BellatrExplorer: An Interactive Random Forest Local Explainability Dashboard

Robbe D'hondt[1,2][0000−0001−7843−2178] (✉) and Celine Vens[1,2][0000−0003−0983−256X]

[1] KU Leuven Campus Kulak, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium
{robbe.dhondt,celine.vens}@kuleuven.be
[2] KU Leuven, imec research group itec

Abstract. This paper presents BellatrExplorer, a dashboard application to interactively explore random forest predictions on the individual instance level. The application is inspired by the recently proposed local interpretability toolbox Bellatrex, that exploits the internal random forest structure to extract 1-3 prototype rules that act as a surrogate model for an instance of interest. BellatrExplorer is aimed at expert users trying to better understand the behavior of their random forest in a specific application, and could allow to uncover potential biases or artifacts arising in model training. Currently, the tool supports random forests for binary classification, regression, and survival analysis tasks. It features (1) intuitive exploration of univariate predictive counterfactuals, (2) analysis of decision tree rules to the individual split level, and (3) a visualisation of the rules extracted by Bellatrex that allow to assess the local interpretation at a glance. The tool is available at https://github.com/robbedhondt/BellatrExplorer/ and a demonstration video can be found at https://itec.kuleuven-kulak.be/bellatrexplorer/.

Keywords: random forest · local explainability · interactive dashboard.

1 Introduction

Random forests [2] are popular machine learning models at the state-of-the-art for tabular data learning. The learning algorithm is based on building an ensemble of decision trees, which are simple rule-based models that recursively partition the data. Randomization across the trees is achieved by training each tree on an independent bootstrapped sample of the original dataset and by using only a random subset of the available features per split. In theory, this makes the decision process of the random forest fully transparent. However, the inherent explainability is limited due to the typically large number of trees (100 or more) and their depth (typically deeper than a 'normal' decision tree[3]).

[3] As the trees in a random forest are built using only a subset of the features per split (weak learners), they can grow deeper to capture more complex patterns in the dataset. Overfitting is mitigated by the ensemble nature of the random forest — averaging multiple trees naturally reduces the variance problems that deep decision trees present.

Several tools have been proposed in the literature to open up the random forest black box, both on a population level (global explainability) as well as on an individual prediction level (local explainability). Here, we focus on the recently proposed model-specific local explainability toolbox Bellatrex [3]. For a given sample, Bellatrex extracts 1 to 3 representative prototype rules from the random forest that, when taken together, closely approximate the prediction of the full ensemble.

Many dashboard applications integrating these model explainability tools already exist. Two notable examples are modelStudio [1] and `explainerdashboard`[4], that allow the user to select a set of graphs to evaluate any general-purpose model based on a validation of its performance and on model-agnostic explainability toolboxes.

For random forests, two model-specific dashboards are of interest. The first one, RfX [5], focuses on global explainability through icicle plots and two-dimensional embeddings of the decision trees. The second one, by Gurung et al. [6], probably comes the closest to our work. However, it is focused only on binary classification problems, whereas we also tackle regression and survival analysis. Additionally, their choices of visualisations represents a mix of global explainability (statistics like the number of times each feature is used at each split and feature split points) and local explainability (surrogate tree construction and counterfactual generation using the actionable tweaking algorithm). This is very different from our choices of visualizations that focus on local explainability, giving detailed information on the rule level and allowing counterfactual exploration (rather than focusing on generation). Finally, in Gurung et al. the target audience is mostly non-expert, whereas here the proposed dashboard focuses on the machine learning engineer familiar with the random forest algorithm.

2 Dashboard components

A screenshot of the dashboard is shown in Figure 1. In this section, we discuss the different components of the dashboard and the possibilities of the application in more detail.

Modeling: In the modeling pane, the random forest is set up. First, a built-in dataset is selected, or the user uploads their own dataset. One of the columns of this dataset is selected as the target variable. This variable represents a binary classification, regression, or survival analysis (through random survival forests [7]) task. The final button starts the training of the random forest.

Instance selection: Once the random forest finished training, this section is populated with a slider for each feature in the dataset. The slider options are the percentiles of that feature in the training data, but the options are presented and spaced in the scale of the original feature distribution. In the background of each slider, a linear gradient indicates the change in predicted value of the random forest when moving one of the sliders to a certain position.

[4] See `https://explainerdashboard.readthedocs.io/` (last visited 2025-04-28).

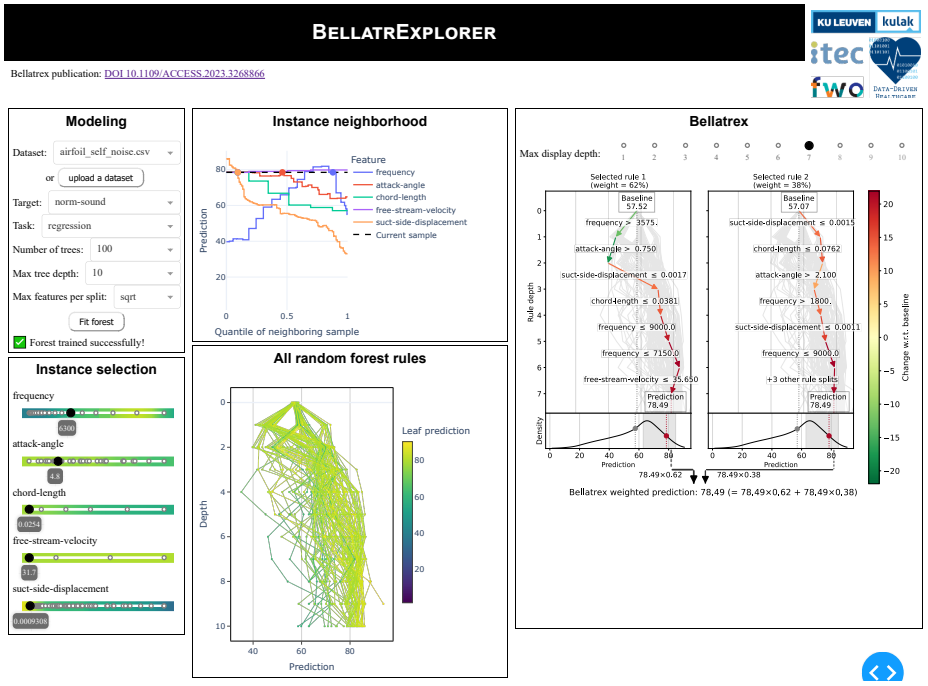

Fig. 1. Screenshot of the dashboard application.

Instance neighborhood: This panel shows the information from the background gradient of the sliders in a more quantitative way. In essence, this plot shows for each feature the univariate effect of changing that feature from the current sample to any percentile of interest (shown on the x axis). The current sample is indicated by a circle for each feature, at the current prediction and the corresponding currently selected percentile. Double-clicking one of the features in the legend isolates the line for that feature, allowing a better visual inspection. Additionally, hovering over any point in the graph gives extra information about the raw feature value at the quantile of interest.

All random forest rules: This panel shows the decision path for the current sample for each tree in the random forest. The sample starts at the root node for each tree with a baseline prediction based on the bootstrapped sample for that tree. With each split, the sample works its way down each tree, gradually changing its predicted value based on the prototype function[5] that aggregates all the training samples in that node. The decision path of each tree is colored according to the final prediction of that tree (in the leaf node). Hovering over a

[5] For regression and binary classification, this is simply the mean of the target. For survival analysis, we compute the mean survival time (by integrating the survival function that is computed at each node with the trapezoidal rule).

node in the plot results in a tooltip giving more information on what split was made in that node and all the nodes before it (partial rule path).

Bellatrex rules: The final panel displays the rules selected by Bellatrex. The original Bellatrex publication [3] presented the selected prototype rules to the user in a textual format. Here, we display them in a similar way as in the "All random forest rules" panel[6], a visualisation that we have used before in the context of multiple sclerosis progression predictions [4]. A separate graph for this is preferred, as it allows the user to view the selected rules with all their splits at a single glance.

3 Future work

As Bellatrex can be applied to any random forest ensemble, we aim to also provide an interface in the future allowing users to upload their own random forest model (from any machine learning package). Additionally, the dashboard could also be extended to the multi-target setting (multi-label classification, multi-target regression, and multi-event survival analysis) by showing linked rulepaths for each target. Finally, the dashboard can then be evaluated in a user study for any further refinements.

Acknowledgments. This work was funded by Research Fund Flanders (FWO fellowship 1S38025N and grant G0A2120N) and supported by the Flemish government (through the AI Research Program). We thank Arthur Cremelie for implementing a first version of the dashboard.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baniecki, H., Biecek, P.: modelStudio: Interactive Studio with Explanations for ML Predictive Models. Journal of Open Source Software **4**(43), 1798 (2019), https://doi.org/10.21105/joss.01798
2. Breiman, L.: Random forests. Machine learning **45**, 5–32 (2001)
3. Dedja, K., Nakano, F.K., Pliakos, K., Vens, C.: BELLATREX: Building Explanations Through a LocaLly AccuraTe Rule EXtractor. IEEE Access **11**, 41348–41367 (2023). https://doi.org/10.1109/ACCESS.2023.3268866
4. D'hondt, R., Dedja, K., Aerts, S., Van Wijmeersch, B., Kalincik, T., Reddel, S., Havrdova, E.K., Lugaresi, A., Weinstock-Guttman, B., Mrabet, S., et al.: Explainable time-to-progression predictions in multiple sclerosis. Computer Methods and Programs in Biomedicine p. 108624 (2025)
5. Eirich, J., Münch, M., Jäckle, D., Sedlmair, M., Bonart, J., Schreck, T.: Rfx: A design study for the interactive exploration of a random forest to enhance testing procedures for electrical engines. In: Computer Graphics Forum. vol. 41, pp. 302–315. Wiley Online Library (2022)

[6] For more information about how to interpret the plot, see https://itec.kuleuven-kulak.be/a-guide-to-bellatrex/.

6. Gurung, R., Lindgren, T., Boström, H.: An interactive visual tool to enhance understanding of random forest predictions. In: European Conference on Data Analysis (ECDA) (2019)
7. Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S.: Random survival forests. Ann. Appl. Stat (2008)