Guide through jungle of models! forester: An R package to automatically select between tree-based models

Anna Kozak, Szymon Szmajdziński, Hoang Thien Ly, Przemysław Biecek

Faculty of Mathematics and Information Science, Warsaw University of Technology

Introduction

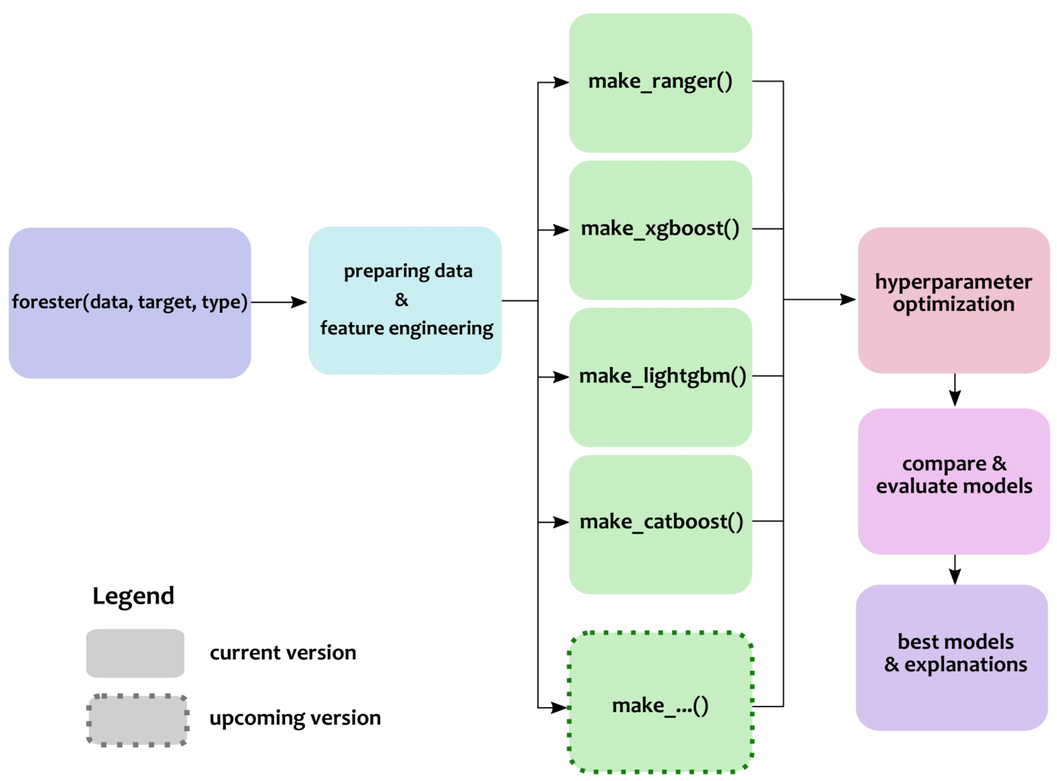
Designing a machine learning model for a specific task is an arduous, time-consuming process. To simplify this process, we introduce the R package **forester** that offers tools to automatically test various tree-based models without pre-processing the data. In particular, we focus on studies in the healthcare domain. The specification of such data is the small data size and dealing with missing data. At the same time, we make sure that our models are responsible by using explanatory machine learning methods - the **DALEX** package. Remember that responsible models first and foremost require interaction between physicians and data scientists.

Benefits of forester package

1. **No requirements for data** - There is no need to create particular object for each model. The package deals with common data structures, such as: data frames, matrices, data tables. It partly performs feature engineering so the users do not have to.
2. **Simple user interface** - One function with three parameters, that is all it takes to create the model.
3. **Automatic hyperparameter optimization** - Tuple of hyperparameters is automatically optimized and selected to train models.
4. **Comparing and selecting best model** - **Forester** package is able to make comparisons between metrics of built models and choose the best one.
5. **Providing explanations** - Explanation plays a crucial role in eliminating reluctance and increasing trust for decision makers while using model's results. With the assistance of **DALEX** package, **forester** enables users to create explanations in both local and global levels.

Package structure

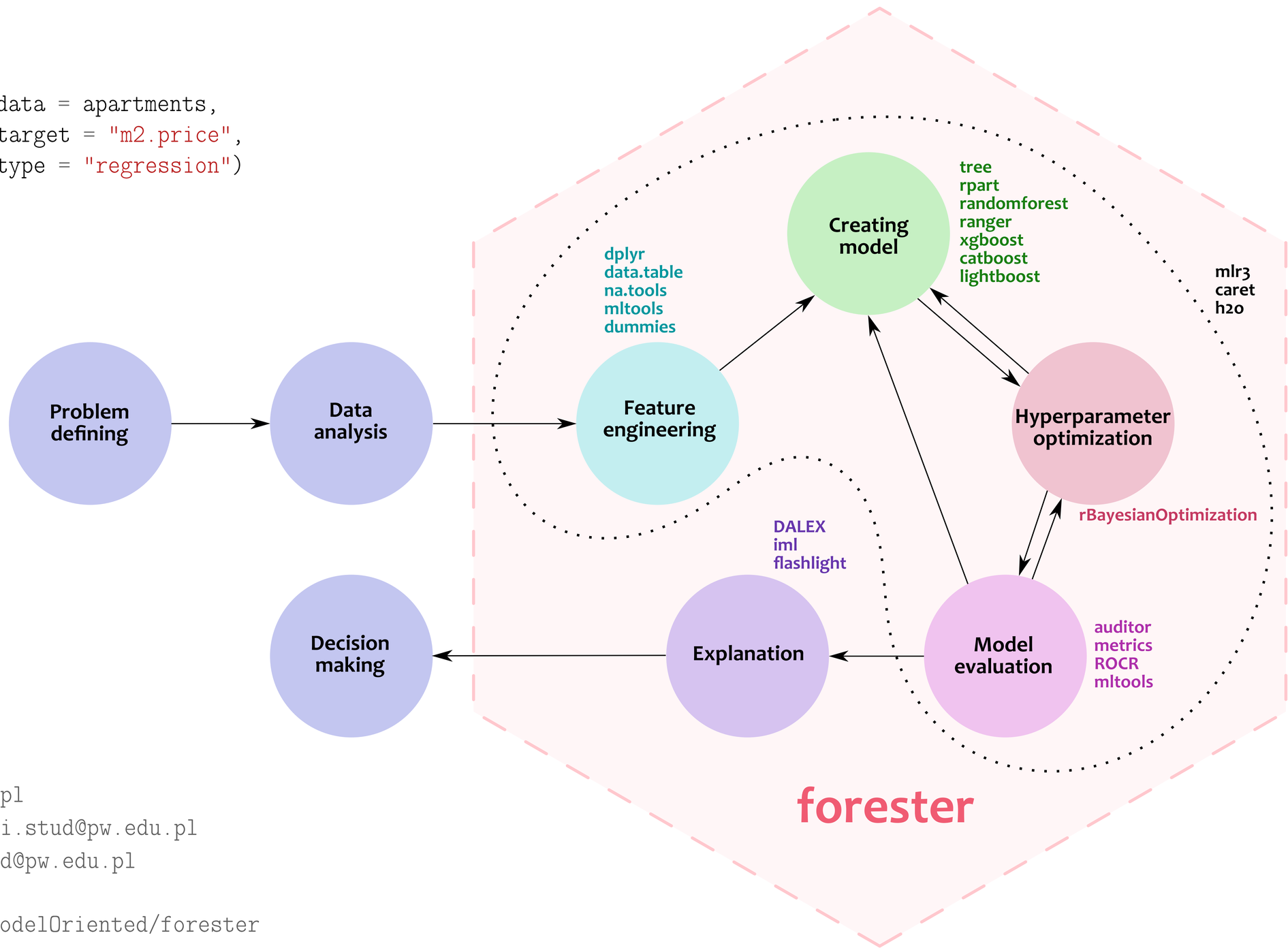
With functions in **forester** package, users can create tree-based model in an unified, simple formula. With only three parameters, user can create various models. First parameter is data with no requirements of pre-processing. After that, user has to specify the name of target column and type of the task, classification or regression. **Forester** automatically does the rest.



Structure of forester package.

```
library(forester)
data("apartments")
best_model <- forester(data = apartments,
                       target = "m2.price",
                       type = "regression")
```

Role of forester package in general Machine Learning pipeline.



Contact info

- ✉ anna.kozak@pw.edu.pl
- ✉ szymon.szmajdzinski.stud@pw.edu.pl
- ✉ hoang_thien.ly.stud@pw.edu.pl

- https://github.com/ModelOriented/forester