Interpretable Features for Local Explanations of Machine Learning Models

Presentation · May 2019

DOI: 10.13140/RG.2.2.23260.23684

1 author:

Mateusz Staniak
University of Wroclaw

Interpretable Features for Local Explanations of Machine Learning Models

Mateusz Staniak, MI^2 Data Lab @ Warsaw University of Technology

Ljubljana, 16 V 2019

About me

- First year PhD student in Computer Science.
- Current interests:
  - Interpretable Machine Learning,
  - Proteomics ,
  - ML applications in Biology and Medicine,
  - Automation in Exploratory Data Analysis and Model Exploration.
- Background: Mathematics (Statistics & Probability theory).
- https://github.com/mstaniak

MI2DataLab/live

Local Interpretable (Model-agnostic) Visual Explanations - model visualization for regression problems and tabular data based on LIME method. Available on CRAN

R ★ 27 ⑂ 3

autoEDA-resources

A list of software and papers related to automatic/fast Exploratory Data Analysis

HTML ★ 36 ⑂ 5

ModelOriented/localModel

LIME-like explanations with interpretable features based on Ceteris Paribus curves. Now on CRAN.

R ★ 7 ⑂ 2

mi2-warsaw/PISAoccupations

Shiny app and package for exploring data from PISA study.

R ⑂ 1

The need for interpretability

Before

5 years of web logs + ML

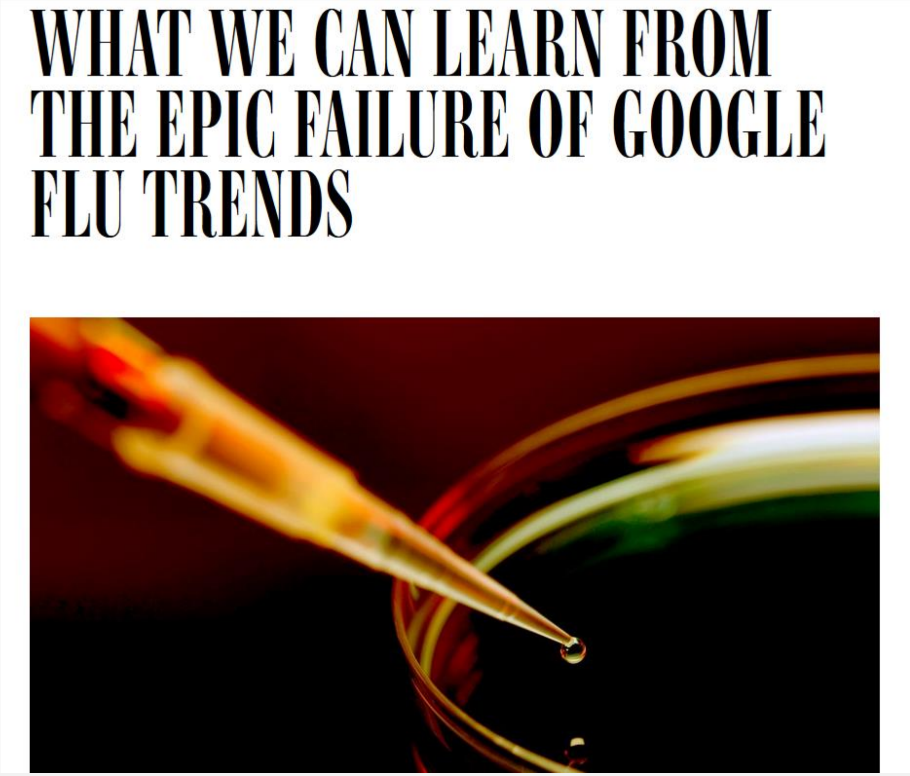
=

proved to be a more useful and timely indicator [of flu] than government statistics with their natural reporting lags

- Viktor Mayer-Schönberger and Kenneth Cukier ,

Big Data: A Revolution That Will Transform How We Live, Work and Think

After

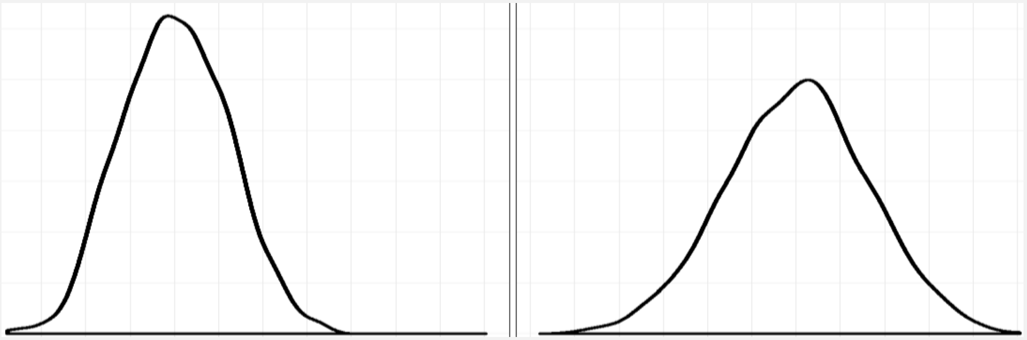


wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

Machine Learning models are vulnerable to:

- Biased training (and other data quality issues),
- Concept drift,
- Unmeasurable objectives (Fairness, Lawfulness).

Females vs Male frequencies

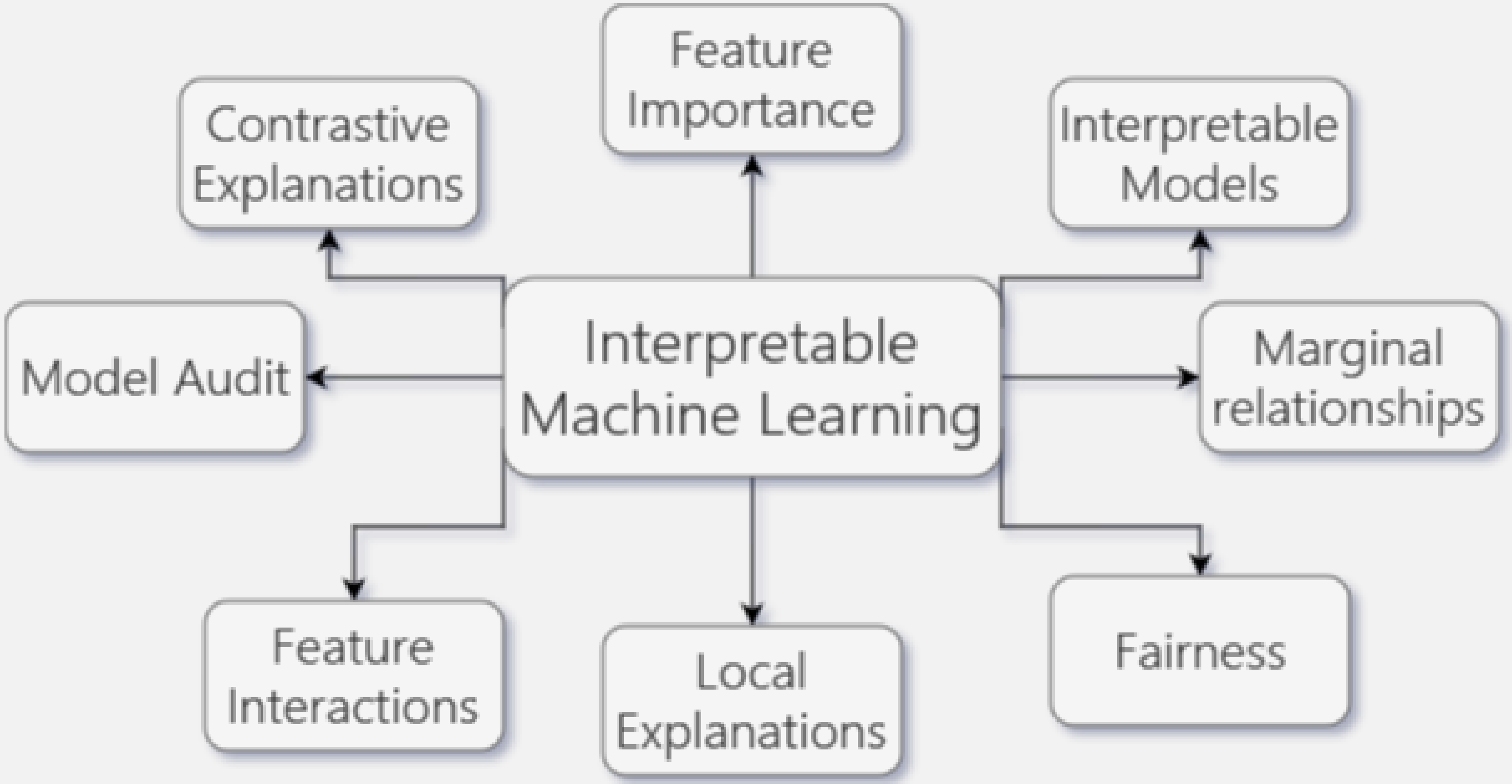


Training data vs validation data distribution

Amazon scraps secret AI recruiting tool that 'didn't like women'

- **Amazon ended job recruiting service that was reportedly biased against women**
- **It was created by Amazon's Edinburgh team in 2014 to automatically sort CVs**
- **The AI taught itself to downgrade resumes that included words like 'women's'**

Interpretable Machine Learning



Types of explanations

- Intrinsic vs **post-hoc.**
- Model-specific vs **model-agnostic.**
- Global vs **local** (model-level vs instance-level).

Intrinsic vs post-hoc

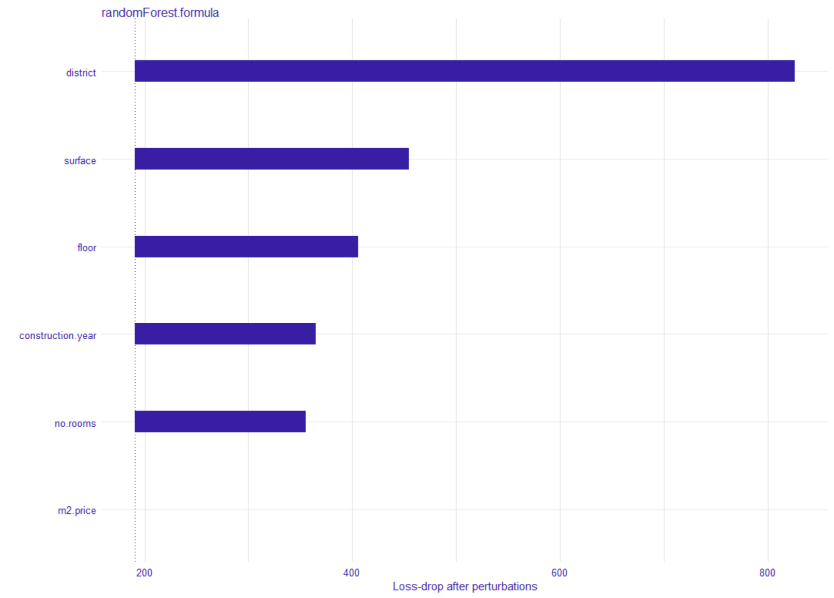
- Intrinsic explanations are based on algorithm design
  (model form, training explanations and model jointly).

```
lm(formula = m2.price ~ ., data = apartments)

Residuals:
   Min     1Q Median     3Q    Max
-247.5 -202.8 -172.8  381.4  469.0

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         5020.1391   682.8721   7.352 4.11e-13 ***
construction.year     -0.2290     0.3483  -0.657   0.5110
surface              -10.2378     0.5778 -17.720  < 2e-16 ***
floor                -99.4820     3.0874 -32.222  < 2e-16 ***
no.rooms             -37.7299    15.8440  -2.381   0.0174 *
districtBielany       17.2144    40.4502   0.426   0.6705
districtMokotow      918.3802    39.4386  23.286  < 2e-16 ***
districtOchota       926.2540    40.5279  22.855  < 2e-16 ***
districtPraga        -37.1047    40.8930  -0.907   0.3644
districtSrodmiescie 2080.6110    40.0149  51.996  < 2e-16 ***
districtUrsus         29.9419    39.7249   0.754   0.4512
districtUrsynow      -18.8651    39.7565  -0.475   0.6352
districtWola         -16.8912    39.6283  -0.426   0.6700
districtZoliborz     889.9735    40.4099  22.024  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Post-hoc explanations only require predict interface
  (the ability to obtain model predictions for specified examples).

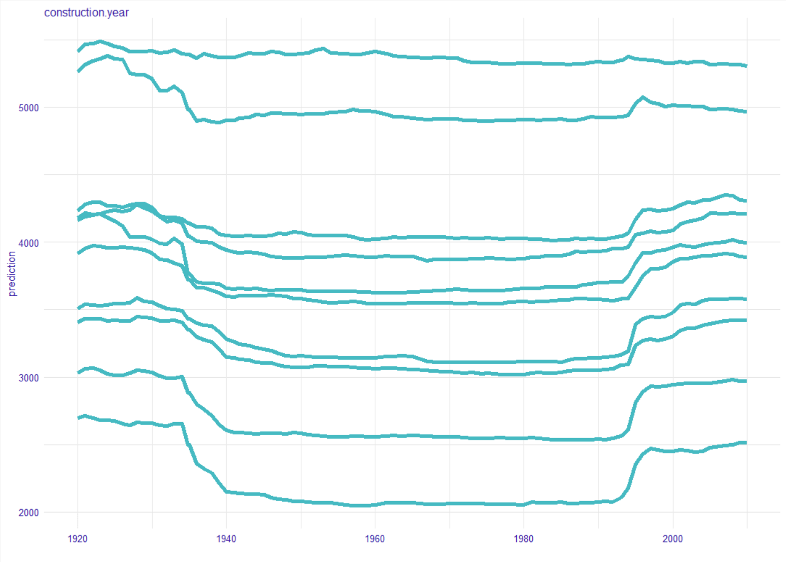


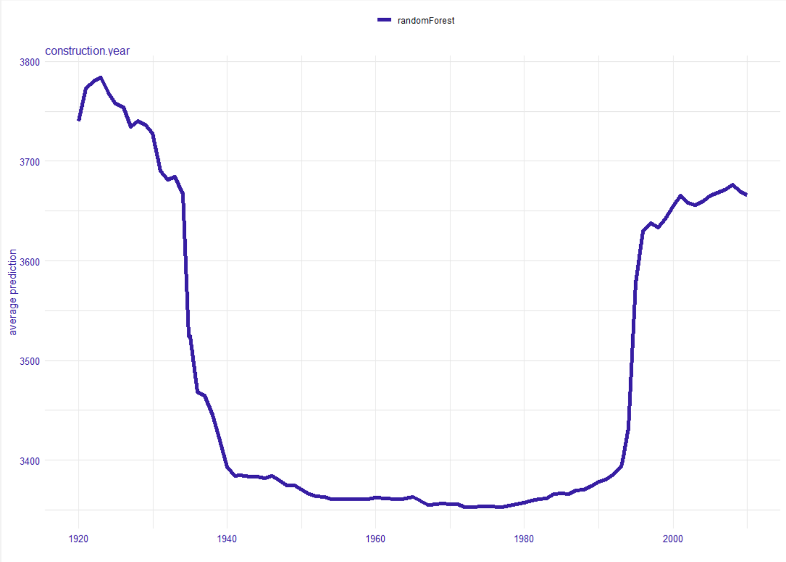
Model-agnostic Approach

	Definition	Example	Comments
Model-agnostic explanations	Do not use knowledge about the specific algorithm	Permutation-based variable importance	Do not require model re-fitting
Model-specific explanations	Assume that a specific algorithm was used to fit the model	Average minimum depth in a random forest	Can be more accurate

Local vs Global Explanations

- Local explanations are concerned with a single observation and its prediction.
- Global explanations are concerned with the model as a whole.
  Global explanations are often aggregation of local explanations (e.g. mean).





Local Explanations

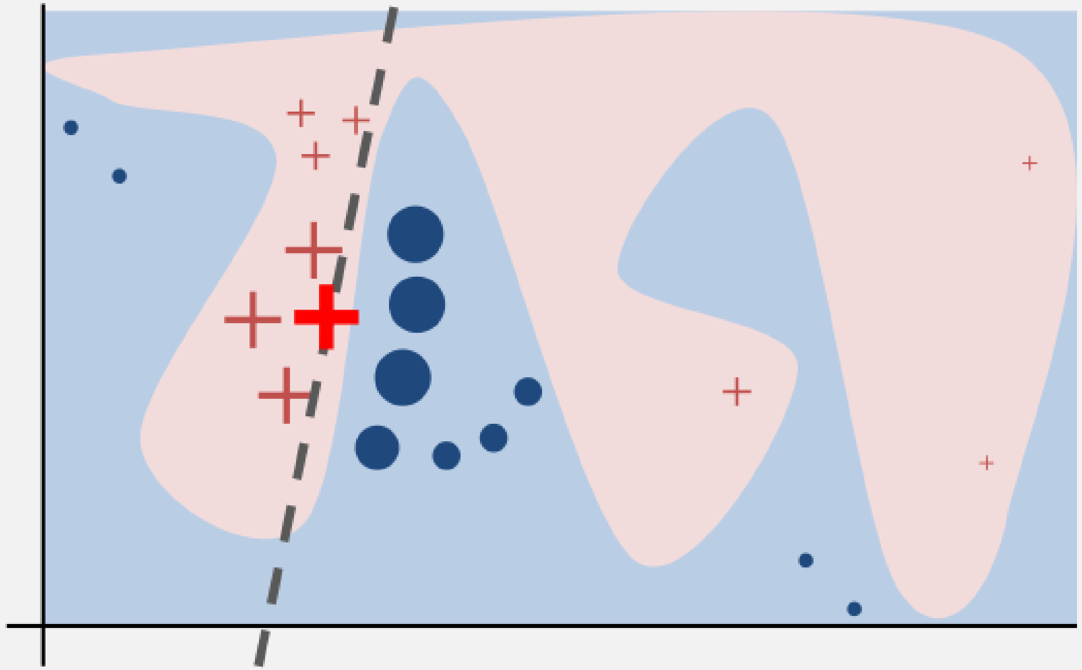
Approaches to Local Explanations

- What-If analysis (marginal response of the model when changing a single variable for a single observation):
  - Ceteris Paribus profiles (Individual Conditional Expectation).
- Local surrogate models (aka LIME – fitting an interpretable model locally):
  - LIME and its modifications (aLIME, k-LIME, localSurrogate),
  - LIVE and localModel (versions of LIME developed at MI2 Data Lab).
- Example-based explanations
  - Contrastive explanations
  - Prototypes and criticism
- Prediction decomposition (attributing additive scores to features).
  - EXPLAIN,
  - Shapley Values,
  - Break Down and iBreakDown (methods connected to Shapley Values developed at MI2 Data Lab)

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.

Local Interpretable Model-agnostic Explanations



[1]

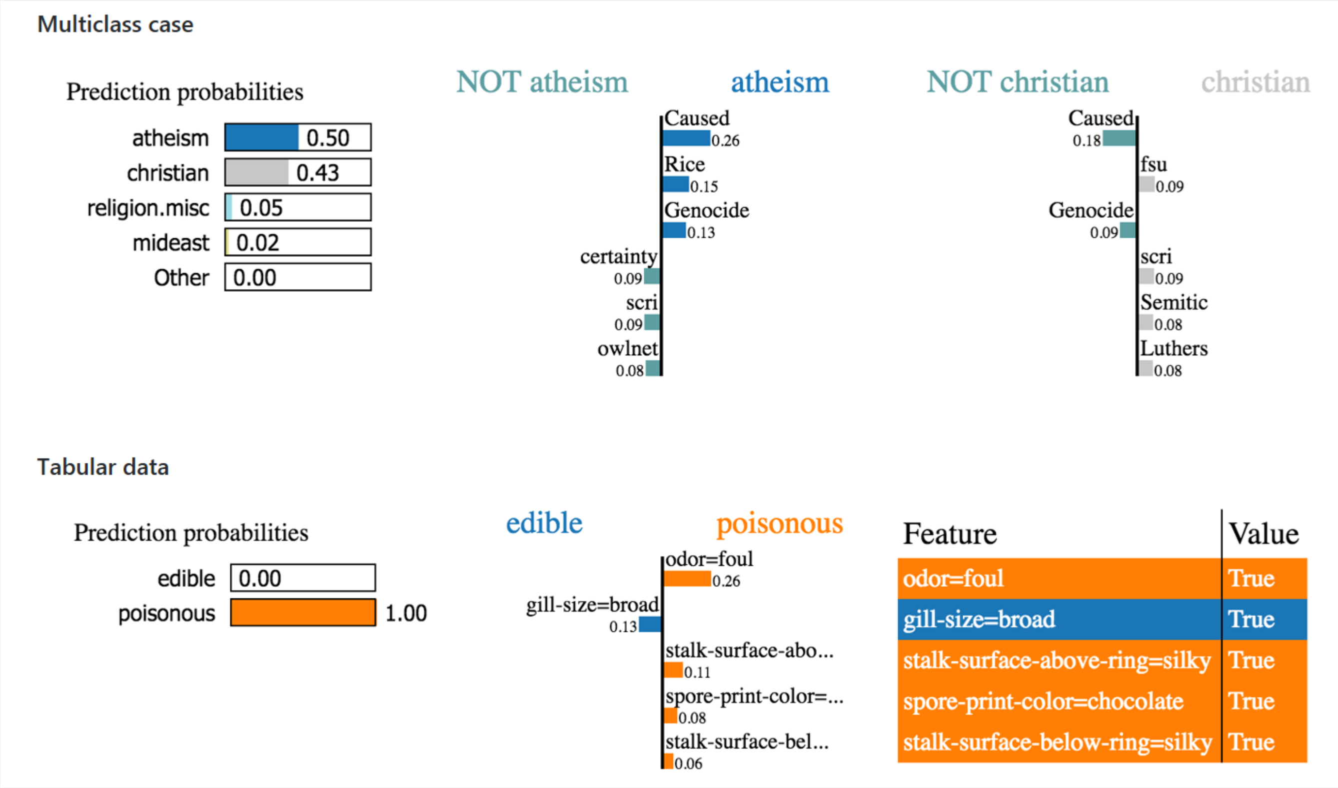
M. T. Ribeiro, S. Singh, C. Guestrin, „«Why Should I Trust You?»: Explaining the Predictions of Any Classifier", *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016

- 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 explanations



https://github.com/marcotcr/lime

X_1	…	X_p
M	…	0.11
F	…	-0.25
…	…	…
U	…	0.887

$f(x)$

y
0.87
0.14
…
0.54

1)

Z_1	…	Z_p
X_1 = M	…	$X_p < 0.2$
X_1 = F, U	…	$X_p < 0.2$
…	…	…
X_1 = F, U	…	$X_p > 0.2$

X_1 = F, U	…	$X_p < 0.2$

2)

X_1 = F, U	…	$X_p < 0.2$

X_1 = M	…	$X_p < 0.2$

X_1 = F, U	…	$X_p > 0.2$

X_1 = M	…	$X_p > 0.2$

…

X_1 = F, U	…	$X_p < 0.2$

Z_1	…	Z_p
X_1 = M	…	$X_p < 0.2$
X_1 = F, U	…	$X_p > 0.2$
X_1 = M	…	$X_p > 0.2$
…	…	…
X_1 = F, U	…	$X_p < 0.2$

3)

Z_1	…	Z_p
X_1 = M	…	$X_p < 0.2$
X_1 = F, U	…	$X_p > 0.2$
X_1 = M	…	$X_p > 0.2$
…	…	…
X_1 = F, U	…	$X_p < 0.2$

X_1	…	X_p
M	…	0.111
F	…	1.27
M	…	0.887
…	…	…
F	…	-0.2

4)

f(x)

f(x)
0.93
0.77
0.122
…
0.64

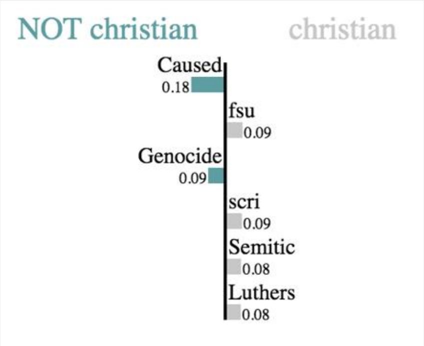
5)

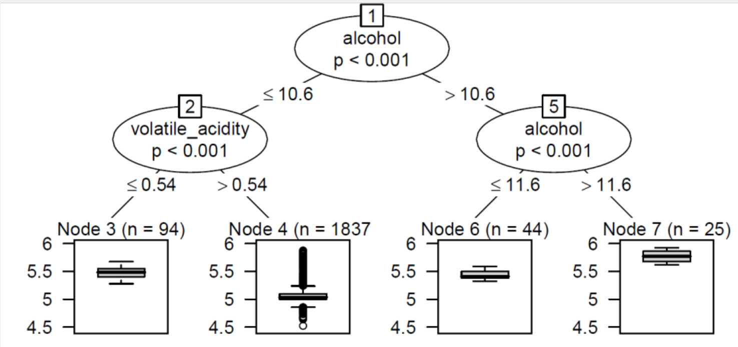
Z_1	…	Z_p	f(x)
$X_1 = M$	…	$X_p < 0.2$	0.93
$X_1 = F, U$	…	$X_p > 0.2$	0.77
$X_1 = M$	…	$X_p > 0.2$	0.122
…	…	…	…
$X_1 = F, U$	…	$X_p < 0.2$	0.64

g(z)

g(z)
0.90
0.81
0.07
…
0.641

6) g(z)





Some remarks on LIME for tabular data

- Most of the work so far focused on step 2) – the sampling:
  - Laugel et al. 2018: the neighbourhood must include the decision boundary,
  - Adhikari et a. 2018: the neighbourhood must include enough data points from both classes,
  - Tan et al. 2019: sampling introduces significant uncertainty.
- For image or text step 3) is trivial, but for tabular data and non-trivial interpretable input spaces, the inverse transformation is a problem.
- For tabular data, step 1) is important, but often features are not transformed.

Interpretable Features

https://www3.cs.stonybrook.edu/~leman/courses/13CSE512/images/joke1.png


SUPPORT VECTOR MACHINES
REPEA
POWE
LAWS
END DUALITY GAP
Map Reduce
Map Reuse
Map Recycle
BAYESIANS AGAINST DISCRIMINATION
FREE VARIABLES!
BAN GENETIC ALGORITHMS


SUPPORT
VECTOR
MACHINES
LAWS
Map Reduce
Map Reuse
Map Recycle
FREE
BAN
GENETIC
ALGORITHMS

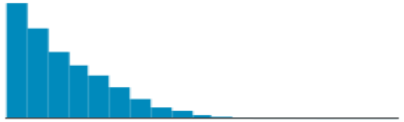

SUPPORT
VECTOR
MACHINES
LAWS
Map Reduce
Map Reuse
Map Recycle
FREE
ALGORITHMS


LAWS
GAP
FREE
BAN
GENETIC


SUPPORT
VECTOR
MACHINES
LAWS
GAP
FREE
BAN
GENETIC


SUPPORT VECTOR MACHINES
REPEA
POWE
LAWS
END DUALITY GAP
Map Reduce
Map Reuse
Map Recycle
BAYSIANS AGAINST DISCRIMINATION
FREE VARIABLES!
BAN GENETIC ALGORITHMS

Tabular data

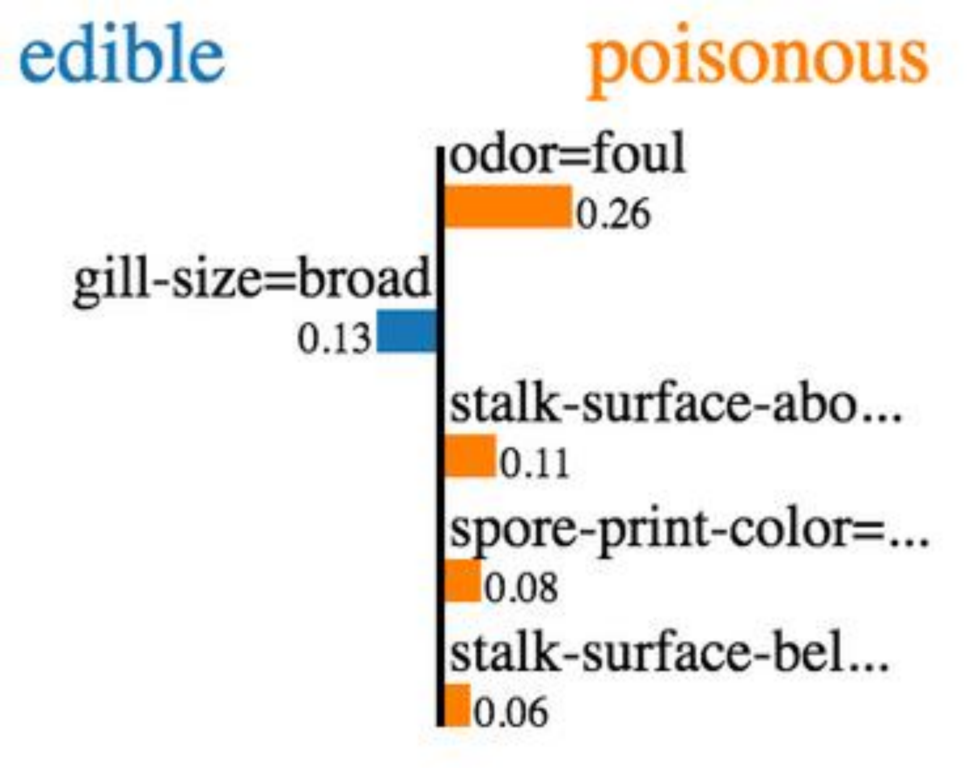
A Region	A Country	# Distance	# Points	A Glider
PACA 25% Rhône-Alpes 13% Other (28) 61%	France 96% Espagne 2% Other (18) 2%	 25 1.25k	 21.2 1.2k	Pegase 16% Duo Discus 8% Other (178) 77%
	Afrique du Sud	568.280029	511.959991	Duo Discus X
	Afrique du Sud	398.700012	402.049988	JS1 18m
	Namibie	482.380005	398.660004	Nimbus 4D (< 750 kg)

https://www.kaggle.com/ezgliding/netcoupe-flight-metadata

Existing Approaches for Tabular Data

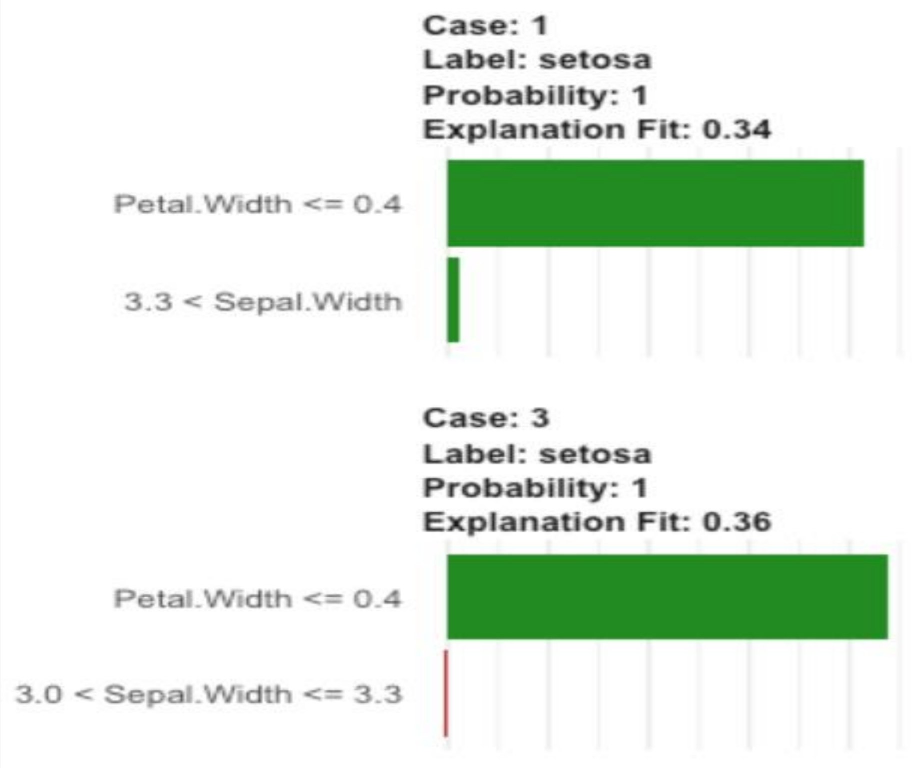
Discretized features

- lime library (Python) – Marco Tulio Ribeiro



https://github.com/marcotcr/lime

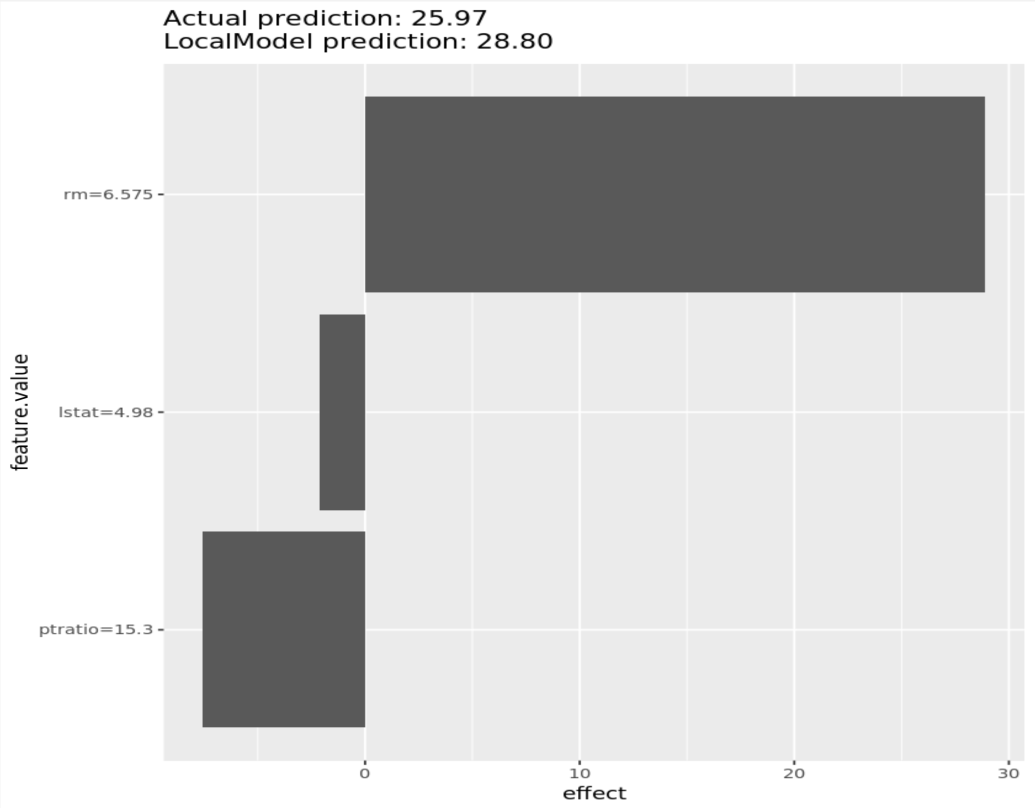
- lime package (R) – Thomas Lin Pedersen



https://github.com/thomasp85/lime

Continuous features

- iml package (R) – Christoph Molnar



https://github.com/christophM/iml

- live package (R) – Mateusz Staniak

Variable	N	Estimate		p
fixed_acidity	2000		0.14 (0.14, 0.15)	<0.001
volatile_acidity	2000		−1.43 (−1.46, −1.40)	<0.001
citric_acid	2000		−0.66 (−0.69, −0.63)	<0.001
residual_sugar	2000		0.00 (−0.00, 0.01)	0.9
chlorides	2000		−2.57 (−2.71, −2.43)	<0.001
free_sulfur_dioxide	2000		0.00 (0.00, 0.00)	<0.001
total_sulfur_dioxide	2000		0.00 (0.00, 0.00)	<0.001
density	2000		−25.69 (−28.09, −23.30)	<0.001
pH	2000		−0.82 (−0.85, −0.79)	<0.001
sulphates	2000		2.56 (2.53, 2.60)	<0.001
alcohol	2000		0.20 (0.19, 0.20)	<0.001
(Intercept)			30.32 (27.93, 32.71)	<0.001

−20100 102030

https://github.com/MI2DataLab/live

Navigation

Current Issue
Accepted articles
Archive
R News
News and Notes
Submissions
Reviews and Proofreading

The R Journal: article published in 2018, volume 10:2

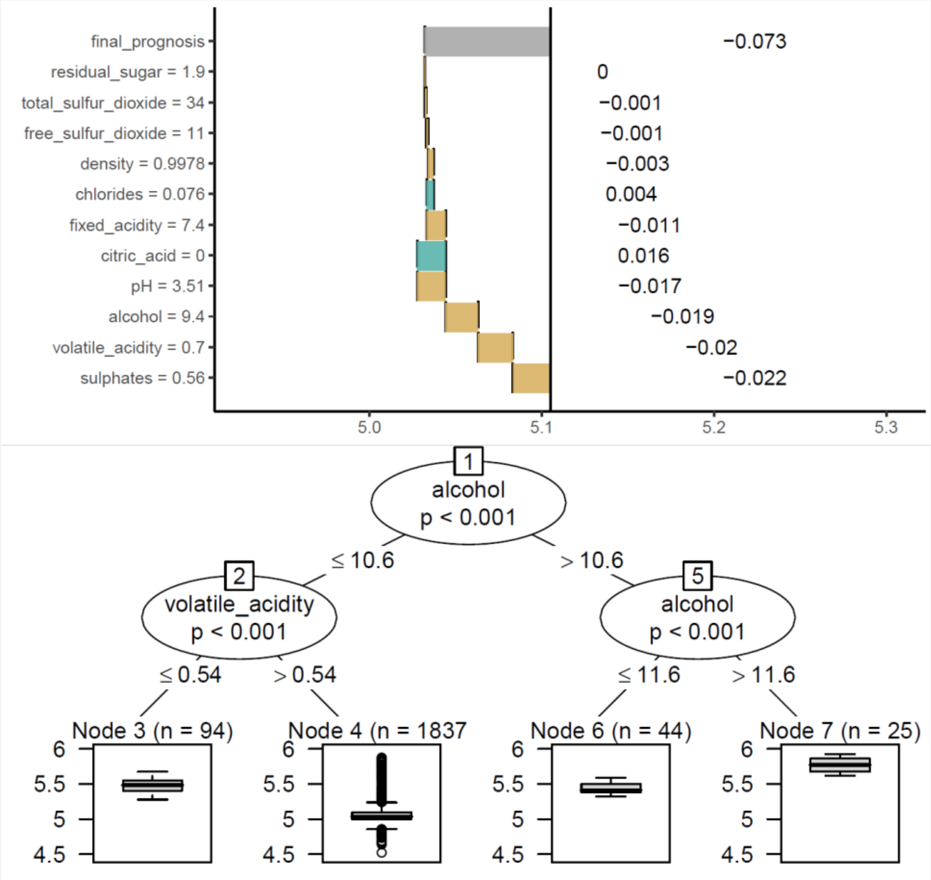
Explanations of Model Predictions with live and breakDown Packages

Mateusz Staniak and Przemysław Biecek , *The R Journal* (2018) 10:2, pages 395-409.

Abstract Complex models are commonly used in predictive modeling. In this paper we present R packages that can be used for explaining predictions from complex black box models and attributing parts of these predictions to input features. We introduce two new approaches and corresponding packages for such attribution, namely live and breakDown. We also compare their results with existing implementations of state-of-the-art solutions, namely, lime (Pedersen and Benesty, 2018) which implements Locally Interpretable Model-agnostic Explanations and iml (Molnar et al., 2018) which implements Shapley values.

LIVE: Local Interpretable Visual Explanations

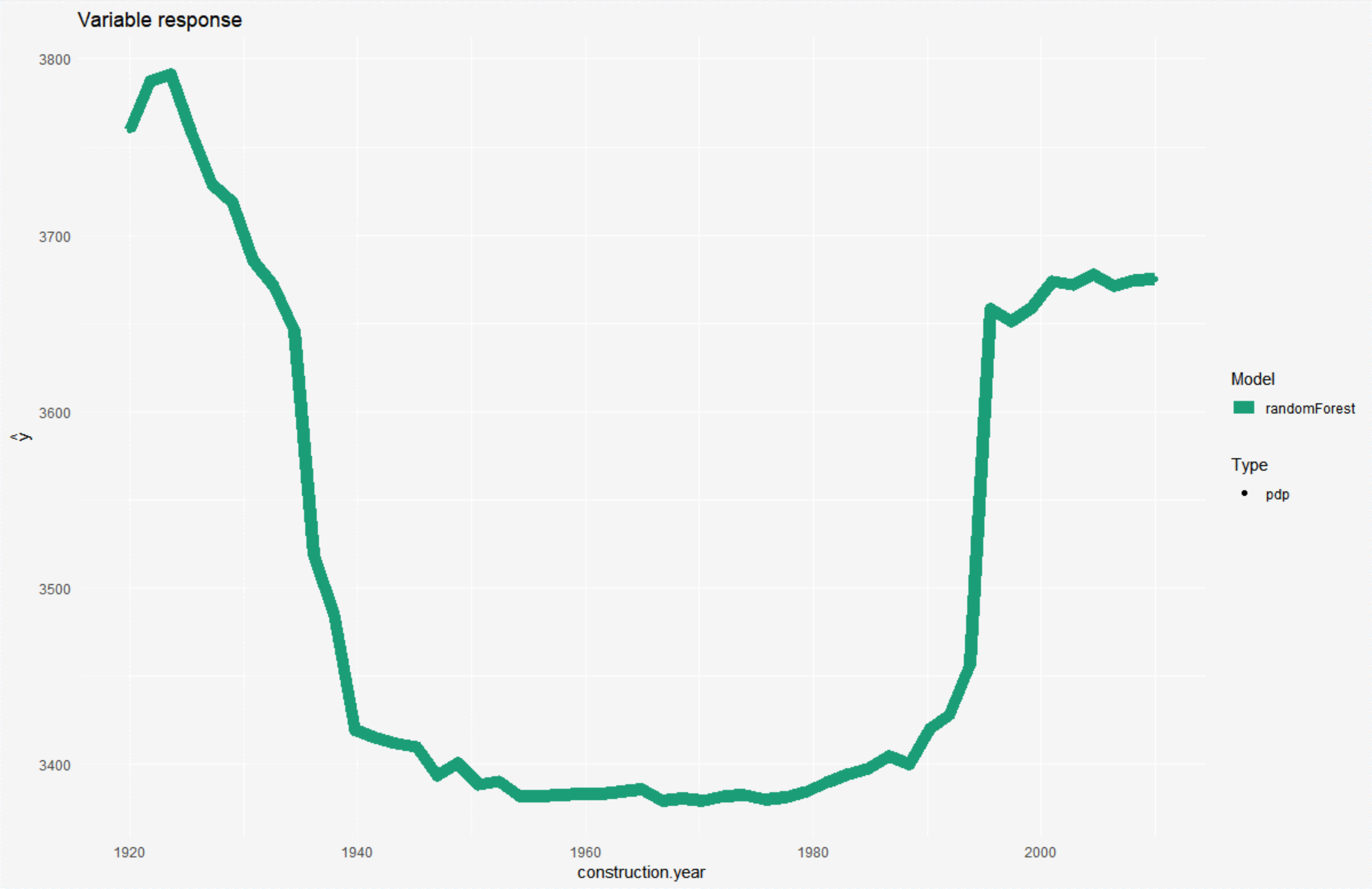
- LIME adapted to tabular data and regression problems.
- Emphasis on model visualization.
- No discretization is performed.
- Different methods of sampling are available. Default: change of one feature per observation.
- High flexibility (for example, any model supported by mlr package can be an explanation,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



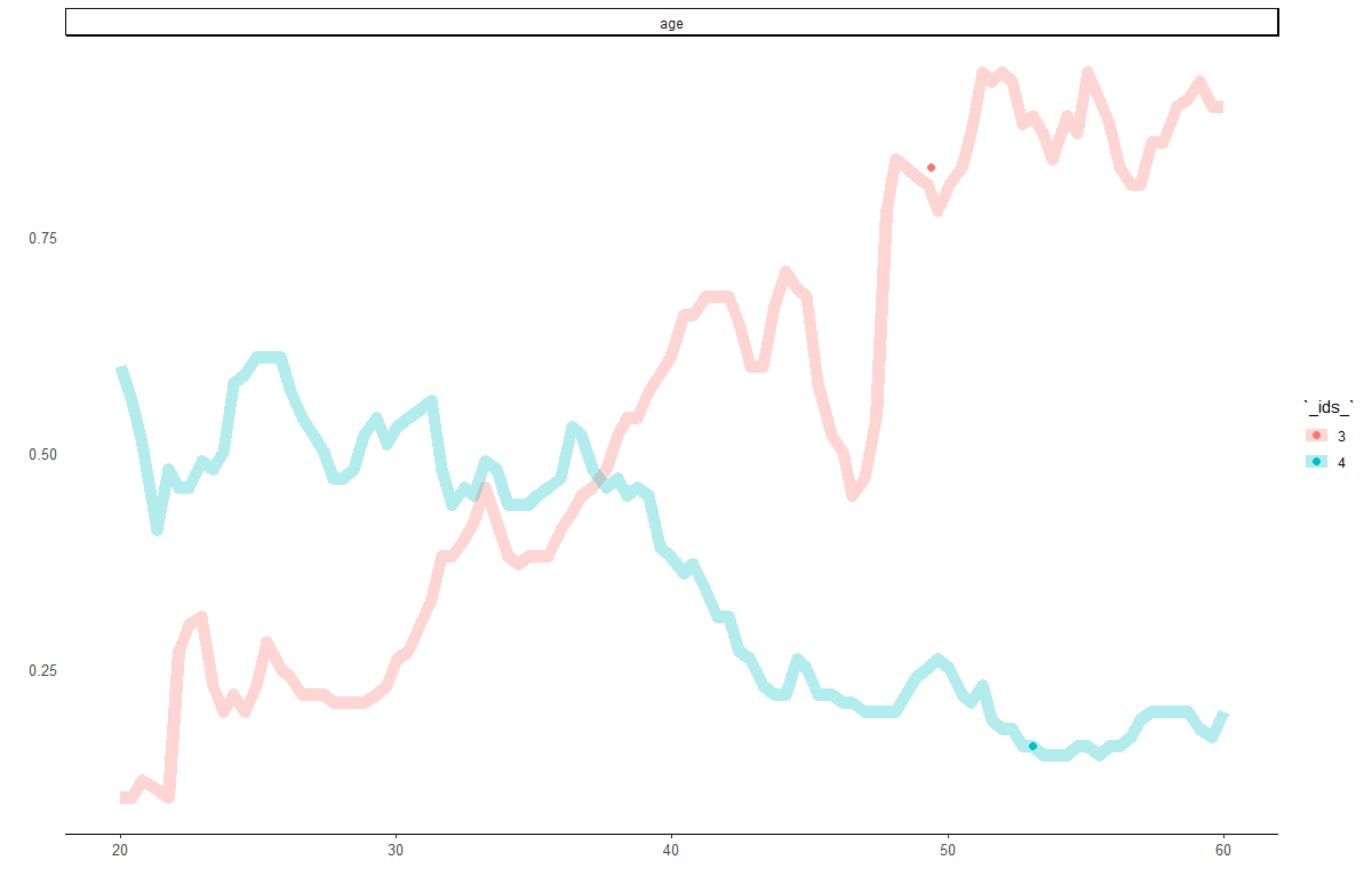
A New Approach:

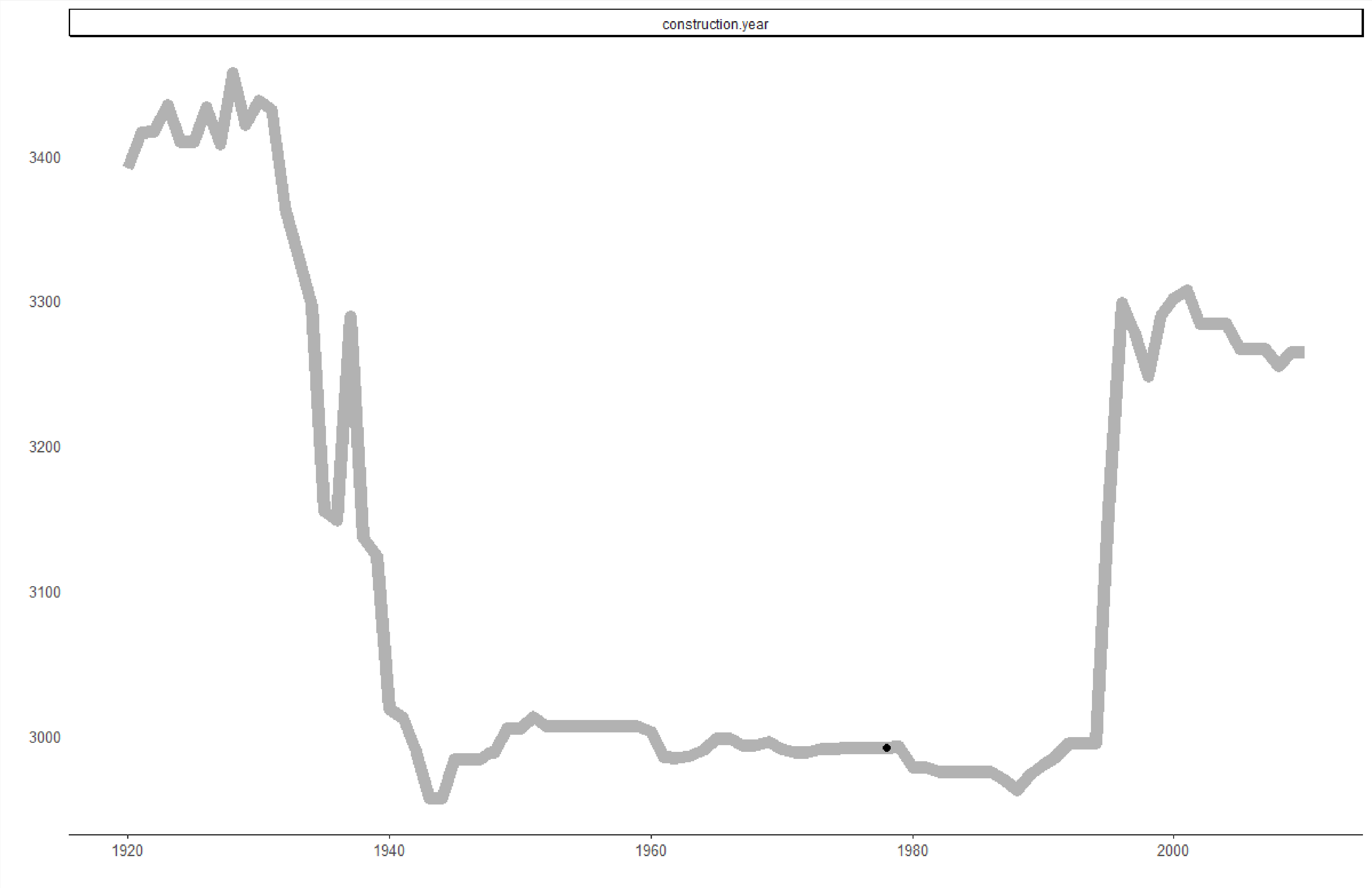
Use our knowledge about the model behaviour

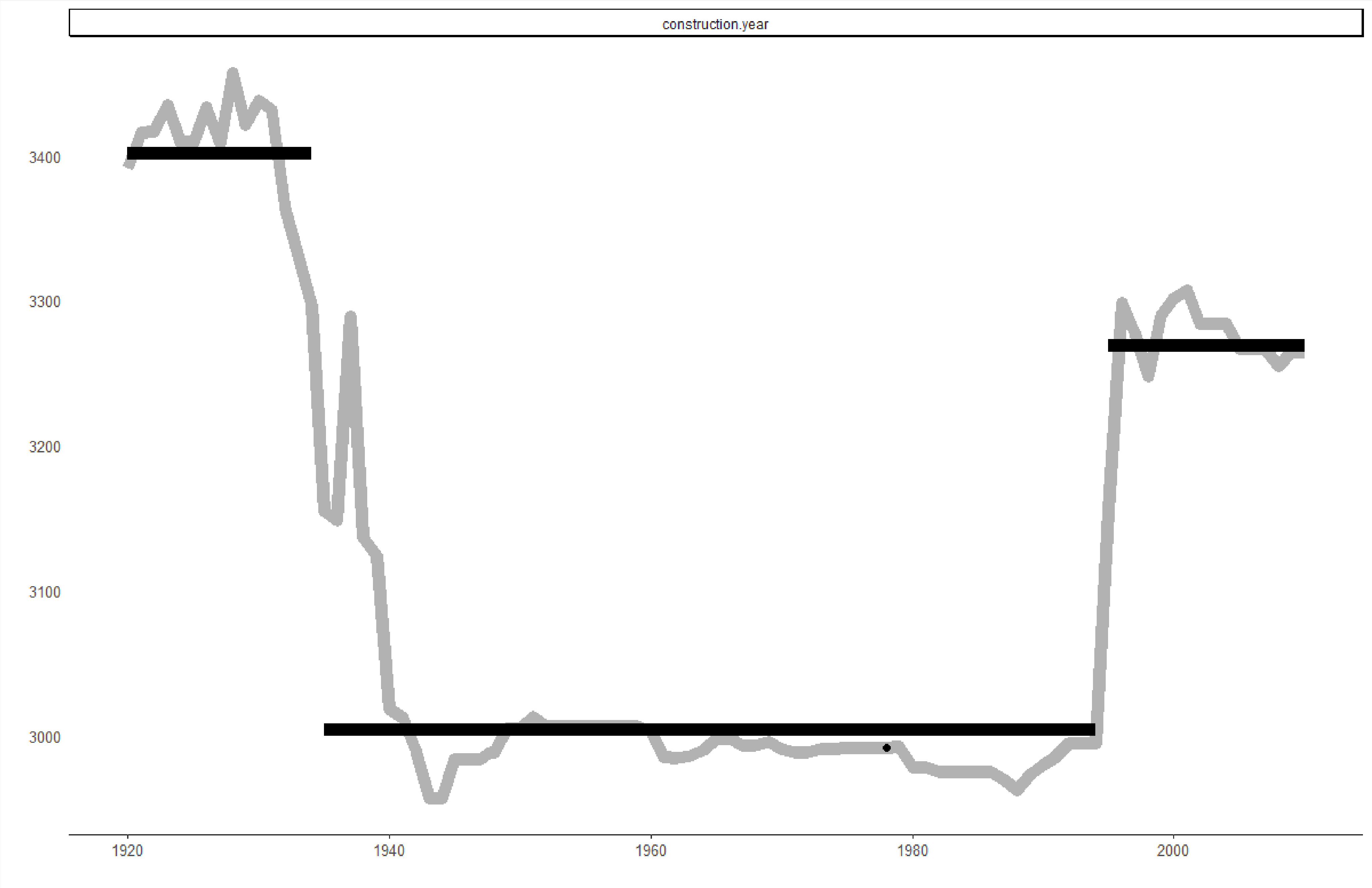
Partial Dependence Plots

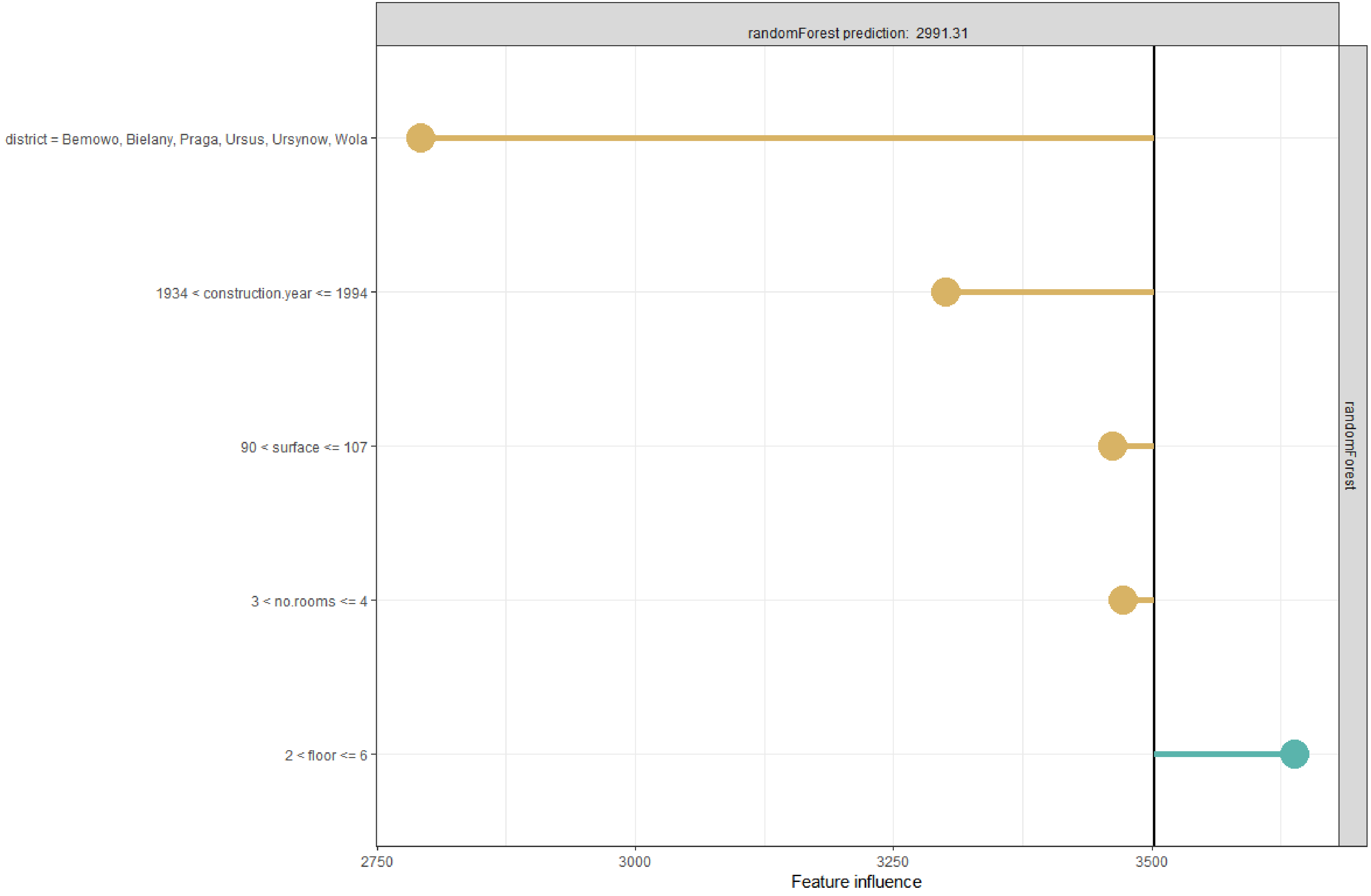


Ceteris Paribus Profiles




construction.year
3400
3300
3200
3100
3000
1920
1940
1960
1980
2000


construction.year
3400
3300
3200
3100
3000
1920
1940
1960
1980
2000

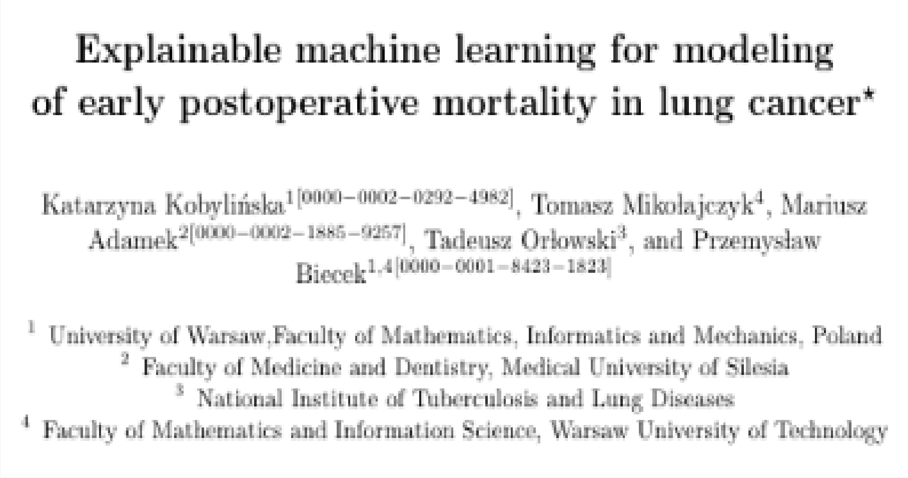

randomForest prediction: 2991.31
district = Bemowo, Bielany, Praga, Ursus, Ursynow, Wola
1934 < construction.year <= 1994
90 < surface <= 107
3 < no.rooms <= 4
2 < floor <= 6
randomForest
2750
3000
3250
3500
Feature influence

The Future of localModel

- Dimensionality reduction while creating interpretable features:
  - can provide explanations based on groups of features (for example correlated features),
  - can improve computations (Shapley Values to estimate effects of new features?),
  - method: discretization of multidimensional Ceteris Paribus profiles? Random permutations as in images?
- Large-scale comparison of LIME-variants (fidelity, stability, uncertainty).

The Future of localModel

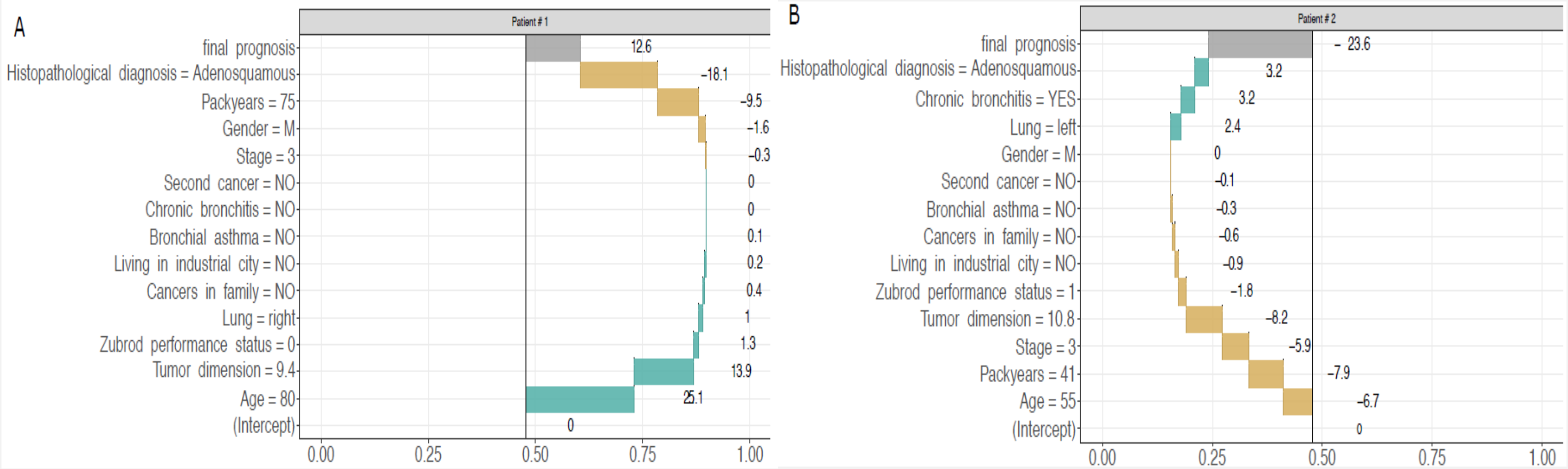
- Potential application in biostats:
  - Lung cancer data
    - Continuation of research done with Break Down methodology



Explainable machine learning for modeling of early postoperative mortality in lung cancer*

Katarzyna Kobylińska[1][0000−0002−0292−4982], Tomasz Mikołajczyk[4], Mariusz Adamek[2][0000−0002−1885−9257], Tadeusz Orłowski[3], and Przemysław Biecek[1,4][0000−0001−8423−1823]

[1] University of Warsaw,Faculty of Mathematics, Informatics and Mechanics, Poland
[2] Faculty of Medicine and Dentistry, Medical University of Silesia
[3] National Institute of Tuberculosis and Lung Diseases
[4] Faculty of Mathematics and Information Science, Warsaw University of Technology



  - Metabolomics data
    - 110 metabolites as lung cancer predictors (lung cancer vs inflamation)
    - apply local explanation to find local feature effects and identify groups of patients for which markers are important

Summary

- IML techniques helps explore, compare and maintain Machine Learning models.
- Explanations of individual predictions rely on good interpretable features.
- For tabular data, the notion of an *interpretable feature* is not clear.
- We propose a method of creating interpretable features based on conditional behaviour of the model.

More resources

- https://github.com/ModelOriented/localModel – R implementation of the described methodology.
- iBreakDown – Shapley-like explanations with first-order interactions: https://arxiv.org/abs/1903.11420
- https://github.com/olagacek/SAFE, https://github.com/ModelOriented/xspliner - tools for feature extraction from complex models.
- https://pbiecek.github.io/DALEX_docs/ - introduction to Interpretable Machine Learning and DALEX family of packages.
- https://github.com/mi2datalab - tools for IML built by MI^2 Data Lab.
- http://mi2.mini.pw.edu.pl/ - MI^2 Data Lab website.

References

- Biecek, P., 2018. DALEX: explainers for complex predictive models, Journal of Machine Learning Research 19(84):1−5, 2018.
- 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
- Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., 2013. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. arXiv:1309.6392 [stat].
- Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 4765–4774.

References

- Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.
- 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. Mateusz Staniak and Przemysław Biecek , The R Journal (2018) 10:2, pages 395-409.
- Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." arXiv Preprint arXiv:1706.07269. (2017)

References

- T. Laugel, X. Renard, M.-J. Lesot, C. Marsala, i M. Detyniecki, „Defining Locality for Surrogates in Post-hoc Interpretablity", *arXiv:1806.07498 [cs, stat]*, cze. 2018
- A. Adhikari, D. M. J. Tax, R. Satta, i M. Fath, „Example and Feature importance-based Explanations for Black-box Machine Learning Models", *arXiv:1812.09044 [cs]*, grudz. 2018.
- D. Alvarez-Melis i T. S. Jaakkola, „On the Robustness of Interpretability Methods", *arXiv:1806.08049 [cs, stat]*, cze. 2018.
- H. Fen, Tan, K. Song, M. Udell, Y. Sun, i Y. Zhang, „Why should you trust my interpretation? Understanding uncertainty in LIME predictions", *arXiv:1904.12991 [cs, stat]*, kwi. 2019.
- Molnar, C., Bischl, B., Casalicchio, G., 2018. iml: An R package for Interpretable Machine Learning. JOSS 3, 786. https://doi.org/10.21105/joss.00786

Backup slides

MI2 Data Lab

- Joins students (MSc, PhD) and researchers from University of Warsaw (top university in Poland) and Warsaw University of Technology.
- Head: Przemysław Biecek, PhD (http://biecek.pl).
- Website: https://mi2-warsaw.github.io/
  (includes information about members, grants and publications).

MI2 Data Lab: areas of research

- Interpretable Machine Learning,
- -omics and medical data analysis,
- Statistical software engineering,
- Bioinformatics,
- NLP, text mining,
- Image data,
- Data visualization.

DALEX: Explainers for Complex Predictive Models in R

Przemysław Biecek PRZEMYSLAW.BIECEK@GMAIL.COM
Faculty of Mathematics and Information Science, Warsaw University of Technology
75 Koszykowa Street, Warsaw, Poland
Samsung Research Poland

Explainable machine learning for modeling of early postoperative mortality in lung cancer*

Katarzyna Kobylińska[1 [0000−0002−0292−4982]], Tomasz Mikołajczyk[4], Mariusz Adamek[2 [0000−0002−1885−9257]], Tadeusz Orłowski[3], and Przemysław Biecek[1,4 [0000−0001−8423−1823]]

[1] University of Warsaw,Faculty of Mathematics, Informatics and Mechanics, Poland
[2] Faculty of Medicine and Dentistry, Medical University of Silesia
[3] National Institute of Tuberculosis and Lung Diseases
[4] Faculty of Mathematics and Information Science, Warsaw University of Technology

MI^2 DataLab

Warsaw, Poland http://mi2.mini.pw.edu.pl

Repositories 24 People 17 Teams 1

iBreakDown: Uncertainty of Model Explanations for Non-additive Predictive Models

A Preprint

Alicja Gosiewska
Faculty of Mathematics and Information Science
Warsaw University of Technology
alicjagosiewska@gmail.com
https://orcid.org/0000-0001-6563-5742

Przemyslaw Biecek
Faculty of Mathematics, Informatics and Mechanics
University of Warsaw
Faculty of Mathematics and Information Science
Warsaw University of Technology
przemyslaw.biecek@gmail.com
https://orcid.org/0000-0001-8423-1823

Explanations of Model Predictions with live and breakDown Packages

by Mateusz Staniak and Przemysław Biecek

Abstract Complex models are commonly used in predictive modeling. In this paper we present R packages that can be used for explaining predictions from complex black box models and attributing parts of these predictions to input features. We introduce two new approaches and corresponding packages for such attribution, namely **live** and **breakDown**. We also compare their results with existing implementations of state-of-the-art solutions, namely, **lime** (Pedersen and Benesty, 2018) which implements *Locally Interpretable Model-agnostic Explanations* and **iml** (Molnar et al., 2018) which implements *Shapley values*.

LAS: Language Agnostic System for Question Answering

3 Author(s) Dominika Basaj ; Barbara Rychalska ; Anna Wroblewska **View All Authors**

Robotic Process Automation of Unstructured Data with Machine Learning

Anna Wróblewska*,†, Tomasz Stanisławek*,†, Bartłomiej Prus-Zajączkowski*,†, Łukasz Garncarek†
*Faculty of Mathematics and Information Science, Warsaw University of Technology
ul. Koszykowa 75, Warszawa, Poland
†Applica.ai
ul. Wiślana 8, Warszawa, Poland

Article

Prediction of Signal Peptides in Proteins from Malaria Parasites

Michał Burdukiewicz [1], Piotr Sobczyk [2], Jarosław Chilimoniuk [3], Przemysław Gagat [3], Paweł Mackiewicz [3,*]

1. Faculty of Mathematics and Information Science, Warsaw University of Technology, 00-661 Warszawa, Poland; michalburdukiewicz@gmail.com
2. Department of Mathematics, Wrocław University of Technology, 50-370 Wrocław, Poland; Piotr.Sobczyk@pwr.edu.pl
3. Department of Genomics, University of Wrocław, 50-383 Wrocław, Poland; jaroslaw.chilimoniuk@gmail.com (J.C.); przemyslaw.gagat@uwr.edu.pl (P.G.)

* Correspondence: pamac@smorfland.uni.wroc.pl

TRIM28 and Interacting KRAB-ZNFs Control Self-Renewal of Human Pluripotent Stem Cells through Epigenetic Repression of Pro-differentiation Genes

Urszula Oleksiewicz [17] • Marta Gładych [17] • Ayush T. Raman [17] • Holger Heyn • Elisabetta Mereu • Paula Chlebanowska • Anastazja Andrzejewska • Barbara Sozańska • Neha Samant • Katarzyna Fąk • Paulina Auguścik • Marcin Kosiński • Joanna P. Wróblewska • Katarzyna Tomczak • Katarzyna Kulcenty • Rafał Płoski • Przemysław Biecek • Manel Esteller • Parantu K. Shah • Kunal Rai • Maciej Wiznerowicz • Show less • Show footnotes

What is an explanation?

- An answer to a „Why?” question (Miller 2017).
- Helps understand the model.
- Can be:
  - a plot,
  - a summary statistic,
  - an observation (an example),
  - a model,
  - a model parameter.