# Investigating the Efficiency of Tree-based Models for Tabular Data with *forester* Package

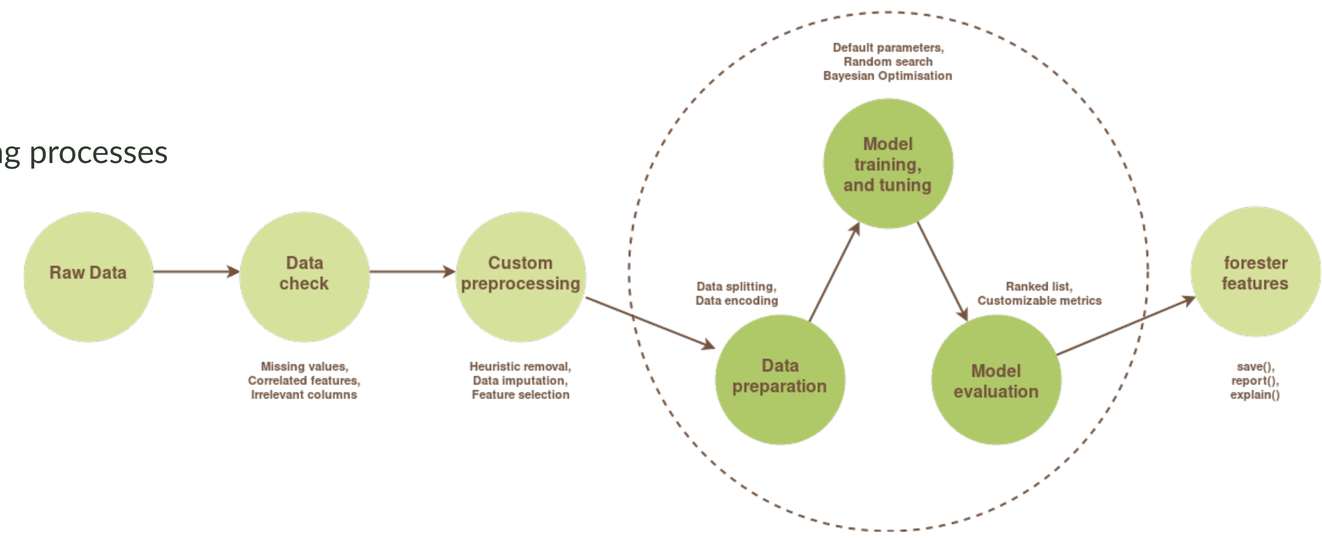
Hubert Ruczyński, Anna Kozak

Warsaw University of Technology

## What is *forester*?

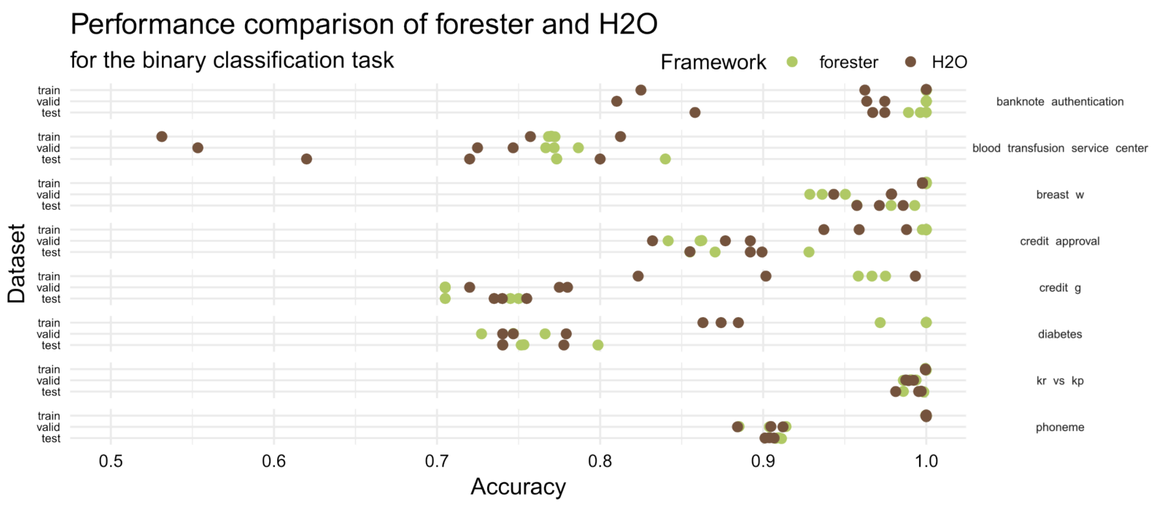
The *forester* package is **an AutoML tool in R** that wraps up all machine learning processes into a single `train()` function, which includes:

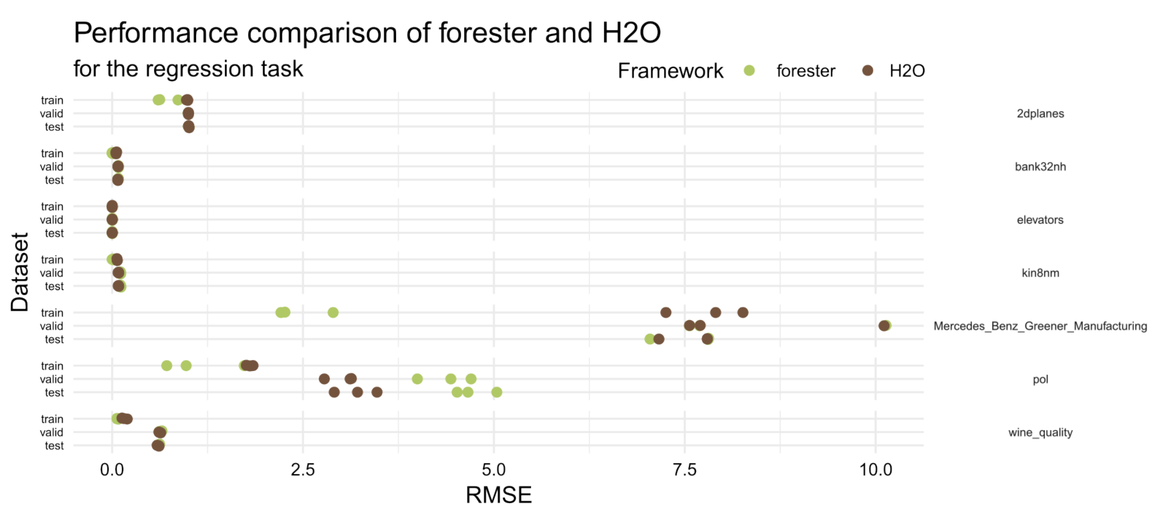
1. rendering a brief **data check** report,
2. **preprocessing** initial dataset enough for models to be trained,
3. **training** 5 tree-based models with default parameters, random search and Bayesian optimization,
4. **evaluating** them and providing a ranked list.




## Simple doesn't mean worse!

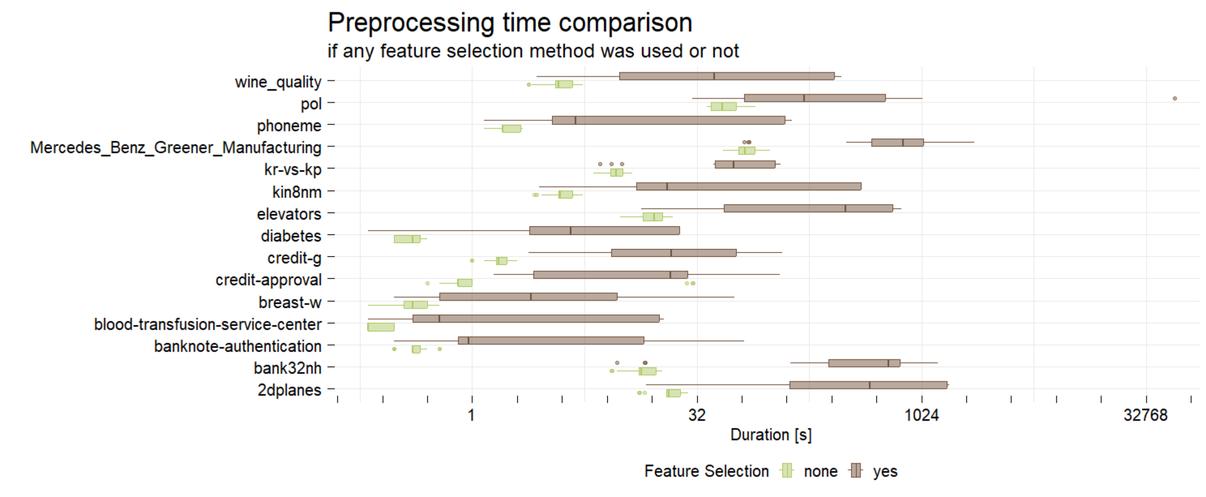
According to our experiments, the forester package achieves competitive results in much shorter time in comparison to well-known H2O AutoML tool. We have compared their performance on 8 binary classification, and 7 regression tasks, and the calculations were repeated 3 times for each dataset and framework.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

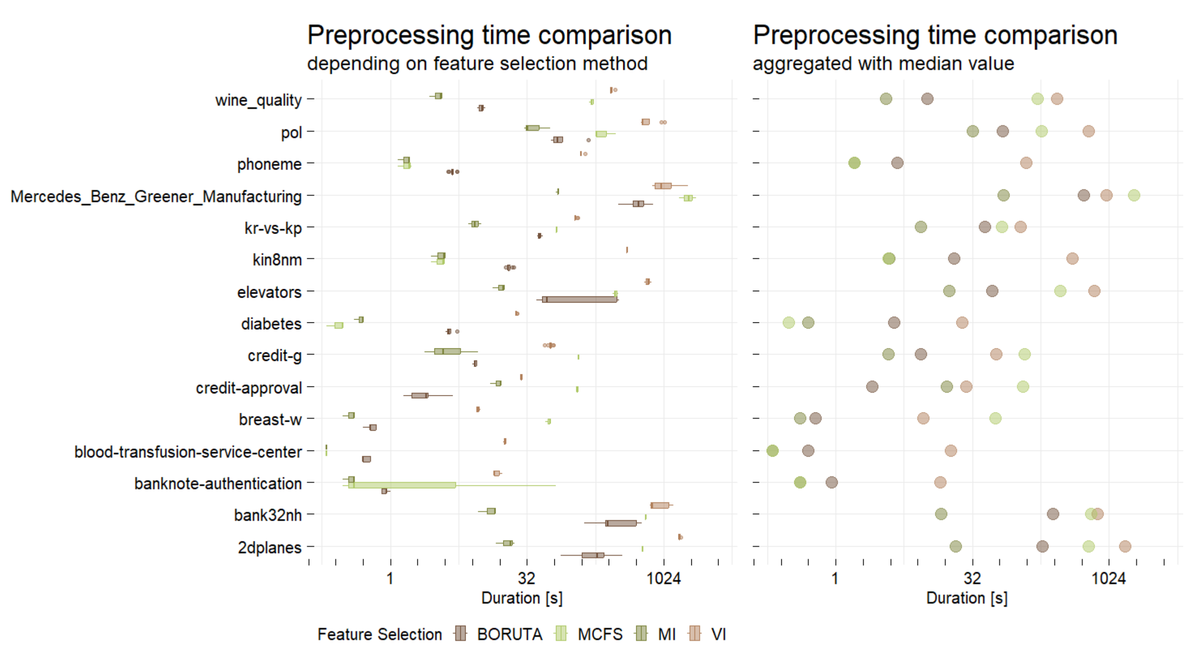



## Do tree-based models need preprocessing?

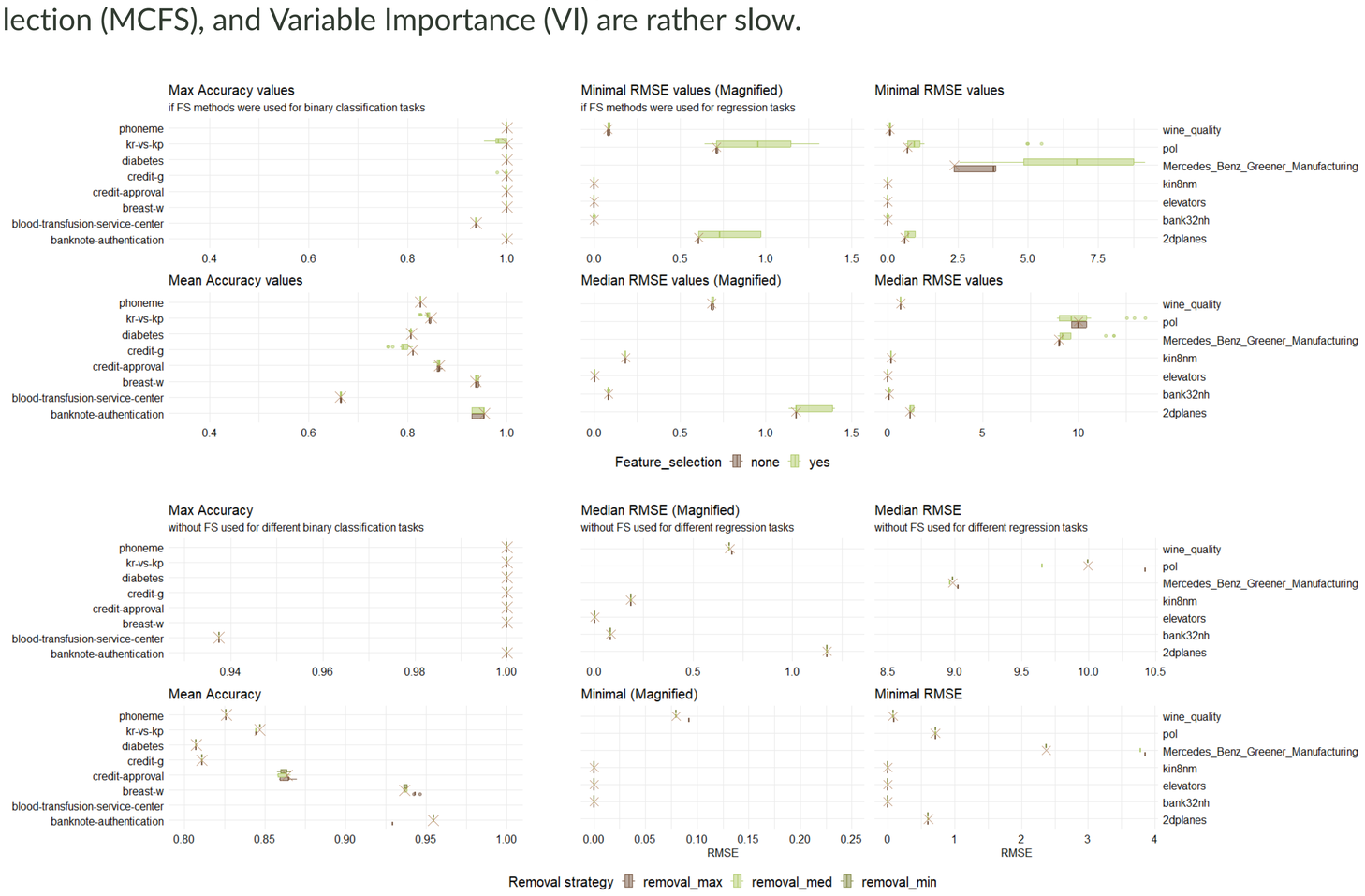
Conducted ablation study used the custom preprocessing module in order to determine how various data preparation approaches impact the final performance of the *forester* on the datasets mentioned before. The analysis focuses on both sheer performance, as well as time complexity of the pipeline.




## Time complexity

The most important findings consider feature selection (FS) methods. They are the most expensive part of preprocessing, and their execution times differ significantly between the methods. The right plot shows us that Mutual Information (MI) based selection method, and BORUTA are relatively fast, whereas Monte Carlo Feature Selection (MCFS), and Variable Importance (VI) are rather slow.




## Feature selection impact on performance

FS methods are responsible for unstable results, and in most cases, its usage leads to worse results than for baseline methods marked with **X**. In some cases however, with FS methods we can obtain better results.
When we consider preprocessing strategies based on heuristic removals, the results are less significant, but in most cases lead to enhancements of the results.

## Contact info

hruczynski21@interia.pl
https://github.com/ModelOriented/forester

## References

A. Kozak and H. Ruczyński. forester: A Novel Approach to Accessible and Interpretable AutoML for Tree-Based Modeling. In *AutoML Conference 2023 (ABCD Track)*, 2023. URL https://openreview.net/forum?id=Q3DWpGoX7PD.