In Search of Interpretable Features to Explain Decisions of Black Box Models

Mateusz Staniak (Warsaw University of Technology)

Joint work with Przemysław Biecek (MI^2 Data Lab)

Google Flu Trends

5 years of web logs + Machine Learning

=

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*

WHAT WE CAN LEARN FROM THE EPIC FAILURE OF GOOGLE FLU TRENDS

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

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'**
- **Amazon now uses a 'much-watered down version' of the recruiting engine to help with some rudimentary chores, such as removing duplicate resumes**

By REUTERS
PUBLISHED: 04:02 GMT, 10 October 2018 | **UPDATED:** 16:55 GMT, 10 October 2018

Share

994
shares

156
View comments

Amazon's 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.

But the firm was ultimately forced to end the project after it found the system had taught itself to prefer male candidates over females.

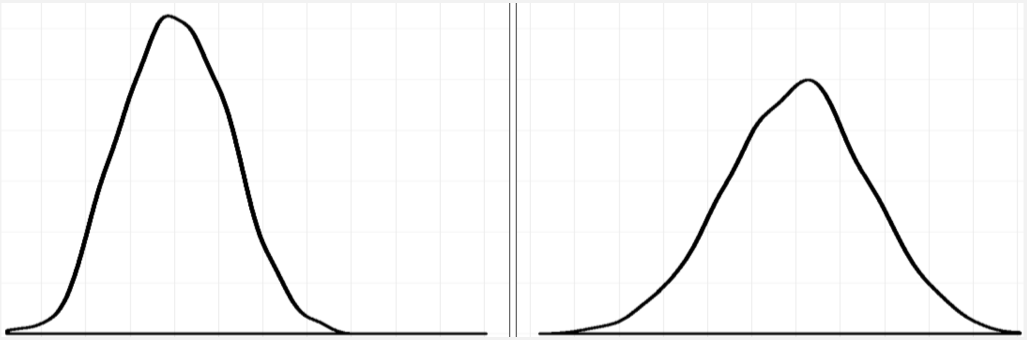
https://www.dailymail.co.uk/sciencetech/article-6259205/Amazon-scraps-secret-AI-recruiting-tool-showed-bias-against-women.html

Machine Learning models are vulnerable to:

- Biased training

Females vs Male frequencies

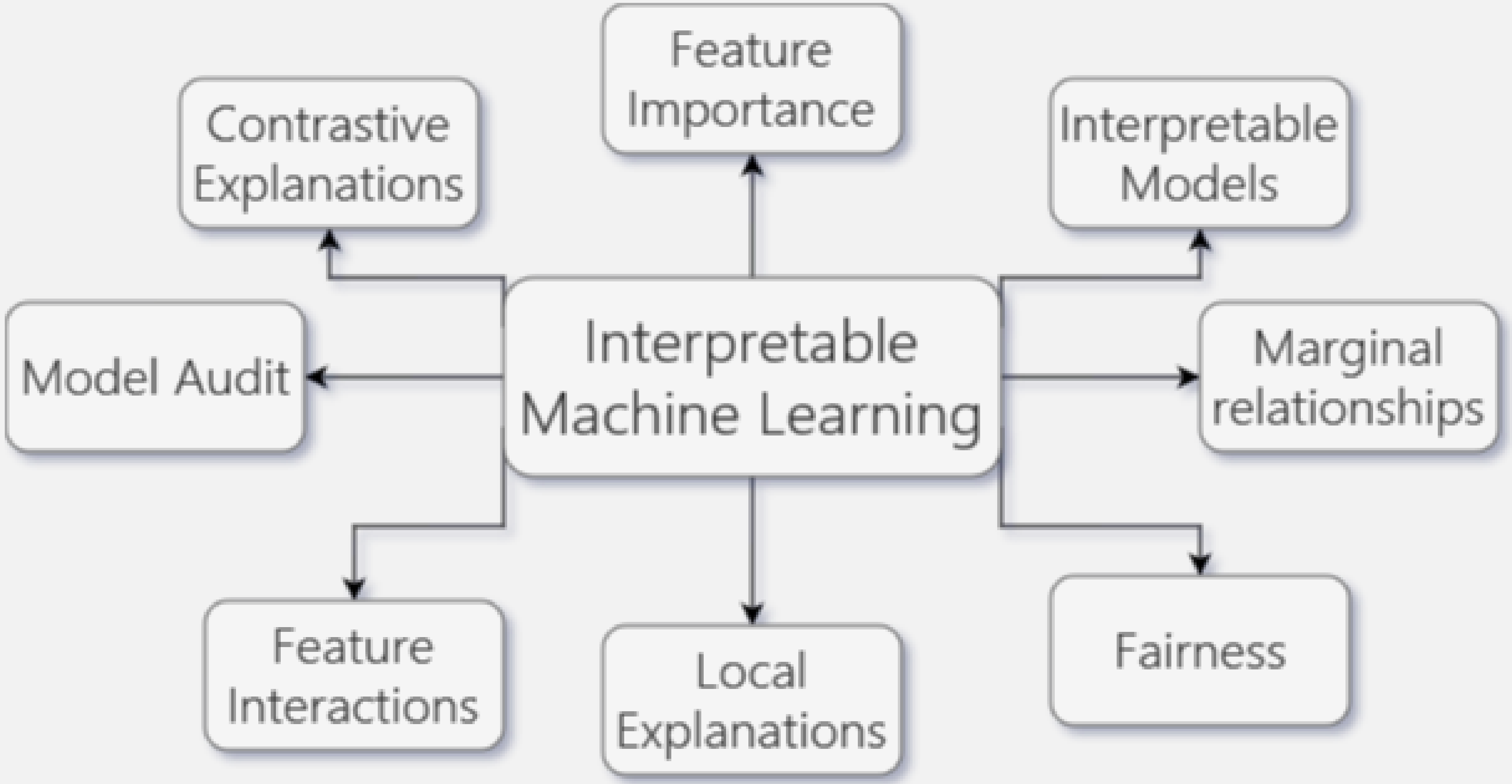
- Concept drift



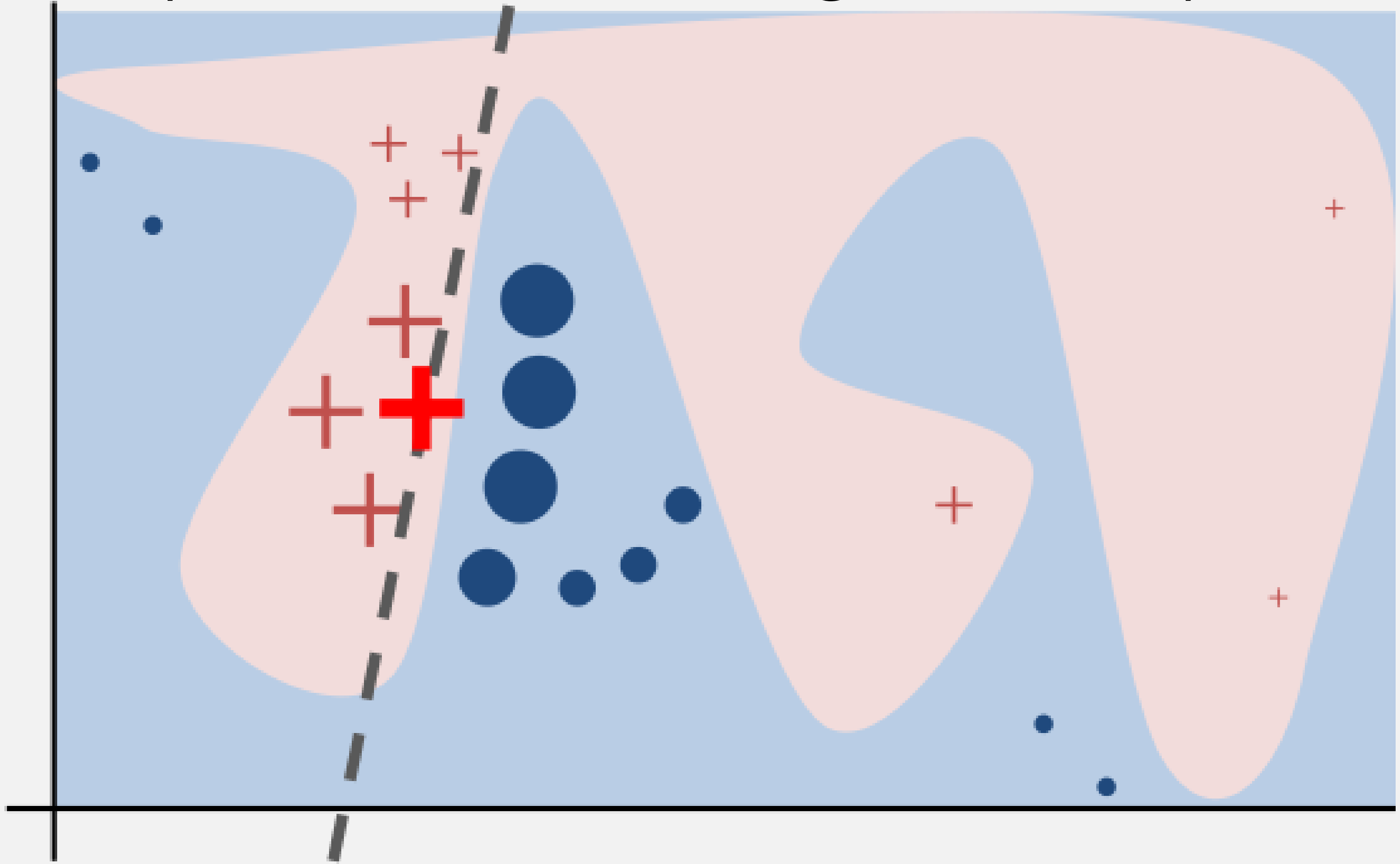
Training data vs validation data distribution

- Unmeasurable objectives

Fairness, Lawfulness, ...


Feature Importance
Contrastive Explanations
Interpretable Models
Model Audit
Interpretable Machine Learning
Marginal relationships
Feature Interactions
Local Explanations
Fairness

LIME
(Local Interpretable Model-Agnostic Explanations)



Interpretable Features

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

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


SUPPORT
VECTOR
MACHINES
REPEA
POWE
LAWS
Map Reduce
Map Reuse
Map Recycle
BAN
GENETIC
ALGORITHMS
AGAI


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


REPEA
POWE
LAWS
GAP
AGAI
FREE
BAN
GENETIC


SUPPORT
VECTOR
MACHINES
REPEA
POWE
LAWS
GAP
AGAI
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

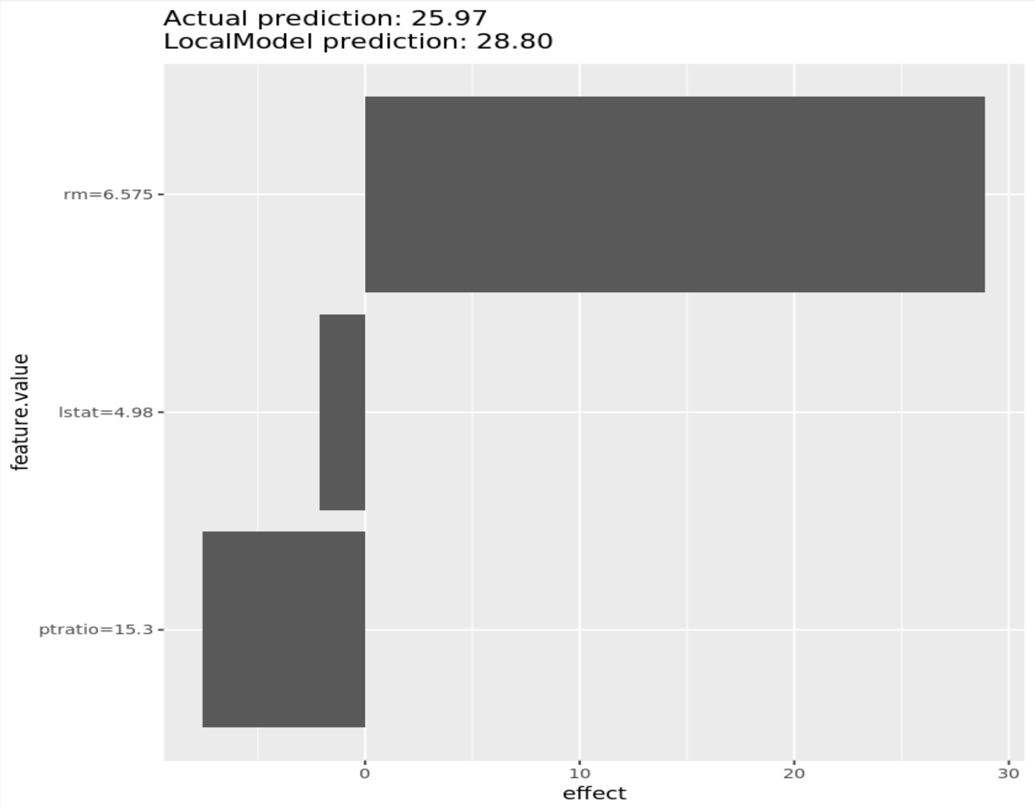
A Region	A Country	# Distance	# Points	A Glider
PACA 25% Rhône-Alpes 13% Other (28) 61%	France 96% Espagne 2% Other (18) 2%	25 1.25k	21.2 1.2k	Pegase 16% Duo Discus 8% Other (178) 77%
	Afrique du Sud	568.280029	511.959991	Duo Discus X
	Afrique du Sud	398.700012	402.049988	JS1 18m
	Namibie	482.380005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

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

Existing Approaches for Tabular Data

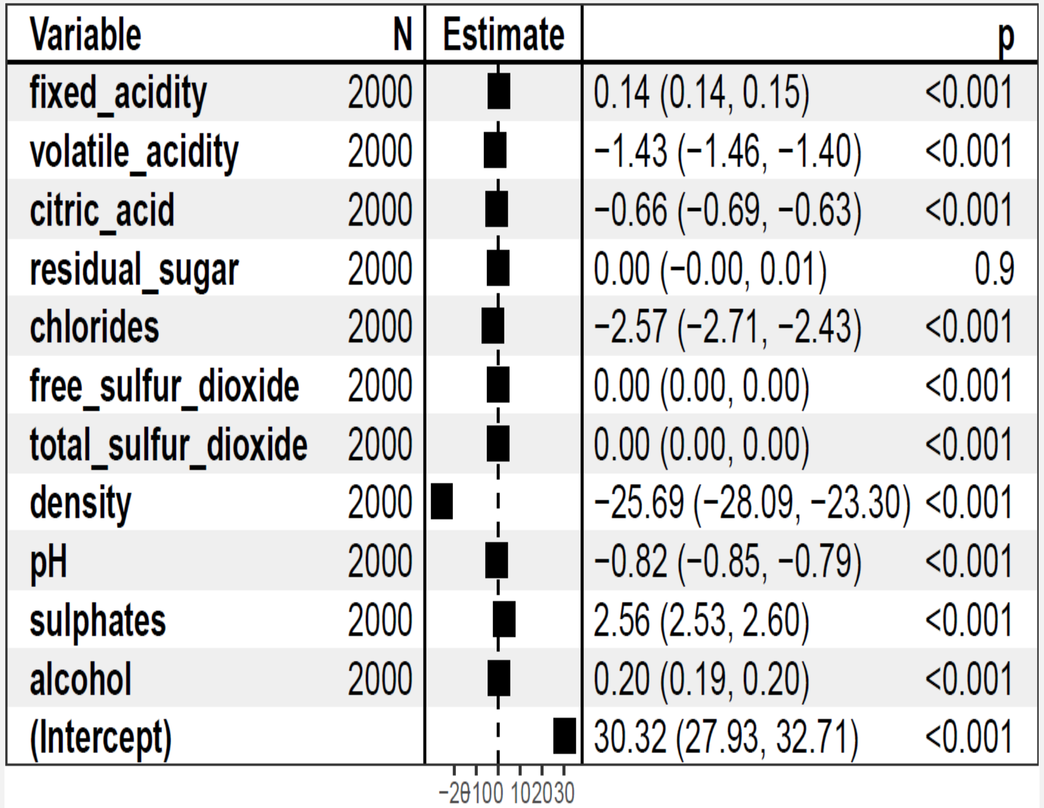
Continuous features

- iml package (R) – Christoph Molnar



https://github.com/christophM/iml

- live package (R) – Mateusz Staniak

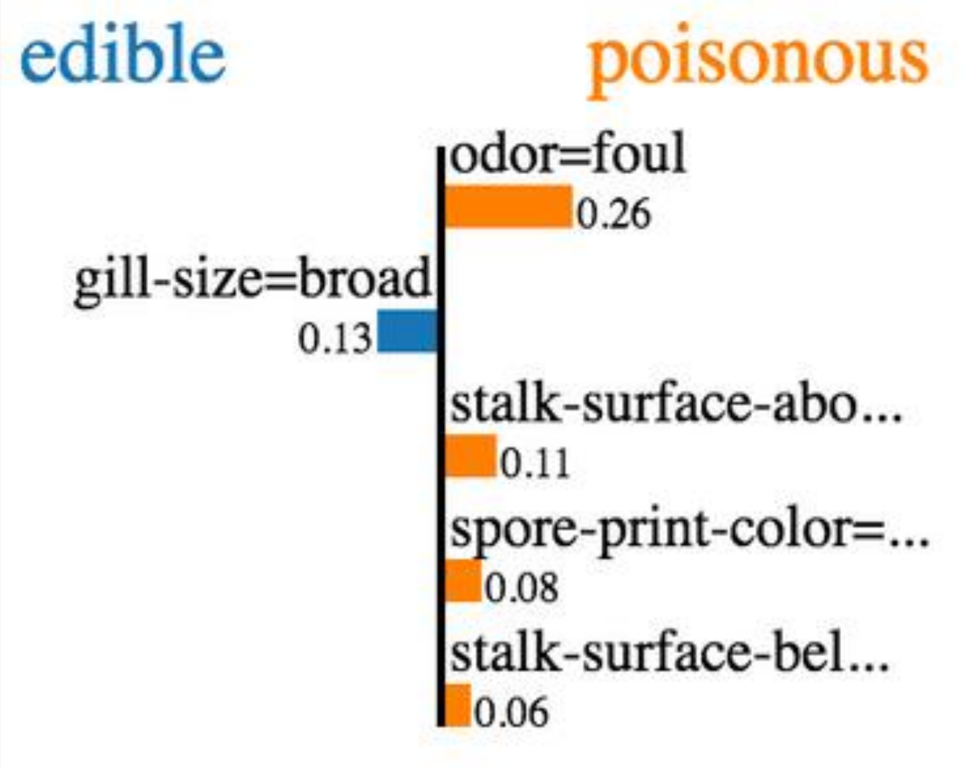


Variable	N	Estimate		p
fixed_acidity	2000		0.14 (0.14, 0.15)	<0.001
volatile_acidity	2000		−1.43 (−1.46, −1.40)	<0.001
citric_acid	2000		−0.66 (−0.69, −0.63)	<0.001
residual_sugar	2000		0.00 (−0.00, 0.01)	0.9
chlorides	2000		−2.57 (−2.71, −2.43)	<0.001
free_sulfur_dioxide	2000		0.00 (0.00, 0.00)	<0.001
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

https://github.com/MI2DataLab/live

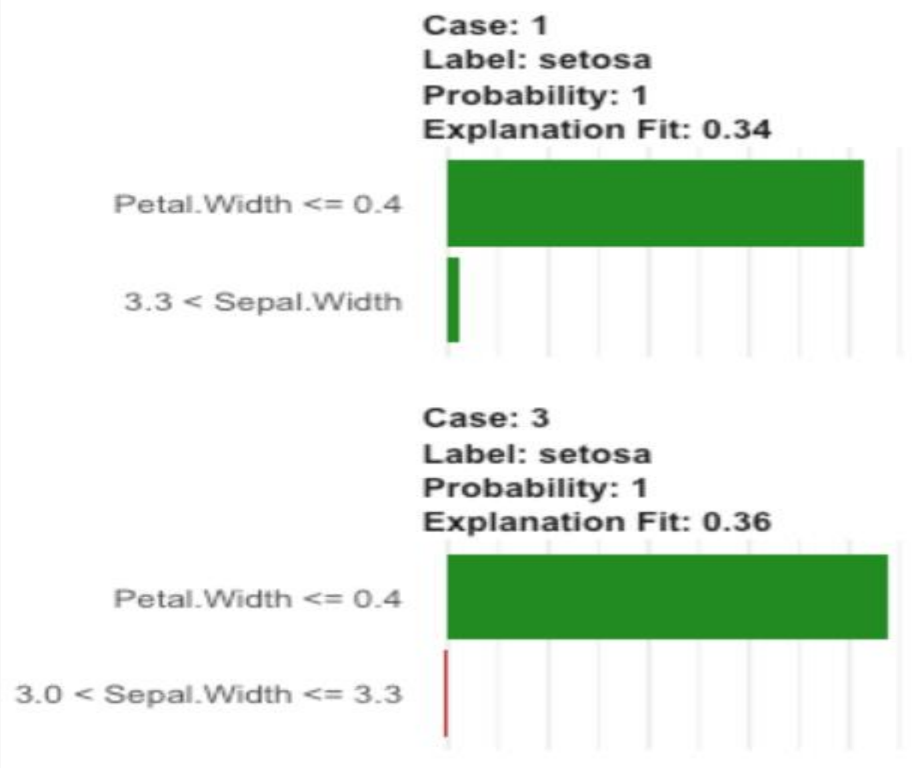
Discretized features

- lime library (Python) – Marco Tulio Ribeiro



https://github.com/marcotcr/lime

- lime package (R) – Thomas Lin Pedersen

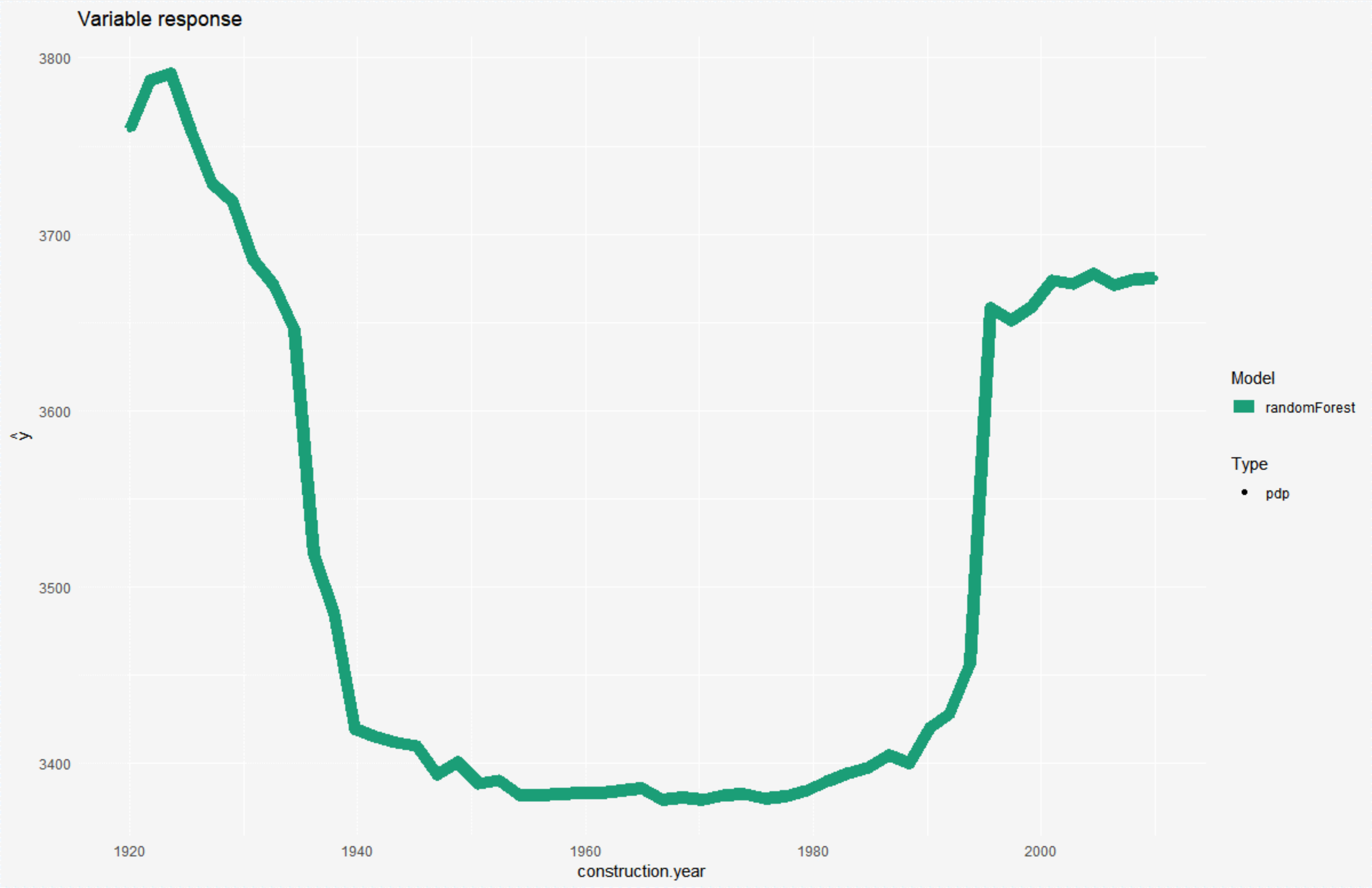


https://github.com/thomasp85/lime

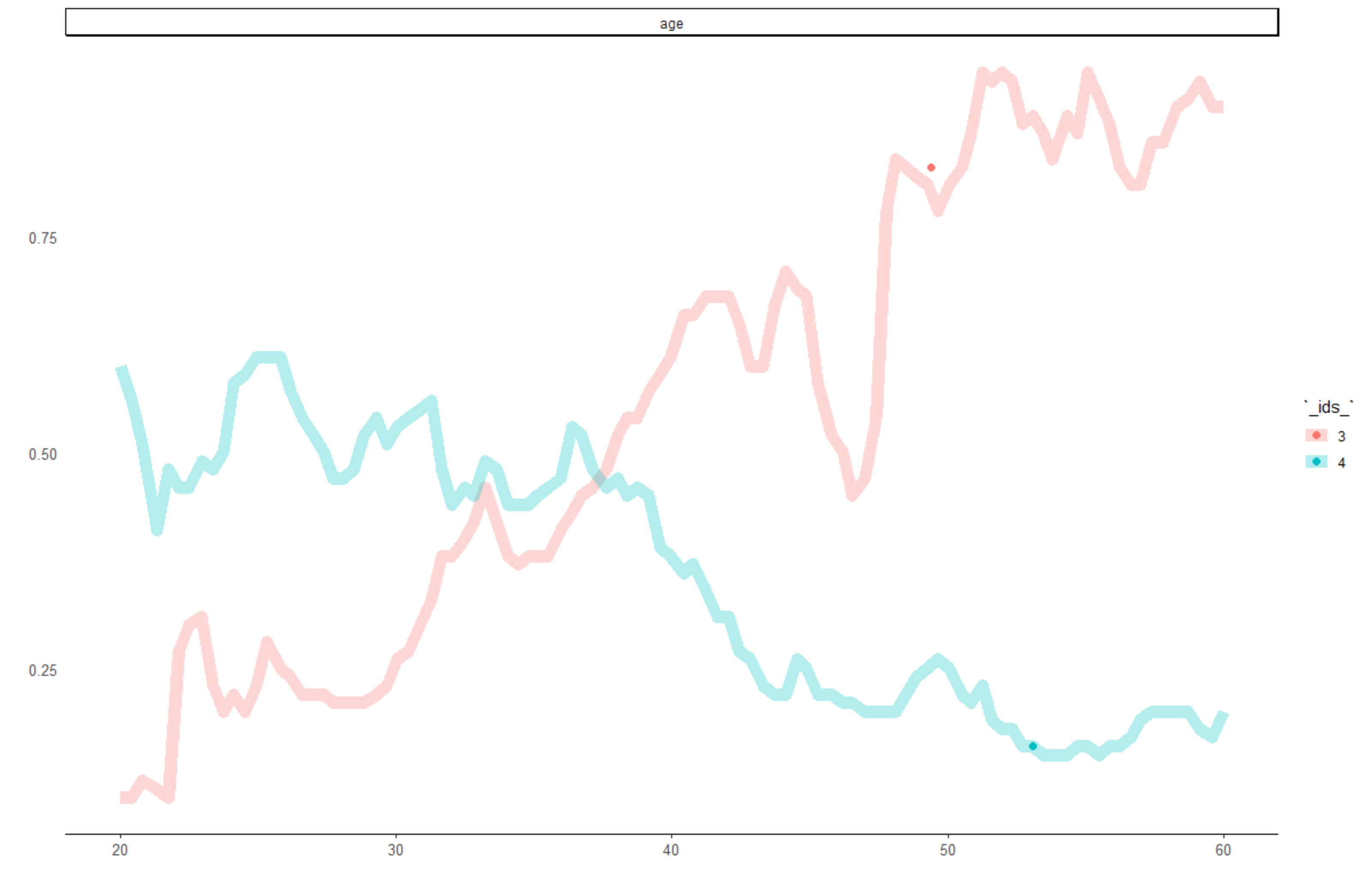
A New Approach:

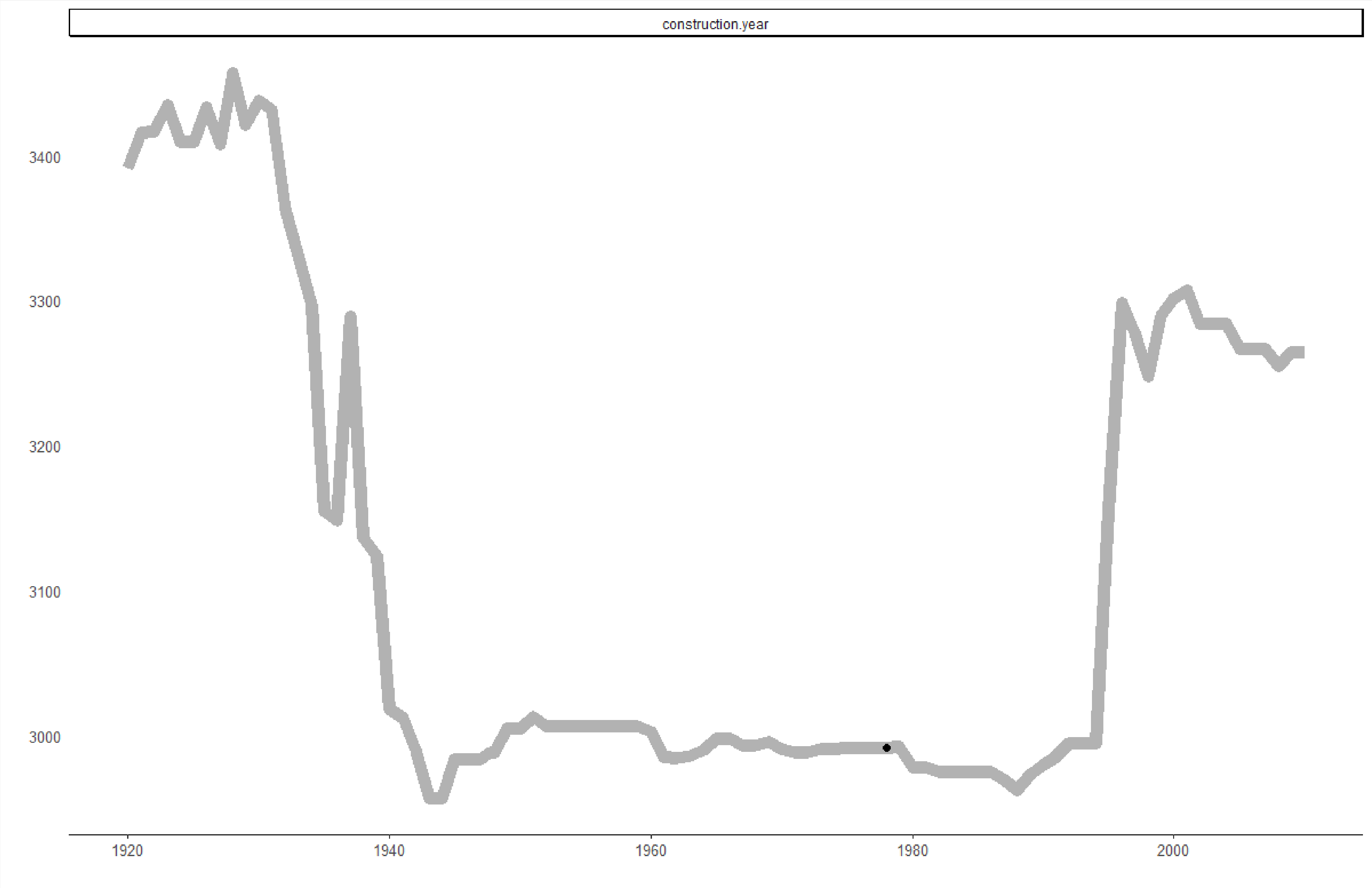
Use our knowledge about the model behaviour

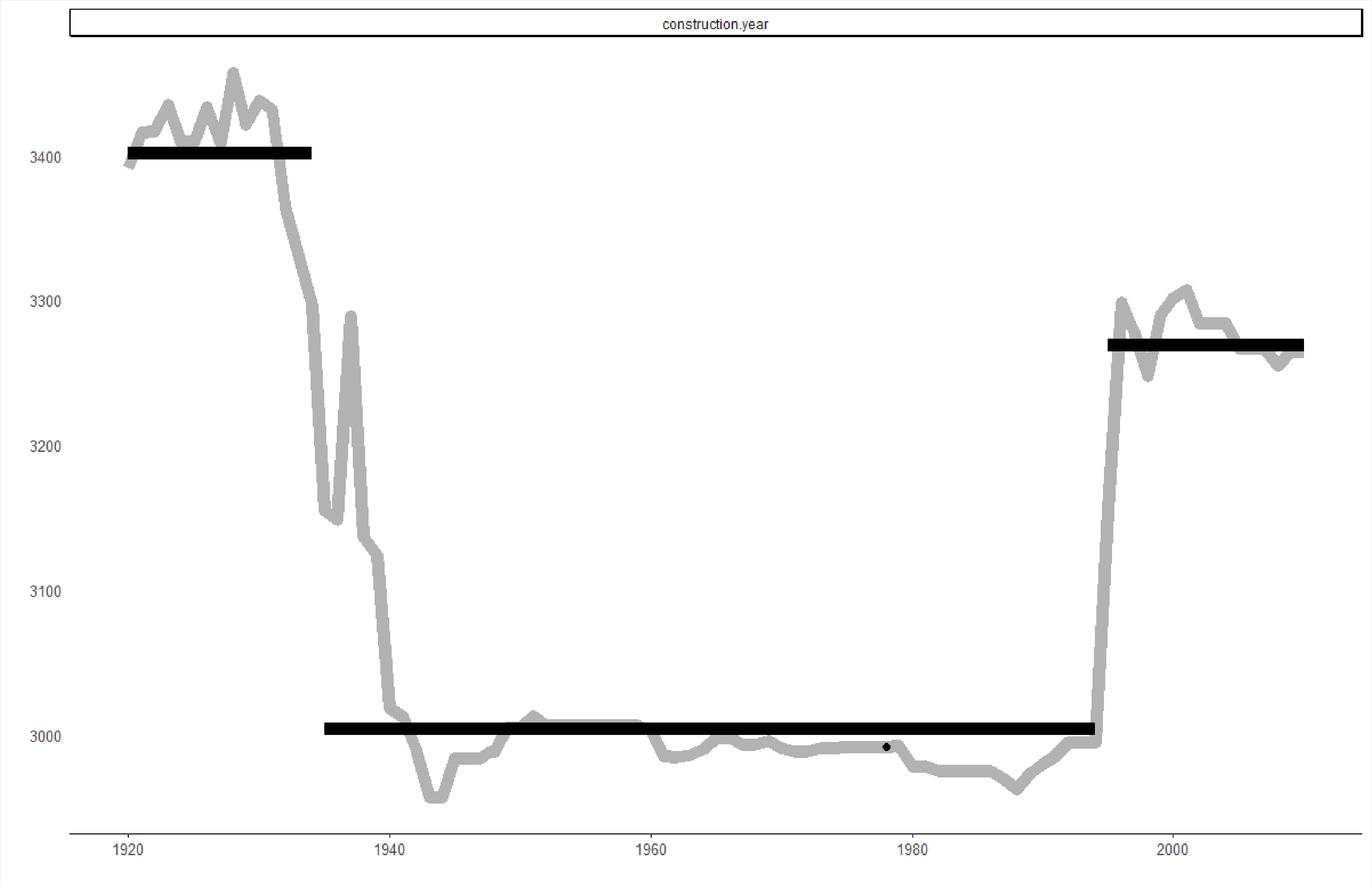
Partial Dependence Plots

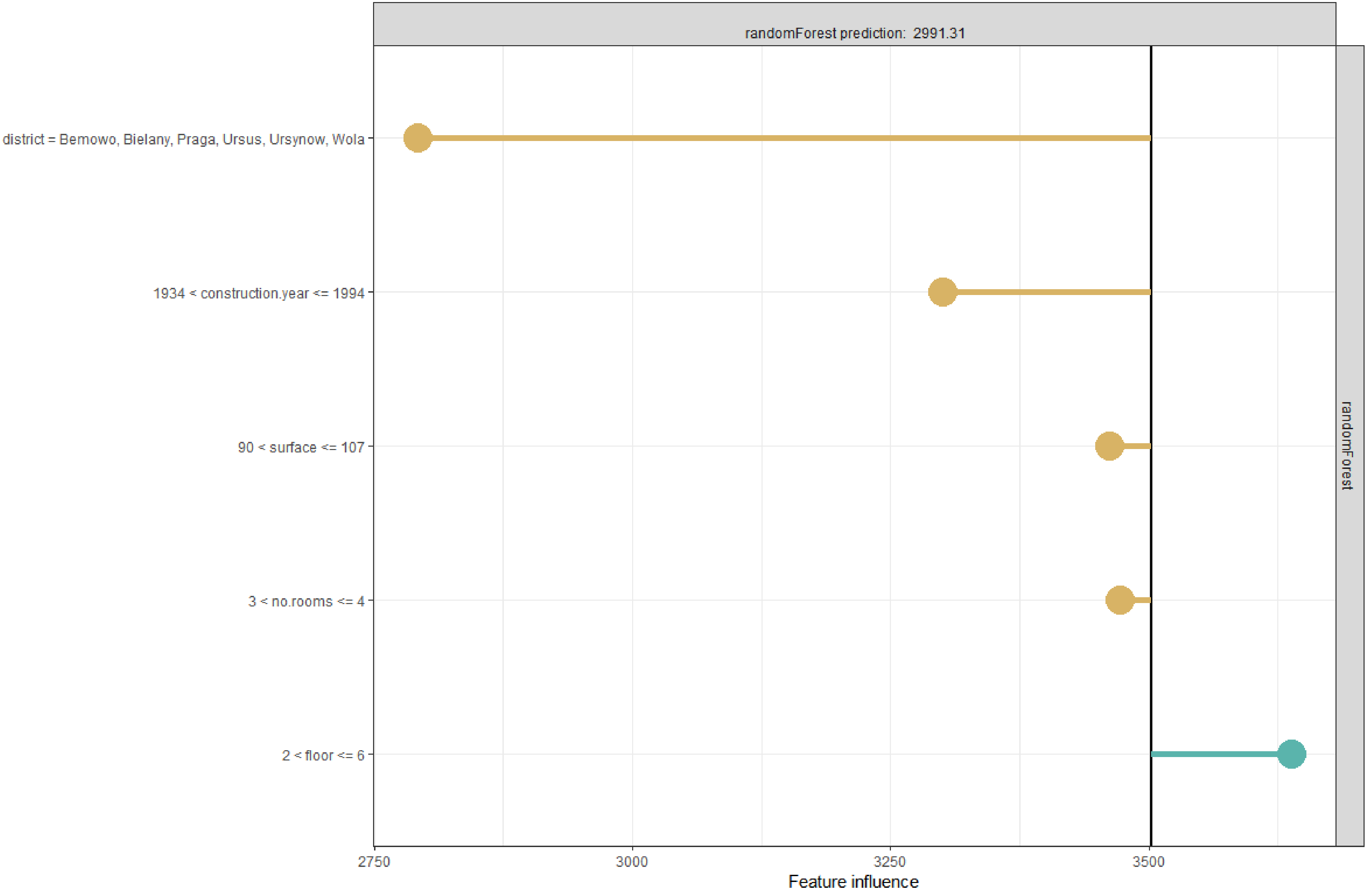


Ceteris Paribus Profiles




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
2750
3000
3250
3500
Feature influence
randomForest

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 described methodology.
- https://github.com/mi2datalab - tools for IML built by MI^2 Data Lab.
- https://pbiecek.github.io/DALEX_docs/ - introduction to Interpretable Machine Learning and DALEX family of packages.
- http://mi2.mini.pw.edu.pl/ - MI^2 Data Lab website.
- https://github.com/olagacek/SAFE, https://github.com/ModelOriented/xspliner - tools for feature extraction from complex models.

Acknowledgement

This work was financially supported by NCN Opus grant 2016/21/B/ST6/02176 and by NCN Opus grant 2017/27/B/ST6/01307.

References

- 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].
- 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
- Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "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, KDD '16. ACM, New York, NY, USA, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778
- Staniak, M., Biecek, P., 2018. Explanations of model predictions with live and breakDown packages. arXiv:1804.01955 [cs, stat].