forester: an R package for automated building of tree-based machine learning models

Anna Kozak, Adrianna Grudzień, Hubert Ruczyński, Patryk Słowakiewicz

MI2.AI Group, Faculty of Mathematics and Information Science, Warsaw University of Technology

Introduction

A significant amount of time is spent on building models with high performance. Selecting the appropriate model structures, optimising hyperparameters and explainability are only part of the process of creating a machine learning-based solution. Despite the wide range of structures considered, tree-based models are champions in competitions or hackathons. So, aren't tree-based models enough? They are, and that's why **we want to fully automate the process of training tree-based models so that even the newcomers can easily build**, train and understand these powerful prediction tools. At the same time, **the experienced users gain a powerful tool for making high-quality baseline models** for new tasks, they start working with.

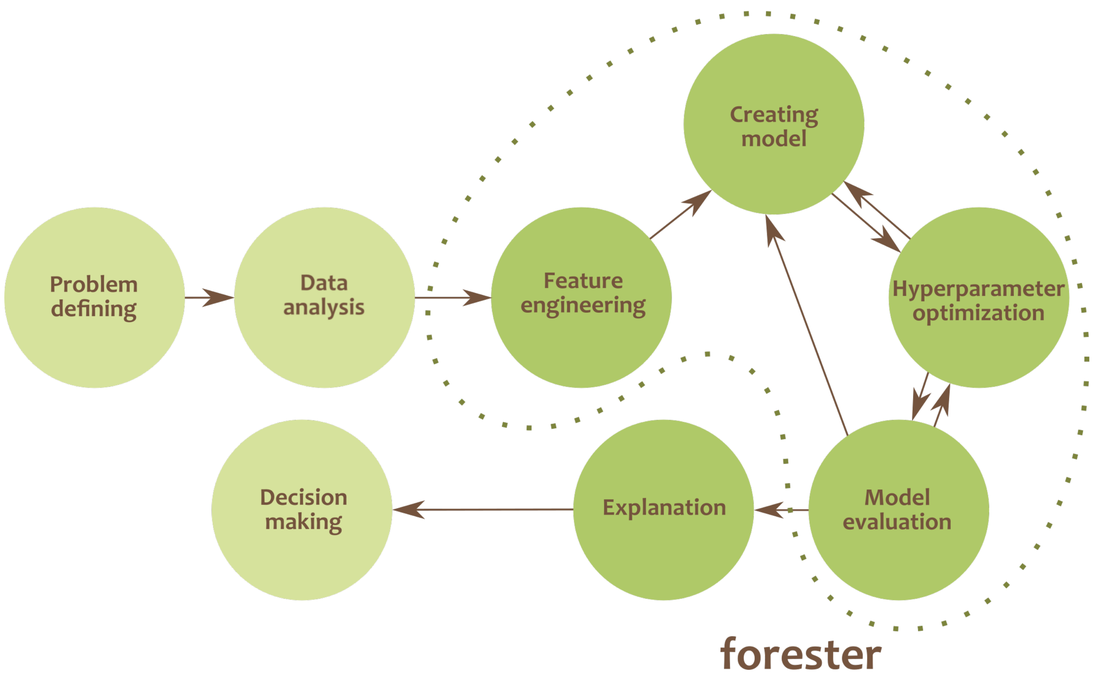
What is the *forester?*

The *forester* is **an autoML tool in R** that wraps up all machine learning processes into a single `train()` function, which includes:

1. rendering a brief data check report,
2. preprocessing initial dataset enough for models to be trained,
3. training 5 tree-based models with default parameters, random search and Bayesian optimisation,
4. evaluating them and providing a ranked list.

However, that's not everything that the *forester* has to offer. Via additional functions, the user can easily explain created models with the usage of *DALEX* or generate one of the predefined reports including:

1. information about the dataset,
2. in-depth parameters of trained models,
3. visualisations comparing the best models,
4. explanations of the aforementioned models.

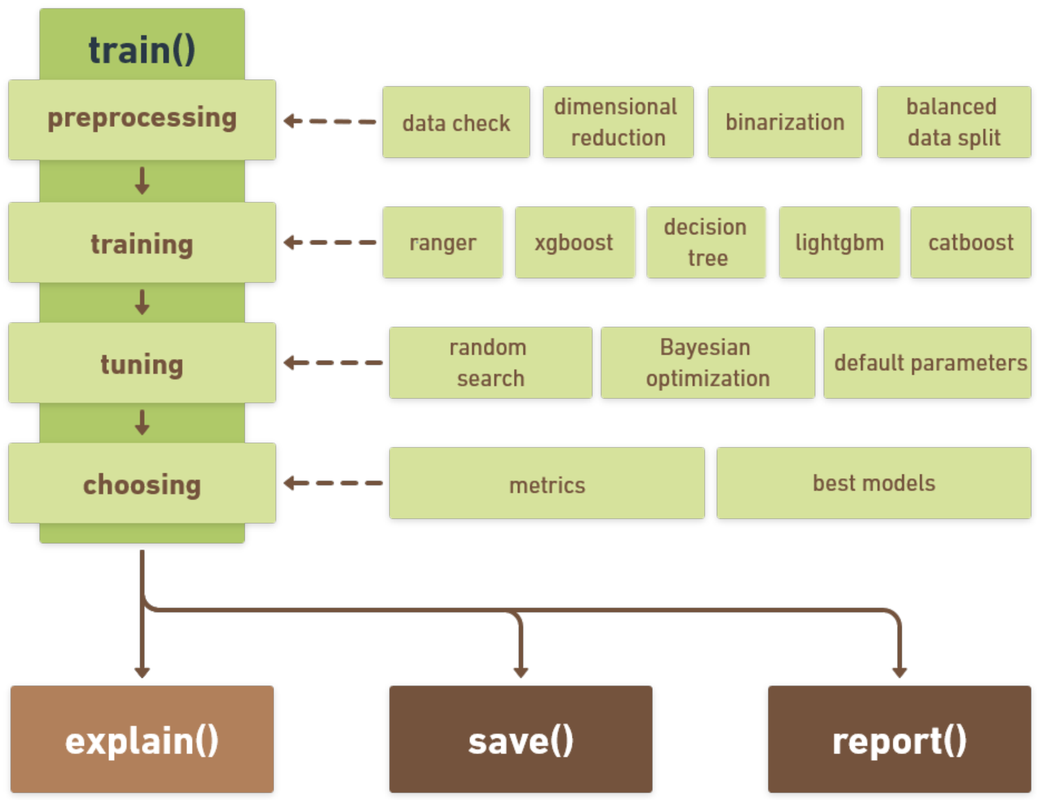


Why tree-based models?

Tree-based models, especially **XGBoost are extremely popular amongst winners in Kaggle competitions** and they firmly show their superiority with tabular data, not only in terms of fast computations. Moreover, the researchers also prove that **tree-based models are superior to deep learning neural networks** because they don't suffer from uninformative columns presence and are not biased toward overly smoothed solutions.

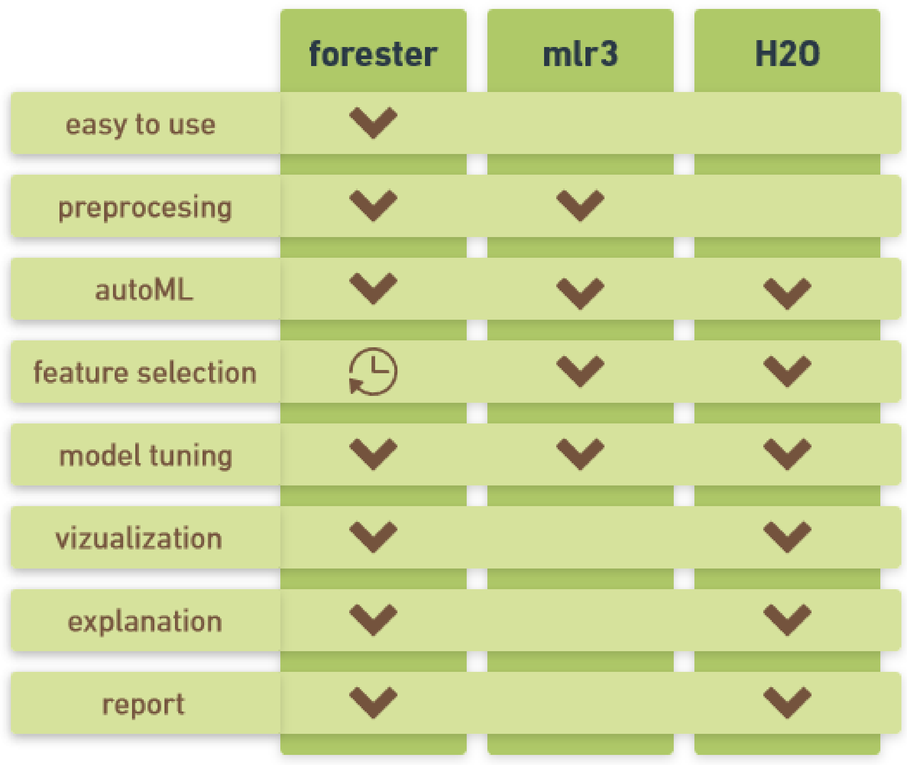
Package structure

With functions in *forester* package users can create a well-tuned tree-based model with a unified, simple formula. With the usage of only two required parameters: the raw, not preprocessed dataset and target column name, the user is able to achieve satisfying results. The *forester* automatically handles the "ugly" part for you.



For whom is this package created?

The *forester* is designed for beginners in data science, but also for more experienced users. They get an easy-to-use tool that can be used to prepare high-quality baseline models for comparison with more advanced methods or a set of output parameters for more thorough optimisations. **Tree-based models are created in just one line of code.** The package differentiates itself in this aspect from powerful autoML frameworks like *mlr3* and *H2O*.



Contact info

anna.kozak@pw.edu.pl
grudziena@outlook.com
hruczynski21@interia.pl
slowakiewiczpatryk@outlook.com
https://github.com/ModelOriented/forester