Mathematics of Explainable AI with Applications to Finance

- Alexey Miroshnikov
- Data Science Research Group, Discover Financial Services

- King Abdullah University of Science and Technology, AMCS/STAT Graduate Seminar, November 14, 2024

1. Motivation

Model complexity and interpretability

- Contemporary predictive and generative ML models are complex
  - Neural Networks (NN) and Graph Neural networks (GNN)
  - Gradient Boosting Machines (GBM)
  - Unsupervised and semi-supervised methods (e.g. variational autoencoders)
  - Large-Language models
- Interpretability (explainability) of ML models is crucial for business adoption, model documentation, regulatory oversight, and human acceptance and trust. Crucial in Banking, Insurance, Healthcare.
- Accuracy may come at the expense of interpretability
  - Linear models are easy to interpret, $Y = a_1X_1 + \cdots + a_nX_n$.
  - Nonlinear models (NN, GNN, GBM) are difficult to interpret.

Regulatory requirements

ML models and strategies that rely on ML models are subject to federal laws and regulations, including the Equal Credit Opportunity Act (ECOA), Fair Housing Act (FHA), and Equal Employment Opportunity Act (EEOA).

Financial institutions in the United States (US) are required under the ECOA to notify declined or negatively impacted applicants of the main factors that led to the adverse action.

Determining the factor contributing the most to an outcome of a model may be done via individualized feature attributions.

Common approaches:

Self-interpretable models

Post-hoc model explanations

2. Use cases

Feature attribution for predictive ML models

ML risk models use historical, consumer and consumer reporting information to estimate the probability of default. US Federal regulations require lenders to provide applicants with the primary factors that contribute to an adverse action (i.e., decline).

Marketing campaigns. FI send out letters with descriptions of various products. These campaigns are costly. It makes sense to optimize advertising efforts based on which ad campaign (strategy) will bring in the most customers or has the highest level of engagement. Explainability techniques can be applied to assess the effectiveness of marketing campaigns or strategies to bring in customers.

Explainability methods. There are a variety of mathematical and statistical techniques that quantify the contribution of each element from the input vector to the predictive model output given the distribution of inputs. Game theoretic approaches are popular, as well as models that are inherently interpretable.

Individual feature attributions

Input

- (X, Y), where $X = (X_1, \dots, X_n)$ are features, $Y \in \mathbb{R}$ a response variable on $(\Omega, \mathcal{F}, \mathbb{P})$.
- $x \to f(x) = \mathbb{E}[Y|X = x]$ or $\mathbb{P}(Y = 1|X = x)$ (regressor or classification score).

(Local) Model explainer

Quantifies the contribution of an observation $x = (x_1, x_2, \dots x_n) \sim X$ to the value $f(x)$.

$$\mathbb{R}^n \ni x \to E\left(x; f, X, \mathcal{I}_f\right) = (E_1, E_2, \dots E_n) \in \mathbb{R}^n .$$

Here the model f, the random vector X and model implementation $\mathcal{I}_f$ serve as parameters.

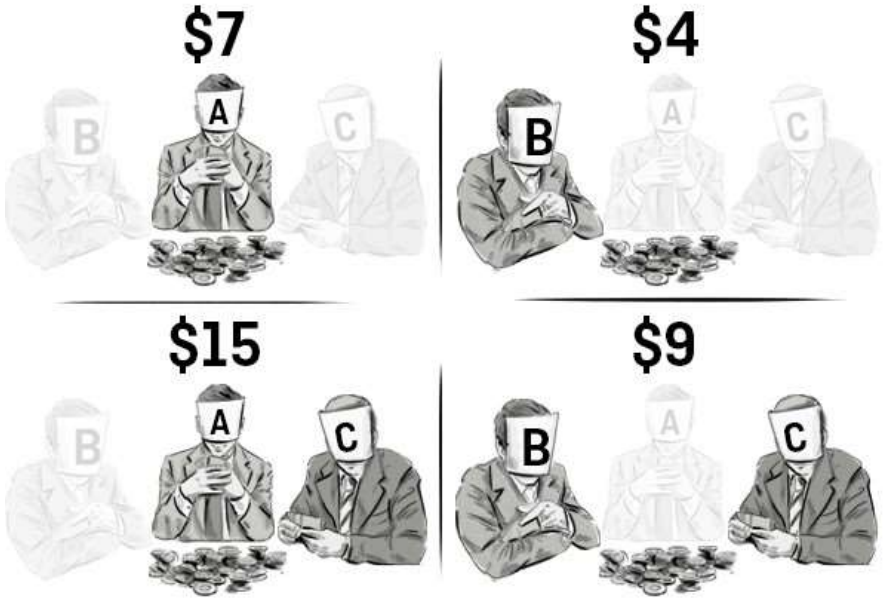
Game-theoretic approaches

- Game theoretic approaches have been explored in Štrumbelj & Kononenko (2014), Lundberg & Lee (2017)
- Cooperative game (N, v)
  - set of players indexed by $N = \{1,2, \dots, n\}$
  - utility $v(S)$, $S \subseteq N$
- Game value

$$(N, v) \to h[N, v] = \{h_i[N, v]\}_{i=1}^{n} \in \mathbb{R}^n$$

- Shapley value (Shapley, 1953)

$$\varphi_i[v] = \sum_{S \subseteq N \setminus \{i\}} \frac{s!(n-s-1)!}{n!} \big(v(S \cup i) - v(S)\big),\ i \in N.$$



[from gametheory.online]

ML games and values

We study game values in the marginalist form

$$h_i[N, v] = \sum_{S \subseteq N \setminus \{i\}} w(S, |N|) \cdot \big(v(S \cup i) - v(S)\big)$$

ML games

$v^{CE}(S, x; X, f) = \mathbb{E}[f(X_S, X_{-S})|X_S = x_s]$ (conditional game)

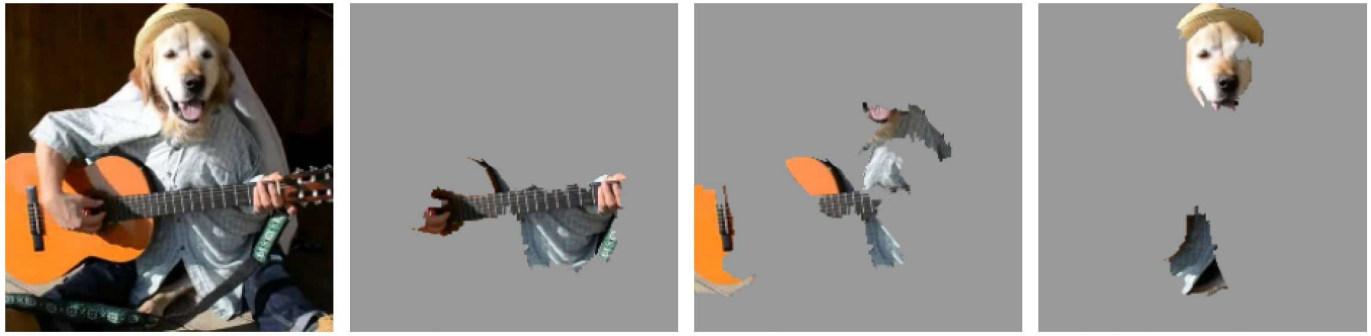
$v^{ME}(S, x; X, f) = \mathbb{E}[f(x_S, X_{-S})]$ (marginal game)

Works

- Fast marginal game value attributions for tree-based models with symmetric trees
  [K. Filom, A.M., K. Kotsiopoulos, A. Ravi Kannan, Foundations of Data Science (2024)]
- On stability of AI explanations based on marginal and conditional game values
  [A.M., K. Kotsiopoulos, K. Filom, A. Ravi Kannan, arxiv:2102.10878 (2024)]
- Sampling algorithms for attributions based on coalitional values such as Owen value
  [K. Kotsiopoulos, A.M., K. Filom, A. Ravi Kannan arxiv:2303.10216 (2023)]

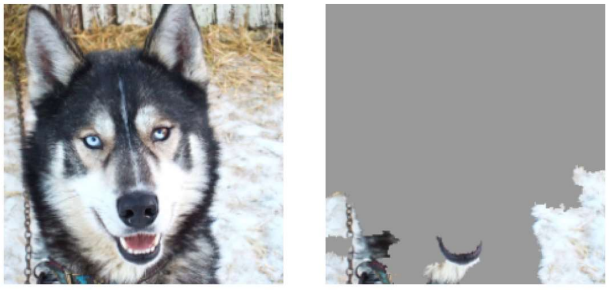
Example: image classification feature attribution

[Ribeiro et al. “Why should I trust you?”]



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google’s Inception neural network. The top 3 classes predicted are “Electric Guitar” ($p = 0.32$), “Acoustic guitar” ($p = 0.24$) and “Labrador” ($p = 0.21$)



(a) Husky classified as wolf (b) Explanation

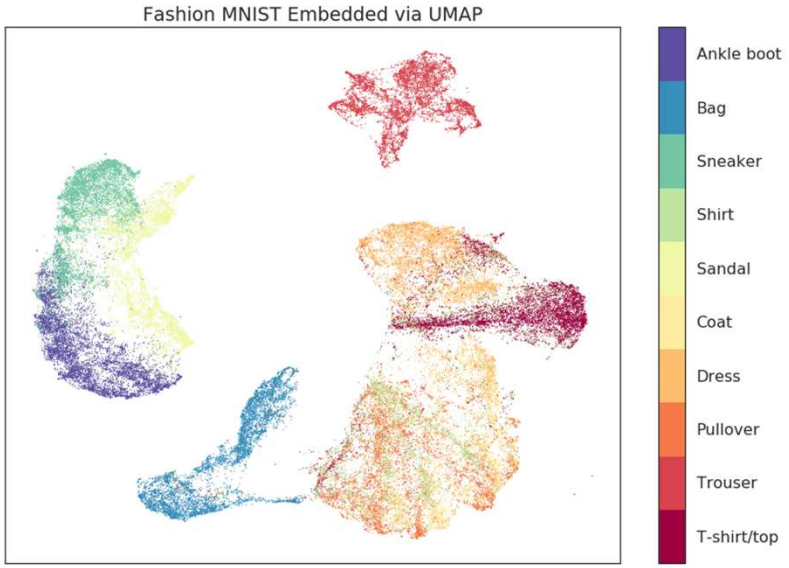
Figure 11: Raw data and explanation of a bad model’s prediction in the “Husky vs Wolf” task.

Clustering methods via manifold learning

Customer behavior. Clustering algorithms are used to categorize customer behavior and segment the datasets of features. To get the intuition and insights about the data.

Manifold learning creates the data embedding that helps with denoising and regularization of the data. In some cases it can help to significantly improve clustering by the regularization and improved representation.

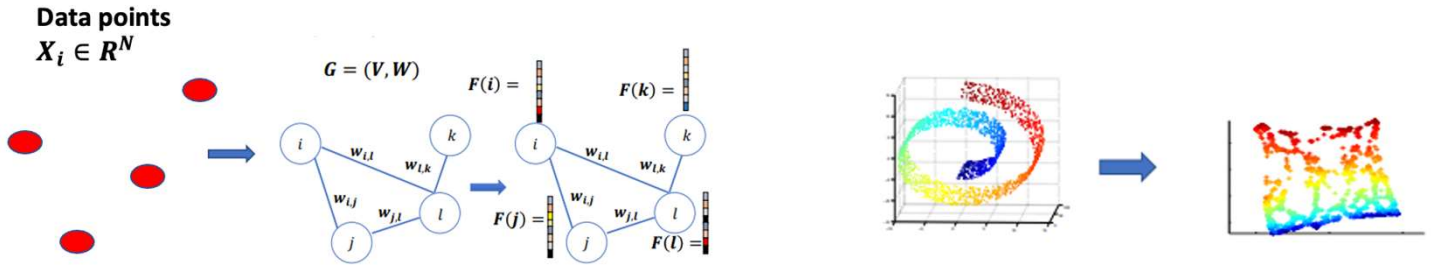
Feature attribution (local and global) allows one to characterize different aspects of population in terms of features. Each predictors has a semantic meaning. Segments can be characterized by features.



https://umap-learn.readthedocs.io/en/latest/supervised.html

Manifold Learning and Clustering

1. Construct a topological representation of data, e.g. KNN-graph or weighted graph like in UMAP using topological data analysis.
2. Initialize the low dimensional representation using e.g. spectral embedding (e.g. the Laplacian eigenmap).
3. Optimize a loss function making the low dimensional representation to have a fuzzy topological representation as close as possible to the original one.
4. (optional) Perform clustering in the embedded space.



- UMAP [McInnes et al 2018], tSNE [van der Maaten & Hinton, 2008], Laplacian eigenmaps [Belkin & Niyogi 2003]

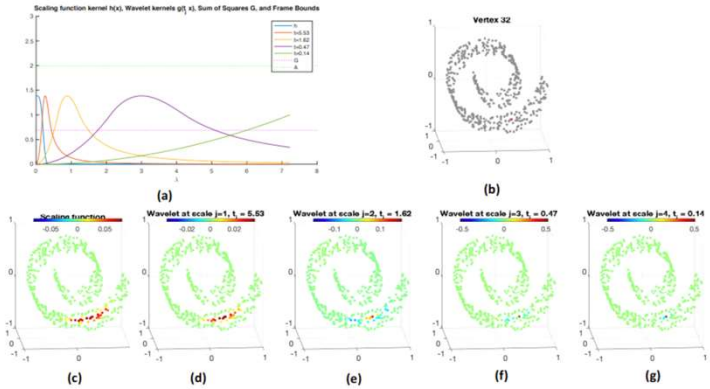
Issues with UMAP

- Laplacian eigenmaps initialization focus on low-frequency overlooking higher-frequency patterns.
- Lack of explicit mapping linking the original high dimensional dataset to its low-dimensional dimensional embedding.

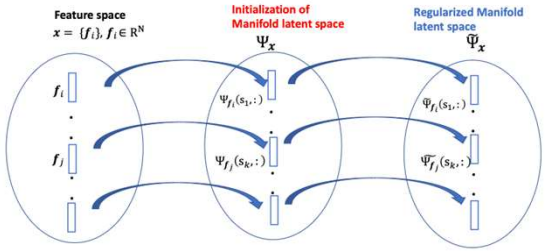
Multi-Scale Graph Embedding Approach for Interpretable Manifold Learning

Our work [S. Deutch, L. Yelibi, A.T. Lin, A. Ravi Kannan, arXiv:2406.02778,2024]

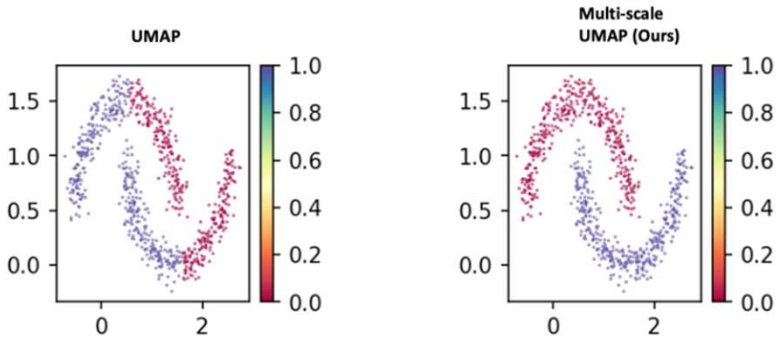
- Spectral graph wavelets (SGW) for multiscale encoding



- Correspondence of original and embedded features



- Accuracy



Method / Accuracy	ARI	AMI
UMAP	0.54	0.51
t-SNE	0.42	0.35
ISOMAP	0.36	0.3
Diffusion Maps	0.25	0.19
HeatGeo	0.54	0.52
MS-IMAP Method 1	0.75	0.73
MS-IMAP Method 2	**0.89**	**0.87**

Table 1:Comparison of clustering performance on the Two Moons datasets.

Generative AI and LLM

Generative AI (genAI) might be helpful for customer support, compliance, and risk management.

- Understanding policies; e.g. helping to make a query into a policy to make the discussion on the phone smooth.
- It is possible to use genAI to scan the documents and generate summaries.
- Sentiment analysis (multi-class analysis) e.g. text messages, phone, etc. This is used to classify the sentiment by using genAI instead of the traditional NLP classifier e.g. Large Language Model (LLM) based transformer.

Interpretability of LLM models

- LLM models trained on vast amount of data and different parts responsible for the output.
- Lack of transparency poses critical challenges when it comes to their adaptation by financial institutions.

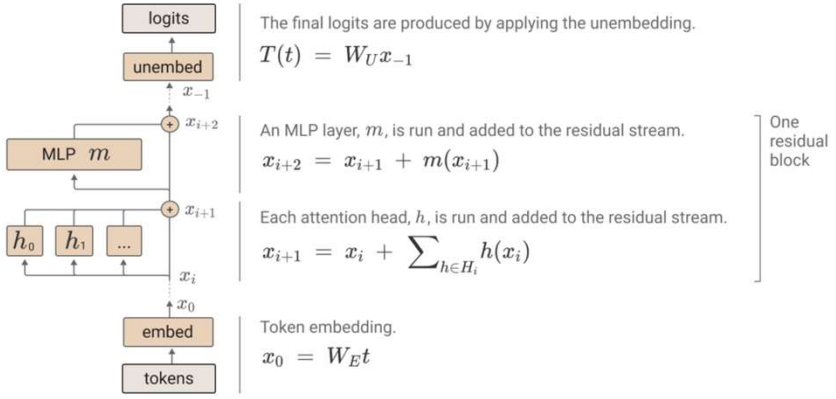
Mechanistic Interpretability

Understanding an LLM at the level of neurons, circuits, and attention heads

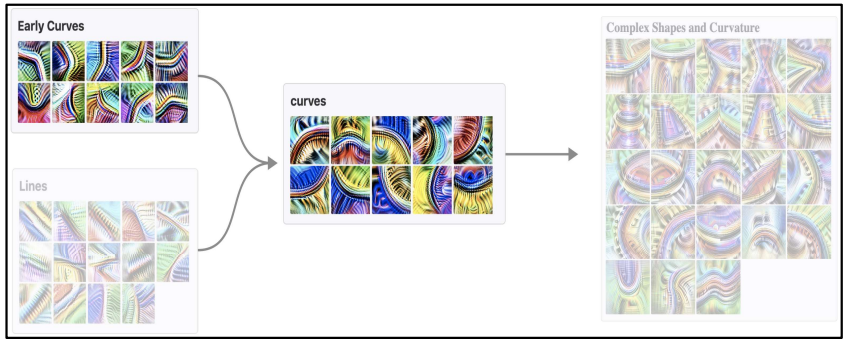
- Micro scale explanation
- Lack of transparency of LLMs
  - Safety challenges such as hallucination, toxicity, unfairness.
  - Misalignment with human values.
- Model pruning
  - Significant cost savings and rapid inference time

A. Golgoon, K. Filom, A. Ravi Kannan, Mechanistic interpretability of large language models with applications to the financial services industry, ACM on AI in Finance, 2024.

- Examples of how algorithmic tasks can be designed for compliance monitoring purposes.



N. Elhage et. al. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*, 2021.

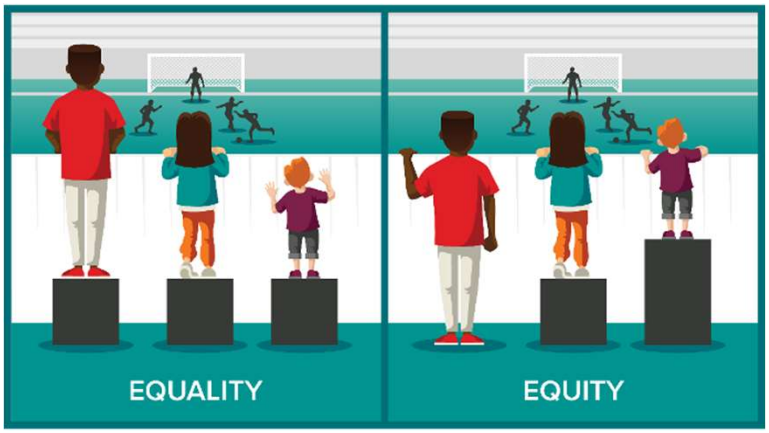


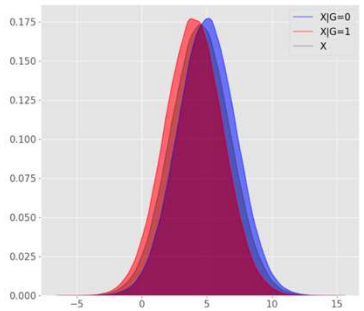
C. Olah et. al, Zoom In: An Introduction to Circuits. Distill, 2020.

Fair Lending

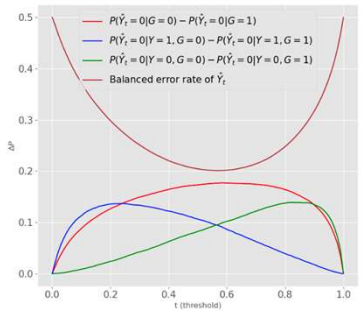
- The decision-making process may lead to certain unintended type of bias impacting sub-populations
- ECOA and FHA laws and regulations prohibit discrimination against protected classes (sub-populations); thus, disparities against the sub-populations must be considered.
- The disparities in the outputs, maybe be measured as differences in probability of default

$$|\mathbb{P}(Y = 0|G = 0) - \mathbb{P}(Y = 0|G = 1)|$$





(a) Subpopulation distributions of X.



(b) Fairness measurements.

ML Fairness Explainability

A.M., K. Kotsiopoulos, R. Franks, A. Ravi Kannan, "Wasserstein-based fairness interpretability framework for machine learning models, Machine Learning (Springer), 2022.

- Global metric

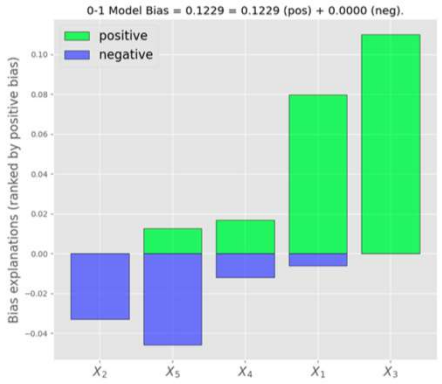
$$\text{Bias}_{W_1}(f|G) = W_1\left(P_{f(X)|G=0}, P_{f(X)|G=1}\right) = \int_0^1 bias_t^C(f|G)dt$$

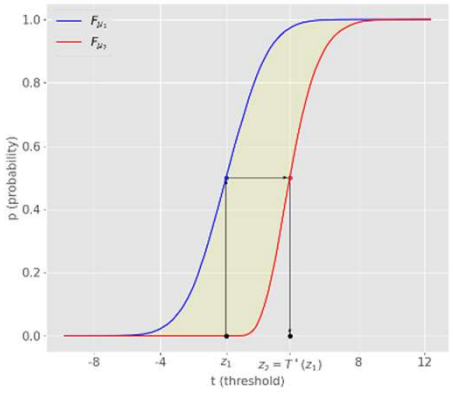
- Marginal bias game

$$v_{bias}(S;f) = Bias_{W_p}(v^{ME}(S;f,X)|X,G)$$

- Bias explanations

$$\varphi_i[v_{bias}(S;f)], i \in N = \{1,2,\dots N\}$$





3. Marginal feature attributions for ML models with oblivious trees

Individual feature attributions

Input

- (X, Y), where $X = (X_1, \dots, X_n)$ are features, $Y \in \mathbb{R}$ a response variable on $(\Omega, \mathcal{F}, \mathbb{P})$.
- $x \to f(x) = \mathbb{E}[Y|X = x]$ or $\mathbb{P}(Y = 1|X = x)$ (regressor or classification score).

(Individual) Model explainer

Quantifies the contribution of an observation $x = (x_1, x_2, \dots x_n) \sim X$ to the value $f(x)$.

$$\mathbb{R}^n \ni x \to E\left(x; f, X, \mathcal{I}_f\right) = (E_1, E_2, \dots E_n) \in \mathbb{R}^n .$$

Here the model f, the random vector X and model implementation $\mathcal{I}_f$ serve as parameters.

ML games and values

ML games

$v^{CE}(S, x; X, f) = \mathbb{E}[f(X_S, X_{-S})|X_S = x_s]$ (conditional game, "true−to−the data")

$v^{ME}(S, x; X, f) = \mathbb{E}[f(x_S, X_{-S})]$ (marginal game, "true-to-the-model")

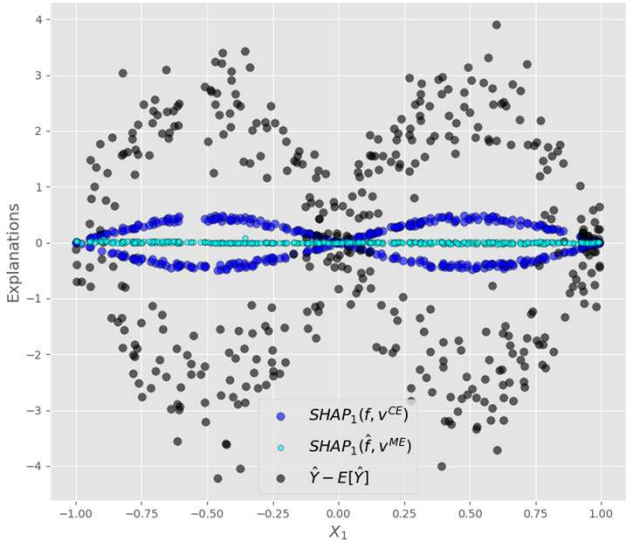
Game value (marginalist form)

$$h_i[N, v] = \sum_{S \subseteq N \setminus \{i\}} w(S, |N|) \cdot \big(v(S \cup i) - v(S)\big)$$

Conditional and marginal explanations of f for feature X_i at x are defined:

- $x \to h_*^{CE}(x) = h_i[N, v_*^{CE}(\cdot, x; f)] \in \mathbb{R}^n$
- $x \to h_*^{ME}(x) = h_i[N, v_*^{ME}(\cdot, x; f)] \in \mathbb{R}^n$

$Y = f(X) = X_2 X_3$, $X_2 = \sin(\pi X_1) + \epsilon$



Obstacles in computing game-theoretic feature attributions

- Features are almost never independent.

  Conditional feature attributions (based on v^{CE}) often differ from the marginal ones (based on v^{ME}).

  - Remedy: Grouping features based on dependencies and using coalitional value (e.g. the Owen value) unifies the two frameworks and yields more stable explanations.
    [M.-Kotsiopoulos-Filom-Ravi Kannan (2024)]

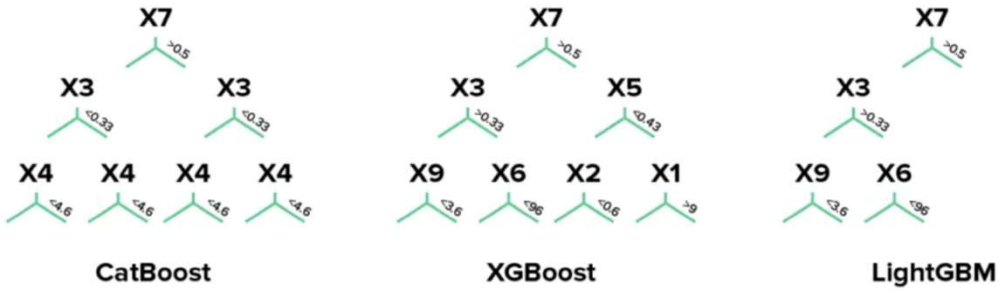
- Formulas for game values usually have exponentially many terms.

  E.g. the formula for Shapley value has 2^{n-1} terms (n can be ≥ 100 in a credit card risk model)

  - Remedy: Monte-Carlo approximation. [Štrumbelj-Kononenko 2010 & 2014], [Kotsiopoulos-Miroshnikov-F.-Ravi Kannan 2023].
  - Remedy: Focusing on a specific type of models.

Solution in a special case: oblivious trees

- ML tree ensembles $f(x) = \sum_j \mathcal{T}_j(x)$, $\mathcal{T}_j$ is a tree (a simple function).
- $f \to h[v^{ME}(f)] = \sum_j h\left[v^{ME}\left(\mathcal{T}_j\right)\right]$ due to linearity
- Type of trees:



Picture from Medium.

- The CatBoost library utilizes **oblivious (symmetric) decision trees** as base learners [Dorogush-Ershov-Gulin 2018].
- Despite this constraint, ensembles of symmetric trees demonstrate competitive predictive power [Ferov-Modrý 2016], [Hancock-Khoshgoftaar 2020].

Main result: a model-specific and inherently-interpretable approach

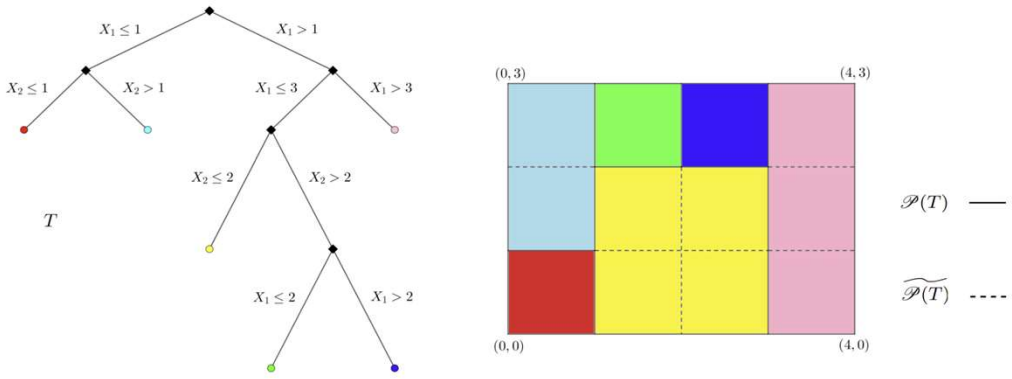
[Filom-M.-Kotsiopoulos-Ravi Kannan 2023] (10.3934/fods.2024021)

Let T be an ensemble of symmetric decision trees of depth d trained on a dataset D.
(Typically, D is large and $\mathrm{d} = \log(\mathcal{L}) \leq 10$). Then:

- Precomputation. There is **an explicit formula** for marginal game values of T solely in terms of the model's parameters with complexity $O(\mathcal{L}^{1.6})$ per leaf.
- Computation. Based on this analytic solution, we designed **an algorithm for estimating marginal feature attributions** of $\mathcal{T}$ according to certain precomputed look-up tables.
- Complexity. The algorithm is **fast. T**he computation complexity of the is $O(|T| \cdot d)$ and **accurate** (variance of error $\sim 1/|D|$). It does not depend on the background dataset!
- Generalization. The formula **can be generalized** for an axiomatically characterized family of game values (including variants of Shapley such as Banzhaf or Owen).
- Proof. combinatorial analysis + null player property (features that do not appear in the tree do not play).

What is special about oblivious (symmetric) trees?

- For a tree $\mathcal{T}$, marginal feature attributions based on a linear game value are piecewise constant, but only with respect to a grid partition $\widetilde{\mathcal{P}(\mathcal{T})}$, which is often finer than the tree's partition $\mathcal{P}(\mathcal{T})$. They coincide when $\mathcal{T}$ is symmetric.
- Game value computations can be simplified by exploiting the symmetry.



Picture from 10.3934/fods.2024021.

References

- J.F. Banzhaf, Weighted voting doesn't work: a mathematical analysis. Rutgers Law Review 19, 317-343, (1965).
- M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 2003.
- H. Chen, J. Danizek, S. Lundberg, S.-I. Lee, True to the Model or True to the Data. arXiv preprint arXiv:2006.1623v1, (2020)
- S. Deutsch, L. Yelibi, A. Tong Lin, A. Ravi Kannan, MS-IMAP -- A Multi-Scale Graph Embedding Approach for Interpretable Manifold Learning.
- K. Filom, A. Miroshnikov, K. Kotsiopoulos, A. Ravi Kannan, On marginal feature attributions of tree-based models. *Foundations of Data Science, AIMS.* Volume 6, Issue 4: 395-467, (2024).
- A. Golgoon, K. Filom, A. Ravi Kannan. Mechanistic interpretability of large language models with applications to the financial services industry.
- P. Hall, B. Cox, S. Dickerson, A. Ravi Kannan, R. Kulkarni, N. Schmidt, A United States Fair Lending Perspective on Machine Learning. Front. Artif. Intell. 4:695301. doi: 10.3389/frai.2021.695301 (2021).
- P. Hall, N. Gill, An Introduction to Machine Learning Interpretability, O'Reilly. (2018).
- S.M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, 31st Conference on Neural Information Processing Systems, (2017).
- L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction, 2018.
- K. Kotsiopoulos, A. Miroshnikov, K. Filom, A. Ravi Kannan, Approximation of group explainers with coalition structure using Monte Carlo sampling on the product space of coalitions and features, *arXiv preprint (2024),* arXiv:2303.10216v2.
- A. Miroshnikov, K. Kotsiopoulos, R. Franks, A. Ravi Kannan, Wasserstein-based fairness interpretability framework for machine learning models, *Machine Learning Journal, Springer.* (2022), https://link.springer.com/article/10.1007/s10994-022-06213-9.
- A. Miroshnikov, K. Kotsiopoulos, A. Ravi Kannan, Mutual information-based group explainers with coalition structure for machine learning model. explanations, arXiv:2102.10878v4 (2022)
- A. Miroshnikov, K. Kotsiopoulos, A. Ravi Kannan, Stability theory of game-theoretic group feature explanations for machine learning models, arXiv:2102.10878v6 (2024)
- L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, pages 2579–2605, 2008.
- L. S. Shapley, A value for n-person games, Annals of Mathematics Studies, No. 28, 307-317 (1953).
- G. Owen, Values of games with a priori unions. In: Essays in Mathematical Economics and Game Theory (R. Henn and O. Moeschlin, eds.), Springer, 76 {88 (1977).
- G. Owen, Modification of the Banzhaf-Coleman index for games with apriory unions. In: Power, Voting and Voting Power (M.J. Holler, ed.), Physica-Verlag, 232-238. and Game Theory (R. Henn and O. Moeschlin, eds.), Springer, 76-88 (1982).
- M.T. Ribeiro, S. Singh and C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier, 22nd Conference on Knowledge Discovery and Data Mining, (2016).
- E. Strumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst., 41, 3, 647-665, (2014).