Interpretacja predykcji metod ML i DM
From Shapley values to SHAP in ML

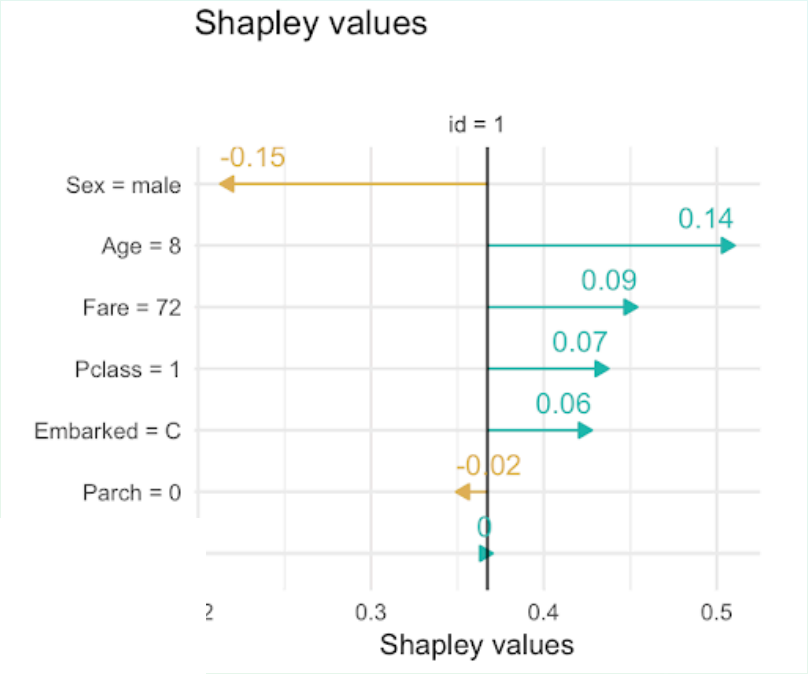
Jerzy Stefanowski

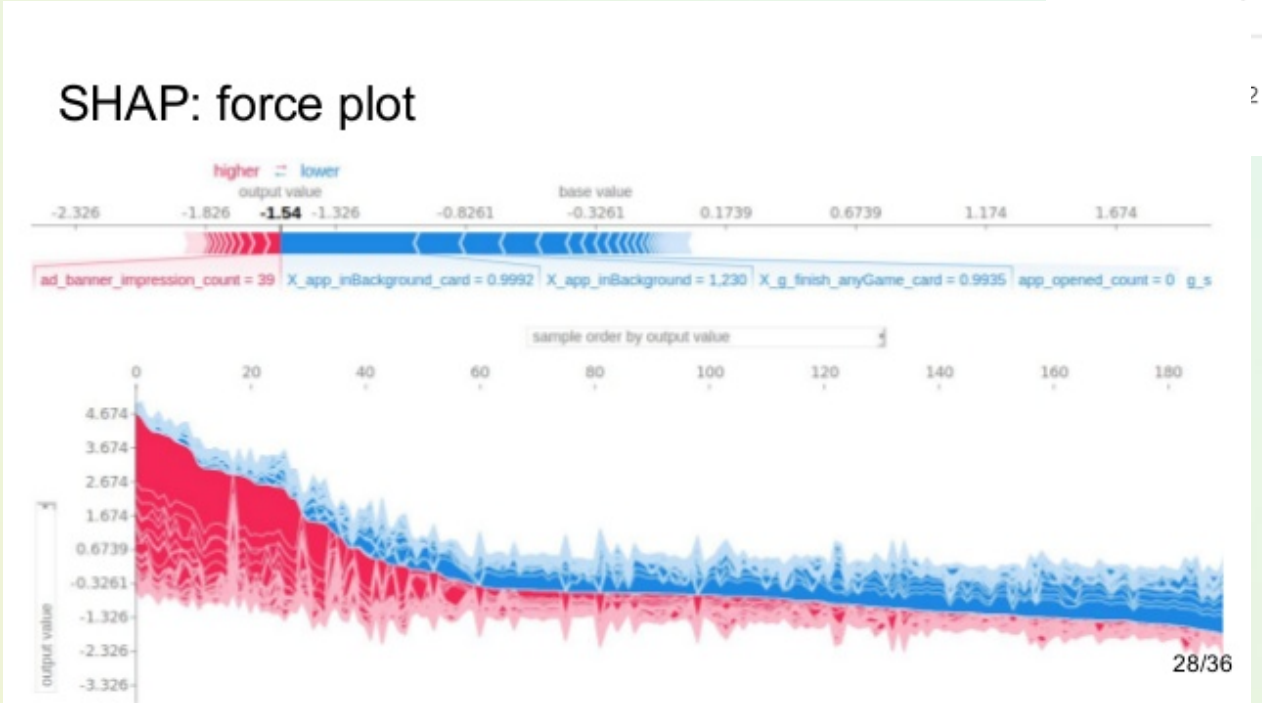
Institute of Computing Sciences,
Poznań University of Technology
Poland

Wersja 2020

Shapley values for prediction models

In nutshell





Plan

- Interpretation of ML predictions
- Set functions → Shapley and Banzhaf indices, Möbius representation
- More on Shapley values
- SHAP approach
- An illustrative example + Lundberg software
- SHAPsummary
- Studying most important conditions, their subsets and their interaction in rules
- Rule interestingness measures
  - Selection of complete rules
  - Medical case studies

Z uwagi na różne (...) – slajdy w języku angielskim

Interpreting Model Predictions

- In machine learning trade off between performance and complexity
  - Complex models – although accurate are often not explainable black boxes
- Needs to explain models / their predictions and relation to data
  - Consider financial or administrative decisions:
    - It is legally required to provide an explanation for why a prediction was made
  - Medicine – needs to interpret a case
  - Studying cases when model fails, incorrect decisions

- Global vs. local explanations

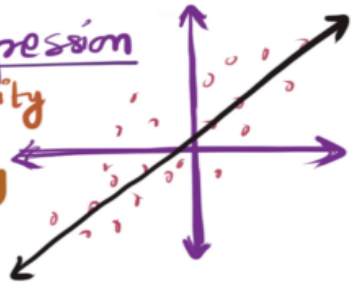
Model Explainability (degree of explainability)

1) linearity
2) monotonicity
} make predictions easier & reliable

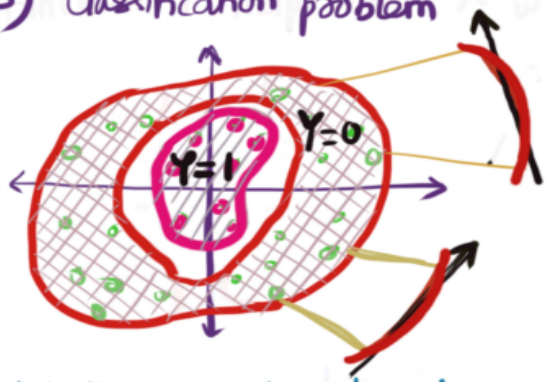
$f: X \to Y$ (mapping from x to y by f)

$f(x) = x\beta$

1) linear regression
global fidelity
= local fidelity



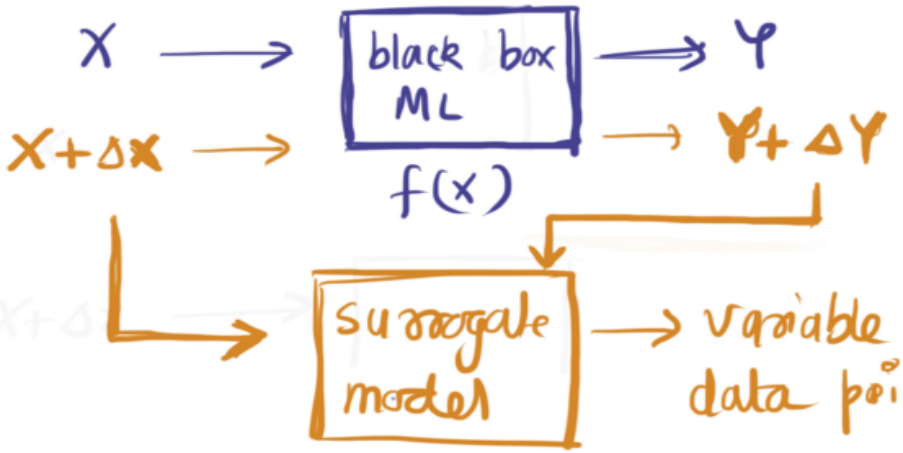
3) classification problem



globally : complex boundary
locally : linear boundaries

2) piere wise linear
local fidelity
≠ global fidelity

global function



Δx : small change in x
Δy : small change in y

Rys. za MC.AI

Lloyd Stowell Shapley

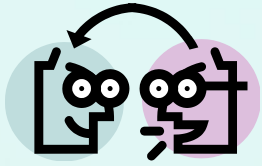
- An American mathematician and Nobel Prize-winning economist (1923-2016).
- He contributed to the fields of mathematical economics and especially game theory
- He came up with this solution concept for a cooperative game in 1953.
- Shapley wants to calculate the contribution of each player in a coalition game.
- Assume there are N players and S is a subset of the N players. Let v(S) be the total value of the S players. When player {i} join the S players, Player i's marginal contribution is $v(S \cup \{i\}) - v(S)$.

Origins of the game co-operation

Shapley, Banzhaf indices / values and Möbius representation

- Previously considered in cooperative games, voting systems, party coalitions and multiple criteria decision aid:
- $X = \{1,2,\ldots,n\}$ a set of elements / players / agents
  A set function $\mu : P(X) \rightarrow [0,1]$
  - A weighted average contribution of agent / element i in all coalitions
  - Conjoint importance of elements $A \subseteq X$
  - Measuring interaction of elements
- Here we mainly focus on Shapley value and its usefulness to explain the impact of particular attributes on performance of the final classifier prediction
  - The Shapley value: It is the average of the marginal contributions across all permutations .

Basics of set functions



- $X = \{1,2,\ldots,n\}$ a set of elements (e.g. players in the game); $P(X)$ – the power set of X = the set of all possible subsets of X
  A set function $\mu : P(X) \rightarrow [0,1]$
- Function μ - a measure satisfying:
  - $\mu(\varnothing) = 0$ and $\mu(X) = 1$
  - $A \subseteq B$ implies $\mu(A) \leq \mu(B)$
  - „1” could be treated as max value
- Interpretation of function μ in a particular problem
  - The profit obtained by players / agents
  - The importance of criteria in MCDA
- Transformations of function μ
  - Shapley and Banzhaf values refer to single elements $i \in X$, but also their interactions, subsets of elements $A \subseteq X$
  - Möbius representation m: $P(X) \rightarrow R$

Möbius representation

Möbius representation m: $P(X) \rightarrow R$

For all $A \subseteq X$:

$$\sum_{B \subseteq A} m(B) = \mu(A)$$

$$m(A) = \sum_{B \subseteq A} \mu(B)(-1)^{|A|-|B|}$$

- m(A) – the contribution given by the conjoint presence of all elements from A to the function μ

All set functions will be illustrated by a toy examples

Illustrative example – Möbius representation

Consider players 1,2,3, where the profits of their actions are
$\mu(\{1\})=5$, $\mu(\{2\})=7$, $\mu(\{3\})=4$ and $\mu(\{1,2\})=15$ (by def. $\mu(\varnothing)=0$)

Calculate $m(\{1\})=5$, $m(\{2\})=7$ and $m(\{1,2\})=15-5-7=3$

Note - $\mu(\{1,2\})=15$ is greater than $\mu(\{1\} + \mu(\{2\})= 5+7$

The contribution coming out from the conjoint presence of {1} and {2} in this coalition and it is equal to $m(\{1,2\})=3$

Illustrative example – Shapley value

- Shapley value – average contribution / importance of element
- Consider X={1,2,3} where the profits of the agent actions are $\mu(\{1\})=5$, $\mu(\{2\})=7$, $\mu(\{3\})=4$, $\mu(\{1,2\})=15$, $\mu(\{1,3\})=12$, $\mu(\{2,3\})=14$ and $\mu(\{1,2,3\})=30$
- How to fairly split the total profit of 30 units among the agents taking into account their contribution?
- Attribute to the conjoint presence of agents $A \subseteq X$, so split equally $m(A)$ among agents

$$m(A)/|A|$$

- Each agent should receive the value (Shapley)

$$\phi_i(\mu) = \sum_{A \subseteq X : i \in A} \frac{m(A)}{|A|}$$

Illustrative example – Shapley value

- X={1,2,3} and profits are $\mu(\{1\})=5$, $\mu(\{2\})=7$, $\mu(\{3\})=4$, $\mu(\{1,2\})=15$, $\mu(\{1,3\})=12$, $\mu(\{2,3\})=14$ and $\mu(\{1,2,3\})=30$

Möbius representations

- $m(\{1\})=5$, $m(\{2\})=7$, $m(\{3\})=4$, $m(\{1,2\})=\mu(\{1,2\})-\mu(\{1\})-\mu(\{2\})=15-5-7=3$, $m(\{1,3\})=3$, $\mu(\{2,3\})=3$ and $m(\{1,2,3\})=\mu(\{1,2,3\})-\mu(\{1,2\})-\mu(\{1,3\})-\mu(\{2,3\})+\mu(\{1\})+\mu(\{2\})+\mu(\{3\})=30-15-12-14+5+7+4=5$

Shapley values for each agent

- $\phi_1(\mu)=m(\{1\})/1+m(\{1,2\})/2+m(\{1,3\})/2+m(\{1,2,3\})/3=5+3/2+3/2+5/3=9.67$
- $\phi_2(\mu)=m(\{2\})/1+m(\{1,2\})/2+m(\{2,3\})/2+m(\{1,2,3\})/3=7+3/2+3/2+5/3=11.67$
- $\phi_3(\mu)=m(\{3\})/1+m(\{1,3\})/2+m(\{2,3\})/2+m(\{1,2,3\})/3=5+3/2+3/2+5/3=9.67$

$$\phi_i(\mu)=\sum_{A\subseteq X: i\in A}\frac{m(A)}{|A|}$$

Other formulations

Shapley value:

$$\Phi_i(\mu) = \sum_{A \subseteq X - \{i\}} \frac{(|X - A| - 1)!|A|!}{|X|!} \cdot [\mu(A \cup \{i\}) - \mu(A)]$$

Banzhaf value:

$$\Phi_{Bi}(\mu) = \frac{1}{2^{|X|-2}} \sum_{A \subseteq X - \{i\}} [\mu(A \cup \{i\}) - \mu(A)]$$

Both interpreted as an averaged contribution of element *i* to all coalitions *A*

Interaction indices (i,j) → Morofushi and Soneda; Roubens

$$I_{MS}(i,j) = \sum_{A \subseteq X - \{i,j\}} \frac{(|X - A| - 2)!|A|!}{(|X| - 1)!} \cdot [\mu(A \cup \{i,j\}) - \mu(A \cup \{i\}) - \mu(A \cup \{j\}) + \mu(A)]$$

$$I_R(i,j) = \frac{1}{2^{n-2}} \sum_{A \subseteq X - \{i,j\}} [\mu(A \cup \{i,j\}) - \mu(A \cup \{i\}) - \mu(A \cup \{j\}) + \mu(A)]$$

Other illustrative examples

Three friends participating in cost of a dinner

$$v(c) = \begin{cases} 80, & \text{if } c = \{A\} \\ 56, & \text{if } c = \{B\} \\ 70, & \text{if } c = \{C\} \\ 80, & \text{if } c = \{A, B\} \\ 85, & \text{if } c = \{A, C\} \\ 72, & \text{if } c = \{B, C\} \\ 90, & \text{if } c = \{A, B, C\} \end{cases}$$

The contribution of A => 51.17

See the calculations of all coallitions in KDDBlog

https://www.kdnuggets.com/2019/12/interpretability-part-3-lime-shap.html

Some extra math axioms

Shapley establishes the following four Axioms in order to achieve a fair contribution:

Axiom 1: Efficiency. The sum of the Shapley values of all agents equals the value of the total coalition

Axiom 2: Symmetry. All players have a fair chance to join the game. That's why it necessary to list all the permutations of the players.

Axiom 3: Dummy. If player i contributes nothing to any coalition S, then the contribution of Player i is zero.

Axiom 4: Additivity. For any pair of games v, w: $\varphi(v+w)=\varphi(v)+\varphi(w)$, where $(v+w)(S)=v(S)+w(S)$ for all S. This property enables us to do the simple arithmetic summation.

Shapley Values

- Nice interpretation and provides more information than basic variable importance / see earlier lectures

However,

- Their computation – needs all permutations.
- Time consuming + calculating all possible coalitions and their outcomes quickly become infeasible as the attributes increase

In 2013, E. Štrumbelj I.Konennko proposed an approximation using Monte-Carlo sampling / another approach than SHAP

Needs for the simplified version or another version for their approximating

From Shapley values to SHAP

Shapley regression value

$$\Phi_i = \sum_{A \subseteq F - \{i\}} \frac{(|F - A| - 1)!|A|!}{|F|!} \cdot [f(A \cup \{i\}) - f(A)]$$

are feature importance for linear models in the presence of multi-collinearity /attributes are correlated/

Predictions of the model -> a model is trained with features A (without) and another model with feature {i}

A- all possible different subsets of features from F

Weights – for all possible subsets sum to exactly one

Moreover – a value for a model predicting an instance *x*

$$\hat{\phi}_j = \frac{1}{M} \sum_{m=1}^{M} \left(\hat{f}(x_{+j}^m) - \hat{f}(x_{-j}^m) \right)$$

Using Shapley value in SHAP

- SHAP – **SH**apley **A**dditive ex**P**lanations is a new approach not an extension of the Shapley value
- Proposed by Lundberg and Lee (NIPS 2016) as a unified approach to explaining the output of machine learning predictors / classifiers / regressors
- Currently popular due to access in software,
- See e.g. authors' repository

  https://github.com/slundberg/shap

Not exactly the previous Shapley values calculation but another way of approximating them and putting in the specialized visualization framework.

SHAP benefits

- **The global interpretability** - SHAP values can show how much each attribute contributes, either positively or negatively to the target variable
  / smth. like variable importance but it is able to show more information
- **The local interpretability** - each instance (its attributes) is described by its local set of SHAP values and can help in studying why this case leads to the given prediction and what is the contribution of its attribute values.
- SHAP values can be calculated for many prediction models

From Shapley values to SHAP

- Shapley sampling values /approximate approach/ -> approximate the effect of removing an attribute from the model by integrating over instances from the training data set.
- SHAP – promotes another model agnostic approximation i.e. Kernel SHAP algorithm
- Moreover special formulations for tree models SHAPTree, deep models DeepSHAP, linear regression, ...-> works in polynomical time
- See notebooks in

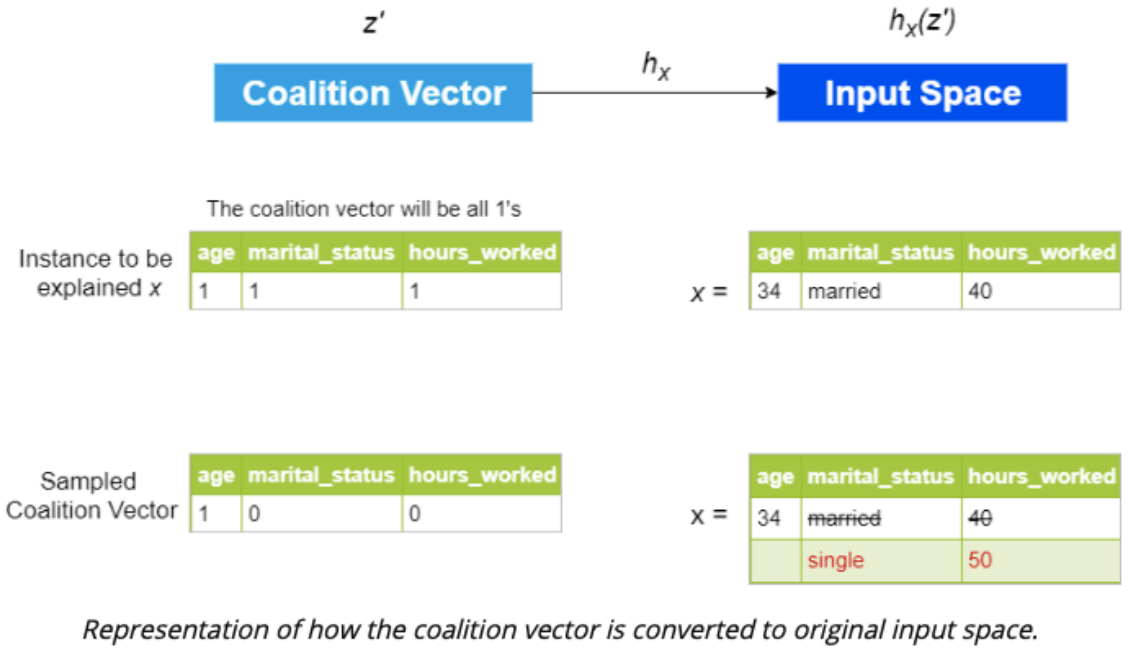
https://github.com/slundberg/shap/blob/master/README.md

SHAP specialized calculation techniques

- Kernel (partly extend linear model inspired like LIME) – do not require the evaluation of all 2^M sets
- Instead an additive attribute model – a weighted linear regression with simplified inputs z and estimation Shapley values by making calculation over a sample of instance predictions

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i,$$

where $z' \in \{0,1\}^M$, *M is the number of simplified input features, and* $\phi_i \in \mathbb{R}$.



Representation of how the coalition vector is converted to original input space.

More mathematics of SHAP

Please refer to a nice book Interpretable Machine Learning
A Guide for Making Black Box Models Explainable.
by Christoph Molnar

Chapter 5.10

Access online

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

SHAP motivating example

Consider a model predicting flat prices, e.g. in one Polish town

- A model uses three attributes: flat size, year of building and a localization (region/district of the city)
- For an offer 40m2, building from 1920 and inside Old City this model predicts price 450.000
- Knowing that average prices in historical data for this city are approx. 400.000, questions are what are the reasons that this model predicts more, what are the impact of particular contributing predictors?
- SHAP values may indicate that it is positively increased due the district (approx. increase 70.000); negative decrease relation age (approx. – 20.000), the m2 size does not present any impact.

More on illustration – Boston housing data

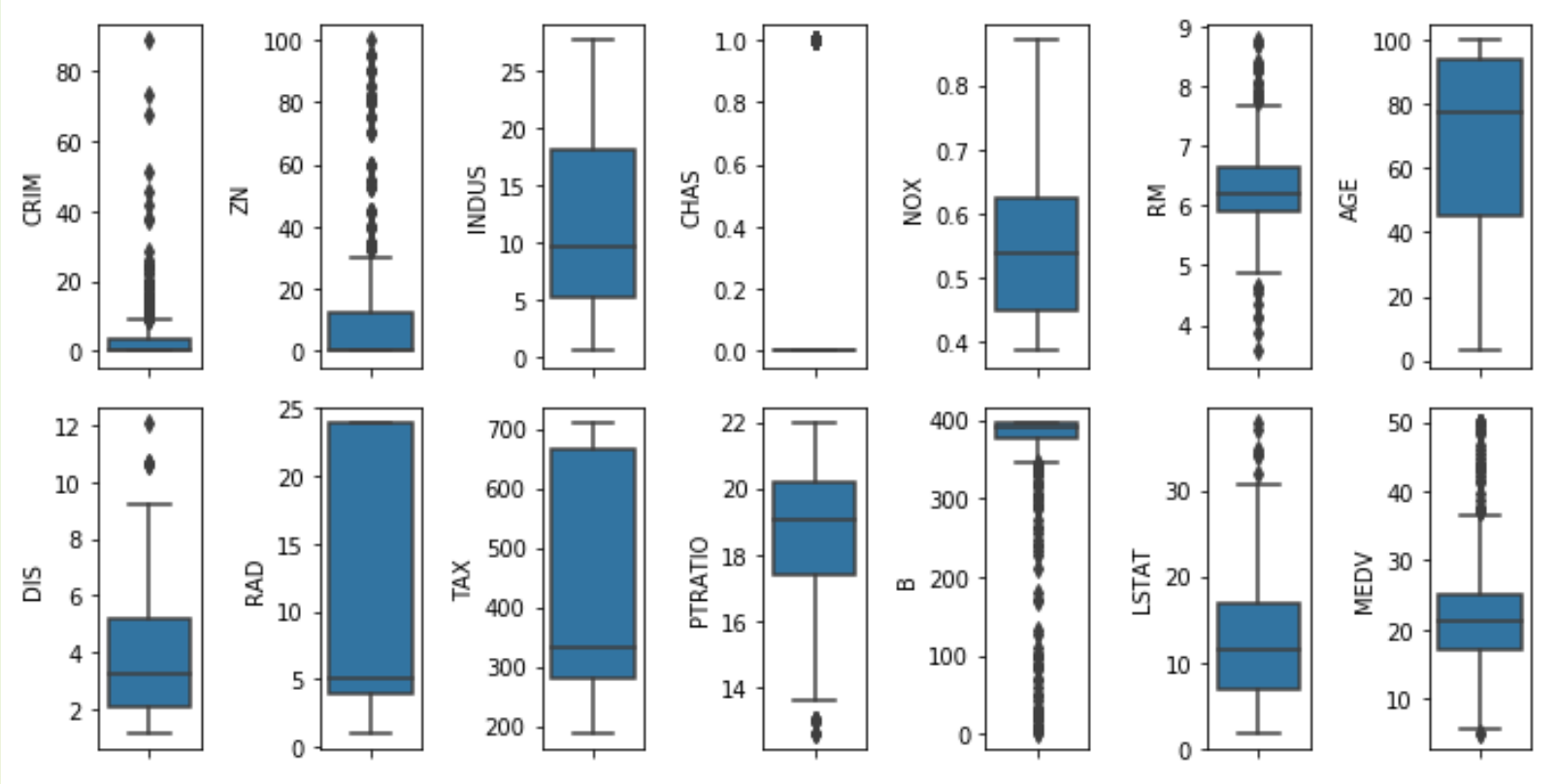
Public dataset 506 cases with 13 attributes and one target (MEDV – the price of the house)

Attributes (translated):

- CRIM – wskaźnik przestępczości na mieszkańca według miasta
- ZN – część działki pod zabudowę mieszkaniową pod działki o powierzchni ponad 25 000 stóp kwadratowych
- INDUS – odsetek niedetalicznych akrów biznesowych na miasto.
- CHAS – zmienna zmienna Charles River (1, jeśli trasa ogranicza rzekę; 0 w przeciwnym razie)
- NOX – stężenie tlenków azotu (części na 10 milionów)
- RM – średnia liczba pokoi na mieszkanie
- AGE – odsetek jednostek zajmowanych przez właścicieli wybudowanych przed 1940 r
- DIS – ważone odległości do pięciu centrów zatrudnienia w Bostonie
- RAD – indeks dostępności do radialnych autostrad
- TAX- pełna stawka podatku od nieruchomości od 10 000 USD
- PTRATIO – stosunek liczby uczniów do nauczycieli według miasta
- B – 1000 (Bk – 0,63) ^ 2, gdzie Bk to odsetek czarnych według miasta
- LSTAT -% niższy status populacji
- MEDV – Mediana wartości domów zajmowanych przez właścicieli w tysiącach dolarów

Prediction – let's learn XGBoost regressor (vs. linear regression)

Data characteristics

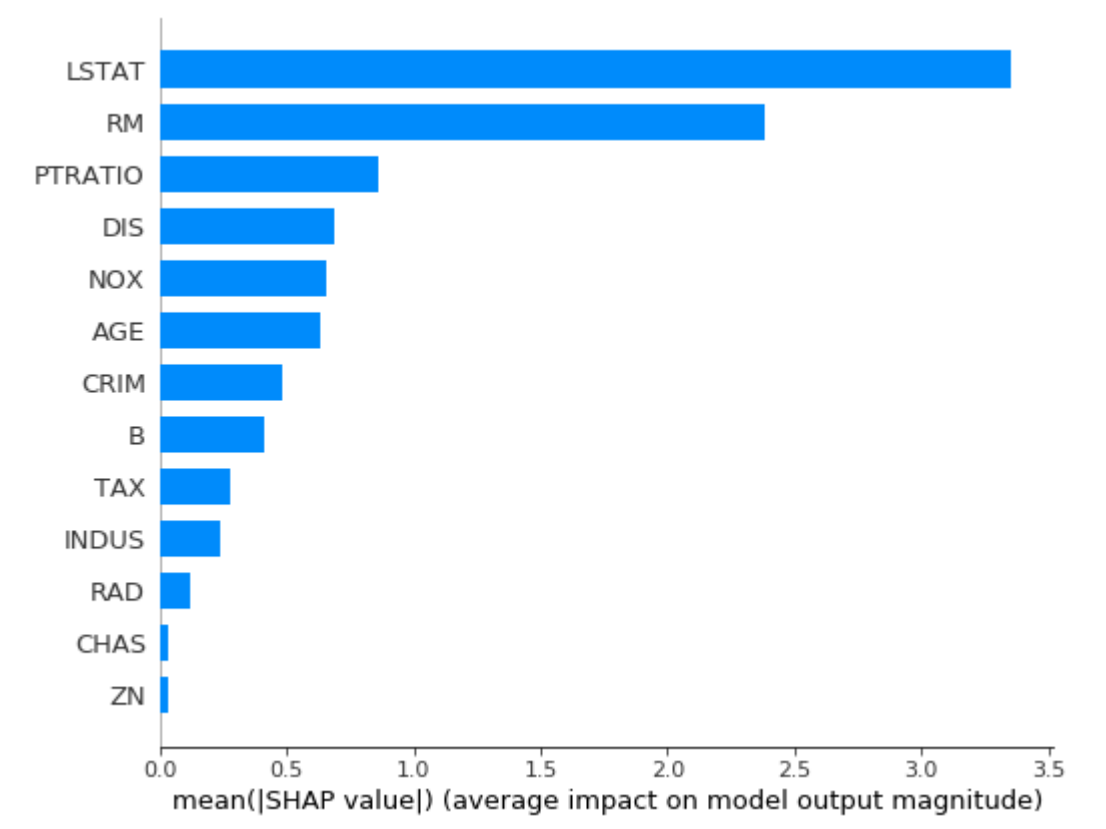


Same attributes skewed with outliers

Global interpretability

Graphs summarizing the impact of attributes on the model prediction

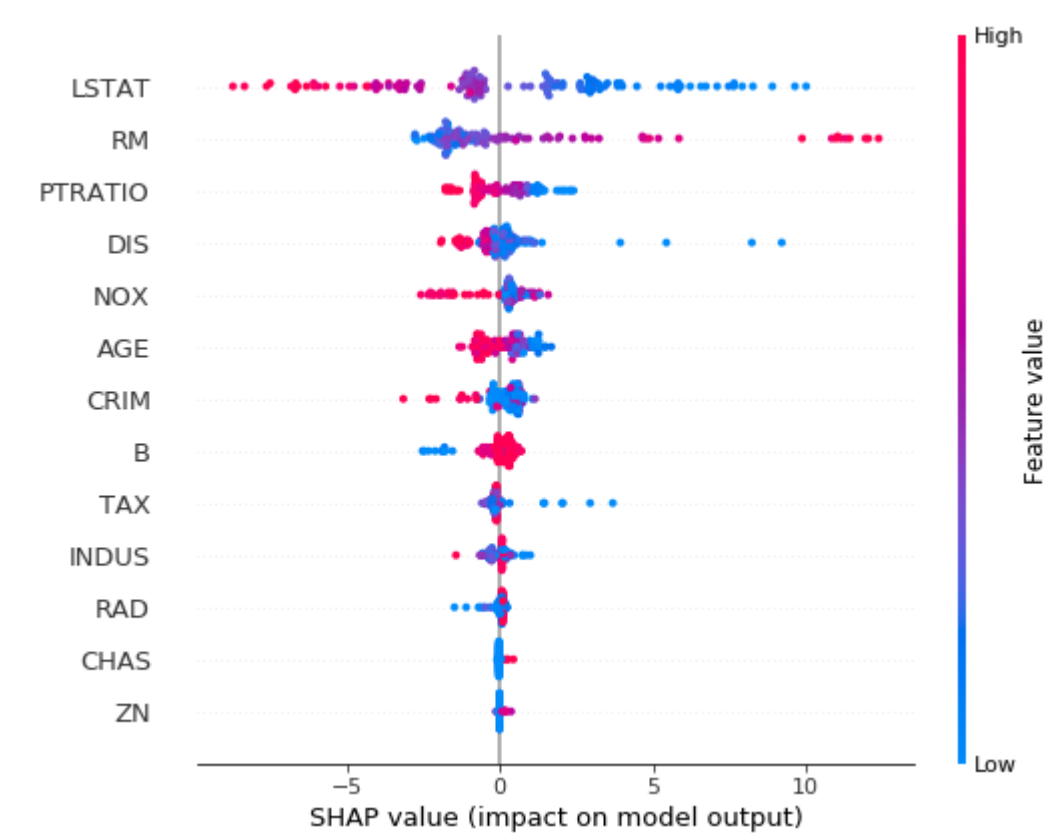
Ranking following the descending Shapley values / the higher the most influential / here LSTAT, RM, … the most influential, while CHAS and ZN the worst.



Global interpretability

The SHAP value plot can further show the positive and negative relationships of the predictors / attributes with the target variable

Red color means an increase while blue show decrease



The SHAP positive and negative impact

This plot is made of all the dots in the train data. It demonstrates the following information:

- Feature importance: Variables are ranked in descending order.
- Impact: The horizontal location shows whether the effect of that value is associated with a higher or lower prediction.
- Original value: Color shows whether that variable is high (in red) or low (in blue) for that observation.

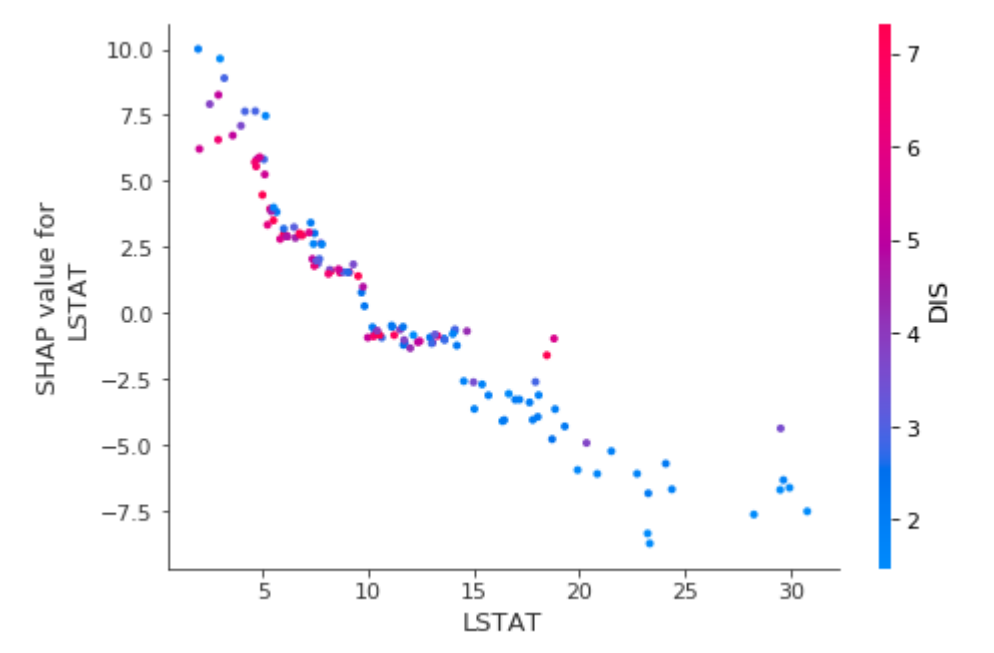
LSTAT : its lower values - high impact on the higher price

RM : contrary relation – its higher values increases the price

SHAP Dependence Plot — Global Interpretability

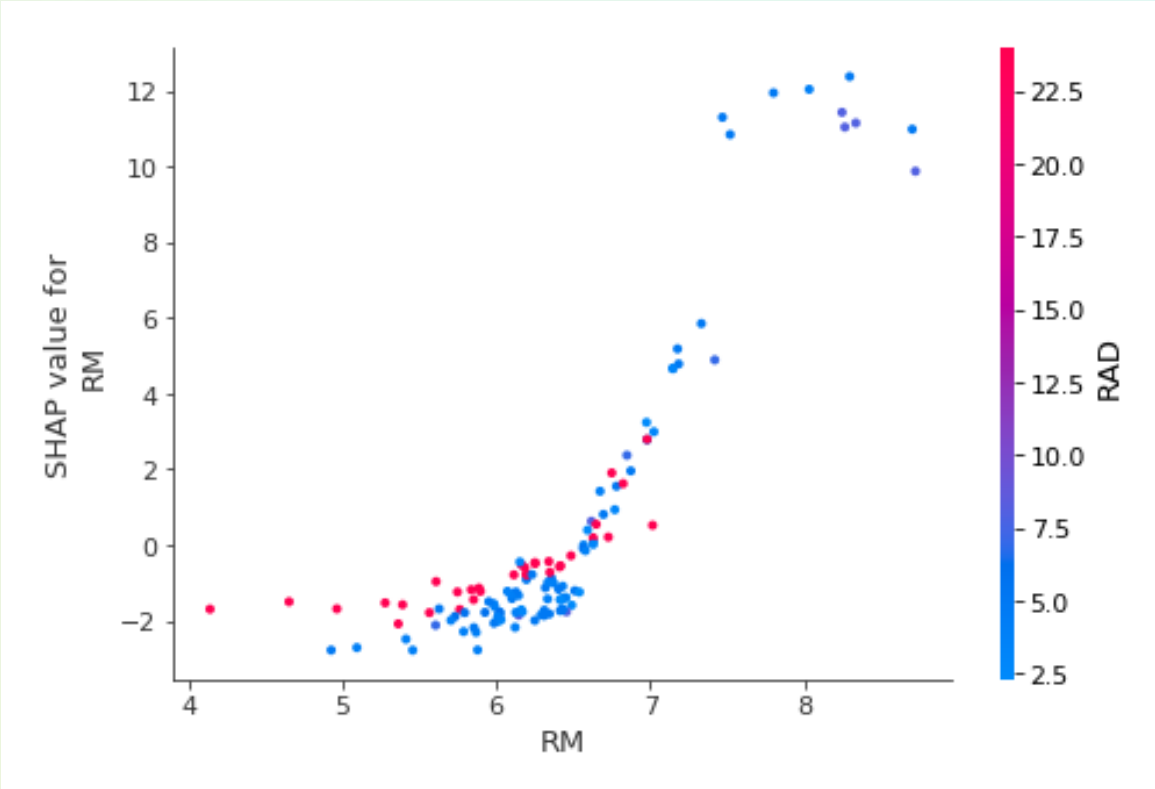
It shows the marginal effect one or two attributes have on the predicted outcome of a machine learning model. Each dot – one instance.

One can check whether the relationship between the target and an attribute is linear, monotonic or more complex



LSTAT is negatively linearly correlated with the output price / DIS is the next correlated attribute; its higher values (red dots) occur for smaller values of LSTATA.

Other dependency plot



The next important attribute RM positively related to output / it is also inter-related with RAD

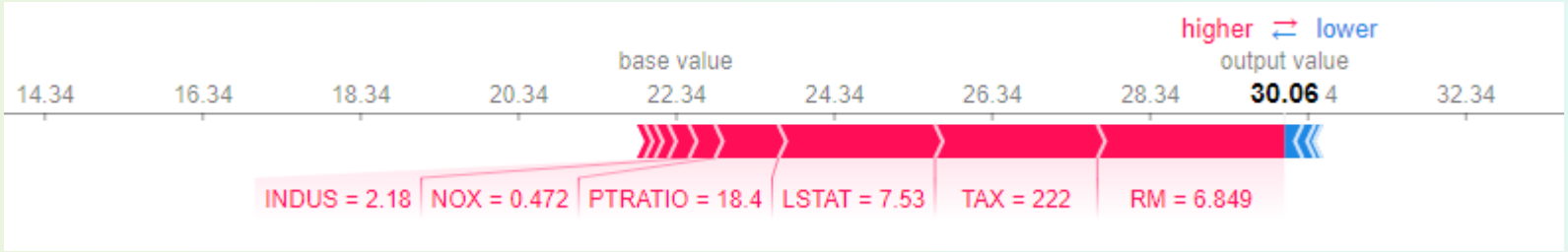
Local interpretability

SHAP determines a separate set of values for each instance in the dataset. Could be useful:

- It allows to explain why model output takes given value for each observation (in case of credit decisions, each rejection/approval can be explained).
- It can determine the observations where a certain variable or a set of variables are more/less predictive, and thus it aids in segmentation;
- It can help in optimizing the model by removing outliers (observations where SHAP values are low for a big number/ all variables)
- It can help in explaining interdependencies between variables at a local/ segment level
- It can help with model exclusions, as missing features have no attributed impact to the model parameters.

Local role of attributes for changing prediction

One of the instances



output value – prediction of a model (30,06) – higher than average values for the all instances (base value 22.34)

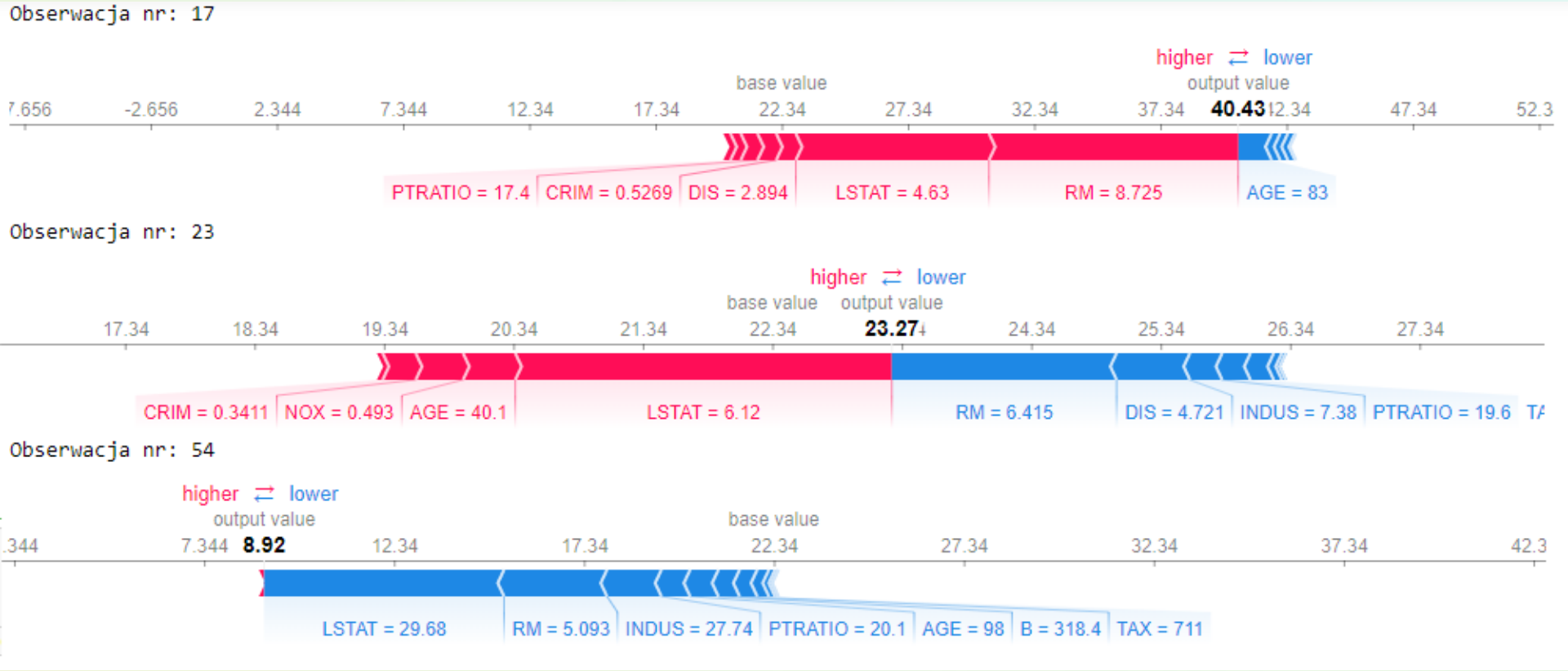
Colors – red – attributes pushing an increase of the output price, while the opposite holds for attribute in blue

The order of attributes follow their Shapley values contribution on the output values / here the most influential RM, TAX then LSTAT and PTRATIO

Compare predictions for more instances

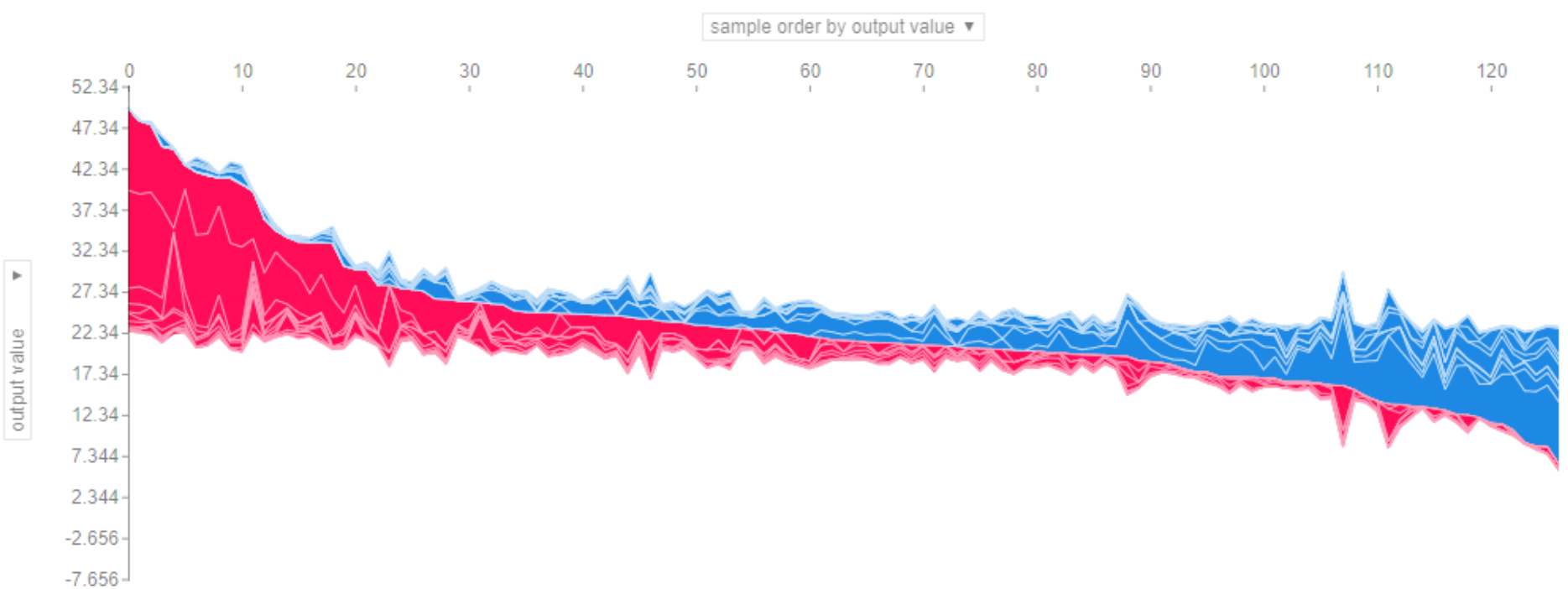
Compare predictions for instances 17, 23 and 54

It help in better analysis particular decisions (e.g. credit policy)

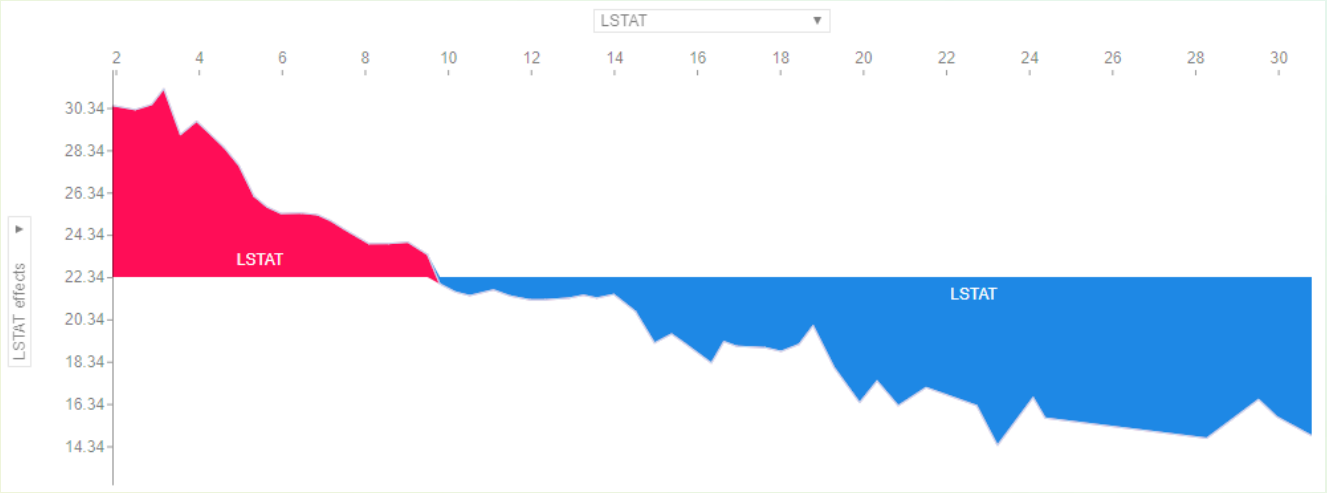


Explaining more instances in the data

Take more local explanations such as the one shown above, rotate them 90 degrees, and then stack them horizontally / could be entire dataset



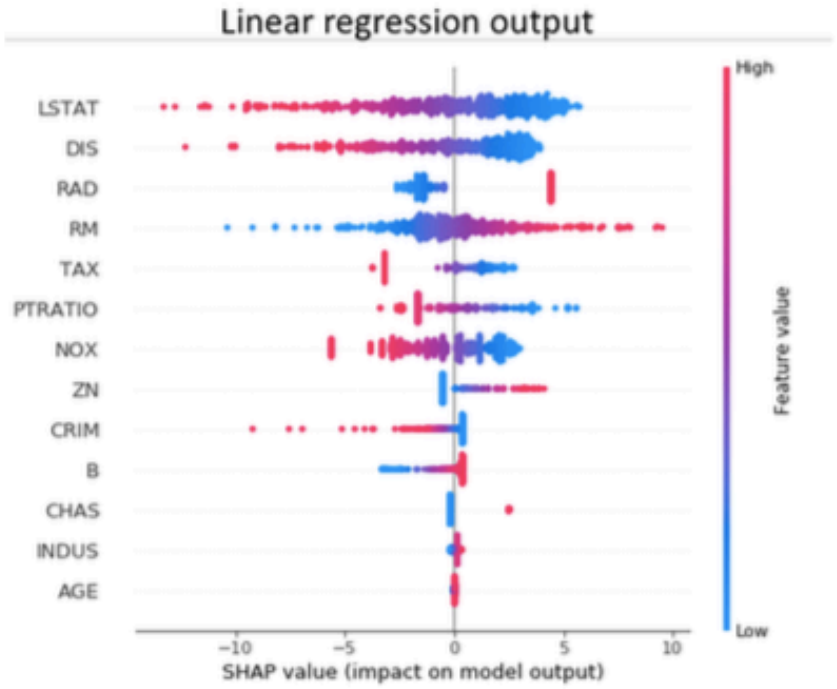
Analysing single attributes

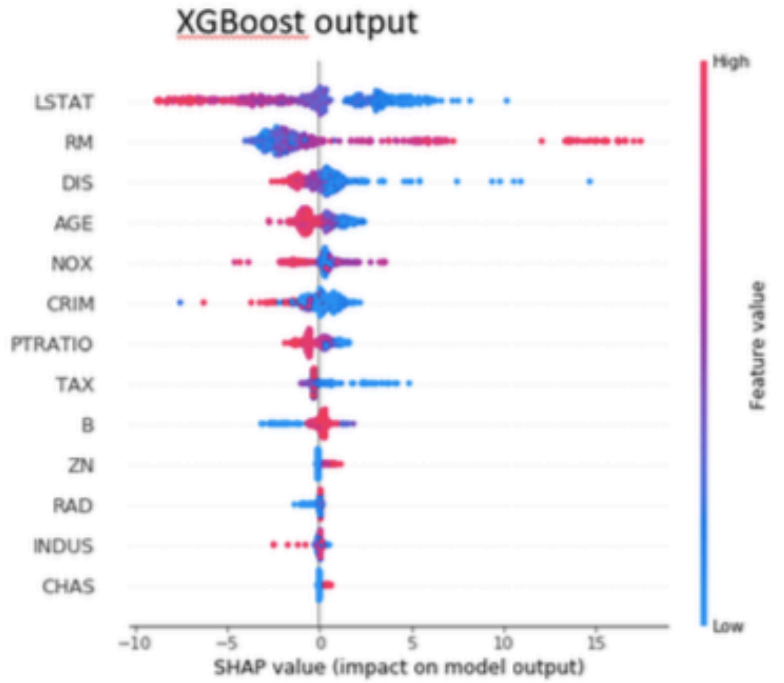


Select variable and make zoom

Compare Xboost vs. linear regression

Global evaluation – some differences in attribute rankings





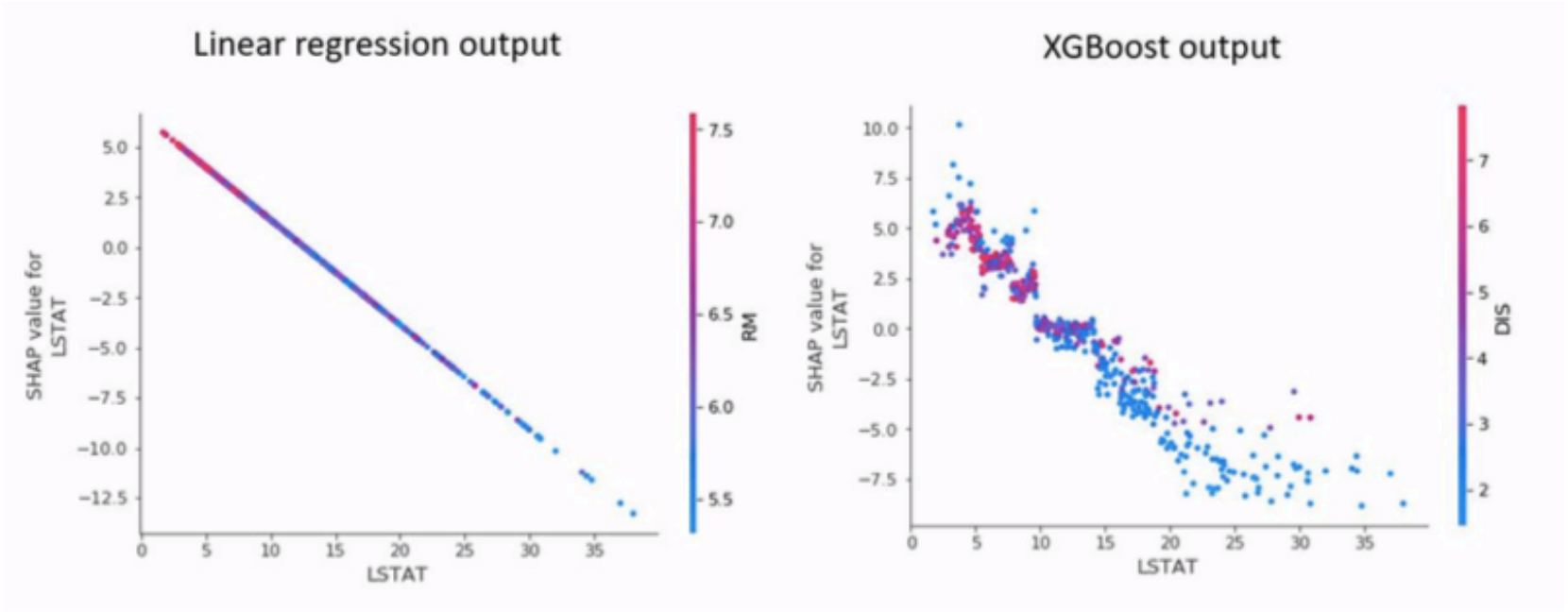
In both models, the LSTAT variable has the most predictive power. The order of importance, further on, is slightly different, given that the regression is constrained in fitting the relationship to a linear one, while the XGBoost can use non-linear components to describe it. This is also apparent in the single-variable graphs, which, in addition to showing the positive/negative relationships to the target, are also showing the form of the relationship.

Linear regression vs.XGBoost

Applying a multiple linear regression to the boston dataset, the resulting R^2 is **76%**, with a RMSE value of **4.67**.

Same values on the XGBoost are **87%** (R^2) and **3.11** (RMSE)[4], showing that the model is better fitted on the data.

Dependency plost - XGBoost more complex relations



The XGBoost model reveals a more complex relationship being captured, and it better explains interactions between variables.

Different predictions of model

Local interpretations for the same instance predictions

SHAP summary

- SHAP values do not identify causality, it has to be handled by special experimental techniques
- Differences to LIME and other agnostic methods

Disadvantages

Shapley values can be misinterpreted and access to data is needed to compute them

Computations still slow

- Several extensions, e.g.
  Novel SHAP variant for capturing pairwise interaction effects

Another usage of Shapley values
Evaluating rule induced from data

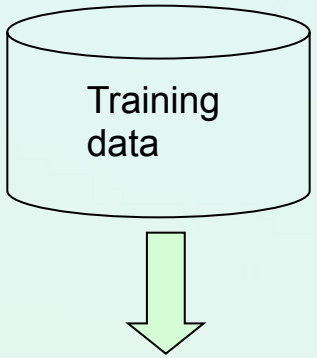
An original proposal of J.Stefanowski et al.
paper 2007

Slides from ECMLPKDD MLLS workshop
invited plenary talk by J.Stefanowski

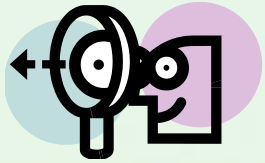
Motivations for interpreting rule patterns

- Description perspective → each rule evaluated individually - possibly an „interesting pattern”.
- Difficulties
  - Too many rules to be analyzed!
    - HSV (122 ob.×11 attr.) → 44 rules
    - Urology (500 ob. × 33 attr.) → 121 rules
- Related works → focus interest on some rules:
  - Subjective vs. objective perspective
  - Rule selection or ordering
    - Studies on rule evaluation measures
  - Interactive browsing

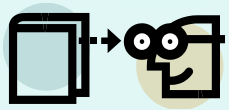
⇨ Need for identification of characteristic attribute value pairs describing patients from particular classes



r1. (A6 = 3) => (D1=1);
r2. (A1=2)&(A2=2)&(A4=2)&(A5 =3) => (D1=1);
r3. (A1 =2)&(A3=1)&(A4=2) & (A =3) => (D1=1)
r4. (A1 = 2) & (A5 = 1) => (D1=2);
r5. (A1 = 2) & (A6 = 1) => (D1=2);
r6. (A2 = 1) & (A4 = 3) => (D1=2);
r7. (A2 = 1) & (A5 = 1) => (D1=2);
.....



Rule interestingness measures



Rule *R*: IF ***P*** THEN ***K***

Objective measures → quantify R with the contingency table (learning data - *n*)

	K	¬*K*	
P	*a*	*c*	n_P
¬*P*	*b*	*d*	$n_{\neg P}$
	n_K	$n_{\neg K}$	*n*

$$sup(R) = a$$

$$G(P \wedge K) = \frac{a}{n}$$

$$conf(R) = \frac{a}{a+c}$$

$$IND(K,Q) = \frac{G(P \wedge K)}{G(P) \cdot G(K)}$$

$$K(K \mid P) = G(P)^{\alpha} \cdot (P(K \mid P) - G(K))$$

Many measures → see McGarry, Geng, L., Hamilton, H.et al. surveys;

- Besides support → Bayesian confirmation measures (*K*,*P*)
- Study impact of the rule premise on its conclusion
  - Refer to class probabilities → imbalance

$$C(P,K) = conf(K \mid P) - P(K)$$

$$N(K,P) = P(P \mid K) - P(P \mid \neg K) = \frac{a}{a+b} - \frac{c}{c+d}$$

I.Szczech, S.Greco, R.Slowinski: Properties of rule interestingness measures. Inf. Scie. 2012

Toward analysing conditions in rules

if $p_1 \wedge p_2 \wedge \ldots \wedge p_n$ *then class K*

if (blacking=medium) ∧ (oil_cons=low) ∧ (horsepower=high)
then (technical condition = good)

Current proposals:

→ Selecting a subset of rules from a larger set of many rules;

→ Focus on a „complete" condition part of a rule!

New view → evaluating an importance of elementary conditions and their interaction within the „if" part of the rule

Our aims:

- To propose a new approach based using set functions → Shapley, Banzhaf indices and Möbius representation
  - Start from a single rule → then generalize to the set of rules
- To verify the approach in rule discovery problems

Adaptation to evaluate conditions in a single rule

- Consider a single rule *if* $p_1 \wedge p_2 \wedge \ldots \wedge p_n$ *then class K*
- Need to analyse its sub-rules *if* $p_{j1} \wedge p_{j2} \wedge \ldots \wedge p_{jl}$ *then class K* such that $\{p_{j1}, p_{j2}, \ldots, p_{jl}\} \subseteq \{p_1, p_2, \ldots, p_n\}$
  - sub-rules are more general than the first rule
- Choice of the characteristic function μ to evaluate a rule?
  - Confidence of the rule $\mu(W,K)=conf(r)$, where W is a set of conditions in r
  - Also – confirmation measures, …
- Then, for $Y \subset W$ we need to adapt set functions
  - $\mu(\varnothing,K)$=? O or class prior

Indices for each condition in a rule

$p_i \in W$ - single condition in rule r, and $|W| = n$

- Shapley value:

$$\Phi_s(p_i, r) = \sum_{Y \subseteq W - \{p_i\}} \frac{(n - |Y| - 1)!|Y|!}{n!} \cdot [\mu(Y \cup \{p_i\}, K) - \mu(Y, K)]$$

- Banzhaf value:

$$\Phi_B(p_i, r) = \frac{1}{2^{n-1}} \sum_{Y \subseteq W - \{p_i\}} [\mu(Y \cup \{p_i\}, K) - \mu(Y, K)]$$

Both values Φ – a weighted contribution of p_i in rules generalized from r
For Shapley value - $\mu(W)$ is shared among all elements of W

Pairs – measures of an interaction resulted from putting p_i and p_j together in all subsets of conditions in rule r:

- Positive – complementary in increasing the confidence
- Negative – putting together provide some redundancy

$$I_{MS}(p_i, p_j) = \sum_{Y \subseteq W - \{p_i, p_j\}} \frac{(n - |Y| - 2)!|Y|!}{(n-1)!} \cdot [\mu(Y \cup \{p_i, p_j\}, K) - \mu(Y \cup \{p_i\}, K) - \mu(Y \cup \{p_j\}) + \mu(Y, K)]$$

Adapted indices for subsets - part 2

Generalized indices for a subset of conditions V⊂W [Grabisch]

Shapley generalized index

$$I_S(V,r) = \frac{1}{2^{n-|V|}} \sum_{Y \subseteq W-V} \frac{(n-|Y|-|V|)!|Y|!}{(n-|V|+1)!} \sum_{L \subseteq V} (-1)^{|V|-|L|} \mu(Y \cup L, K)$$

Banzhaf index of conditions V⊂W

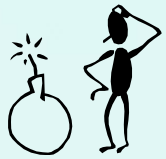
$$I_B(V,r) = \frac{1}{2^{n-|V|}} \sum_{Y \subseteq W-V} \sum_{L \subseteq V} (-1)^{|V|-|L|} \mu(Y \cup L, K_j)$$

Average conjoint contribution of the subset of conditions V⊂W to the confidence of all rules generalized from *r*

The Möbius representation of set functions μ :

$$m(V,r) = \sum_{B \subseteq V} (-1)^{|V-B|} \mu(B,K)$$

An intuitive example



HSV treatment – one of the rules

if (gastric_juice=*medium*)∧(HCL_conc.=*low*) then (*result*=*good*)
conf=1.0 , supp = 13 examples.

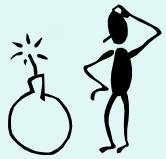
Möbius representation $m(1,2)$ = -0.14493!!!

- Rule generalizations and Möbius representation m:
  - Empty condition part → $m(0)=0$
  - *if* (gastric_juice=*medium*) then (*result* =*good*)
    $m(1)=0.16667$ and conf=0.16667
  - *if* (HCL_conc.=*low*) then (*result* =*good*)
    $m(2)=0.97826$ and conf= 0.97826
- An increase of rule confidence
  1 = m(1) + m(2) + m(1,2)
- Values of Möbius representation show the distribution of confidence among all coalitions of the considered conditions in the subset {(gastric_juice=*medium*),(HCL_conc.=*low*)}

Shapley value for single conditions

φ(gastric_juice=*medium*)=0.0942; φ((HCL_conc.=*low*) =0.908

An intuitive example



HSV treatment – one of the rules

if (gastric_juice=*medium*)∧(HCL_conc.=*low*) then (*result*=*good*)
conf=1.0 , supp = 13 examples.

Möbius representation $m(1,2)$ = -0.14493!!!

- Rule generalizations and Möbius representation m:
  - Empty condition part → $m(0)=0$
  - *if* (gastric_juice=*medium*) then (*result* =*good*)
    $m(1)=0.16667$ and conf=0.16667
  - *if* (HCL_conc.=*low*) then (*result* =*good*)
    $m(2)=0.97826$ and conf= 0.97826
- An increase of rule confidence
  1 = m(1) + m(2) + m(1,2)
- Values of Möbius representation show the distribution of confidence among all coalitions of the considered conditions in the subset {(gastric_juice=*medium*),(HCL_conc.=*low*)}

Shapley value for single conditions

φ(gastric_juice=*medium*)=0.0942; φ((HCL_conc.=*low*) =0.908

Evaluating conditions in ACL rule

if (sex = female)∧(Y1 < 2.75) ∧(PCL∈[3.71,4.13)) *then* (no ACL) conf. =1.0

Sex	Y1	PCL	Banzhaf	Shapley	Mobius	conf.
∅	∅	√	0.43535	0.49575	0.28571	0.2857
∅	√	∅	0.24207	0.30246	0.04651	0.0465
∅	√	√	0.53015	0.53015	0.1766	0.5241
√	∅	∅	0.14139	0.1591	0.1452	0.1452
√	∅	√	0.1135	0.1135	-0.2316	0.2923
√	√	∅	0.1734	0.1734	-0.1034	0.1486
√	√	√	0.72476	0.72476	0.72476	1

Evaluating conditions in a set of rules

- The set of rules $R = \bigcup_{j=1}^{k} R(K_j)$, where *R*(*Kj*) a set of rules having as a consequence class *Kj*
- A given set of conditions Γ_f occur in many rules
- $FM_r(\Gamma_f)$ denote an evaluation of its contribution to the confidence of rule *r*
- The global contribution of Γ_f in a rule set *R* with respect to class *Kj* is calculated as:

$$G_{Kj}(\Gamma_j) = \sum_{r \in R(Kj)} FM_r(\Gamma_f) \cdot \sup(r) - \sum_{s \in \neg R(Kj)} FM_s(\Gamma_f) \cdot \sup(s)$$

- Conditions Γ_f are ranked according to $G_{Kj}(\Gamma_f) \rightarrow$ identify the most characteristic combinations of conditions for rules from a given class
- Computational costs → start from the smallest sets of cond.

An illustrative example

An interest in condition (a7=0) in a set of several rules

It occurs in following rules with conf=1:

R1 if (a3=1)∧(a7=0) ∧(a3=1) then (D=1) sup 1

R2 if (a4=1)∧(a7=0) then (D=1) sup 45

R5 if (a4=0)∧(a7=0) then (D=2) sup 7

(Möbius representation of (a7=0)) in R1, R2 m=0.939
and in R5 m=0.184

A global contribution of (a7=0)

- (D=1) 0.939×1 + 0.939 × 45 = 43.194
- (D=2) 0.184 ×7 = 1.288

Finally $G_{D=1}$(a7=0) = 43.194 – 1.288 = 41.906

Analysis of conditions in buses rules

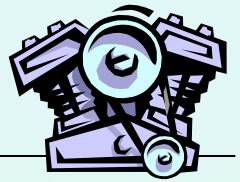


Table 1. Rankings of best conditions according to evaluation measures calculated for "buses" rules

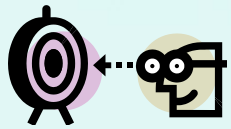
busses in a good technical condition					
Möbius		Shapley		Banzhaf	
condition	value	condition	value	condition	value
comp-press=high	214.34	comp-press=high	116.91	comp-press=high	116.91
torque=high	163.36	torque=high	163.36	torque=high	163.36
blacking=low	161.33	blacking=low	87.86	blacking=low	87.86
oil cons.=low	132.36	oil cons.=low	70.88	oil cons.=low	70.80
MaxSpeed=high	122.66	MaxSpeed=high	63.71	MaxSpeed=high	63.71
busses in a bad technical condition					
Möbius		Shapley		Banzhaf	
condition	value	condition	value	condition	value
torque=low	48.33	torque=low	29.17	torque=low	29.17
blacking=high	46.70	comp-press=low	29.00	comp-press=low	29.00
comp-press=low	29.00	blacking=high	27.98	blacking=high	28.06
oil-cons.=high	27.00	oil-cons.=high	27.00	oil-cons.=high	27.00
summ-cons.=high	26.67	horsepower=low	26.00	horsepower=low	26.66
horsepower=low	26.66	MaxSpeed=low	25	MaxSpeed=low	25

- Pairs of conditions – much lower evaluations e.g. (horsepower=average) and (oil consumption=low) 0.166
- Previous analysis → „good" conditions: high compression pressure, torque, max-speed and low blacking components. Opposite values → characteristic for bad technical conditions. Blacking components in the exhaust gas and oil consumption more important than fuel consumption.

Evaluating conditions in ACL rules

- Diagnosing an anterior cruciate ligament (ACL) rupture in a knee on the basis of magnetic resonance (MR) images (Slowinski K. et al.)
- 140 patients described by 6 attributes
  - age, sex and body side and MR measurements (X, Y and PCL index).
- Patients classified into two classes „1” (with ACL lesion – 100) and „2” (without ACL – 40).
- LEM2 rule induction algorithm → 15 rules (1- 4 elementary conditions with different support, few possible rules).
- Clinical discussion → MR measurements are the most important.
  - In particular PCL< 3.23 (patients with ACL), PCL ≥ 4.53 (without ACL)
  - Other PCL values → combinations with two other attributes age or sex indicate classes.
  - Age below 16.5 years (so children or youth) characteristic for class (without ACL lesion).
  - ACL injury more frequent for men (sportsmen)!

ACL → minimal set of rules



rule 1. *if* (PCLINDEX $<$ 3.225) *then* Class1 [26, 65%]
rule 2. *if* (AGE=[16.5,35))$\wedge$(PCLINDEX=[3.225,3.71)) *then* Class1 [6, 15%]
rule 3. *if* (SEX=MALE) $\wedge$ (SIDE=RIGHT) $\wedge$ (PCLINDEX=[3.225,3.71)) *then* Class1 [3, 7.5%]
rule 4. *if* (AGE=[16.5,35)) $\wedge$ (PCLINDEX=[3.71,4.125)) $\wedge$ (X$\geq$14.5) *then* Class1 [2, 5%]
rule 5. *if* (X=[8.5,11.75)) $\wedge$ (PCLINDEX=[4.125,4.535)) $\wedge$ (SEX=MALE) *then* Class1 [1, 2.5%]
rule 6. *if* (X=[8.5,11.75)) $\wedge$ (PCLINDEX=[3.225,3.71)) $\wedge$ (AGE$\geq$ 35) *then* Class1 [2, 5%]
rule 7. *if* (PCLINDEX=[3.71,4.125)) $\wedge$ (X=[8.5,11.75)) $\wedge$ (SEX=1) *then* Class1 [1, 2.5%]
rule 8. *if* (PCLINDEX$\geq$4.535) *then* Class2 [75, 75%]
rule 9. *if* (SEX=FEMALE) $\wedge$ (PCLINDEX=[4.125,4.535)) *then* Class2 [10,10%]
rule 10. *if* (PCLINDEX=[3.71,4.125)) $\wedge$ (AGE$\geq$ 35) *then* Class2 [6,6%]
rule 11. *if* (X=[11.75,14.5)) $\wedge$ (Y=[2.75,3.75)) $\wedge$ (SEX=FEMALE) *then* Class2 [8, 8%]
rule 12. *if* (SIDE=LEFT) $\wedge$ (X=[11.75,14.5)) $\wedge$ (Y=[2.75,3.75)) *then* Class2 [7, 7%]
rule 13. *if* (PCLINDEX=[3.225,3.71)) $\wedge$ (AGE$\geq$ 35) $\wedge$ (SEX=MALE) *then* Class2 [2, 2%]
rule 14. *if* (AGE$<$16.5) *then* Class2 [14, 14%]
rule 15.*if* (PCLINDEX=[3.225,3.71))$\wedge$(Y=[3.75,4.75))$\wedge$(AGE$\geq$ 35)$\wedge$(SIDE=LEFT) *then* Class2 [1,1%]

Evaluating conditions in ACL rules

With ACL				Without ACL			
Möbius		Shapley		Möbius		Shapley	
PCL < 3.23	18.57	PCL < 3.23	18.57	PCL ≥ 4.53	21.42	PCL ≥ 4.53	21.42
PCL∈[3.23,3.7)	4.87	PCL∈[3.23,3.7)	5.06	Age < 16.5	4.0	Age < 16.5	4.0
(Age∈[16.5,35] & (PCL ∈[3.7,4.1)	1.58	Age∈[16.5,35)	1.63	Sex=female	3.23	Sex=female	2.85
(X1≥14.5) & (PCL ∈[3.7,4.1)	0.54	(X1≥14.5) & (PCL ∈[3.7,4.1)	0,92	PCL∈[4.13,4.5)	2.22	Y1∈[2.75,3.75)	1.84
X1 ∈[8.5,11.8)	0.52	Age∈[16.5,35 & (PCL ∈[3.7,4.1)	0.86	(Age≥35] & (PCL ∈[3.7,4.1)	1.78	(Age≥35] & (PCL ∈[3.7,4.1)	1.78
Sex=male	0.44	Sex=male	0.83	X1∈[11.8,14.5) PCL∈[3.23,3.7)	1.31	X1∈[11.8,14.5)	1.53
Age∈[16.5,35)	0.34	Y1<2,75 & (PCL ∈[3.7,4.1)	0.67	Y1∈[2.75,3.75)	1.28	PCL∈[4.13,4.53	1.48

Subsets of conditions → characteristic description of both diagnostic classes; PCL index with extreme intervals definitely the most important + its other values occur in some pairs, e.g (Age∈[16.5,35]) & (PCL ∈ [3.7,4.1)

Sex and age – young men (often sportsmen)

Evaluating conditions in ACL rules

- Rankings of conditions with respect to Shapley and Banzhaf values – top elements are the same.
- Top ranking with Möbius representation small re-ordering but PCL also dominates
- Pairs of conditions are higher evaluated than in the previous case
- Support for profiles of ACL patients
  - MR measurements are the most important

  Patients with ACL
  - PCL< 3.23 ; (Age∈[16.5,35]) & (PCL ∈ [3.7,4.1)
  - Sex=male and X1 ∈[8.5,11.8)

  Patients without ACL
  - PCL ≥ 4.53
  - Other MR measurements → combinations with two other attributes age or sex indicate classes.
  - Age below 16.5 years (so children or youth) or (age = much older) are characteristic for (without ACL)
- Profiles consistent with the earlier analyses and clinical knowledge

Highly selective vagotomy rules

Highly selective vagotomy (HSV) - laparoscopic surgery for perforated Duodenal Ulcer Disease.

- An attempt to determine indications for surgery treatment;
  - 122 patients described by 11 pre-operating attributes and assigned to 4 target class
  - 44 rules (1- 5 conditions)
- Focus on describing characteristic profiles of patients
- The previous results, e.g. very good prediction – class 1)
  - long or medium duration of the disease,
  - without complications of ulcer or acute haemorrhage from ulcer,
  - medium or small volume of gastric juice per 1 hour (basic secretion),
  - medium volume of gastric juice per 1 hour under histamine,
  - high HCl concentration under histamine.

Evaluating conditions in HSV rules – class 1 (good)

Möbius		Shapley		Banzhaf	
Cond	*Value*	*Cond*	*Value*	*Cond*	*Value*
A6=2	2,34	A6=2	3,85	A6=2	4,01
A9=3	2,31	A4=1	3,41	A4=1	3,57
A4=2	1,89	A4=2	3,16	A4=2	3,08
A4=1	1,58	A9=3	2,59	A9=3	2,72
A2=2	1,27	A2=2	1,65	A2=2	1,88

Möbius		Shapley		Banzhaf	
Cond	*Value*	*Cond*	*Value*	*Cond*	*Value*
A4=1 & A6=2	2,62	A4=1 & A6=2	2,82	A4=1 & A6=2	2,83
A4=1 & A8=1	1,95	A4=1 & A8=1	1,95	A4=1 & A8=1	1,95
A2=2 & A6=2	1,89	A5=2 & A6=1	1,49	A5=2 & A6=1	1,49
A2=2 & A9=3	1,53	A3=3 & A7=2	1,18	A3=3 & A7=2	1,18
A5=2 & A6=1	1,49	A2=2 & A6=2	1,01	A2=2 & A6=2	1,01

Attributes: A2 – age; A4 – complications of ulcer; A6 - volume of gastric juice per h; A9 - HCL concentration after histamine; A5 - HCL concentration; A3 - duration of disease

Subsets of conditions → closer to single conditions

HSV –patient class profiles

- Very good result of HSV (class 1)
  - without complications of ulcer or acute haemorrhage from ulcer,
  - medium or small volume of gastric juice per 1 hour (basic secretion),
  - medium volume of gastric juice per 1 hour under histamine,
  - high HCl concentration under histamine
  - / no medium duration of disease
- Satisfactory result of HSV (class 2)
  - long or medium duration of disease,
  - multiple haemorrhages,
  - medium or small volume of gastric juice per 1 hour (basic secretion),
  - medium volume of gastric juice per 1 hour under histamine,
  - medium or low HCl concentration under histamine
- Unsatisfactory result of HSV treatment (class 3)
  - medium or short duration of the disease,
  - perforation of ulcer,
  - high or small volume of gastric juice per 1 hour (basic secretion),
  - high volume of gastric juice per 1 hour under histamine,
  - No low HCl concentration under histamine condition in the rankings
- Bad result of HSV treatment (class 4)
  - Consistent profile
  - + new condition - low HCl concentration under histamine

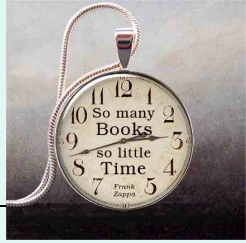
Working with larger set of rules

- „ESWL” – urological data
  - Urinary stones treatment by ESWL extracorporeal shock waves lithotripsy
- 500 patients × 33 attributes classified into two classes (imbalanced) – difficult to analyse (Antczak, Kwias et al. 2000)
- Explore rule induction algorithm → 484 rules (2-7 conditions with different support ≥ 5%, confidence ≥ 0.8).

ESWL rules

- Explore rule induction algorithm → 484 rules (2-7 conditions with different support ≥ 5%, confidence ≥ 0.8).
- Using the set functions we identify:
  - Class 1 → 8 single conditions, 12 pairs
    - (basic dysuric symptoms=1), (crystaluria=1), (location of the concrement=2),(stone size=2), …, (crystaluria=2)&(proteinurine=1), etc.
  - Class 2 → 10 single conditions, 13 pairs
    - (location of the concrement =3), (lumbar region pains=5), (operations in the past=3),…, (crystaluria=3)&(proteinurine=2),.., (cup-concrement=1)&(stone size=2), etc.
- More visible differences in Shapley and Banzhaf rankings ; triples less evaluated than single conditions and pairs.

Extensions to improve computability



- Limitations - computational for rules having more conditions
  - Both time and memory (to store temporary results)
- Possible heuristic approaches:
  - First filter and reduce the set of rules, then evaluate.
  - Iterative analysis, start from single conditions, pairs and work with smaller sets of conditions
- Modify calculations of measures (approximate them)

  M.Sikora: *Selected methods for decision rule evaluation and pruning* (2013)
  - Analyse only single conditions in rules
  - Do not consider all sub-rules (restrict to rules affected by dropping the single condition, or base sub-rules with the single condition)
  - Simpler forms of Baznhaf and Shapley indices

  Possible re-using of best conditions in rule constructive induction

Recap

- SHAP and Shapely Values have a solid theoretical foundation of Game Theory.
- Shapely values guarantee that the prediction is fairly distributed across the different features.
- SHAP connects other interpretability techniques, like LIME and DeepLIFT, to the strong theoretical foundation of Game Theory.
- SHAP has a lightning-fast implementation for Tree-based models, which are one of the most popular sets of methods in Machine Learning.
- SHAP can also be used for global interpretation by calculating the Shapely values for a whole dataset and aggregating them.
- it provides a strong linkage between your local and global interpretation

Currently popular due to implementations

Few references

Mainly

Authors' papers:

- Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems. 2017.
- Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. "Consistent individualized feature attribution for tree ensembles." arXiv preprint arXiv: 1802.03888 (2018).

Nice books:

- Christoph Molnar, "Interpretable Machine Learning: A Guide for making black box models explainable"
- Przemysław Biecek, Explanatory Model Analysis (book under preparation)

Explain Your Model with the SHAP Values – blog TowardsDataScience

KDD Blog

Some Polish inspirations – also to these slides

M.Mamczur blog Wartość Shapley'a – interpretacja modeli blackbox

Thank you for your attention

Questions and remarks?

Contact, remarks:
Jerzy.Stefanowski@cs.put.poznan.pl

or www.cs.put.poznan.pl/jstefanowski