DALEX: How would you explain this prediction?

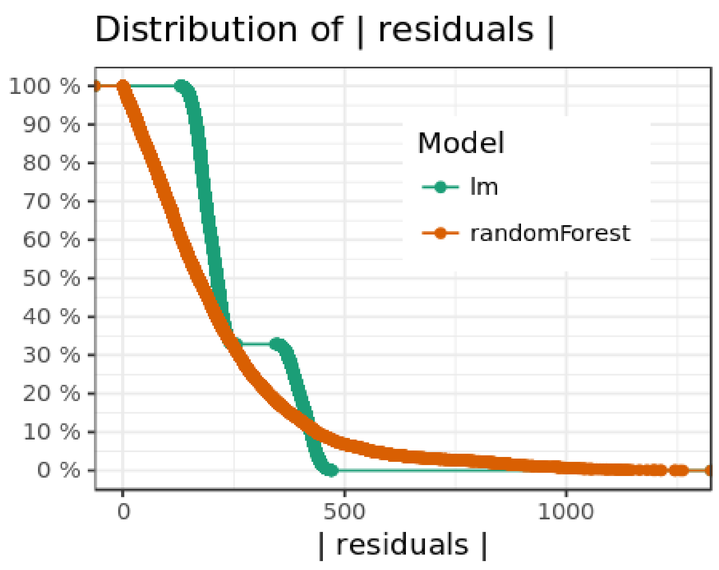
Agnieszka Sitko, Mateusz Staniak, Przemysław Biecek
University of Warsaw | Wroclaw University of Technology, Poland
ag.sitko@gmail.com | mateusz.staniak@math.uni.wroc.pl | przemyslaw.biecek@gmail.com

1. Introduction

Complicated models like neural networks, boosting or bagging have become dominant in terms of their performance, but they mostly remain black boxes. DALEX (Descriptive mAchine Learning EXplanations) is a set of tools that helps to validate, understand and improve complex models. It provides multiple techniques for explaining predictions both locally and globally. Moreover, DALEX is equipped with model-agnostic (inspired by LIME, [1]) methodologies for visual presentation of explanations. DALEX has been used in companies like Disney (platform ESPN+), Trivadis, Gradient for audience segmentation, and KRUK for risk models validation.

2. Model validation

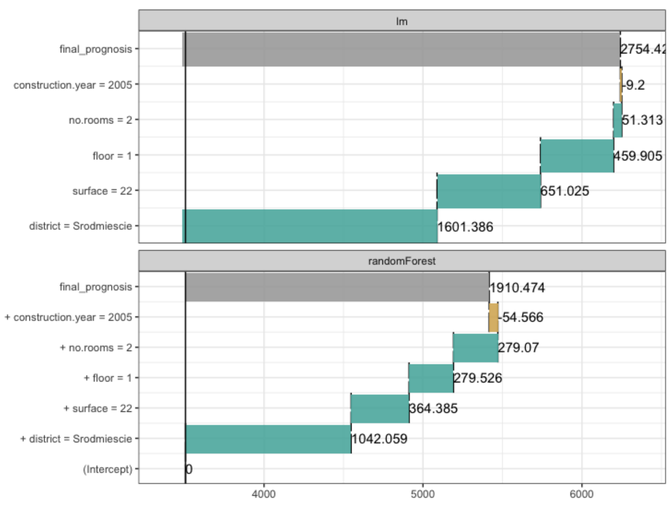
Let's take two models with equal MSE values and analyze their residuals (tail distribution).



Random forest model has in general smaller residuals than linear model, yet a tiny fraction of very large residuals affects the root mean square. Which model is better now?

3. Local explanation

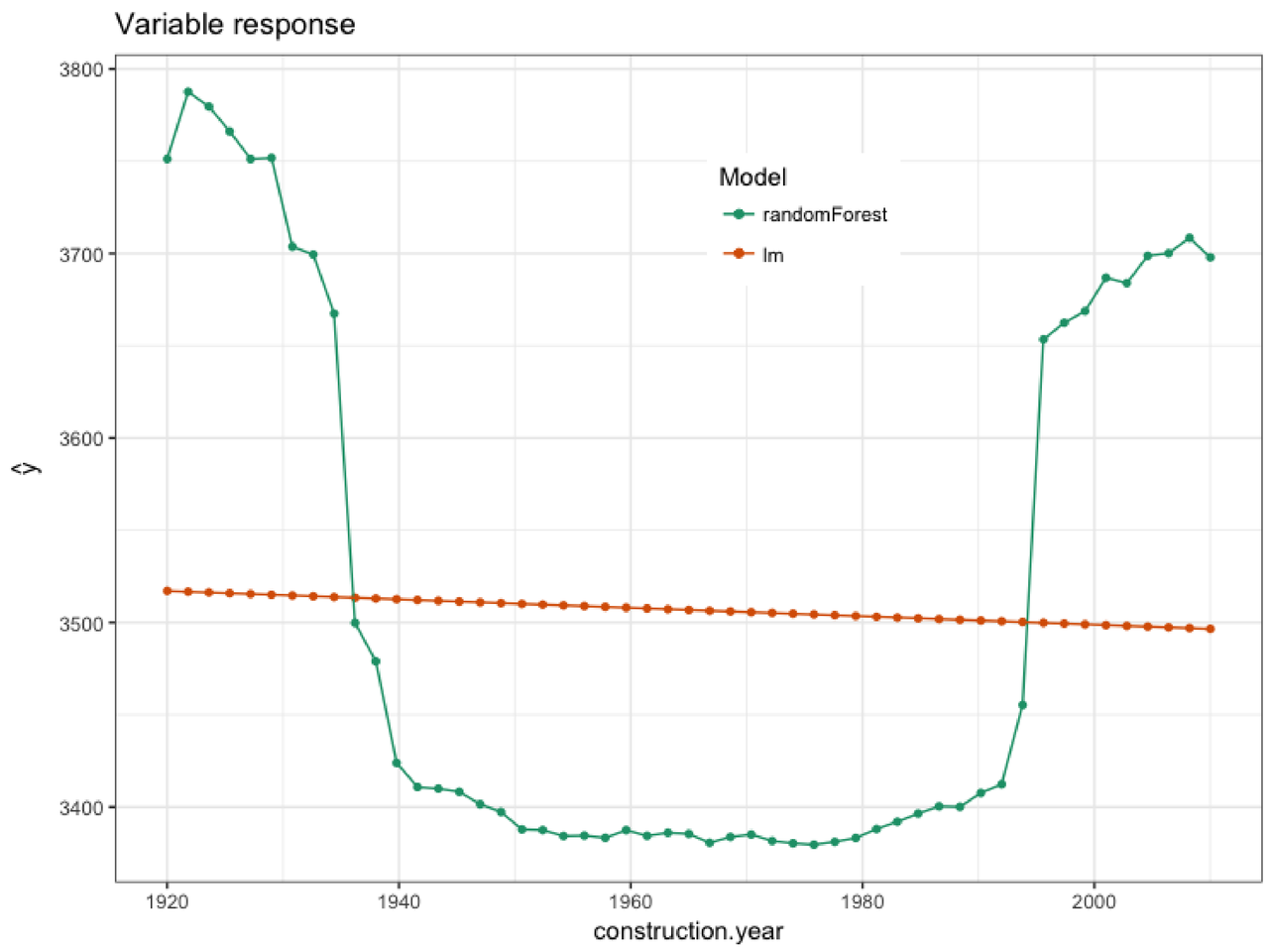
Single prediction explainers are designed to decompose model prediction into parts contributed by separate variables. The model agnostic feature contribution is based on distances to relaxed model predictions. Detailed description may be found in [2].



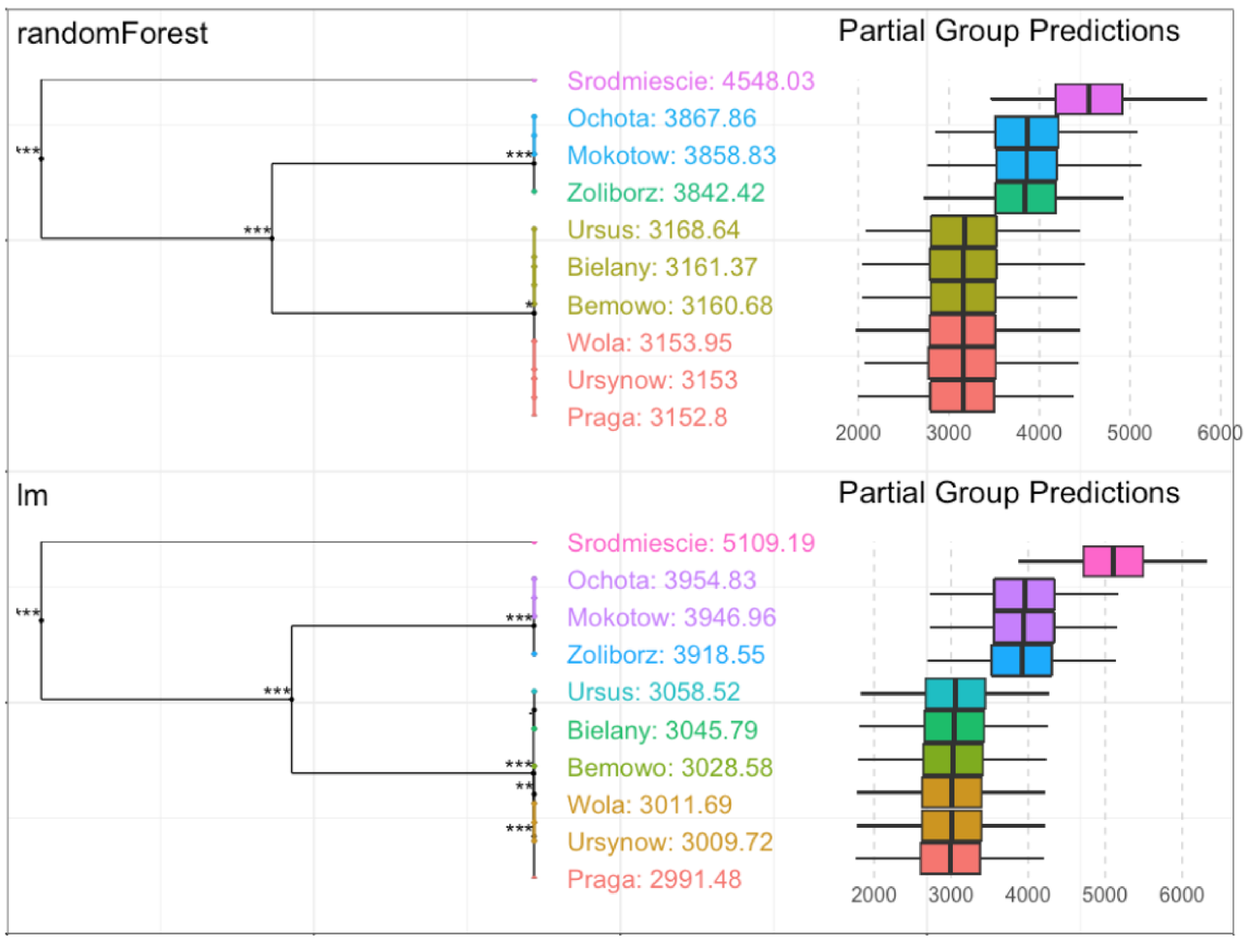
4. Single variable explanations

Single variable explainers are interpolating the conditional effect of a single variable.

We can use Partial Dependence Plot (PDP) to compare two or more models. Below we plot PDP for the linear model against the random forest model. Not surprisingly random forest captures the non-linear relation that cannot be captured by linear models.



Merging Path Plot analyzes model structure and suggests its improvements. In the plot below we observe optimal partition (clusters are represented by colors) of factor levels in terms of model likelihood. Merging procedure is described in [3].



5. Conclusions

DALEX is a set of model-agnostic procedures for model validation, explanation and improvement. It provides a methodology of visualizing complicated relations in a simplified, easy to understand way. DALEX's documentation may be found on Github: https://github.com/pbiecek/DALEX.

6. References

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.

[2] Mateusz Staniak and Przemyslaw Biecek. Explanations of model predictions with live and breakdown packages. 2018.

[3] Agnieszka Sitko and Przemyslaw Biecek. Merge and select: Visualization of a likelihood based k-sample adaptive fusing and model selection. 2017.