SciEn

Proceedings of 8th International Conference on Mechanical, Industrial and Energy Engineering 2024
02-04 January, 2025, Khulna, BANGLADESH

SciEn Conference Series: Engineering Vol. 3, 2025, pp 68-73 https://doi.org/10.38032/scse.2025.3.16

Data-Driven Approach to Predict Future Oil Production of an Oil Field Using Machine Learning Techniques

Ataharuse Samad, Istiaque Muhammad Khan, Md. Shakil Rahaman, Ahmed Sakib, Md. Ashraful Islam*

Department of PME, Chittagong University of Engineering & Technology, Chattogram 4349, Bangladesh

ABSTRACT

Determining accurate future production in the oil and gas industry is increasingly challenging with traditional methods. Decline Curve Analysis often fails to provide precise results, while Reservoir Simulation Models require detailed parameters and are time-consuming due to the constantly changing reservoir conditions during the production period. Today, industries are leveraging machine learning and extensive data from oil wells for predictive purposes, can significantly reduce operational costs and minimize negative environmental impacts. This study aims to predict future production using machine learning algorithms. Specifically, Gradient Boosting Regression, Light Gradient Boosting Regression, and Extreme Gradient Boosting Regression models were developed using the production dataset of the Norwegian Volve oil field (well NO159F-11H) within 12 parameters. These models were trained on 80% of the dataset, while the remaining 20% was reserved for testing purposes. The accuracy of the models was assessed using the coefficient of determination (R^2), which was found to be 99% for both training and testing data across all models. GBR demonstrated the lowest mean absolute error (MAE = 12.810) and root mean square error (RMSE = 17.802) compared to the other two models based on testing value. On the other hand, based on training dataset, XGBoost showed the lowest MAE (1.192) and RMSE (1.671) values. However, the results for well NO159 F-11H show that GBR outperformed the other two methods but this doesn't imply that GBR is always better than XGBoost or LightGBR in all cases. An extensive study was conducted to evaluate the predictive performance of these models, with systematic assessments and hyperparameter adjustments to reliably anticipate the well's performance.

Keywords: Future production, GBR, XGBoost, LightGBR, Volve oil field

1. Introduction

Hydrocarbon production forecasting involves estimating the final recoveries and well life to make decisions in the oil and gas sector, which are essential for flow maintenance and workover plans, future planning, and production/injection situations. Accurate production estimation is difficult to achieve because of several factors, including uncertain and complicated reservoir structures, changes in fluid properties and dynamic production behaviour [1]. Several methods to determine traditional production forecasting include volumetric, material balance, DCA and RSA. Each method requires different types of data for production forecasting and has also limitations that why different models are used for different reservoirs [2].

DCA is useful for conventional reservoirs, fitting decline curves to historical production data. However, it's less suitable for unconventional reservoirs due to inaccurate and missing data. RSA is a sophisticated tool for predicting HC production, based on extensive data including geological features, well parameters, fluid properties, and historical production data.

Compare DCA & RSA with Machine learning (ML) approach, nowadays, ML techniques are increasingly being used in the oil and gas production industry, replacing DCA and RSA for short-term forecasting. DCA is used for short-term forecasting but cannot provide accurate results for long-term predictions due to its reliance on production history data and production rate. [3]. On the other hand, RSA requires a lot of reservoir formation data, which is not always available. In a reservoir, initial conditions are not constant all the time, day by day pressure, temperature, volume and other parameters change with daily production & injection [4]. That's why need updating the simulation model through history matching becomes essential for maintaining prediction accuracy when real-time data becomes accessible. As time progresses, the gathering of huge amounts of data becomes a big issue, since the model's rising complexity makes accuracy increasingly difficult to maintain [5]. ML models can effectively capture intricate interactions among various production-influencing factors, which traditional methods may struggle to handle.

Research has indicated that ML is a useful tool for identifying subtle connections between multiple factors that affect the production of crude gas and oil [6]. Both long-term forecasting and short-term monitoring can benefit from these approaches' superior ability to capture complicated reservoir dynamics, resulting in more precise forecasts [1]. However, ML models have limitations that need to be understood for reliable predictions. An important limitation is that most of the ML models are heavily reliant on large volumes of quality data. Poor datasets bring down the accuracy level in predictions. Also, there is a high tendency of ML models to overfit the training dataset, performing well on seen conditions but poorly on unseen ones. Furthermore, HC production potential from newly discovered wells has been predicted using competitive-learning-based networks, proving the flexibility of ML in solving particular problems.

*Corresponding Author Email Address: ataharusesamad.150618@gmail.com Published By: SciEn Publishing Group

Additionally, the productivity of shale gas reservoirs has been actively analyzed through the use of ML approaches [6].

2. Literature Review

There are many researchers are work on a lot of machine learning model techniques in petroleum engineering sector and production prediction is one of them. There are a lot of machine learning algorithm but all of them are not perfectly work on the oil and gas reservoir [9]. Some of them are provide good accuracy for gas reservoir and some are for oil reservoir.

Q. Cao et al. (2016) focused on data-driven production forecasting using Deep Learning (DL) model, Artificial Neural Networks (ANN) to predicted the production forecasting for existing and new wells using geological data and past production data. They challenged that the model was better for efficient well placement and production estimation than traditional well evaluation processes [8].

Pejman Shoeibi Omrani et al. (2019) discussed deep learning and hybrid approaches applied to short and long-term production forecasting, utilizing different gas field data between a few weeks to several years. The models they employed DCA, Physical model (IPR+VLP), ANN, and Hybrid approach based on Short-term forecasts, Mid-term forecasts & Long-term forecasts. The best accuracy was shown by the ANN model including choke opening information for both 6-month and 1-year production forecasts [1].

Cheng Zhan et al. (2019) identified that Long Short-Term Memory (LSTM) method for production forecasting in unconventional resources, requiring minimal historical data. A total of 300 wells with over two years of production history were selected, and the first three months of data were utilized to train the model and forecast output for the next twenty-one months [7].

Maryam Bagheri et al. (2020) presented new ML techniques such as MLP and SVR for predicting missing data, PCA for identify the essential features and LSTM & SVR for production prediction. In that work, 6% of outliers were eliminated, and over 60% of the anomalous and missing data is effectively identified and imputed [2].

Eduardo Andrés Muñoz Vélez (2020) utilised ANN and GB algorithms to create a model for selecting the optimal EOR method and predicting heavy oil production. They predicted that the model could be experienced for the gas industry [8].

A complete ML technique for forecasting shale gas production using geological and operational factors was developed by Gang Hui (2021). As input variables, there are 13 geological and operational factors derived from well logging, core experiment, and treatment data; the target variable has been set to be the 12-month shale gas production, predicted the production forecast using NN, Extra Trees, GBDT, and LR [6].

2.1 Summary and Implications

The discussion provides good knowledge about future production prediction with ML and DL approach. But all models do not apply to any type of reservoir because the parameters and conditions of conventional and unconventional reservoirs are not the same due to geologic features and formation conditions. The authors figure out in their research which model provides the best accuracy for future prediction by taking other reservoir parameters. Normally, when dealing with massive data in the production phase, we occasionally are unable to decipher the information extracted from the data after seeing it and appropriate AI models are used in the situation [9]. However, DL models are trained using very large and complex data so that the model may not overfit. Computationally expensive, it takes a lot of time while training, and requires special hardware. DL techniques can be better applied to unstructured data, like images or text. While comprehensive, the dataset used in this study does not meet the size and complexity required to fully utilize the potential of deep learning models. Considering all the statements from different points of view, selecting Gradient Boosting regression (GBR), Extreme Gradient Boosting Regression (XGBoost) and Light Gradient Boosting Regression (LightGBR) models to fulfill the purpose of this study. These models fit small datasets, and hence this research will use these types of models. Boosting models can also be used in showing feature importance, which is very vital for interpretability in oil and gas studies where clear results guide decisions. They are computationally efficient, taking less time and resources compared to DL models. The main purpose of this study is to conduct a comparative study based on evaluation and identify the best predictive performance model for production forecasting. Based on an 8:2 ratio, the data points are separated into two sets: the training set and the testing set. This suggests that 10713 of the 13392 data points will be used to train the data-driven models, while the other 2679 data points will be used for training purposes in order to evaluate models' prediction capabilities.

3. Methodology

Data preparation is an essential stage in data analysis processes. It includes handling missing data, data cleaning, feature selection, and normalization to make raw data more appropriate for analysis. It is necessary to remove unnecessary observations, correct structural flaws, and manage undesired outliers [10]. Here, a working flow diagram is drawn in Fig.1 to understand the working flow of the selective model. The selective three model methodology are described below.

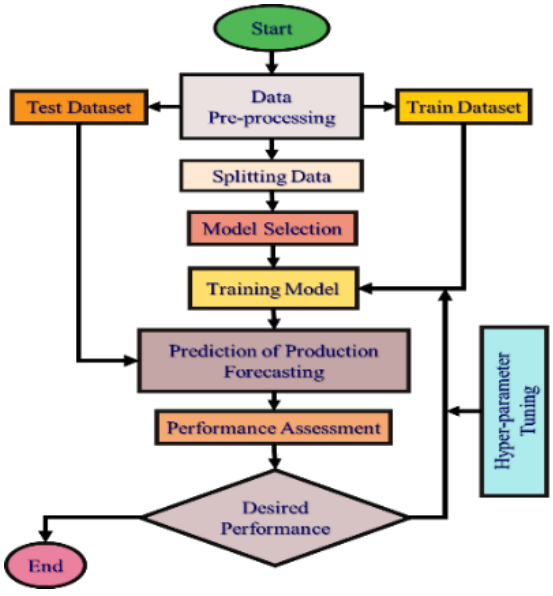

Fig.1 Working Flow Diagram

3.1 Gradient Boosting Regression (GBR)

GBR is a powerful ML technique that has been used for regression tasks. The principle behind it is to combine

several weak learners, usually decision trees, into a strong predictive model. The process initiates with an initializing simple model having an average prediction of the target variable. Further, it defines the loss function that quantifies error in the prediction. In each round, a weak learner is fitted to predict the residuals and the overall model is updated by adding the predictions of the new learner to the existing model, controlled by a learning rate parameter. This continues for some predefined number of iterations or until performance on some holdout validation set stops improving [11]. GBR $F_n(x_t)$ is defined as the sum of n regression trees.

$$F_n(x_t) = \sum_{i=1}^{n} f_i(x_t) \quad (1)$$

where every $f_i(x_t)$ is a regression tree. The succession of trees is constructed successively by estimating the new regression tree $f_{n+1}(x_t)$ using the equation below:

$$argmin \sum_t L(y_t . F_n(x_t) + f_{n+1}(x_t)) \quad (2)$$

Where $L(.)$ is differentiable for loss-function $L(.)$.

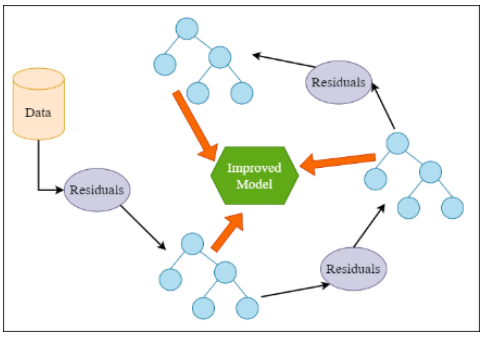

Fig.2 Working flow diagram of GBR model.

3.2 Light Gradient Boosting Regression (LightGBR)

LightGBM is a gradient boosting decision tree algorithm applied for classification, regression, and ranking tasks. It is very good at residual value modeling and prediction with high accuracy and efficiency, preferred in those machine learning applications that need precision data processing. The two most important techniques involved in this algorithm to boost the performance are the Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) [12]. These combined methods make the outputs of LightGBM fast and accurate. It will be capable of handling big data, maintaining parallelism, and ensuring satisfactory accuracy [13].

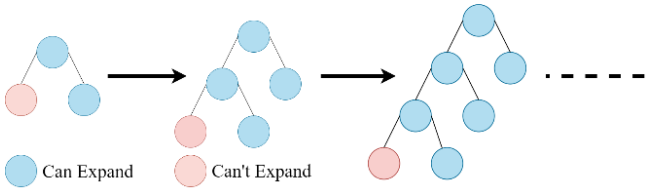

Fig.3 Leaf-wise tree growth in LightGBR

LightGBR ensembles multiple weak regressors to make one strong regressor, where each weak regressor represents distinct features, which lets us know how much each feature is affecting the prediction result. Due to this reason, LightGBR has very good interpretability.

3.3 Extreme Gradient Boosting Regression (XGBoost)

XGBoost is a ML approach to solve regression and classification problems by using a predictive model in the form of a decision tree. It can be distinguished among various implementations of Tree Gradient Boosting [14].

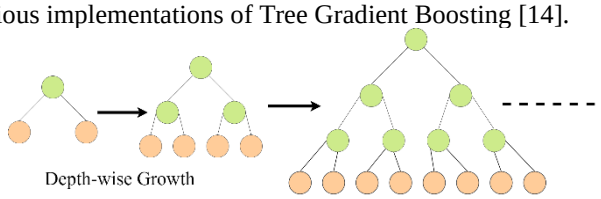

Fig.4 Basic structure of XGBoost tree model.

The main idea in XGBoost is an extension of the basic boosting idea, namely iteratively adding weak trees with different weights while improving the model step by step. The set of trees must approach the residuals of the past forecast as much as feasible, which is shown as follows.

$$\hat{y_t} = \sum_{k=1}^{k} f_k(x_i) \qquad f_k \in F \quad (3)$$

where F is the function space that contains all of the regression trees, x_i is the i-th training sample, and f_k is the score for the k-th tree. It is anticipated that the projected value $\hat{y_t}$ will approximate the genuine valuey_t to the greatest extent feasible while maintaining its capacity for generalization [15]. Below is the formula to calculate objective (L).

$$L = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega(f_t) + constant \quad (4)$$

The loss function, or difference between the predicted and true values, is represented by $l\left(y_i, \hat{y}_i^{(t)}\right)$ in the equation. It can be Any second-order derivable loss function can be used. $\Omega(f_t)$ defines the complexity of the model. The complexity decreases and the generalization ability increases with a decreasing value of $\Omega(f_t)$.

$$\Omega(f_t) = \gamma \mathrm{T} + \frac{1}{2}\lambda\|\omega\|^2 \quad (5)$$

Where γ and λ are constant coefficients, T controls number of tree leaves, ω controls the score of each leaf. Since XGBoost extends the loss function using second order approximation and eliminates the constant component to obtain the simplest goal [16]. As a result, XGBoost is flexible enough to handle many issues and allows faster operation speed and significantly shorter training time.

3.4 Model Performance Matrix

Here, we calculate R-square value (R^2), Mean absolute error (MAE), Root mean squared error (RMSE) to identify the model performance.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (6)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (7)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (8)$$

4. Result and Discussion

The dataset used for predict production in this study is oil production data (well NO159F-11H) of Volvo oil field in Norwegian region. The including parameter are recorded date, on_steam hour, average downhole pressure &

temperature, average differential pressure tubing & annulus pressure, average Choke-size percentage, average wellhead pressure and temperature, choke-size, produced oil, gas & water volume.

Data visualization of machine learning to represent how dataset are correlated with each other. It helps us to identify which parameter and how much influence to the target feature. A data visualization technique called heat-map by Fig.5 to identify the correlation between the input feature to target feature. From heat-map analysis, we can see that the value of coefficient is 0.99 for bore gas volume represent highly correlated with bore oil volume. But bore water volume in not highly correlated with bore oil volume because of -0.35 coefficient value.

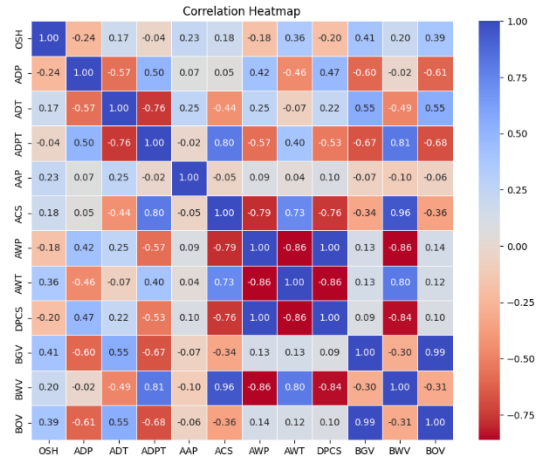

Fig.5 Heat-map representation of correlation between all data-point.

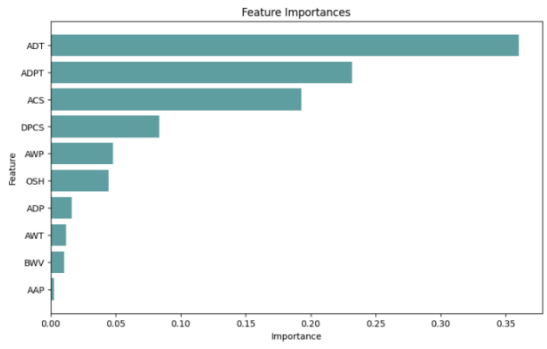

Fig.6 Important feature identification.

To improve model performance, feature importance helps to identify the most important relevant feature from the data set. Due to highly correlated gas production with oil that's why it's not taken into importance identification. This technique helps us to determine the accurate prediction and desired output. The feature importance graph is shown in the Fig.6. Here, Fig.6 showed importance is almost zero for average annulus pressure indicating the zero contribution to the oil production from the reservoir.

Table 1 Evaluation Parameter of three models for training and testing datasets.

Datasets	Matrix	GBR	LightGBR	XGBoost
Training	R^2	0.9999	0.9998	0.9999
	RMSE	2.577	5.621	1.671
	MAE	2.042	3.635	1.192
Testing	R^2	0.9978	0.9977	0.9974
	RMSE	17.802	18.512	19.534
	MAE	12.810	13.306	13.925

The performance matrix of three models are listed in Table 1. The table is represented that all model are strongly performed during the training and testing time. The R-square value was found approximately 99 percent for three models on both training and testing dataset.

The cross-plots comparing actual and predicted oil production for GBR, LightGBR, and XGBoost are shown in Figures 7a, 7b, and 7c, respectively. The red dashed line represents a 45° slope, indicating the ideal line of perfect predictions for all three models. In general, most data points align closely with this line, showing high accuracy across the models.

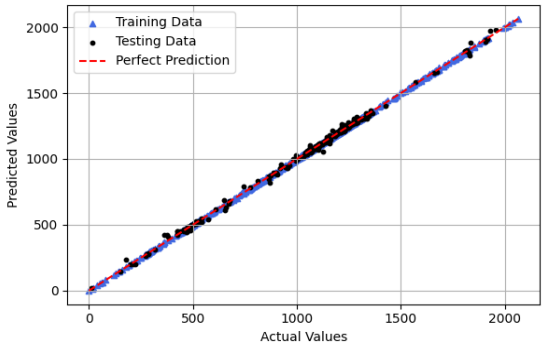

Fig.7a Cross-plot of actual vs predicted oil production for training and testing dataset (GBR)

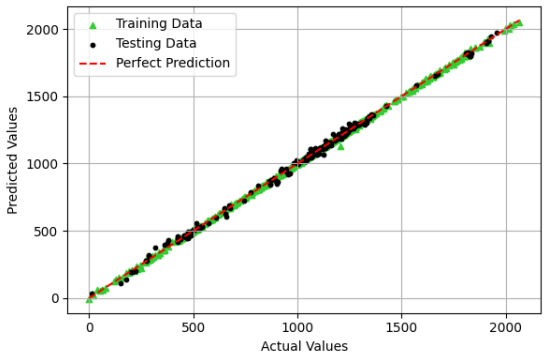

Fig.7b Cross-plot of actual vs predicted oil production for training and testing dataset (LightGBR)

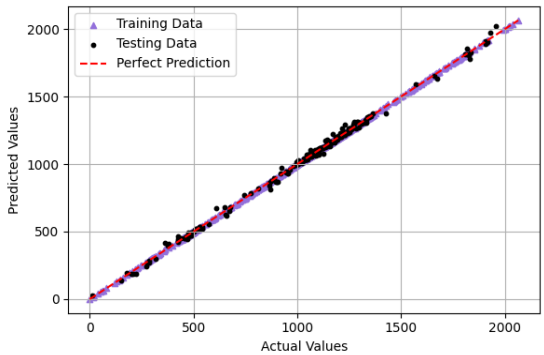

Fig.7c Cross-plot of actual vs predicted oil production for training and testing dataset (XGBoost)

In Figures 7a–7c, the training data points are mostly aligned along the 45° line, which indicates strong training performance for all three models with minimal over- or underestimation. However, Figure 7c (XGBoost) shows a few outliers in the testing data, where some predictions are underestimated - the points that lie below the perfect prediction line and some are overestimated - the points that lie above the perfect prediction line. Although these outliers slightly impact XGBoost's overall testing performance, they

do not significantly affect the model's reliability compared to GBR and LightGBR. Overall, the cross-plots confirm that the training and testing performance of all three models is excellent, with no significant outliers detected.

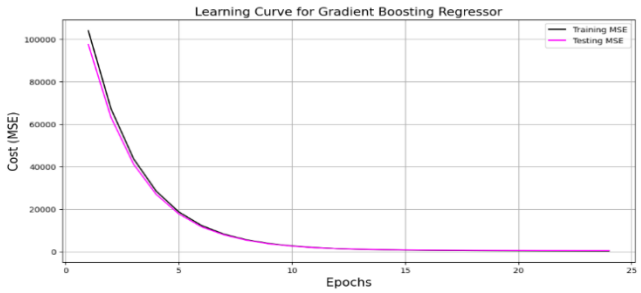

Fig.8a Epoch vs Cost curve for the GBR

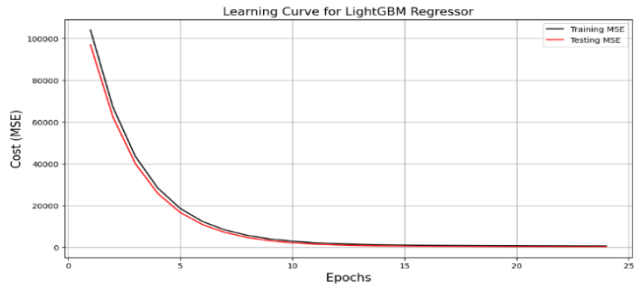

Fig.8b Epoch vs Cost curve for the LightGBR

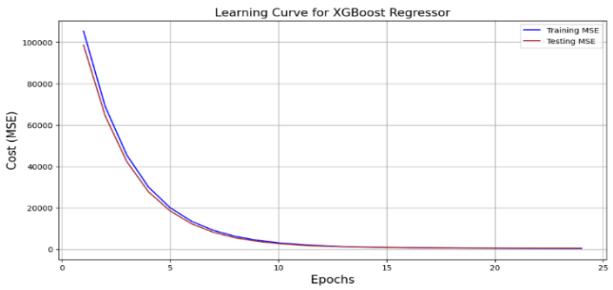

Fig.8c Epoch vs Cost curve for the XGBoost

GBR, LightGBR, and XGBoost learning curves (Figures 8a−8c respectively) show clear trends in model performance over time. Mean Squared Error (MSE) decreases quickly in the first epochs for all three models, then plateaus as the models converge. The training and testing curves are well aligned, and GBR has the lowest final MSE, indicating the best generalization. Contrarily, LightGBM demonstrates a fast reduction in errors in the early epochs and competitive performance in the end, indicating its superior learning speed. Despite being more effective than the other two models, XGBoost converges more slowly and has higher MSE values.

The bar graph in Fig.9 illustrates the performance of different models and helps identify the best model for this dataset. The difference between the training and testing R-squared values for both GBR) and LightGBM is 0.0021, which is smaller than that observed for XGBoost. However, XGBoost achieves the lowest RMSE and MAE during the training phase. In contrast, GBR demonstrates the lowest RMSE and MAE during the testing phase. The Difference in the R-square value for the training and testing dataset is very low, which indicates that these models are run without overfitting.

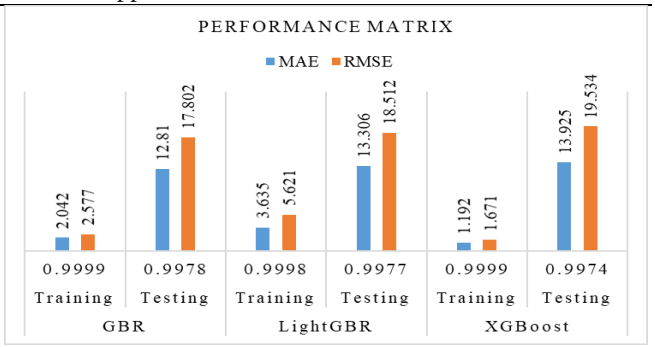

Fig.9 Visualization of Evaluation parameters

The consistent performance of the GBR model across both training and testing sets highlights its robustness and reliability is understood from this figure.

Optimum hyper-parameter tuning like grid search algorithm are used in this study which help to identify this best performance. The hyper-parameter tuning listed on Table 2 and their corresponding value.

Table 2 Hyper-parameter tuning of three models.

Hyper-parameter Tuning	GBR	LightGBR	XGBoost
objective		regression	Squared_error
loss	Squared_error		
n_estimators	220	300	70
Learning_rate	0.2	0.2	0.3
max_depth	4	7	7

To compare the models' predictability, Fig.10 display the actual versus predicted oil production.

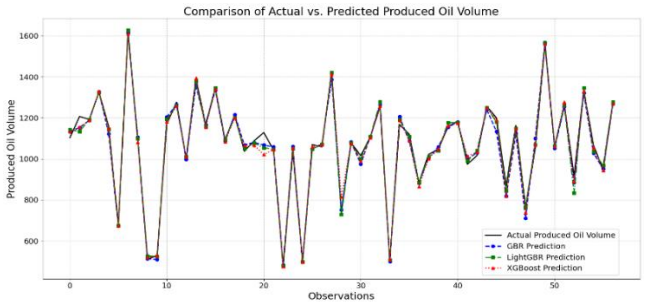

Fig.10 Comparison of actual and predicted oil production for three models.

In these figures, the black line represents actual oil production, while the three dashes-colored lines represent the predicted oil production from three different models respectively, based on daily records. Upon close examination, the predicted lines almost overlap with the actual production line. However, a deeper analysis reveals fluctuations between the actual and predicted lines. The GBR model shows the least fluctuation, whereas the LightGBR and XGBoost models exhibit the more. Based on this comparison, we can conclude that the GBR model is the most accurate for this dataset.

5. Conclusion

In this paper, the primary goal was to identify the best ML technique for predicting oil production using historical data and key input parameters. The analysis of GBR, LightGBR, and XGBoost models reveals their strengths and limitations in predicting oil production. The heat-map analysis shows a high correlation between bore gas volume

and oil volume, while bore water volume shows a low correlation. GBR and LightGBR show strong training and testing performance, achieving approximately 99% R-squared values. Cross-plots show high accuracy with GBR and LightGBR aligning closely with the ideal 45° line. XGBoost, while accurate, has some outliers where predictions are slightly underestimated or overestimated, affecting its reliability. Learning curve analysis reveals that GBR achieves the lowest final MSE, showcasing excellent generalization. LightGBR demonstrates rapid initial error reduction and competitive performance, indicating its superior learning speed. XGBoost converges more slowly and has higher MSE values, reflecting a need for more tuning to enhance generalization. GBR is the most reliable model due to its superior testing accuracy, robust generalization, and stable predictions, while XGBoost has higher testing RMSE. Therefore, GBR is recommended as the best model for predicting oil production.

6. Acknowledgement

First of all, I express my endless thanks to Almighty ALLAH for His great blessings on me and my parents for unwavering support, encouragement and also their intellectual ideas and regular guidance.

References

[1] P. S. Omrani *et al.*, "Deep Learning and Hybrid Approaches Applied to Production Forecasting," *Abu Dhabi International Petroleum Exhibition & Conference*, Nov. 2019.

[2] M. Bagheri *et al.*, "Data Conditioning and Forecasting Methodology using Machine Learning on Production Data for a Well Pad," *Offshore Technology Conference*, May 2020.

[3] D. Han, S. Kwon, H. Son, and J. Lee, "Production forecasting for shale gas well in transient flow using machine learning and decline curve analysis," *Proceedings of the SPE/AAPG/SEG Asia Pacific Unconventional Resources Technology Conference*, Jan. 2019.

[4] Y. Li and Y. Han, "Decline Curve Analysis for Production Forecasting Based on Machine Learning," *SPE Symposium: Production Enhancement and Cost Optimisation*, Nov. 2017.

[5] Q. . Cao, R. . Banerjee, S. . Gupta, J. . Li, W. . Zhou, and B. . Jeyachandra, "Data Driven Production Forecasting Using Machine Learning," *SPE Argentina Exploration and Production of Unconventional Resources Symposium*, May 2016.

[6] E. A. M. Vélez, F. R. Consuegra, and C. A. B. Arias, "EOR Screening and Early Production Forecasting in Heavy Oil Fields: A Machine Learning Approach," *SPE Latin American and Caribbean Petroleum Engineering Conference*, Jul. 2020.

[7] C. Zhan, S. Sankaran, V. LeMoine, J. Graybill, and D.-O. S. Mey, "Application of machine learning for production forecasting for unconventional resources," *Proceedings of the 7th Unconventional Resources Technology Conference*, Jan. 2019.

[8] E. A. M. Vélez, F. R. Consuegra, and C. A. B. Arias, "EOR Screening and Early Production Forecasting in Heavy Oil Fields: A Machine Learning Approach," *SPE Latin American and Caribbean Petroleum Engineering Conference*, Jul. 2020.

[9] E. Naqa and M. J. Murphy, "What is machine learning?," in *Springer eBooks*, 2015, pp. 3–11.

[10] B. A. Juliussen, J. P. Rui, and D. Johansen, "Algorithms that forget: Machine unlearning and the right to erasure," *Computer Law & Security Review*, vol. 51, p. 105885, Sep. 2023.

[11] *Data Preprocessing in Data Mining*, 1st ed., vol. 72. springer, 2015. [Online]. Available: https://link.springer.com/book/10.1007/978-3-319-10247-4

[12] J. Cai, K. Xu, Y. Zhu, F. Hu, and L. Li, "Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest," *Applied Energy*, vol. 262, p. 114566, Feb. 2020.

[13] Shehadeh, O. Alshboul, R. E. A. Mamlook, and O. Hamedat, "Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression," *Automation in Construction*, vol. 129, p. 103827, Jul. 2021.

[14] Balamwar, R. Mitra, M. K. Tiwari, and P. Verma, "Prediction and Analysis of Seasonal Dynamic Metal Consumption using Aggregated LightGBM - A Case Study," *IFAC-PapersOnLine*, vol. 55, no. 10, pp. 725–730, Jan. 2022.

[15] L. Zhang, W. Bian, W. Qu, L. Tuo, and Y. Wang, "Time series forecast of sales volume based on XGBoost," *Journal of Physics Conference Series*, vol. 1873, no. 1, p. 012067, Apr. 2021.

[16] X. Dairu and Z. Shilong, "Machine Learning Model for Sales Forecasting by Using XGBoost," *IEEE*, Jan. 2021.

NOMENCLATURE

OSH	: On Steam Hour, hour
ADP	: Average Downhole Pressure, bar
ADT	: Average Downhole Temperature, ^{o}C
ADPT	: Average Differential Pressure Tubing, bar
AAP	: Average Annulus Pressure, bar
ACS	: Average Choke-Size
AWP	: Average Wellhead Pressure, ^{o}C
AWT	: Average Wellhead Temperature, ^{o}C
DPCS	: Differential Pressure Choke-Size
BGV	: Bore Gas Volume, sm3/days
BWV	: Bore Water Volume, sm3/days
BOV	: Bore Oil Volume, sm3/days
NN	: Neural Network.
GBDT	: Gradient Boosting Decision Tree.
LR	: Liner Regression
GBR	: Gradient Boosting Regression
LightGBR	: Light Gradient Boosting Regression
XGBoost	: Extreme Gradient Boosting Regression