Local Interpretability of Machine Learning Models

Mateusz Staniak,
Hasselt, 12 XII 2018

What is Interpretable Machine Learning?

- New (and growing) area of research:
  - First work: IME and EXPLAIN (Robnik-Sikonja, 2010),
  - Breakthrough: LIME (Tulio Ribeiro, 2016),
  - Diverse new research: explanations based on
    - game theory,
    - approximations and (mathematical/real) analysis,
    - asking questions,
    - surrogate models,
    - and more.
- Also known as Explainable Artificial Intelligence (xAI).

Different faces of IML

- Building explainable methods.
- Model exploration & maintenance.
- Explaining (post-hoc) *black box* models.
- Knowledge extraction from complex models.

Motivation & basic of IML

IBM

IBM Watson Health | Life sciences | Oncology | Value-based care | Government | Imaging | Blog

Watson for Oncology: 90% concordance with tumor board recommendation

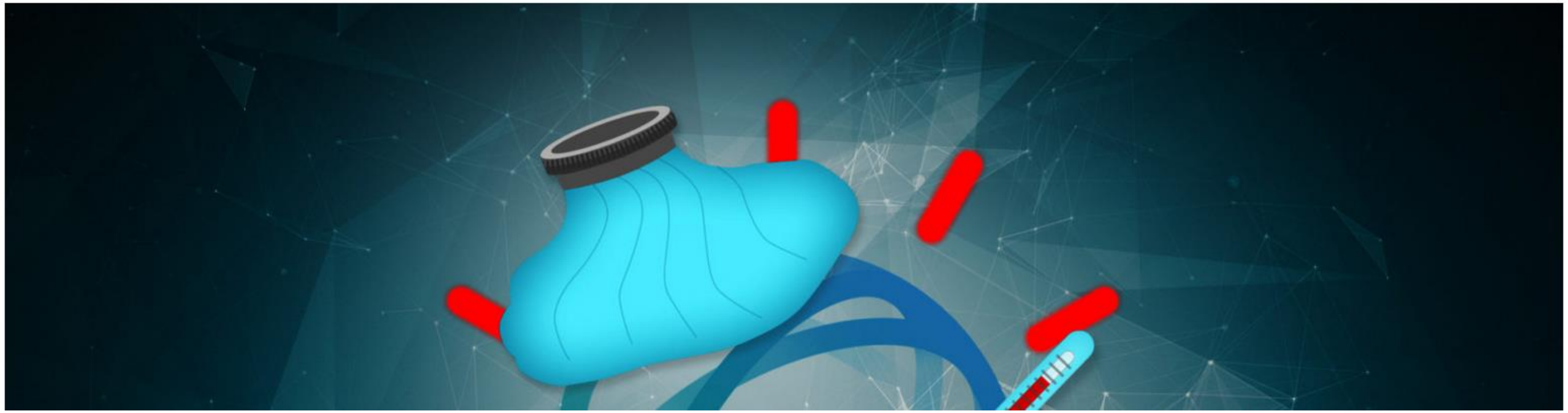
https://www.youtube.com/watch?v=fAiRqM44hgM

EXCLUSIVE

STAT+

IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show

By CASEY ROSS @caseymross and IKE SWETLITZ @ikeswetlitz / JULY 25, 2018



https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/

Machine Learning Failure: Amazon Scraps Biased Recruiting Tool

October 11, 2018 by Jeffrey Dastin

PRINT EMAIL

REUTERS

Amazon.com machine-learning specialists uncovered a big problem: Their new recruiting engine did not like women.

The team had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters.

Automation has been key to Amazon's e-commerce dominance, be it inside warehouses or driving pricing decisions. The company's experimental hiring tool used artificial intelligence to give job candidates scores ranging from one to five stars—much like shoppers rate products on Amazon, some of the people said.

"Everyone wanted this holy grail," one of the people said. "They literally wanted it to be an engine where I'm going to give you 100 resumes, it will spit out the top five, and we'll hire those."

Gender bias was not the only issue. Problems with the data… meant that unqualified candidates were often recommended for all manner of jobs.

https://www.carriermanagement.com/news/2018/10/11/185221.htm

Problems & questions in IML

- How well does the model perform? (**Model performance**)
- Which variables are most important in the model? (**Feature importance**)
- What is the relationship between predictors and response? (**Variable response** / effect)
- What factors drive a particular prediction? (**Local explanations**)
- Does the model discriminate against some group? (**Fairness**)

Types of explanations

- Intrinsic vs **post-hoc.**

- Model-specific vs **model-agnostic.**

- Global vs **local** (model-level vs instance-level).

Global vs local explanations

- Local explanations are concerned with a single observation and its prediction.
- Global explanations are concerned with the model as a whole.
  - Example: decomposition of prediction into sum of scores is a local explanation. Variable importance is a global explanation.
  - Global explanations are often aggregation of local explanations (e.g. mean).

Model-agnostic approach

- Explanations that use knowledge about the specific model (e.g. algorithm) are called model-specific.
- Explanation that make no assumptions about the model structure are called model-agnostic. They work for any model and only require *predict* interface. In particular, such explanations
  - do not require re-fitting the model,
  - can be used with model ensembles, model-stacking etc.
- Example:
  - average minimum depth in a random forest as a variable importance measure (model-specific),
  - permutation variable importance (uses only column permutation and predict interface – model-agnostic).

Local Explanations in the DALEXverse

Joint work with Przemysław Biecek

Variable Explainers
package: pdp, ALEPlot, **factorMerger**

Structure Explainers
package: randomForest

Model Management
package: **archivist**

Model Diagnostic Tools
package: **auditor**, ggfortify

Model Performance Explainers
package: **auditor,** ROCR, caret, mlr

Model Predictions Explainers
package: **breakDown**, **live**, shapleyr, lime

13

https://github.com/pbiecek/DALEX_docs/blob/master/workshops/eRum2018/Workshop_eRum_2018_part1.pdf

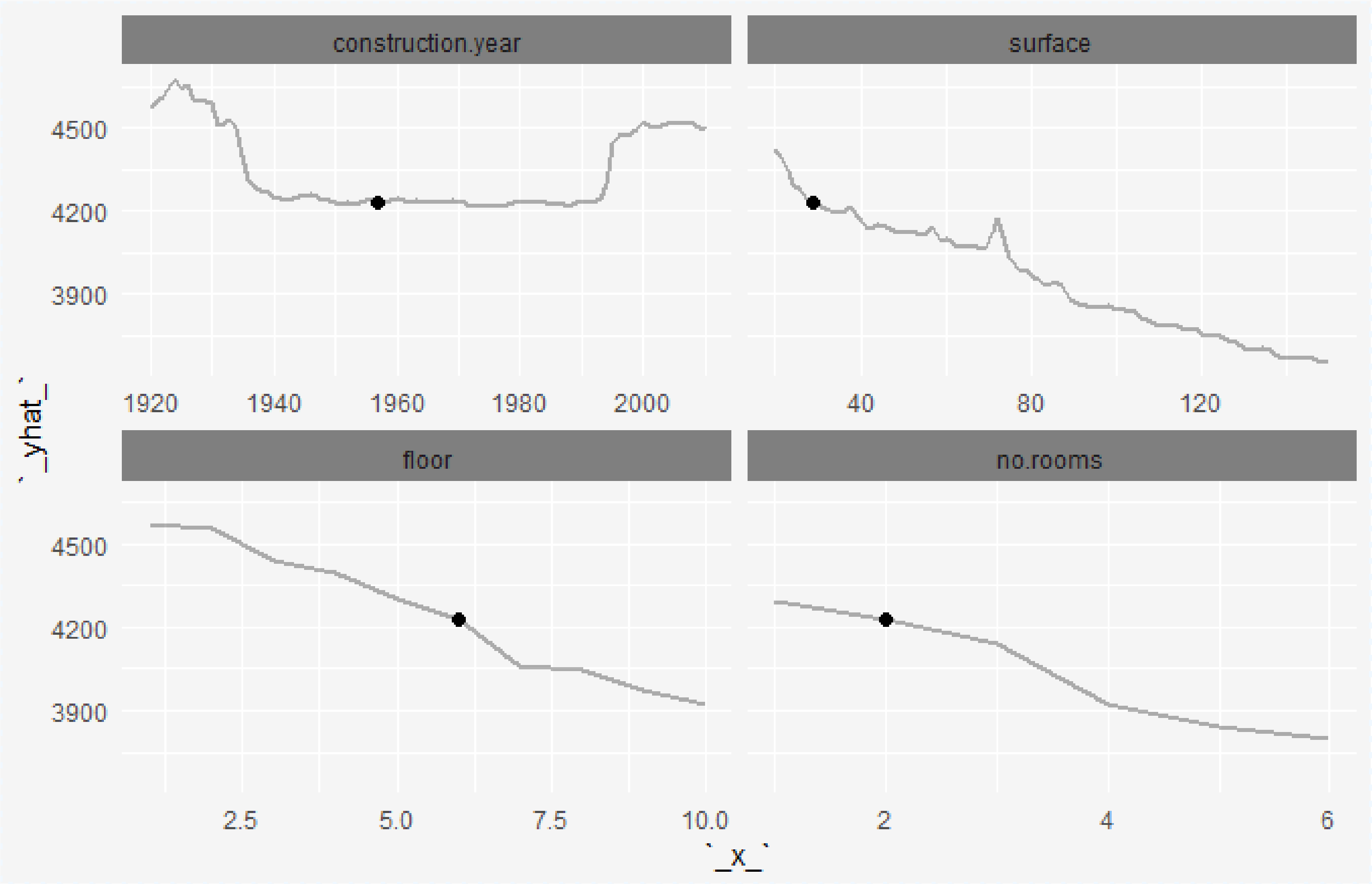
Approaches to local explanations

- What-If analysis (marginal response of the model when changing a single variable for a single observation):
  - Ceteris Paribus profiles.
- Local surrogate models (aka LIME – fitting an interpretable model locally):
  - LIME and its modifications (aLIME, k-LIME, localSurrogate),
  - LIVE.
- Prediction decomposition (attributing additive scores to features).
  - IME,
  - Break Down,
  - Shapley Values.

Ceteris Paribus

- Plots of alternative scenarios (how to change the model decision?)
- We draw a relationship between model response and a single variable, while keeping values of other features unchanged.
- Also know as ICE.

Ceteris Paribus example



Ceteris Paribus plots…

- … are local PDP (point-wise average of CP profiles is PDP).
- … shows, how much a variable must change to change the model outcome.
- … allows us to investigate model stability and local quality of the fit (after adding ground truth to the plot).

Ceteris Paribus profiles

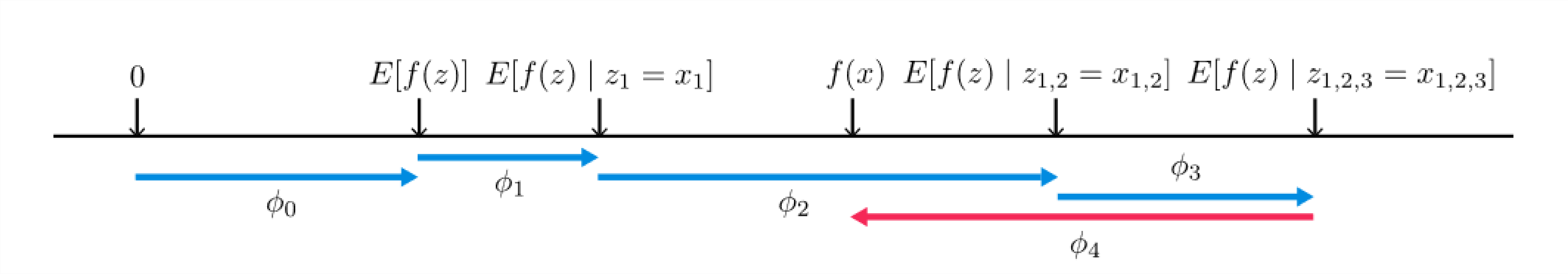
- … are local PDP (point-wise average of CP profiles is PDP).
- … shows, how much a variable must change to change the model outcome.
- … allows us to investigate model stability and local quality of the fit (after adding ground truth to the plot).

Explanations of model predictions with live and breakDown packages

- recently accepted to The R Journal

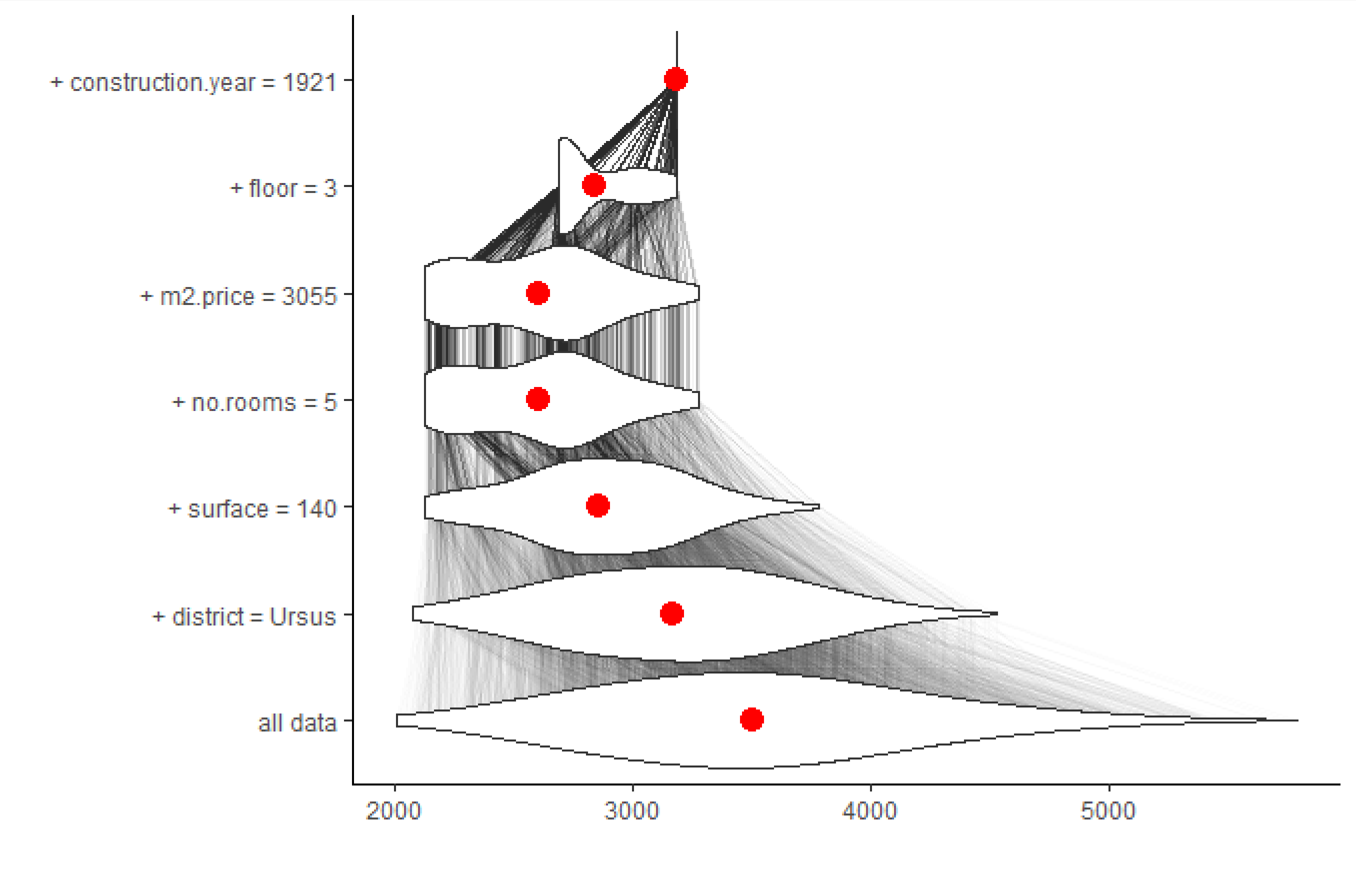
Prediction decompositions

- Different ideas:
  - non-sequential conditioning – IME,
  - sequential condititiong:
    - Break Down: single path
    - Shapley values: average over different paths



http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

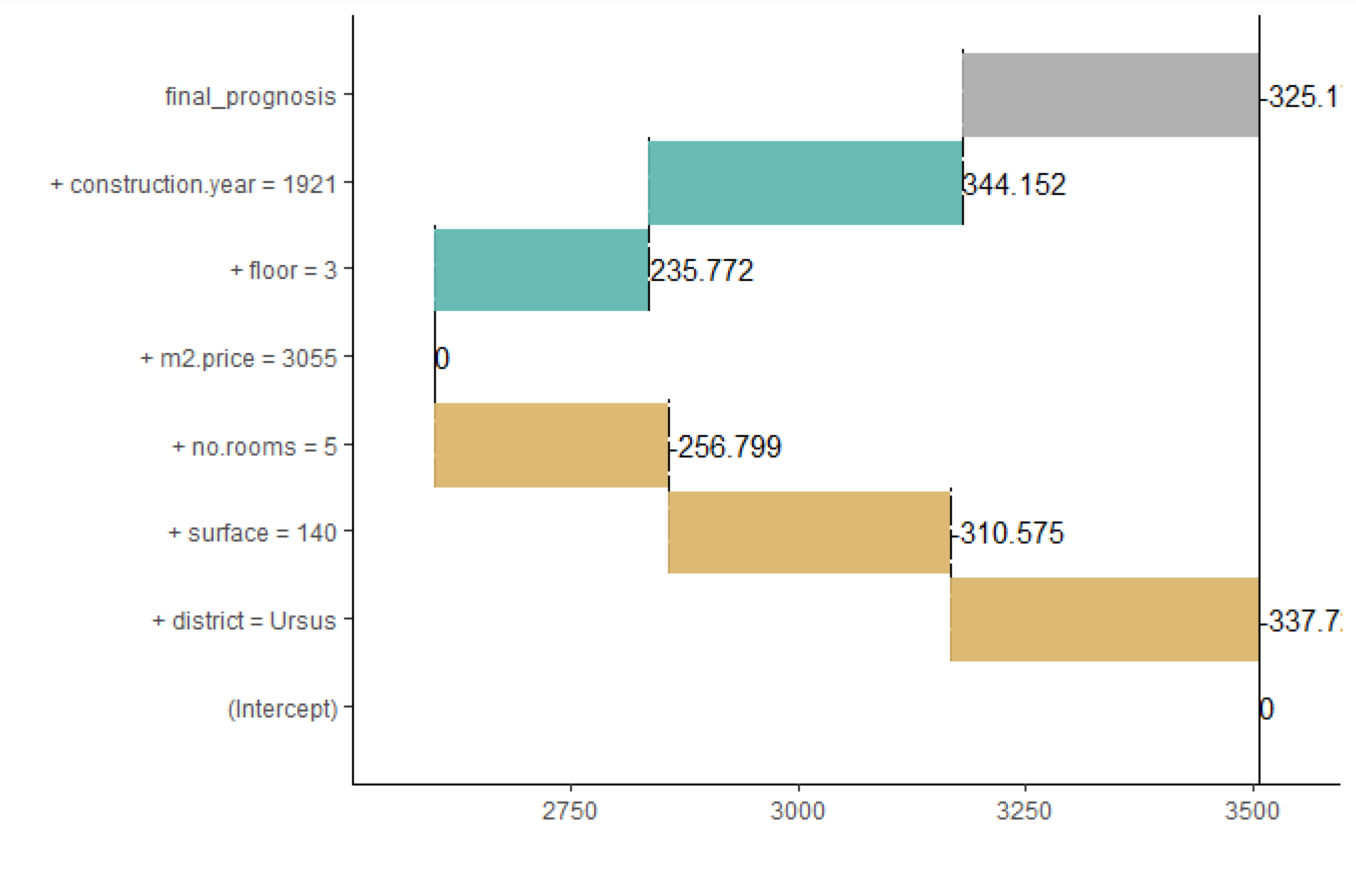
Break Down intuition



Break Down algorithm

1. We start with the average model prediction.
2. We fix a value of each variable (as a candidate).
3. We choose the variable which resulted in the biggest change.
4. We repeat the procedure for the other variables until all values are fixed.

Break Down example



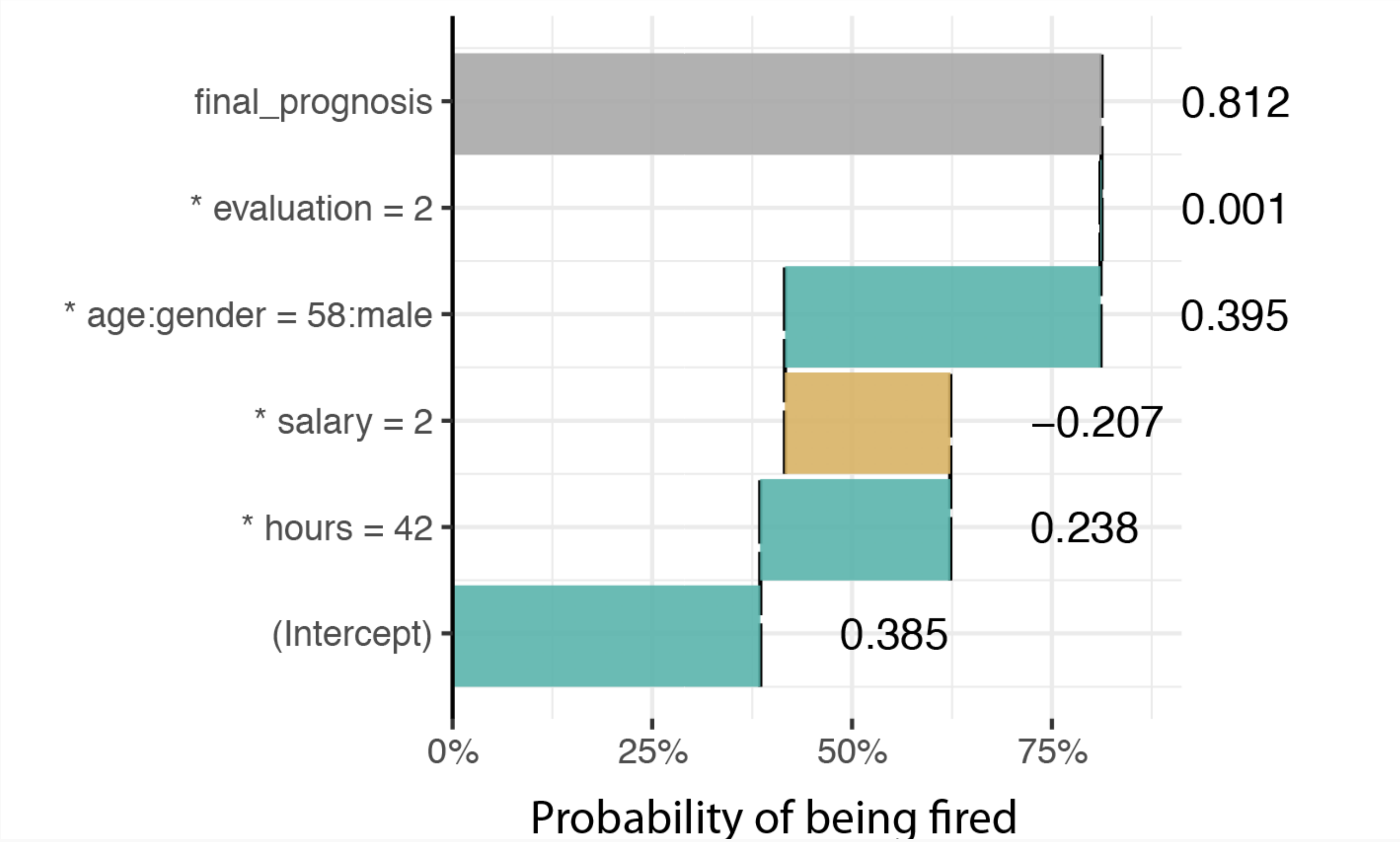
Break Down with interactions

- Break Down can be used to detect local interactions.
- Joint effect of two features can be compared to their additive effects:

$$score_2(f, x^*, (i,j)) = \left| E[f(X)|X_i = x_i^*, X_j = x_j^*] - E[f(X)|X_i = x_i^*] - E[f(X)|X_j = x_j^*] + E[f(X) \right.$$

- Features of pairs of features are chosen such that each feature occurs only once in the resulting explanation.

Break Down with interactions



Taken from ***Predictive Models: Visualisation, Exploration and Explanation With examples in R and Python***
Przemyslaw Biecek and Tomasz Burzykowski

Local Explanations: applications

An application of explainer methods breakdown and Ceteris Paribus to understand statistical models built on lung cancer data

Autor

November 8, 2018

Abstract

Advanced machine learning algorithms called black box models are more and more common in predictive analytics. They are easy to build and are a good and robust tool to work with large databases. Contrarily, they may be hard to understand when applied to a specific observation. This article presents two methods, which allow to find and present visually how much a single feature contributes to a final survival prognosis of a patient and how a change in each feature would affect the model response.

K. Kobylińska, M. Adamek, P. Biecek

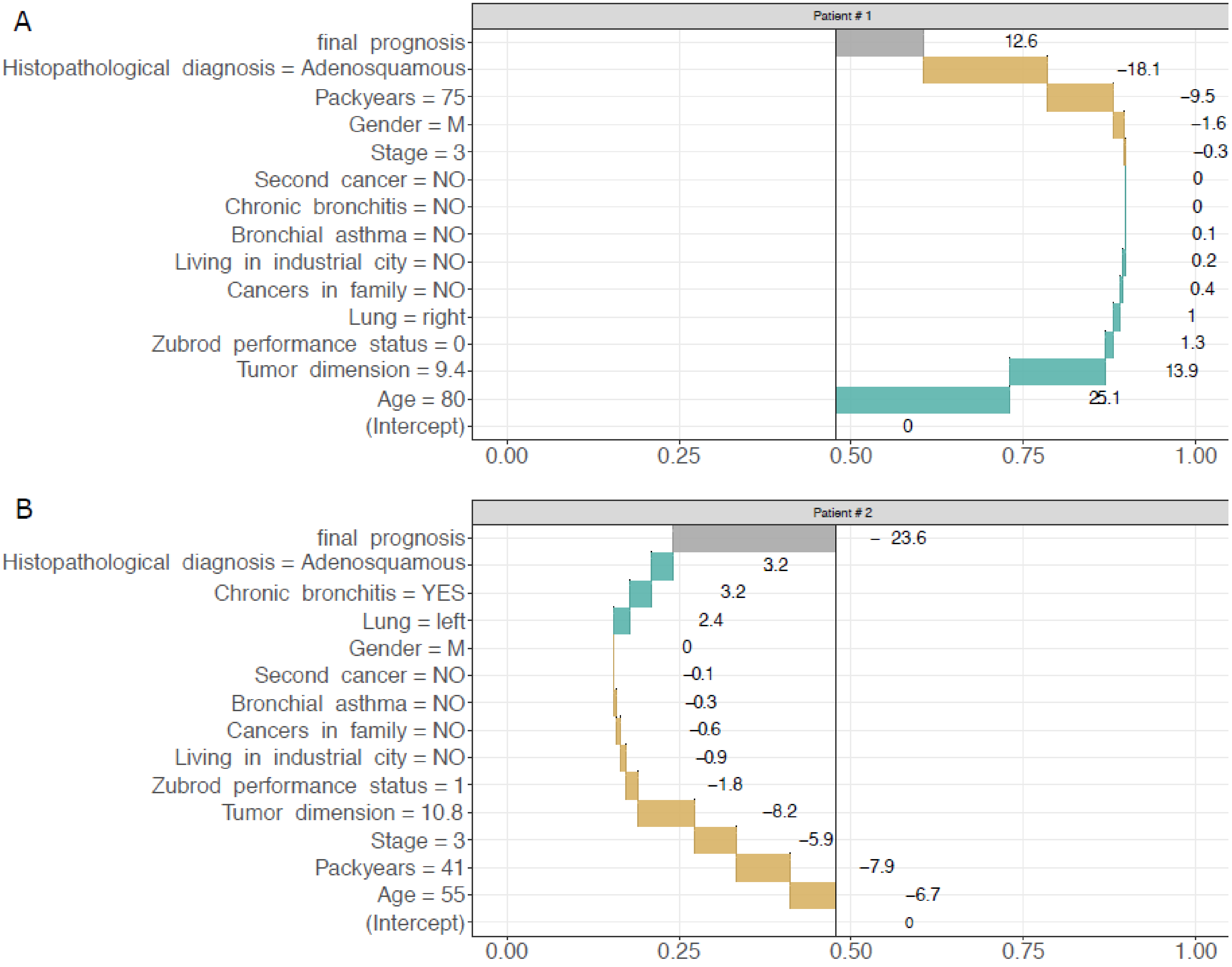
Local Explanations: applications

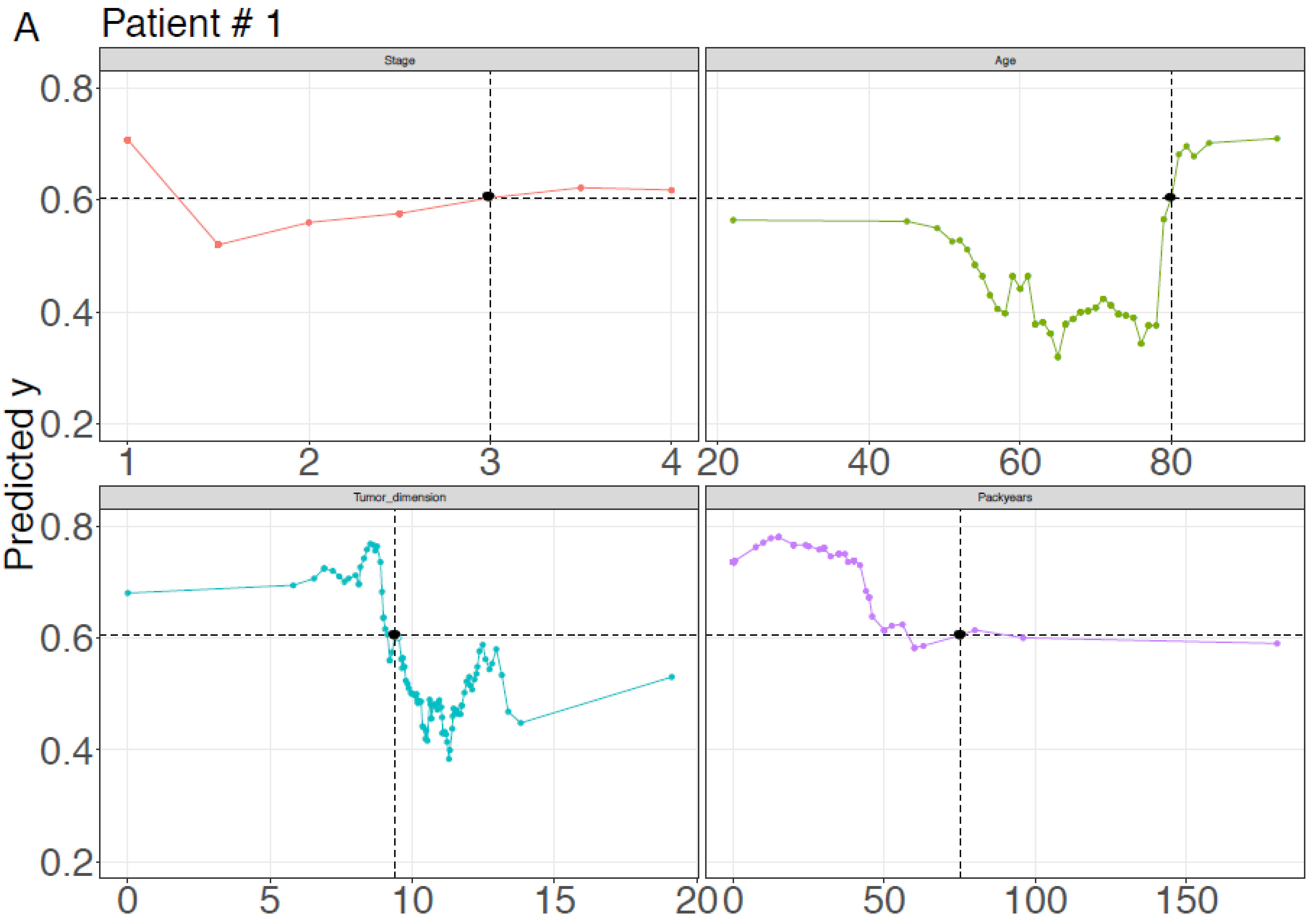
- Unique dataset: all cases of operable lung cancer in Poland from period of 12 years (35445 patients), 337 variables (risks, previous surgeries, histopathological descriptions, pre-surgical evaluation, etc.)
- The goal: prediction of 3-year survival - counting from the beggining of the treatment - based on features available at that moment.
- Two models were compared:
  - Logistic regression – AUC 0.69
  - Random forest – AUC 0.89

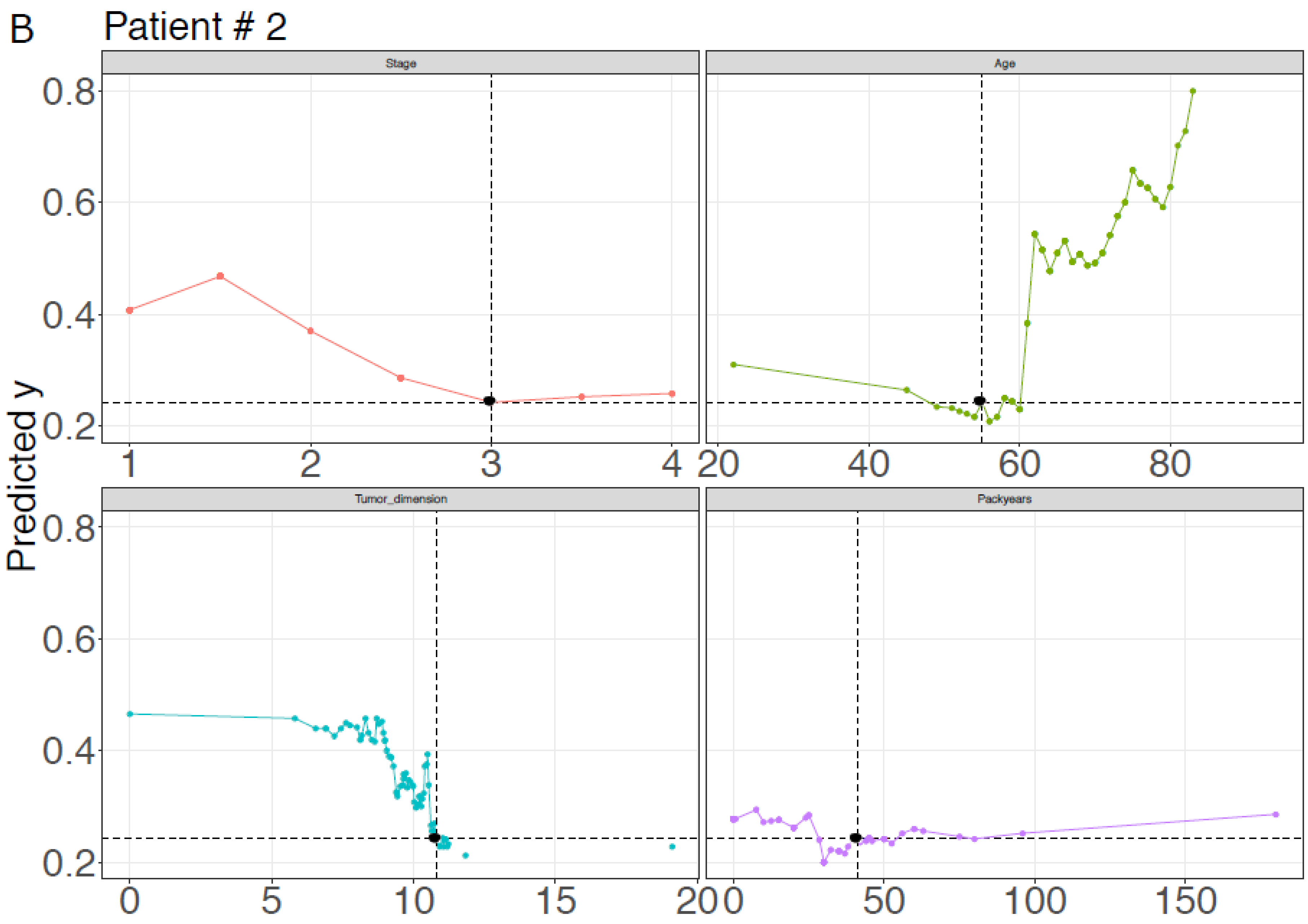
Local Explanations: applications

Table 1: Exact data for two patients that are selected to present the break down and Ceteris Paribus methodology. Last two rows show the random forest and logistic regression predictions for that observations.

	Patient 1	Patient 2
Lung	right	left
Zubrod performance status	1	0
Stage	3	3
Histopathological diagnosis	Adenosquamous carcinoma	Adenosquamous carcinoma
Second cancer	NO	NO
Living in industrial city	NO	NO
Chronic bronchitis	NO	YES
Bronchial asthma	NO	NO
Gender	M	M
Age	80	55
Tumor dimension	9.4	10.8
Cancers in family	NO	NO
Packyears	75	41
Survived 3 years	1	1
Random Forest prediction	0.6	0.24
Logistic Regression prediction	0.49	0.17


A
Patient # 1
final prognosis
Histopathological diagnosis = Adenosquamous
Packyears = 75
Gender = M
Stage = 3
Second cancer = NO
Chronic bronchitis = NO
Bronchial asthma = NO
Living in industrial city = NO
Cancers in family = NO
Lung = right
Zubrod performance status = 0
Tumor dimension = 9.4
Age = 80
(Intercept)
12.6
−18.1
−9.5
−1.6
−0.3
0
0
0.1
0.2
0.4
1
1.3
13.9
25.1
0
0.00
0.25
0.50
0.75
1.00
B
Patient # 2
final prognosis
Histopathological diagnosis = Adenosquamous
Chronic bronchitis = YES
Lung = left
Gender = M
Second cancer = NO
Bronchial asthma = NO
Cancers in family = NO
Living in industrial city = NO
Zubrod performance status = 1
Tumor dimension = 10.8
Stage = 3
Packyears = 41
Age = 55
(Intercept)
− 23.6
3.2
3.2
2.4
0
−0.1
−0.3
−0.6
−0.9
−1.8
−8.2
−5.9
−7.9
−6.7
0
0.00
0.25
0.50
0.75
1.00


A
Patient # 1
Stage
Age
Tumor_dimension
Packyears
Predicted y
0.8
0.6
0.4
0.2
1
2
3
4
20
40
60
80
0
5
10
15
20
0
50
100
150

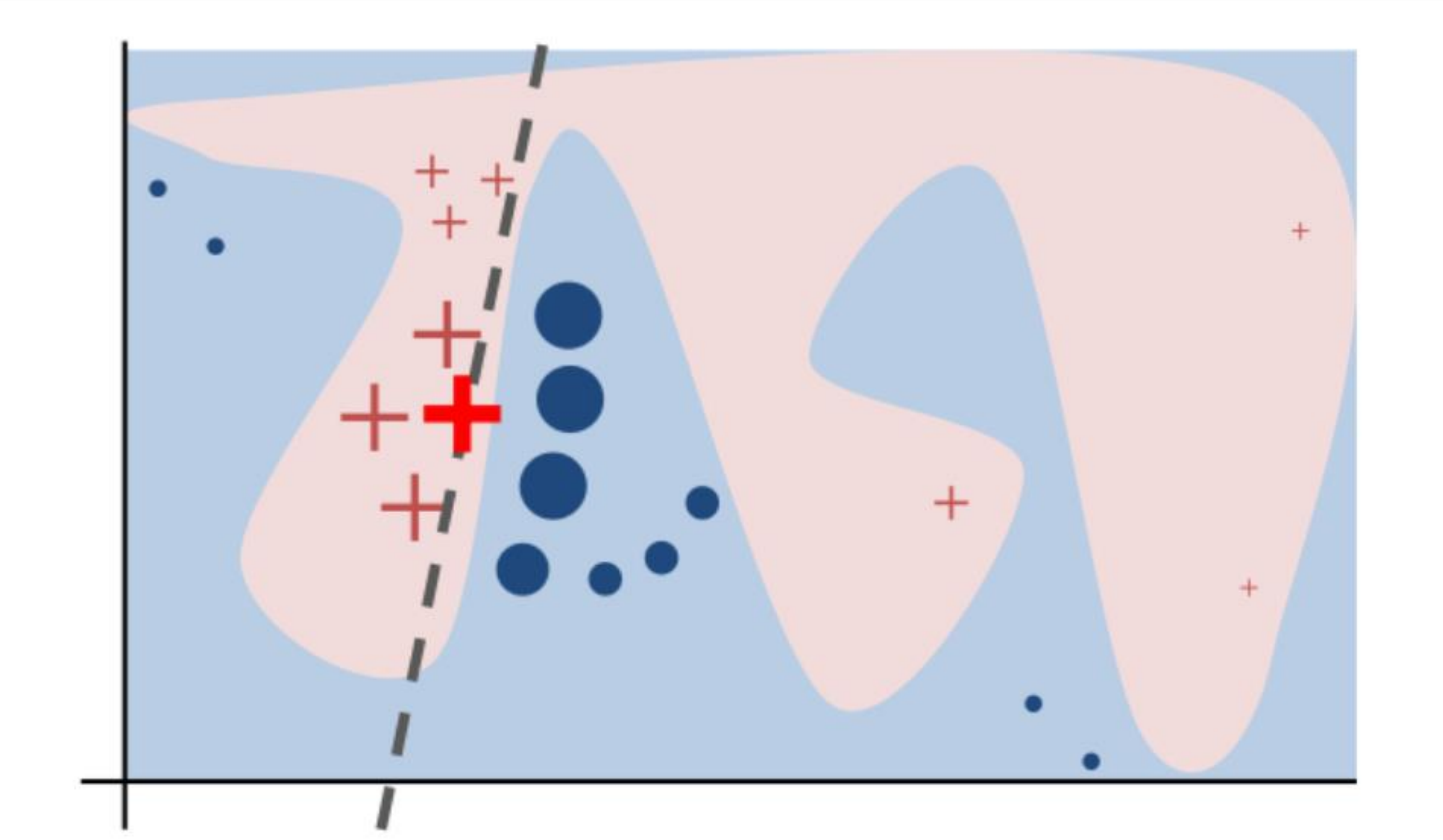


B
Patient # 2
Stage
Age
Tumor_dimension
Packyears
Predicted y
0.8
0.6
0.4
0.2
1
2
3
4
20
40
60
80
0
5
10
15
20
0
50
100
150

Local surrogate models

- Complex model is approximated with a simpler model (e.g. linear regression) locally.
- Original idea: LIME (2016) – examples in image and text analysis.
- First, a dataset of new observations similar to the explained is created.
- Then, model predictions are calculated for this new dataset.
- Simple model is fitted to these predictions.

LIME intuition



https://arxiv.org/pdf/1602.04938.pdf

LIME

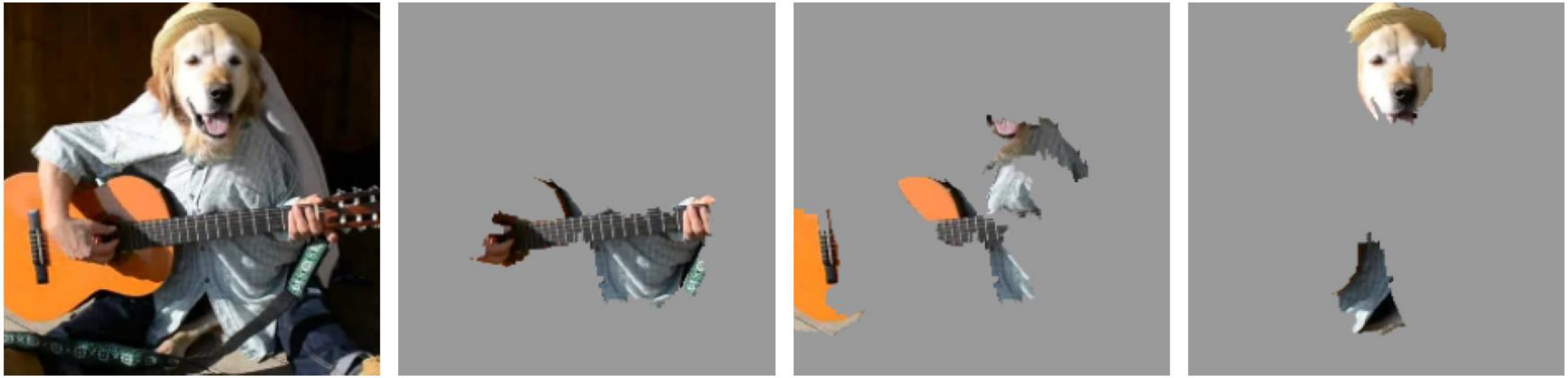
- Optimization problem:

$$\mathcal{L}(f, g, \pi_x) = \sum_{z,z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$$

  - f is the explained model,
  - g is the explanation model,
  - z is a interpretable representation of x,
  - π is a distance measure (a kernel).

LIME explanation

LIME interpretable inputs



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

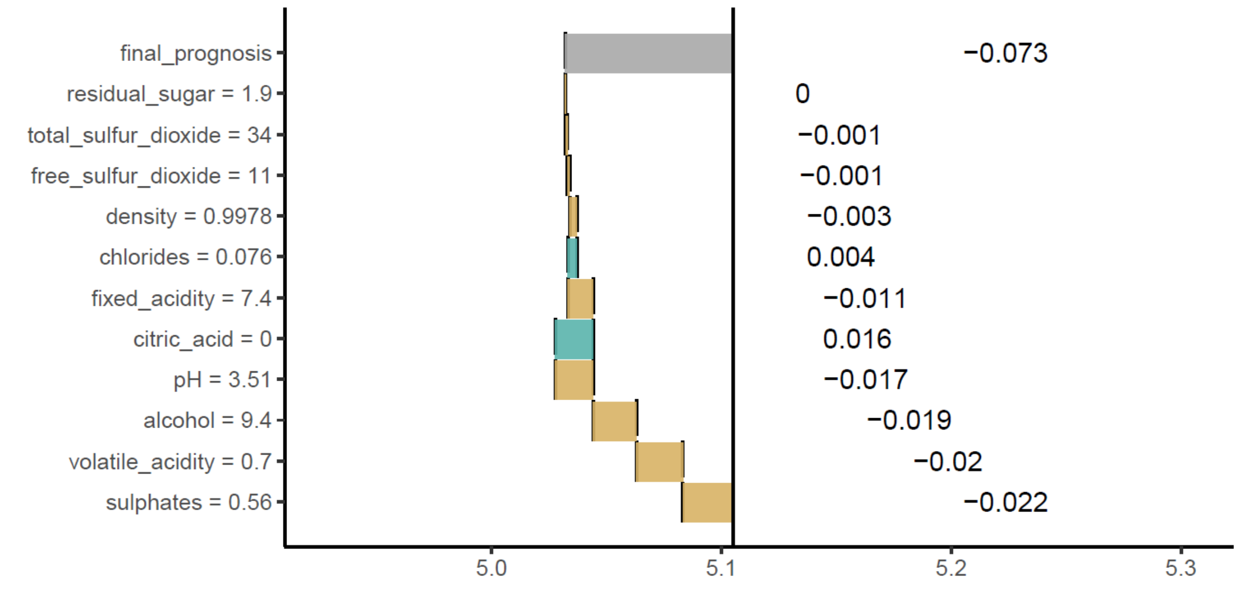
https://arxiv.org/pdf/1602.04938.pdf

LIVE: Local Interpretable Visual Explanations

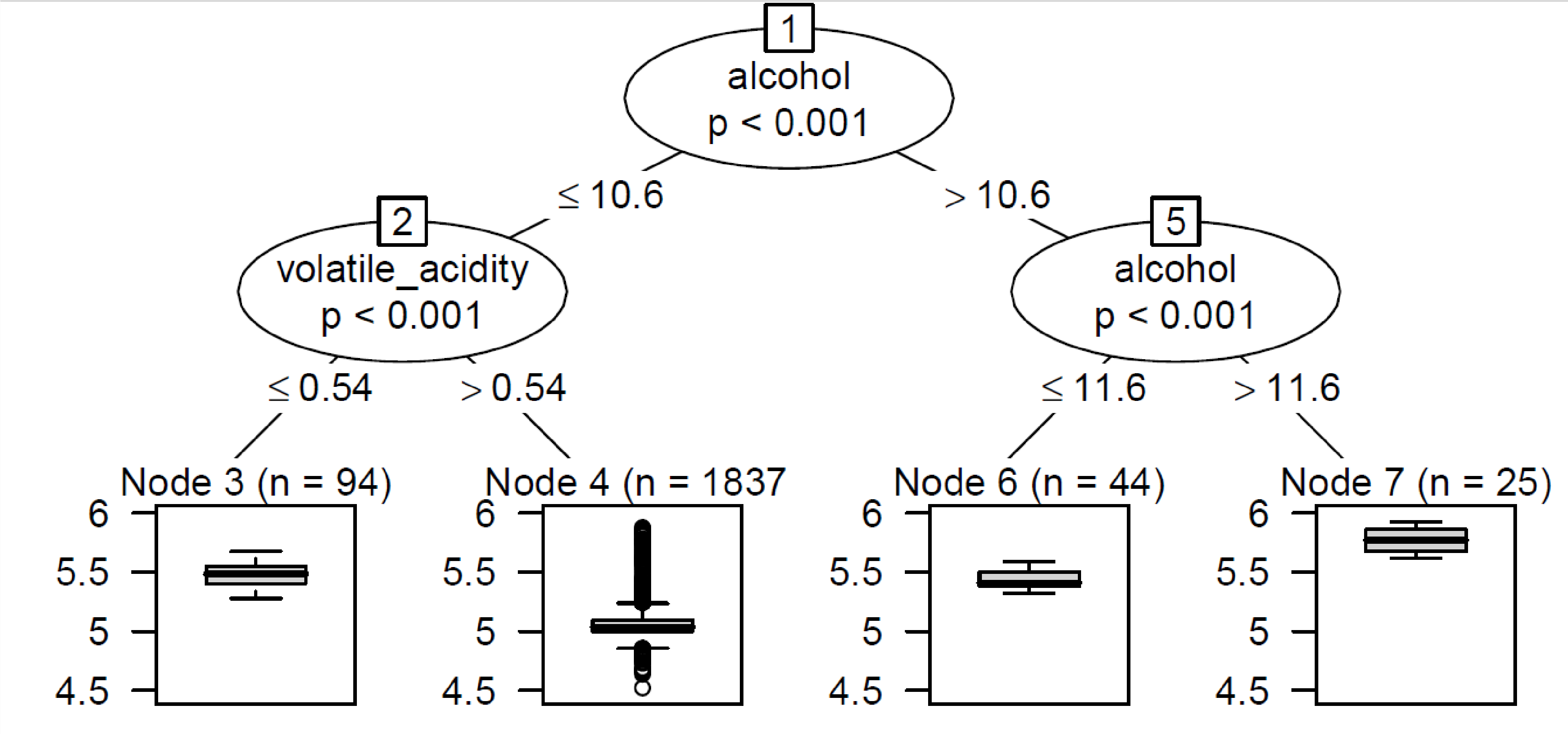
- LIME adapted to tabular data.
- Emphasis on model visualization.
- No discretization is performed.
- Different methods of sampling are available. Default: change of one feature per observation.
- High flexibility (for example, any model supported by mlr package can be an explanation).

LIVE explanations

Variable	N	Estimate		p
fixed_acidity	500		0.10 (0.08, 0.12)	<0.001
volatile_acidity	500		−1.47 (−1.64, −1.29)	<0.001
citric_acid	500		−0.54 (−0.64, −0.44)	<0.001
residual_sugar	500		0.01 (−0.04, 0.06)	0.664
chlorides	500		1.12 (0.32, 1.91)	0.006
free_sulfur_dioxide	500		−0.01 (−0.01, −0.00)	<0.001
total_sulfur_dioxide	500		0.00 (−0.00, 0.00)	0.417
density	500		−27.45 (−41.49, −13.41)	<0.001
pH	500		−0.29 (−0.42, −0.16)	<0.001
sulphates	500		1.19 (1.08, 1.30)	<0.001
alcohol	500		0.23 (0.21, 0.26)	<0.001
		−40 −30 −20 −10 0		



LIVE explanations



MI2 Data Lab

- http://mi2.mini.pw.edu.pl – group website.
- https://github.com/mi2datalab - R & Python libraries developed by the group, in particular related to xAI.
- https://github.com/ModelOriented - new R & Python libraries of DrWhy? Project (successor to DALEX).
- Funding: NCN grant DALEX (2018-2020).
- Project related to NLP, healthcare (e.g. lung cancer) and more.

References

- Biecek, P., 2018. DALEX: explainers for complex predictive models. arXiv:1806.08915 [cs, stat].
- Goodman, B., Flaxman, S., 2017. European Union regulations on algorithmic decision-making and a “right to explanation.” AI Magazine 38, 50. https://doi.org/10.1609/aimag.v38i3.2741
- Grudziaz, A., Gosiewska, A., Biecek, P., 2018. survxai: an R package for structure-agnostic explanations of survival models. Journal of Open Source Software 3, 961. https://doi.org/10.21105/joss.00961
- Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B., 2017. What do we need to build explainable AI systems for the medical domain? arXiv:1712.09923 [cs, stat].
- Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 4765–4774.

References

- Lundberg, S.M., Nair, B., Vavilala, M.S., Horibe, M., Eisses, M.J., Adams, T., Liston, D.E., Low, D.K.-W., Newman, S.-F., Kim, J., Lee, S.-I., 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature Biomedical Engineering 2, 749. https://doi.org/10.1038/s41551-018-0304-0
- Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.
- Robnik-Šikonja, M., Bohanec, M., 2018. Perturbation-Based Explanations of Prediction Models, in: Zhou, J., Chen, F. (Eds.), Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent. Springer International Publishing, Cham, pp. 159–175. https://doi.org/10.1007/978-3-319-90403-0_9
- Staniak, M., Biecek, P., 2018. Explanations of model predictions with live and breakDown packages. arXiv:1804.01955 [cs, stat].

Thank you for your attention

Time for discussion!