Data-Driven Machine Learning Approach in Reservoir Parameter Prediction

Vivian O. Oguadinma[1], ThankGod Ujowundu[2], Chibuzo V. Ahaneku[3]
[1] Dievoc Integrated
[2] TotalEnergies
[3] University of Malta

Abstract

In the realm of reservoir engineering, the application of machine learning has emerged as a transformative force, offering unprecedented insights into reservoir parameter characterization. In this study, we present a comprehensive analysis of four distinct machine learning models, namely Bagging, Extra Tree Regressor, XGBoost, and Ridge, to elucidate their efficacy in predicting permeability, a critical parameter for reservoir characterization. Our findings reveal an understanding of each model's performance. The Bagging model, while demonstrating an impressive trained accuracy of 0.99, exhibits some uncertainty in high permeability predictions, casting slight shadows on its applicability for reservoir characterization. In contrast, the Extra Tree Regressor model outshines the Bagging model with a trained accuracy of 100% and a prediction accuracy of 99.8%. It boasts lower absolute and absolute percentage errors, reinforcing its ability in permeability prediction. However, the XGBoost model takes a unique approach by emphasizing the density-corrected log over gamma-ray and sonic logs. Despite achieving remarkable trained and predicted data accuracy exceeding 99%, its reliance on the corrected density log introduces a mean absolute percentage error above 10, warranting closer scrutiny. In contrast, the Ridge model struggles, evident from its high AIC reading, signifying its limited compatibility with permeability prediction. Joint plots and LMplot analyses further showcases model behaviors. The Extra Tree model exhibits a 99% confidence interval, underscoring its reliability with minimal underpredictions. Conversely, the Bagging and Ridge models show susceptibility to high uncertainties in permeability predictions, particularly at extreme values. Our study concludes that the Extra Tree Regressor model excels in permeability prediction, with potential applications in reservoir interval assessments. The XGBoost model, while competent in sandstone reservoir prediction, bears a higher uncertainty burden. The Bagging and Ridge models, due to their uncertainty challenges, are less suitable for non-reservoir and sandstone reservoir interval predictions. High permeability correlations with elevated porosity, reduced water saturation, and lower gamma ray readings highlight the reservoir intervals' distinct characteristics. These observations underscore the reliability of our models and their potential contributions to reservoir engineering practices.

Keywords: Permeability Prediction, Machine Learning, Reservoir Characterization, Feature Engineering, Model Evaluation.

1. Introduction

The exploration and production of hydrocarbon reservoirs demand precise characterization of reservoir parameters to optimize recovery and mitigate uncertainties (Nwaezeapu, et al., 2018). Over the years, reservoir geoscientists have employed various methods, from analytical techniques to empirical models, to estimate these parameters. However, the inherent complexities of subsurface reservoirs have necessitated the evolution of reservoir characterization approaches.

Well log signatures can show several aspects of the formation's lithology. Research has recently shifted its attention to the prediction of reservoir parameters using log curve data (Song et al., 2021a, Ibekwe et al., 2023; Pwavoa et al., 2023; Oguadinma et al., 2023).

Corresponding author: vo.oguadinma@deivocintegrated .com

The foundation for characterizing reservoir attributes and reservoir modeling is the reservoir physical parameters, which primarily comprise porosity, permeability, water saturation, and oil saturation (Li et al., 2016; Wang et al., 2019; Song et al., 2021a; Oguadinma et al., 2016; Oguadinma et al., 2017). In particular, a reservoir's pore space plays a significant role in the buildup and movement of hydrocarbons. It is also essential for the development of reservoirs that hold hydrocarbons. The porosity of a rock is a direct indicator of its hydrocarbon-storage capacity. It plays a significant role in the assessment of reservoirs and is crucial to the discovery and growth of oil and gas fields. Another essential metric for determining a reservoir's properties and calculating its oil reserves is the water saturation, gas, and oil in the reservoir.

Porosity and saturation can be measured using one of two recognized methods (Song et al., 2020; Wang et al., 2020, 2022; Song et al., 2021b). Using rock slices or cores, the first method is a direct estimation that gets the physical parameter data directly. Although this procedure is more accurate and is frequently used in laboratories, it is expensive and time-consuming. The other approach is an indirect measurement, which involves approximating the saturation and porosity using function approximation, statistical, and geological approaches (Wyllie et al., 1956; Raymer et al., 1980; Oguadinma et al., 2014; Oguadinma et al., 2021; Ibekwe et al., 2023). The primary focus of early reservoir parameter prediction logging techniques was on linear data. The physical parameters are determined by the use of empirical formulas and linear equations, a purely quantitative approach that ignores the Reservoir actual environments.

Traditionally, reservoir estimation has relied on core data, well logs, and geological information (Oguadinma et al., 2014). However, these conventional approaches often face challenges in capturing the inherent heterogeneity and complexities of subsurface formations (Aniwetalu et al., 2018). In addition, the relationship between the well logging data and reservoir parameters is nonlinear. The traditional regression analysis methods are difficult to achieve satisfactory results. Therefore, exploring a novel method for reservoir parameter prediction is particularly necessary for the development of unconventional and complex oil and gas fields.

Recent advancements in artificial intelligence (AI) and machine learning (ML) have provided a paradigm shift in reservoir geoscience and engineering, offering promising tools to enhance the accuracy and efficiency of parameter characterization.

Permeability, porosity, and saturation are among the reservoir metrics that have been obtained by certain researchers. In order to forecast permeability, Akande et al. (2015) suggested an artificial neural network (ANN) based on the correlation feature selection. The outcomes demonstrate that permeability may be predicted using fewer features with this method. Komarialaei and Salahshoor (2012) also predicted permeability using the ANN model and principal component analysis (PCA). The method's practicality is demonstrated by the experimental results. Hadi and Sadegh (2016) used an intelligent technique based on seismic characteristic data to estimate porosity. The random forest method was presented by Song et al. (2016) to estimate seismic reservoirs. It was discovered that the method has a certain stability and accuracy and is less affected by noisy data.

Predicting the permeability based on well logs using machine learning algorithms is a feasible and alternative method. However, the accuracy cannot fully meet the requirement. Therefore, the complex nonlinear relationship is still needs to be further explored.

This study represents a pivotal advancement in the field of reservoir geoscience by harnessing the power of machine learning techniques. It focuses on the essential reservoir parameter permeability and explores the predictive capabilities of several ML models. Permeability, a critical property influencing fluid flow within a reservoir, serves as a keystone in assessing reservoir behavior and optimizing production strategies.

Corresponding author: vo.oguadinma@deivocintegrated .com

In this study, we evaluate the performance of four distinct ML models: Bagging, Extra Tree Regressor, XGBoost, and Ridge. Our analysis extends beyond conventional evaluation metrics, delving into model interpretability and prediction reliability. The evaluation process is bolstered by comparing the models' performance across a diverse range of reservoir scenarios, including sandstone and non-reservoir intervals.

2. Methodology

2.1: Study Area and Data Collection

The dataset used for this study is a full well log suit comprising of Gamma ray log; shallow, true and medium resistivity logs; neutron log; density log; permeability log, porosity log and water saturation (SW) log were collected and quality checked using the python data processing mechanism as explained below.

2.2: Data Preprocessing

Before analysis, the collected data underwent a rigorous preprocessing phase. Missing values in well logs were imputed using the mean imputation method. Outliers were detected and removed using the Z-score-based outlier detection method. Categorical variables, such as lithology and facies, were encoded using one-hot encoding.

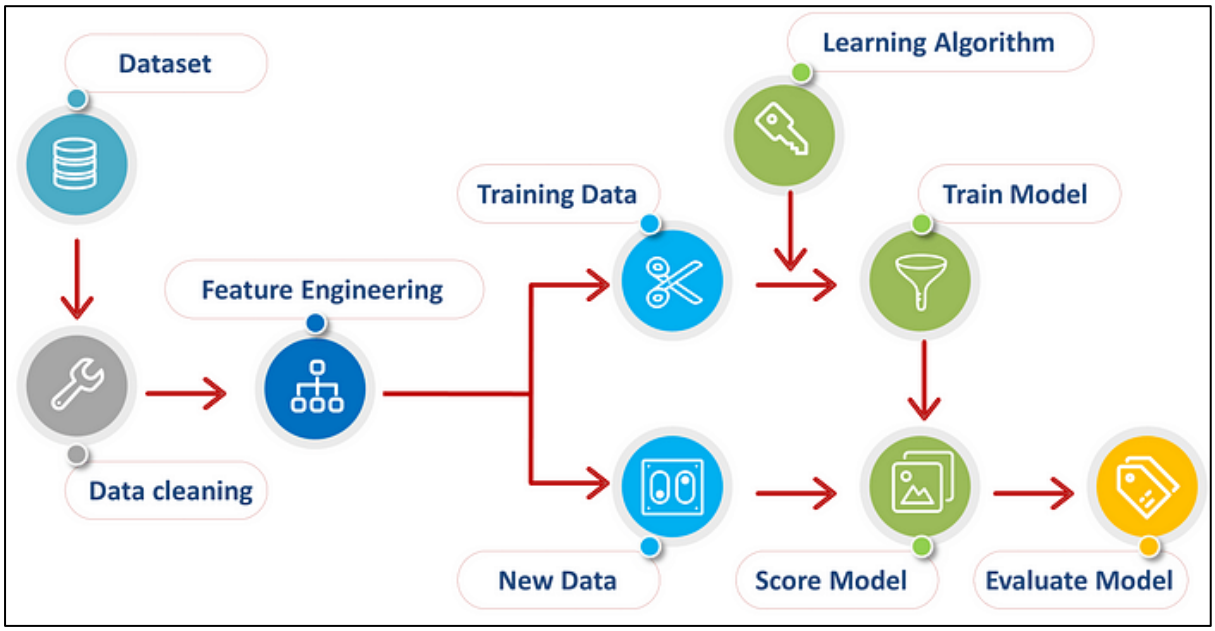

Fig. 1: Machine learning data pre-processing technique (Chakure, 2019)

2.3: Feature Selection and Engineering

Feature selection was performed to identify the most relevant variables for predicting permeability. This involved using a combination of correlation analysis and feature importance scores obtained from machine learning models. Additionally, new features were engineered to capture potential interactions between variables.

2.4: Machine Learning Models

To predict permeability, porosity and water saturation, we employed several machines learning algorithms, including:

Corresponding author: vo.oguadinma@deivocintegrated .com

2.4.1: **Bagging Model:**

This ensemble model was used to harness the predictive power of multiple base estimators. The Bagging principle is primarily a process-driven technique rather than one represented by a specific mathematical equation. However, the core idea of Bagging can be understood through the mathematical concepts underlying the process.

***Bootstrap Sampling*:** The process of creating bootstrap samples involves randomly selecting data points from the original dataset with replacement. The mathematical representation for bootstrap sampling is as follows:

Let N be the total number of data points in the dataset.

A bootstrap sample of size N (the same size as the original dataset) is created by randomly selecting N data points from the original dataset with replacement. This can be represented mathematically as:

$$D^{*}_{i} = (x_1, y_1), (x_2 y_2) \dots \dots \dots (xN, yN) \quad \text{.....Eqn. 1}$$

where D_i^* is the bootstrap sample, (x_j, y_j) represents the j-th data point, and N is the total number of data points.

Aggregation: Bagging combines the predictions of multiple base models. The mathematical representation for aggregating the predictions can vary depending on the type of problem (classification or regression) and the method used (e.g., voting, averaging, or weighted averaging).

For classification, a common aggregation method is to use the mode (most frequent class) of the individual predictions.

For regression, the predictions are typically averaged to obtain the final prediction.

Mathematically, the aggregation process can be represented as follows:

For classification: Let C_i^* be the predicted class from the i-th base model, and the final prediction is given by:

$$C_{final} = model(C_1^*, C_2^* \dots \dots C_m^*) \quad \text{.....Eqn. 2}$$

where m is the number of base models.

For regression: Let Y_i^* be the predicted value from the i-th base model, and the final prediction is given by:

$$Y_{final} = \frac{1}{m} \sum_{i=1}^{m} Y_i^* \quad \text{.....Eqn. 3}$$

where m is the number of base models.

It's important to note that while Bagging's core idea can be represented in these mathematical terms, the primary value of Bagging lies in the process of creating diverse subsets and combining predictions, rather than in a single mathematical equation. In this study, we applied the Bagging regression approach.

Corresponding author: vo.oguadinma@deivocintegrated .com

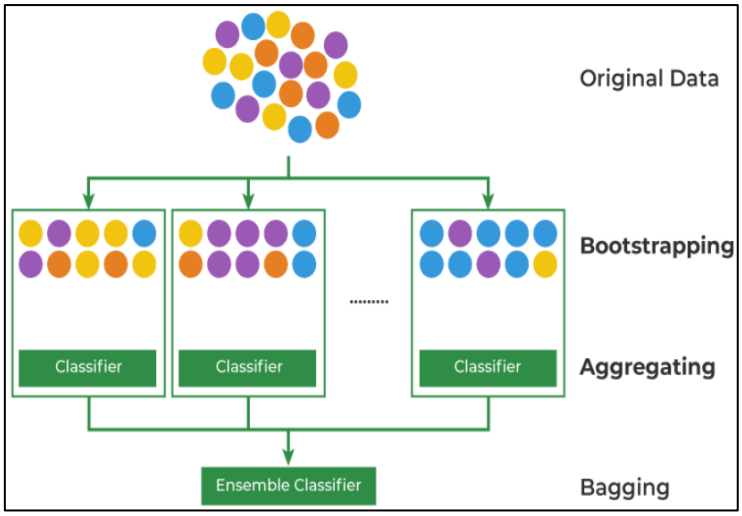

Fig. 2: Bagging helps improve accuracy and reduce overfitting, especially in models that have high variance (GreeksforGreeks)

2.4.2: **Extra Tree Regressor Model:**

The Extra Trees algorithm was chosen for its ability to handle noisy data. This model, which is a variation of Random Forest, is an ensemble learning method for regression tasks. Random forest (RF) is an ensemble machine learning approach proposed by Breiman (2001), which has the advantages of interpretability, convenience, and fast calculating speed. The Extra Trees algorithm primary principle is to build multiple decision trees using different subsets of the data and then combine their predictions. While there isn't a single mathematical equation that encapsulates the entire model, in this study, we provide a method that relies on the principle of averaging the predictions from multiple decision trees. The mathematical equation to express the prediction made by the Extra Trees Regressor is as follows:

For each individual tree *i* in the ensemble:

$$Y_i^* = Tree_i(X) \quad \text{.....Eqn. 4}$$

Where:

Y_i^* is the predicted value from the *i*-th tree.
$Tree_i(X)$ represents the prediction made by the *i* -th decision tree for the input data X.

To make a final prediction for the Extra Trees Regressor, you simply average the predictions from all individual trees in the ensemble:

$$Y_{final} = \frac{1}{n} \sum_{i=1}^{n} Y_i^* \quad \text{.....Eqn. 5}$$

Where:

Y_{final} is the final prediction made by the ensemble of Extra Trees Regressor.
n is the total number of trees in the ensemble.
Y_i^* is the prediction made by the *i*-th tree.

This averaging process helps improve the stability and accuracy of the regression model by reducing the variance and overfitting. Each tree in the ensemble contributes to the final prediction, and the

Corresponding author: vo.oguadinma@deivocintegrated .com

randomness introduced during the tree-building process adds diversity to the model, making it less prone to overfitting.

2.4.3: **XGBoost Model**:

XGBoost (Extreme Gradient Boosting) is a popular machine learning algorithm that is particularly effective for both classification and regression tasks. The key principle behind XGBoost is gradient boosting, which involves creating an ensemble of decision trees and minimizing a specific loss function. While XGBoost involves multiple mathematical equations, I'll provide an overview of the main equations that support its principle.

The central equation is the calculation of the prediction made by the XGBoost model:

$$y\hat{}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i) \quad \text{.....Eqn. 6}$$

Where:

$y\hat{}_i$ is the predicted value for the *i*-th data point.
x_i is the feature vector for the *i-th* data point.
$\phi(x_i)$ is the prediction function.
K is the total number of base learners (individual decision trees).
$f_k(x_i)$ is the prediction made by *the k-th* base learner for the *i-th* data point.

The prediction is a sum of the predictions from all the base learners.

XGBoost involves a specific type of gradient boosting, which minimizes a loss function. The primary loss function for regression problems in XGBoost is the squared error loss.

$$L(\phi) = \sum_{i=1}^{N} \left((y_{i\_} \, y\hat{}_i \,)^2 \right) \quad \text{.....Eqn. 7}$$

Where:

L(ϕ) is the loss function.
N is the total number of data points.
yi is the actual target value for the *i*-th data point.
y^i is the predicted value for the *i*-th data point.

To minimize this loss function, XGBoost uses a combination of regularization terms and a second-order approximation of the loss function. These terms help prevent overfitting and fine-tune the model.

2.4.4: **Ridge Model**:

The Ridge Regression model is a linear regression model with a regularization term that helps prevent overfitting. The principle of Ridge Regression is to minimize the sum of squared differences between the observed and predicted values while also penalizing the magnitudes of the model's coefficients (parameters). The mathematical equation for Ridge Regression is as follows:

Corresponding author: vo.oguadinma@deivocintegrated .com

Objective Function of Ridge Regression:

Ridge Regression seeks to minimize the following objective function:

Objective Function = $\sum_{i=1}^{N}(y_{i-}\ y\hat{}_{i})^2 + \ \alpha\sum_{j=1}^{p}\beta_j^2$ .....Eqn. 8

Where:

N is the number of data points.
y_i is the observed target value for the *i*-th data point?
y^i is the predicted value for the *i*-th data point.
α (alpha) is the regularization hyperparameter, which controls the strength of the penalty on the coefficients.
p is the number of features (predictor variables).
βj represents the regression coefficients for the *j*-th feature.

The objective function consists of two parts:

The first part (equation 9) is the ordinary least squares (OLS) loss, which measures the goodness of fit.

$\sum_{i=1}^{N}((y_{i-}\ y\hat{}_{i})^2)$.....Eqn. 9

The second part (equation 10) is the L2 regularization term, also known as the Ridge penalty. It discourages the model from having large coefficients by adding a penalty based on the square of the coefficients.

$(\alpha\sum_{j=1}^{p}\beta_j^2)$.....Eqn. 10

The Ridge Regression model aims to find the coefficients (βj) that minimize this combined objective function. The regularization term encourages the model to have small coefficients, effectively reducing the impact of individual features. The hyperparameter α controls the trade-off between fitting the data well and keeping the coefficients small.

The Ridge Regression adds an L2 regularization term to the linear regression objective function, which helps prevent overfitting by penalizing large coefficients. The goal is to find the coefficients that minimize the sum of squared differences between observed and predicted values while keeping the coefficients as small as possible.

2.5: **Methods for Model Evaluation**

To evaluate the performance of the models, we used a combination of metrics, including:

Accuracy: A measure of the model's ability to correctly predict the target values.

Mean Absolute Error (MAE): A measure of the absolute differences between predicted and actual values.

Corresponding author: vo.oguadinma@deivocintegrated .com

Mean Absolute Percentage Error (MAPE): A measure of the percentage difference between predicted and actual values.

AIC (Akaike Information Criterion): A measure of the model's goodness of fit.

Each model's accuracy and error metrics were computed and compared to select the best-performing model for permeability, porosity and SW prediction.

2.6: **Visualization**

Various visualization techniques were applied to provide insights into the model predictions. These included joint plots and LM plots, which were used for example, to visualize the relationships between predicted and actual permeability values, as well as confidence intervals.

3. Result and Interpretation

3.1: Feature Importance

Feature selection was performed to identify the most relevant variables for predicting permeability. And from this chart, Volume of shale (VSH) and Gamma ray (GR) logs are the most relevant logs for the desired purpose.

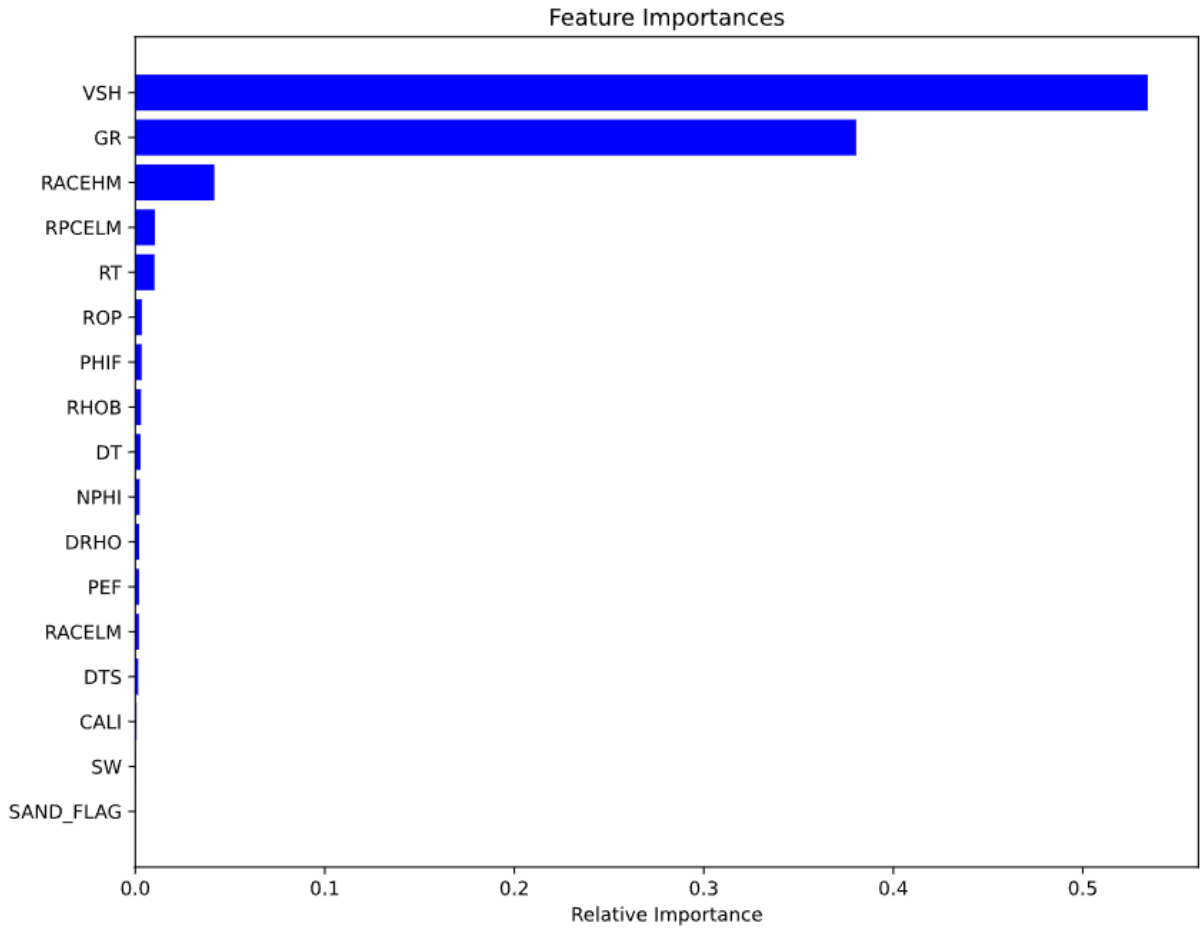

Fig. 3: Result from feature selection showing chart, where Volume of shale (VSH) and Gamma ray (GR) logs as the most relevant logs for reservoir parameter prediction.

Corresponding author: vo.oguadinma@deivocintegrated .com

3.2 Model Evaluation

3.2.1: Bagging Model

There is a trained accuracy of 0.99 which is almost 1.0. This shows that the model acts well in training and predicting data as well. The predicted data is 0.94 which shows that the model is a robust fit when used to predict permeability (Figure. 5). The mean absolute error of 5.7md in terms of permeability and mean absolute percentage error of 8.7 which is less than 10 is a good attribute desired in a well-performing model. The AIC is in +4000 as opposed to the former in +6000 obtained from the ridge model confirming the good fit of the model (Table 1 and Figure 4).

Table 1: Error metrics value readings from the four trained and tested models.

Models	Train_Score	Test_Score	AIC	Mean Absolute Error	Root Mean Squared Error	Mean Absolute Percentage Error
Ridge Regression	0.1232	0.1037	6206.36	135.24	357.37	21917.01
Extra Tree Regression	1.0000	0.9989	3342.84	2.34	12.34	6.17
XGBoost	0.9999	0.9921	4080.75	5.50	33.50	15.94
Bagging	0.9971	0.9923	4094.63	5.88	33.17	7.49

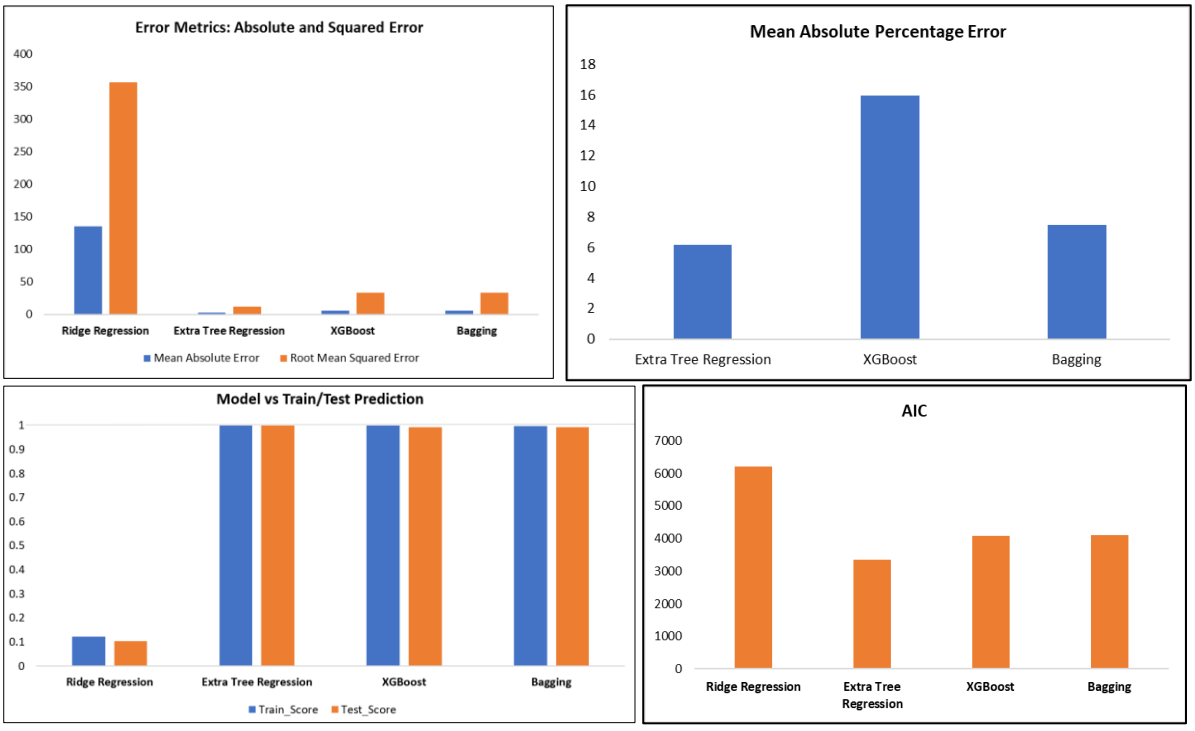

Fig. 4: Charts of the error metrics across the four machine learning models applied to this study. Observe how the ET model ranks least in all the error metrics with the Ridge model ranking comparatively high.

Corresponding author: vo.oguadinma@deivocintegrated .com

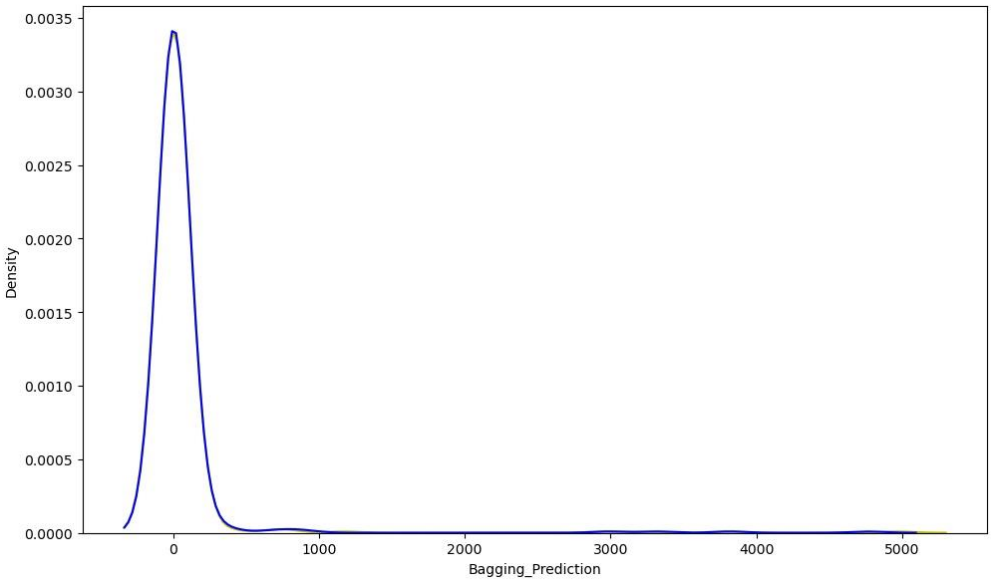

Fig. 5: The bagging model overlaps the actual prediction just as much as the XGB model with a slight discrepancy at the toe of the plot.

3.2.2: Extra Tree Regressor Model

The trained accuracy obtained from this model is 100% with a prediction accuracy of 99.8%. The absolute error and absolute percentage errors are less than the values obtained from the Bagging model. The AIC reading is lesser and even better. Thus the Extra Tree Regressor model performs better than the Bagging model as evident in the chart output of Figure 5.

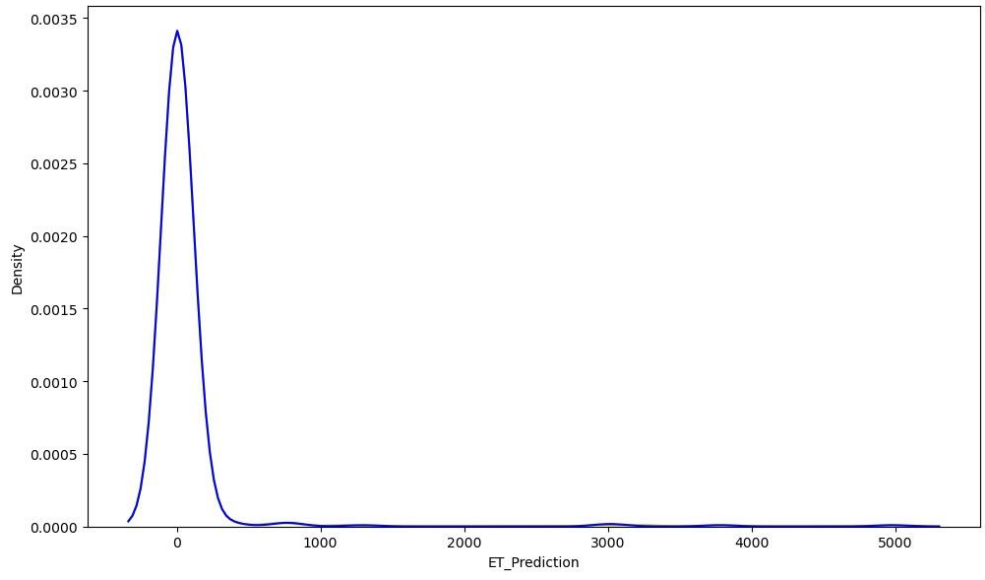

Fig. 6: The Extra Trees model overlaps the actual prediction almost perfectly.

3.2.3: XGBoost Model

From the XGB model analysis in comparison to the Extra Tree Regressor, it is clear that the Extra Trees Regressor model is the best in permeability prediction (Figure. 6 and 7). XGboost model made the density-corrected log more important than the gamma-ray and sonic logs. However, corrected density

Corresponding author: vo.oguadinma@deivocintegrated .com

doesn't have a good correlation to permeability (Figure 4) as such it is interpreted to have added to the large value of mean absolute percentage error which is above 10 displayed by the model. Although the trained and predicted data from the model show an accuracy of over 99%, the error specifics are important in making the final conclusions. Many researchers and authors only make conclusions based on accuracy and mean absolute error, this study went a step ahead. Because the XGboost made the error with the corrected density log, the value of accuracy is in doubt

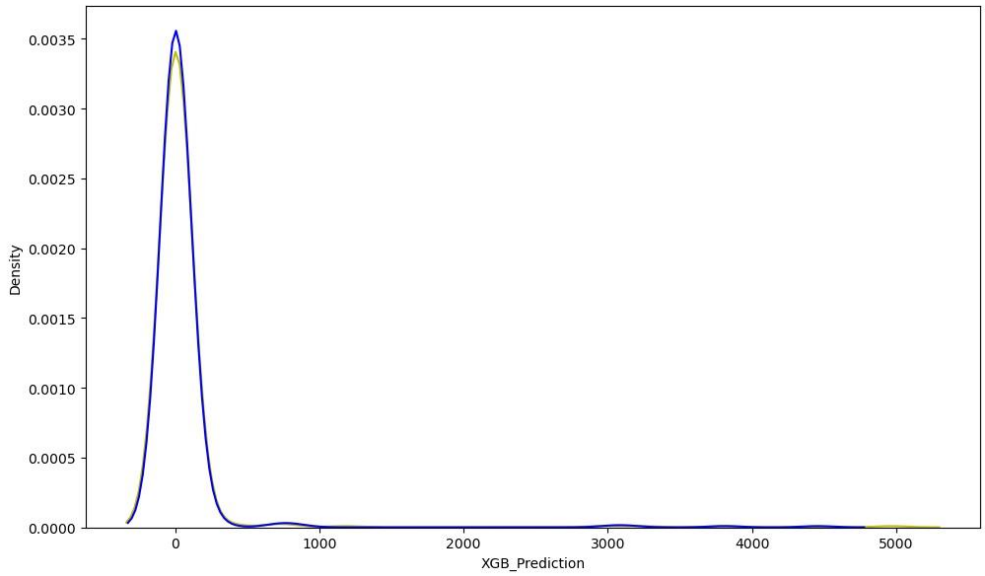

Fig. 7: The XGB model overlaps the actual prediction but is not as perfect as the Extra Tree model. The yellow curve represents the actual predictions while the blue curve stands for model predictions

3.2.4: Ridge Model

From Table 1, the value of AIC is very high for the Ridge model; so it doesn't have the potential to fit well with the model in terms of permeability prediction. Furthermore, the error metrics such as the mean error, the mean absolute error, and the mean percent absolute error values are very high; and with the lowest accuracy value.

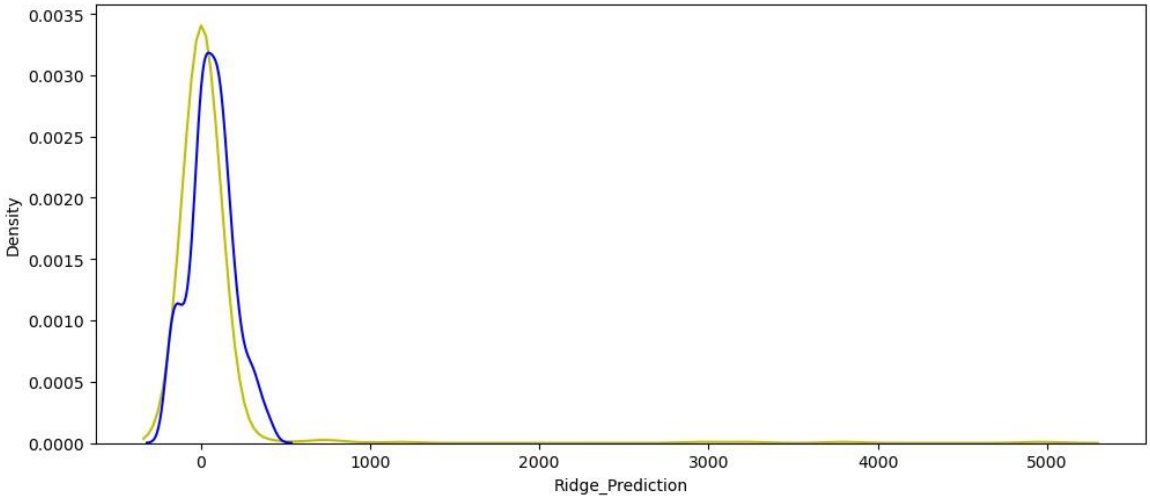

Fig. 8: This graphed prediction shows that the Ridge model is underpredicting. It hardly matches the shape and value of the actual prediction. The yellow curve represents the actual predictions while the blue curve stands for model predictions.

Corresponding author: vo.oguadinma@deivocintegrated .com

3.3: Joint plot

3.3.1: Extra Tree model

The ET model result shows a 99% confidence interval of the Extra Trees prediction model. This model shows better and almost perfect prediction when compared to other models. It shows negligible evidence of underpredictions and the error metrics values are desirable as the mean absolute percentage error is a little above 10 and has the lowest mean absolute error value of 2 (Fig. 9 and Table 1).

3.3.2: Bagging

This result from the Bagging model shows a 90% confidence interval. Although at high predictions the model is good and overlaps, the shaded areas show the uncertainty regions and it proves that at high values of around 500md of permeability, there is a high chance of uncertainty when the Bagging model is used. And this will be a problem in terms of reservoir characterization (Figure 9 and Table 1).

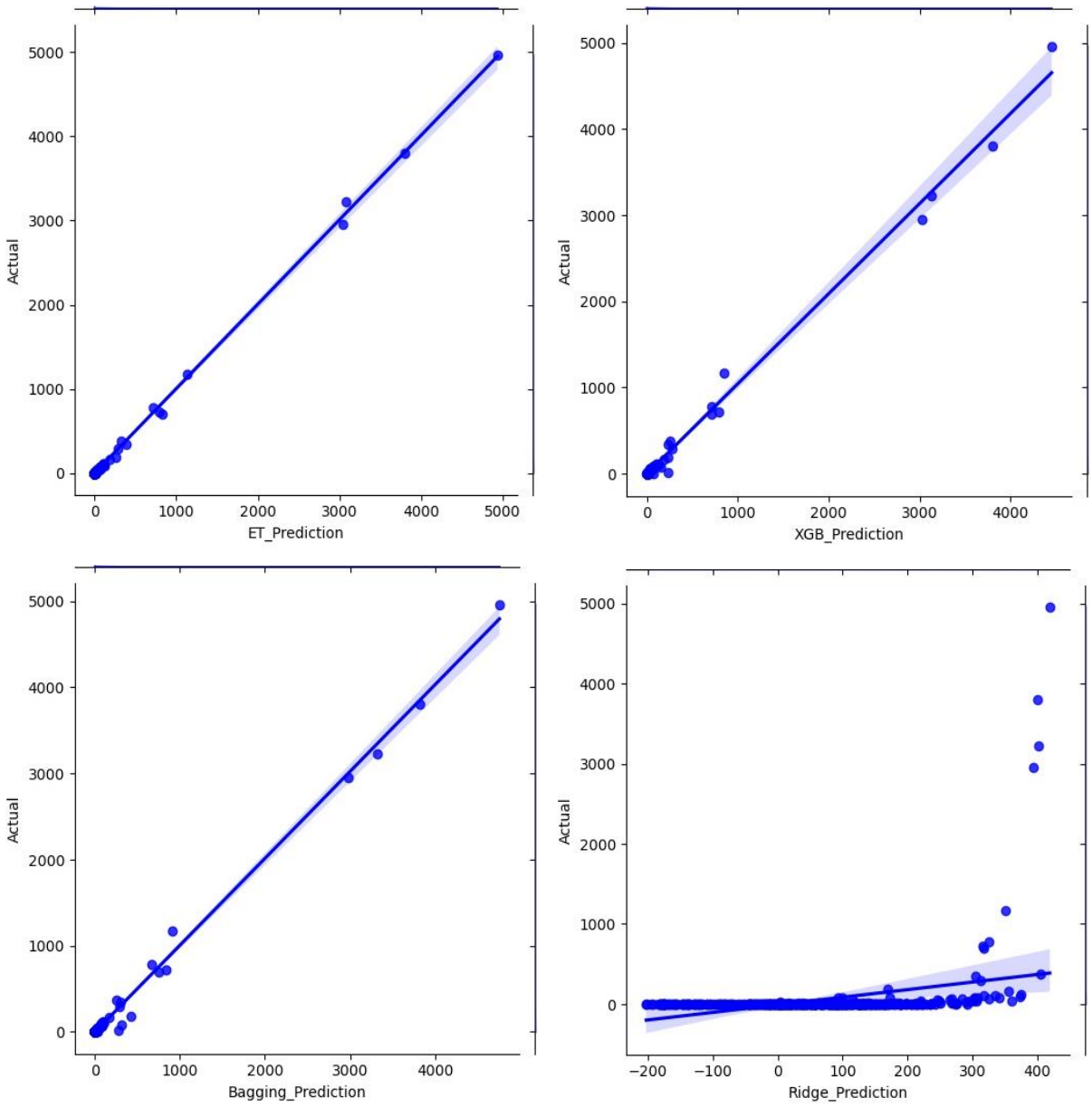

Fig. 9: These images are the results from the Joint Plots extracted from Extra tree regression, XGB, Bagging, and Ridge models. Observe the line of fitness for each model.

Corresponding author: vo.oguadinma@deivocintegrated .com

3.3.3: XGB

The result of the XGB prediction model shows a 95% confidence interval. At about 2000md, there are some errors and under-predictions at the interval. This is not really a problem because 2000md in itself is already a high permeability value. This model is better than the Bagging model because of the lesser mean square error and higher accuracy value it has compared to the Bagging model.

3.3.4: Ridge model

The Ridge prediction model shows a very low confidence interval. It has less than the desired prospectivity of being used in permeability prediction.

3.4: LMplot

3.4.1: Extra Trees model

From the line graph (Figure 10), there is almost a perfect correlation between the points and the lines. This observation indicates that the ET model has an accurate predicting power for non-reservoir and sandstone reservoir intervals (Table 2). This model will give higher accuracy and lesser error compared to other models. It is recommended for reservoir interval predictions. Although, there could be intercalations of thin-bedded shales within the sandstone reservoir interval. This model is also perfect for predicting homogeneous reservoir intervals.

3.4.2: XGB model

There is no perfect but good enough relationship between the point data and line in the graph in the XGB model. Although, it has an above-average potential for predicting the sandstone reservoir interval but very little potential for non-reservoir prediction due to high uncertainty. Thus, can be recommended for reservoir interval prediction following the ET model as the best (Figure. 10; Table 2).

3.4.3: Bagging model

The Bagging model shares similar attributes as the XGB model in terms of line-point data relationships. Just like the XGB model, it has little potential to predict the non-reservoir and sandstone reservoir interval due to high uncertainty (Figure. 10; Table 2).

3.4.4: Ridge model

With a lot of negative correlations, the Ridge model cannot be used or suggested for use in reservoir or non-reservoir prediction because of the overwhelming uncertainty that it shows.

Corresponding author: vo.oguadinma@deivocintegrated .com

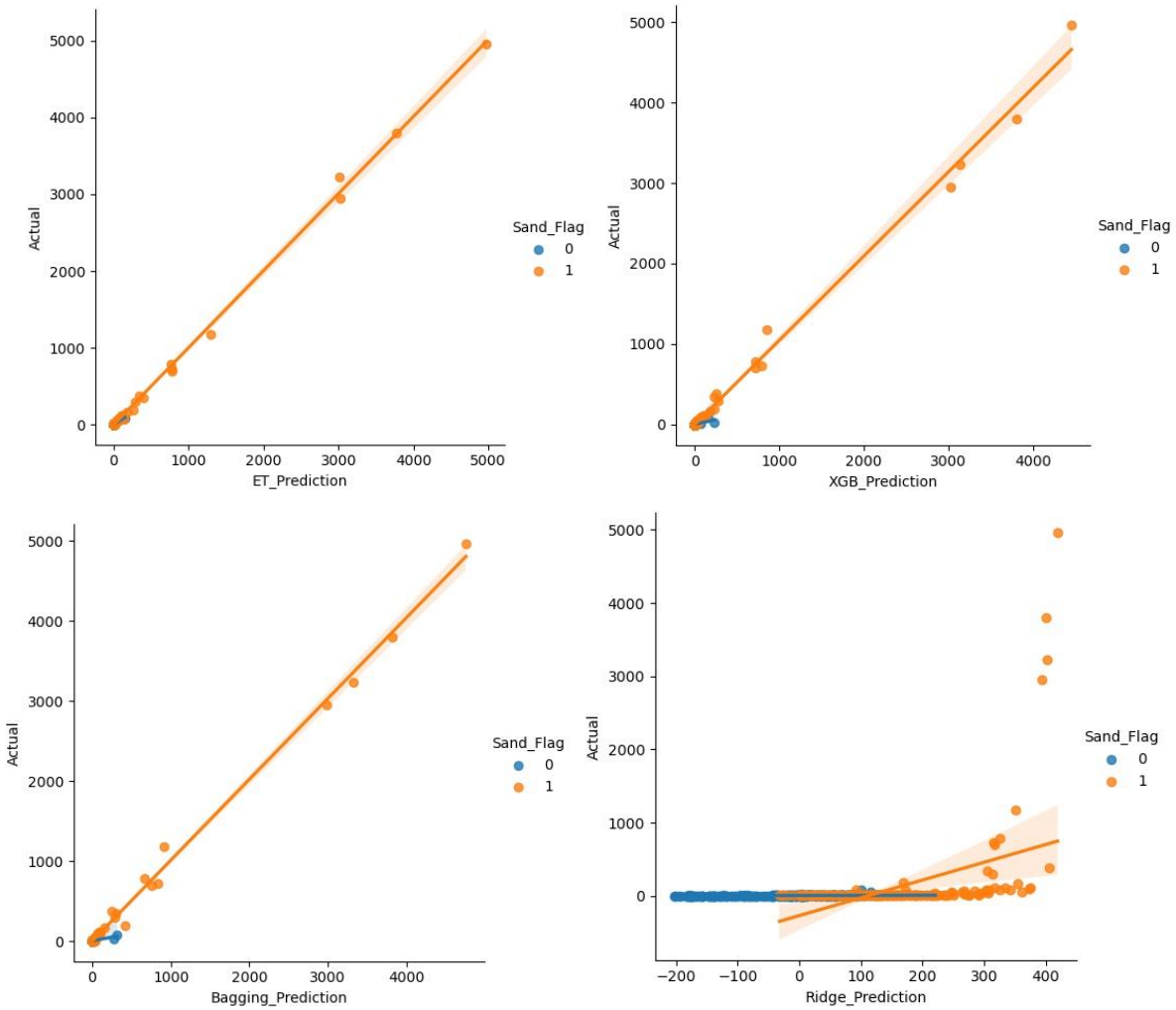

Fig. 10: These images are the results from the LM Plots extracted from Extra tree regression, XGB, Bagging and Ridge models. Observe the line of fitness for each model.

Table 2: Showing the four regression models and their value predictions. On the Sand_Flag column, the sandy reservoir section is denoted with 1 while the non-reservoir section is represented with 0.

Depth	Actual	Ridge Prediction	ET Prediction	XGB Prediction	Bagging Prediction	Sand Flag
3486.3	17.4284	137.18944	8.589059	8.7285	3.54883	1
3539.9	0.2844	135.240938	0.816454	1.477946	1.39068	1
3506.4	1.146	163.504303	0.948433	1.183534	1.16375	1
3474.1	6.7549	118.817115	11.022542	5.437419	6.98632	1
3599.9	0.001	141.339461	0.001	0.00675	0.001	0
3538.9	0.8532	140.797857	0.940918	0.027954	0.96274	1
3626.1	0.0031	-5.233577	0.00513	0.00199	0.00252	0
3428.1	73.5602	308.031912	77.659202	84.491486	63.3087	1
3477.1	1.2462	37.381466	1.373014	1.253118	1.51327	0
3545.4	0.2385	94.04899	0.461615	1.240877	0.2543	1

Corresponding author: vo.oguadinma@deivocintegrated .com

4.0: Discussion

4.1: Performance of Machine Learning Models

The primary objective of this study was to predict permeability and reservoir sections in using machine learning models and evaluate their performance. Four distinct models, including Bagging, Extra Trees Regressor, XGBoost, and Ridge, were employed in this endeavor. The results indicate that the Extra Trees Regressor model outperforms the other models in terms of accuracy and predictive power.

The Extra Trees Regressor exhibited a trained accuracy of 100%, indicating its exceptional performance in training and predicting permeability values. It achieved a high prediction accuracy of 99.8%, further affirming its robustness. The mean absolute error (MAE) of 5.7md and mean absolute percentage error (MAPE) of 8.7% demonstrate the model's accuracy in estimating permeability. Moreover, the model's AIC value, which is significantly lower than that of the other models, reinforces its suitability for permeability prediction.

Comparatively, the Bagging model also showed promise with a trained accuracy of 99% and a predicted accuracy of 94%. However, its error metrics, including MAE and MAPE, were slightly higher than those of the Extra Trees Regressor model. The Ridge model, on the other hand, exhibited underpredicting behavior with discrepancies in both shape and value when compared to the actual prediction. Its high AIC value indicates a lack of fit to the data, making it unsuitable for permeability prediction.

The XGBoost model, despite achieving high trained and predicted accuracy scores, exhibited shortcomings when evaluated based on error metrics. It attributed excessive importance to the corrected density log, which does not correlate strongly with permeability. As a result, the XGBoost model yielded a relatively high MAPE, casting doubt on the accuracy of its predictions.

The Ridge model from the table is over-predicted with values over 100 (Table 2), as such can't be used for reservoir parameter prediction.

4.2: Implications for Reservoir Characterization

The choice of the Extra Trees Regressor model as the top-performing model holds significant implications for reservoir characterization. Accurate permeability prediction is crucial for identifying reservoir intervals with high production potential (Figure 11 and 12). The Extra Trees Regressor's ability to achieve near-perfect predictions and its robustness in handling uncertainties make it a valuable tool in this regard.

The joint plots and LM plots provided visual insights into the models' performance. The Bagging model demonstrated a 90% confidence interval, indicating high uncertainty at high permeability values. This uncertainty could pose challenges in reservoir characterization, particularly when dealing with high-permeability reservoirs. The XGBoost model exhibited a 95% confidence interval with minor errors at extremely high permeability values, which are already indicative of favorable reservoir intervals. However, it introduced uncertainties due to its high MAPE.

In contrast, the Extra Trees Regressor model presented a 99% confidence interval with negligible evidence of underpredictions. This level of precision and reliability is particularly valuable for identifying and characterizing reservoir intervals. The Ridge model's low confidence interval and significant negative correlations further affirm its unsuitability for reservoir prediction.

Corresponding author: vo.oguadinma@deivocintegrated .com

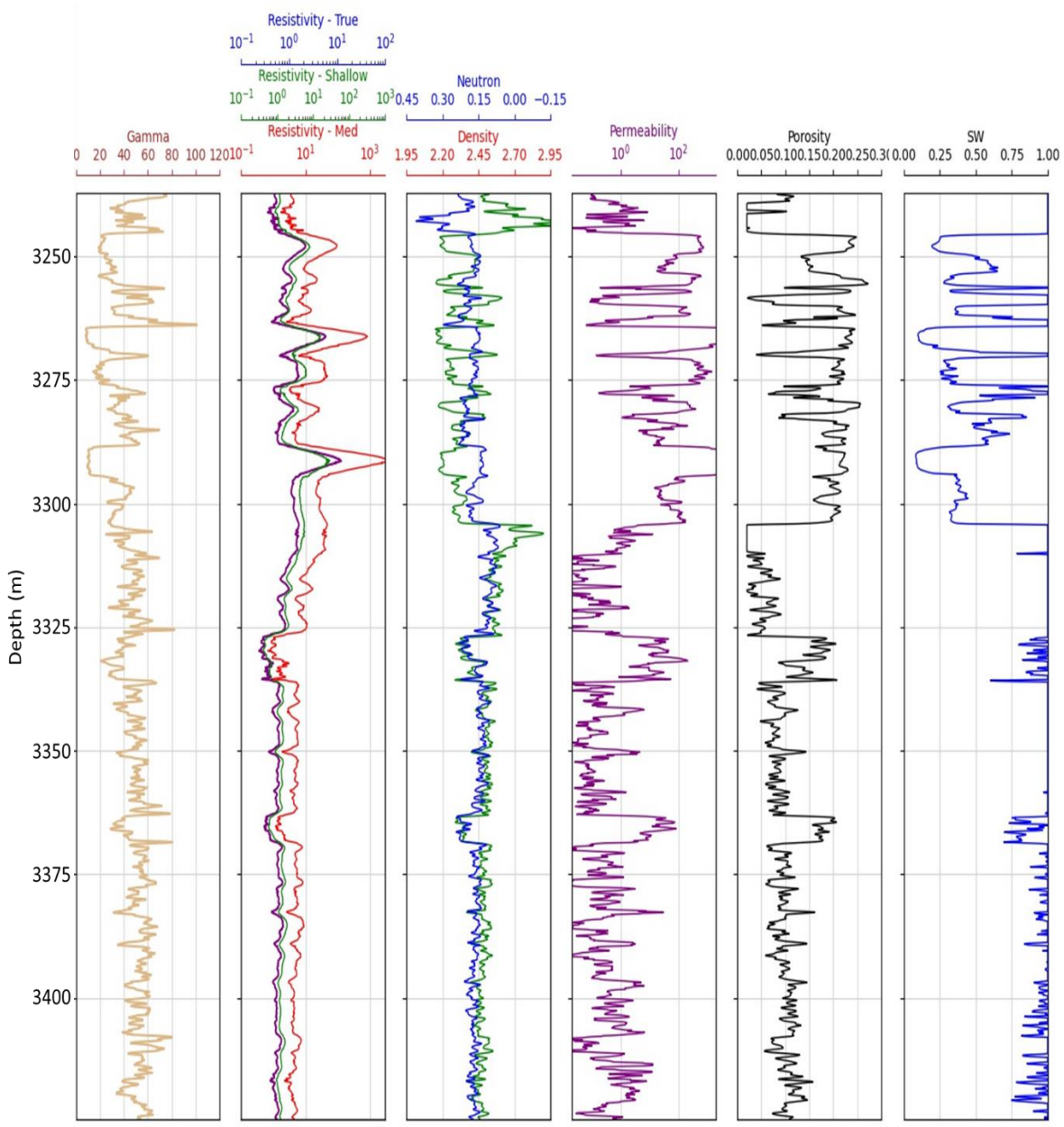

Fig. 11: Log signatures of Gamma ray, resistivity, neutron, permeability, porosity and water saturation logs used for this study and showing areas of low gamma ray reading, high resistivity, permeability, porosity reading and reas of low water saturation.

Corresponding author: vo.oguadinma@deivocintegrated .com

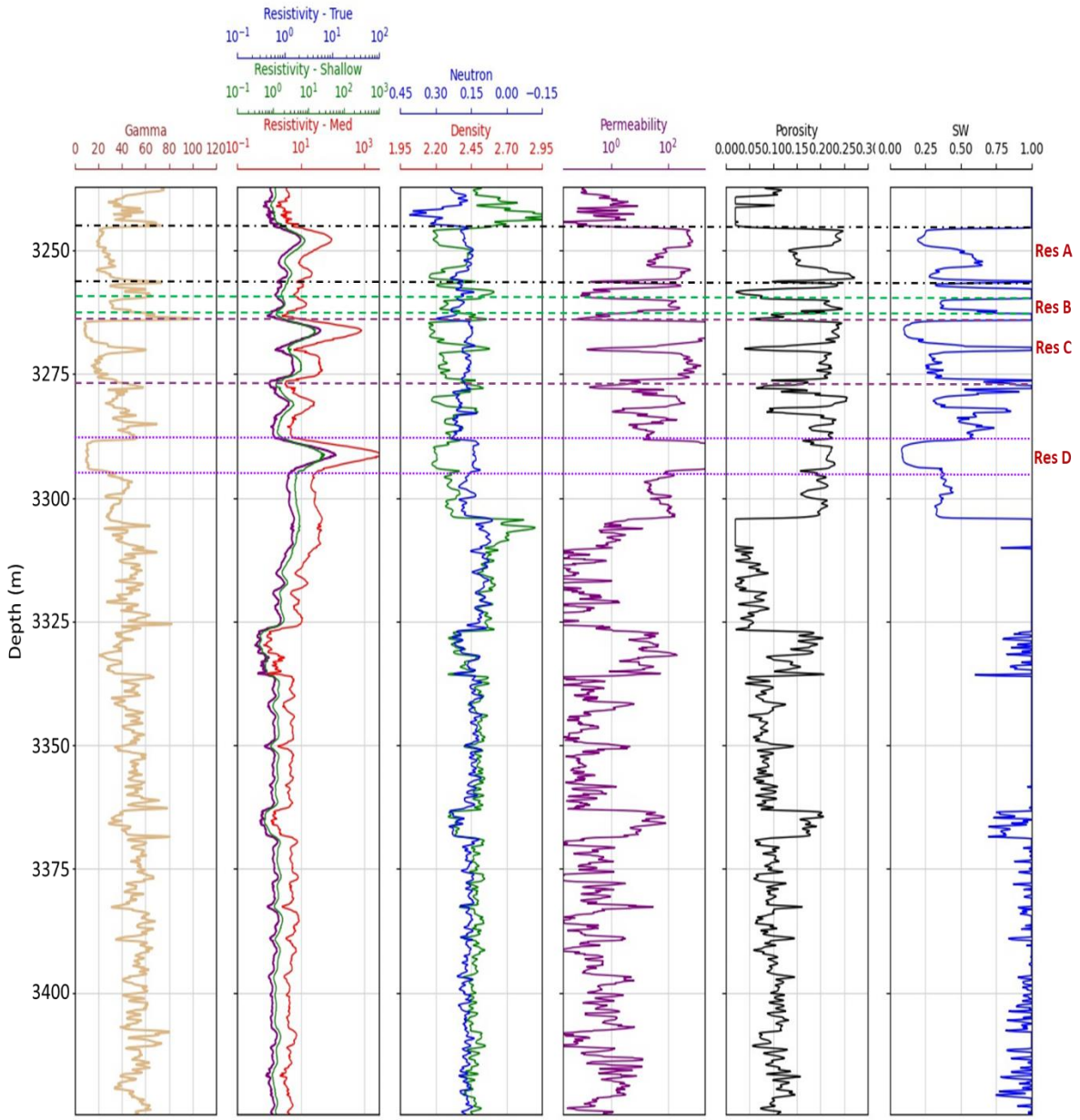

Fig. 12: Log signatures of Gamma-ray, resistivity, neutron, permeability, porosity, and water saturation log showing the identified reservoirs within the area of study which fall within low gamma-ray reading, high resistivity, permeability, porosity reading, and areas of low water saturation.

The fact that high permeability matches high porosity, low water saturation, and low gamma ray match the reservoir intervals shows that the model works well and model predictions can be trusted.

4.3: **Limitations and Future Directions**

It is essential to acknowledge the limitations of this study. The models' performance may vary in different geological settings, necessitating further validation on diverse datasets and in varying reservoir types. Additionally, while we focused on the accuracy and error metrics of the models, future research could explore interpretability and uncertainty quantification, providing deeper insights into model predictions.

Corresponding author: vo.oguadinma@deivocintegrated .com

Conclusion

The present study represents a significant step towards enhancing our understanding of reservoir parameter characterization using state-of-the-art machine learning techniques. Through the comprehensive evaluation of four machine learning models, namely Bagging, Extra Trees Regressor, XGBoost, and Ridge, we have gained valuable insights into their performance and applicability in predicting permeabilitya pivotal factor in reservoir characterization.

Our findings reveal that the Extra Trees Regressor model emerges as the most promising candidate for accurately predicting permeability. With a trained accuracy of 100% and a prediction accuracy of 99.8%, this model exhibits exceptional robustness in both the training and prediction phases. The mean absolute error (MAE) of 5.7md and mean absolute percentage error (MAPE) of 8.7% attest to its precise estimations, while the significantly lower AIC value compared to other models underscores its superior fit to the data.

Conversely, the Bagging model demonstrates a strong performance with a trained accuracy of 0.99 and a predicted accuracy of 0.94. While it shows potential for permeability prediction, its slightly higher MAE and MAPE values suggest a marginally reduced accuracy compared to the Extra Trees Regressor model. The XGBoost model, despite its impressive trained and predicted accuracy, faces challenges due to its disproportionate emphasis on the corrected density log, which exhibits a weaker correlation with permeability. Consequently, the model yields a relatively high MAPE, casting doubt on its reliability.

In contrast, the Ridge model falls short in multiple aspects, displaying significant discrepancies in shape and value when compared to actual predictions. Its high AIC value indicates an inadequate fit to the data, rendering it unsuitable for permeability prediction.

These findings hold paramount significance for reservoir characterization endeavors. Accurate permeability prediction is essential for identifying reservoir intervals with high production potential. The Extra Trees Regressor model, with its near-perfect predictions and robustness against uncertainties, stands out as an invaluable tool for reservoir characterization in this context. The visual insights provided by joint plots and LM plots further reaffirm its precision and reliability.

This study underscores the transformative potential of machine learning models, particularly the Extra Trees Regressor, in permeability prediction for reservoir characterization. Accurate permeability assessments lay the foundation for identifying reservoir intervals with high production prospects, contributing to the optimization of hydrocarbon exploration and production endeavors. As we continue to delve deeper into the applications of machine learning in geosciences, the insights gleaned from this research pave the way for more sophisticated and reliable reservoir characterization techniques.

References

Aigbe A., Delhaye-Prat V., Ujowundu T., (2008). Complex Submarine Canyon Fills in the Ancient Record of Onshore Niger Delta. NAPE Conference, Abuja November 16 - 21.

Akande, K., Owolabi, T., and Olatunji, S. (2015). Investigating the Effect of Correlation-Based Feature Selection on the Performance of Neural Network in Reservoir Characterization[J]. J. Nat. Gas Sci. Eng. 27, S1875510015301074. doi:10.1016/j.jngse.2015.08.042

Aniwetalu, et al. Spectral analysis of Rayleigh waves in the Southeastern part of Niger Delta, Nigeria. Int J Adv Geosci. 2018;6:51-6. Available: http://dx.doi.org/10.14419/ijag.v6i1.8776

Breiman, L. (2001). Random forest. Mach. Learn. 45, 5–32. doi:10.1023/a:1010933404324

Corresponding author: vo.oguadinma@deivocintegrated .com

Busari Olarewaju, Thankgod Ujowundu, Vincent Delhaye-Prat, Christian Seyve (2008). Modifications induced by growth-faults on the sequence stratigraphic architecture of the Miocene series from the Niger Delta, Nigeria. NAPE Conference, Abuja November 16 - 21.

Chakure, Afroz (2019). Machine learning data pre-processing technique. https://hardtasksin.wordpress.com/2019/06/14/data-preprocessing/

Delhaye-Prat, V., Ladipo, K., Aigbe, A., Ujowundu, T., Busari, O., & Seyve, C. (2009). Seismic Geomorphology of Ancient Submarine Canyon Systems – Implication for Prospectivity of the Niger Delta, Nigeria. doi:https://doi.org/10.3997/2214-4609.201400564

Hadi, F., and Sadegh, K. (2016). Prediction of Porosity and Water Saturation Using Pre-stack Seismic Attributes: a Comparison of Bayesian Inversion and Computational Intelligence Methods[J]. Comput. Geosciences 20 (5), 1075–1094. doi:10.1007/s10596-016-9577-0

Ibekwe , K. N., M. Arukwe , C., Ahaneku , C. V., Onuigbo , E., Omoareghan , J. O., Lanisa , A., & Oguadinma , V. O. (2023). The Application of Seismic Attributes in Fault Detection and Direct Hydrocarbon Indicator in Tomboy Field, Western-Offshore Niger Delta Basin. Journal of Energy Research and Reviews, 14(2), 9–23. https://doi.org/10.9734/jenrr/2023/v14i2279

Ibekwe KN, Arukwe C, Ahaneku C, et al. Enhanced hydrocarbon recovery using the application of seismic attributes in fault detection and direct hydrocarbon indicator in Tomboy Field, western-Offshore Niger Delta Basin. Authorea; 2023.

Ibekwe KN, Oguadinma VO, Okoro VK, Aniwetalu E, Lanisa A, Ahaneku CV. Reservoir characterization review in sedimentary basins. Journal of Energy Research and Reviews. 2023;13(2):20-28.

Joshua Pwavodi, Ibekwe N. Kelechi, Perekebina Angalabiri, Sharon Chioma Emeremgini, Vivian O. Oguadinma, Pore pressure prediction in offshore Niger delta using data-driven approach: Implications on drilling and reservoir quality, Energy Geoscience, Volume 4, Issue 3, 2023.

Komarialaei, H., and Salahshoor, K. (2012). The Design of New Soft Sensors Based on PCA and a Neural Network for Parameters Estimation of a Petroleum Reservoir[J]. Liquid Fuels Tech. 30 (22), 12. doi:10.1080/10916466.2010.512899

Li, T., Song, H., Wang, J., Wang, Y., and Killough, J. (2016). An Analytical Method for Modeling and Analysis Gas-Water Relative Permeability in Nanoscale Pores with Interfacial Effects. Int. J. Coal Geology. 159, 71–81. doi:10.1016/j.coal.2016.03.018

Nwaezeapu VC, Tom IU, David ETA, Vivian OO. Hydrocarbon Reservoir Evaluation: a case study of Tymot field at southwestern offshore Niger Delta Oil Province, Nigeria. Nanosci Nanotechnol. 2018;2(2).

Oguadinma et al. An integrated approach to hydrocarbon prospect evaluation of the Vin field, Nova Scotia Basin. S.E.G. technical program expanded Abstracts. International Exposition and Annual Meeting, Dallas, Texas. 2016;99-110. Available:10.1190/segam2016- 13843545.1

Oguadinma et al. Lithofacies and Textural Attributes of the Nanka Sandstone (Eocene): proxies for evaluating the Depositional Environment and Reservoir Quality. J Earth Sci Geotech Eng. 2014;9660:4(4)2014:1-16ISSN: 1792-9040 (print).

Oguadinma et al. Study of the Pleistocene submarine canyons of the southeastern Niger delta basin: tectonostratigraphic evolution and infilling Conference/Reunion des sciences de la Terre, Lyon, France; 2021.

Oguadinma et al. The art of integration: A basic tool in effective hydrocarbon field appraisal, Med-GU Conference, Istanbul. Turkey; 2021.

Oguadinma V, Okoro A, Reynaud J, Evangeline O, Ahaneku C, Emmanuel A et al. The art of integration: A basic tool in effective hydrocarbon field appraisal. Mediterranean Geosciences Union Annual Meeting; 2021.

Corresponding author: vo.oguadinma@deivocintegrated .com

Oguadinma VO, Aniwetalu EU, Ezenwaka KC, Ilechukwu JN, Amaechi PO, Ejezie EO. Advanced study of seismic and well logs in the hydrocarbon prospectivity of Siram Field, Niger delta basin. Geol Soc Am Admin Programs. 2017;49. DOI: 10.1130/abs/2017AM-296312

Raymer, L. L., Hunt, E. R., and Gardner, J. S. (1980). "An Improved Sonic Transit Time-to-Porosity Transform," in SPWLA 21st Annual Logging Symposium, Lafayette, Louisiana, July, 1980 (OnePetro).

Song, H., Liu, C., Lao, J., Wang, J., Du, S., and Yu, M. (2021). Intelligent Microfluidics Research on Relative Permeability Measurement and Prediction of Two-phase Flow in Micropores[J]. Geofluids 2021, 1–12. doi:10.1155/2021/1194186

Song, H., Xu, J., Fang, J., Cao, Z., Yang, L., and Li, T. (2020). Potential for Mine Water Disposal in Coal Seam Goaf: Investigation of Storage Coefficients in the Shendong Mining Area. J. Clean. Prod. 244, 118646. doi:10.1016/j.jclepro.2019.118646

Song, H., Zhang, J., Ni, D., Sun, Y., Zheng, Y., Kou, J., et al. (2021). Investigation on In-Situ Water Ice Recovery Considering Energy Efficiency at the Lunar South Pole. Appl. Energ. 298, 117136. doi:10.1016/j.apenergy.2021.117136

Song, J., Gao, Q., and Li, Z. (2016). Application of Random Forests for Regression to Seismic Reservoir Prediction[J]. Oil Geophys. Prospect. 51 (6), 1202–1211. doi:10.13810/j.cnki.issn.1000-7210.2016.06.021

Ujowundu Thankgod , Vincent Delhaye-Prat, Abimbola Aigbe (2008). NAPE Conference, Abuja November 16 - 21.

Vivian OO, Kelechi IN, Ademola L, et al. Reservoir and sequence stratigraphic analysis using subsurface data. ESS Open Arch. February 09; 2023.

Vivian OO, Kelechi IN, Ademola L, et al. Submarine canyon: A brief review. ESS Open Arch; 2023.

Wang M, Feng D, Li D and Wang J (2022). Reservoir Parameter Prediction Based on the Neural Random Forest Model. Front. Earth Sci. 10:888933. doi: 10.3389/feart.2022.888933

Wang, J., Song, H., and Wang, Y. (2020). Investigation on the Micro-flow Mechanism of Enhanced Oil Recovery by Low-Salinity Water Flooding in Carbonate Reservoir. Fuel 266, 117156. doi:10.1016/j.fuel.2020.117156

Wang, J., Song, H., Rasouli, V., and Killough, J. (2019). An Integrated Approach for Gas-Water Relative Permeability Determination in Nanoscale Porous media. J. Pet. Sci. Eng. 173, 237–245. doi:10.1016/j.petrol.2018.10.017

Wyllie, M. R. J., Gregory, A. R., and Gardner, L. W. (1956). Elastic Wave Velocities in Heterogeneous and Porous Media. Geophysics 21 (1), 41–70. doi:10.1190/1.1438217

Corresponding author: vo.oguadinma@deivocintegrated .com