Learning with explainable trees

Tree-based models are among the most popular and successful machine learning algorithms in practice. New tools allow us to explain the predictions and gain insight into the global behavior of these models.

Although seldom mentioned in news articles, they are often the favorites of machine learning practitioners: tree-based models. These non-linear models are able to learn complex relationships in data and have a long history in machine learning research, as well as a successful track record in a wide range of practical applications [1]. Especially when it comes to the analysis of datasets without strong multi-scale temporal or spatial structures, e.g., tabular-style data, tree-based models regularly outperform other approaches, including deep models [2]. In addition to their excellent performance, high computational and data efficiency, and ease of use, tree-based models are also often considered to be more interpretable --a property of high practical relevance-- than deep neural networks and kernel methods.

But are trees really so easy to interpret? Simple decision trees certainly have this property. Here, visualizing the decision path through the tree suffices to understand how and why the model arrives at its prediction (see Fig. 1). However, this analysis becomes rather unpractical for state-of-the-art ensemble tree models such as random forests or large gradient-boosted decision trees. The involvement of multiple trees in the decision process and the complexity of the decision path make the interpretation of the results very difficult for the human user. Therefore, state-of-the-art tree models are practically black boxes.

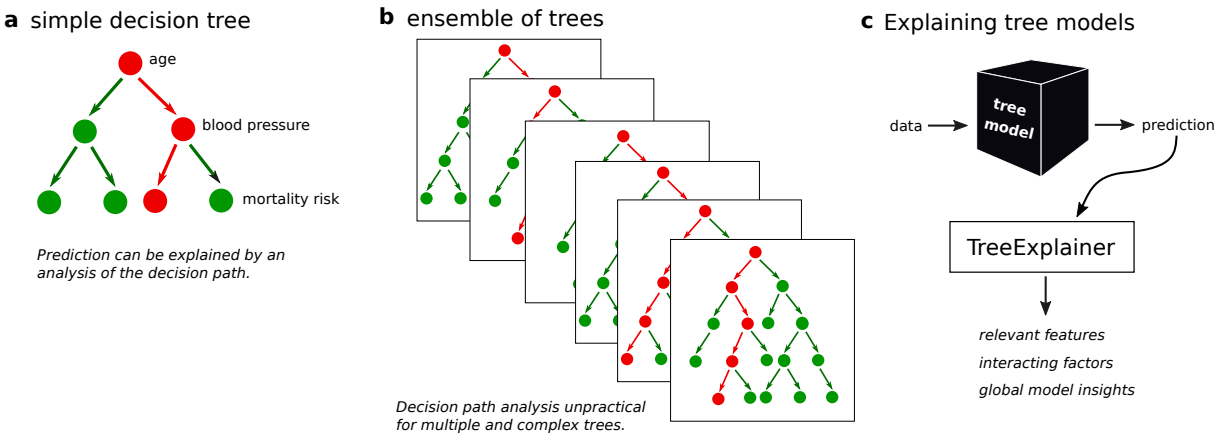

Fig. 1 | **Explanation of tree-based models**. **a**, Simple decision trees can be easily understood by visualizing the decision path. **b**, Due to their complexity, state-of-the-art ensemble tree models are practically black boxes. **c**, TreeExplainer extracts relevant features and finds interaction effects in tree-based models.

The field of explainable AI (XAI) has recently developed various techniques for explaining individual decisions of complex machine learning models [3]. These explanations quantify the contribution of individual input variables (e.g., image pixels) to the overall model decision, allowing one to find out what the model is paying attention to when, e.g., classifying an image. Most of this XAI research has focused on explaining deep neural networks; only few works have transferred XAI concepts to other types of machine learning algorithms, e.g., k-means

clustering [4]. Despite their practical relevance, tree-based models have so far not been in the focus of this research.

In a recent paper, published in *Nature Machine Intelligence*, Lundberg et al. proposed TreeExplainer [5], a general method to explain the individual decisions of a tree-based model in terms of input contributions. By building on the classic game-theoretic Shapley values [6], TreeExplainer has a solid mathematical foundation and various desirable properties. One key contribution of Lundberg et al. is the derivation of a polynomial time algorithm for the exact computation of Shapley values in tree-based models. (For general models, the exact computation of Shapley values is NP-hard.) This exactness allows the user to compute explanations with theoretical guarantees, e.g., consistency. Another significant innovation is the extension of the concept of explanation to the analysis of feature interactions. This is an important step forward in explainable AI research as it enables one to measure interaction effects. For instance, in genetics it can be crucial to distinguish between (predicted) outcomes for which two genes are relevant independently of each other, and outcomes for which their joint presence is causal. TreeExplainer allows one to distinguish between these two cases. Finally, Lundberg et al. developed a new set of tools for a global analysis of model behavior based on the explanations provided by TreeExplainer. The practical value of these tools is demonstrated in three medical machine learning problems.

In one example, the analysis of a tree model trained on mortality risk factors in the general US population reveals that seemingly irrelevant risk factors can be extremely important for specific individuals (rare high magnitude effects). Furthermore, the model analysis uncovers a meaningful interaction between the risk factors of blood pressure and age. It shows that an early age onset of high blood pressure is a much larger risk factor than a late age onset, a hidden interaction which is consistent with medical knowledge. It is very difficult or even impossible to obtain such complex insights out of a tree-based model without explanations.

By making tree-based models explainable, Lundberg et al. bring the advantages of XAI to many practical applications. The proposed analysis tools will help us to better understand what our models *really* do, why they arrive at their predictions and why they sometimes fail. This is important to foster trust in AI and to unmask Clever Hans predictors [7], i.e., models which only pretend to solve the task, but base their predictions on artifacts, irrelevant features or spurious correlations. The ability to explain tree-based models will further increase their popularity, e.g., in applications legally requiring a “right to explanation” [8]. And if at some time all models become explainable by default, then we will probably be asked: “What was it like to use these tree-based methods, when they were still black boxes?”

References

1. Kaggle. The State of ML and Data Science 2017. (2017). url: https://www.kaggle.com/surveys/2017.
2. Chen, T. & Guestrin, C. *Proc. of the 22nd ACM SIGKDD*, pp. 785–794 (2016).
3. Samek, W. et al. (Eds.) *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Springer, 2019)
4. Kauffmann, J. et al. *arXiv:1906.07633* (2019)
5. Lundberg, S. et al. *Nat. Mach. Intell.* (2020).
6. Shapley, L. *Contributions to the Theory of Games 2.28*, pp. 307–317 (1953).
7. Lapuschkin, S. et al. *Nat. Comm.* **10**, 1096 (2019)

8. Bryce, G. & Flaxman, S. *AI Magazine* **38**, 350-57 (2017).

Author information

Affiliations
Machine Learning Group, Fraunhofer Heinrich Hertz Institute
Berlin, Germany
Wojciech Samek

Corresponding author
Correspondence to Wojciech Samek