説明可能性に対するレコード再構築リスクの実験的評価

當麻 僚太郎[1,a)]　菊池 浩明[2,b)]

概要：機械学習モデルは近年，様々な領域で活用されている．しかし，多くのモデルは複雑な内部構造を持ち，モデルの公平性や透明性を担保することが難しいブラックボックスなモデルである．そこで，機械学習モデルの振舞いを説明する技術として，説明可能 AI（XAI）が注目されている．多くの Machine-Learning-as-a-Service（MLaaS）プラットフォームでは，入力特徴量と出力との間の関係性を説明する，SHAP や LIME などの XAI 技術が提供されている．一方で，2022 年に，Luo らによって，Shapley 値に基づくモデル説明からは本来秘匿されている入力データの属性を再構築できるリスクが示されている．しかしながら，XAI の説明の質に対して，再構築リスクがどう変化するかは未だに明らかでない．そこで，本研究では，SHAP と LIME の説明可能性を変化させたとき，レコード再構築リスクがどのように変化するかを評価した．オープンデータセットを用いて実験を行い，説明可能性が高いとき，レコード再構築リスクが高くなることを明らかにした．

キーワード：説明可能性，レコード再構築攻撃，SHAP，LIME

Empirical Evaluation of Record Reconstruction Risk for Explainability

Ryotaro Toma[1,a)]　Hiroaki Kikuchi[2,b)]

Abstract: In recent years, machine learning models have been widely applied across various domains. However, many of these models possess complex internal structures and are often regarded as black boxes, making it difficult to ensure fairness and transparency. To address this issue, explainable AI (XAI) techniques have attracted attention to interpret model behavior. Many Machine-Learning-as-a-Service (MLaaS) platforms now offer XAI tools such as SHAP and LIME, which aim to explain the relationship between input features and model outputs. On the other hand, Luo et al. (2022) demonstrated that model explanations based on Shapley values can potentially reveal sensitive attributes of the original input data, posing a privacy risk. However, how the quality of explanations affects the risk of record reconstruction remains unclear. In this study, we evaluate how record reconstruction risk changes with the level of explainability provided by SHAP and LIME. Through experiments on public datasets, we show that higher explainability tends to correspond to increased reconstruction risk.

Keywords: XAI, explainability, record reconstruction attack, SHAP, LIME

1. はじめに

機械学習モデルは近年，金融 [1] や医療 [2]，E コマース [3] といった重要なユースケースで利用されている．特に，深層ニューラルネットワークやアンサンブルモデルのような種類のモデルは，複雑な内部構造が「ブラックボックス」であり，その動作を内部的に解析することや，モデルがどのようにして意思決定に至ったかを把握することが困難である．そのため，モデルの透明性を保証し，入力特徴量とモデル出力との関係を説明するために説明可能 AI（XAI）技術が注目されている [4], [5].

一方で，XAI には説明に用いられたプライベートな入

[1] 明治大学大学院先端数理科学研究科
Graduate School of Advanced Mathematical Sciences, Meiji University
[2] 明治大学総合数理学部
School of Interdisciplinary Mathematical Sciences, Meiji University
[a)] cs242022@meiji.ac.jp
[b)] kikn@meiji.ac.jp

力を漏洩するリスクがあることが知られている．例えば，2022 年に Luo ら [6] は Shapley 値によるモデル説明から本来秘匿されているモデルへの入力レコードを推論出来ることを示した．しかしながら，XAI の説明可能性を低下させたとき，レコード再構築リスクがどう変化するかは未だに明らかでない．

そこで，本研究では，Shapley 値 [7] と LIME[8] によるモデル説明の一部をマスクしたとき，Luo ら [6] の手法を基にしたレコード再構築リスクがどう変化するかを明らかにする．また，説明のマスクによる説明可能性の低下を Yeh ら [9] の infidelity を用いて評価し，説明可能性とレコード再構築リスクのトレードオフを明らかにする．この提案方式を，三つのオープンデータセット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 値は 1953 年に Shapley によって提案 [7] された，協力ゲーム理論において各プレイヤーの貢献度を定量化する指標である．n 個の入力ベクトル $\boldsymbol{x} = (x_1, \ldots, x_n)$ に対するモデル出力 $f(\boldsymbol{x})$ の Shapley 値を $\boldsymbol{s} = (s_1, \ldots, s_n)$ と表す．本研究では $\boldsymbol{s}$ を $f(\boldsymbol{x})$ のモデル説明とする．

$N = \{1, 2, \ldots, n\}$ を特徴量のインデックス集合，S を N の部分集合，$\boldsymbol{x}^{(0)}$ を Shapley 値の計算に用いる参照サンプル，$\phi(\boldsymbol{x}; \boldsymbol{x}^{(0)}, f) = (s_1, \ldots, s_n)$ を Shapley 値を計算する写像とする．参照サンプルは Shapley 値の計算において欠損値の代わりとなる値であり，平均値やランダムサンプリングされた値が用いられる．すなわち，$\boldsymbol{x}_{[S]} = ((x_{[S]})_1, \ldots, (x_{[S]})_n)$ を S に対応する入力ベクトルとし，$i = 1, \ldots, n$ について

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

と定義する．

このとき，Shapley 値 s_i は

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

で定義する．

2.2 LIME

LIME（Local Interpretable Model-agnostic Explanations）は 2016 年に Ribeiro らによって提案 [8] された，各入力特徴量の貢献度を説明する手法である．この手法では，線形回帰や決定木，ルールベースなどの解釈が容易なモデルを用いて，ブラックボックスモデル f の振舞いを入力ベクトル $\boldsymbol{x} = (x_1, \ldots, x_n)$ の周りで近似する．本研究では，説明モデル g を線形モデル $g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b$ 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ことを想定する．G をモデル g の値域とし，$\Omega(g)$ をモデル g の係数ベクトル $\boldsymbol{w}$ に含まれる非ゼロ要素の個数とする．

$\mathcal{L}$ を説明モデル g を訓練するための損失関数とし，

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

で定義する．ここで，$\pi_{\boldsymbol{x}}(\boldsymbol{z})$ は関数 $\pi_{\boldsymbol{x}}(\boldsymbol{z}) = e^{-D(\boldsymbol{x}, \boldsymbol{z})^2 / \sigma^2}$ であり，距離 $D(\boldsymbol{x}, \boldsymbol{z})$ に基づいて損失の重みを決定する．また，$\boldsymbol{z}$ は入力 $\boldsymbol{x}$ と同様の n 次元ベクトルであり，単純化ベクトル $\boldsymbol{z}'$ は $\boldsymbol{z}$ の要素のうち $\boldsymbol{x}$ と一致する箇所を 1，そうでない箇所を 0 とするバイナリベクトルである．集合 Z は $\boldsymbol{z}$ と $\boldsymbol{z}'$ の定義域である．

このとき，説明モデル g^* は目的関数

$$g^* = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_{\boldsymbol{x}}) + \Omega(g)$$

によってフィッティングされる．

2.3 XAI のプライバシーリスク

XAI にはプライバシーリスクが存在する．例えば，メンバーシップ推論攻撃 [10], [11], [12]，モデル抽出攻撃 [10], [13], [14], [15]，属性推論攻撃 [6], [16]，敵対的攻撃 [10], [17] などがある．2024 年に Liu らは，SHAP や LIME，Grad-CAM などの 7 種類の XAI に対してメンバーシップ推論攻撃を提案 [12] し，XAI が訓練データの情報を漏洩することを示している．また，2023 年に Yan らは XAI-aware Model Extraction Attack（XaMEA）を提案 [14] し，XAI がモデルの情報を漏洩することを示している．

Luo ら [6] は 2022 年に，Shapley 値を用いた属性推論攻撃を提案し，Shapley 値によるモデル説明のプライバシーリスクを調査している．攻撃者は MLaaS 上のサービスにデータを送信し，データのモデル説明として Shapley 値を受け取ることができる．さらに，攻撃者はサービスから他のユーザの Shapley 値を盗むことができる．この仮定の下で，攻撃者はターゲットユーザが入力したレコードの再構築を行う．

2.4 関連研究

2.4.1 説明可能 AI（XAI）技術

ブラックボックスモデルを説明する XAI 技術は大域的な手法と局所的な手法に大別される．

大域的な手法では，モデルの全体的な振舞いを説明し，特徴量の重要度を算出する．例えば，2016 年に Datta らが提案した Qualitative Input Influence（QII）[18] は，すべての入力レコードにおいて入力特徴量がモデル出力に与える影響を定量的に評価する手法である．また，2020 年に Covert らは，Shapley 値 [7] に基づいて特徴量の重要度を算出する Shapley Additive Global Importance（SAGE）

を提案 [19] している．これらの手法は特定の入力レコードに依存しない大域的な入力特徴の重要度を示すが，個々の入力レコードに対するモデルの振舞いを説明することができない．

一方, 局所的な手法では, 各入力レコードに対して特徴量の重要度を与える．2016 年に Ribeiro らが提案した LIME [8] や, 2017 年に Shrikumar らが提案した DeepLIFT [20], 同じく 2017 年に Selvaraju らが提案した Grad-CAM [21], 2018 年に Ribeiro らが提案した Anchors [22] など, 多くの手法が提案されている．また，Shapley 値に基づいた手法として，2014 年に Štrumbelj らが提案したサンプリング手法 [23] や，2017 年に Lundberg らが提案した SHAP [24] がある．Lundberg らは，SHAP，LIME，DeepLIFT などの XAI 手法をまとめた Additive Feature Attribution Methods クラスを提案している．

Amazon SageMaker [25] や Microsoft Azure [26], Google Cloud Platform [27] などの主要な MLaaS プラットフォームでは多くの XAI 手法が提供されている．

2.4.2 XAI の説明可能性

様々な XAI 手法に対して包括的に説明可能性を評価するための指標がいくつか提案されている．

Samek らが 2016 年に提案した MoRF（Most Relevant First）と LeRF（Least Relevant First） [28] は，画像分類モデルに対して得られた各画素や領域の重要度ヒートマップを評価する手法である．ヒートマップを画像内の位置の順序付き集合 $\mathcal{O} = (r_1, r_2, ..., r_L)$ として定義する．ここで各位置 r_p は画像上のピクセル位置を表す二次元ベクトルである．この順序付けに従って領域を順次摂動する処理を MoRF と呼び，再帰的に

$$\boldsymbol{x}^{(0)}_{\mathrm{MoRF}} = \boldsymbol{x}$$
$$\forall 1 \le k \le L : \boldsymbol{x}^{(k)}_{\mathrm{MoRF}} = g(\boldsymbol{x}^{(k-1)}_{\mathrm{MoRF}}, r_k)$$

で定義する．ここで関数 g は，画像 $\boldsymbol{x}^{(k-1)}_{\mathrm{MoRF}}$ 内の指定された位置 r_k の情報を取り除く摂動関数である．LeRF は順序 $\mathcal{O}$ を逆順に利用して同様に

$$\boldsymbol{x}^{(0)}_{\mathrm{LeRF}} = \boldsymbol{x}$$
$$\forall 1 \le k \le L : \boldsymbol{x}^{(k)}_{\mathrm{LeRF}} = g(\boldsymbol{x}^{(k-1)}_{\mathrm{LeRF}}, r_{L+1-k})$$

で定義する．

また，Yeh ら [9] は 2019 年に，入力に対するモデル説明の非忠実度を測定する infidelity を提案している．d 次元入力ベクトル $\boldsymbol{x}$，ブラックボックスモデル f，説明関数 Φ，確率測度 μ_I を持つ確率変数 $I \in \mathbb{R}^d$ が与えられたとき，Φ の infidelity を

$$INFD(\Phi, f, \boldsymbol{x}) = \mathbb{E}_{I\sim\mu_I}[(I^T\Phi(f,\boldsymbol{x}) - (f(\boldsymbol{x}) - f(\boldsymbol{x}-I))^2)]$$

と定義する．I は $\boldsymbol{x}$ 周辺の摂動を表し，いくつかの手法を用いることができる．

- ベースラインとの差：$I = \boldsymbol{x} - \boldsymbol{x}_0$
- ベースラインとの差の部分集合：任意の部分集合 $S_k \subseteq [d]$ に対して，$I_{S_k} = \boldsymbol{x} - \boldsymbol{x}[\boldsymbol{x}_{S_k} = (\boldsymbol{x}_0)_{S_k}]$
- ノイズ付きベースラインとの差：$I = \boldsymbol{x} - \boldsymbol{z}_0$，ここで $\boldsymbol{z} = \boldsymbol{x}_0 + \epsilon$, ϵ は平均 0 のノイズで例えば $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- 複数のベースラインとの差：$I = \boldsymbol{x} - \boldsymbol{x}_0$，ここで $\boldsymbol{x}_0$ は複数の値を取る確率変数

本研究では，ベースライン $\boldsymbol{x}_0$ をゼロベクトルとし，ランダムに選んだ 100 個の部分集合 $S_1, \cdots, S_{100} \subseteq [n]$ に対して infidelity を計算した平均値をその説明の infidelity として用いる．infidelity が低いほど，その説明は質の高い忠実な説明として評価される．

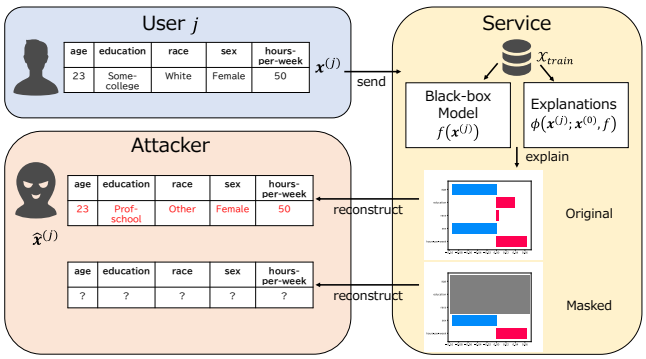


図 1: システムモデルの概要図

3. 問題設定

3.1 マスクされたモデル説明からのレコード再構築攻撃

モデル説明の一部をマスクしたときの攻撃リスクや説明可能性の影響は未だに明らかでない．そこで本研究では，説明ベクトルのマスクがどのようにレコード再構築攻撃に影響を与えるかを調査する．レコード再構築攻撃 [6] を次のように定義する．

f と ψ をそれぞれブラックボックスモデル，攻撃モデルとする．$\mathcal{X}_{train}$ をブラックボックスモデル f を訓練するための訓練データセット，$\mathcal{X}_{aux}$ を攻撃モデル ψ を訓練するための補助データセット，$\mathcal{X}_{test}$ をテストデータセットとする．$\boldsymbol{x}^{(j)} = (x^{(j)}_1, ..., x^{(j)}_n)$ をユーザ $j = 1, \ldots, m$ の入力ベクトルとし，$\boldsymbol{x}^{(0)}$ を参照サンプル，$\boldsymbol{s}^{(j)} = \phi(\boldsymbol{x}^{(j)}; \boldsymbol{x}^{(0)}, f)$ を入力ベクトル $\boldsymbol{x}^{(j)}$ の説明ベクトルとする．全ての $\boldsymbol{x}_{aux} \in \mathcal{X}_{aux}$ について説明ベクトルを生成したデータセット $\mathcal{S}_{aux} = \{\phi(\boldsymbol{x}_{aux}; \boldsymbol{x}^{(0)}, f) \,|\, \boldsymbol{x}_{aux} \in \mathcal{X}_{aux}\}$ とブラックボックスモデル f が与えられたとき，攻撃者は攻撃モデル $\psi : \mathcal{S}_{aux} \to \mathcal{X}_{aux}$ を訓練する．ターゲットユーザ j の説明ベクトル $\boldsymbol{s}^{(j)}$ が与えられたとき，対応する入力ベクトル $\boldsymbol{x}^{(j)}$ を推測することをレコード再構築攻撃と呼ぶ．本研究の概要図を **図 1** に示す．

これらの設定の下で，説明ベクトルのマスクがレコード

再構築リスクに与える影響を実験的に評価する．また，説明ベクトルのマスクが引き起こす説明可能性の低下を Yeh らの infidelity [9] を用いて評価し，説明可能性とレコード再構築リスクのトレードオフを明らかにする．

3.2 攻撃リスク評価

レコード再構築リスクを評価するために，攻撃者の平均絶対誤差（MAE）と攻撃成功率（Success Rate; SR）の二つの指標を用いる．

3.2.1 攻撃者の MAE

MAE は絶対誤差の平均である．m 行 n 列のデータセット $\mathcal{X}$ と再構築したデータセット $\hat{\mathcal{X}}$ に対する攻撃者の MAE を

$$MAE(\hat{\mathcal{X}}, \mathcal{X}) = \frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} |\hat{\mathcal{X}}_i^{(j)} - \mathcal{X}_i^{(j)}|$$

で定義する．

3.2.2 攻撃成功率 SR

SR は全ての入力特徴のうち正確に再構築できた特徴の比率を表す．離散値に対しては元の値と一致していれば再構築できたとみなし，連続値に対しては絶対誤差が一定の閾値を下回っているとき再構築できたとみなす．データセット $\mathcal{X}$ に対して再構築したデータセット $\hat{\mathcal{X}}$ の SR は

$$SR(\hat{\mathcal{X}}, \mathcal{X}) = \frac{success(\hat{\mathcal{X}}, \mathcal{X})}{mn} \tag{1}$$

で定義する．ここで，$success(\hat{\mathcal{X}}, \mathcal{X})$ は正確に再構築された入力特徴の数である．

3.3 説明可能性評価

モデル説明の質（説明可能性）を評価するために, モデル説明のマスク率（Masked Rate; MR）と Yeh らの infidelity [9] を用いる．

3.3.1 モデル説明のマスク率 MR

Samek らの MoRF と LeRF [28] に基づいて，n 次元説明ベクトル $\boldsymbol{s}$ をマスクする．テストデータセット $\mathcal{X}_{test}$ の各レコード $\boldsymbol{x}_{test}^{(i)} = (x_1^{(i)}, x_2^{(i)}, ..., x_n^{(i)})$ $(i = 1, ..., |\mathcal{X}_{test}|)$ に対して Shapley 値や LIME による説明ベクトル $\boldsymbol{s}_{test}^{(i)} = (s_1^{(i)}, s_2^{(i)}, ..., s_n^{(i)})$ を生成し，全レコードについてそれぞれ平均を取ったグローバルな重要度ベクトル $\bar{\boldsymbol{s}}_{test} = (\bar{s}_1, \bar{s}_2, ..., \bar{s}_n)$ に基づいて順序付けられたインデックスの集合 $\mathcal{O} = (r_1, r_2, ..., r_n)$ を定義する．この $\mathcal{O}$ に基づいて，

$$\boldsymbol{s}_{\mathrm{MoRF}}^{(0)} = \boldsymbol{s}$$
$$\forall 1 \leq k \leq n : \boldsymbol{s}_{\mathrm{MoRF}}^{(k)} = g(\boldsymbol{s}_{\mathrm{MoRF}}^{(k-1)}, r_k)$$

としてマスクする．ここで，関数 g を説明ベクトル $\boldsymbol{s}_{\mathrm{MoRF}}^{(k-1)}$ の要素 r_k を 0 に置き換えるマスク関数とする．LeRF についても同様に

$$\boldsymbol{s}_{\mathrm{LeRF}}^{(0)} = \boldsymbol{s}$$
$$\forall 1 \leq k \leq n : \boldsymbol{s}_{\mathrm{LeRF}}^{(k)} = g(\boldsymbol{s}_{\mathrm{LeRF}}^{(k-1)}, r_{n+1-k})$$

としてマスクする．このとき，マスク率 MR は MoRF と LeRF のいずれにおいても同様に

$$MR(\boldsymbol{s}_{\mathrm{MoRF|LeRF}}^{(k)}) = \frac{k}{n}$$

と定義する．

表 1: 実験に用いた三つのデータセット

Dataset	No. of Records	Classes	No. of Features
Adult [29]	48,842	2	14
Bank Marketing [30]	45,211	2	16
Credit Card Client [31]	30,000	2	24

4. 実験

4.1 データセット

表 1 に実験に用いたデータセットを示す．データセット中に含まれるカテゴリ変数はすべて One-Hot エンコーディングを用いて二値の連続値に変換する．また，各データセットのうち 6 割を訓練データセット $\mathcal{X}_{train}$，2 割をテストデータセット $\mathcal{X}_{test}$ とし，残りの 2 割から $|\mathcal{X}_{aux}| = 100, 200, 400, 800, 1600$ を選んで実験を行う．最終的に，それぞれの $|\mathcal{X}_{aux}|$ に対する評価指標の値を平均する．

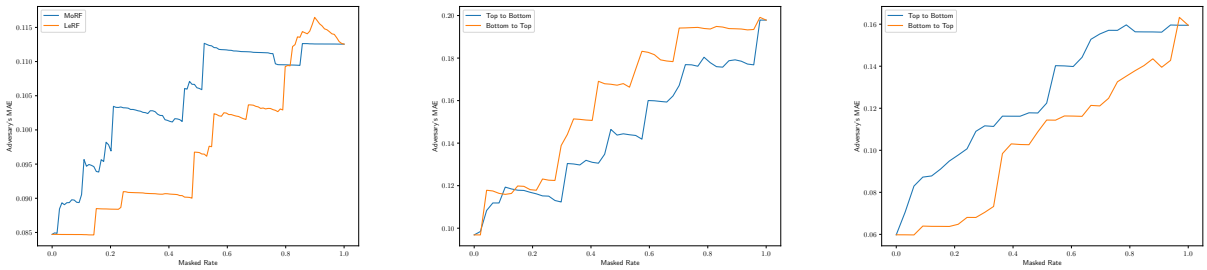
4.2 実験設定

説明対象のブラックボックスモデル f は PyTorch [32] を用いて，N 次元の入力層と 1 次元の出力層を持ち，$2N$ 個のニューロンを含む隠れ層を二つ持つ二値分類モデルとして実装した．活性化関数は出力層のみ sigmoid であり，他の層は全て ReLU を用いた．

Shapley 値の計算には, 近似手法として Kernel SHAP [24] を用いた．また，実装にあたって Lundberg らの提供する Python ライブラリ SHAP [24] を用いた．LIME の実装には Ribeiro らの提供する Python ライブラリ LIME [8] を用いた．

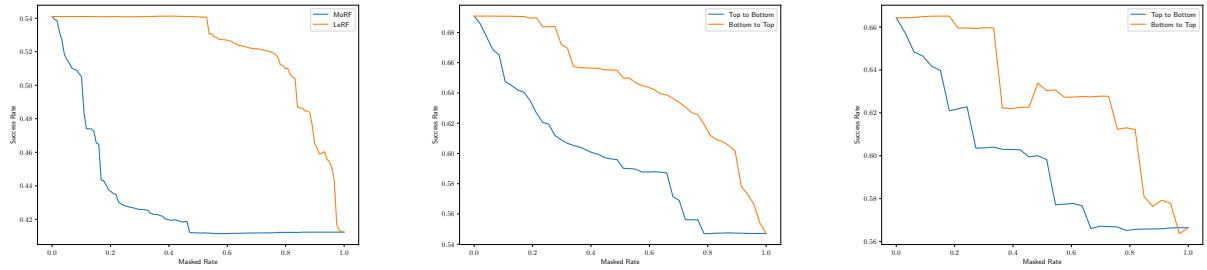
4.3 結果 1：マスク率とレコード再構築リスク

Shapley 値に対する MR とレコード再構築リスクの実験結果を **図 2**，**図 3** に，LIME に対する実験結果を **図 4**，**図 5** に示す．XAI の違いやデータセットの違いによらず，MR が高くなるほど攻撃者の MAE は上昇し，SR は低下した．しかしながら，グラフの振舞いは SR よりも攻撃者の MAE の方が不安定に変化し，特に LIME に対する攻撃者の MAE について，MR が高くなるにつれて MAE が低下している領域が見られた．



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

図 2: Shapley 値に基づく説明ベクトルの一部をマスクしたときの攻撃者の MAE



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

図 3: Shapley 値に基づく説明ベクトルの一部をマスクしたときの SR

表 2: 説明ベクトルの infidelity とレコード再構築リスクの相関係数

XAI	評価指標	Adult	Bank Marketing	Credit Card Client
Shapley 値	MAE	−0.30	0.14	0.14
	SR	0.19	−0.19	0.04
LIME	MAE	−0.24	0.27	0.34
	SR	0.22	−0.24	−0.10

4.3.1 結果 2：説明の infidelity とレコード再構築リスク

Shapley 値に対する infidelity とレコード再構築リスクの実験結果を **図 6**, **図 7** に, LIME に対する実験結果を **図 8**, **図 9** に示す．また，それぞれの XAI，データセット，評価指標の組み合わせについて，相関係数をまとめたものを **表 2** に示す．結果として，攻撃者の MAE と SR の双方において相関係数の絶対値は最大でも 0.3 程度と弱く，レコード再構築リスクと説明の infidelity との間に明確な関係は見られなかった．

4.3.2 結果 3：説明の infidelity とマスク率

説明のマスクによる infidelity の変化 **図 10**，**図 11** に示す．図はどちらも Adult データセットに対する結果である．これらの結果から，説明の MR が高いほど infidelity が高い，すなわちモデル説明の質が低下することがわかる．

4.3.3 考察

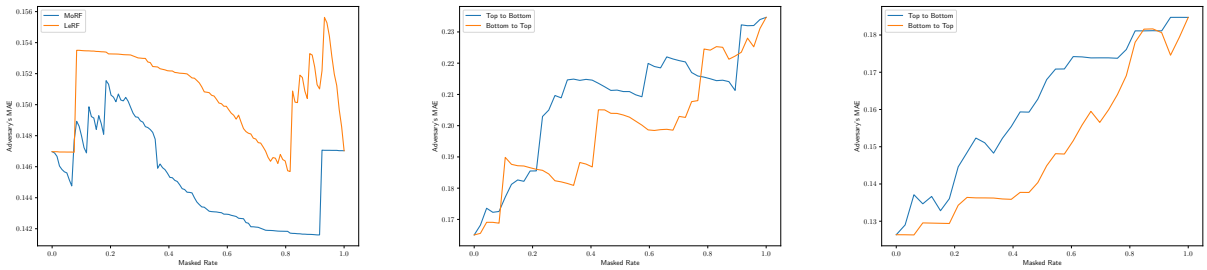
どの実験結果も基本的に MR が高くなるほど攻撃リスクが低下する傾向を示していたが，一部では MR が高いにもかかわらず攻撃リスクが高い部分が見られた．例えば，図 4（a）では，MR が 0.1 から 0.8 までの間で MR が高くなるにつれて攻撃者の MAE が低下している領域があった．一方で，図 5（a）では同じ領域での SR は変化が小さかった．これは，Adult データセットにおいて重要な特徴量はおよそ 20 % 程度であり，残りの 80 % は削除しても問題のない特徴量であるため，重要でない情報が削除されることによって予測をしやすくなり，攻撃者の MAE が下がったと考えられる．

また，MR が高くなるほど攻撃リスクが低下した一方で，MR が高くなるほど infidelity が高い，すなわちモデル説明の質が低下した．したがって，説明ベクトルのマスク量によってモデル説明の質とレコード再構築リスクのトレードオフを調節することができる．

5. おわりに

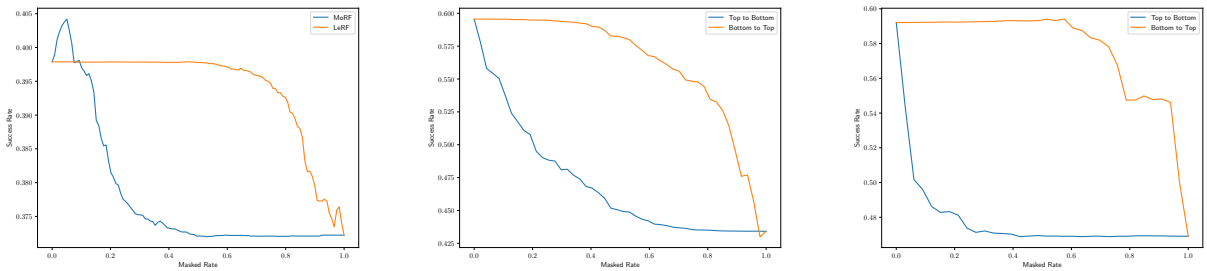
モデル説明をマスクしたときのレコード再構築リスクと infidelity を実験的に評価した．XAI 手法の違いやデータセットに違いによらず，マスク率が高いほどレコード再構築リスクが低下した．また，説明ベクトルのマスク量によって説明可能性と安全性のトレードオフを制御できることを示唆した．

今後の課題として，プライバシー保護手法による XAI の説明可能性の低下とレコード再構築リスクの関係を調査することが挙げられる．



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

図 4: LIME に基づく説明ベクトルの一部をマスクしたときの攻撃者の MAE



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

図 5: LIME に基づく説明ベクトルの一部をマスクしたときの SR

参考文献

[1] Chen, X.-Q., Ma, C.-Q., Ren, Y.-S., Lei, Y.-T., Huynh, N. Q. A. and Narayan, S.: Explainable artificial intelligence in finance: A bibliometric review, *Finance Research Letters*, Vol. 56, pp. 104–145 (2023).

[2] Sakai, A., Komatsu, M., Komatsu, R., Matsuoka, R., Yasutomi, S., Dozen, A., Shozu, K., Arakaki, T., Machino, H., Asada, K., Kaneko, S., Sekizawa, A. and Hamamoto, R.: Medical Professional Enhancement Using Explainable Artificial Intelligence in Fetal Cardiac Ultrasound Screening, *BIOMEDICINES*, Vol. 10, No. 3 (2022).

[3] Khrais, L. T.: Role of Artificial Intelligence in Shaping Consumer Demand in E-Commerce, *Future Internet*, Vol. 12, No. 12 (2020).

[4] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *NATURE MACHINE INTELLIGENCE*, Vol. 1, No. 5, pp. 206–215 (online) (2019).

[5] Chen, J., Song, L., Wainwright, M. and Jordan, M.: Learning to Explain: An Information-Theoretic Perspective on Model Interpretation, *Proceedings of the 35th International Conference on Machine Learning* (Dy, J. and Krause, A., eds.), Proceedings of Machine Learning Research, Vol. 80, PMLR, pp. 883–892 (online), available from ⟨https://proceedings.mlr.press/v80/chen18j.html⟩ (2018).

[6] Luo, X., Jiang, Y. and Xiao, X.: Feature Inference Attack on Shapley Values, *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, New York, NY, USA, Association for Computing Machinery, pp. 2233–2247 (2022).

[7] Shapley, L. S.: *A Value for n-Person Games*, Contributions to the Theory of Games, Vol. 2, pp. 307–318, Princeton University Press (1953).

[8] Ribeiro, M. T., Singh, S. and Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA, Association for Computing Machinery, pp. 1135–1144 (2016).

[9] Yeh, C.-K., Hsieh, C.-Y., Suggala, A.S., Inouye, D.I., and Ravikumar, P.: On the (In)fidelity and Sensitivity of Explanations, *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, NIPS'19, Red Hook, NY, USA, Curran Associates Inc., pp. 10967–10978 (2019).

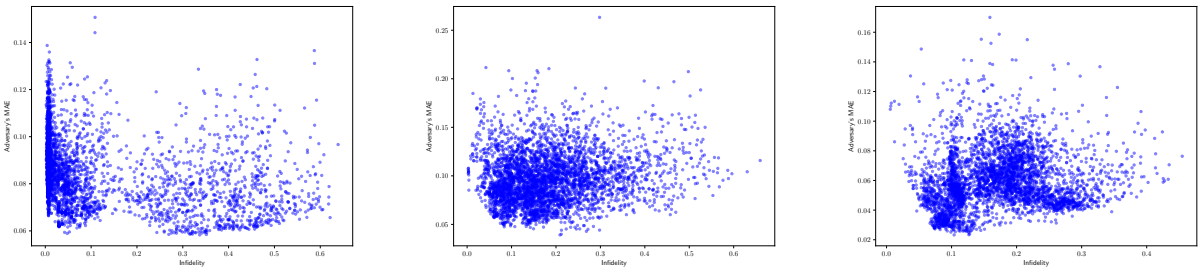
[10] Kuppa, A. and Le-Khac, N.-A.: Adversarial XAI Methods in Cybersecurity, *IEEE Transactions on Information Forensics and Security*, Vol. 16, pp. 4924–4938 (2021).

[11] Shokri, R., Strobel, M. and Zick, Y.: On the Privacy Risks of Model Explanations, *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, New York, NY, USA, Association for Computing Machinery, pp. 231–241 (2021).

[12] Liu, H., Wu, Y., Yu, Z. and Zhang, N.: Please Tell Me More: Privacy Impact of Explainability through the Lens of Membership Inference Attack , *2024 IEEE Symposium on Security and Privacy (SP)*, Los Alamitos, CA, USA, IEEE Computer Society, pp. 4791–4809 (2024).

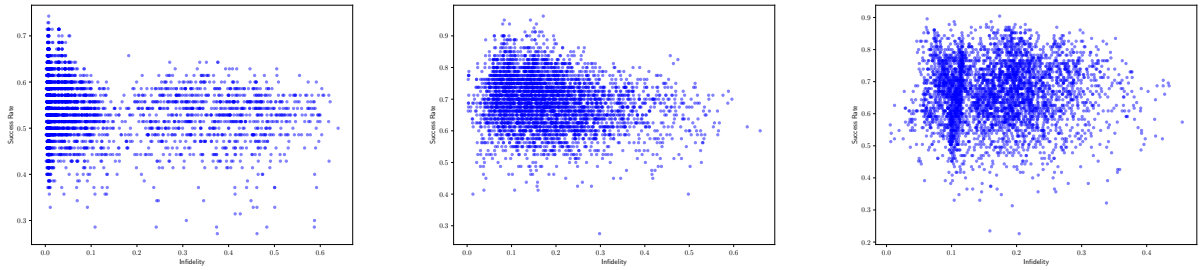
[13] Yan, A., Hou, R., Liu, X., Yan, H., Huang, T. and Wang, X.: Towards explainable model extraction attacks, *International Journal of Intelligent Systems*, Vol. 37, No. 11, pp. 9936–9956 (2022).

[14] Yan, A., Huang, T., Ke, L., Liu, X., Chen, Q. and Dong, C.: Explanation leaks: Explanation-guided model extraction attacks, *Inf. Sci.*, Vol. 632, No. C, pp. 269–284



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

図 6: Shapley 値に基づく説明ベクトルの infidelity と攻撃者の MAE



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

図 7: Shapley 値に基づく説明ベクトルの infidelity と SR

(2023).

[15] Olatunji, I. E., Rathee, M., Funke, T. and Khosla, M.: Private Graph Extraction via Feature Explanations, *Proceedings on Privacy Enhancing Technologies*, Vol. 2023, No. 2 (2023).

[16] Fredrikson, M., Jha, S. and Ristenpart, T.: Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, New York, NY, USA, Association for Computing Machinery, pp. 1322–1333 (2015).

[17] Baniecki, H. and Biecek, P.: Adversarial attacks and defenses in explainable artificial intelligence: A survey, *Information Fusion*, Vol. 107, p. 102303 (2024).

[18] Datta, A., Sen, S. and Zick, Y.: Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems, *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 598–617 (2016).

[19] Covert, I. C., Lundberg, S. and Lee, S.-I.: Understanding global feature contributions with additive importance measures, *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, Curran Associates Inc. (2020).

[20] Shrikumar, A., Greenside, P. and Kundaje, A.: Learning important features through propagating activation differences, *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, JMLR.org, pp. 3145–3153 (2017).

[21] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626 (2017).

[22] Ribeiro, M. T., Singh, S. and Guestrin, C.: Anchors: high-precision model-agnostic explanations, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18, AAAI Press (2018).

[23] Štrumbelj, E. and Kononenko, I.: Explaining prediction models and individual predictions with feature contributions, *Knowl. Inf. Syst.*, Vol. 41, No. 3, pp. 647–665 (2014).

[24] Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, Red Hook, NY, USA, Curran Associates Inc., pp. 4768–4777 (2017).

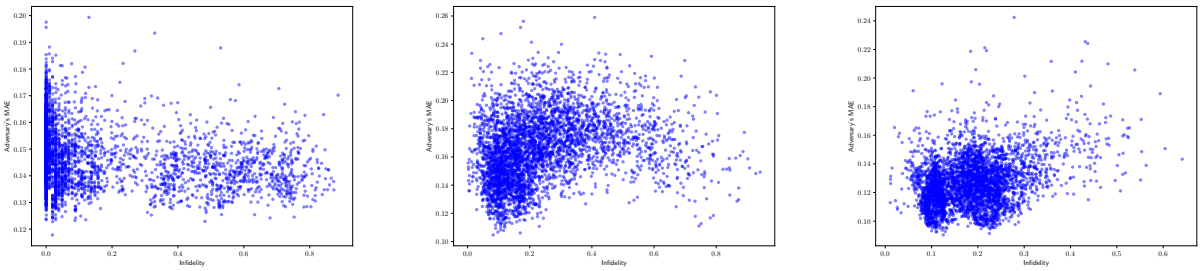
[25] Amazon SageMaker Documentation: Model Explainability, Amazon Web Services (online), available from ⟨https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-model-explainability.html⟩ (accessed 2024-04-19).

[26] Azure Machine Learning Documentation: Model Explainability, Microsoft Azure (online), available from ⟨https://learn.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability⟩ (accessed 2024-04-19).

[27] Google Cloud: Introduction to Vertex Explainable AI, Google Cloud (online), available from ⟨https://cloud.google.com/vertex-ai/docs/explainable-ai/overview⟩ (accessed 2024-04-19).

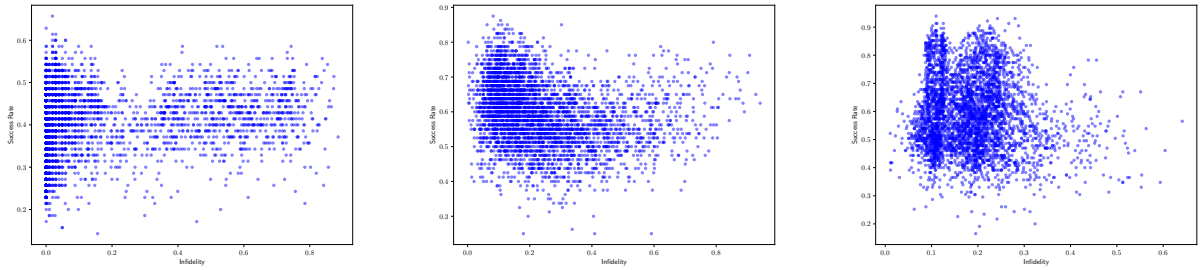
[28] Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K.-R.: Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems, 28(11):2660–2673 (2016).

[29] Becker, B. and Kohavi, R.: Adult, UCI Ma-



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

図 8: LIME に基づく説明ベクトルの infidelity と攻撃者の MAE



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

図 9: LIME に基づく説明ベクトルの infidelity と SR

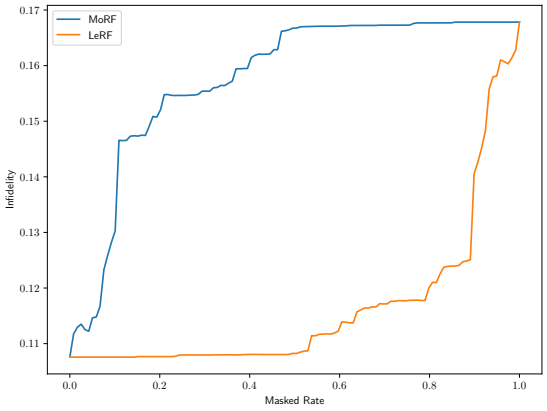


図 10: Shapley 値に基づく説明ベクトルのマスク率 MR と infidelity

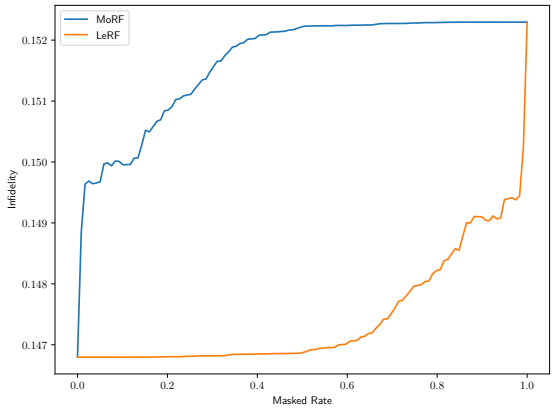


図 11: LIME に基づく説明ベクトルのマスク率 MR と infidelity

chine Learning Repository (online), available from ⟨https://doi.org/10.24432/C5XW20⟩ (accessed 2024-04-19).

[30] Moro, S., Rita, P. and Cortez, P.: Bank Marketing, UCI Machine Learning Repository (online), available from ⟨https://doi.org/10.24432/C5K306⟩ (accessed 2024-04-19).

[31] Yeh, I.-C.: Default of Credit Card Clients, UCI Machine Learning Repository (online), available from ⟨https://doi.org/10.24432/C55S3H⟩ (accessed 2024-04-19).

[32] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. and Chintala, S.: *PyTorch: an imperative style, high-performance deep learning library*, Curran Associates Inc. (2019).