Predicting Oil Recovery Factor in The Gulf of Mexico using Public Records from Bureau of Safety and Enivronmental Enforcement(BSEE)

Carlos Avila-Salazar

2025-04-28

1 Summary

This report presents a data-driven approach to predicting the oil recovery factor (ORF) using various machine learning models. ORF is a key metric in petroleum engineering that represents the proportion of recoverable oil from a reservoir relative to its total original oil in place (OOIP). It is a crucial factor in determining the efficiency of extraction methods and the overall profitability of an oil field. Higher ORF values indicate that a greater percentage of the oil in the reservoir can be extracted, which is influenced by geological characteristics, fluid properties, and recovery techniques. ORF can be estimated by analyitcal methods, e.g. material balances or numerical techniques such as reservoir simulation. This in particular can be very expensive and time consuming, requiring input from multiple disciplines, e.g. geo-physicists, geologists, petro-physicists, engineers, etc. These techniques need accurate data that is not usually available in early stages of the field development. At this stage, data based methods can be used to get ORF estimates that are representative of the population being evaluated and inform business decisions.

The dataset used for this study, referred to as Sands Atlas 2020, was sourced from the Bureau of Safety and Environmental Enforcement (BSEE) available at the BSEE website. It contains geological and reservoir engineering parameters collected from various offshore oil fields. The

dataset includes key features such as:

Total Net Thickness (THK): Measures the total thickness of reservoir rock that contributes to oil production.

Porosity: Represents the proportion of void space in the rock that can store hydrocarbons.

Water Saturation (SW): Indicates the fraction of pore space occupied by water rather than hydrocarbons.

Permeability: Measures how easily fluids can flow through the reservoir rock, expressed in millidarcies (mD).

Weighted Average Initial Pressure (PI): Represents the reservoir pressure before production starts, affecting oil flow in pounds per square inch (psi).

Oil API Gravity (API): A measure of oil density; higher values indicate lighter oil that flows more easily.

Gas-Oil Ratio (GOR): The volume of gas produced per barrel of oil, affecting reservoir pressure and recovery efficiency in hydrocarbon reservoirs, with values expressed in thousand cubic feet per barrel (mcf/bbl) to indicate the gas content relative to oil production.

The dataset was preprocessed to remove missing values and retain only numerical features for statistical analysis and machine learning modeling. To improve predictive accuracy, data entries where ORF values were zero were removed, as they could distort the model’s ability to learn meaningful patterns. Several models, including Random Forest, Linear Regression, Decision Tree, and LOESS, were trained and evaluated to determine the most effective predictive method.

Model performance was assessed using Root Mean Squared Error (RMSE), a common metric for measuring prediction accuracy in regression problems. Among the models evaluated, Random Forest achieved the lowest test RMSE (0.1049734), indicating the best generalization performance. LOESS (0.1060428) and Linear Regression (0.1068171) followed closely, while the Decision Tree model performed the worst, with a test RMSE of 0.1116143, highlighting its tendency to overfit due to high variance. Notably, Random Forest also had the lowest training RMSE (0.0894715), suggesting a strong fit to the training data.

2 Data Wrangling

The dataset used in this analysis was obtained from the Bureau of Safety and Environmental Enforcement (BSEE) and was downloaded as a ZIP file containing multiple CSV and Excel files. The first step involved extracting the files and identifying the primary dataset for analysis. Once the data was loaded, initial cleaning steps were performed to remove missing values. There is a variety of variables, numerical and categorical variables, in consultation with a subject-matter expert, a subset of variables was selected to train the models.

3 Exploratory Data Analysis

To better understand the dataset and identify key patterns, exploratory data analysis was conducted using scatter plots and histograms. The scatter plots examine the relationships between oil recovery factor (ORF) and various reservoir parameters, highlighting potential correlations. Meanwhile, the histograms provide insights into the distribution of each feature, allowing us to assess their variability and suitability for modeling.

The relationship between total net thickness (THK) and ORF shows a slight positive trend, suggesting that thicker reservoirs may contribute to higher recovery. In contrast, porosity and ORF exhibit a weak correlation, indicating that porosity alone may not be a strong predictor of oil recovery. Similarly, water saturation (SW) and ORF display little to no correlation, suggesting that water saturation does not significantly influence ORF in a linear manner. Permeability and ORF do not present a clear trend, aligning with the highly variable distribution observed in the permeability data.

The scatter plot of initial pressure (PI) versus ORF highlights a high variance in initial pressure and its effect on recovery, making it difficult to establish a strong linear relationship. On the other hand, API gravity and ORF demonstrate a moderate positive trend, suggesting that higher API gravity values may contribute to better oil recovery. Lastly, gas-oil ratio (GOR) and ORF reveal a weak negative trend, indicating that higher gas content might slightly reduce oil recovery.

Overall, the scatter plots reinforce the need for non-linear modeling techniques, as simple linear relationships are not evident in most of the variables. Given these insights, Random Forest or LOESS regression may be more effective than traditional linear models in capturing the complex interactions influencing oil recovery.

The histograms generated for each reservoir parameter provide valuable insights into their distributions and potential impacts on modeling. By visualizing these distributions after filtering out zero values, we can better understand the behavior of each feature and make informed decisions about preprocessing and model selection.

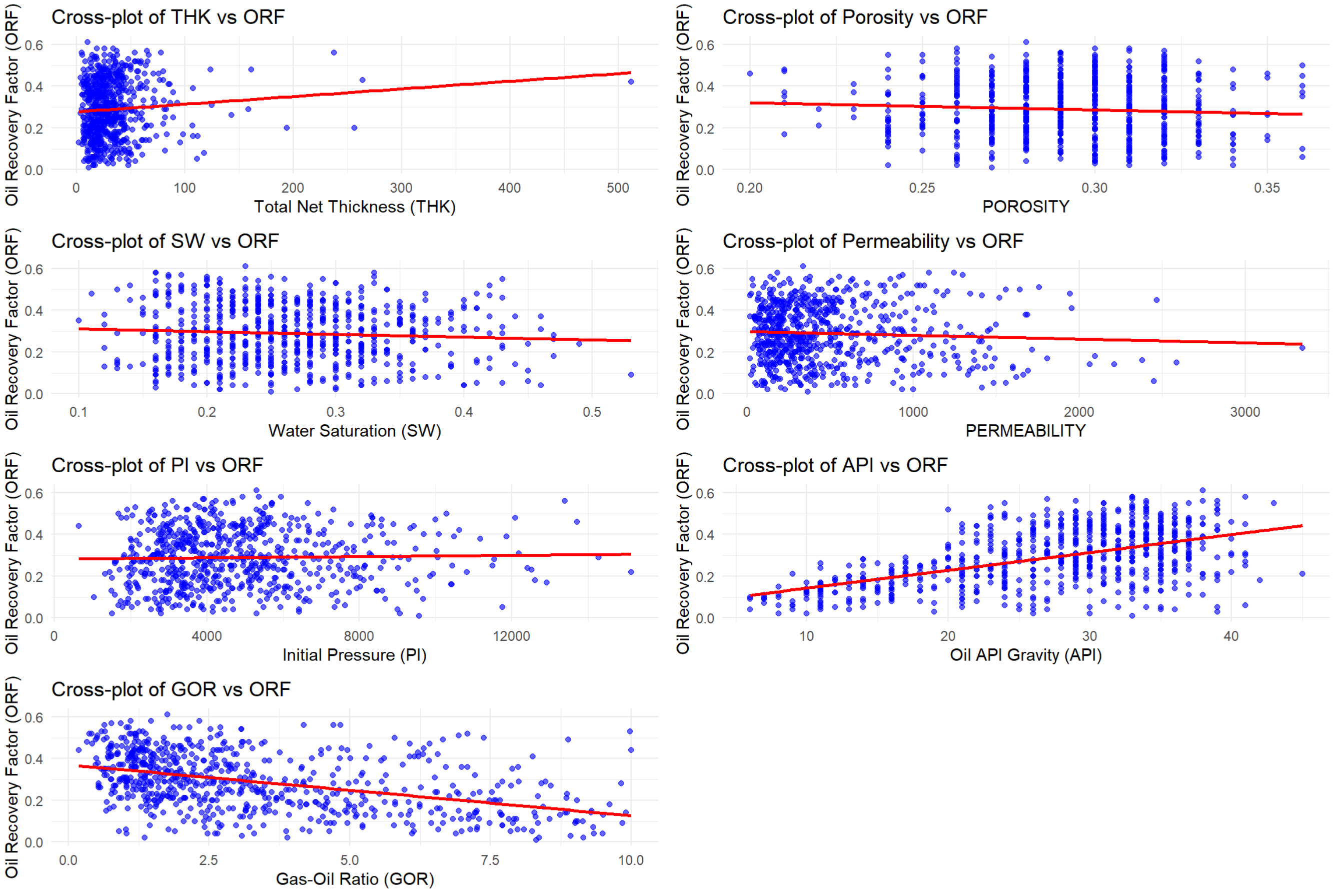
The total net thickness (THK) histogram shows a right-skewed distribution, with most values concentrated between 5 and 50 feet. The P10, P50, and P90 percentiles (11.50, 26.26, and 56.92, respectively) indicate that a majority of the reservoirs have relatively thin pay zones, while a small fraction exhibit significantly greater thickness. Similarly, permeability (mD) exhibits a strong right skew, with values ranging from 108.77 at P10 to 1103.55 at P90. This suggests that while some reservoirs exhibit low permeability, a minority have extremely high permeability, which could significantly influence oil production rates. Gas-oil ratio (GOR) follows a similar trend, with values spanning from 0.997 to 7.23 Mcf/bbl, indicating significant variability in gas content. In this case, GOR was not used for modeling.

Other parameters, such as porosity and water saturation (SW), show relatively balanced distributions. The porosity histogram presents a near-normal distribution, with values ranging between 0.26 and 0.32 at the P10 and P90 levels. With a median (P50) of 0.29, most reservoir samples exhibit moderate porosity, suggesting consistent rock quality across different reservoirs. Similarly, water saturation values, which range from 0.17 to 0.36, indicate variability in fluid saturation across different wells, but its overall distribution suggests it can be used in its raw form. Weighted average initial pressure (PI) displays a wide range, from around 2,478.63 psi at P10 to 8,022.62 psi at P90. This high variance suggests that pressure is highly reservoir-dependent and may influence oil mobility.

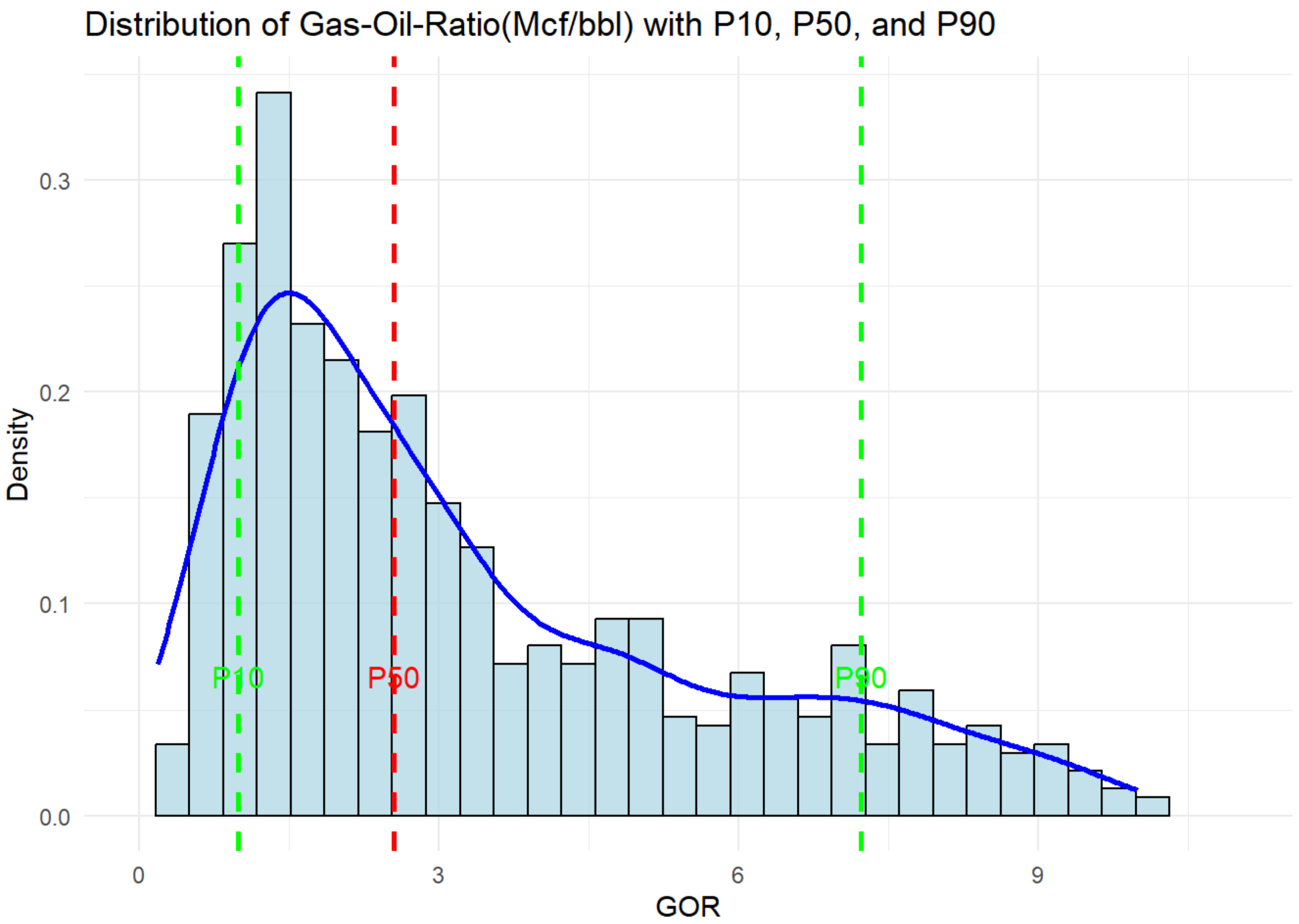
Two parameters, weighted average oil API gravity and oil recovery factor (ORF), had significant reductions in data after filtering out zeros. API gravity was reduced by 72.64%, leaving data that primarily falls between 14.00 and 36.00, with a median of 29. API gravity is crucial in determining oil quality, and its heavily reduced dataset suggests that missing data imputation or handling should be considered before including it in the model. Similarly, ORF, which serves as the target variable, experienced a 73.64% reduction, resulting in values ranging from 0.11 to 0.47. The right skew of ORF indicates that while some reservoirs have high recovery efficiency, most exhibit moderate to low recovery factors. This distribution suggests that advanced modeling techniques, such as ensemble methods, may be required to better capture the complex relationships affecting oil recovery. Random forest was used in modeling for this reason.

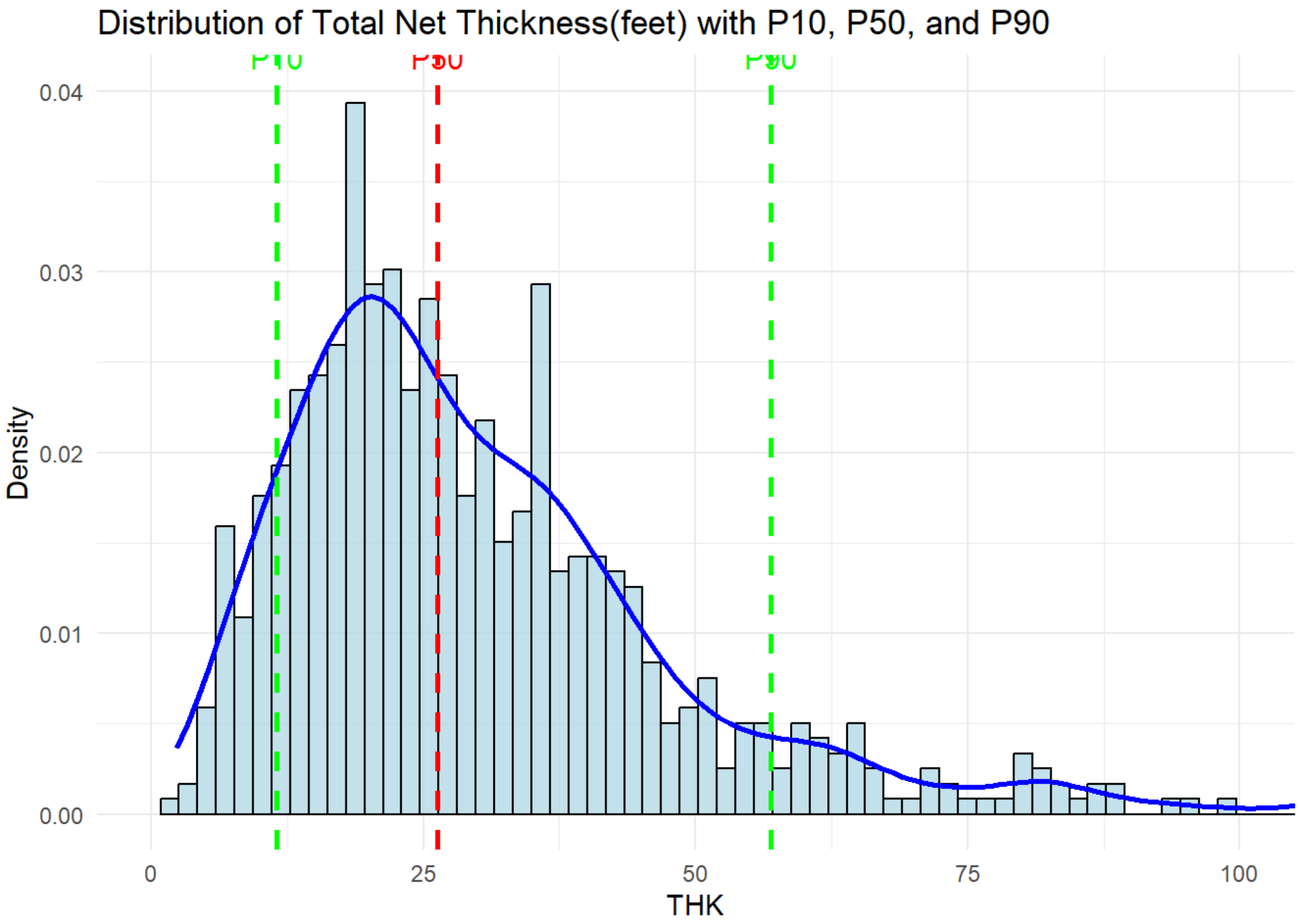
Overall, the histograms reveal that several parameters, including permeability, GOR, and initial pressure, exhibit high variance and skewness. The large data reductions for API gravity and ORF indicate missing data challenges that need to be addressed. Given these insights, non-linear models such as Random Forest or LOESS are likely to perform better than linear models, as they can capture the complex interactions observed in the data.

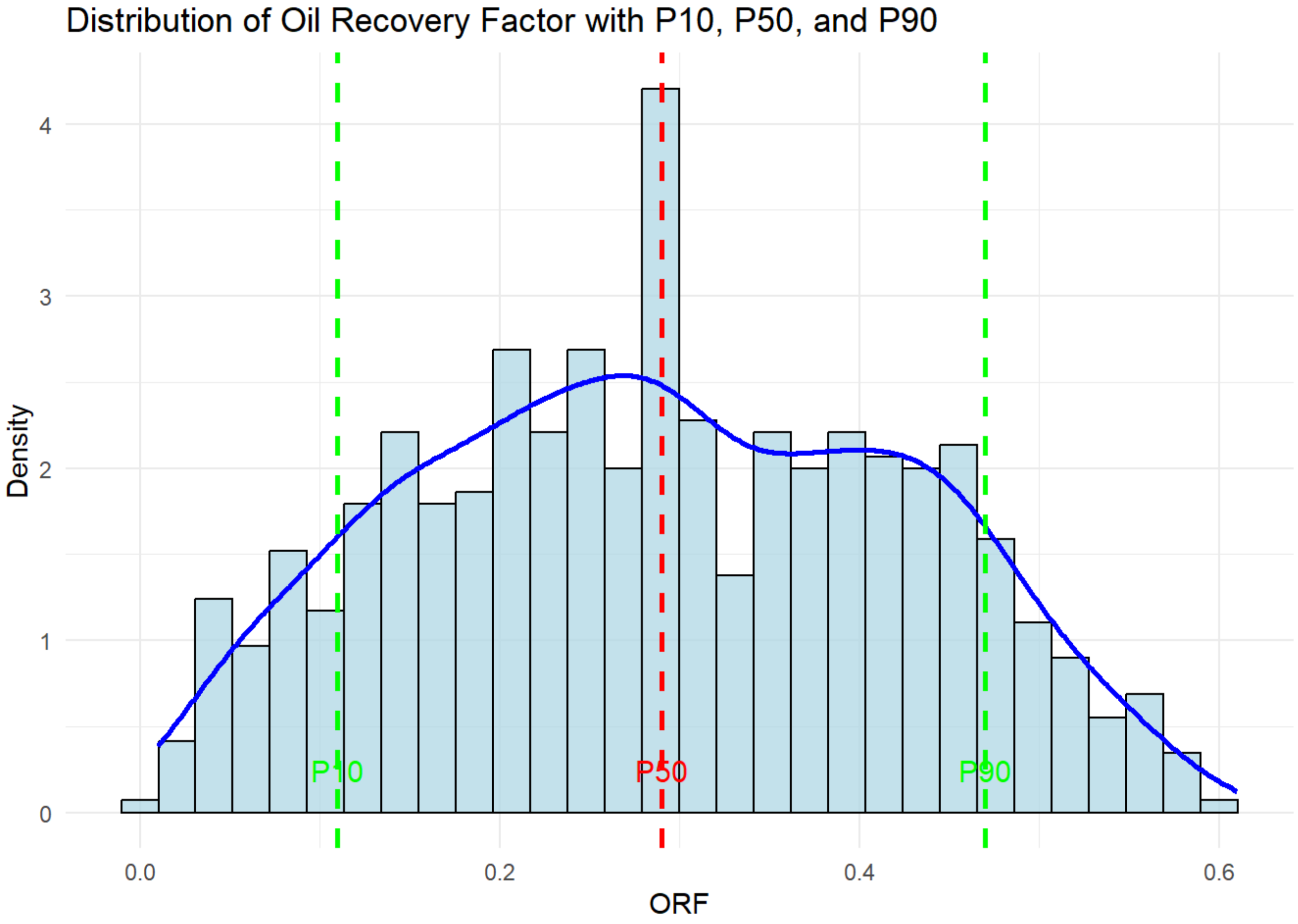
3.1 Scatter Plots of Parameters vs. ORF

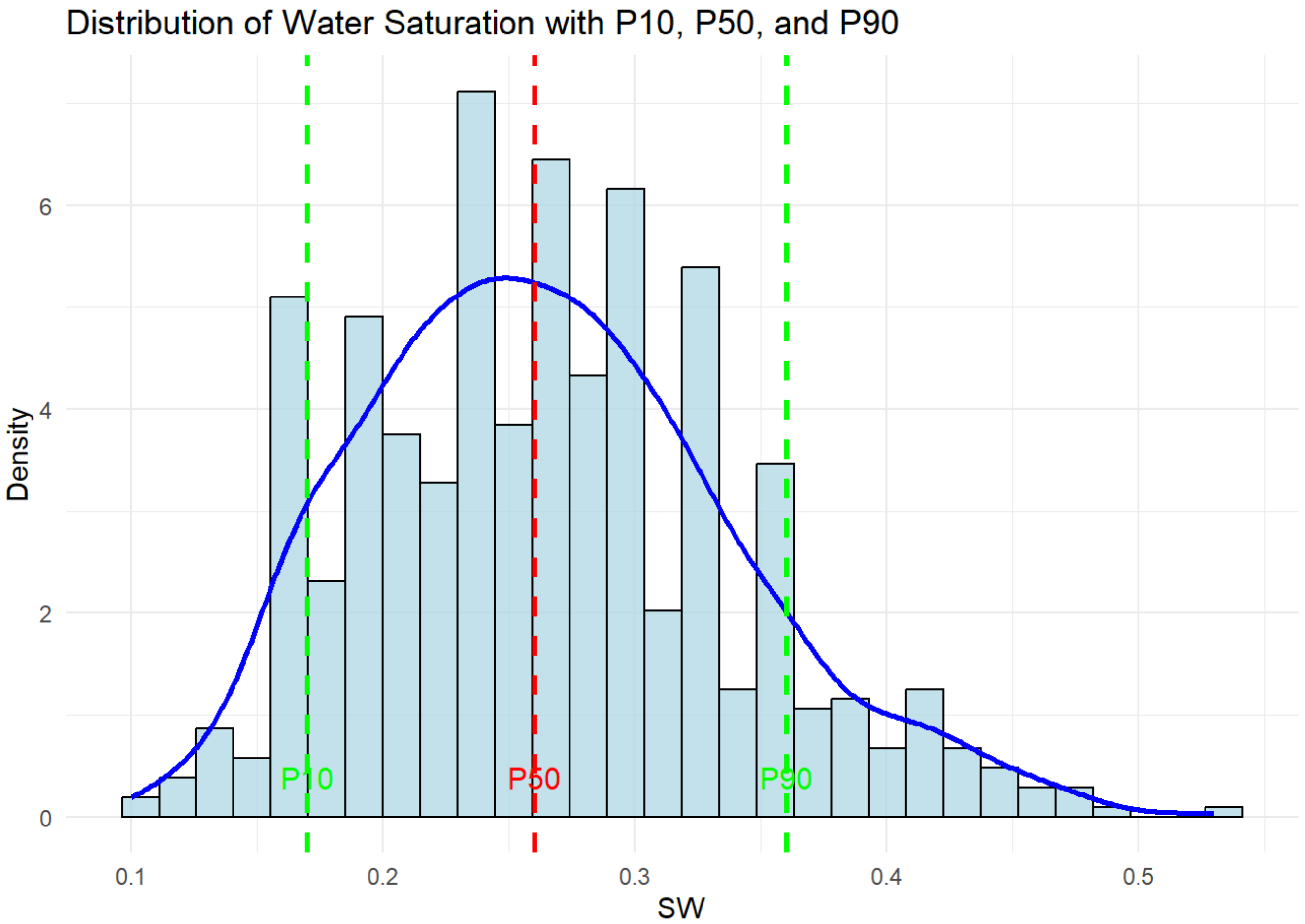


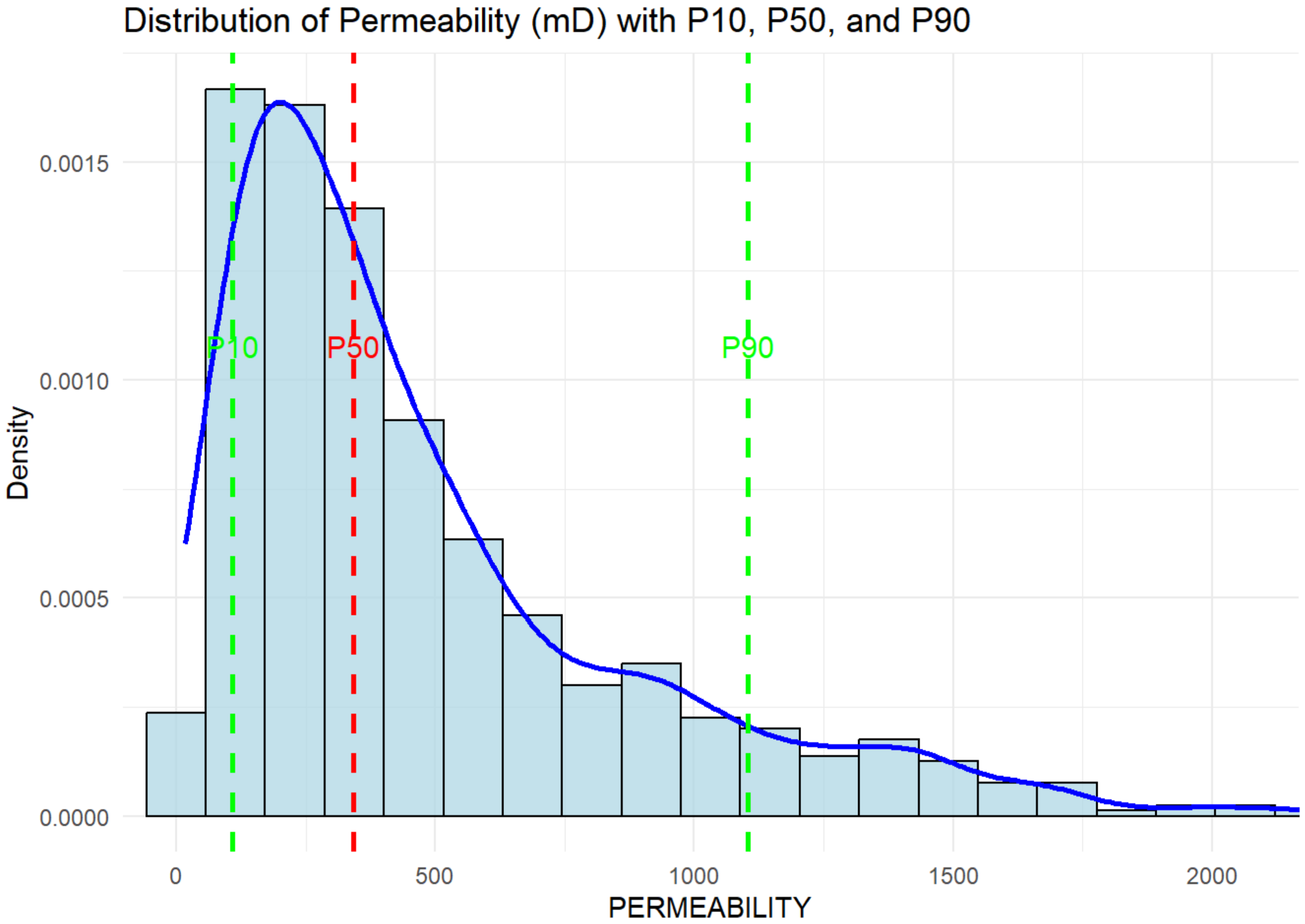
3.2 Histograms of Distributions

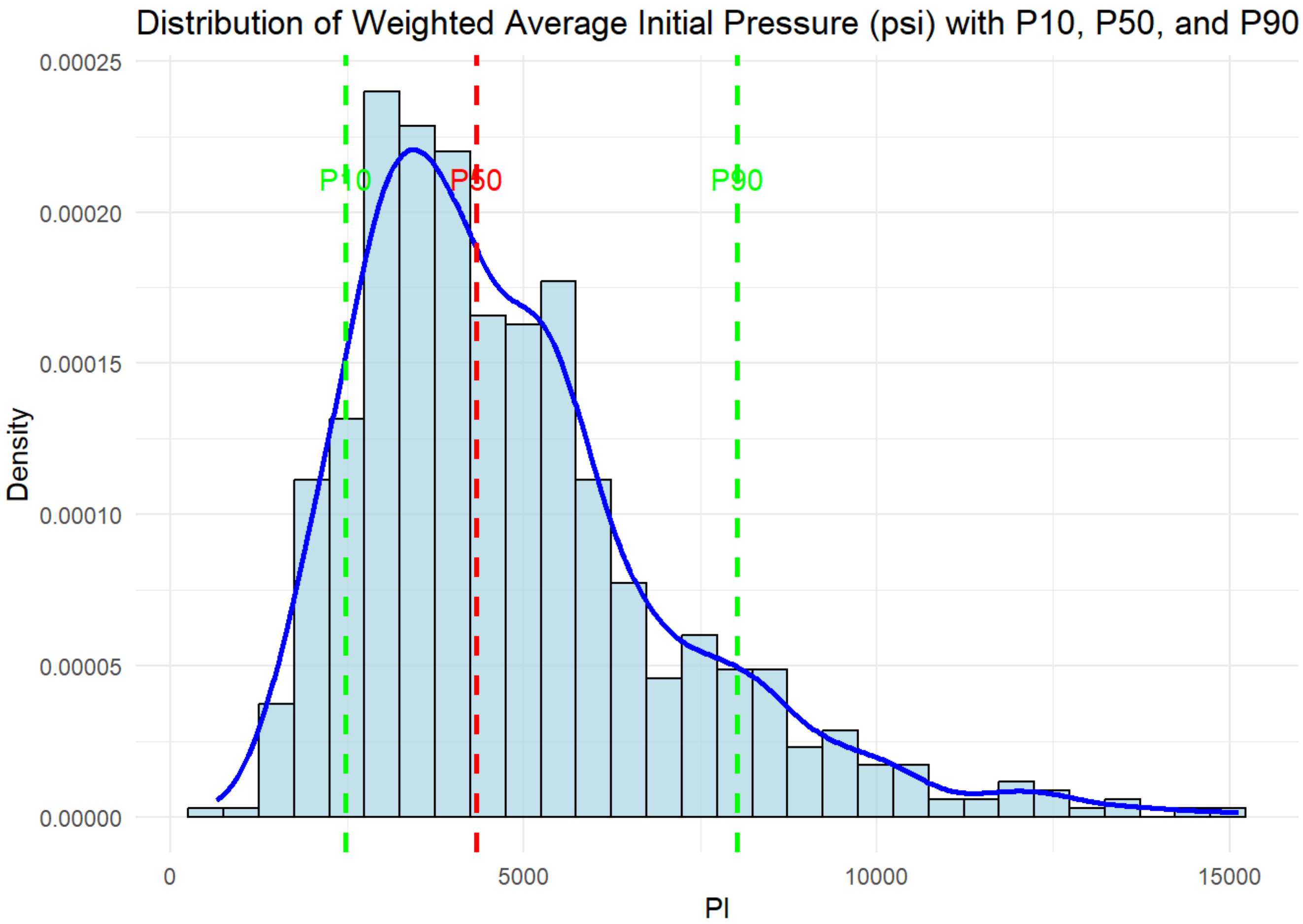


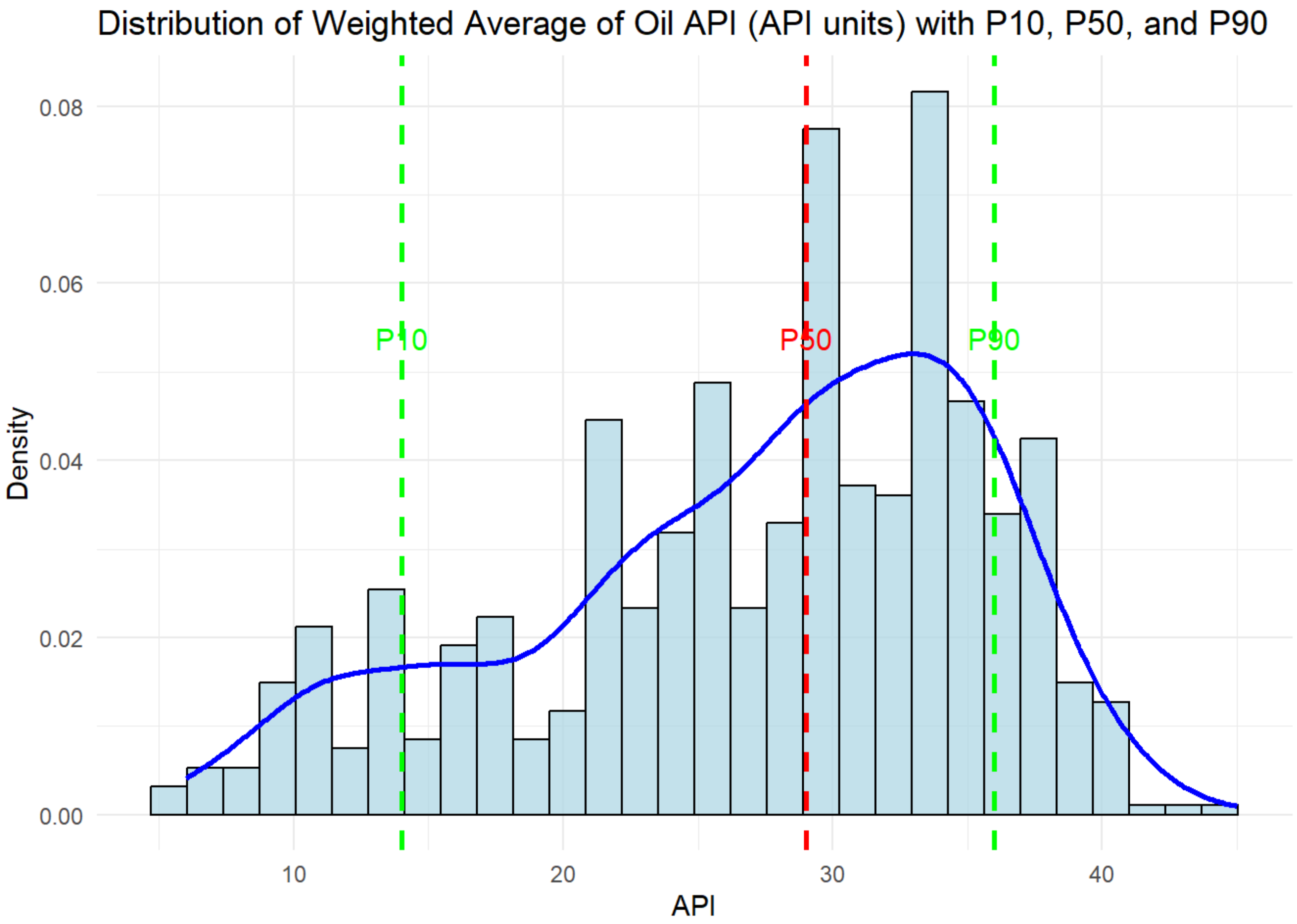

Distribution of Total Net Thickness(feet) with P10, P50, and P90
P10
P50
P90
Density
0.04
0.03
0.02
0.01
0.00
0
25
50
75
100
THK

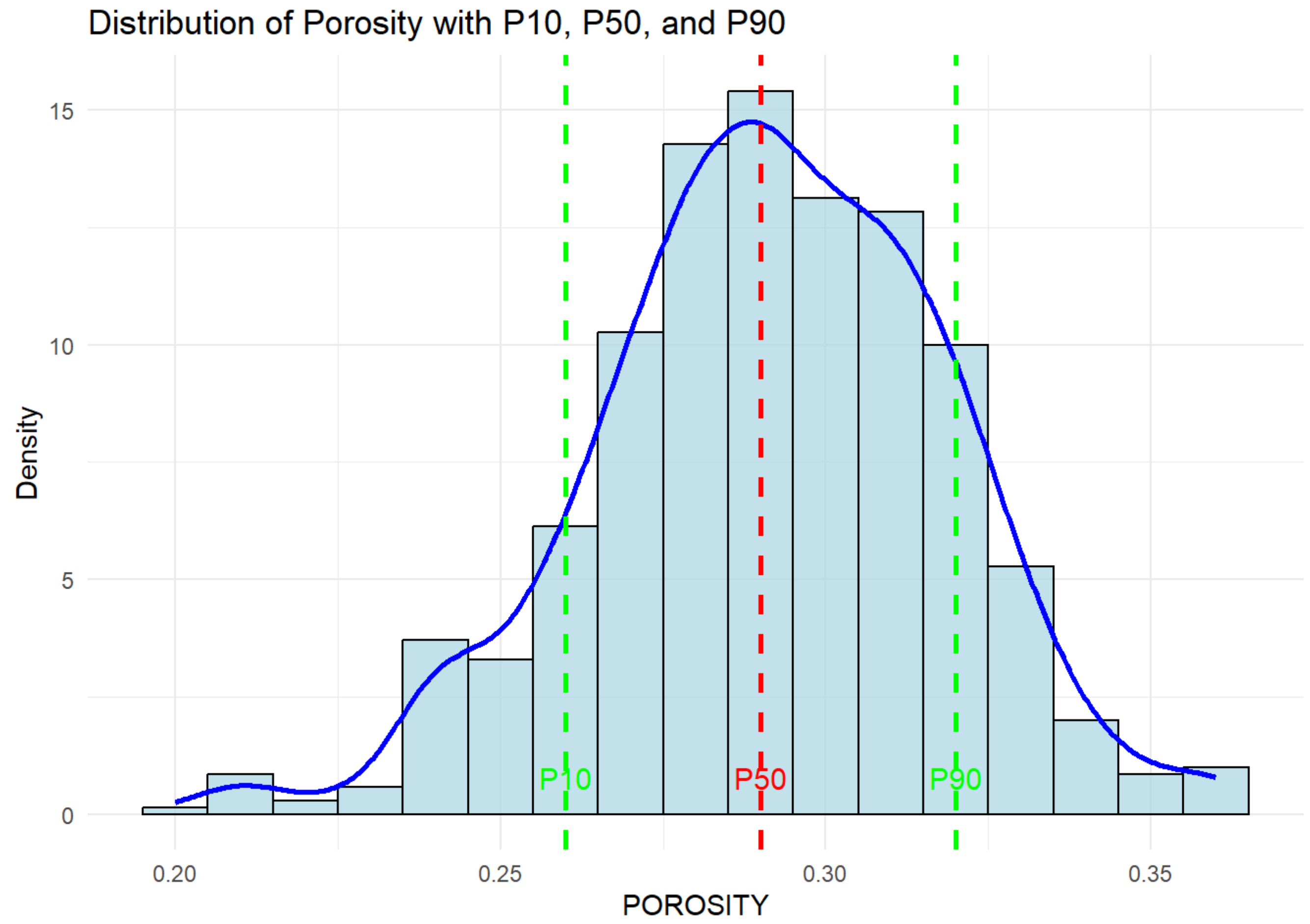

Distribution of Oil Recovery Factor with P10, P50, and P90
Density
4
3
2
1
0
P10
P50
P90
0.0
0.2
0.4
0.6
ORF


Distribution of Water Saturation with P10, P50, and P90
Density
0
2
4
6
P10
P50
P90
0.1
0.2
0.3
0.4
0.5
SW


Distribution of Permeability (mD) with P10, P50, and P90
Density
0.0000
0.0005
0.0010
0.0015
P10
P50
P90
0
500
1000
1500
2000
PERMEABILITY


Distribution of Weighted Average Initial Pressure (psi) with P10, P50, and P90
Density
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
P10
P50
P90
0
5000
10000
15000
PI


Distribution of Weighted Average of Oil API (API units) with P10, P50, and P90
Density
0.08
0.06
0.04
0.02
0.00
P10
P50
P90
10
20
30
40
API



Percentile Splits for Each Parameter

Parameter	P10	P50	P90
GOR	0.9966044	2.551781	7.226816
THK	11.5000000	26.260000	56.920000
ORF	0.1100000	0.290000	0.470000
SW	0.1700000	0.260000	0.360000
PERMEABILITY	108.7700000	342.000000	1103.550000
PI	2478.6300000	4322.630000	8022.620000
API	14.0000000	29.000000	36.000000
POROSITY	0.2600000	0.290000	0.320000

4 Modeling

To predict the Oil Recovery Factor (ORF), multiple machine learning models were implemented, including Random Forest, Linear Regression, Decision Tree, and LOESS regression. These models were selected based on their ability to capture complex relationships between geological and reservoir parameters and ORF.

The dataset used for modeling was preprocessed to ensure the quality and reliability of the input features. This involved removing missing values, retaining only numerical features, filtering out entries where ORF was zero, and selecting relevant predictor variables such as Total Net Thickness (THK), Porosity, Water Saturation (SW), Permeability, Weighted Average Initial Pressure (PI), and Oil API Gravity (API). A train-test split of 80%-20% was applied to the cleaned dataset to evaluate model performance effectively.

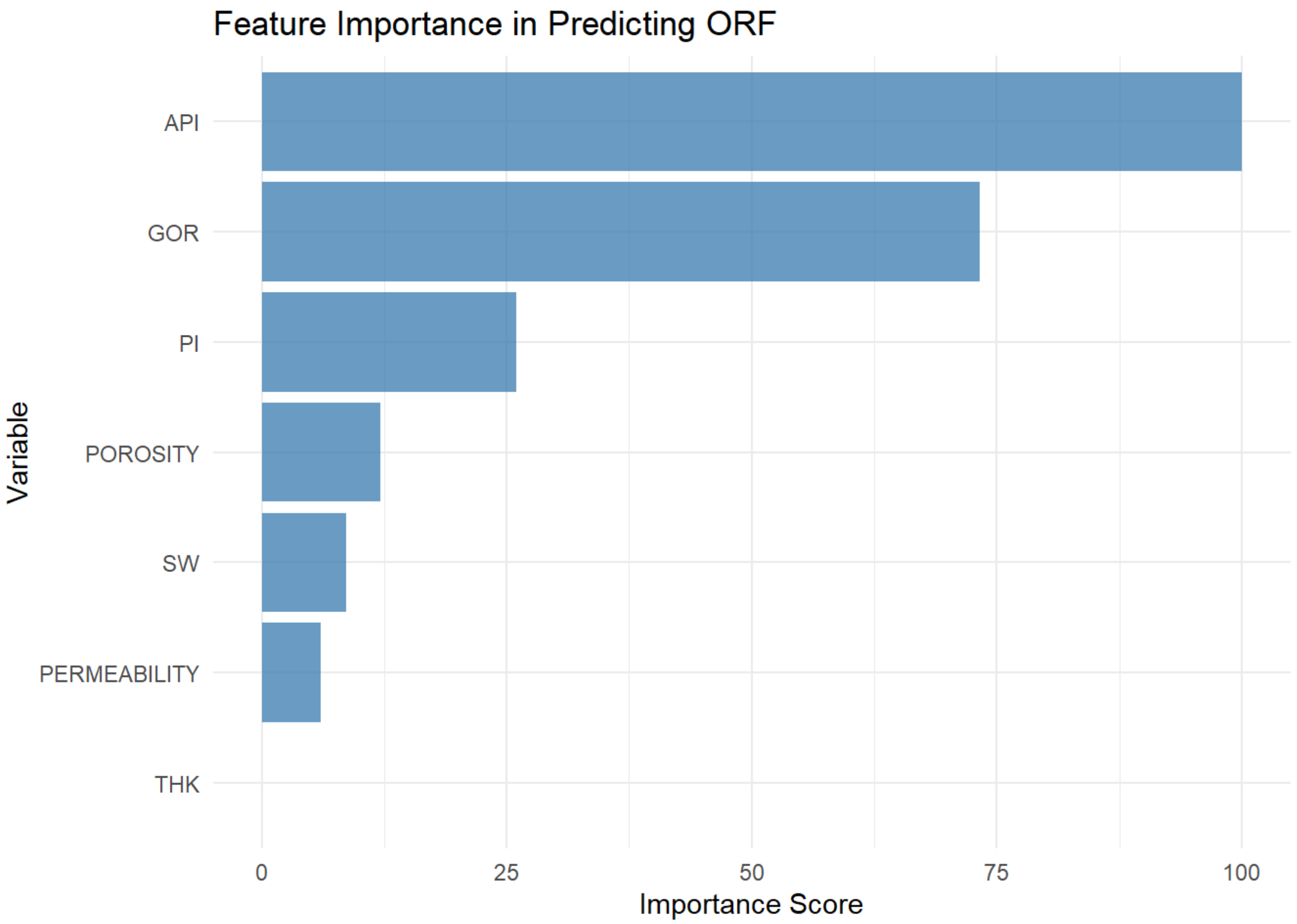
For model training, a Random Forest model was trained with 5-fold and 10-fold cross-validation, using a range of tuning parameters to optimize performance. Feature importance analysis was conducted to identify the most influential variables in predicting ORF. A multiple linear regression model was also trained to assess the linear relationships between predictor variables and ORF, with cross-validation used to mitigate overfitting. A decision tree model was built to capture nonlinear relationships in the data, where the complexity parameter (cp) was tuned to balance model simplicity and predictive accuracy. Additionally, LOESS regression was applied using the best predictor variable identified through correlation analysis to capture localized trends in the data.

The performance of each model was assessed using Root Mean Squared Error (RMSE) on both training and test sets. The Random Forest model achieved the lowest RMSE, indicating strong predictive performance and robustness to nonlinearities. Linear Regression showed moderate accuracy but struggled with capturing complex interactions, while the Decision Tree model provided interpretable decision rules but was prone to overfitting. LOESS regression was effective for localized trend detection but less generalizable. The results suggest that ensemble methods like Random Forest outperform simpler models in predicting ORF. Further hyperparameter tuning and feature engineering could improve model accuracy and interpretability.

Random Forest 10-Fold Cross-Validation Results

Metric	Value
RMSE	0.1068780
R-squared	0.3995738
MAE	0.0859198

5 Random Forest Importance Variable Plot



6 Linear Regression Coefficients

The table below presents the estimated coefficients, standard errors, t-values, and p-values for the regression model used to predict oil recovery factor (ORF):

Feature	Estimate	Std. Error	T-value	P-Value
(Intercept)	0.3841	0.0739	5.1954	0.0000
THK	0.0002	0.0002	0.8867	0.3756
POROSITY	-0.2813	0.2186	-1.2868	0.1987
SW	-0.4791	0.0957	-5.0052	0.0000
PERMEABILITY	-0.0001	0.0000	-3.7274	0.0002
PI	0.0000	0.0000	2.4126	0.0162
API	0.0067	0.0006	11.5259	0.0000
GOR	-0.0195	0.0021	-9.3039	0.0000

7 Results

To evaluate model performance in predicting oil recovery factor (ORF), several regression models were trained and tested. The table below presents the Root Mean Squared Error (RMSE) for each model on both the training and test datasets:

RMSE Comparison of Different Models

Model	Train_RMSE	Test_RMSE
Random Forest	0.0894715	0.1049734
Linear Regression	0.1051881	0.1068171
Decision Tree	0.1017232	0.1116143
LOESS	0.1149228	0.1060428

Among the tested models, Random Forest achieved the lowest training RMSE (0.0895) and the lowest test RMSE (0.1050), demonstrating strong predictive performance with minimal overfitting. This suggests that Random Forest effectively captured the complexities in the training data while generalizing well to unseen data. Linear Regression and LOESS performed similarly on the test set, with LOESS (0.1060) slightly outperforming Linear Regression (0.1068). The Decision Tree model had the highest test RMSE (0.1116), indicating it struggled to capture the underlying patterns in the data as effectively as the other models.

To further evaluate the reliability of Random Forest, 10-fold cross-validation was performed. The results showed an average RMSE of 0.1069, an R-squared value of 0.3996, and a Mean Absolute Error (MAE) of 0.0859. These metrics suggest that while Random Forest performs well, there is room for improvement in explaining variance in the data, as indicated by the moderate R-squared value. The tuning parameter mtry was held constant at 2, which may warrant further optimization to enhance performance.

An analysis of feature importance, as shown in the variable importance plot, reveals that API and GOR are the most influential factors in predicting ORF, followed by PI and POROSITY. The prominence of API and GOR suggests that fluid properties play a crucial role in determining ORF, which aligns with domain expectations. Variables such as SW, PERMEABILITY, and THK had minimal impact on the model’s predictions. This insight suggests that future model refinements could involve focusing on the most influential variables while potentially reducing less impactful features to streamline computations.

These results highlight the advantage of using ensemble methods such as Random Forest for prediction tasks, as it outperformed the other models in both training and test accuracy. However, further hyperparameter tuning and feature selection could enhance its performance even further. Additionally, the comparable performance of Linear Regression and LOESS suggests that some relationships in the data may be well approximated using simpler models.

8 Conclusion

- Random Forest performed best, achieving the lowest training RMSE (0.0895) and test RMSE (0.1050), indicating strong predictive performance with minimal overfitting.
- 10-fold cross-validation confirmed model reliability, with an average RMSE of 0.1069, an R^2 value of 0.3996, and an MAE of 0.0859, though the moderate R^2 suggests room for improvement.
- Feature importance analysis revealed that API and GOR were the most significant predictors of ORF, followed by PI and POROSITY. Variables such as SW, PERMEABILITY, and THK had minimal impact.
- Linear Regression and LOESS showed similar performance, suggesting that some relationships in the data may be well-approximated using simpler models.
- The Decision Tree model had the highest test RMSE (0.1116), indicating it struggled to capture complex patterns as effectively as ensemble methods.
- Further improvements can be made by optimizing hyperparameters (e.g., tuning `mtry`), refining feature selection, and exploring additional ensemble techniques.

References

BSEE Repository (https://www.data.bsee.gov/GGStudies/Files/2020%20Atlas%20Update.zip (https://www.data.bsee.gov/GGStudies/Files/2020%20Atlas%20Update.zip))