Explainability of stacking model for prediction of corporate CO2 emissions

NGUEMKAM TEBOU Ingrid Pamela [1]

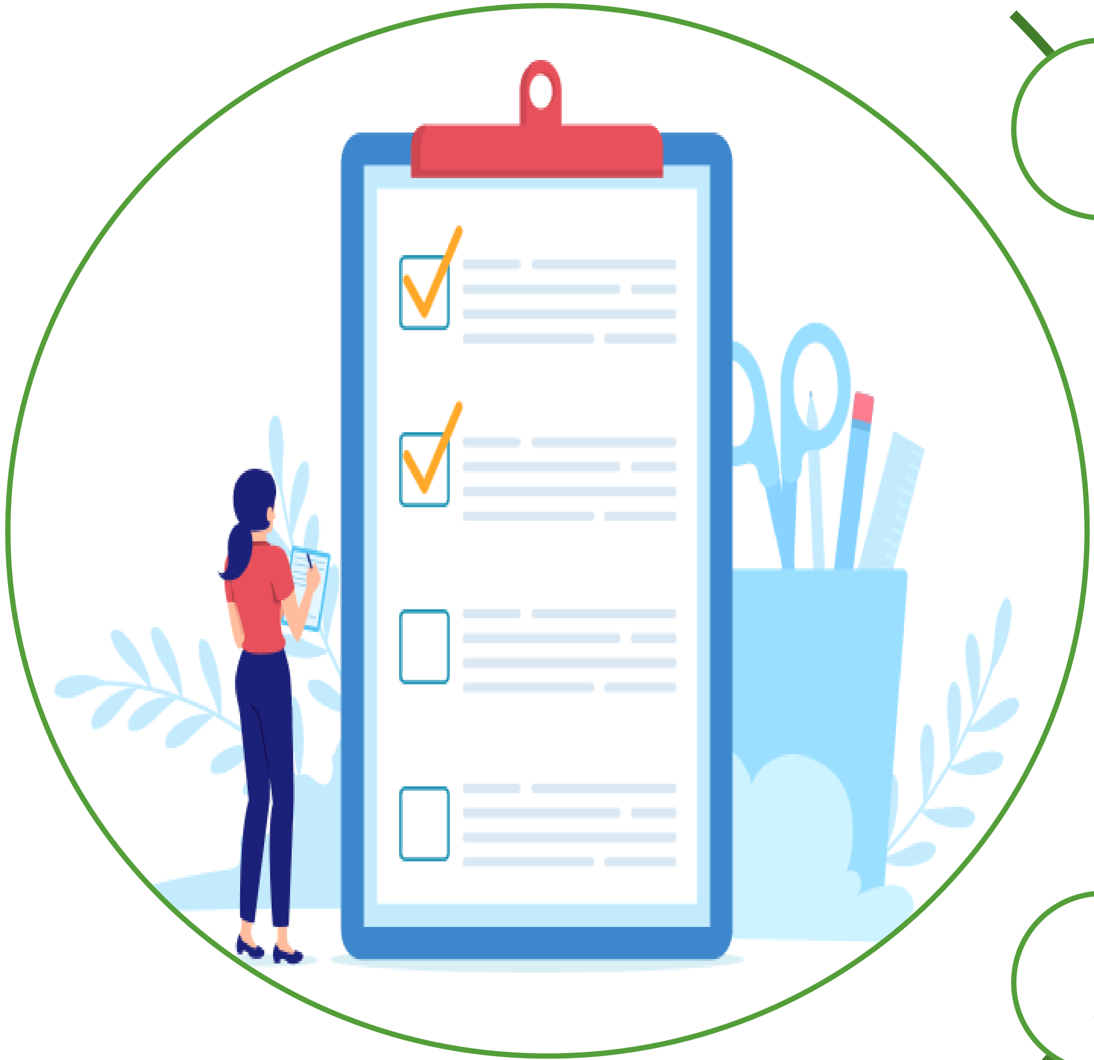
TSOPZE Norbert [1]

TCHUENTE Dieudonné [2]

[1] University of yaoundé 1, Yaoundé, Cameroon
[2] Toulouse Business School, Toulouse, France

December *2024*

Plan



1 Context and definitions

2 State of the art

3 Our approach

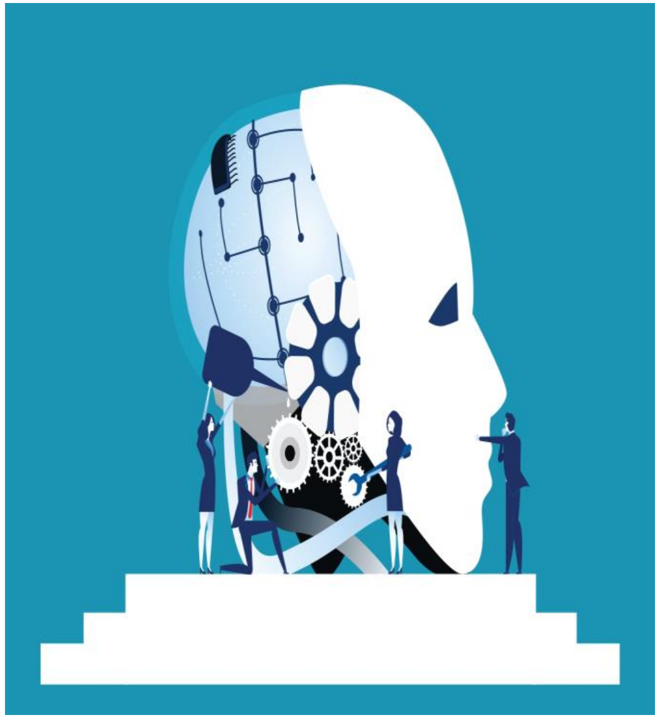
4 Experiments

5 Conclusion and perspectives

Context

Artificial Intelligence (AI)

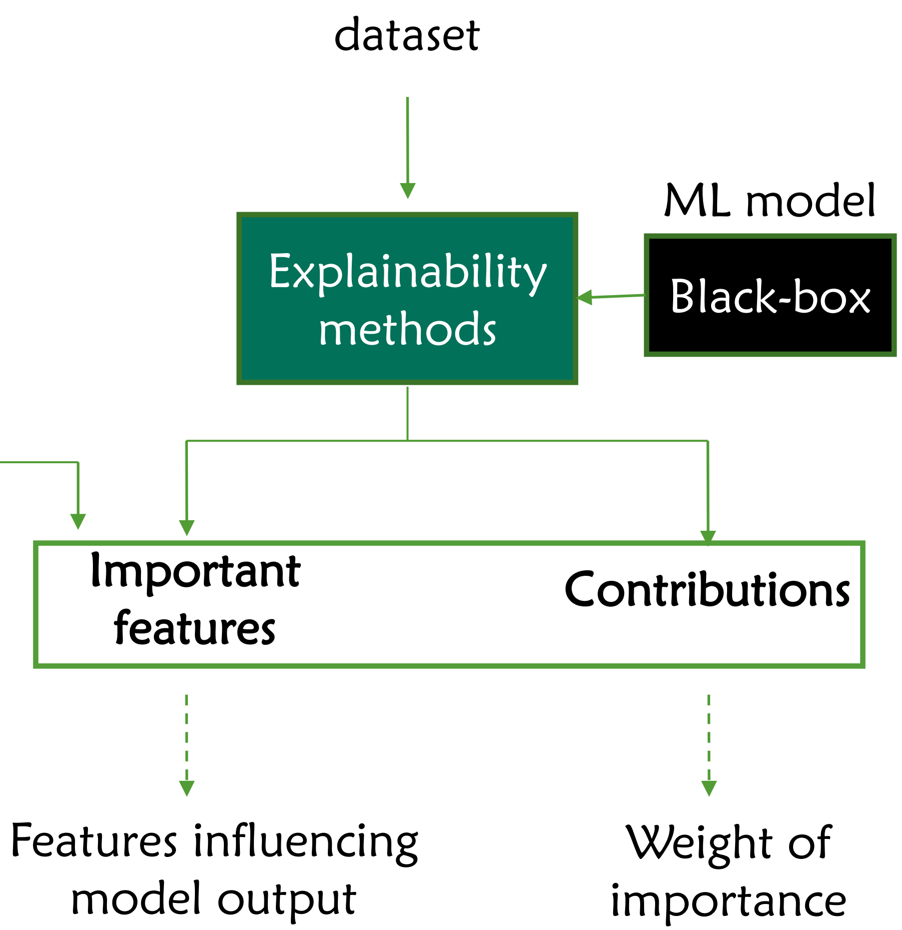
- Artificial Intelligence widely used
- Efficient Machine Learning models but bigger complexity
- Prediction logic of model's output dificult to unsderstand by users
- Problem of trustworthy



Explainable Artificial Intelligence (XAI)

Definitions

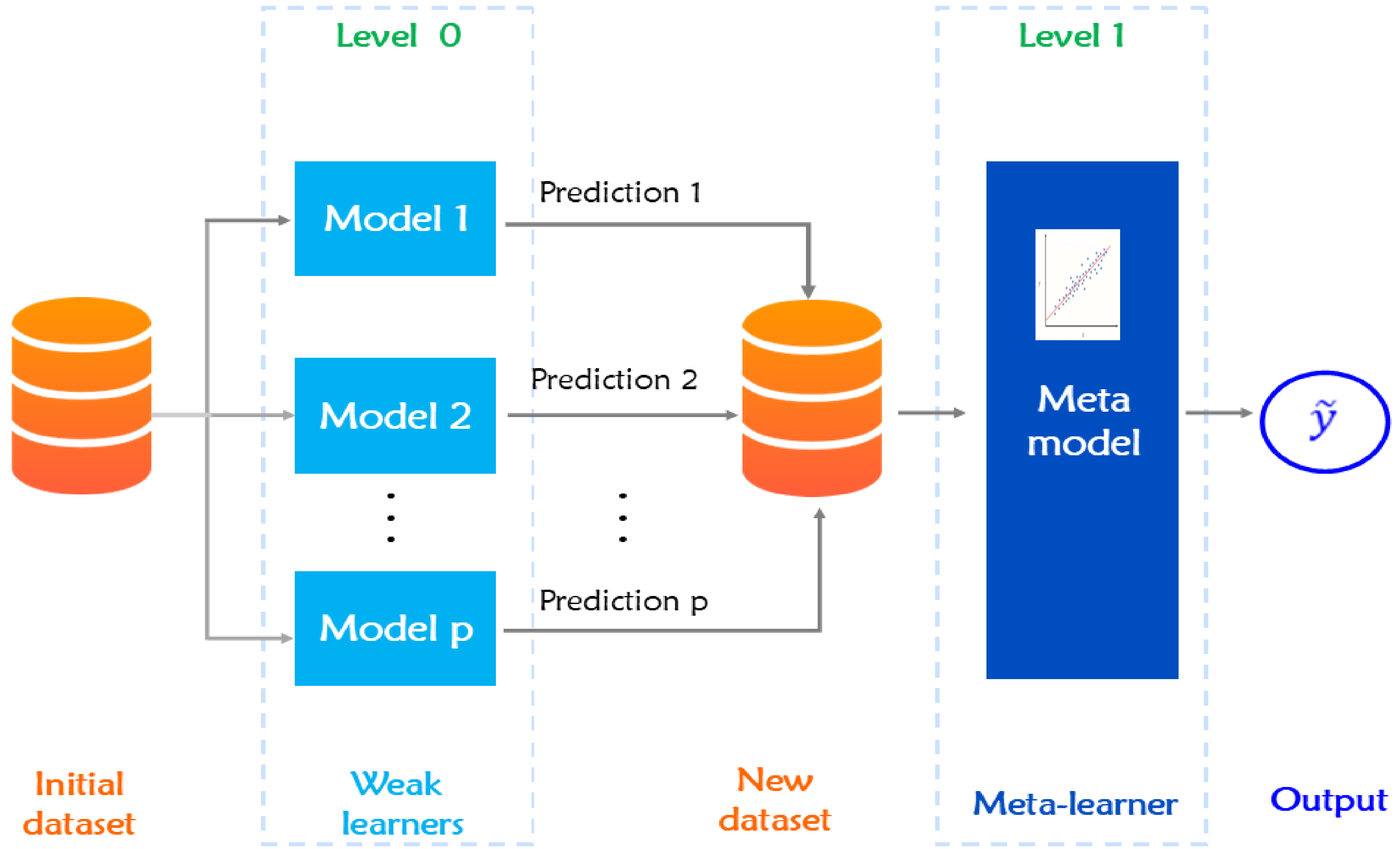
- Artificial Intelligence
  - Explainable Artificial Intelligence
    - Explainability
    - Explanation
      - Global explanation
      - Local explanation
    - Explain a prediction
    - Feature based explainability



State of art: Explainability methods

Authors	Principle	Method	Advantages	Limits
Bach et al. (2015)	Redistribute the prediction starting from the output layer of the network and back-propagating to the input layer	**LRP**	- Conservation of contributions, - Stable explanation	Limited to neural networks
Ribeiro et al. (2016)	Locally approximate the blackbox model by a simpler model using the neighbourhood of the instance to be explained	**LIME**	- Simple, - Can be applied to any model	- Quality of explanation depends on the choice of neighbourhood, - Explanation may be unstable
Lundberg et Lee. (2017)	Using Shapley values as variable contributions by transposing game theory to machine learning	**SHAP**	- Fair distribution of contributions, - Can be applied to any model,	- Long execution time, - Explanation may be unstable

Stacked Generalization Model (stacking)



Two levels :

1- Base learners ∼ individual predictions

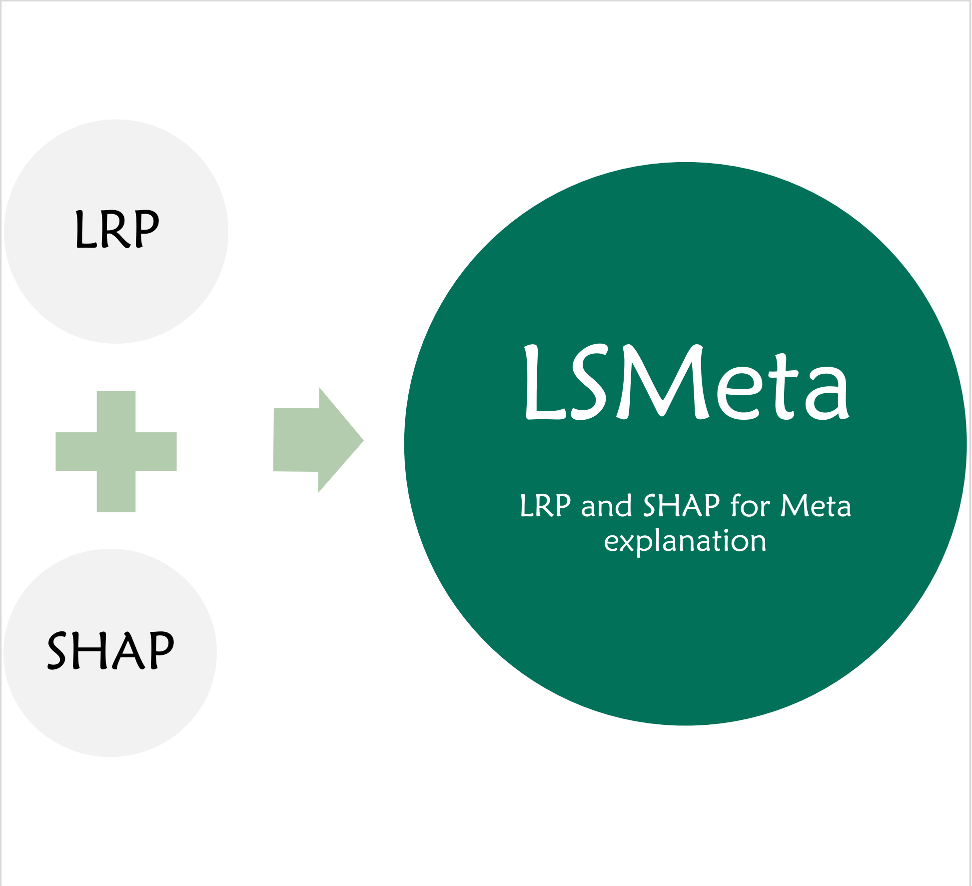
2- Meta-learner ∼ combining predictions

Research question

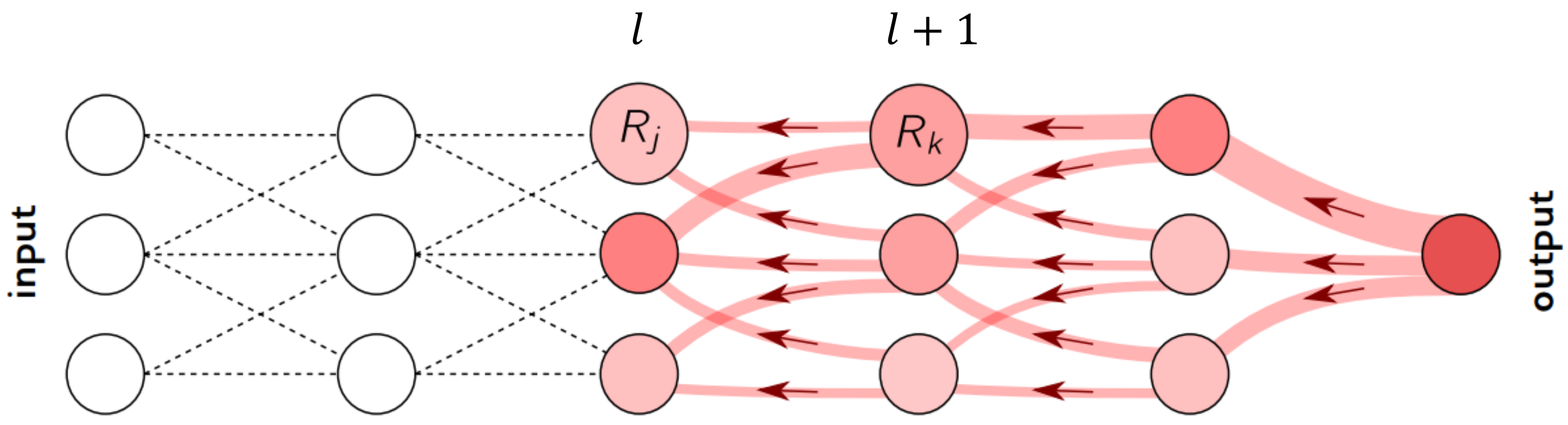
Taking into account the contributions of all the base learners could provide better explanations

Given an output $\tilde{y}$ provided by the meta-learner on an input x, how to provide contributions of the attributes of x on the calculation of $\tilde{y}$, taking into account the contributions of all the base learners ?

Proposition



Layerwise Relevance Propagation-LRP



Source – G. Montavon et al, 2019

Propagation formula

$$R_{j\leftarrow k}^{(l,l+1)} = R_k^{(l+1)} \cdot \frac{a_j w_{jk}}{\sum_h a_h w_{hk}}$$

With

$$\begin{cases} a_{j=output\ of\ neuron\, j} \\ w_{jk=weight\ between\ neurons\, j\, and\ k} \\ R_{k=Relevance\ of\ neuron\ k} \\ l=layer\ of\ neuron\ j \\ l+1=layer\ of\ neuron\ k \end{cases}$$

SHapley Additive exPlanations - SHAP

- Transposing game theory into machine learning
- In game theory : shapley value is the solution to fairly distribute a payoff among players
- Game = explain prediction, players = variables, payoff = predicted value
- Shapley values as variable contributions

$$\varphi_{val}(i) = \sum_{S \subseteq F\{i\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \Delta_{val}(i, S)$$

With

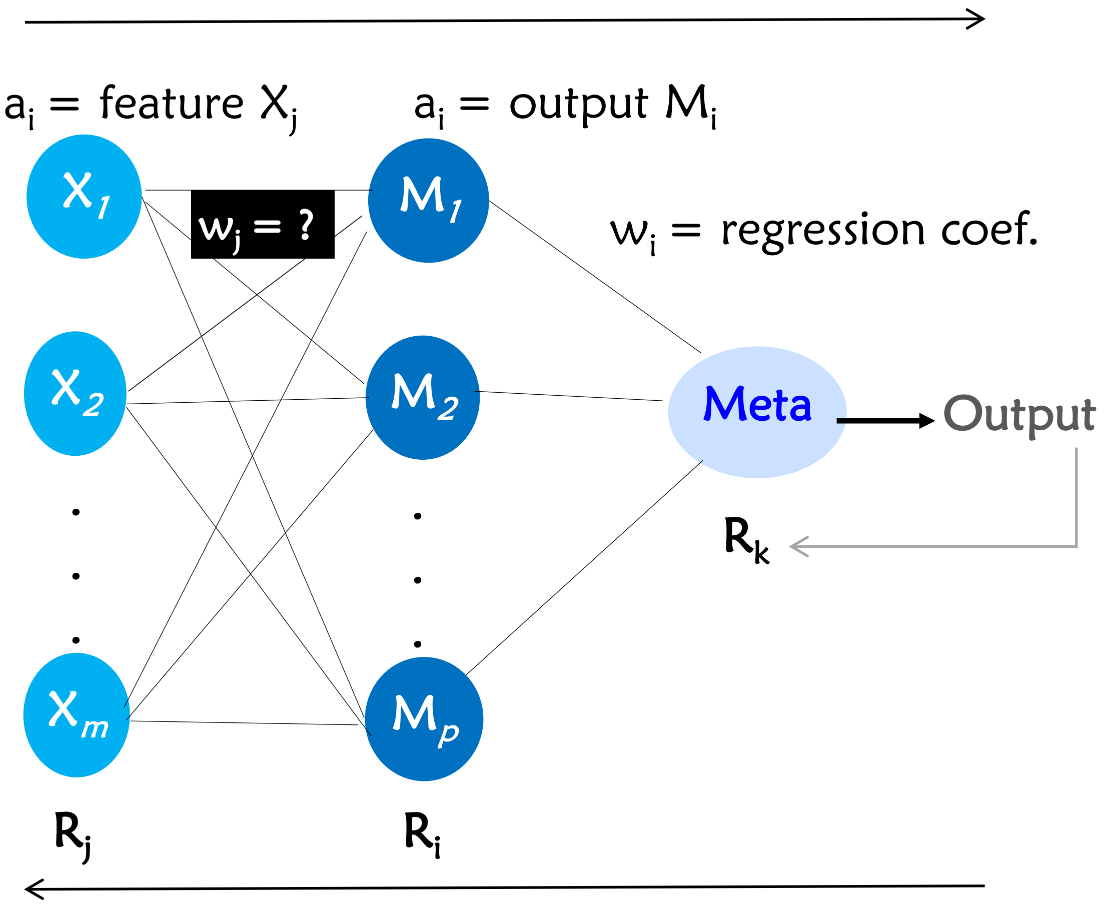
$$\begin{cases} \varphi_{val}(i) = Shapley\ values \\ S = coalition \\ p = number\ of features \end{cases}$$

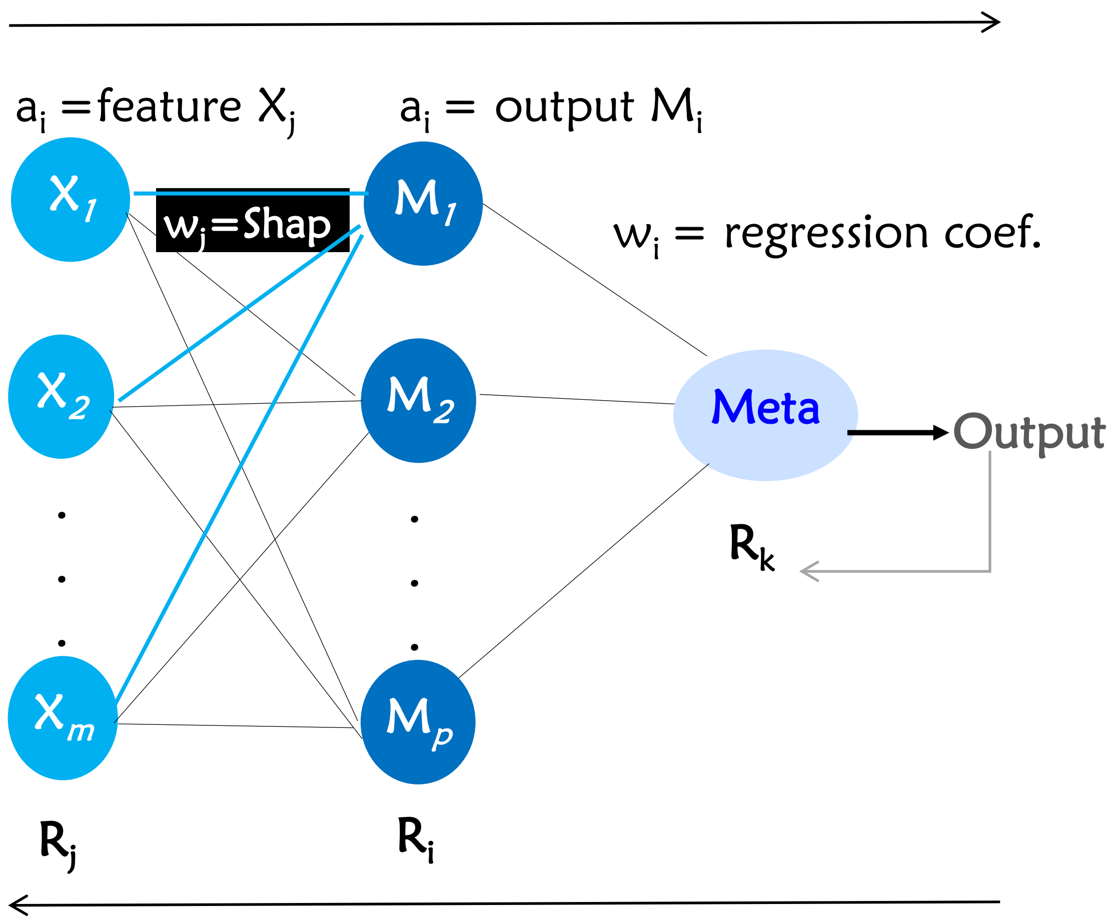
$$\Delta_{val}(i, S) = val(S \cup \{i\}) - val(S)$$

Our approach : Description 1/2

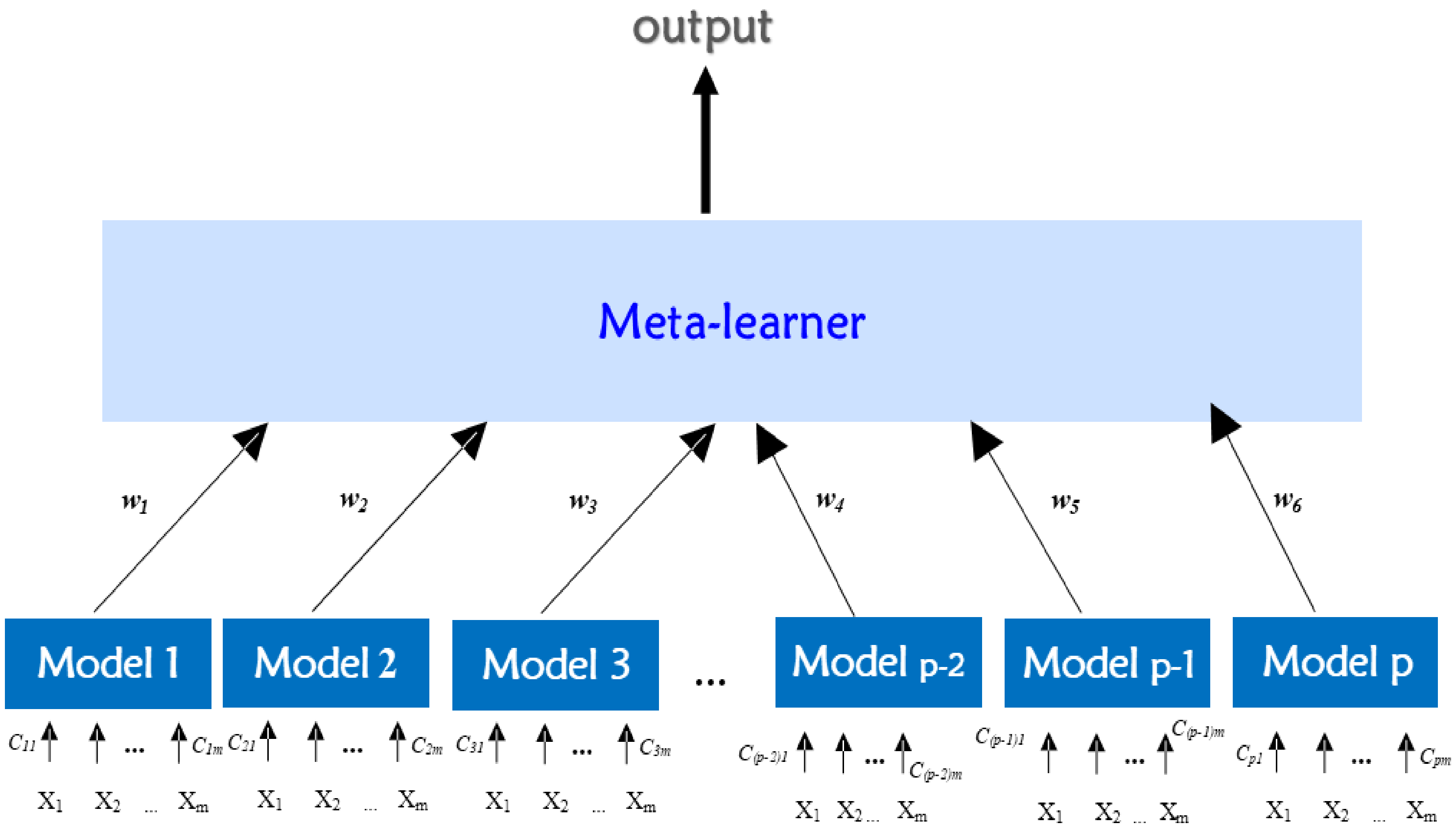
LRP
$$R_{i\leftarrow k}^{(l,l+1)} = R_k^{(l+1)} \cdot \frac{a_i w_{ik}}{\sum_h a_h w_{hk}}$$

a_j=output of neuron j
w_{jk}=weight between neurons j and k
R_k=Relevance of neuron k
l=layer of neuron j
$l+1$=layer of neuron k





Our approach : Description 2/2



Steps:

1- Contributions of base learners on prediction of meta-learner - LRP

2- Contributions of the features to those of the base learners - SHAP

3- Aggregation of contributions - LRP

Our approach : Properties

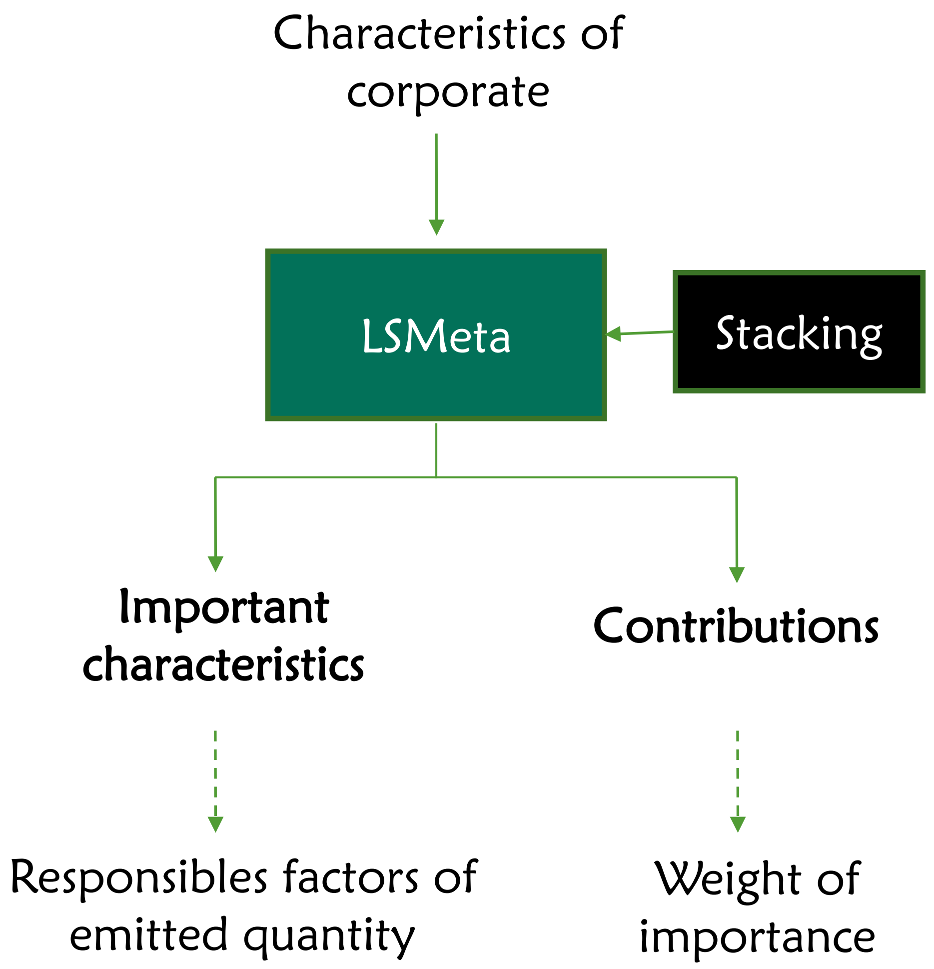
- Local and Global Explainability
- Specific to Stacking model
- Exponential complexity

$$\mathcal{O}(p.2^m) \text{ with } C_{SHAP} = \mathcal{O}(2^m)$$

Experiments

Dataset: Thompson Reuters ESG

dataset on the quantity of CO_2 emitted by some Corporate around the World based on their environmental, societal and Governance characteristics



Experiments

Dataset

Thompson Reuters ESG

Preprocessing (14531 x 113 $\longrightarrow$ 14531 x 53) : *MICE*

Stacked Generalization *(Nguyen et al., 2021)*

Base learners : OLS, Elastic Net, KNN, Random Forest, XGBoost et MLP

Meta-learner : mean, Elastic Net, OLS

Evaluation metrics

Prediction $MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \tilde{y}_i|$ $\sim\sim$ $R^2 = 1 - \frac{\sum(y_i - \tilde{y}_i)^2}{\sum(y_i - \bar{y})^2}$

	Mean of predictions	Meta-Elastic Net	Meta-OLS
MAE	0.161	0.158	**0.116**
R^2	0.758	0.739	**0.856**

LSMeta – Local Explanation

Example : Aggreko PLC

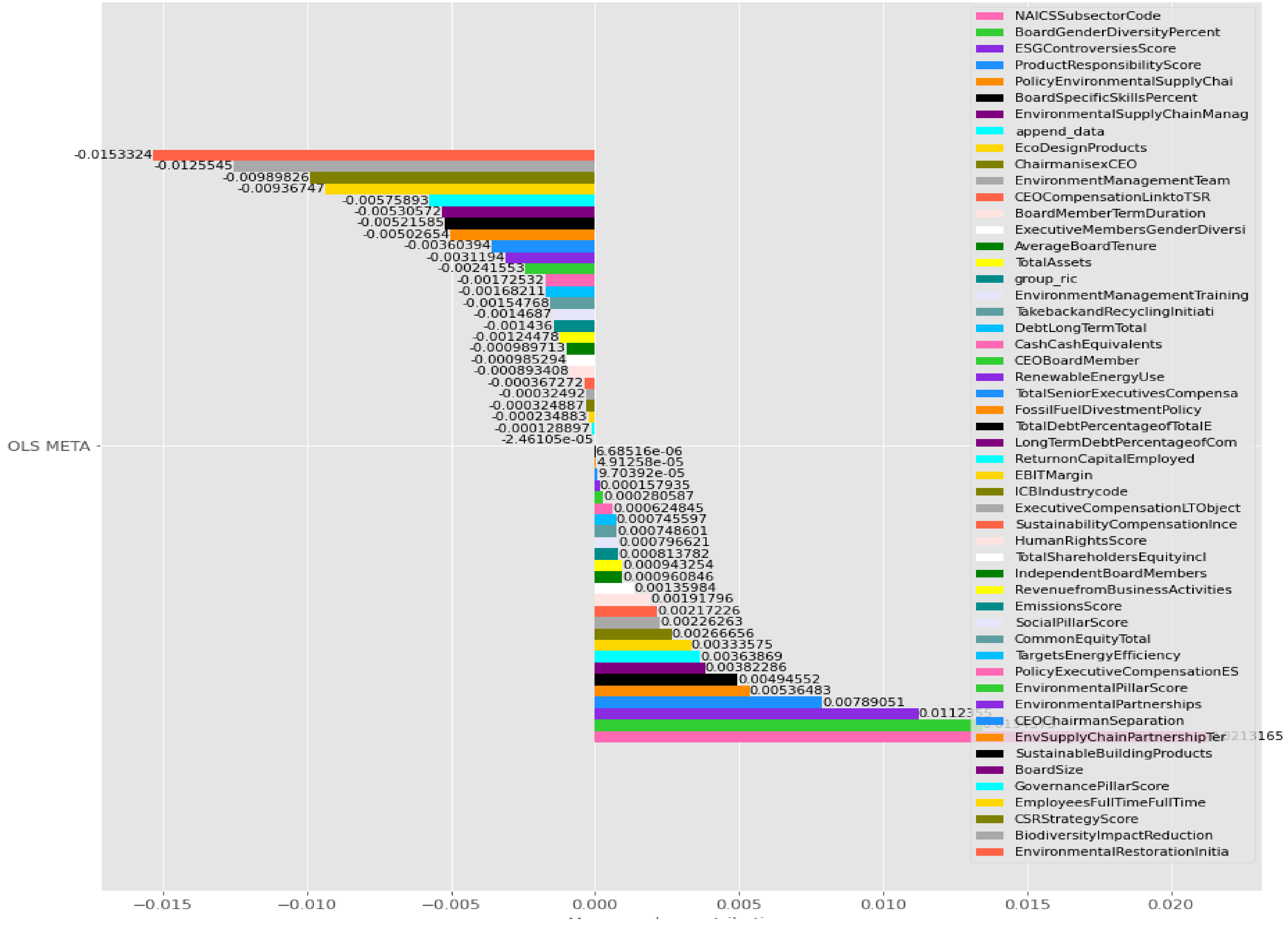
Features	Contribution
GovernancePillarScore	1,29E-03
PolicyExecutiveCompensationES	1,25E-03
CEOChairmanSeparation	6,83E-04
FossilFuelDivestmentPolicy	6,53E-04
CEOChairmanSeparation	5,97E-04

Table – Top-5 Features with positive *contributions*

Features	Contribution
ExecutiveMembersGenderDiversi	-1,54E-04
ESGControversiesScore	-1,60E-04
EnvironmentalSupplyChainManag	-2,32E-04
PolicyEnvironmentalSupplyCha	-2,89E-04
ChairmanisexCEO	-5,58E-04

Table – Top-5 Features *with negatives contributions*

LSMeta – Global Explanation



+

NAICSSubsectorCode, BoardGenderDiversityPercent , ESGControversiesScore, ProductResponsibilityScore, PolicyEnvironmentalSupplyChai

-

EnvironmentalRestorationInitia, BiodiversityImpactReduction, CSRStrategyScore, EmployeesFullTimeFullTime, GovernancePillarScore

Experiments

Evaluation metrics

Explainability

Fidelity score : $\frac{1}{n}\sum_{i=1}^{n} |\tilde{y}_i - \tilde{y}_{i\setminus A}|$

$\sim\sim$

Stability score: Variable Stability Index *(Visani et al., 2020)*

Evaluation – fidelity score

Evaluation process

1. Top-5 of most importants features
2. Average of fidelity scores

	SHAP	LSMeta
Average fidelity	0.216 (±0.742)	0.096 (±0.256)

Table – Average fidelity scores

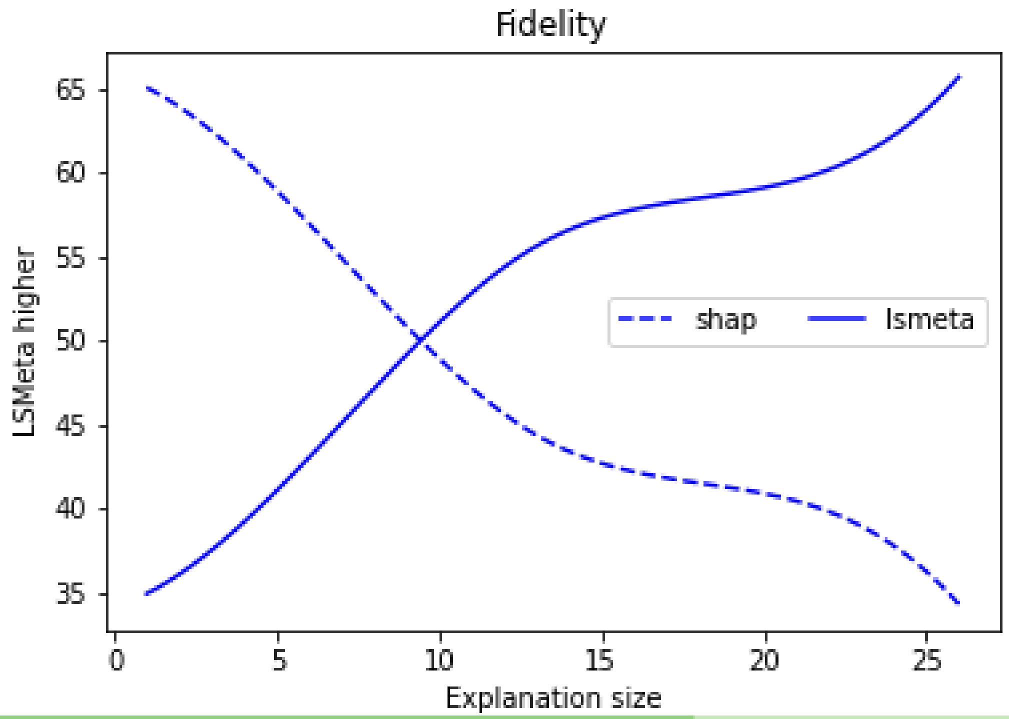
Evaluation- fidelity & stability

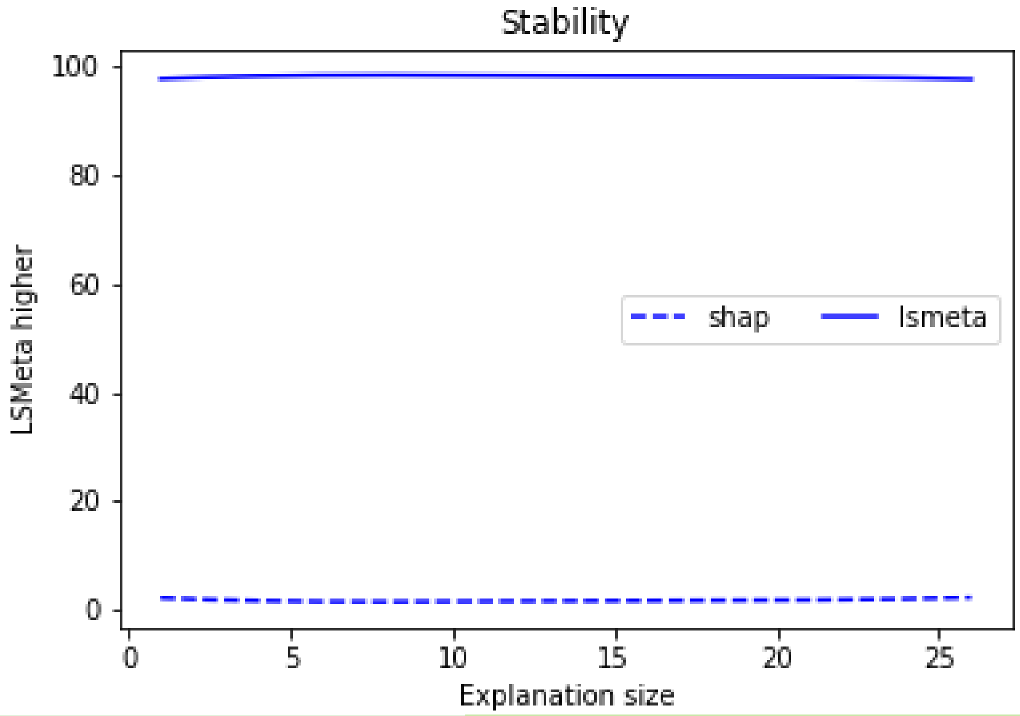
Evaluation process

1. Calculation of the fidelity/stability of each observation
2. Comparison of scores between SHAP and LSMeta
3. Counting the number of times LSMeta is bigger
4. Percentage

Number of important features	1	1/10	1/4	1/3	1/2
Fidelity LSMeta >= SHAP (%)	34.25	41.07	55.62	58.17	65.67
Stability LSMeta >= SHAP (%)	97.83	98.42	98.38	98.31	97.76

Table – Proportion of individual fidelity/stability scores





Conclusion and perspectives

- Explain prediction made by meta-learner
- Stacked Generalization Model and LSMeta (combining LRP and SHAP)
- Dataset of Corporate CO_2 emissions
- Stable explanations and faithful to model output
- Environmental implications

    - Identification of important factors of CO_2 emissions

    - Identification of high and low risk corporate

Perspectives

- Consult experts about the explanations provided
- Use other variants of LRP and SHAP
- Identify the types of problem/data suitable for LSMeta

Paper

Ingrid Pamela Nguemkam Tebou, Norbert Tsopze and Dieudonné Tchuente.

Explaining the predictions of a meta-learner: case of corporate CO2 emissions.

- *Accepted to Conférence de Recherche en Informatique, édition 2023*
- DOI :

  https://link.springer.com/chapter/10.1007/978-3-031-63110-8_8

References

- Matthew Brander and Gary Davis. Greenhouse gases, co2, co2e, and carbon : What do all these terms mean. *Econometrica, White Papers*, 2012
- Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifer decisions by layer-wise relevance propagation. *PloS one*, 10(7) :e0130140, 2015.
- Scott M Lundberg and Su-In Lee. A unifed approach to interpreting model predictions. Advances in neural information processing systems, 30, 2017
- Quyen Nguyen, Ivan Diaz-Rainey, and Duminda Kuruppuarachchi. Predicting corporate carbon footprints for climate finance risk analyses : a machine learning approach. Energy Economics, 95 :105129, 2021.
- Matthieu Bellucci, Nicolas Delestre, Nicolas Malandain, and Cecilia Zanni-Merk. Towards a terminology for a fully contextualized xai. *Procedia Computer Science*, 192 :241–250, 2021.
- Amina Adadi and Mohammed Berrada. Peeking inside the black-box : a survey on explainable artifcial intelligence (xai). *IEEE access*, 6 :52138–52160, 2018

Thank you

For your kind attention

End