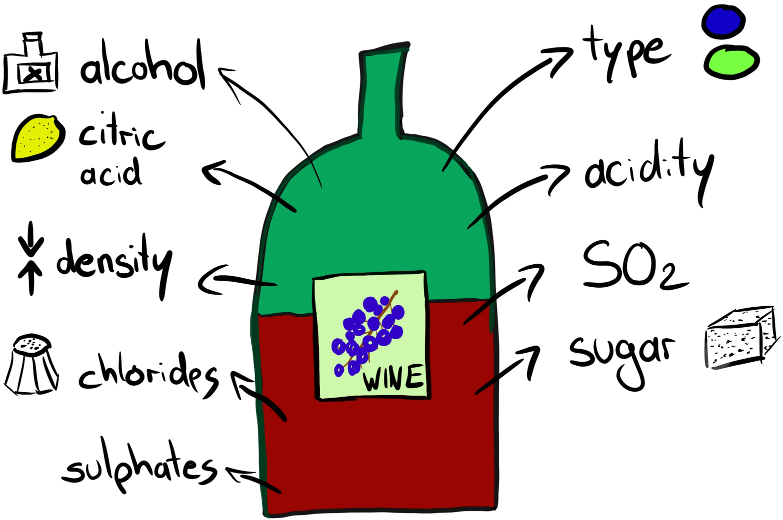
Explainable AI (XAI)

Christoph Molnar

13. Institutstag - July 05, 2019

Let's Predict Wine Quality



alcohol
citric acid
density
chlorides
sulphates
type
acidity
SO_2
sugar
WINE

Disclaimer: I am NOT a wine expert!

How can we develop a wine quality prediction program?

Programing vs. Machine Learning

Machine Learning (supervised)

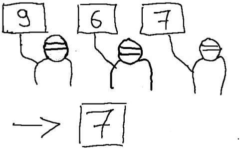
Machine Learning (supervised)

Machine Learning (supervised)

Step 1: Find data

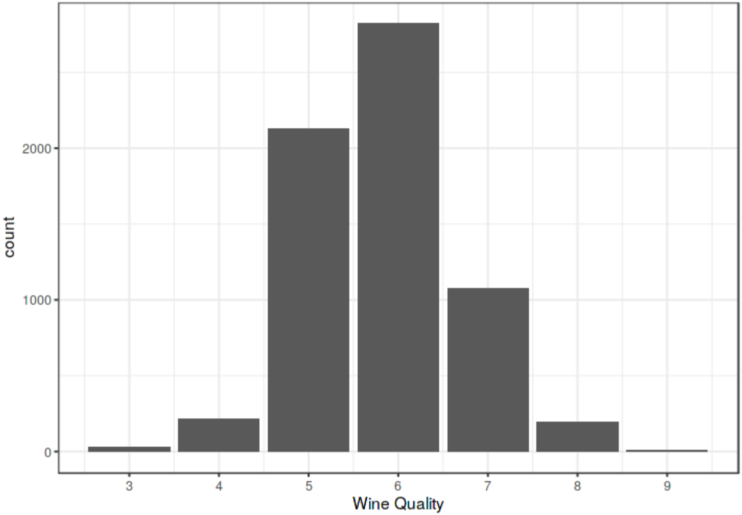
Wine Dataset

- 6500 red and white Portuguese "Vinho Verde" wines
- Features: Physicochemical properties
- Quality assessed by blind tasting, from 0 (very bad) to 10 (excellent)



P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Wine Quality Distribution



Step 2: Apply Machine Learning

Random Forest

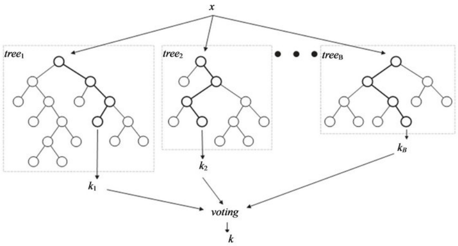
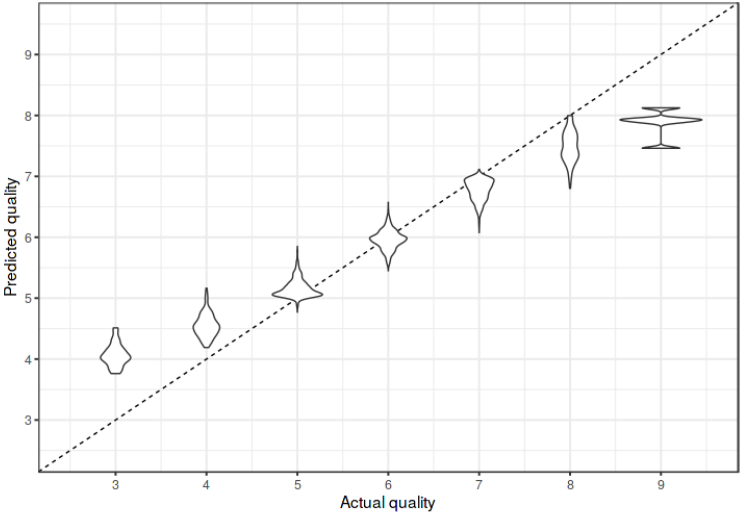


Image: http://www.hallwaymathlete.com/2016/05/introduction-to-machine-learning-with.html

Train Random Forest to Predict Quality

Mean absolute error on test data (cross-validated): 0.44

Prediction vs. Actual Quality



Step 3: Profit

We want to know:

- Which wine properties are the most predictive for quality?
- How does a property affect the predicted wine quality?
- Can we extract a “Rule of Thumb” from the black box?
- Why did a wine get a certain prediction?
- How do we have to change a wine to achieve a different prediction?

Looking inside the black box

Which features are important?

Permutation Feature Importance

original

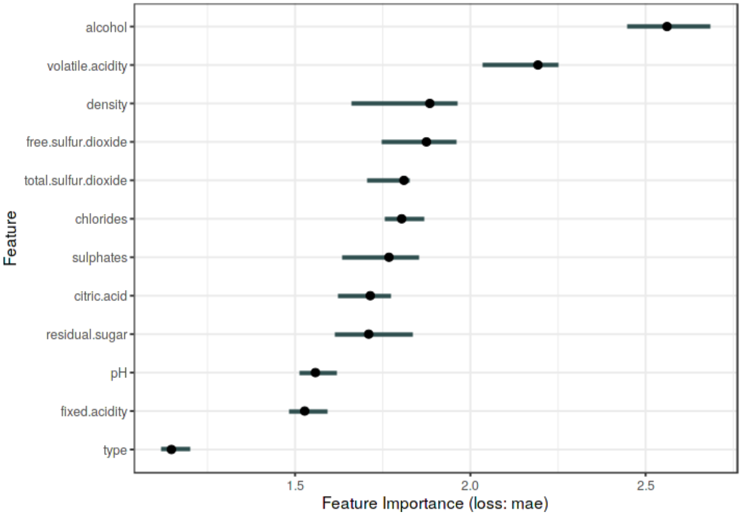
x_1	$\dots$	x_j	$\dots$	x_p
3		1.4		6.0
5		1.2		7.2
$\dots$		$\dots$		$\dots$
6		2.0		8.9

$\Rightarrow$

shuffled x_j

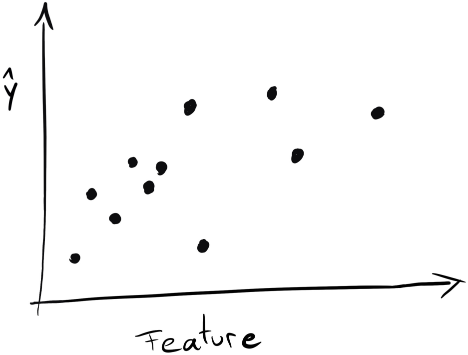
x_1	$\dots$	x_j	$\dots$	x_p
3		2.0		6.0
5		1.4		7.2
$\dots$		$\dots$		$\dots$
6		1.2		8.9

Which features are important?

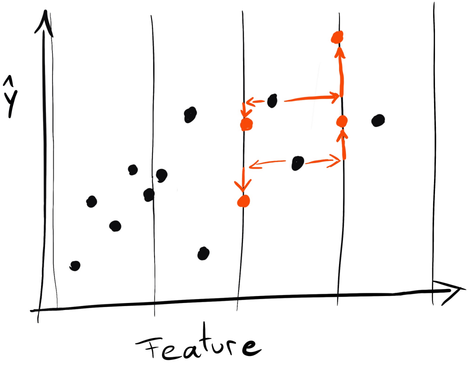


How do features affect predictions?

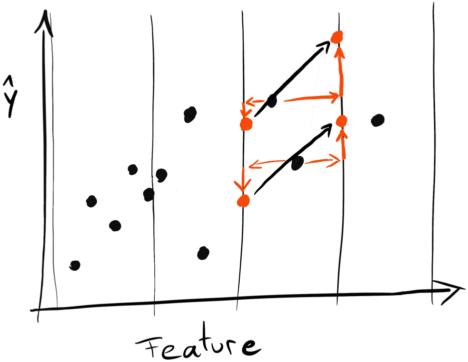
Accumulated Local Effects



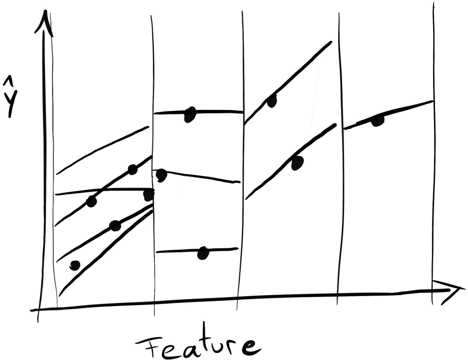
Accumulated Local Effects



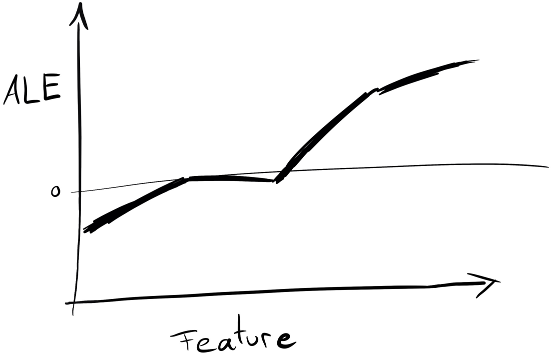
Accumulated Local Effects



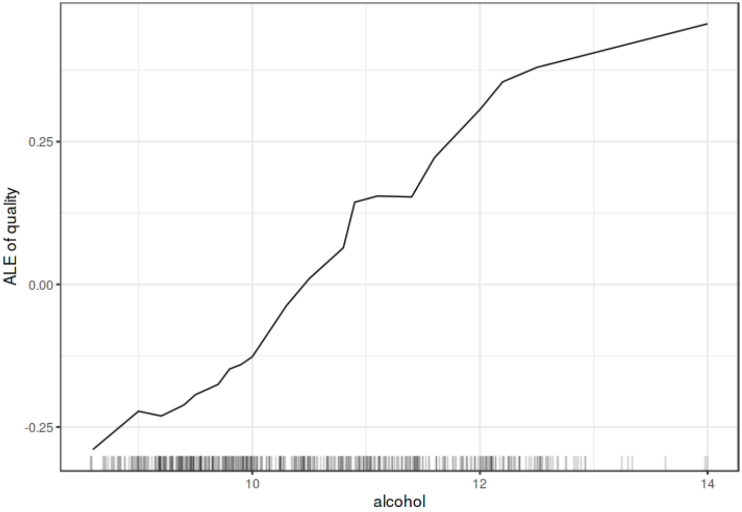
Accumulated Local Effects



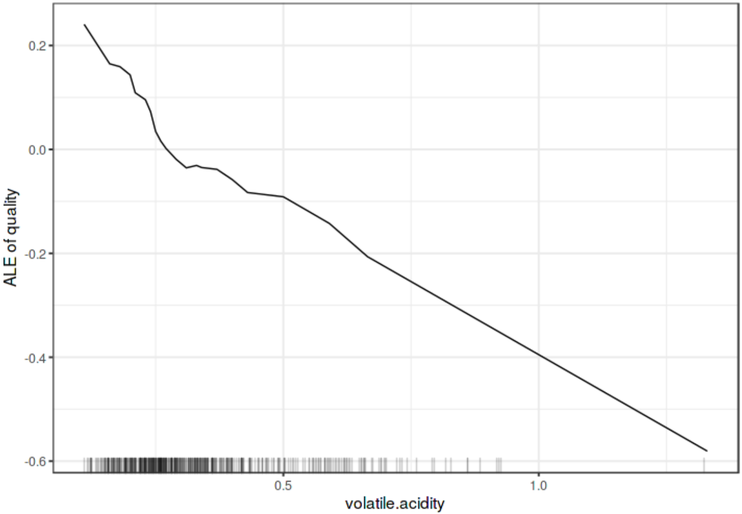
Accumulated Local Effects



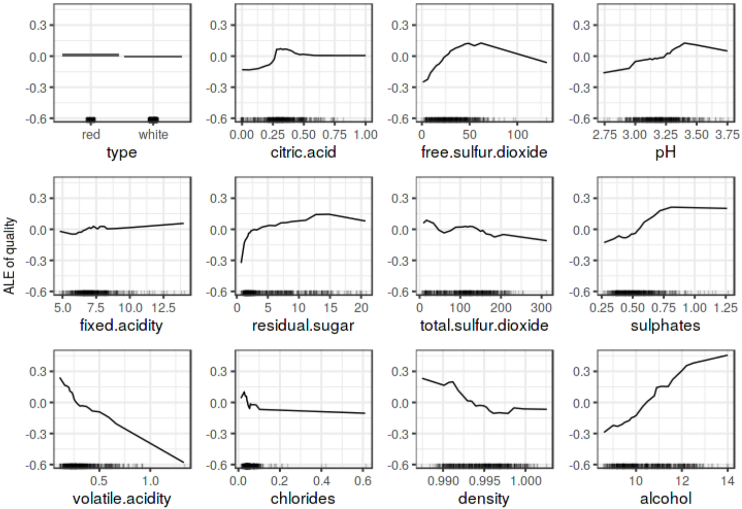
Effect of Alcohol



Effect of Volatile Acidity



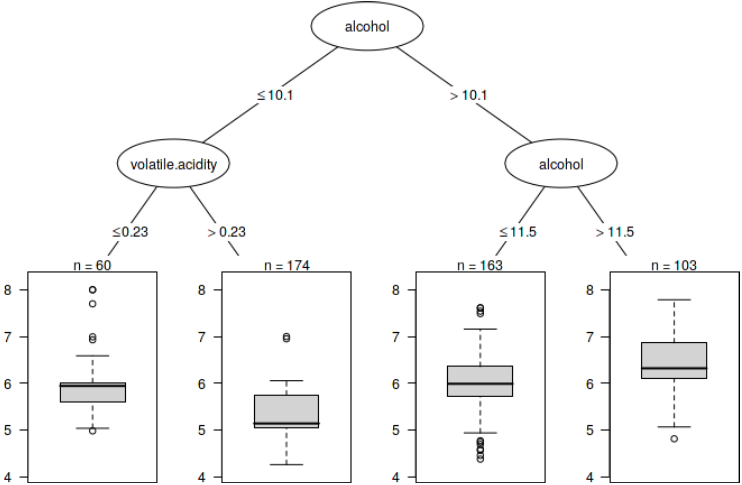
How do features affect predictions?



Rule of thumb for wine quality?

Surrogate Model

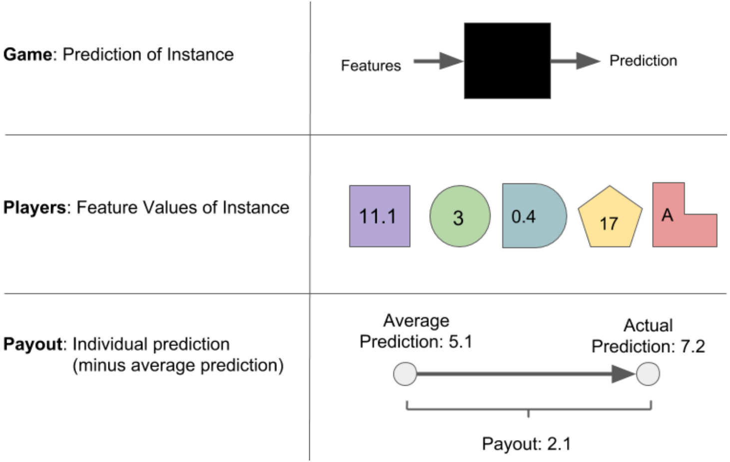
Surrogate Model



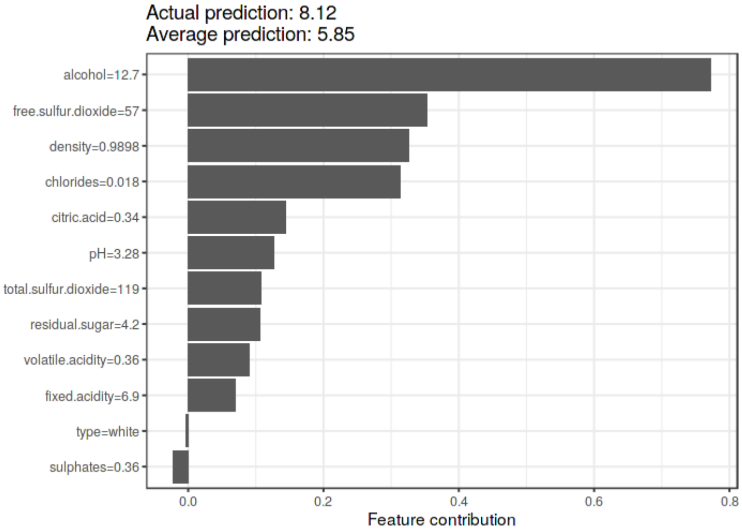
Tree explains 37.36% of black box prediction variance.

Explain individual predictions

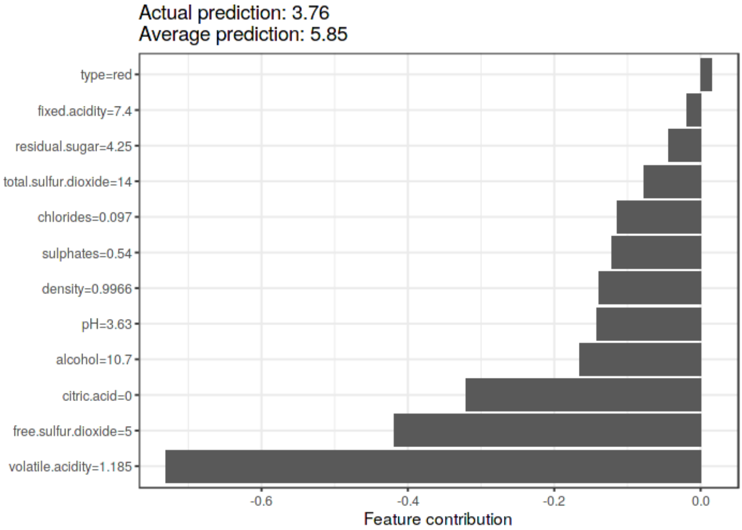
Shapley Value



Explain best wine



Explain worst wine



Improve worst wine?

Counterfactual Explanations

Counterfactual Explanations

Counterfactual Explanations

Improve worst wine?

How do we get the wine above predicted quality of 5?

- Decreasing volatile acidity to 0.2 yields predicted quality of 5.09
- Decreasing volatile acidity to 1.0 and increasing alcohol to 13% yields predicted quality of 5.01

Why interpretability?

Interested in learning more?

More on interpretable machine learning in my book
http://christophm.github.io/interpretable-ml-book/.

Backup

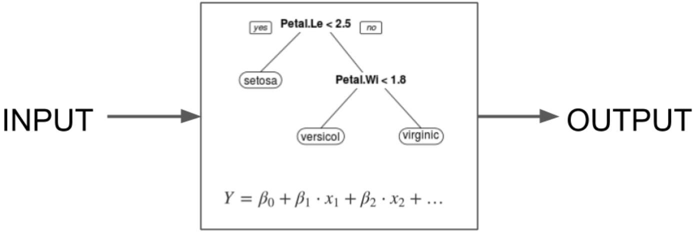
Units in Wine dataset

- fixed acidity g(tartaric acid)/dm^3
- volatile acidity: g(acetric acid/dm^3)
- citric acid: g/dm^3
- residual sugar: g/dm^3
- chlorides: g(sodium chloride)/dm^3
- free sulfur dioxide: mg/dm^3
- total sulfur dioxide: mg/dm^3
- density$>$ g/cm^3
- pH
- sulphates: g(postassium sulphate) / dm^3
- alcohol vol.%

What tools do we have?

Interpretable Models

Interpretable Models



Intepretable Model: Linear Regression

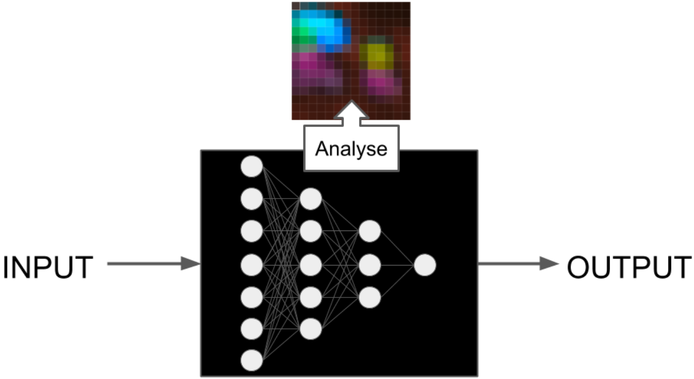
Intepretable Model: Decision Tree

Interpretable Model: Decision Rules

IF $90m^2 \leq$ size $< 110m^2$ AND location $=$ "good" THEN rent is between 1540 and 1890 EUR

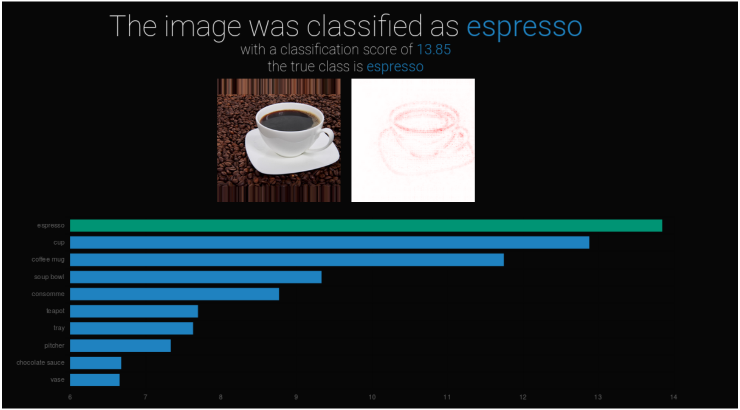
Model-specific Methods

Model-specific Methods



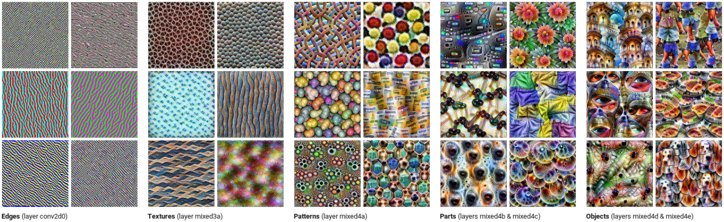
Model-specific Methods

Layerwise Relevance Propagation (LRP)



Bach, Sebastian, et al. “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation.” PloS one 10.7 (2015): e0130140.

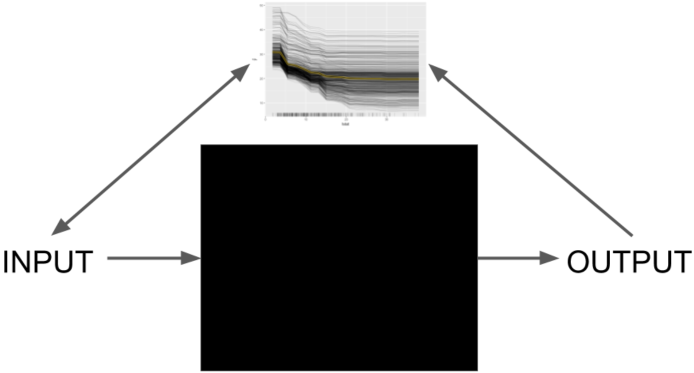
Model-specific Methods



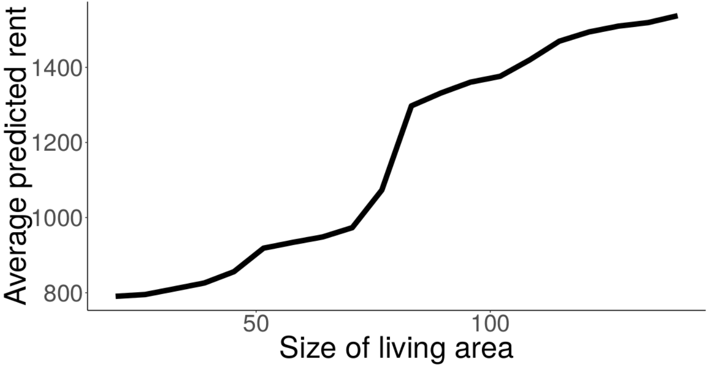
https://distill.pub/2017/feature-visualization/

Model-agnostic Methods

Model-agnostic Methods



Model-agnostic Methods



Model-agnostic Methods: Global Surrogate

Model-agnostic Methods: Local Surrogate