Leveraging Formal Methods for Efficient Explainable AI

Amira Jemaa, Adnan Rashid, Sofiène Tahar

Department of Electrical and Computer Engineering
Concordia University, Montreal, QC, Canada

{*a_jem, rashid, tahar*}*@encs.concordia.ca*

In modern data-driven decision-making, transparency and clarity in Machine Learning (ML) models are critical. Explainable Artificial Intelligence (XAI) [1] fulfills this demand by offering human-understandable explanations for the predictions and decisions made by complex Artificial Intelligence (AI) algorithms. XAI bridges the gap between algorithmic processes and end-users, enabling stakeholders to trust and comprehend AI system decisions, thus promoting accountability, fairness, and ethical use of technology.

Heuristic approaches like SHAP [2], LIME [3], and Anchor [4] are commonly used to explain predictions from non-interpretable ML models. However, these methods lack global guarantees, often providing locally valid explanations that fail to hold true across the entire instance space. This limitation highlights a need for more rigorous approaches, particularly those based on formal methods, to ensure robust and reliable explanations.

To fulfill this need, various tools, such as XReason [5], Silas [6], and PyXAI [7], have been recently introduced to enhance the trust in the explainability of ML models. For instance, XReason utilizes XGBoost Classifier [8] alongside SAT solvers and SMT methods to provide detailed, instance-level abductive explanations. Similarly, Silas employs Random Forest [9] for robust explainability, focusing on feature importance, and uses SMT in their work. PyXAI, a recent addition to the field that specializes in explaining models like RandomForest Classifier, XGBoost Classifier, XGBoost Regressor, and LGBM Regressor using an abductive approach at the instance level and employs SAT solvers.

In this research, we have chosen XReason due to its open-source availability, allowing for extensive customization and enhancement. Additionally, XReason's robust integration of SAT-solvers provides a comprehensive and reliable framework for generating detailed, instance-level abductive explanations, making it an ideal choice for our needs. Building on XReason, our approach uses Maximum Satisfiability (MaxSAT) solvers [10] to generate explanations with the fewest essential features. We have enhanced XReason by implementing class explanations and integrating it with the LightGBM (LGBM) Classifier [11], an advanced gradient boosting framework. Our focus is on translating complex decision rules into concise yet informative explanations. Unlike heuristic methods, this approach ensures explanations are clear and efficient across all instances, offering dependable insights into model predictions.

The process of generating explanations using formal methods involves several steps. A

trained ML model, which is inherently non-interpretable, is encoded into a formal representation using techniques like MaxSAT. This encoding allows the solver to identify the most important features of an instance, ensuring globally consistent and verifiable explanations across the entire instance space. Moreover, the class explanations are crucial as they provide insights into how different classes are predicted by the model, and hence enhancing the understanding of the underlying patterns and decision-making processes specific to tabular data. By encoding and generating explanations for different classes, users gain a comprehensive view of the model's behavior and predictions. In order to validate our approach, we conducted several experiments using the segmentation dataset [12], which consists of 210 samples with 19 features of tabular data. Despite variations in the dataset specifics, a thorough analysis confirms the accuracy of all explanations, showcasing the robustness of the proposed methods.

In a second effort, we have investigated the integration of LGBM Classifier with XReason, which resulted in improvements in explanation generation. In fact, LGBM explanations consistently provided shorter and more concise explanations compared to those of XGBoost. Specifically, the average length of MaxSAT explanations for LGBM is 5.51, slightly lower than XGBoost's 5.68. The median explanation lengths were identical for both models, which was equal to 5, but LGBM exhibits a lower standard deviation of 1.88 compared to XGBoost's 2.16, indicating that LGBM provides more consistent explanations across different instances. Moreover, the percentage of shortest explanations using MaxSAT for LGBM is 49.52%, closely matching XGBoost's 50.47%, but with reduced explanation variability. The improvements in consistency and conciseness without sacrificing accuracy underscore the advantages of using LGBM in conjunction with formal methods for XAI.

In summary, integrating formal methods into XAI frameworks like XReason, coupled with the enhancements introduced, represents a significant advancement in the field. This tool delivers robust, efficient, and interpretable explanations, making it indispensable for the development of trustworthy AI systems. Figure 1 depicts an abstract overview of the XReason tool, where our contributions are identified with yellow boxes. Future work aims to further refine XReason and expand its applicability across various ML tasks and industrial applications. Additionally, efforts will be directed towards addressing adversarial attacks in XAI, with the goal of enhancing the robustness and reliability of the explanations provided by the tool.

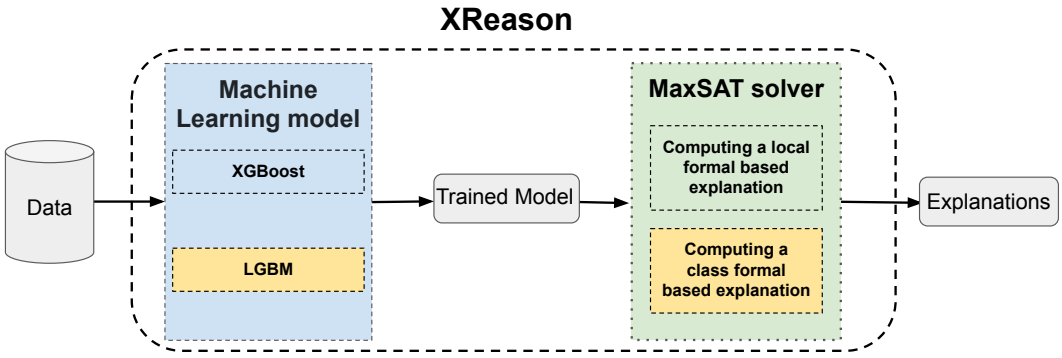

Figure 1: The explanation process

References

[1] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM 2016, pp. 1135–1144.

[2] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, 2017, pp. 4765–4774.

[3] M. T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you? explaining the predictions of any classifier, in: SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 1135–1144.

[4] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: AAAI Conference on Artificial Intelligence, volume 32, 2018, pp. 1528–1535.

[5] XReason, [online assessed 2024]. URL: `https://github.com/alexeyignatiev/xreason`.

[6] Silas, [online assessed 2024]. URL: `https://www.depintel.com/silas_download.html`.

[7] PyXAI, [online assessed 2024]. URL: `https://www.cril.univ-artois.fr/pyxai/`.

[8] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[9] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[10] A. Ignatiev, Y. Izza, P. J. Stuckey, J. Marques-Silva, Using maxsat for efficient explanations of tree ensembles, in: AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 3776–3785.

[11] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, LightGBM: A highly efficient gradient boosting decision tree, Advances in Neural Information Processing Systems 30 (2017) 3149–3157.

[12] Segmentation dataset, [online assessed 2024]. URL: https://epistasislab.github.io/pmlb/profile/segmentation.html.