|  |  |
| --- | --- |
| Please fill in the name of the event you are preparing this manuscript for. | SPE Asia Pacific Oil & Gas Conference and Exhibition 2022 |
| Please fill in your 6-digit SPE manuscript number. | SPE-201xxx-MS |
| Please fill in your manuscript title. | Applying Data Analytics & Machine Learning Methods for Recovery Factor Prediction and Uncertainty Modelling |
| Please fill in your author name(s) and company affiliation.   |  |  |  |  | | --- | --- | --- | --- | | Given Name | Middle Name | Surname | Company | | Munish |  | Kumar |  | | Kanna |  | Swaminathan |  | | Aizat |  | Rusli |  | | Abel |  | Thomas-Hy |  | |  |  |  |  | |  |  |  |  | |  |  |  |  | | |
| This template is provided to give authors a basic shell for preparing your manuscript for submittal to an SPE meeting or event. Styles have been included (Head1, Head2, Para, FigCaption, etc.) to give you an idea of how your finalized paper will look before it is published by SPE. All manuscripts submitted to SPE will be extracted from this template and tagged into an XML format; SPE’s standardized styles and fonts will be used when laying out the final manuscript. Links will be added to your manuscript for references, tables, and equations. Figures and tables should be placed directly after the first paragraph they are mentioned in. The technical content of your paper WILL NOT be changed. Please start your manuscript below. | |

Abstract

The estimation of recoverable hydrocarbons, or field recovery factor (RF), is a critical variable for Oil and Gas (O&G) companies to identify profitable investments, plan and optimise field development, manage/ monitor ongoing production, and rank commercial and technical decision making. However, determination of RF remains one of the greatest uncertainties in O&G projects, particularly in the early stages of field life.

RF prediction is difficult because a huge number of variables can affect the final RF of a field, some of which O&G operators have no control over, such as fluid and rock properties. Others important parameters include completion methods if enhanced oil recovery or waterflooding was done. In early field life, insufficient production data coupled with subsurface uncertainty make RF prediction more of an “art” than a science, and it is often the experience of the operator combined with analogue studies that is used to determine RF. However, operator experience can be biased, and in most cases, there is insufficient data from analogue fields to narrow the RF range.

Utilizing techniques of big data manipulation and machine learning, we has studied 1447 sandstone and carbonate fields extracted from an extensive worldwide database to (a) deconstruct and identify key variables that impact production profiles for the field and (b) create machine learning models that can predict RF based on said key variables. Our data set consists of ~200,000 real world data points. This data set was ‘pythonically’ cleaned and sorted such that only variables where measurements are present for 80% of the cases are kept. Where missing data was present, ERCE utilized regression models to augment the data set. Big data techniques were also used to automatically remove, rename, and combine data sets to reduce data granularity, saving ~1000 man-hours were manual cleaning attempted for the database.

Our work aims to show that not all variables influence RF equally; any machine learning model should therefore be built with variables that have the greatest influence on RF yet have the lowest pairwise correlation. The paper will demonstrate the accuracy of the RF prediction using our machine learning tool and will also discuss recommendations on the applicability of the procedure in different field types, and with different data availability.

Introduction

Recovery Factor (RF) is one the most critical, and sometimes subjective, inputs towards determine field reserve. Despite its significance, there is no clear approach to calculate or estimate given the variety of factors governing its value. Currently, evaluation of RF is the culmination of technical as well as non-technical aspects (analogs, experience based). It considers both qualitative and quantitative parameters in the evaluation. As such, there is a lot of uncertainties associated with predicting RF values. This is especially in the case of early field life when there is insufficient production data and subsurface uncertainty. In early field life, analogs and operators’ experience are usually used to predict RFs. This method is not reliable due to the insufficient data and biasness from operators. Hence, there is a need to improve the process of evaluating RFs. As such, this paper aims to determine important variables that affects RF prediction as well as develop machine learning models that predicts RF based on the variables identified, focusing on oil bearing reservoirs.

Literature Review

Literature Review (theory and reference)

Collinearity is a condition where a predictor variable correlates with another predictor variable strongly. Therefore, changes in one of variable would cause changes in the other, giving a highly unstable end model. This also makes it difficult to determine which variables are significant if the model fluctuates greatly. Additionally, overfitting of model occurs due to highly correlatable variables dominating the output.

Fundamentally, all 3 algorithms are linked to decision trees. Decision trees can overfit, which is why it is necessary to either stop at a certain level or stop once the nodes reach a certain size (“Stopping criteria”). It is difficult to decide when it is best to stop training by just looking at the learning curve for

training by itself . Alternatively, one can delete nodes (“Pruning”). The best method however is known as an “ensemble average” i.e., data set is split into numerous small sets (bootstrapping), a decision tree is created for each, giving N learned models (XL), which can then be combined to give an improved model that is robust to overfitting and outliers. Theoretically, Catboost can handle categorical as well as numeric data. Does this without requiring the conversion of categorical to dummy variable. This was supported in the results obtained. RFR is insensitive to outliers, able to work on large number of variables and hard to overfit. However, it requires a lot of CPU memory.

Ensemble Modelling is improving the predictive capability of ML an be done by “averaging” different interpreted models. Bootstrap Aggregating (Bagging) – take (homogeneous) “weak learners” of the same type (same variables each time, but random subset), pass a prediction through each of them and average the result. Each “weak learner” is independent

Boosting – “Strong learner” = Weighted sum of (homogeneous) “Weak Learners” with higher importance given to models that were difficult to predict in the previous step.

Blending/ Stacking – combining different “weak learners” – i.e. multiple models are trained to predict the outcome and a meta-model is created that uses the predictions from those models as an input along with the original features.

Methodology

The machine learning approach will be described in this section. The approach is divided into 6 stages: data collection, data preparation, choosing of learning algorithm, model training, model testing and evaluation of model. The dataset used was open-source, global, consisted of both carbonate and sandstone reservoirs, all fluid types and included recovery mechanism. It had a total of 342 columns and 1447 rows.

Data collection and preparation can be group together under “Exploratory data analysis”. Exploratory data analysis is the initial investigation on the dataset to determine patterns and outliers through statistics and data visualization. Firstly, the attributes of the variables in dataset are defined. Statistics and data visualization are then obtained via univariate, bivariate and multivariate analysis. The previous step allows the identification of missing values, abberants and outliers. Hence, data cleaning is carried out to account for these factors that can affect (something). Data cleaning involves removing duplicates, fixing indexes, replacing incorrect characters, normalization of names, etc. Missing values were filled with regression-based values achieved through applying domain knowledge where a reservoir or Petrophysics uses industry learnt relationships and applies them as regressional trends to create pseudo-data. Lastly, correlation matrix is obtained to provide information on any collinearity between the variables. After all the above steps are carried out, final input variables that are used in the machine learning model are obtained.

Supervised learning approached is applied in which the raw inputs are split into a train-test set (70% - 30% split). The “train” set is used to build the model while the “test” set is used to test the trained model. We “validate” the model using a totally new, never-before-seen data set. A total of 20 models were tested as a first pass\* using 10-fold cross validation; these models were ranked using the mean absolute error (MAE), the mean squared error (MSE) and the root mean squared error (RMSE). The formulas are described as

where yi is the predicted value, xi is the true (measured?) value and n is the total number of data points. The input variables were then ranked for the top 3 best models. To account for idiosyncrasies in the data (noise, patterns, outliers, etc.), k-fold cross-validation was ran to validate the stability of the model. The model was tested with 5, 10 and 15 folds.

The models were then tuned by varying hyperparameters. A hyperparameter is a characteristic of a model that is external to the model and whose value cannot be estimated from data. The value of the hyperparameter has to be set before the learning process begins. These “hyperparameters” can be optimized via a random grid search to regression error (MSE and RMSE) metrics. The predictive capability of machine learning was improved through “averaging” different interpreted models. Bootstrap An optimal solution that is a blend of all the tuned models was obtained.

Results

This section will cover the results. When testing for collinearity, the following correlation matrix has been obtained. Highly correlated values are defined as having coefficients larger than 0.7.

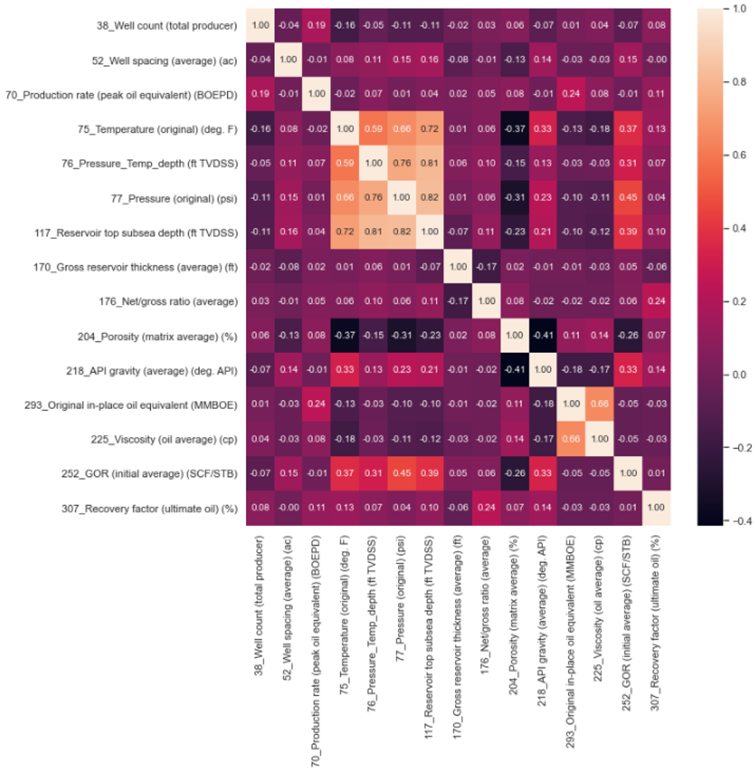


Figure 1: Correlation Matrix – Pairwise Correlation

From the plot, the following 3 parameters show collinearity: Temperature, Reservoir top subsea depth and pressure. As these parameters do not add additional information to the predictive model, data from “Reservoir top subsea depth” and “Pressure\_Temp\_Depth” was eliminated to allow for a more stable model.

After data collection and data cleaning, final input variables were determined and summarized in table 1. The dataset comprises of 7 categorical data types, and 12 numerical data types. The total data size is 436 values per column, for a total of 8,284 data points which was approximately 4% of the original data base.

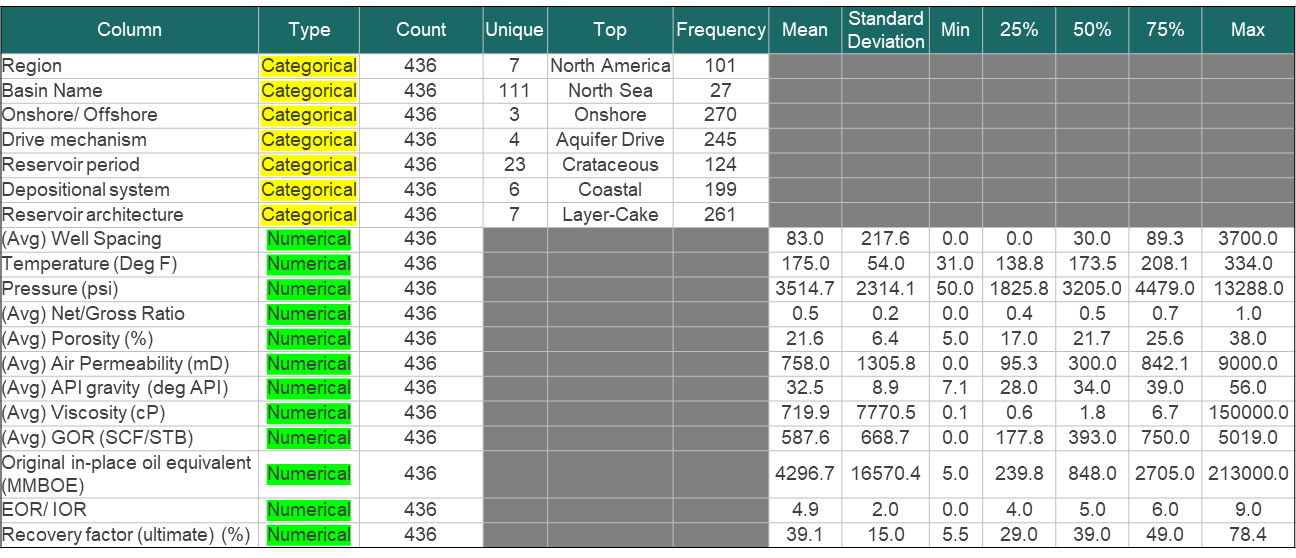


Table 1: Breakdown of Input Variables used in the ML algorithm

When ranking the 20 models using MAE, MSE and RMSE, it was found that MSE and RMSE performed better against outliers. The best 3 approaches found were Category Boosting (CatBoost), the Random Forest (RFR) and the Gradient Boosting Regressor (GBR). The results for each model are tabulated in Table 2.

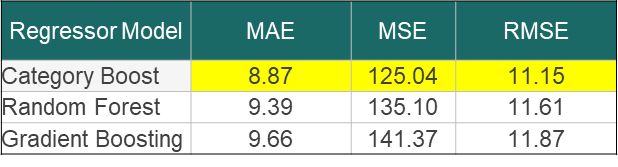


Table 2: Results of approaches

With reference to Table 2, Catboost achieved the lowest values in all error tests and GBR produced the highest values of error. When the top 13 features for each decision tree model were ranked, the following results were obtained.

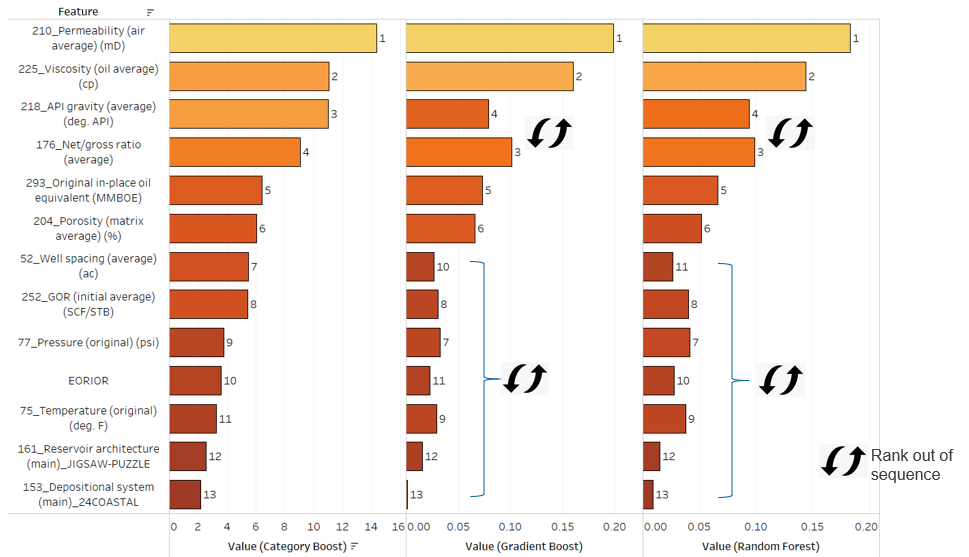


Figure 2: Ranks of Top 13 Features

From figure 2, Permeability, viscosity, API, NTG, OOIP and Porosity appear to be the the 6 most important variables. The remaining variables change in importance ranking depending on the algorithm employed. It was observed that categorical data is more important for the CatBoost algorithm than it is for the more “numerically weighted decision tree”. As each model considers each feature differently, it poses a challenge when choosing an optimal model that gives an outcome of low bias and variance.

Results of K-fold Cross-validation showed that RFR performs better at 5 K-folds while the other models perform better at 10 K-folds. When tuning the model, it was observed that optimization plateaus and more runs or more cross-validation will not reduce the error further.

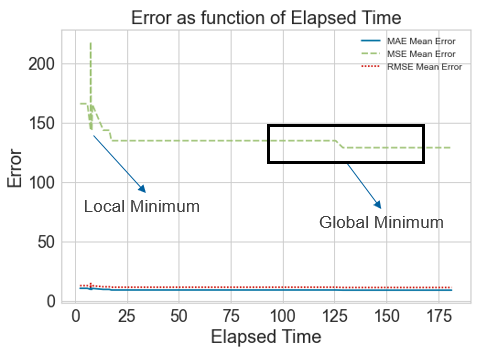
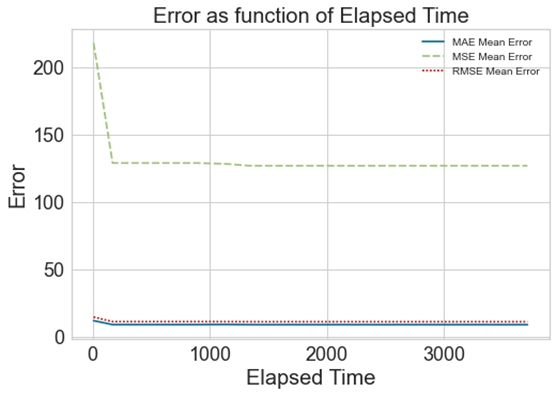


Figure 3:

For 10-fold cross-validation i.e. 10 subsets of the full training data, running it 50x means the optimization algorithm tunes the hyperparameters 50 times per model (using gradient descent). The total run time is 180s, and the MAE obtained was 9.12. Repeating this for 1,000 fits results in a run time of >1 hour , but the MAE only reduces to 9.07. Furthermore, the solution converges rapidly at ~50 iterations in the hyperparameter grid search.

Table 3 presents the parameters tuned. (What does optimized solution mean?)

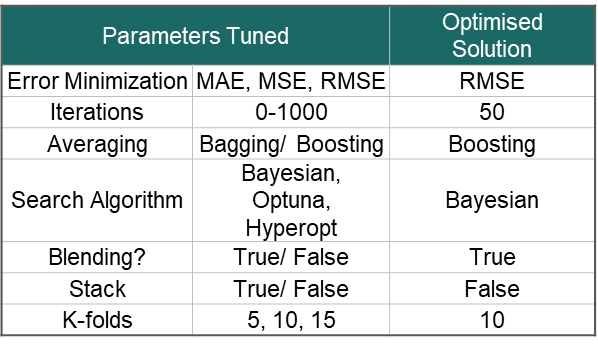


Table 3: Summary of Parameters Tuned

Results of the error test showed that varying hyperparameters decreased regression error.

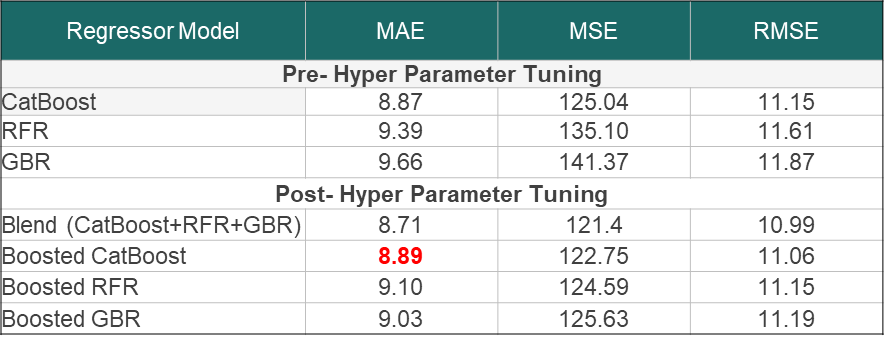


Table 4: Error Comparison between Regression Models

As tabulated in Table 4, most values for all error test for post-hyper parameter tuning decreased from pre-hyper parameter tuning. An increase was seen in the MAE value for boosted Catboost after varying hyperparameters. However, the increase was a 0.02 difference which is negligible. As such, tuning of hyperparameters decreases regression error which leads to a more accurate model. Furthermore, the blended model achieved the lowest error values for MAE, MSE and RMSE. This suggests that the blended model was the most accurate amongst the model tested. Hence, the optimal solution was a blend of all the tuned models. This was further supported by the R2 value obtained. As seen in Figure 4, both train and test set had R2 (adjusted R2?) value of 0.982. This suggests that the best fit line is able to fit most of the data in the regression model.

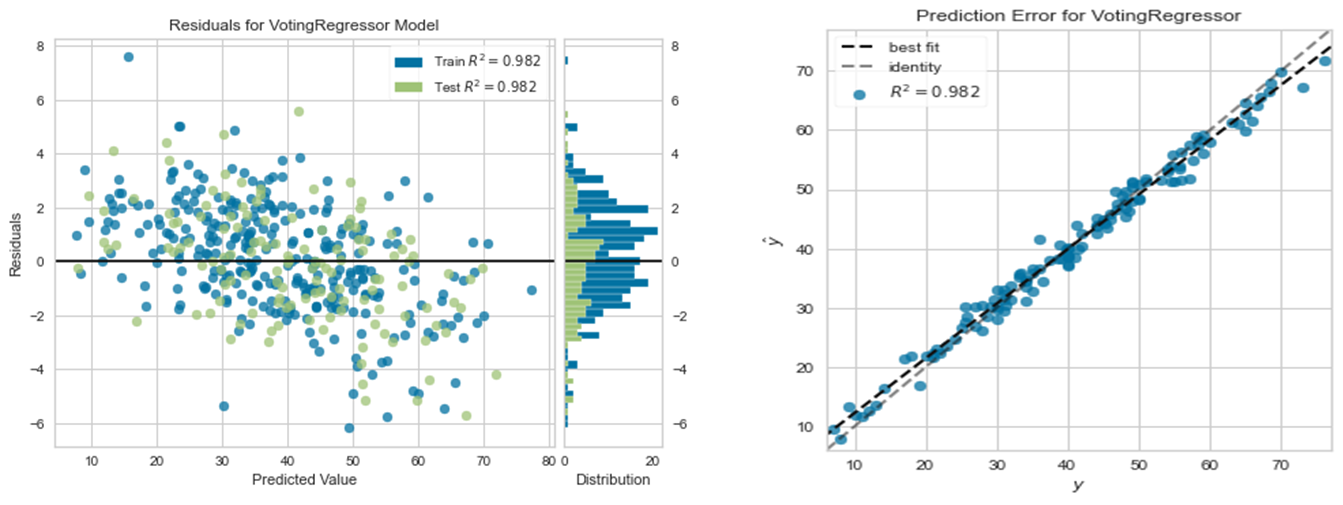


Figure 4:

Therefore, with the high R2 value and lowest values of error, it can be concluded that the optimal solution is blend of all the tuned models.

Discussion

Discussion; slide7 (talk about why domain knowledge is preferred in this case?); slide 8; slide 9; slide 14; slide 16; slide 18; slide 19

(Explain results) This section will explain the findings obtained.

The alternative to handle missing values is to drop the missing variables. However, when this was done, we ended up with only 712 numerical data points or 0.37% of the original data set. Hence, applying domain knowledge yield better results. Domain knowledge also informs us that NTG & Porosity are independent variables required to evaluate STOIIP.

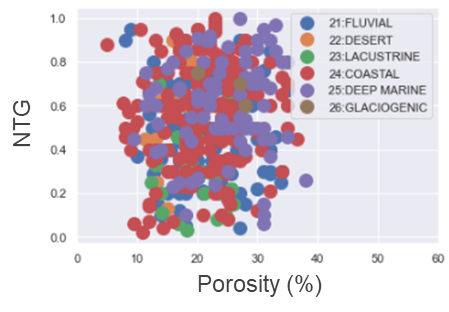
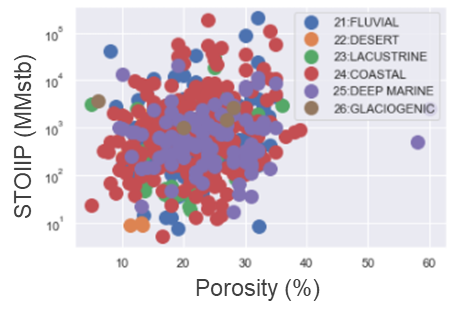


Figure 3: STOIIP vs Porosity plot and NTG vs Porosity plot

Therefore, while the 2 plots in Figure 3 look similar, correlation for one is possible, but makes no sense for the other (geologically). As such, missing values in NTG cannot be corrected with porosity.

In an optimalization problem, there is a need to minimize the function. As such, the gradient is used to determine the direction in which the parameters should move in. However, as gradient is not indicative of how much to move, iterations are done through moving a fixed amount. Theoretically, when the number of iterations increases, the solution moves towards the minima and set of optimal parameters. This improves the performance of the model. In reality, the solution converges rapidly at ~50 iterations in the hyperparameter grid search; adding more iterations to improve model performance is counter-productive since it only increases the training time.

Conclusion

Conclusion

Reference

[1] Xue Ying 2019 J. Phys.: Conf. Ser. 1168 022022