Synthetic Benchmarks for Scientific Research in Explainable Machine Learning

Yang Liu
Abacus.AI

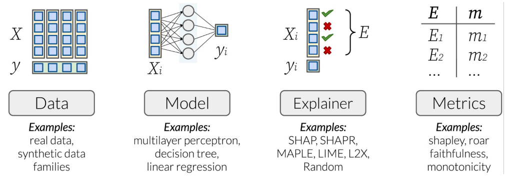
Sujay Khandagale
Abacus.AI

Colin White
Abacus.AI

Willie Neiswanger
Stanford University

Introduction

- Machine learning models are growing more complex.
- Their applications become more high-stakes.
  - credit scoring, loan approval, criminal recidivism.
- Many types of explainers have been proposed.
- Local feature attribution is one of the most popular type.
  - SHAP, LIME, SHAPR, MAPLE.
- We propose a synthetic benchmark suite to evaluate local feature attribution explainers and simulate real datasets.
- We evaluate six popular explainers and identify their failure modes.



Evaluation Metrics

Datapoint $\boldsymbol{x} \sim \mathcal{D}$ Feature set $S \subseteq \{1, \cdots, D\}$

Feature weights $\boldsymbol{w}$ A set of *i* least important features $S^-(\boldsymbol{w}, i)$

$$p\left(\boldsymbol{x}' \sim \mathcal{D}\left(\boldsymbol{x}_S\right)\right) = p\left(\boldsymbol{x}' \sim \mathcal{D} \mid x_i' = x_i \text{ for all } i \in S\right)$$

$$\text{faith}- = \text{Pearson}\left(\left|\mathbb{E}_{\boldsymbol{x}' \sim \mathcal{D}(\boldsymbol{x}_{F\setminus i})}[f(\boldsymbol{x}')] - f(\boldsymbol{x})\right|_{1 \le i \le D}, [w_i]_{1 \le i \le D}\right)$$

$$\delta_i^- = \mathbb{E}_{\boldsymbol{x}' \sim \mathcal{D}(\boldsymbol{x}_{S^-(\boldsymbol{w}, i+1)})}[f(\boldsymbol{x}')] - \mathbb{E}_{\boldsymbol{x}' \sim \mathcal{D}(\boldsymbol{x}_{S^-(\boldsymbol{w}, i)})}[f(\boldsymbol{x}')],$$

$$\text{mono}- = \frac{1}{D-1} \sum_{i=0}^{D-2} \mathbb{I}_{|\delta_i^-| \le |\delta_{i+1}^-|}$$

Summary of **10 evaluation metrics**.

Metric	Type	Model evaluations	Retrain	Linearity
faith+/-	correlation	$\Theta(D)$		✓
mono+/-	ranking	$\Theta(D)$		✓
roar-faith+/-	correlation	$\Theta(D)$	✓	✓
roar-mono+/-	ranking	$\Theta(D)$	✓	✓
shapley-mse	accuracy	$\Theta(2^D)$		
shapley-corr	correlation	$\Theta(2^D)$		

Synthetic Datasets

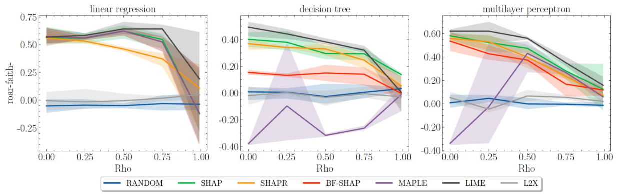
- Conditional expectations are needed to compute metrics.
- They are hard to compute for real-world datasets
- Synthetic datasets enable accurate sampling of conditional distributions.
- We implement 2 feature types:
  - **Multivariate Gaussian** and **Mixture of Gaussians**
- Three dataset types:
  - **Linear**, **Nonlinear Additive**, **Piecewise Constant**

Experimental Results

Multivariate Gaussian Dataset with Piecewise Constant Labels, Decision Tree model (ρ=0).

	RANDOM	SHAP	BF-SHAP	SHAPR	LIME	MAPLE	L2X
faith+(↑)	$-0.028_{\pm 0.022}$	$\mathbf{0.922}_{\pm 0.020}$	$0.887_{\pm 0.031}$	$0.918_{\pm 0.039}$	$0.859_{\pm 0.035}$	$0.626_{\pm 0.050}$	$-0.004_{\pm 0.100}$
faith-(↑)	$-0.022_{\pm 0.023}$	$0.970_{\pm 0.006}$	$0.937_{\pm 0.017}$	$\mathbf{0.977}_{\pm 0.004}$	$0.918_{\pm 0.010}$	$0.647_{\pm 0.045}$	$0.002_{\pm 0.080}$
mono+(↑)	$0.538_{\pm 0.012}$	$\mathbf{0.720}_{\pm 0.018}$	$0.676_{\pm 0.027}$	$0.719_{\pm 0.019}$	$0.667_{\pm 0.032}$	$0.712_{\pm 0.008}$	$0.562_{\pm 0.024}$
mono-(↑)	$\mathbf{0.467}_{\pm 0.006}$	$0.433_{\pm 0.019}$	$0.449_{\pm 0.027}$	$0.435_{\pm 0.012}$	$0.428_{\pm 0.014}$	$0.440_{\pm 0.017}$	$0.430_{\pm 0.040}$
roar-faith+(↑)	$0.003_{\pm 0.028}$	$0.461_{\pm 0.095}$	$0.496_{\pm 0.016}$	$0.468_{\pm 0.082}$	$\mathbf{0.585}_{\pm 0.046}$	$-0.429_{\pm 0.018}$	$0.045_{\pm 0.060}$
roar-faith-(↑)	$0.008_{\pm 0.049}$	$0.581_{\pm 0.024}$	$0.535_{\pm 0.067}$	$0.559_{\pm 0.026}$	$\mathbf{0.621}_{\pm 0.019}$	$-0.339_{\pm 0.013}$	$0.052_{\pm 0.038}$
roar-mono+(↑)	$0.474_{\pm 0.016}$	$0.747_{\pm 0.028}$	$\mathbf{0.771}_{\pm 0.015}$	$0.730_{\pm 0.022}$	$0.707_{\pm 0.024}$	$0.425_{\pm 0.009}$	$0.500_{\pm 0.027}$
roar-mono-(↑)	$0.492_{\pm 0.019}$	$0.721_{\pm 0.032}$	$0.683_{\pm 0.038}$	$0.713_{\pm 0.044}$	$\mathbf{0.745}_{\pm 0.020}$	$0.471_{\pm 0.016}$	$0.451_{\pm 0.041}$
shapley-corr(↑)	$0.001_{\pm 0.014}$	$0.992_{\pm 0.005}$	$0.956_{\pm 0.007}$	$\mathbf{0.998}_{\pm 0.001}$	$0.955_{\pm 0.009}$	$0.735_{\pm 0.038}$	$0.073_{\pm 0.084}$
shapley-mse(↓)	$1.134_{\pm 0.040}$	$0.003_{\pm 0.001}$	$0.008_{\pm 0.001}$	$\mathbf{0.000}_{\pm 0.000}$	$0.026_{\pm 0.001}$	$0.071_{\pm 0.007}$	$0.188_{\pm 0.022}$

- No single explainer outperforms the rest consistently across metrics and ML models
- Explainers are generally more effective in explaining linear models
- Explainer performance drop as features become more correlated
- MAPLE failed on faith- with both decision tree and MLP models by often predicting important features as least important



Explanation faith- for three types of ML models: linear regression, decision tree, MLP

Real-world Dataset Simulation

- Wine dataset:
  - 11 continuous features, 1 categorical output, ~5000 data points.
- Simulation process:
  - Compute empirical covariance matrix
  - Use the covariance matrix to generate features
  - Use a k-nearest neighbor model to generate labels
- Validation process:
  - Compute Jensen-Shannon distance between real and synthetic datasets
  - Train ML models on either real or simulated wine dataset
  - Generate explanations for ML models trained on real data, simulated data
  - Compute mean squared error between the two sets of explanations

Mean squared error between explanations for predictions of models trained on real and simulated wine dataset.

Model	SHAP	LIME	MAPLE	L2X	Random
Linear	0.028 ± 0.009	0.047 ± 0.016	0.027 ± 0.009	0.0009 ± 0.0001	
Tree	0.047 ± 0.003	0.009 ± 0.001	0.052 ± 0.012	0.0008 ± 0.0001	1.988 ± 0.001
MLP	0.028 ± 0.003	0.037 ± 0.008	0.040 ± 0.002	0.0008 ± 0.0001	

Explainer performance on simulated wine dataset across metrics.

	RANDOM	SHAP	LIME	MAPLE	L2X
faith- (↑)	$0.012_{\pm 0.011}$	$\mathbf{0.461}_{\pm 0.034}$	$0.237_{\pm 0.031}$	$-0.007_{\pm 0.036}$	$-0.010_{\pm 0.032}$
faith+ (↑)	$0.025_{\pm 0.038}$	$0.488_{\pm 0.023}$	$\mathbf{0.595}_{\pm 0.022}$	$0.556_{\pm 0.021}$	$0.055_{\pm 0.035}$
mono- (↑)	$0.490_{\pm 0.004}$	$0.502_{\pm 0.010}$	$0.500_{\pm 0.013}$	$\mathbf{0.506}_{\pm 0.011}$	$0.492_{\pm 0.001}$
mono+ (↑)	$0.523_{\pm 0.010}$	$\mathbf{0.556}_{\pm 0.012}$	$0.539_{\pm 0.005}$	$0.513_{\pm 0.008}$	$0.522_{\pm 0.008}$
shapley-corr (↑)	$0.011_{\pm 0.027}$	$\mathbf{0.815}_{\pm 0.024}$	$0.692_{\pm 0.019}$	$0.669_{\pm 0.007}$	$0.035_{\pm 0.055}$
shapley-mse (↓)	$1.032_{\pm 0.022}$	$\mathbf{0.014}_{\pm 0.003}$	$0.032_{\pm 0.005}$	$0.041_{\pm 0.001}$	$0.055_{\pm 0.001}$

Conclusions

- Synthetic datasets enable:
  - Quantitative evaluation of feature attribution methods
  - Simulation of real datasets and benchmark explainers
- The best choice of explainer depends on metrics, ML model, and dataset type
- Some explainers fail in unexpected ways
- GitHub: https://github.com/abacusai/xai-bench
- Full workshop paper: https://arxiv.org/abs/2106.12543