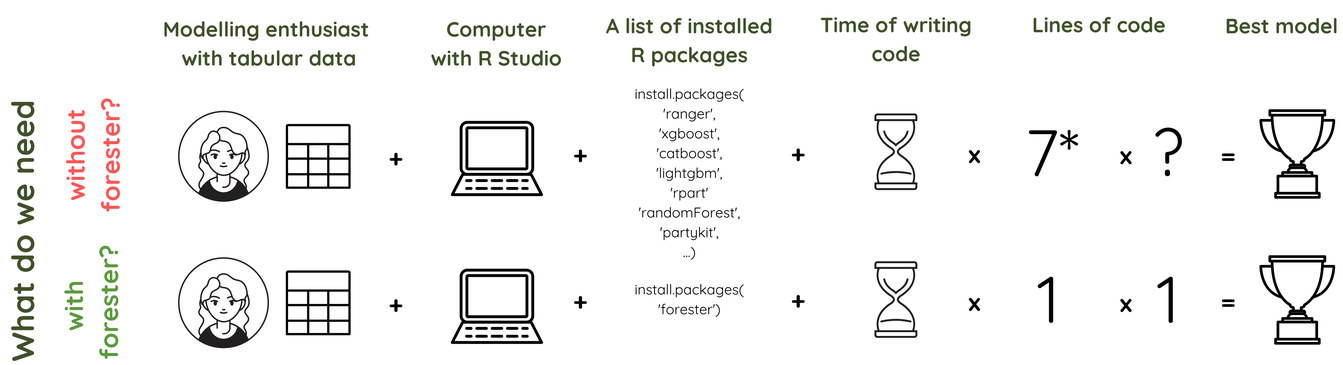
forester: A Novel Approach to Accessible and Interpretable AutoML for Tree-Based Modeling

Anna Kozak [1], Hubert Ruczyński[1]

[1]Warsaw University of Technology

Let's talk about AutoML, tree-based models, explainable AI (XAI), and exploratory data analysis (EDA)!

How to build tree-based models in R?



* dependent on the number of packages used

What is *forester*?

- Full automation of the process of training tree-based models,
- No demand for ML expertise,
- Powerful tool for making high-quality baseline models for experienced users.

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

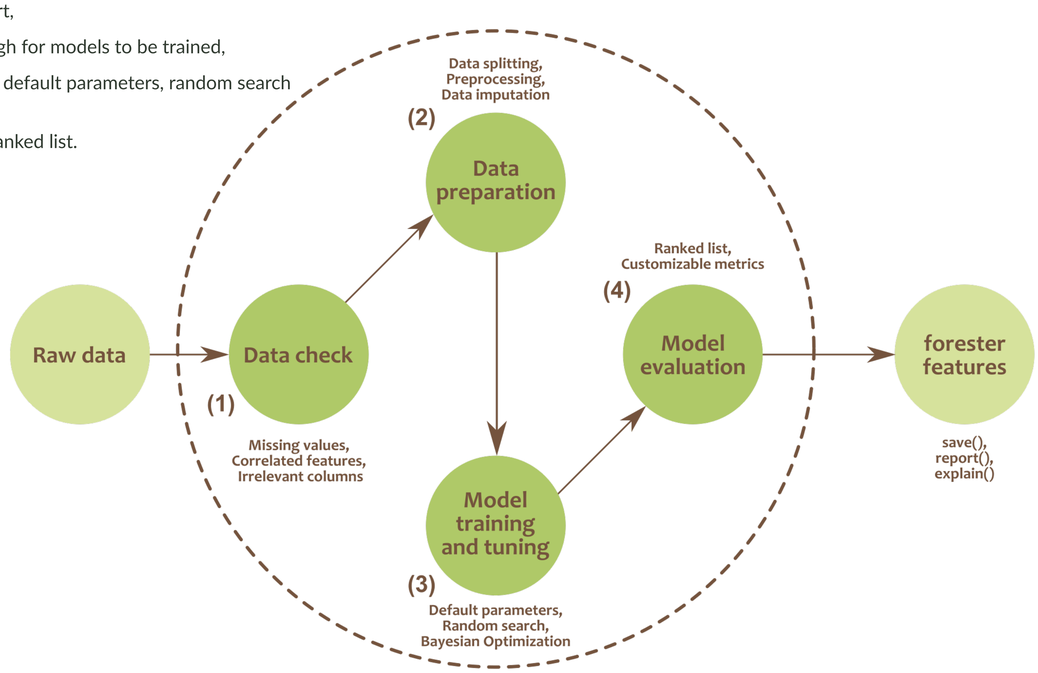
1. Rendering a brief **data check** report,
2. **Preprocessing** initial dataset enough for models to be trained,
3. **Training** 5 tree-based models with default parameters, random search and Bayesian optimisation,
4. **Evaluating** them and providing a ranked list.

For whom is this package created?

The *forester* package 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.

How to use it?

```
library(forester)
data('lisbon')
train_output <- train(lisbon, 'Price')
```

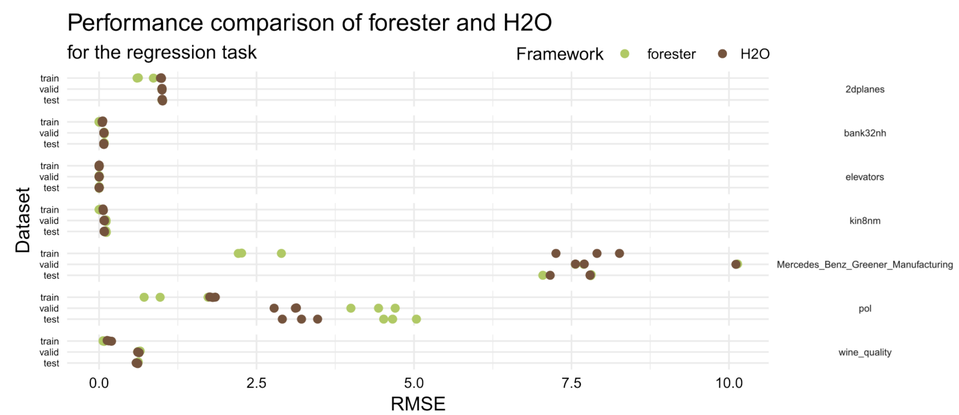


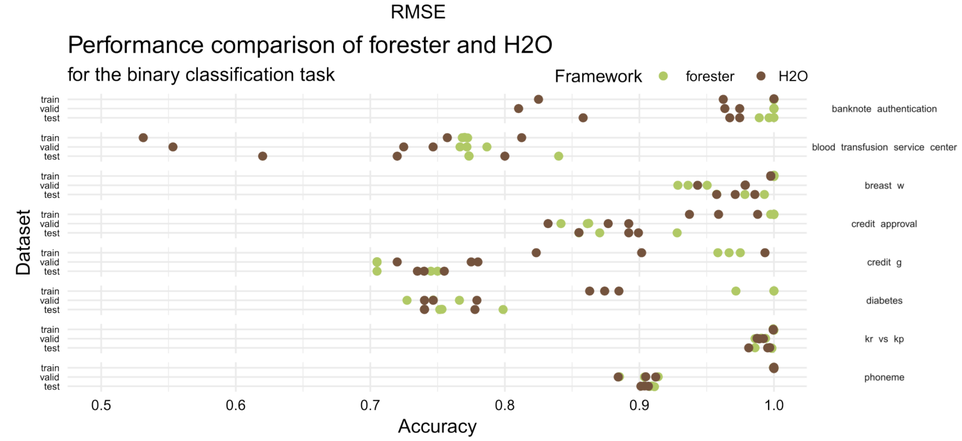
Prepare meaningful report less than in 60 seconds!

As data scientists, we are fully aware that there are some time expensive processes in out work. One of them is creating a report with meaningful results. That's why one of the most powerful *forester* feature, which makes it a efficient tool for both experienced users and the newcomers, is a `report()` function. This single-line command is designed to **provide a holistic view on the outcomes of the ML process** happening inside of the *forester*.

Simple doesn't mean worse!

According to our experiments, the forester package achieves competitive results in much shorter time in comparison to well-known H2O AutoML tool. We have compared their performance on 8 binary classification, and 7 regression tasks, and the calculations were repeated 3 times for each dataset and framework. The forester outperformed H2O most of the times, even though the latter package's training lasted 2 times longer on average.





Contact info

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

Paper

GitHub

References

L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=Fp7__phQszn.

A. Kozak, H. Ruczyński, P. Słowakiewicz, A. Grudzień, and P. Biecek. *forester: Quick and Simple Tools for Training and Testing of Tree-based Models*, 2023. URL https://github.com/ModelOriented/forester. R package version 1.1.4.