AI モデルの説明可能性 Shapley 値からの属性推定リスクの評価とその対策

當麻 僚太郎 †

明治大学総合数理学部 先端メディアサイエンス学科 菊池研究室 †

1 はじめに

近年，機械学習モデルは金融や雇用などの重要な領域で活用されることが増えている [1, 3]．多くのモデルはニューラルネットワークやアンサンブルモデルなどの複雑な構造を持つため，入力に対する挙動がブラックボックスである．そのため，モデルの公平性や透明性を保証し，モデルの出力に対して説明を与えるための説明可能性技術 eXplainable AI（XAI）が注目されている [1, 2]．

機械学習モデルを用いた商品サービスを提供する基盤である Machine Learning as a Service（以下，MLaaS）プラットフォームでは，様々な説明可能性技術を用いた説明を提供している．特に Shapley 値 [12] を基にした説明は，Amazon Web Services [4] や Microsoft Azure [5] などの主要な MLaaS プラットフォームで提供されている．例えば，図 1 では Amazon SageMaker Studio [6] を用いて各入力ベクトルに対する Shapley 値ベクトルと特徴量ごとに Shapley 値の絶対値を平均した大域的な説明を得ている．

しかし，2022 年に Luo ら [7] は Shapley 値に基づく説明から本来秘匿されているモデルへの入力属性を推論出来ることを示した．Luo ら [7] は，最小勾配法による属性推定アルゴリズム ψ を提案している．一方，説明モデル f に多くの機械学習アルゴリズムがあるように，ψ にも多くの可能性がある．特に，モデル f と ψ の間には相関があり，属性推定リスクの評価は自明ではない．

そこで，本研究では，Luo ら [7] の手法を基にして，各説明変数と目的変数間の相関や攻撃者が採用するアルゴリズム ψ の違いに対してどのように属性推定リスクが変化するかを明らかにする．特に，f と ψ が線形回帰モデルのときには，Shapley 値から正確にプライベートな特徴ベクトルの推定が可能であることを示す．また，Shapley 値と同様の局所的な説明手法である LIME[13] に対するプライバシーリスクを調査し，Luo ら [7] の攻撃手法が Shapley 値以外の説明可能性技術に対して有効であることを示す．この提案方式を，3 つのデータセット

†Kikuchi Laboratory, Department of Frontier Media Science, School of Interdisciplinary Mathematical Science, Meiji University.

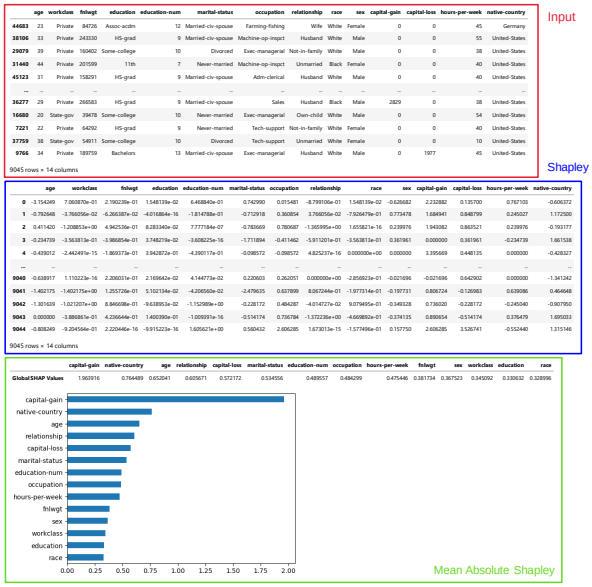


図 1: Amazon SageMaker を用いた Shapley 値の計算例

Adult [9]，Bank Marketing [10], Credit Card Client [11] について適用した結果を報告する．

2 基本定義

2.1 Shapley 値

Shapley 値 [12] は協力ゲーム理論において連携プレイヤー間で利益を分配するための協調作業を定量化するために，1953 年に Shapley に提案された指標である．本研究では，n 特徴の入力 $\boldsymbol{x} = (x_1, \ldots, x_n)$ に対するモデルの出力 $f(\boldsymbol{x})$ の局所的な説明として Shapley 値ベクトル $\boldsymbol{s} = (s_1, \ldots, s_n)$ を与える．

特徴量のインデックス集合を $N = \{1, 2, \ldots, n\}$, N の部分集合を S, Shapley 値を計算するために参照するデータのサンプルを $\boldsymbol{x}^0$ とする．S に対応する入力サンプルを $\boldsymbol{x}_{[S]} = ((x_{[S]})_1, \ldots, (x_{[S]})_n)$ とする．ここで，$i = 1, \ldots, n$ について，

$$(x_{[S]})_i = \begin{cases} x_i & \text{if } i \in S, \\ x_i^0 & \text{otherwise.} \end{cases} \tag{1}$$

例えば，$\boldsymbol{x} = (\mathbf{2}, \mathbf{5}, \mathbf{1}, \mathbf{3})$，$\boldsymbol{x}^0 = (0, 3, 2, 1)$，$S = \{2, 3\}$ としたとき，$\boldsymbol{x}_{[S]} = [0, \mathbf{5}, \mathbf{1}, 1]$ である．このとき，特徴量

に対する Shapley 値 s_i は，

$$s_i = \frac{1}{n!} \sum_{S \subseteq N \setminus \{i\}} |S|!(n - |S| - 1)! \left(f(\boldsymbol{x}_{[S \cup \{i\}]}) - f(\boldsymbol{x}_{[S]})\right) \quad (2)$$

で定められる．$\boldsymbol{s} = \phi(\boldsymbol{x}; \boldsymbol{x}^0, f) = (s_1, \ldots, s_n)$ を Shapley 値を与える写像とする．

2.2 Shapley 値の計算例

入力サンプル $\boldsymbol{x} = (1.5, \text{True}, A)$ に対し，参照サンプル $\boldsymbol{x}^0 = (-0.4, \text{False}, B)$ を用いて Shapley 値を計算する例を示す．ここで，攻撃対象のモデル f は表 1 の通りに出力する．

表 1: 入力データに対するモデル f の出力

	x_1	x_2	x_3	$f(x)$
$\boldsymbol{x}$	1.5	True	A	0.8
$\boldsymbol{x}^0$	−0.4	False	B	0.6
$\boldsymbol{x}_{[\{1\}]}$	1.5	False	B	0.9
$\boldsymbol{x}_{[\{2\}]}$	−0.4	True	B	0.8
$\boldsymbol{x}_{[\{3\}]}$	−0.4	False	A	0.2
$\boldsymbol{x}_{[\{1,2\}]}$	1.5	True	B	1.0
$\boldsymbol{x}_{[\{1,3\}]}$	1.5	False	A	0.6
$\boldsymbol{x}_{[\{2,3\}]}$	−0.4	True	A	0.4

このとき，Shapley 値 $\boldsymbol{s} = (s_1, s_2, s_3)$ は式 (2) によりそれぞれ計算される．

$$\begin{aligned} s_1 = {} & \frac{0!(3-0-1)!}{3!}(f(\boldsymbol{x}_{[\{1\}]}) - f(\boldsymbol{x}_{[\{\}]})) \\ & + \frac{1!(3-1-1)!}{3!}(f(\boldsymbol{x}_{[\{1,2\}]}) - f(\boldsymbol{x}_{[\{2\}]})) \\ & + \frac{1!(3-1-1)!}{3!}(f(\boldsymbol{x}_{[\{1,3\}]}) - f(\boldsymbol{x}_{[\{3\}]})) \\ & + \frac{2!(3-2-1)!}{3!}(f(\boldsymbol{x}_{[\{1,2,3\}]}) - f(\boldsymbol{x}_{[\{2,3\}]})) \quad (3) \\ = {} & \frac{1}{3}(0.9 - 0.6) + \frac{1}{6}(1.0 - 0.8) \\ & + \frac{1}{6}(0.6 - 0.2) + \frac{1}{3}(0.8 - 0.4) \\ = {} & \frac{2}{6} \approx 0.33 \end{aligned}$$

s_2，s_3 も式 (3) と同様にして，$\boldsymbol{s} = (0.33, 0.18, -0.32)$ と計算される．$\boldsymbol{s}$ の総和はおよそ 0.2 であり，これは $f(\boldsymbol{x}) - f(\boldsymbol{x}^0)$ に等しい．正の Shapley 値 s_1，s_2 に対応する属性 x_1，x_2 はモデル f の出力を増加させるように働き，負の Shapley 値 s_3 に対応する属性 x_3 はモデル f の出力を減少させるように働くことを示す．

2.3 LIME

Local Interpretable Model-agnostic Explanations (LIME) は，Ribeiro ら [13] によって提案された，Shapley 値と同様に入力サンプルごとに説明を生成する手法である．n 特徴の入力 $\boldsymbol{x} = (x_1, \ldots, x_n)$ が与えられたとき，その周辺のデータに対するモデル f のふるまいを，線形モデルや決定木，ルールベースなどの解釈が容易なモデル g で近似する．本研究では説明モデル g に線形モデル $g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$ を採用しているものとし，その係数ベクトル $\boldsymbol{w}$ が説明ベクトルとして与えられるものとする．

説明モデル g を学習するための損失関数 $\mathcal{L}$ は

$$\mathcal{L}(f, g, \pi_{\boldsymbol{x}}) = \sum_{\boldsymbol{z}, \boldsymbol{z}' \in Z} \pi_{\boldsymbol{x}}(\boldsymbol{z})(f(\boldsymbol{z}) - g(\boldsymbol{z}'))^2$$

と定義される．ここで，$\pi_{\boldsymbol{x}}(\boldsymbol{z}) = e^{-D(\boldsymbol{x},\boldsymbol{z})^2/\sigma^2}$ は距離 $D(\boldsymbol{x}, \boldsymbol{z})$ に応じた重みであり，σ はそのパラメタである．$\boldsymbol{z}$ は $\boldsymbol{x}$ と同じ n 次元ベクトルであり，$\boldsymbol{z}'$ は $\boldsymbol{z}$ の一部の特徴量を抜き出したベクトルである．Z は $\boldsymbol{z}$ と $\boldsymbol{z}'$ の組の集合である．

モデル g の取りうる集合を G，$\boldsymbol{w}$ の非ゼロ要素の数を $\Omega(g)$ としたとき，説明モデル g は目的関数 $\xi(\boldsymbol{x}) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_{\boldsymbol{x}}) + \Omega(g)$ を最小化する g^* が学習される．

2.4 LIME の計算例

Shapley 値の計算例と同様に，入力サンプル $\boldsymbol{x} = (1.5, \text{True}, A)$ に対して LIME を計算する例を示す．表 2 は，説明モデル g を学習するために用いるデータ $\boldsymbol{z}$，$\boldsymbol{z}'$ である．

表 2: 説明モデル g の学習に用いるデータ $\boldsymbol{z}$

	z_1	z_2	z_3	z'_1	z'_2	z'_3	$f(\boldsymbol{z})$
$\boldsymbol{z}^1$	1.5	True	A	1.5	1	1	0.8
$\boldsymbol{z}^2$	−0.4	False	B	−0.4	0	0	0.6
$\boldsymbol{z}^3$	0.1	False	A	0.1	0	1	0.3
$\boldsymbol{z}^4$	0.8	True	C	0.8	1	0	0.9
$\boldsymbol{z}^5$	−1.1	True	A	−1.1	1	1	0.2

ここで，距離を求める関数を

$$D(\boldsymbol{x}, \boldsymbol{z}) = \sqrt{(x_1 - z_1)^2 + |x_2 = z_2|^2 + |x_3 = z_3|^2}$$

とし，$\pi_{\boldsymbol{x}}(\boldsymbol{z})$ のパラメタ $\sigma = 2$ とする．説明モデル g を線形モデル $g(\boldsymbol{z}') = w_1 z'_1 + w_2 z'_2 + w_3 z'_3 + b$ とすると，損

失関数は

$$\begin{aligned}\mathcal{L}(f,g,\pi_{\boldsymbol{x}}) &= \pi_{\boldsymbol{x}}(z^1)(f(z^1)-g(z'^1))^2 \\ &\quad + \pi_{\boldsymbol{x}}(z^2)(f(z^2)-g(z'^2))^2 \\ &\quad + \pi_{\boldsymbol{x}}(z^3)(f(z^3)-g(z'^3))^2 \\ &\quad + \pi_{\boldsymbol{x}}(z^4)(f(z^4)-g(z'^4))^2 \\ &\quad + \pi_{\boldsymbol{x}}(z^5)(f(z^5)-g(z'^5))^2 \\ &= (1.5w_1+w_2+w_3+b-0.8)^2 \\ &\quad + 0.2(-0.4w_1+b-0.6)^2 \\ &\quad + 0.5(0.1w_1+w_3+b-0.3)^2 \\ &\quad + 0.7(0.8w_1+w_2+b-0.9)^2 \\ &\quad + 0.2(-1.1w_1+w_2+w_3+b-0.2)^2 \end{aligned} \tag{4}$$

である．ここで，w_1，w_2，w_3，b で $\mathcal{L}(f,g,\pi_{\boldsymbol{x}})$ を偏微分して，

$$\frac{\partial}{\partial w_1}\mathcal{L}(f,g,\pi_{\boldsymbol{x}}) = 5.94w_1+3.66w_2+2.66w_3+3.6b-3.24 = 0$$

$$\frac{\partial}{\partial w_2}\mathcal{L}(f,g,\pi_{\boldsymbol{x}}) = 3.68w_1+3.8w_2+2.4w_3+3.8b-2.94 = 0$$

$$\frac{\partial}{\partial w_3}\mathcal{L}(f,g,\pi_{\boldsymbol{x}}) = 2.66w_1+2.4w_2+3.4w_3+3.4b-1.98 = 0$$

$$\frac{\partial}{\partial b}\mathcal{L}(f,g,\pi_{\boldsymbol{x}}) = 3.62w_1+3.8w_2+3.4w_3+5.2b-3.48 = 0$$

このとき，$w_1 \approx 0.23, w_2 \approx 0.13, w_3 \approx -0.30, b \approx 0.61$ であり，説明ベクトルとして $\boldsymbol{w} = (0.23, 0.13, -0.30)$ が得られる．

3 基本原理

3.1 Feature Inference Attack on Shapley Values [7]

3.1.1 システムモデル

Luo ら [7] は，サービス事業者が機密の学習データセット $\mathcal{X}_{train}$ に基づいてブラックボックスモデル f を訓練し，MLaaS プラットフォーム上に展開するシステムモデルを仮定している．その実験概要図を図 2 に示す．

ユーザはプライベートな入力サンプル $\boldsymbol{x}$ を送信し，モデルの出力 $\hat{\boldsymbol{y}} = (y_1, \ldots, y_c)$ と n 個の説明値のベクトル $\boldsymbol{s} = (s_1, \ldots, s_n)$ を得る．ただし，c は正解ラベルの数である．$c > 2$ のとき対応する説明ベクトルは本来 c 個分得られるが，ここでは $c = 1$ とする．

3.1.2 攻撃者

攻撃者は $\mathcal{X}_{train}$ と同じ分布に従う補助データセット $\mathcal{X}_{aux}$ を持っていると仮定する．全ての $\boldsymbol{x}_{aux} \in \mathcal{X}_{aux}$ をモデル f に送信し，対応する説明データ $\mathcal{S}_{aux}$ を得る．そして，$\psi : \mathcal{S}_{aux} \to \mathcal{X}_{aux}$ が誤差 $L(\psi(\mathcal{S}_{aux}), \mathcal{X}_{aux})$ を最小化するように訓練する．プライベートな入力 $\boldsymbol{x}$ の推測値は，与えられた Shapley 値 $\boldsymbol{s}$ を用いて $\hat{\boldsymbol{x}} = \psi(\boldsymbol{s})$ とする．攻撃者の属性推論を図 3 に示す．このアルゴリズムを Algorithm 1 に示す．

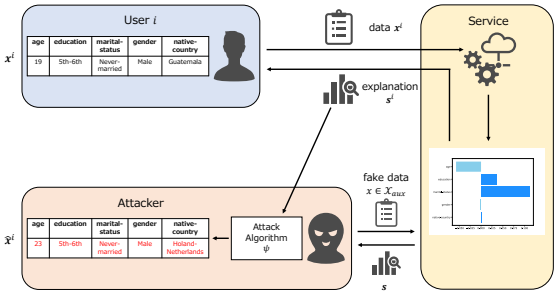


図 2: 全体概要図

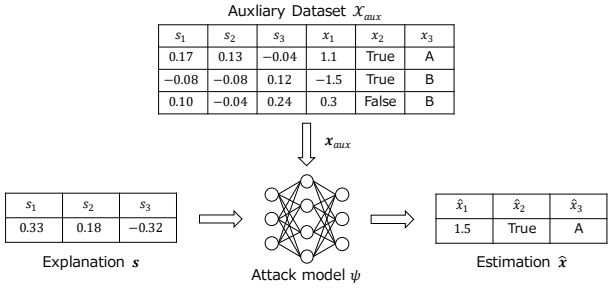


図 3: 属性推論攻撃の概要図

3.1.3 属性推論攻撃の例

10 行 5 列のサンプルデータに対して属性推論攻撃を行う例を示す．データの生成は，標準正規分布に従う独立な 3 つの乱数列 n_1，n_2，n_3 を用いて，$x_1 = n_1$，$x_2 = n_2$，$x_3 = n_1 n_2$，$x_4 = n_2 n_3$，$y = x_1 - x_3 x_4$ とする．生成したデータを表 3 に示す．

データの 1〜5 行目を $\mathcal{X}_{test}$，6〜10 行目を $\mathcal{X}_{train}$ とする．Shapley 値は $\mathcal{X}_{train}$ の各行を参照サンプルとし，それぞれ求めた Shapley 値の平均，すなわち $\boldsymbol{s} = \frac{1}{5}\sum_{j=6}^{10}\phi(\boldsymbol{x}, \boldsymbol{x}^j)$ とする．

例として，$\mathcal{X}_{train}$ をフィッティングした線形回帰モデル f について，$\mathcal{X}_{test}$ に対する Shapley 値 $\mathcal{S}_{test}$ を表 4 に示す．モデル f と推定アルゴリズム ψ の組み合わせに対する MAE を表 5 に示す．f と ψ がどちらも線形モデルのとき，誤差 0 で正確に入力特徴を推論出来た．

Algorithm 1 補助データセットを用いた推定 [7]

Input: ブラックボックスモデル f，補助データセット $\mathcal{X}_{aux}$，学習率 α，攻撃対象の Shapley 値ベクトル $\boldsymbol{s}$

Output: 推論されたプライベートな入力 $\hat{\boldsymbol{x}}$

$\mathcal{S}_{aux} \leftarrow \phi(\mathcal{X}_{aux}; f)$
$\theta_\psi \leftarrow \mathcal{N}(0, 1)$
for each epoch **do**
  for each batch **do**
    $loss \leftarrow 0$
    $B \leftarrow$ randomly select a batch of samples
    for $j \in 1, \ldots, |B|$ **do**
      $(\hat{\boldsymbol{x}}_{aux})^j \leftarrow \psi((\boldsymbol{s}_{aux})^j; \theta_\psi)$
      $loss \leftarrow loss + L((\hat{\boldsymbol{x}}_{aux})^j, (\boldsymbol{x}_{aux})^j)$
    end for
    $\theta_\psi{}' \leftarrow \theta_\psi - \alpha \nabla_{\theta_\psi} loss$
  end for
end for
$\hat{\boldsymbol{x}} \leftarrow \psi(\boldsymbol{s}; \theta_\psi)$
return $\hat{\boldsymbol{x}}$

表 3: サンプルデータ

	x_1	x_2	x_3	x_4	y
$\mathcal{X}_{test}$	1.8	0.1	0.3	−0.4	1.9
	0.4	1.5	0.6	1.0	−0.2
	1.0	0.8	0.7	0.7	0.5
	2.2	0.1	0.3	−0.1	2.2
	1.9	0.4	0.8	1.0	1.1
$\mathcal{X}_{train}$（$\mathcal{X}_{aux}$）	−1.0	0.3	−0.3	−0.5	−1.2
	1.0	1.5	1.4	0.1	0.9
	−0.2	−0.2	0.0	0.0	−0.2
	−0.1	0.3	0.0	0.5	−0.1
	0.4	−0.9	−0.4	−1.3	−0.1

3.2 線形モデルに対する説明可能性と属性推定リスク

補題 1. f を線形回帰による説明モデルとする．任意の $i \in N$，$S \subseteq N \setminus \{i\}$，参照ベクトル $(x_1^0, x_2^0, \ldots, x_n^0)$ について，

$$f(\boldsymbol{x}_{[S\cup\{i\}]}) - f(\boldsymbol{x}_{[S]}) = \beta_i(x_i - x_i^0) \tag{5}$$

である．

証明) 説明モデル f を $f(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$ と表

表 4: Shapley 値 $\mathcal{S}_{test}$

	s_1	s_2	s_3	s_4
$\boldsymbol{x}^1$	1.30	0.02	0.06	−0.04
$\boldsymbol{x}^2$	0.28	−0.29	0.18	0.34
$\boldsymbol{x}^3$	0.72	−0.13	0.21	0.26
$\boldsymbol{x}^4$	1.59	0.02	0.06	0.04
$\boldsymbol{x}^5$	1.37	−0.04	0.25	0.34

表 5: モデル f と推定アルゴリズム ψ の組み合わせに対する MAE

		MAE				
f	ψ	x_1	x_2	x_3	x_4	平均
線形回帰	線形回帰	0.00	0.00	0.00	0.00	0.00
線形回帰	決定木	0.82	1.24	0.74	1.18	1.00
決定木	線形回帰	0.69	0.52	0.41	0.53	0.54
決定木	決定木	0.68	1.16	0.82	0.54	0.80
平均		0.55	0.73	0.49	0.59	

すと，線形モデルのため，

$$\begin{aligned} f(\boldsymbol{x}_{[S\cup\{i\}]}) - f(\boldsymbol{x}_{[S]}) &= \beta_0 + \sum_{k\in S\cup\{i\}} \beta_k x_k + \sum_{k\in N\setminus(S\cup\{i\})} \beta_k x_k^0 \\ &\quad - \beta_0 + \sum_{k\in S} \beta_k x_k + \sum_{k\in N\setminus S} \beta_k x_k^0 \\ &= \beta_i(x_i - x_i^0) \end{aligned}$$

□

命題 2. f を線形モデルによる説明モデル，ψ を線形モデルによる推定アルゴリズムとする．$n < |\mathcal{X}_{aux}|$ のとき，ψ による推定の MAE = 0 である．

証明) 補題 1 より，s_i について

$$\begin{aligned} s_i &= \frac{1}{n!} \sum_{S\subseteq N\setminus\{i\}} |S|!(n - |S| - 1)! f(\boldsymbol{x}_{[S\cup\{i\}]}) - f(\boldsymbol{x}_{[S]}) \\ &= \frac{1}{n!} \sum_{S\subseteq N\setminus\{i\}} |S|!(n - |S| - 1)! \beta_i(x_i - x_i^0) \\ &= \lambda_i(x_i - x_i^0) \end{aligned}$$

ここで $\lambda_i = \frac{1}{n!} \sum_{S\subseteq N\setminus\{i\}} |S|!(n - |S| - 1)!\beta_i$ をまとめた項である．したがって，ψ による推定モデルは，

$$\begin{aligned} \hat{x}_i &= \alpha_0 + \alpha_1 s_1 + \cdots + \alpha_n s_n \\ &= \alpha_0 + \alpha_1(\lambda_1(x_1 - x_1^0)) + \cdots + \alpha_n(\lambda_n(x_n - x_n^0)) \\ &= \alpha_0 - \sum_{k=1}^{n} \alpha_k \lambda_k x_k^0 + \alpha_1\lambda_1 x_1 + \cdots + \alpha_n\lambda_n x_n \\ &= \gamma_0 + \gamma_1 x_1 + \cdots + \gamma_n x_n \end{aligned}$$

と，$x_1, \ldots, x_n$ の線形式で与えられる．ただし，ここで $\gamma_i = \alpha_i\lambda_i$，$\gamma_0 = \alpha_0 - \sum_{k=1}^{n} \alpha_k\lambda_k x_k^0$ をまとめた項である．

表 6: 使用データセット

データセット	レコード数	クラス	特徴量
Adult [9]	48,842	2	14
Bank Marketing [10]	45,211	2	16
Credit Card [11]	30,000	2	24

$\mathcal{X}_{aux}$ が十分に大きく，$n+1$ 以上の行数があるならば，最小二乗法により，誤差なく $\gamma_1, \ldots, \gamma_n$ が算出される．□

結果のところ，線形式によるモデルの説明データを提供すると，属性推定リスクが上がることを意味している．

4 提案方式

想定する攻撃は先行研究と同様に，図 2 とする．実験に用いるデータセットを表 6 に示す．

この設定下において，説明モデル f や攻撃者が採用する最適化アルゴリズム，各説明変数と目的変数間の相関などに対する属性推定リスクを明らかにすることを目的とする．また，LIME に対する属性推定リスクを調べる．

4.1 評価指標

属性推定リスクの評価に用いる指標は，MAE と攻撃成功率 SR の 2 つである．

4.1.1 MAE

MAE（Mean Absolute Error），すなわち，$\ell_1$loss は誤差の絶対値の平均を取る．m 行 n 列のデータセット $\boldsymbol{x}$ に対する推定データ $\hat{\boldsymbol{x}}$ の MAE は

$$\ell_1(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} |\hat{x}_i^j - x_i^j| \tag{6}$$

で与えられる．

4.1.2 攻撃成功率 SR

SR（Success Rate）は攻撃によって正しく推定された入力特徴量の割合を表す．質的変数に対しては推定カテゴリが一致している時，量的変数に対しては推定値と真の値との誤差の絶対値が閾値以下である時成功と判定する．$\boldsymbol{x}$ と推定 $\hat{\boldsymbol{x}}$ の SR は

$$SR(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \frac{success(\hat{\boldsymbol{x}}, \boldsymbol{x})}{mn} \tag{7}$$

で与えられる．ここで，推定に成功した入力特徴の個数を $success(\hat{\boldsymbol{x}}, \boldsymbol{x})$ とする．

4.2 実験方法

4.2.1 実験 1

Shapley 値と LIME 値のブラックボックスモデル f に対する属性推定リスクを調べる．攻撃対象となるブラックボックスモデル f はニューラルネットワーク（NN），ランダムフォレスト（RF），勾配ブースティング木（GBDT），カーネル SVM（SVM）の 4 種類である．NN は Pytorch[8] で実装し，n 次元の入力層と c 次元の出力層を持ち，ニューロン数 $2n$ の隠れ層を 2 つ持つ．活性化関数は出力層のみ softmax でそれ以外は ReLU を用いる．RF，SVM，GBDT は sklearn で実装した．RF と GBDT の木の数と最大の深さは，それぞれ $(100, 5)$, $(100, 3)$ とする．その他のパラメータは全てデフォルトの値とする．RG-E は $\mathcal{X}_{aux}$ に基づく経験分布からランダムに予測したときの結果である．

攻撃者が持つデータセット $\mathcal{X}_{aux}$ の行数を変化させたときに，どのように MAE と SR が変化するかを調べる．Shapley 値に対する属性推論の結果を図 4，5 に，LIME に対する属性推論の結果を図 6，7 に示す．

4.2.2 実験 2

攻撃者が採用する最適化アルゴリズムを変化させたときの属性推定リスクを調査する．攻撃者の採用する推定アルゴリズムとは，攻撃者が Algorithm 1 において攻撃モデル ψ のパラメータ θ_ψ を更新する式 (8)

$$\theta_\psi \leftarrow \theta_\psi - \alpha \nabla_{\theta_\psi} loss \tag{8}$$

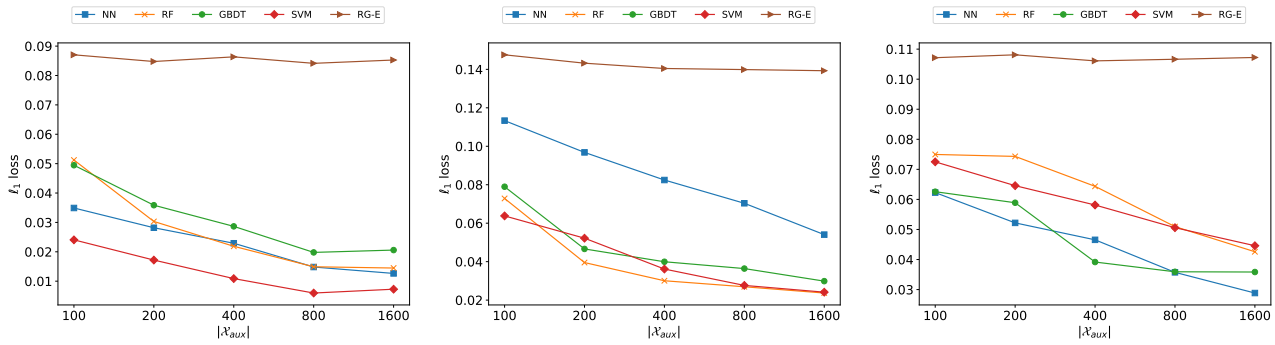
の変種を意味する．本研究では，勾配降下法ベースの最適化アルゴリズムとして，SGD [14]，Momentum [15]，RMSprop [16]，Adam [17] の 4 種類を調べる．推定モデル ψ は先行研究と同じく，特徴量の数 n に対して隠れ層のニューロン数 $4n$，出力層のニューロン数 n のニューラルネットワークとし，活性化関数は全てで sigmoid 関数を用いる．実装は Pytorch で行い，SGD の学習率 $\eta = 0.01$，Momentum（すなわち SGD のうち $momentum \neq 0$ のもの）の学習率 $\eta = 0.01$，$momentum = 0.9$ と指定したもの以外は全てデフォルトのパラメータを用いる．

実験の結果を図 8 に示す．

4.3 実験結果と考察

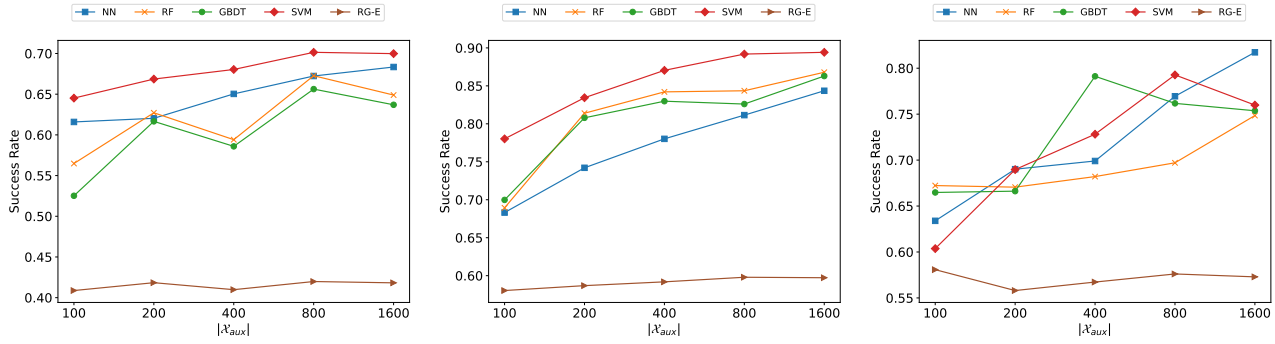
4.3.1 結果 1

図 4，5 において，データセットの行数 $|\mathcal{X}_{aux}|$ が増えるにしたがって $\ell_1$loss が下がり SR が上がる，すなわち，



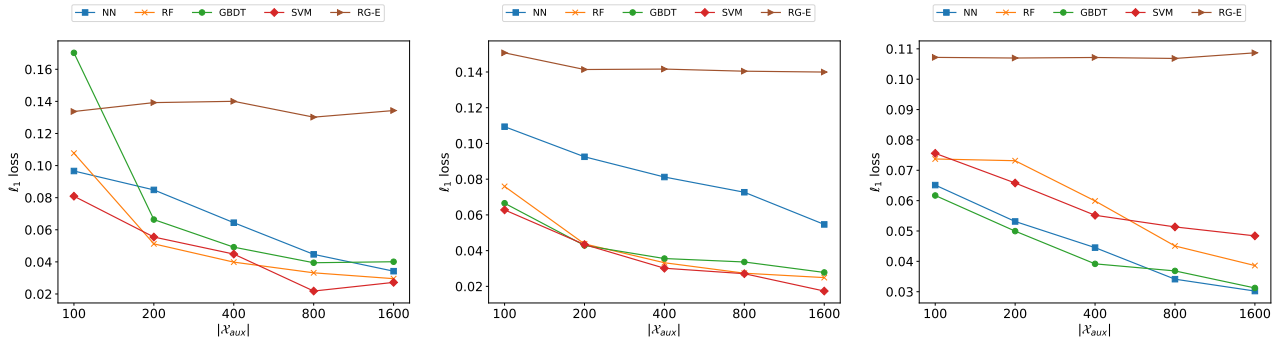
(a) Adult **(b)** Bank Marketing **(c)** Credit Card

図 4: 補助データセットの大きさ $|\mathcal{X}_{aux}|$ と各モデル f についての Shapley 値からの属性推論攻撃の MAE



(a) Adult **(b)** Bank Marketing **(c)** Credit Card

図 5: 補助データセットの大きさ $|\mathcal{X}_{aux}|$ と各モデル f についての Shapley 値からの属性推論攻撃の SR



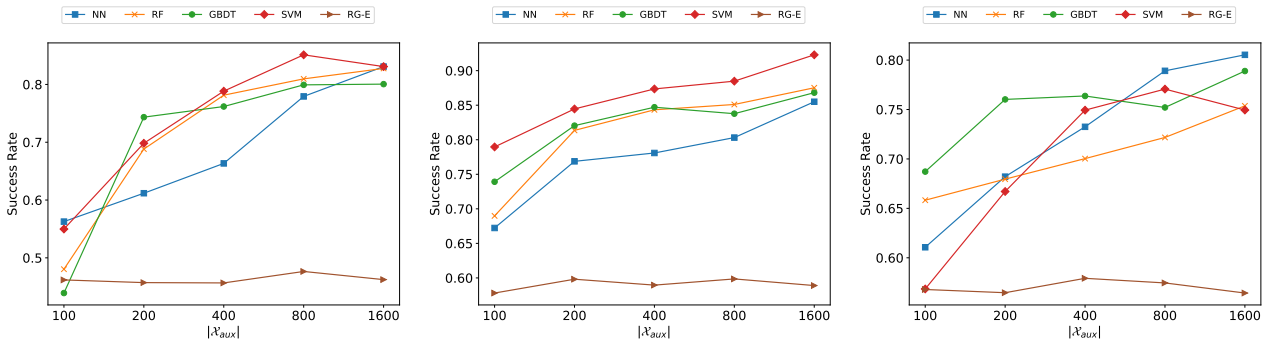
(a) Adult **(b)** Bank Marketing **(c)** Credit Card

図 6: 補助データセットの大きさ $|\mathcal{X}_{aux}|$ と各モデル f についての LIME からの属性推論攻撃の MAE

属性推定リスクが上がった．ただし，SR の値の大きさはデータセットによって異なり，属性推定リスクはデータに依存することが示唆された。

図 6，7 において，参照データの大きさ $|\mathcal{X}_{aux}|$ に対する推定リスクの傾向は，Shapley 値と同様であった．ほとんどの場合についてランダムな予測より小さい誤差で属性推定されたが，Adult データセット，$|\mathcal{X}_{aux}| = 100$, f が GBDT の場合のみ Shapley 値と異なり推定誤差がランダムな予測を上回った．

Shapley 値と LIME それぞれについての平均 SR を表 7 に示す．補助データセットの大きさ $|\mathcal{X}_{aux}|$ が小さいときは Shapley 値の属性推定リスクの方がより高いが，$|\mathcal{X}_{aux}|$ が大きいときは LIME の属性推定リスクの方がより高い．これは，LIME の出力 $\boldsymbol{w}$ からの元の入力 $\boldsymbol{x}$ の推定には学習データ $|\mathcal{X}_{aux}|$ が必要であることに依る．したがって，学習データが少ない時の属性推定は Shapley 値より難しいが，学習データが多い時は LIME より Shapley 値の方が難しい．



(a) Adult **(b)** Bank Marketing **(c)** Credit Card

図 7: 補助データセットの大きさ $|\mathcal{X}_{aux}|$ と各モデル f についての LIME からの属性推論攻撃の SR

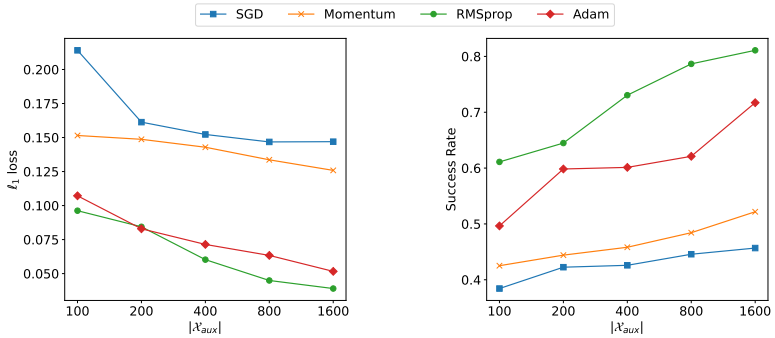


図 8: 攻撃者が採用する最適化アルゴリズムに対する，補助データセットの大きさを変化させたときの MAE と SR の変化

表 7: Shapley 値と LIME の属性推定リスク比較

	平均 SR	
	$\lvert\mathcal{X}_{aux}\rvert = 100$	$\lvert\mathcal{X}_{aux}\rvert = 1600$
Shapley 値	**0.65**	0.77
LIME	0.62	**0.83**

4.3.2 結果 2

図 8 の MAE と SR に共通して，SGD が最も属性推定の精度が低く，RMSprop が最も高い．全ての最適化アルゴリズムで，$|\mathcal{X}_{aux}|$ が増加するにつれて MAE は減少した．また，Adam と RMSprop は MAE が小さく，SGD と Momentum は MAE が大きい傾向が見られた．同様に，$|\mathcal{X}_{rand}|$ が増加するにつれて SR も増加した．Adam と RMSprop は SGD，Momentum と異なり，モデルの訓練中に学習率の調整を行う手法である．したがって，学習率の調整を行うことで属性推定の精度が高くなる．

5 おわりに

Luo ら [7] の手法に基づき，Shapley 値と LIME の属性推定リスクを調べた．オープンデータを用いた実験結果より，全ての説明モデル f に対して，どちらもランダムな予測よりも高い精度で属性推定された．また，Shapley 値と LIME の双方で，補助データセットの大きさが増えるにつれて攻撃精度が増加する傾向が見られた．さらに，f と ψ が線形モデルのとき，Shapley 値から正確にプライベートな入力特徴量の推定が可能であることを証明した．

属性推定のリスクを抑えるために，公開する Shapley 値や LIME の値にノイズを加えることを提案する．また，2022 年に Bozorgpanah ら [18] はデータそのものを匿名加工や差分プライバシーによって保護しても，ある程度であれば Shapley 値の有用性を損なわないことを報告している．そのため，データと説明ベクトルに対する加工によって属性推定リスクを下げられることが期待される．

今後の課題として，説明ベクトルにノイズを加えたときの属性推定リスクの調査や Shapley 値と LIME 以外の説明可能性技術に対する属性推定リスクの調査が挙げられる．

参考文献

[1] Cynthia Rudin, “Stop explaining black box machine learning models forhigh stakes decisions and use interpretable models instead”, *Nature Machine Intelligence* 1, 5, 206-215, 2019.

[2] Jianbo Chen, et al. “Learning to Explain: An Information-Theoretic Perspective on Model Interpretation”, In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July*, Vol. 80. PMLR, 882-891, 2018.

[3] Zest AI, “Why picking the right AI-credit decisioning partner matters”, Zest AI Insights. (Accessed November 3, 2023. https://www.zest.ai/insights/why-zaml-makes-your-ml-platform-better)

[4] Amazon Web Services, Inc, “Amazon SageMaker Clarify Model Explainability”, Amazon SageMaker Documentation. (Accessed November 3, 2023. https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-model-explainability.html)

[5] Microsoft, “Model interpretability”, Azure Machine Learning Documentation. (Accessed November 3, 2023. https://learn.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability)

[6] Amazon Web Services, Inc. “Amazon SageMaker Studio”, Amazon SageMaker Documentation. (Accessed November 3, 2023. https://docs.aws.amazon.com/sagemaker/latest/dg/studio-updated.html)

[7] Xinjian Luo, Yangfan Jiang, and Xiaokui Xiao, “Feature Inference Attack on Shapley Values”, In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS ’22), November, Los Angeles, CA, USA*. ACM, New York, NY, USA, pp.1-15, 2022.

[8] Adam Paszke, et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”, In Advances in Neural Information Processing Systems 32 (pp. 8024–8035). Curran Associates, Inc. 2019.

[9] Becker Barry, and Kohavi Ronny, “Adult”, UCI Machine Learning Repository. (https://doi.org/10.24432/C5XW20)

[10] Sérgio Moro, Paulo Cortez, and Paulo Rita, “A data-driven approach to predict the success of bank telemarketing”, Decision Support Systems 62, 22–31, 2014.

[11] I-Cheng Yeh and Che-hui Lien, “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients”, Expert systems with applications 36, 2, 2473–2480, 2009.

[12] Lloyd S Shapley, “A value for n-person games”, Vol. 2. Princeton UniversityPress, 303-317, 1953.

[13] Marco Ribeiro, Sameer Singh, and Carlos Guestrin, ““Why Should I Trust You?”: Explaining the Predictions of Any Classifier”, In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pages 97–101, San Diego, California. Association for Computational Linguistics, 2016.

[14] Bottou, Léon, “Online Algorithms and Stochastic Approximations”., Cambridge University Press, ISBN 978-0-521-65263-6, 1998.

[15] Ilya Sutskever, James Martens, George Dahl, Geoffrey Hinton, “On the importance of initialization and momentum in deep learning”, In Proceedings of the 30th international conference on machine learning (ICML-13). Vol. 28. Atlanta, GA. pp. 1139-1147, 2013.

[16] Geoffrey Hinton, “Coursera Neural Networks for Machine Learning Lecture 6”, (https://www.cs.toronto.edu/ tijmen/csc321/slides/lecture_slides_lec6.pdf)

[17] Diederik P. Kingma, Jimmy Ba, “Adam: A Method for Stochastic Optimization”, ICLR 2015, 2015

[18] Bozorgpanah, A., Torra, V., and Aliahmadipour, L, “Privacy and Explainability: The Effects of Data Protection on Shapley Values”, Technologies 10, 6, 125, 2022.