Local model-agnostic explanations

Szymon Bobek

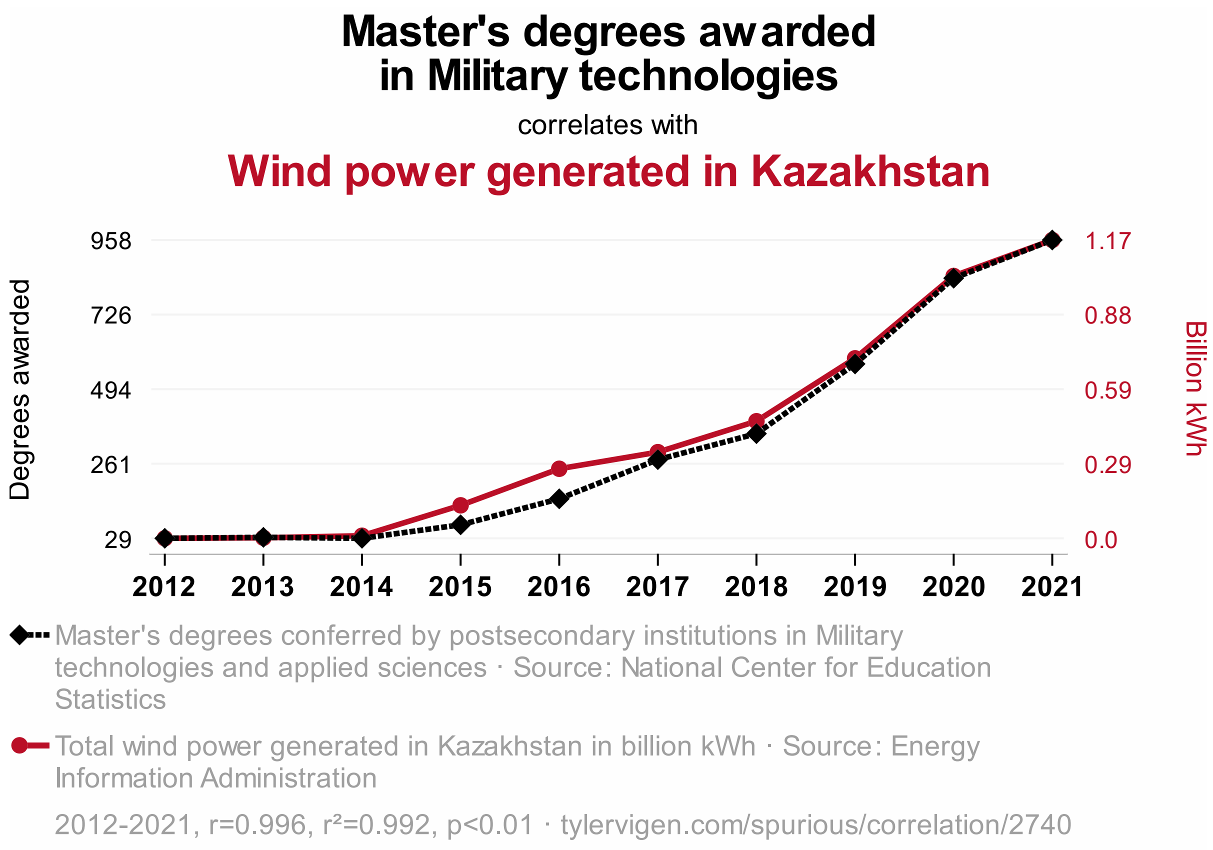
Jagiellonian University
2024

JAGIELLONIAN UNIVERSITY
IN KRAKÓW

https://geist.re

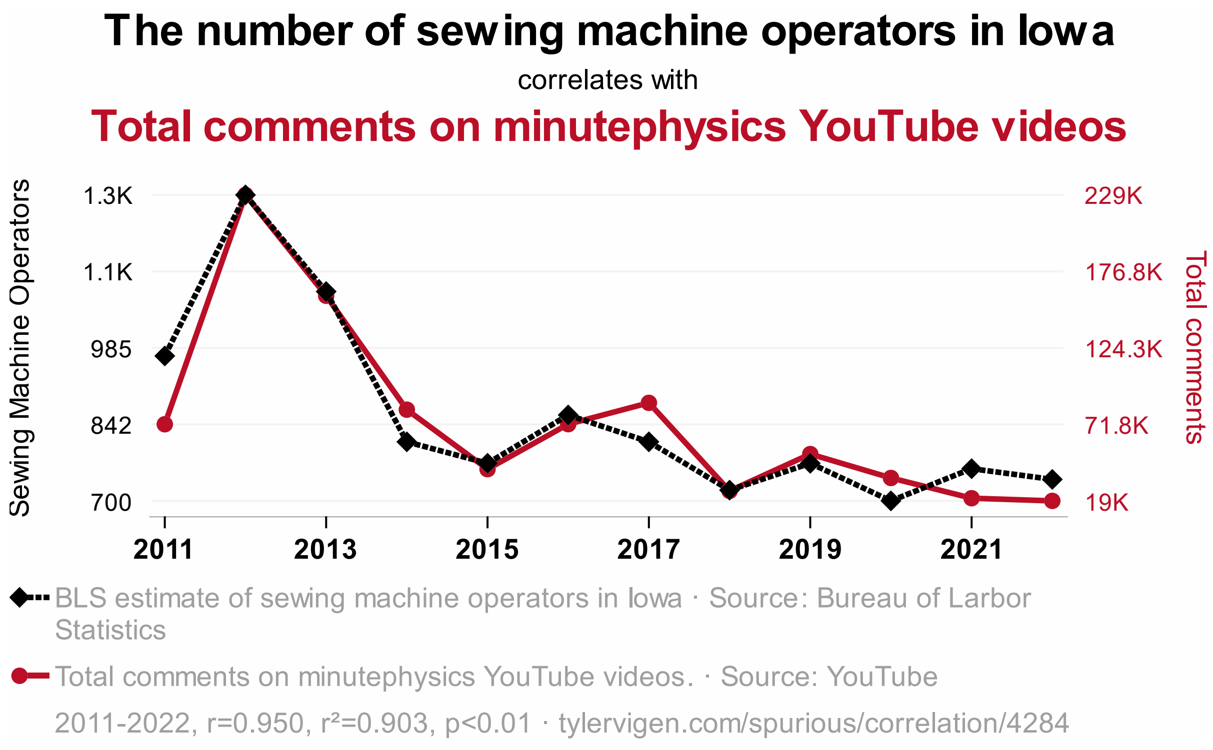
Fun with XAI - Correlation does not imply causation

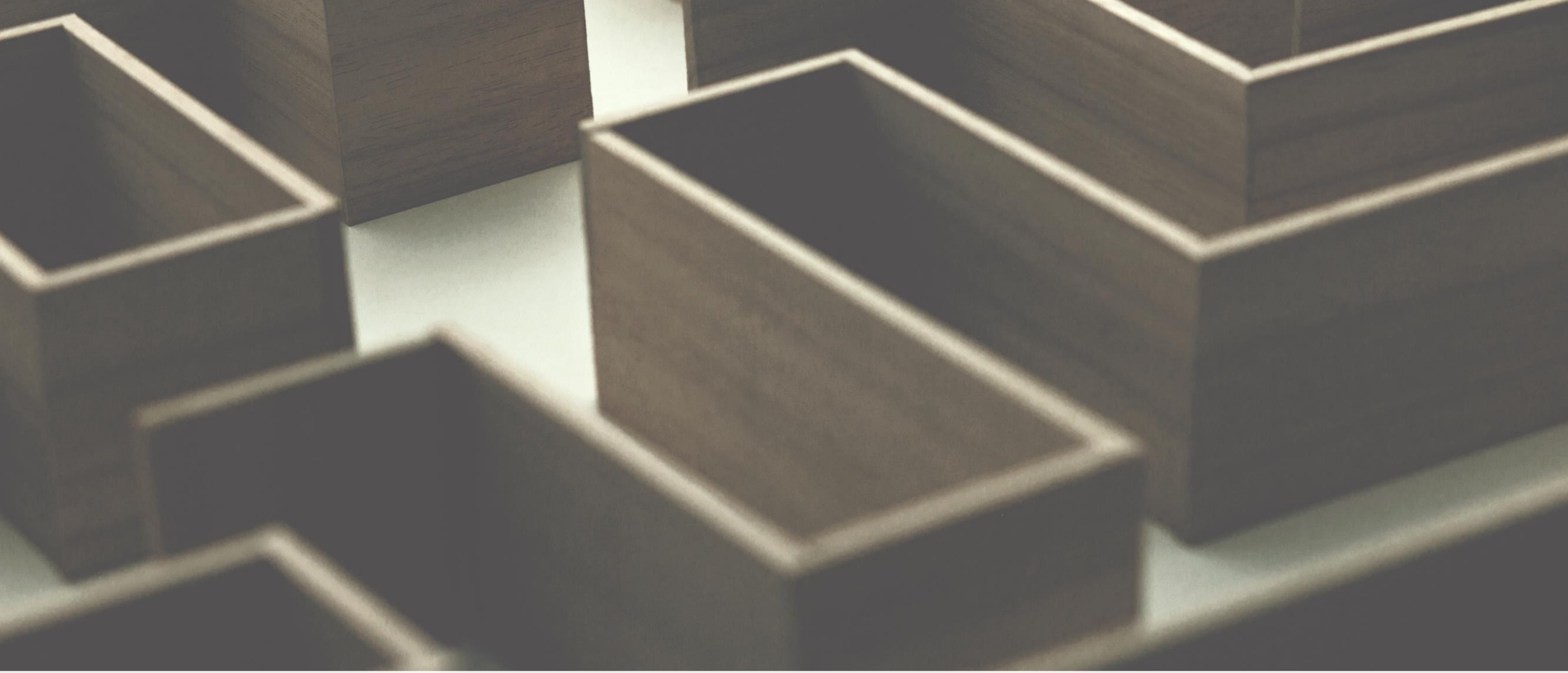
- It is not possible to legitimately deduce a cause-and-effect relationship between two events or variables solely on the basis of an observed association or correlation between them
- Most of ML methods and scientific evidence is based upon correlation of variables
- Explainable AI is not an exception
- All models are wrong, but some are useful (and some are not in some cases)



Fun with XAI - Correlation does not imply causation

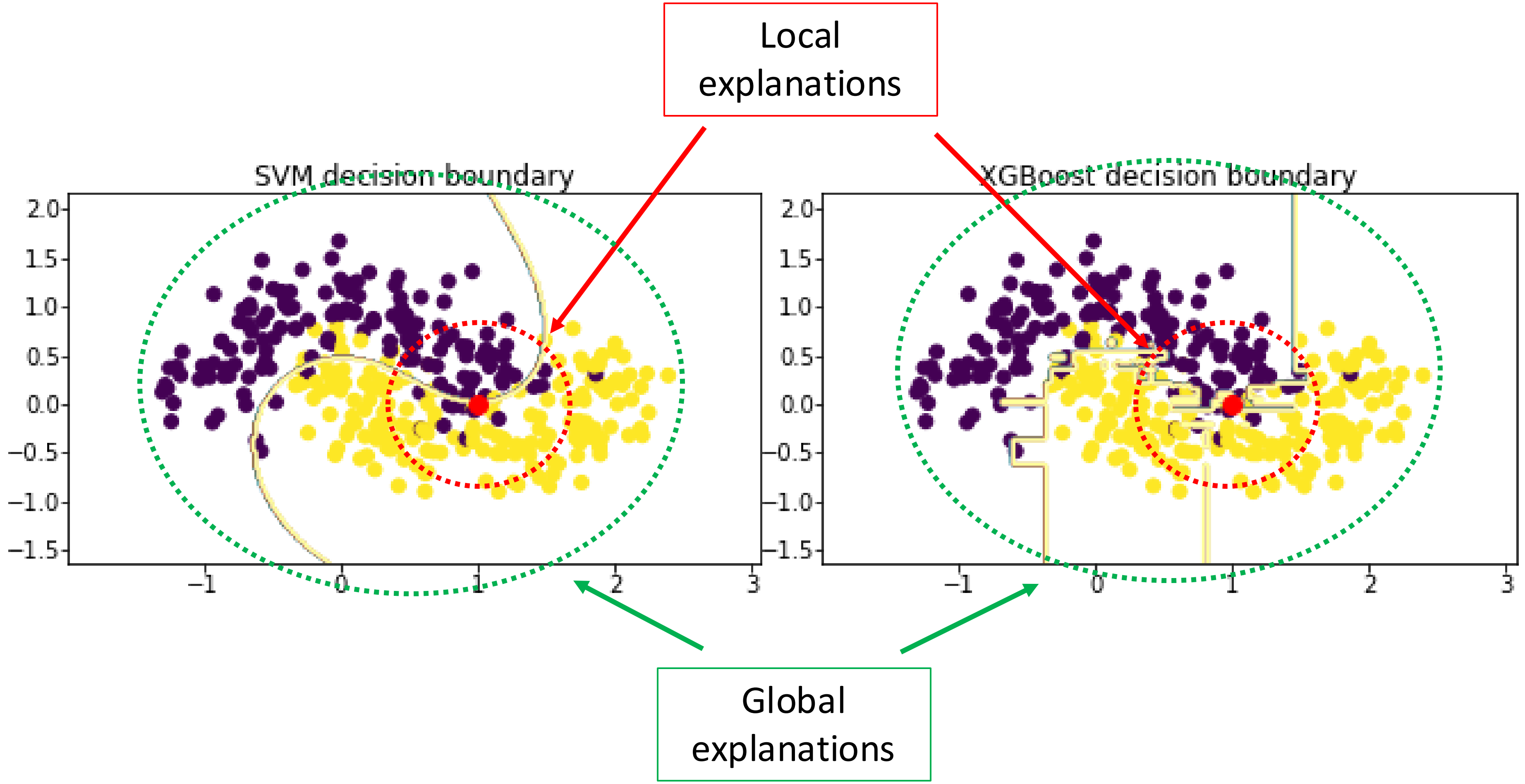
- It is not possible to legitimately deduce a cause-and-effect relationship between two events or variables solely on the basis of an observed association or correlation between them
- Most of ML methods and scientific evidence is based upon correlation of variables
- Explainable AI is not an exception
- All models are wrong, but some are useful (and some are not in some cases)





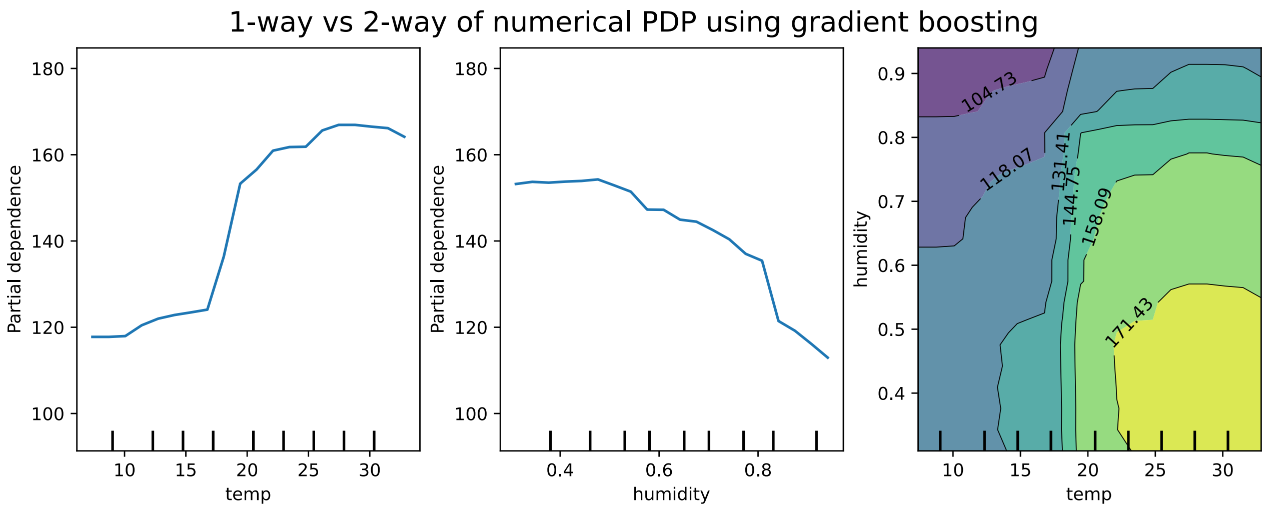
Local, model-agnostic explanations

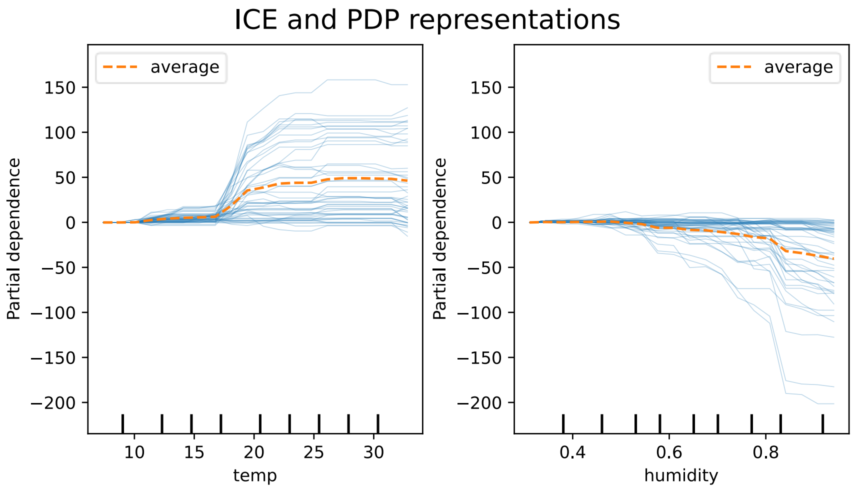
Local vs Global explanations



Individual Conditional Expectation are local?

- (ICE) plots display one line per instance that shows how the instance’s prediction changes when a feature changes.
- For convenience we start from 0 by subtracting from all plots the prediction of the lower value of the feature of consideration
- The average of ICE curves from the PDP
- It is even easier to spot if there are interactions captured by model. If the ICE curves are not parallel, there are some interactions
- They give more insight into data, as average may cancel out some opposite effects





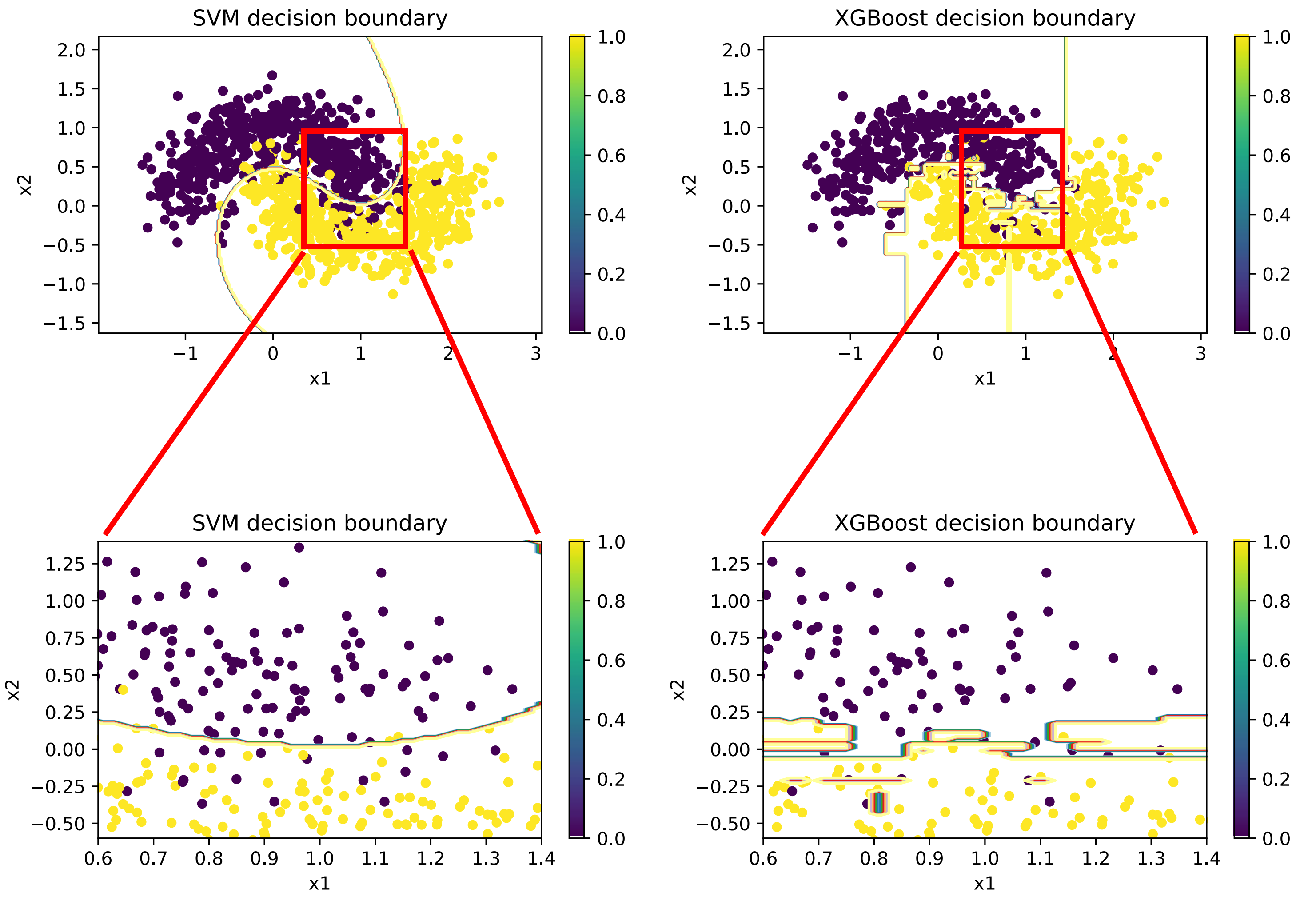
$$\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)})$$

$$\hat{f}^{(i)}_{cent}(x_S) = \hat{f}^{(i)} - \hat{f}(x^a, x_C^{(i)})$$

Anchor point, usually the lower value of a feature we are plotting

LIME

Locally, the decision boundary is simpler



- In this approach we focus on explaining an instance
- "Zooming in" we can fit inherently interpretable model that will approximate the decision of the blackbox one
- The assumption is not always valid. There are models which have complex decision boundary even locally
- Term "Locally" is vague. The locality is subjective
- When zooming in, we are limiting the number of samples that can be used for training
- What in case of instances that are far from the distribution?

Local Model-Agnostic Surrogate Model

Locality-aware loss function

Black-box model

Local surrogate model

$$\phi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Interpretable model

Size of the neiborhood (locality)

Complexity of interpretable model

Why should I trust you?

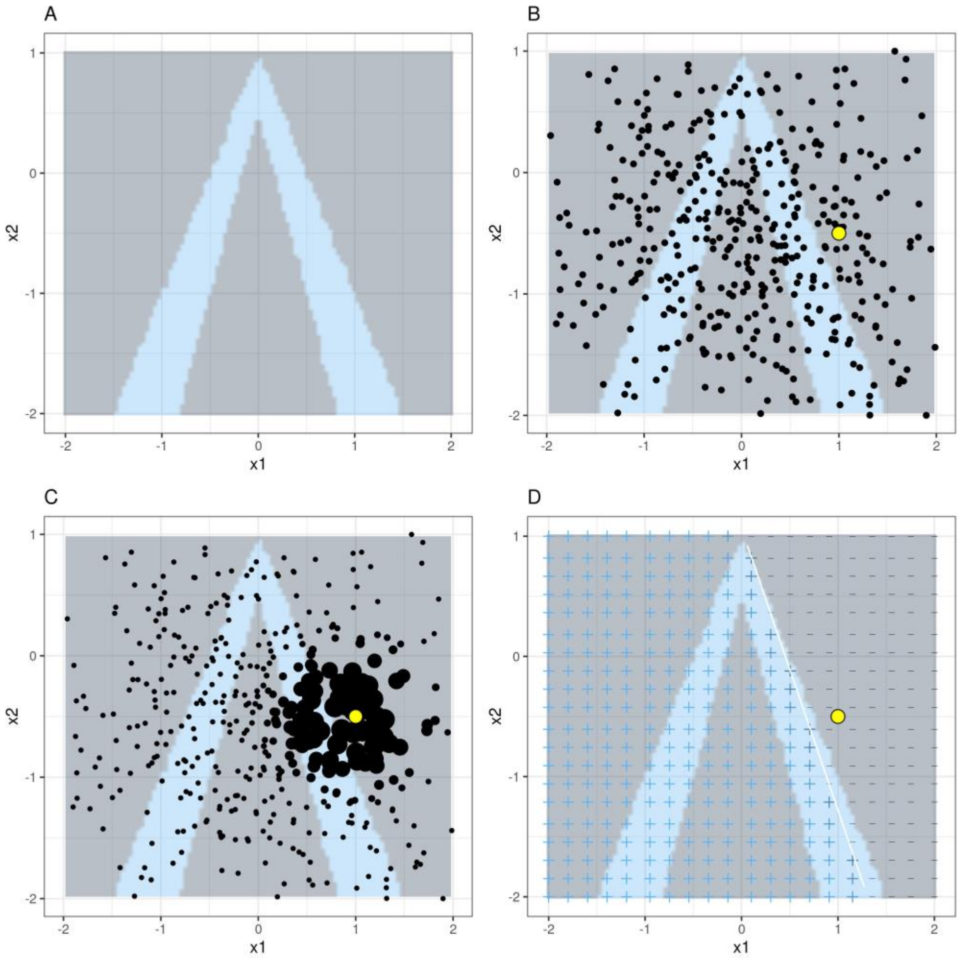
MSE for regression

Black-box model

$$\phi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Linear regression with Lasso

Exponential smoothing Kernel



Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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). Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

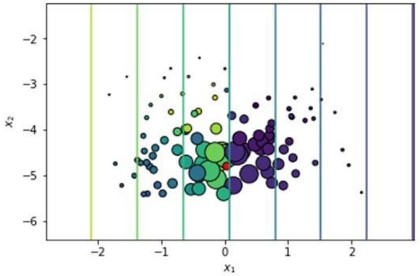
Why should I trust you?

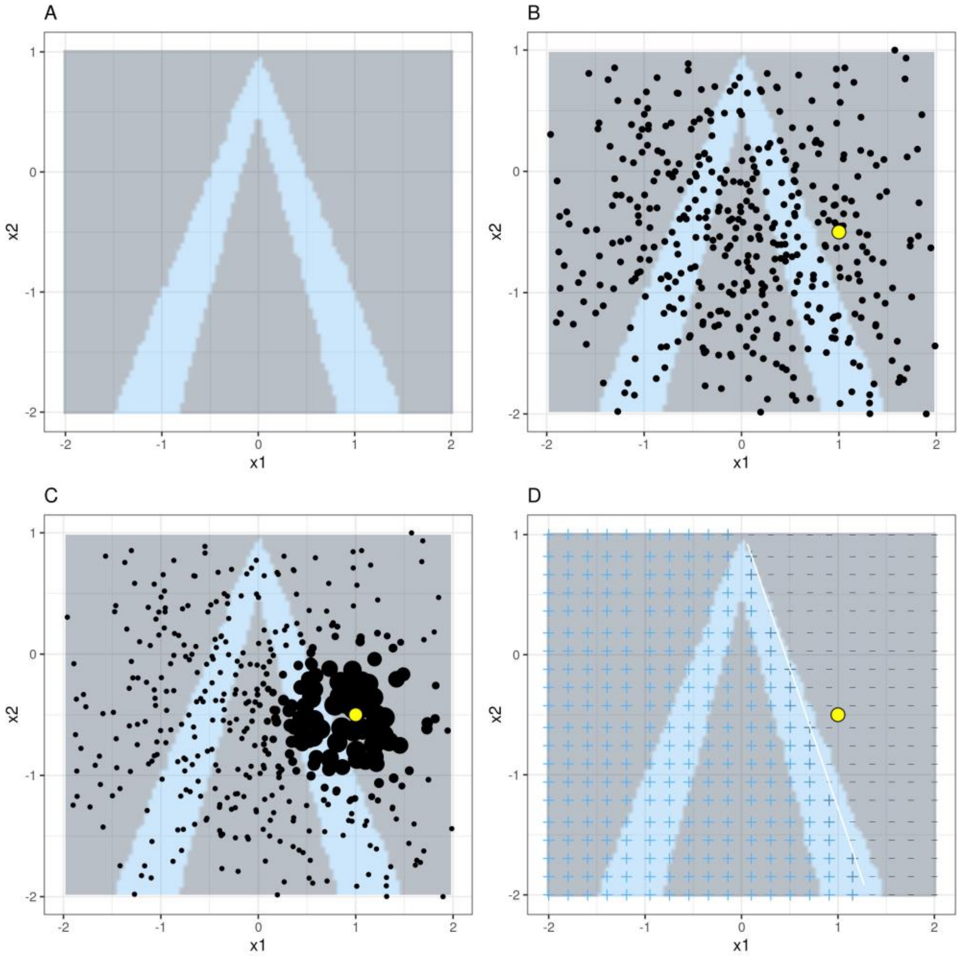
- Given an instance of interest, sample around the instance with probability *N(0,1)* to generate *z* new samples
- Weight samples using predefined kernel, in case of LIME it is exponential smoothing with default kernel width = 0.75

$$K(\boldsymbol{x}, z_i) = \sqrt{\frac{e^{-||\boldsymbol{x}-z_i||^2}}{\sigma^2}}$$

- For the generated, weighted dataset obtain probabilities from blackbox model
- Fit LASSO regression for that probabilities (!)

Fitting regression on probabilities gives us very nice interpretation + "actionability"

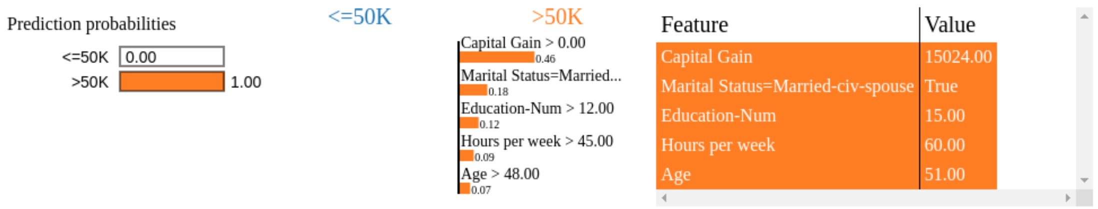


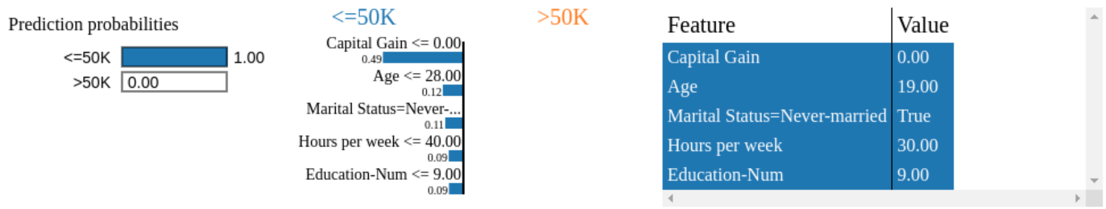


Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

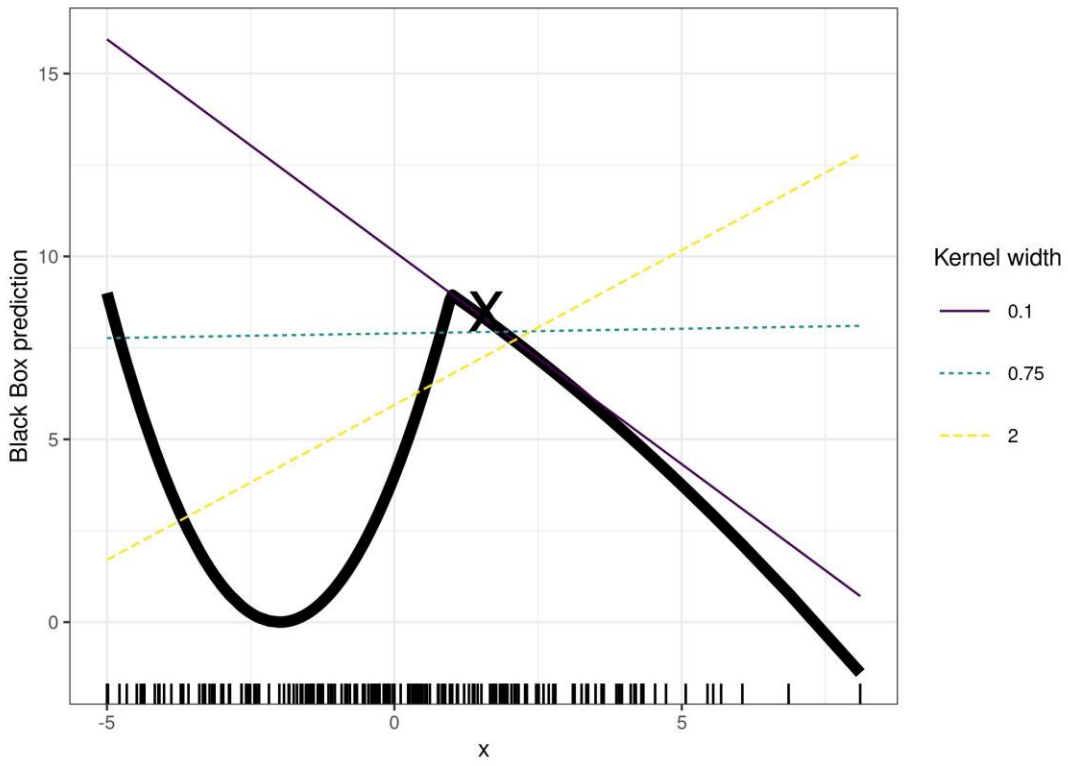
Interpretation and kernel width

- In certain cases, the feature importance might depend on the kernel size
- The values represent the importance of a feature according to Ridge Lasso trained on probabilities



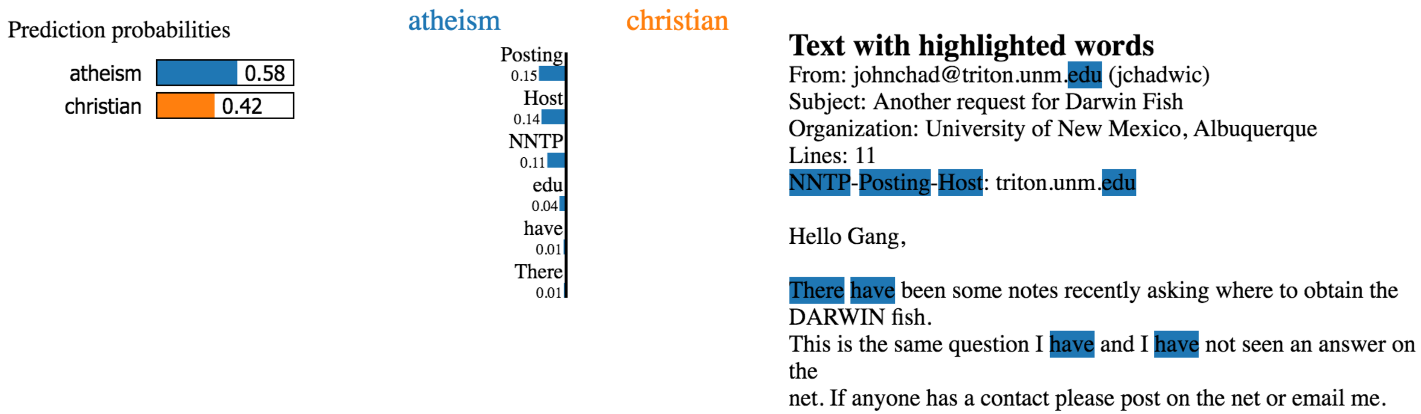


- Kernel size might impact the feature importance



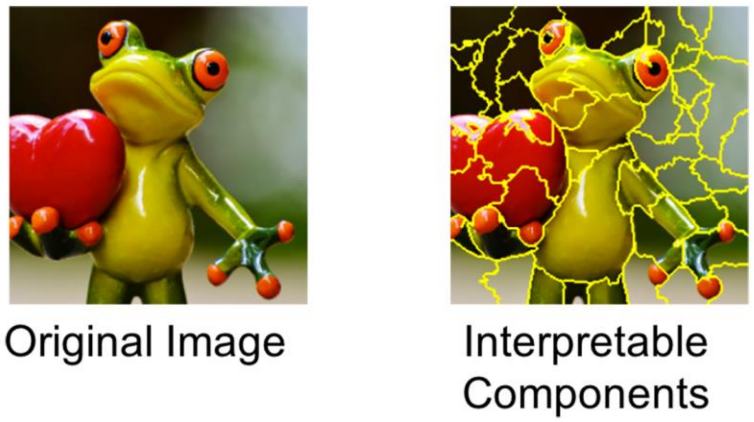
LIME for text

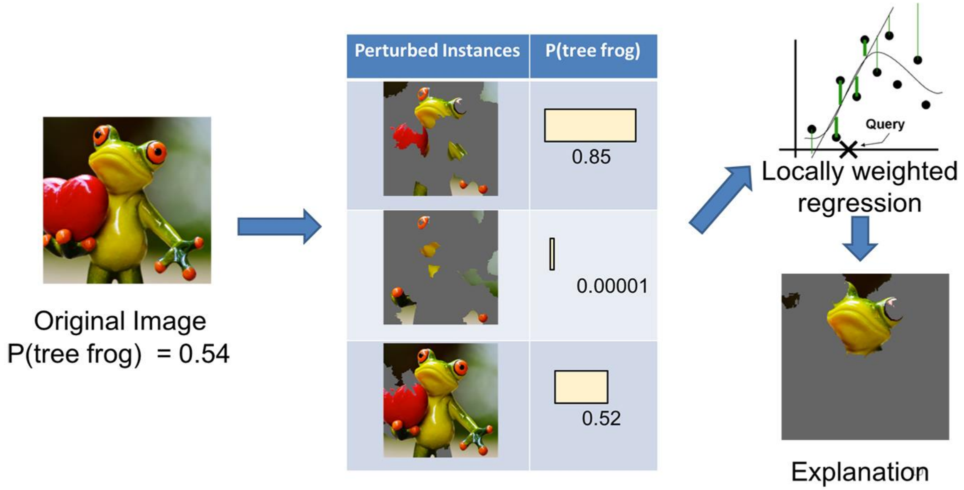
Posting	NNTP	Host	There	have	edu	;)	prob	weight
1	0	1	1	0	0	1	0.17	0.57
0	1	1	1	1	0	1	0.17	0.71
1	0	0	1	1	1	1	0.99	0.71
1	0	1	1	1	1	1	0.99	0.86
0	1	1	1	0	0	1	0.17	0.57



- We perturb text by removing words (i.e. using OHE notation and zero-ing out words by random)
- We predict class for each of the perturbed sentences
- The "weight" is calculated as 1 minus the proportion of words that were removed, for example if 1 out of 7 words was removed, the proximity is 1 - 1/7 = 0.86
- We train Ridge LASSO on this weighted instances and probabilites

LIME for images





- We create interpretable components by generating superpixels
- We generate pertubed data by replacing superpixels with average (or gray) color
- For each of the perturbed instances we calculate probability of being in particular class
- We weigh the instances according to the similarity to the original image
- We train Ridge LASSO on that dataset

Pros and cons

- Advantages
  - Simple to implement and relatively easy to interpret results
  - LIME is one of the few methods that works for tabular data, text and images.
  - The quality of explanations can be measured with a usage of fidelity
  - Many implementations
  - Relatively fast
- Disadvantages
  - The kernel width might be problematic
  - The multidimensional data suffers from dimensionality curse
  - It is possible to fool it by building classifier that recognizes perturbed and original data and behaves differently

Shapley Values

Lloyd Shapley

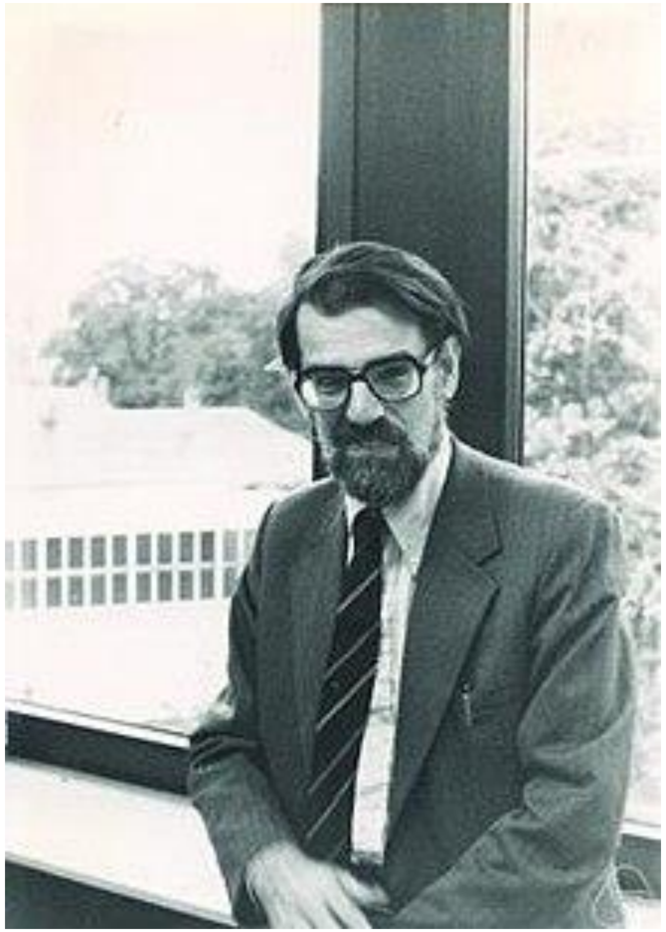
- Nobel-Prize winning economist
- In 1953 he publishes *"A value for n-person games*" where he introduced concept which became known as Shapley Values
- The question he tried to answer was: In a cooperative game, how each of the players contribute to the final win/loose?

the representation of Lemma 3 and obtain the formula:

$$\phi_i[v] = \sum_{\substack{R \subseteq N \\ R \ni i}} c_R(v)/r \qquad \text{(all } i \in N) . \tag{10}$$

Inserting (8) and simplifying the result gives us

$$\phi_i[v] = \sum_{\substack{S \subseteq N \\ S \ni i}} \frac{(s-1)!(n-s)!}{n!} v(S) - \sum_{\substack{S \subseteq N \\ S \not\ni i}} \frac{s!(n-s-1)!}{n!} v(S) \qquad \text{(all } i \in N) . \tag{11}$$



Intuition behind Shapley Values

- Imagine we have three students preparing a project they will earn points for
- Teacher said that they will be given points for each part of the project and they should split the given reward between themselves
- Students decided that equal split is not fair, because they share different competences and skills and contributed differently to the final grade

Student	Points earned	Comment
None	0	No students, no points
{Alice}	15	Alice knows ML
{Bob}	25	Bob knows ML but also XAI
{Charlie}	38	He has little knowledge on XAI and ML, but is a good programmer and fast learner so he can gain skills
{Alice, Bob}	25	They will earn the same amount as Bob only, but they can split tasks
{Alice, Charlie}	41	Alice can do her part, then Charlie will finish
{Bob, Charlie}	51	Bob and Charlie will do ML and XAi, but with Charlie's programming skills thay will do this better
{Alice, Bob, Charlie}	51	They will earn the same amount of points as Bob and Charlie , but have time to go for a beer

Marginal contribution (what is coallition's benefit from user participation)

Addition	To Coalition	Points before	Points after	Marginal contribution	Permutations
Alice	Empty coalition	0	15	15	Alice, Bob, Charlie
Alice	Empty coalition	0	15	15	Alice, Charlie, Bob
Alice	{Bob}	25	25	0	**Bob**, Alice, Charlie
Alice	{Charlie}	38	41	3	**Charlie**, Alice, Bob
Alice	{Bob, Charlie}	51	51	0	**Bob, Charlie,** Alice
Alice	{Charlie,Bob}	51	51	0	**Charlie, Bob,** Alice

$$\varphi_{\text{Alice}} = \frac{1}{6}(2*15 + 1*0 + 1*3 + 2*0) = 5.5$$

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{\overbrace{|S|}^{\text{Size of a coalition}}!\,(n-|S|-1)!}{\underbrace{n!}_{\text{Number of all possible coalitions}}} \underbrace{[v(S \cup \{i\}) - v(S)]}_{\text{Marginal contribution of } i}$$

Machine Learning Interpretation

- Player is a feature value
- Coalition is a set of features' values
- Payout function is a prediction minus average (expected)
- An empty coalition is *no features coallition* – an average prediction
- We simulate removing feature by sampling its value from *background data*

Player

	age	education	relationship	sex
0	39	Bachelors	Not-in-family	Male
1	50	Bachelors	Husband	Male
2	38	HS-grad	Not-in-family	Male
3	53	11th	Husband	Male
4	28	Bachelors	Wife	Female
...	...	...	...	...
48836	33	Bachelors	Own-child	Male
48837	39	Bachelors	Not-in-family	Female
48839	38	Bachelors	Husband	Male
48840	44	Bachelors	Own-child	Male
48841	35	Bachelors	Husband	Male

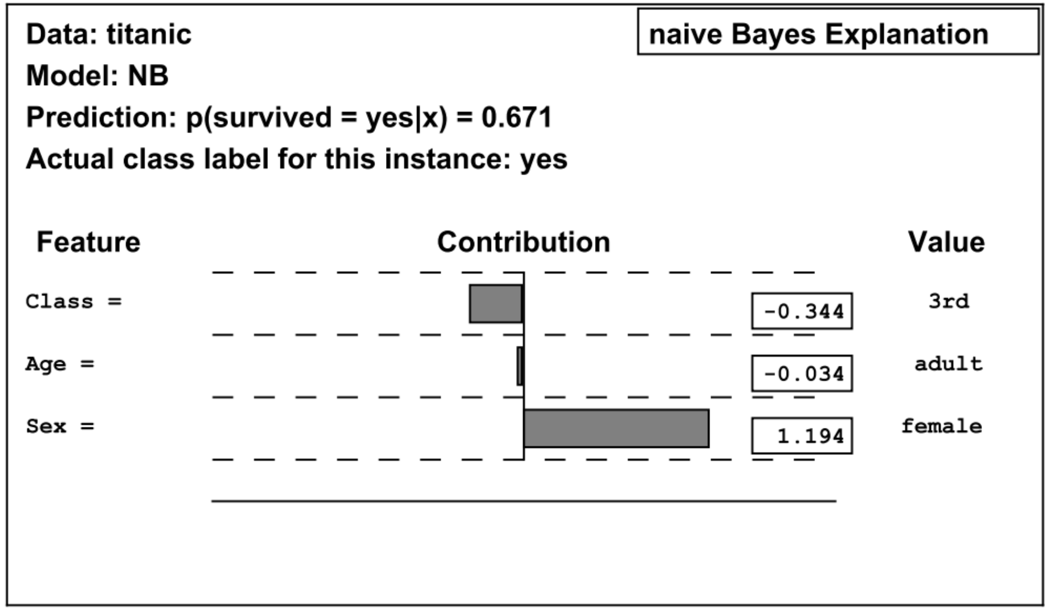
Payout is the output of the model (i.e. probability of being in one of the classes minus the average probabiltiy)

Coalition

Coalition with "missing" age=38

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!}\,[v(S \cup \{i\}) - v(S)]$$

Shapley values for Machine learning are not new

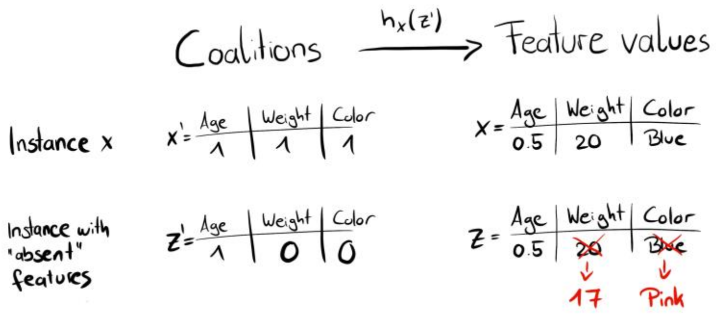


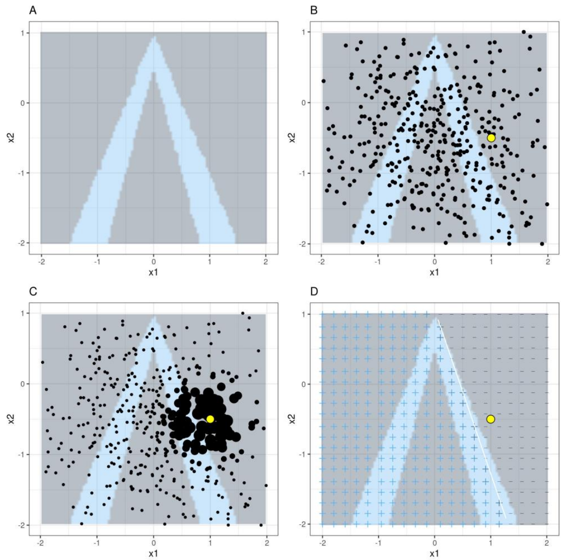
Erik Strumbelj and Igor Kononenko. 2010. An Efficient Explanation of Individual Classifications using Game Theory. J. Mach. Learn. Res. 11 (3/1/2010), 1–18.

- First proposed by Strumbelj and Kononenko in 2010
- They calculate SV by permutaiton sampling
- Later (2014) they improved their work by employing Monte Carlo sampling
- Their work did not gain popularity

Kernel SHAP

- Calculating exact Shapley values requires generating 2^p permutations of all features' values, where p is number of features values...which is a lot...
- Instead we can try to approximate the exact Shapley values with other methods





$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

$$L(\hat{f}, g, \pi_x) = \sum_{z' \in Z} [\hat{f}(h_x(z')) - g(z')]^2 \pi_x(z')$$

$$\pi_x(z') = \frac{(M-1)}{\binom{M}{|z'|}|z'|(M-|z'|)}$$

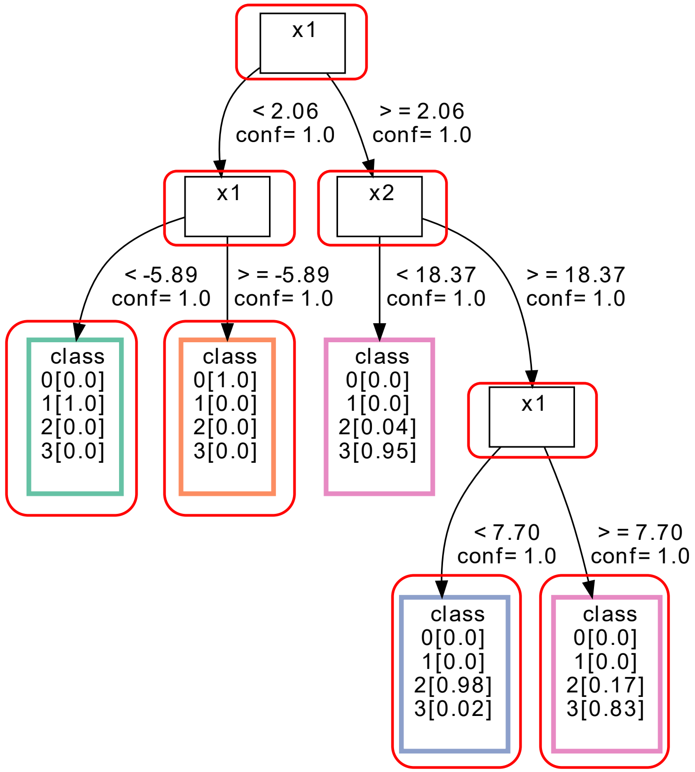
M -- maximum coalition size. The highest weight get large and small coalitions (we can learn more about single feature effect from 1-element coalition as well as from M-1 elelement coalition)

Tree-SHAP

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n-|S|-1)!}{n!}\left[v(S \cup \{i\}) - v(S)\right]$$

$$v(S \cup \{i\}) = E\left[f(x)|x_{S \cup \{i\}})\right] \qquad v(S) = E\left[f(x)|x_S)\right]$$

- We redefine Shapley values equation in terms of conditional expectation
- If S is empty we use weighted (by the num of samples) average prediction from all terminal notes
- If S contains all features, then use the node that the particular instance falls into
- If S contains some features we ignore unreachable nodes. Unreachable means reaching it contradicts value sin x_S



x1	x2
3.0	19

Axioms of Shapley Values

- Efficiency – SHAP values add up to the centered prediction

$$\sum_{j=1}^{p} \phi_j = \hat{f}(x) - E_X(\hat{f}(X))$$

- Symmetry – if two feature values contribute equally, their contribution should be equal. Order is irrelevant

$$v(S \cup \{j\}) = v(S \cup \{k\}) \Leftrightarrow \phi_j = \phi_k$$

- Dummy – Features not affecting the predicion receive SHAP valuess equal 0

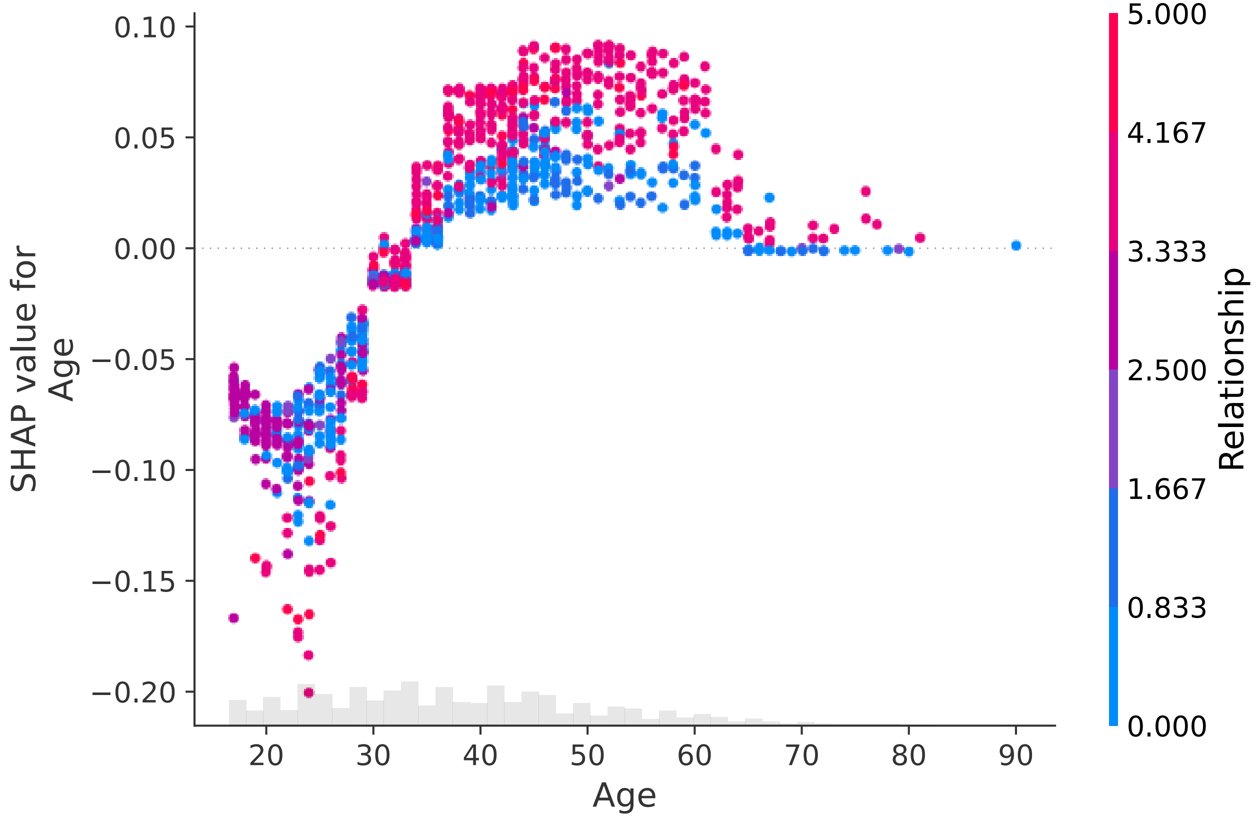
$$val(S \cup \{j\}) = val(S) \Leftrightarrow \phi_j = 0$$

- Additivity – Additive predictions correspond to additive SHAP values

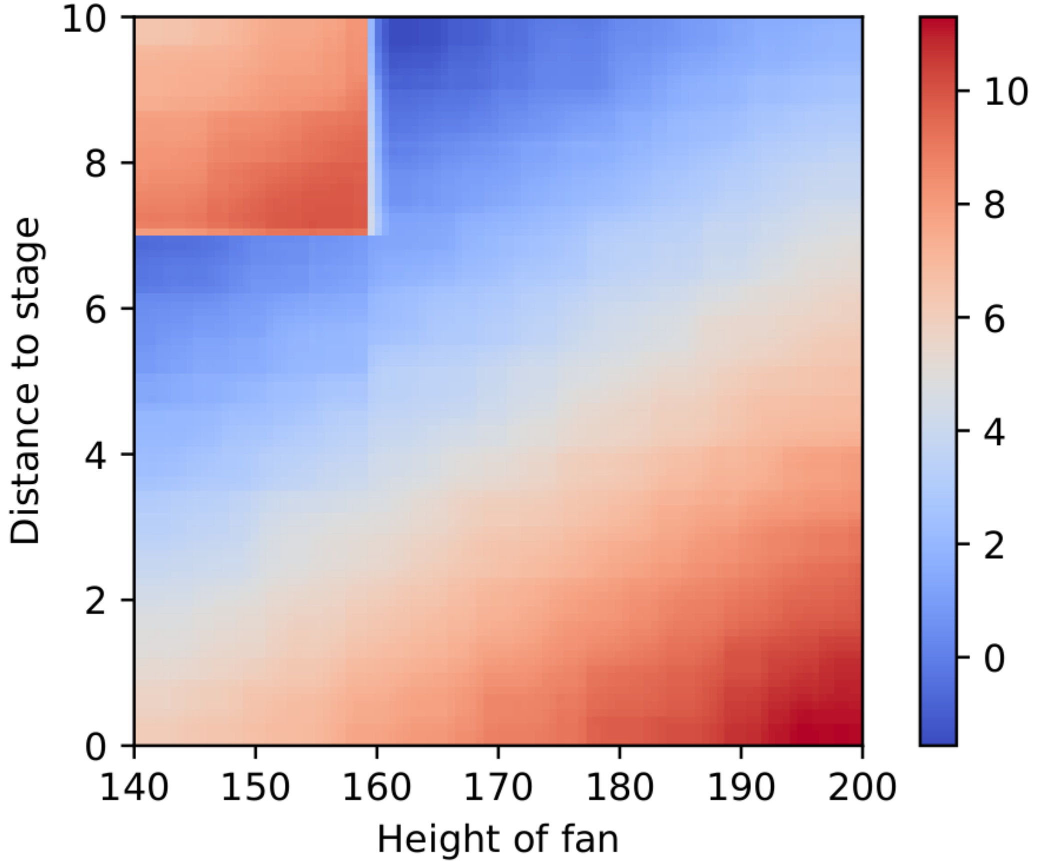
$$\phi_j^{AB} = \phi_j^{A} + \phi_j^{B}$$

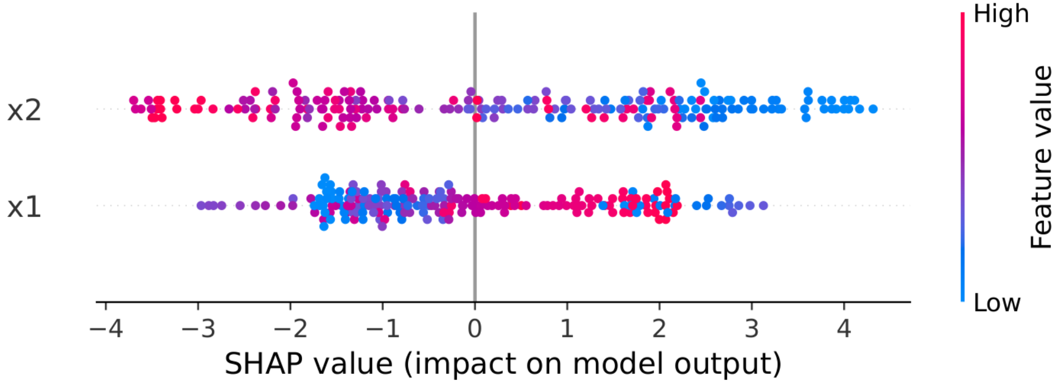
Interpreting SHAP- Scatter plots

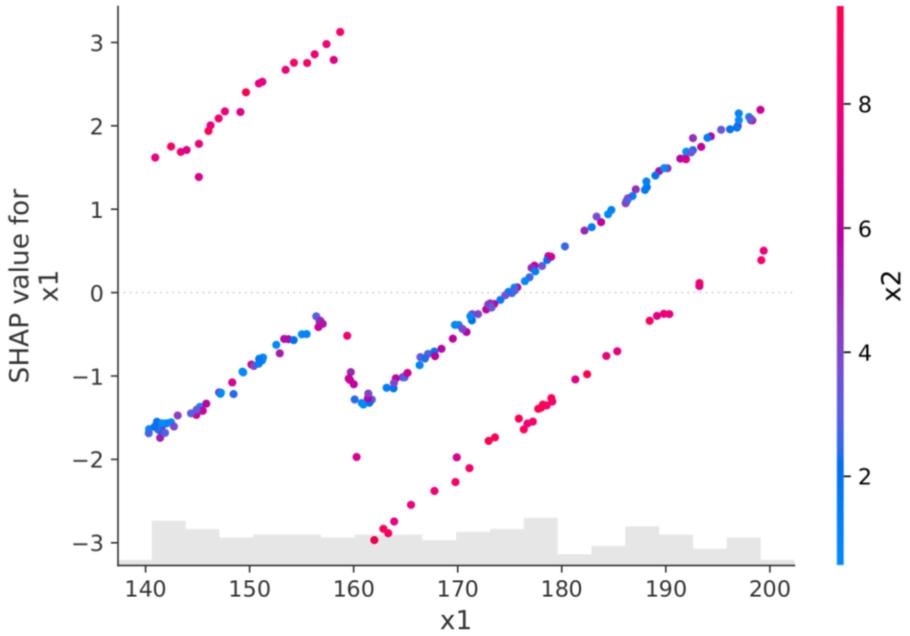
- Each point represent Shapley value for a given feature's value
- Horizontal patterns represent interactions – for instance there is an interaction between Relationship and Age
- After 30 the instances "in relationship" are more likely to earn more money
- While PDP and ALE plots show average effects, SHAP dependence also shows the variance on the y-axis.



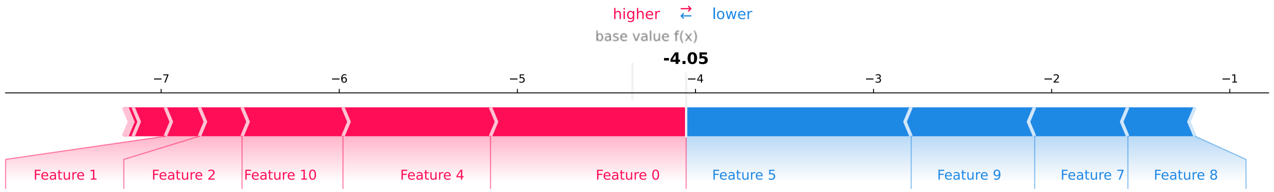
More on SHAP interactions

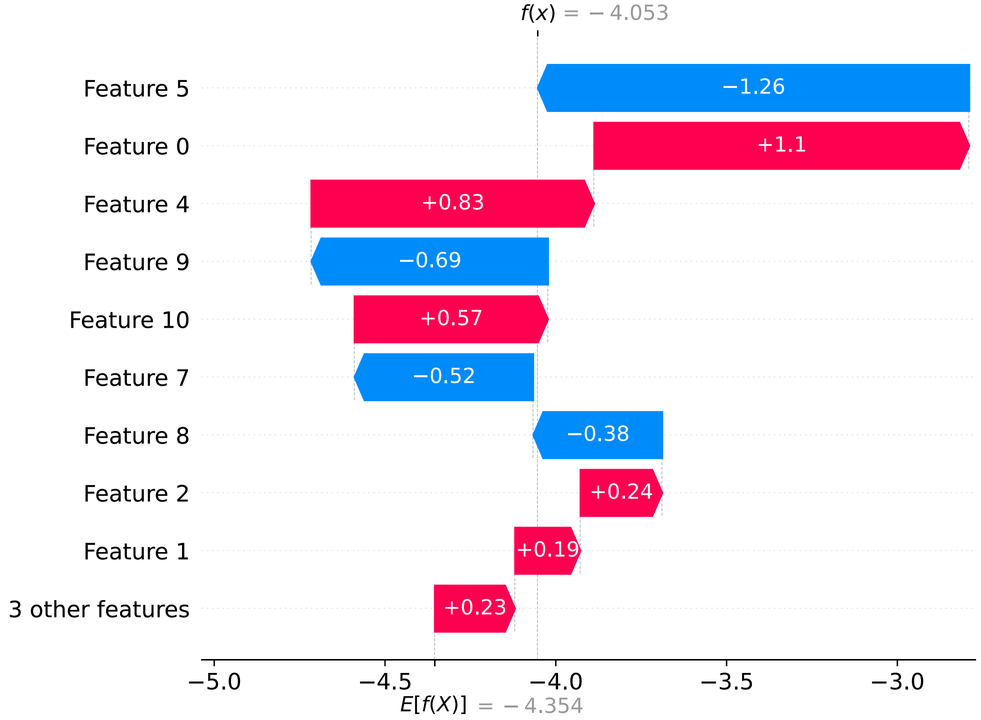


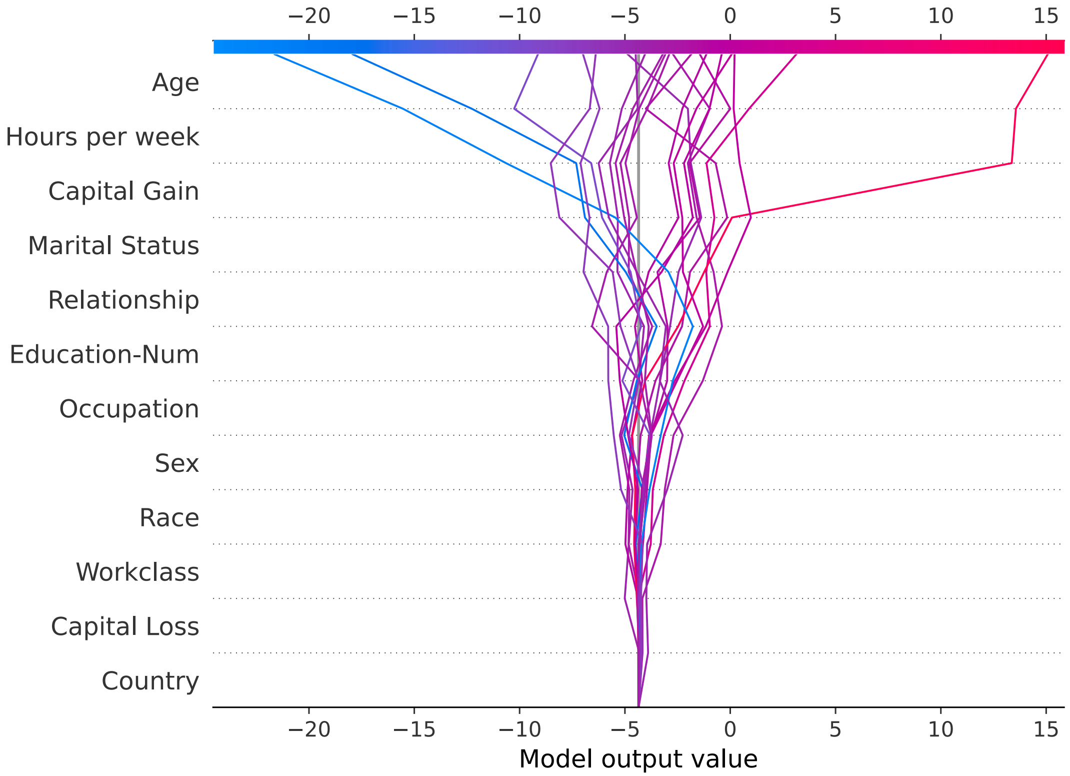




Interpreting SHAP – Force- and Decision plots





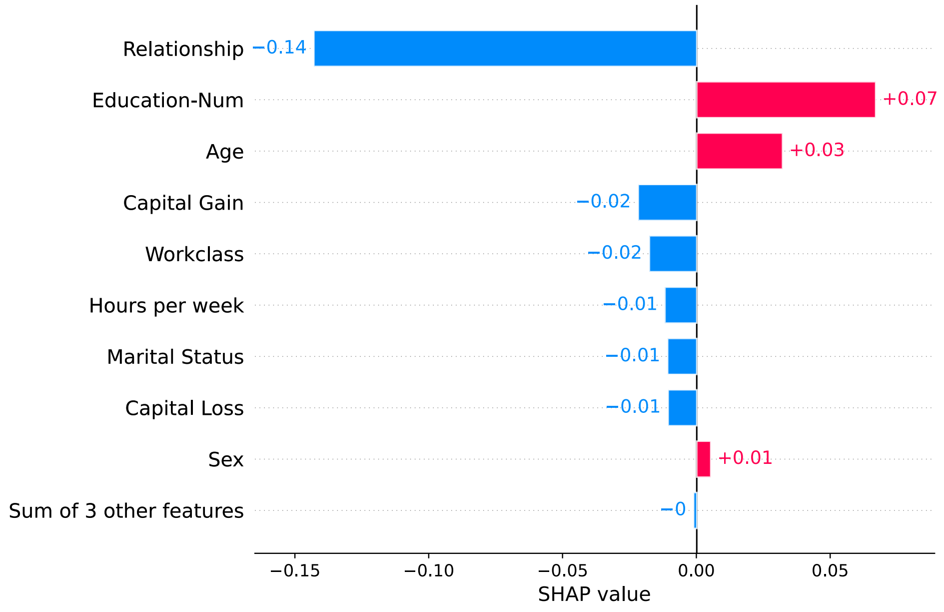


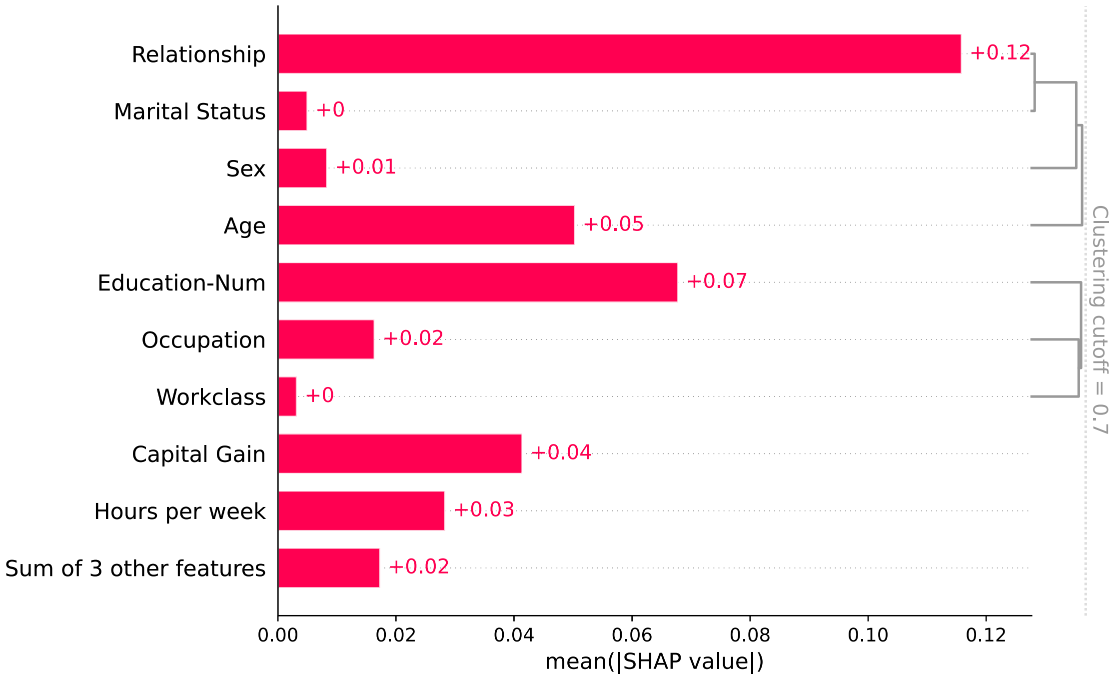
For single preinstance

For multiple preinstances

Interpreting SHAP – Barplots and Beeswarm

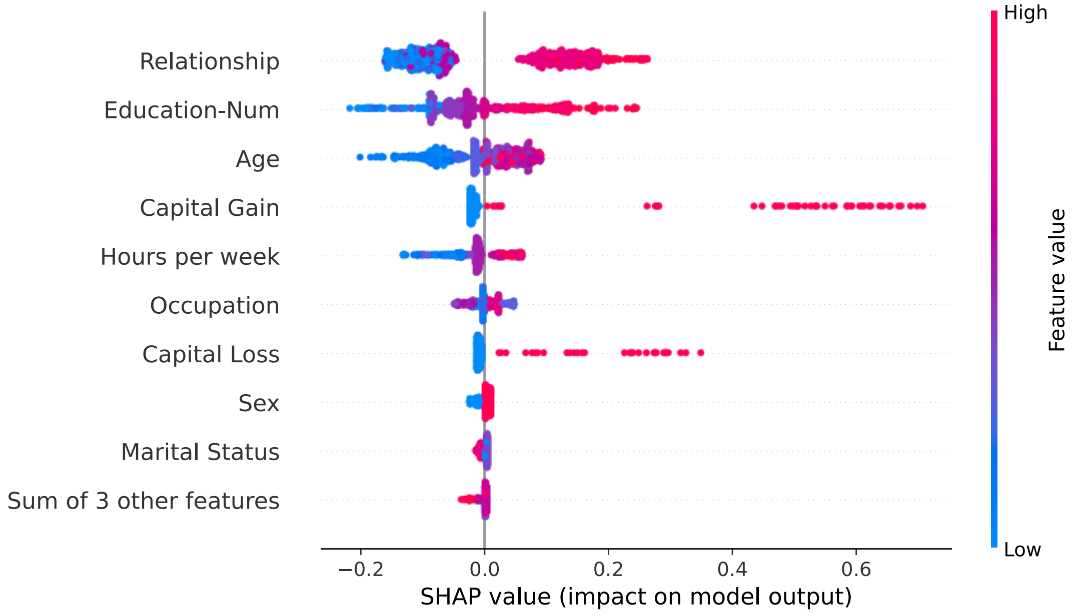
- Redundant features can be clustered
- This is much better indicator of redundancy than correlation

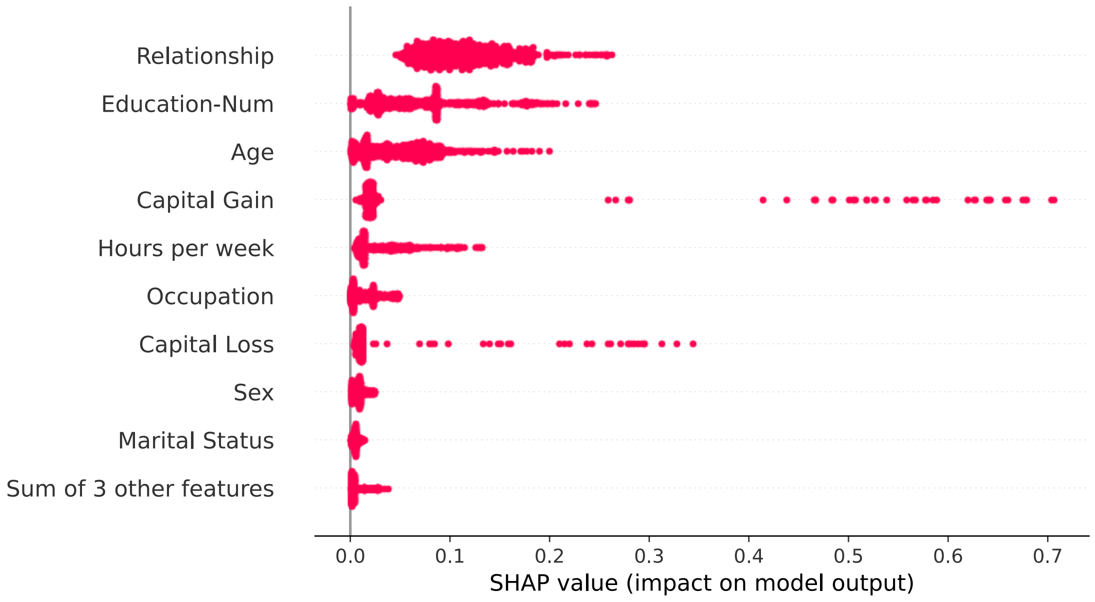




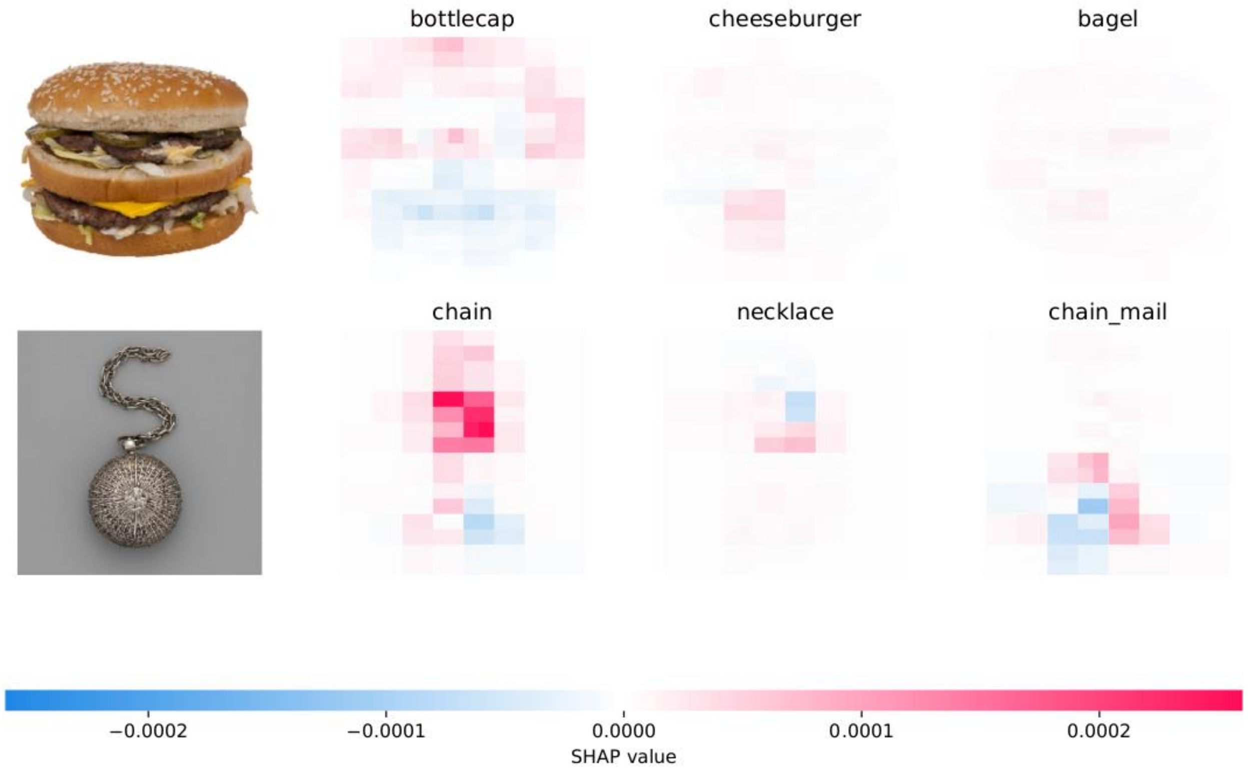
Interpreting SHAP – Barplots and Beeswarm

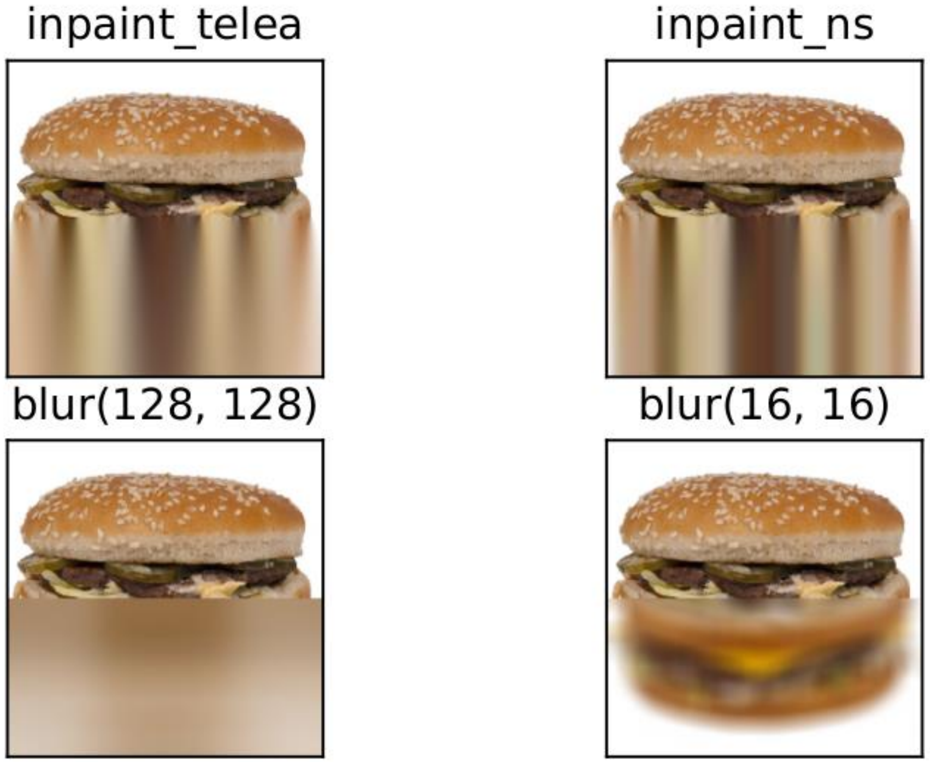
- It is a combination of waterfall plot and scatter plot



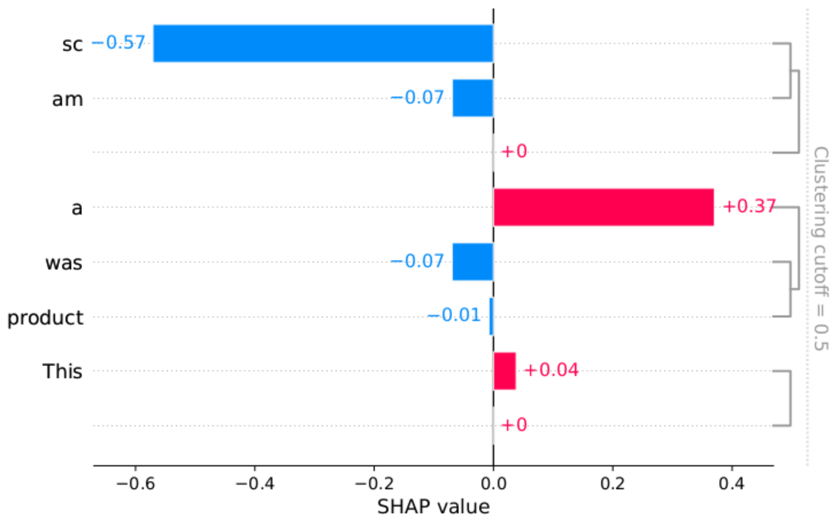


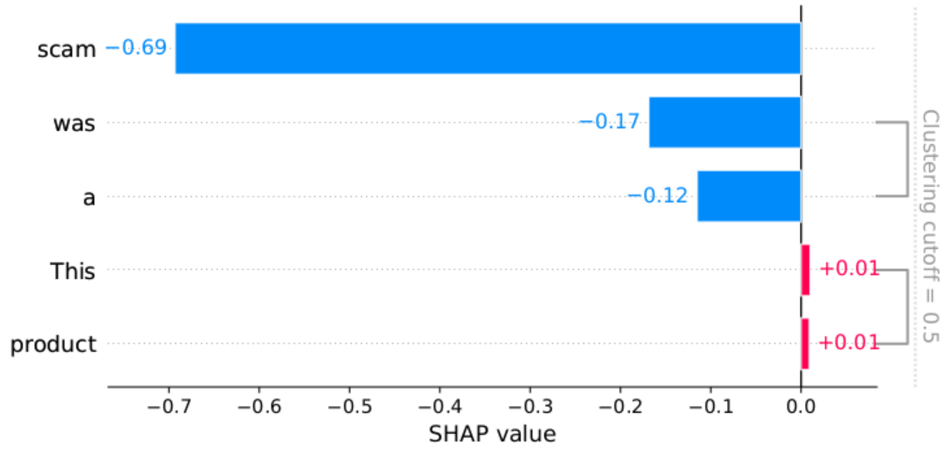
Shap for images

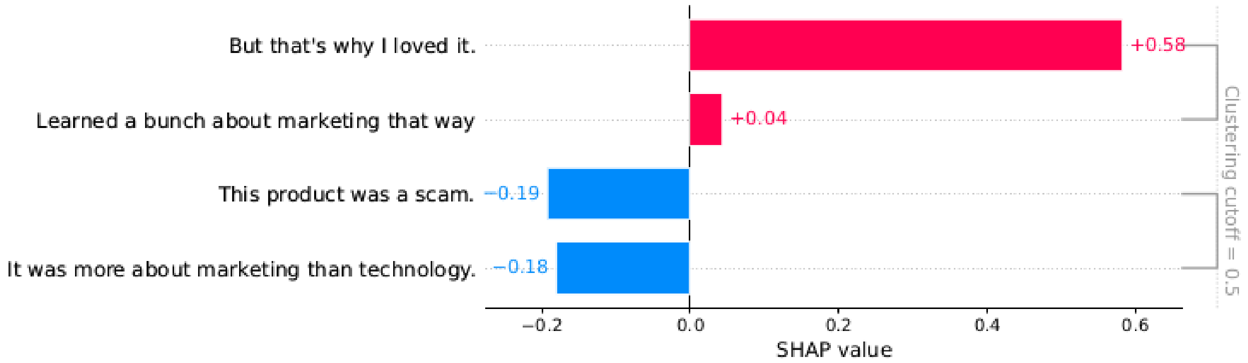




Shap for Text







Pros and Cons

- Advantages
  - Solid mathematical theory
  - Model agnostic (to some extent)
  - It provides local and global explanations
  - Fast implementation for Trees and Deep NN
  - Nice visualizations (including text)
- Disadvantages
  - Background data is an elephant in the room
  - They are not actionable nor they are surrogate models!
  - KernelShap ignores feature dependence (feature generation of unlikely instances)
  - Implementation is... evolving

Thank you for your attention!

JAGIELLONIAN UNIVERSITY
IN KRAKÓW

https://geist.re