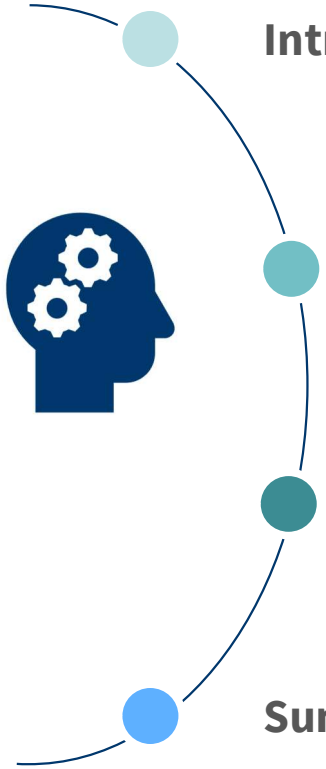
Business Analytics & Data Science

Interpretable Machine Learning & XAI

Stefan Lessmann

Agenda



Introduction

Global Explanation Methods
Surrogate models, permutation importance, partial dependence, and ICE

Local and Example Based Explanations
LIME, SHAP values, counterfactual and adversarial examples

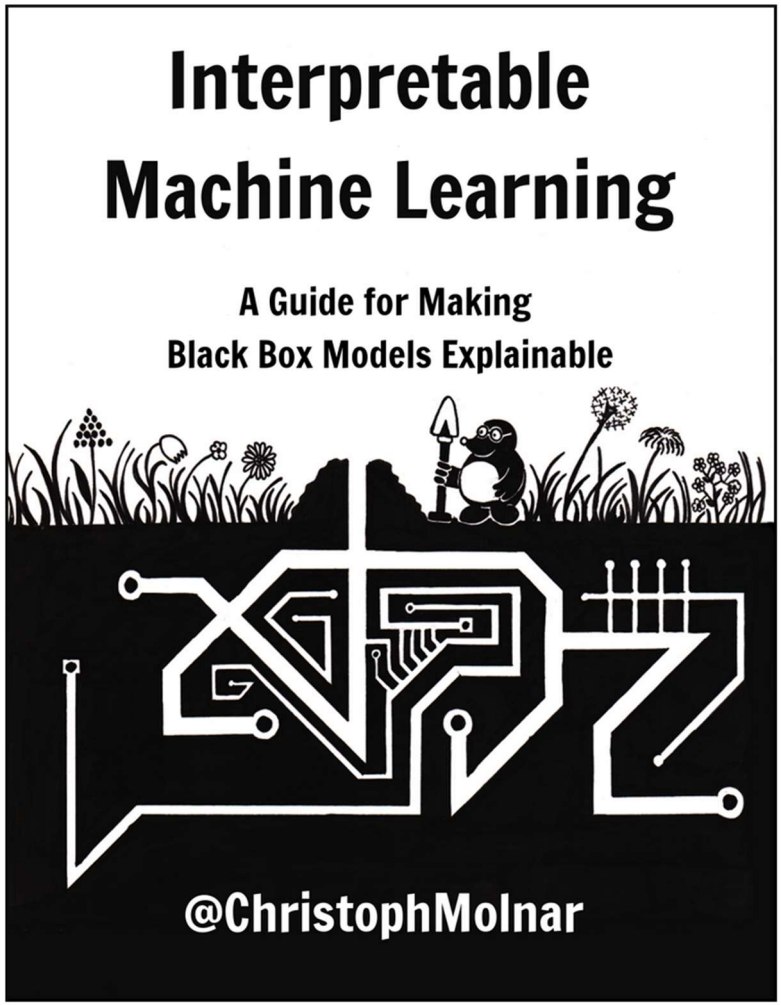
Summary

Interpretable Machine Learning (IML)

Free online book by C. Molnar

- **Comprehensive and readable coverage of recent developments in IML**
- **Basis for several topics and examples in this chapter**
- **Online version available for free**

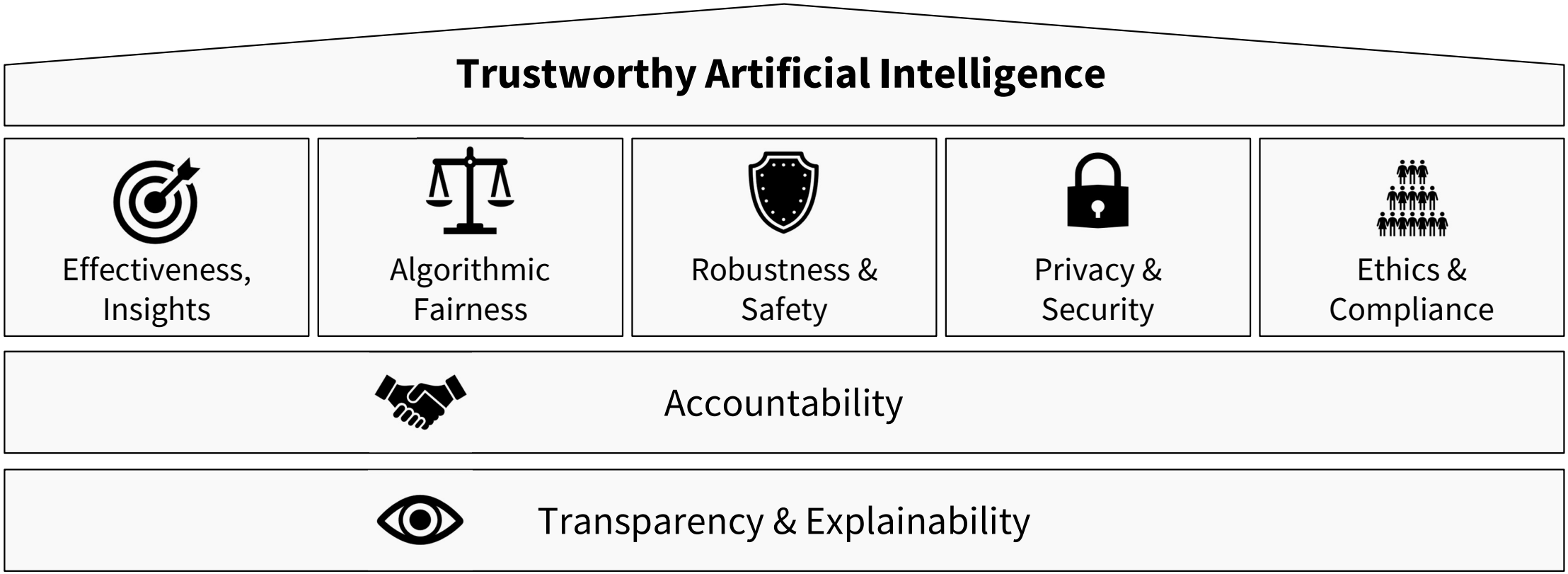
https://christophm.github.io/interpretable-ml-book/



Introduction

The Bigger Picture

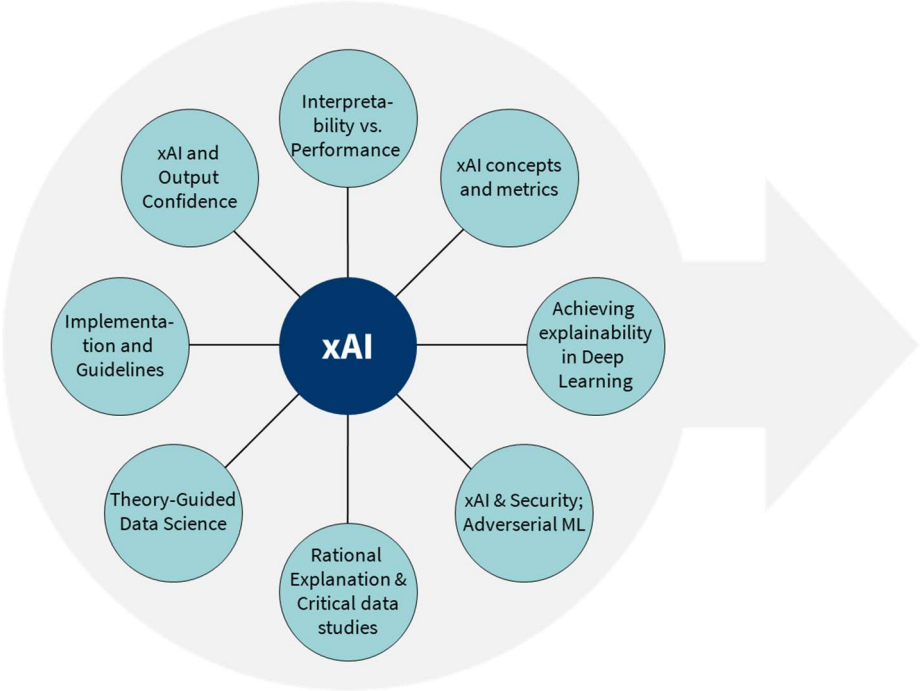
Interpretable ML in context

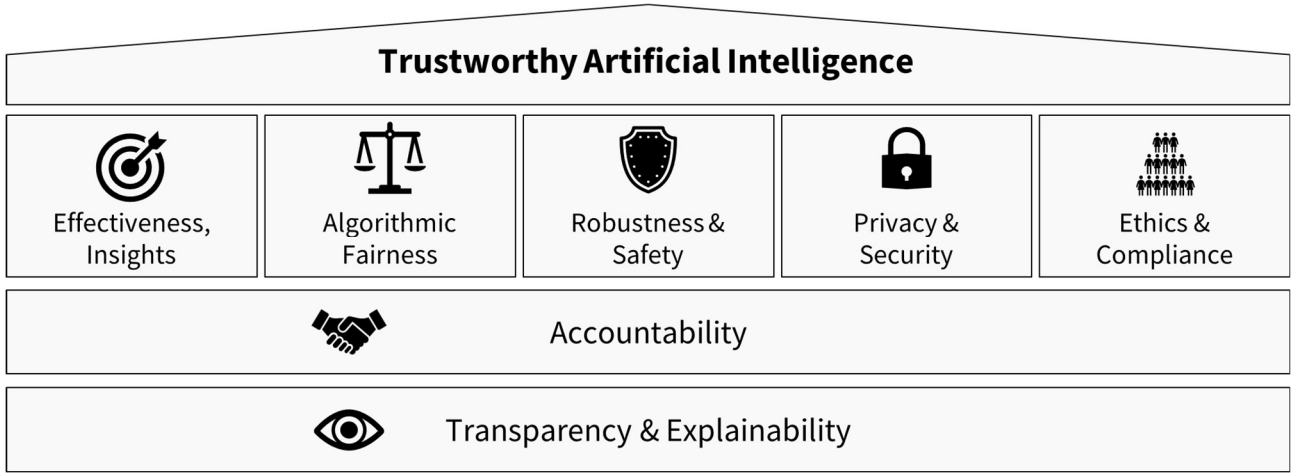


Quellen: Barredo et al. (2020); Doran et al. (2017); Doshi-Velez & Kim (2017); Hoffman et al. (2019); Langer et al. (2021)
https://www.research.ibm.com/artificial-intelligence/trusted-ai/

The Bigger Picture

Interpretable ML / xAI as key enabler for trustworthy AI

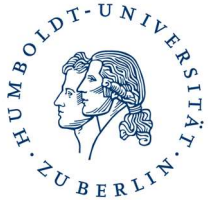




Quellen: Barredo et al. (2020); Doran et al. (2017); Doshi-Velez & Kim (2017); Hoffman et al. (2019); Langer et al. (2021)
https://www.research.ibm.com/artificial-intelligence/trusted-ai/

Target Audience in xAI

Explanations must consider domain knowledge



Users

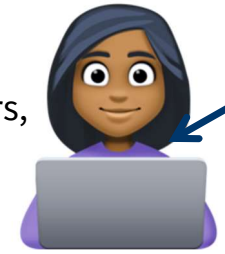
- Line managers, doctors, insurance agents, police, …
- Trust the model, gain knowledge, assess justifiability

Regulatory entities/agencies

- EU, GDPR, BaFin, …
- Fairness, Compliance, Audit

Product owners

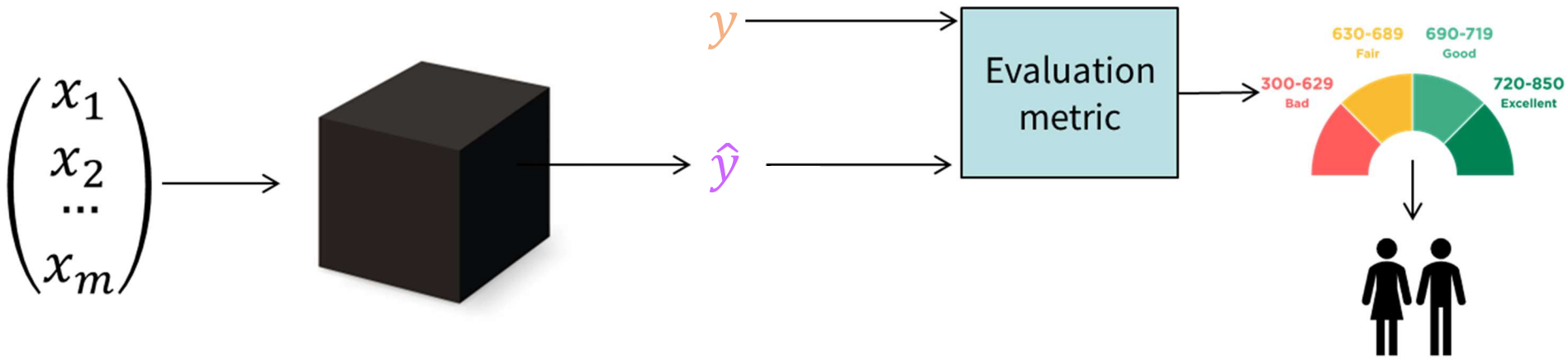
- Data scientists, developers, analysts, …
- Ensure/improve product efficiency, research , new functions



Users affected by AI decisions

- Clients, patients, applicants, accused, …
- Understand their situation, verify fair decisions

Why Interpretability Matters

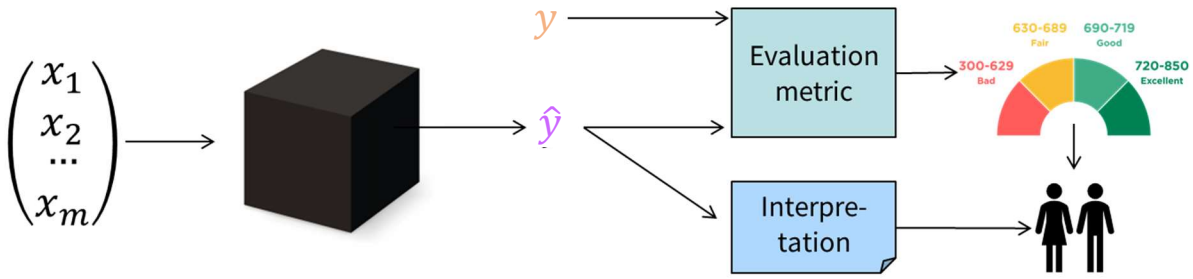


x_1 = Bureau Score
x_2 = Debt/Income
...
x_m = Age

Image based on Lipton (2016)

Why Interpretability Matters

- Machine learning increasingly used in mission-critical applications
- Measures of (predictive) performance provide an incomplete picture
- Gain insight into model-inferred patterns to identify ways to improve
- Confirm model acts in a manner that agrees with domain knowledge
- Auditing and robustness
- Comply with legal requirements and codes of conduct
  - Explain algorithmic decisions (e.g. EU GDPR Recital 71)
  - Verify that model outputs are not biased (e.g. do not discriminate against protected groups)



Lipton (2016)

Desiderata of an Interpretable Model

Based on Lipton (2016)

- **Trust**
  - Somehow covered by accuracy is understood as robustness
  - Trust in performance does not indicate free of bias (e.g., predictive policing)
- **Causality**
  - ML models are correlational
  - But interpretable ML model might provide clues about structural relationships one could test
- **Transferability**
  - Training versus deployment population: change over time and due to model deployment
  - Interpretation to help detect model vulnerability to failure and manipulation (e.g., adversarial examples)
- **Informativeness**
  - Output (i.e., prediction) often not enough to solve a problem (e.g., decision support context)
  - Reveal model inferred structure of the feature to target relationships and/or feature interactions
- **Fairness**
  - Interpretations to ensure whether model recommendations/decisions conform to ethical standards
  - Recidivism predictions are used to determine who to release and who to detain

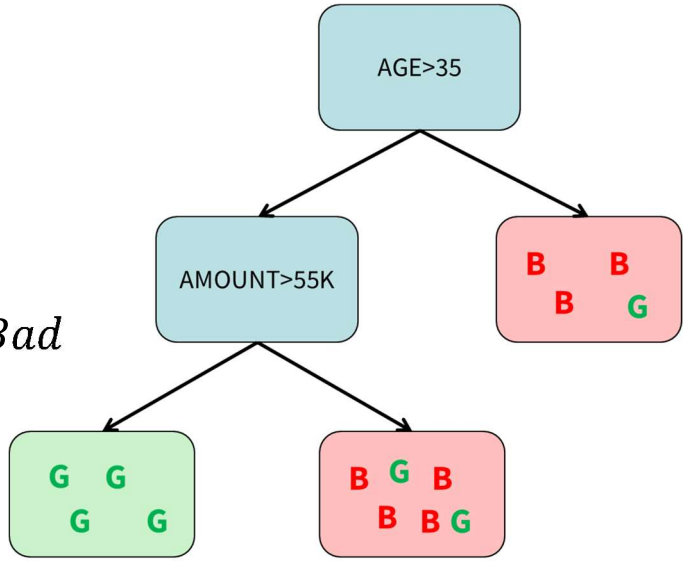
Scope of Interpretable Machine Learning

Various options to distinguish different approaches

- **Intrinsic interpretability versus post hoc explanation**
  - ML models with simple structure are intrinsically interpretable
  - Post hoc explanation methods are applied to fully trained models to provide
    - Feature summary statistics and/or visualizations
    - Example-based explanations
    - An intrinsically interpretable model, which explains some black-box model
- **Model-specific versus model-agnostic explanation methods**
- **Global versus local explanation methods**
  - Interpretation of the entire model
  - Interpretation of a single prediction
- **Entirely different topic: algorithmic transparency**

Intrinsically Interpretable (White-Box) Models

- **Linear models**
  - Linear and additive function of feature values and their weights
  - Interpretability ranking: linear regression > generalized linear models > generalized additive models
  - $\hat{y}(\boldsymbol{x}) = \phi\big(E(y|\boldsymbol{x})\big) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$
- **Classification and regression trees**
  - Set of simple decision rules
  - Interpretability decreases with depth
  - $If\ AGE > 35\ Then\ (If\ AMOUNT < 55\ Then\ Good\ Else\ Bad)\ Else\ Bad$
- **Interpretability of white-box models**
  - Understand how feature values cause model predictions
  - Understand how the forecast of a specific data instance emerges



*GLM: generalized linear model; GAM: generalized additive model

Global Explanation Methods

Surrogate models, permutation importance, partial dependence, and ICE

Surrogate Models / Pedagogical Rule Extraction

Use a white-box model to explain a black-box model

i	Product	List price	Age	Industry	…	Resale price [$]
1	Dell XPS 15’	2,500	36	Mining	…	347
2	Dell XPS 17’	3,000	36	Health	…	538
3	HP Envy 17’	1,300	24	Office	…	121
4	HP EliteBook 850	1,900	36	Mining	…	172
5	Lenovo Yoga 13’	1,100	12	Office	…	266

$f^{BB}(\boldsymbol{X})$

$\hat{Y}^{\text{BB}}$
315
475
144
201
188

1) Learn BB model and obtain forecasts

i	Product	List price	Age	Industry	…	$\hat{Y}^{\text{BB}}$
1	Dell XPS 15’	2,500	36	Mining	…	315
2	Dell XPS 17’	3,000	36	Health	…	475
3	HP Envy 17’	1,300	24	Office	…	144
4	HP EliteBook 850	1,900	36	Mining	…	201
5	Lenovo Yoga 13’	1,100	12	Office	…	188

2) Form training set for WB model:
BB forecasts becomes WB target

$$\hat{Y}^{WB} = w_0 + w_1 \text{Product} + w_2 \text{ListPrice} + w_3 \text{Age} + w_4 \text{Industry} + \cdots$$

3) Chose WB model (e.g., linear regression) and fit using the new training set

4) The fitted and interpretable WB model approximates how the BB model relates features to predictions.

Surrogate Models / Pedagogical Rule Extraction

Use a white-box model to explain a black-box model

- **White-box model provides an explanation of the black box through approximating the functional relationship the black box model inferred**
- **Trade-off between quality of explanation and interpretability**
  - WB model can only approximate BB model behavior
  - If WB model approximates too well it will lose interpretability (e.g., decision tree with many levels)
- **Examine adequacy of the surrogate model (see, e.g., Baesens et al., 2003)**
  - Complexity of the surrogate model. Is it interpretable?
    - Number of terms in the linear model
    - Conciseness of the rule set
  - Statistical measures of fit
    - R^2 for regression settings
    - Fidelity for classification: (a+d)/(b+c)

		BB classification	
		$\hat{Y}_{BB} = 1$	$\hat{Y}_{BB} = 0$
WB classification	$\hat{Y}_{WB} = 1$	a	b
	$\hat{Y}_{WB} = 0$	c	d

Multi-Stage Models

Combine interpretable white-box model with black-box model

i	Product	List price	Age	Industry	…	Resale price [$]
1	Dell XPS 15’	2,500	36	Mining	…	347
2	Dell XPS 17’	3,000	36	Health	…	538
3	HP Envy 17’	1,300	24	Office	…	121
4	HP EliteBook 850	1,900	36	Mining	…	172
5	Lenovo Yoga 13’	1,100	12	Office	…	266

1) Estimate white-box model

$$Y = f^{WB}(X) + \epsilon^{WB}$$

Apply WB model

$\hat{Y}^{WB} = f^{WB}(X)$	ϵ^{WB}
315	347-315=32
475	538-475=63
144	121-144=-23
201	172-201=-29
188	266-188=78

2) Calculate residuals

i	Product	List price	Age	Industry	…	ϵ^{WB}
1	Dell XPS 15’	2,500	36	Mining	…	32
2	Dell XPS 17’	3,000	36	Health	…	63
3	HP Envy 17’	1,300	24	Office	…	-23
4	HP EliteBook 850	1,900	36	Mining	…	-29
5	Lenovo Yoga 13’	1,100	12	Office	…	78

3) Fit black-box model to residuals

$$\epsilon^{WB} = f^{BB}(X) + \epsilon^{BB}$$

Apply BB model

$\hat{Y}^{BB} = f^{BB}(X)$	$\hat{Y} = \hat{Y}^{WB} + \hat{Y}^{BB}$	ϵ
29	344=315+29	347-344=3
66	541=475+66	538-541=-3
-18	126=144-18	121-126=-5
-26	175=201-26	172-175=-3
76	276=188+76	266-264=2

4) Integrated both models’ forecasts

$$\hat{Y} = \hat{Y}^{WB} + \hat{Y}^{BB} = f^{WB}(X) + f^{BB}(X)$$

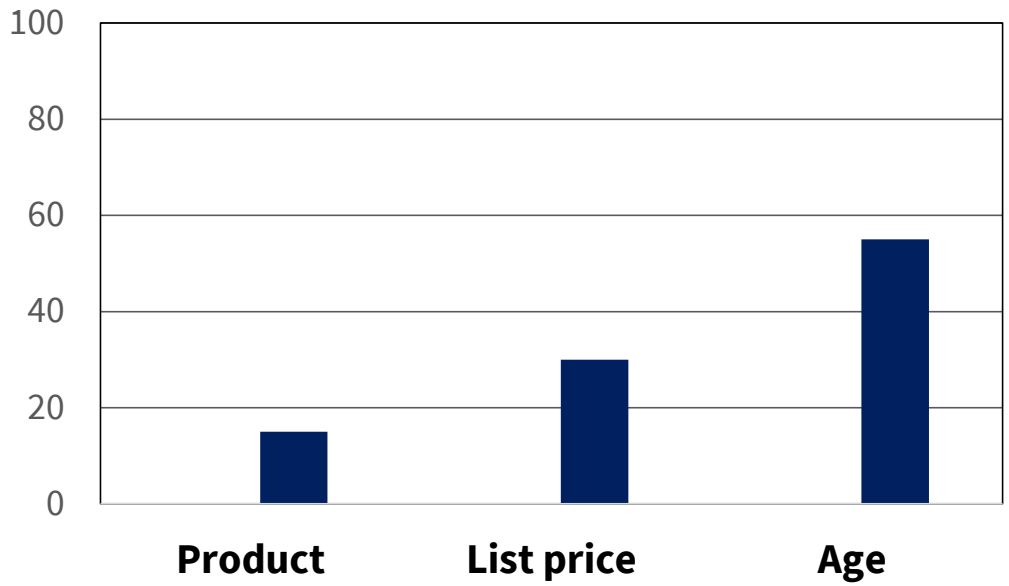
See, e.g., Kraus & Feuerriegel (2019)

Permutation-Based Feature Importance

A feature is important if accuracy decreases after corrupting that feature

- **Learner-agnostic way to judge the relevance of features**
- **Produces a feature ranking**
- **Importance score does not tell whether feature is positively/negatively correlated with the target**

Product	List price	Age	…	Resale price
Dell XPS 15’	2,500	36	…	347
Dell XPS 15’	2,500	24	…	416
Dell XPS 17’	3,000	36	…	538
HP Envy 17’	1,300	24	…	121
HP EliteBook	1,900	36	…	172
Lenovo Yoga 13’	1,100	12	…	266
…	…	…	…	…



Permutation-Based Feature Importance

A feature is important if accuracy decreases after corrupting that feature

- **Proposed in Breiman's (2001) paper on random forest**
- **Easily extendible to any type of predictive model**
- **Algorithm based on Fisher et al (2019)**
  - Input:
    Trained model f, feature matrix X, target vector y, error measure L(y,f).
  - Estimate the original model error e^{orig} = L(y, f(X)) (eg MSE)
  - For each feature j = 1,…,p do:
    - Generate feature matrix X^{perm} by permuting feature j in the data X. This breaks the association between feature j and the true outcome y.
    - Estimate error e^{perm} = L(Y,f(X^{perm})) based on the predictions of the permuted data.
    - Calculate permutation feature importance $FI^j = e^{perm}/e^{orig}$. Alternatively, the difference can be used: $FI^j = e^{perm} - e^{orig}$
  - Sort features by descending FI.

Permutation-Based Feature Importance

Concluding remarks

- **Training versus test data to compute feature importance**
- **Advantages**
  - Easy to understand
  - Accounts for feature interactions
  - Does not require costly re-estimation of a model (c.f. methods that delete features)
- **Disadvantages**
  - Requires labelled data
  - Possibly large variance with random permutation
  - Suffers from feature correlation
    - Biased toward unrealistic data instances
    - Correlated features ‘share’ importance, which might underestimate their merit
- **When using Random Forest, make sure you actually look at permutation importance (as opposed to Gini importance, see Strobl et al., 2007)**

Partial Dependence Plot (PDP)

Examine the marginal effect of a feature on model predictions

- **Proposed together with GBM in Friedman (2001)**
  - Depict how predictions change with changes in a feature when keeping everything else constant
  - Predictions correspond to forecasts of
    - The actual outcome variable in regression
    - Class-membership probabilities in classification
- **Formal definition of the partial dependence function**

$$\hat{f}_{x_s}(\boldsymbol{x}_s) = E_{\boldsymbol{x}_c}\left[\hat{f}_{x_s}(\boldsymbol{x}_s, \boldsymbol{x}_c)\right] = \int \hat{f}(\boldsymbol{x}_s, \boldsymbol{x}_c) d\mathbb{P}(\boldsymbol{x}_c)$$

  - With $\boldsymbol{x}_s$ denoting the features for which partial dependence is plotted, $\boldsymbol{x}_c$ the remaining features, and $\hat{f}$ the machine learning model
  - The features in $\boldsymbol{x}_s, \boldsymbol{x}_c$ make up the whole feature space
  - Usually the set S comprises one or two features

Partial Dependence Plot (PDP)

May reveal linear, monotonic or more complex relationship

- **Marginalizing over the features C gives a function that depends only on S, interactions with other features included**

Partial Dependence Plot (PDP)

Definition and estimation

- **Formal definition of the partial dependence function**

$$\hat{f}_{x_s}(\boldsymbol{x}_s) = E_{\boldsymbol{x}_c}\left[\hat{f}_{x_s}(\boldsymbol{x}_s, \boldsymbol{x}_c)\right] = \int \hat{f}(\boldsymbol{x}_s, \boldsymbol{x}_c) d\mathbb{P}(\boldsymbol{x}_c)$$

  - With $\boldsymbol{x}_s$ denoting the features for which partial dependence is plotted, $\boldsymbol{x}_c$ the remaining features, and $\hat{f}$ the machine learning model
  - The features in $\boldsymbol{x}_s, \boldsymbol{x}_c$ make up the whole feature space
  - Usually the set S comprises only one or two features

- **Estimate partial dependence function by calculating averages in the training set**

$$\hat{f}_{x_s}(\boldsymbol{x}_s) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}\left(\boldsymbol{x}_s, \boldsymbol{x}_c^{(i)}\right)$$

  - With $\boldsymbol{x}_c^{(i)}$ denoting the actual feature values from the data set and n the number of instances

Partial Dependence Plot (PDP)

PDP calculations exemplified

Raw data

i	Y	$\hat{p}(Y = 1 \mid X)$	**CREDIT AMOUNT** X_1	**SALARY** X_2	**DURATION** X_3
1	**1**	0.45	$25,000	$75,000	2 - 5 years
2	**1**	0.55	$25,000	$65,000	< 2 years
3	**1**	0.33	$50,000	$55,000	>5 years
4	**0**	0.53	$25,000	$45,000	< 2 years
5	**0**	0.79	$50,000	$111,000	2 - 5 years

Partial Dependence Plot (PDP)

PDP calculations exemplified

Raw data

i	Y	$\hat{p}(Y = 1\|X)$	CREDIT AMOUNT X_1	SALARY X_2	DURATION X_3
1	1	0.45	$25,000	$75,000	2 - 5 years
2	1				
3	1				
4	0				
5	0				

i	Y	$\hat{p}(Y = 1\|X)$	CREDIT AMOUNT X_1	SALARY X_2	DURATION X_3	
1	1	Compute model prediction	$25,000	$75,000	2 - 5 years	Average
2	1		$25,000	$65,000	< 2 years	
3	1		$25,000	$55,000	>5 years	
4	0		$25,000	$45,000	< 2 years	
5	0		$25,000	$111,000	2 - 5 years	
1	1		$50,000	$75,000	2 - 5 years	Average
2	1		$50,000	$65,000	< 2 years	
3	1		$50,000	$55,000	>5 years	
4	0		$50,000	$45,000	< 2 years	
5	0		$50,000	$111,000	2 - 5 years	

Partial Dependence Plot (PDP)

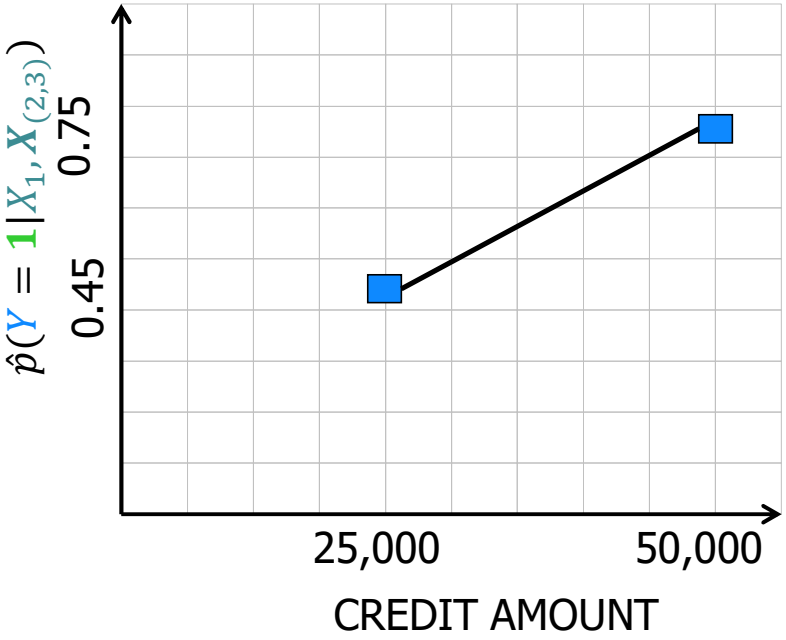
PDP calculations exemplified

Raw data

i	Y	$\hat{p}(Y=1\|X)$	CREDIT AMOUNT X_1	SALARY X_2	DURATION X_3
1	1	0.45	$25,000	$75,000	2 - 5 years
2	1				
3	1				
4	0				
5	0				

i	Y	$\hat{p}(Y=1\|X)$	CREDIT AMOUNT X_1	SALARY X_2	DURATION X_3
1	1		$25,000	$75,000	2 - 5 years
2	1				rs
3	1				rs
4	0				rs
5	0	Compute model prediction			ars
1	1		$50,000	$75,000	2 - 5 years
2	1		$50,000	$65,000	< 2 years
3	1		$50,000	$55,000	>5 years
4	0		$50,000	$45,000	< 2 years
5	0		$50,000	$111,000	2 - 5 years

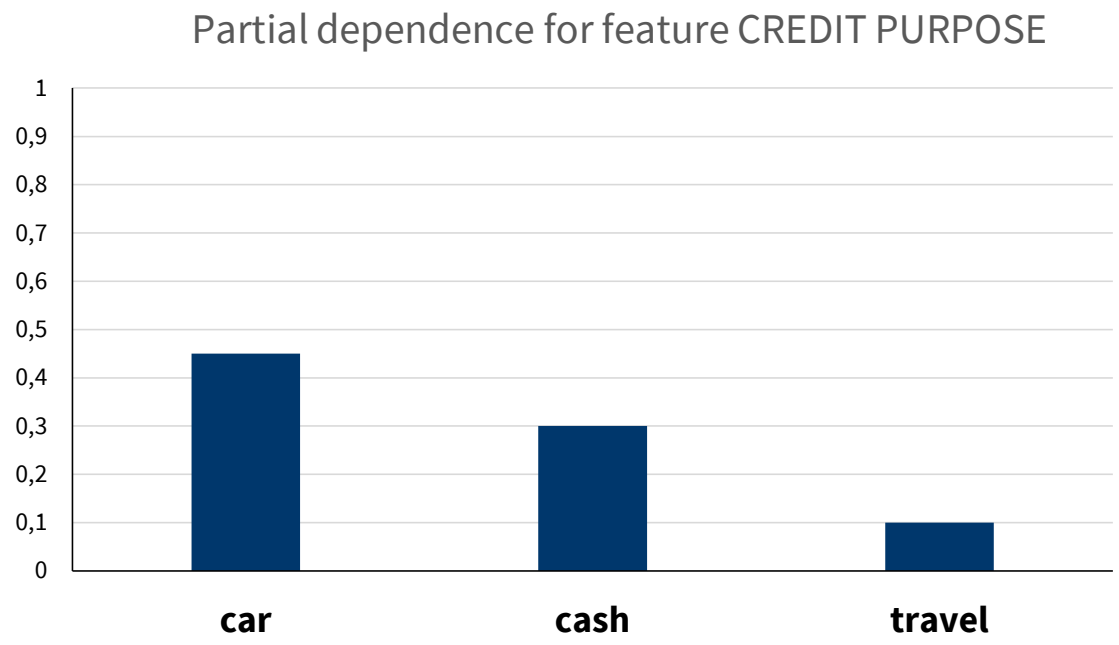
$\hat{p}(Y=1\|X_1, X_{(2,3)})$	CREDIT AMOUNT X_1
0.45	$25,000
0.75	$50,000



Partial Dependence Plot (PDP)

Categorical features: force all data instances to have the same level

- **For categorical features, force all data instances to have the same level**
  - Consider feature credit purpose with levels *car*, *cash*, and *travel*
  - To compute value for *car*, replace category level for all data instances with this value and average over model forecasts
  - Proceed in the same way with the other levels



Partial Dependence Plot (PDP)

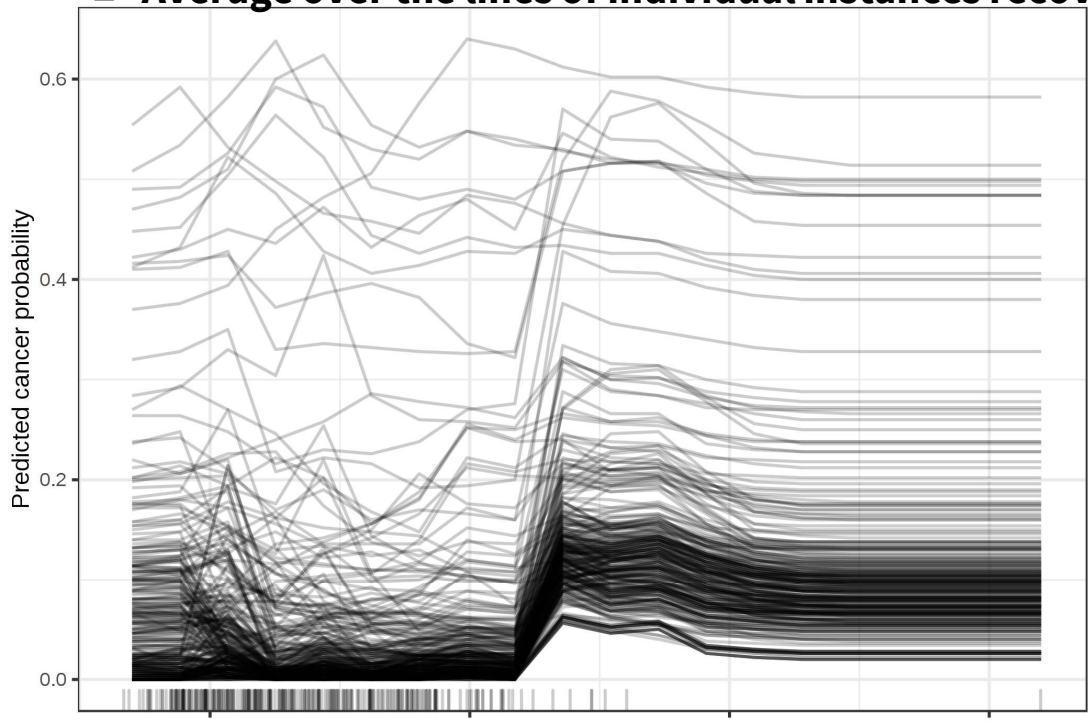
Concluding remarks and appraisal

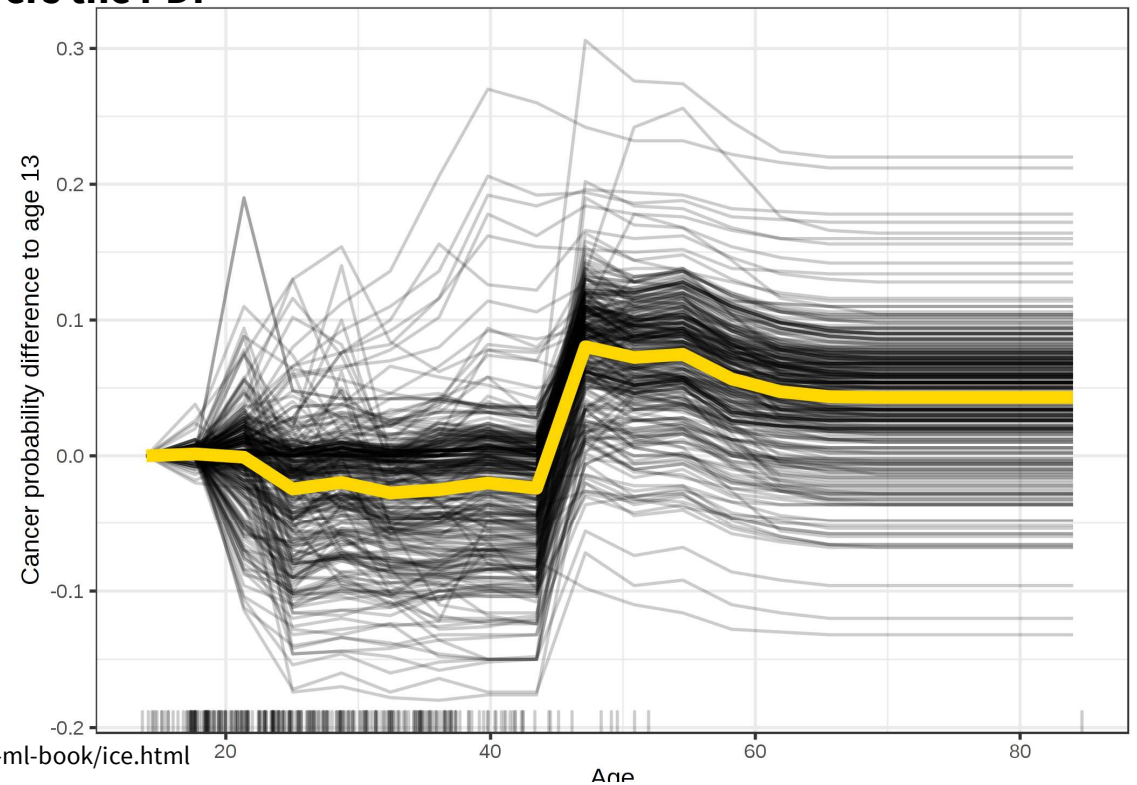
- **Easy to understand and implement**
- **Causal interpretation of the model prediction**
- **Make sure to examine feature distribution**
- **Does not scale to more than two features per plot**
- **Assume that features in set S and set C are not correlated**
  - Correlation can lead to implausible data instances during averaging
  - Possible remedy: Accumulated Local Effect (ALE) plots (Apley & Zhu, 2016)
- **Heterogeneous effects may be overlooked**
  - Say feature has pos. correlation with outcome for males and neg. correlation with females
  - Average effect of the feature is zero → PDP plot shows a horizontal line
  - Individual Conditional Expectation Curves (ICE) can uncover interaction effects

Example of ICE and centered ICE Plot

Depict changes in an instance's prediction with changes in one feature

- **Actually not a global but a local method (see below)**
- **Vary values of focal feature over a grid while keeping values of all other features fixed**
- **Average over the lines of individual instances recovers the PDP**





https://christophm.github.io/interpretable-ml-book/ice.html

Local and Example Based Explanations

LIME, SHAP values, counterfactual and adversarial examples

Local Interpretable Model-Agnostic Explanations (LIME)

Ribeiro et al. (2016)

- **Interpretable local surrogate model explains individual-level predictions**
- **Credit scoring example**

Bureau score (BS): 600
Debt-to-income (DI): 25%
Age: 28
… …

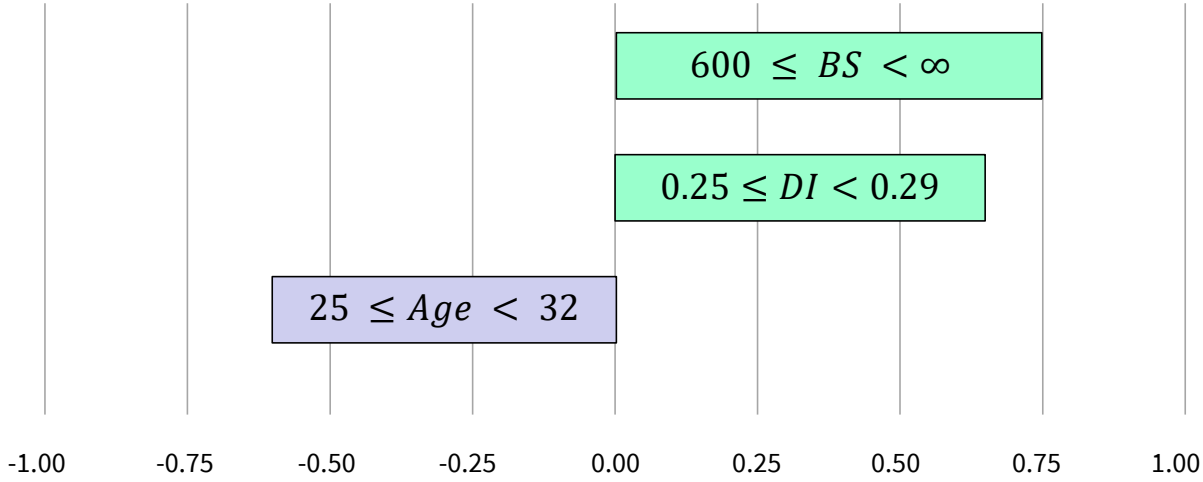
Features

True Class: good risk

Model estimate $\hat{p}(y_i = 1|\boldsymbol{x}_i)$

$f(\boldsymbol{x}_i)$: 0.87

$g(\tilde{\boldsymbol{x}}_i)$: 0.80



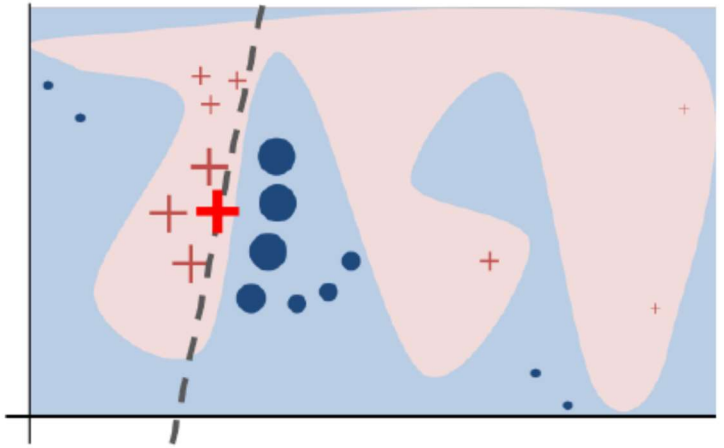
Notation

Target variable	$y \in \{0,1\}$	Black-box model	$f: \mathbb{R}^p \rightarrow [0,1]$
Positives	$(y = 1)$	White-box model	$g \in \mathcal{G}$
Negatives	$(y = 0)$	Interpretable representation of $\boldsymbol{x}$	$\tilde{\boldsymbol{x}} \in \mathbb{R}^{\tilde{p}}$
Application	$x \in \mathbb{R}^p$		
Training data	$\mathcal{D} = \{y_i, \boldsymbol{x}_i\}_{i=1}^n$		

Local Interpretable Model-Agnostic Explanations (LIME)

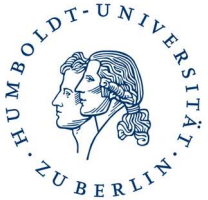
Characteristic features

- **Model-agnosticism**
  - No assumptions about the black-box model and thus works with any model
  - Based on inputs and outputs of the black-box
- **Locality**
  - Explains the prediction of an individual data instance
  - Approximates black-box model by an interpretable model in the neighborhood of that instance
- **Interpretability**
  - Uses some interpretable model (see above) as explanation model
  - Incorporates feature selection



Ribeiro et al (2016)

Fidelity-Interpretability Trade-Off



- **Formalization of the minimization problem**

$$\xi(\boldsymbol{x}_i) = \underset{g \in \mathcal{G}}{\operatorname{argmin}}\, \Omega(g) + \mathcal{L}\left(f, g, \pi_{\boldsymbol{x}_i}\right)$$

  - $\boldsymbol{x}_i$: Instance
  - f: Black-box model with prediction function $f(\boldsymbol{x}_i)$
  - $\mathcal{G}$: Set of interpretable functions / model classes (e.g., linear model)
  - $\pi_{\boldsymbol{x}_i}$: Function defining a neighborhood of instance $\boldsymbol{x}_i$
  - $\Omega(g)$: Complexity of explanation g
  - $\mathcal{L}\left(f, g, \pi_{\boldsymbol{x}_i}\right)$: Loss function measuring the accuracy of the approximation of f by g
  - $\xi(\boldsymbol{x}_i)$: LIME explanation

- **Alternative implementations for image, text, and structured data**

Pseudo-Code Representation

- **Inputs:**
  - Decision instance ($\boldsymbol{x}_i$), Black-box model (f)
  - Other meta-parameters
- **Generate synthetic data points $\boldsymbol{z}_j$ in the neighborhood of $\boldsymbol{x}_i$**
- **Get black-box prediction $f(\boldsymbol{z}_j)\ \forall j$**
- **Estimate interpretable model $g \in \mathcal{G}$**
  - Training data: $\left\{f(\boldsymbol{z}_j), \boldsymbol{z}_j, \pi_{\boldsymbol{x}_j}(\boldsymbol{z}_j)\right\}_{j=1}^{J}$
  - Loss function: $\min_{g \in \mathcal{G}} \Omega(g) + \mathcal{L}(f, g, \pi_{\boldsymbol{x}_i})$
- **Explain black-box prediction $f(\boldsymbol{x}_i)$ by white-box approximation $g(\boldsymbol{x}_i)$**

Local Interpretable Model-Agnostic Explanations (LIME)

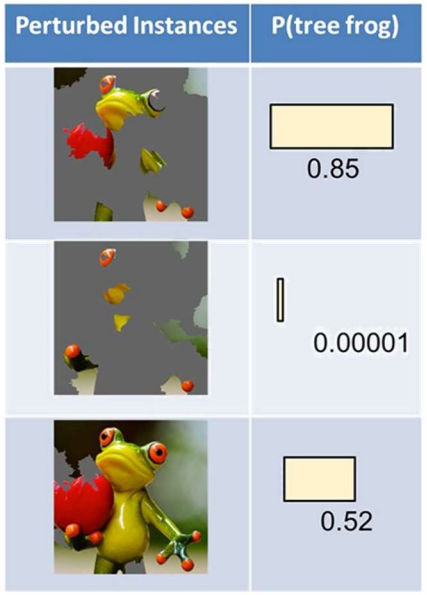
LIME supports structured and unstructured data

- **Perturbation also depends on the type of data**
  - For text data, perturbation works by masking single words
  - For image data, perturbation works by switching super-pixels on or off
  - For tabular data, LIME perturbs features individually, drawing from a normal distribution with mean and standard deviation taken from that feature

Original Image

Interpretable Components

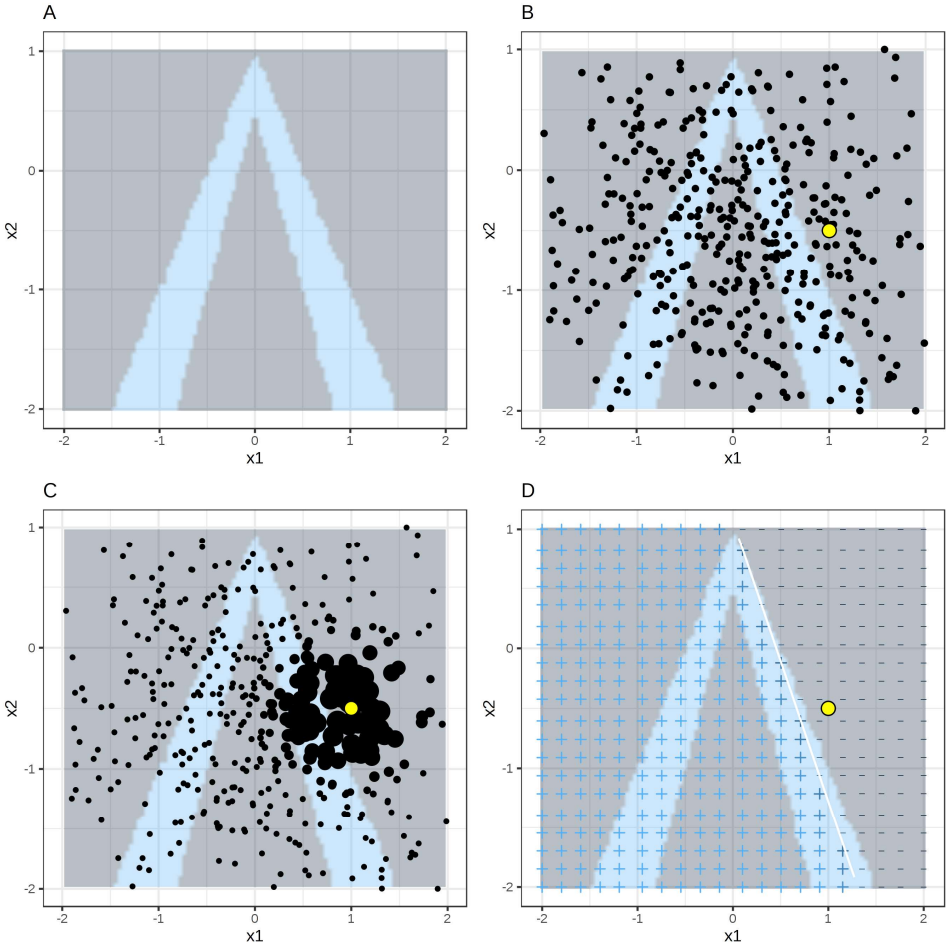
Original Image
P(tree frog) = 0.54

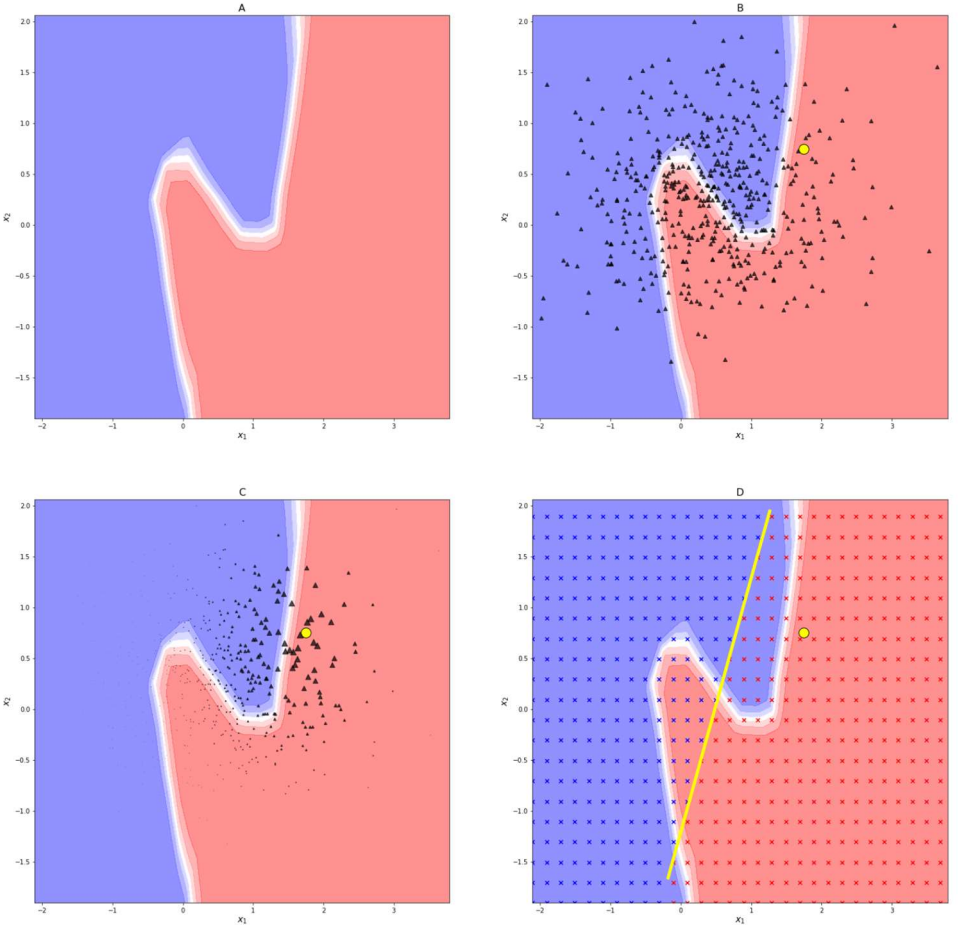


https://towardsdatascience.com/understanding-how-lime-explains-predictions-d404e5d1829cv

Local Interpretable Model-Agnostic Explanations (LIME)

LIME for structured tabular data



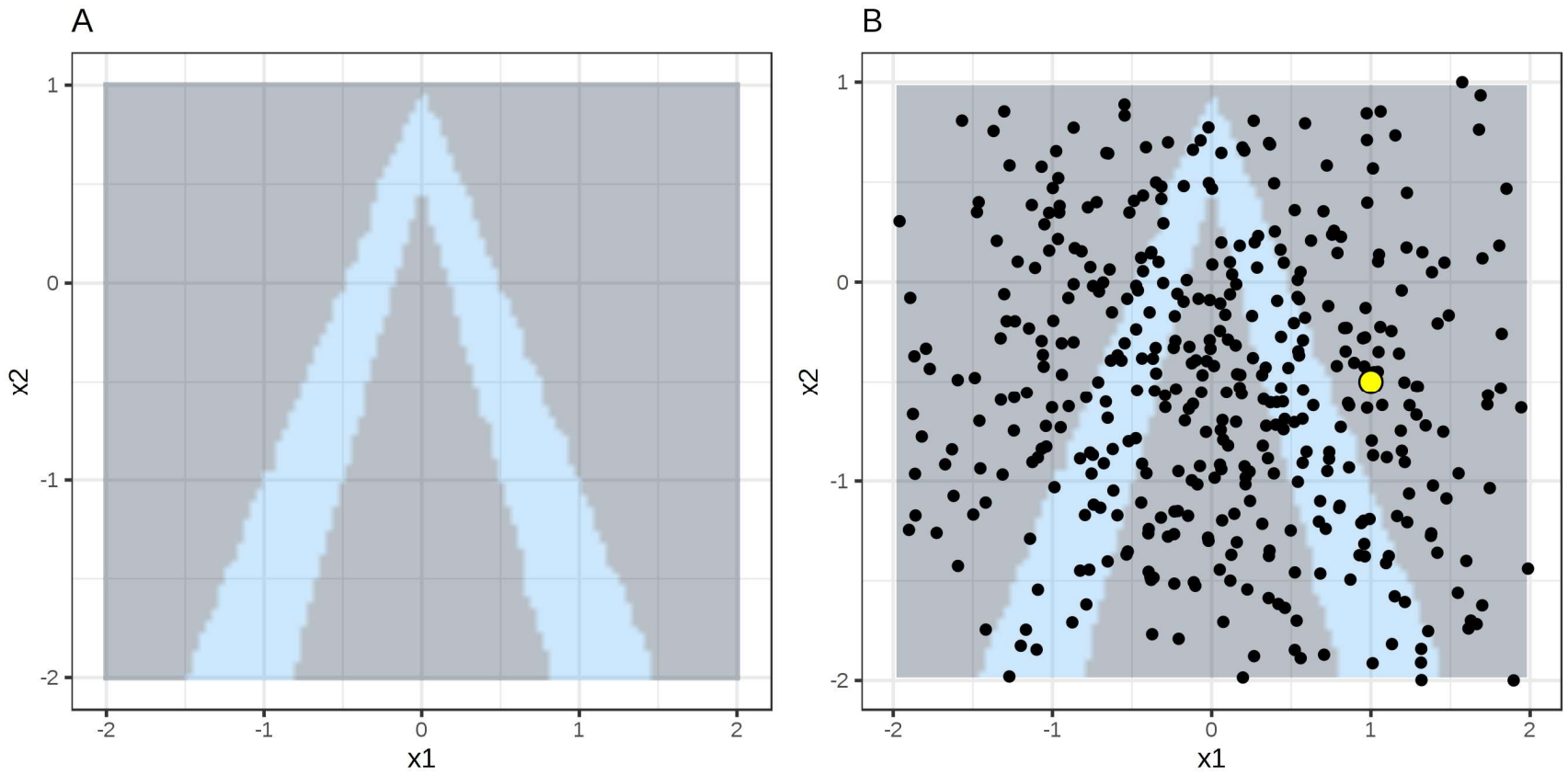


https://christophm.github.io/interpretable-ml-book/lime.html#fn37

https://github.com/stefanlessmann/lime_from_scratch

Local Interpretable Model-Agnostic Explanations (LIME)

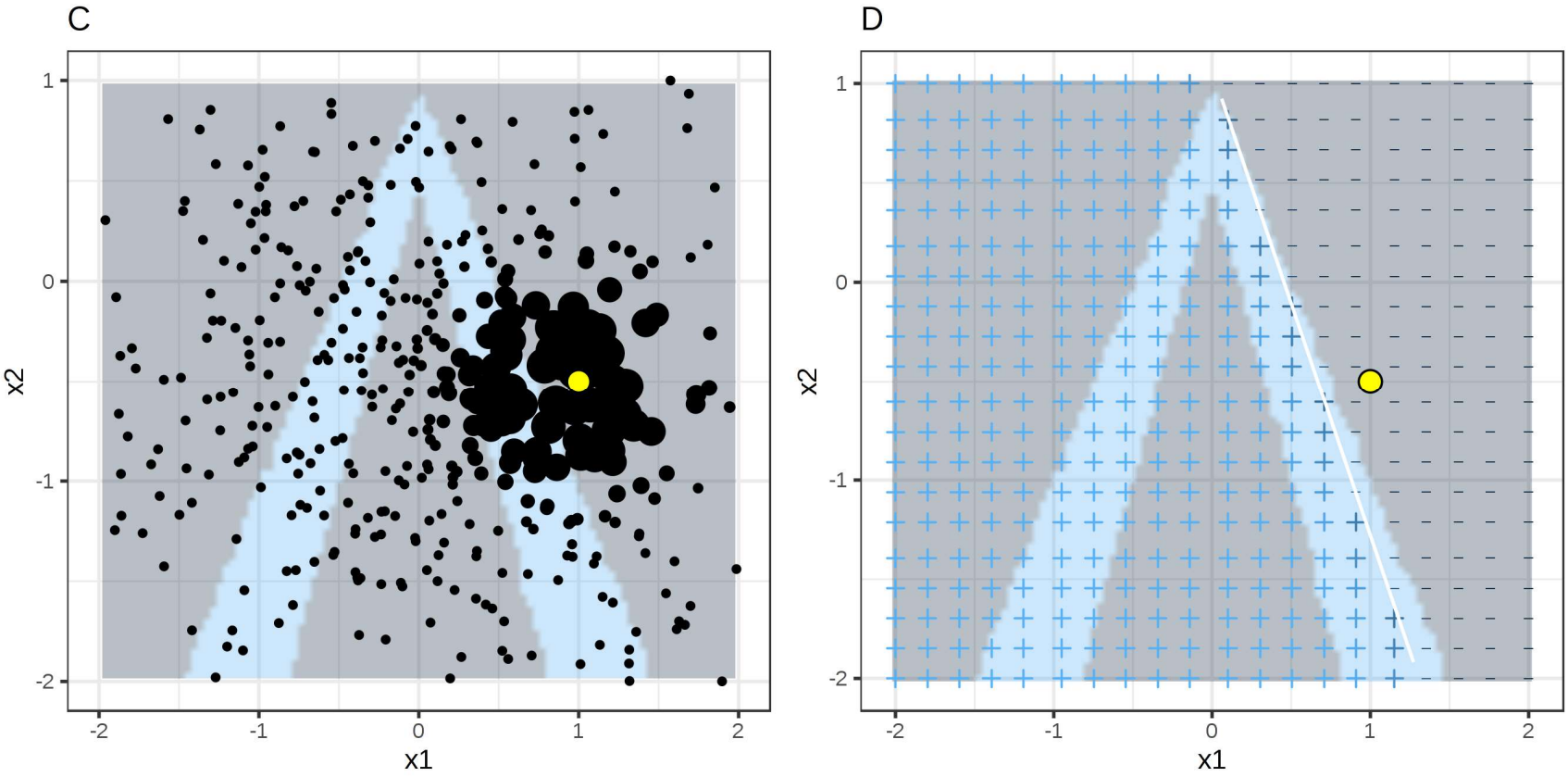
LIME for structured tabular data



https://christophm.github.io/interpretable-ml-book/lime.html#fn37

Local Interpretable Model-Agnostic Explanations (LIME)

LIME for structured tabular data



https://christophm.github.io/interpretable-ml-book/lime.html#fn37

Local Interpretable Model-Agnostic Explanations (LIME)

Concluding remarks and appraisal

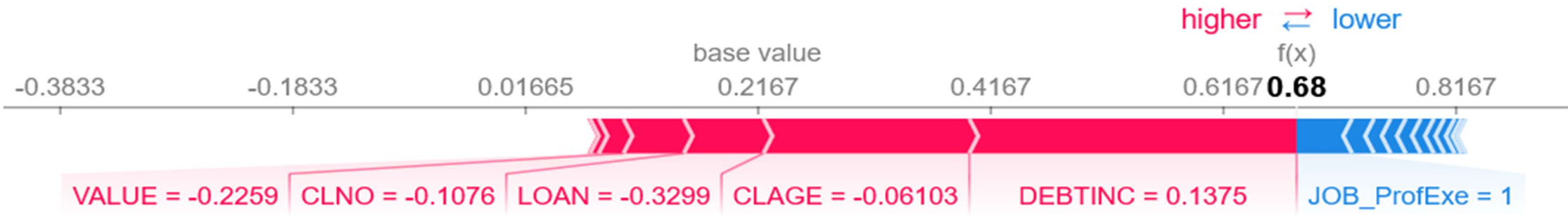
- **Easy to use and good software support (e.g., Python library *lime*)**
- **Independence between prediction and local surrogate model**
  - Could change (e.g. update) prediction model while keeping explanation model
  - Could also use different, more interpretable features for the explanation model
- **Concise explanation**
  - Feature selection results in human-friendly explanation. Easy to explain prediction to laymen
  - Fidelity measure gives good idea how reliable the interpretable model explains the black-box
- **Defining a meaningful neighborhood is difficult**
- **Sampling data instances from a Gaussian while ignoring feature correlation**
  - Can result in the generation of unlikely data instances
  - Simulation studies also indicate high variance in explanations (e.g., Alvarez-Melis et al., 2018)

https://christophm.github.io/interpretable-ml-book/lime.html#fn37

SHAP (SHapley Additive exPlanations)

Lundberg & Lee (2017)

- **Goal is to explain the prediction of an instance x by computing the contribution of each feature to that prediction**
- **Credit scoring example (https://github.com/Humboldt-WI/bads/blob/master/tutorials/10_nb_interpretable_ml.ipynb)**



SHAP (SHapley Additive exPlanations)

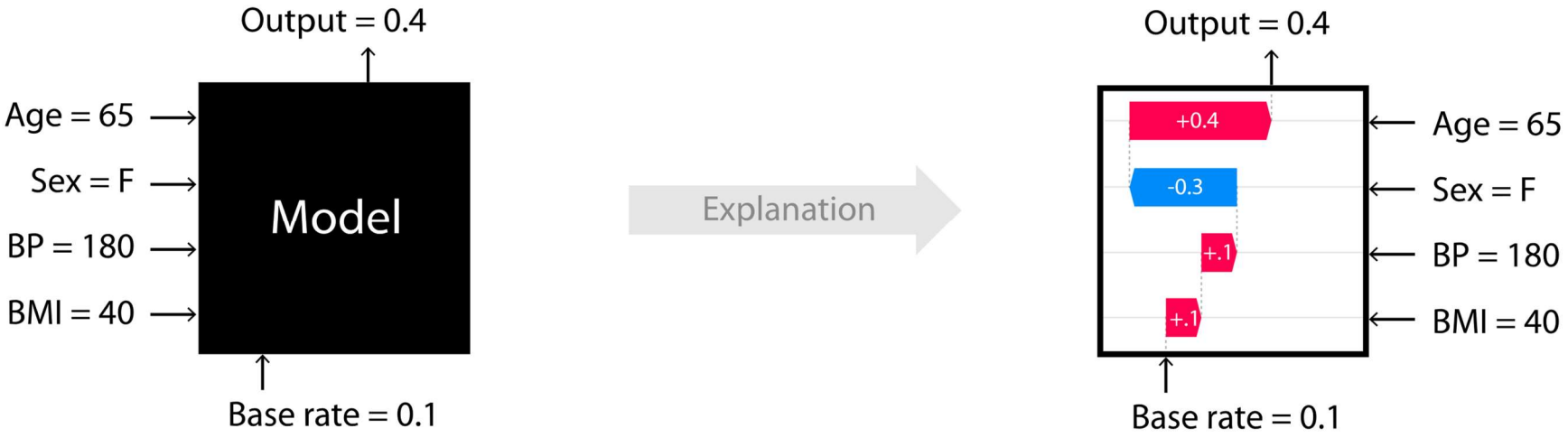
Explain prediction of x by computing the contribution of each feature to that prediction

- **Based on the Shapley value from cooperative game theory**
  - Fair payout of a player in a cooperative game
  - Shapely value of a player is the difference between the payout with her playing and w/o her playing
  - Average Idea is to average over all possible coalitions of players
- **Applications to supervised ML**
  - Features act as player
  - Forecast act as game payout
  - Study how a feature value impacts forecast (i.e., how *payout* changes)
- **Computationally very costly and typically intractable**
- **Contributions of Lundberg & Lee (2017, 2020)**
  - Model agnostic approach to approximate Shapely values (KernelSHAP)
  - Fast approximations for specific models (tree-based & deep learning)

Many options to analyze global importance, interactions, … with SHAP

Check out: https://github.com/slundberg/shap

SHAP



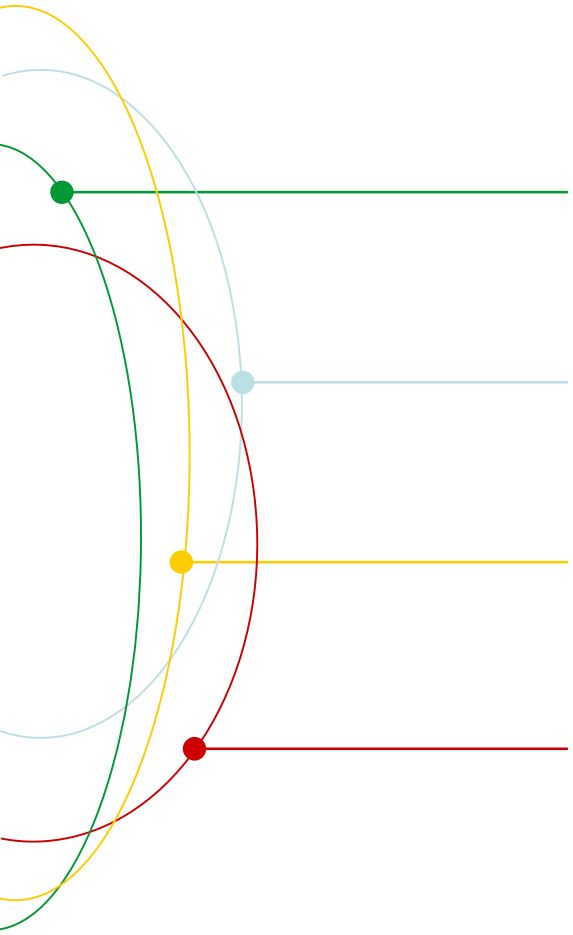
Recent IML Developments

See Molnar (2019) for an introduction

- **Anchors (Ribeiro et al., 2018)**
  - Local explanation method that finds a decision rule that “anchors” the prediction sufficiently
  - A rule anchors a prediction if changes in other feature values do not affect the prediction
- **Counterfactual examples**
  - Local explanation method that treats feature values as causes of an instance’s prediction
  - A counterfactual explanation describes the smallest change to the feature value(s) that changes the prediction to a predefined output
    - Counterfactual examples are new/artificial instances
    - Possibly many and contradictory counterfactuals
- **Adversarial examples**
  - As counterfactual examples but with the aim to deceive the model (eg min. change to make an error)
  - Much current research in the scope of computer vision (eg misinterpret a sign in self-driving cars)

Summary

Summary



Learning goals

- Option space for interpreting ML models
- Understanding of specific approaches

Findings

- Crucial to report and diagnose ML-based systems
- Local versus global interpretation
- Example-based techniques
- Model-specific vs. –agnostic explanation methods
- Often use interpretable models as component
- Working of permutation importance, PDP, LIME

What next

- Finish last exercise notebook
- Summary and Q & A
- BADS exam

Literature

- Alvarez-Melis, D., & Jaakkola, T. S. (2018). *On the Robustness of Interpretability Methods*. ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden.
- Apley, D. W., & Zhu, J. (2016). Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *Archive Preprint,* arXiv:1612.08468v2.
- Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science,* 49(3), 312-329.
- Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research,* 20(177), 1-81.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*, 1189-1232.
- Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine,* 38(3), 50-57.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning* (2nd ed.). New York: Springer.
- Kraus, M., & Feuerriegel, S. (2019). Forecasting remaining useful life: Interpretable deep learning approach via variational Bayesian inferences. Decision Support Systems, 125, 113100.
- **Molnar, C. (2019) *Interpretable Machine Learning*. E-Book. https://christophm.github.io/interpretable-ml-book/**
- Lipton, Z. C. (2016). The Mythos of Model Interpretability. *ICML 2016 Workshop on Human Interpretability in Machine Learning (WHI 2016)*, New York, NY, USA.
- Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. Advances in Neural Information Processing Systems, pp. 4765-4774.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "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 (ACM KDD2016), ACM: New York, NY, USA.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). *Anchors: High-Precision Model-Agnostic Explanations*. Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18).
- Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics,* 8(1), 1-20.

Thank you for your attention!

Stefan Lessmann

Chair of Information Systems
School of Business and Economics
Humboldt-University of Berlin, Germany

Tel. +49.30.2093.5742
Fax. +49.30.2093.5741

stefan.lessmann@hu-berlin.de
http://bit.ly/hu-wi

www.hu-berlin.de