Local Explanations of Complex Machine Learning Models

Mateusz Staniak[1], Michał Kuźba[1, 2], Przemysław Biecek[1,2]

[1]Faculty of Mathematics and Information Science, Warsaw University of Technology

[2]Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw

Why explain machine learning models?

From online ad targeting to aiding medical and legal decision making, complex machine learning models such as deep neural networks are pervasive in the digital era. However, decision made by the black box models can't be justified and explained, which often makes them untrustworthy. In this poster, we present methods that help explain decision made by machine learning models. Each of these methods is implemented as an R package that belongs to the **DALEX**[2] family of packages. Python versions are available for **Break Down** (`https://github.com/MI2DataLab/pyBreakDown`) and **Ceteris Paribus** (`https://github.com/ModelOriented/pyCeterisParibus/`).

How features impact a particular prediction?

Complex machine learning models can learn nonlinear relationships with many interactions. For such models, same feature can have a different influence on different predictions and there are usually no built-in methods of providing feature importance.
LIVE package[4] can be used to approximate a black box model around a single prediction with a simple model like linear regression to obtain feature effects. This approach is based on [5].

Variable		N	Estimate		p
age		5000		354,62 (351,40, 357,84)	<0,001
sex	female	4592		Reference	
	male	408		-1715,91 (-1826,76, -1605,06)	<0,001
bmi		5000		-145,51 (-155,47, -135,56)	<0,001
children		5000		-137,38 (-183,28, -91,47)	<0,001
smoker	no	4825		Reference	
	yes	175		10966,38 (10803,93, 11128,84)	<0,001
region	northeast	219		Reference	
	northwest	4371		1134,91 (988,52, 1281,30)	<0,001
	southeast	209		609,45 (408,43, 810,48)	<0,001
	southwest	201		-1073,33 (-1276,39, -870,27)	<0,001
(Intercept)				-2598,40 (-2929,28, -2267,52)	<0,001

Figure 1: The presented plot is a visual representation of a linear model fitted by `live` package: coefficients are plotted against features to show, which of them have a strong influence on the prediction.

How does a single feature influence the prediction?

Ceteris Paribus profiles convey more information about the relationship between a feature and the response than variable effects provided as a single score. They show how the prediction would be affected if we changed a value of one feature while keeping other features unchanged.

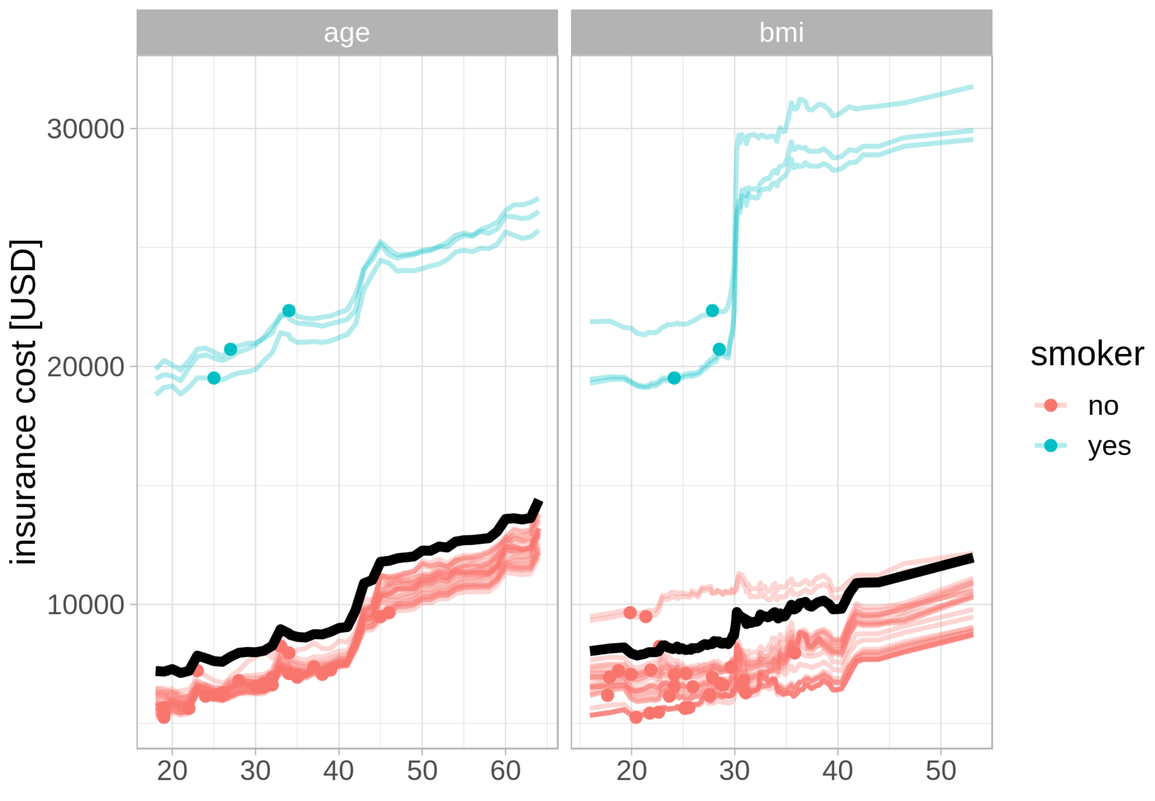

Figure 2: Ceteris Paribus plots describing the effect of BMI on the price of insurance. Much higher price is estimated for smokers. The shape of this relationship also strongly varies between smokers and non-smokers, indicating an interaction between BMI and smoking.

How to find interaction between the features?

The **Break Down** package allows user to find local interactions. This method decomposes the prediction into scores that can be attributed either to a single feature (additive effect) or to two features (interaction). Computational details can be found in [4].

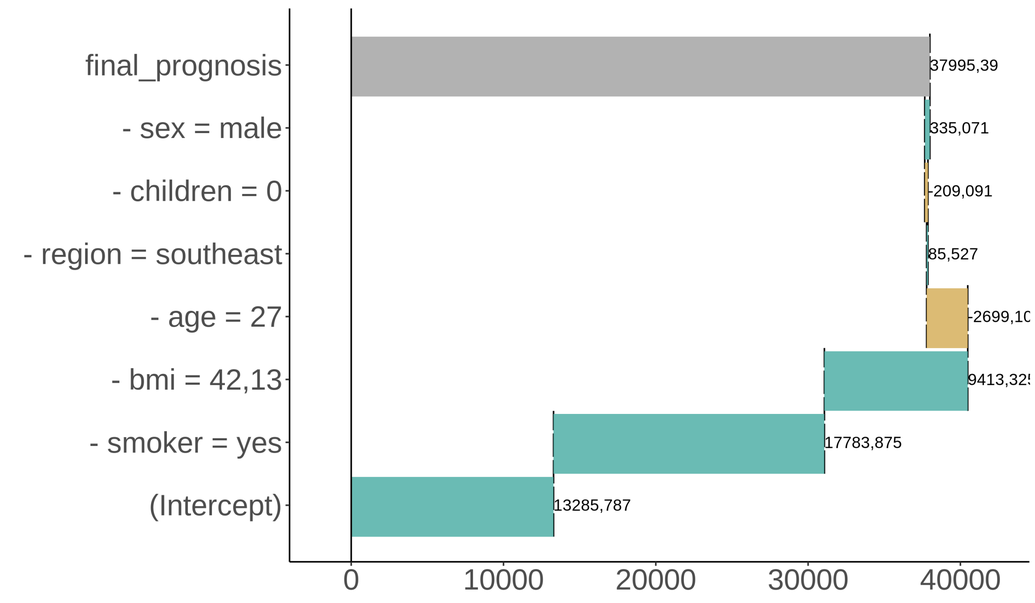

Figure 3: Break Down plot presents how prediction is decomposed into a sum of scores. Color and labels indicate if the influence is positive or negative.

How well does the model fit locally?

Wangkardu plots, based on Ceteris Paribus profiles, help discover issues with stability and fit by plotting the ground truth and Ceteris Paribus curves for observations most similar to the explained one. Large differences from actual response indicate poor fit while variation in profiles indicates lack of stability.

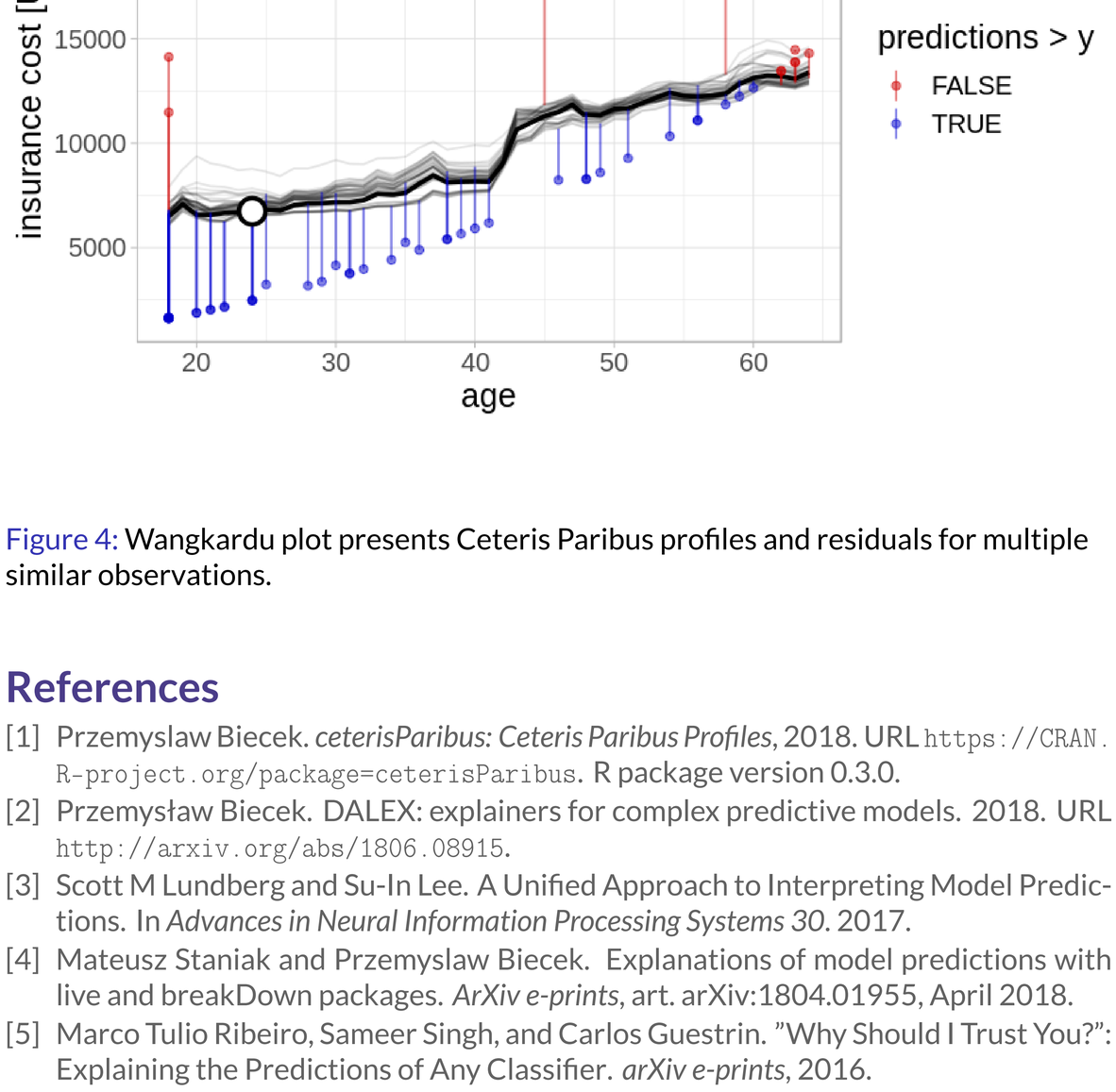

Figure 4: Wangkardu plot presents Ceteris Paribus profiles and residuals for multiple similar observations.

References

[1] Przemyslaw Biecek. *ceterisParibus: Ceteris Paribus Profiles*, 2018. URL `https://CRAN.R-project.org/package=ceterisParibus`. R package version 0.3.0.

[2] Przemysław Biecek. DALEX: explainers for complex predictive models. 2018. URL `http://arxiv.org/abs/1806.08915`.

[3] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*. 2017.

[4] Mateusz Staniak and Przemyslaw Biecek. Explanations of model predictions with live and breakDown packages. *ArXiv e-prints*, art. arXiv:1804.01955, April 2018.

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv e-prints*, 2016.

Acknowledgements

This work was financially supported by NCN Opus grant 2016/21/B/ST6/0217.

- ✉ `m.staniak@mini.pw.edu.pl`
- ✉ `michal.kuzba@students.mimuw.edu.pl`