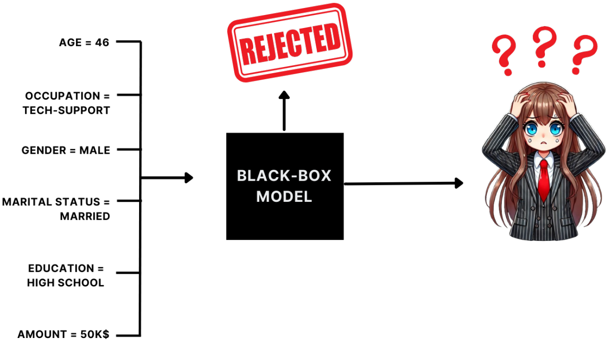
Affiner SHAP : Améliorer la stabilité grâce à la sélection de voisins en couches

Gwladys Kelodjou[1], Laurence Rozé[2], Véronique Masson[1], Luis Galárraga[1], Romaric Gaudel[1], Maurice Tchuente[3], Alexandre Termier[1]

[1]Univ Rennes, Inria, CNRS, IRISA - UMR 6074
[2]Univ Rennes, INSA Rennes, CNRS, Inria, IRISA - UMR 6074
[3]Sorbonne University, IRD, University of Yaoundé I, UMI 209 UMMISCO

CAp 2024 - AAAI 2024

Why should I trust you?



AGE = 46

OCCUPATION = TECH-SUPPORT

GENDER = MALE

MARITAL STATUS = MARRIED

EDUCATION = HIGH SCHOOL

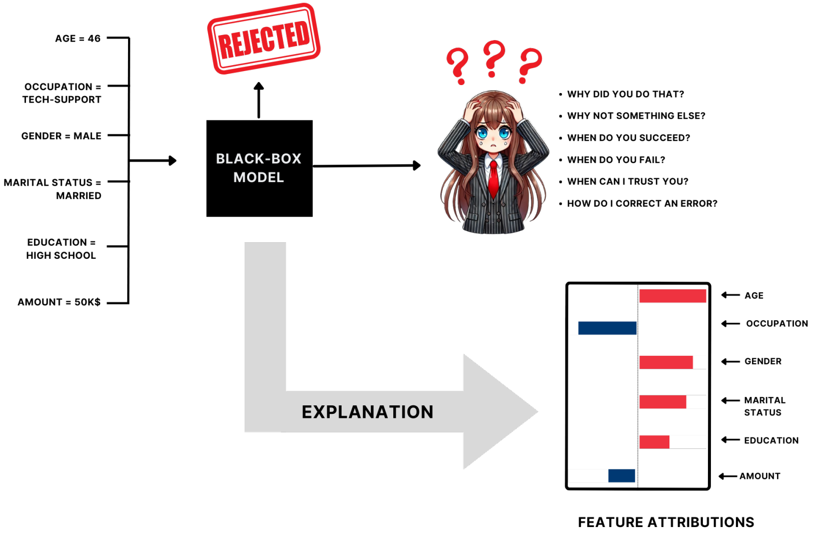
AMOUNT = 50K$

BLACK-BOX MODEL

REJECTED

- WHY DID YOU DO THAT?
- WHY NOT SOMETHING ELSE?
- WHEN DO YOU SUCCEED?
- WHEN DO YOU FAIL?
- WHEN CAN I TRUST YOU?
- HOW DO I CORRECT AN ERROR?

Local post-hoc explainability



AGE = 46

OCCUPATION = TECH-SUPPORT

GENDER = MALE

MARITAL STATUS = MARRIED

EDUCATION = HIGH SCHOOL

AMOUNT = 50K$

REJECTED

BLACK-BOX MODEL

- WHY DID YOU DO THAT?
- WHY NOT SOMETHING ELSE?
- WHEN DO YOU SUCCEED?
- WHEN DO YOU FAIL?
- WHEN CAN I TRUST YOU?
- HOW DO I CORRECT AN ERROR?

EXPLANATION

FEATURE ATTRIBUTIONS

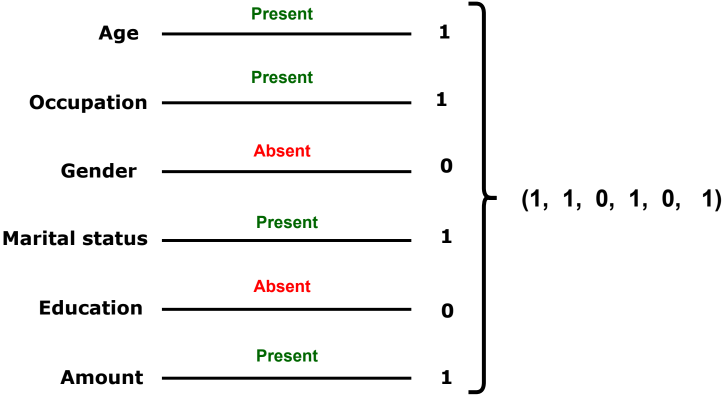
SHAP value

- **Shapley value :**
  - Fairly distribute the gain obtained by multiple players collaborating in a game.
  - Considers different coalitions to determine each player's contribution.
- **SHAP (SHapley Additive exPlanations)[1] :** Applies Shapley value to determine how much each feature contributed to the model's decision.
  - Exact computation of SHAP values is challenging.
- **Kernel SHAP :** Model-agnostic approximation of SHAP values using linear regression.

[1] Lundberg, S. M. ; and Lee, S.-I. **«A unified approach to interpreting model predictions»**. NIPS 2017

Background
Coalition

Coalition : subset of features.



Age — Present — 1

Occupation — Present — 1

Gender — Absent — 0

Marital status — Present — 1

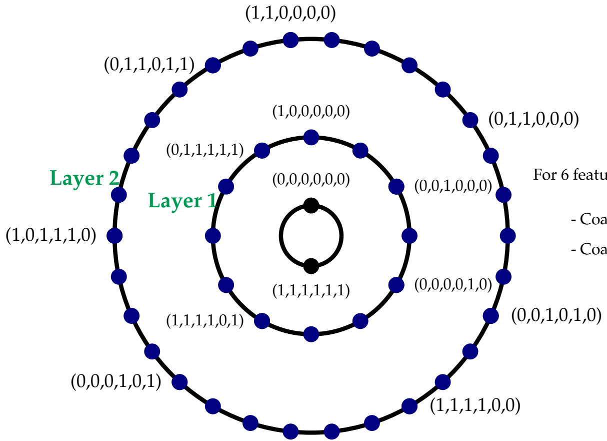
Education — Absent — 0

Amount — Present — 1

(1, 1, 0, 1, 0, 1)

A coalition is also referred to as a **neighbor**.

Layer-wise neighbor selection



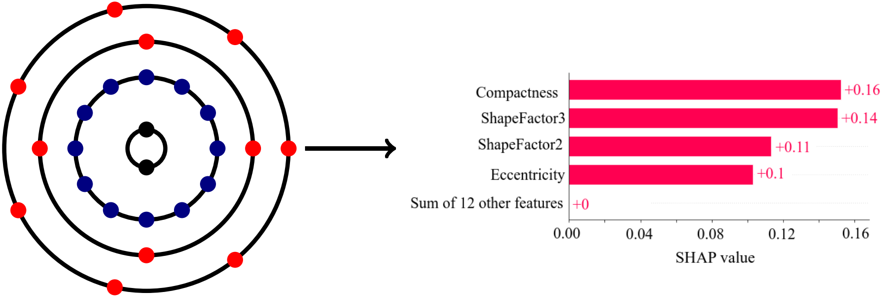
For 6 features, **stratum 2** contains **42 coalitions** :

- Coalitions of layer 1 (**12**)
- Coalitions of layer 2 (**30**)

- **Layer** : set of coalitions sharing the same number of present or absent features.
- **Stratum** : cumulative set of complete layers.

Kernel SHAP

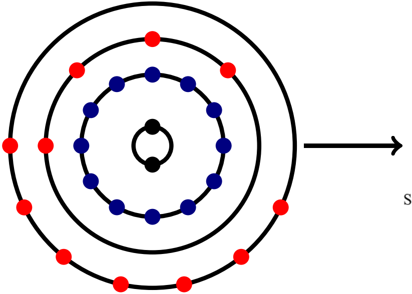
Budget : determines the number of coalitions to use.

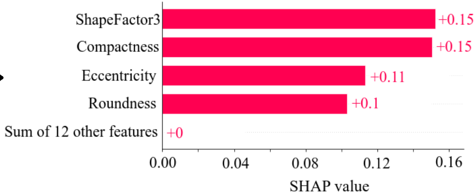


- First generate coalitions from lower layers.
- Randomly samples from subsequent layers if the budget is not exhausted.

Kernel SHAP suffers from stability issues

Different executions lead to various explanations.





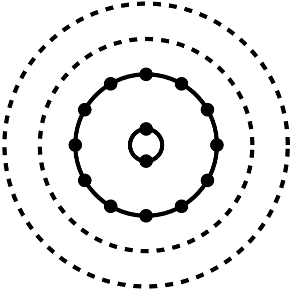
Kernel SHAP suffers from stability issues

Stability : The ability to reproduce the same explanation.

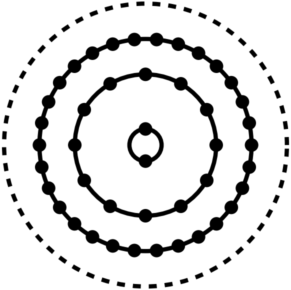
Contribution 1 : Neighbor selection

Achieving Kernel SHAP's stability : ST-SHAP

Set the budget to consider only complete stratums.



Stratum 1 full

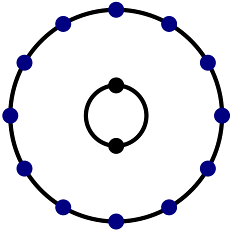


Stratum 2 full

Summary of experimental results

Complete stratums lead to **stable** and **high-quality** explanations.

Contribution 2 : Stratum 1 attributions



Use only Stratum 1 coalitions

Attribution values with Stratum 1

For any feature $j \in N = \{1, \cdots, M\}$ (the set of all features), attribution ϕ_j with Stratum 1 is :

$$\phi_j = \tilde{\phi}_j + \frac{1}{M}\left(f(N) - f(\emptyset) - \sum_{i=1}^{M} \tilde{\phi}_i\right)$$

where for any i, $\tilde{\phi}_i = \frac{f(\{i\}) - f(\emptyset) + f(N) - f(N \setminus \{i\})}{2}$ and M the number of features.

First stratum Attribution Analysis
Properties

Stratum 1 Attribution Properties

- **LES family**[1,2] :
  - **L**inearity
  - **E**fficiency
  - **S**ymmetry
- Missingness
- **Execution time :** $O(M)$[3]

SHAP values Properties

- Local Accuracy
  - Efficiency
- Missingness
- Consistency
  - Null effect
  - Linearity
  - Symmetry
- **Execution time :** $O(2^M)$

[1] Ruiz, L. M. ; Valenciano, F. ; and Zarzuelo, J. M. 1998. **«The family of least square values for transferable utility games»**. Games and Economic Behavior

[2] Condevaux, C. ; Harispe, S. ; and Mussard, S. 2022. Fair and Efficient Alternatives to Shapley-based Attribution Methods. ECML PKDD

[3] M is the number of features in the example to be explained.

To summary

- Eliminating the random step in Kernel SHAP leads to explanation stability.

- Removing randomness still maintains high-quality explanations.

- Using only Stratum 1 achieves complete stability and good explanations.

Shaping Up SHAP: Enhancing Stability through Layer-Wise Neighbor Selection

Gwladys Kelodjou[1], Laurence Rozé[2], Véronique Masson[1], Luis Galárraga[1], Romaric Gaudel[1], Maurice Tchuente[3], Alexandre Termier[1]

[1]Univ Rennes, Inria, CNRS, IRISA - UMR 6074 | [2]Univ Rennes, INSA Rennes, CNRS, Inria, IRISA - UMR 6074 | [3]Sorbonne University, IRD, University of Yaoundé I, UMI 209 UMMISCO

gwladys.kelodjou@irisa.fr

Thank You For Your Attention !

Paper

Visit our poster for more details !