pureGAM: Learning an Inherently Pure Additive Model
(KDD 2022)

Samsung Software Developer Community
Korea Vision & Robotics
Eunjung Choi
2023.08.26

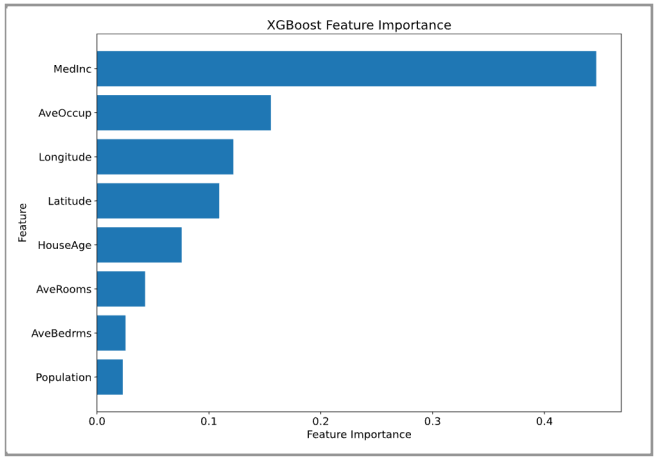
Background: XAI methods in Tabular data

- **설명가능성(Interpretability)**

  AI 모델의 예측 결과를 사용한 변수(feature) 혹은 변수-상호작용(feature-interaction)으로 설명할 수 있는 능력

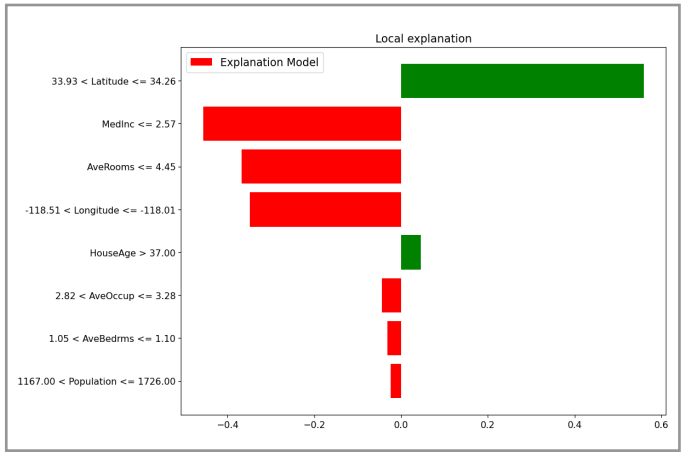
- **단일 변수 기반 설명가능성의 한계**
  - **제한된 설명가능성**: 단일 변수의 feature importance 만으로는 예측 결과에 대한 완전한 설명가능성을 제시하기 어려움
  - **변수-상호작용 무시**: 실제 모델의 예측 결과는 변수들 간의 복잡한 상호작용으로 이루어질 수 있음

Model-intrinsic XAI

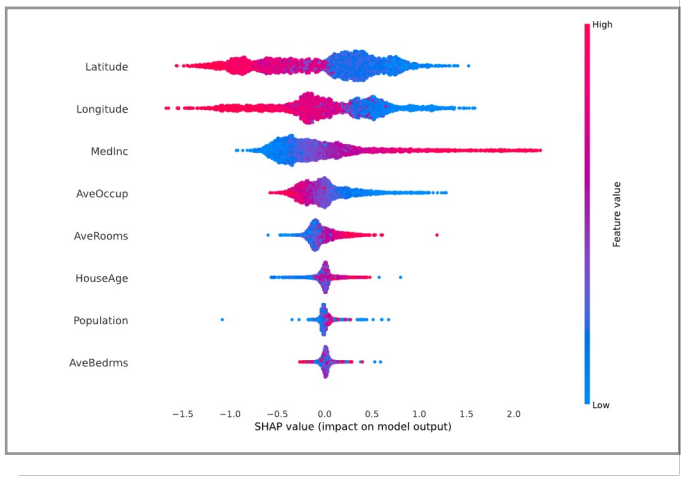


XGBoost Feature Importance

Post-hoc XAI



LIME



SHAP

Background: Generalized Additive Model (GAM)

- **Linear Model:** $y \sim N$ 일 때 효과적으로 적합
- **GLM (Generalized Linear Model):** $y \sim$ some exponential family distribution 을 위해 link function $g(\cdot)$ 을 사용하는 선형 모델
- **Additive Model:** 비선형성을 포착하기 위해 변수와 모수(parameter)의 비선형 결합을 허용하는 shape function 사용

  개별 변수에 대한 single model $f_i(x_i)$ 의 합으로 예측 결과를 나타내므로 additive model(가법 모델) 이라고 함

  ex. $y = x_1 + x_2^2 + \sqrt{x_3} + \log(x_4) + \exp(x_5) + \epsilon, \text{where } \epsilon \sim N(0,1)$

- Additive model들은 비선형성을 포착하면서, 예측 결과를 개별 변수에 대해서 분해할 수 있기 때문에 accuracy와 interpretability를 동시에 챙길 수 있음

Model	Form	Intelligibility	Accuracy
Linear Model	$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$	+++	+
Generalized Linear Model	$g(y) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$	+++	+
Additive Model	$y = f_1(x_1) + ... + f_n(x_n)$	++	++
Generalized Additive Model	$g(y) = f_1(x_1) + ... + f_n(x_n)$	++	++
Full Complexity Model	$y = f(x_1, ..., x_n)$	+	+++

- XGBoost
- Random Forest
- Deep Neural Network

Table 1: From Linear to Additive Models.

Accuracy와 Interpretability의 trade-off 관계

Intelligible Models for Classification and Regression (KDD 2012)

Background: Generalized Additive Model (GAM)

- **GAM (Generalized Additive Model)**

$$g(y) = \sum_{i=1}^{p} f_i(x_i) = f_1(x_1) + \cdots + f_p(x_p)$$

  - g 가 identity function 이면 $\mathrm{E}(Y) = \sum f_i(x_i) \Rightarrow$ regression task $\quad * g(y) = g(E[Y|x_1, \cdots, x_p])$
  - g 가 logistic function 이면 $\log\frac{1}{1-p} = \sum f_i(x_i) \Rightarrow$ classification task
  - shape function 을 사용해 feature와 target간의 비선형적 관계를 잘 포착하여 linear model보다 성능이 뛰어나고 개별 변수 연산의 합으로 예측결과를 나타내므로 complex model 보다 설명가능성 역량 또한 뛰어남
  - 하지만 변수간 interaction을 나타내지는 않음

- **GA^{2}M (Generalized Additive Model with pairwise interactions)**

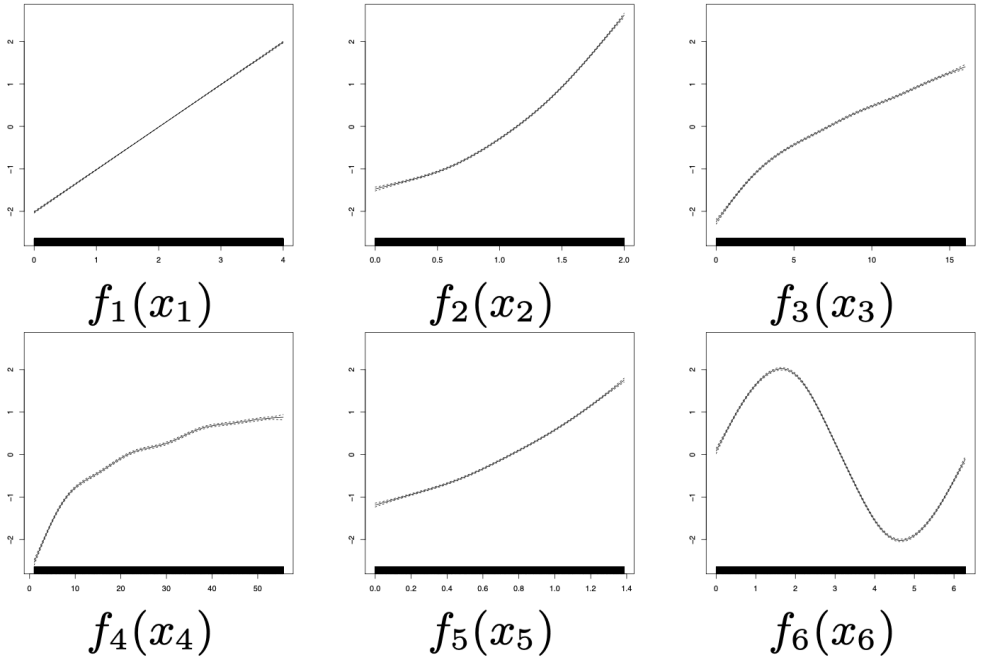
$$g(y) = \sum_{u \in D_2} f_u(x_u) = \sum f_i(x_i) + \sum f_{i,j}(x_i, x_j)$$

$* D_2$: 집합 $\{1, \cdots, p\}$ 의 모든 부분집합들 중 원소의 개수가 2 이하인 집합들의 모임

  - GAM에 pairwise interaction term을 추가해 변수간 interaction까지 고려

Background: Generalized Additive Model (GAM)

- **GAM의 설명가능성**
  - shape function을 시각화하여 예측결과에 미치는 개별 변수의 기여도를 직관적으로 알 수 있음
  - ex. $y = x_1 + x_2^2 + \sqrt{x_3} + \log(x_4) + \exp(x_5) + \epsilon, \text{where } \epsilon \sim N(0,1)$ 를 모형으로 적합시켰을 때,
    - y축: $f_i(x_i)$
    - x축: x_i



Problem Statement

- **Identifiability challenge (식별 가능성)**

$$g(y) = f_1(x_1) + f_{1,2}(x_1, x_2)$$
$$if, h_{1,2}(x_1, x_2) := f_1(x_1) + f_{1,2}(x_1, x_2)$$
$$g(y) = 0 + h_{1,2}(x_1, x_2)$$

- Additive model에서 각 shape function의 영향력을 이해하기 위해서는 각 shape function들이 고유하게 식별 가능해야함
- 주어진 예시에서 interaction과 main effect를 합친 $h_{1,2}$ 를 사용해서 모델의 같은 예측을 다르게 분해할 수 있음
- 이러한 비고유성(non-uniqueness)은 모델의 해석가능성을 저해하는 문제

- 일반적으로 shape function에 제약을 주는 방식으로 identifiability 달성
- 하지만 선행 연구들은 본 논문에서 제시하는 identifiability 조건 (pureness condition) 을 만족하지 못함

pureGAM Overview

- **pureGAM an inherently pure additive model of both main effects and higher-order interactions**

  - **Identifiability**: shape function에 pureness condition이라는 제약 조건을 걸어 identifiability challenge 해결
  - **Simplicity (설명가능성의 관점에서):**
    - (Occam's razor) 동일한 결과를 내는 경우, 더 간단하거나 더 적은 가정을 필요로 하는 설명이나 가설이 선호되어야 한다
    - pureness condition으로 f_u 가 중첩된 저차원의 $f_v(v \subset u)$ 를 흡수할 수 없도록 함
    - 즉, pureness condition 하에서 $f_{1,2,3}(x_1, x_2, x_3)$ 은 $f_{1,2}(x_1, x_2)$ 에 대해서 pure 하다고 말함 (식별 가능함)
    - 동일한 성능을 낸다면 더 간단한 component가 해석하기 쉬움
  - **Unification:** pureness condition을 만족하는 pure coding method로 categorical 과 numerical feature 모두 지원
  - **Optimality:** 유일해이자 최적해를 가지는 새로운 training method, joint learning strategy 도입

Models	pureGAM($k = 3$)	pureGAM($k = 2$)	GAMI-Net	EBM	XGBoost
RMSE	**0.0128 ± 0.0019**	0.0415 ± 0.0012	0.0417 ± 0.0005	0.0447 ± 0.0018	0.0227 ± 0.0015

Performance results with higher-order interactions

pureGAM Formula

Model	Formula	
GAM	$g(y) = \sum_{i=1}^{p} f_i(x_i)$	단변량 변수의 연산합으로 prediction 도출
GA^{k}M	$g(y) = \sum_{u \in \mathcal{D}_k} f_u(x_u)$	GAM에 k차 interaction term 추가
pureGAM	$g(y) = \sum_{u \in \mathcal{D}_k} f_u(x_u)$ s.t. $\forall u \in \mathcal{D}_k, f_u(x_u)$ is pure	GA^{k}M에 pureness condition(제약) 추가 → $f_u(x_u)$ is identifiable

(1) $g(y) = g\left(\mathbb{E}[Y|x_1, \cdots, x_p]\right)$.

(2) p: The total number of features.

(3) $\mathcal{D}_k$: The set of all non-empty subsets of $\{1, \cdots, p\}$ with cardinality $\leq k$.

(4) The pureness (hierarchical orthogonality) condition of f_u:

$$\int f_u(x_u) h_v(x_v) w(x) dx = 0, \forall v \subset u, \forall h \in \mathcal{L}^2(\mathbb{R}^v).$$

pureGAM Modeling

- **pureGAM**

$$E(Y|\boldsymbol{x}) = \sum_{u \in D_k} f_u(x_u) \quad \text{s.t.} \; \forall u \in D_k, f_u(x_u) \text{ is pure.}$$

- **pureness condition**
  - 모든 함수는 weighted L^2 space의 함수 → 함수간 내적이 가능한 유일한 공간이기 때문
  - $k-$variate subfunction으로 분해할 때 x의 확률 분포 $w(x)$를 가중치로 고려

DEFINITION 1. *We define "$f_u(x_u)$ is pure", if*

$$\forall v \subset u, \forall h_v \in \mathcal{L}^2(\mathbb{R}^v), \int f_u(x_u) h_v(x_v) w(x) dx = 0. \qquad (2)$$

In other words, f_u is orthogonal to any subfunction h_v (i.e. $v \subset u$) w.r.t. inner product $\langle f, g \rangle_w := \int f(x) g(x) w(x) dx$.

  - f_u 가 모든 부분함수들 h_v 와 직교하면 f_u is pure (identifiable)

pureGAM Modeling

- **pureGAM**

$$E(Y|\boldsymbol{x}) = \sum_{u \in D_k} f_u(x_u) \quad \text{s.t.} \; \forall u \in D_k, f_u(x_u) \text{ is pure.}$$

- **pureness condition**
  - $u \neq i$ 인 x_i 에 대해서 적분했을 때 0이 되어야함

DEFINITION 2. *We say $f_u(x_u)$ is pure if and only if:*

$$\forall i \in u, \int f_u(x_u) w_u(x_u) dx_i = 0. \qquad (3)$$

pureGAM Modeling

- **pureGAM**

$$E(Y|\boldsymbol{x}) = \sum_{u \in D_k} f_u(x_u) \quad \text{s.t.} \forall u \in D_k, f_u(x_u) \text{ is pure.}$$

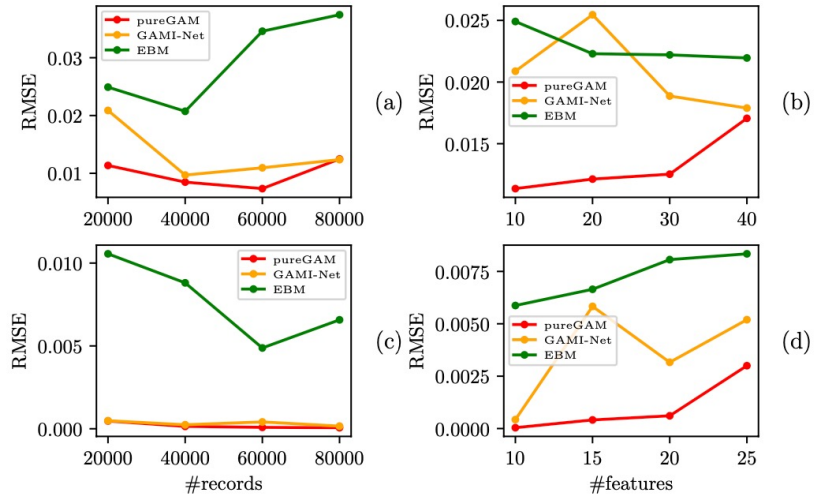
- **pureGAM optimization**
  - g: identity function

$$\{f_u(x_u)\,|u \in \mathcal{D}_k\}$$

$$= \text{argmin}_{\{g_u \in \mathcal{L}^2\}_{u \in \mathcal{D}_k}} \int \left(\sum_{u \in \mathcal{D}_k} g_u(x_u) - \mathbb{E}(Y|x) \right)^2 w(x)\,dx$$

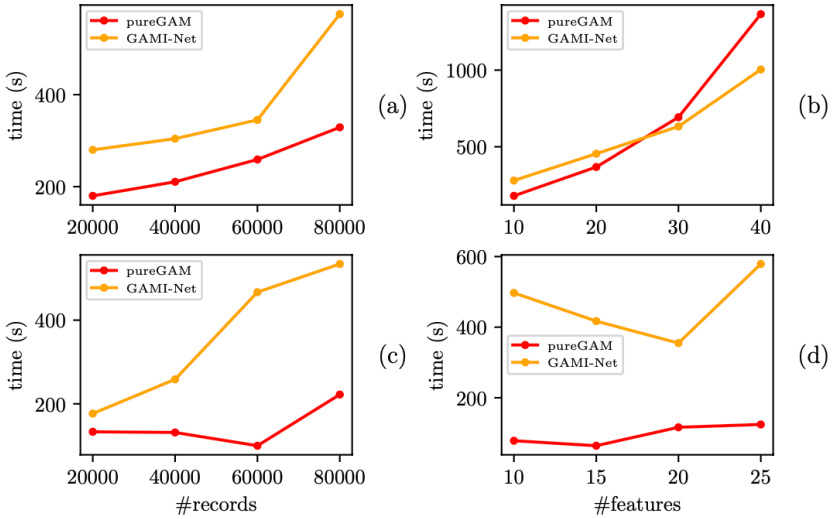
$$s.t. \forall u \in \mathcal{D}_k, \forall i \in u, \int f_u(x_u)\, w_u(x_u)\, dx_i = 0 \qquad (4)$$

Evaluation



(a)(c): Varying #records with #features=10; (b)(d): Varying #features with #records=20,000;
(a)(b): features are all numerical; (c)(d): features are all categorical.

Figure 3: Performance of pureGAM/GAMI-Net/EBM on synthetic datasets.



(a)(c): Varying #records with #features=10; (b)(d): Varying #features with #records=20,000;
(a)(b): features are all numerical; (c)(d): features are all categorical.

Figure 5: Results of training time on synthetic data.

Models	pureGAM($k = 3$)	pureGAM($k = 2$)	GAMI-Net	EBM	XGBoost
RMSE	**0.0128 ± 0.0019**	0.0415 ± 0.0012	0.0417 ± 0.0005	0.0447 ± 0.0018	0.0227 ± 0.0015