Machine learning as a tool to predict the mass of oil from well logs

Marcelo Guarido and Daniel Trad

ABSTRACT

Oil saturation is the measure of the amount of oil inside the porosity of a reservoir rock. Its calculation, usually from core analysis, is an important quantity that helps to characterize the reservoir. In this work, we are not predicting the actual oil saturation due to the lack of information for the wells gathered, but the fraction of mass of oil in the core. Most of this report is focused on the data preparation prior to modeling, as our variables and targets came from two different measurement sources (well logs and core analysis), and in how to create a valid workflow to make features and targets compatible to each other. In the end, we show how to select an appropriate machine learning model to predict the target, which need to be one with non-linear properties, and how to interpret the feature importance. To predict the fraction of mass of oil, the induction log ILD is the one that brings most of the information, but it needs to be combined with other logs for the prediction to make sense. The metric used to evaluate the models was the R^2, and the best model had a score of 0.82.

INTRODUCTION

Knowledge of oil and water saturation distribution in a oilfield is of great significance for the latter development (Yue et al. 2018), and the estimation can be done by direct petrophysical models (with fair knowledge of the relations between logs and saturation) or, more recently, with the use of machine learning models (Zhang et al. 2019). Here we will focus on the machine learning side of it.

Machine learning algorithms usage on geoscience, engineering, and petrophysics data is in ascension this last decade. Maybe the most common application is for facies classification, by the use of ensemble classifiers (Bestagini, Lipari, and Tubaro 2017; Zhang and Zhan 2017; Caté et al. 2017), neural networks (Silva et al. 2014), and support-vector machines (Caté et al. 2017; Alexsandro, P. Carlos, and Geraldo 2017; Wrona et al. 2018). Guarido (2019) showed that deeply analyzing the data before modeling can provide insights for data engineering, and he applied polar coordinates transformation of the features by realizing circular relationship with the target. Machine learning has broad applications in geophysics, such as in FWI, by using convolutional neural networks for salt identification (Lewis and Vigh 2017; Guarido, Li, and Cova 2018), and FLEXWIN for time-window selection (Chen et al. 2017). It is possible to find works on trace interpolation using support-vector regression (Jia and Ma 2017), or by Monte-Carlo approximations (Jia, Yu, and Ma 2018). Deep neural networks are used by Araya-Polo et al. (2017) for fault detection and by Araya-Polo et al. (2018) for tomography. Nearest neighbors (k-NN) can be implemented to help on CMP velocity analysis (Smith 2017). Russell, Ross, and Lines (2002) combine NN with AVO. In petrophysics, Ahmadi and Chen (2019) shows that hybrid methods provides higher accuracy than single models, but the latter are more robust. Many others machine learning algorithms applications can be found in the literature.

Oil and water saturation also have a fairly long list of machine learning applications to help in interpretation, but we will narrow the list a few. Zhang et al. (2019) uses RNN (recurrent neural networks), most specifically the LSTM (long-short term memory) approach to estimate water saturation with good performance. Khan, Tariq, and Abdulraheem (2018) also goes on the deep learning side by training a neural networks model to predict water saturation in complex lithologies with high accuracy. Kapoor (2017) shows that ensemble tree methods, such as random forest and gradient boosting, can predict oil and water saturation in the oil sands more accurately than empirical estimations.

In this paper, we were not able to work on the estimation of oil saturation in the oil field, as this information is missing on public data (even with the header pointing that it exists). But we were able to retrieve the fraction of mass of oil presented in the core analysis for several wells, and the goal became to estimate it from the well logs. This paper is not focused on the machine learning algorithms used for the predictions, but on how a *data science* project proceeds. Usually statisticians and data scientists expend from 80% to 90% of the time preparing the data for the modeling, and here we want to explore this need. In general, a data science project follows a work flow that starts with the *data description*, where each feature and target must be deeply understood, followed by the *data preparation*, when the data is cleaned, filtered, processed, and prepared for analysis and modeling. Later, the *data analysis* step is very import to understand, from a series of different plottings, any relationship between features and target. The *modeling* comes as one of the final parts, were predictions and model selection are evaluated. Finally, we need to go for the *interpretation* of the model and predictions, trying to understand the physics behind the outputs and using our field knowledge to determine if the model makes sense.

DATA DESCRIPTION

A *data science* project consists on a series of steps that usually follow the same order. For the project to be successful, it is important for the data scientist to have deep understanding of the data before preparation and modeling. Knowledge is gathered from literature and papers, as well as, in the case of this paper, from acquisition reports. Without this kind of understanding, the data scientist may not be able to identify issues in the data.

Table 1: Features and target descriptions.

Name	Description
WELL	Well name
DEPTH	Measured depth (m)
RHOB	Bulk density
DPHI	Density porosity
NPHI	Neutron porosity
SP	Spontaneous potential
GR	Gamma ray
DT	Sonic interval transit time
ILD	Deep induction resistivity
PE	Photoelectric effect
MASS OIL	Fraction of mass of oil in the core
POROSITY	Core analysis porosity

The data used for this report is from the *Athabasca* oil field (Figure 1) and, initially,

well logs and core analysis from *548* wells were provided (table 1).

Figure 1: Map of the Athabasca oil field. Source: Wikipedia

Well logs and core analysis for the 548 wells were acquired using the *GeoScout* software, and all the information used is from public data (meaning this work is reproducible). The well logs, for each well, have a `.las` file containing all the information required. However, most of the wells do not have all the wireline measurements requested, even if they exist in the `.las` file header. Less than 10% of the files actually contain the wireline logs in an desirable quantity (missing is $< 25\%$). Working with public data can be challenging, as companies still keep some of the data private, as intellectual property, and also the public data can be incomplete and/or have several issues. This is a very common challenge for data scientists.

Initially, the main objective of this paper was to estimate the *oil saturation* of the well, by matching the measured ones during core analysis. However, the downloaded data comes with this information empty for **all** the wells. The only information about oil content in the core is the *fraction of mass of oil*, and the objective of this work is now to predict it.

DATA PREPARATION

After gaining all the necessary knowledge about the input dataset the data scientist needs to proceed to the *data preparation*, analysis, and modeling.

Working with data for the same well from two different sources (wireline logs and core analysis) brings challenges in the data preparation. The first one is the sample interval of

the measurements. The well logs have a depth sample interval of around 15.25cm (with small variations for different wells), as the core analysis is done in a very irregular sample interval, varying from 0.1m to 3.9m in the dataset. This requires the data (logs and cores) to be resampled to the same sample interval. For convenience, all the data was resampled to 15cm. But we had to keep in mind that the core analysis was not done for the same length as the well logs. Mostly of the `MASS OIL` column has empty entries (that could mean zero or just not measured, and not sure to be zero). So, for the well logs, the resampling was done with a linear interpolation, as for the *fraction of mass of oil*, the interpolation was done to fill the length of the core sample with the same value, and making it zero when the entry is blank. That is not a big issue, as we will only use depth where the mass of oil is measured to do the modeling.

The second challenge is to address the problem that the cores samples and well logs were not acquired in the same "depth domain", i.e., that the depth references are different. A way to get around this issue is to select a compatible measurement from both datastes and compute the difference in shift between them. We made this comparison by using `DPHI` (density porosity) from the well logs, and the *porosity* measured in the core. Remember that we are assuming that both measurements are compatible (which is an acceptable assumption). This kind of analysis also helps us to QC the data acquired. Sometimes a `WELL NAME` contains well logs and core analysis from different wells, probably caused by human error when uploading the data to the database. When it happens, `DPHI` and core porosity have no visual match, and were removed from our dataset.

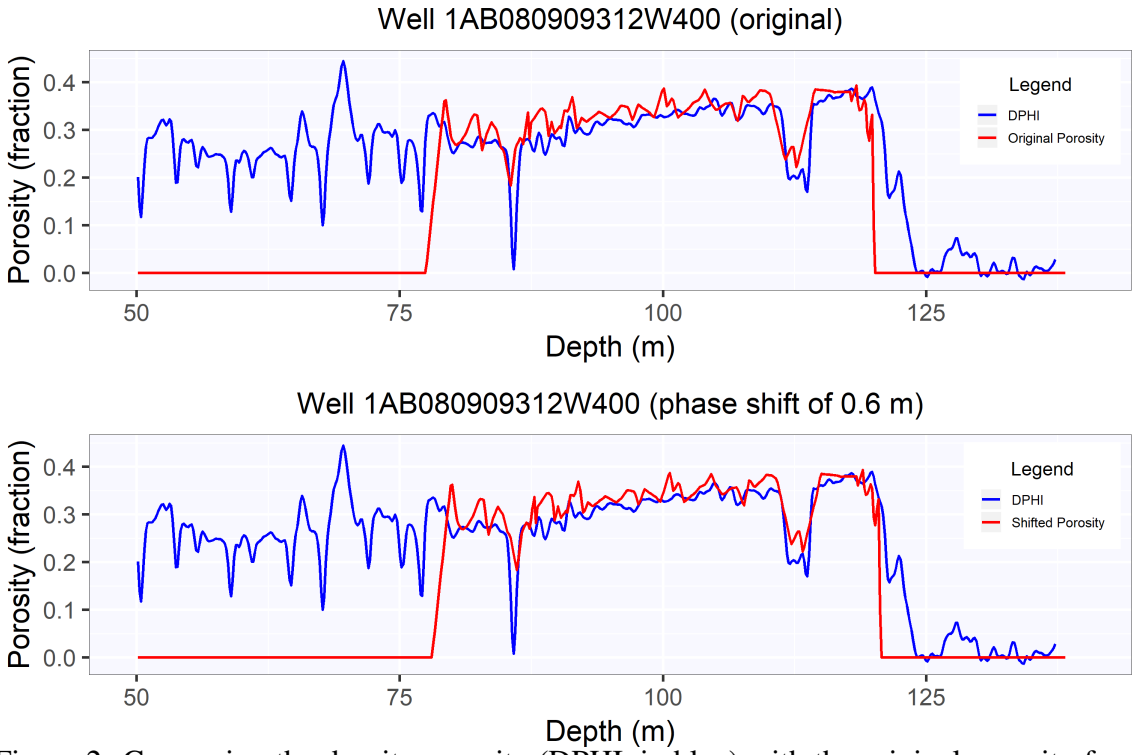

Figure 2: Comparing the density porosity (DPHI, in blue) with the original porosity from cores (top, in red) and with the shifted core porosity (bottom, in red). The shift applied was of 0.6m.

Figure 2 shows, on the top, the curves of the density porosity (DPHI, in blue) overlapped by the core porosity (in red) measured for the well. The 0's in the core porosity curve mean no data. At least in the region where both curves were measured, it is possible to note they follow a similar trend, mostly on lower frequency events (noticeable around the depth of 110m) but slightly shifted from each other. This means that the curves have different depth reference, and the measures of the fraction of mass of oil will be shifted relative to the logs, and the predictions can be strongly damaged by it.We estimate the distance between the curves and apply a phase shift based on a running window over the curves, computing the distance of the segments for each window (by minimizing the squared difference), and selecting the median of all calculated distances (in number of points) to be the used for the phase shift of the whole well. In the bottom plot of the Figure 2 is the result of the shifted core porosity applied. The calculated distance was of 0.6m. The same phase shift is applied on the measures of mass of oil. This way, both data are on the same depth reference.

With all the filtering applied as listed above, the number of useful wells was reduced from 548 to only 50 at this point. This shows how hard is to find reliable data from public databases (a common problem in any study field). Wireline logs are of high cost, as drilling or well production must be stopped during the measurements for several days.

DATA ANALYSIS

Now that the data preparation is complete, the *data analysis* is the next step in a data science project. It is the moment to plot the data and look for any insights you may have before modeling. Most of the times, you can recognize important expected patterns and identify if predictions make sense.

Figure 3 has one of the wells curves and mass of oil (fraction) presented in a single plot. This well is a good example to start some visual analysis. The mass of oil for this well varies from values close to 0 to 0.15, and it is possible to observe that those variations follow some trends of several well logs. The *photoelectric effect* (PE) tends to assume higher values when the mass of oil is smaller, and the trend is followed to a higher frequency content (thinner spikes). The *gamma-ray* (GR) shows to follow a lower frequency similarity to the mass of oil, and higher gamma-ray measurements represents low mass of oil. Similar analysis is possible from the *spontaneous potential* (SP), but not as accurate as GR, and the trend tends to invert as the well goes deeper. *Neutron porosity* (NPHI) is apparently pointing locations with high variance of the mass of oil. *Density porosity* (DPHI) shows to be following the mass of oil behavior quite closely, but not as accurate on higher frequencies. It is important to notice that the *bulk density* (RHOB) and the *density porosity* (DPHI) are basically a mirror of each other, which makes sense, as the density porosity is extracted from the bulk density values (Tittman and Wahl 1965). As those two logs are redundant information, only DPHI will be used for modeling. With so many logs following different trends of the mass of oil, we expect that a regression model will do a fairly good job to predict the mass of oil.

The distribution of values of the fractions of mass of oil is shown in Figure 4, where the `x-axis` is the binned mass of oil, and the `y-axis` is the counting for each bin. It looks to be a *bimodal distribution*, as it contains two modes (two peaks). The first peak is

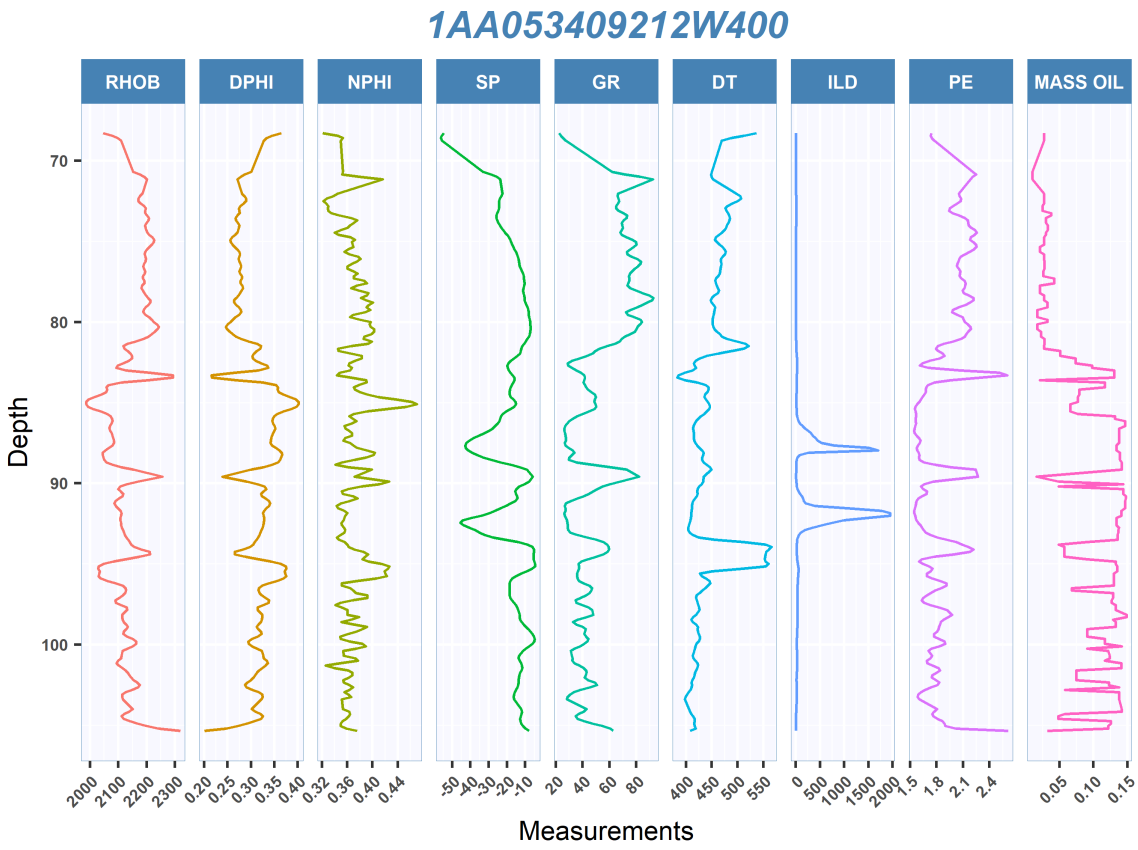

Figure 3: Example of wireline logs and mass of oil (fraction) from one of the provided wells.

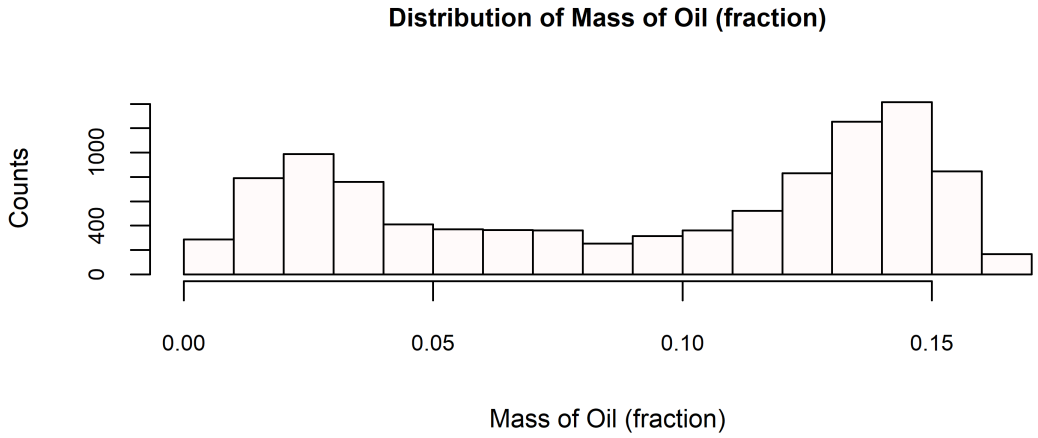

Figure 4: Distribution of mass of oil (fraction of core total mass) over all the wells.

around 0.03, which represents a fairly low amount of oil, and may be measurements on the limits of the reservoir , and the second peak is around 0.15, which might be the zone of the reservoir. The bimodal behavior does not look connected to the porosity of the reservoir, as its distribution (from DPHI, in Figure 5) has a single mode around 0.3. This means that in locations where the mass of oil is lower, the mass of gas and/or water may be higher.

Porosity alone will not be a perfect indication of mass of oil, but we expect it to be one of the most important features in the predictions.

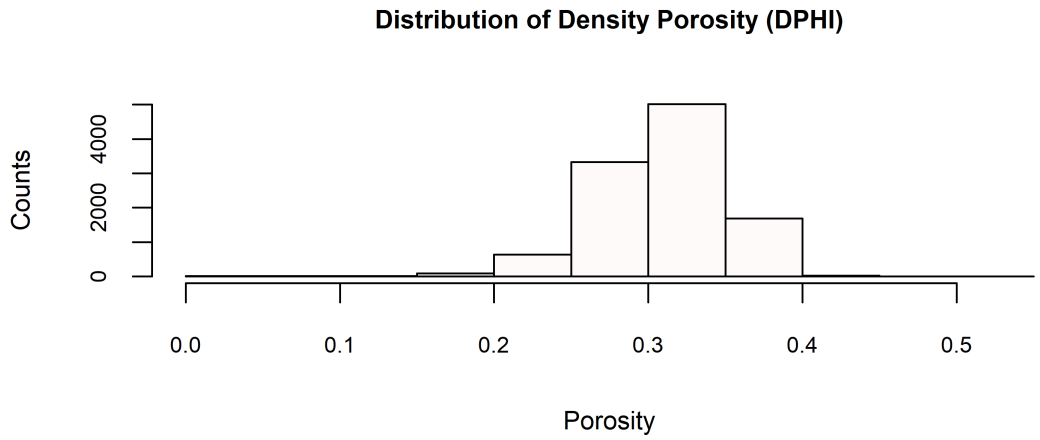

Figure 5: Distribution of density porosity (DPHI) over all the wells.

A simple way to evaluate any correlation between features (well logs) and target (fraction of mass of oil) is by visualizing them within the *pairs plot* panel. This panel consists in three types of plots: the lower left part are the scatter plots between all the variables and target provided; the upper right portion has the density plots of the scatter plots from the lower part; and the on diagonal sits the distribution of each feature and target. However, as pointed by the distribution of mass of oil (Figure 4), there are two modes (two distributions) combined. A single correlation for two distributions will not look simple to evaluate, so we split the pairs plot for places where the mass of oil (frac) is smaller than 0.08 (Figure 6) and for places where the target is greater or equal 0.08 (Figure 7). Let's interpret the pairs plot from Figures 6 and 7. The lower part of the pairs plot are simply the scatter plot (in blue) of the pairs of features and targets, the diagonal is the distribution of each feature and the target, as the upper part is the density plot from the scatter plots.

Even with the split plots relative to the fraction of mass of oil, the pairs plots have very similar trends. DPHI, GR, DT, ILD, and PE look to have, at least, some minor linear correlation with the mass of oil. Also, DPHI values tend to be higher when the mass of oil is larger. Those are very positive observations, and now we need to check if a machine learning model can translate it to precise predictions.

MODELING

Now that we have a good understanding of our dataset, and that we fixed and aligned the data the best we could, the next step in our data science project is the *modeling*.

Any supervised learning model is trained on a subset of the dataset (usually 80% of the data) and the model is evaluated on the rest of the data (the remaining 20%). As we have 50 wells, we will split 40 for the *training set*, and 10 for the *testing set*. We will not use any criteria to select the wells (random split).

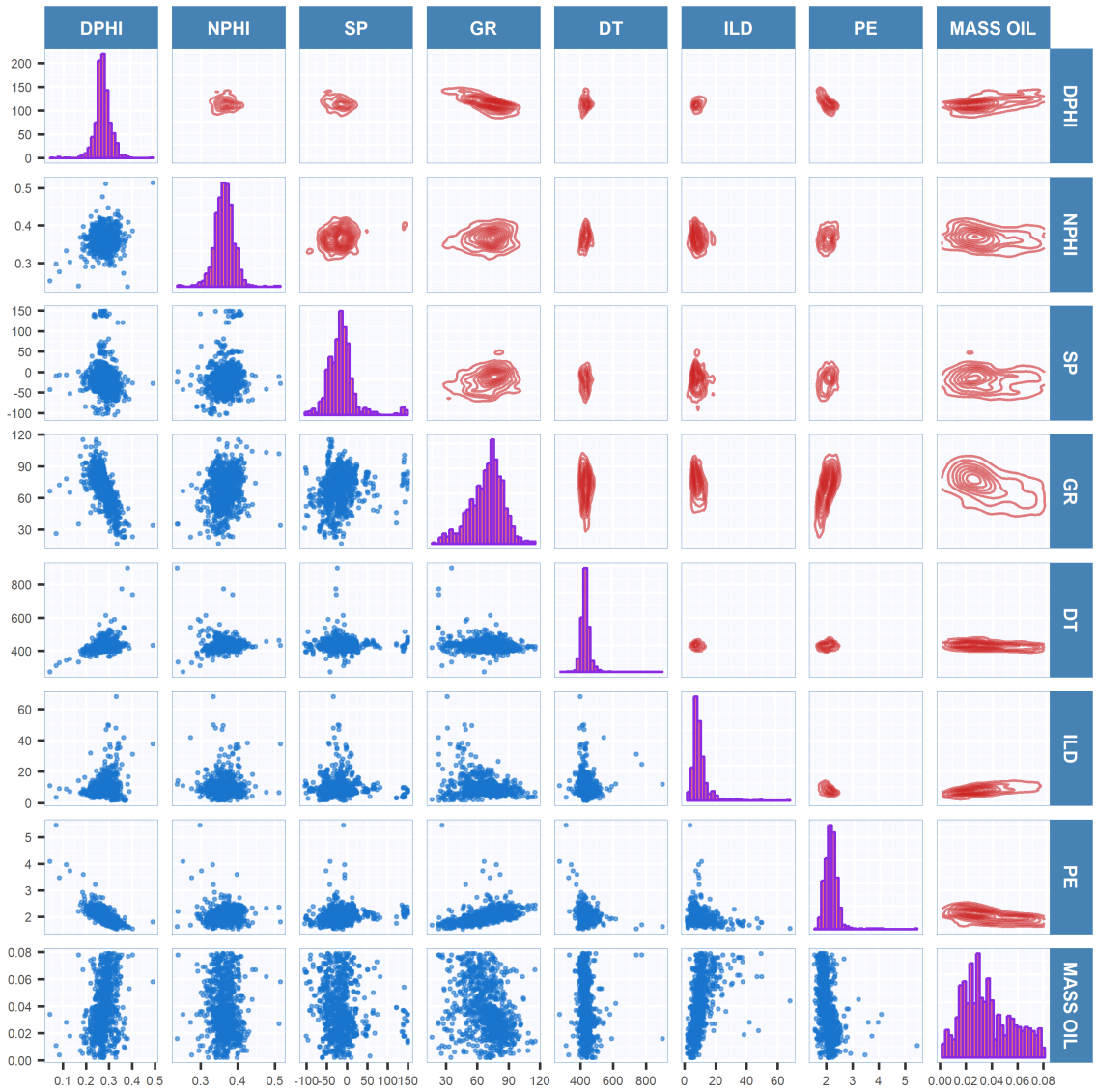

Figure 6: Pairs plot of all features and target, considering only the fraction of mass of oil smaller than 0.08.

It was observed before that some of the well logs present some level of linear correlation with the mass of oil. This brings us an important question: can a simple linear regression with multiple features be a good predictor of mass of oil?

Figure 8 shows the predictions using a linear regression model. On the left is shown the correlation between true data and predictions, with a R^2 of 0.52. On the right are two examples of true mass of oil (in blue) and predictions (in red) in specific wells. A perfect R^2 correlation value is supposed to be 1, as 0 means no correlations between target and predictions, and negative values means that the model predictions are worse than a constant line. For this model, the R^2 is arguably acceptable (depending on your goals). For this work, we consider it as acceptable if your goal is to predict major trends of mass of oil in your data. In the two example wells, the predictions fairly achieve this goal, but they would not be acceptable as a valid interpretation of the well. In other words, the linear

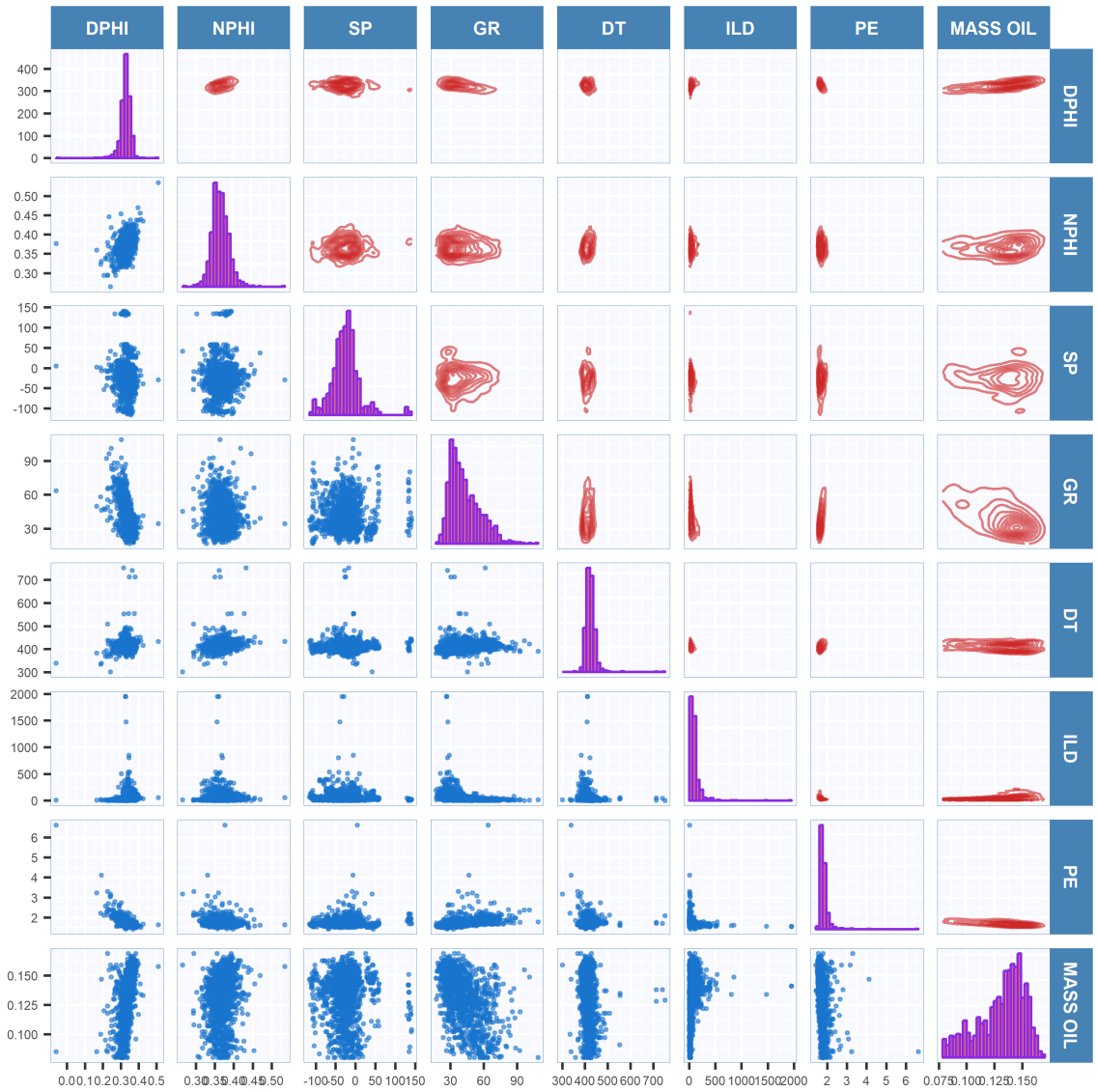

Figure 7: Pairs plot of all features and target, considering only the fraction of mass of oil greater or equal than 0.08.

regression is not suitable for precise predictions of fractions of mass of oil (our goal here).

At this point, with the partial failure of the linear regression to make predictions, we can assume that the correlation between features and target is non-linear. With this information in mind, we have two options to help us to improve the predictions: 1) apply a polynomial feature engineering on the features or 2) select a model that has non-linear properties. In this work, we will be going through the second option.

The *gradient boosting regressor* (Guarido 2018) is a model with non-linear properties, and sounds to be a better option in this case, and Figure 9 shows its predictions. Looking at the two example wells, the predictions did a great job, when compared to the linear regression in Figure 8, and the predicted mass of oil (red) matches the true values (blue) closer. The correlation between predictions and target also have a better trend, dispersing

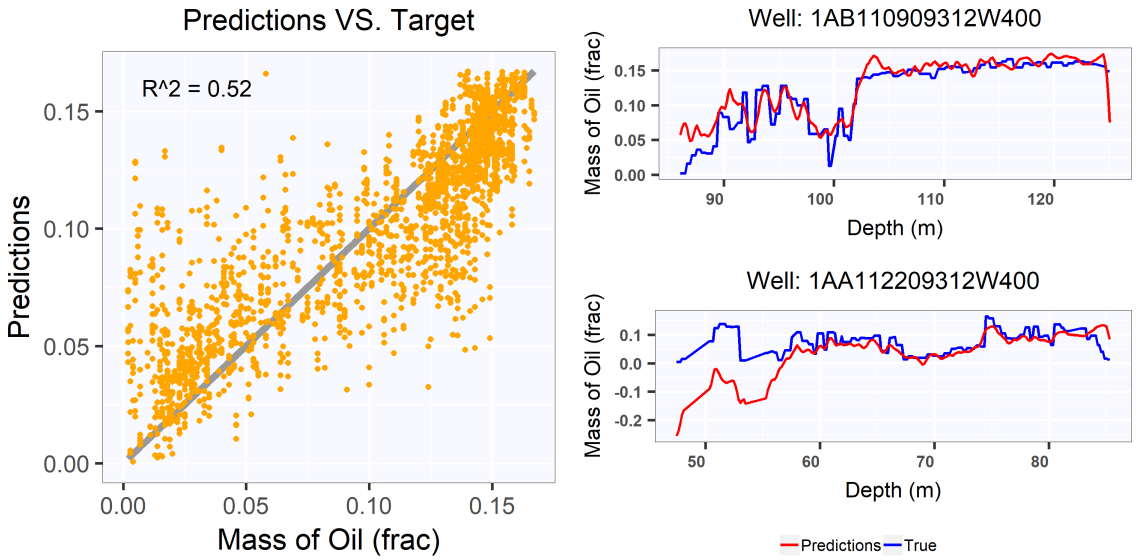

Figure 8: Predictions using a linear regression model. On the left is shown the crossplot between true data and predictions, with a R^2 of 0.52. On the right are two examples of true mass of oil (in blue) and predictions (in red) in specific wells.

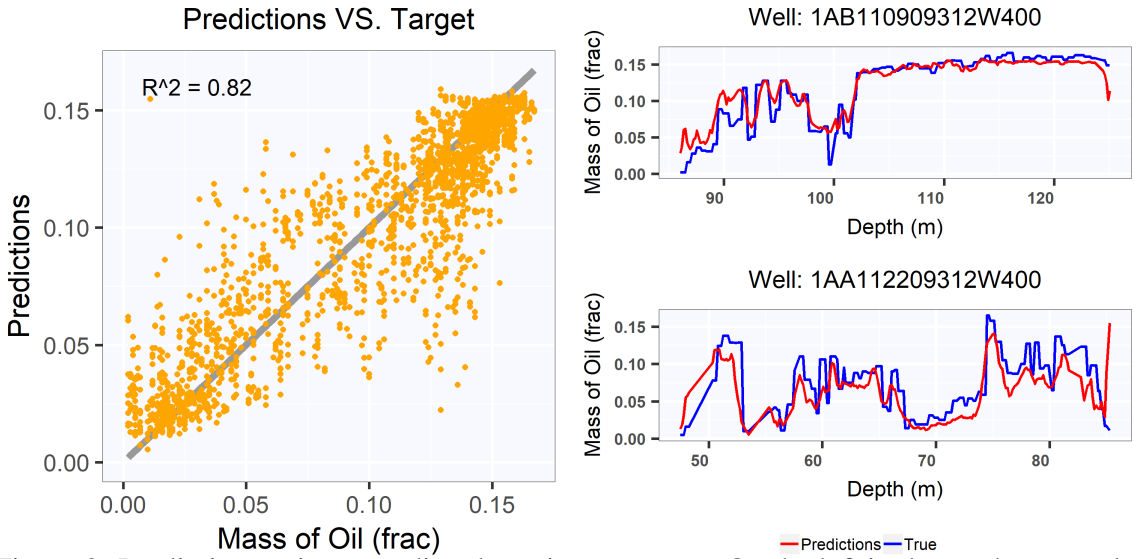

Figure 9: Predictions using a gradient boosting regressor. On the left is shown the crossplot between true data and predictions, with a R^2 of 0.82. On the right are two examples of true mass of oil (in blue) and predictions (in red) in specific wells.

around the "perfect line" (gray), and now the R^2 is 0.82. It is a huge improvement of the predictions, and that is thanks to the non-linear properties of the selected model.

There are other models with non-linear properties that could be used for the predictions. The most famous one is the *neural networks*, where activation functions in the neurons create non-linear criteria. However, a neural networks model contains weights that are hard (if not impossible) to interpret, meaning it is not straight forward to understand the importance that each feature had for the predictions. As we want to use interpretable models, we will keep with the gradient boosting for the predictions.

INTERPRETATION

The final step for our data science project is to interpret the trained model. Remember that we did a pre-evaluation of the wells logs and how important would they be to predict the fraction of mass of oil, by analyzing the pairs plots from Figures 6 and 7 and checking linear correlations with the target, even if minor linear ones.

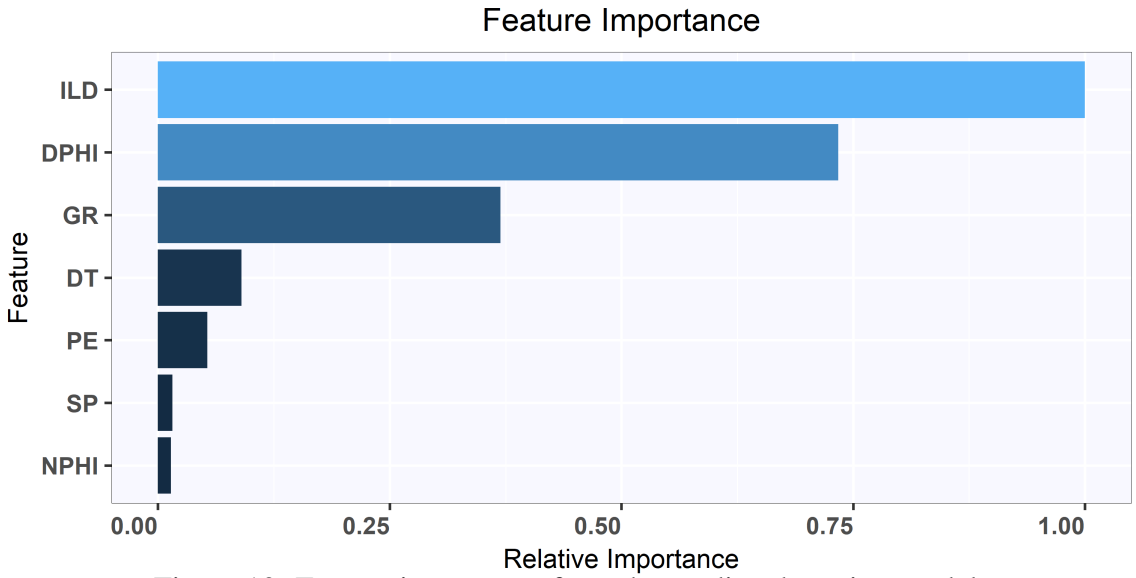

Figure 10: Feature importance from the gradient boosting model.

Now we can evaluate the features importance from the gradient boosting model in Figure 10. Our predictions before sound to be reasonably correct, as we pointed DPHI, GR, DT, ILD, and PE as features with at least minor linear correlation with the mass of oil, and they are the top 5 important features. Some surprises still raised: we were expecting that DPHI (porosity measure) would be the top position in importance, but it was surpassed by ILD (induction log). ILD measures the formation resistivity, and its combination with DPHI and GR as major players, and DT and PE as fine tunes in the model, we could train a model with high confidence predictions. SP and NPHI shows to include almost no information to the model, and training a model without these two logs gave us the predictions as before.

CONCLUSIONS

Oil saturation is an important parameter in the reservoir characterization, and in this paper we showed how it can be calculated with high precision. Well, actually, we showed how to calculate the fraction of mass of oil in the core, not the oil saturation, and the reason for that is the lack of information on public core analysis data.

We presented the work in a way to mirror a data science research project, that starts from the data understanding, to the final machine learning modeling and interpretation. We showed that the public data can be mostly unusable depending of the goal of a project. Here, we started with a data set containing 548 wells and after data cleaning and quality control, the number of wells was reduced to 50. One important procedure pre-modeling was to match in depth information that comes from different sources, in this cases the well logs

with the core analysis. We used variables from both data sets that have compatible measure goals (DPHI and core porosity) to estimate and fix the depth shift between the curves. After the data was clean and processed, data analysis came in place as a quality control measure, insights of features importance and redundancy (the reason we removed RHOB from the modeling), and in a way to understand the data so deeply that we could "predict" our predictions. It was possible to evaluate if the correlation of the well logs with the mass of oil is linear by testing model without and with non-linear properties, and the latter proved to be more appropriate. We use a gradient boosting model for the predictions, and the model showed to be robust and highly accurate with the estimations, with a measured $R^2 = 0.82$. We wrapped up the project by interpreting the model and its features importance. The induction log (IDL) showed that the measurement of the formation resistivity is the most valuable to estimate the mass of oil, followed by the measure of porosity (DPHI) and gamma-ray (GR). Seismic time measurements (DT) and photoelectric effect (PE) were used by the model as refinement parameters.

In the end, we showed that a data science research project procedure is very broad and can be applied to the Oil & Gas industry as a support tool for interpretation.

ACKNOWLEDGMENTS

The authors thank the sponsors of CREWES for continued support. This work was funded by CREWES industrial sponsors and NSERC (Natural Science and Engineering Research Council of Canada) through the grant CRDPJ 461179-13. We also thank GLJ Petroleum Consultants, specially Bill Spackman and Michael Morgan, for technical support and data acquisition. Finally, we thank Soane Mota dos Santos for all the knowledge share during very useful conversations.

REFERENCES

Ahmadi, Mohammad Ali, and Zhangxing Chen. 2019. "Comparison of Machine Learning Methods for Estimating Permeability and Porosity of Oil Reservoirs via Petro-Physical Logs." *Petroleum* 5 (3): 271–84. https://doi.org/10.1016/j.petlm.2018.06.002.

Alexsandro, G. C., A. C. da P. Carlos, and G. N. Geraldo. 2017. "Facies Classification in Well Logs of the Namorado Oilfield Using Support Vector Machine Algorithm." In, 1853–8. 15th International Congress of the Brazilian Geophysical Society & EXPOGEF, Rio de Janeiro, Brazil, 31 July-3 August 2017. https://doi.org/10.1190/sbgf2017-365.

Araya-Polo, Mauricio, Taylor Dahlke, Charlie Frogner, Chiyuan Zhang, Tomaso Poggio, and Detlef Hohl. 2017. "Automated Fault Detection Without Seismic Processing." *The Leading Edge* 36 (3): 208–14. https://doi.org/10.1190/tle36030208.1.

Araya-Polo, Mauricio, Joseph Jennings, Amir Adler, and Taylor Dahlke. 2018. "Deep-Learning Tomography." *The Leading Edge* 37 (1): 58–66. https://doi.org/10.1190/tle37010058.1.

Bestagini, Paolo, Vincenzo Lipari, and Stefano Tubaro. 2017. "A Machine Learning Approach to Facies Classification Using Well Logs." In, 2137–42. SEG Technical Program Expanded Abstracts 2017. https://doi.org/10.1190/segam2017-17729805.1.

Caté, Antoine, Lorenzo Perozzi, Erwan Gloaguen, and Martin Blouin. 2017. "Machine Learning as a Tool for Geologists." *The Leading Edge* 36 (3): 215–19. https://doi.org/10.1190/tle36030215.1.

Chen, Yangkang, Judith Hill, Wenjie Lei, Matthieu Lefebvre, Jeroen Tromp, Ebru Bozdag, and Dimitri Komatitsch. 2017. "Automated Time-Window Selection Based on Machine Learning for Full-Waveform Inversion." In, 1604–9. SEG Technical Program Expanded Abstracts 2017. https://doi.org/10.1190/segam2017-17734162.1.

Guarido, Marcelo. 2018. "Machine Learning in Geoscience: Facies Classification with Features Engineering, Clustering, and Gradient Boosting Trees." *CREWES Research Report* 30: 13.1–13.23.

———. 2019. "Machine Learning Strategies to Perform Facies Classification." *GeoConvention 2019 Abstracts*.

Guarido, Marcelo, Junxiao Li, and Raúl Cova. 2018. "Machine Learning in Geoscience: Using Deep Learning to Solve the Tgs Salt Identification Challenge." *CREWES Research Report* 30: 14.1–14.12.

Jia, Yongna, and Jianwei Ma. 2017. "What Can Machine Learning Do for Seismic Data Processing? An Interpolation Application." *Geophysics* 82 (3): V163–V177. https://doi.org/10.1190/geo2016-0300.1.

Jia, Yongna, Siwei Yu, and Jianwei Ma. 2018. "Intelligent Interpolation by Monte Carlo Machine Learning." *Geophysics* 83 (2): V83–V97. https://doi.org/10.1190/geo2017-0294.1.

Kapoor, Gagan. 2017. "Estimating Pore Fluid Saturation in an Oil Sands Reservoir Using Ensemble Tree Machine Learning Algorithms." *Saint Mary's University*.

Khan, Mohammad Rasheed, Zeeshan Tariq, and Abdulazeez Abdulraheem. 2018. "Machine Learning Derived Correlation to Determine Water Saturation in Complex Lithologies." *Society of Petroleum Engineers*. https://doi.org/doi:10.2118/192307-MS.

Lewis, Winston, and Denes Vigh. 2017. "Deep Learning Prior Models from Seismic Images for Full-Waveform Inversion." In, 1512–7. SEG Technical Program Expanded Abstracts 2017. https://doi.org/10.1190/segam2017-17627643.1.

Russell, Brian, Christopher Ross, and Larry Lines. 2002. "Neural Networks and Avo." *The Leading Edge* 21 (3): 268–314. https://doi.org/10.1190/1.1885507.

Silva, Adrielle, Irineu Lima Neto, Abel Carrasquilla, Roseane Misságia, Marco Ceia, and Nathaly Archilha. 2014. "Neural Network Computing for Lithology Prediction of Carbonate- Siliciclastic Rocks Using Elastic, Mineralogical and Petrographic Properties." In, 1055–8. 13th International Congress of the Brazilian Geophysical Society; EXPOGEF, Rio de Janeiro, Brazil, 26-29 August 2013. https://doi.org/10.1190/sbgf2013-218.

Smith, Kenneth. 2017. "Machine Learning Assisted Velocity Autopicking." In, 5686–90. SEG Technical Program Expanded Abstracts 2017. https://doi.org/10.1190/segam2017-17684719.1.

Tittman, J., and J. S. Wahl. 1965. "The Physical Foundations of Formation Density Logging (Gamma-Gamma)." *Geophysics* 30 (2): 284–94.

Wrona, Thilo, Indranil Pan, Robert L. Gawthorpe, and Haakon Fossen. 2018. "Seismic Facies Analysis Using Machine Learning." *Geophysics* 83 (5): O83–O95. https://doi.org/10.1190/geo2017-0595.1.

Yue, Ming, Weiyao Zhu, Hongyan Han, Hongqing Song, Yunqian Long, and Yu Lou. 2018. "Experimental Research on Remaining Oil Distribution and Recovery Performances After Nano-Micron Polymer Particles Injection by Direct Visualization." *Fuel* 212: 506–14.

Zhang, Licheng, and Cheng Zhan. 2017. "Machine Learning in Rock Facies Classification: An Application of Xgboost." In, 1371–4. International Geophysical Conference, Qingdao, China, 17-20 April 2017. https://doi.org/10.1190/IGC2017-351.

Zhang, Qitao, Chenji Wei, Yuhe Wang, Shuyi Du, Yuanchun Zhou, and Hongqing Song. 2019. "Potential for Prediction of Water Saturation Distribution in Reservoirs Utilizing Machine Learning Methods." *Energies* 12: 3597.