# 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

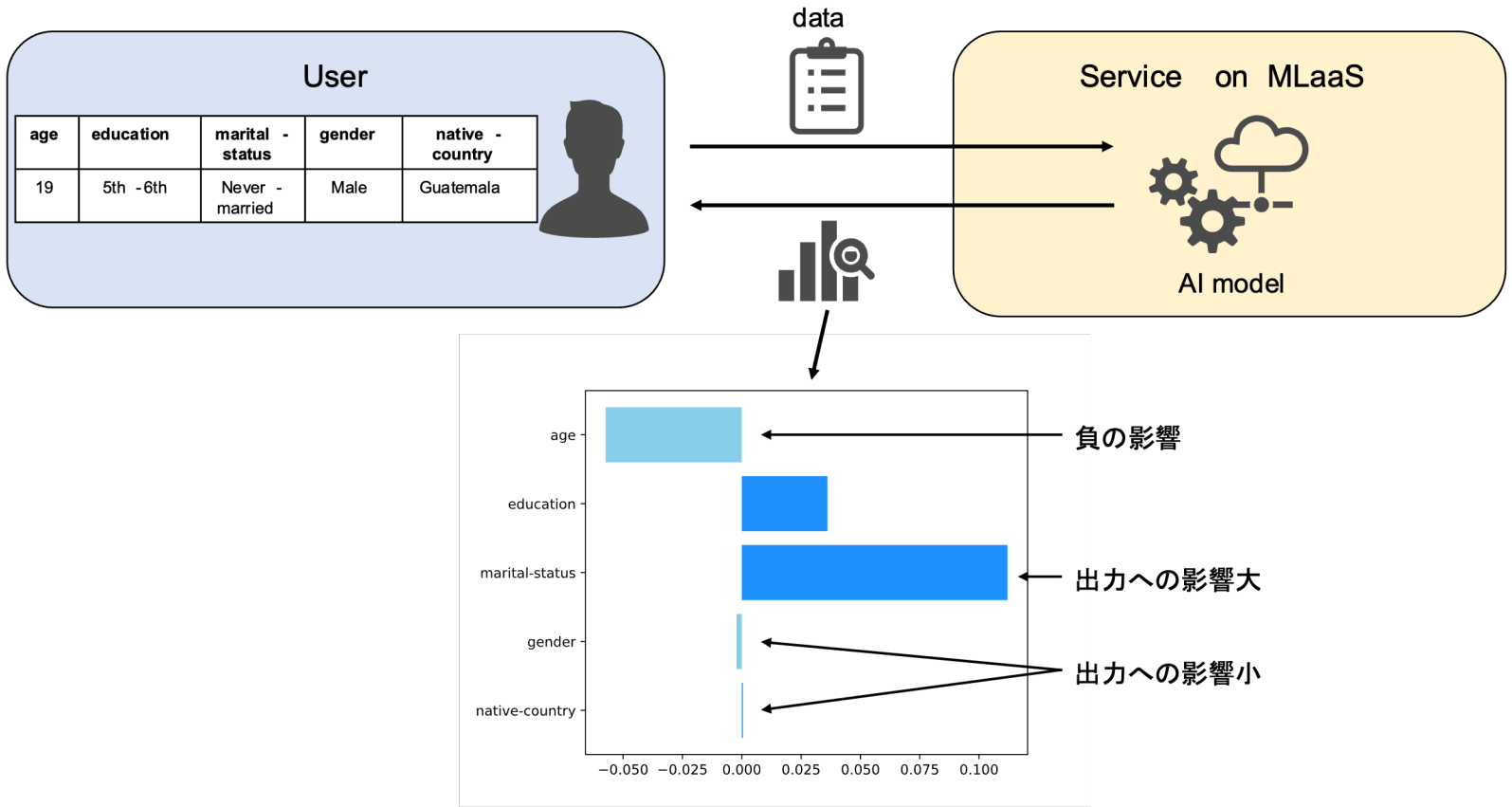
當麻 僚太郎，菊池 浩明
明治大学

# 説明可能性技術（XAI）の重要性

- 機械学習モデルが意思決定に利用されることが増えている
- モデルの多くはブラックボックス
- モデルの公平性や透明性を保証し，出力に説明を与える必要

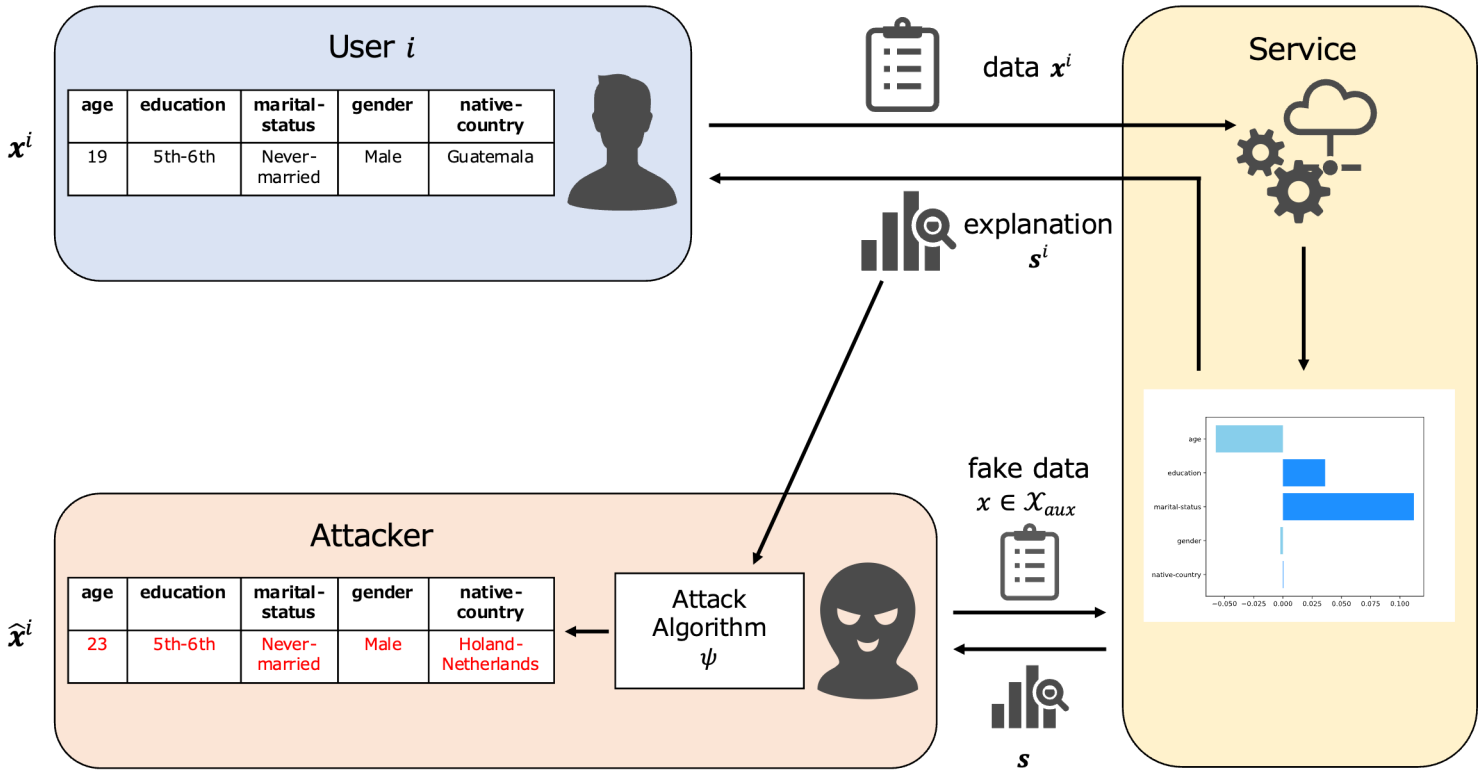
# XAIの例： Shapley値 [Shapley 1953]

- 説明可能性技術として多くのMLaaSサービスで提供される
  - AWSやGCPなど



# 特徴量推論攻撃 [Luo 2022]

- 機密データセット $\mathcal{X}_{train}$ に基づいて学習したブラックボックスモデル $f$ から得られるXAIの値 $\boldsymbol{s}^i$ を基に，入力されたユーザの特徴量 $\boldsymbol{x}^i$ を推論する



# 特徴量推論攻撃 [Luo 2022]

- 攻撃者 1
  - 訓練データセット $\mathcal{X}_{train}$ 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
  - 全ての $x_{aux} \in \mathcal{X}_{aux}$ について対応する説明データ $\mathcal{S}_{aux}$ から攻撃アルゴリズム $\psi: \mathcal{S}_{aux} \rightarrow \mathcal{X}_{aux}$ を訓練する

- 攻撃者2
  - 補助データセット $\mathcal{X}_{aux}$ を持っていない

# 特徴量推論攻撃の例

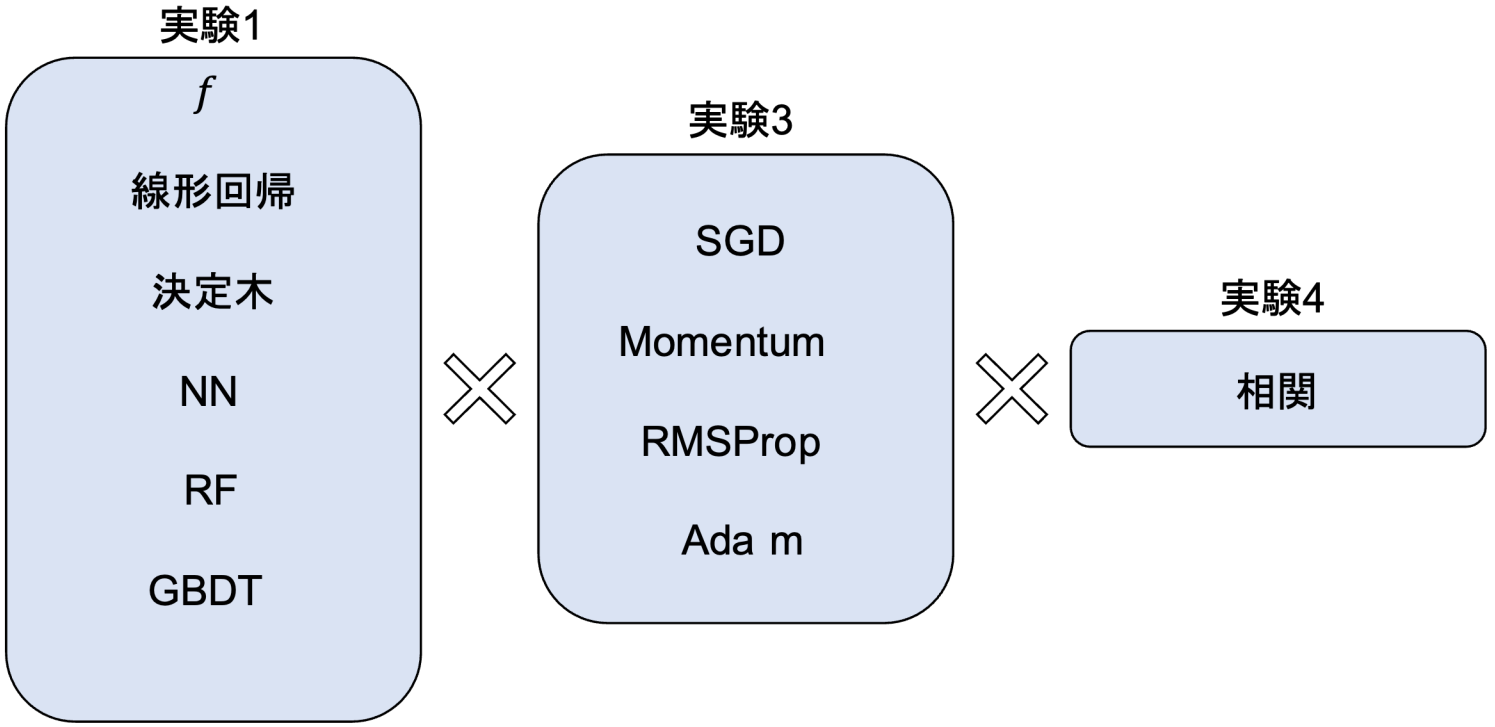
| | | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ | |
|---|---|---|---|---|---|---|---|
| $\mathcal{X}_{test}$ | $\boldsymbol{x}$ | 1.8 | 0.1 | 0.3 | −0.4 | 1.9 | → $\hat{y} = f(\boldsymbol{x})$ |
| $\mathcal{X}_{aux}$ | | −1 | 0.3 | −0.3 | 0.5 | | |
| Shapley値 | $\boldsymbol{s}$ | 1.30 | 0.02 | 0.06 | −0.04 | | |
| 推定 | $\hat{\boldsymbol{x}}$ | 2.0 | 0.2 | 0.1 | −0.1 | | ← $\hat{\boldsymbol{x}} = \psi(\boldsymbol{s}, \boldsymbol{x}^0, f)$ |
| MAE | | 0.2 | 0.1 | 0.2 | 0.3 | | |

# [Luo 2022] の問題点

- ブラックボックス学習モデル $f$ に依存する推定リスクが不明
- 攻撃アルゴリズム $\psi$ の学習方式に依存する推定リスクが不明
- 各説明変数と目的変数との間の相関に依存する推定リスクが不明

# 提案方式

- 学習モデル $f$ に対するShapley値 $s$ からの属性推定 $\psi$ のリスクを明らかにする
- どの組み合わせのリスクが高いか？

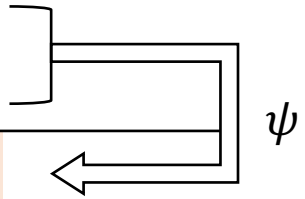


# 線形回帰モデルの特徴量推論脆弱性

- モデル $f$ が線形モデルであるときShapley値からプライベートな入力特徴を誤差なく推論出来る

**命題.** $f$ を線形モデルによる説明モデル，$\psi$ 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，$\psi$ による推定の $\mathrm{MAE} = 0$ である．

| | | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|---|---|
| $\mathcal{X}_{test}$ | $\boldsymbol{x}$ | 1.8 | 0.1 | 0.3 | −0.4 | 1.9 |
| $\mathcal{X}_{aux}$ | $\boldsymbol{x}^0$ | −1 | 0.3 | −0.3 | 0.5 | |
| Shapley値 | $\boldsymbol{s}$ | 1.30 | 0.02 | 0.06 | −0.04 | |
| 推定 | $\hat{\boldsymbol{x}}$ | 1.8 | 0.1 | 0.3 | −0.4 | |
| MAE | | 0.0 | 0.0 | 0.0 | 0.0 | |



# 実験方法

- モデル $f$ や $\psi$ の推定に使うアルゴリズムに対する属性推定リスクを調べる

- 評価指標
  - Mean Absolute Error （MAE）
    - $\ell_1(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left|\hat{x}_i^j - x_i^j\right|$
  - Success Rate (SR)
    - 推定に成功した入力特徴量の割合
    - $SR(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \frac{success(\hat{\boldsymbol{x}},\boldsymbol{x})}{mn}$

表 6　使用データセット

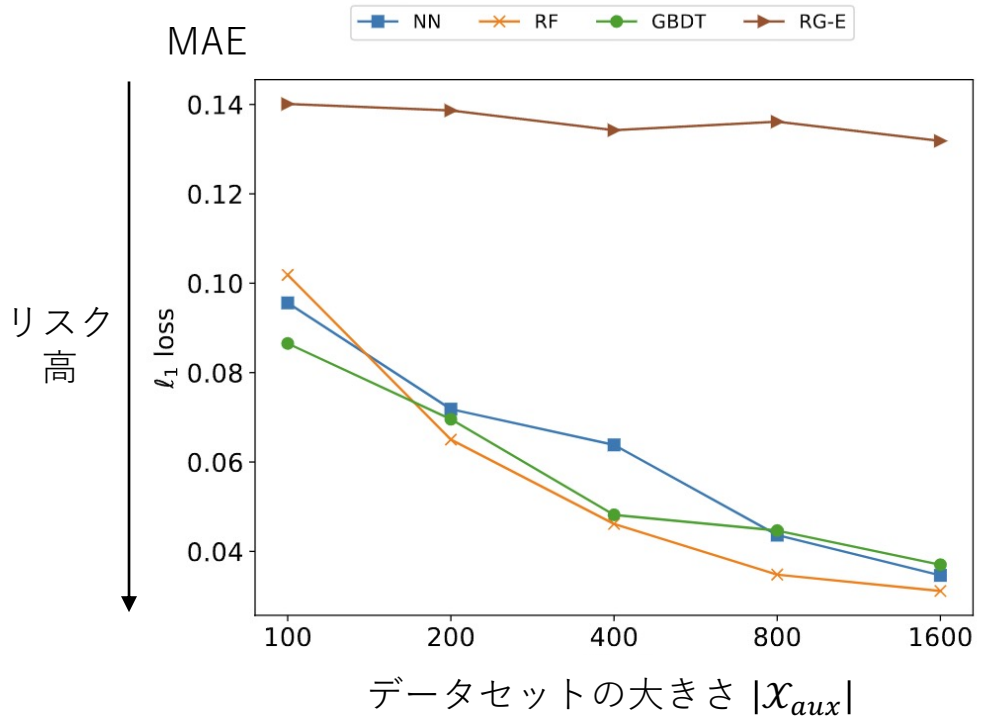
| データセット | レコード数 | クラス | 特徴量 |
|---|---|---|---|
| Adult[10] | 48842 | 2 | 14 |

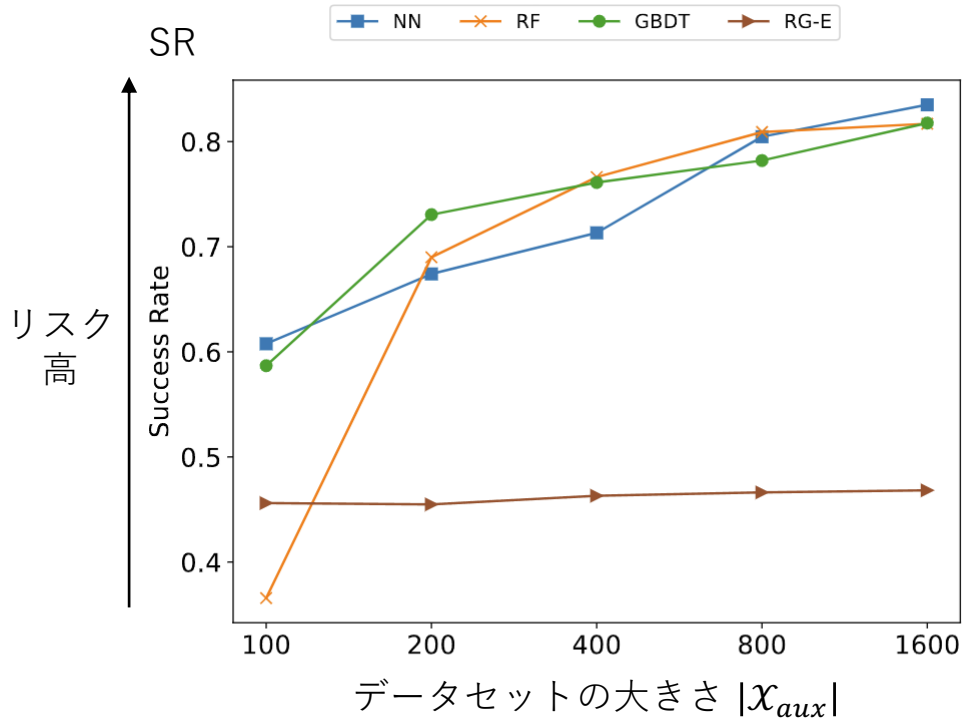
# 実験1，2，3，4

1. モデル $f$ （NN，RF，GBDT）に対する属性推定リスク
2. $f$ を線形モデル
3. 学習方法（SGD，Ｍｏｍｅｎｔｕｍ，ＲＭＳProp，Ａｄａｍ）
4. 説明変数の相関係数についての評価
   - 相関係数の計算は，目的変数が質的変数なので，説明変数が量的変数のときは相関比，質的変数のときはCramerの連関係数を用いる

# 結果1 – 攻撃者1

- データセットの行数 $|\mathcal{X}_{aux}|$ が大きくなるにつれて，MAEとSRの両方で推定リスクが上がった



# 結果2

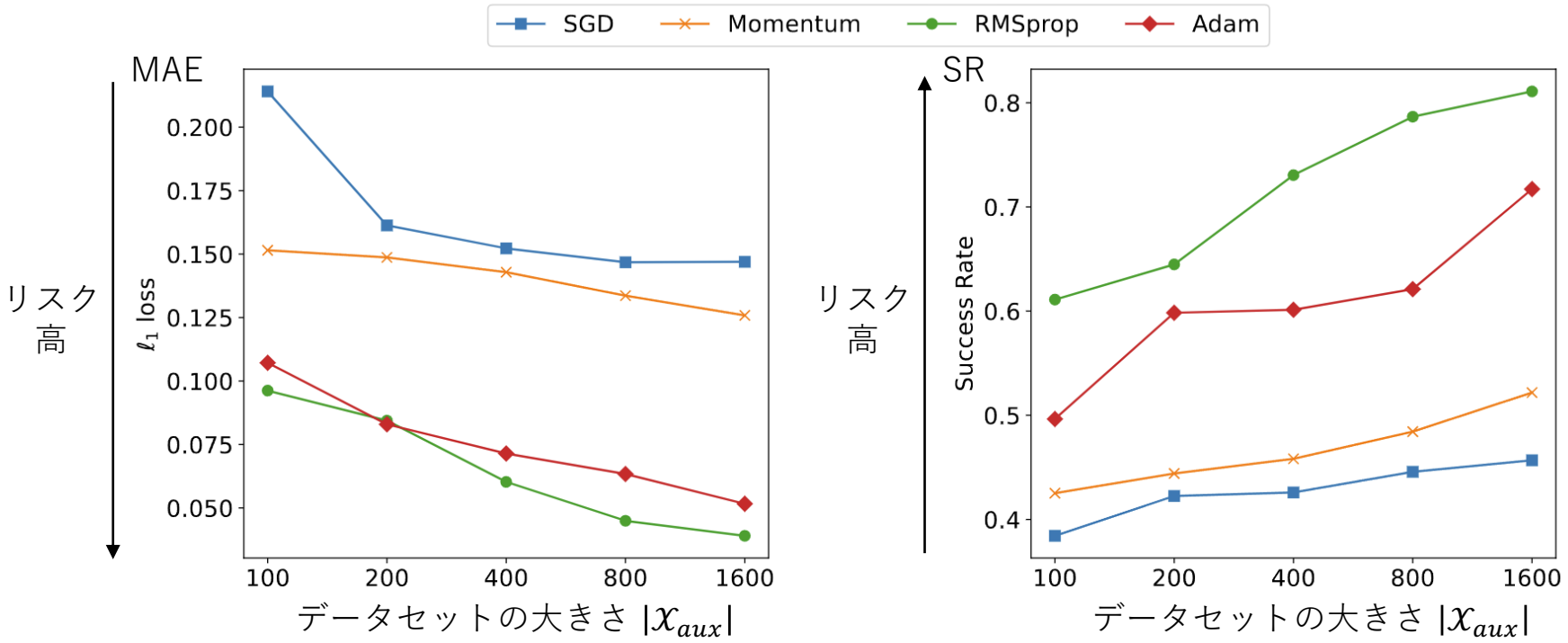
- モデル $f$ が線形モデルのときに攻撃が成功することを確認する
- 実際に，すべての列に対してSRが1となった

表 7　$\mathcal{X}_{test}$ の各列に対する SR

| 列 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SR | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

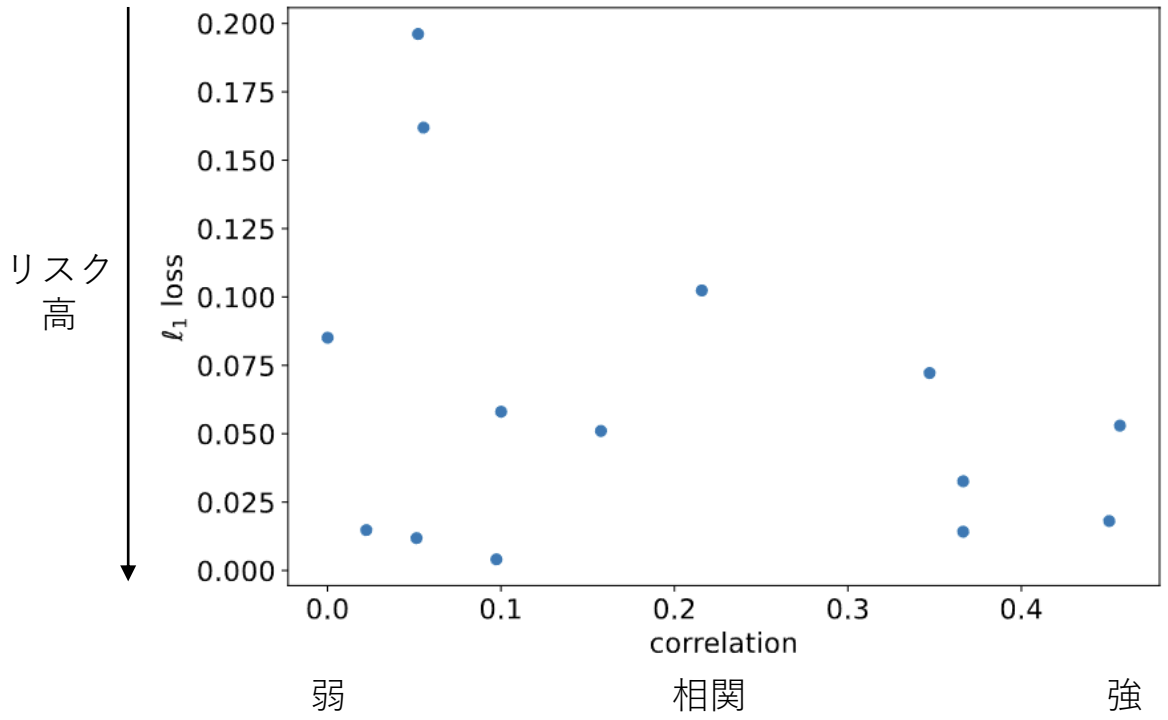
# 結果3

- MAEとSRのどちらも，SGDが最も属性推定の精度が低く，RMSPropが最も高かった



# 結果4 – 攻撃者 1

- 相関係数が大きい（$> 0.3$）列のMAEが低い値（$< 0.1$）に集中していた



# 結論

- $f$ がNN，RF，GBDTの間で推定リスクが特に高いものはなかった
- $f$ が線形モデルのとき誤差なく推定される
- $\psi$ の学習方法によって推定リスクが変わる
- 相関係数が大きい属性は推定リスクが高い

# 考察

- $\psi$ はShapley値から入力を推論するため，Shapley値にノイズを加えることで $\psi$ の精度を下げられる

- 今後の課題
  - Shapley値にノイズを加えたときの推定リスクの調査
  - Shapley値以外のXAIに対する推定リスクの調査