Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Predicting Subsurface Lithology: a Machine Learning Approach with Explainable Insights

Danilo Vomlel[1]
DCC/UFMG, Belo Horizonte, MG

In petroleum engineering, reservoir characterization is a crucial activity that guides decision-making in the exploration and development of reserves. A key aspect of this process is determining lithological successions [4]. This can be accomplished through direct methods, such as cuttings analysis or core sampling, which, while highly accurate, are often expensive and time-consuming. In contrast, indirect methods, such as well logging, infer lithological properties from subsurface measurements, offering a more cost-effective and efficient alternative.

Indirect methods rely on the expertise of geologists and geophysicists, who interpret well logs—such as gamma-ray, resistivity, sonic, density, and neutron logs—to estimate the most probable lithofacies in a given region. These interpretations are informed by prior lithological knowledge, emphasizing the potential of machine learning techniques to automate and enhance this process. A machine learning model can learn to map well log data to lithological facies, improving efficiency while maintaining or even surpassing the accuracy of manual inference.

The objective of this study is to develop and evaluate a machine learning model for lithofacies classification and to employ explainability techniques [1] to explain the model's predictions. By validating the results against established geological knowledge, this approach aims to uncover patterns in the data, generate actionable insights, and bridge the gap between domain expertise and data-driven methods. Ultimately, this work seeks to advance the efficiency and accuracy of reservoir characterization.

For this study, the FORCE dataset [2], comprising data from 118 wells in the North Sea and 12 distinct lithologies—primarily sandstones and shales—was utilized. The XGBoost [3] algorithm was selected for its ability to capture non-linear relationships through a decision-tree-based architecture. Of the 118 wells, 90 were used for training, with the remaining wells were reserved for testing.

With class balancing applied, the model achieved an accuracy of 59%, whereas without balancing, the accuracy increased to 70%. By incorporating additional features, such as moving averages and spectral transformations of the time series, the model's accuracy reached a maximum of 78%.

To interpret the model's predictions, the SHAP (SHapley Additive exPlanations) [5] method was employed. SHAP is a popular explainability tool that assigns an importance value to each feature for a specific prediction by calculating its contribution to the model output. This approach provides a clear understanding of how each input feature influences the predictions.

The GR variable was the most relevant feature for most classes, while NPHI (neutron) was important for others, such as Sandstone. According to the SHAP values, classes like Chalk and Limestone exhibit very similar behavior across features. However, some classes were identified with greater precision, as shown in Figure 1, with Sandstone increasing its predicted probability under low NPHI and GR values, and the Shale class increasing its probability for high GR, NPHI, and RHOB values.
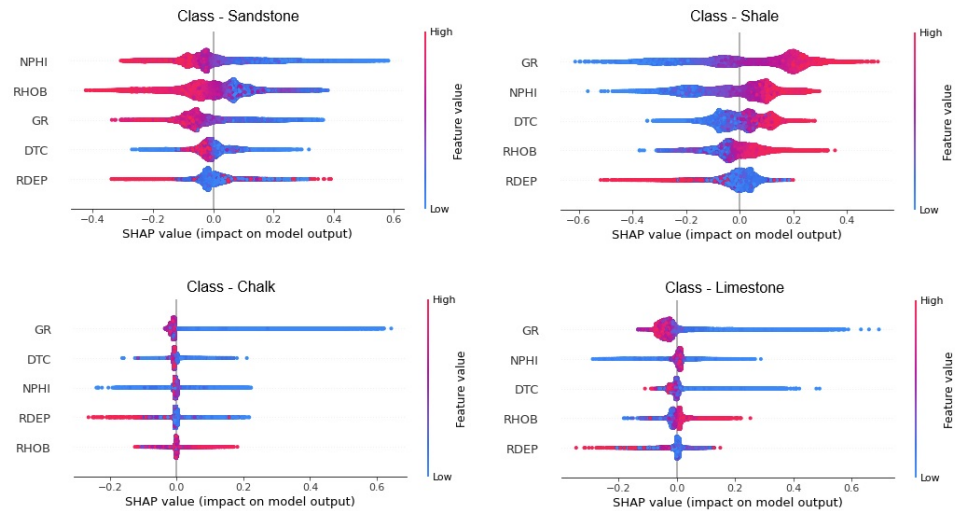
---

[1]dvomlel@ufmg.br

2

Figure 1: SHAP Values for Sandstone, Shale, Chalk and Limestone Across Features in a XGBoost Model for Lithological Classification. Source: Created by the author.

The study demonstrates the utility of combining advanced machine learning techniques, feature engineering, and interpretability tools for lithology prediction. The insights derived from SHAP not only validate the model against geological knowledge but also enhance the understanding of how well log data correlates with lithological classes, paving the way for more robust and interpretable reservoir characterization workflows.

Acknowledgments

References

[1] S. Ali, T. Abuhmed, S. Sappagh, K. Muhammad, J. M. A. Moral, R. Confalonieri, R. Guidotti, J. D. Ser, N. D. Rodríguez, and F. Herrera. "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence". In: **Information Fusion** 99 (2023), p. 101805. DOI: 10.1016/2023.101805.

[2] P. Bormann, P. Aursand, and F. Dilib. **FORCE Machine Learning Competition**. Nov. 2020. DOI: 10.5281/zenodo.4351156. URL: https://github.com/bolgebrygg/Force-2020-Machine-Learning-competition.

[3] T. Chen and C. Guestrin. "Xgboost: A scalable tree boosting system". In: **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[4] M. K. Dubois, G. C. Bohling, and S. Chakrabarti. "Comparison of four approaches to a rock facies classification problem". In: **Computers & Geosciences** 33.5 (2007), pp. 599–617. DOI: 10.1016/j.cageo.2006.08.011.

[5] S. M. Lundberg and S. Lee. "A unified approach to interpreting model predictions". In: **Advances in neural information processing systems** 30 (2017). DOI: 10.48550/1705.07874.