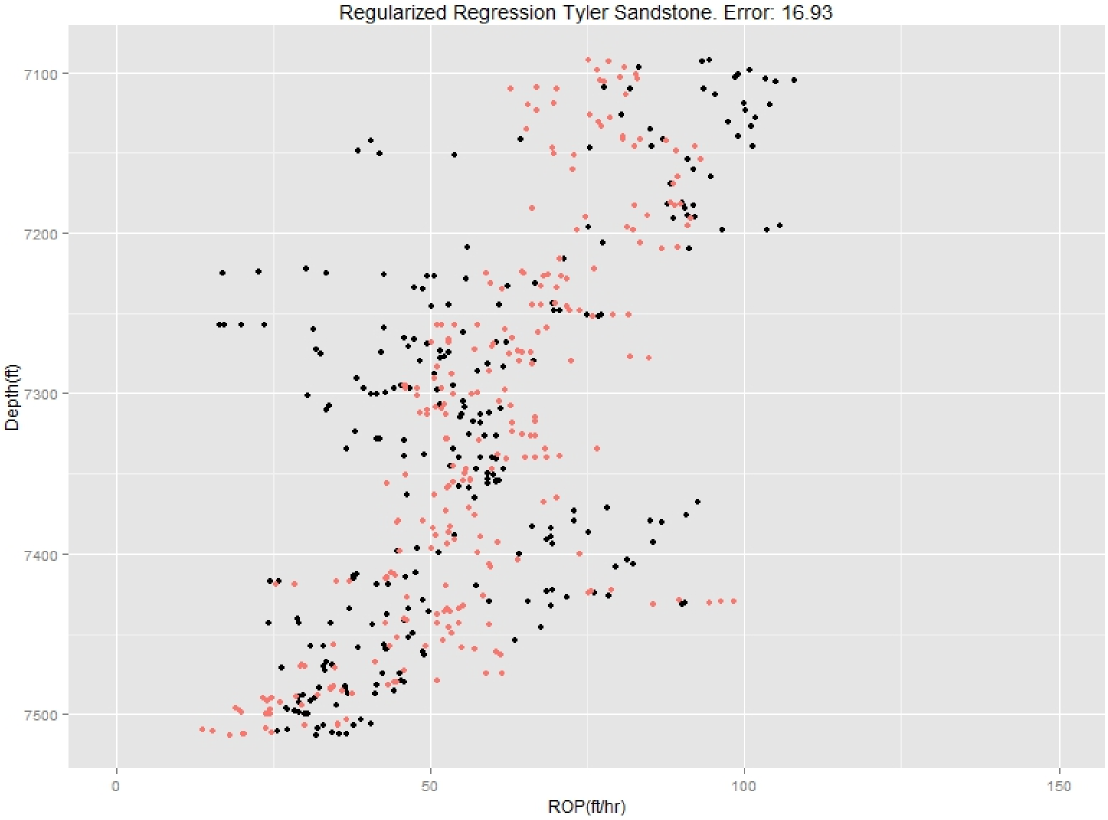


Figure 13: ROP prediction using ridge regression in Tyler Sandstone

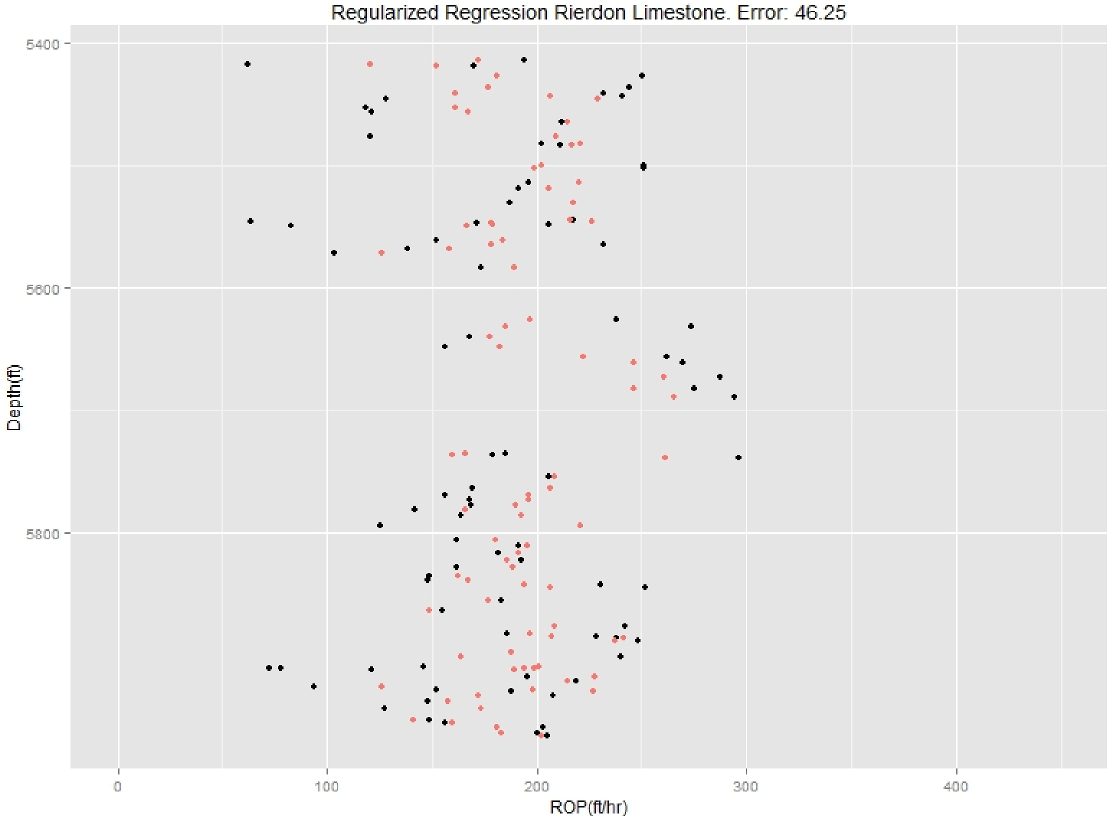


Figure 14: ROP prediction using ridge regression in Rierdon Limestone

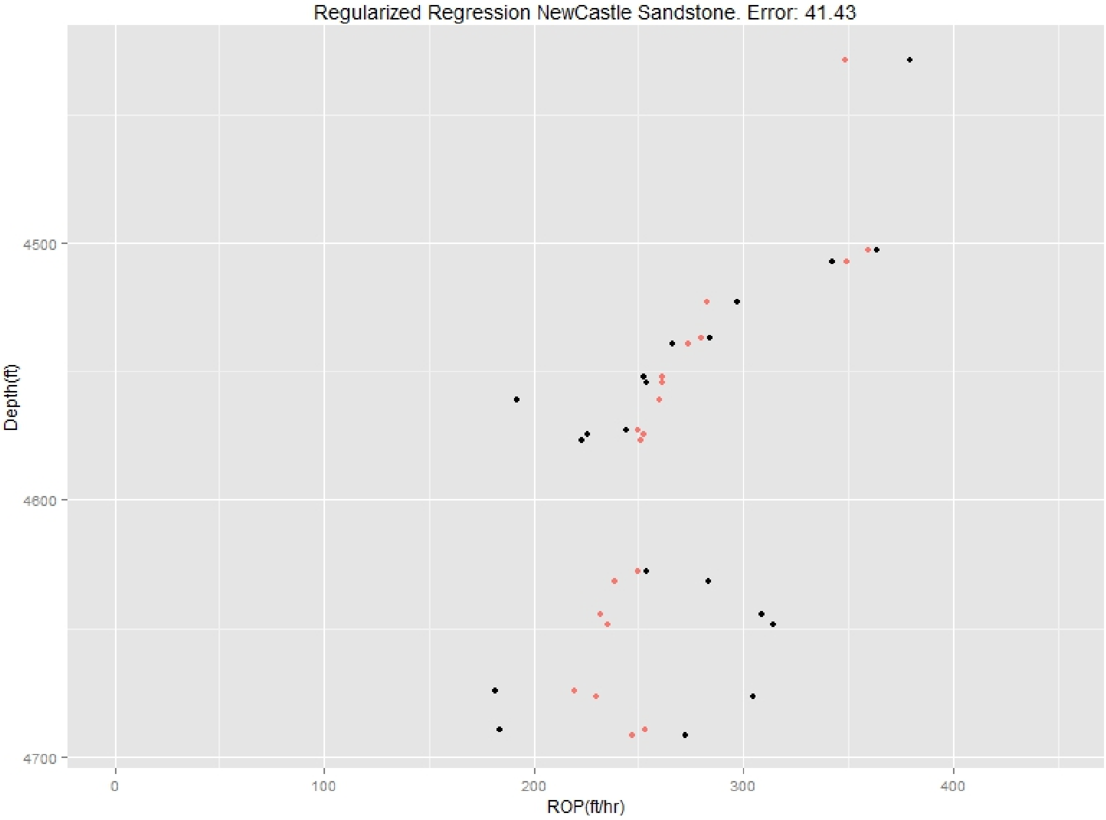


Figure 15: ROP prediction using ridge regression in Newcastle Sandstone

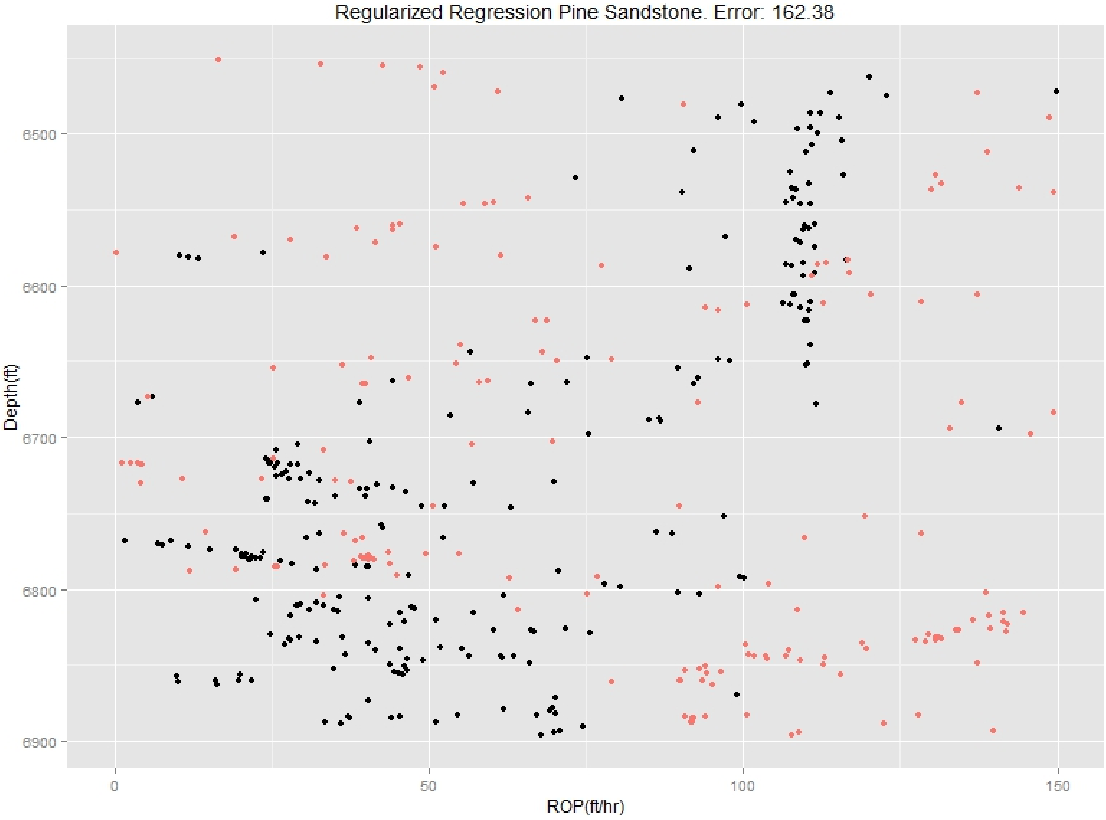


Figure 16: ROP prediction using ridge regression in Pine Sandstone

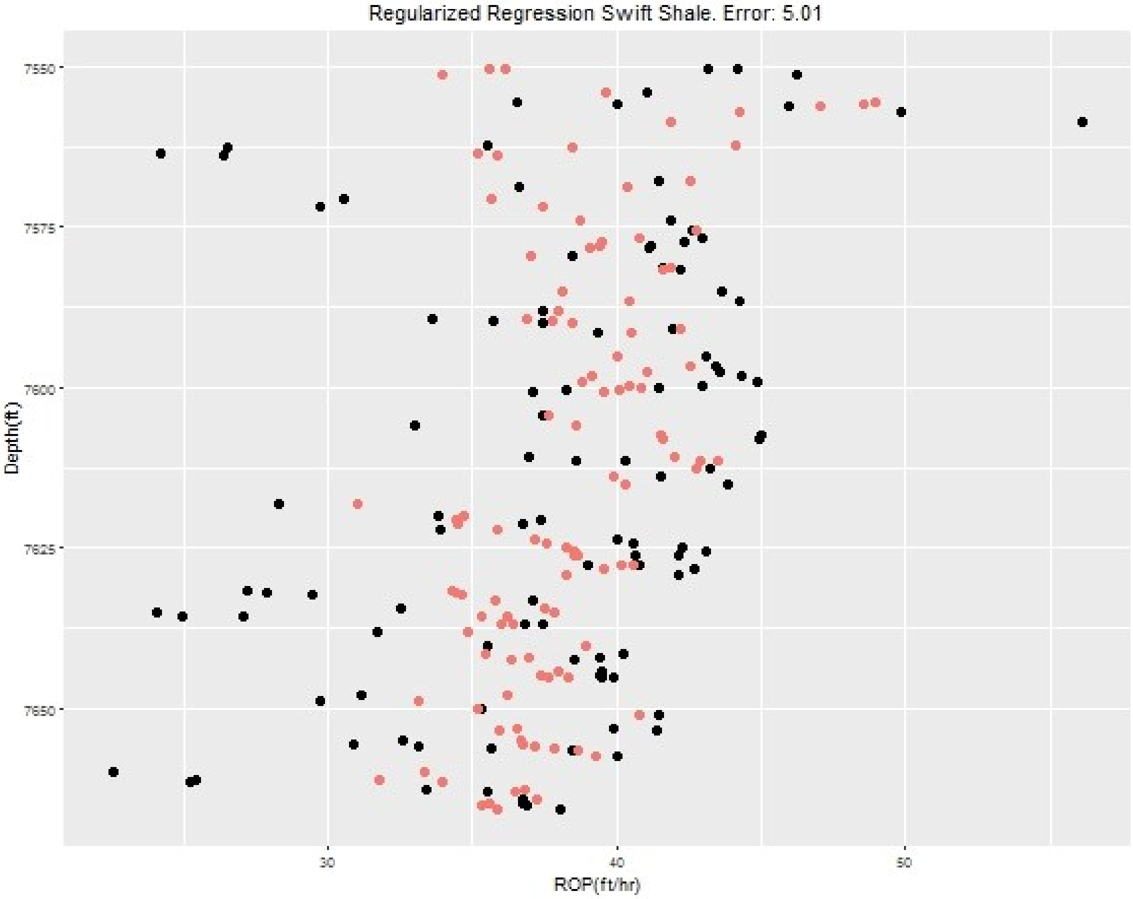


Figure 17: ROP prediction using ridge regression in Swift Shale

Ridge regression works well in reducing the prediction errors for certain formations: Rierdon Limestone and Swift Shale. The RMSE in some formations increases in compared to OLS regression. Figure 18 shows lasso regression for the Tyler sandstone.

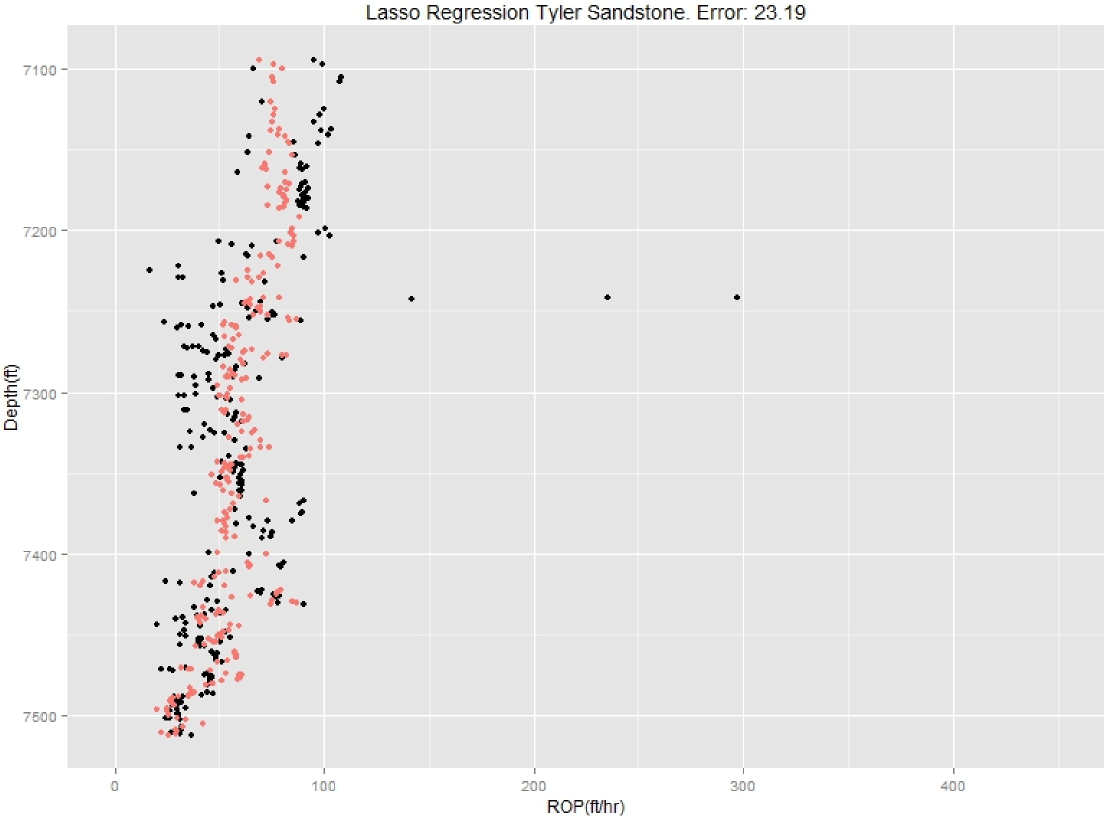


Figure 18: ROP prediction in Tyler Sandstone using Lasso

4.3: Principal Components Regression

The idea behind the principal component regression is to perform principal component analysis (PCA) on the data, then perform regression on the eigenvectors of principal components. PCA can be used to reduce the number of features while retaining as much variance explained by the data as possible. PCA transforms the axes of the data to a different scale, one which requires fewer predictors. PCA can be used to project high dimension data to a lower dimension. Lower dimensional data are more advantageous and computationally efficient. For example, three dimensional data can be visualized with a plot, while data with eight dimensions cannot.

PCA is used to determine components of the data set which contribute maximum variance, and should be used when there is a requirement to reduce the number of features or reduce dimensions. PCA can be used to reduce noise in data sets (not unknown in drilling data). Principal components regression (PCR) involves performing PCA on the data set to retain a certain number of features, subsequently using these retained features for regression. This process relies on the premise that PCA retains as much variance as possible in the data, curbs noise in the data and ensures that the original data can be represented by the eigenvectors of PCA. PCR and partial least square (PLS) have been explored in depth by Mevik and Wehrends (2007).

PCR is used to predict ROP. Cross validation is used to ensure that the optimum number of components maybe retained to ensure the lowest RMSE. It is seen that PCA beyond 3 components usually represents the variance of the data in case of this data set. Table 2 summarizes the variance explained for varying components for the dataset. Table 2 confirms that 3-4 components are adequate for the data used in this thesis.

Number of components	**1**	**2**	3	4	5	6	7	8	9	10
Tyler										
Percentage Variance Retained (%)	86	97	100	100	100	100	100	100	100	100
Pine										
Percentage Variance Retained (%)	59	86	99	100	100	100	100	100	100	100
Rierdon										
Percentage Variance Retained (%)	81	97	99	100	100	100	100	100	100	100
Swift										
Percentage Variance Retained (%)	89	96	99	100	100	100	100	100	100	100

Table 2: Percentage variance captured using PCA for different rock formations

ROP is predicted using PCR (3 components). Figures 19 - 23 show ROP predictions using PCR. The error in this cases is lower than that seen in OLS regression.

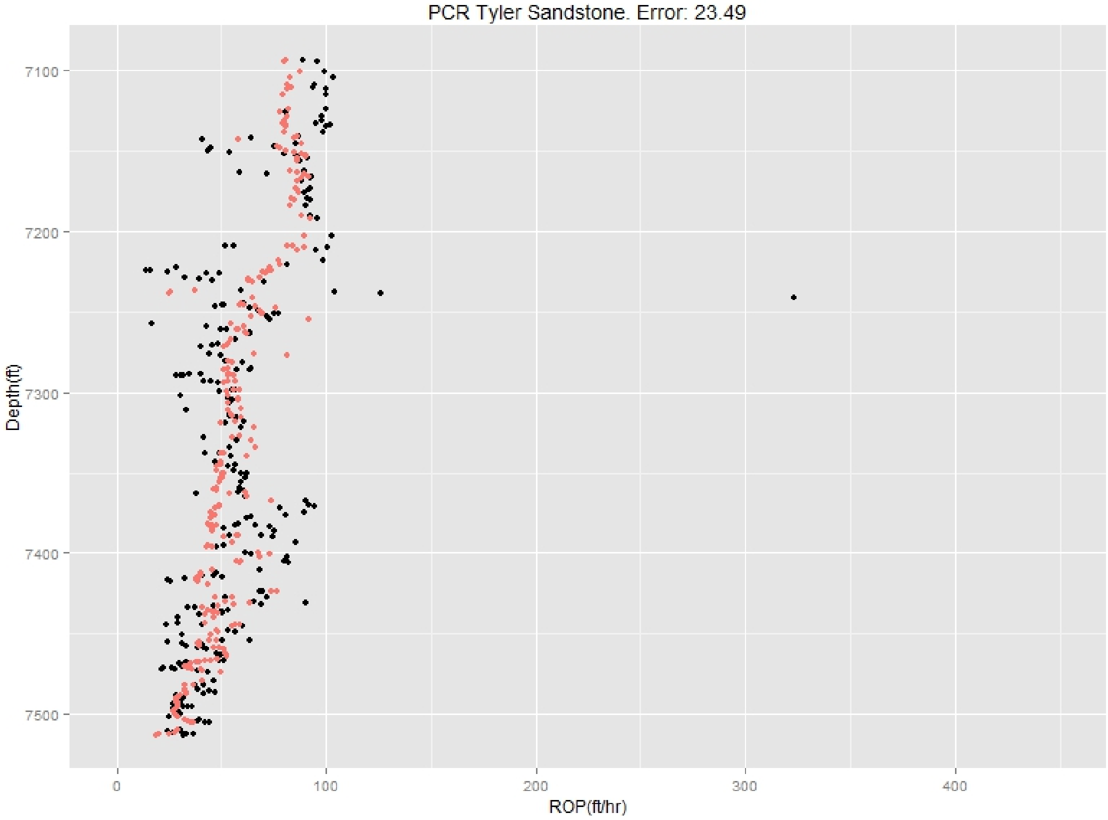


Figure 19: ROP prediction using PCR for Tyler Sandstone

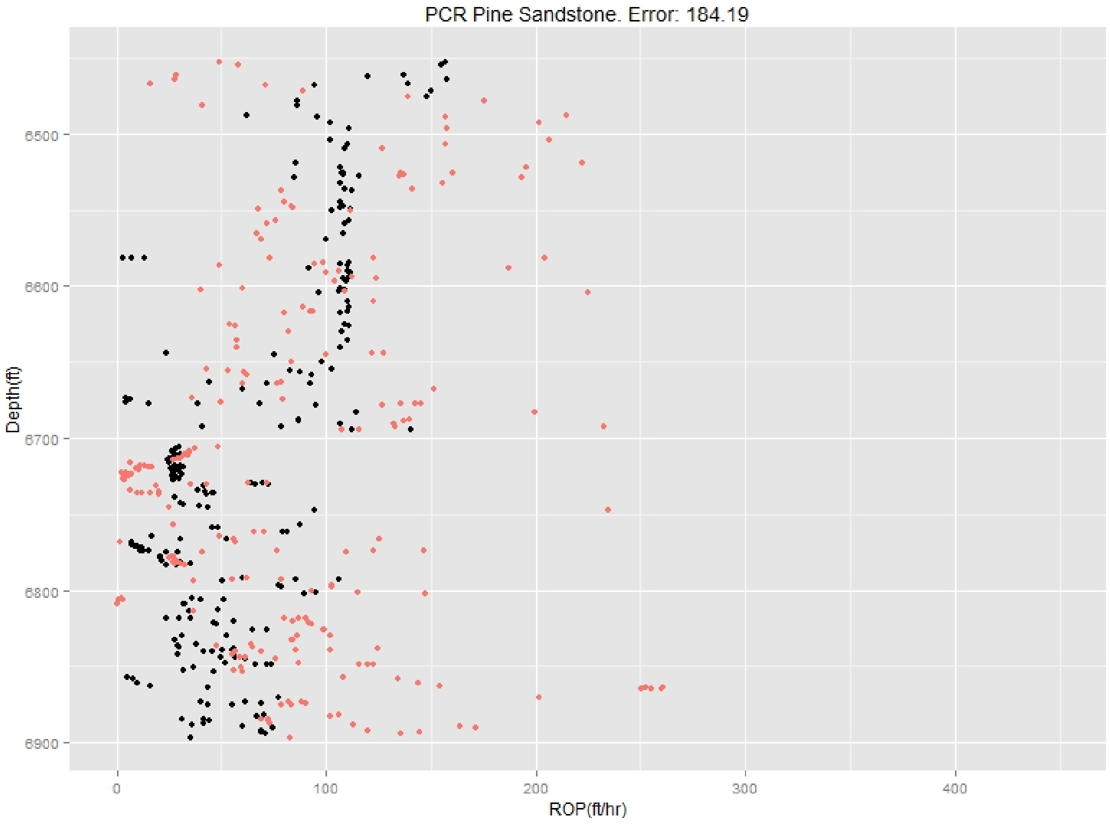


Figure 20: ROP prediction using PCR for Pine Sandstone

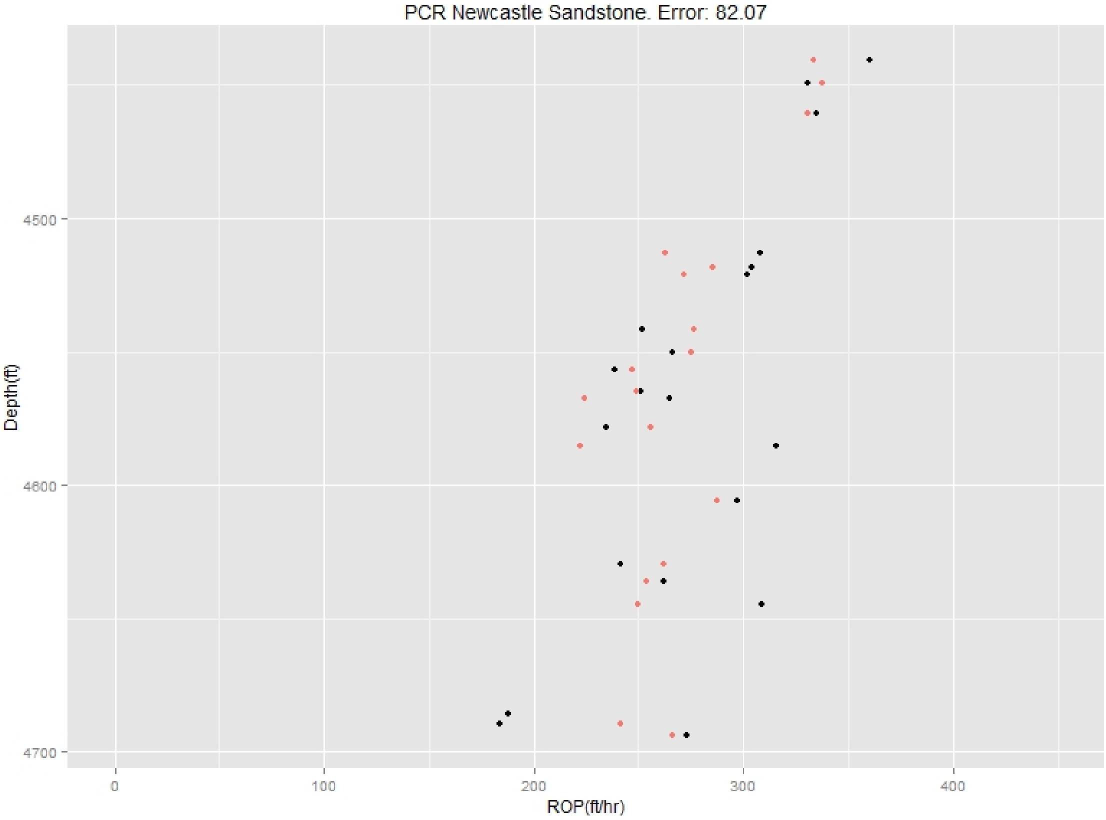


Figure 21: ROP prediction using PCR for Newcastle Sandstone

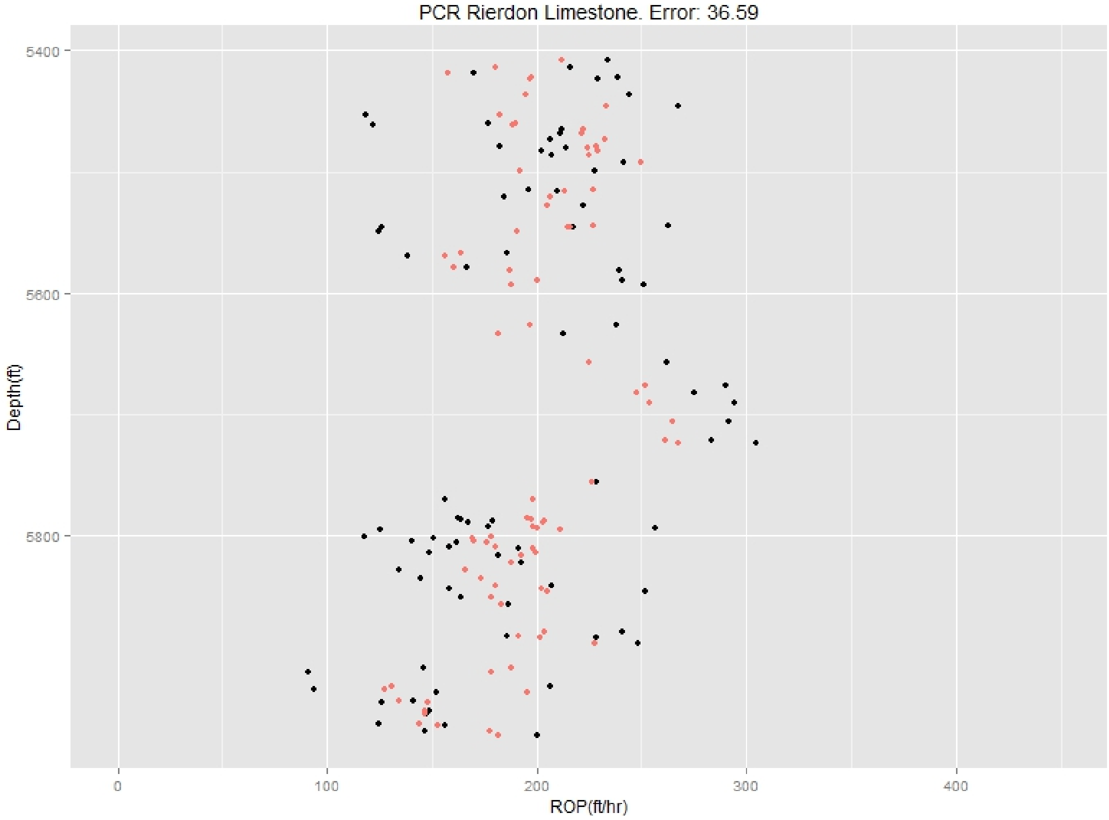


Figure 22: ROP prediction using PCR for Rierdon Limestone

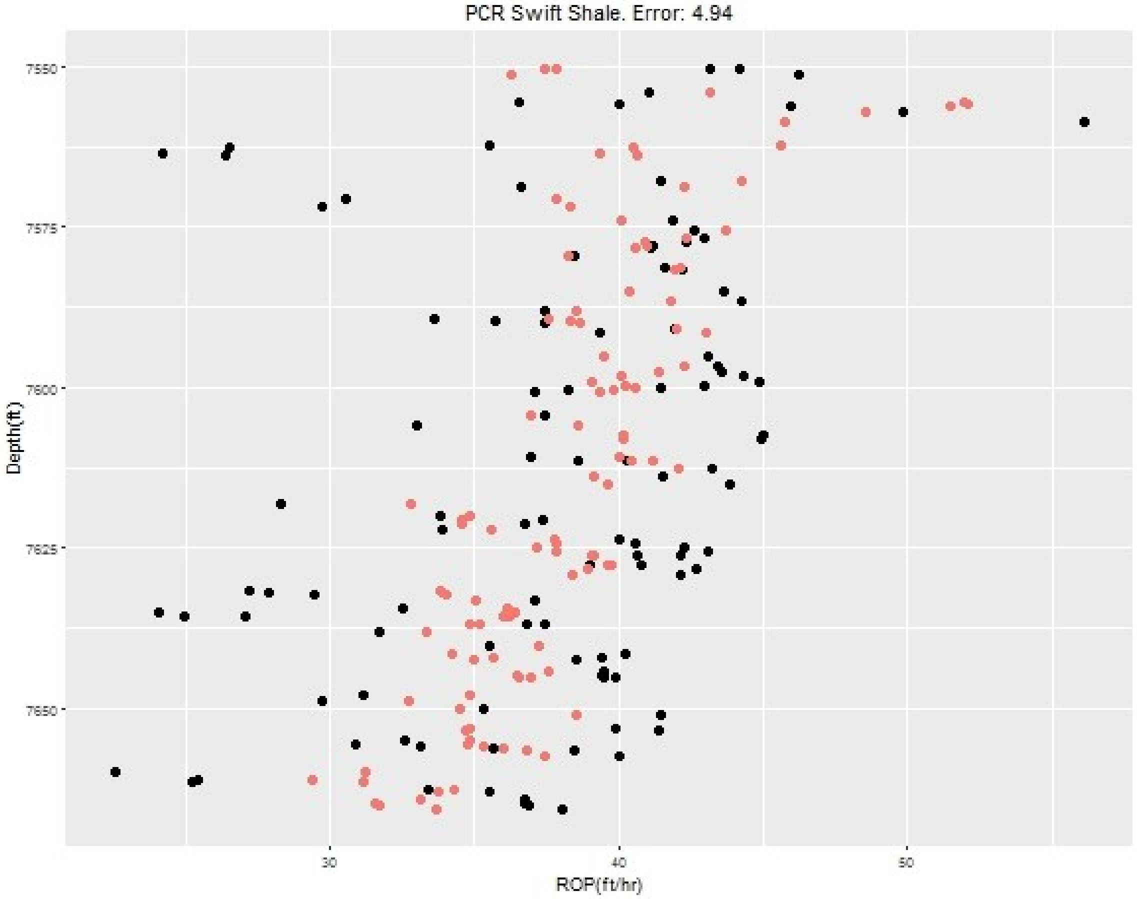


Figure 23: ROP prediction using PCR for Swift Shale

4.4: Regression Algorithm Selection

From Figures 9-18 it is apparent that no single method outperforms everything else. Figure 24 shows a bar plot and Table 3 provides a summary of the total RMSE for different regression techniques. This confirms that no single regression method outperforms all others. However, computationally PCR is better than OLS which is faster than regularized regression.

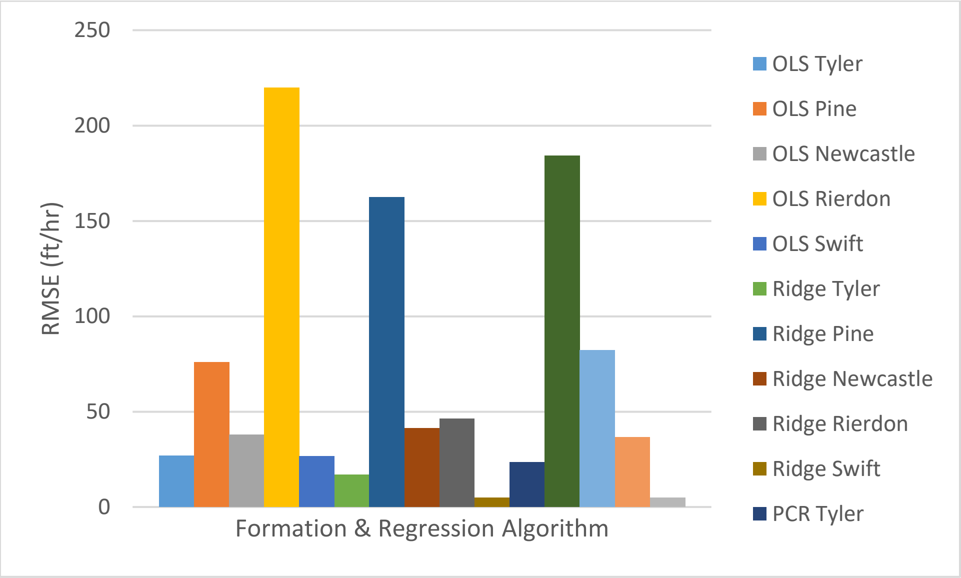


Figure 24: Error comparison for ROP prediction using different regression methods

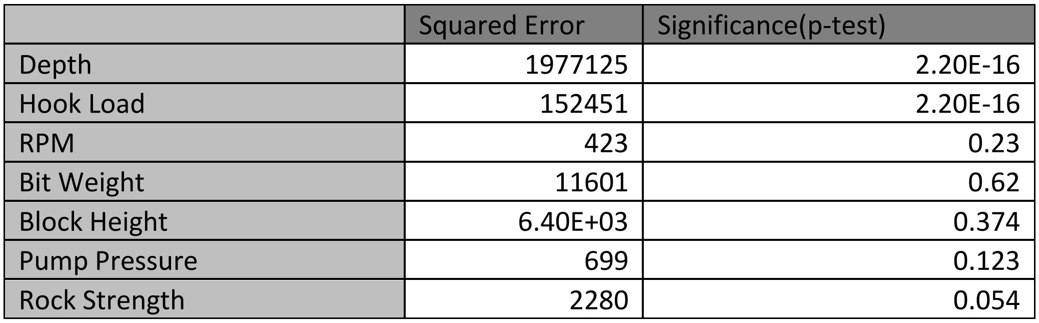
Formation	Regression Algorithm	Average RMSE (ft/hr)	Average Error Percentage (%)
Tyler	OLS	26.72	46.22
	Ridge regression	16.93	29.28
	PCR	23.49	40.63
Pine	OLS	75.99	105.79
	Ridge regression	162.38	226.06
	PCR	184.19	256.42
Newcastle	OLS	38.04	13.85
	Ridge regression	41.43	15.09
	PCR	82.07	29.89
Rierdon	OLS	219.86	116.04
	Ridge regression	46.25	24.41
	PCR	36.59	19.31
Swift	OLS	26.68	28.99
	Ridge regression	5.02	5.45
	PCR	4.94	5.37

Table 3: Summary of ROP prediction using different regression techniques

[2]Chapter 5: Nonlinear Prediction Methods

Only linear methods of prediction have been discussed so far. But in reality data are seldom linear, which requires the use of non-linear prediction algorithms. Determination of non-linearity of data is possible by comparing slopes of fit of linear and nonlinear regressor as suggested by Cheng et al. (2006). An ANOVA test carried out on the dataset has been summarized in Table 4 below. This is used to determine the linearity of the data (Cheng et al., 2006). If the significance column in table 4 is higher than 0.05, this would indicate a deviation from linearity. The data in Table 4 concludes that parameters such as RPM, bit weight and pump pressure - which are important input variables from an engineering standpoint - are not linearly related to the ROP. This chapter will introduce nonlinear methods such as trees. Random forests will be derived as a modification of trees, yielding an excellent algorithm for prediction. Pros and cons of each method will be examined as ROP in different formations are predicted and compared to the predictions in the previous chapters.

[2] Hegde,C.M., Wallace S.P. and Gray, K.E. (2015b). Use of regression and bootstrapping in drilling: inference and prediction. Presented at SPE Middle East Intelligent Oil & Gas Conference & Exhibition, Abu Dhabi, United Arab Emirates, 15-16 September. SPE-176791.
The author of this thesis was the primary author of the paper



	Squared Error	Significance(p-test)
Depth	1977125	2.20E-16
Hook Load	152451	2.20E-16
RPM	423	0.23
Bit Weight	11601	0.62
Block Height	6.40E+03	0.374
Pump Pressure	699	0.123
Rock Strength	2280	0.054

Table 4: ANOVA test on input data

5.1: TREES

Tree methods can be used either for classification or regression (prediction of response variable). In this chapter trees have been employed for regression or prediction of ROP. A tree includes a flowchart like structure, in which an input variable or feature is evaluated at each node as shown in Figure 25. To simplify such complex data relationships, the approach taken by trees is to partition the data into smaller (more manageable) sections, as illustrated in Figure 26. The sub divisions can be partitioned again, which constitutes recursive partitioning, until the sub divisions can be fit with simple linear models. Trees are fast (useful for real-time predictions) and easy to understand (Figure 25). A sample decision tree was built using the model input features and is shown in Figure 25. The decision tree is built by determining the best input features (in terms of entropy of the model (James et. al, 2014)). The topmost node (Flowrate <374) is the criterion being evaluated. A positive evaluation leads to the left branches and a negative evaluation leads to the right branches. Consecutive evaluations of input parameters lead to a prediction from the tree. For example, on the left branch if flowrate is less than 374 gpm and RPM is less than 63, the tree will return a ROP prediction of 513 ft/hr. A random forest is built using multiple trees by bootstrapping (James et al. (2013) provide a good summary) the training set. At each tree node, the number of feature vectors available is randomized: by selecting a subset of the total number of features available for prediction. This helps increase the accuracy of the algorithm by de-correlating the feature vectors (prediction is based on all feature vectors as opposed to the ones with the highest correlation to the data). More details on the specifics of the random forest algorithm can be found in the paper written by Breiman (2001), and a simplified easy to read explanation is found in the book published by James et al. (2013).

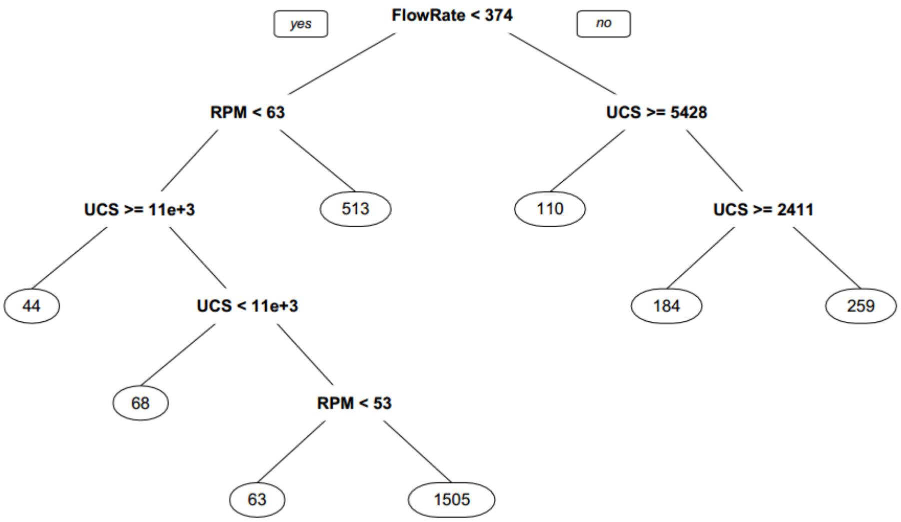


Figure 25: Simple tree diagram for Tyler sandstone formation using surface measurements as input parameters (Hegde et al., 2017)

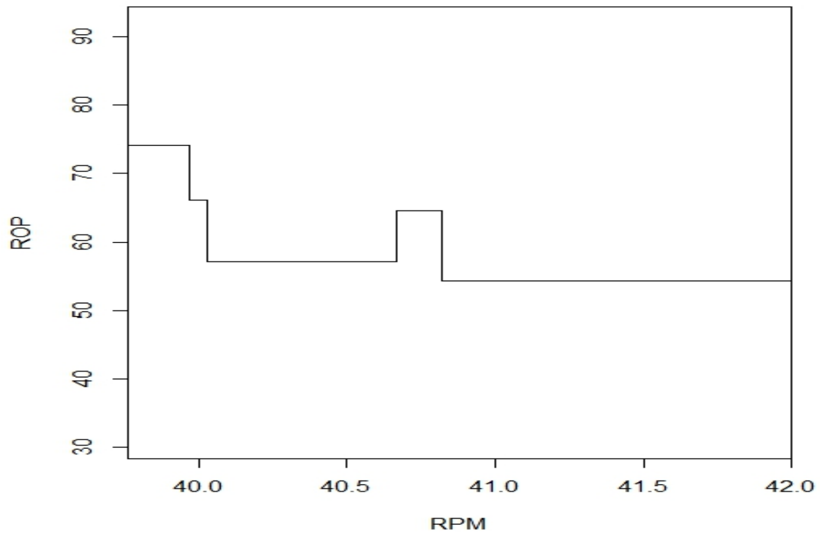


Figure 26: Data partition for a simple tree involving prediction of ROP using RPM as a predictor (Hegde et al., 2015)

Trees were used to predict ROP in specific formations (Tyler, Rierdon, Pine and Swift). Figure 27 shows a sample tree built on the Tyler formation data. ROP for Tyler formation has been visualized and compared to measured data in Figure 28.

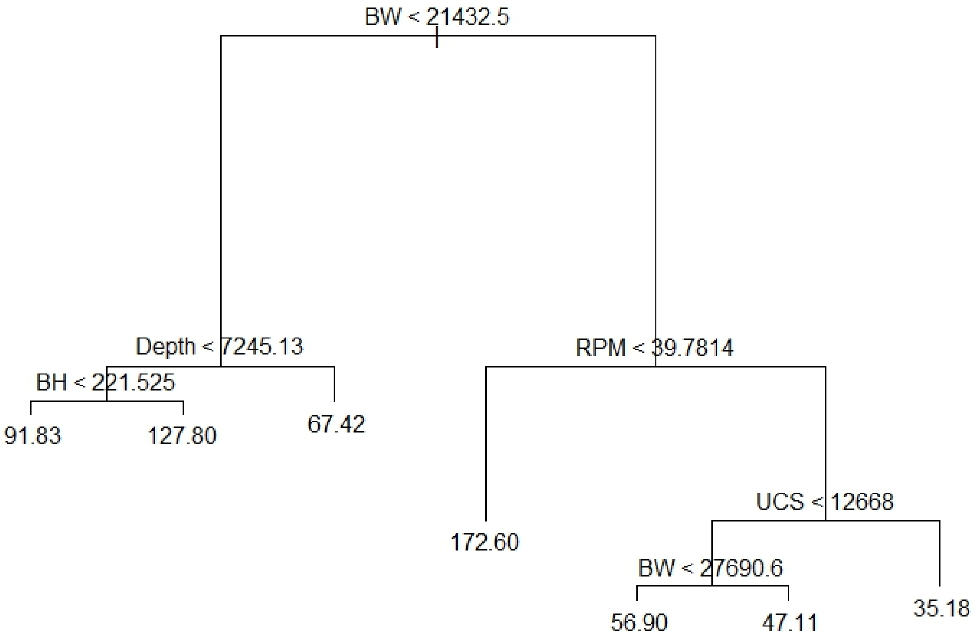


Figure 27: Tree on Tyler Sandstone Data for ROP Prediction

Figure 27 shows the tree built on Tyler sandstone (in the same manner as before). Splits between two variables are seen at each node. The bottom of the tree beyond which there aren't any more splits is called a 'leaf'. The vertical growth can (and should) be controlled - termed as pruning a tree. The ROP is predicted using the tree shown in Figure 27 where input parameters are evaluated to make a prediction (as explained previously with respect to Figure 25). These predictions have been shown in Figure 28. The RMSE and R^2 of these predictions are low.

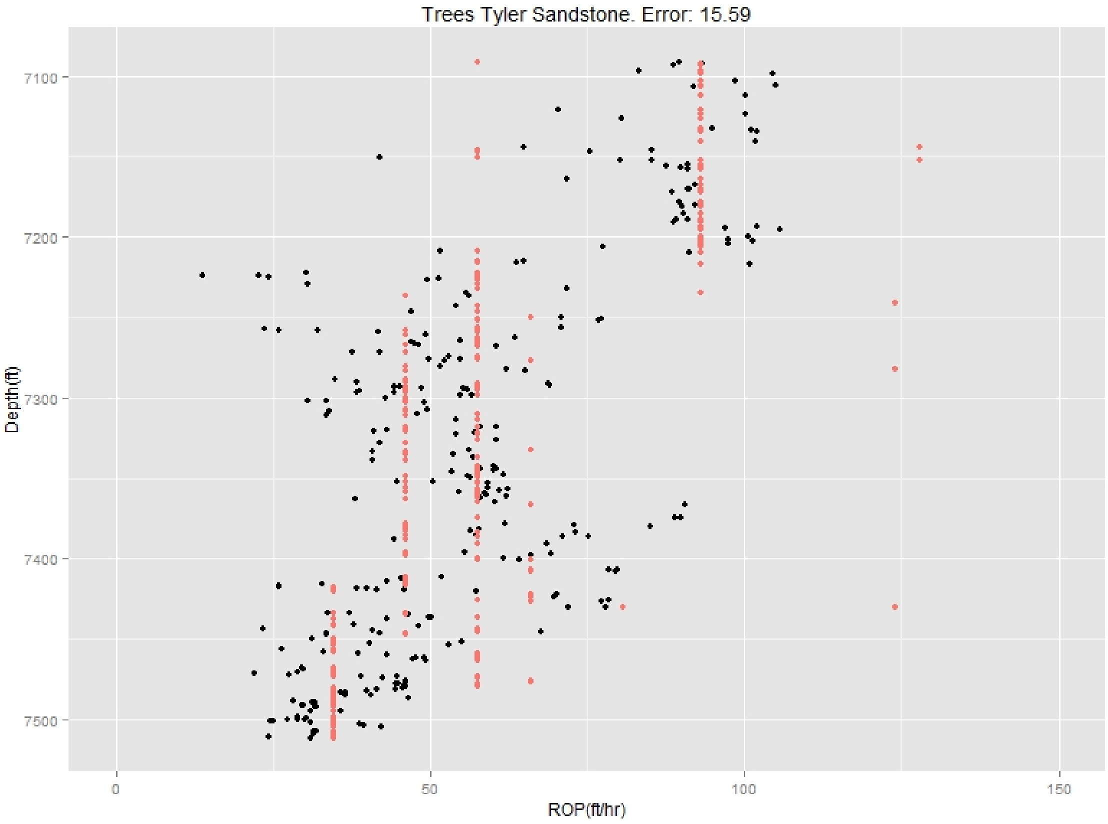


Figure 28: ROP prediction using trees in Tyler Sandstone

Shortcomings of trees include their accuracy and variance of predictions, which can be improved by pruning a tree. Pruning involves controlling the vertical depth of a tree which can help decrease the error due to prediction. Pruning lengths are determined using cross validation. Figure 29 plots the deviance (sum of squared error) against the number of features for each tree. From the figure it is easy to conclude that the best result is achieved with a 6 split tree.

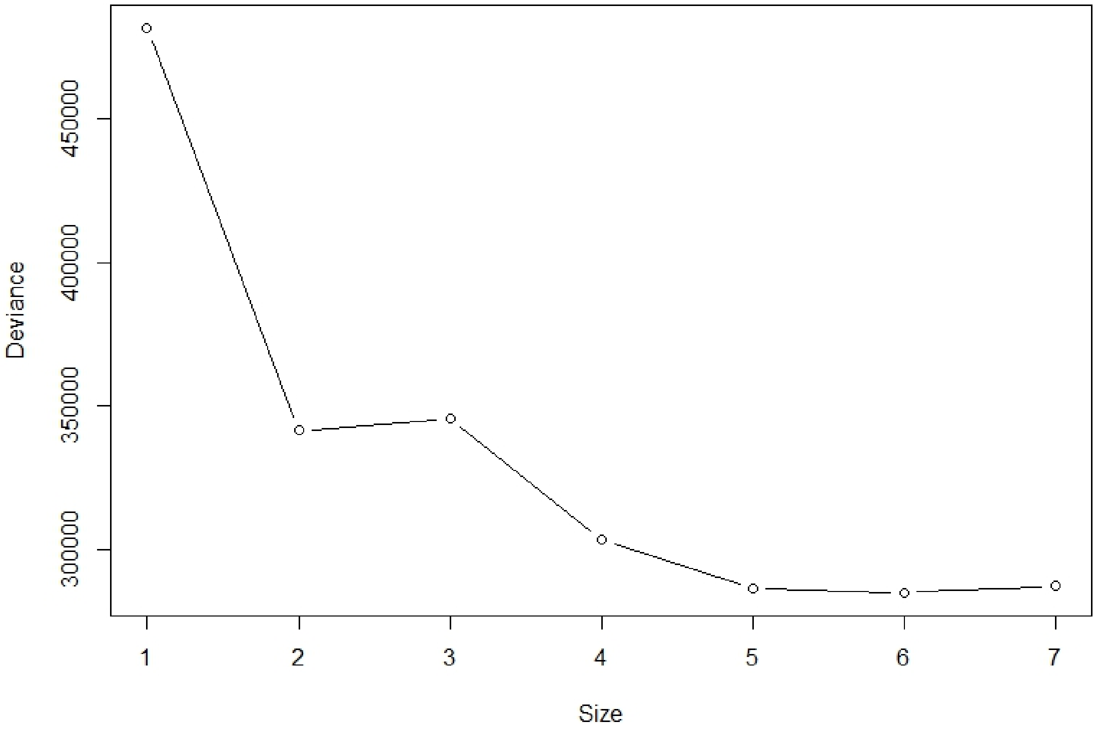


Figure 29: Deviance versus Size of Tree for ROP using Trees in Tyler Sandstone

5.2: TREES VERSUS LINEAR MODELS

The predictor equation of linear regression has been compared to that of trees (Equation 2 vs Equation 5). They are fundamentally different:

$$\text{ROP} = \sum_{n=1}^{N} f_n x_n,$$

(Equation 2)

$$\text{ROP} = \sum_{m=1}^{M} g_m . 1_{(X \in R_m)},$$

(Equation 5)

where R_m is a partition of feature space (as shown in Figure26) and g_m are constants determined by reducing the sum of squared error. The better model depends on the situation at hand. If ROP can be approximated well with a linear model, linear regression will outperform trees. However, if the data is non-linear and complex, trees may do a better job of predicting ROP.

5.3: BAGGING

The bootstrap is very powerful technique which can be used to improve the prediction capabilities of trees. Trees suffer from high variance in prediction. Reducing this variance would make a tree based algorithm a powerful predictor. If a dataset is split into two parts, and decision trees are grown on either half, they both would yield vastly different trees (high variance in prediction). A combined predictor would be the average of both these trees. In contrast, a procedure with low variance will yield similar results if applied repeatedly to partitioned data sets. The bootstrap can be used to sample 'B' number of data sets from the sole data set (discussed in Chapter 3). A tree can be grown on each new dataset, and averaged to reduce the variance, overcoming the main disadvantage of trees.

Given a set of *n* independent observations $P_1,...,P_n$, each with variance σ^2, the variance of the mean P of the observations is given by σ^2/n. Hence, averaging a set of observations reduces variance. If *n* number of trees could be created and averaged, this would greatly reduce the variance of trees. One way to reduce the variance (and increase the prediction accuracy of trees) is to take average a number of trees. Since these trees have to pertain to the same population, bootstrapping can be used to sample multiple data sets from the population (or original dataset).

In this approach, we generate '*B*' different bootstrapped training data sets. These are unique training sets generated from the original training set. Trees are then trained on each training set and finally averaged to get a final model. This is called bagging. While bagging can improve predictions for many different statistical learning models it is very useful for decision trees. To apply bagging to regression trees, *B* regression trees are constructed using *B* bootstrapped training sets which are then averaged. These trees have to be grown deep (not pruned) so that they have high variance and low bias. Averaging these *B* trees reduces the variance. Bagging has been demonstrated to give high improvements in accuracy by combining together hundreds or even thousands of trees into a single procedure (James et al., 2014).

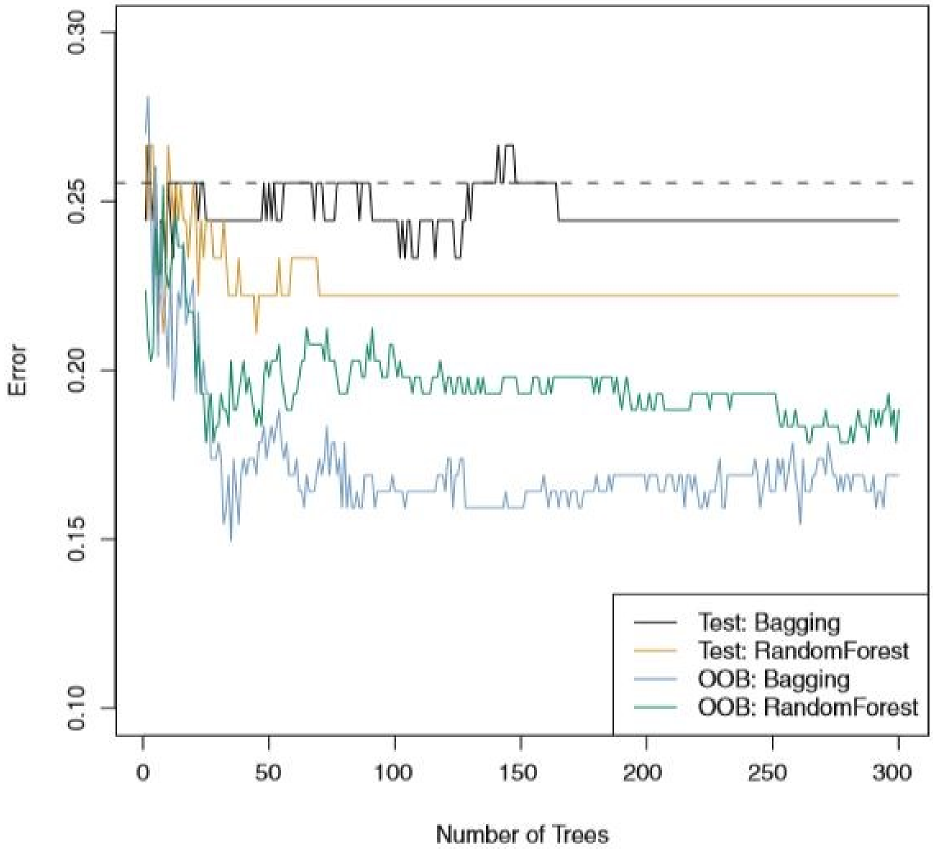


Figure 30: Number of trees versus error for different methods (James et al., 2014)

An easy way to estimate the test error of a bagged model, without the need to perform cross-validation is by estimating the out-of-bag (OOB). It has been shown that each bagged tree makes use of around two-thirds of the observations (Breiman, 1996). The remaining third of the observations can be used to evaluate the model's error. This is called the OOB error which is a valid estimate of the test error in the bagged tree model.

5.4: Random Forests

Random forest is an extension of bagging bearing additional advantages making it a powerful prediction algorithm. At each node a random sample of input features is considered to construct the decision tree (as opposed to all features). This has an effect of de-correlating the trees, which helps reduce variance and improve prediction accuracy. By using reduced number of predictors (each tree is forced to use a small number predictors), which forces all features (even those with a low correlation to ROP) to contribute to the prediction of ROP. Though counter intuitive input parameters with low correlation must contribute to prediction. Previously examined methods such as regression, trees, and bagging take “more” contributions from the input parameters that are correlated better with ROP, thereby masking the parameters with low correlation. This has actually been shown to affect the accuracy of the prediction (Figure 30) where random forest performs better than bagging for the same dataset. Since this algorithm stresses on importance of all input parameters (those with low and high correlation with ROP), it makes feature selection (selection of input parameters) even more important and integral to the success of the algorithm. De-correlating the input features helps in creating more randomized trees on the bootstrapped sample, which when averaged produce a better predictor for ROP.

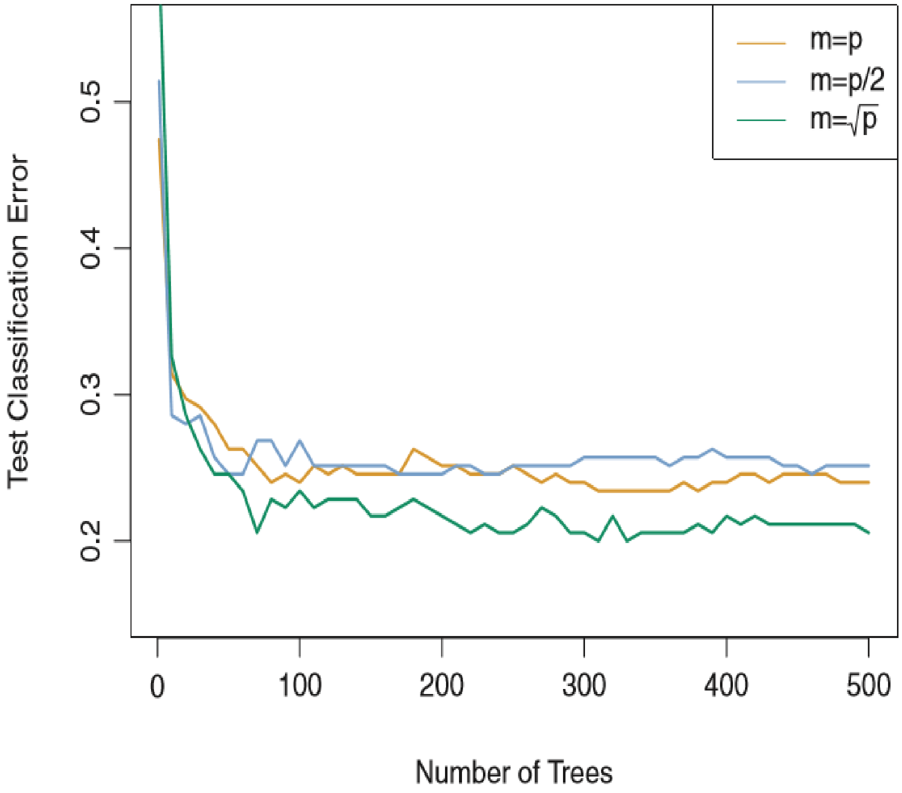


Figure 31: Selection of *m* in random forests so as to achieve least error (James et al., 2014)

Figure 31 shows selection of the parameter *m* in random forests, denoting the number of input features considered (at each split) for growing a tree. Test classification error is the error in classification as defined by a confusion matrix, explained well for the classification of torque by Hegde et al. (2015c). A rule of thumb (in the machine learning circles) is to use an *m* equal to the square root of the number of input features *p* (Figure 31). Cross validation can be used to select an *m* based on lowest error.

5.5: Evaluation of Non Linear Methods

Trees can be improved using bootstrapping (bagging or random forests). Another algorithm based on trees is boosting. In this case trees are sequentially grown based on the result from the previous step. A tree is grown, upon whose residuals another tree maybe grown to improve predictions. This process is repeated until there is no more benefit (no further improvement of RMSE). This algorithm is more complicated than bagging or random forests. It has to be validated carefully to avoid over-fitting. Random forests is simpler and more robust. James et al. (2014) provide a great introduction to bagging. Figure 32 shows a comparison of prediction error for boosting and random forests, where boosting helps achieve a lower error than random forest predictions.

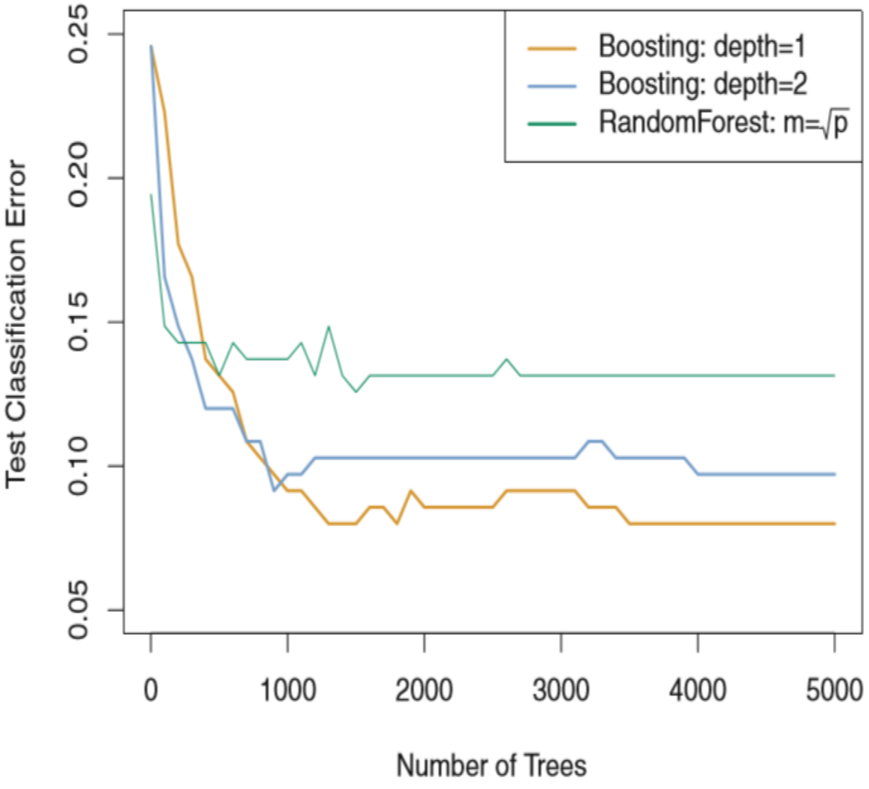


Figure 32: Comparison of boosting versus random forests (James et al., 2014)

ROP predictions of bagging and random forest algorithm in the Tyler sandstone formation have been compared in Figure 33. Both algorithms performed much better than trees as seen in Figure 34.

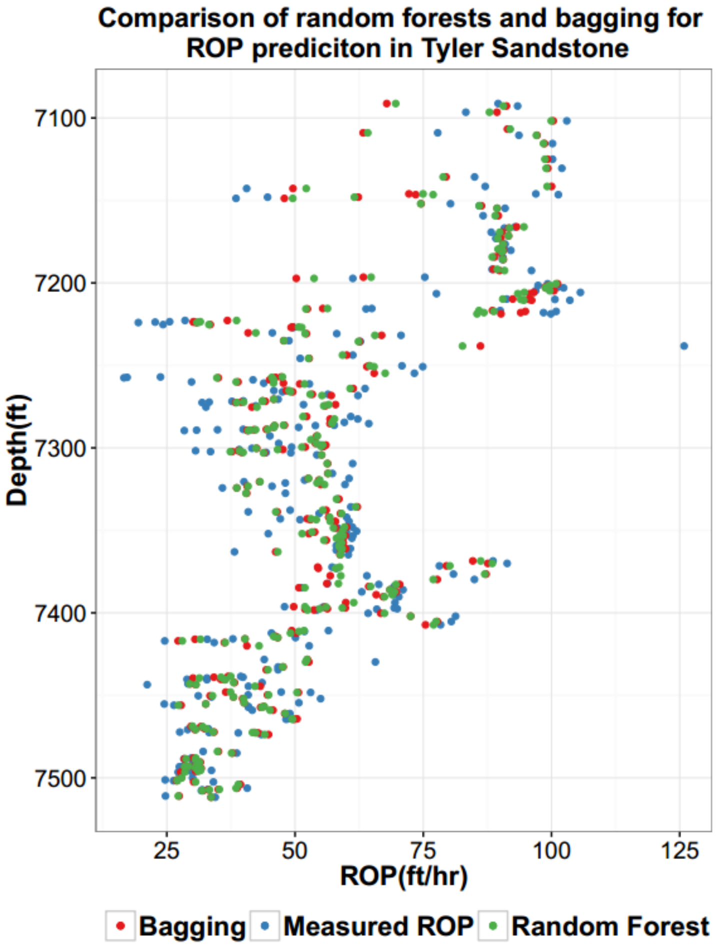


Figure 33: Comparison of random forests to bagged trees for ROP prediction in Tyler sandstone

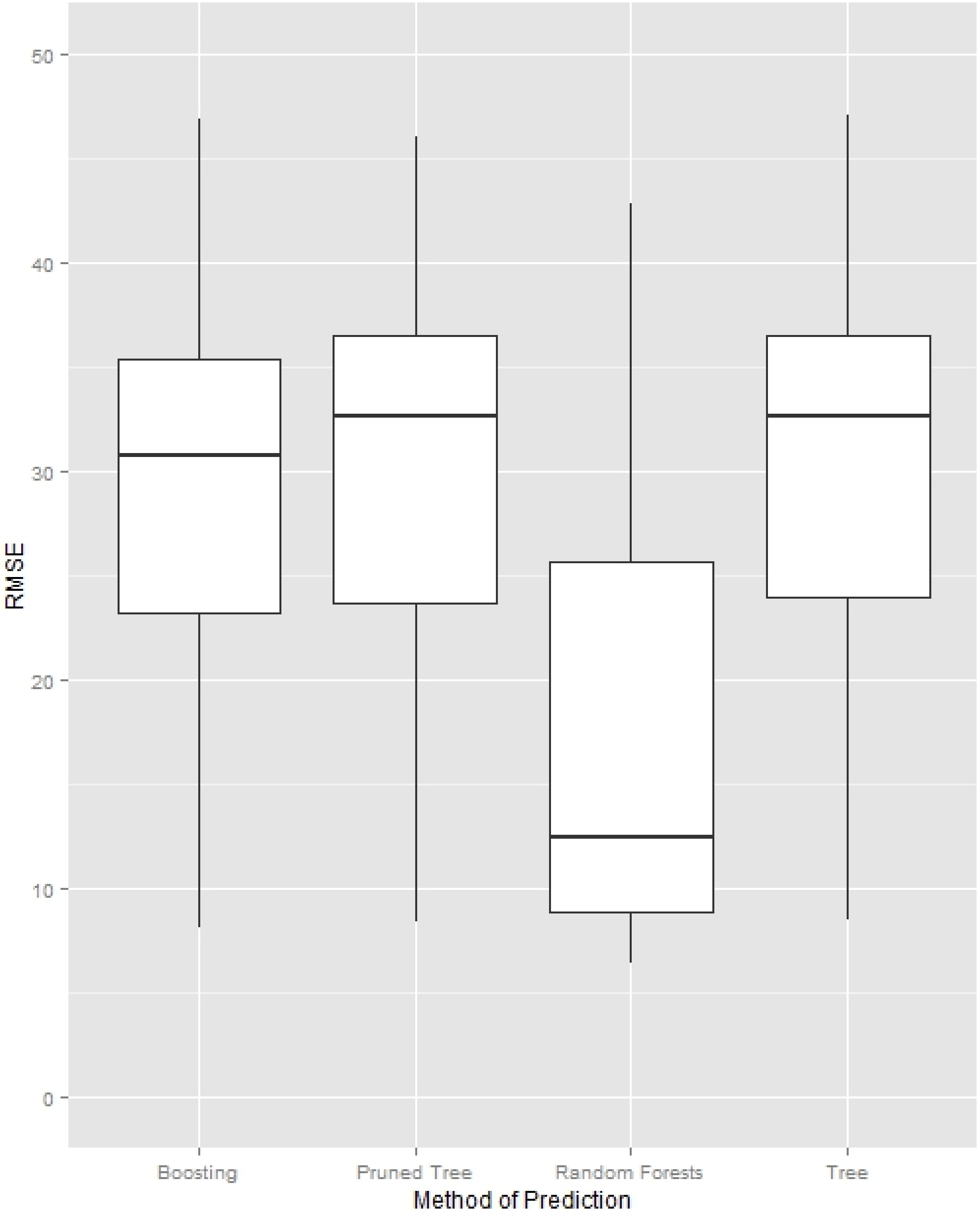


Figure 34: Box plot of RMSE using different methods for ROP prediction in sandstone formations (Hegde et al., 2015)

Figure 34 shows a box plot comparing the RMSE for ROP prediction in all formations using boosting, trees and random forests. Random forests performs better than the other algorithms in predicting ROP.

[3]Chapter 6: Parametric Analysis of ROP Models

The accuracy of statistical learning models predominantly relies on the quality, range, and volume of the training data. This section is dedicated to parametric analysis related to the training data: its range and volume for efficient ROP prediction. We evaluate the change in accuracy of ROP prediction based on changing the type and size of the training set relative to the test set.

6.1: Type of Training Data

Since the accuracy of statistical learning models is largely dependent on the training data, this section aims to evaluate three different types of training data illustrated in Figure 6. The first kind of training data (case 1) is the data obtained while drilling the formation in question– formation specific training data. The second kind of training data is data obtained from preceding formations and data from the current formation. The third kind of training set (case 3) is the data obtained while drilling preceding formations (or upper levels), which are used to predict ROP in a different formation: for example, using Broom Creek drilling data to predict ROP in Tyler formation (as shown in Figure 35 as case 3). Case 3 is a situation which is encountered when the bit enters a new formation, and no

[3] Hegde,C.M., Wallace S.P. and Gray, K.E. (2015b). Use of regression and bootstrapping in drilling: inference and prediction. Presented at SPE Middle East Intelligent Oil & Gas Conference & Exhibition, Abu Dhabi, United Arab Emirates, 15-16 September. SPE-176791.
The author of this thesis was the primary author of the paper

prior data for that formation is available. Case 2 is a combination of case 1 and case 3. One training set is better than the other if more accurate ROP predictions are made when a model is built on it: training sets are evaluated using the normalized error of ROP prediction. Intuitively one can expect case 1 to be a better training set than case 2 – because case 1 has formation specific drilling data (or relevant data). However, case 2 contains data from other formations as well as the relevant data. This extra data (partially relevant) gets equal preference – by the algorithm – in building the data-driven model, which decreases the accuracy of models built on case 2. Case 2 is expected to be better than case 3 since it has some formation specific relevant data, whereas case 3 has data from other formations.

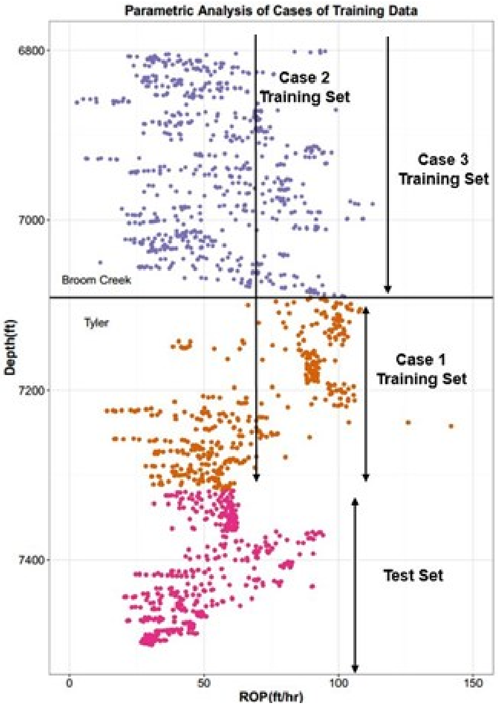


Figure 35: Illustration of test and training sets for parametric evaluation of nature of training set

Three training sets were used to build a statistical learning model which was evaluated for ROP prediction errors on the same test set. The best training set can be determined by comparing these prediction errors. Figures 36 and 37 show the test set errors for the three different cases of training sets. As expected, case 1 outperforms cases 2 and 3 for ROP prediction. Case 2 performs better than case 3 since it contains some data from the formation in question.

The Ratcliffe is the only formation where training data from case 2 and case 3 outperform case 1. This may be hypothesized due to the thickness of the formation. This formation has 67 ft of data, which makes it a very thin formation. The sparsity of available data in the formation causes higher error rates in case 1 as compared to cases 2 & 3. Case 2 performs better than case 3 since it includes some formation specific data, which has been shown to help achieve more accurate models. These indicate that for thin bedded formations it is better to include training data from previous formations.

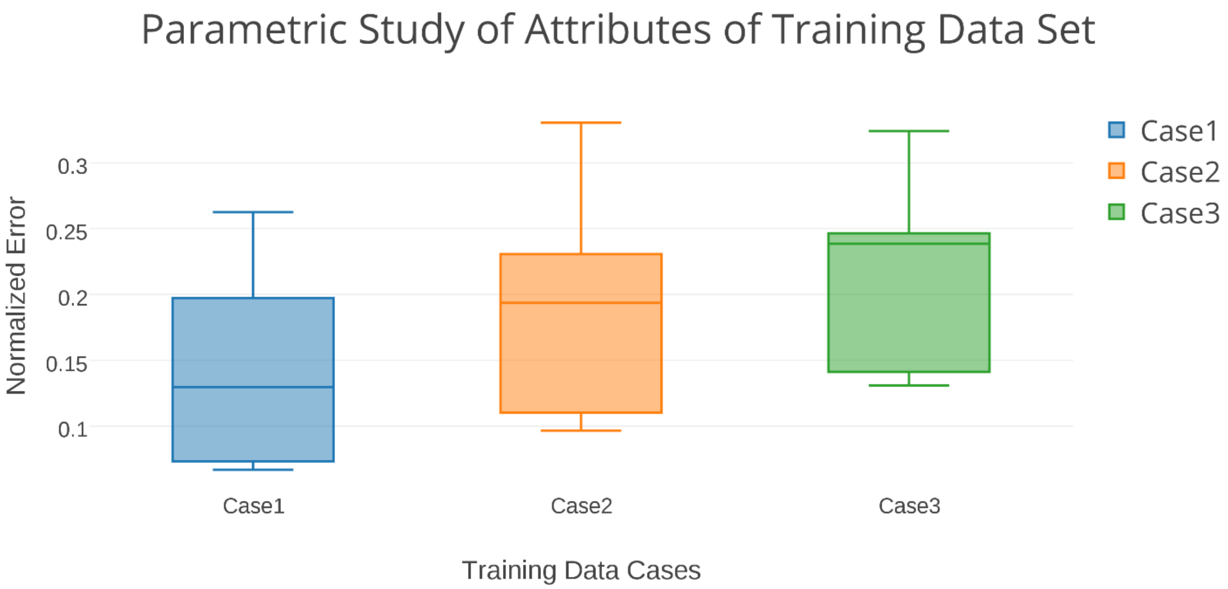


Figure 36: Box plot of normalized errors in different formations for changing training set attributes

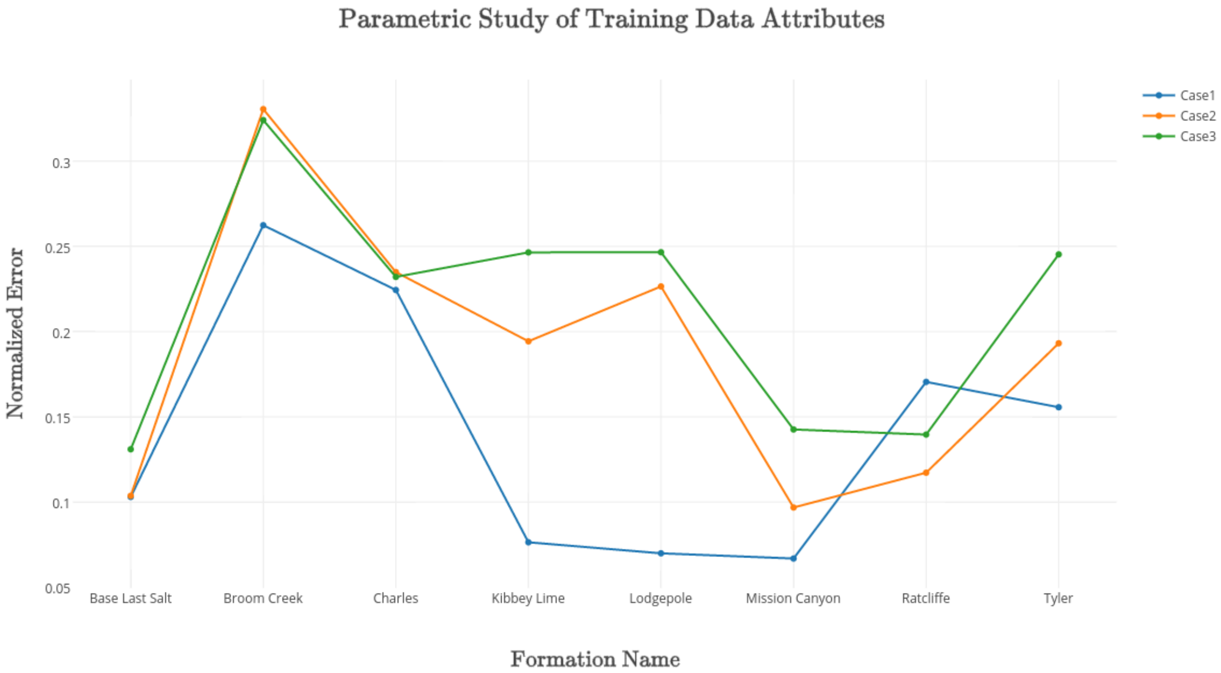


Figure 37: Line plot of normalized error in different formations for changing training set attributes

6.2: Study of Training-test Set Ratio in ROP Prediction

The drilling data in each formation is partitioned into training and test sets for ROP prediction. Increasing the length of the training set should improve the accuracy of the statistical learning model since more data would be available for learning. The optimum size of the training set depends on the formation as well as the data dependent.

The size of the training set relative to the test set have been changed for each formation; the ROP prediction error for each case was recorded. The training set was changed in size, varying its length from 10% to 90% of the size of the test set and the average prediction error is compared. Figure 38 shows the results obtained from this parametric study. A statistical learning model (random forest algorithm) has been used to predict ROP in each case. Figure 38 shows a decrease in error with an increase in the training-test set ratio, which indicates that an increase in the length of the training set produces an increase in accuracy (as expected). The accuracy desired (say a normalized error ratio of 0.2) can be easily computed from the plot in Figure 38. Figure 39 illustrates the decreasing error trend with an increase in training data in the form of a box plot.

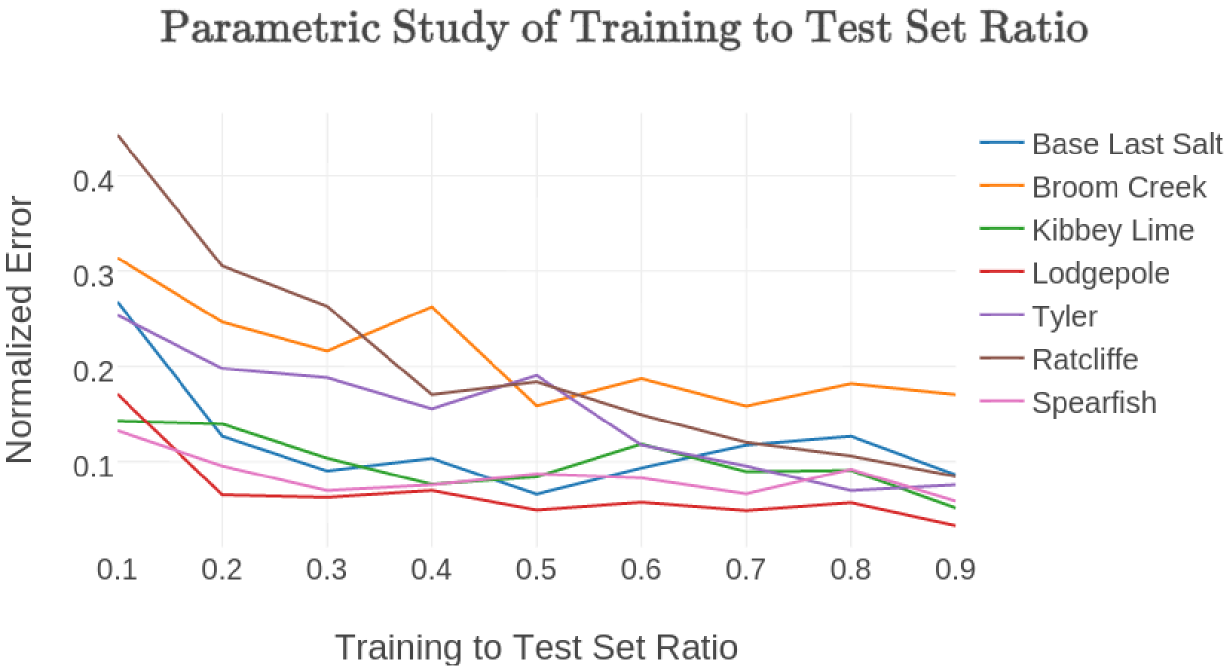


Figure 38: Parametric study of training-test set ratio in ROP prediction

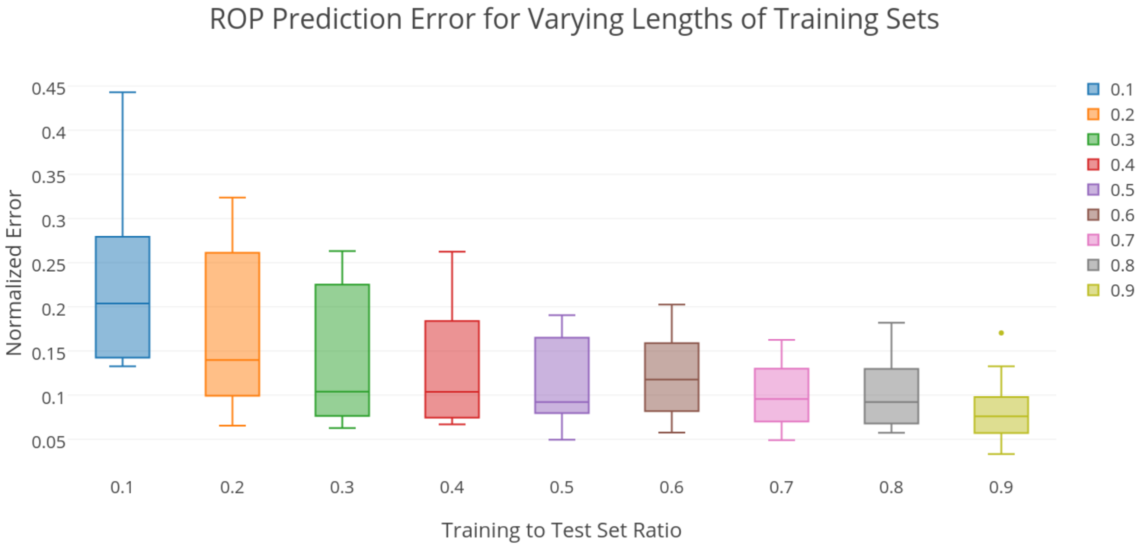


Figure 39: Box plot to visualize parametric study of training-test set ratio in ROP prediction

6.3: Optimal Training Sets

The plots in this section provide some insight into practical applications of statistical learning models in drilling. Training sets are more reliable and efficient for statistical leaning models when constrained to the formation of interest (case 1). Optimal training-test set ratios vary depending on required accuracy and formation. If an error rate of 0.2 or 20% is assumed to be required, then a ratio of 0.2 between training and test set length remains sufficient for most formations. A lower error rate requires a larger volume of training set data, pushing the training-test set ratio to 0.3-0.5 in a few cases as seen in Figure 9. In some cases (Tyler and Ratcliffe) higher ratios like 0.7 maybe necessary for low error rates of 10%. In one case (Broom Creek) a low error rate <10% is not possible for any ratio of training-test set data. The results indicate that in most cases 20-30% of the formation depth is sufficient to obtain an accurate model.

Chapter 7: ROP Analysis and Optimization

In the previous chapters different statistical learning models were introduced for prediction of ROP during drilling. Since drilling is a complex process, ROP prediction is not a simple task (Dashevskiy et al., 2013). Certain processes are controllable on the surface, whereas some of the input parameters cannot be controlled. Parameters such as weight-on-bit, RPM pf the bit and flow rate can be changed on the fly during drilling. Strength of the rock, pore pressure of the formation, and its thickness are examples of some uncontrollable input parameters. This chapter will explore techniques to change controllable parameters based on model recommendations to improve the ROP (or maximize it). This thesis will assume that the maximum attainable ROP will be the best ROP. The previous chapters have covered ROP prediction using surface measured input features. Some of the input features (ones that can be controlled on the surface: weight on bit, speed of bit rotation or RPM and pump pressure) can be used in conjunction with the ROP model to find the best settings to improve ROP. A random forests-based ROP model is used as the ROP model for optimization.

7.1: Variables And Spread of Data

Tyler formation has been used for ROP evaluation in this section. Surface controllable input parameters -WOB, pump pressure, RPM - have been varied to maximize ROP. Wallace et al. (2015) provide a framework to incorporate such a workflow in drilling. The authors introduce an "optimization score" which computes the percent of drilling efficiency based on an "optimum" scenario set by the user. In this case the optimum scenario would be one with the highest ROP (given a set of conditions like formation, strength of the rock etc.) as computed by the ROP prediction model. Theoretically all input parameters can be optimized to obtain the best set of parameters for ROP calculation, however, the percentage increase in ROP beyond two or three parameters are not worth the computational effort.

7.2: ONE DIMENSIONAL OPTIMIZATION

One dimensional optimization optimizes one input feature used in the ROP prediction model while keeping all other input features constant. It is important to note that values of the features varied have a limited threshold. The threshold is determined by field conditions since it is dangerous to extrapolate outside the range of data. In this thesis, the threshold does not exceed the range of the input feature in the training set. This way the predicted ROP is a realistic prediction which can be achieved while drilling the formation in question. A brute force algorithm (running all possible simulations and choosing the best) was used for optimization to compute the ideal value of the input feature. Since the model used is statistical in nature (and its shape unknown), the search for global maxima is not simple (it's easy to mistake a local maximum for the global maximum). A simple loop can be used to find the ideal setting for ROP optimization. The feature in question is varied keeping all other input features constant. This is used to calculate the ideal settings of the input feature. These settings can be plugged into the ROP model which will give an estimate of the improved ROP.

Figures 40, 41 and 42 shows the improvement in ROP with a change of input features (RPM, WOB and pressure). Optimization of RPM yields a much higher predicted

ROP as opposed to conventional drilling (Figure 40). Figure 41 shows the improvement in ROP when weight on bit is optimized. Figure 42 shows the ROP change on optimizing mud flowrate. The predicted ROP in each of the Figures (40, 41 and 42) have some “spikes” (sudden increases / jumps) which can be ruled as outliers by looking at the values of RPM, weight-on-bit (WOB), and flowrate which do not indicate a stick slip or excessive torsional vibrations. Predicting stick-slip in drilling or excessive vibrations is out of scope for this thesis, and will be discussed by the author in future work.

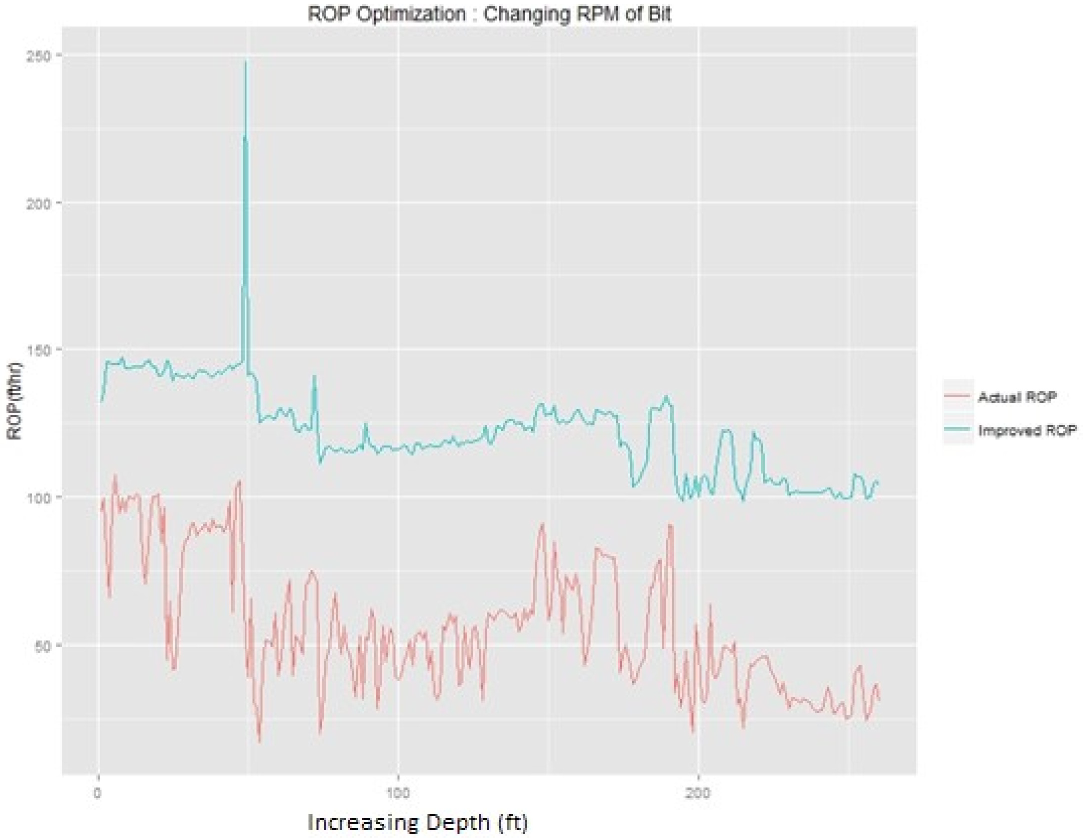


Figure 40: RPM of Bit Optimized to improve ROP in Tyler Formation

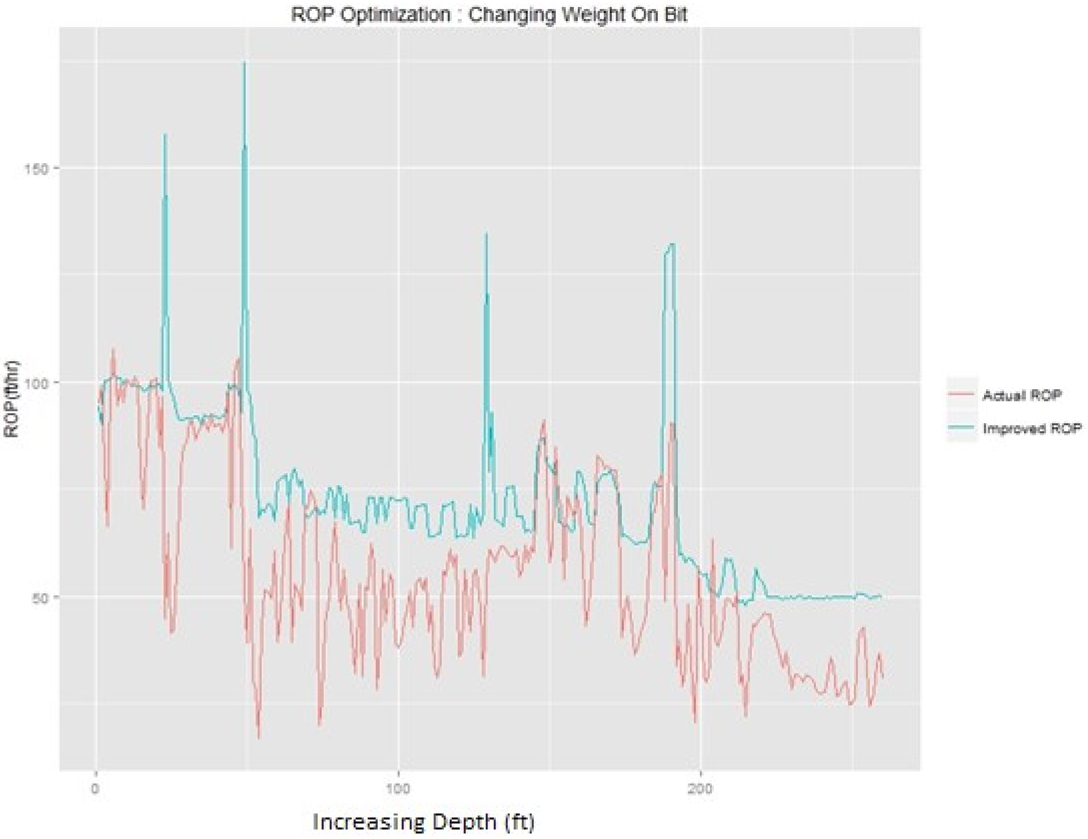


Figure 41: Optimization of weight on bit to improve ROP during drilling in Tyler Formation

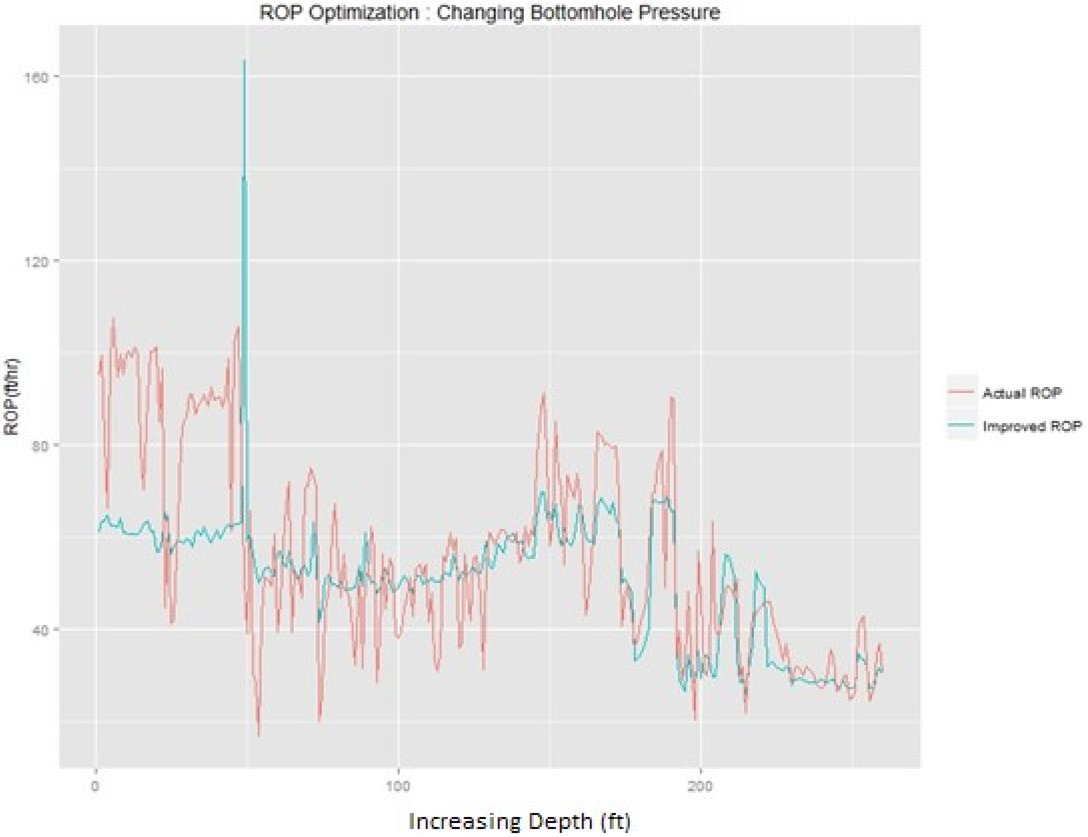


Figure 42: Bottom hole pressure optimized to stabilize ROP during drilling in the Tyler Formation

ROP is a complex function (determined using statistical learning algorithms) of its input features. These input features are coupled and do not act independently. Since more than one input feature is controllable on the surface, it is worthwhile to look at optimizing multiple features at once.

7.3: Two Dimensional Optimization

Two dimensional optimization refers to optimizing two input features on the surface in an effort to improve ROP. For a given set of input data two given features are varied while others are set to be a constant so that ROP may be maximized. The algorithm used for evaluation will be a brute force algorithm as before. Figure 38 plots the results of two dimensional optimization.

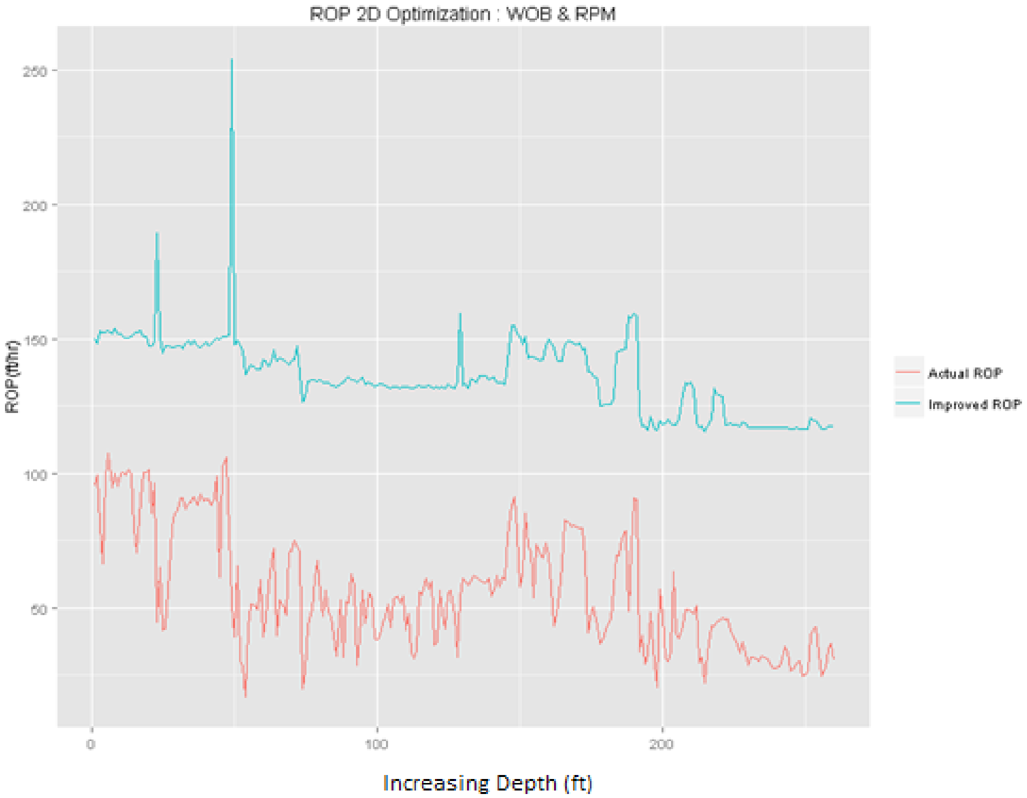


Figure 43: Weight on bit and bit rotation speed optimized to improve ROP during drilling in the Tyler Formation

Figure 43 illustrates optimization of weight on bit and RPM of the bit, which yields a better result with higher ROP as compared to the case when each individual parameters was optimized. Since a brute force algorithm is being employed for optimization, computational efficiency of algorithms are of the order N^3.

7.4: THREE DIMENSIONAL OPTIMIZATION

Three different features are optimized in in an effort to improve ROP. ROP was optimized by changing RPM of the bit, weight on bit and pump pressure (Figure 44). These variables were used since they had the highest individual covariance with ROP. Mean optimized ROP was 133 ft/hr which is much higher than the measured mean ROP of 57.52 ft/hr. This is a 137.5% increase in ROP if ideal values of WOB, RPM and pump pressure are used during drilling. Table 5 provides of a summary of different input features and their optimization of ROP.

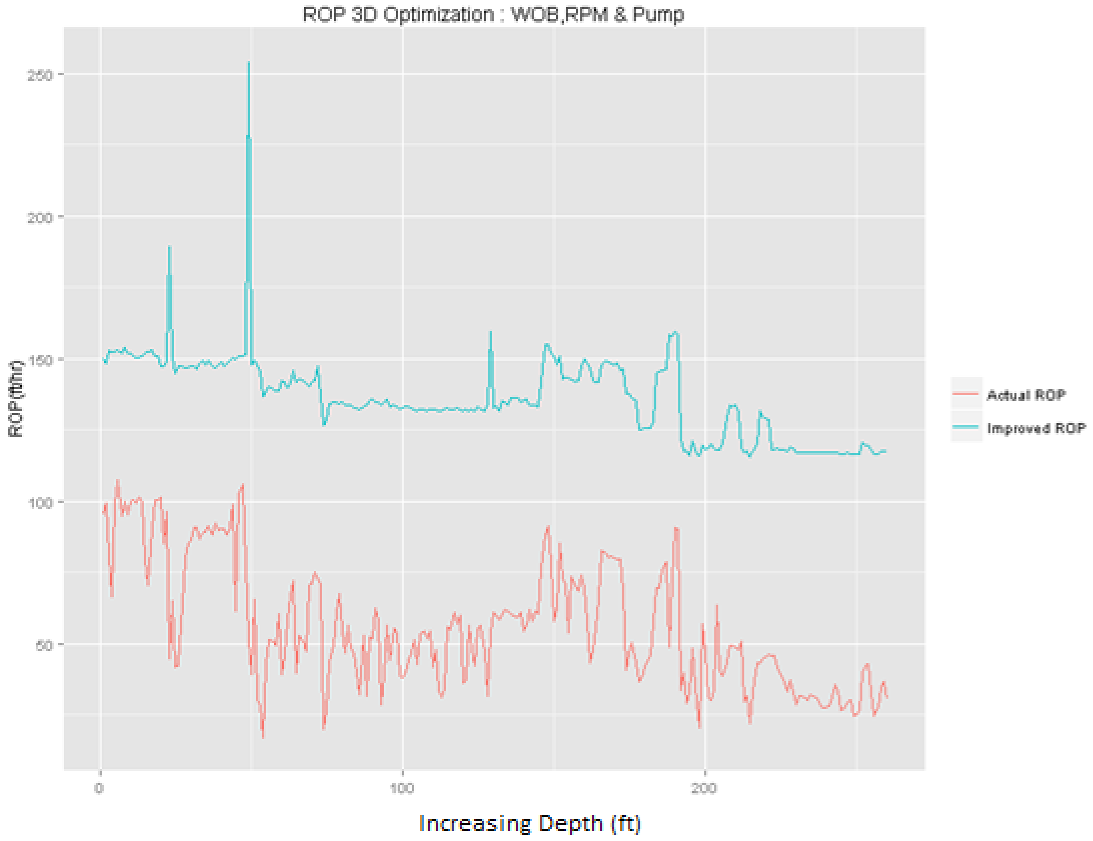


Figure 44: Three dimensional optimization of ROP in the Tyler Formation where WOB, RPM and mud flow rate are varied

Number of Features	**Features Optimized**	**Average Optimized ROP (ft/hr)**	**Percentage Increase of ROP (%)**
1	WOB	68	21%
1	RPM	113	101.7%
1	Pump Pressure	62	10.7%
2	WOB & RPM	128	128.57%
3	WOB, RPM & Pump Pressure	133	137.5%

Table 5: ROP Optimization using optimizing different number of features in the Tyler Sandstone Formation

This chapter shows the applications of data analytics and statistical learning to improve drilling efficiency. Here the role of statistical learning is to create an accurate prediction method which can be levied to set ideal surface inputs for the best ROP. A brute force algorithm (running all possible simulations and choosing the best) was used in this chapter for optimizing input parameters, however a better optimization scheme can be adopted to reduce computational time. Better optimization techniques must be investigated for field applications of this technique; One example is the Boender et al. 1982) developed an algorithm which can be used to find the global minima in case of a black-box (or unknown) function.

7.5: ROP OPTIMIZATION AND RATE OF DRILLING

ROP measures the rate of penetration which is essentially how fast or slow a well is being drilled. Given the nature of drilling, a post drill analysis will serve beneficial for drilling adjoint (or pad) wells. Time saved by drilling faster can be easily computed. Optimizing or increasing ROP by setting ideal surface parameters help reduce active drilling time of a well.

The time saved for drilling the whole well has been plotted in Figure 40. The data set was divided into smaller sets of 100 ft for each formation. For example, if the bit is at a depth of 5000 ft the training set is composed of data collected from 5000 ft -5100 ft. This data is used to training and develop a ROP model. This model is used to compute the ideal inputs (by optimizing two input features – ROP and WOB) for the test length (5000 ft – 5100 ft). Time to drill through a given section can be calculated in taking the inverse of ROP in that section. Figure 45 shows the ROP prediction for the entire data set. These ROP prediction models are used to compute and set idea parameters. Figure 46 shows the amount of time that can be saved by drilling with ideal parameters is 22 hours which comes out to 11.7% of total active drilling time.

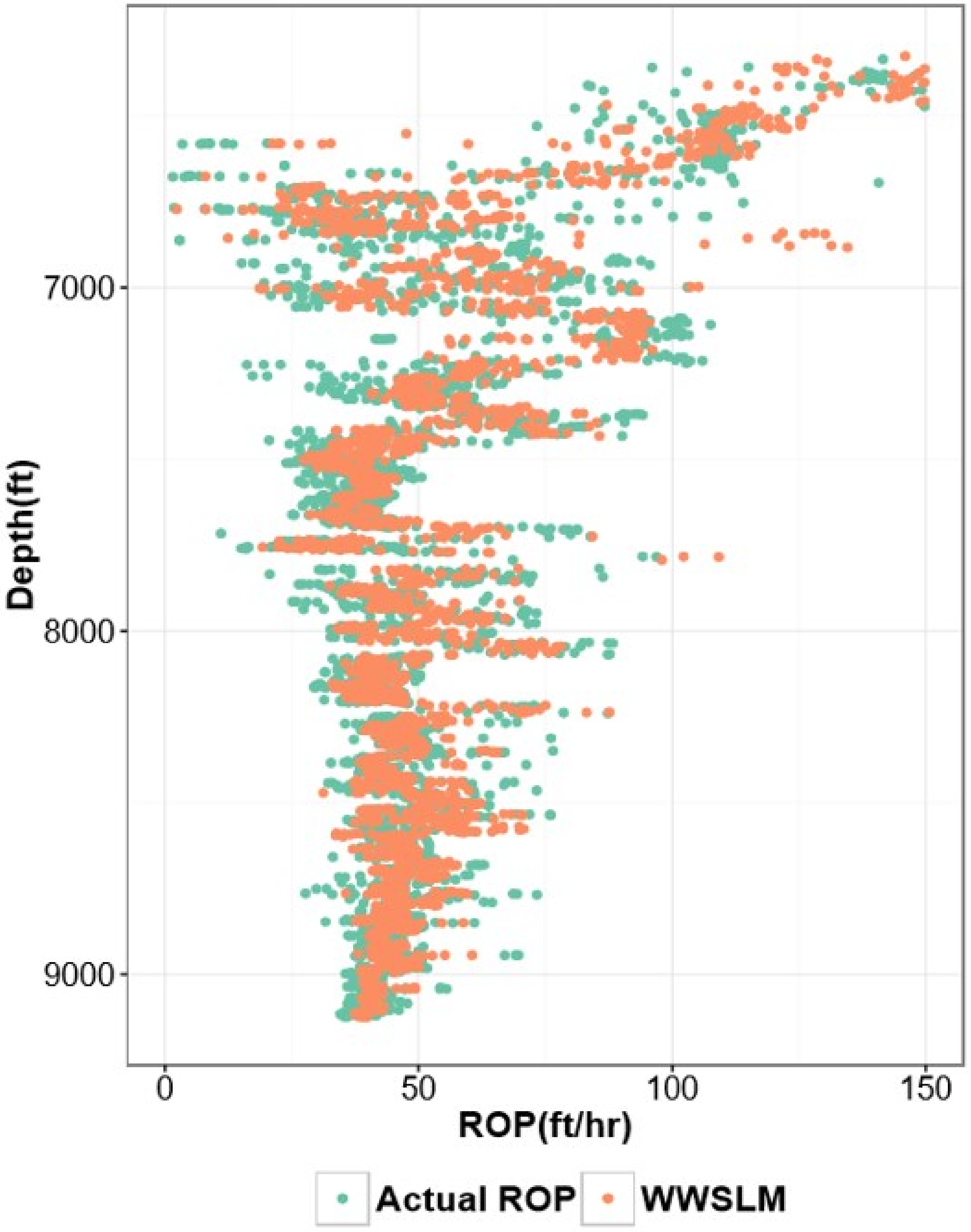


Figure 45: ROP prediction for entire well using the Wider Windows Statistical Learning Model (WWSLM) ROP model

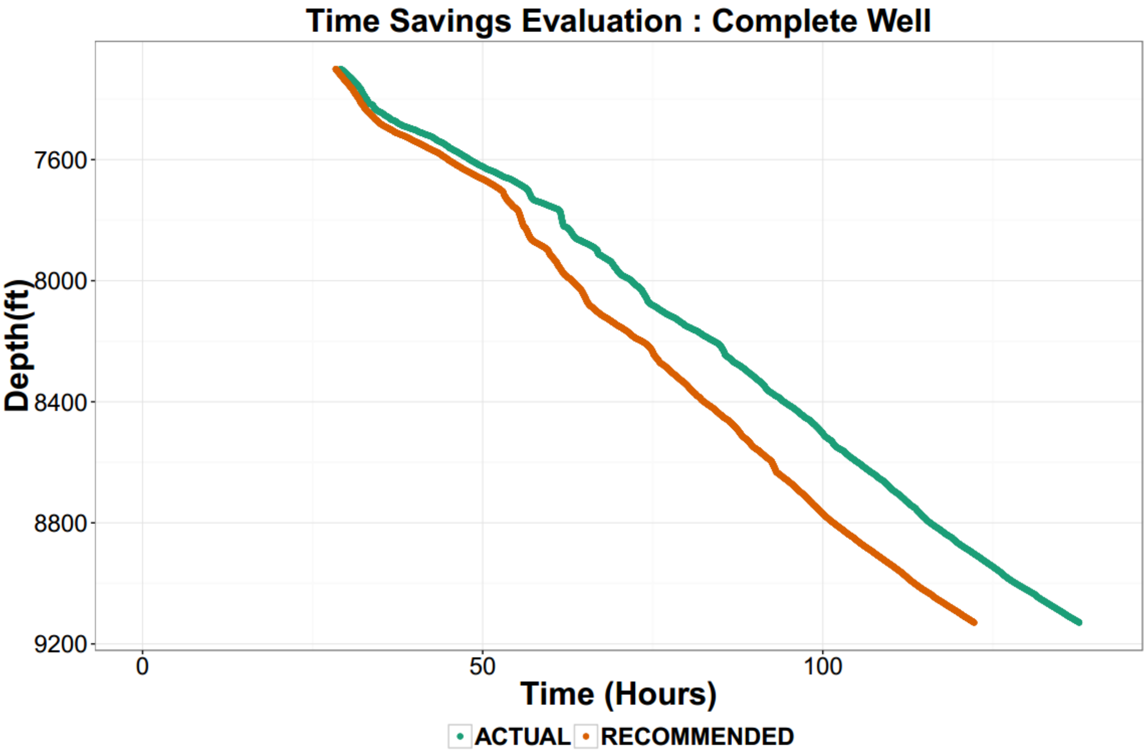


Figure 46: Time saved with ROP Optimization (22 hours)

Chapter 8: Future Research and Continuing Work

There are continuing efforts underway within the Drilling Parametrics group of Wider Windows that are directly related to this project. This project will continue to expand the Wider Windows group's understanding of the phenomena that affect drilling performance in the downhole environment.

Based on the success of the Wider Windows Statistical Learning Model in predicting drilling performance based only on the surface-readable input parameters, additional work will expand the WWSLM to include torque, MSE, and effects of vibration. A thorough comparison with traditional ROP models used in the industry is of interest. Higher versions of WWSLM are being developed which will include newer algorithms to curb the shortcomings of methods outlined in this thesis. Ensemble methods to improve accuracy has been a subject of research which will be addressed in future work. Better optimization algorithms to reduce computational time will be a part of future research. MSE is the parameter commonly used in industry to optimize drilling. By including all of these parameters a more comprehensive model can be developed which can address drilling optimization in a more robust fashion.

Chapter 9: Conclusions

This thesis set out to investigate application of the statistical learning to predict drilling parameters. ROP is identified here as a key parameter to be predicted and optimized. Importance of data visualization was discussed. Data management and its importance has been discussed in building statistical learning models. Splitting of data sets, model assessment and over fitting were duly addressed. Procedures to avoid overfitting were discussed which is important to any data analytics and statistical learning project. Auxiliary tools such as cross validation and bootstrapping were introduced. ROP was predicted using simple linear methods. Improvements were made to linear methods in an effort to increase their accuracy. Nonlinear methods were introduced as a technique to model non-linear data. Regularized regression was introduced for data with highly correlated data features. Computationally faster methods were introduced using PCA regression. A comparison of the regression techniques included an analysis of each regression model on various formations. Conclusions indicated that no single regression outperformed all others. However, based on needs of speed and accuracy different methods can be used when required.

Nonlinear methods were introduced to model and overcome the accuracy limitations of linear models. Trees, bagging and random forests were introduced as nonlinear algorithms used for ROP prediction. Trees have high variance and aren't very accurate. Bootstrapping can be used on trees so as to create an ensemble of trees, which are accurate and low in variance (bagging). The available input parameters at each tree split can be randomized to de-correlate the trees: random forests. Random forests predictions are extremely accurate and showed a high R^2 value. They are the best prediction method in terms of RMSE and R^2. Random forests were used to predict ROP resulting in a mean error of 13% of the measured data. A parametric study was conducted to evaluate the type and volume of data required to make accurate predictions using statistical learning models. It was concluded that data collected in the same formation as that of the test set is the best training set. The amount of data required is formation dependent. An error rate of sub 15% is generally acquired with a training set of lengths less than 30% of the total length of the formation.

ROP predictions on the training set were used to optimize ROP on the test set. Ideal input features were determined in order to achieve the highest ROP while drilling a formation. ROP was optimized by varying or “setting” one, two, and three input features. Increasing the number of input features optimized increased the average ROP, however this came with a great increase in computational time. A balance approach optimized two parameters where optimization of WOB and RPM saved 22 hours of active drilling time for the entire well.

In conclusion, statistical learning techniques and data analytics show promise in drilling optimization. They can be used for accurate prediction of ROP and simulation optimization of ROP. This is just the first step towards drilling optimization.

List of Acronyms

ANOVA: Analysis of Variance

BHA: Bottom Hole Assembly

BIC: Bayesian Inversion Criterion

BW: Bit Weight

CSV: Comma Separated Values

FEM: Finite Element Method

UCS: Unconfined Compressive Strength

IAP: Industrial Affiliate Program

LOOCV: Leave One Out Cross Validation

LWD: Logging While Drilling

MD: Measured Depth (ft)

MSE: Mechanical Specific Energy

MWD: Measurement While Drilling

NDB: Natural Diamond Bit

NPT: Non-Productive Time

OLS: Ordinary Least Squares

OOB: Out of Bag

PCA: Principal Component Analysis

PCR: Principal Components Regression

PDC: Polycrystalline Diamond Compact

PDM: Positive Displacement Motor

PLS: Partial Least Squares

RMSE: Root-Mean-Square Error

ROP: Rate of Penetration

RPM: Rotations per Minute

SLM: Statistical Learning Model

SPE: Society of Petroleum Engineers

T&D: Torque and Drag

TVD: Total Vertical Depth

WOB: Weight on Bit

WWSLM: Wider Windows Statistical Learning Model

References

Aadnoy, B. S., Fazaelizadeh, M., & Hareland, G. 2010. A 3D analytical model for wellbore friction. *Journal of Canadian Petroleum Technology*, *49*(10), 25-36.

Boender, C. G. E., Kan, A. R., Timmer, G. T., & Stougie, L. (1982). A stochastic method for global optimization. *Mathematical programming*, *22*(1), 125-140.

Bingham, M. G. 1965. A New Approach to interpreting rock drillability. Oil & Gas J. 94-101.

Bilgesu, H. I., Tetrick, L.T., Altmis, U., Mohaghegh, S., Ameri, S. 1997. A new approach for the prediction of rate of penetration (ROP) values. SPE Eastern Regional Meeting. Society of Petroleum Engineers.

Bourgoyne Jr, A. T., and F. S. Young Jr. 1974. A multiple regression approach to optimal drilling and abnormal pressure detection. *Society of Petroleum Engineers Journal* 14(04): 371-384.

Buntine, W. (1992). Learning classification trees. *Statistics and computing*,*2*(2), 63-73.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application* (Vol.1). Cambridge university press.

Dashevskiy, D., Macpherson, J. D., Dubinsky, V., & McGinley, P. (2007).*U.S. Patent No. 7,172,037*. Washington, DC: U.S. Patent and Trademark Office.

Dunlop, J., Isangulov, R., Aldred, W., Arismendi Sanchez, H., Sanchez Flores, J.L., Alarcon Herdoiza, J., Belaskie, J., Luppens, J.C. 2011. Increased rate of penetration through automation. SPE/IADC 139897.

Eckel, J. R. (1967). Microbit studies of the effect of fluid properties and hydraulics. Journal of Petroleum Technology, pp. 541-546.

Gjelstad, G., Hareland, G., Nikolaisen, K. N., and Bratli, R. K. (1998). The method of reducing drilling costs more than 50 percent. SPE/ISRM Eurock. Trondheim, Norway, July 8-10.

Hareland, G., and P. R. Rampersad. 1994. Drag-bit model including wear. SPE Latin America/Caribbean Petroleum Engineering Conference. Society of Petroleum Engineers.

Hegde,C.M., Wallace S.P. and Gray, K.E. (2015a). Using trees, bagging and random forests to predict rate of penetration during drilling. Presented at SPE Middle East Intelligent Oil & Gas Conference & Exhibition, Abu Dhabi, United Arab Emirates, 15-16 September. SPE-176792.

Hegde,C.M., Wallace S.P. and Gray, K.E. (2015b). Use of regression and bootstrapping in drilling: inference and prediction. Presented at SPE Middle East Intelligent Oil & Gas Conference & Exhibition, Abu Dhabi, United Arab Emirates, 15-16 September. SPE-176791.

Hegde,C.M., Wallace S.P. and Gray, K.E. (2015c). Real time prediction and classification of torque and drag during drilling using statistical learning methods. Presented at SPE Eastern Regional Conference, Morgantown, West Virginia, USA, 13-15 October. SPE-177313.

James, G., Witten, D., Hastie, T., & Tibshirani, R. 2013. *An introduction to statistical learning*. New York: springer.

Johancsik, C.A., D.B. Friesen, and Rapier Dawson. 1984. Torque and drag in directional wells-prediction and measurement. *Journal of Petroleum Technology* 36(6): 987-992.

Jahanbakhshi, R., R. Keshavarzi, and A. Jafarnezhad. 2012. Real-time prediction of rate of penetration during drilling operation in oil and gas wells. 46th US Rock Mechanics/Geomechanics Symposium. American Rock Mechanics Association.

Kitchel, B. G., Moore, S. O., Banks, W. H., & Borland, B. M. (1997). Probabilistic drilling cost estimating. SPE Computer Applications, 9(04), 121-125.

Lesage, M., I.G. Falconer, and C.J. Wick. 1988. Evaluating drilling practice in deviated wells with torque and weight data. *SPE Drilling Engineering* 3.03: 248-252.

Maidla, E.E., Wojtanowicz, A. K. 1987. Field method of assessing borehole friction for directional well casing. Presented at the Middle East Oil Show, Manama, Bahrain, March. SPE 15696.

Maurer, W. C. (1962). The “Perfect-Cleaning” theory of rotary drilling. Journal of Petroleum Technology, pp. 1270-1274.

Mevik, B. H., & Wehrens, R. (2007). The pls package: principal component and partial least squares regression in R. *Journal of Statistical Software*,*18*(2), 1-24.

Motahhari, H. R., Hareland, G., and James, J. A. (2010). Improved drilling efficiency technique using integrated PDM and PDC Bit parameters. Journal of Canadian Petroleum Technology, v. 49, no. 10, pp. 45-52.

Newman, K.R, and R Procter. 2009. Analysis of hook load forces during jarring. IADC/SPE Drilling Conference and Exhibition. Amsterdam, Netherlands, March.

Newman, K. R. Finite element analysis of coiled tubing forces. Society of Petroleum Engineers, SPE 89502, SPE/ICoTA Coiled Tubing Conference and Exhibition, Houston, Texas, March 2004.

Nygaard, R., Hareland, G., Budiningsih, Y., Terjesen, H. E., and Stene, F. (2002). Eight years experience with a drilling optimization simulator in the north sea. IADC/SPE Asia Pacific Drilling Technology. Jakarta, Indonesia, September 11.

Rampersad, P. R., Hareland, G., and Boonyapaluk, P. (1994). Drilling optimization using drilling data and available technology. III Latin American/Caribbean Petroleum Engineering Conference. Buenos Aires, Argentina, April 27-29

Saldivar, B., Boussaada, I., Mounier, H., Mondie, S., & Niculescu, S. I. (2014, August). An overview on the modeling of oilwell drilling vibrations. In *World Congress* (Vol. 19, No. 1, pp. 5169-5174).

Soares, C. (2015). Development and applications of a new system to analyze field data and compare rate of penetration (ROP) models. M.S Thesis, The University of Texas at Austin.

Walker, B. H., Black, A. D., Klauber, W. P., Little, T., and Khodaverdian, M. (1986). Roller-bit penetration rate response as a function of rock properties and well depth. 61st Annual Technical Conference and Exhibition of the Society of Petroleum Engineers. New Orleans, LA, USA, October 5-8.

Wallace, Hegde and Gray (2015). System for real time drilling performance optimization and automation based on statistical learning methods. Presented at SPE Middle East Intelligent Oil & Gas Conference & Exhibition, Abu Dhabi, United Arab Emirates, 15-16 September. SPE 176804.

Warren, T. M. (1987). Penetration-rate performance of roller-cone bits. SPE Drilling Engineering, pp. 9-18.

Winters, W. J., Warren, T. M., and Onyia, E. C. (1987). Roller bit model with rock ductility and cone offset. 62nd Annual Technical Conference and Exhibition of the Society of Petroleum Engineers. Dallas, TX, USA, September 27-30.