Social ranking for feature selection

L. Gourvès, S. Moretti, S. Tamby

CNRS UMR7243 - LAMSADE, Université Paris-Dauphine
stefano.moretti@dauphine.fr

ANR GATSBII kickoff meeting
Toulouse
30-31 January, 2025

Motivations of XAI: feature selection

We need techniques to eXplain/interpret AI models:

- For justifying decisions taken according to a recommendation provided by a black box
- For discovering new relationships between features

Motivations of XAI: feature selection

We need techniques to eXplain/interpret AI models:

- For justifying decisions taken according to a recommendation provided by a black box
- For discovering new relationships between features
- If an AI model is too difficult to interpret due to too many features, one can try to reduce the number of features...

Motivations of XAI: feature selection

We need techniques to eXplain/interpret AI models:

- For justifying decisions taken according to a recommendation provided by a black box
- For discovering new relationships between features
- If an AI model is too difficult to interpret due to too many features, one can try to reduce the number of features...
- ... without reducing too much the performance of the model...

Motivations of XAI: feature selection

We need techniques to eXplain/interpret AI models:

- For justifying decisions taken according to a recommendation provided by a black box
- For discovering new relationships between features
- If an AI model is too difficult to interpret due to too many features, one can try to reduce the number of features...
- ... without reducing too much the performance of the model...
- ... filtering out features that are strongly correlated

Feature selection: ranking the features by order of importance.

Shapley value for feature selection

- Applications of the Shapley value (Shapley (1953))) range from pioneer studies on linear regression analysis (Israeli (2007)) and on feature selection for classification models (Cohen et al. (2005))

Shapley value for feature selection

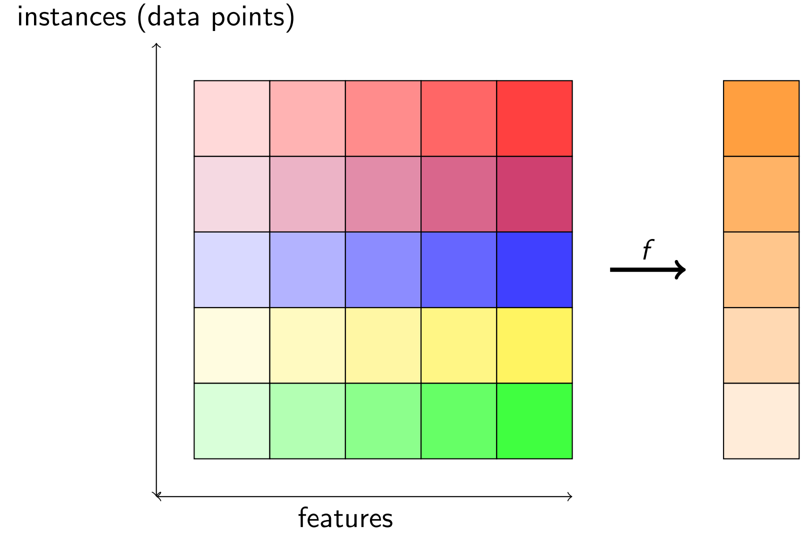
- Applications of the Shapley value (Shapley (1953))) range from pioneer studies on linear regression analysis (Israeli (2007)) and on feature selection for classification models (Cohen et al. (2005))
- to very popular applications like the SHapley Additive exPlanations (SHAP) (Lundberg and Lee (2017); Lundberg et al. (2020)) and the Shapley Additive Global importancE (SAGE) (Covert et al. (2021)).

Shapley value for feature selection

- Applications of the Shapley value (Shapley (1953))) range from pioneer studies on linear regression analysis (Israeli (2007)) and on feature selection for classification models (Cohen et al. (2005))
- to very popular applications like the SHapley Additive exPlanations (SHAP) (Lundberg and Lee (2017); Lundberg et al. (2020)) and the Shapley Additive Global importancE (SAGE) (Covert et al. (2021)).
- some studies have recently raised important concerns about the capability of the Shapley value to rank features based on their relevance in constructing simplified models

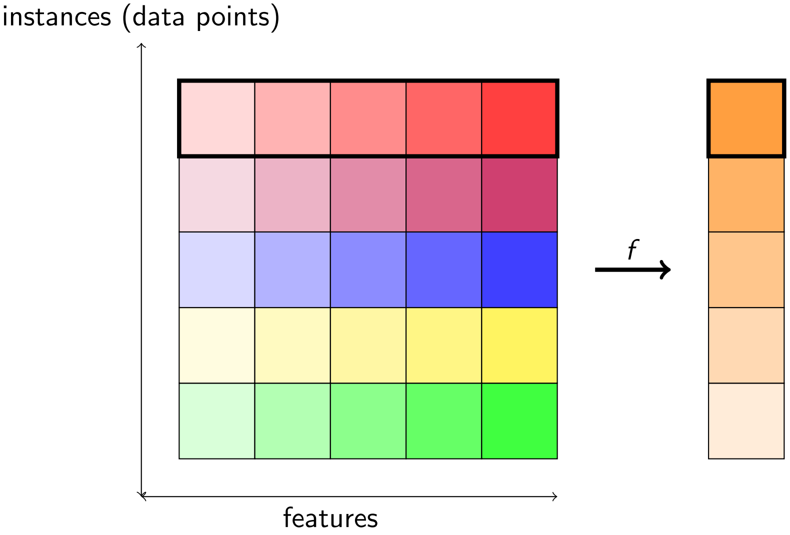
Notations

A dataset X and function f use by the ML model trained on X



Notations

A dataset X and function f use by the ML model trained on X



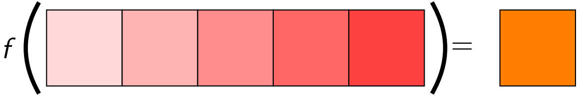
Feature selection as a coalition game

- players: finite set of features N
- coalitions: subset of features (2^N)
- $v(S)$, is the evaluation function on coalition $S \in 2^N$: total deviation of perturbed predictions in the noisy dataset $X_{\bar{\mathbf{x}}_S}$ from the prediction $f(\mathbf{x})$

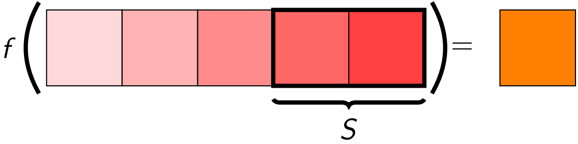
$$v(S) = -\sum_{p \in M} |f(\mathbf{x}^p) - f(\mathbf{x})|. \tag{1}$$

where M is the set of instances in the dataset and $\mathbf{x}^p \in X_{\bar{\mathbf{x}}_S}$, with $p \in M$, in the noisy dataset $X_{\bar{\mathbf{x}}_S}$

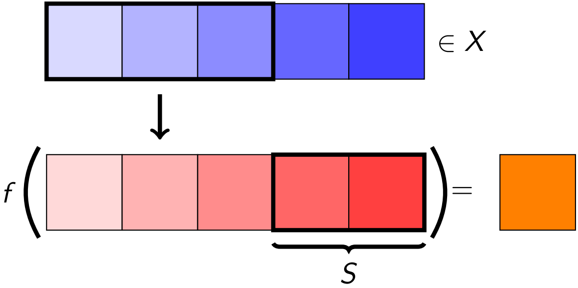
Perturbed data



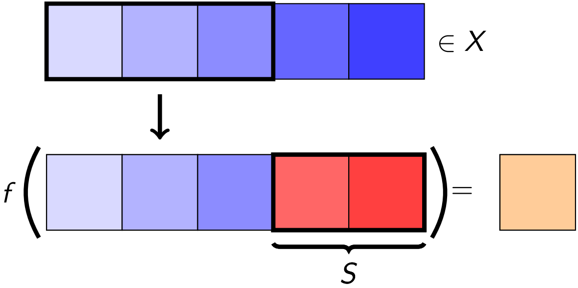
Perturbed data



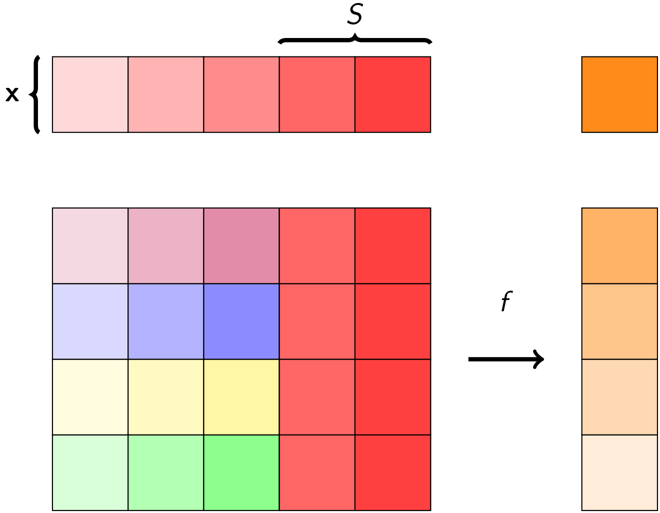
Perturbed data



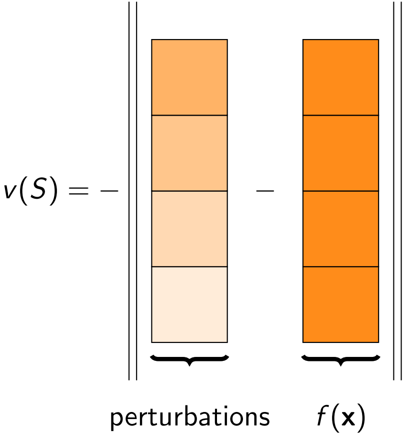
Perturbed data



We sample multiple perturbations...



… and we estimate the value of a coalition by averaging the errors



Shapley value

For any evaluation function (e.f.) $v \in \mathcal{E}^N$, the Shapley value is the vector $\phi(v) = (\phi_1(v), \ldots, \phi_n(v))$ such that

$$\phi_i(v) = \sum_{S \in 2^N : i \notin S} \frac{s!(n-s-1)!}{n!} (v(S \cup \{i\}) - v(S)) \qquad (2)$$

for each $i \in N$, where $s = |S|$ is the cardinality of coalition S.
The Shapley value is the only one-point solution that satisfies the above four properties *i*), *ii*),*iii*) and *iv*) for one-point solutions on the class of evaluation functions $\mathcal{E}^N$ (Shapley (1953)).

i) *efficiency*: $\sum_{i \in N} \psi_i(v) = v(N) - v(\emptyset)$;

ii) *symmetry*: for any $i, j \in N$ such that $v(S \cup \{i\}) = v(S \cup \{j\})$ for all $S \in 2^{N \setminus \{i,j\}}$, then $\psi_i(v) = \psi_j(v)$;

iii) *null player*: for any $i \in N$ such that $v(S \cup \{i\}) - v(S) = 0$ for all $S \in 2^N$, then $\psi_i(v) = 0$;

iv) *additivity*: $\psi(v) + \psi(w) = \psi(v + w)$ for all e.f.s $v, w \in \mathcal{E}^N$.

Ranking: symmetry and strict desirability

- for features selection we wish to rank features according to their relevance in determining the prediction of the whole ML model (with all features).
- Two features i and j are symmetric if $v(S \cup \{i\}) = v(S \cup \{j\})$ for all coalitions $S \in 2^{N \setminus \{i,j\}}$. We consider i and j equally relevant.

Ranking: symmetry and strict desirability

- for features selection we wish to rank features according to their relevance in determining the prediction of the whole ML model (with all features).
- Two features i and j are symmetric if $v(S \cup \{i\}) = v(S \cup \{j\})$ for all coalitions $S \in 2^{N \setminus \{i,j\}}$. We consider i and j equally relevant.
- feature i is strictly more desirable than feature j if
  - $v(S \cup \{i\}) \geq v(S \cup \{j\})$ for all coalitions $S \in 2^{N \setminus \{i,j\}}$
  - and $v(T \cup \{i\}) > v(T \cup \{j\})$ for some $T \in 2^{N \setminus \{i,j\}}$.

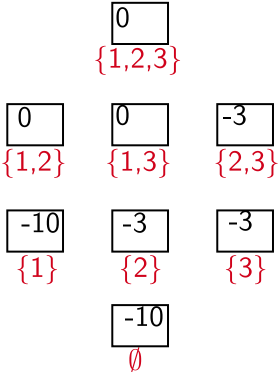
  We consider i strictly more relevant than j.

Ranking: symmetry and strict desirability

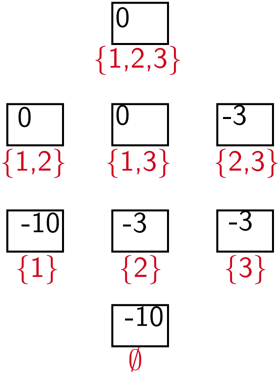
- for features selection we wish to rank features according to their relevance in determining the prediction of the whole ML model (with all features).
- Two features i and j are symmetric if $v(S \cup \{i\}) = v(S \cup \{j\})$ for all coalitions $S \in 2^{N \setminus \{i,j\}}$. We consider i and j equally relevant.
- feature i is strictly more desirable than feature j if
  - $v(S \cup \{i\}) \geq v(S \cup \{j\})$ for all coalitions $S \in 2^{N \setminus \{i,j\}}$
  - and $v(T \cup \{i\}) > v(T \cup \{j\})$ for some $T \in 2^{N \setminus \{i,j\}}$.

  We consider i strictly more relevant than j.
- The Shapley value (and the lex-cel) align with symmetry and strict desirability

Example (secret holder, Fryer et al. (2021))

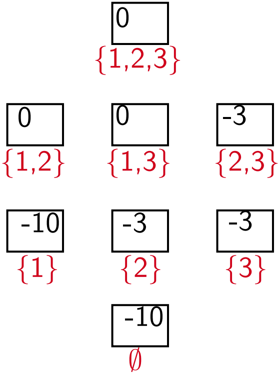


Example (secret holder, Fryer et al. (2021))



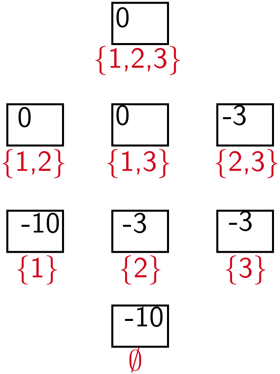
- 2 and 3 are symmetric.

Example (secret holder, Fryer et al. (2021))



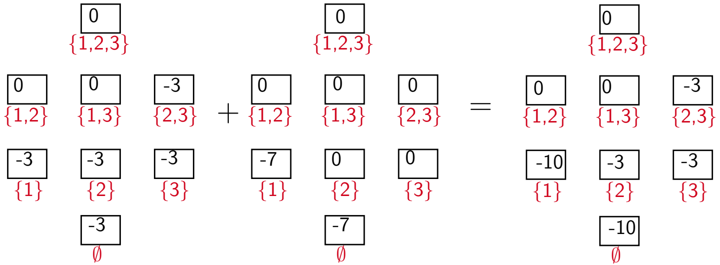
- 2 and 3 are symmetric.
- but are they more important than 1?

Example (secret holder, Fryer et al. (2021))



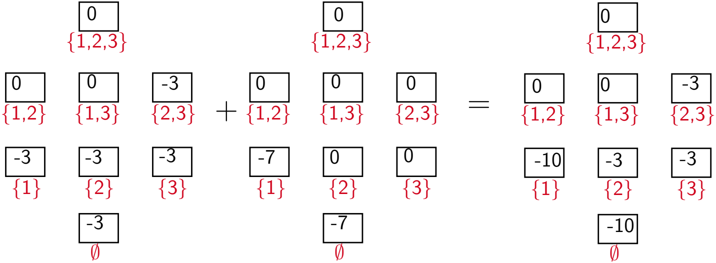
- 2 and 3 are symmetric.
- but are they more important than 1?
- Notice that 2 and 3 are "redundant features", while 1 is necessary to get the optimal prediction performance $v(N)$.

Additivity doesn't work...



$\text{Sh}=\left(2, \frac{1}{2}, \frac{1}{2}\right)$ $\quad$ $\text{Sh}=\left(0, \frac{7}{2}, \frac{7}{2}\right)$ $\quad$ $\text{Sh}= (2, 4, 4)$

Additivity doesn't work...



$\text{Sh}=\left(2, \frac{1}{2}, \frac{1}{2}\right)$ $\qquad$ $\text{Sh}=\left(0, \frac{7}{2}, \frac{7}{2}\right)$ $\qquad$ $\text{Sh}= (2, 4, 4)$

- no compelling reason to linearly combine the (opposite) critical roles played by features in the two game on the left

New properties

Definition (Coalitional Anonymity)

Let $i, j \in N$, $v, v' \in \mathcal{E}^N$ and a bijection π on $2^{N \setminus \{i,j\}}$ be such that, for all $S, T \in 2^{N \setminus \{i,j\}}$

$$v(S \cup \{i\}) \geq v(T \cup \{j\}) \Leftrightarrow v'(\pi(S) \cup \{i\}) \geq v'(T \cup \{j\}). \tag{3}$$

A ranking solution $R : \mathcal{E}^N \to \mathcal{R}(N)$ satisfies the *coalitional anonymity* property if it holds that

$$i \ R^v \ j \Leftrightarrow i \ R^{v'} \ j.$$

Definition (Independence from the Worst Set (IWS))

We say that a ranking solution $R : \mathcal{E}^N \to \mathcal{R}(N)$ satisfies the property of *independence from the worst set* if for any evaluation function $v \in \mathcal{E}^N$ such that coalitions in 2^N are partitioned into equivalence classes

$$\Sigma_1^v > \Sigma_2^v > \cdots > \Sigma_m^v$$

with $m \geq 2$, and $i, j \in N$ such that $iP^v j$, then it holds $iP^{v'} j$ for any evaluation function $v' \in \mathcal{E}^N$ such that coalitions in 2^N are partitioned into equivalence classes

$$\Sigma_1^{v'} > \Sigma_2^{v'} > \cdots > \Sigma_{m-1}^{v'} > \Sigma_m^{v'} > \cdots > \Sigma_p^{v'},$$

with $\Sigma_k^v = \Sigma_k^{v'}$ for all $k = 1, \ldots, m-1$.

Coalitional anonymity

- Only the position in the coalitional ranking matters (and not the composition of coalitions)

Coalitional anonymity

- Only the position in the coalitional ranking matters (and not the composition of coalitions)
- For instance, the ranking between features i and j based on an evaluation function v

$$\ldots \; v(i,k) = v(j,k) > v(i) = v(j)$$

should be as in v' with

$$\ldots \; v'(i) = v'(j,k) > v'(i,k) = v'(j)$$

Independence from the worst set

- A strict ranking is not affected by a modification of the ranking of worst coalitions.

Independence from the worst set

- A strict ranking is not affected by a modification of the ranking of worst coalitions.
- For instance, if one decides that a feature i should be ranked strictly better than a feature j in

$$\ldots \; v(i,k) > v(j,k) > v(i) = v(j)$$

i should be ranked strictly better than j also in v' with

$$\ldots \; v'(i,k) > v'(j,k) > {\color{red} v'(j) > v'(i)}$$

Sym and StDes plus CA and IWS = lex-cel

- Coalitional Anonymity: if there is no a priori assumption on the number of features to be selected, the size of coalitions fulfilling a certain level of prediction performance should not influence the relevance ranking of features.
- Independence from the Worst Set: when a decision is taken on whether selecting feature i or j first, a change affecting coalitions with the smallest performance in predictions has no impact on the decision.

Sym and StDes plus CA and IWS = lex-cel

- Coalitional Anonymity: if there is no a priori assumption on the number of features to be selected, the size of coalitions fulfilling a certain level of prediction performance should not influence the relevance ranking of features.
- Independence from the Worst Set: when a decision is taken on whether selecting feature i or j first, a change affecting coalitions with the smallest performance in predictions has no impact on the decision.

Theorem (based on Aleandri et al. (2024))

Lex-cel is the unique ranking solution fulfilling properties of symmetry, strict desirability, coalitional anonymity and independence from the worst set.

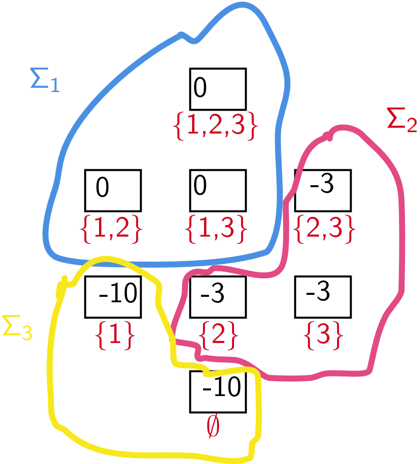
Lex-cel

- Consider and e.f. v and partition coalitions in equivalence classes arranged in descending order according to v

$$\Sigma_1^v > \Sigma_2^v > \ldots > \Sigma_m^v$$

- We denote as i_k^v the number of sets in Σ_k^v that contain the element i, with $k = 1, \ldots, m$.
- Let $\boldsymbol{\theta}^v(i)$ be the m-dimensional vector $\boldsymbol{\theta}^v(i) = (i_1^v, \ldots, i_m^v)$ associated with v.
- The lex-cel ranking solution (Bernardi et al. (2019)) is the map $R_{le} : \mathcal{E}^N \to \mathcal{R}(N)$ such that $i \ R_{le}^v \ j \Longleftrightarrow \boldsymbol{\theta}^v(i) \quad \geq_L \quad \boldsymbol{\theta}^v(j)$ for any $v \in \mathcal{E}^N$ and $i, j \in N$.

Example

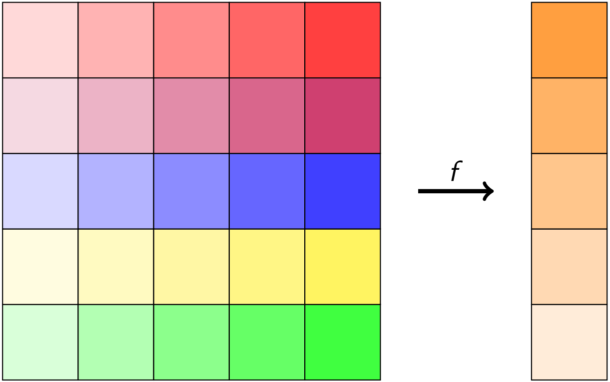


- $\boldsymbol{\theta}^v(1) = (3, 0, 1)$, $\boldsymbol{\theta}^v(2) = \boldsymbol{\theta}^v(3) = (2, 2, 0)$
- So the lex-cel ranking is: $1 \; P_{le} \; 2 \; I_{le} \; 3$

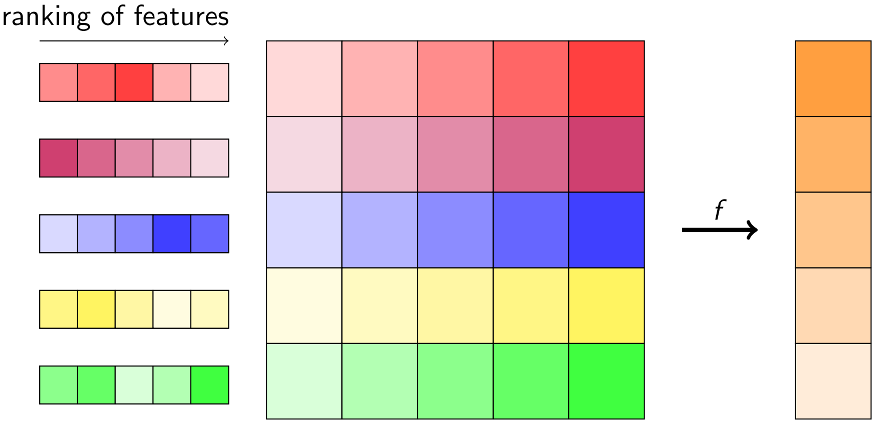
leXAI *vs.* SHAP

- $leXAI(v, k)$ and $SHAP(v, k)$ select the first k features according to lex-cel and the Shapley value on v. respectively
- we conducted computational experiments on public datasets and compared errors produced by the selected features according to the two methods.

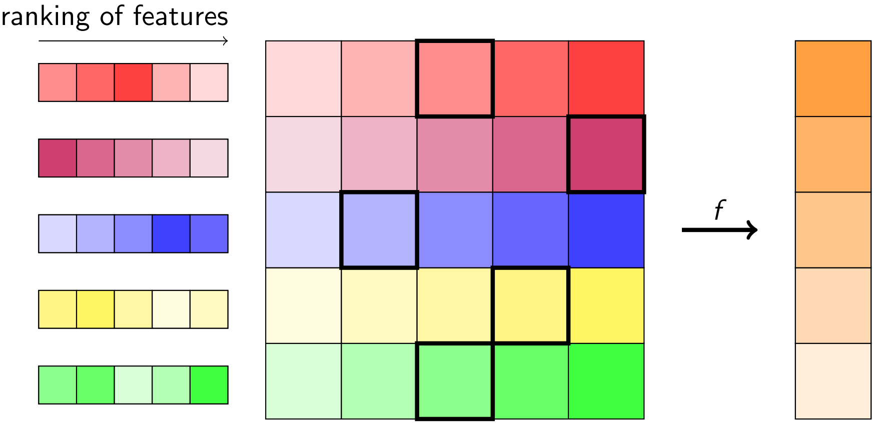
Experiments



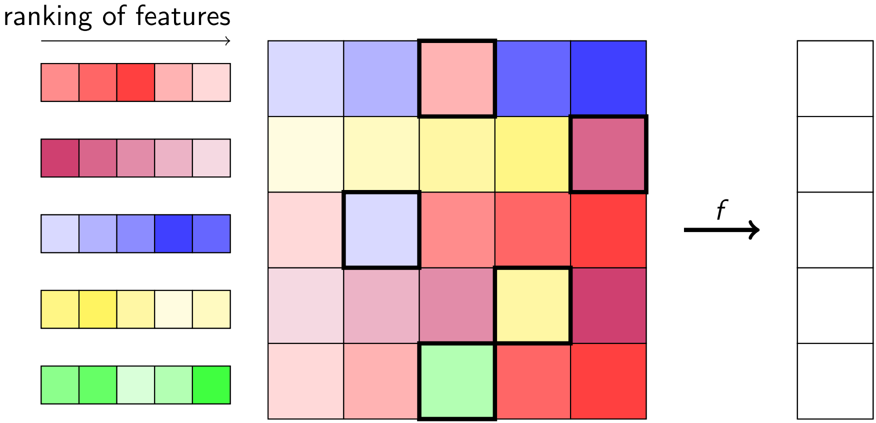
Experiments



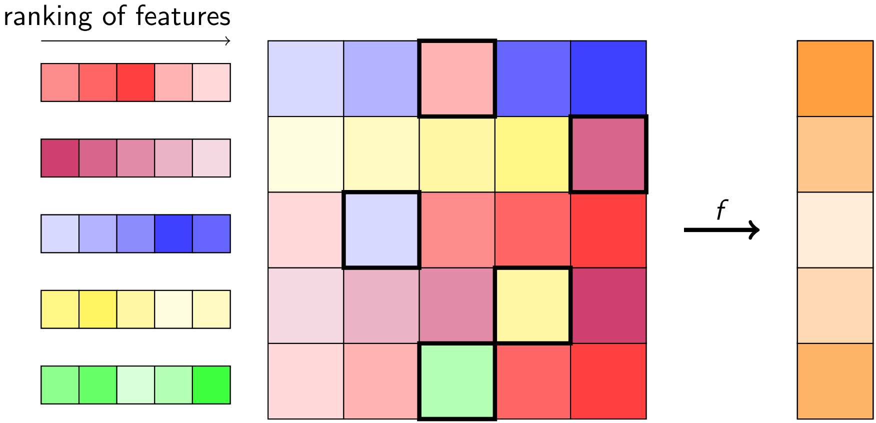
Experiments



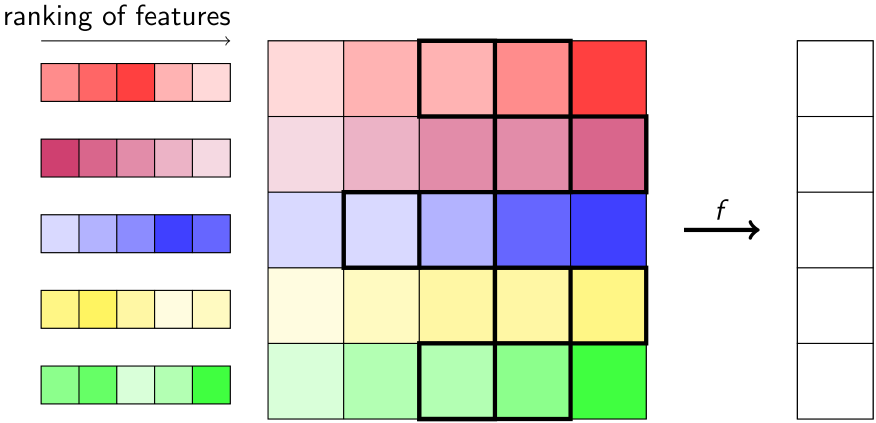
Experiments



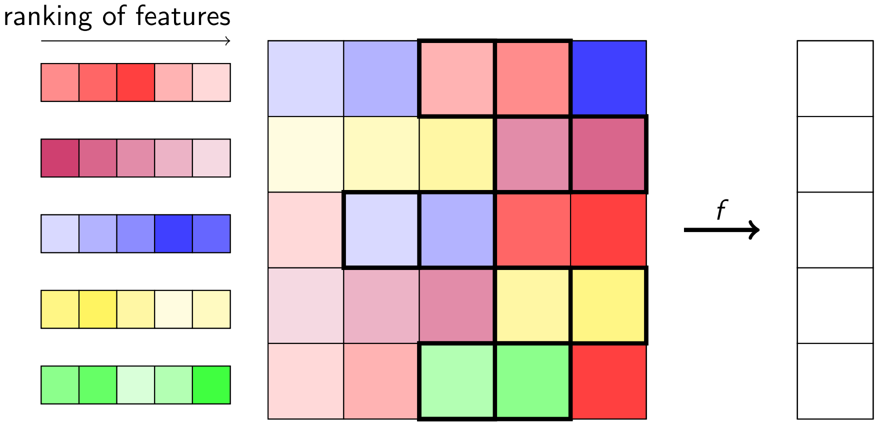
Experiments



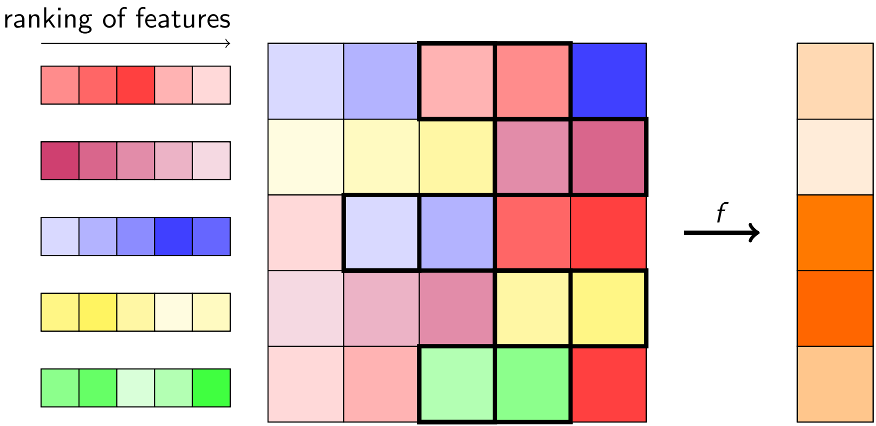
Experiments



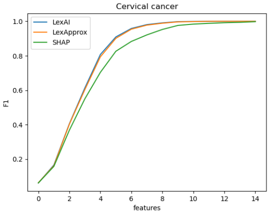
Experiments

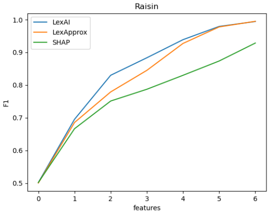


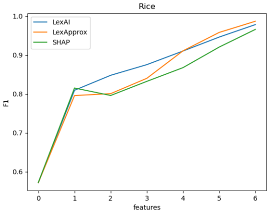
Experiments

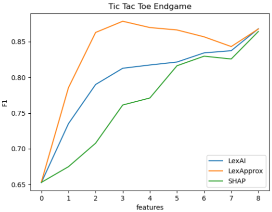


Classification task

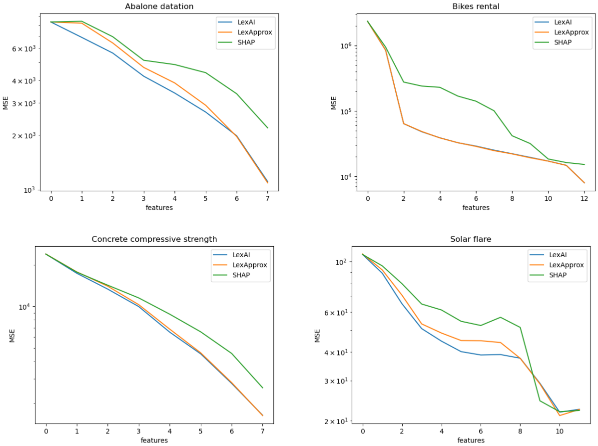








Regression task



Computation time

name	$\|M\|$	$\|N\|$	LeXAI	LeXAI approx	SHAP
Cervical	858	15	1325.0	1.0	707.8
Raisin	900	7	4.7	0.6	59.0
Rice	3810	7	19.5	2.3	250.0
Tic Tac Toe	957	9	19.9	0.7	137.8

name	$\|M\|$	$\|N\|$	LeXAI	LeXAI approx	SHAP
bike	731	13	287.0	0.7	587.1
abalone	4177	8	42.9	2.9	482.3
flare	322	12	49.1	0.2	247.1
concrete	1030	8	10.4	0.7	90.8

An attempt to approximate lex-cel

Proposition

Let $v \in \mathcal{E}^N$ be a monotonic evaluation function such that $v(N \setminus \{i\}) \neq v(N \setminus \{j\})$ for all $i, j \in N$ with $i \neq j$. Then,

$$i \; P_{le}^{v} \; j \Leftrightarrow v(N \setminus \{j\}) > v(N \setminus \{i\}) \tag{4}$$

for all $i, j \in N$.

An attempt to approximate lex-cel

Proposition

Let $v \in \mathcal{E}^N$ be a monotonic evaluation function such that $v(N \setminus \{i\}) \neq v(N \setminus \{j\})$ for all $i, j \in N$ with $i \neq j$. Then,

$$i \; P^v_{le} \; j \Leftrightarrow v(N \setminus \{j\}) > v(N \setminus \{i\}) \tag{4}$$

for all $i, j \in N$.

Notice that we can rewrite condition (4) in Proposition 3.1 for any $i, j \in N$ as the equivalent one

$$i \; P^v_{le} \; j \Leftrightarrow M_i(v) > M_j(v)$$

where $M_i(v) = v(N) - v(N \setminus \{i\})$ and $M_j(v) = v(N) - v(N \setminus \{j\})$ (known as the *marginal index* Owen (2013); Hwang and Liao (2010))

Conclusions

- to control the selection of redundant/unnecessary features, don't use the Shapley value (or other methods satisfying additivity)

Conclusions

- to control the selection of redundant/unnecessary features, don't use the Shapley value (or other methods satisfying additivity)
- to simplify the model maintaining high prediction quality, features' excellence should be awarded

Conclusions

- to control the selection of redundant/unnecessary features, don't use the Shapley value (or other methods satisfying additivity)
- to simplify the model maintaining high prediction quality, features' excellence should be awarded
- lex-cel can be an option

Further works

- Handling a large number of features
- Global explanation
- Specialization for some families of models
- quantify the relevance
- add compensations

Further works

- Handling a large number of features
- Global explanation
- Specialization for some families of models
- quantify the relevance
- add compensations

Thank you!

Michele Aleandri, Felix Fritz, and Stefano Moretti. Desirability and social rankings. *arXiv preprint arXiv:2404.18755*, 2024.

Giulia Bernardi, Roberto Lucchetti, and Stefano Moretti. Ranking objects from a preference relation over their subsets. *Social Choice and Welfare*, 52(4):589–606, 2019. doi: 10.1007/s00355-018-1161-1. URL https://hal.archives-ouvertes.fr/hal-02191137.

Shay Cohen, Eytan Ruppin, and Gideon Dror. Feature selection based on the shapley value. *other words*, 1(98Eqr):155, 2005.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021.

Daniel Fryer, Inga Strümke, and Hien Nguyen. Shapley values for feature selection: The good, the bad, and the axioms. *Ieee Access*, 9:144352–144360, 2021.

Yan-An Hwang and Yu-Hsien Liao. Consistency and dynamic approach of indexes. *Social Choice and Welfare*, 34(4):679–694, 2010.

Osnat Israeli. A shapley-based decomposition of the r-square of a linear regression. *The Journal of Economic Inequality*, 5: 199–212, 2007.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

Guillermo Owen. *Game theory*. Emerald Group Publishing, 2013.

L.S. Shapley. A value for n-person games. In Kuhn H. and Tucker A.W., editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, 1953.