Association for the Advancement of Artificial Intelligence

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



AAAI 2024

Context

- Machine Learning techniques are widely used in various domains.

- Concerns about fairness and trustworthiness motivate interest in **explainability**.

    - Addressing machine learning model opacity through **explanations**.

Local post-hoc explainability

Post-hoc local explainability focuses on explaining decisions for single instances.

Context

Local post-hoc explainability approaches :

- LIME[1], SHAP[2], DeepLIFT[3], etc. :
- Providing feature attribution scores.

Some Qualities of a Good Explanation

- **Interpretability** : The degree of understanding of the explanation by a user.
- **Stability** : The ability to reproduce the same explanation.
- **Fidelity** : The ability of the explanation model to accurately mimic the black-box model.

[1] Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. «**"Why should i trust you?" Explaining the predictions of any classifier**». In Proceedings of the 22nd ACM SIGKDD

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

[3] Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. «**Learning important features through propagating activation differences**». In International conference on machine learning, PMLR

Context

- **SHAP (SHapley Additive exPlanations)** : Applies cooperative game theory to reveal feature contributions.

- **Cooperative game theory** : fairly distribute the gain obtained by multiple players collaborating in a game.

- **Shapley value** : Consider different coalitions in which the player is present or absent to determine his contribution.

Context

- **SHAP values** are the **Shapley values** of a conditional expectation function of the original model.

- Exact computation of SHAP values is challenging.

- **Kernel SHAP** is a linear-regression-based approximation of SHAP values.
    - Widely used for its model-agnostic nature.

Problem

Kernel SHAP's estimator **suffers from stability issues**.

Related Work

- Metrics for stability in LIME explanations are introduced by Visani et al. (2022) [1].
- OptiLIME framework (2020)[2] allows users to customize the level of stability and fidelity in LIME explanations.
- S-lime (2021)[3] proposes a framework to determine the number of perturbation points required to ensure stable LIME explanations.
- The concept of **robustness** is introduced by Alvarez et al. (2018)[4], highlighting challenges in popular methods like LIME and SHAP.

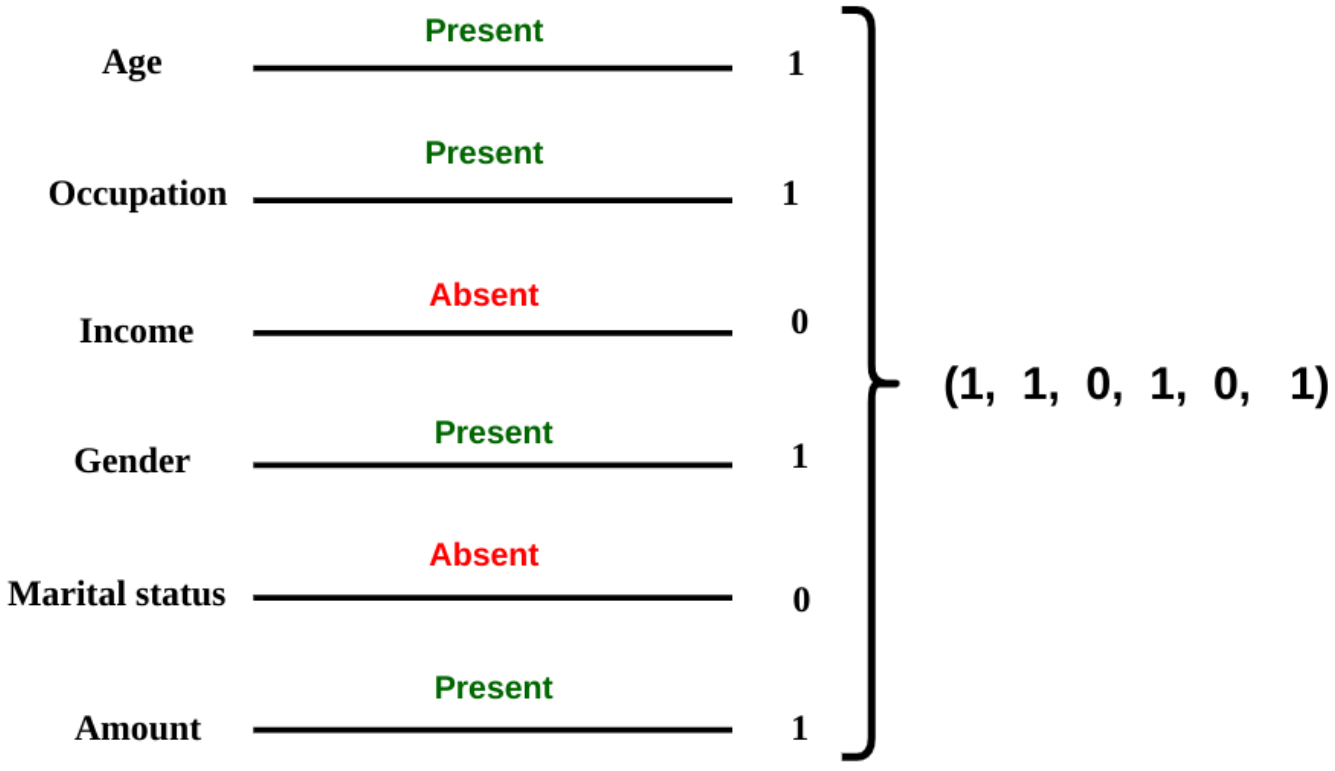
[1] Visani, G. et al. 2022. «**Statistical stability indices for LIME : Obtaining reliable explanations for machine learning models**». Journal of the Operational Research Society

[2] Visani, G.; Bagli, E.; and Chesani, F. 2020. **OptiLIME : Optimized LIME Explanations for Diagnostic Computer Algorithms**. CIKM Workshops

[3] Zhou, Z.; Hooker, G.; and Wang, F. 2021. «**S-lime : Stabilized-lime for model explanation**». KDD

[4] Alvarez-Melis, D.; and Jaakkola, T. S. 2018. «**On the Robustness of Interpretability Methods**». ICML Workshop

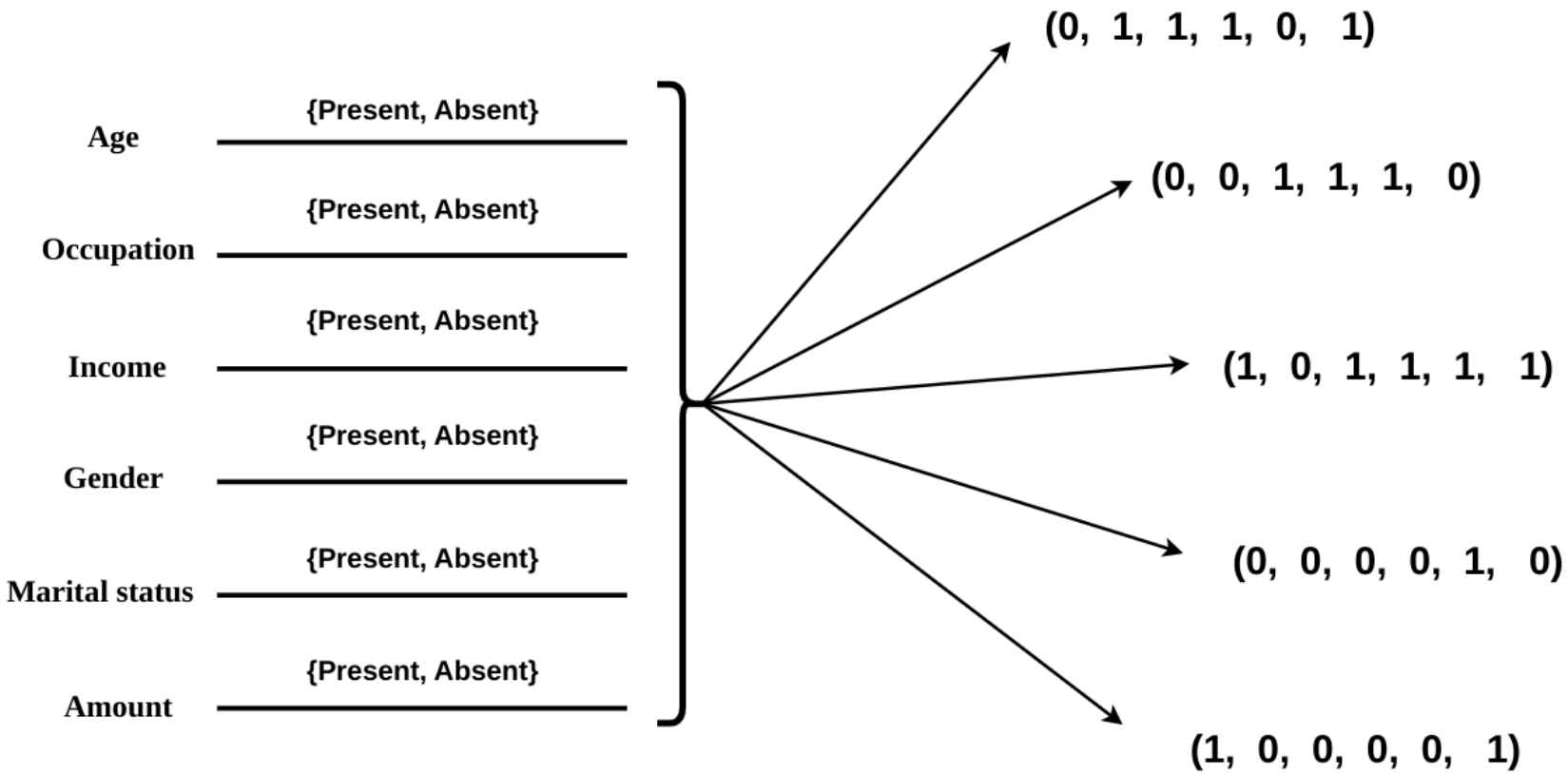
Coalition : subset of features.



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

Background
Coalition

Determining the contributions of each feature requires considering multiple coalitions.

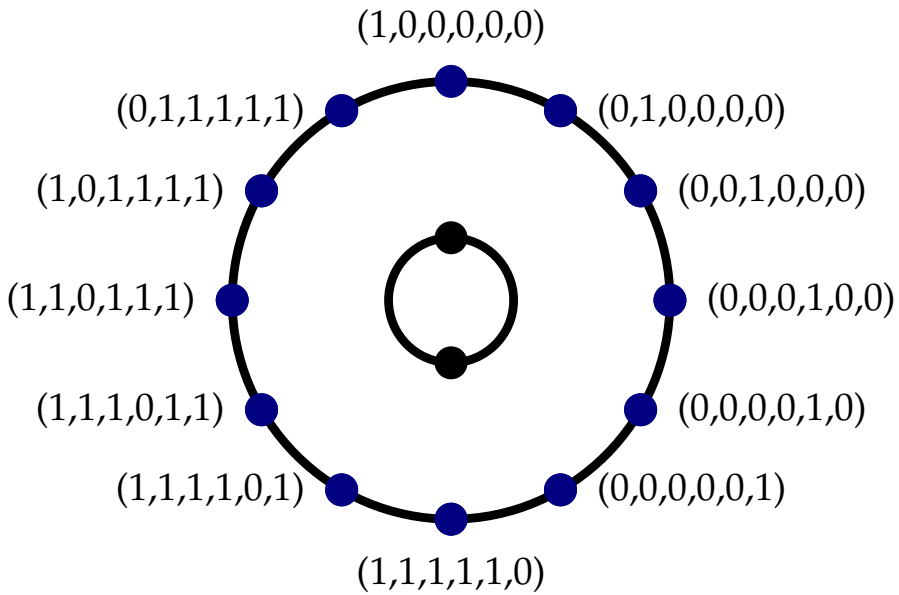


Computing SHAP values involves considering all possible coalitions.

Layer-Wise Neighbor Selection

Layer

Set of coalitions sharing the same number of features present or features absent.



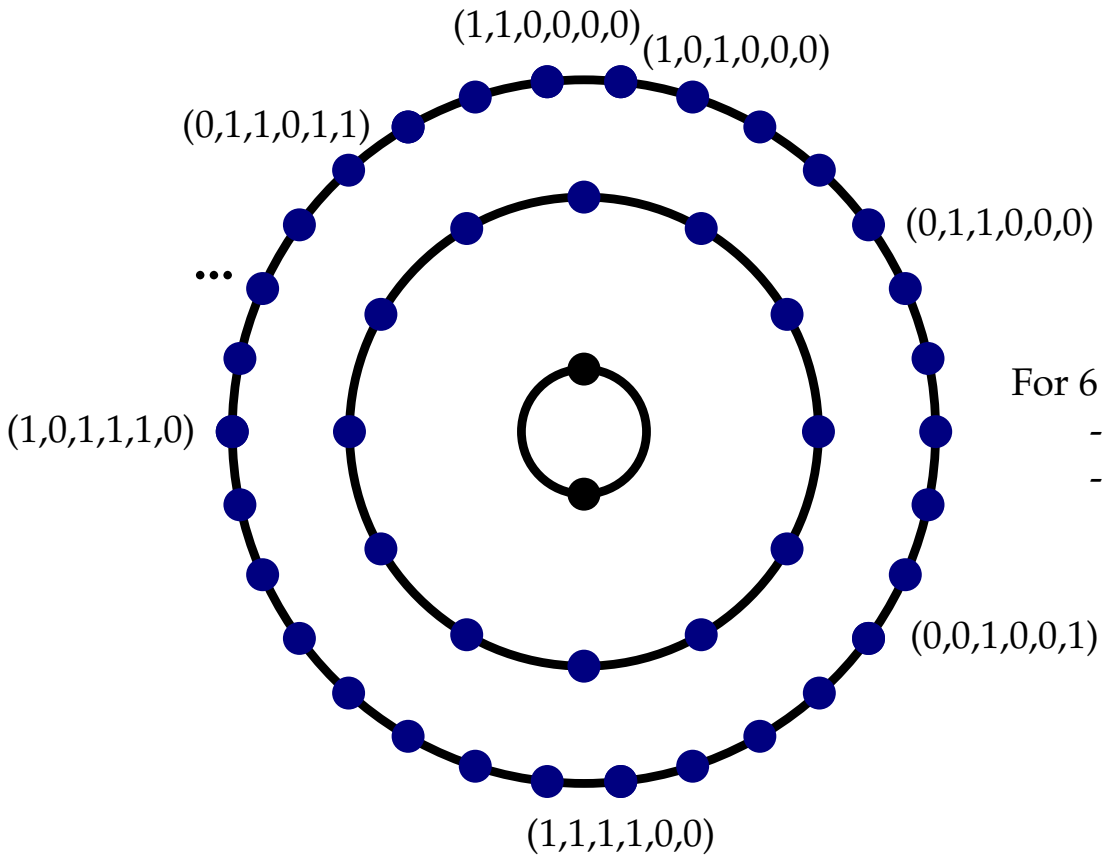
For 6 features, **layer 1** contains **12 coalitions :**

- All coalitions with a single feature present
- All coalitions with a single feature absent

Layer 1

Coalitions with a single feature present or with a single feature absent.

Layer-Wise Neighbor Selection

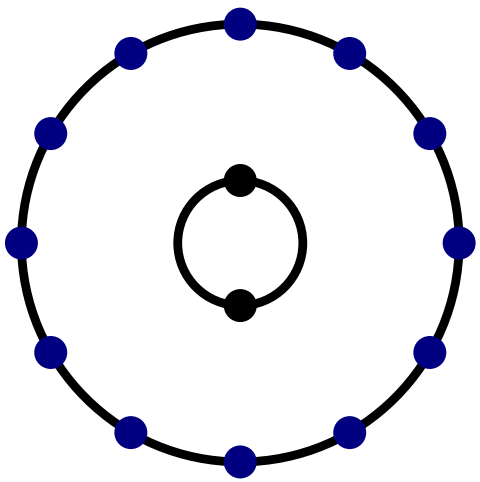


For 6 features, **stratum 2** contains **42 coalitions** :
- Coalitions of layer 1 (**12**)
- Coalitions of layer 2 (**30**)

- **Layer 2** : Coalitions with a two features present or two features absent.
- **Stratum 2** : Layer 1 + Layer 2.

Kernel SHAP

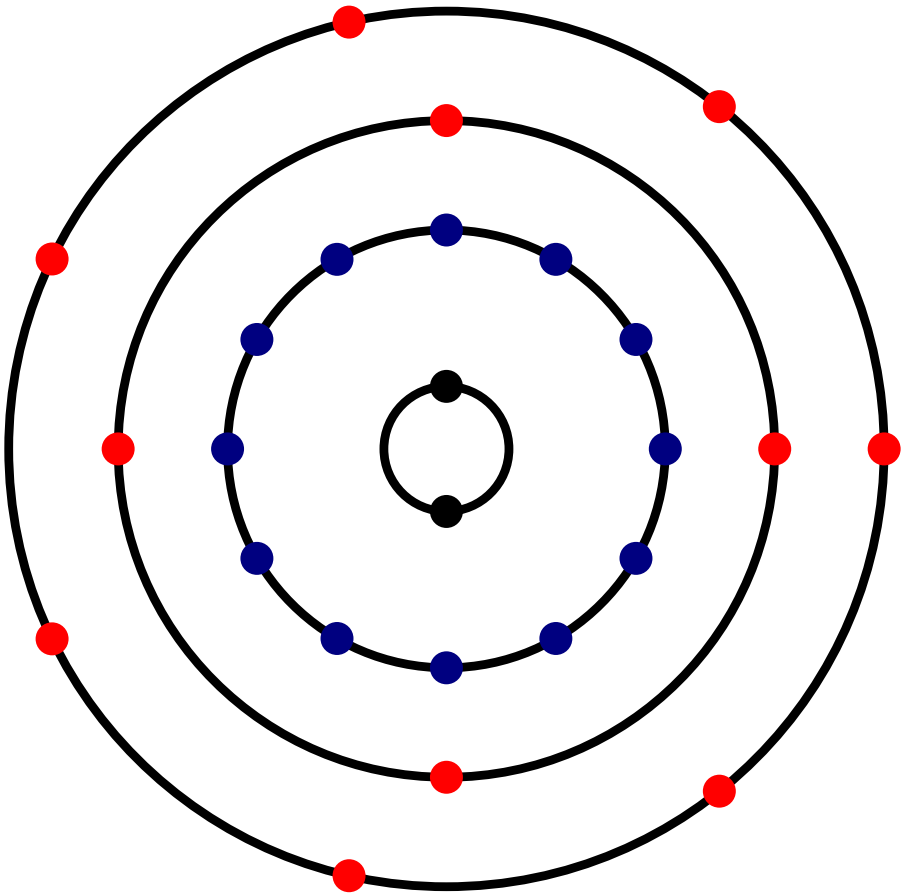
- Samples a subset of coalitions.
- **Budget** : determines the size of the coalition sample.



Stratum 1 full !

First generate coalitions from lower layers.

Kernel SHAP



Random sampling within layers 2 and 3

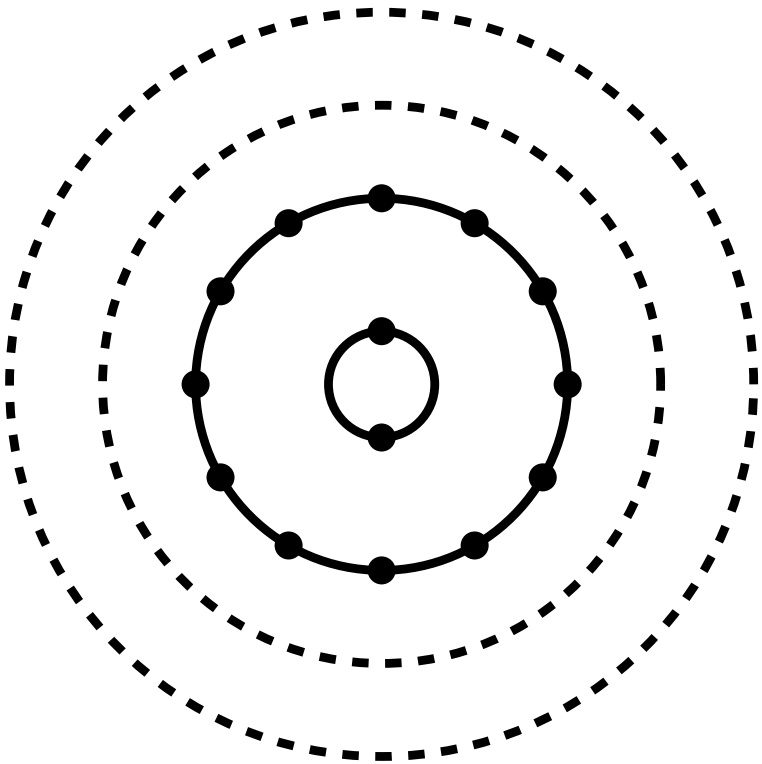
Randomly samples from subsequent layers if the budget is not exhausted.

Kelodjou et al.

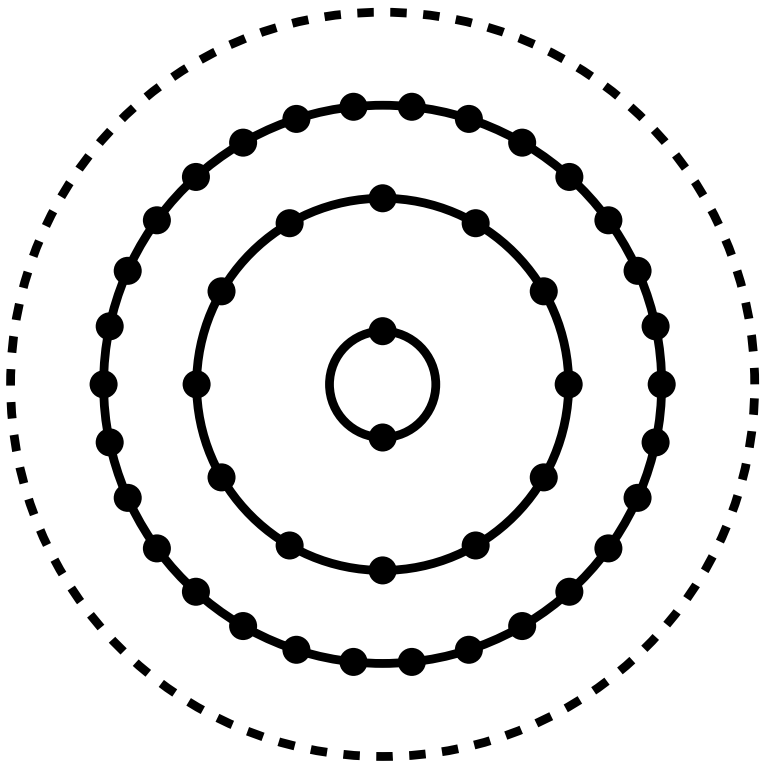
Contribution 1 : Neighbor selection

Achieving Kernel SHAP's stability

Set the budget to consider only complete stratums.

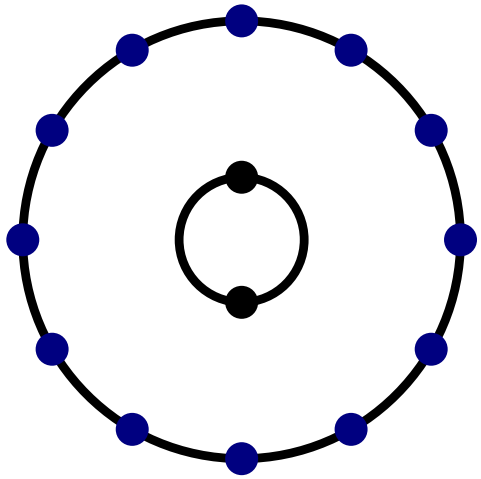


Stratum 1 full



Stratum 2 full

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.

Kelodjou et al.

First stratum Attribution Analysis
Properties

Stratum 1 Attribution Properties

- **LES family**[2,3] :
  - **Linearity**
  - **Efficiency**
  - **Symmetry**
- Missingness
- **Execution time :** $O(M)$

SHAP values Properties

- Local Accuracy
  - Efficiency
- Missingness
- Consistency
  - Null effect
  - Linearity
  - Symmetry
- **Execution time :** $O(2^M)$[1]

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

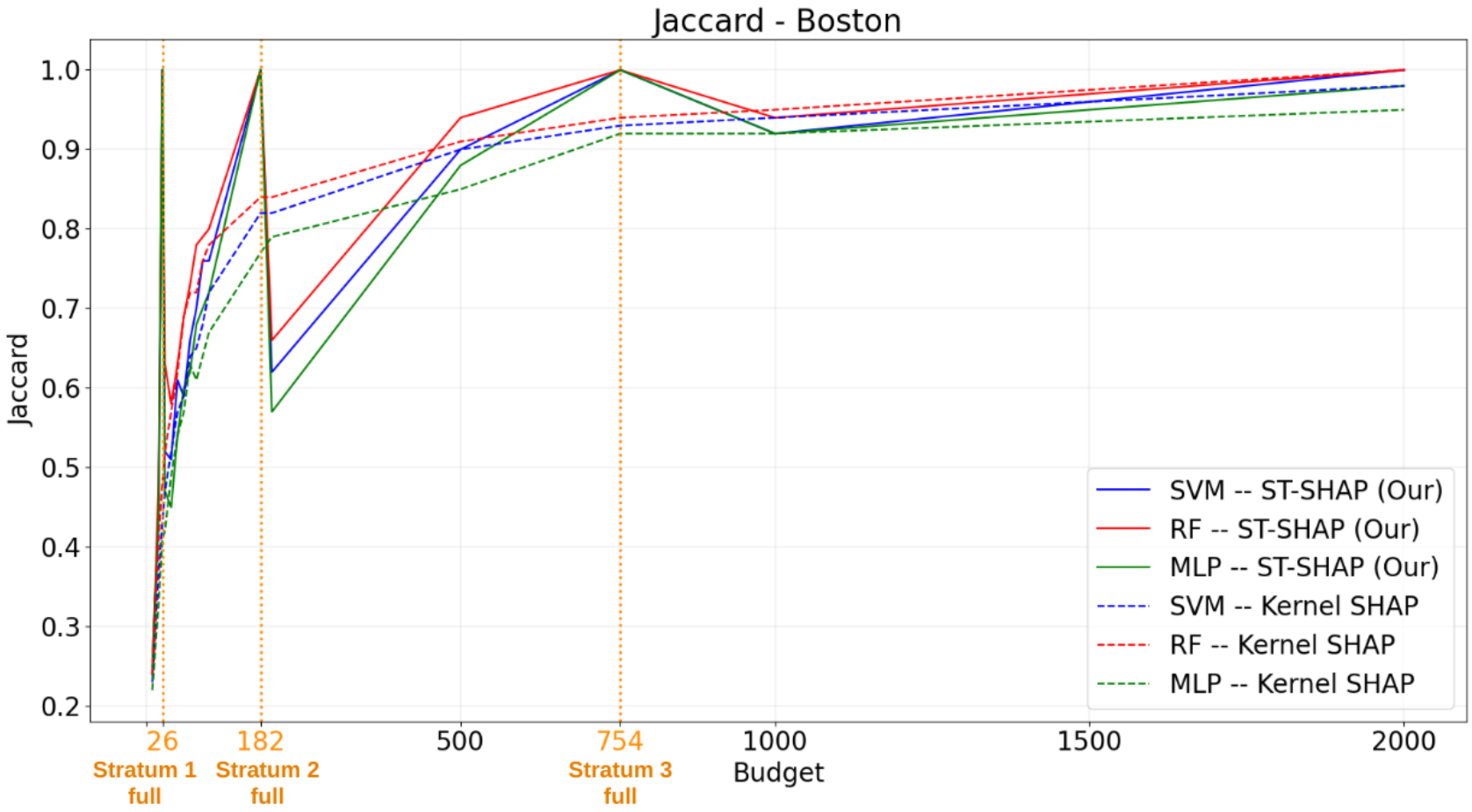
[2] Ruiz, L. M.; Valenciano, F.; and Zarzuelo, J. M. 1998. «**The family of least square values for transferable utility games**». Games and Economic Behavior

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

Experiments
Protocol

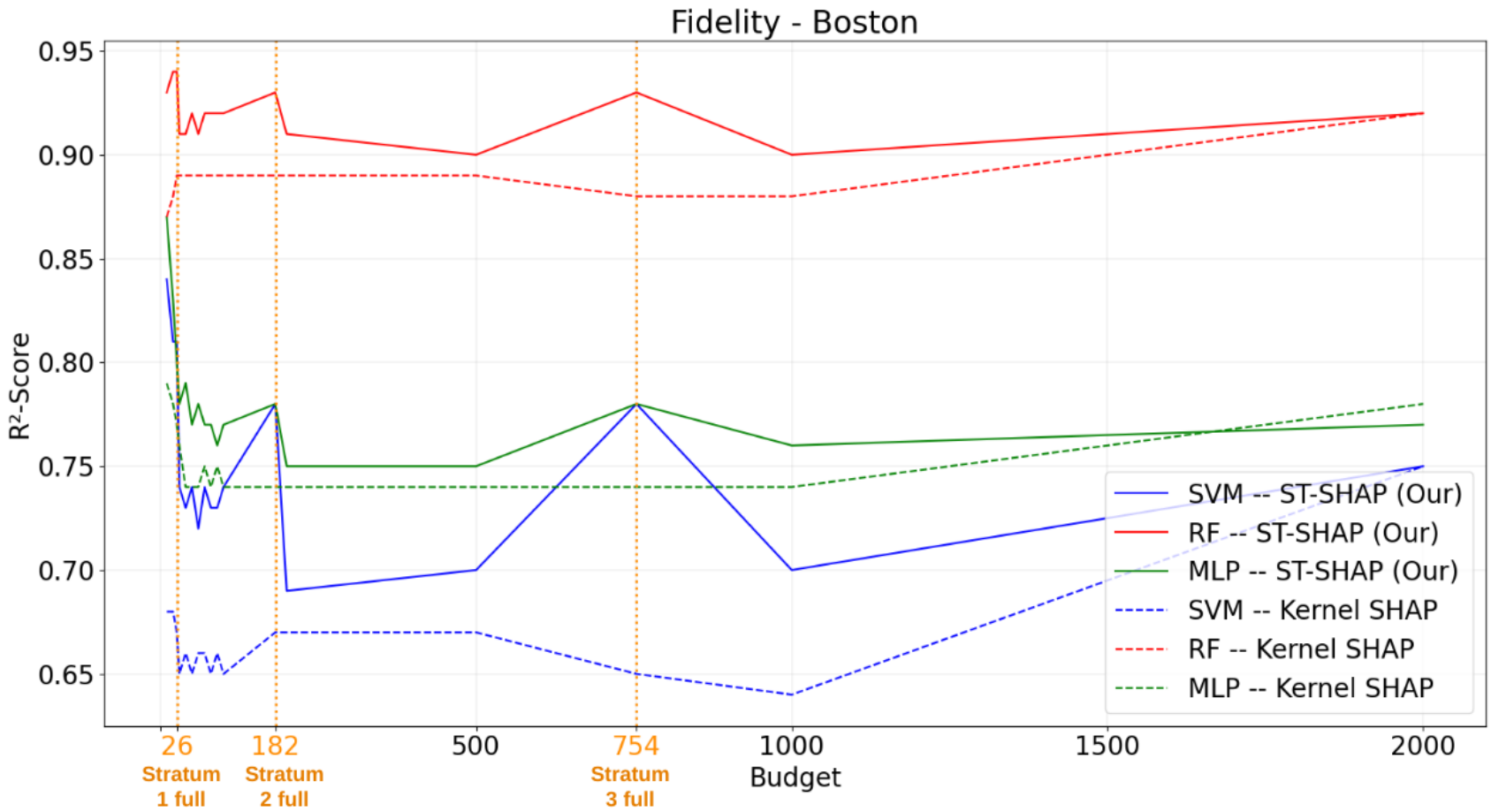
- **Black-box models** : Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Multi-layer Perceptron (MLP).
- **Datasets** : Boston, Movie, Credit Card, Adult Income, Dry Bean, HELOC, Spambase, Wisconsin Breast Cancer Database.
- **Various budgets** to assess stability and fidelity.
- **Metrics** :
  - **Jaccard coefficient** for stability measurement.
  - R^2**-score** and **Accuracy** for fidelity measurement.
- Code is publicly available at `https://github.com/gwladyskelodjou/st-shap`.

Experiments
Results



Complete stratums $\Longrightarrow$ stable output.

Experiments
Results



No impact on explanation fidelity observed with our methods : remains highly accurate.

Kelodjou et al.

Conclusion

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

- Removing randomness still maintains high fidelity.

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

Future work

- Properties of black-box models that guarantee good approximations of the SHAP values.

- Budget - Black-Box complexity Relationship for Explanation Optimization.