MACHINE LEARNING INTERPRETABILITY METHODS APPLIED TO CALIBRATION MODELS DEVELOPED ON NEAR INFRARED SPECTROSCOPIC DATA

Dr. Astrid MALECHAUX[1], Jordane POULAIN[1], Dr. Sylvie ROUSSEL[1]

[1]Ondalys, 4 rue Georges Besse, 34830 Clapiers, France

Keywords: Machine Learning, Explainable ML, Interpretability, SVM, ANN

In the past decades, Machine Learning (ML) models have become more and more complex, leading to improvements in their predictive performance. However, these models can often be described as "black boxes", in the sense that it is very difficult to explain how results are obtained by a model from the input data. As complex Machine Learning models are increasingly used to make decisions, for instance in industrial or medical applications, there is a growing need to improve their interpretability in order to have greater confidence in their results [1], providing the so-called Explainable AI (Artificial Intelligence).

This study is focused on the interpretability of Machine Learning models after model calibration on near-infrared spectroscopic data, a.k.a. the "post-hoc explanation" of ML models.

Several regression models such as Support Vector Machines (SVM), Artificial Neural Networks (ANN) and Extreme Gradient Boosting (XGBoost), have been trained and compared to the classical PLS calibration models.

The study is applied to the freely available Benchmark dataset consisting of Near Infrared spectra of minced meat samples recorded with the FOSS Tecator Infratec Food and Feed Analyzer in the 850-1050nm wavelength range [2].

Various ML model interpretability algorithms currently applied to non-spectroscopic data have been tested and compared: Local Interpretable Model-agnostic Explanations [3], Shapley Additive Explanations [4] and pseudo-samples prediction [5]. They have also been compared to the regression coefficients of an intrinsically interpretable Partial Least Squares model. While the applicability to spectroscopic data of pseudo-samples prediction has been demonstrated [5], Local Interpretable Model-agnostic Explanations and Shapley Additive Explanations are theoretically better suited to uncorrelated variables [6].

The results indicate that the outputs of the interpretability methods tested appear to be in good agreement with the Partial Least Squares regression coefficients. Thus, they allow the identification of wavelengths ranges used by "black box" Machine Learning models. Moreover, they could be used as indicators of the risk of overfitting for complex models developed on spectroscopic data.

In conclusion, this study shows the potential for applying different Machine Learning model interpretability methods to Near Infrared spectroscopic data. Further work on different NIR datasets will be conducted to confirm these results.

Astrid MALECHAUX, Jordane POULAIN, Sylvie ROUSSEL. Machine learning interpretability methods applied to calibration models developed on near infrared spectroscopic data. ICNIRS 2025, June 2025, Rome, Italia.

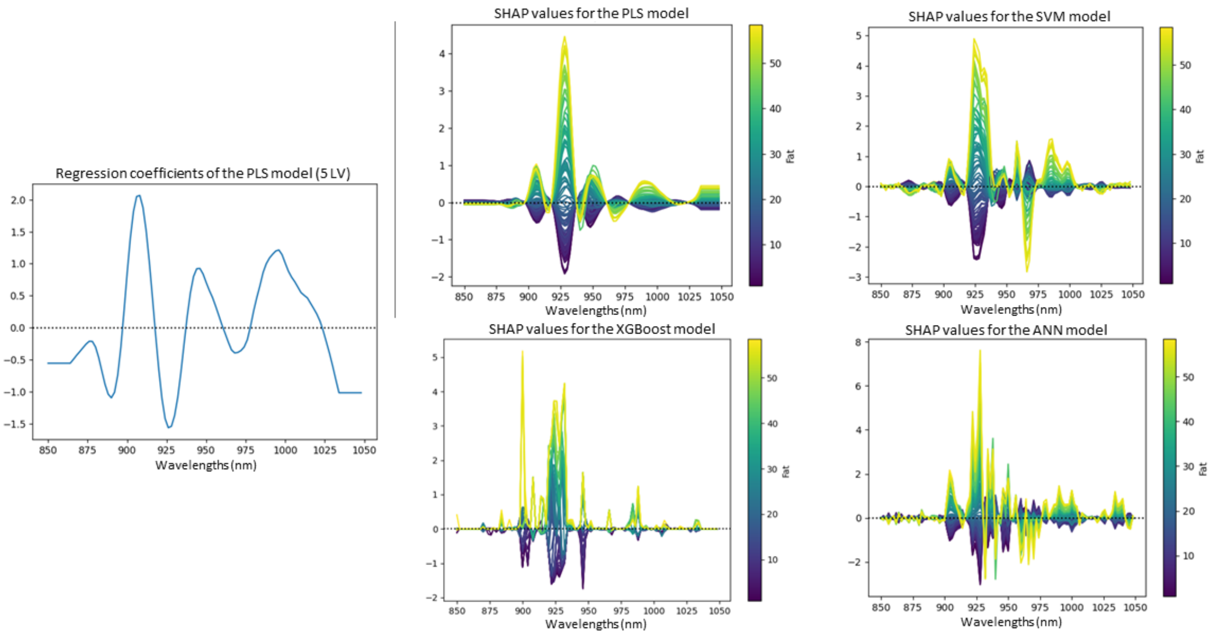

Figure 1: Example of results obtained on the calibration spectra with the Shapley Additive Explanations method for different Machine Learning models

ACKNOWLEDGMENTS

Ondalys wishes to acknowledge the Occitanie Region for the funding of the R&D projects on Explainable AI/ML.

[1] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable AI: A Review of Machine Learning Interpretability Methods, Entropy, 2021, 23, 18.

[2] H. H. Thodberg, Tecator meat sample dataset, StatLib Datasets Archive, 1995.

[3] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, 1135-1144.

[4] S. M. Lundberg, S. Lee, A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems, 2017, 30, 4765-4774.

[5] G. J. Postma, P. W. T. Krooshof, L. M. C. Buydens, Opening the kernel of kernel partial least squares and support vector machines, Analytica Chimica Acta, 2011, 705(1-2), 123-134.

[6] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, K. Lekadir, G. Menegaz, A perspective on explainable artificial intelligence methods: SHAP and LIME, Advanced Intelligent Systems, 2024, 2400304.