# vivo: local variable importance via oscillations

Anna Kozak[1], Przemysław Biecek[1]
[1]Faculty of Mathematics and Information Science, Warsaw University of Technology

## Interpretability

When a model has many features and plotting all one-dimensional summary statistics is troublesome, **vivo** indicates which variables are worth paying attention to. The **vivo** is an R package which calculates instance level feature importance (measure of local sensitivity). The feature importance is based on Ceteris Paribus profiles and can be calculated in a few variants.

## Ceteris Paribus Profiles

Ceteris Paribus is a latin phrase meaning *"other things held constant"* or *"all else unchanged"*. Ceteris Paribus Plots show how the model response depends on changes in a single input variable, keeping all other variables unchanged. They work for any Machine Learning model and allow for model comparisons to better understand how a **black box model** works.

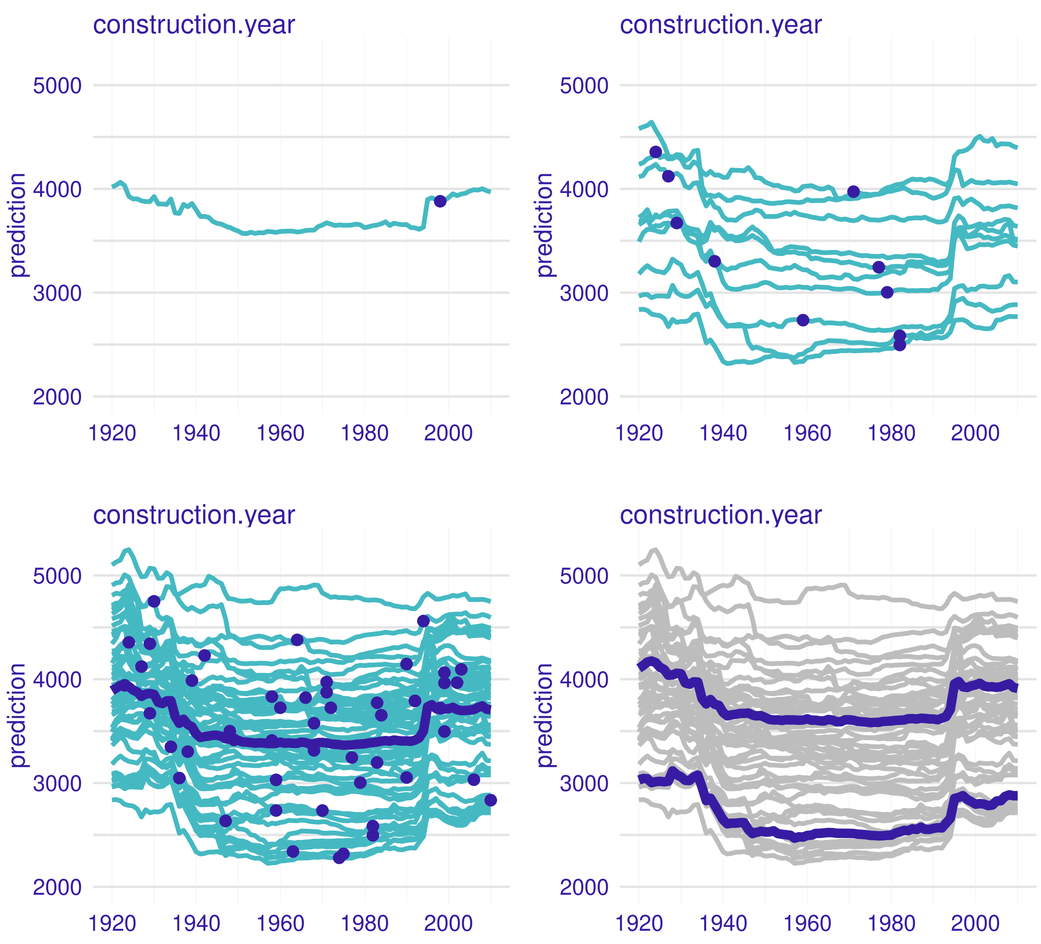



Figure 1: The first plot shows the Ceteris Paribus profile for a single observation. The plot on the right shows the profiles for a few observations. In the lower left corner we have profiles for observations and a line showing their aggregation - a partial dependency plot. The last plot shows the aggregation of profiles using clustering.

## Methodology

The our measure of local variable importance is based on the oscillations of the Ceteris Paribus profiles. In particular, the larger the deviation along the corresponding Ceteris Paribus profile, the larger influence of an explanatory variable on prediction at a particular instance. For a variable that exercises little or no influence on model prediction, the profile will be flat or will barely change.

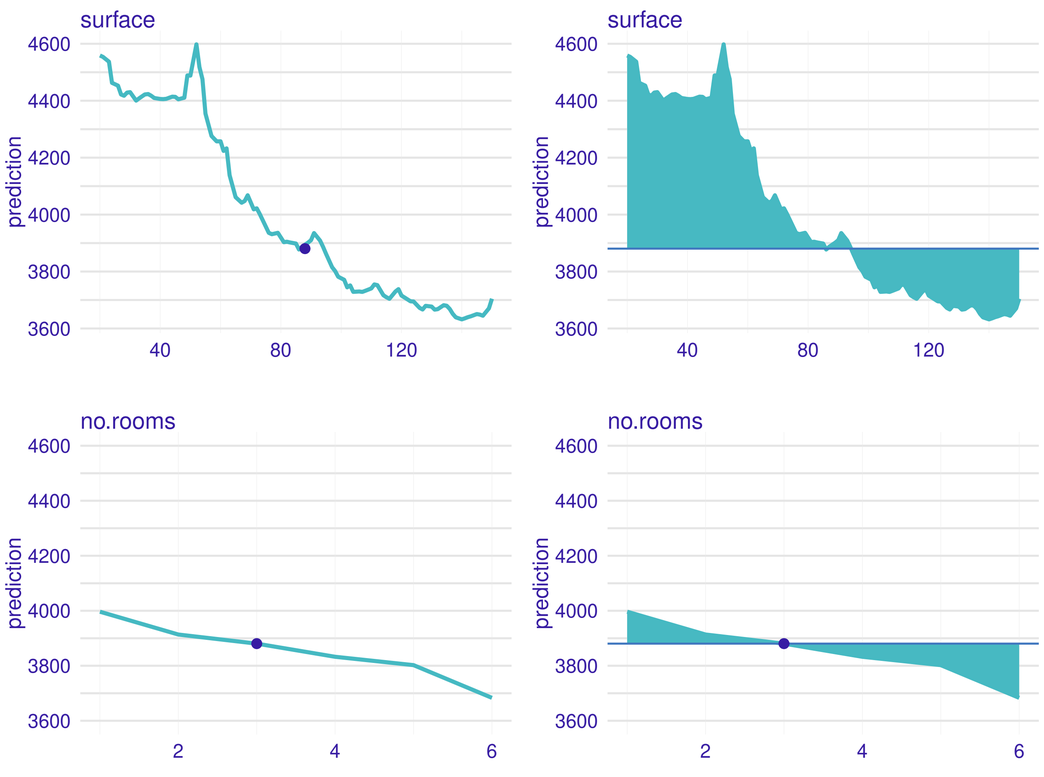
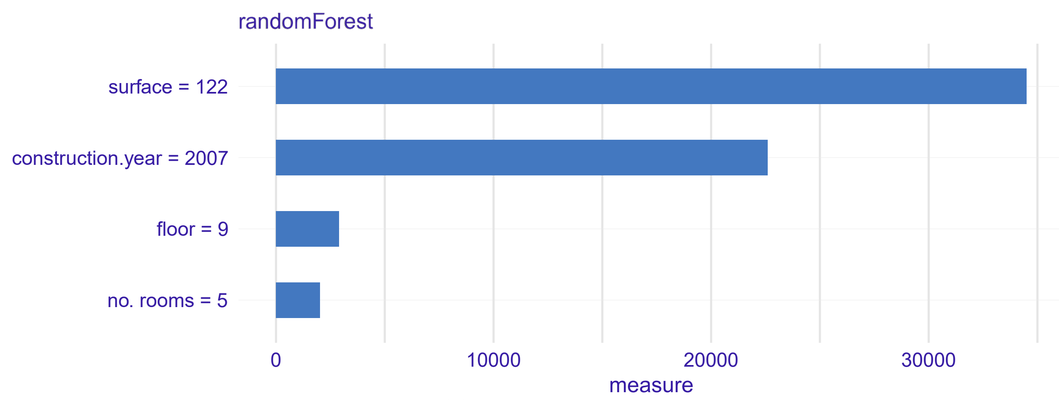


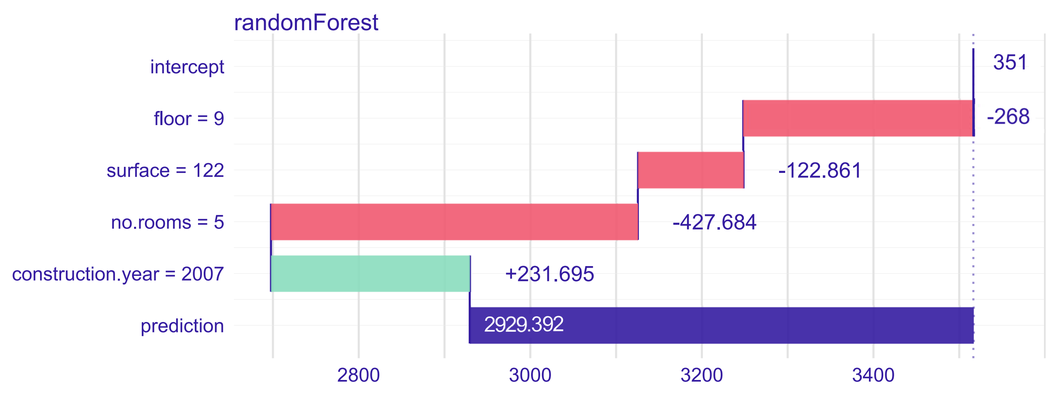

Figure 2: The value of the colored area is our measure. The larger the area, the more important is the variable.

## Comparison of the proposed measure with LIME, iBreakDown and SHAP

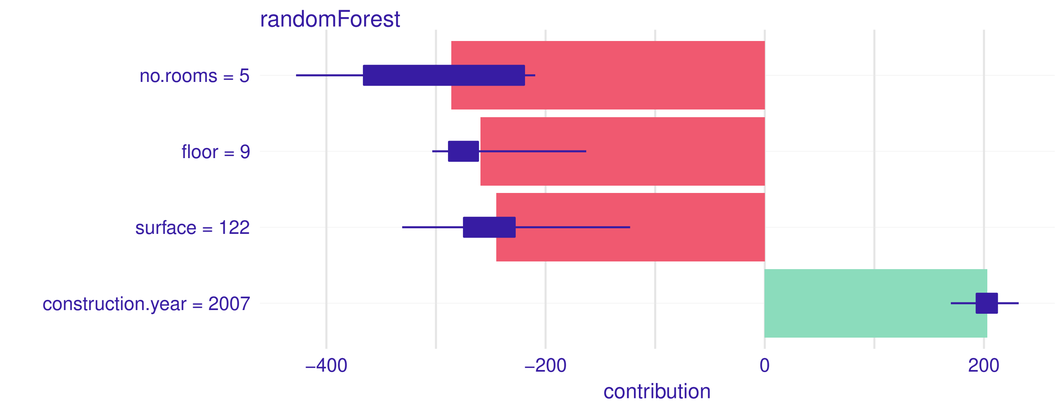
Below is a comparison of methods of local importance of variables based on the black box model - random forest. The vivo, iBreakDown and LIME measures indicate other variables as significant. This only confirms the essence of using various tools when explaining **black box models**.



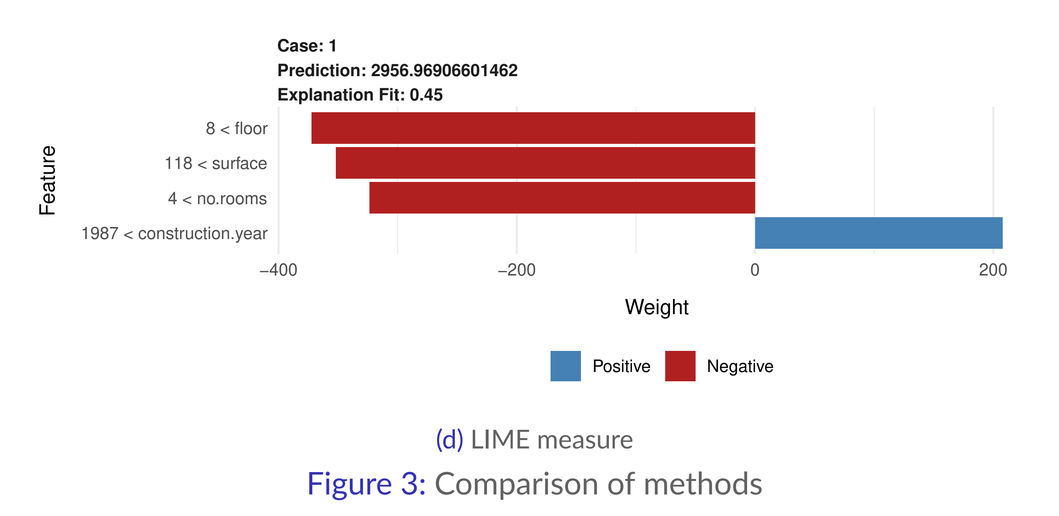

(a) vivo measure




(b) iBreakDown measure




(c) SHAP values




(d) LIME measure

Figure 3: Comparison of methods

## Details

Oscillations of Ceteris Paribus profiles are easy to interpret and understand. By using the average of oscillations, it is possible to select the most important variables for an instance prediction. This method can easily be extended to two or more variables.

## References

[1] Przemysław Biecek. *ingredients: Effects and Importances of Model Ingredients*, 2019. URL https://cran.r-project.org/web/packages/ingredients/index.html.

[2] Przemysław Biecek and Tomasz Burzykowski. *Predictive Models: Explore, Explain, and Debug*. 2019. URL https://pbiecek.github.io/PM_VEE/.

[3] Alicja Gosiewska and Przemyslaw Biecek. iBreakDown: Uncertainty of Model Explanations for Non-additive Predictive Models, 2019. URL https://arxiv.org/abs/1903.11420v1.

[4] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. *Why Should I Trust You?: Explaining the Predictions of Any Classifier*, page 1135–1144. ACM Press, 2016. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778. URL http://dl.acm.org/citation.cfm?doid=2939672.2939778.