AIモデルの説明可能性LIMEとShapley値からの属性推定リスクの評価と比較

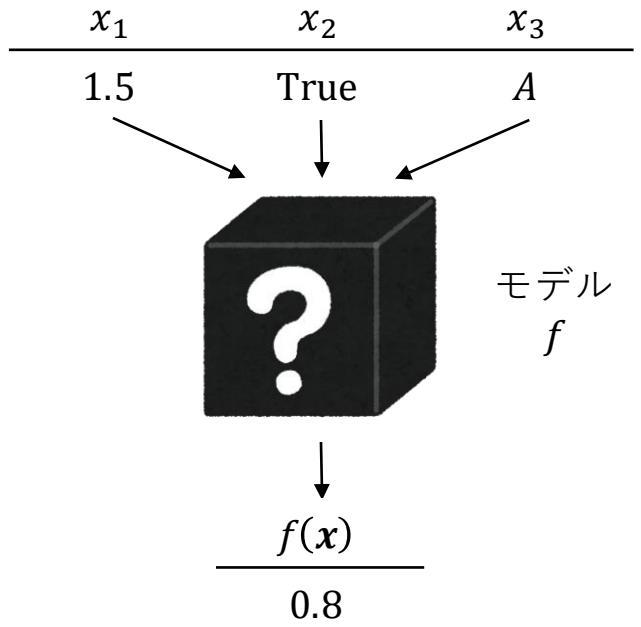
當麻僚太郎　菊池浩明
明治大学

背景

- 様々な機械学習モデル
  - ニューラルネットワークやランダムフォレスト，線形モデルなど

- 多くのモデルはブラックボックスである

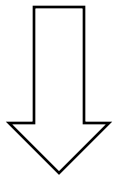
- 透明性や公平性が要求されている
  - GDPR第22条「説明を受ける権利」
  - ➢XAIによる説明



XAIの例

- Shapley値 [Shapley 1953]

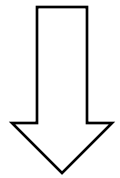
	x_1	x_2	x_3	$f(\boldsymbol{x})$
$\boldsymbol{x}$	1.5	True	A	0.8
$\boldsymbol{x}^1$	-0.4	False	B	0.6
$\boldsymbol{x}^2$	0.1	False	A	0.3
$\boldsymbol{x}^3$	0.8	True	C	0.9
$\boldsymbol{x}^4$	-1.1	True	A	0.2



	s_1	s_2	s_3
$\boldsymbol{s}$	0.32	0.10	-0.12

- LIME [Ribeiro 2016]

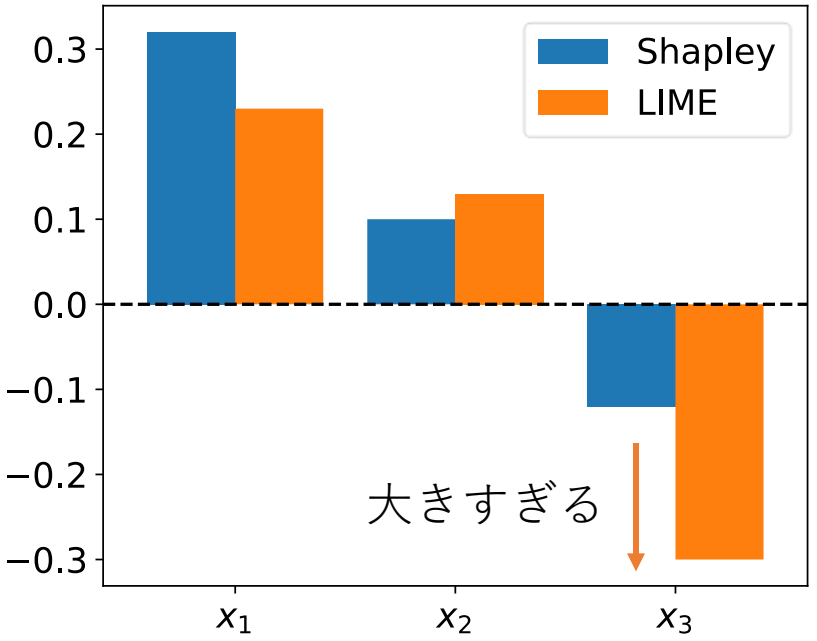
	x_1	x_2	x_3	$f(\boldsymbol{x})$
$\boldsymbol{x}$	1.5	True	A	0.8
$\boldsymbol{x}^1$	-0.4	False	B	0.6
$\boldsymbol{x}^2$	0.1	False	A	0.3
$\boldsymbol{x}^3$	0.8	True	C	0.9
$\boldsymbol{x}^4$	-1.1	True	A	0.2



	w_1	w_2	w_3
$\boldsymbol{w}$	0.23	0.13	-0.30

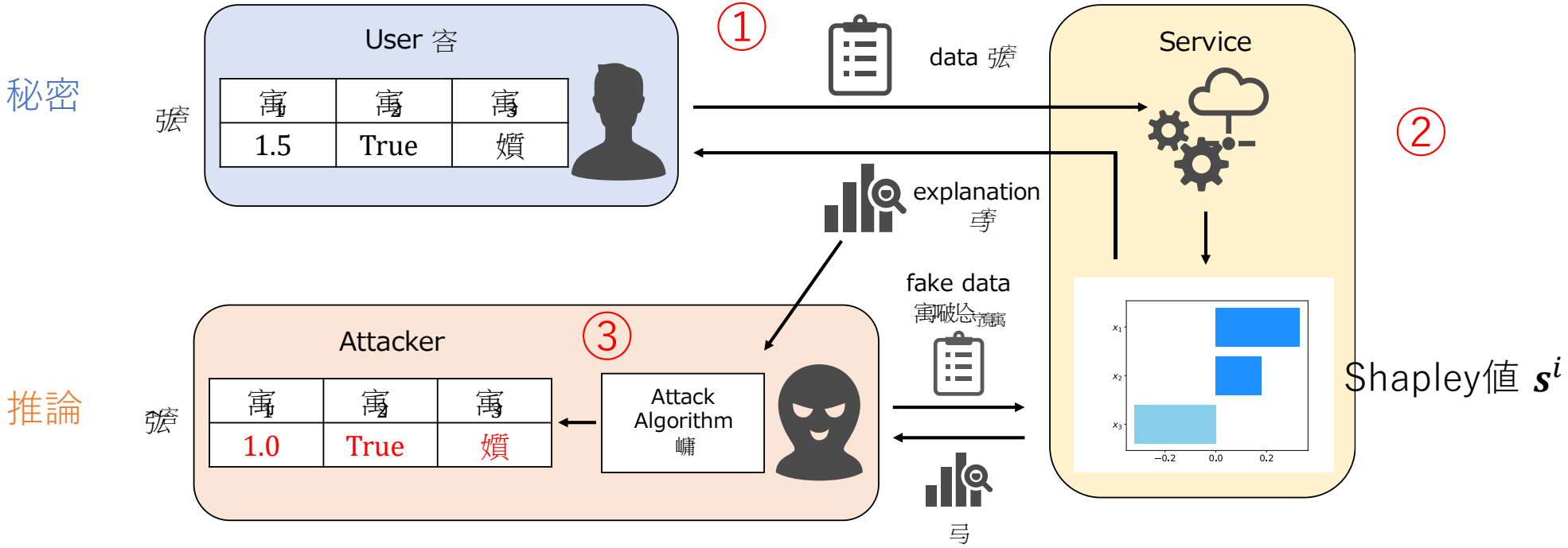
Shapley値とLIMEの比較

- モデル $f(x) = \frac{1}{1+\exp(-x_1-x_2-x_3)}$
- 入力 $\boldsymbol{x} = (1.5, \mathrm{True}, A) = (1.5, 1, -1)$ としたので，x_2 と x_3 の影響度は正負は逆だが同程度のはず
- この例では，Shapley値の方が適切



XAIの課題： 属性推論攻撃 [Luo 2022]

1. ユーザがサービス上のモデルにデータ $\boldsymbol{x}^i$ を送信
2. モデルの出力 $f(\boldsymbol{x}^i)$ とShapley値ベクトル $\boldsymbol{s}^i$ を算出
3. 攻撃者が窃取した $\boldsymbol{s}^i$ から攻撃モデル ψ により $\hat{\boldsymbol{x}}^i$ を推論



Research Questions

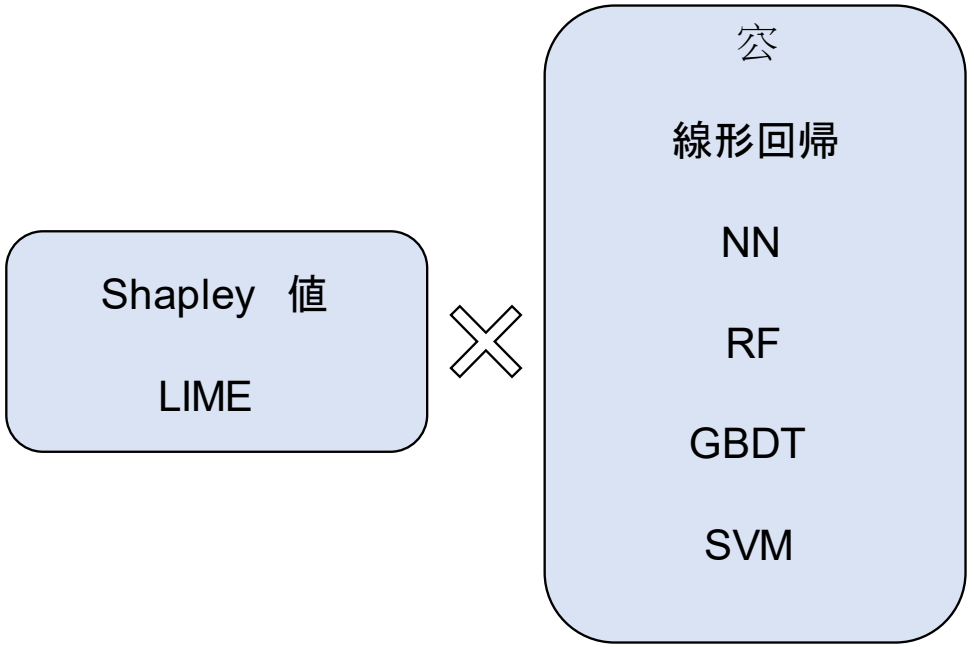
1. Shapley値とLIME，属性推論攻撃に対してより脆弱なのはどちらか？

2. どの機械学習モデルが脆弱か？

提案方式

- 学習モデル f に対するXAI： Shapley値とLIMEからの属性推定 ψ のリスクを明らかにする

Q. どの組み合わせのリスクが高いか？



研究方法

- モデル f （LR，NN，RF，GBDT，SVM）

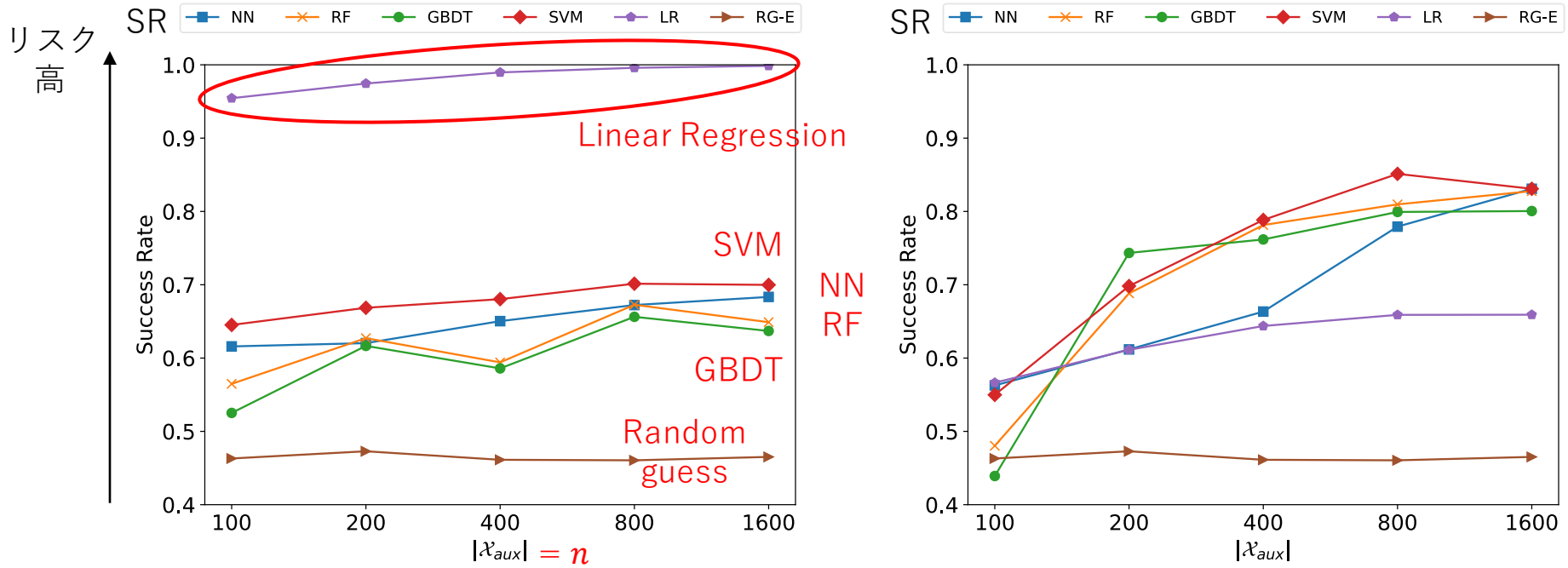
- 評価指標
  - 攻撃者のMAE
    - $\ell_1(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left|\hat{x}_i^j - x_i^j\right|$
  - 攻撃成功率SR
    - 推定に成功した入力特徴量の割合
    - $SR(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \frac{success(\hat{\boldsymbol{x}}, \boldsymbol{x})}{mn}$

表 4.1: 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
Bank Marketing [11]	45211	2	16
Credit Card [12]	30000	2	24
	m		n

結果1：属性推定リスク

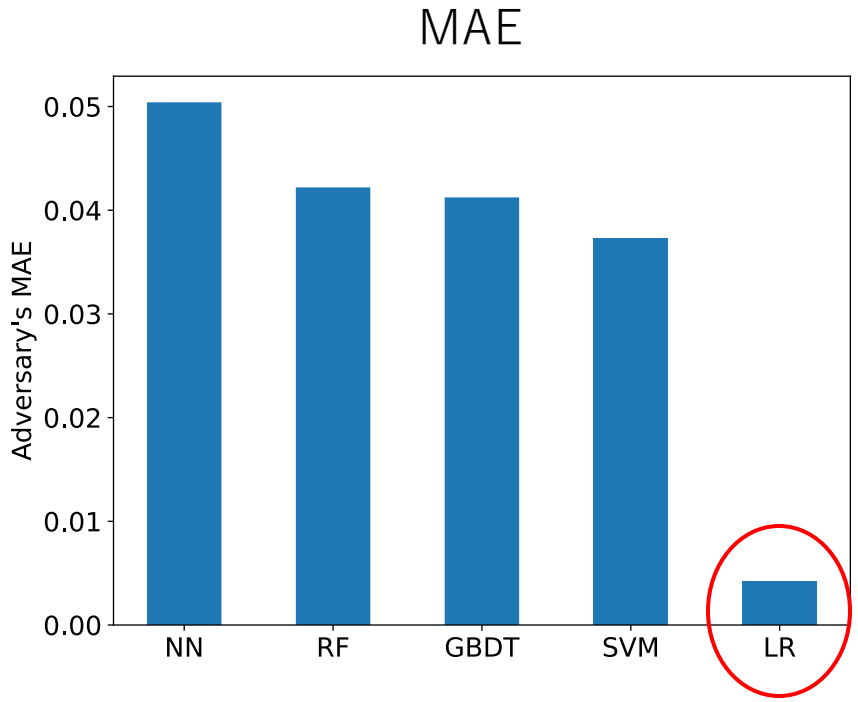
- レコード数 $|\mathcal{X}_{aux}| = n$ が大きくなるにつれて，推定リスクが上がった

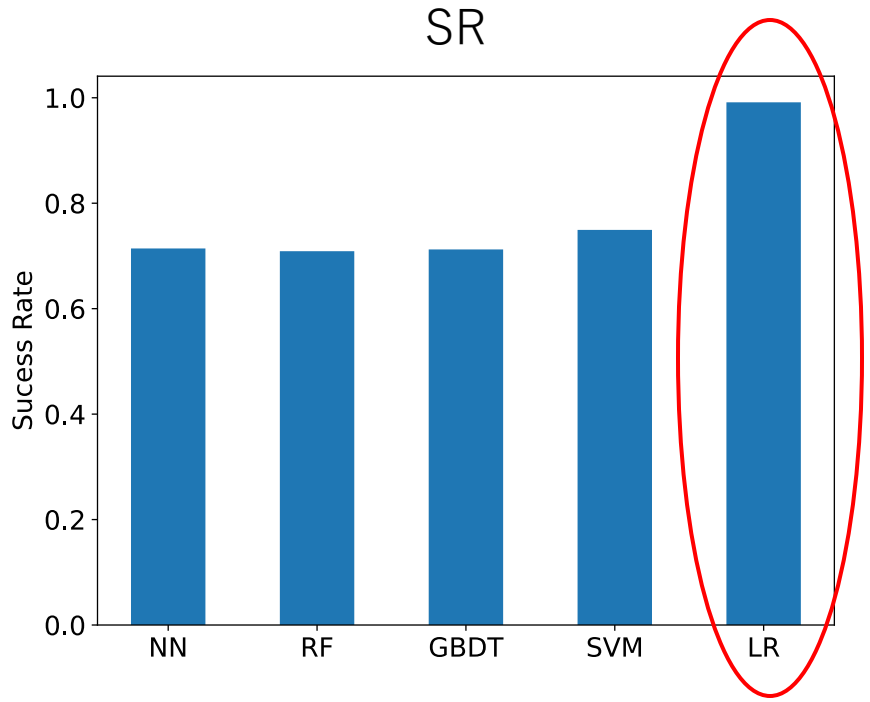


Shapley値　　LIME

結果2：モデルごとの脆弱性

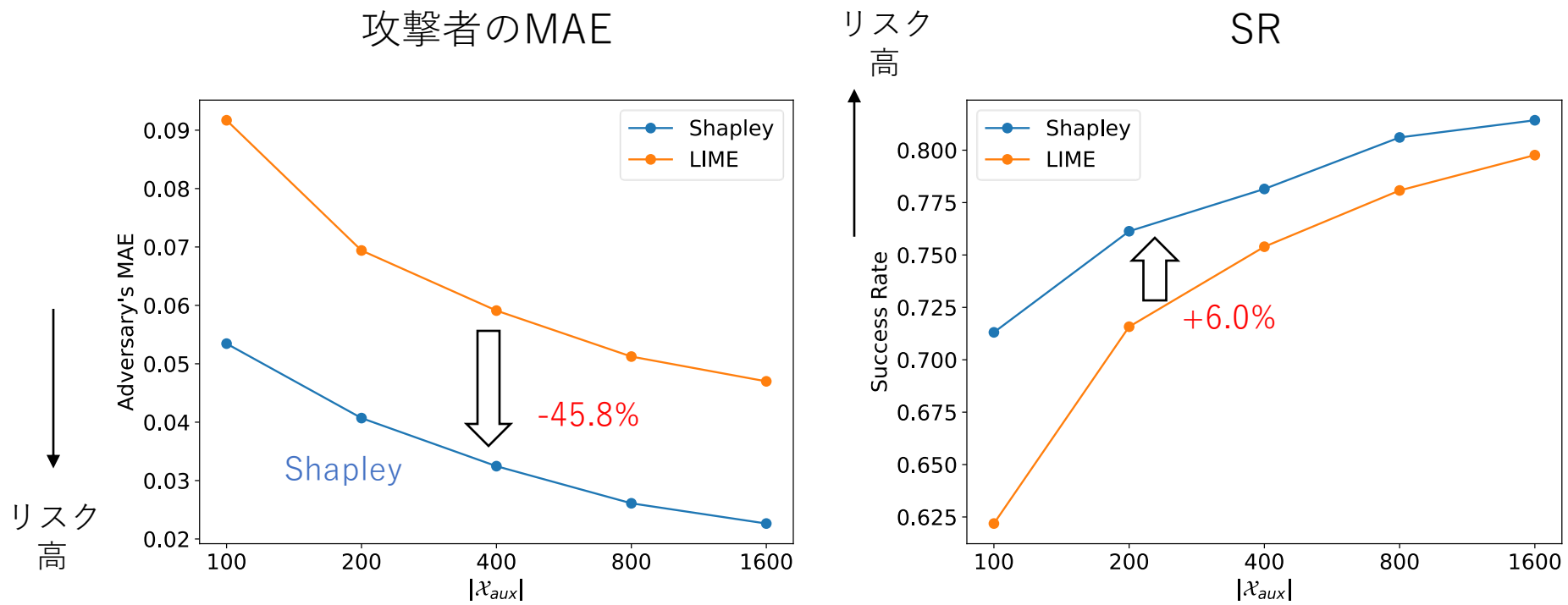
- モデル f が線形モデルのときShapley値から属性推論される





結果3：Shapley値とLIMEの比較

- 補助データセットの行数 $|\mathcal{X}_{aux}|$ に対するMAEとSRの変化



考察： 線形モデルの脆弱性

補題 1

f を線形回帰モデル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，

$$f\left(x_{[S\cup\{i\}]}\right) - f\left(x_{[S]}\right) = \beta_i\left(x_i - x_i^0\right)$$

である．

- 補題より，Shapley値 $s_i = \sum_{S\subseteq N\setminus\{i\}} \frac{|S|!(n-|S|-1)!}{n!}\left(f\left(\boldsymbol{x}_{[S\cup\{i\}]}\right) - f\left(\boldsymbol{x}_{[S]}\right)\right) = \frac{|S|!(n-|S|-1)!}{n!}\beta_i\left(x_i - x_i^0\right)$ となり，属性 i に対してShapley値 s_i が線形式で求まる

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Shapley値からの攻撃者の $\mathrm{MAE} = 0$ である．

- f が線形モデルのとき，補助データセットの行数が特徴量の数 n より十分多ければ，攻撃者は入力ベクトルを誤差なく推論出来る

結論

1. Shapley値とLIMEどちらの方がより脆弱か？
   - ➢Shapley値の方が脆弱（相対誤差で46%推定リスクが高い）

2. どの機械学習モデルが脆弱か？
   - ➢線形モデルとShapley値の組み合わせは脆弱
   - ➢それ以外のモデルは推定リスクに差がない

- 対策
  - モデル f へのリクエスト数制限
  - ノイズを加えた説明ベクトル [Watson 2022]
  - 属性推定リスクを低減したXAI手法の提案