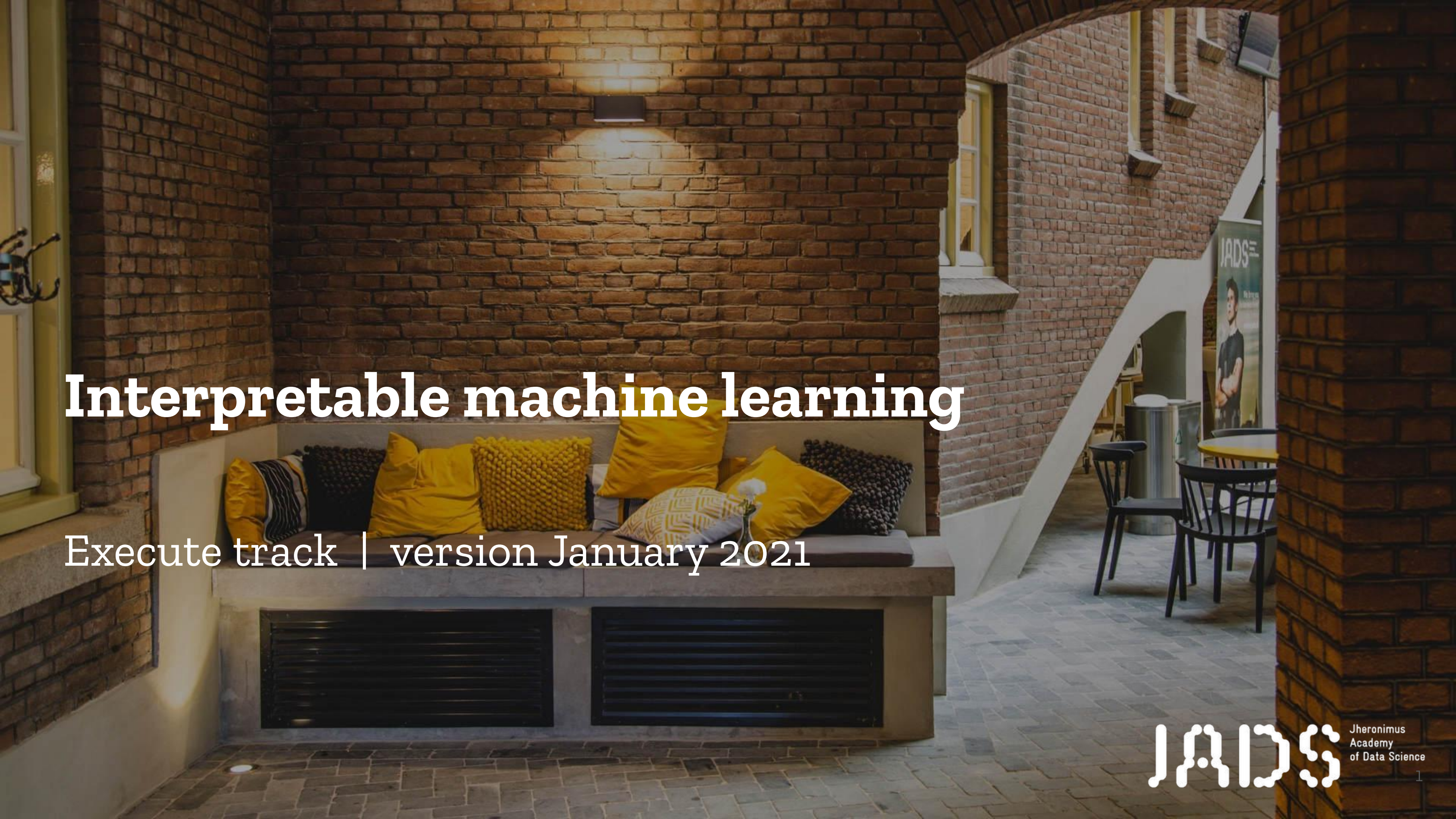


Interpretable machine learning

Execute track | version January 2021

Jheronimus Academy of Data Science

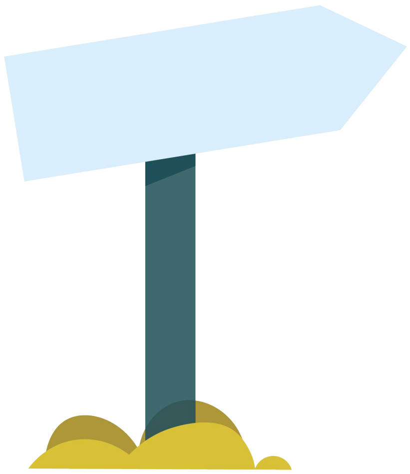
Attribution and copyright notice

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

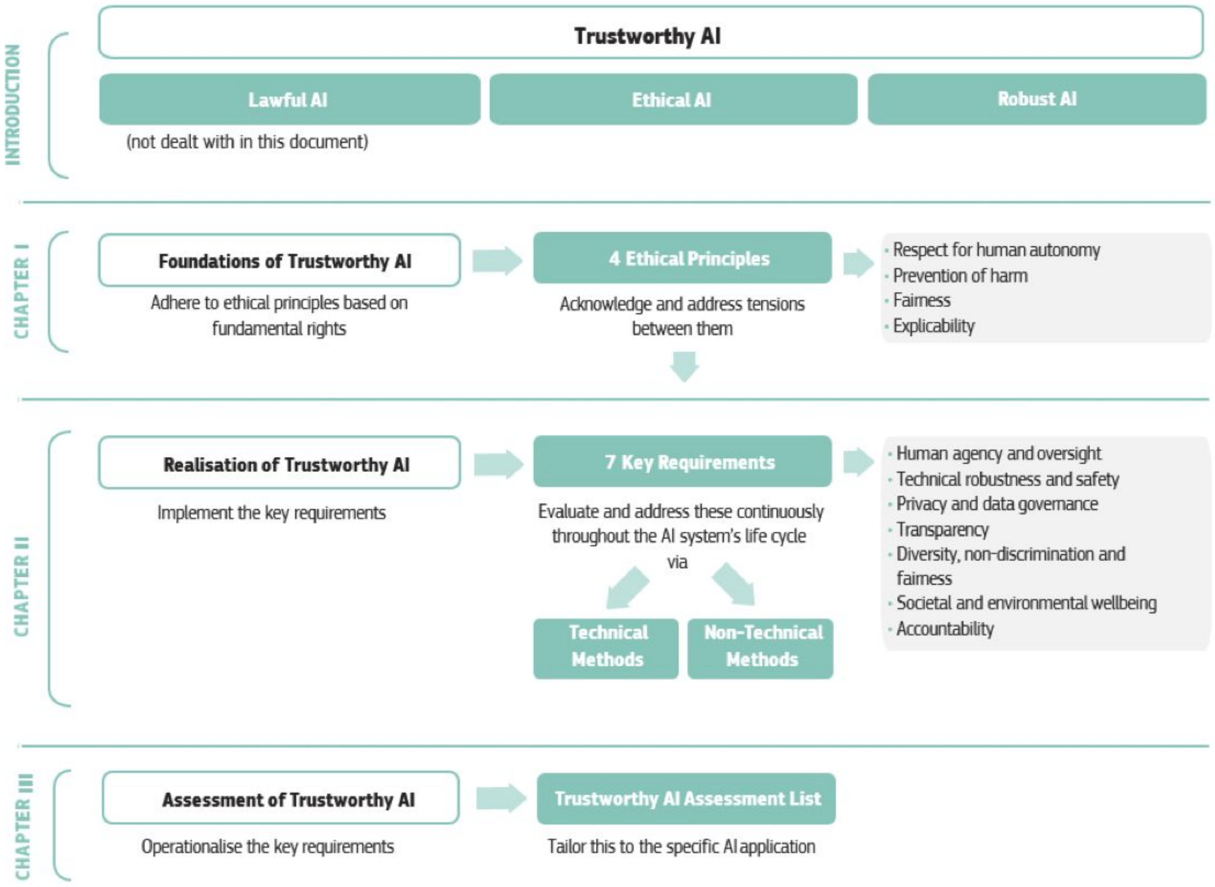
- Interpretable Machine Learning, by Christoph Molnar (referenced as IML)
- Several open access journal papers (referenced individually)

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



source: EU Ethics 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

source: EU Ethics Guidelines for Trustworthy AI

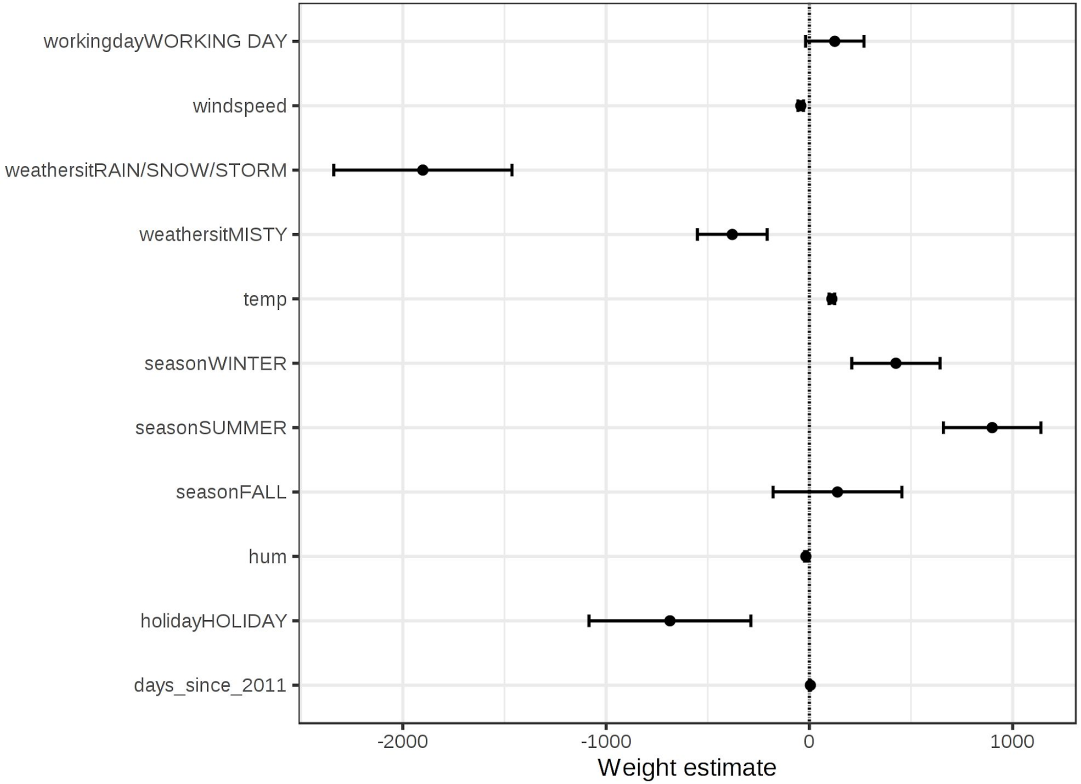
Good, human-friendly explanations

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

Source: IML section [2.6 Human-friendly Explanations | Interpretable Machine Learning]

Interpretable models
(easy)

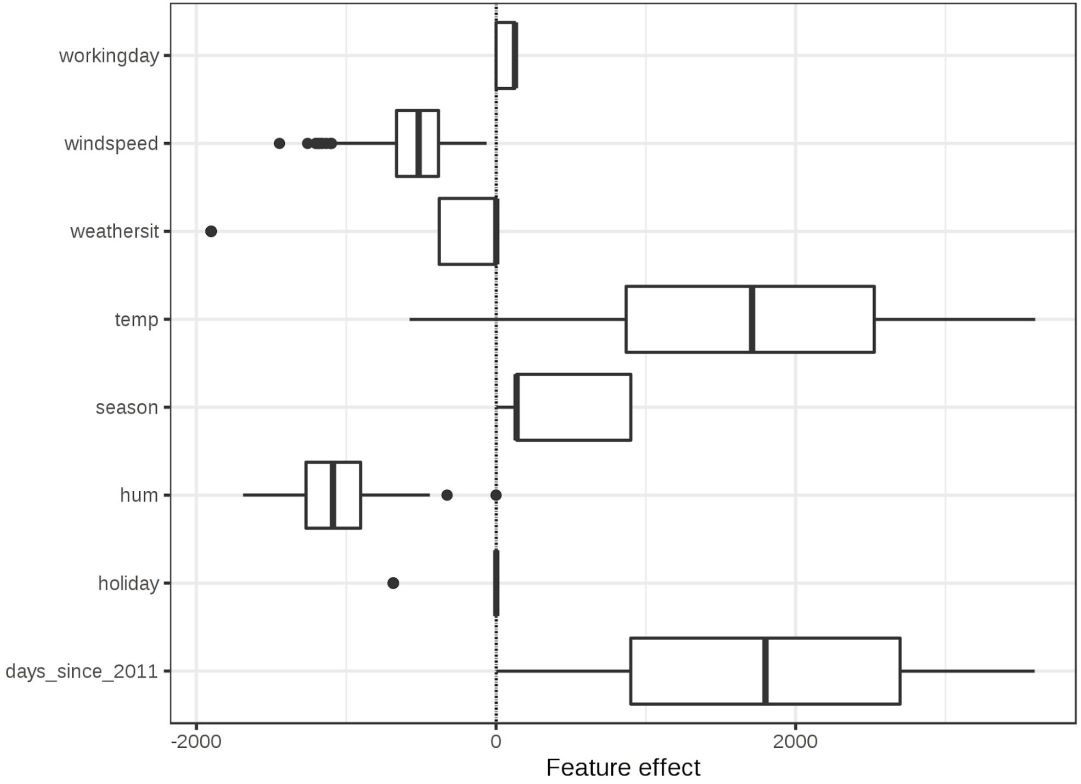
Linear regression: weight plot



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

Source: IML section [4.1 Linear regression]()

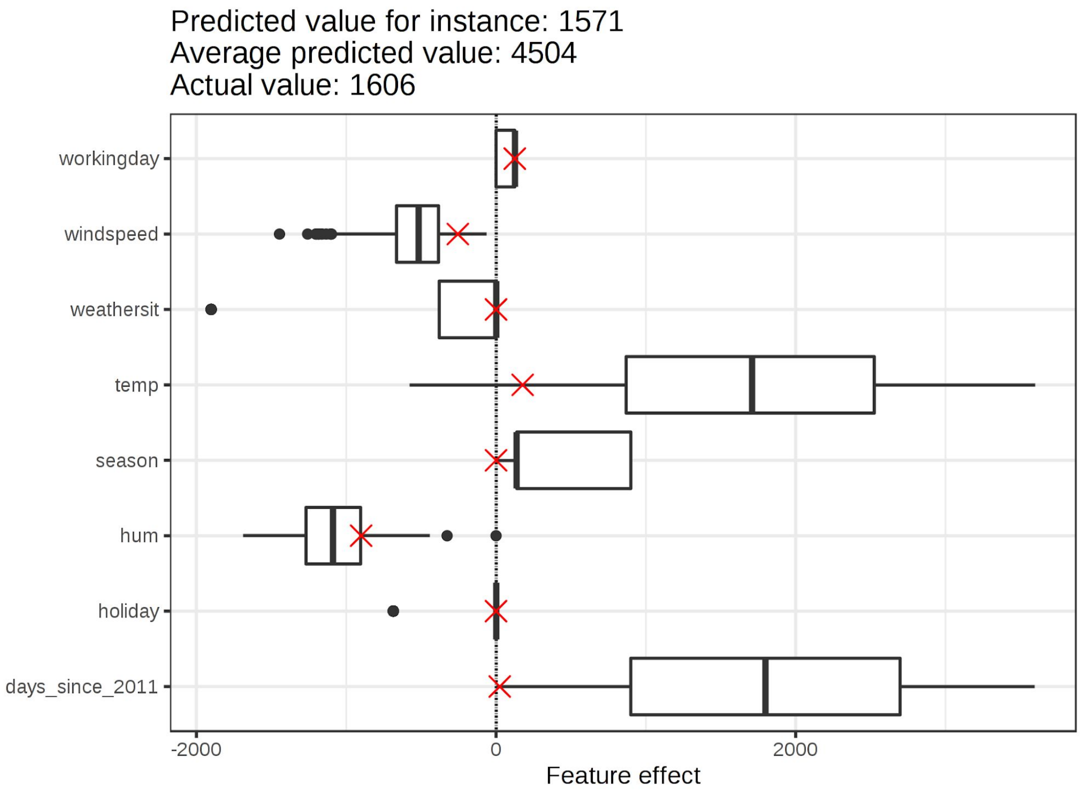
Linear regression: effect plot



- Shows effect = *weights * actual feature values*

Source: IML section [4.1 Linear regression]()

Linear regression: effect plot with individual prediction



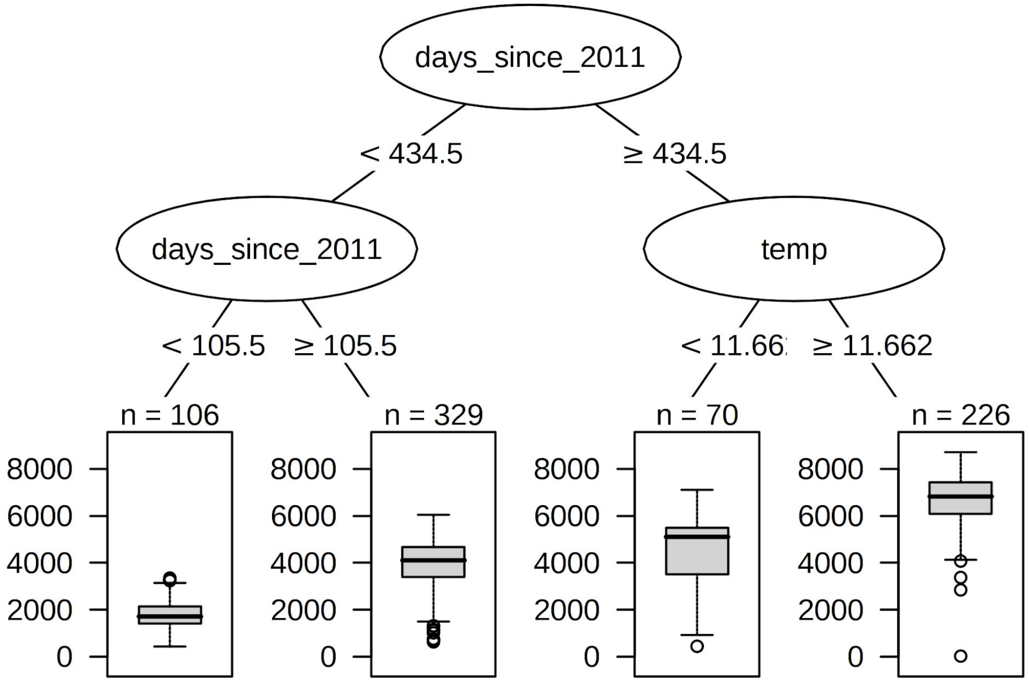
- Shows how feature effects influence individual predictions compared to average prediction

Source: IML section [4.1 Linear regression]

Interpretability of linear models

Advantages	Disadvantages
Transparent how predictions are generated	Non-linearity or interactions have to be added manually
Selective explanation with Lasso	Poor predictive performance
Widely accepted	Interpretation can be unintuitive for correlated features
Guarantee that you have optimal weights (as long as assumptions of linear models is met)	

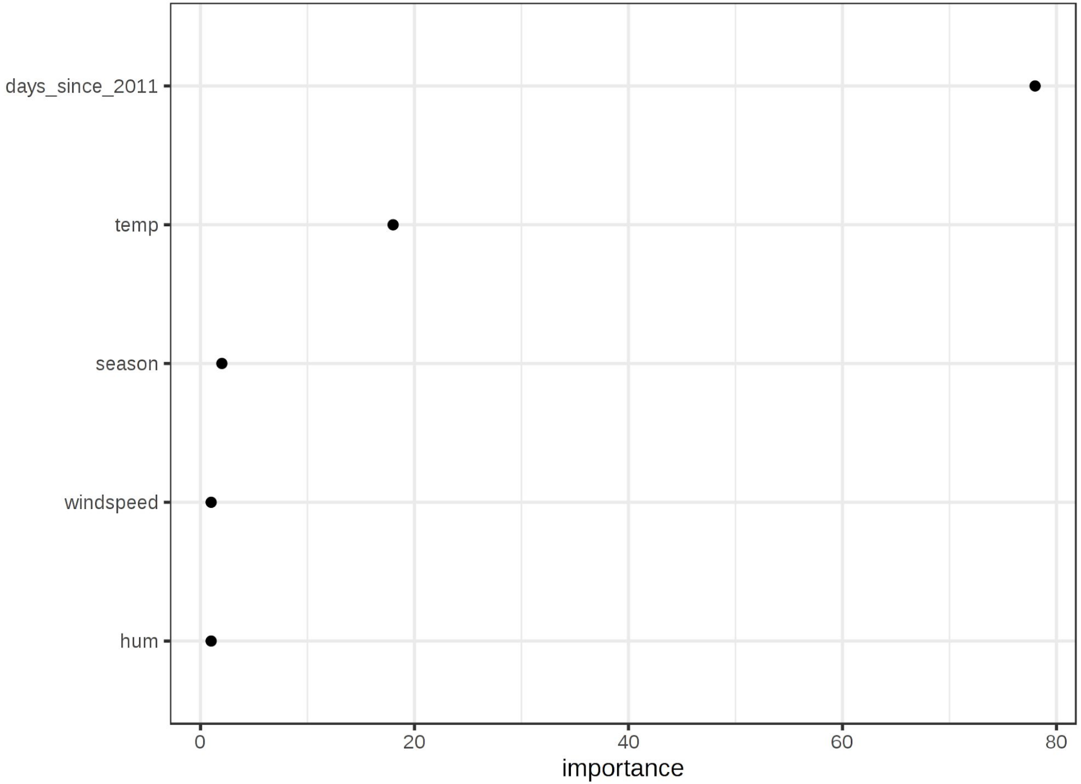
Decision tree



- So what has bigger influence: time of temperature?

Source: IML section [4.4 Decision tree]

Decision tree: feature importance



Source: IML section 4.4 Decision tree

Beware Default Random Forest Importances

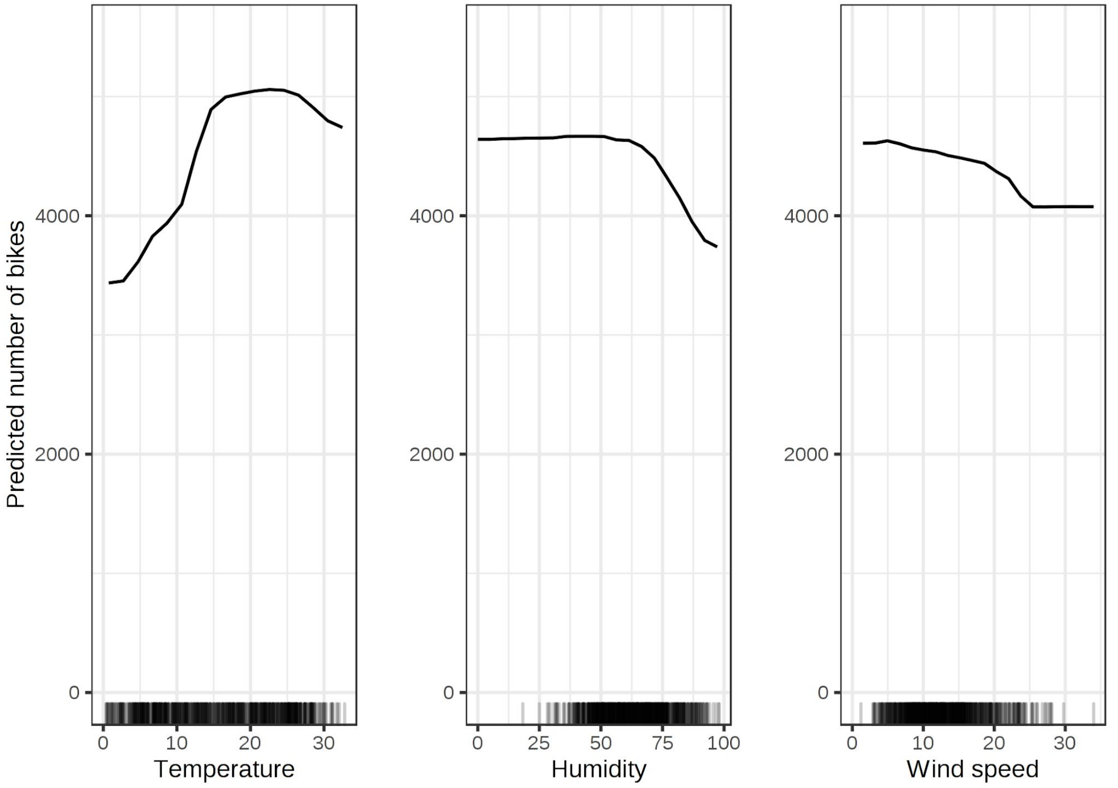
- Default method based on mean decrease in impurity → can be very wrong
- Better solution: permutation importance
- Python library: rfpimp
- Credit: Prof Terence Parr, University of San Francisco

Interpretability of decision trees

Advantages	Disadvantages
Ideal for capturing interactions	Linear relationships are not shown efficiently (multiple steps in nodes)
Predictions are made in groups	Lack of smoothness
Tree structure is intuitive visualisation	Tree are unstable (inherent to greedy algorithm)
Trees create good human-friendly explanations	Long (deep) tree as not easily interpreted

Model-agnostic Methods (intermediate)

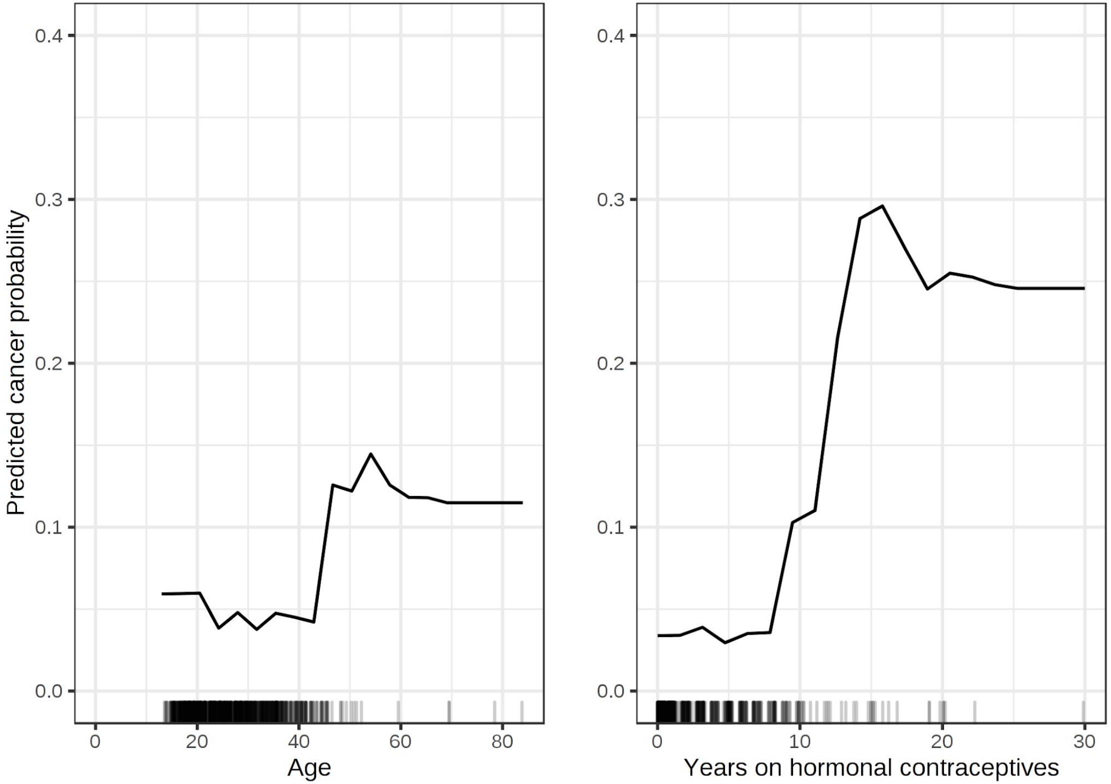
Partial dependence plot (PDP): regression



Source: IML section [5.1 Partial dependence plot]

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

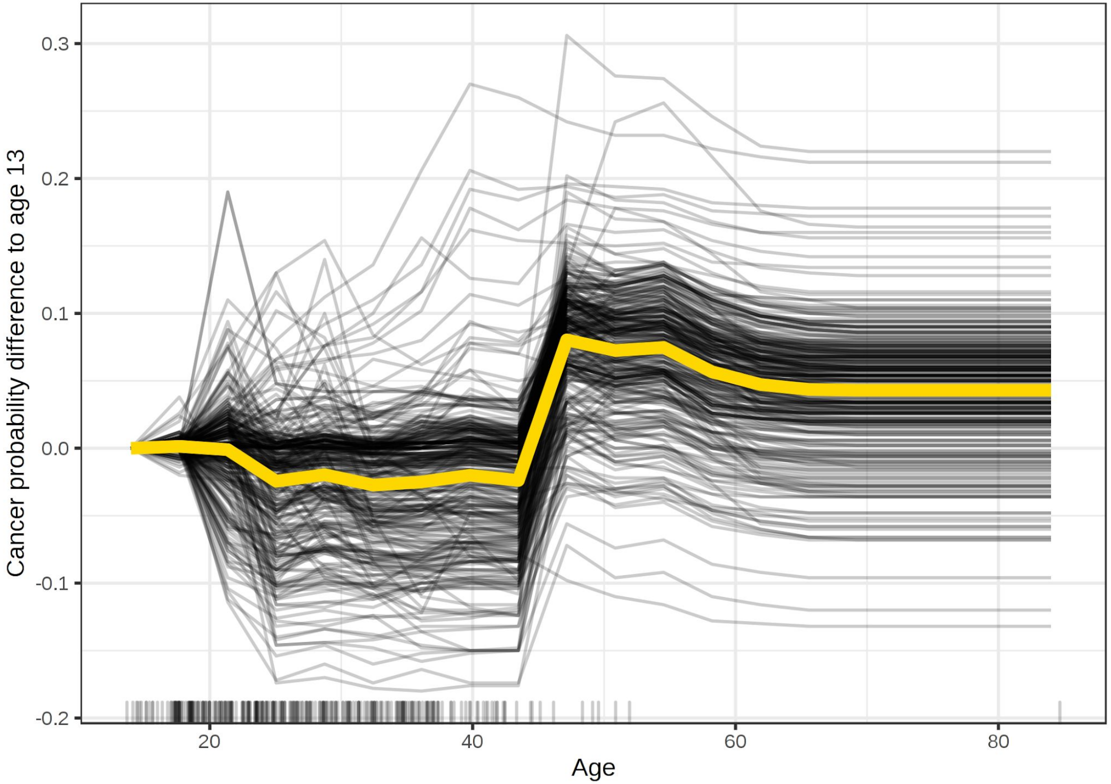
Partial dependence plot (PDP): classification



Source: IML section [5.1 Partial dependence plot]

- Similar approach with predicted probability on the y-axis

Individual conditional expectation (ICE): classification



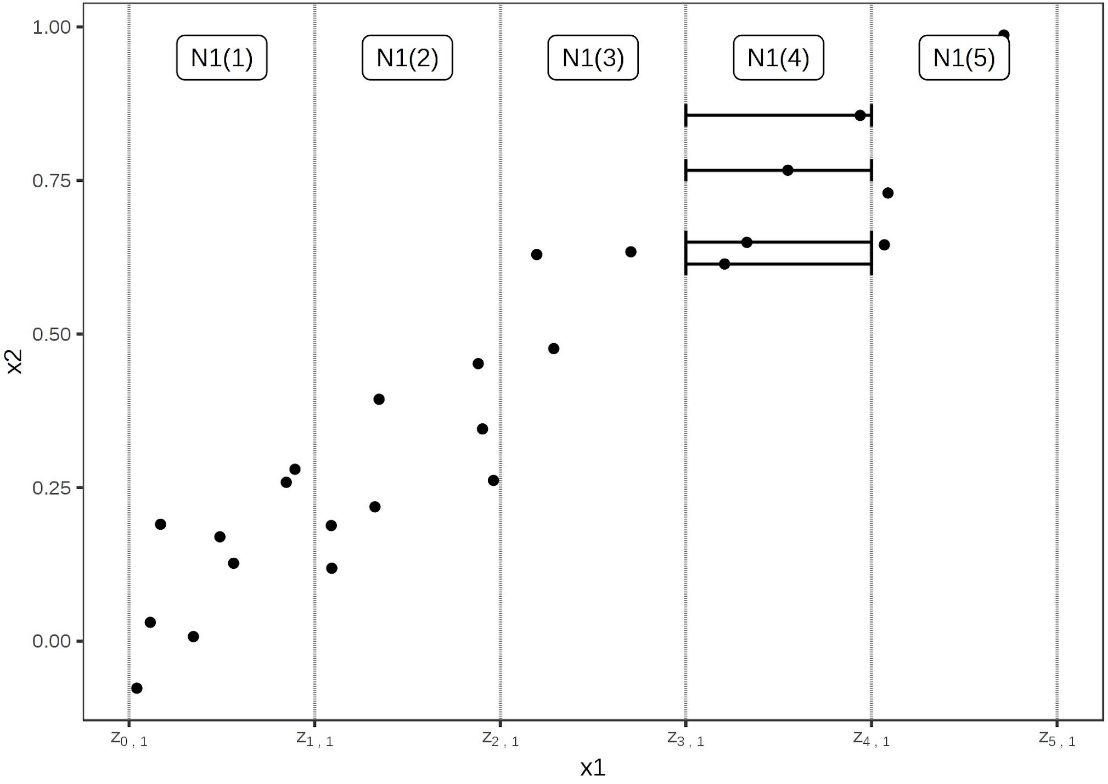
Source: IML section [5.2 Individual conditional expectation]

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

Using PDP and ICE

Advantages	Disadvantages
Computation is intuitive	**Assumption of independence of features often doesn't apply**
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	Heterogeneous effects are hidden because average marginal effect is taken → this is tackled in ICE
Easy to implement	
Has causal interpretation	

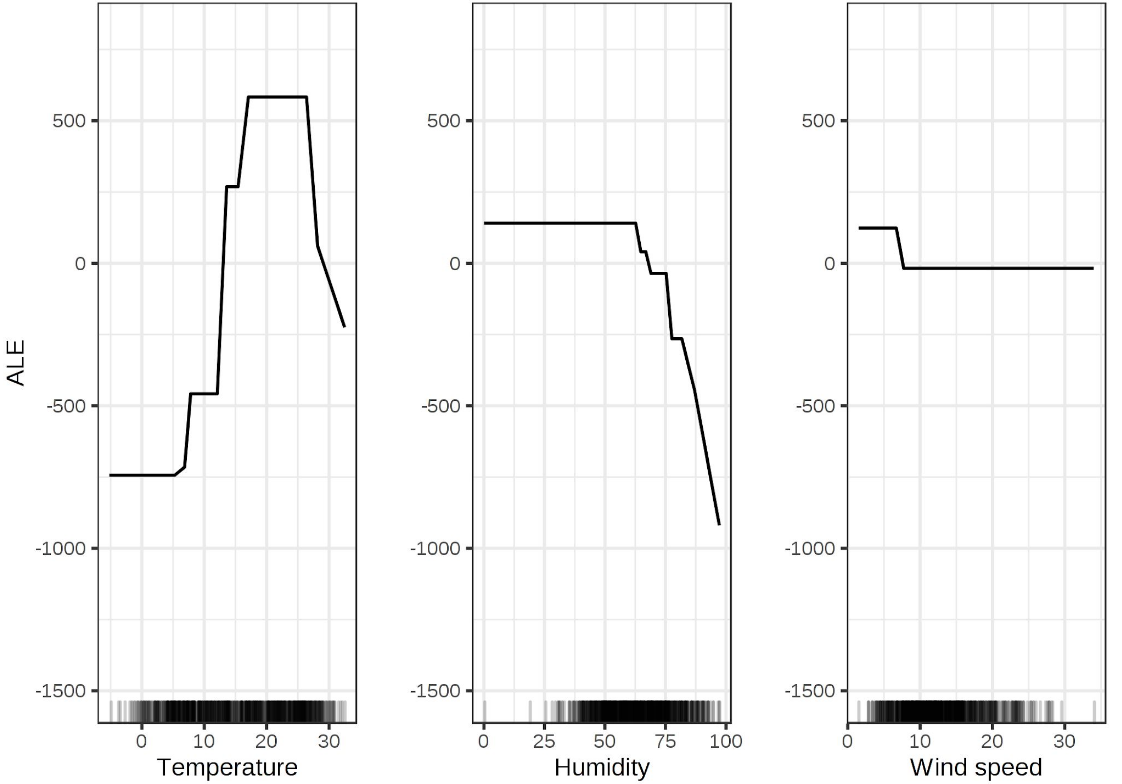
Accumulated local effects (ALE): regression

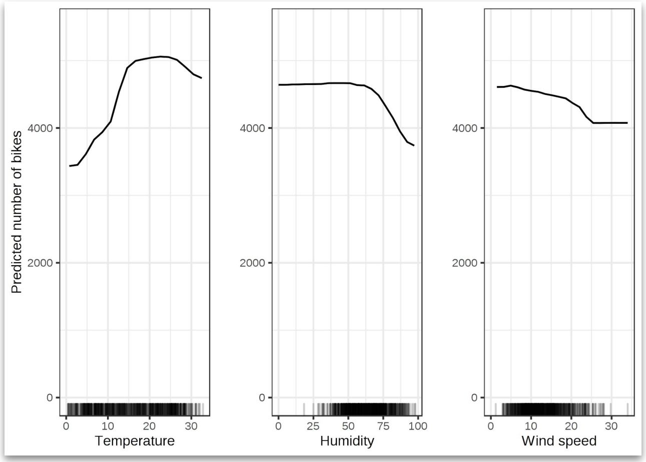


Source: IML section [5.3 Accumulated local effects]

- 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



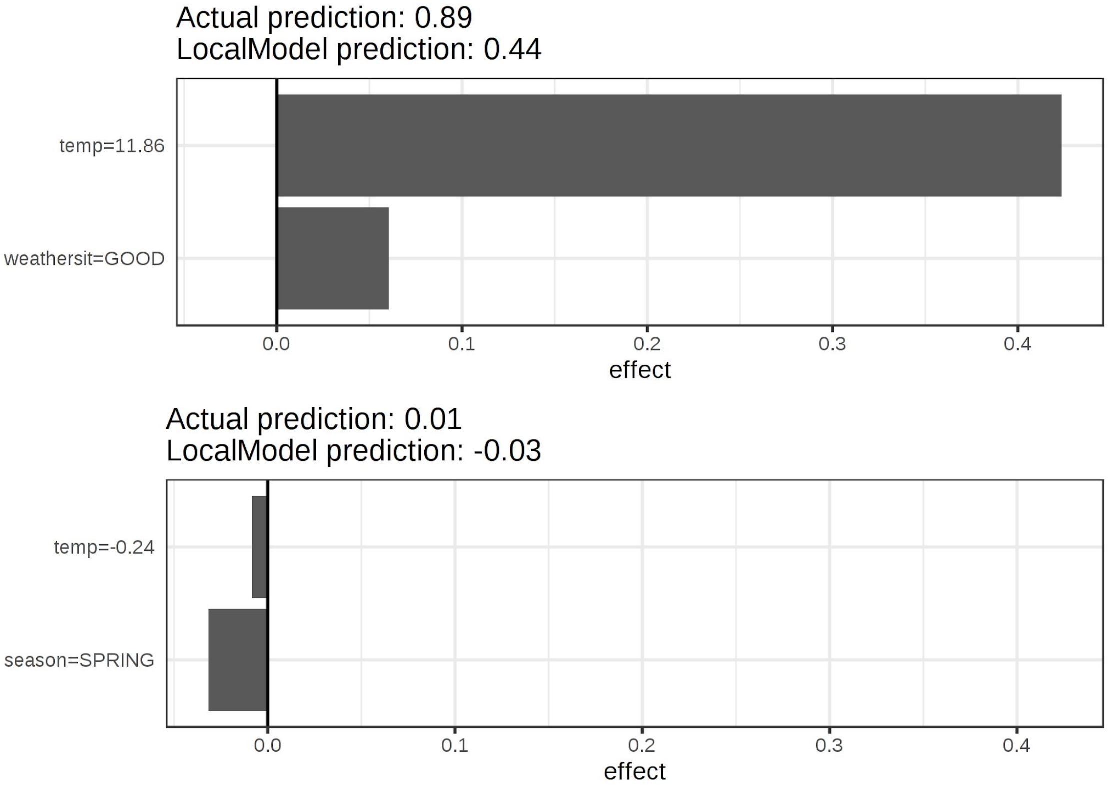


Source: IML section [5.3 Accumulated local effects]()

Using ALE

Advantages	Disadvantages
ALE plots are unbiased	ALE plots can become a bit shaky
ALE plots are faster to compute	ALE plots are not accompanied by ICE curves like PDP
The interpretation of ALE plots is clear	Implementation of ALE plots is much more complex
ALE plots are centered at zero.	Interpretation remains difficult when features are strongly correlated

Local surrogate (LIME)



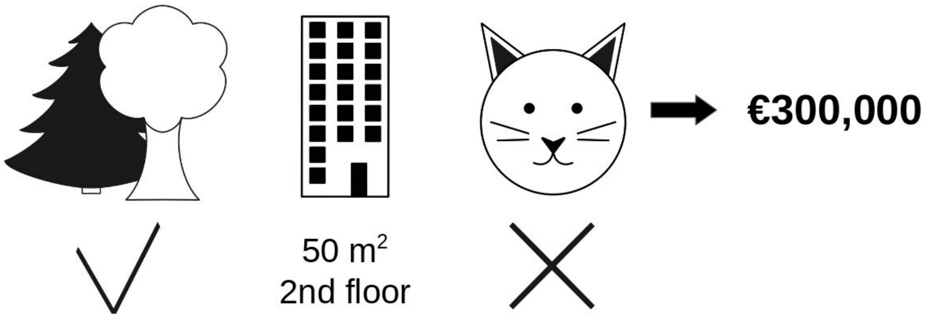
- Treat original model as black-box
- Train local surrogate model using interpretable model

Source: IML section [5.7 Local surrogate (LIME)]

Using LIME

Advantages	Disadvantages
Even if you replace the underlying machine learning model, you can still use the same local, interpretable model for explanation	The correct definition of the neighborhood is a very big, unsolved problem when using LIME with tabular data.
Local surrogate models benefit from the literature and experience of training and interpreting interpretable models.	Another really big problem is the instability of the explanations.
When using Lasso or short trees, the resulting explanations are short (= selective) and possibly contrastive.	Can be manipulated to hide biases
LIME is one of the few methods that works for tabular data, text and images.	

Shapley values: prediction as a collaborative game



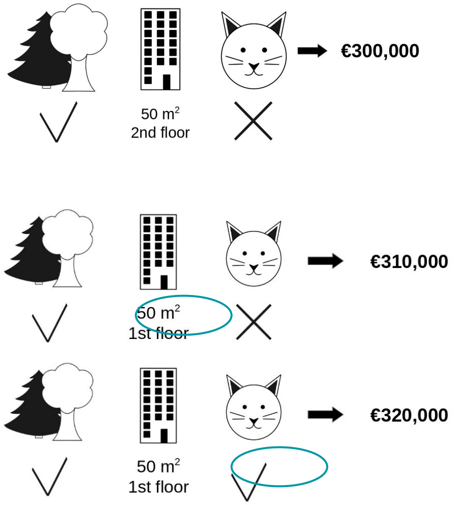
Feature	Value
park-nearby	True
area	50 m^2
floor	2nd
cat-ban	True

prediction	**300.000**
average prediction	310.000
explain difference	**-10.000**

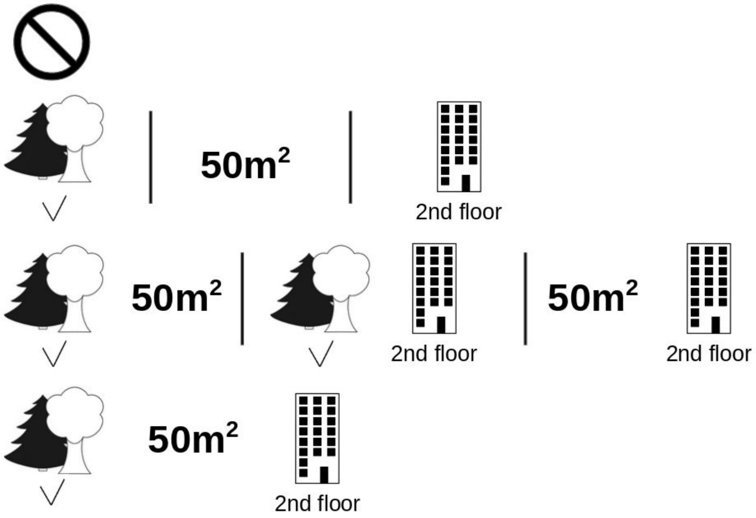
Source: IML section [5.9 Shapley values]

Compare random draws of a coalition, for all coalitions

Random sampling of one coalition

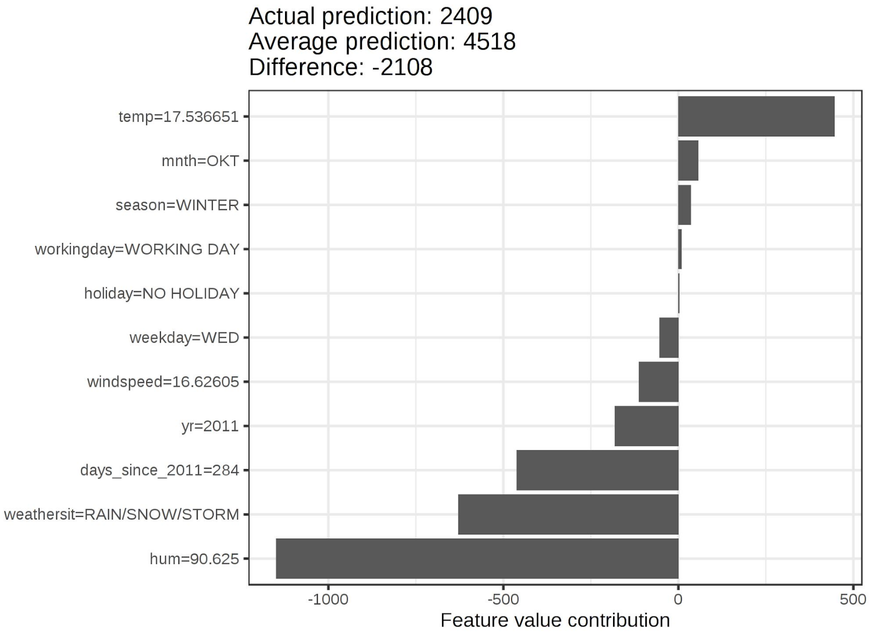


Repeat for all possible coalitions



Source: IML section [5.9 Shapley values]

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.

Source: IML section [5.9 Shapley values]

Using Shapley values

Advantages	Disadvantages
Based on solid theory, maybe only full explanation which is legally acceptable in the EU	Computationally heavy, requires approximations
The difference between the prediction and the average prediction is fairly distributed among the feature values of the instance	Easily misinterpreted
Allows contrastive explanations	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

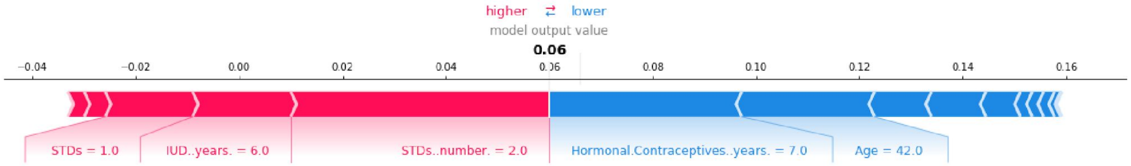
LIME
Use simple, local estimator to calculate Shapley values

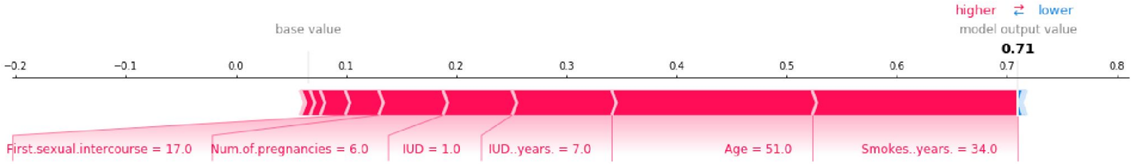
SHAP

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

Source: IML section [5.10 SHAP]

SHAP: single predictions

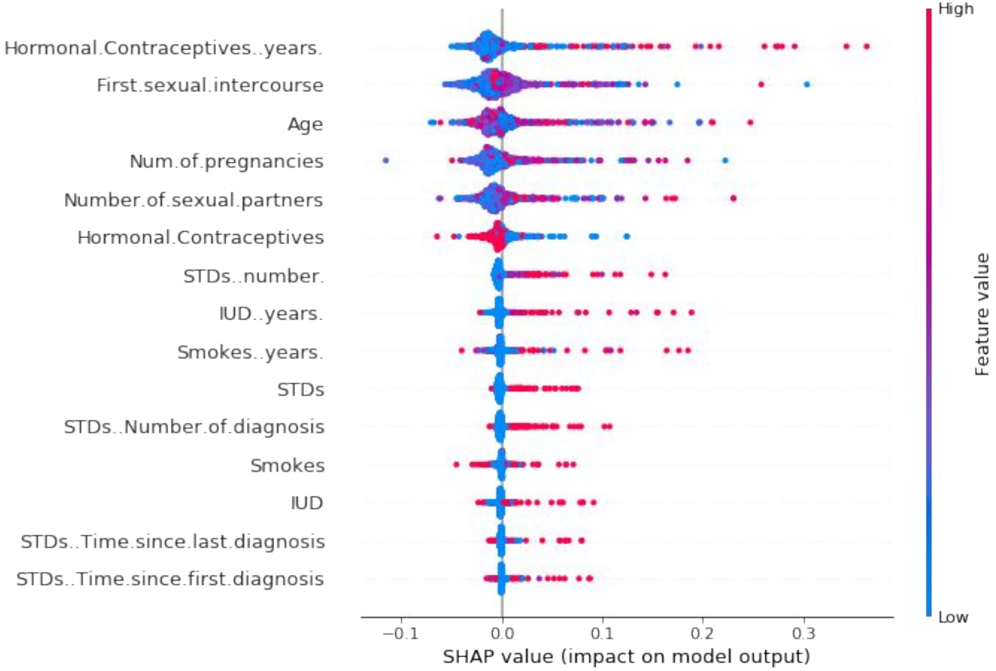




- Top: low predicted risk **0.06.** Risk increasing effects such as STDs are offset by decreasing effects such as age.
- Bottom: high predicted risk **0.71**. Age of 51 and 34 years of smoking increase her predicted cancer risk.

Source: IML section [5.10 SHAP]

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

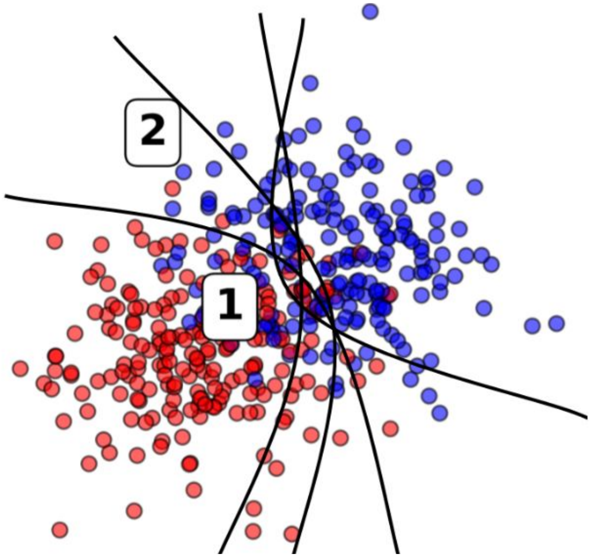
Source: IML section [5.10 SHAP]

Using SHAP

Advantages	Disadvantages
Similar advantages as Shapley values	... and similar disadvantages, too
Fast computation	KernelSHAP is slow
Possible to make global explanations	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

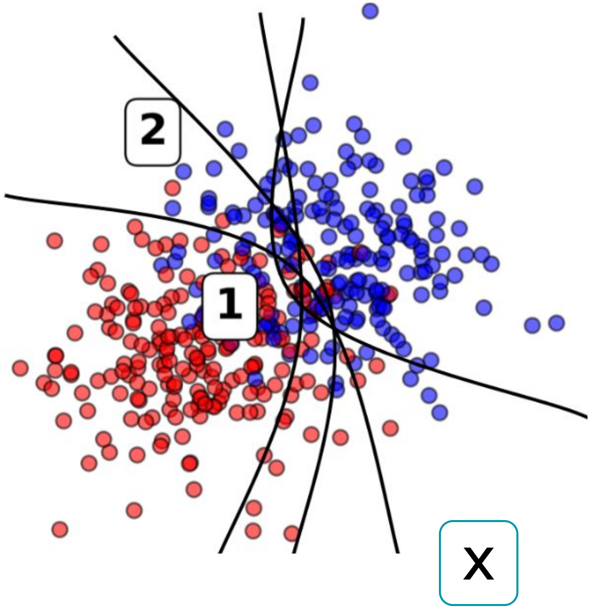


Aleatoric uncertainty (1) captures noise inherent in the observations, as shown in by the overlapping classes. This uncertainty may be resolved by adding another feature.

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.

Uncertainty estimation for classification and risk prediction on medical tabular data, Pacmed (2020)

How to estimate uncertainty, particularly epistemic?

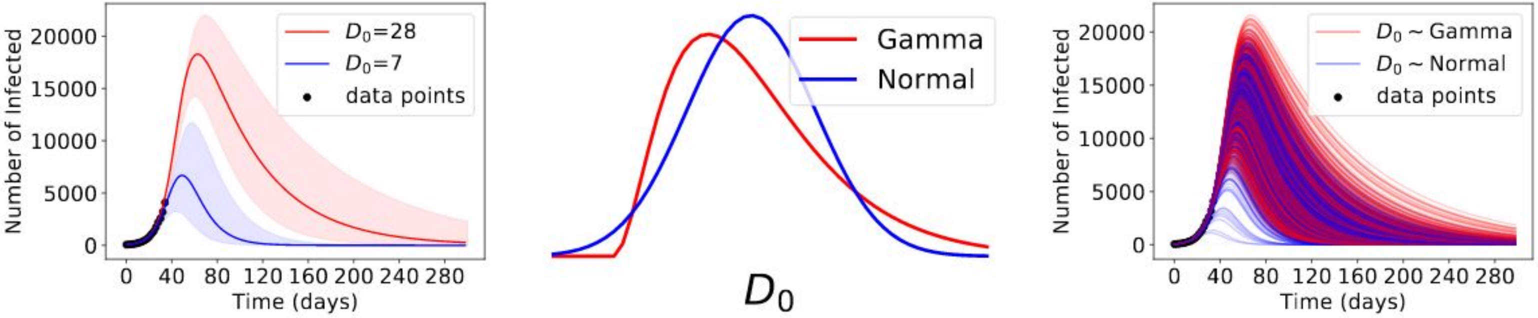


- Estimating epistemic uncertainty relates to detecting when ***not*** to use your model (out-of-domain detection, OOD) e.g. at point X

- Current state of the art "*... are barely as good as random guessing when trying to identify OOD*".

[Out-of-domain detection for electronic patient records, GitHub repository Pacmed Labs]

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

[Underspecification in machine learning, 40 people from Google (November 2020)]

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