A investigation of data quality in reservoir characterizations using Machine Learning

Alesson M. Torres[1], Cristiane O. Faria[2], Karla Figueiredo[2]

[1]*PPG-CComp, UERJ*
São Francisco Xavier, 524, Maracanã, 20550-900, Rio de Janeiro, Brazil
alesson.mansur@pos.ime.uerj.br
[2]*Instituto de Matemática e Estatística - UERJ*
São Francisco Xavier, 524, Maracanã, 20550-900, Rio de Janeiro, Brazil
cofaria@ime.uerj.br, karlafigueiredo@ime.uerj.br

Abstract. Inferring the capacity of reservoirs is one of the essential tasks in the oil and gas exploration process. The characterizations of transport and storage are crucial for reservoir evaluations and, therefore, depend on the permeability and porosity. Although estimating the permeability in porous media is challenging since experimental data gathering is very costly, estimations are not accurate. Machine Learning (ML) methods have been applied to predict the permeability in oil-producing areas as cost-effective and quick characterization strategies. However, the quality predictions of ML algorithms depend on the available data quality and the algorithm parameters optimization. In this work, in order to have a comprehensive understanding, we investigate the permeability inference employing algorithms as Multivariate Linear Regression, Decision Tree Regression, Support Vector Machines (SVM) and Multilayer Perceptron (MLP). The ML approach was constructed and tested via data samples experimentally gathered from Australia and Papua New Guinea region. Data pre-processing metrics are optimized. The most relevant feature was analyzed and optimized parameters improved the inferences as expected. The mean squared error and root mean squared error for the test set are on the order of 0.0066 and 0.0811, respectively, indicating that our results are very promising.

Keywords: Machine Learning, Regression, Permeability

1 Introduction

One of the most important concerns in petroleum engineering is the reservoir characterization. According to Ahmadi and Chen [1], permeability plays an essential role in any reservoir evaluation plan because it is a crucial property for oil wells storage and transport characterizations. Chehrazi and Rezaee [2] discuss that permeability is a key parameter in determining the economic value of hydrocarbon accumulation. Since it regulates the directions of the reservoir fluids and the flow through porous media, Ahmadi and Chen [1] indicate that accurate estimation of permeability is essential for the improvement of oil/gas recovery, CO_2 sequestration, selection of cost-effective production schemes, and optimization of oil well placement. Suitable magnitudes for permeability can be experimentally obtained from core samples or well logs. However, since the coring process is very costly and time intensive, the oil industry requires the development of cost-effective and quick approaches for reservoir evaluation and characterization. Chen et al. [3] and Ahmadi and Chen [1] discussed that to achieve these requirements, statistical analysis, and machine learning algorithms had been widely employed as alternative approaches to estimating the permeability from petrophysical logs and the data from the existing cores.

One of the most challenging concerns about the use of Machine Learning as a modeling approach is data quality. Gudivada et al. [4] enumerate that poor-quality data can be manifested in the form of missing data, duplicate data, highly correlated variables, a large number of variables, and outliers. It can pose significant problems for building Machine Learning models. Consequently, low-quality data can affect negatively the permeability estimations performed by the Machine Learning predictors and not provide results suitable to be used in real problems.

In this paper, we investigate the quality of the data provided by *Porosity and Permeability Database - Record*

1990/88 [5] employing four Machine Learning algorithms to estimate the oil reservoirs permeability. They are: Regularized Multivariate Regression (Ridge) (Montgomery et al. [6]), Decision Tree Regressor (DTR) (Breiman et al. [7]), Support Vector Regressor (SVR) (Awad and Khanna [8]) and Multilayer Perceptron (MLP) (Alpaydin [9]). In addition, data pre-processing techniques and RReliefF feature selection algorithm are employed for data preparation and dimensionality reduction, respectively.

The next sections are organized as follows: Section 2.1 describes the data collecting and data preprocessing steps; Section 2.2 revises the employed Machine Learning algorithms; Section 2.3 establishes the grid search approach; Section 2.4 states the features selection approach; Section 2.5 presents the evaluation metric for model selection; and Section 2.6 details the performance evaluation metrics for the chosen model. Finally, Section 3 analyse the obtained results and Section 4 concludes the paper.

2 Materials and Methods

2.1 Data Acquisition

The assessments were executed with data provided by *Porosity and Permeability Database - Record 1990/88* [5], which is a report consisting of important petro-physical data regarding 551 oil reservoirs from Australia. From the information provided by the report, we gathered the following features: (i) region; (ii) longitude; (iii) latitude; (iv) sample depth; (v) porosity; (vi) if the reservoir is onshore or offshore; and (viii) horizontal permeability. It resulted in 6278 samples.

The exclusion criterion was the missing values: if at least one feature is null, the sample is excluded. It reduced the dataset to 2060 samples, which were split into $1648(80\%)$ samples for training and validation and $412(20\%)$ samples for testing. Since longitude and latitude features are in different ranges from the other ones, all dataset was normalized with mean equal to zero and deviation equal to one.

2.2 Machine Learning Algorithms

Decision Tree Regressor (DTR). Decision Trees are supervised learning methods applied in both regression and classification. The first algorithms are from 1980's and one of the most popular of them is called CART (Classification And Regression Trees). It builds a decision tree by partitioning the feature space into several sub-spaces taking a recursive binary splitting, such that the samples with similar target values are grouped. A common criterion to evaluate the node quality in Decision Tree Regressors (DTR) is the Mean Squared Error (MSE). More details can be obtained in Breiman et al. [7], James et al. [10], and Hastie et al. [11].

Regularized Multivariate Regression (Ridge). Montgomery et al. [6] define a Multivariate Regression model as a regression model that involves more than one independent variable. The dependent variable Y is described as a linear function of k independent variables or regressors: $Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n$. The parameters $b_i, i = 0, 1, \ldots, k$ are called the regression coefficients. This model permits the computation of a regression coefficient b_i for each independent variable X_i and this mathematical operation is usually performed by the least squares method. When the data is nonorthogonal poor estimates of the regression can be obtained. It results in a vector of least squares parameter estimates too far from the average. Consequently, the absolute values of the least squares estimates are too large and they are very unstable. One way to reduce this problem is to shrink the regression coefficients by imposing a penalty on their size. This approach is called Ridge Regression and it can be understood as a regularized multivariate regression. More details can be obtained in Montgomery et al. [6], in James et al. [10], and in Hastie et al. [11].

Support Vector Regressor (SVR). The Support Vector Regressor (SVR) model is a generalization of Support Vector algorithms proposed by Cortes and Vapnik [12] to become applicable to regression problems. According to Awad and Khanna [8], the SVR has been proven to be an effective tool in real-value function estimation. A supervised-learning approach trains a symmetrical loss function, which equally penalizes high and low misestimates. The Vapnik's-insensitive approach is employed to build a flexible tube of minimal radius symmetrically around the estimated function, such that the absolute values of errors less than a certain threshold are ignored both above and below the estimate. It means that points outside the tube are penalized, but those within the tube, either above or below the function, receive no penalty. They also explain that the main advantages of the SVR are its computational complexity that does not depend on the dimensionality of the input space, followed by its excellent generalization capability, with high prediction accuracy. More details can be obtained in Awad and Khanna [8], in Cortes and Vapnik [12], in Smola and Scholkopf [13], and in Hastie et al. [11].

Multilayer Perceptron (MLP). According to Alpaydin [9], a Multilayer Perceptron (MLP) is a feedforward network with intermediate or hidden layers between the input and the output layers. Abirami and Chitra [14] describe the MLP process as follows: The input layer receives the input data to be processed. The output layer performs the required task such as prediction and classification. An arbitrary number of hidden layers placed in between the input and output layer is the actual computational engine of the MLP. Like a feed forward network, the data flow in the forward direction from the input to the output layer. The neurons in the MLP are trained with the backpropagation learning algorithm. MLP's are designed to approximate any continuous function and can solve problems that are not linearly separable. The major use cases of the MLP are pattern classification, recognition, prediction and approximation. When used for regression, the network can approximate nonlinear functions of the input. More details can be obtained in Awad and Khanna [8], and in Alpaydin [9].

2.3 Machine Learning Models Fine-tuning

In order to fine-tune the Machine Learning models, several hyper-parameter values were evaluated by employing the grid search approach with standard 5-fold Cross-Validation on the training set. The optimum hyper-parameter set for a model is chosen by evaluating the best R^2 score value in the fine-tuning process. The evaluated parameters' set for the models are presented in Appendix.

2.4 Feature Selection Investigaton - RReliefF

Aiming to improve the evaluation metrics and the models generalization, a feature selection investigation was employed. The chosen approach is the Regressional ReliefF (RReliefF) algorithm proposed by Robnik-Šikonja and Kononenko [15]. According to Robnik-Šikonja and Kononenko [16], the RReliefF algorithm extends the Relief algorithm for regression. A Relief-based feature selection algorithm calculates a feature score for each feature. This score can then be applied to rank and select top scoring features. Usually, higher scores features represent higher contribution features. So the feature selection starts by removing the lower score features.

The results were compared for: **Scenario 1:** evaluating the models with no feature exclusion; **Scenario 2:** evaluating the models excluding one feature with lower contribution; and **Scenario 3:** evaluating the models excluding two features with the lowest contributions. The results were be compared for each scenario.

2.5 Model Selection Metric

The fine-tuned parameters of each model were employed on the training steps. After the models training, we evaluated them by the coefficient of determination (also known as R-squared or R^2) metric. Chicco et al. [17] compared usual regression analysis metrics and discussed the advantages of the coefficient of determination as a more informative metric, defined as: let n be the number of samples, $\hat{y}_i$ be the ith predicted value and y_i the corresponding ith ground truth. Therefore, R^2 is given by:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{1}$$

where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the mean of the collected values. Chicco et al. [17] also explain that the coefficient of determination can be understood as the variance proportion predictable from the independent variables in the dependent variable. The best value is $R^2 = 1$ and the worst value is when R^2 tends to $-\infty$.

2.6 Model Evaluation Metrics

In order to evaluate the quality of the results performed by the chosen model on the test set, we employed the Mean Squared Error (MSE) and the Root Mean Squared Error (RMSE) as evaluation metrics. More details about these metrics can be obtained in Chicco et al. [17]. The R^2 score was employed as a model evaluation metric as well.

3 Results and Discussion

The models were implemented with the open-source package Scikit-learn (Pedregosa et al. [18]) provided for Python programming language. The first step we sort the feature importance score given by the RReliefF algorithm. Table 1 shows the importance rank for each of the six features. Lower ranks indicate lower feature importance.

Table 1. Features importance ranking.

Feature	Depth	Porosity	Longitude	Latitude	Region	Continental
Rank	6	5	4	3	2	1

After sorting the features by the RReliefF algorithm, we performed the grid search approach for the four models in the three different scenarios: **Scenario 1** using all six (6) features; **Scenario 2** using five (5) features, removing Continental feature; and **Scenario 3** using four (4) features, removing Continental and Region features. The optimized parameters obtained by the grid search approach under the three different scenarios are presented for DTR and Ridge models in Table 2 and for SVR and MLP models in Table 3.

Table 2. Fine-tuned parameters for DTR and Ridge models

DTR				**Ridge**			
	Scenario				Scenario		
Parameters	**1**	**2**	**3**	**Parameters**	**1**	**2**	**3**
max_depth	3	10	7	fit_intercept	False	False	False
max_features	2	4	3	alpha	4.0	4.0	6.5
min_samples_split	5	50	50	solver	‘saga’	‘sag’	‘sag’
min_samples_leaf	5	15	15	tol	0.01	0.01	0.01

Table 3. Fine-tuned parameters for SVR and MLP models.

SVR				**MLP**			
	Scenario				Scenario		
Parameters	**1**	**2**	**3**	Parameters	**1**	**2**	**3**
kernel	‘poly’	‘poly’	‘rbf’	hidden_layer_sizes	150	100	200
gamma	1.0	5.0	5.0	activation	‘relu’	‘relu’	‘relu’
coef0	5.0	5.0	0.0	solver	‘adam’	‘lbfgs’	‘adam’
degree	2	2	-	alpha	0.0005	0.0005	0.001
C	5.0	5.0	1.0	learning_rate	‘invscaling’	‘constant’	‘constant’
epsilon	0.025	0.025	0.025	learning_rate_init	0.01	0.001	0.001
				momentum	0.9	0.5	0.1
				nesterovs_momentum	True	False	False
				beta_1	0.9	0.82	0.82

After grid searching the parameters, the models were trained. Table 4 presents the R^2 score for the training and validation dataset considering the three scenarios previously described. The validation R^2 score is the mean of the five scores obtained by the standard 5-fold cross-validation.

The best metric values presented in Table 4 were performed by the DTR model. So it was chosen as the most appropriate permeability predictor.

Table 4. Train and validation R^2 score values for all models.

Model	Scenario	Train	Validation
DTR	1	0.1945	0.1945
	2	0.2429	0.1910
	3	0.2713	0.1968
SVR	1	0.0911	0.0647
	2	0.0402	0.0182
	3	0.1793	0.0874
MLP	1	0.1692	0.1026
	2	0.1575	0.1123
	3	0.1222	0.0960
Ridge	1	0.0912	0.1056
	2	0.0894	0.1060
	3	0.0880	0.1055

After model selection, we performed predictions by employing the test dataset to the trained DTR model. Table 5 presents the performance metric values performed by the model when using 6, 5 and 4 features, respectively.

Table 5. Performed metrics values for DTR model on test dataset.

Metric	**Scenario 1**	**Scenario 2**	**Scenario 3**
MSE	0.0073	0.0068	0.0066
RMSE	0.0852	0.0822	0.0811
R^2 score	-0.0238	0.0473	0.0732

The results presented in Table 5 have particular characteristics. Firstly, one can see the negative value for the R^2 metric for Scenario 1. Usually, this means that the Machine Learning model is potentially poor to model the target variable, as described in Pedregosa et al. [18]. However, the MSE and RMSE metrics are on the order of 0.0073 and 0.0852, which indicates plausible results. So, despite the negative value for R^2, the model is performing good predictions on the test dataset.

Secondly, the performance metric values improve as the number of features is decreased. The obtained results are better for Scenario 2 and even better for Scenario 3. It indicates that the feature selection approach employing the RReliefF algorithm is a good strategy. It decreases the model complexity and improves the results. So, Scenario 3 results in the highest positive value for the R^2 score.

The feature selection approach is crucial for the achievement of good estimations. Using the most important predictors variables as described in Scenario 3, we obtain the MSE and the RMSE metrics for the test set on the order of 0.0066 and 0.0811, respectively. These results are supported by literature according to Male and Duncan [19] since one of the most important physical properties for permeability prediction is porosity, which also depends on latitude, longitude, and depth. Then, as a low-cost predictor for reservoir characterizations, these features can perform reasonable estimations for permeability.

4 Conclusions

In this work we investigate the data quality in reservoir characterizations by modeling the horizontal permeability using Machine Learning regression models. The hyper-parameters were optimized using the grid search approach and the features were selected by the RReliefF algorithm. The best permeability predictions were performed by the Decision Tree Regressor model, which performed MSE and RMSE on the order of 0.0066 and 0.0811, respectively, for the test set. We observe that the results were improved after selecting the most important features. Since our results are very promising, we can conclude that the experimentally gathered data and petrophysical logs can be used with Machine Learning models to provide low-cost estimators for reservoir characterizations.

However, the limitation of our study is the data depreciation, since our database is from 1990. The geological characteristics of the reservoirs can be different nowadays. So, in order to propose more reliable results, we intend to investigate the quality of our model predictions on more recent databases as future works.

Acknowledgements. This work was held with financial support of Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES).

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] M. A. Ahmadi and Z. Chen. Comparison of machine learning methods for estimating permeability and porosity of oil reservoirs via petro-physical logs. *Petroleum*, vol. 5, pp. 271–284, 2019.
[2] A. Chehrazi and R. Rezaee. A systematic method for permeability prediction, a petrofacies approach. *Journal of Petroleum Science and Engineering*, vol. 82, pp. 1–16, 2012.
[3] L. Chen, C. Ren, L. Li, Y. Wang, B. Zhang, Z. Wang, and L. Li. A comparative assessment of geostatistical, machine learning, and hybrid approaches for mapping topsoil organic carbon content. *International Journal Geo-Infomation*, vol. 8, n. 4, 2019.
[4] V. Gudivada, A. Apon, and J. Ding. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software*, vol. 10, n. 1&2, pp. 1–20, 2017.
[5] Australia. Bureau of mineral resources, geology and geophysics (1990). program summary, 1990.
[6] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis*. Wiley, New Jersey, USA, 5th ed edition, 2012.
[7] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. The Wadsworth & Brooks/Cole statistics/probability series, Monterey, CA, USA, 1984.
[8] M. Awad and R. Khanna. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Apress Open, Berkeley, CA, USA, 1st edition, 2015.
[9] E. Alpaydin. *Introduction to Machine Learning*. The MIT Press, Massachusetts, USA, 2th ed edition, 2014.
[10] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer Texts in Statistics, Stanford, CA, USA, second edition, 2013.
[11] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, Stanford, CA, USA, second edition, 2017.
[12] C. Cortes and V. N. Vapnik. Support-vector networks. *Machine Learning*, vol. 20, n. 3, pp. 273–297, 1995.
[13] A. J. Smola and B. Scholkopf. A tutorial on support vector regression. *Statistics and Computing*, vol. 14, pp. 199–222, 2004.
[14] S. Abirami and P. Chitra. Chapter fourteen - energy-efficient edge based real-time healthcare support system. In P. Raj and P. Evangeline, eds, *The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases*, volume 117 of *Advances in Computers*, pp. 339–368. Elsevier, 2020.
[15] M. Robnik-Šikonja and I. Kononenko. An adaptation of relief for attribute estimation in regression. *Machine Learning: Proceedings of the Fourteenth International Conference*, vol. , pp. 296–304, 1997.
[16] M. Robnik-Šikonja and I. Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, vol. 53, pp. 23–69, 2003.
[17] D. Chicco, M. Warrens, and G. Jurman. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, vol. 7:e623, 2021.
[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. P., R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[19] F. Male and I. J. Duncan. Lessons for machine learning from the analysis of porosity-permeability transforms for carbonate reservoirs. *Journal of Petroleum Science and Engineering*, vol. 187, pp. 11, 2019.

Appendix

This work employed the following scikit learn models:

- sklearn.svm.SVR() [1]
- sklearn.neural_network.MLPRegressor() [2]
- sklearn.linear_model.Ridge() [3]
- sklearn.tree.DecisionTreeRegressor() [4]

Tables 6 and 7 present the parameters values evaluated on grid search approach for DecisionTreeRegressor and Ridge models and for SVR and MLPRegressor models, respectively. More details about the parameters can be obtained in Pedregosa et al. [18].

Table 6. Grid Search parameters for DTR and Ridge models

DTR		Ridge	
Parameters	Values	Parameters	Values
max_depth	2, 3, 4, 5, 6, 7, 8, 9, 10	fit_intercept	False, True
max_features	2, 3, 4, 5, 6	alpha	0.1, 1.0, 4.0, 5.0, 6.5, 10.0, 25.0, 50.0
min_samples_split	5, 6, 8, 10, 15, 20, 25, 30, 40, 50	solver	'svd', 'lsqr', 'sparse_cg', 'sag', 'saga'
min_samples_leaf	2, 5, 8, 10, 12, 15, 20	tol	0.01, 0.001, 0.0001

Table 7. Grid Search parameters for SVR and MLP models.

SVR		MLP	
Parameters	Values	Parameters	Values
kernel	'linear', 'rbf', 'sigmoid', 'poly'	hidden_layer_sizes	100, 150, 200
gamma	0.1, 0.5, 1.0, 5.0	activation	identity, logistic, tanh, relu
coef0	0.0, 0.1, 0.5, 1.0, 5.0	solver	lbfgs, sgd, adam
degree	2, 3	alpha	0.0001, 0.0005, 0.001
C	0.05, 0.1, 0.5, 1.0, 5.0	learning_rate	constant, invscaling, adaptive
epsilon	0.025, 0.05, 0.1, 0.5, 1.0, 5.0	learning_rate_init	0.0001, 0.001, 0.01
		momentum	0.1, 0.5, 0.9
		nesterovs_momentum	True, False
		beta_1	0.82, 0.9

[1] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
[2] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html
[3] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html
[4] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html