

Interpretable machine learning

Daniel Kapitan | version January 2021

Attribution and copyright notice

This lecture is based on the following material available in the commons:

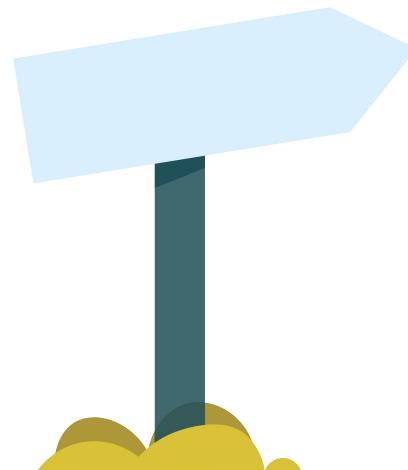
- [Interpretable Machine Learning](#), by Christoph Molnar (referenced as IML)
- Several open access journal papers (referenced individually)

© by [Daniel Kapitan](#), *Interpretable Machine Learning*.

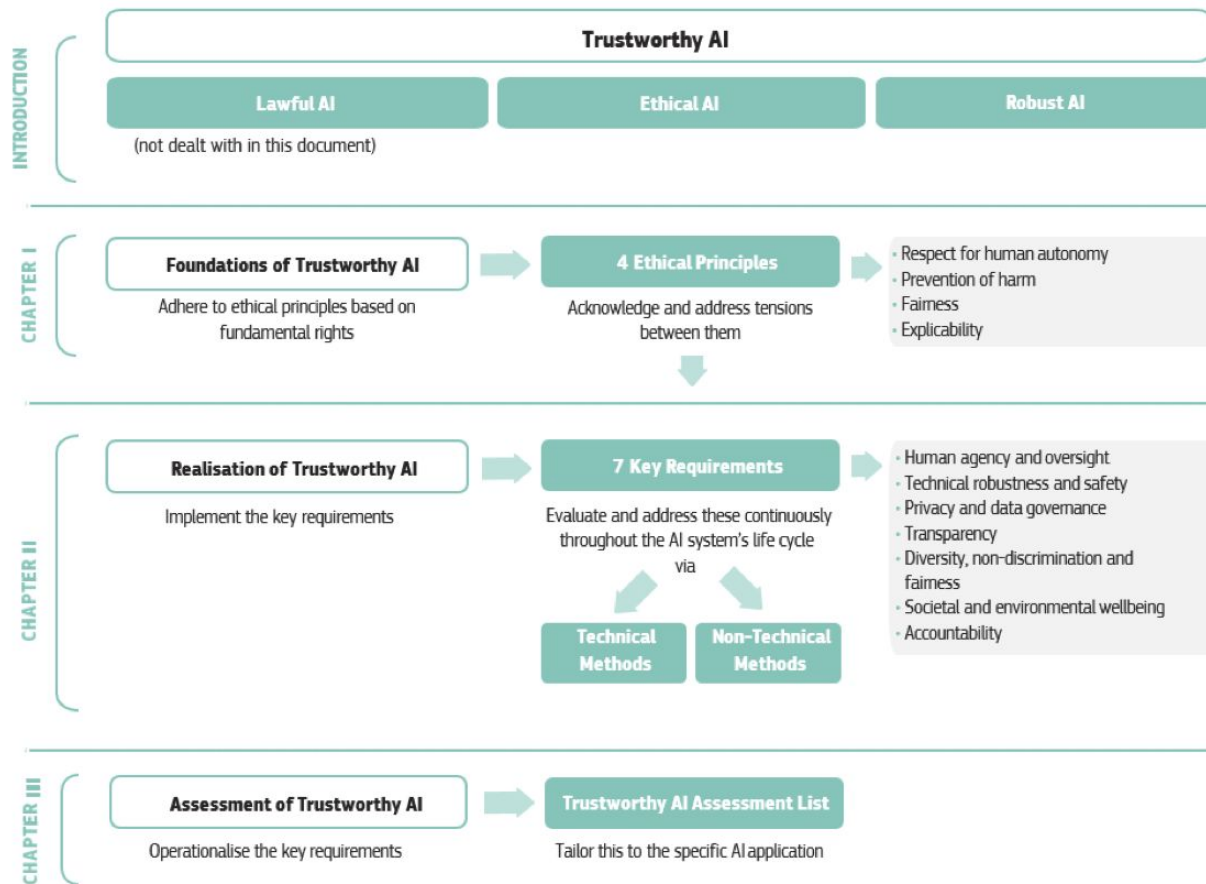
This work is licensed under a
[Creative Commons Attribution-ShareAlike 4.0 International License](#).

Learning objectives

- _ Understand different concepts of interpretability and their strengths and weaknesses
- _ Know how to apply general model-agnostic methods for interpreting black box machine learning models for tabular data
- _ Know how to choose which interpretation method is most suitable for your machine learning project



European guidelines for trustworthy AI



Today's scope: interpretability

7 Key Requirements

Implications for interpretability

Human agency and oversight

Human-friendly explanations to facilitate oversight

Technical robustness and safety

Reliability and reproducibility of results

Privacy and data governance

Quality and integrity of data

Transparency

Traceability, explainability and communication

Diversity, non-discrimination and fairness

Avoid unfair bias and stakeholder participation

Societal and environmental wellbeing

n/a

Accountability

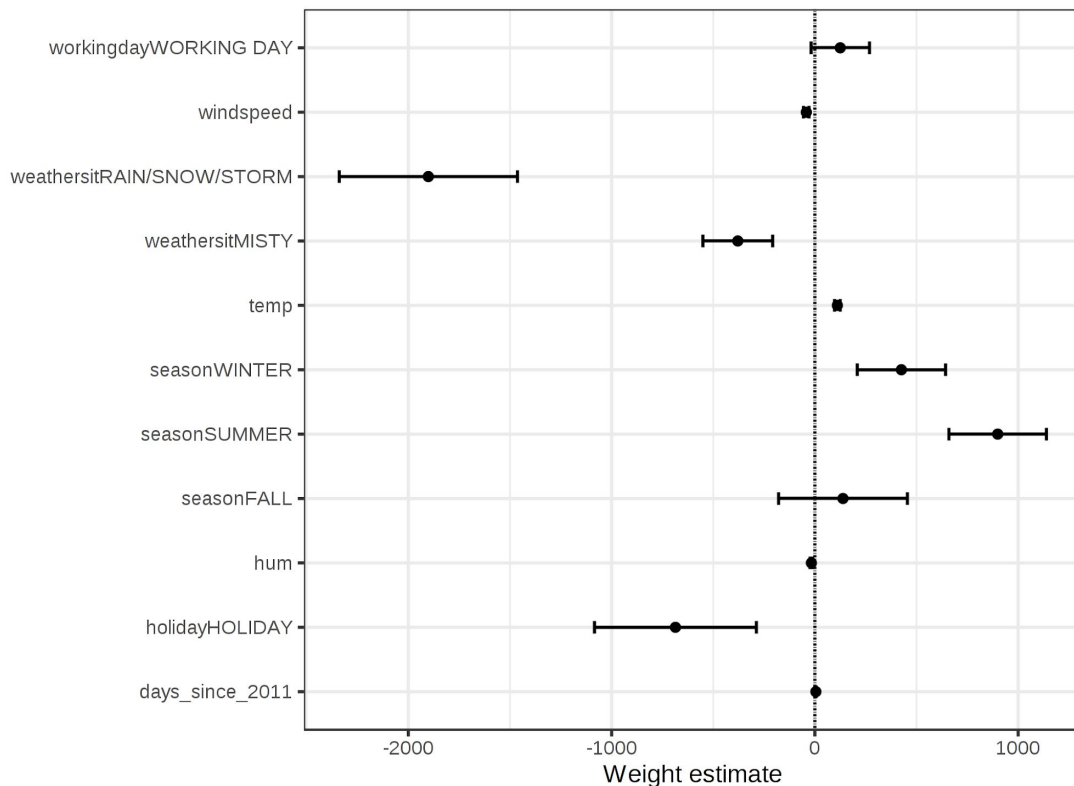
Auditability of results

Good, human-friendly explanations

Explanations are	Implications for interpretable machine learning
... contrastive	Use examples
... selective	Make explanations short, dare to simplify
... social	Involve domain experts!
... focused on the abnormal	Highlight outliers in features
... truthful	Minimise prediction errors
... consistent with prior beliefs of the explainee	Difficult to integrate in modeling, e.g. non-linearity vs. monotonicity
... general and probable	Use feature support as measure for generality

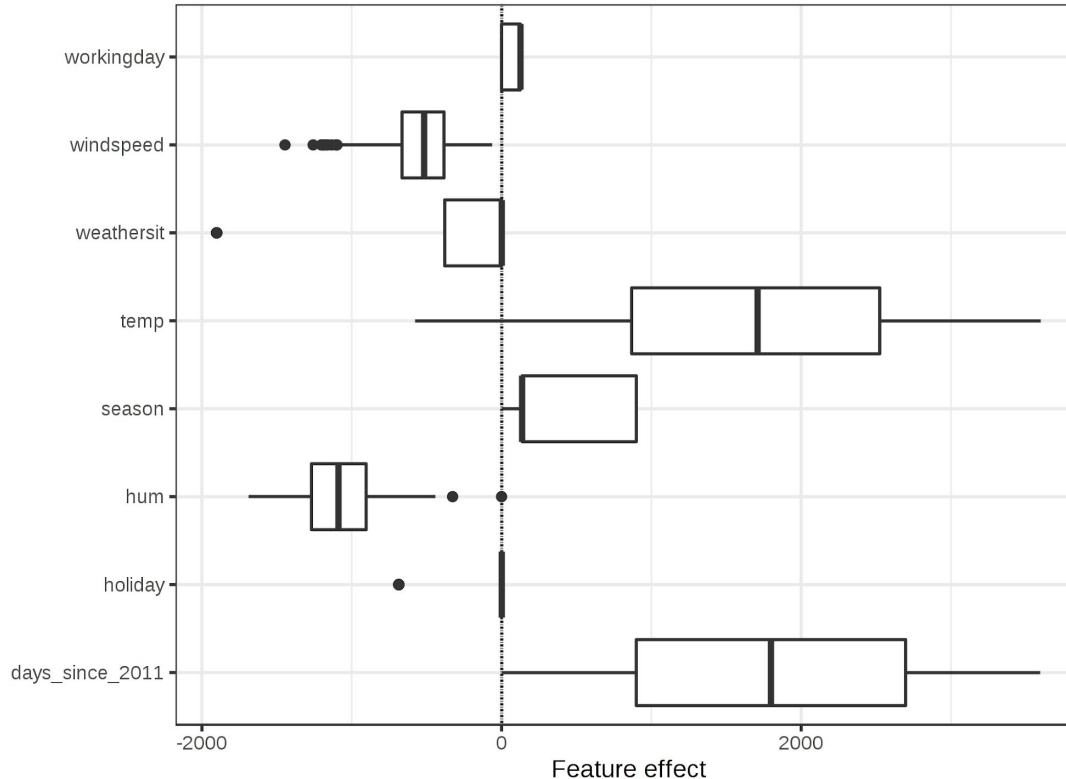
Interpretable models (easy)

Linear regression: weight plot



- Shows weights with confidence intervals
- Works better with normalized features!

Linear regression: effect plot



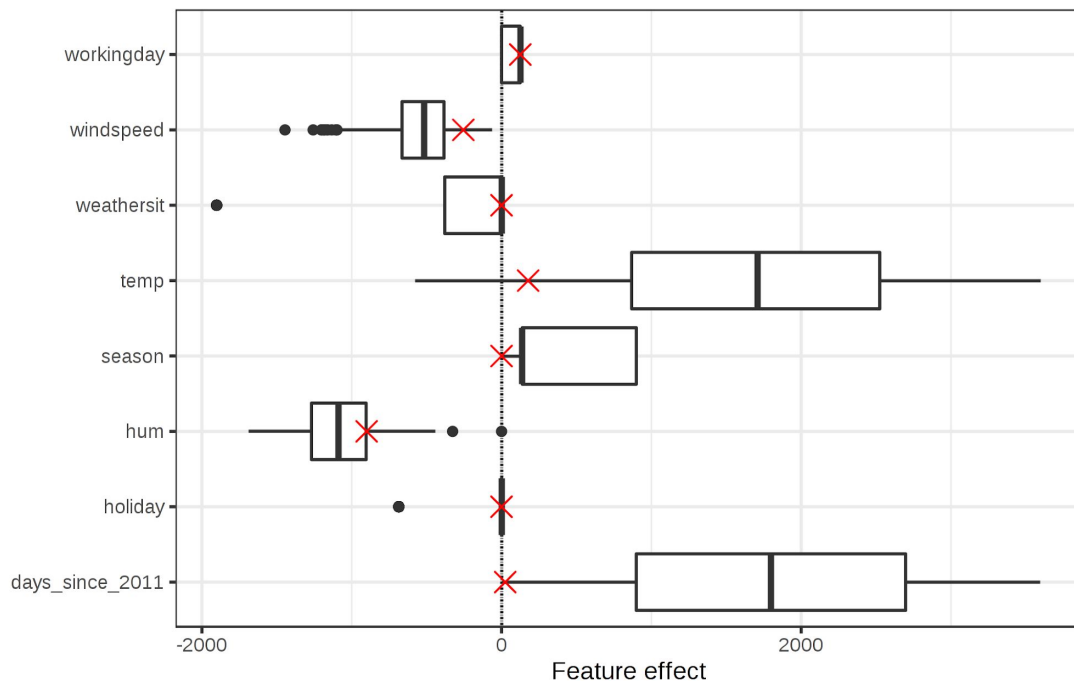
- Shows *effect* = $\text{weights} * \text{actual feature values}$

Linear regression: effect plot with individual prediction

Predicted value for instance: 1571

Average predicted value: 4504

Actual value: 1606



- Shows how feature effects influence individual predictions compared to average prediction

Interpretability of linear models

Advantages

Transparent how predictions are generated

Selective explanation with Lasso

Widely accepted

Guarantee that you have optimal weights (as long as assumptions of linear models is met)

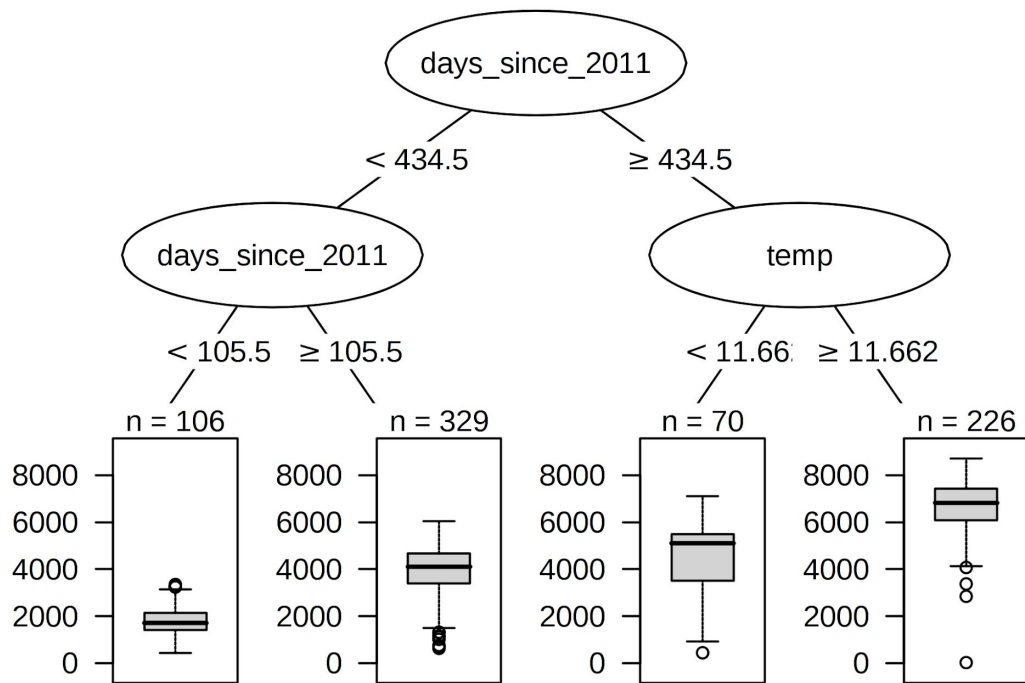
Disadvantages

Non-linearity or interactions have to be added manually

Poor predictive performance

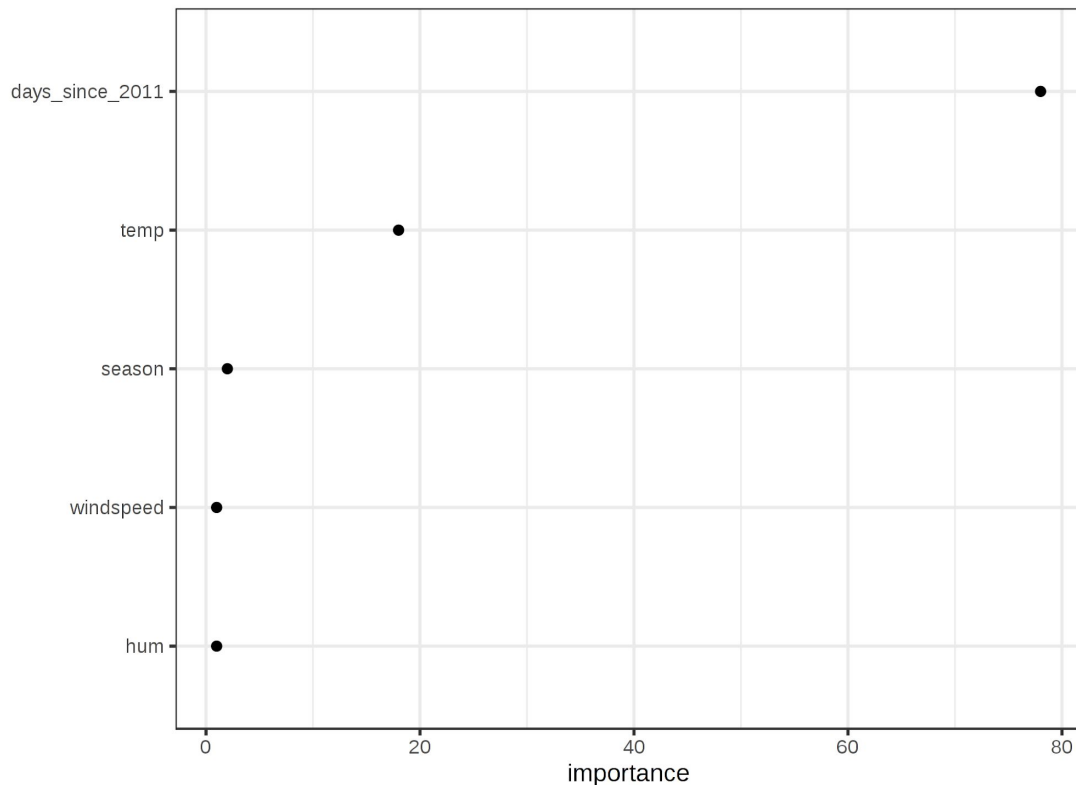
Interpretation can be unintuitive for correlated features

Decision tree



- So what has bigger influence: time of temperature?

Decision tree: feature importance



Beware Default Random Forest Importances

- Default method based on mean decrease in impurity → can be very wrong
- Better solution: permutation importance
- Python library: [rfpimp](#)
- Credit: Prof [Terence Parr](#), University of San Francisco

Interpretability of decision trees

Advantages

Ideal for capturing interactions

Predictions are made in groups

Tree structure is intuitive visualisation

Trees create good human-friendly explanations

Disadvantages

Linear relationships are not shown efficiently (multiple steps in nodes)

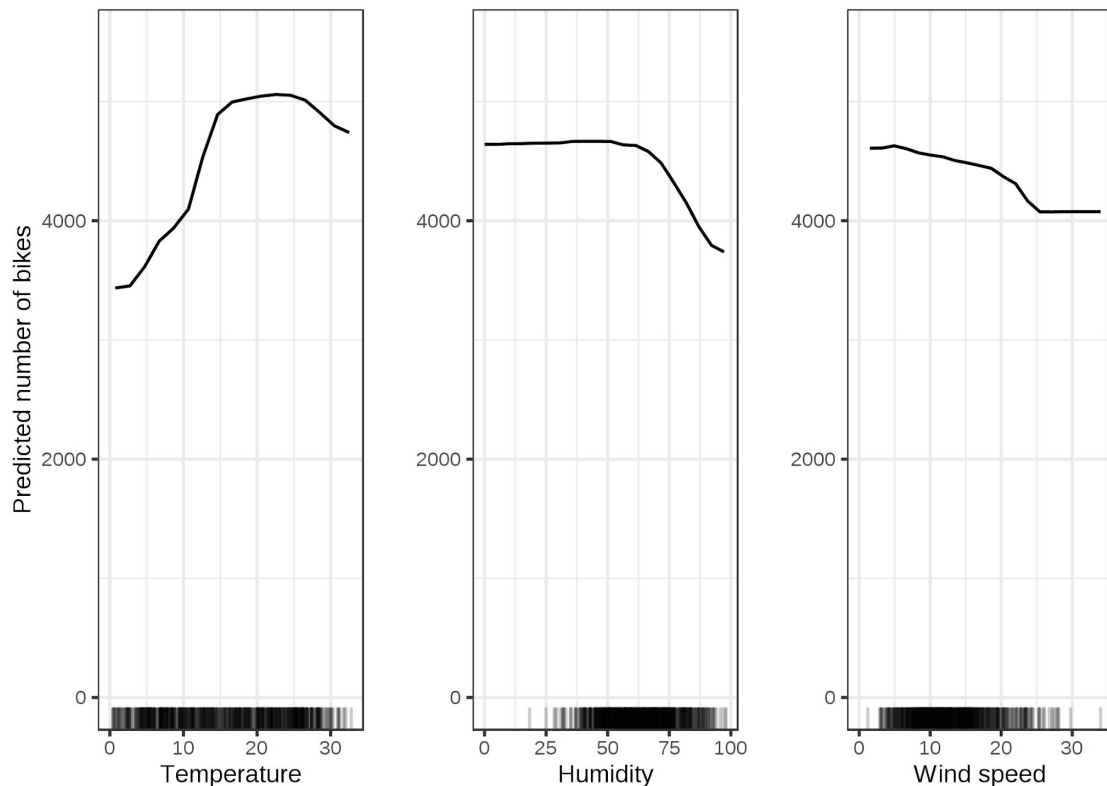
Lack of smoothness

Tree are unstable (inherent to greedy algorithm)

Long (deep) tree as not easily interpreted

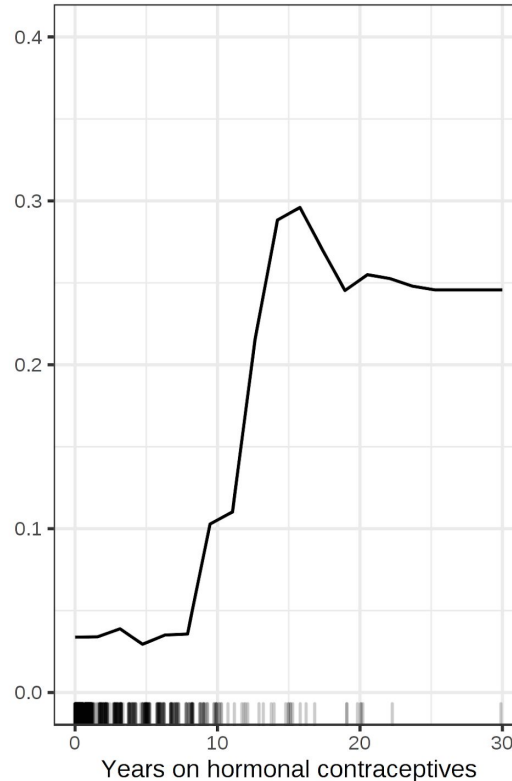
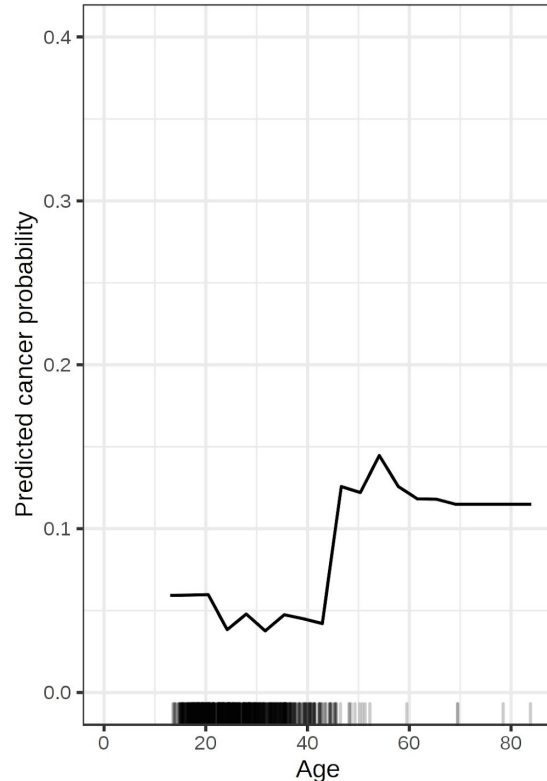
Model-agnostic Methods (intermediate)

Partial dependence plot (PDP): regression



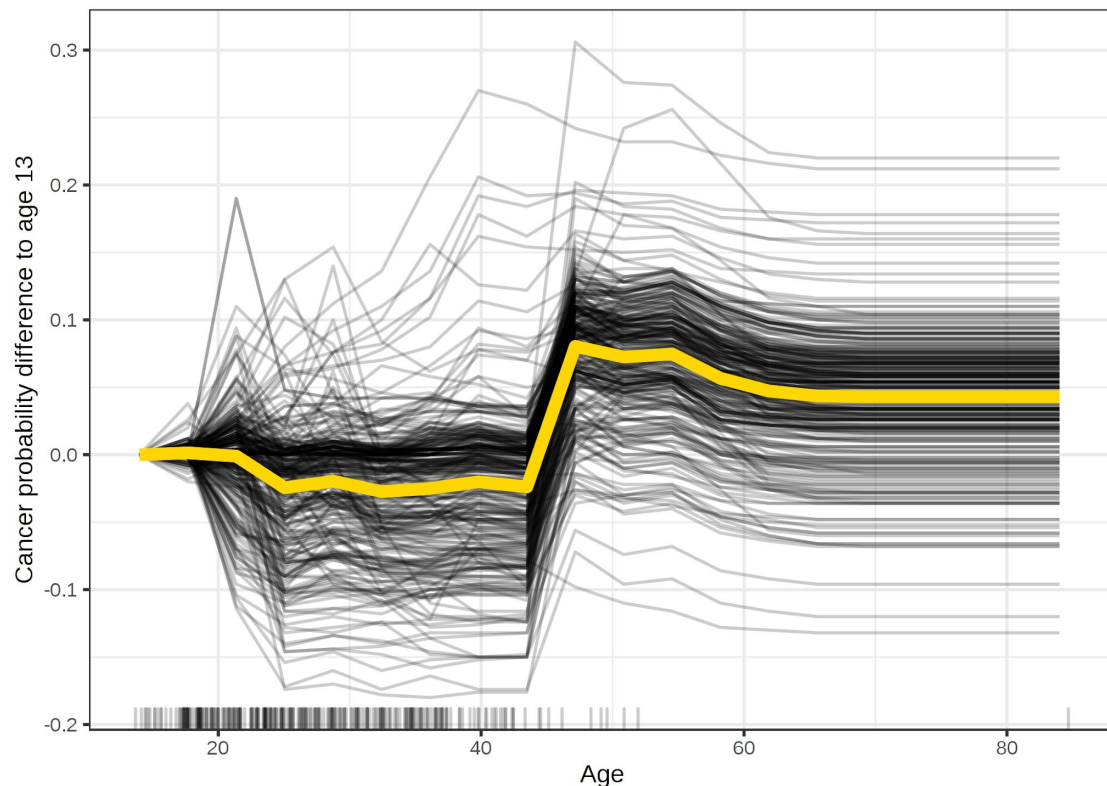
- Select features X_S from the total set of features X_1, \dots, X_P
- Define X_C as the complementary set of features
- The partial dependence plot is defined as:
$$\hat{f}_{x_S}(x_S) = E_{x_C} [\hat{f}(x_S, x_C)] = \int \hat{f}(x_S, x_C) d\mathbb{P}(x_C)$$
- Partial dependence plots are calculation using Monte Carlo simulations

Partial dependence plot (PDP): classification



- Similar approach with predicted probability on the y-axis

Individual conditional expectation (ICE): classification



- Same principle as PDP, but now with single line per observation

Using PDP and ICE

Advantages

Computation is intuitive

Assuming no correlations with other features, interpretation is clear: PDP shows how the average prediction in your dataset changes when the j -th feature is changed

Easy to implement

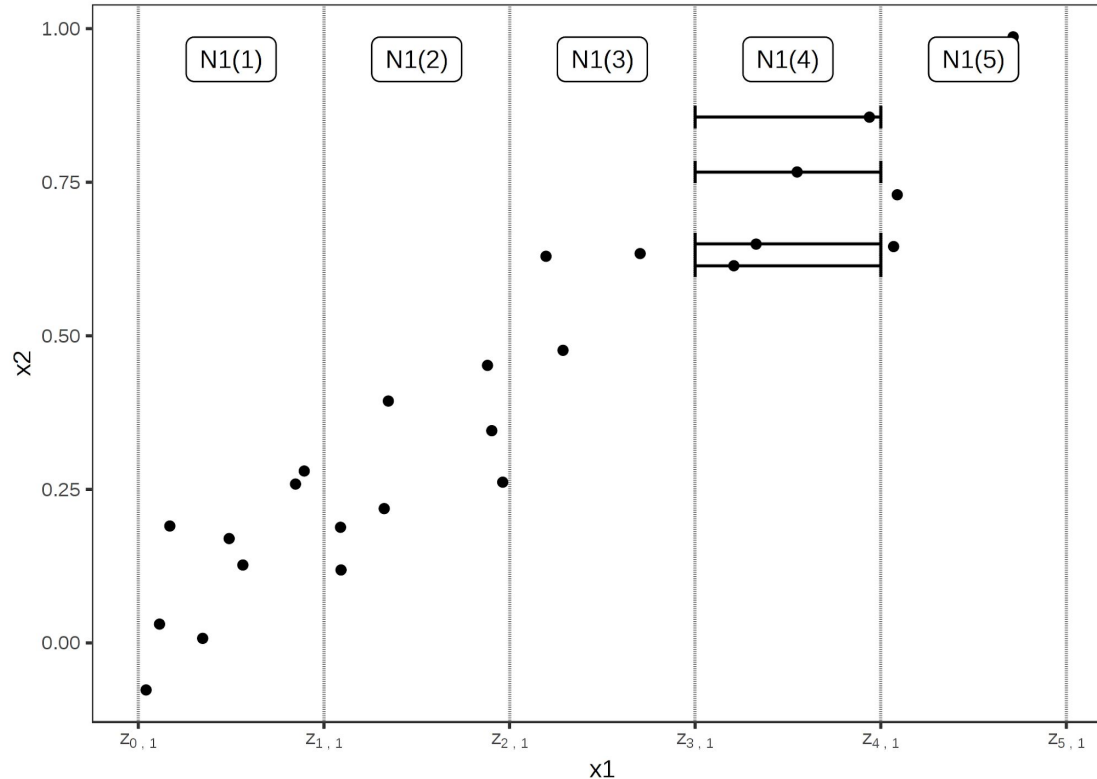
Has causal interpretation

Disadvantages

Assumption of independence of features often doesn't apply

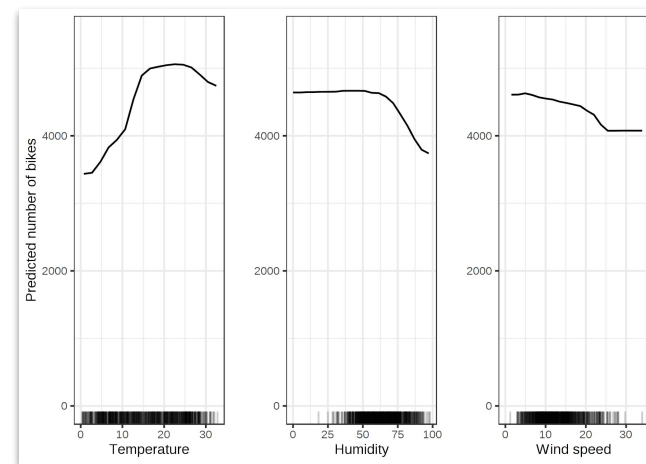
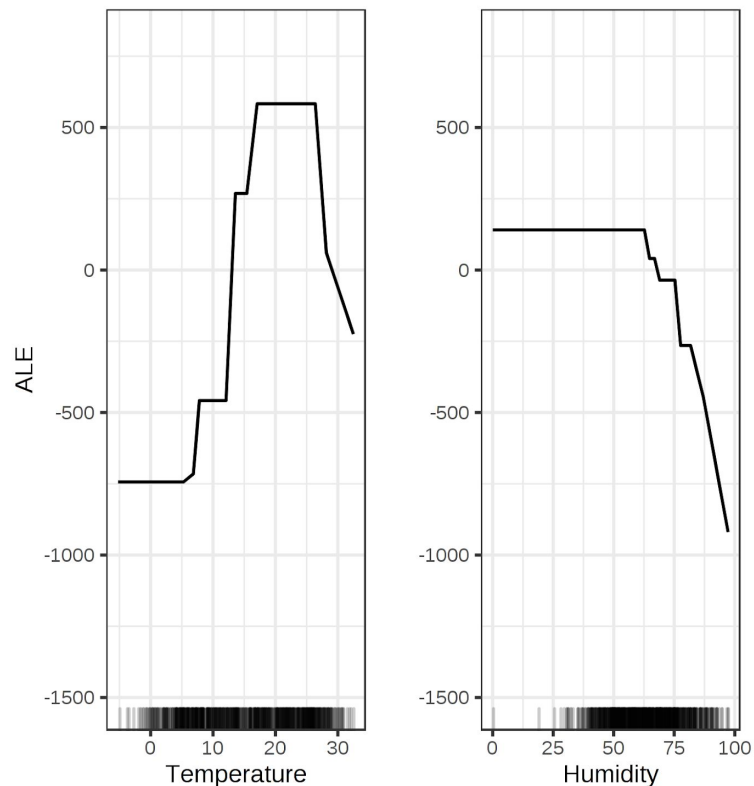
Heterogeneous effects are hidden because *average* marginal effect is taken → this is tackled in ICE

Accumulated local effects (ALE): regression



- Limitation PDP: can't handle collinearity since it calculates average effect across feature space even when there aren't any observations
- ALE solves this by changing predictions in a small "window" of the feature around v for data instances in that window

Accumulated local effects: regression



Using ALE

Advantages

ALE plots are unbiased

ALE plots are faster to compute

The interpretation of ALE plots is clear

ALE plots are centered at zero.

Disadvantages

ALE plots can become a bit shaky

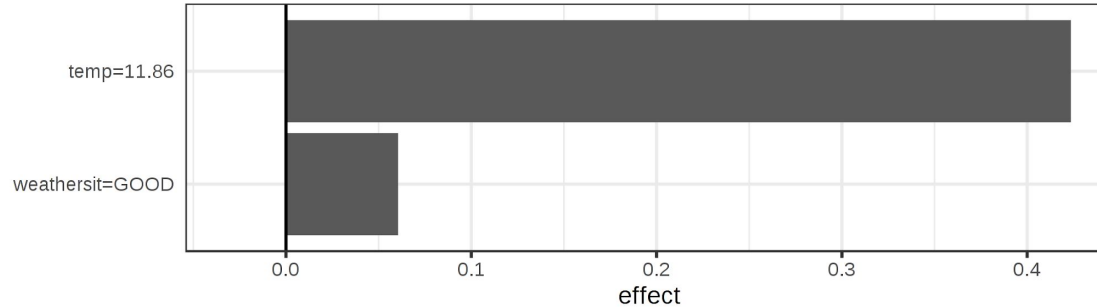
ALE plots are not accompanied by ICE curves like PDP

Implementation of ALE plots is much more complex

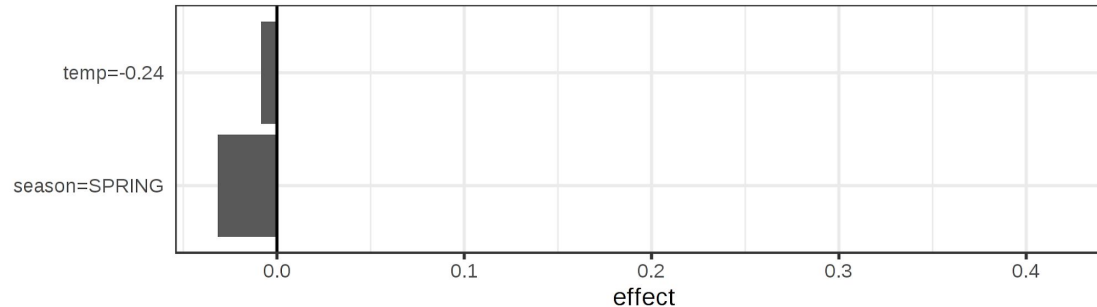
Interpretation remains difficult when features are strongly correlated

Local surrogate (LIME)

Actual prediction: 0.89
LocalModel prediction: 0.44



Actual prediction: 0.01
LocalModel prediction: -0.03



- Treat original model as black-box
- Train local surrogate model using interpretable model

Using LIME

Advantages

Even if you replace the underlying machine learning model, you can still use the same local, interpretable model for explanation

Local surrogate models benefit from the literature and experience of training and interpreting interpretable models.

When using Lasso or short trees, the resulting explanations are short (= selective) and possibly contrastive.

LIME is one of the few methods that works for tabular data, text and images.

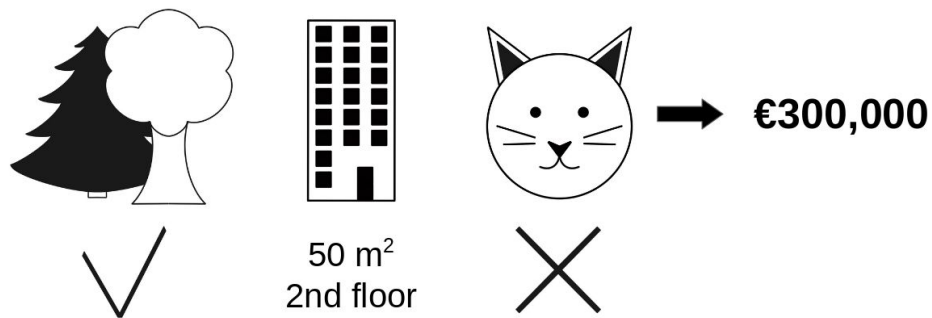
Disadvantages

The correct definition of the neighborhood is a very big, unsolved problem when using LIME with tabular data.

Another really big problem is the instability of the explanations.

Can be manipulated to hide biases

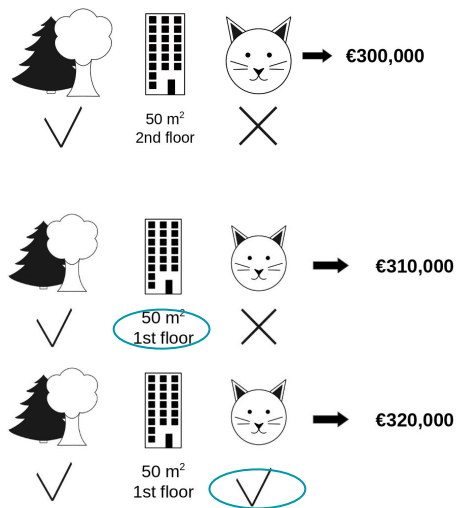
Shapley values: prediction as a collaborative game



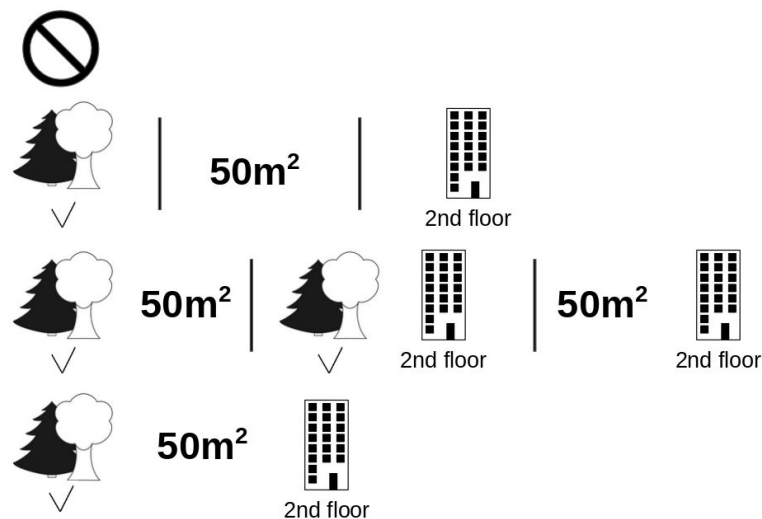
Feature	Value
park-nearby	True
area	50 m ²
floor	2nd
cat-ban	True
<hr/>	
prediction	300.000
average prediction	310.000
explain difference	-10.000

Compare random draws of a coalition, for all coalitions

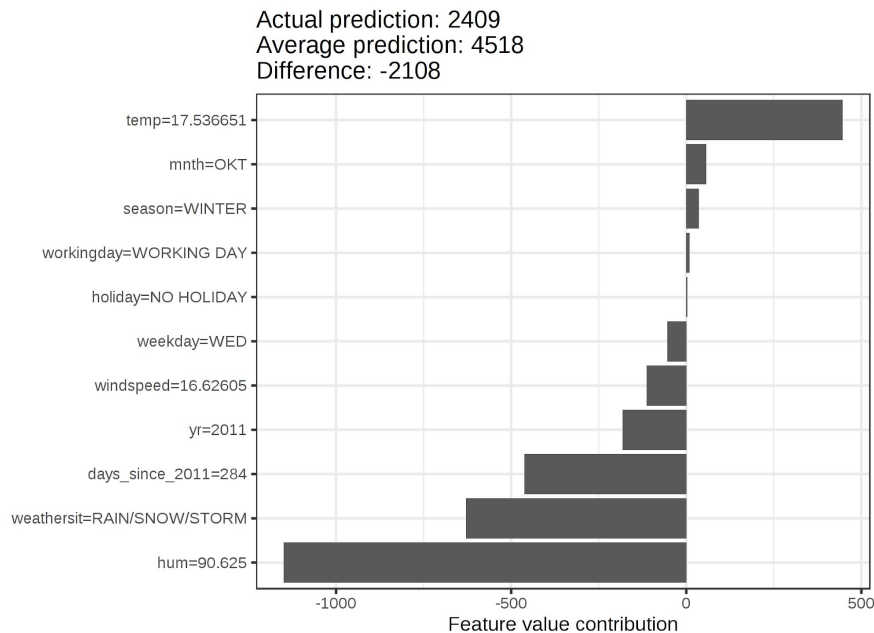
Random sampling of one coalition



Repeat for all possible coalitions



Shapley values for a random forest



- Prediction day 285:
- The sum of Shapley values yields the difference of actual and average prediction (-2108).
- **NB:** The Shapley value is the average contribution of a feature value to the prediction in different coalitions. The Shapley value is NOT the difference in prediction when we would remove the feature from the model.

Using Shapley values

Advantages

Based on solid theory, maybe only full explanation which is legally acceptable in the EU

The difference between the prediction and the average prediction is fairly distributed among the feature values of the instance

Allows contrastive explanations

Disadvantages

Computationally heavy, requires approximations

Easily misinterpreted

Not suitable for sparse explanations: always uses all features

Need access to the data to calculate Shapley value for new instance (can be solved with synthetic data)

Suffers from inclusion of unrealistic data instances when features are correlated.

SHAP: SHapley Additive Explanations

Shapley values

Estimate average contribution of feature in coalitions through permutation

LIME

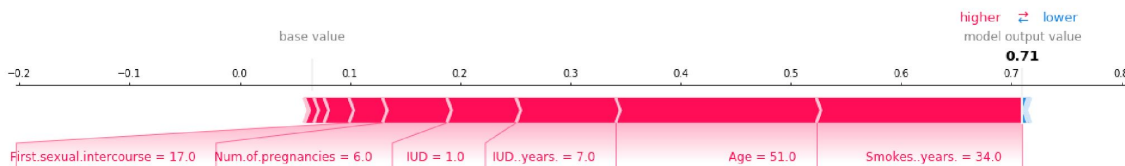
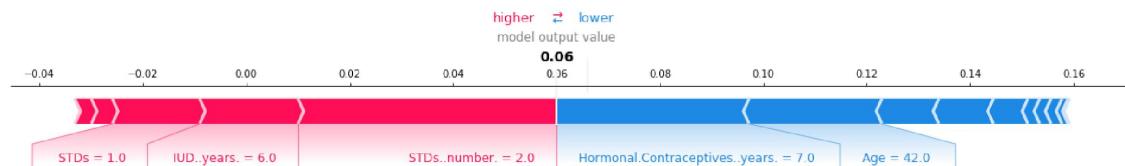
Use simple, local estimator to calculate Shapley values



SHAP

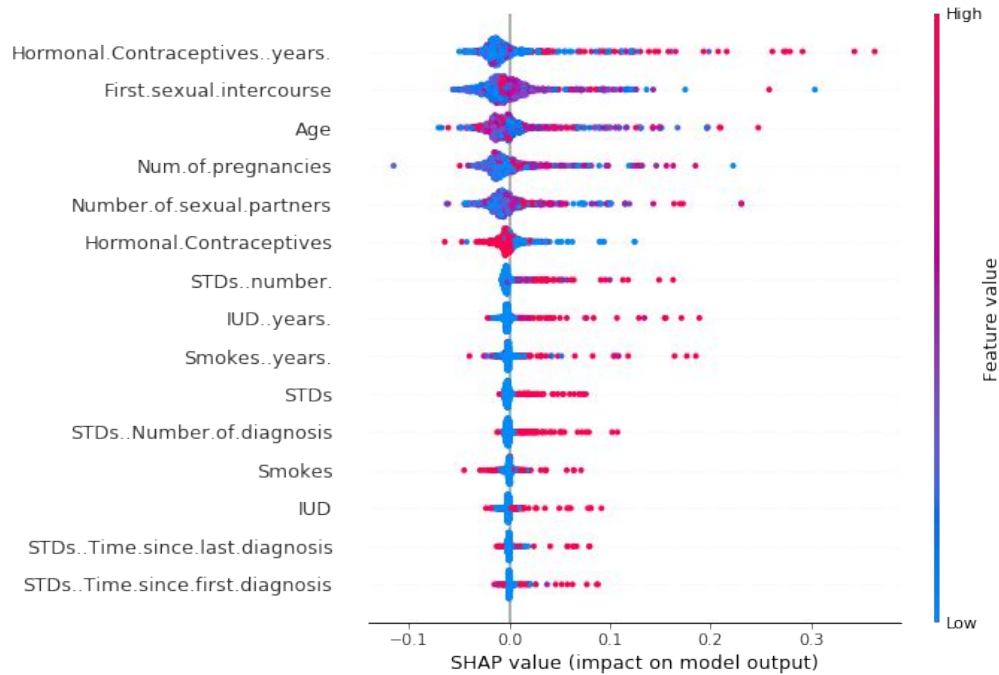
- **KernelSHAP:** sample and fit linear model for coalitions
- **TreeSHAP:** specific estimation for tree ensembles

SHAP: single predictions



- Top: low predicted risk **0.06**. Risk increasing effects such as STDs are offset by decreasing effects such as age.
- Bottom: high predicted risk **0.71**. Age of 51 and 34 years of smoking increase her predicted cancer risk.

SHAP: summary plot



- SHAP summary plot combines feature importance with feature effect
- Low number of years on hormonal contraceptives reduce the predicted cancer risk, a large number of years increases the risk.
- Repeated reminder: All effects describe the behavior of the model and are not necessarily causal in the real world

Using SHAP

Advantages

Similar advantages as Shapley values

Fast computation

Possible to make global explanations

Disadvantages

... and similar disadvantages, too

KernelSHAP is slow

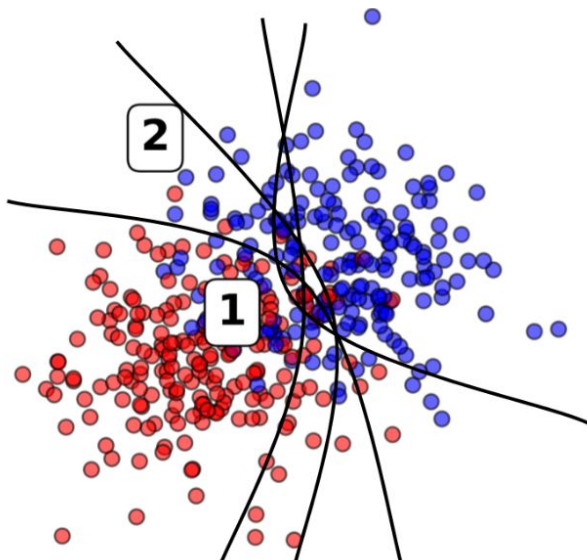
KernelSHAP ignores features dependence, most other permutation-based methods have this problem

TreeSHAP can produce unintuitive feature contributions

Uncertainty

(hard / unsolved)

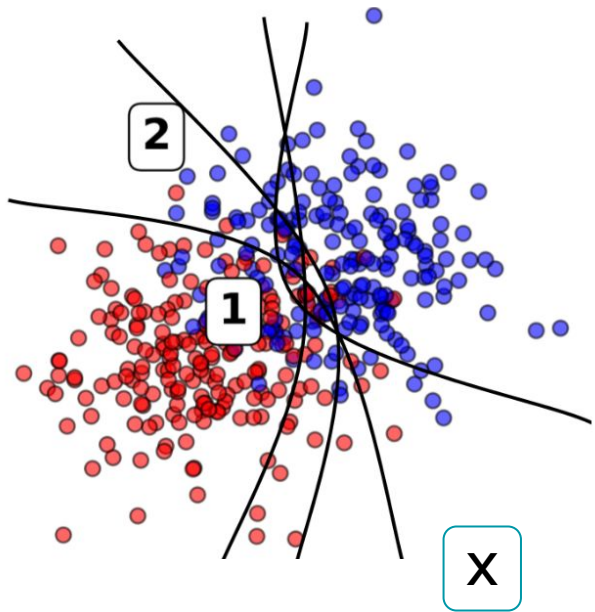
Aleatoric and epistemic uncertainty



Aleatoric uncertainty (1) captures noise inherent in the observations, as shown in by the overlapping classes. This uncertainty may be resolved by adding another feature.

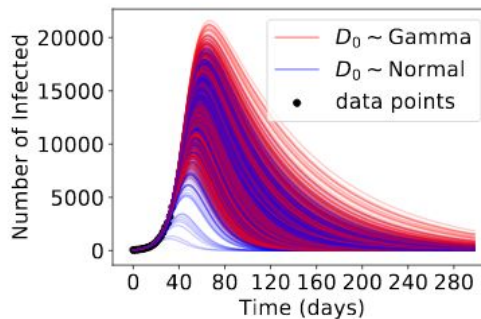
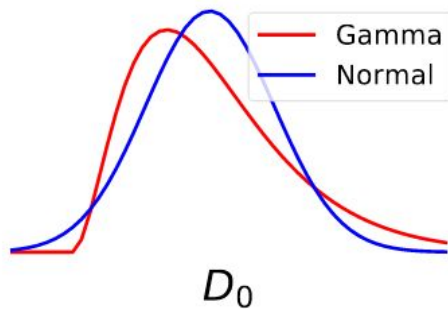
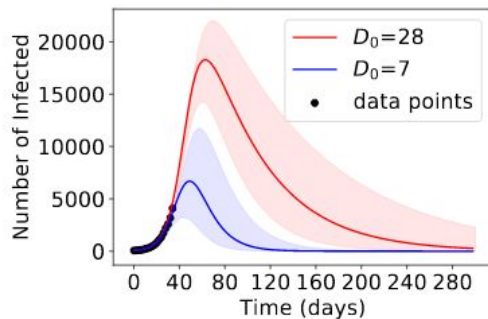
Epistemic uncertainty (2) accounts for uncertainty in the model. Even a good model is uncertain about the decision boundary, which is caused by a lack of data.

How to estimate uncertainty, particularly epistemic?



- Estimating epistemic uncertainty relates to detecting when **not** to use your model (out-of-domain detection, OOD) e.g. at point X
- Current state of the art “... are barely as good as random guessing when trying to identify OOD”.

A final word of caution: be aware of underspecification



- _ Triggered by discussions trying to predict the peak of COVID infections during the first wave
- _ Recent article named a “potential wrecking ball” on practical use of ML

Recap: Learning objectives

- _ Understand different concepts of interpretability and their strengths and weaknesses
- _ Know how to apply general model-agnostic methods for interpreting black box machine learning models for tabular data
- _ Know how to choose which interpretation method is most suitable for your machine learning project

