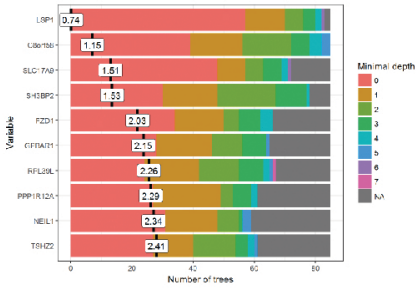
Reg. No: 2017/27/B/ST6/01307; Principal Investigator: dr hab. inż. Przemysław Biecek

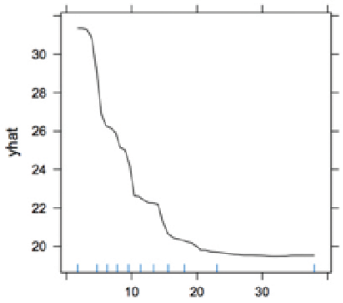
DALEX: Desciptive and model Agnostic Local EXplanations

Global



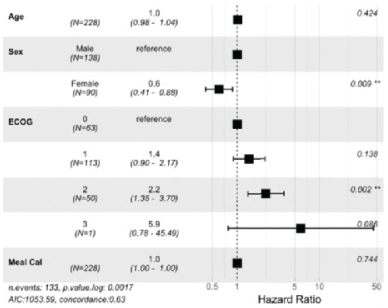
Explains global variable importances through composition plots

Marginal



Explains conditional response for a single variable through partial dependency plots

Local



Explains a single model response through a white-box approximation

Research project objectives

Black boxes are complex machine learning models, for example deep neural network, an ensemble of trees of high-dimensional regression model. They are commonly used due to they high performance. But how to understand the structure of a black-box, a model in which decision rules are too cryptic for humans?

The aim of the project is to create a methodology for such exploration. To address this issue we will develop methods, that: (1) identify key variables that mostly determine a model response, (2) explain a single model response in a compact visual way through local approximations, (3) enrich model diagnostic plots.

Research project methodology

This project is divided into three subprojects - local approximations od complex models (called LIVE), explanations of particular model predictions (called EXPLAIN) and conditional explanations (called CONDA).

The planned approach was pre-testsed by pilot studies [Biecek2017, Dąbrowska2017, Paluszyńska2017, Staniak2017, Sitko2017]. For random forest we have created a package randomForestExplainer [Paluszyńska2017]. Our method explores the structure of a random forest and presents importances of variables. The importance is measured as a gain in the model performance or distribution of depth in trees' ensemble. We will extend this tool to other ensembles of classifiers obtained from boosting or model stacking method. Local explanations focus on a single model prediction. Our preliminary effort is the live package (Local Interpretable Model-agnostic Visual Explanations) [Staniak2017], which extends the LIME method (local approximations by a simple white box).

Marginal explanations are similar, but describe an effect of a single input variable or a pair of input variables. As a cornerstone we will use partial dependence plots [Greenwell 2017] and their adaptations for qualitative variables as implemented in factorMerger [Sitko2017].

Expected impact on the development of science

Explanations of black boxes have fundamental implications for the field of predictive and statistical modelling. The advent of big data forces imposes usage of black boxes that are easily able to overperform classical methods. But the high performance itself does not imply that the model is appropriate Thus, especially in applications to personalized medicine or some regulated fields, one should scrutinize decision rules incorporated in the model. New methods and tools for exploration of black-box models are useful for quick identification of problems with the model structure and increase the interpretability of a black-box.

[Biecek2017] Przemysław Biecek, Marcin Kosiński (2017). archivist: An R Package for Managing, Recording and Restoring Data Analysis Results Journal of Statistical Software. 82 (22)

[Dąbrowska2017] Aleksandra Dąbrowska, Alicja Gosiewska, Przemysław Biecek "MLExpResso: Integrative analyses and visualization" (2017) https://github.com/geneticsMiNIng/MLGenSig/

[Paluszyńska2017] Aleksandra Paluszyńska, Przemyslaw Biecek "BlackBoxOpener: an R package" (2017) https://github.com/geneticsMiNIng/BlackBoxOpener

[Sitko2017] Agnieszka Sitko, Przemysław Biecek "factorMerger: an R package" (2017). Submitted to Journal of Computations and Graphical Statistics https://arxiv.org/abs/1709.04412

[Staniak2017] Mateusz Staniak, Przemysław Biecek. "Local Interpretable (Model-agnostic) Visual Explanations" (2017) https://github.com/MI2DataLab/live

[Ribeiro2016] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. "Why Should I Trust You?: Explaining the Predictions of Any Classifier." (2016). https://arxiv.org/abs/1602.04938

[Greenwell2017] Brandon Greenwell. "pdp: An R Package for Constructing Partial Dependence Plots." (2017) https://journal.r-project.org/archive/2017/RJ-2017-016/index.html