|  |  |
| --- | --- |
|  | |
| **Restricted** | **SR.15.11128** |
| **SweetSpots Identification for UNC Plays Through Machine Learning Method**  **by**  **Mingqi Wu (PTD/TASE)**  **Ligang Lu (PTI/RP)**  **Jan Limbeck (PTI/RP)**  **Detlef Hohl (PTI/RP)** | |
|  | |
|  | |
|  | | |
|  | | |
| This document is classified as Restricted. Access is allowed to Shell personnel, designated Associate Companies and Contractors working on Shell projects who have signed a confidentiality agreement with a Shell Group Company. 'Shell Personnel' includes all staff with a personal contract with a Shell Group Company. Issuance of this document is restricted to staff employed by a Shell Group Company. Neither the whole nor any part of this document may be disclosed to Non-Shell Personnel without the prior written consent of the copyright owners. | | |
|  | | |
| Copyright Shell Global Solutions (US) Inc., 2014. | | |
|  | | |
| **Shell Global Solutions (US) Inc., Houston** | | |
|  | | |
| Further electronic copies can be obtained from the Global Information Centre. | | |

# Executive summary

Reducing the exploration cost and improving the productivity of oil and gas reserviors, in particular the unconventional (UNC) plays are increasingly important and vital to Shell’s business. The existing geological and petro-physical methodologies and related simulation technologies that achieved great success in conventional reservoirexploration and production have not been as effective when applied to UNC exploration and production. It remains highly challenging to accurately and rapidly characterize the high estimated ultimate recovery (EUR) regions of a UNC play with early exploration data to provide guidance on where to drill and how to complete new production wells.

To address the above challenges, the Department of Computation and Modelling (PTI/RP) has initiated this project to investigate and develop machine learning techniques based data driven methodologies for UNC exploration and production. The project has two phases to achieve two sets of objectives:

**Phase I** Objectives are to investigate if machine learning techniques are feasible and effective for data analytics to reveal the underlying relationships among the subsurface geological features and the oil production in a UNC play, such as the Eagle Ford play; and to identify key subsurface geological variables that are highly correlated with production or EUR; and further to develop a machine learning based methodology and workflow to build predictive models to identify the high productivity areas (sweetspots) for drilling new wells s and understand the minimum amount of the data needed to establish reliable predictions.

**Phase II** Objectives are, building on the Phase I work, to investigate how the machine learning techniques can provide data analytics to understand how the drilling and completion will impact well productivity; and to uncover the complex correlations among the variables and productivities; to identify the top factors that can guide the drilling and completion process in order to improve the production.

focuses on our work to achieve the Phase I objectives that addresses the challenge of “where to drill” for UNC plays. It describes our solution--the ML based data analytics methodlogy and workflow as well as our results and findings in.

In the completed Phase I work, we have developed an effective machine learning based data analytics methodology and workflow to identify sweet spots of drilling locations and their associated production estimations. Our key results and findings are:

1. based is useful for UNC play exploration. Our developed methodology and workflow can tingexisiting s;
2. Through our evaluation of multiple ML techinsques, we found that the ML algorithm, *Random Forest*, gives the better results;
3. Performance analyese of the predictive modelling against various training data size are obtained to understand the minimum data amount required in order to achieve reliable and accurate prediction results;
4. Using only a few key variables to build a predictive model can provide even better predictive results; their ranges for high productivity may offer insights for drilling and completion process.

**Table of contents**

[Executive summary II](#_Toc416107829)

[1. Introduction 1](#_Toc416107830)

[2. Data and Workflow 2](#_Toc416107831)

[2.1. Data 2](#_Toc416107832)

[2.2. Workflow 3](#_Toc416107833)

[3. Random Forests and BART 5](#_Toc416107834)

[3.1. Decision Trees 5](#_Toc416107835)

[3.2. Random Forests 6](#_Toc416107836)

[3.3. Bayesian Additive Regression Trees 7](#_Toc416107837)

[4. Results 8](#_Toc416107838)

[4.1. Random Forest 8](#_Toc416107839)

[4.1.1. Classification approach 8](#_Toc416107840)

[4.1.2. Regression approach 19](#_Toc416107841)

[4.2. BART 27](#_Toc416107842)

[4.2.1. Parameter tuning 27](#_Toc416107843)

[4.2.2. Variable importance 28](#_Toc416107844)

[4.2.3. Prediction 29](#_Toc416107845)

[5. Discussion and future work 30](#_Toc416107846)

[6. Acknowledgements 31](#_Toc416107847)

[References 32](#_Toc416107848)

[Bibliographic information 33](#_Toc416107849)

[Report distribution 34](#_Toc416107850)

**List of tables**

[Table 1 RF OOB error rate at different subset size of predictors randomly sampled at each split with the number of trees fixed at 1000. 10](#_Toc416107851)

[Table 2 RF classification results in ratio with 5-fold cross-validation. 15](#_Toc416107852)

[Table 3 RF classification results in number of wells with 5-fold cross-validation. 16](#_Toc416107853)

[Table 4 RF classification results in ratio with different fold cross-validation. 16](#_Toc416107854)

[Table 5 RF classification results in ratio with 5-fold cross-validation using only four important predictors: true vertical depth, Tmax, S2 and S3. 19](#_Toc416107855)

[Table 6 RF OOB MSE at different subset size of predictors randomly sampled at each split with the number of trees fixed at 1000. 21](#_Toc416107856)

[Table 7 RF regression results in number of wells with 5-fold cross-validation. 22](#_Toc416107857)

[Table 8 RF regression results with different fold cross-validation. 23](#_Toc416107858)

[Table 9 RF regression results with 5-fold cross-validation using only 6 important predictors: true vertical depth, S2, Tmax, Chlorite in clay(X-ray diffraction), Ro measured, GRI water filled porosity. 25](#_Toc416107859)

[Table 10 BART regression results in number of wells with 5-fold cross-validation. 29](#_Toc416107860)

[Table 11 Methods comparisons in terms of true positive number. 30](#_Toc416107861)

**List of figures**

[Figure 1 Overview of Eagle Ford data sets. 2](#_Toc416107862)

[Figure 2 Heat map of Eagle Ford oil producers based on normalized 12-month production data. 3](#_Toc416107863)

[Figure 3 Workflow to build predictive model based on core and production data 4](#_Toc416107864)

[Figure 4 Illustration of decision tree 5](#_Toc416107865)

[Figure 5 RF OOB error rate with different number of trees at various subset sizes, ,of predictors randomly sampled at each split (a) (b) (c) . 10](#_Toc416107866)

[Figure 6 A variable importance plot for the EF data set. Variable importance is computed using the permutation importance measure normalized by the maximum score. 11](#_Toc416107867)

[Figure 7 A variable importance plot for the EF data set. Variable importance is computed using the gini importance measure normalized by the maximum score. 12](#_Toc416107868)

[Figure 8 Correlation heat map among the 31 predictors under consideration. 13](file:///C:\Apps\projects\DataMiningUNC\Docs\rept\SweetSpot.docx#_Toc416107869)

[Figure 9 Correlation heat map for overlapped predictors among the top 10 important predictors identified by permutation measure and gini measure. 14](file:///C:\Apps\projects\DataMiningUNC\Docs\rept\SweetSpot.docx#_Toc416107870)

[Figure 10 Predicted sweetspots by RF classification approach using all predictors. 15](file:///C:\Apps\projects\DataMiningUNC\Docs\rept\SweetSpot.docx#_Toc416107871)

[Figure 11 RF classification results in ratio at different percentage of training data. 17](file:///C:\Apps\projects\DataMiningUNC\Docs\rept\SweetSpot.docx#_Toc416107872)

[Figure 12 RF classification results in ratio with top K important predictors (permutation measure). 18](file:///C:\Apps\projects\DataMiningUNC\Docs\rept\SweetSpot.docx#_Toc416107873)

[Figure 13 RF classification results in ratio with top K important predictors (gini measure). 18](file:///C:\Apps\projects\DataMiningUNC\Docs\rept\SweetSpot.docx#_Toc416107874)

[Figure 14 RF OOB MSE with different number of trees at various subset sizes, , of predictors randomly sampled at each split (a) (b) (c). 20](#_Toc416107875)

[Figure 15 A variable importance plot for the EF data set. Variable importance is computed using the permutation-based MSE reduction measure normalized by the maximum score. 21](#_Toc416107876)

[Figure 16 Predicted sweetspots by RF regression approach using all predictors 22](file:///C:\Apps\projects\DataMiningUNC\Docs\rept\SweetSpot.docx#_Toc416107877)

[Figure 17 RF regression results at different percentage of training data. 23](file:///C:\Apps\projects\DataMiningUNC\Docs\rept\SweetSpot.docx#_Toc416107878)

[Figure 18 RF regression results with top K important predictors (permutation-based MSE measure). 24](#_Toc416107879)

[Figure 19 Partial dependence plot for three selected important predictors. 26](file:///C:\Apps\projects\DataMiningUNC\Docs\rept\SweetSpot.docx#_Toc416107880)

[Figure 20 Out-of-sample RMSE with different number of trees. 27](#_Toc416107881)

[Figure 21 A variable importance plot for the EF data set. Variable importance is computed using the variable inclusion proportions measure normalized by the maximum score. 28](#_Toc416107882)

[Figure 22 BART regression results at different percentage of training data. 29](#_Toc416107883)

# 

# Introduction

Reducing the oil and gas exploration cost and improving well productivity becomes both strategically important and economically vital to Shell. The department of Computation and Modeling (PTI/PR) has funded and launched this project to investigate and develop machine learning based new methodologies that can be applied to unconventional (UNC) gas and oil plays to characterize and identify the top productive areas (sweet spots) for new drilling locations and to guide the best selection of the drilling and completion parameters to improve the success rate and the estimated ultimate recovery (EUR). This report describes our work and findings from the Phase I work of the project—sweetspotting.

It is well recognized in the oil and gas industry that existing geological and petro-physical methodologies and simulation techniques are not as effective as in conventional play exploration and production when applied to UNC plays. Better methodologies are highly demanded to improve the cost effectiveness of the UNC exploration and production. On the other hand, the emerging modern machine learning techniques have shown good effectiveness in tackling complex and data intensive problems, where conventional technologies often become ineffective. Therefore this project aims to investigate and develop machine learning based new methodologies to improve the effectiveness and reduce the cost of UNC exploration and production. The specific technical challenges addressed by the first phase of this project are to uncover the underlying cross-correlations among the UNC subsurface geological features and EUR at well locations, and to build effective predictive models that can be used to predict the new well EURs, and to identify the key features as the predictors, and to understand the minimum amount of data (in terms of well numbers) required to produce reliable and accurate predictions for rapidly classifying the most promising areas and business value of a UNC play.

Machine learning concerns data driven algorithmic techniques that can incorporate pertinent information from input data sources and learn the underlying relationships and patterns, often complex and unknown, among the data sources and how do they affect the interested target variables. Thus machine learning provides a promising means to tackle the complex exploration and production problems arising from UNC plays, wherein the underlying geophysics and petrophysics are not well understood. Our Phase I work of this project has investigated how to apply modern machine learning techniques on the Eagle Ford data set to uncover the relationships among the subsurface geological features and well productivity and further to build predictive models to identify the sweet spots for new well locations and their EURs. Specifically, in the following, Section 2 will describe the Eagle Ford data sets and our methodology’s workflow; Then Section 3 will depict the used machine learning algorithms: Decision Trees, Random Forest and Bayes Additive Regression Trees; Section 4 will present the results and findings from our Phase I work. Finally, Section 5 will provide further discussion of the findings and future work.

# Data and Workflow

## Data

The Eagle ford data sets used for this study were obtained from Shell internal business unit; Figure 1 provides an overview of this data, it includes core data from **83** wells (red dots), which represent the geological parameters information, and monthly production data from **2631** oil producers (black dots). In total, there are 31 geological parameters under consideration, additionally, completion parameters, such as lateral length of each producer, are available for production normalization.

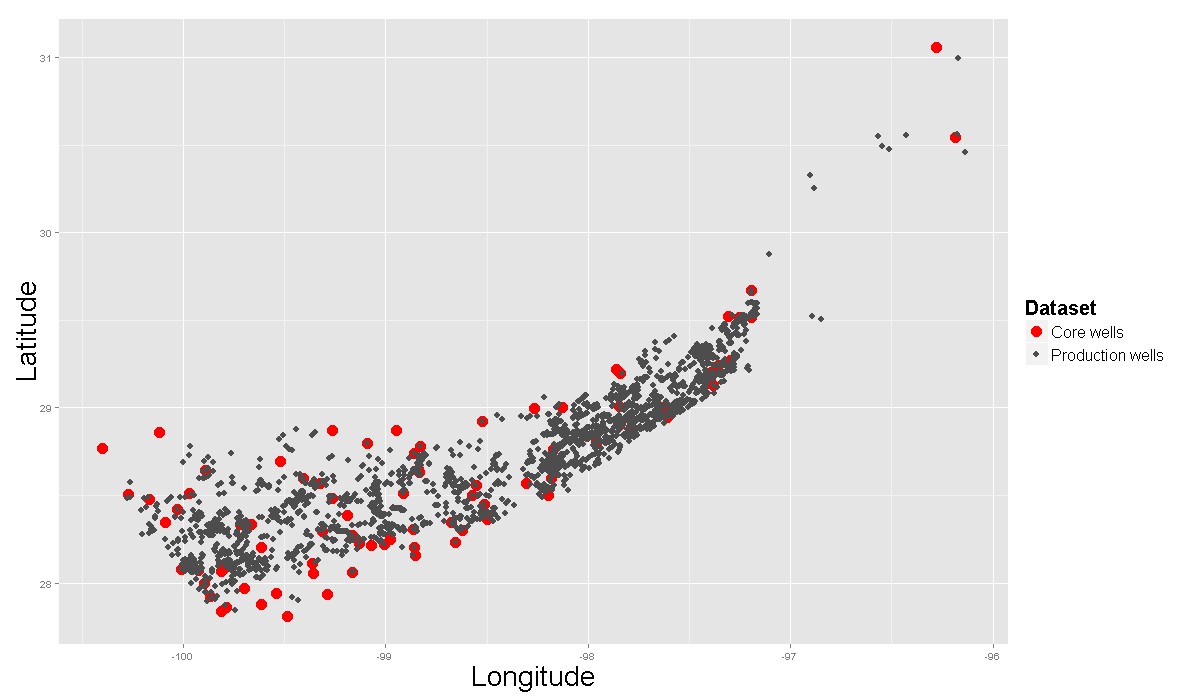


Figure Overview of Eagle Ford data sets.

Figure 2 shows the heat map of Eagle Ford oil producers based on first 12-month production data normalized by lateral length of each producer. Please note, first cumulative 12-month production has been selected as a proxy of EUR. As indicated by the production color bar, locations with red or close to red color represent the so called “sweetspot”, i.e. top quartile most productive areas of shale oil plays. The objective of this study is to build a predictive model with machine learning algorithms that may help to identify these sweetspots based on the values of the geological parameters at the corresponding locations.

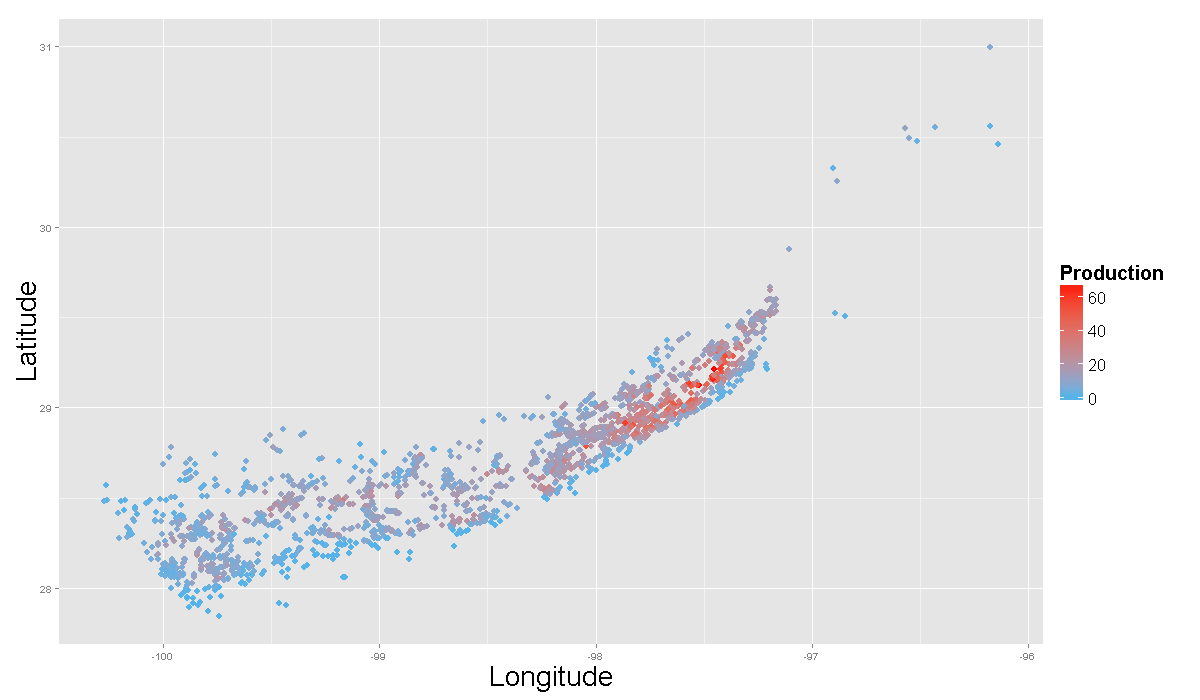


Figure Heat map of Eagle Ford oil producers based on normalized 12-month production data.

## Workflow

Generally speaking, the workflow to build a predictive model with machine learning algorithm from core samples and monthly production data involves two key technical stages: (i) Data interpolation

(ii)Predictive modeling. The former requires knowledge of geostatistical techniques and the latter uses machine learning techniques.

* Data interpolation

In this phase, the inputs are the aggregated core data, and the outputs are krigged maps for each geological parameter. At each core location, there are multiple core samples obtained from different depth, before kriging, for each parameter, its measurements need to be first aggregated or summarized into one value to represent the corresponding geology characteristics at this location.

* Predictive modeling

Basically, given the krigged maps of each geological parameter from data interpolation, the geology information at each producer location is obtained. With the geology information and production data at each producer’s location, the aim of stage II is to learn from these combined information and build a predictive model through machine learning techniques.

The flowchart of the two stage workflow is summarized in Figure 3, the focus of this report is on stage II, building predictive model from krigged geological parameters value and production data. For stage I, how to summarize core samples and interpolate the aggregate values will be described in a follow up report.

**Data Interpolation**

**Predictive Modeling**

Figure Workflow to build predictive model based on core and production data

# Random Forests and BART

In this report, we study the performance of Random Forests (Breiman, 2001) [1], one of the most popular and successful “off-the-shelf” machine learning algorithms, and Bayesian Additive Regression Trees (BART) in building a predictive model for shale oil plays sweetspots identification.

Random Forests (RF) has many attractive features, such as fast, scalable, and easy to implement without many tuning parameters, but the main reasons that it has been selected for this study are due to the following benefits:

* It is a non-parametric method without any distribution assumptions on the predictors, response and residuals, and is robust against outliers.
* It can handle well nonlinear relationships between predictors and response and complex interactions among predictors.
* It provides effective methods for estimating missing data.

RF is an ensemble learning algorithm, which is also called ensemble decision trees method. It operates by growing many decision trees, and the final prediction is a summarized result from all the decision trees.

## Decision Trees

Decision trees are predictive model that use a set of binary rules to calculate a target value, it can be applied to both classification and regression problems.

X4 ≥ t4

X4 < t4

X3 ≥ t3

X3 < t3

X2 ≥ t2

X1 ≥ t1

X2 < t2

X1 < t1

C4

C3

C2

C1

C5

Figure Illustration of decision tree

A decision tree has a tree-like structure with nodes, branches and leafs. Each node in the tree specifies a test of a predictor of the instance, and each branch descending from that node corresponds to one of the possible values for this predictor. The decision tree starts from a node called “root”, which has two outgoing branches, but no incoming branches. All other nodes in the tree have two outgoing branches and exactly one incoming branch. Leaves are known as terminals of decision trees, it is connected to a node with an incoming branch. Each leaf is assigned to one class for classification or the mean of response value for those instances falling into this leaf at training stage for regression. As illustrate in Figure 4, circles, arrows and squares represent nodes, branches and leafs of the decision trees. Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests of the predictors along the branches. Given a trained decision tree, it can predict the response of a new instance based on its attributes (predictors).

## Random Forests

Single decision trees often suffer from high variance [2], that means if we split the training data into two parts at random, and fit a decision tree to both halves, the results that we get could be quite different.

RF attempts to mitigate the problem by growing many decision trees and introducing a “voting system” to help in decision making. For classification, the forest chooses the class with the majority votes from all the trees; for regression, the final results will be the average value of the prediction from each tree. Each tree in RF algorithm is grown as follows [3]:

1. If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning.

In the above procedure, if we set m=M in step 2, the algorithm is called *bagging* or *bootstrap aggregation*. For each tree, it uses a bootstrapped training sample and considers all available predictors in building the decision tree. Bagging’s idea is to reduce estimation variance through averaging the models trained from bootstrap samples, however, sometimes there are high correlation exist among the bagged trees, which make the variance reduction very limited. In contrast, RF decorrelates the trees by only considering a subset of the predictors at each split during the tree building process. This subtle change leads to improvement in model accuracy.

## Bayesian Additive Regression Trees

Motivated by the success of ensemble-of-trees method, a Bayesian version of “sum-of-trees” model was developed recently by Chipman George and McCulloch, called Bayesian Additive Regression Trees (BART) [4]. BART distinguishes itself from other ensemble-of-trees method by using a sum of regression trees to model an unknown function developed under the Bayesian framework. The key idea of BART is to impose a prior that regularizes the fit of each tree, such that every tree model only has small effect, in another word, each tree is a “weak learner”, and only explain a small and different portion of . BART has demonstrated substantial promise in a wide variety of simulations and real word applications [5]. Compared with RF algorithm, BART not only provides single point estimation for regression, but also can compute uncertainty estimates via quantiles of the posterior samples, that is the reason BART is being identified as a potential algorithm for this project.

# Results

The business goal of this project is to identify “sweetspot”, i.e. top quartile most productive areas, of shale oil plays through machine learning techniques. This objective can be treated either as a classification or regression problem. For classification approach, the production wells will be divided into two classes according to their first 12-month production: top quartile (Q1) and the rest (~Q1), then the task is to train a model to classify producers into this two classes; for regression approach, first, a model that is able to predict the first 12-month production will be build; second, the producers will be ranked according to their predicted values; third, top quartile producers will be identified as “sweetspot” based on the ranking from step 2.

As introduced in section 3, RF and BART are decision trees based methods, therefore they can be applied to both classification and regression problems. The results from both approaches using RF and BART to identify “sweetspot” of Eagle Ford data set are studied, compared and summarized in this section.

## Random Forest

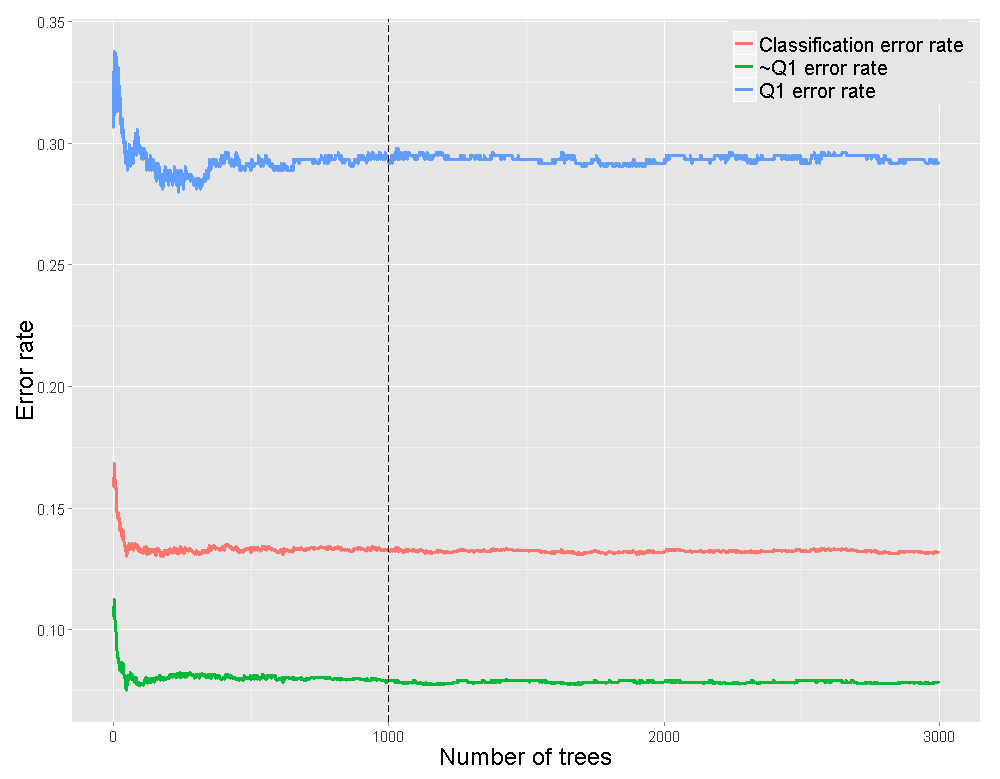
### Classification approach

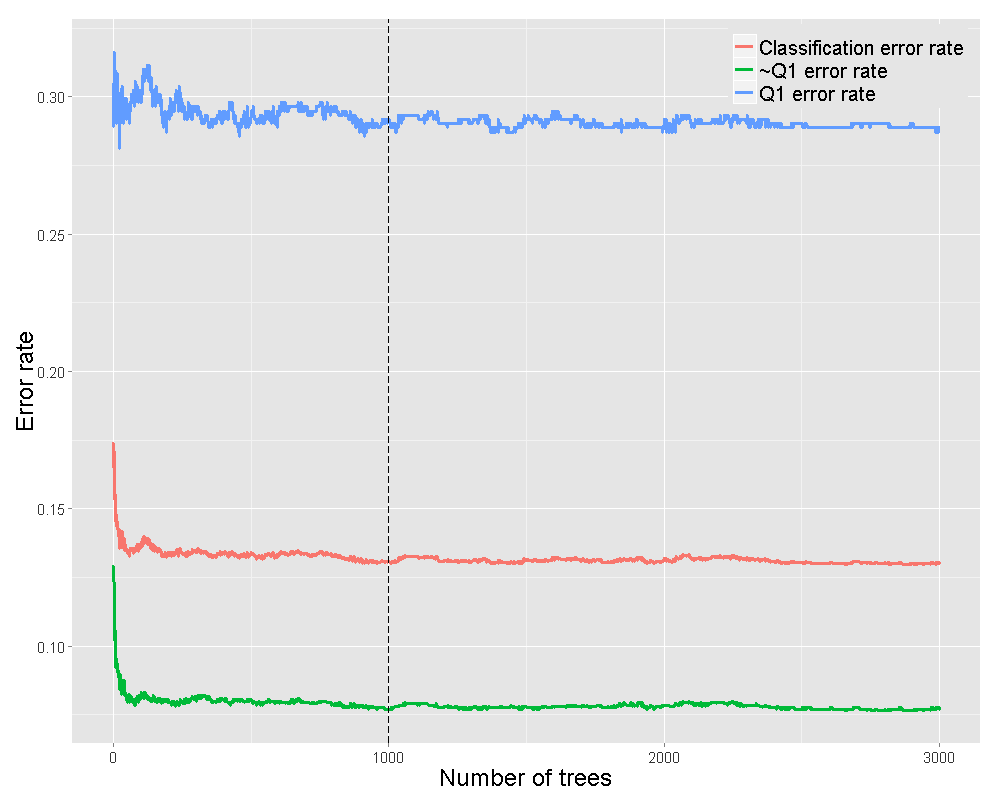
#### Parameter tuning

To find the optimum parameters for RF, there is no need to perform cross-validation or the validation set approach, RF automatically provide *out-of-bag* (OOB) error estimation, which can be used as unbiased estimate of the test error. Recall for RF, each tree is constructed from a bootstrapped subset of the original data. One can show that on average, each bagged tree makes use of around two-thirds of the observations. The remaining one-third of the observations are not used to fit a given tree are referred to as the OOB observations [2]. The OOB observations are like a test set for this tree. Put each OOB observation to those trees that left them out in the construction process to get predictions, and average them, an OOB prediction for the observation is obtained, from which an OOB error rate can be computed. The overall OOB error rate is computed as the average over the OOB error rate of all observations.

* Number of trees

Figure 5 shows the OOB error rate with different number of trees at various subset sizes of predictors randomly sampled at each split for the Eagle Ford dataset. The error rates under investigation include overall classification error rate, class Q1 error rate (false positive) and class ~Q1(false negative) error rate. From the plots, the error rates get stable for all the three scenarios once the number of trees is greater or equal to 1000. After 1000, even more trees are added, the OOB error rates do not decrease any more, which means RF does not suffer from overfit problem. The number of trees will be fixed at 1000 for the rest of this study.





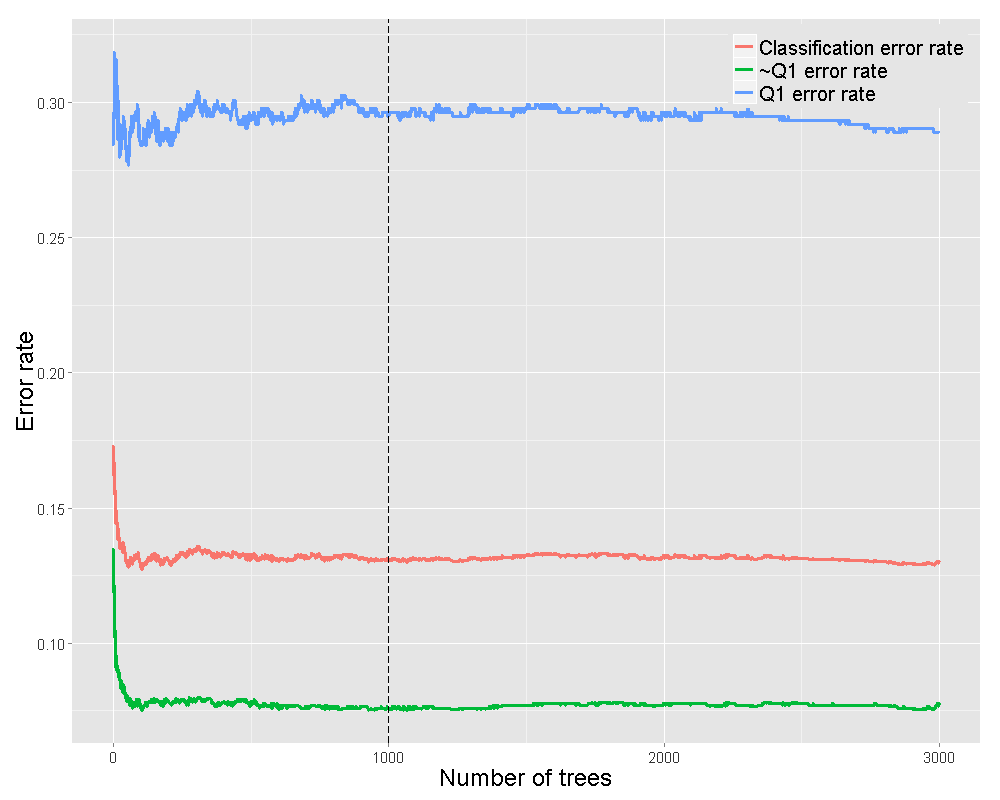


Figure RF OOB error rate with different number of trees at various subset sizes, ,of predictors randomly sampled at each split (a) (b) (c) .

* Subset size of predictors randomly sampled at each split

With the number of trees fixed at 1000, the OOB error rates are calculated at different subset size of predictors randomly sampled at each split  **.** Three sizes have been tested based on Breiman’s suggestions [6]:

* + , about half of the default value for classification;
  + , default value for classification ();
  + , twice the default for classification;

Table RF OOB error rate at different subset size of predictors randomly sampled at each split with the number of trees fixed at 1000.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Overall Error rate | ~Q1 (false negative) | Q1  (false positive) |
| 3 | 0.1311 | 0.0770 | 0.2933 |
| 5 | **0.1307** | **0.0760** | **0.2948** |
| 10 | 0.1311 | 0.0755 | 0.2979 |

The results are summarized in Table 1. It shows the error rates are all very similar at the three testing size, with achieving slightly smaller value, thereafter, is set as 5 for this study.

#### Variable importance

There are two variable importance measures in the RF algorithm, namely *permutation importance* and *Gini importance*.

* Permutation importance

This measure estimates the importance of a predictor by checking how much prediction error (misclassification rate for classification and MSE for regression) increases when OOB data for that predictor is permutated while all other predictors are unchanged. To get stable results, the RF was run 50 times on the Eagle Ford data set, and the average importance scores for each predictor were obtained and summarized in Figure 6. The top 5 important predictors are Ro calculated, Tmax, True vertical depth, GRI water filled porosity and S2.

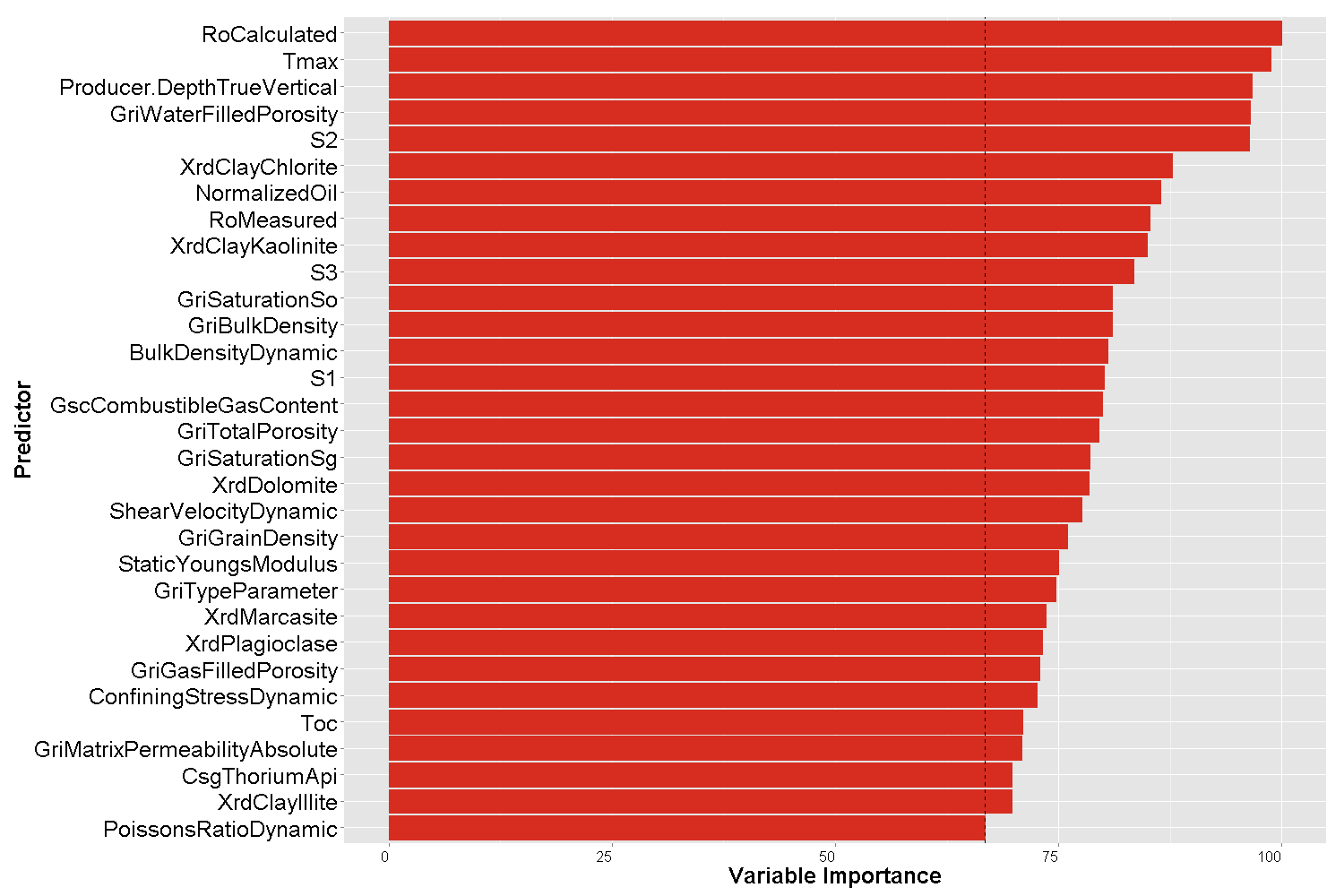


Figure A variable importance plot for the EF data set. Variable importance is computed using the permutation importance measure normalized by the maximum score.

* Gini importance

The Gini impurity index is defined as

(1)

where is the number of classes in target predictor, and is the ratio of this class. As described by Breiman and Cutler on their RF website [3], every time a split of a node is made on a variable, the gini impurity criterion for the two descendent nodes is less than the parent node. Adding up the gini decreases for each individual variable over all trees in the forest gives a fast variable importance that is often very consistent with the permutation importance measure.

Similar to the permutation importance study, the RF was run 50 times on the Eagle Ford data set, and the average gini importance scores for each predictor were obtained and summarized in Figure 7. The top 5 important predictors are True vertical depth, S3, Xrd Dolomite, GRI Matrix Permeability and GriTypeParameter.

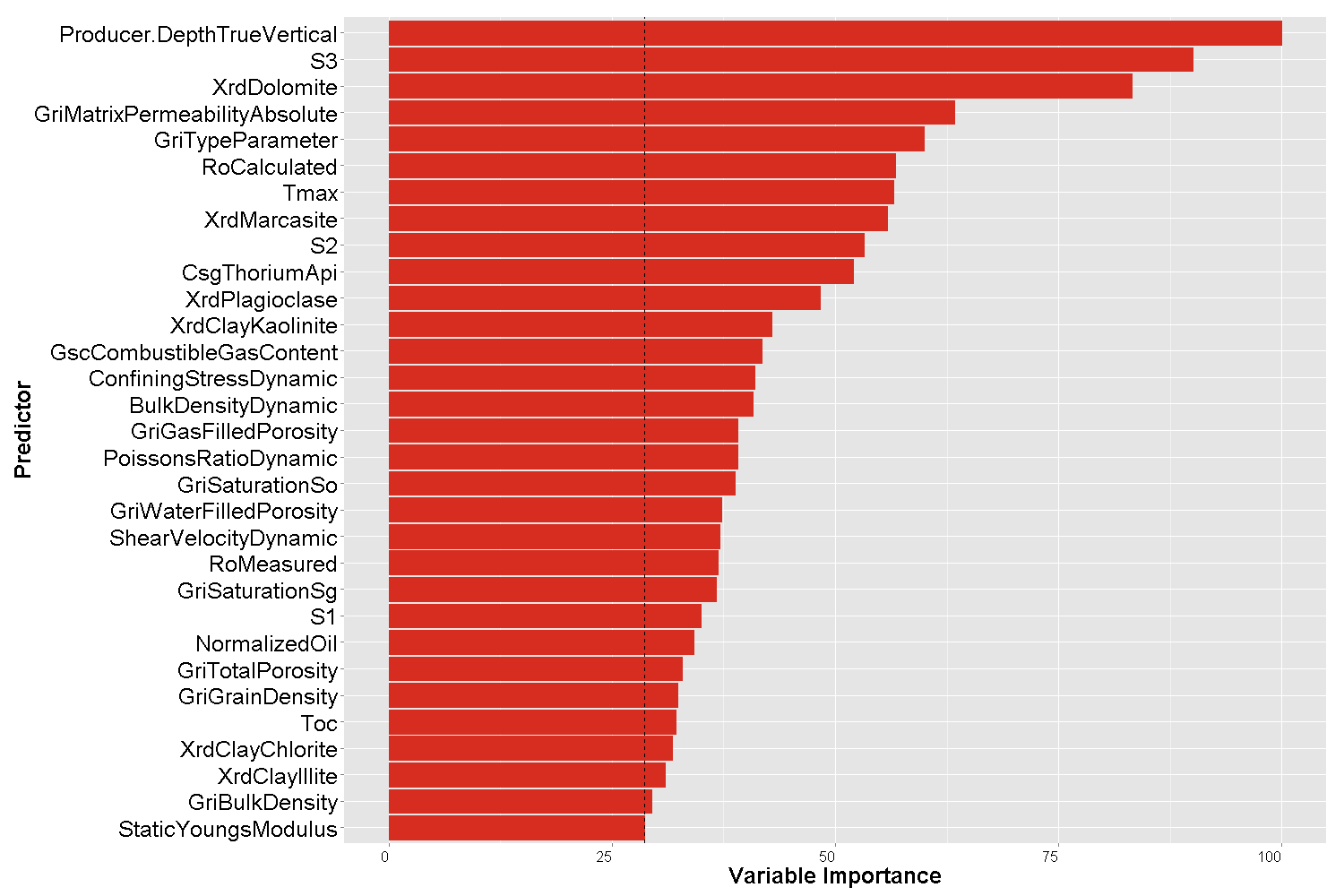


Figure A variable importance plot for the EF data set. Variable importance is computed using the gini importance measure normalized by the maximum score.

* Variable selection

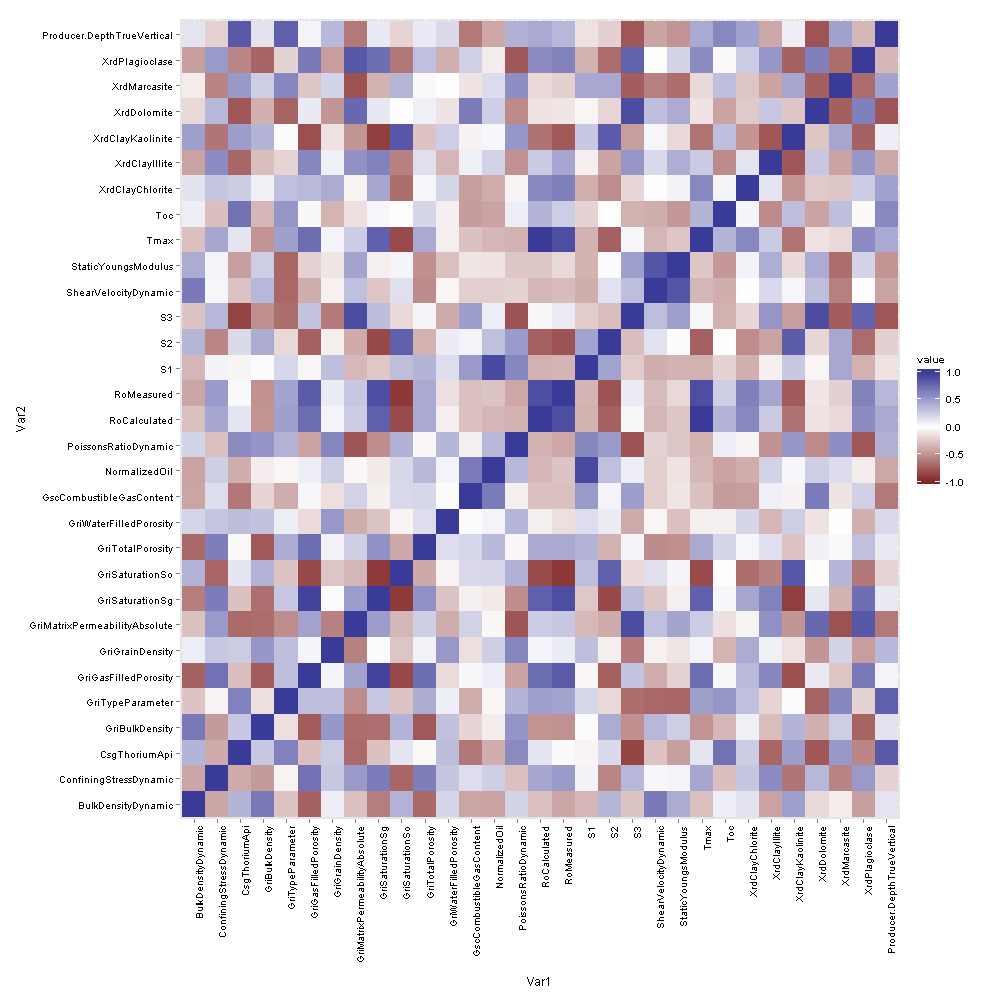


Figure 8 Correlation heat map among the 31 predictors under consideration.

As suggested by Carolin Strobl et al. [7], for predictors of the same type (all continuous or all unordered categorical with the same number of categories), permutation importance and gini importance are all adequate importance measures. However, cautious should be taken when there are highly correlated predictors exist; importance measure tends to give more weight to correlated predictors. Figure 8 shows the heat map of correlation among the 31 predictors under consideration, it shows some predictors are indeed highly correlated with correlation colored by either dark brown or blue. The *Party on* R package [7] tackles this problem by providing conditional variable importance measure, however, it is very expensive in terms of computational cost and has problem when the number of predictors are large.

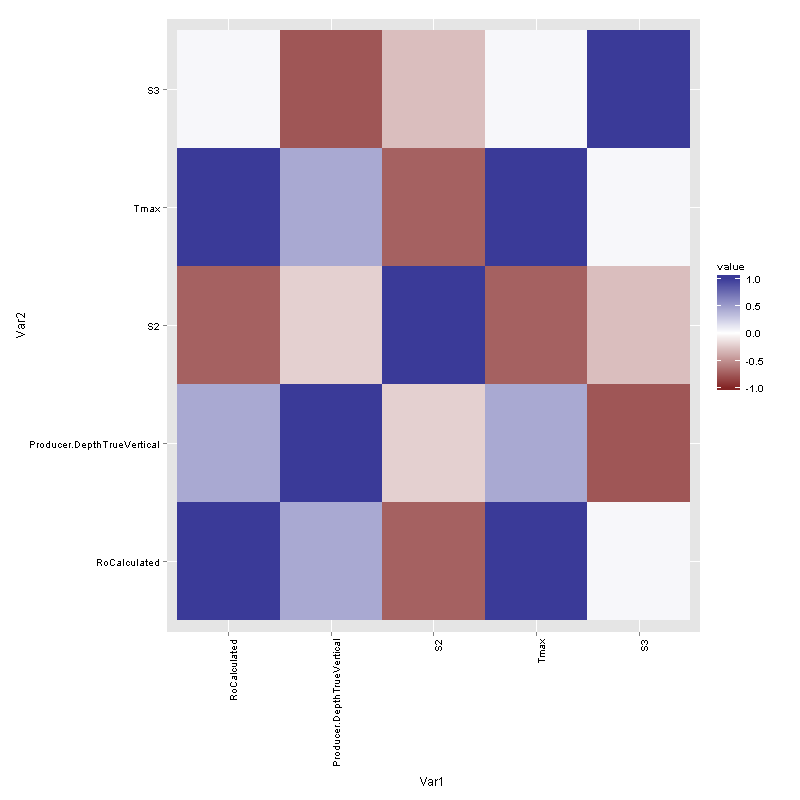


Figure 9 Correlation heat map for overlapped predictors among the top 10 important predictors identified by permutation measure and gini measure.

To identify a subset important predictors, a simple empirical approach is taken: by comparing the top 10 important predictors obtained from permutation importance and gini importance measures, 5 overlapped predictors are identified, True vertical depth, Tmax, Ro calculated, S2 and S3; Further, the correlation matrix is calculated for the 5 predictors, its corresponding heat map is shown in Figure 9. From the plot, Tmax and Ro calculated are identified as highly positive correlated predictors with correlation equal to 1.0. The final subset important predictors are determined by removing Ro calculated, {True vertical depth, Tmax, S2 and S3}. Its prediction performance will be checked in the next section.

#### Prediction

* Prediction based on all 31 predictors

RF algorithm was first applied to the EF data set with 5-fold cross-validation (CV) approach and the following setting: number of trees=1000; .

Table RF classification results in ratio with 5-fold cross-validation.

|  |  |  |
| --- | --- | --- |
| True Positive | True Negative | Accuracy |
| 0.69 | 0.92 | 0.86 |

As indicated by the numeric value of Table 2 and Table 3, the overall accuracy, (ture positive + ture negative)/total number test object, is about 86%, however, accuracy is not a reliable metric for this study given the number of samples in each class (Q1 vs. ~Q1) is unbalanced. *More attention should be paid to true positive and true negative rate***.** Please note, the number of producers assigned to Q1 is not constrained to 25% of the total producers.

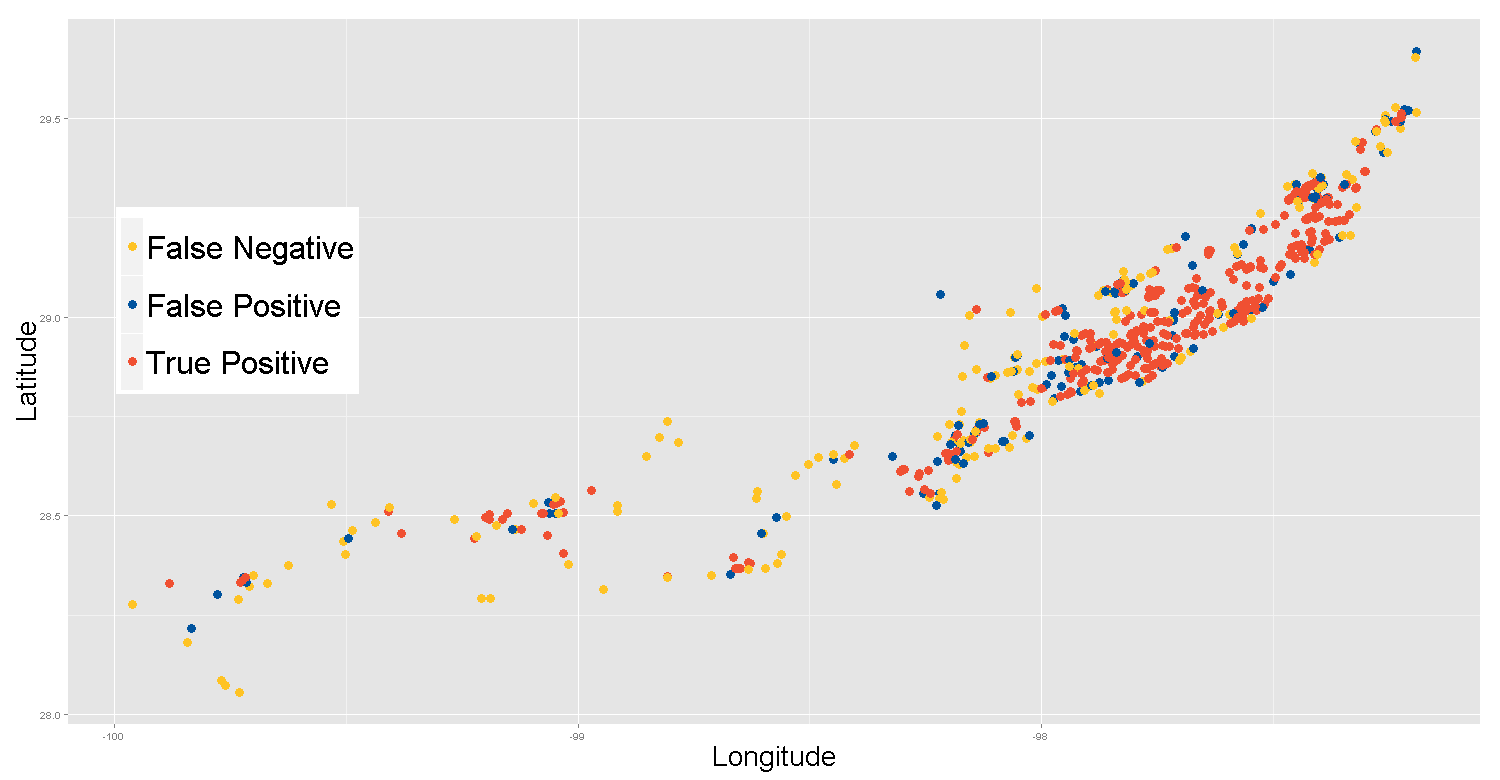


Figure 10 Predicted sweetspots by RF classification approach using all predictors.

Figure 10 shows the sweetspots predicted by RF classification approach using all the predictors, the red dots represent true positive, i.e. those true top quartile producers identified by the predictive model; the blue dots represents false positive, i.e. those non-top quartile producers but identified as top quartile by the predictive model; yellow dots represents false negative, i.e. those true top quartile producers but missed by the predictive model.

Table RF classification results in number of wells with 5-fold cross-validation.

|  |  |  |
| --- | --- | --- |
| True Positive | False Positive | False Negative |
| 451 | 153 | 207 |

Next, to study the effect of percentage of training data on the performance of predictive model, RF classification approach was applied to the EF data set with different fold CV. Please note, when the training percentage is small, reverse CV technique is used. The results are summarized in Table 4 and Figure 11.

Table RF classification results in ratio with different fold cross-validation.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CV | Training (%) | # of Wells in Training | True Positive | True Negative | Accuracy |
| Reverse 10-fold | 10 | 263 | 0.56 | 0.91 | 0.82 |
| Reverse 5-fold | 20 | 526 | 0.62 | 0.92 | 0.84 |
| Reverse 3-fold | 33 | 868 | 0.64 | 0.92 | 0.85 |
| 2-fold | 50 | 1316 | 0.67 | 0.92 | 0.85 |
| 3-fold | 67 | 1763 | 0.70 | 0.91 | 0.86 |
| 5-fold | 80 | 2105 | 0.71 | 0.92 | 0.87 |
| 10-fold | 90 | 2368 | 0.70 | 0.92 | 0.87 |

Detail examining Table 4 and Figure 11 reveals that as the percentage of training data increases, the true negative rate does not change much, oscillating between 0.91 and 0.92, which is understandable given more samples are in ~Q1 class; however as the percentage of training data increases, the true positive rate does change, it dramatically increases when the percentage of training data changes from 10% to 20%, and then gradually increases (almost linear) as the training percentage changes from 20% to 70%, after that, it reach a plateau area. The possible conclusions for true positive rate from these observations are (a) the gain from increasing training data set percentage in the range of 10% to 20% is significant; (b) the gain from increasing training data set percentage in the range of 20% to 70% is linear; (c) not much gained after the training data set percentage reaches 70%. One explanation for (c) could be due to the unbalance samples in each class (Q1 vs. ~Q1), when the training data set above 70%, not much producers from Q1 are in the test data set, which cause the true positive rate saturated.

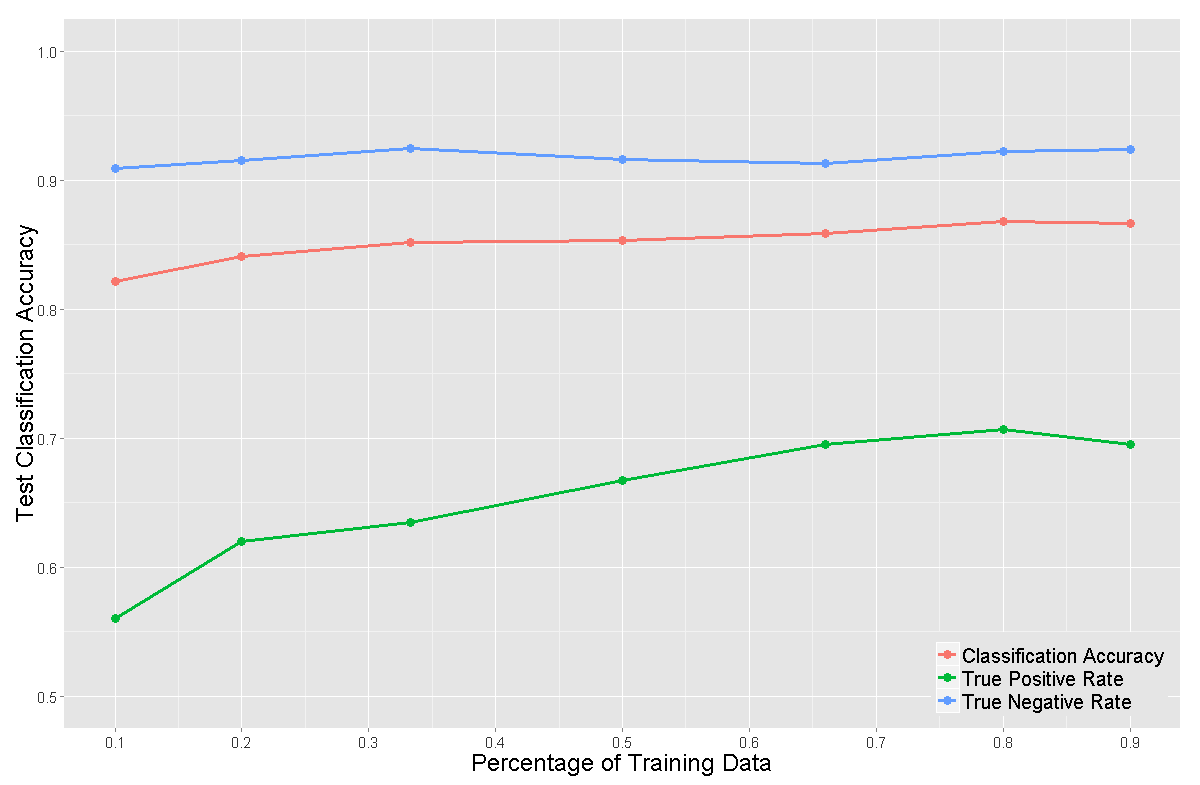


Figure 11 RF classification results in ratio at different percentage of training data.

* Prediction based on select predictors

In section 4.1.1.2, the predictors have been ranked based on two importance measures permutation importance and Gini importance. Here, the performance of RF classification approach will be first examined by using only the top important predictors with 5-fold cross-validation. The results are summarized in Figure 12 and Figure 13 for permutation and gini importance measures respectively. The plots showed that, when the number of predictors is small, generally as more important predictors added to the model, the true positive rate, true negative rate and accuracy increases, however, when K greater than 6 and 5 for permutation and gini importance measures, the performance reaches a plateau areas, in another words, only top 6 predictors for permutation measure and 5 for gini importance measure are needed to get a predictive model as good as using more predictors.

Second, the performance of RF classification approach using only a subset important predictors, {True vertical depth, Tmax, S2 and S3}, identified in section 4.1.1.2 is studied with 5-fold cross-validation. The results are summarized in Table 5, compared with the results in Table 2 using all the 31 predictors, the performance for the 4 predictors model is even better, which proves the simple empirical variable selection approach indeed works for this study in identifying a parsimonious model with good prediction performance.

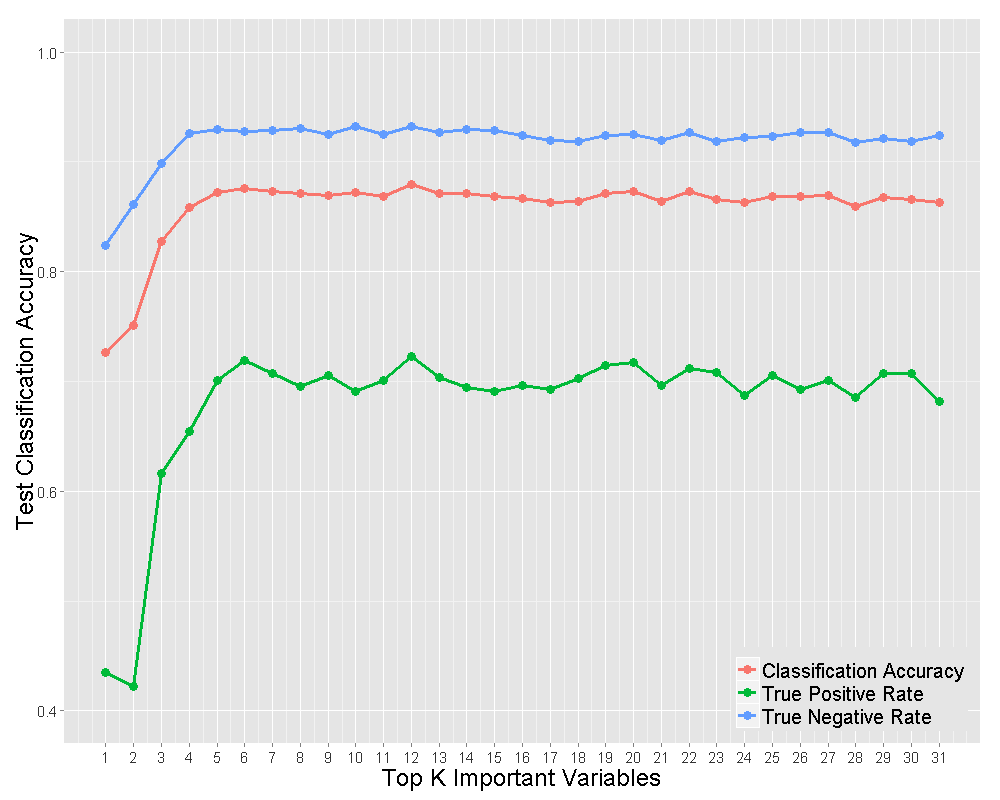


Figure 12 RF classification results in ratio with top K important predictors (permutation measure).

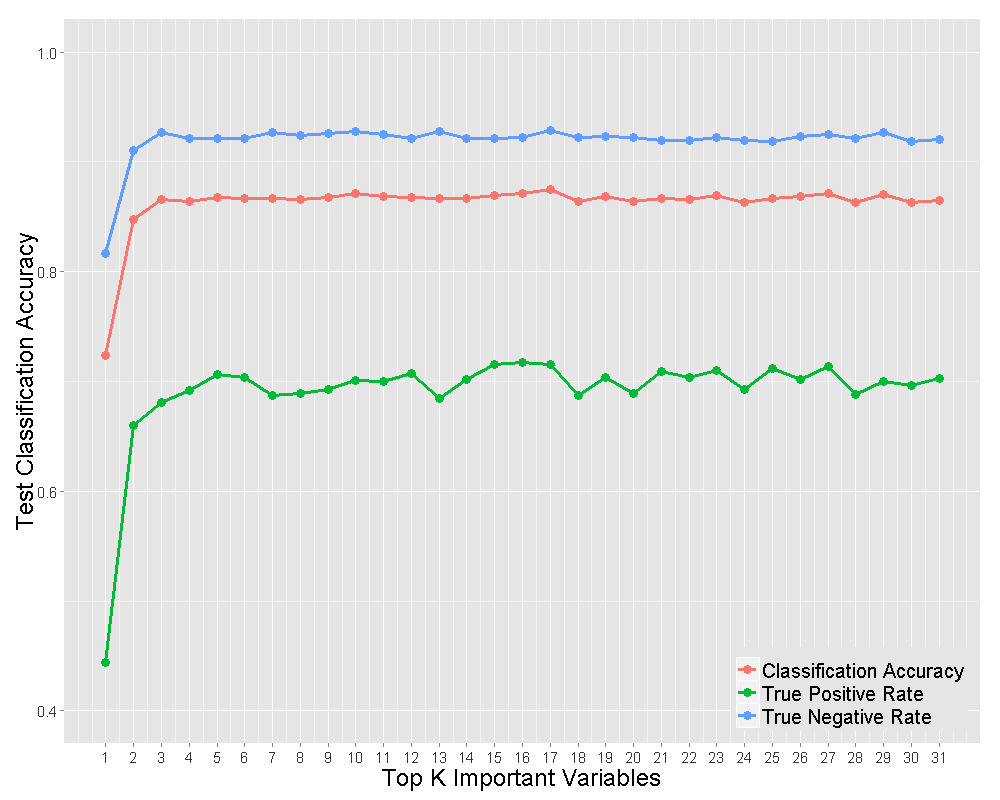


Figure 13 RF classification results in ratio with top K important predictors (gini measure).

Table RF classification results in ratio with 5-fold cross-validation using only four important predictors: true vertical depth, Tmax, S2 and S3.

|  |  |  |
| --- | --- | --- |
| True Positive | True Negative | Accuracy |
| 0.71 | 0.93 | 0.87 |

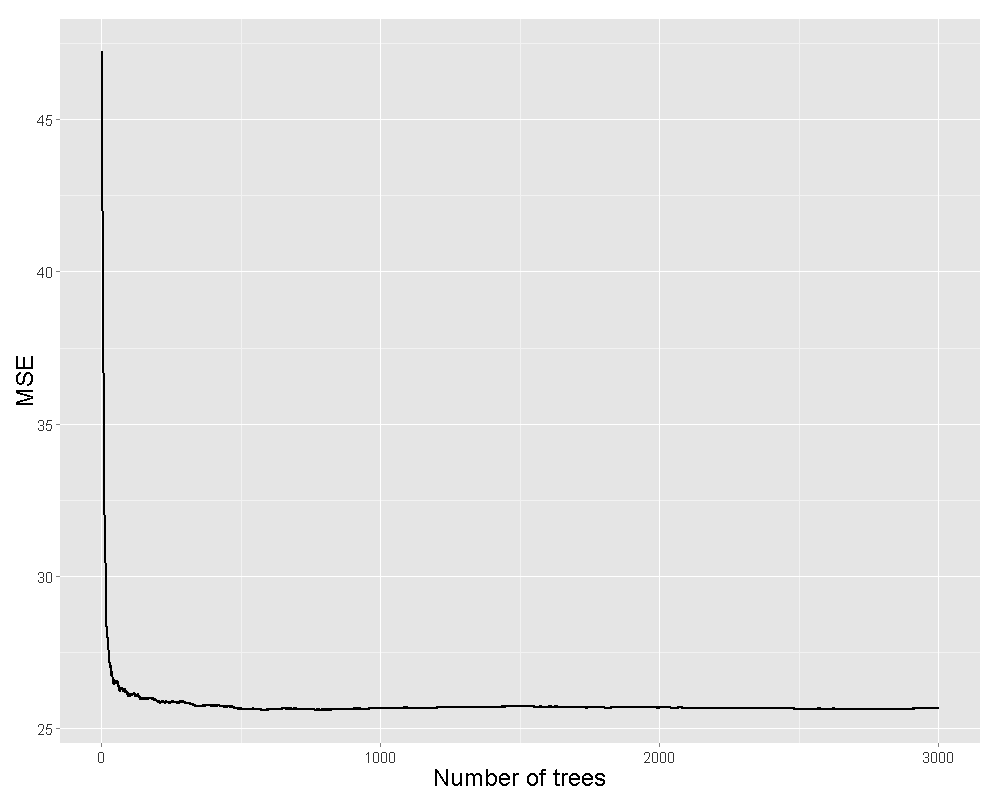
### Regression approach

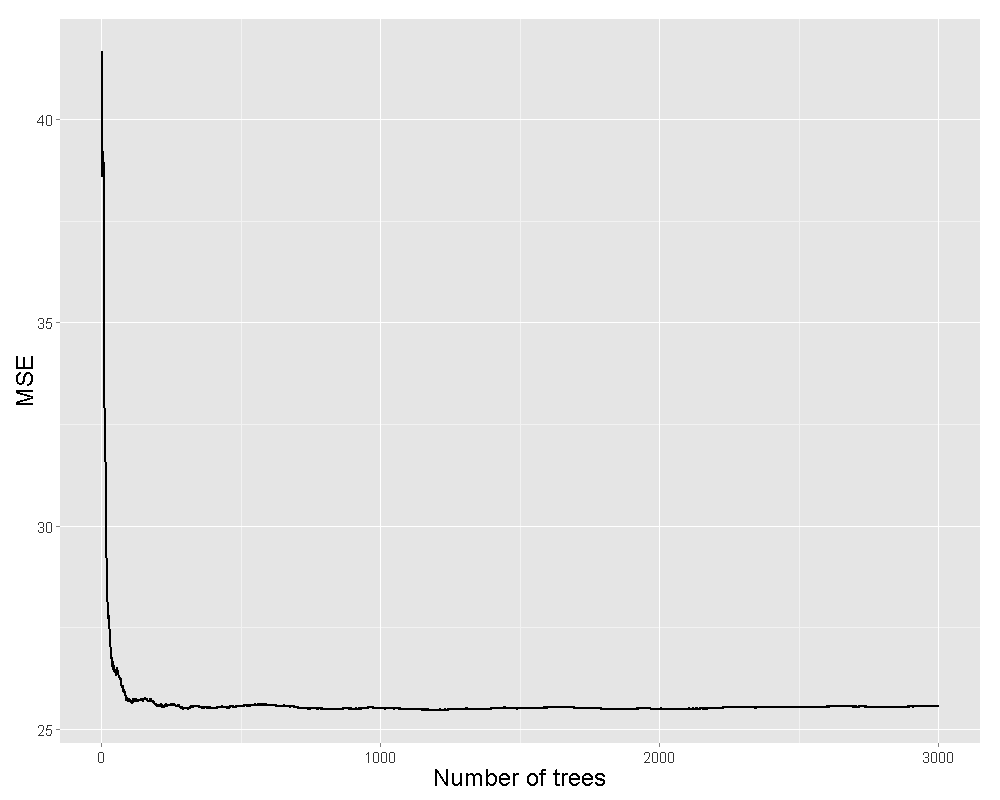
#### Parameter tuning

Similar to the classification approach, OOB mean square error (MSE) is used for RF regression parameter tuning.

* Number of trees

Figure 14 shows the OOB MSE with different number of trees at various subset sizes of predictors randomly sampled at each split for the Eagle Ford dataset. From the plots, the MSE gets stable for all the three scenarios once the number of trees is greater or equal to 1000. After 1000, even more trees are added, the OOB MSE does not decrease any more, which means RF regression does not suffer from overfit problem. The number of trees will be fixed at 1000 for the rest of this study.





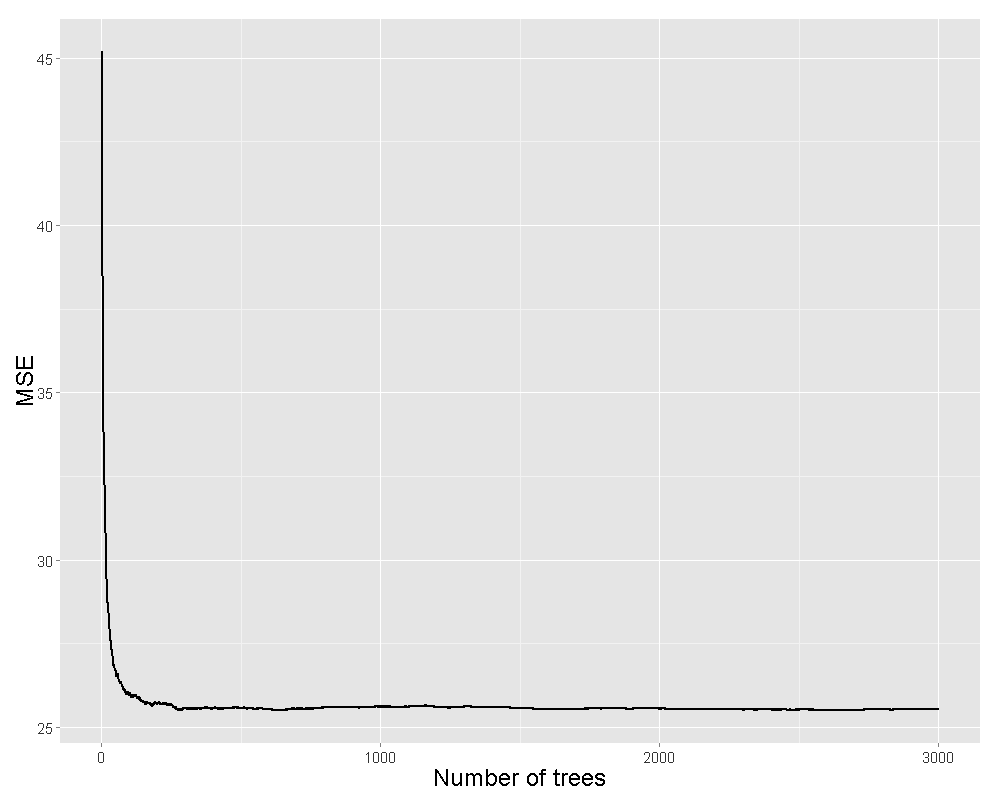


Figure RF OOB MSE with different number of trees at various subset sizes, , of predictors randomly sampled at each split (a) (b) (c).

* Subset size of predictors randomly sampled at each split

With the number of trees fixed at 1000, the OOB MSE is calculated at different subset size of predictors randomly sampled at each split  **.**  Ten sizes have been tested from 2 to 20 incrementing by 2. Note the default value for RF regression , which is equal to total number of predictors divided by 3.

The results are summarized in Table 6. It shows the MSE achieves the minimum value with in the testing range. Thereafter, is set as 12 for RF regression study.

Table RF OOB MSE at different subset size of predictors randomly sampled at each split with the number of trees fixed at 1000.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 |
| MSE | 25.96 | 25.74 | 25.77 | 25.55 | 25.54 | **25.52** | 25.54 | 25.60 | 25.57 | 25.56 |

#### Variable importance

Permutation-based MSE reduction has been adopted as the state-of-the-art approach for the variable importance metric of RF regression [8]. To get stable results, the RF was run 50 times on the Eagle Ford data set, and the average importance scores for each predictor were obtained and

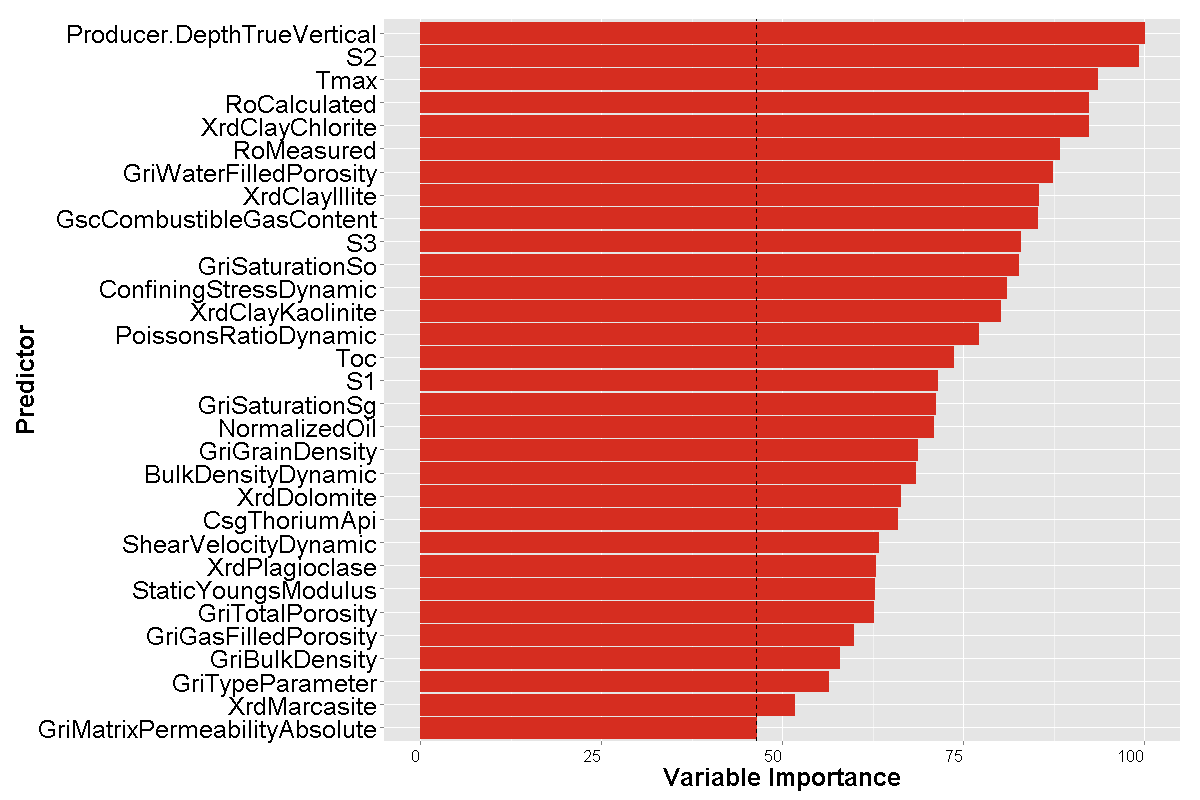


Figure A variable importance plot for the EF data set. Variable importance is computed using the permutation-based MSE reduction measure normalized by the maximum score.

summarized in . The top 7 important predictors are True vertical depth, S2, Tmax, Ro calculated, Chlorite in clay(X-ray diffraction), Ro measured and GRI water filled porosity. As suggestion by , Tmax and Ro calculated are highly correlated with correlation equal to 1.0. Therefore, after removing Ro calculated, a subset of important predictors is obtained {True vertical depth, S2, Tmax, Chlorite in clay(X-ray diffraction), Ro measured, GRI water filled porosity}. This subset is very similar to the one identified in the classification approach, {True vertical depth, Tmax, S2 and S3}, with three overlapped predictors, {*True vertical depth, S2 and Tmax*} which are worth of further investigation.

#### Prediction

* Prediction based on all 31 predictors

RF regression algorithm was first applied to the EF data set with 5-fold cross-validation approach and the following setting: number of trees=1000; . A model that is able to predict the first 12-month production was obtained firstly; second, the producers will be ranked according to their predicted values; third, top quartile producers will be identified as “sweetspot” based on the ranking from step 2. Due to the predict-ranking procedure, the number of false positive will always be equal to the number of false negative.

Table RF regression results in number of wells with 5-fold cross-validation.

|  |  |  |
| --- | --- | --- |
| True Positive | False Positive | False Negative |
| 484 | 174 | 174 |

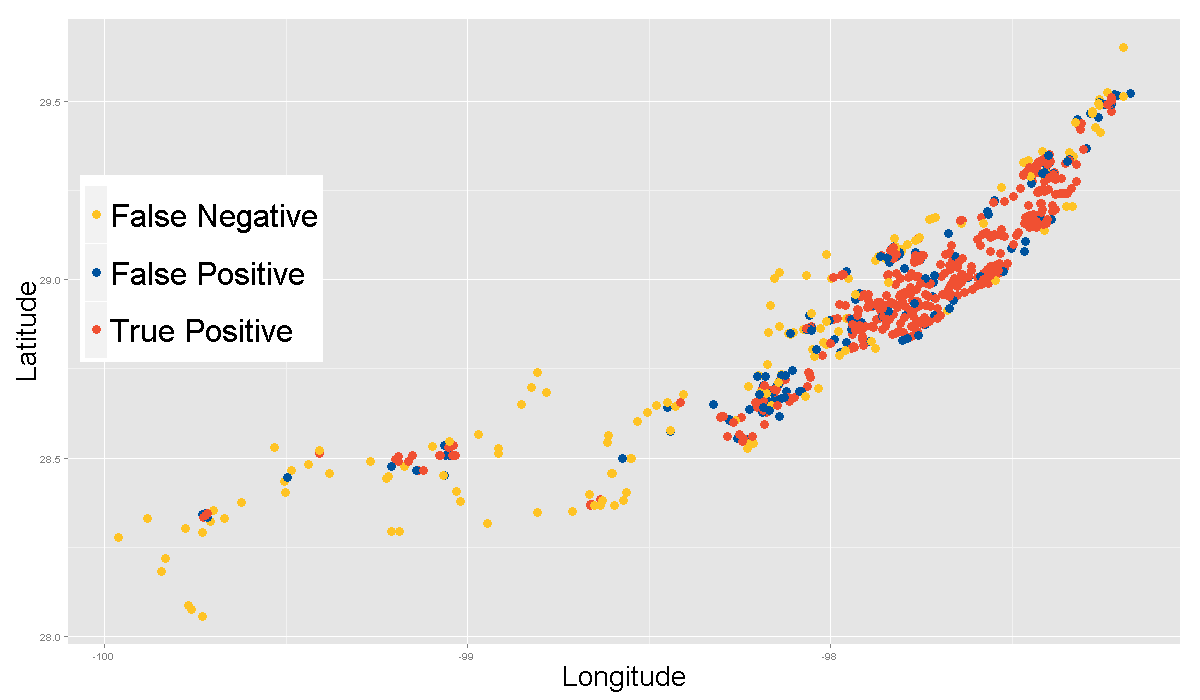


Figure 16 Predicted sweetspots by RF regression approach using all predictors

The results are summarized in Table 7 with overall MSE equal to 26.9. Compared with the classification approach’s results (Table 3), both the true positive number of wells increases and the false negative number of wells decreases by 33, but the false positive number of wells increases by 21. Figure 16 shows the sweetspots predicted by RF regression approach using all the predictors, the color setting is the same as the RF classification approach (Figure 10), and the pattern of these two approaches is very similar.

Table RF regression results with different fold cross-validation.

|  |  |  |  |
| --- | --- | --- | --- |
| CV | Training (%) | # of Wells in Training | RMSE |
| Reverse 10-fold | 10 | 263 | 6.73±0.23 |
| Reverse 5-fold | 20 | 526 | 6.06±0.20 |
| Reverse 3-fold | 33 | 868 | 5.64±0.10 |
| 2-fold | 50 | 1316 | 5.46±0.01 |
| 3-fold | 67 | 1763 | 5.31±0.32 |
| 5-fold | 80 | 2105 | 5.09±0.28 |
| 10-fold | 90 | 2368 | 5.07±0.38 |

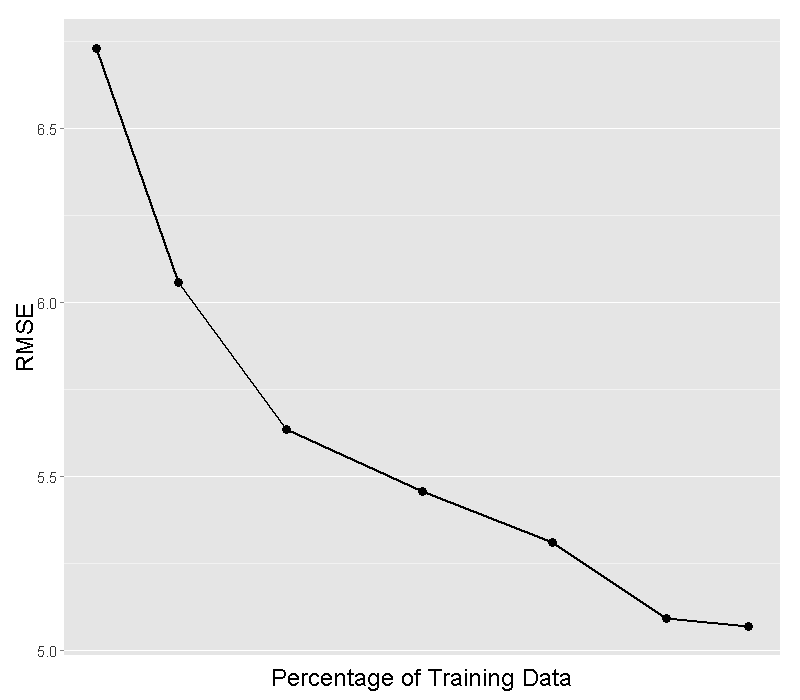


Figure 17 RF regression results at different percentage of training data.

Next, to study the effect of percentage of training data on the performance of predictive model, RF regression approach was applied to the EF data set with different fold CV. Please note, when the training percentage is small, reverse CV technique is used. The results are summarized in Table 8 and Figure 17. Detail examining reveals that as the percentage of training data increases, the MSE dramatically decreases in the range of 10% - 80%; after 80%, it reaches a plateau area.

* Prediction based on select predictors

Again, the performance of RF regression approach will be first examined by using only the top important predictors with 5-fold cross-validation. The results are summarized in Figure 18 for permutation-based MSE reduction importance measure.

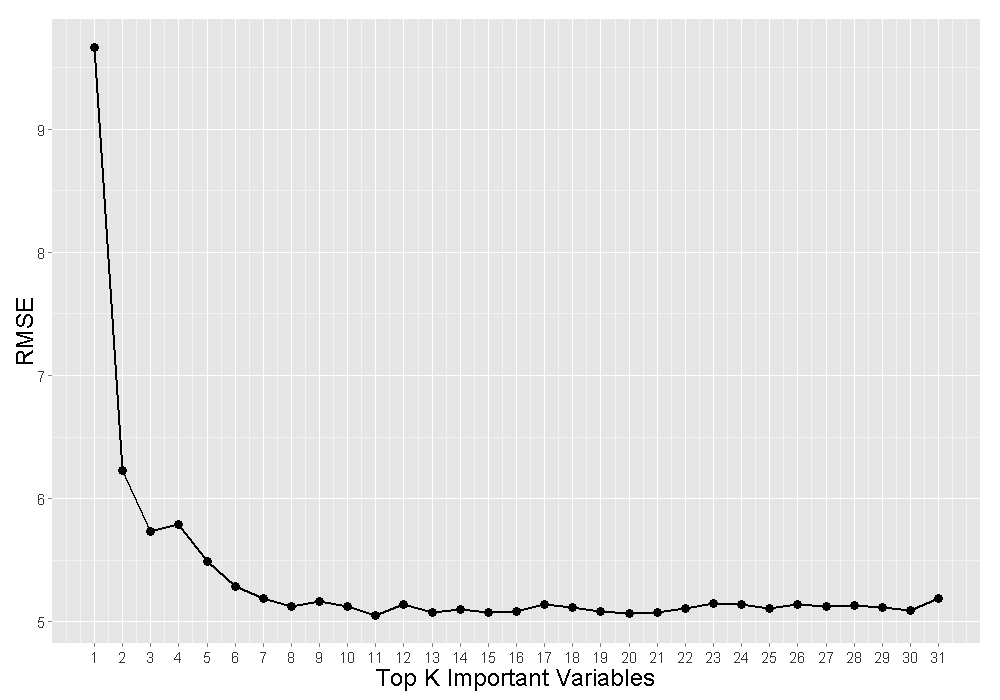


Figure RF regression results with top K important predictors (permutation-based MSE measure).

The plot showed that, when the number of predictors is small, generally as more important predictors added to the model, the MSE reduces significantly, however, when K greater than 7, the performance reaches a plateau areas and MSE is oscillating within a small range, in another words, only top 7 predictors for permutation measure are needed to get a predictive model as good as using more predictors.

Second, the performance of RF regression approach using only a subset important predictors, {True vertical depth, S2, Tmax, Chlorite in clay(X-ray diffraction), Ro measured, GRI water filled porosity} identified in section is studied with 5-fold cross-validation. The results are summarized in with overall MSE equal to 26.3, compared with the results in using all the 31 predictors, the performance for the 6 predictors model is even better, with true positive number increases to 490 from 484; false positive and false negative number decreases from 174 to 168.

Table RF regression results with 5-fold cross-validation using only 6 important predictors: true vertical depth, S2, Tmax, Chlorite in clay(X-ray diffraction), Ro measured, GRI water filled porosity.

|  |  |  |
| --- | --- | --- |
| True Positive | False Positive | False Negative |
| 490 | 168 | 168 |

For the most interesting three important predictors, {*True vertical depth, S2 and Tmax*}, their partial dependence plots are shown in . These plots are graphical visualizations of the marginal effect of the given predictors with the average effect of all other predictors from the model. As indicated by these plots, peak productivity is achieved at certain range of each predictor:

* + Tmax ~ (460, 472);
  + True vertical depth ~ (12000, 12800);
  + S2 ~ around 5.

These ranges need to be discussed further with domain experts and may guide sweetspots identification as well as completion engineering parameters optimization.

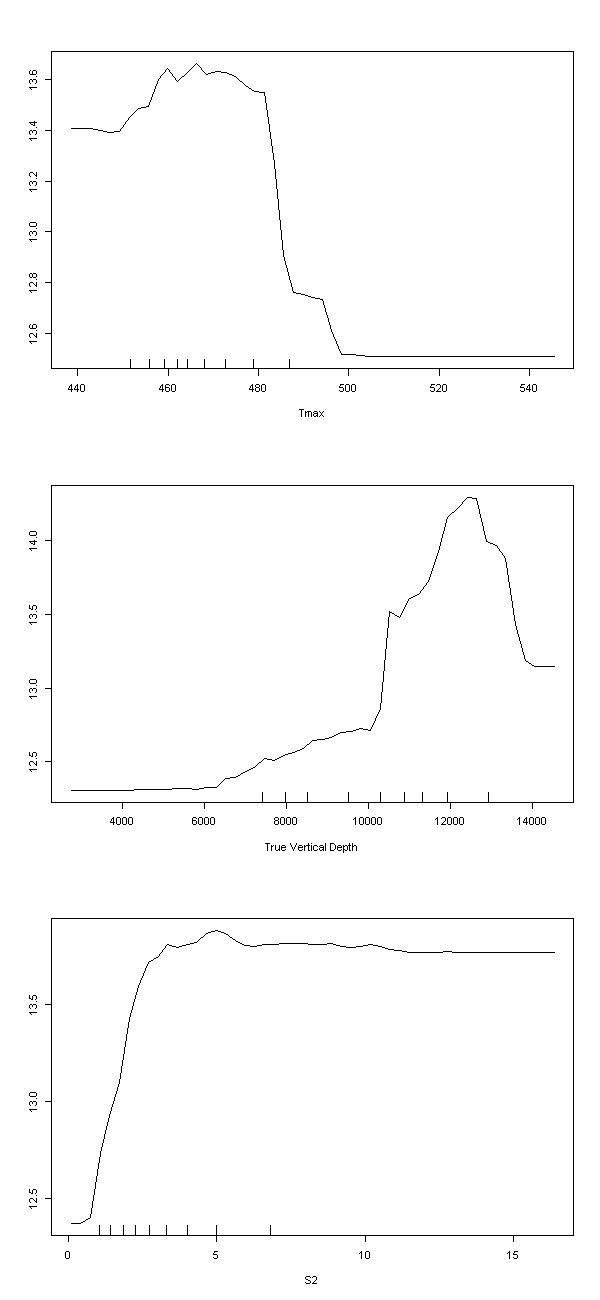


Figure Partial dependence plot for three selected important predictors.

## BART

As mentioned in section 3.3, the main reason that BART is introduced into this study is due to its capability to compute uncertainty estimates via quantiles of the posterior samples for regression problem. Therefore, in this section, the use of BART will focus on regression approach only.

### Parameter tuning

R package bartMachine [9] provides some useful functions to help turning the parameters, the *rmse\_by\_num\_trees( )* and *bart\_machine\_cv( )* are used here for parameter tuning.

* Number of trees

Figure 20 shows the out-of-sample root mean square error (RMSE) with different number of trees for the Eagle Ford dataset. From the plots, the minima of RMSE is located at number of trees equal to 150, after that, the RMSE increases as more trees are added. For this study, the number of trees will be fixed at 150.

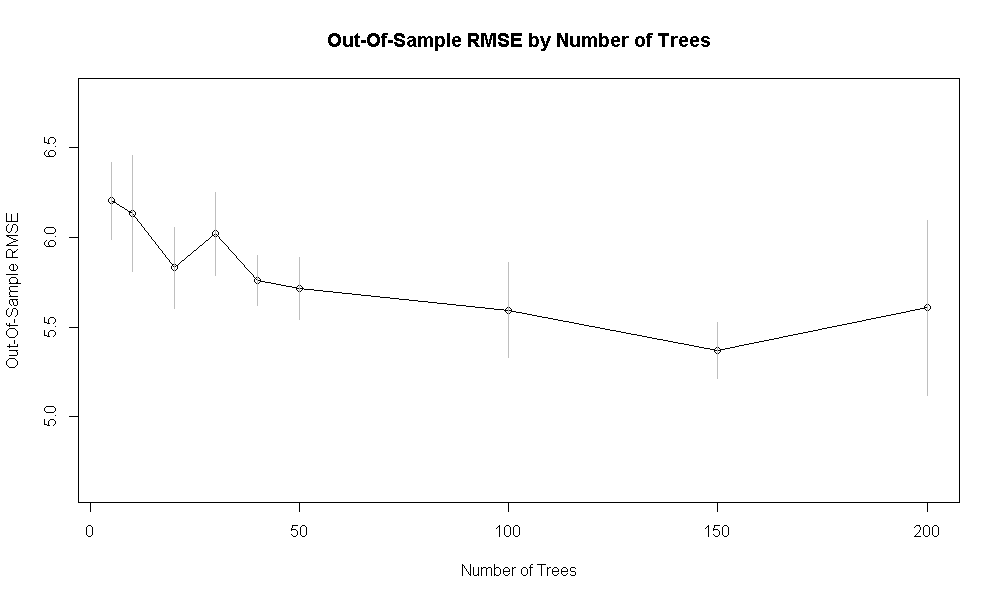


Figure Out-of-sample RMSE with different number of trees.

* Hyperparameters

*bart\_machine\_cv( )* can do grid-searching over a set of hyperparameter combination using cross-validation approach, with the default grid search setting, the “winning” model with lowest out-of-sample RMSE over a 5-fold cross-validation is *k=3; nu=3; q=0.9.* For more detail, please refer to [9].

### Variable importance

bartMachine implements the variable selection procedure developed in Bleih et al. (2014) [10], which makes use of the “variable inclusion proportions”, i.e. the proportion of times each predictors is chosen as a splitting rule divided by the total number of splitting rules appearing in the model [5]. With the local threshold selection rule, the predictors ranking is summarized in Figure 21. The results are pretty consistent with RF regression approach, {*True vertical depth, S2 and Tmax*} are among the top 5 important variables.

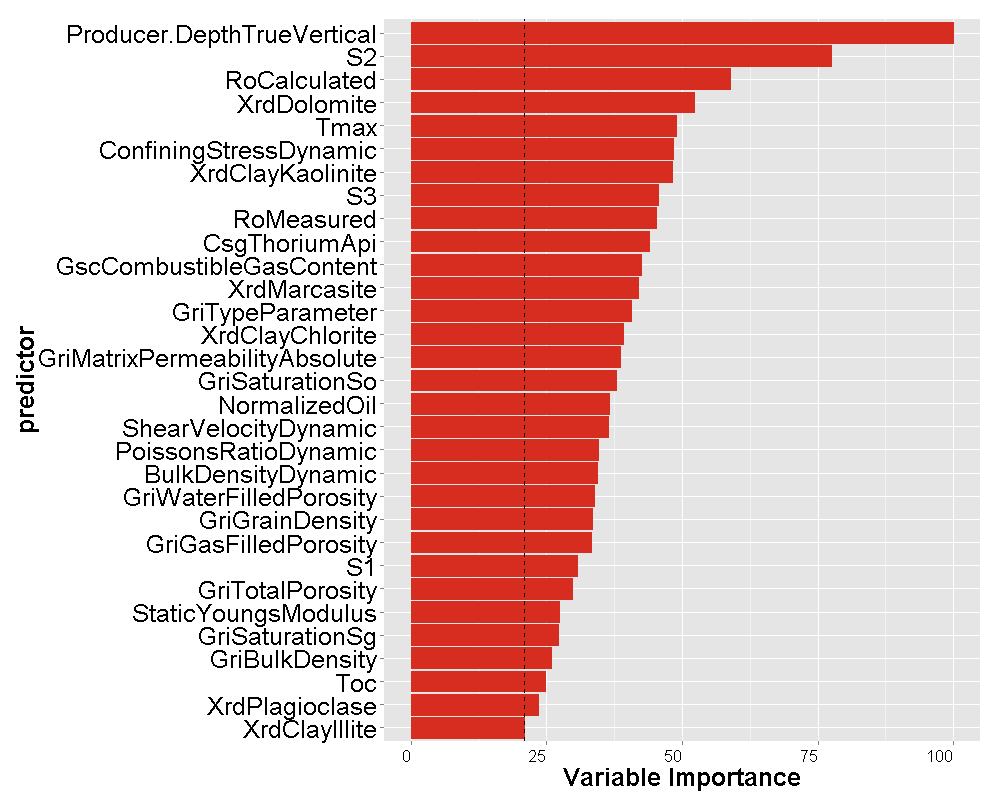


Figure A variable importance plot for the EF data set. Variable importance is computed using the variable inclusion proportions measure normalized by the maximum score.

### Prediction

As a method under Bayesian frame work, the computational cost of BART is very expensive; therefore, our study is limited to the prediction performance of the model with all 31 predictors.

BART regression algorithm was applied to the EF data set with 5-fold cross-validation approach and the following setting: number of trees=150, *k=3; nu=3; q=0.9*. Follow the regression predict-ranking procedure, the BART results are summarized in Table 10 with overall MSE equal to 30.0. It is much worse than RF regression results (see Table 7).

Table BART regression results in number of wells with 5-fold cross-validation.

|  |  |  |
| --- | --- | --- |
| True Positive | False Positive | False Negative |
| 465 | 193 | 193 |

Next, to study the effect of percentage of training data on the performance of predictive model, BART regression approach was applied to the EF data set with different fold CV as the RF regression approach. The results are summarized in Figure 22. Detail examining reveals that as the percentage of training data increases, the MSE dramatically decreases in the whole range of study 10% - 90%.

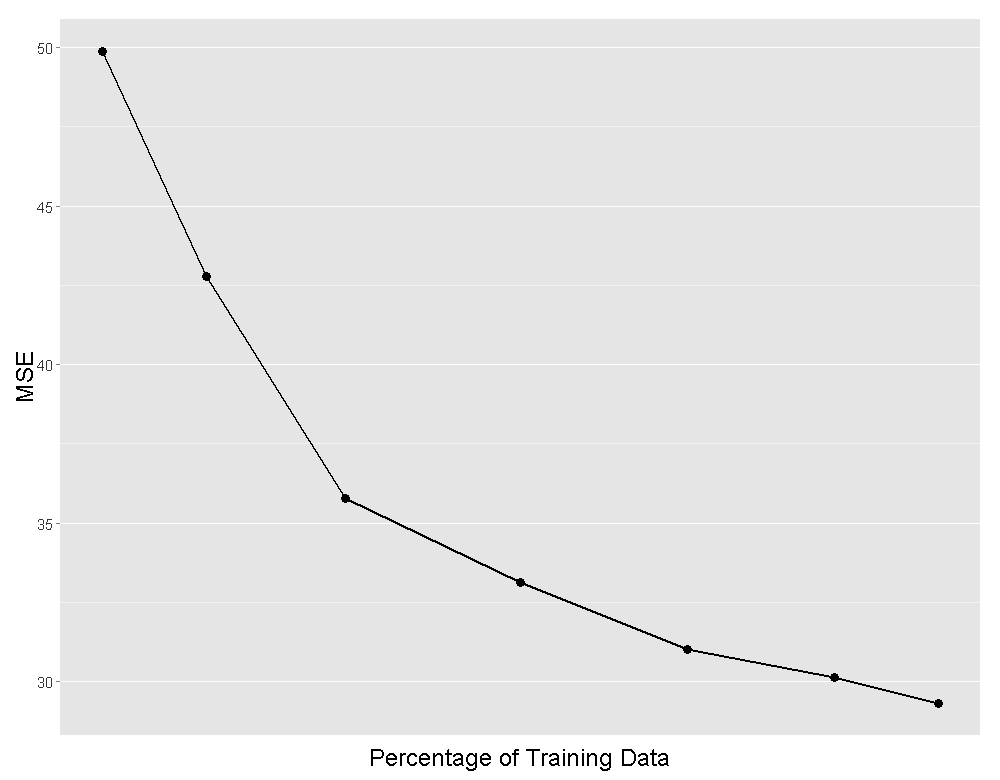


Figure BART regression results at different percentage of training data.

# Discussion and future work

This study investigated the performance of three machine learning approaches in building a predictive model for shale oil plays sweetspots identification, namely RF classification approach, RF regression approach and BART. Their results comparison in terms of true positive number, the number of correctly identified sweetspots, are summarized in Table 11. The random draw and rule based methods are list for comparison purpose only. For random draw model, it views the correct number of sweetspots identified from pure guessing follows a hypergeometric distribution, and the expectation value is 165 for top quartile identification; Rule based method is an empirical approach, defining a score system with a few geological parameters identified by the domain experts.

Table Methods comparisons in terms of true positive number.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Random Draw | Rule Based | RF Classification | RF Regression | BART Regression |
| 165 | 226 | 451 | 484 | 465 |

Table 11 reveals that the performance of these methods in terms of true positive number ranks as the following:

RF Reg. > BART Reg. > RF Classification > Rule Based > Random Draw.

However, one thing to remember that BART as a Bayesian method could offer uncertainty estimates via quantiles of the posterior samples.

For this work, there is still much space for improvement and debate, are first 12-month production only normalized by lateral length a good proxy for EUR? Could we include more completion engineering parameters to improve the normalization? The well logs data are omit in this study, how to use them and integrate with the current frame work? How to interpret the partial dependence plots of the important variables and use their ranges for high productivity to guide drilling and completion process....

Finally, the authors want to point out, this report only covers the stage II study of **Phase I**, its results may need to be reevaluated when integrates with stage I’s findings. Especially, training wells and test well may need to be grouped by the distance to core wells, such that there is no leaking information from training set to test set.

# Acknowledgements

The authors thank …

# References

[] Breiman L. (2001). Random Forests, Machine Learning, **45**(1), 5-32.

[] James, G., Witten, D., Hastie, T., Tibshirani, R (2013). An Introduction to Statistical Learning with Applications in R, Springer.

[] <https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm>

[] Chipman H, George E. and McCulloch R (2010). BART: Bayesian Additive Regression Trees, The Annals of Applied Statistics, **4**(1), 266-298.

[] Kapelner, A and Bleich, J. (2014). bartMachine: A Powerful Tool for Machine Learning, ArXiv e-prints.

[6] Liaw, A. & Wiener, M (2002). Classification and Regression by randomForest, R News,   
Vol. 2/3, 18.

[7] Strobl, C., Hothorn, T. and Zeileis A. (2009). Party on!, Thr R Journal, Vol. 1/2, 14.

[8] Gromping U. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest, The American Statistician, Vol. 63, No.4., 308.

[9] <http://CRAN.R-project.org/package=bartMachine>

[10] Bleich J, Kapelner A, Jensen S, George E (2014). Variable Selection for BART: An Appli-

cation to Gene Regulation. Annals of Applied Statistics, 8(3), 1750.

# Bibliographic information

|  |  |
| --- | --- |
| Classification | Restricted |
| Report Number | SR.xx.xxxx |
| Title | SweetSpots for UNC Plays Through Machine Learning Method |
| Author(s) | M. Wu (PTD/TASE)  L. Lu (PTI/RP)  J. Limbeck (PTI/RP)  D. Hohl (PTI/RP) |
| Keywords |  |
| Date of Issue | June 2015 |
| Period of Work | Jan 2015 - May 2015 |
| US Export Control | EAR 99 |
| WBSE Code | ZZPT/022701/010101 |
| Reviewed by |  |
| Approved by | P. Jonathan (PTD/TASE) |
| Sponsoring Company / Customer | D. Hohl (PTI/RP) |
| Issuing Company | Shell Global Solutions (US) Inc.  Shell Technology Center Houston  3333 Highway 6 South  Houston TX 77082-3101  USA |
|  |  |

# Report distribution

|  |  |
| --- | --- |
| **Electronic distribution (PDF)** | |
| *Name, Company, Ref. Ind.* | *PDF* |
| PT Information Services, PTT/TIKE, [PT-Information-Services@Shell.com](mailto:PT-Information-Services@Shell.com) | 1 |
| Wu, Mingqi GSUSI-PTD/TASE | 1 |
|  |  |
|  |  |
| Jonathan, Philip GSUK-PTD/TASE | 1 |
|  |  |
| Gibbs, Richard A GSUSI-PTD/TASE | 1 |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
| Randell, David GSUK-PTD/TASE | 1 |
| Theis, Winfried GSNL-PTD/TASE | 1 |
| Paleja, Rakesh J GSUK-PTD/TASE | 1 |
| Bierman, Stijn M GSNL-PTD/TASE | 1 |
| Jones, Wayne R GSUK-PTD/TASE | 1 |
| Demel, Seth S GSUSI-PTD/TASE | 1 |
| Kettl, Stena E GSUSI-PTD/TASE | 1 |
| Erdmann, Tashi P GSNL-PTD/TASE | 1 |
| Gonzalez Martinez, Jose Maria M GSNL-PTD/TASE | 1 |
| de Noord, Onno E GSNL-PTD/TASE | 1 |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

The copyright of this document is vested in Shell Global Solutions (US) Inc., Houston, Texas, USA. All rights reserved.  
Neither the whole nor any part of this document may be reproduced, stored in any retrieval system or transmitted in any form or by any means (electronic, mechanical, reprographic, recording or otherwise) without the prior written consent of the copyright owner. Shell Global Solutions is a trading style used by a network of technology companies of the Shell Group.