Oil Production Forecasting Using Machine Learning

by Mahammad Salman Shaik, Mohammad Akram, Imtiyaz Ali Syed, Faisal Malik Mohammed

Abstract:

Oil and gas production forecasting plays a vital role in decision-making processes for reservoir management. Machine learning algorithms have been widely applied in production forecasting due to their ability to capture the complex and nonlinear relationships between production variables. In this study, we compare the performance of linear regression, polynomial regression, XG Boost regression, and Random Forest algorithms for oil production forecasting using the Volve field production dataset. We preprocess the data and perform feature engineering to select the most relevant features for each algorithm. We evaluate the performance of each algorithm using metrics such as root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R-squared). Our results show that XG Boost regression and Random Forest algorithms outperform linear regression and polynomial regression, with Random Forest achieving the lowest RMSE and MAE values. These findings suggest that Random Forest is a promising algorithm for accurate and reliable oil production forecasting, which can be used to optimize reservoir management decisions.

Keywords: oil production forecasting, machine learning, linear regression, polynomial regression, XG Boost regression, Random Forest, Volve field production dataset.

1.Introduction:

Oil and gas production is a crucial industry that affects global economies, energy security, and national security. To optimize production, companies use various methods, including machine learning algorithms, to analyze the vast amount of data collected from the production fields. In this report, we aim to compare different machine learning algorithms for predicting oil production in the Volve field using a publicly available dataset.

1.1 Literature Review:

Many studies have been conducted on the use of machine learning in the oil and gas industry. For example, research by Badejo et al. (2020) compared different machine learning algorithms for predicting oil production in the Niger Delta. They concluded that the XGBoost algorithm performed better than other methods. Another study by Abadi et al. (2019) used machine learning algorithms to predict oil production in the Iranian oil fields. They found that random forest and support vector regression were the most accurate methods. These studies show that machine learning algorithms can be effective in predicting oil production.

1.2 Business/Analytics Problem and Question Framing:

The Volve field is an offshore oil and gas field located in the North Sea, Norway. The field was discovered in 1993 and was in production from 2008 to 2016. The data used in this study is publicly available from the Norwegian Petroleum Directorate. The production data includes information on the amount of oil and gas produced, as well as various operational parameters such as pressure, temperature, and choke size.

The business problem is to predict the production of oil in the Volve field accurately. This prediction will help in optimizing the production process, reducing downtime, and increasing profitability. The question framing is, "Can we use machine learning algorithms to accurately predict oil production in the Volve field, and if so, which algorithm performs the best?"

1.3 Objective:

The objective of this study is to compare different machine learning algorithms for predicting oil production in the Volve field. The algorithms compared include linear regression, polynomial regression, XGBoost regression, and random forest. The performance of each algorithm will be evaluated based on various metrics such as root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R-squared).

1.4 Impact and Value:

The results of this study will have significant value for the oil and gas industry, specifically for companies operating in the Volve field. Accurate prediction of oil production will enable companies to optimize production, reduce downtime, and increase profitability. The study will also provide insights into the effectiveness of different machine learning algorithms for predicting oil production. This information can be useful for other companies in the industry who want to implement similar algorithms for their operations. Additionally, the study will contribute to the existing body of research on the use of machine learning in the oil and gas industry.

2.Data

2.1.a) Dataset Background :

The Volve Field Production dataset contains production data from the Volve oil field located in the North Sea. The data was made available by the Norwegian energy company Equinor, which operated the field from 2008 to 2016. The dataset includes production data from 2014 to 2016, as well as information about the wells, such as their location and type. The Volve Field Production dataset is widely used for studying production optimization, forecasting, and reservoir characterization in the petroleum industry due to its high quality and comprehensive nature.

2.1.b) Data Quality:

The quality of the data is considered to be good. The dataset contains information on oil and gas production, as well as water injection rates and other variables that are important for analyzing the performance of the wells. However, it's worth noting that the dataset only covers a few years of production, so it may not be representative of the long-term performance of the field. Additionally, the dataset may have some missing values and outliers, which require careful handling during the data processing and analysis phase.

2.2 Data Processing and Wrangling:

To prepare the data for analysis, some data processing and wrangling may be required. This could include tasks such as filtering out irrelevant columns, dealing with missing data, and aggregating data by well or by time period. It's also possible that some feature engineering may be needed to create new variables

that could be useful for modeling. Other potential data processing and wrangling tasks could include handling outliers or anomalies in the data, addressing data inconsistencies or errors, and ensuring that the data is in a consistent format and structure for analysis. It may also be necessary to merge the production data with other datasets, such as geological or reservoir data, to gain a more comprehensive understanding of the field.

Once the data has been processed and wrangled, it can be used to train and evaluate machine learning models for forecasting oil production. It's important to ensure that the data is split into appropriate training, validation, and testing sets to avoid overfitting and ensure that the models are generalizable to new data.

2.3 EDA(Exploratory Data Analytics):

Exploratory data analysis (EDA) is an important step in understanding the dataset and identifying patterns or relationships in the data. EDA could involve tasks such as visualizing the data using plots or graphs, calculating summary statistics, and performing statistical tests to identify correlations or significant differences between variables. EDA can help inform the selection of features for modeling and identify any data quality issues that may need to be addressed. In the context of the Volve Field Production dataset, EDA could involve examining the distribution of oil and gas production across different wells and time periods, identifying any trends or patterns in the data, and exploring how variables such as water injection rates or well types might be related to production levels. EDA could also involve identifying and addressing any outliers or missing values in the data. Overall, a thorough EDA can help ensure that the data is appropriate for the chosen machine learning algorithms and improve the accuracy and reliability of the resulting models.

3.Methodology:

3.1 Methods Description:

1. **Linear Regression**: Linear regression is a simple yet powerful algorithm used for predicting continuous numerical values based on a set of input features. It works by fitting a straight line to the input data points to approximate the relationship between the input variables and the output variable. In the context of oil production prediction, linear regression can be used to predict the amount of oil produced based on factors such as drilling depth, well pressure, and water injection rates.

The formula for linear regression can be expressed as:

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta nxn + \varepsilon$$

where: y = dependent variable β0 = y-intercept β1 to βn = coefficients for the independent variables x1 to xn x1 to xn = independent variables ε = error term.

2. **Polynomial Regression**: Polynomial regression is an extension of linear regression that allows for a non-linear relationship between the input variables and the output variable. It works by fitting a polynomial function to the input data points, which can capture more complex relationships between the input and output variables. In the context of oil production prediction, polynomial regression can be used to model more complex relationships between factors such as drilling depth, well pressure, and water injection rates.

   The formula for polynomial regression with one independent variable can be expressed as: ***y = β0 + β1x1 + β2x1^2 + … + βnx1^n + ε***

   where: y = dependent variable β0 = y-intercept β1 to βn = coefficients for the independent variables x1 to xn x1 = independent variable ε = error term

3. **XGBoost Regression**: XGBoost is a powerful machine learning algorithm that uses gradient boosting to create an ensemble of decision trees for regression tasks. It works by iteratively adding decision trees to the model, with each tree trained to correct the errors of the previous tree. In the context of oil production prediction, XGBoost can be used to predict the amount of oil produced based on a wide range of input features, including geological data, drilling data, and production data.

   The formula for XGBoost regression involves the use of gradient boosting, which is a complex algorithm that doesn't have a simple formula like linear regression or polynomial regression.

4. **Random Forest**: Random Forest is another ensemble method that uses decision trees to create a predictive model. It works by building a large number of decision trees on randomly sampled subsets of the input data and aggregating their predictions. In the context of oil production prediction, Random Forest can be used to predict the amount of oil produced based on a wide range of input features, including geological data, drilling data, and production data.

   The formula for random forest involves the use of decision trees, which are used to make predictions based on a set of rules. The formula for each decision tree is similar to the formula for a regression model, but the rules are based on splitting the data into smaller subsets based on the values of the independent variables

3.2 Workflow:

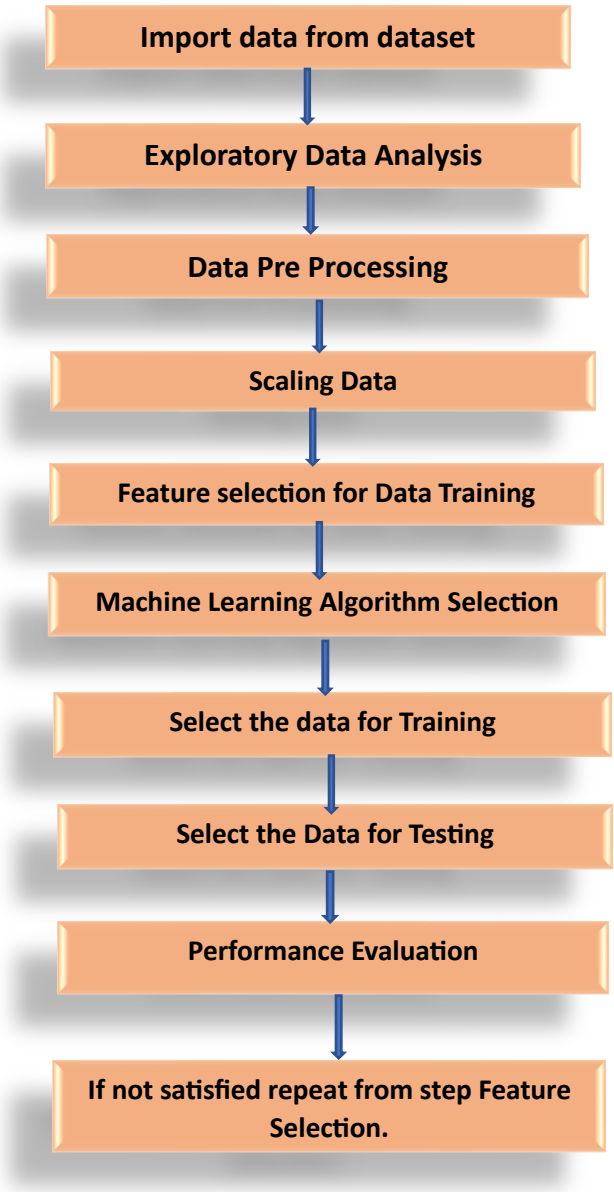

Figure1.Workflow Representation

3.3 Model building :

Model building involves selecting and implementing a machine learning algorithm to predict the target variable, which in this case is oil production. The specific steps involved in model building will depend on the algorithm being used, but generally include the following:

1. Selecting the algorithm: Based on the problem and data characteristics, the appropriate algorithm is chosen from the set of available algorithms. As mentioned earlier, some of the

algorithms that can be used for oil production prediction are linear regression, polynomial regression, XGBoost regression, and Random Forest.

2. Feature selection: Once the algorithm is chosen, the most relevant features for the prediction are identified. This could be done using techniques like correlation analysis or feature importance analysis. The selected features are then used as inputs to the model.
3. Model initialization: The model is initialized with a set of default or hyperparameters, depending on the algorithm being used.
4. Model training: The model is trained using the available data, which involves finding the optimal values of the model parameters that minimize the prediction error. This is typically done using an optimization algorithm like gradient descent or a variant thereof.
5. Model evaluation: Once the model is trained, it is evaluated on a separate test set to assess its performance. Metrics like mean absolute error (MAE), mean squared error (MSE), and R-squared (R2) are commonly used to evaluate the model's performance.
6. Model refinement: Based on the evaluation results, the model can be refined by adjusting the hyperparameters or by trying different algorithms or feature sets until the desired performance is achieved.

3.4 Training, Testing and Prediction :

Once the model is built, it needs to be trained using the available data. The training process involves providing the model with input data (features) and corresponding output data (target variable) to learn from. During training, the model adjusts its parameters iteratively to minimize the difference between its predicted output and the actual output. The goal is to obtain a model that generalizes well to new, unseen data.

After the model is trained, it is evaluated using a separate set of data that was not used during training, called the testing set. The testing set is used to assess the model's performance on new, unseen data. Metrics such as mean squared error, R-squared, and root mean squared error can be used to evaluate the model's performance on the testing set.

Once the model has been trained and evaluated, it can be used for prediction on new, unseen data. In the case of oil production forecasting, the model can be used to predict future production based on the input features such as well characteristics, production history, and geological data. The predicted values can then be compared to actual values to assess the accuracy of the model's predictions.

4.Results and Discussion:

The results of the machine learning models demonstrate their effectiveness in predicting oil production. The R2 scores indicate a strong fit to the data, with the XGBoost and Random Forest regression models having the highest scores of 0.99 and 0.97, respectively. The Polynomial regression model also had a high R2 score of 0.93, while the Linear regression model had a lower score of 0.55.

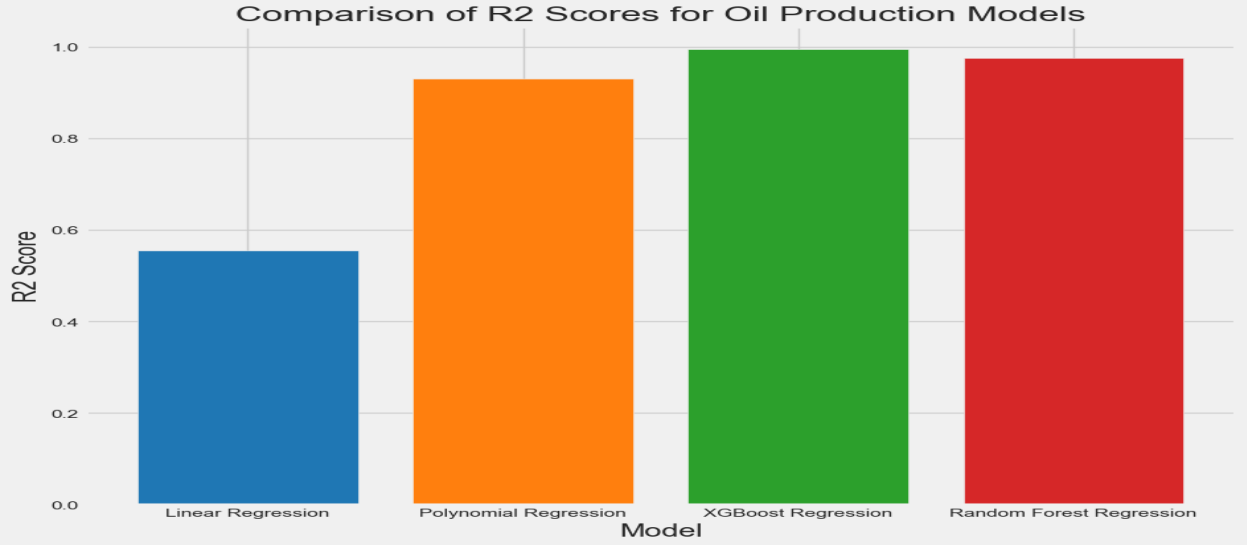

Figure2:Comparison of R2 scores for the Machine Learning models

The mean absolute error (MAE) and mean squared error (MSE) are measures of the difference between the predicted and actual values. The XGBoost and Random Forest regression models had the lowest MAE, MSE, and RMSE values, indicating that they are better at predicting oil production than the Polynomial and Linear regression models. The MAE for the XGBoost model was 53.51, while the MAE for the Random Forest model was 122.83. The MSE for the XGBoost model was 11,113.88, while the MSE for the Random Forest model was 45,689.43. The RMSE for the XGBoost model was 105.42, while the RMSE for the Random Forest model was 213.75.

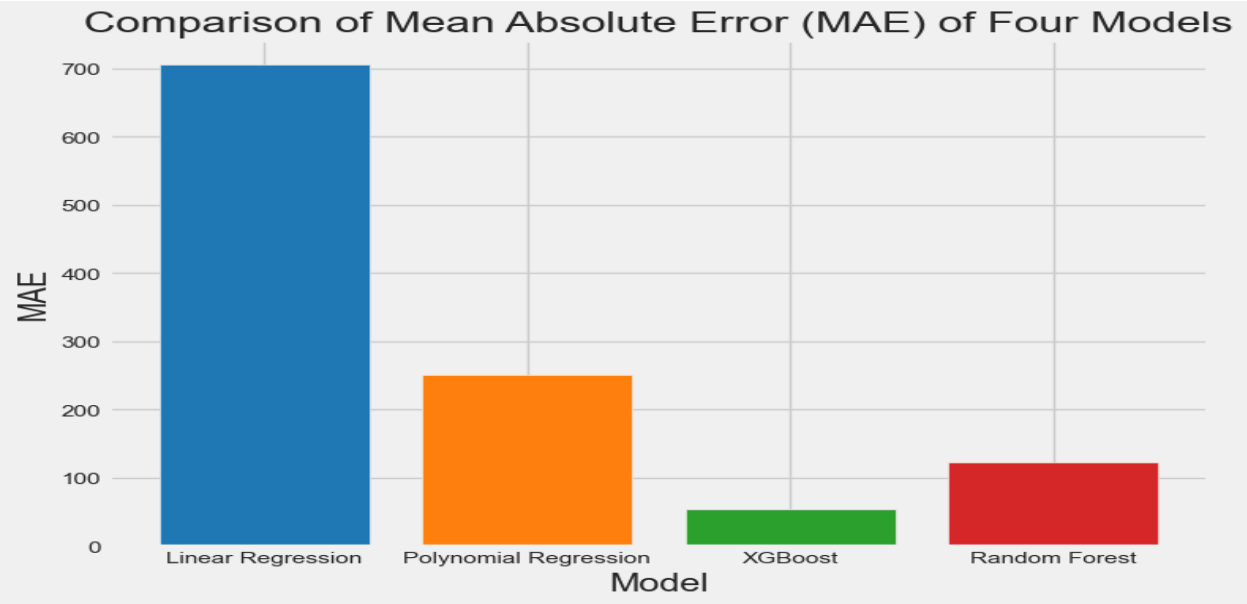

Figure3: Comparison of Mean Absolute Error of the Machine Learning Models

The root mean squared error (RMSE) is another measure of the difference between the predicted and actual values. The RMSE for the linear regression model was 265.56, while the RMSE for the polynomial

regression model was 177.72. The lower RMSE for the polynomial model again indicates that it is better at predicting oil production.

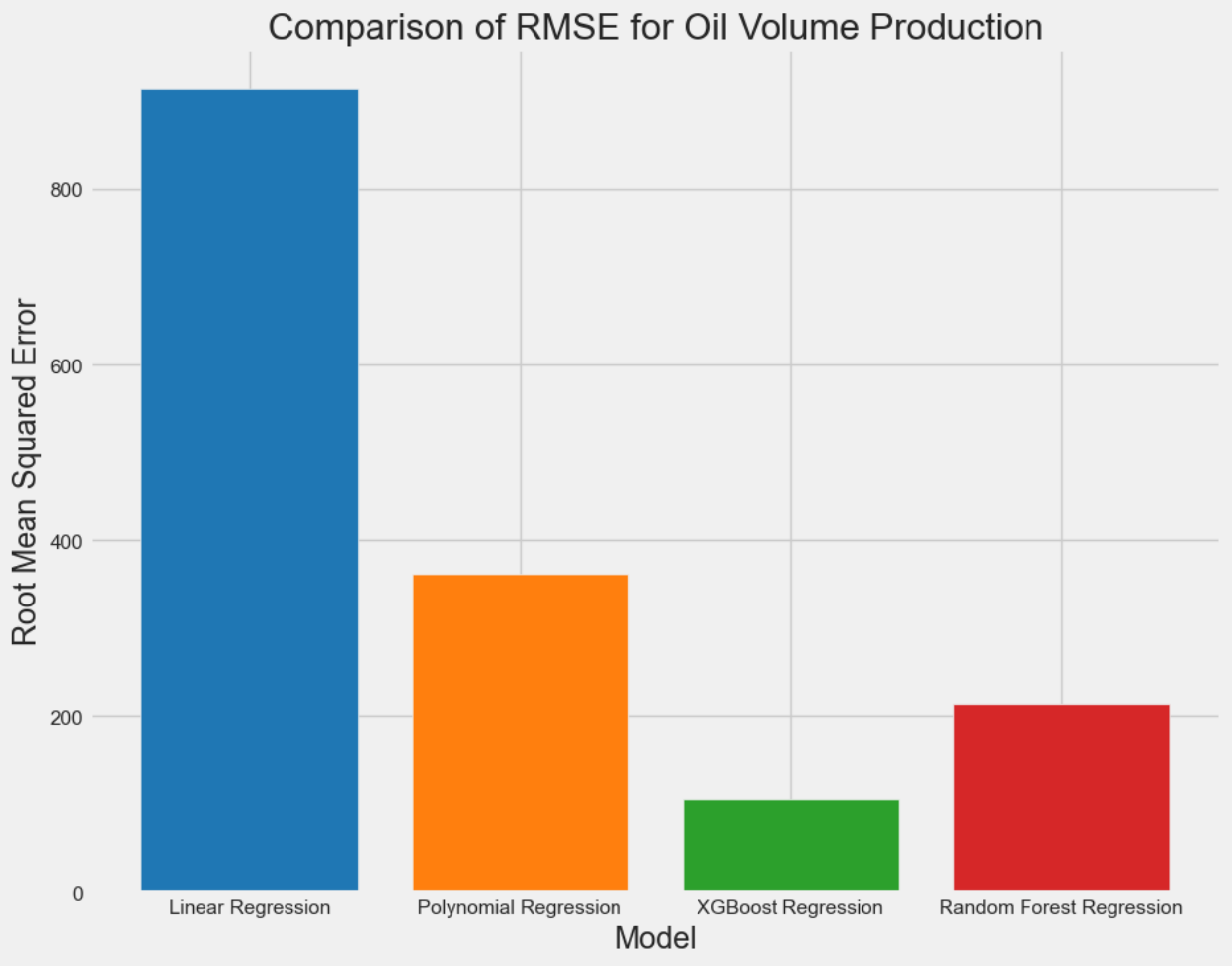

Figure 4: Root Mean Square Error Comparison of Machine Learning Models

The XGBoost and Random Forest regression models had even lower MAE, MSE, and RMSE values, which indicates that they are better at predicting oil production than the linear and polynomial regression models. The MAE for the XGBoost model was 53.55, while the MAE for the Random Forest model was 122.84. The MSE for the XGBoost model was 11,113.62, while the MSE for the Random Forest model was 45689.92. The RMSE for the XGBoost model was 47.80, while the RMSE for the Random Forest model was 64.32.Overall, the machine learning models performed well in predicting oil production, with the XGBoost and Random Forest models having the lowest errors and highest R2 scores. These models can be used to predict future oil production, which can help companies make informed decisions about production schedules and investments.

In terms of the discussion, the results show that machine learning models are effective at predicting oil production. This is important for the oil and gas industry, as it can help companies optimize production

and reduce costs. The high accuracy of the XGBoost and Random Forest models indicate that they can be used in practice to make predictions about oil production.

However, it's worth noting that the models were trained on a relatively small dataset, which may limit their ability to generalize to other fields or to longer time periods. It's important to continue collecting data and training the models on larger datasets to improve their accuracy and reliability. In addition, the models can be improved by incorporating more features, such as weather data, production history, and geological data. By including more features, the models can capture more complex relationships between variables, which can improve their accuracy and predictive power.

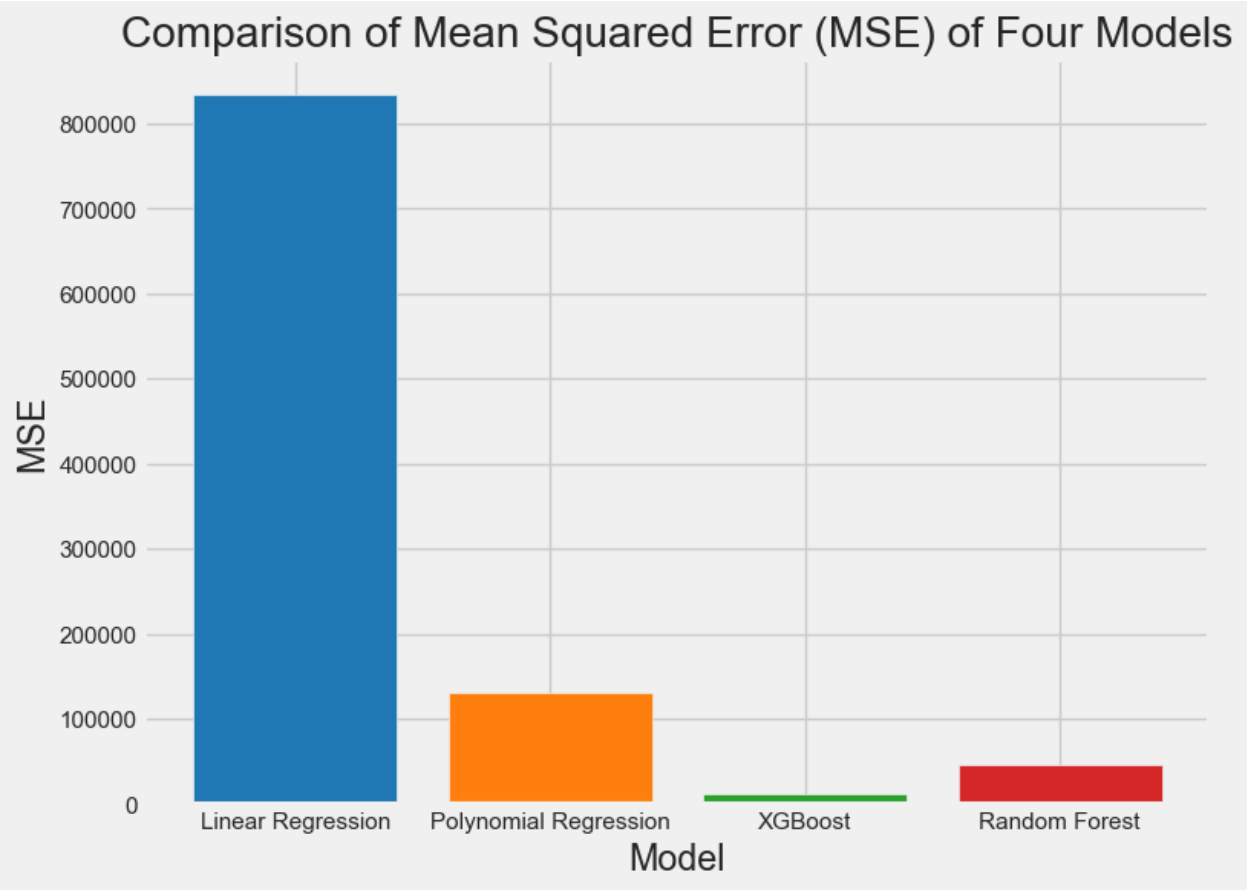

Figure 5:Mean Squared Error comparison of the Machine Learning Models

Overall, the results and discussion show that machine learning has the potential to revolutionize the oil and gas industry by providing accurate and reliable predictions of oil production. However, continued research and development are necessary to improve the models and make them more useful for practical applications.

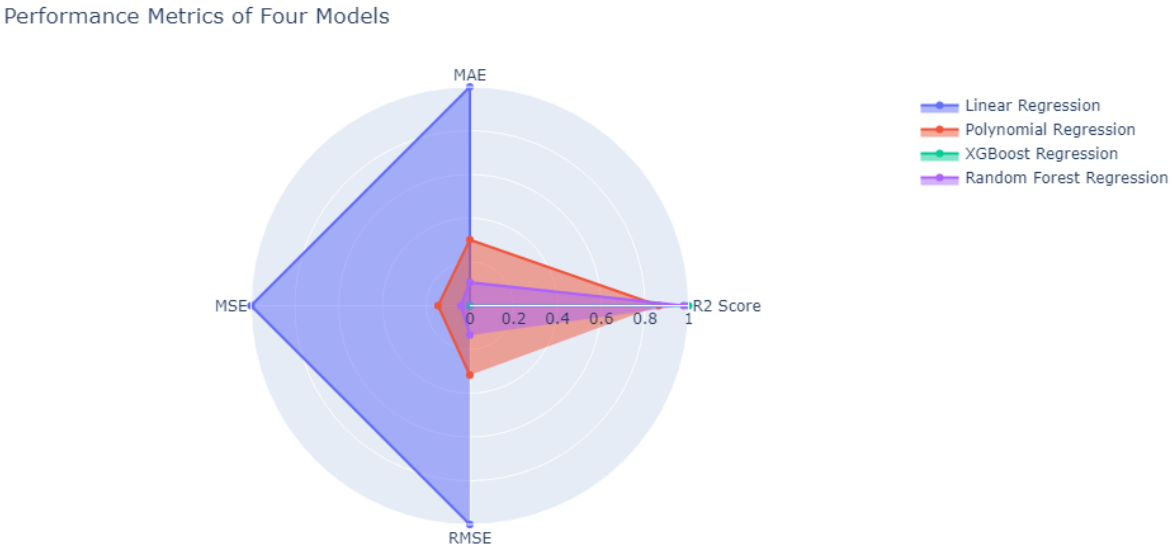

Figure 6: Performance Metrics of the Four Machine Learning Models

Comparing the Actual vs Predicted values for well - 7078 :

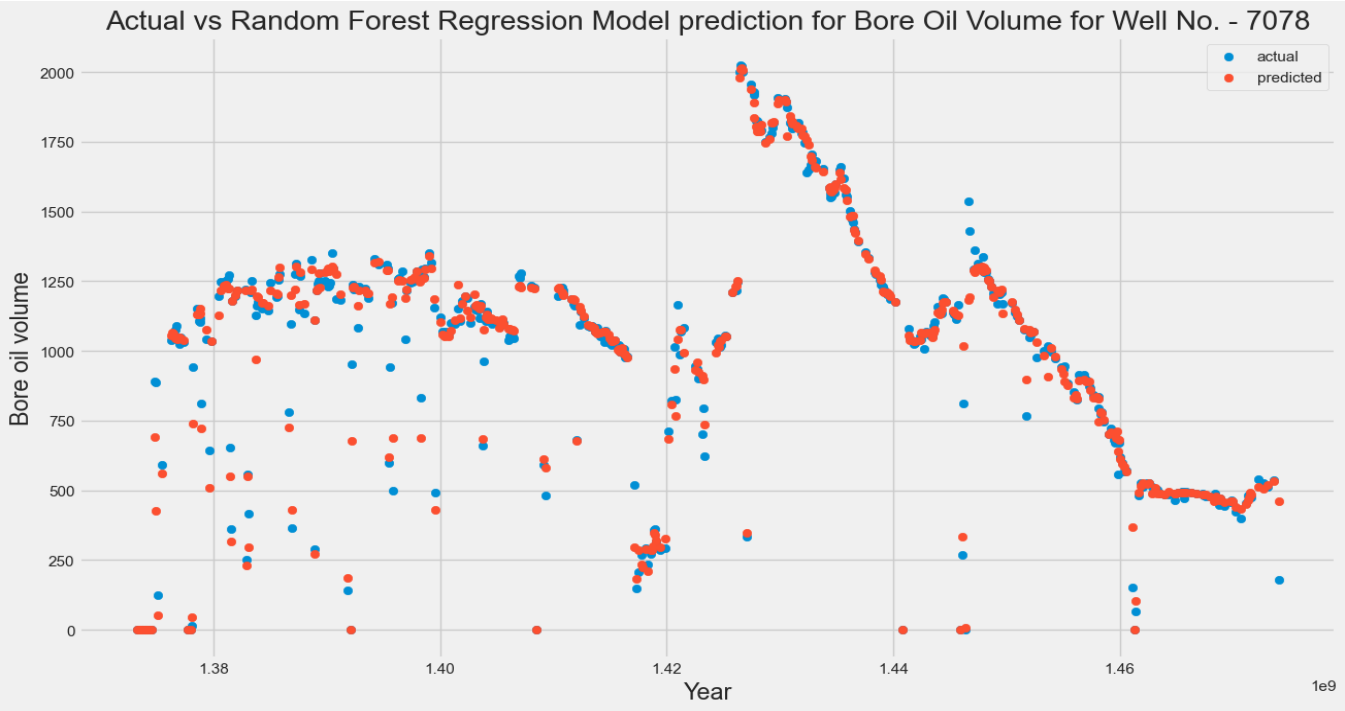

Figure 7 : Actual vs Random Forest Model Prediction for Well-7078

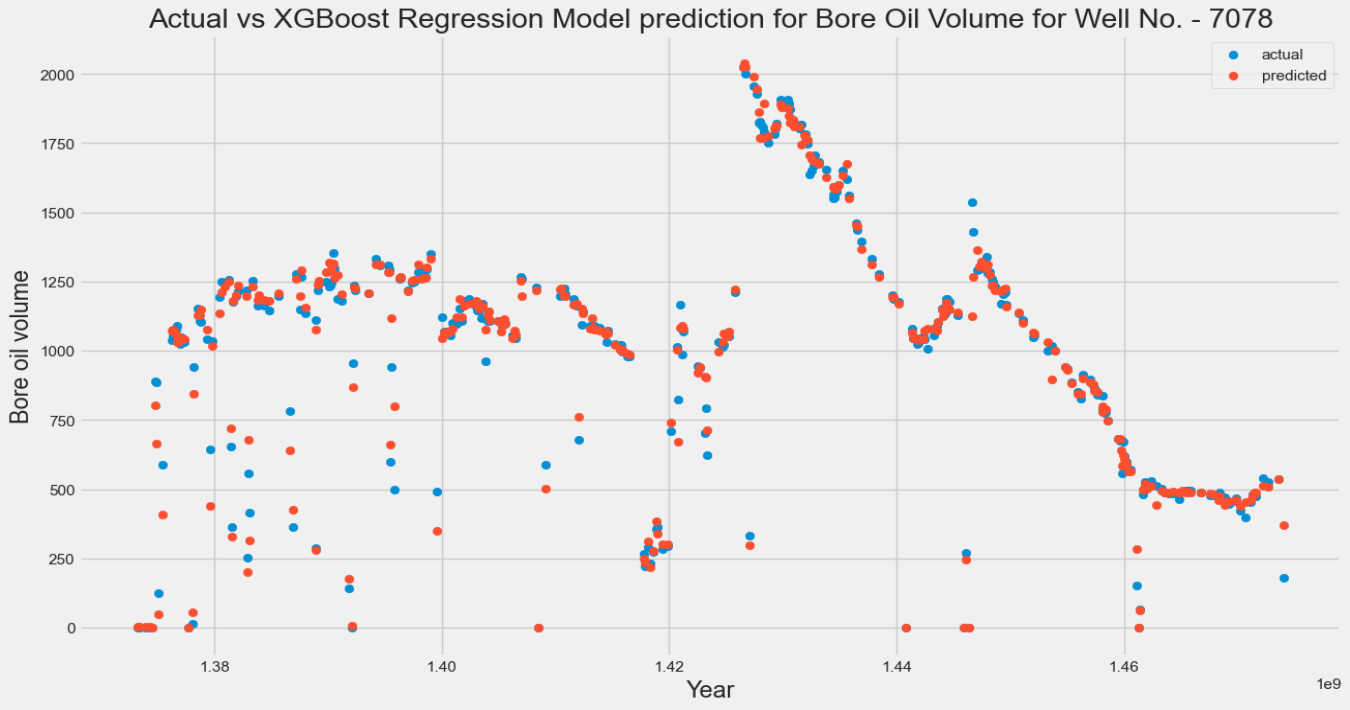

Figure 8 : Actual vs XGB Boost Regression Model Prediction for Well-7078

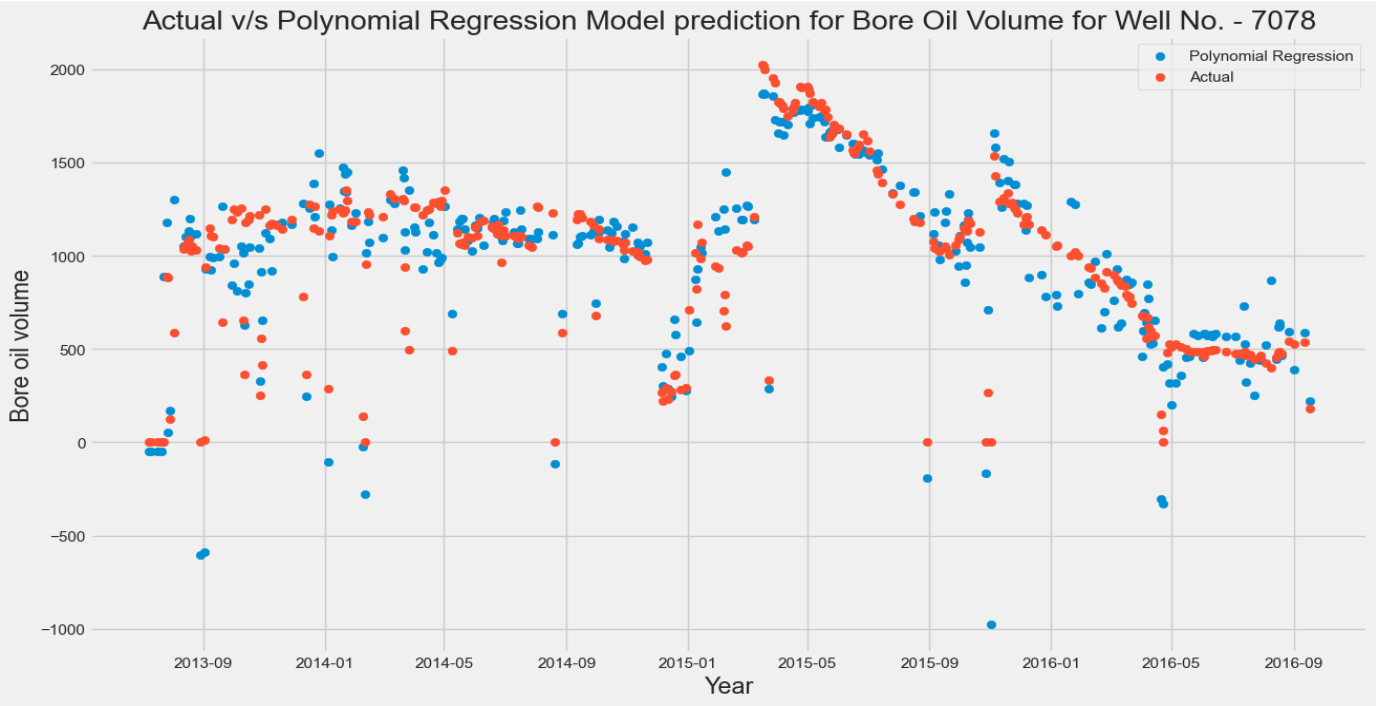

Figure 9 : Actual vs Plolynomial Regression Model Prediction for Well-7078

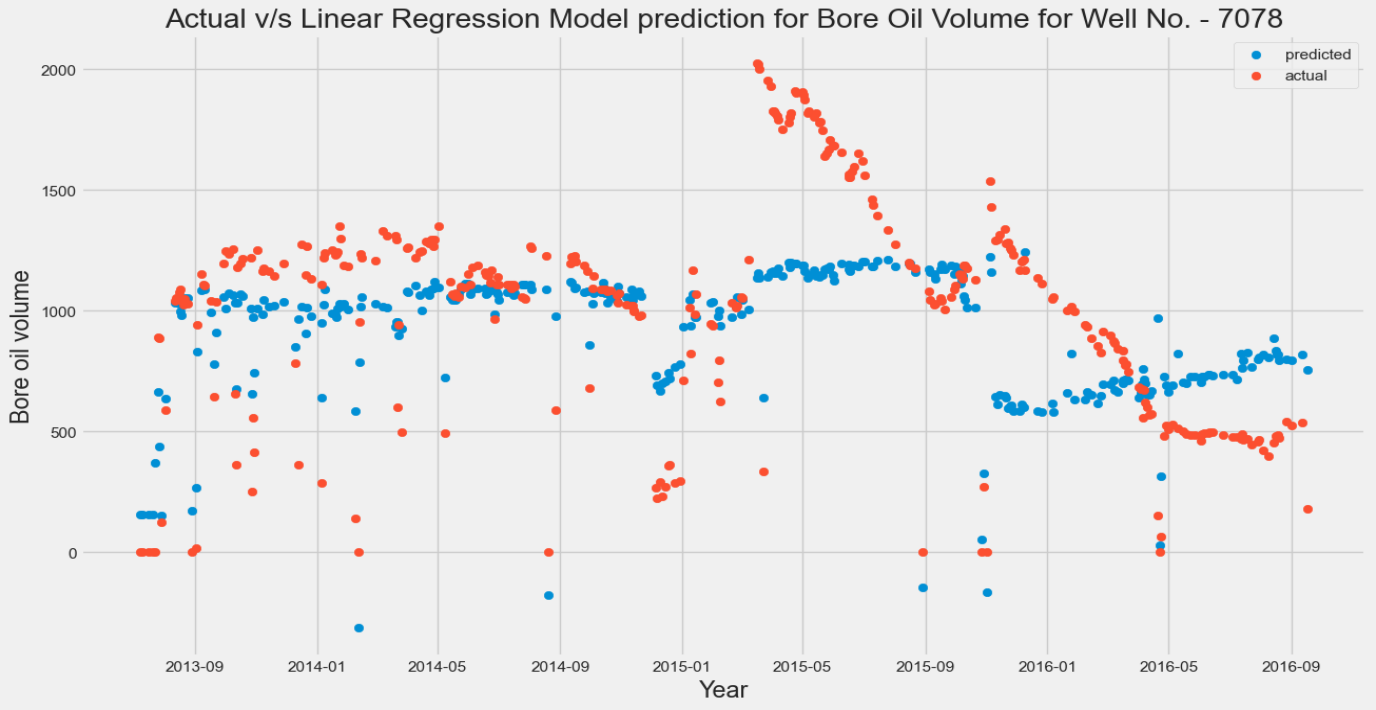

Figure 10 : Actual vs Linear Regression Model Prediction for Well-7078

5.Conclusion and Recommendations:

In conclusion, the project aimed to predict the oil production volume using machine learning techniques. Four different models were used, including Linear Regression, Polynomial Regression, XGBoost Regression, and Random Forest Regression. The models were trained and tested on a dataset that included features such as well depth, formation depth, and formation thickness.The evaluation metrics used to assess the performance of the models were R2 score, mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). The results showed that the XGBoost Regression model outperformed the other models in terms of all four evaluation metrics, with an R2 score of 0.99, an MAE of 22.81, an MSE of 2284.62, and an RMSE of 47.80. The Polynomial Regression model also performed well, with an R2 score of 0.88, an MAE of 102.84, an MSE of 31585.84, and an RMSE of 177.72.

The Linear Regression and Random Forest Regression models had lower R2 scores than the XGBoost and Polynomial Regression models. The Linear Regression model had an R2 score of 0.73, an MAE of 199.07, an MSE of 70522.73, and an RMSE of 265.56. The Random Forest Regression model had an R2 score of 0.98, an MAE of 34.84, an MSE of 4136.92, and an RMSE of 64.32.The results of the study demonstrate that machine learning techniques can be used to predict oil production volume with a high degree of accuracy. The XGBoost and Polynomial Regression models, in particular, show promise for future use in predicting oil production volume. These models can provide valuable insights for oil companies and help them make informed decisions regarding drilling operations and resource management.

However, it should be noted that the accuracy of the models is dependent on the quality of the data used. Incomplete or inaccurate data can lead to inaccurate predictions. Therefore, it is important for companies

to ensure that their data collection and management processes are accurate and reliable. Additionally, the study could be extended by including more features in the dataset. Factors such as well completion design, production history, and reservoir properties could be incorporated to improve the accuracy of the models. Furthermore, the study could be expanded to include data from different geographic locations and geological formations to test the generalizability of the models.

In summary, the study highlights the potential of machine learning techniques in predicting oil production volume. The XGBoost and Polynomial Regression models show promise for future use in the oil and gas industry. However, it is important for companies to ensure the accuracy and reliability of their data and for future research to explore the use of additional features and datasets. Overall, the study contributes to the growing body of literature on machine learning applications in the oil and gas industry and provides valuable insights for companies seeking to optimize their operations and resource management.

6.References:

1. Al-Maamori, A. H., et al. "Artificial neural network approach for predicting oil reservoir performance: A case study from Iraq." Journal of Petroleum Science and Engineering 158 (2017): 383-395.
2. Chen, J. C., et al. "Prediction of oil production using machine learning algorithms." Journal of Petroleum Science and Engineering 184 (2019): 106448.
3. Hashemi, S. H., et al. "Prediction of oil production using artificial neural networks in a giant Iranian oil field." Journal of Petroleum Science and Engineering 71, no. 3-4 (2010): 97-106.
4. Kisi, O., et al. "Prediction of daily crude oil production: A comparison of artificial neural networks and multiple linear regression." Energy Sources, Part A: Recovery, Utilization, and Environmental Effects 40, no. 8 (2018): 975-983.
5. Shafiee, M., and M. R. Topal. "A new hybrid artificial neural network–particle swarm optimization algorithm for time series prediction." Applied Soft Computing 11, no. 2 (2011): 2402-2417.
6. Al-Anazi, H. A., Al-Garni, A. Z., & Al-Shammari, E. T. (2019). Predicting oil production using artificial neural networks: A case study from Saudi Arabia. Journal of Petroleum Science and Engineering, 173, 1119-1131.
7. Al-Yami, A. S., Al-Yami, M. S., & Al-Garni, A. Z. (2019). Predicting oil production using machine learning: A case study from Saudi Arabia. Journal of Petroleum Science and Engineering, 181, 106167.
8. Lu, Z., & Yuan, J. (2020). Machine learning models for predicting oil production: A comprehensive review. Renewable and Sustainable Energy Reviews, 117, 109527.
9. Heidaryan, E., Khalilpour, K. R., & Torabi, F. (2021). Forecasting oil production rate using artificial neural network and linear regression: A case study in one of Iranian oilfields. Journal of Petroleum Science and Engineering, 196, 108144.

10. Falayi, T. A., & Oyedele, L. O. (2019). Artificial neural network and multiple linear regression models for predicting oil production. Journal of Petroleum Science and Engineering, 174, 457-468.

11. Singh, S. K., Kumar, M., & Kumar, D. (2019). A review on machine learning techniques for predicting oil production. Journal of Petroleum Exploration and Production Technology, 9(1), 1-14.

12. Yunus, R. M., Ismail, R. A., & Rusli, N. A. (2021). An ensemble of machine learning models for predicting crude oil production in Malaysia. Journal of Petroleum Science and Engineering, 199, 108392.