DOI:10.1190/segam2020-3428275.1

Machine Learning model interpretability using SHAP values: application to a seismic facies classification task

David Lubo-Robles[1], Deepak Devegowda[1], Vikram Jayaram[2], Heather Bedle[1], Kurt J. Marfurt[1], and Matthew J. Pranter[1]

[1]The University of Oklahoma; [2]Pioneer Natural Resources Company

Summary

Machine learning models have been widely used by geoscientists to accelerate their interpretation and highlight hidden patterns in their data. However, as the complexity of the model increases, the interpretation of the results can become quite challenging. The SHAP technique provides a measure of the importance of each of the input seismic attributes on the model's output. We illustrate the value of the SHAP technique using a tree-based machine learning implementation trained to distinguish between Mass Transport Deposits (MTDs) and salt seismic facies in a Gulf of Mexico survey.

Introduction

In the last decade, supervised and unsupervised machine learning techniques such as Self-organizing Maps (SOM), Generative Topographic Maps (GTM), *k*-means, and neural networks (Hampson et al., 2001; Roy et al., 2014; Zhao et al., 2015; Gupta et al., 2018; Pires de Lima et al., 2019; Gupta et al., 2020) have been applied to classify seismic facies, core lithofacies, and to predict well log properties as well as several other applications. However, with the increasing complexity of these models, the interpretation of their results can become quite challenging.

To address this situation, we apply a recently derived technique from the machine learning community called SHAP (Lundberg and Lee., 2017; Lundberg et al., 2018) to study the impact that a suite of candidate seismic attributes has in the predictions of a Random Forest architecture trained to differentiate salt from MTDs facies in a Gulf of Mexico seismic survey.

SHapley Additive exPlanations (SHAP)

Typically, machine learning models are viewed as 'black-box' algorithms that tend to provide limited insight on input-output relationships. There is a lack of interpretability and formalisms that demonstrate feature importance, both globally and locally, to supervised learning of labelled data. SHapley Additive exPlanations (SHAP) are a recent development that enable quantitative estimation of model interpretability (Lundberg and Lee, 2017, Lundberg et al., 2018).

SHAP use concepts from cooperative game theory, thereby assigning each attribute an importance value based on its impact on the model prediction when the feature is present or not during the SHAP estimation (Lundberg and Lee, 2017; Lundberg et al., 2018; Lundberg et al., 2020; Molnar, 2020). In order to explain complex models, SHAP use a linear additive feature attribute method as a simpler explanation model:

$$f(a) = g(a') = \emptyset_0 + \sum_{j=1}^{J} \emptyset_j a'_j \qquad (1)$$

where, $f(a)$ is the original machine learning model we want to explain, $g(a')$ is the simpler explanation model, J is the number of simplified input seismic attributes, $\emptyset_j$ are the SHAP values measured across all possible input orderings, a'_j is the simplified input vector that indicates if a particular seismic attribute is present or not during the estimation, and $\emptyset_0$ is associated with the model prediction when all the attributes are not considered in the estimation (Lundberg and Lee, 2017; Lundberg et al., 2018, Molnar, 2020).

Data description

The 3D seismic survey was acquired by PGS and is located in the Gulf of Mexico, offshore Louisiana, and covers an area of approximately 8000 km^2 (3089 mi^2) (Qi et al., 2016). For this study, the seismic volume was cropped, consisting of 500 inlines, 840 crosslines, and record length of 2 s. Figure 1 shows two salt diapirs (Salt #1 and Salt #2) characterized by low amplitude, chaotic reflectivity in the seismic survey.

Workflow

In order to perform our supervised seismic facies classification task to differentiate between salt and MTDs (Figure 2), we evaluate four candidate seismic attributes selected based on geologic insight: coherence, GLCM entropy, total energy, and reflector convergence. Following Qi et al. (2016) and Lubo-Robles et al. (2019), we apply a 3D Kuwahara filter to the seismic attributes as a preconditioning step for classification to smooth the internal response of the seismic facies and better define the edges between them.

To generate the training dataset for our Random Forest classifier, we pick a suite of polygons enclosing the target facies. In Figure 1, we show a vertical section through the seismic volume in which we extract the training voxels of the salt diapirs (falling within the red polygons), and MTDs (falling within the blue polygons) seismic facies from the four candidate seismic attributes.

In the learning phase of the classifier, we use an 80-20 train/test split, where 80% of the picked voxels are associated with the training data, while 20% of the voxels belong to the validation dataset. Furthermore, because seismic attributes are, in general, characterized by super-Gaussian or Poisson distributions (Walden 1985; Honorio et al., 2014; Lubo-Robles and Marfurt, 2019), a robust normalization scheme is applied to avoid any bias related to different units between the candidate seismic attributes.

In order to compute the SHAP values of the model, we use the TreeExplainer implementation developed by Lundberg et al. (2018), which provides a fast and exact computation of the Shapley values for tree-based machine learning implementations.

Finally, we use SHAP force plots to analyze the model's prediction at a particular voxel, and SHAP global feature importance and SHAP summary plots to study the global behavior of the model (Lundberg et al., 2018; Lundberg et al., 2020; Molnar, 2020).

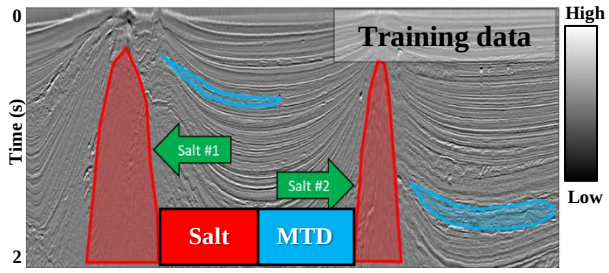

Figure 1. Training dataset generation. A suite of polygons is picked enclosing the target salt and MTDs seismic facies.

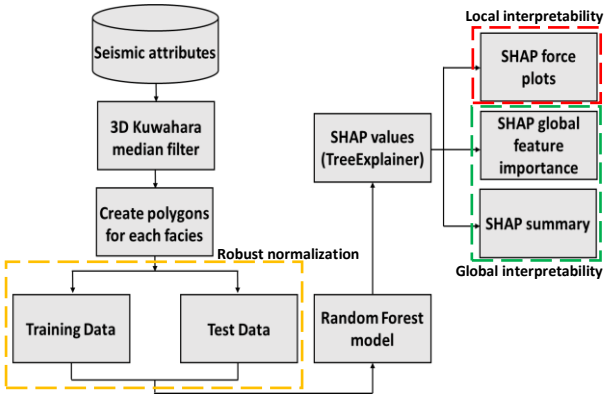

Figure 2. SHAP workflow to analyze the model's prediction.

SHAP application

After training our Random Forest classifier for differentiating between salt and MTDs, we obtain a robust model characterized by accuracy of 98% and 99.4%, and F1-score of 99% and 99.6% in the validation and training dataset respectively. Therefore, the model has good Bias-Variance and Precision-Recall tradeoffs.

Global interpretability

The SHAP global feature importance plot provides a means to analyze the importance of each input attribute in the classification by combining several local explanations of the model (Lundberg et al., 2020). Attributes associated with high average magnitude SHAP values have a greater impact on the classification than features characterized by low average SHAP values (Lundberg et al., 2020; Molnar, 2020).

In Figure 3a, we show the SHAP global feature importance plot for our Random Forest classifier. We note that the four input candidate attributes have the same impact on both seismic facies. Moreover, the highest contribution to the classification is given by the total energy, followed by the coherence, GLCM entropy, and reflector convergence when differentiating between salt and MTDs.

One intrinsic limitation of the SHAP global importance plot is that it does not take into consideration feature effects (Lundberg et al., 2020). To address this issue, we use the SHAP summary plot to analyze the attribute importance together with the magnitude and direction of an attribute's effect (Lundberg et al., 2020; Molnar, 2020).

In Figure 3b, we show the SHAP summary plot of the classifier for the salt seismic facies. The collection of dots in the figure represent individual data points. Each feature (or predictor) in the column is arranged in decreasing order of importance, so the SHAP values (on the x-axis) get progressively smaller down the column (Lundberg et al., 2020; Molnar, 2020). To interpret Figure 3b, we focus on the variable total energy. The SHAP values corresponding to total energy range from negative to larger positive values for the different data points. For the positive SHAP values (these points have a strong influence on the classification of salt), the points are associated with low (colored blue) value of the feature. This indicates that low total energy values are a key characteristic of the salt seismic facies.

Likewise low values of coherence and high values of GLCM entropy and reflector convergence are associated with positive SHAP values and therefore, the salt seismic facies On the other hand, high values of total energy and coherence, and low values of GLCM entropy and reflector convergence decrease the probability of a particular voxel being classified as salt, and increase the probability of having an MTD.

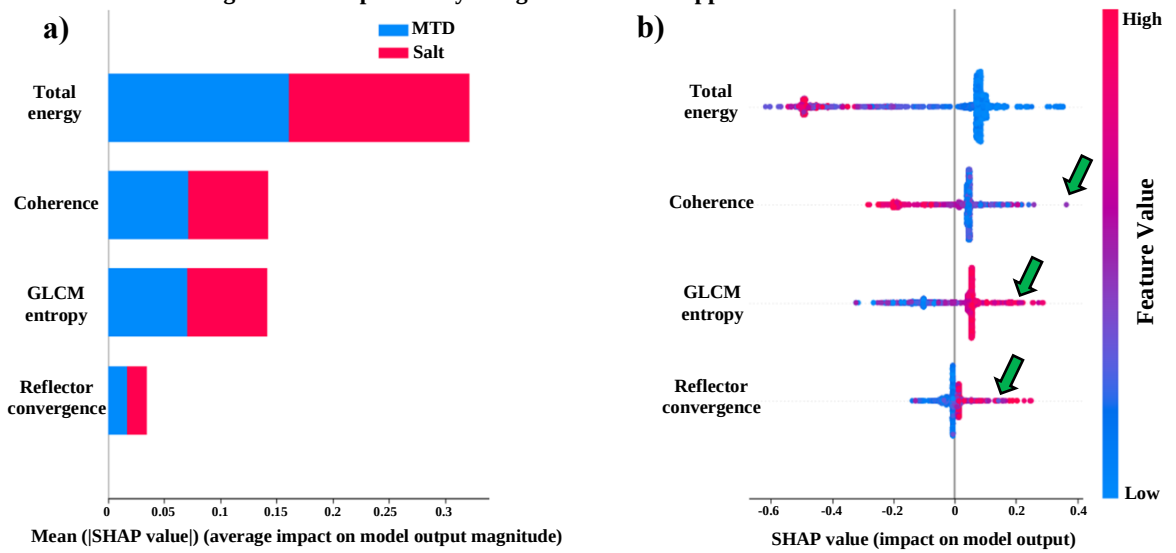

Figure 3. Global interpretation of the Random Forest classifier using SHAP values (a) SHAP global feature importance plot. From four candidate seismic attributes, the highest contribution is associated with the total energy, followed by the coherence, GLCM entropy, and reflector convergence when classifying between salt and MTDs seismic facies. Also, attributes show the same impact on both facies. (b) SHAP summary plot for the salt seismic facies. SHAP positive values increase the probability of having a salt seismic facies, and they are associated with low values of total energy and coherence, and high values of GLCM entropy and reflector convergence. Some outliers characterized by high coherence, or low GLCM entropy or reflector convergence are also visible (green arrows).

Finally, some outliers characterized by high coherence, or low GLCM entropy or reflector convergence (Figure 3b, green arrows) might be classified as salt.

Local interpretability

Local explanations allow to analyze the model classification for selected data points (Lundberg et al., 2020). These are illustrated with SHAP force plots (Lundberg et al., 2018). Following Molnar (2020), SHAP values are associated with different "forces" that increase or decrease the model's prediction. Each prediction starts from the base value, which is given by the average of all probabilities for each seismic facies present in the dataset if none of the input attributes are known (Lundberg and Lee, 2017; Molnar, 2020). In this study, the salt and MTDs seismic facies have a base value probability of 0.82 and 0.18 respectively.

Analyzing the force plot of sample #15 for the salt seismic facies (Figure 4a), we observe that with no inputs whatsoever, the probability of sample #15 being a salt is 82%. This is where the plot begins. Adding reflector convergence (R.C.) slightly increase the probability of sample #15 being a salt facies. Adding coherence (Coh.) pushes the probability to 87%, adding GLCM entropy (GLCM Ent.) pushes it further to 93% and adding total energy (T.E.) pushes it to 100%. It is important to note that the classification of sample #15 as a salt was aided by low values of total energy and coherence, and high values of GLCM entropy and reflector convergence (Table 1).

In Figures 4b and 4c, we study a false negative sample where its actual class is salt, but it was classified as MTD with a 66% probability. We note that this sample is characterized by relatively high values of coherence and total energy and low values of GLCM entropy and reflector convergence, which matches the general behavior expected by an MTD (Figure 3b). For this reason, in Figure 4b, we observe that reflector convergence, GLCM entropy, and total energy push the prediction up from 18 to 66% of being an MTD, while pushing the prediction of being a salt facies from a base value of 82% to a 34% (Figure 4c).

Finally, in sample #256, we note that a value of coherence above the mean tends to push the classification towards the seismic facies. Further research is needed in order to explain how the interaction between coherence with the other candidate seismic attributes tends to increase the probability of having a salt seismic facies in this sample.

Attribute	Mean
Total energy	112380
Coherence	0.58
GLCM entropy	3.78
Reflector convergence	0.22

Table 1. Mean values for each seismic attribute in the training dataset sorted by their importance obtained after implementing SHAP. Values are shown without robust normalization for interpretation purposes.

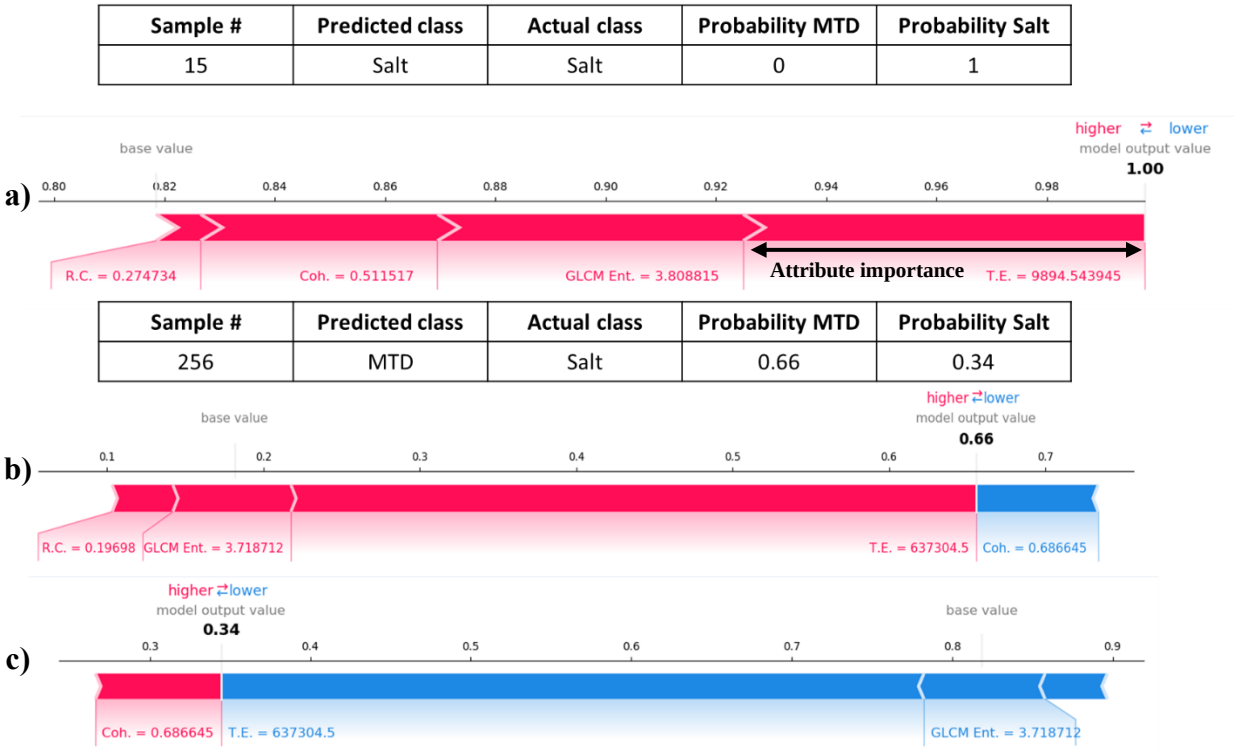

Figure 4. SHAP forces plots for local interpretability (a) Force plot for sample #15 classified correctly as a salt facies. Low values of total energy and coherence and high values of GLCM entropy and reflector convergence push the probability up from 82 to 100% (b) Force plot for sample #256, which is classified as MTD with 66%, but belongs to salt seismic facies. This sample is characterized by relatively high values of coherence and total energy, and low values of GLCM entropy and reflector convergence matching the expected behavior of an MTD (c) Force plot for sample #256 showing a 34% probability of having a salt facies. A coherence value above the mean is pushing the classification towards the salt seismic facies. Further study is required to explain the interaction between the seismic attributes. Attribute values are shown without robust normalization for interpretation purposes.

Conclusions and future work

We successfully applied a SHAP implementation to a Random Forest classifier trained to differentiate between salt and MTDs in a Gulf of Mexico dataset as a means to study seismic attribute importance and the machine learning model predictions. From four input candidate attributes, we determined that the most significant impact in the model is given by the total energy followed by the coherence, GLCM entropy, and reflector convergence. Moreover, we analyze in detail how each attribute's effect impacts the model's output at a local scale, where low values of total energy and coherence and high values of GLCM entropy and reflector convergence tend to increase the probability of particular sample of being classified as a salt seismic facies.

For future work, we will include other texture, non-parallelism, and spectral attributes in the classification to analyze how they affect the model prediction. Also, we will study the force plots of correctly classified MTD and false positive samples, where MTDs are being misclassified as salt by the model. Moreover, we plan to study the SHAP dependence plots to examine how the input seismic attributes interact with each other, and we will apply SHAP to non-binary tasks.

Acknowledgments

We thank PGS for providing the seismic data for use in research and education. Also, we would like to thank the sponsors of the Attribute Assisted Seismic Processing and Interpretation (AASPI) consortium for their support.

REFERENCES

Gupta, I., C. Rai, D. Devegowda, and C. Sondergeld, 2018, Use of data analytics to optimize hydraulic fracture locations along borehole: Petrophysics, **59**, 811–825.

Gupta, I., N. Tran, D. Devegowda, V. Jayaram, C. Rai, C. Sondergeld, and H. Karami, 2020, Looking ahead of the bit using surface drilling and petrophysical data: Machine-learning-based real-time geosteering in volve field: SPE, doi: https://doi.org/10.2118/199882-PA.

Hampson, D. P., J. S. Schuelke, and J. A. Quirein, 2001, Use of multiattribute transforms to predict log properties from seismic data: Geophysics, **66**, 220–236, doi: https://doi.org/10.1190/1.1444899.

Honorio, B., A. Sanchetta, E. Pereira, and A. Vidal, 2014, Independent component spectral analysis: Interpretation, **2**, no. 1, SA21–SA29, doi: https://doi.org/10.1190/INT-2013-0074.1.

Lubo-Robles, D., and K. J. Marfurt, 2019, Independent component analysis for reservoir geomorphology and unsupervised seismic facies classification in the Taranaki Basin, New Zealand: Interpretation, **7**, no. 3, SE19–SE42, doi: https://doi.org/10.1190/INT-2018-0109.1.

Lubo-Robles, D., T. Ha, S. Lakshmivarahan, and K. J. Marfurt, 2019, Supervised seismic facies classification using probabilistic neural networks: Which attributes should the interpreter use?: 89th Annual International Meeting, SEG, Expanded Abstracts, 2273–2276, doi: https://doi.org/10.1190/segam2019-3216841.1.

Lundberg, S., and S. Lee, 2017, A Unified Approach to Interpreting Model Predictions: 31st Conference on Neural Information Processing Systems.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S. Lee, 2020, From local explanations to global understanding with explainable AI for trees: Nature Machine Intelligence, **2**, 56–67.

Lundberg, S. M., G. G. Erion, and S. Lee, 2018, Consistent individualized feature attribution for tree ensembles: arXiv preprint arXiv:1802.03888.

Lundberg, S. M., B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S. Newman, J. Kim, and S. Lee, 2018, Explainable machine-learning predictions for the prevention of hypoxaemia during surgery: Nature Biomedical Engineering, **2**, 749–760.

Molnar, C., 2020, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable: https://christophm.github.io/interpretable-ml-book/shap.html, accessed on 15 April 2020.

Pires de Lima, R., F. Suriamin, K. J. Marfurt, and M. Pranter, 2019, Convolutional neural networks as aid in core lithofacies classification: Interpretation, **7**, no. 3, SF27–SF40, doi: https://doi.org/10.1190/INT-2018-0245.1.

Qi, J., T. Lin, T. Zhao, F. Li, and K. J. Marfurt, 2016, Semisupervised multiattribute seismic facies analysis: Interpretation, **4**, no. 1, SB91–SB106, doi: https://doi.org/10.1190/INT-2015-0098.1.

Roy, A., A. S. Romero-Peláez, T. J. Kwiatkowski, and K. J. Marfurt, 2014, Generative topographic mapping for seismic facies estimation of a carbonate wash, Veracruz Basin, southern Mexico: Interpretation, **2**, no. 1, SA31–SA47, doi: https://doi.org/10.1190/INT-2013-0077.1.

Walden, A. T., 1985, Non-Gaussian reflectivity, entropy, and deconvolution: Geophysics, **50**, 2862–2888, doi: https://doi.org/10.1190/1.1441905.

Zhao, T., V. Jayaram, A. Roy, and K. J. Marfurt, 2015, A comparison of classification techniques for seismic facies recognition: Interpretation, **3**, no. 4, SAE29–SAE58, doi: https://doi.org/10.1190/INT-2015-0044.1.