Romanian Journal of Petroleum & Gas Technology
Vol. V (LXXVI) • No. 2/2024

PHYSICS-DRIVEN FEATURE CREATION TO IMPROVE MACHINE LEARNING MODELS PERFORMANCE FOR OIL PRODUCTION RATE PREDICTION

Eghbal Motaei [1]

Seyed Mehdi Tabatabai [1]

Tarek Ganat [2]

Ahmad Khanifar [1]

Sulaiman Dzaiy [3]

Timur Chis [3]

[1] Petroleum Engineering Department, Petronas Carigali SDN BHD, Malaysia
[2] Sultan Qaboos University, Oman
[3] Petroleum-Gas University of Ploiesti, Romania
email (corresponding author): mehdi.tabatabai@gmail.com

DOI: 10.51865/JPGT.2024.02.22

ABSTRACT

This paper aims to develop a machine learning-based model for oil production rate prediction. The significance of feature dimension reduction is addressed by applying well-established approaches like Principal Component Analysis (PCA) and the proposed physics-driven feature creation technique. The physics-driven features, derived from experience or analytical modeling, introduce physical relevance and improve model quality. The study focuses on oil production prediction using a dataset that includes reservoir permeability, wellbore skin, reservoir pressure, net pay thickness, water cut, and well-liquid production rate. Several machine learning techniques, such as SVM, k-NN, Decision Tree, Random Forest, and linear regression, were constructed using PCA feature selection. The models were tuned and validated using k-fold cross-validation. The same models were then built using physics-driven features, and their performance metrics were compared. The results show significant improvement when applying the proposed physics-driven feature creation, compared to PCA. Over 10-fold cross-validation, PCA improved the R^2 performance metric by 10% (from 70% to 77%), while physics-driven features increased it by 20% (from 70% to 90% on average). The Random Forest and linear regression models outperformed the others, particularly when built based on physics-driven features. Additionally, models based on physics-driven features exhibited less sensitivity to data splits for learning and testing, proving more reliable with better performance metrics compared to those using original features.

Keywords: Oil rate prediction, Feature Engineering, Principal Component Analysis, Artificial Intelligence, Machine Learning

INTRODUCTION

The digital transformation has introduced numerous Artificial Intelligence (AI) approaches to Decline Curve Analysis (DCA) for oil production rate forecasts. DCA, a well-established method since 1944 [1], has proven more promising than time-consuming numerical or analytical approaches, even with current advancements in computational capacity. The primary advantage of DCA is its time efficiency, allowing for simple application across hundreds of wells [2]. Since the early 2000s, machine learning techniques have been extensively applied to oil production rate estimation [3], aiming for time-reliable predictions using datasets generated from analytical or numerical simulation models [4]. Machine learning has proven to be powerful in numerous geoscience applications [5,6,7], including oil production prediction [8]. Researchers have employed various machine learning techniques for this purpose, such as Support Vector Machine (SVM) [9,10] and Random Forest (RF) [11]. The latter study utilized SVM and RF to predict oil production rates as part of supply chain optimization.

In machine learning approaches, feature selection typically involves choosing variables based on their correlation with target values, ensuring that the model is built using the most impactful data [10]. A well-established technique for feature selection and dimension reduction is Principal Component Analysis (PCA). This method is commonly used to replace raw data features with new, hybrid features that capture the most significant variations in the dataset [12]. While conventional dimension reduction approaches like Principal Component Analysis (PCA) often risk compromising machine learning model quality, this study introduces a novel technique that aims to reduce feature dimensions while simultaneously enhancing model performance. The key innovation lies in bridging the gap between dimension reduction and model quality improvement by formulating physical phenomena into features. This is achieved by generalizing the dependency relationships between measured features and target values, effectively incorporating domain-specific knowledge into the feature engineering process.

In many conventional approaches, such as Principal Component Analysis (PCA), dimension reduction often comes at the cost of decreased machine learning model quality. This study aims to introduce a novel technique that achieves dimension reduction while simultaneously improving model quality. The key innovation lies in bridging model quality improvement and dimension reduction by formulating physical phenomena into features. This is accomplished by generalizing the dependency relationships between measured features and the target value. Physics-driven feature creation is applicable to any engineering case where a formulable phenomenon or well-established practical understanding exists. This approach has been used under various names, such as knowledge-driven features [13]. The significance of these features lies in their ability to transform raw data into a physical space, thereby strengthening the feature dependency on the target value.

A clear example of this concept is the classification of overweight individuals using two measured features: height and weight. In this case, Body Mass Index (BMI) can effectively replace these two features, reducing dimensions while maintaining or even improving model quality. BMI itself serves as an indicator of weight class, and its range can be directly correlated to the target value.

This paper introduces a similar concept, applying analytical or practical correlations of interdependent features to create new, more meaningful features. These engineered features replace raw features in building more efficient machine learning models. By incorporating domain knowledge into feature engineering, this approach aims to enhance model performance while reducing dimensionality. An additional significant advantage of physics-driven feature creation is the reduced dependency of model quality on data selection. In conventional approaches like PCA, the input data is crucial; if the samples don't adequately cover all data ranges, the machine learning output can be highly erroneous. However, physics-driven feature creation mitigates this issue. In this approach, features are maintained within physical possibilities in the new feature domain. The feature transformation is consistently guided by physical phenomena, which serves as a safeguard, ensuring that outputs remain within the engineering domain. This inherent constraint helps avoid erroneous predictions in cases of out-of-range inputs.

DATA SOURCE AND DATA QUALITY CONTROL

This study was conducted on a sandstone reservoir featuring more than 10 production strings with multiple production layers. The raw dataset comprises permeability, skin factor, reservoir pressure, water cut, net pay thickness, and well rate. Permeability was calculated from petrophysical logs, with the arithmetic average of permeability values used for the full net pay interval. Net pay thickness was determined based on shale content cut-off. Water cut was extracted from well test reports, while the skin factor was calculated from well test results, including pressure build-up analysis or model regression.

Post-data collection, all raw data was input into an analytical well model for quality control purposes. Data points that could not be matched within a 20% tolerance using the analytical well model were considered outliers and removed from the analysis. The ranges of all parameters in the dataset are visualized in Figure 1. This data exploration reveals that the permeability of the wells ranges from 18 to 42 mD, the water cut is between 5% to 30%, the skin factor is ranged from 12 to 16, the net pay thickness is less than 6 meters, the reservoir pressure is less than 2000 psi (divided by 100 in the plot to fit the y-scale), and the production rate is less than 211 stb/d (divided by 10 in the plot to fit the y-scale). This visualization provides a comprehensive overview of the data distribution across all features.

To analyze the features correlation, Pearson's correlation coefficient is used. Pearson's correlation coefficient is a measure of the linear correlation between two independent variables within a dataset. It is calculated as a parameter between -1 and +1, where the absolute value represents the degree of correlation between the two variables. The sign of the coefficient indicates whether the correlation is direct (+) or inverse (-). Mathematically, Pearson's correlation coefficient is calculated by dividing the covariance of the two variables by their respective standard deviations [14]. This statistical measure provides insight into the strength and direction of the linear relationship between the features in the dataset.

Figure 1. *Data Exploratory Analysis for dataset. Some parameters were rescaled to fit into the scale such as oil production rate which is divided by 10 and pressure which is divided by 100.*

MATERIALS AND METHODS

The paper aims to introduce an approach for feature creation based on the physical correlation between the raw features and target values. To achieve this, three different machine learning models are constructed:

1. The first model uses the raw features as input.
2. The second model uses features derived by Principal Component Analysis (PCA).
3. The third model utilizes physics-driven features.

Finally, the performance metrics of each model are addressed and compared. Figure 2 represents the general workflow applied for all three machine learning models. The target variable in this study is the current production rate, which is dependent on the most recent water cut, the latest effective permeability, recent skin, measured reservoir pressure and active net production pay interval.

FEATURE ANALYSIS

Original Features

The statistics of the available data are presented in Table 1. This table also includes the Pearson correlation coefficient between each feature and the target value. As shown in Table 1, the reservoir pressure has the highest Pearson correlation coefficient of 0.68, indicating a strong positive linear relationship with the target variable. Reservoir permeability is the second most important feature, with a direct correlation coefficient of 0.48. Net pay thickness has a moderate positive correlation of 0.45.

On the other hand, water cut has a negative correlation coefficient of -0.17, suggesting an inverse relationship with the target. The least correlated feature is skin factor, with a correlation coefficient of -0.09. Despite the varying correlation strengths, it was decided to include all five raw features in the modeling process. Even the low correlation feature, skin factor, was retained for analysis, as there was no feature with no impact on the target variable.

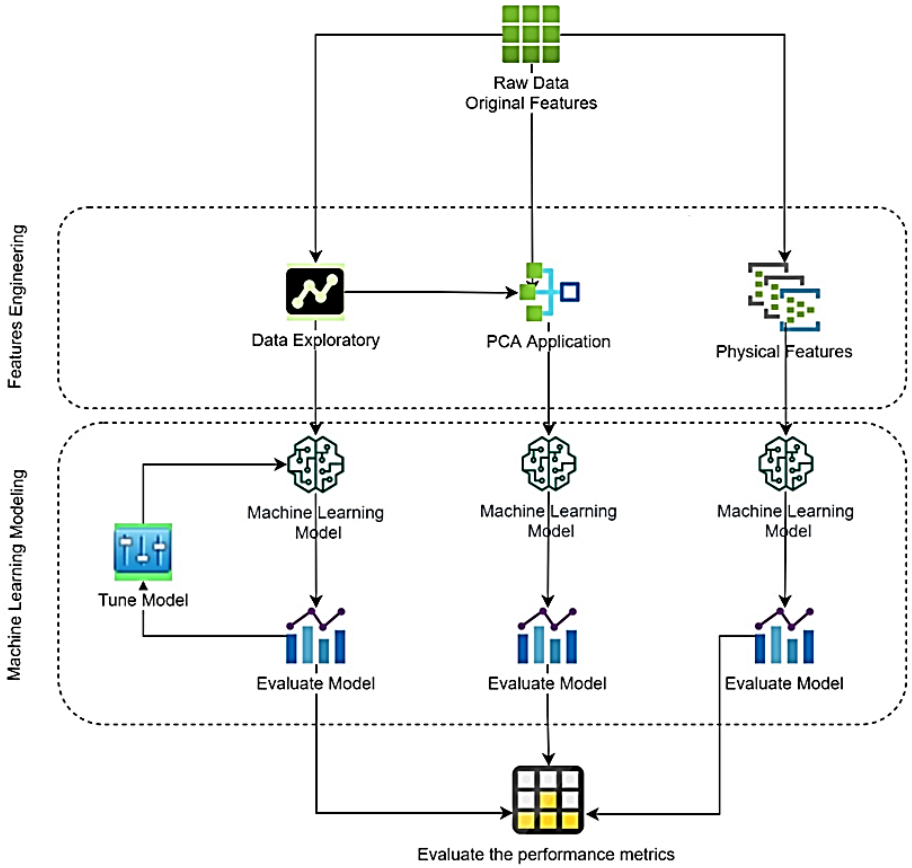

Figure 2. *Diagram of machine learning models for the three models: 1) normal feature selection, 2) feature transform by PCA, and 3) using derived physics-driven features.*

Table 1. *Statistical summary and feature correlations with oil rate for a dataset comprising 86 samples from 10 wells.*

Feature	Mean	Mode	Median	Min.	Max.	Pearson coefficient
Permeability	30	18	30	18	42	**0.48**
Net pay	4	3	4	2	6	**0.45**
Skin	14	12	14	12	16	**-0.09**
Pressure	1482	1684	1517	890	1981	**0.68**
Water cut	18	13	16	6	30	**0.17**
Rate	78	70	71	24	211	Target

Principal Component Analysis

Principal Component Analysis (PCA) is done on many machine learning applications to reduce the features and replace the original features with new features that are more correlated and dimensionally also reduced to make the models more efficient [12]–[18]. As the aim is to reduce the features count, in model construction with PCA, the three principal components are used in the modelling construction.

Principal Component Analysis (PCA) is commonly applied in various machine learning applications to reduce the number of features and replace the original features with new ones that are more correlated. This dimensionality reduction helps make models more efficient [15]–[21]. To achieve feature reduction, three principal components are utilized in the model construction with PCA.

Physics-driven features

Referring to the diffusivity equation which explains the fluid flow in porous media through the below established equation (1) [22]:

$$\nabla^2 P = \frac{\emptyset \mu C}{k} \frac{\partial P}{\partial t} \tag{1}$$

Where P is the pressure in psi; $\emptyset$ is porosity in fraction; μ is viscosity in cP; C is total system compressibility in psia-1; k is reservoir permeability in mD; and $\frac{\partial P}{\partial t}$ is pressure gradient across the time in Psia.s-1.

The above equation obtained by combining material balance equation, equation of state, and Darcy law and it carries the assumptions of those three main equations, especially on the fluid compressibility which makes it limited to only liquid flow and must be modified for gas wells. Solving the above system equation in the pseudo steady state model will lead to equation (2) (Stewart, 2011):

$$P - P_{wf} = \frac{q_o B\, \mu}{0.0078\ k\,.\,h} \ln\left(\frac{r_e}{r_w} - 0.75 + S\right) \tag{2}$$

Where: P is the pressure in psi; μ is viscosity in centipoise (cP); r_e is external reservoir radius in feet; r_w is wellbore radius in feet; k is reservoir permeability in milliDarcy (mD); S is wellbore skin, dimensionless; h is the net pay thickness in ft; B is formation volume coefficient, dimensionless in rbl/Stb; P_{wf} is flowing pressure in psi, and q_o is oil production rate in Stb/day.

The above equation serves as a physical guide for predicting oil production using the defined new features. Two new features, *A* and *B*, are derived to replace the original ones. Feature *A* is based on the linear relationship of the total liquid flow rate, as represented in equation (2). To convert this into an oil rate prediction, the water cut parameter must be applied. Water cut is defined as the ratio of water rate to total liquid production rate, expressed as a percentage. To transform the liquid production rate into an oil production rate, the term (100 - wc) is divided by the liquid rate. Thus, equation (2) can be rearranged for oil production rate:

$$P - P_{wf} = \frac{q_o\, B\, \mu}{0.0078\ k\,.\,h\ (100 - wc)} \ln\left(\frac{r_e}{r_w} - 0.75 + S\right) \tag{3}$$

In which the wc is water cut in percentage and q_o is oil production rate.

Now, based on equation (3), the first physical feature can be driven as:

$$A = k\,.\,h\,.\,P.\,(100 - wc) \tag{4}$$

The new feature *B*, is extracted:

$$B = \frac{P}{S} \tag{5}$$

Equations (4) and (5) are introduced in this study as physics-driven features derived from the insights gained from equation (2). Feature *A* is based on equation (2), which originates from the pseudo-steady state equation, indicating that oil flow rate is proportional to reservoir permeability, net pay thickness, and reservoir pressure. Feature *B* is defined based on the inverse relationship of the skin factor, as described in equation (3), and direct relationship of fractional flow of oil. In the machine learning model, both features *A* and *B* are utilized to represent the physical relationship between the measured well test parameters (referred to as features) and the target variable oil rate.

Modelling Methodology

After data processing and feature selection and creation, the next step was to develop the machine learning model. Five types of machine learning algorithms – Support Vector Machine (SVM), Linear Regression, k-Nearest Neighbors (k-NN), Decision Tree, and Random Forest – were selected to evaluate potential model improvements using physics-driven features. Cross-validation with 10 folds was employed in constructing all models. For each subset, 85% of the data was used for training and 15% for testing.

The *k*-fold in the ensembles refers to the cross-validation process itself, indicating 10 random splits of the data. In this study, 10 folds represent 10 different iterations, during which the model's quality may fluctuate; the average quality across these iterations will be used as the performance metric. The same concept will also apply when evaluating the introduction of the physics-based features. Data was cleaned for outliers and subsequently loaded, followed by an analysis of Pearson correlation for each dataset to select features for modeling. The machine learning model was then established, tested, and scored using the 10-fold cross-validation process. The workflow of this modeling approach is illustrated in Figure 3.

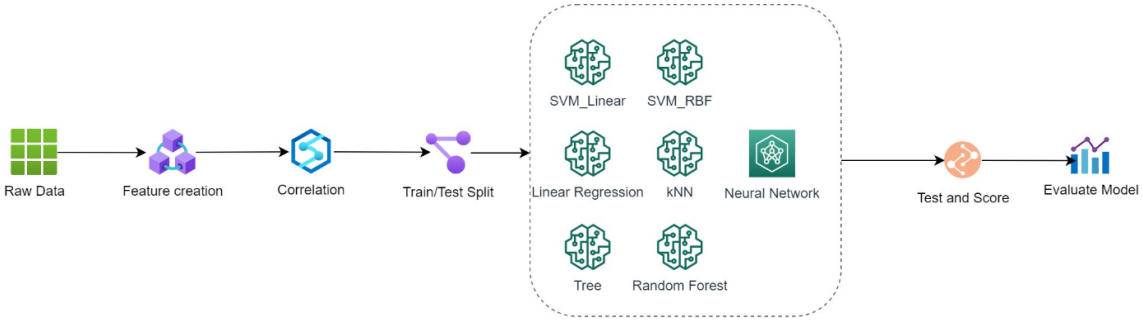

Figure 3. *General workflow for machine learning application*

Support Vector Machine Model

The Support Vector Machine (SVM) model demonstrates strong performance in identifying hyperplanes to separate and classify data. It functions as a Support Vector Classifier (SVC) and is enhanced by applying a feature space through various kernels [23]. These kernels can be linear, Gaussian Radial Basis Function (RBF), polynomial,

sigmoid, or other types. In this study, the SVM model utilizes two kernels: linear and RBF. Default values for all hyperparameters are employed across the three different models.

Linear Regression Model

A linear regression model is a classical supervised machine learning approach [14] known for its efficiency. The algorithm aims to fit a straight line that minimizes the error between the predicted target values and the measured training values. Essentially, it performs curve fitting. However, when dealing with multiple features, it becomes an optimization task to minimize the error, treating it as a cost function influenced by all given features. The model is particularly effective with continuous data types.

Decision Tree and Random Forest (RF) models

Classifiers as Decision Tree and Random Forest (RF) models have been applied in petroleum engineering to predict well productivity [24] and have previously been used for assisted history matching [25], demonstrating the efficiency of these models for prediction tasks. A Decision Tree utilizes a tree-like structure to make decisions aimed at achieving a specific goal. It accomplishes this by splitting the dataset based on features to meet the classification objective and minimize error. In contrast, a Random Forest consists of multiple decision trees, each built from the same dataset but using different subsets. This approach averages the votes from the individual trees and reports the most common outcome among the predicted values.

Model Setup

All machine learning models were trained using 10 data folds in the cross-validation process. For each dataset, this involves splitting the data into 10 distinct sets, resulting in the training, testing, and scoring of 10 different models. Each dataset is divided into training and testing subsets for the evaluation of each module.

Model Testing and Evaluation

Predicting the oil production rate is a regression problem. Several performance metrics can be employed, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2. In this study, R^2 is utilized as the primary performance measure for the model, defined based on the Normalized Mean Squared Error, as indicated in the equations 6 and 7 below [26]:

$$F(X) = NMSE = \frac{MSE}{variance\ (y_k)} = \frac{\sum_{i=1}^{n}(\hat{y}_k - y_k)^2}{\sum_{i=1}^{n}(\hat{y}_k - \bar{y}_k)^2} \tag{6}$$

$$R^2 = 1 - NMSE \tag{7}$$

All models will be evaluated for each fold based on the R^2 value, and the final R^2 score of each model will be used for model assessment. In this study, three primary models were developed, with the objective of evaluating these models to analyze the impact of features on enhancing their performance metrics.

RESULTS AND DISCUSSION

Feature Analysis Results

Principal Component Analysis

In this study, PCA is utilized to generate new features with higher correlations. Through PCA, five new features were introduced: PC1, PC2, PC3, PC4, and PC5, with a minimum correlation coefficient of 0.03 and a maximum correlation coefficient of 0.755, as depicted in Figure 4. These new features show improved correlation coefficients compared to the original features. While PC1, PC2, PC3, and PC4 exhibit strong correlations, PC5 has a very low correlation and was excluded from the modeling due to its minimal impact on the model relative to the original features.

Figure 4 illustrates the correlations of the new features from the PCA analysis, highlighting the sensitivity or impact of each feature on the oil production rate target. Based on the defined color range, permeability is identified as the most impactful parameter, while skin is noted as the least impactful among the original features.

Figure 4. *Detailed Pearson Correlation for all features including original, PCA, and physics-driven features.*

Physics-driven Features

The correlation coefficients between features A and B and the target value are calculated to be 0.96 and 0.67, respectively. Details of the correlation for all features are presented in Table 2. There is a clear reduction in feature count from five original features – permeability, skin, pressure, water cut, and net pay – to two features, A and B. This reduction leads to an improvement of up to 96% in Pearson correlation, highlighting the significance of the synthetic features derived from the governing physics of the production phenomenon in porous media.

Throughout this work, all machine learning models utilized 10 folds for training and evaluation. Each fold corresponds to a model trained using a specific split of the data for training and testing. In these 10 folds, all models have dedicated test data, while the training data overlap, as illustrated in Figure 5. For example, in fold 1, the first 15% of the data is used for testing, while in the second fold, the subsequent 15% is used. This approach demonstrates the model quality in relation to the data splits.

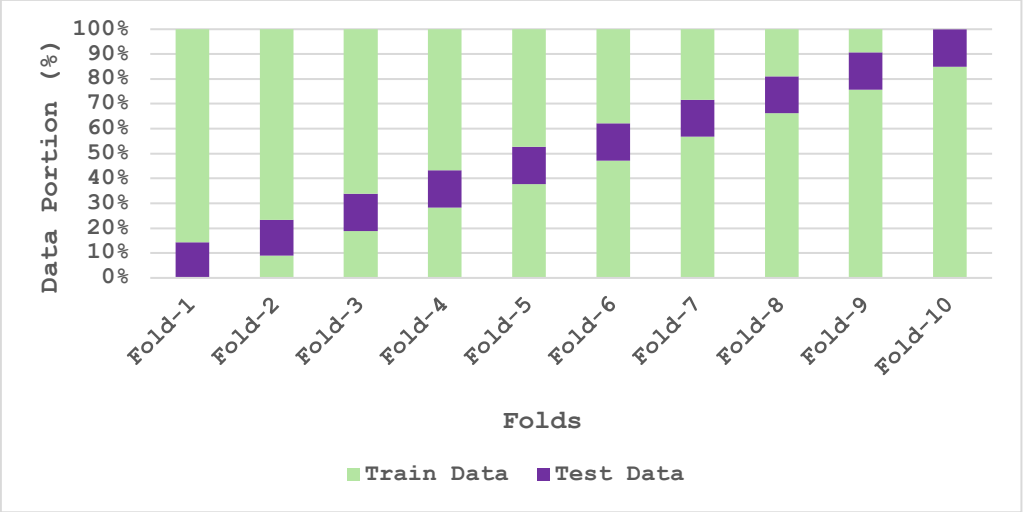

Figure 5. *Schematic of machine learning model configuration with 10 folds that utilizes the 85% of data as training and 15% as testing.*

Support Vector Machine Model Results

The results of the SVM model for linear and RBF kernels are presented in Figure 6. The Y-axis shows the rate predicted by SVM, while the X-axis represents the true measured data. A unit slope trend indicates the best model quality with an R2 value of 1.0, while deviations from this slope demonstrate discrepancies between the model's predictions and measured oil rates. For the linear kernel, the model's performance improves significantly when using physics-driven features compared to both PCA and original features. Interestingly, PCA did not enhance the SVM model with a linear kernel.

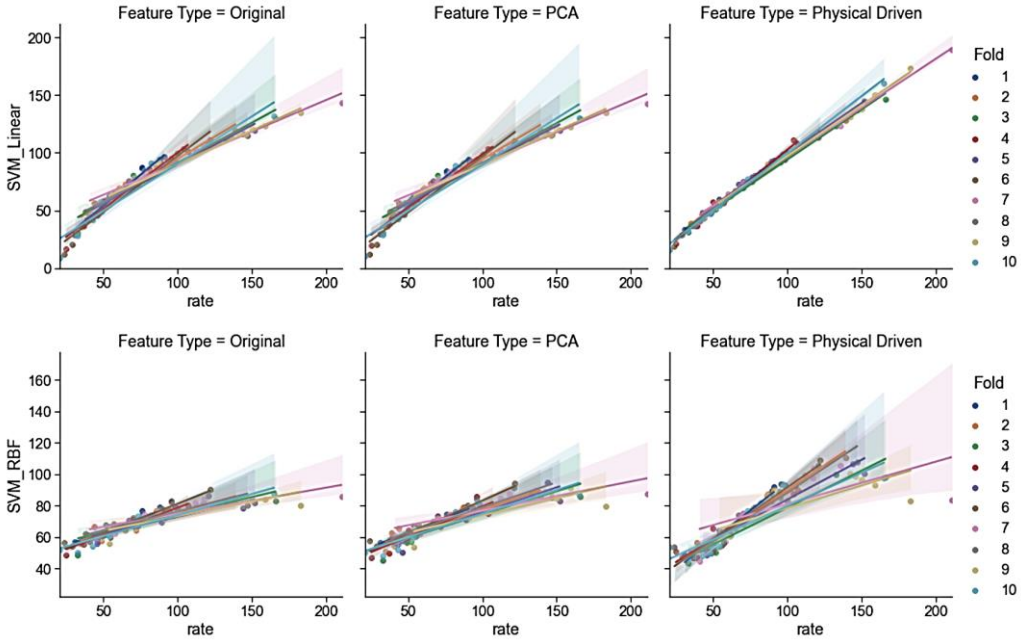

Figure 6. *SVM model quality comparison across feature types: original, PCA, and physics-driven. The Y-axis represents the oil rate predicted by the SVM model, while the X-axis shows the measured oil production rate. A unit slope (diagonal line) indicates high model quality with an R2 value of 1.0. Deviations from this slope represent lower R2 values and reduced model performance.*

In contrast, the SVM model with an RBF kernel shows low quality across all three feature types. However, the PCA-based model demonstrates significant improvement over the original features, while the physics-driven features only slightly enhance model quality. The SVM models using original features (K, S, P, WC, and h) and PCA are sensitive to data split and fold. Conversely, the SVM model with synthetic physics-driven features (A and B) maintains consistent quality across folds when using a linear kernel. However, the radial basis kernel becomes highly sensitive to fold changes.

Linear Regression Model Results

In this study, the model performance using original features is already reasonably high, with an R2 value of 0.928. The PCA approach did not improve upon this R2 metric. However, the physics-driven features significantly enhanced the linear regression model performance, increasing the R2 value to 0.991. Another improvement brought about by the new features is the reduced sensitivity of the model to cross-validation folds. As illustrated in Figure 7, the model performance remains stable across different folds when using physics-driven features, regardless of the specific fold chosen.

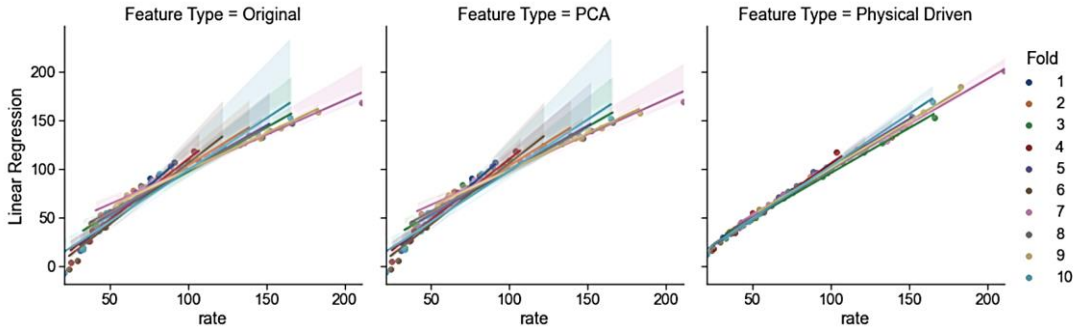

Figure 7. *Linear Regression model for all feature types. Physics-driven features outperform other data folds in terms of sensitivity to fold and accuracy in model performance measures. X-axis is the real measured oil rate and y-axis is the predicted model results from machine learning models.*

K-Nearest Neighbours (kNN) Model Results

In the kNN model setup, we consistently used 10 neighbors with Euclidean Metric across all three models to ensure consistency and isolate the impact of features on model predictivity. The kNN approach applied to the original features yielded unsatisfactory results. However, as illustrated in Figure 8, PCA significantly improved its performance. The physics-driven features demonstrated similar improvement, achieving an even higher R2 value and exhibiting less sensitivity to dataset folds. Notably, in the kNN model, all folds maintained consistent quality regardless of the feature set used.

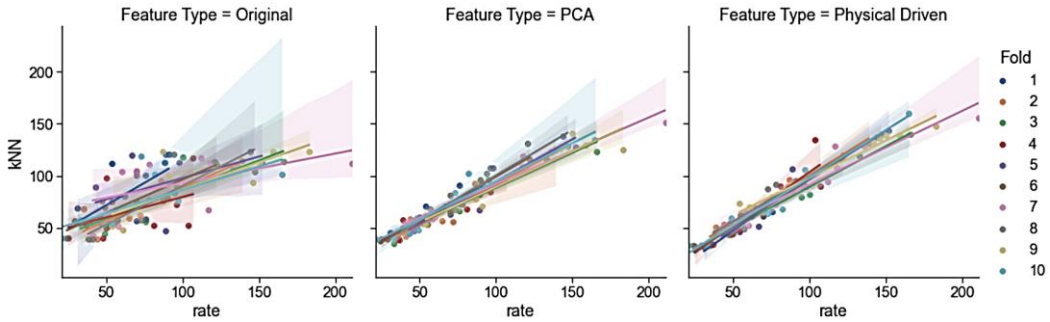

Figure 8. *kNN machine learning model for all feature types. The model performance metrics of the 3 models are 0.5, 0.87, and 0.92 for original, PCA, and Physics-driven features, respectively.*

Decision Tree and Random Forest (RF) Models Results

Following standard procedure, we maintained a consistent setup across all three models. Figure 9 illustrates the results of these models. Both the models constructed using original features and PCA-derived features demonstrate low quality and high sensitivity to folds. In contrast, the model utilizing physics-driven features shows significantly improved performance, achieving a high-quality prediction with an R2 value of 0.95. This physics-driven model also exhibits greater stability across different folds.

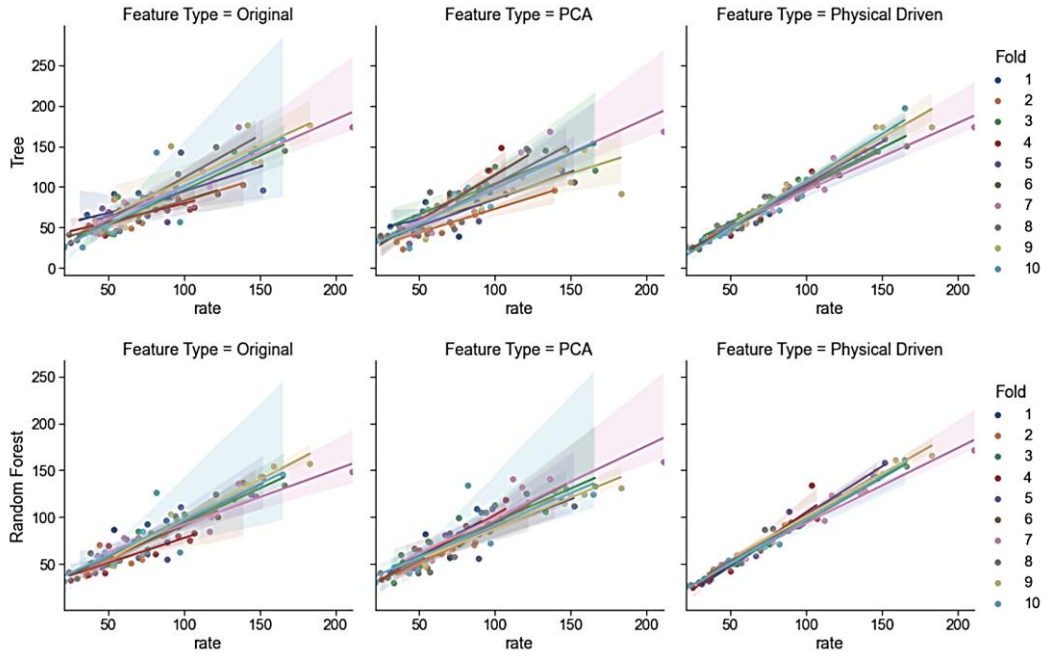

Figure 9. *Decision Tree and Random Forest machine learning models for all feature types. A reliable model prediction could be done through new features compared to original and PCA features.*

Model Performance Metrics

The main performance metrics of the models, including RMSE, MAE, and R2, are used to evaluate their effectiveness, as summarized in Table 2. Among all machine learning approaches, most have shown improvement with the introduction of physics-driven features. While PCA enhances the model's performance metrics, especially the R2 value, the implementation of physics-driven features significantly improves this performance metric across all models. Figure 10 presents an infographic comparing the improvement of PCA over original features and the improvement coefficient over PCA. As illustrated, PCA increases model efficiency by reducing the number of active features from five original features and improving the R2 value by 7% (from 70% to 77%). In contrast, the model based on physics-driven features improves the R2 value by 20% (from 70% to 90%) on average.

A closer examination of the kNN model reveals that the original model prediction is disappointing, with an R2 value of 50%. Applying PCA significantly enhances the kNN model's efficiency, achieving an R2 of 87%. However, the proposed physics-driven approach further improves the quality to an excellent margin. Among all machine learning models, linear regression demonstrates the best performance with original features. PCA

did not improve this already excellent model. However, the physics-driven approach further enhanced it to a super-accurate model, improving the R2 value from 93% to 99%. A similar trend is observed in SVM with a linear kernel, where PCA showed no improvement, but the proposed model increased performance from 87% to 98%.

Table 2. *Summary of machine learning model performance metrics for all feature types for cross validation model*

Model	RMSE	MAE	R^2	Feature Type
Tree	20.1	15.4	74%	Original
SVM_RBF	31.9	22.0	35%	
SVM_Linear	14.3	9.0	87%	
Random Forest	15.5	11.5	85%	
Linear Regression	10.6	8.0	93%	
kNN	27.9	21.3	50%	
Tree	20.8	16.0	72%	PCA
SVM_RBF	30.2	20.3	41%	
SVM_Linear	14.4	9.0	87%	
Random Forest	16.9	12.5	82%	
Linear Regression	10.5	7.9	93%	
kNN	14.1	10.0	87%	
Tree	9.1	6.5	95%	Physics-driven
SVM_RBF	25.0	13.4	60%	
SVM_Linear	5.2	3.4	98%	
Random Forest	7.1	4.4	97%	
Linear Regression	3.7	2.7	99%	
kNN	11.1	7.8	92%	

Model	MSE	RMSE	MAE	R2	Feature Type
Average of all Models	459.5	20.1	14.5	70%	Original
	358.2	17.8	12.6	77%	PCA
	154.1	10.2	6.4	90%	Physics-based

Figure 10. *Improvement of overall machine learning model performance metric over 10 folds of cross-validation from original 70% to 77% and 99% for PCA and Physics-driven features, respectively.*

Figure 11 provides a closer view of the linear regression model, illustrating the improvement achieved through the proposed feature creation. It's important to note that for models using physics-based features, an 85-15 split (85% training, 15% test data) was employed to ensure model accuracy. The reported R2 values as performance metrics are based on the test portion of the data. Another noteworthy finding is that for all machine learning models, except SVM with RBF kernel, the R2 performance metric improved to above 92% using the proposed feature selection model.

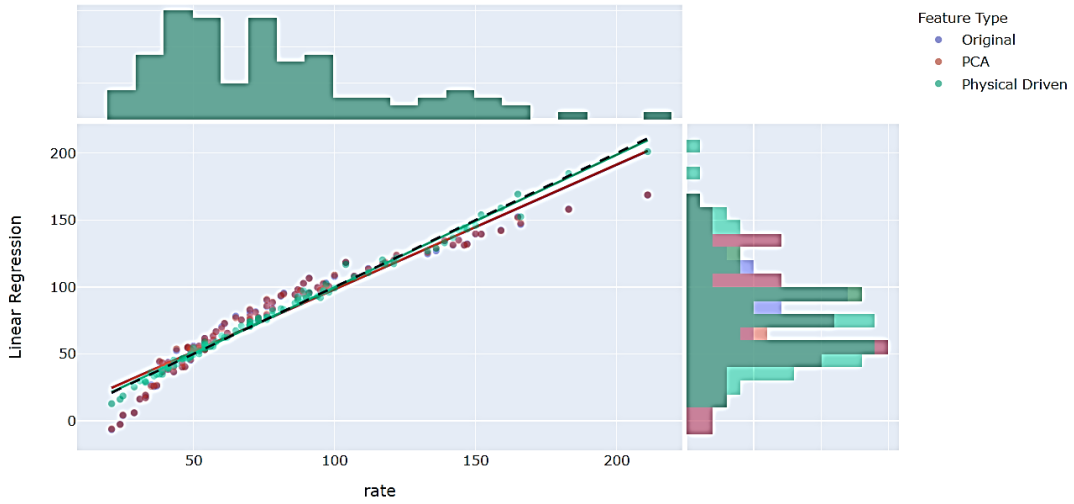

Figure 11. *Improvement of linear regression model using new proposed feature creation*

CONCLUSIONS

This paper proposes a feature creation approach that leverages understanding gained from the physical relationship between the target value and measured features to develop more efficient models. Machine learning models built on these physics-driven features demonstrate reduced sensitivity to data folds or splits for learning and testing, making them more reliable compared to those using original features. Key conclusions from the study include:

- Models constructed using physics-driven features consistently outperform others in rate prediction.
- Physics-driven features effectively reduce feature dimensionality.
- Machine learning models based on physics-driven features exhibit lower sensitivity to data folds.
- This approach competes favorably with well-established methods such as PCA.
- Regardless of the specific machine learning algorithm, physics-driven features improve model quality.
- In most cases, this approach enhances model quality to achieve high efficiency (R2 > 0.9), with the exception of SVM using RBF kernel.

Recommendations for Future Works

The model's applicability can be extended to more complex cases with higher dimensions, exploring its scalability and effectiveness in diverse scenarios. To fully realize the potential of this approach, we recommend conducting comprehensive hyperparameter tuning for machine learning models at each step, rather than relying solely on settings optimized for original features. This process should be applied equally to models using original, PCA-derived, and physics-driven features to ensure a fair comparison.

It's important to note that the reduction of features through this physics-driven approach may alter the optimal model configuration. The transformed feature space might require

different model architectures or learning strategies to fully leverage the new representation. Therefore, careful consideration should be given to re-optimizing models with the new features, as the best configuration for the original feature set may not be optimal for the physics-driven features.

REFERENCES

[1] Ukwu A.K., Onyekonwu M.O., Ikiensikimama S.S., Decline curve analysis using combined linear and nonlinear regression, *Soc. Pet. Eng. - SPE Niger. Annu. Int. Conf. Exhib. NAICE 2015*, 2015, doi: 10.2118/178295-ms.

[2] Meribout M., Azzi A., Ghendour N., Kharoua N., Khezzar L., AlHosani E., Multiphase Flow Meters Targeting Oil & Gas Industries, *Meas. J. Int. Meas. Confed.*, vol. 165, p. 108111, 2020, doi: 10.1016/j.measurement.2020.108111.

[3] Yeten B., Durlofsky L.J., Khalid A., Optimization of Smart Well Control, SPE Conference, Calgary, Alberta, Canada, 2002, doi: 10.2523/79031-ms.

[4] Zhong Z., Sun A.Y., Wang Y., Ren B., Predicting field production rates for waterflooding using a machine learning-based proxy model, *J. Pet. Sci. Eng.*, vol. 194, Nov. 2020, doi: 10.1016/j.petrol.2020.107574.

[5] Tabatabai S.M., Chis T., Jugastreanu C., Formation evaluation in low resistivity low contrast (LRLC) shaly sand thin lamination; forward modeling and inversion optimization using genetic algorithm; *Romanian Journal of Petroleum & Gas Technology*, vol. 3, no. 1, pp. 83-97, 2022, DOI: 10.51865/JPGT.2022.01.09

[6] Robail F., Sanyal S., Gazali M.I., Azudin A.N.B.M.N., Tabatabai S.M., Wulandar R., Zulfikar L., When machine learning helps to understand a facies controlled two ways tilted contact in a large carbonate reservoir, International Petroleum Technology Conference, Feb 2024, DOI: 10.2523/IPTC-23843-MS

[7] Daud M.S.H., Tabatabai M.S., Wong F.K., Cascaded machine learning in NMR: unveiling a continuous grain size distribution approach for tackling sand production challenges, SPWLA 65th Annual Logging Symposium, Rio de Janeiro, Brazil, May 18-22, 2024.

[8] Zhang Z., Li H., Zhang D., Reservoir characterization and production optimization using the ensemble-based optimization method and multi-layer capacitance-resistive models, *J. Pet. Sci. Eng.*, vol. 156, pp. 633–653, 2017.

[9] Davtyan A., Rodin A., Muchnik I., Romashkin A., Oil production forecast models based on sliding window regression, *J. Pet. Sci. Eng.*, vol. 195, Dec. 2020.

[10] Fulford D.S., Bowie B., Berry M.E., Bowen B., Turk D.W., Machine learning as a reliable technology for evaluating time/rate performance of unconventional wells, *SPE Econ. Manag.*, vol. 8, no. 1, pp. 23–39, Jan. 2016, doi: 10.2118/174784-PA.

[11] Asala H.I., Chebeir J.A., Manee V., Gupta I., Dahi-Taleghani A., Romagnoli J.A., An integrated machine-learning approach to shale-gas supply-chain optimization and refrac candidate identification, in *SPE Reservoir Evaluation and Engineering*, 2019, vol. 22, no. 4, pp. 1201–1224, doi: 10.2118/187361-PA.

[12] Aïfa T., Neural network applications to reservoirs: Physics-driven models and data models, *Journal of Petroleum Science and Engineering*, vol. 123. Elsevier, pp. 1–6, Nov. 01, 2014, doi: 10.1016/j.petrol.2014.10.015.

[13] Yeomans C.M., Shail R.K., Grebby S., Nykänen V., Middleton M., Lusty P.A.J., A machine learning approach to tungsten prospectivity modelling using knowledge-driven feature extraction and model confidence, *Geosci. Front.*, vol. 11, no. 6, pp. 2067–2081, 2020, doi: 10.1016/j.gsf.2020.05.016.

[14] Vrabie I. *et al.*, Digital twin for downhole pressure gauges: Model and field case study, *Soc. Pet. Eng. - SPE Russ. Pet. Technol. Conf. 2020, RPTC 2020*, 2020, doi: 10.2118/201926-ms.

[15] Kim J., Kim S., Park C., Lee K., Construction of prior models for ES-MDA by a deep neural network with a stacked autoencoder for predicting reservoir production, *J. Pet. Sci. Eng.*, vol. 187, Apr. 2020, doi: 10.1016/j.petrol.2019.106800.

[16] Patel R.G., Trivedi J., Rahim S., Li Z., Initial Sampling of Ensemble for Steam-Assisted-Gravity-Drainage-Reservoir HistoryMatching, *J. Can. Pet. Technol.*, vol. 54, no. 6, pp. 424–441, 2015, doi: 10.2118/178927-PA.

[17] Trehan S., Oil Field Production using Machine Learning CS 229 Project Report.

[18] Lee K., Lim J., Ahn S., Kim J., Feature extraction using a deep learning algorithm for uncertainty quantification of channelized reservoirs, *J. Pet. Sci. Eng.*, vol. 171, pp. 1007–1022, Dec. 2018, doi: 10.1016/j.petrol.2018.07.070.

[19] Srikonda R., Rastogi A., Oestensen H., Increasing facility uptime using machine learning and physics-drivenhybrid analytics in a dynamic digital twin, *Proc. Annu. Offshore Technol. Conf.*, vol. 2020-May, 2020, doi: 10.4043/30723-ms.

[20] Scheidt C., Caers J., Representing spatial uncertainty using distances and kernels, *Math. Geosci.*, vol. 41, no. 4, pp. 397–419, 2009, doi: 10.1007/s11004-008-9186-0.

[21] Aïfa T., Neural network applications to reservoirs: Physics-driven models and data models, *J. Pet. Sci. Eng.*, vol. 123, pp. 1–6, 2014, doi: 10.1016/j.petrol.2014.10.015.

[22] Stewart G., *Well Test Design and Analysis*. 2011.

[23] Hastie T., Tibshirani R., James G., Witten D., *An Introduction to Statistical Learning*, *Springer Texts*, vol. 102, 2006.

[24] Attanasi E.D., Freeman P.A., Coburn T.C., Well predictive performance of play-wide and Subarea Random Forest models for Bakken productivity, *J. Pet. Sci. Eng.*, vol. 191, no. December 2019, p. 107150, 2020, doi: 10.1016/j.petrol.2020.107150.

[25] Aulia A., Jeong D., Saaid I.M., Kania D., Shuker M.T., El-Khatib N. A., A Random Forests-based sensitivity analysis framework for assisted history matching, *J. Pet. Sci. Eng.*, vol. 181, no. July, p. 106237, 2019, doi: 10.1016/j.petrol.2019.106237.

[26] Hayder G., Solihin M.I., Mustafa H.M., Modelling of River Flow Using Particle Swarm Optimized Cascade-Forward Neural Networks: A Case Study of Kelantan River in Malaysia, doi: 10.3390/app10238670.

Received:December 2024; Revised:December 2024; Accepted:December 2024; Published:December 2024