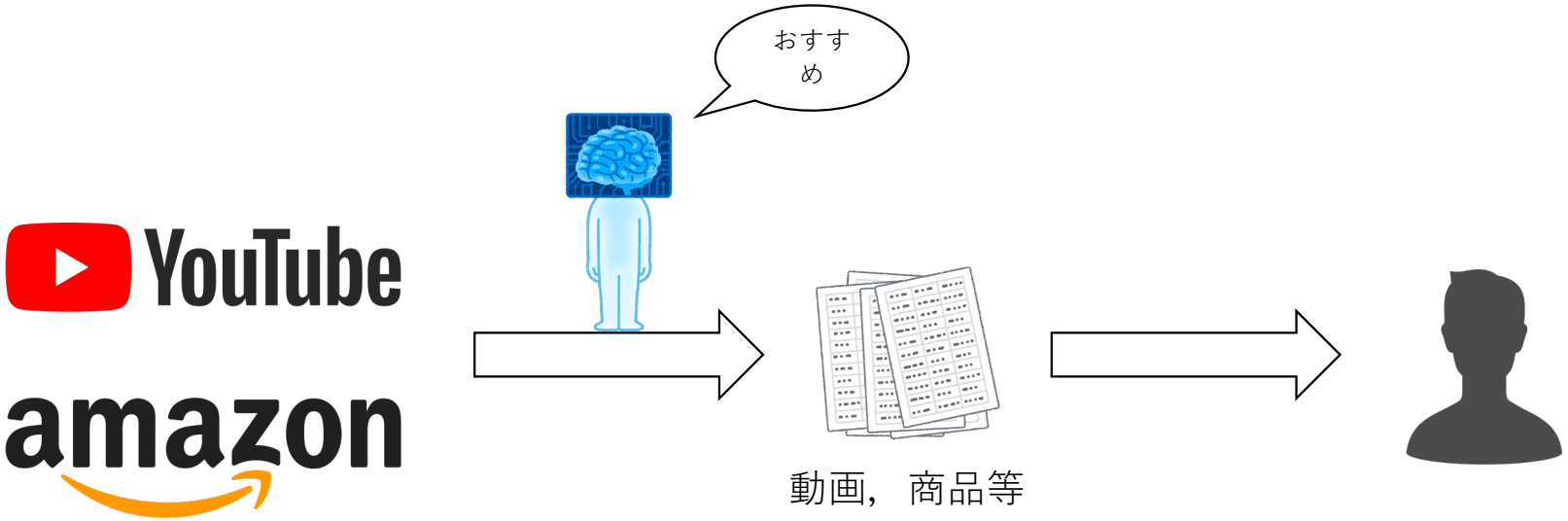
AIモデルの説明可能性Shapley値からの属性推定リスクの評価とその対策

當麻 僚太郎

説明可能性技術（XAI）の重要性

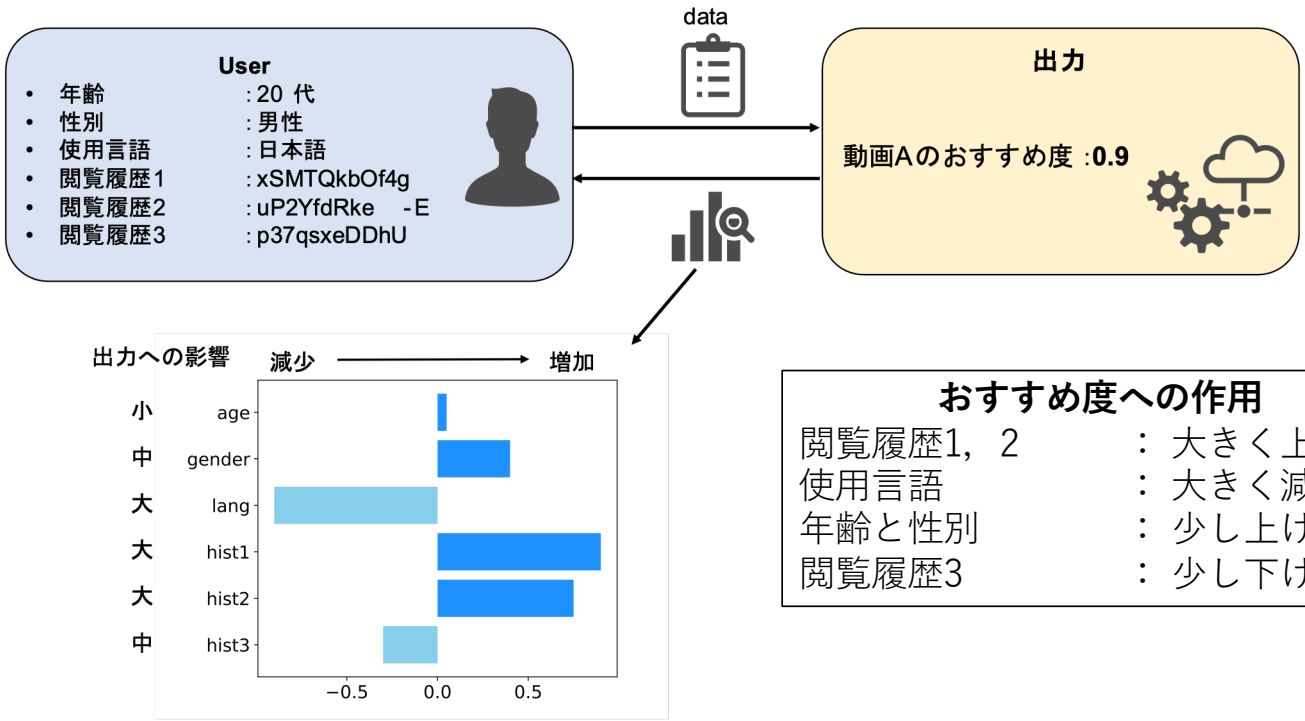


リコメンドの仕組みが不透明
・どんなデータを使っている？
・人種や性別などの属性は影響している？
・データのどの部分からおすすめが決まる？

→ **XAIによる説明**

XAIの例

- Shapley値 [Shapley 1953]
- LIME [Ribeiro 2016]

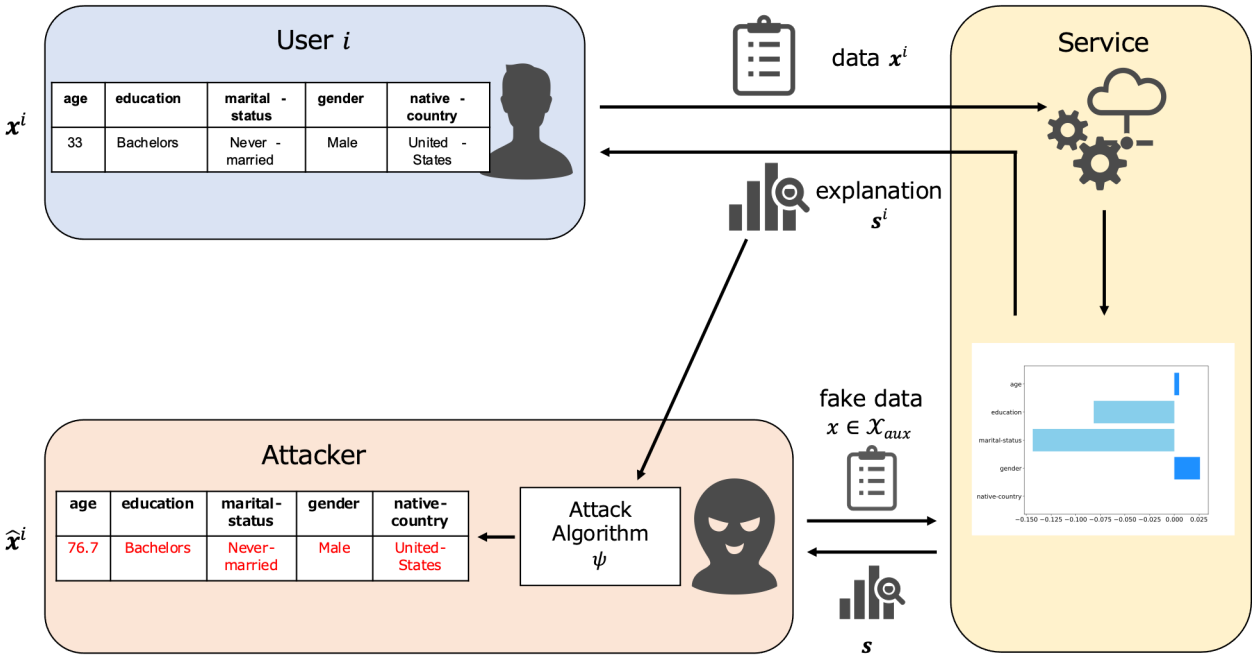


おすすめ度への作用

閲覧履歴1，2	： 大きく上げる
使用言語	： 大きく減らす
年齢と性別	： 少し上げる
閲覧履歴3	： 少し下げる

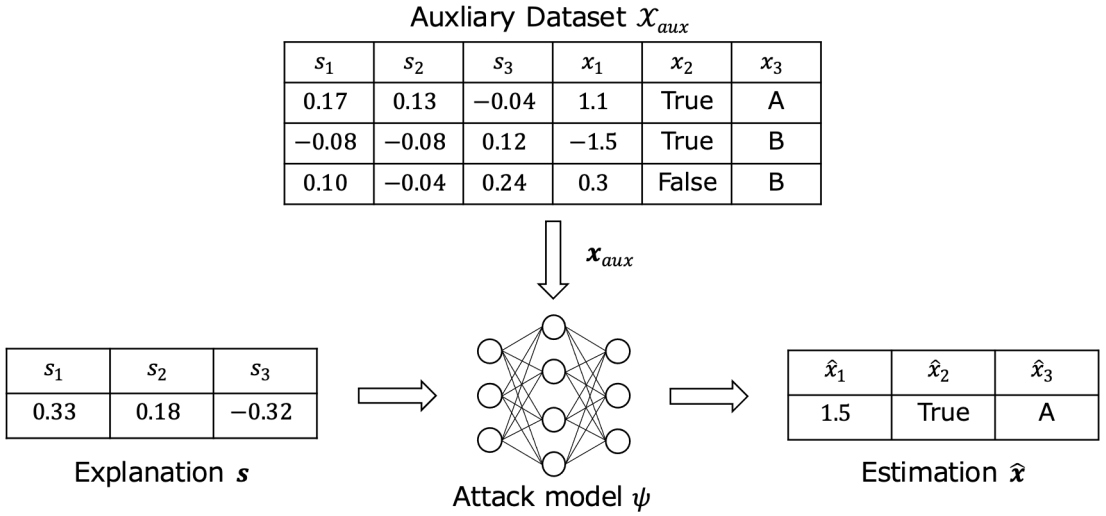
XAIの課題： 特徴量推論攻撃 [Luo 2022]

- 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



攻撃者

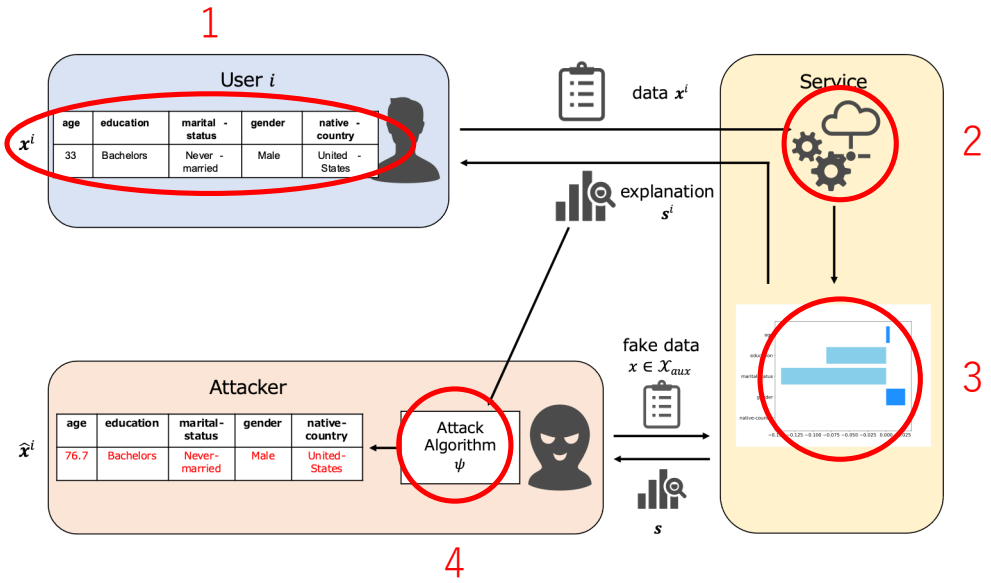
- 訓練データセット $\mathcal{X}_{train}$ と同じ分布に従う補助データセット $\mathcal{X}_{aux}$ を持っていると仮定
- 全ての $x_{aux} \in \mathcal{X}_{aux}$ について対応する説明データ $\mathcal{S}_{aux}$ から攻撃アルゴリズム ψ: $\mathcal{S}_{aux} \rightarrow \mathcal{X}_{aux}$ を訓練する



[Luo 2022]の問題点

- 属性推定リスクがどんな時に増加するのか不明

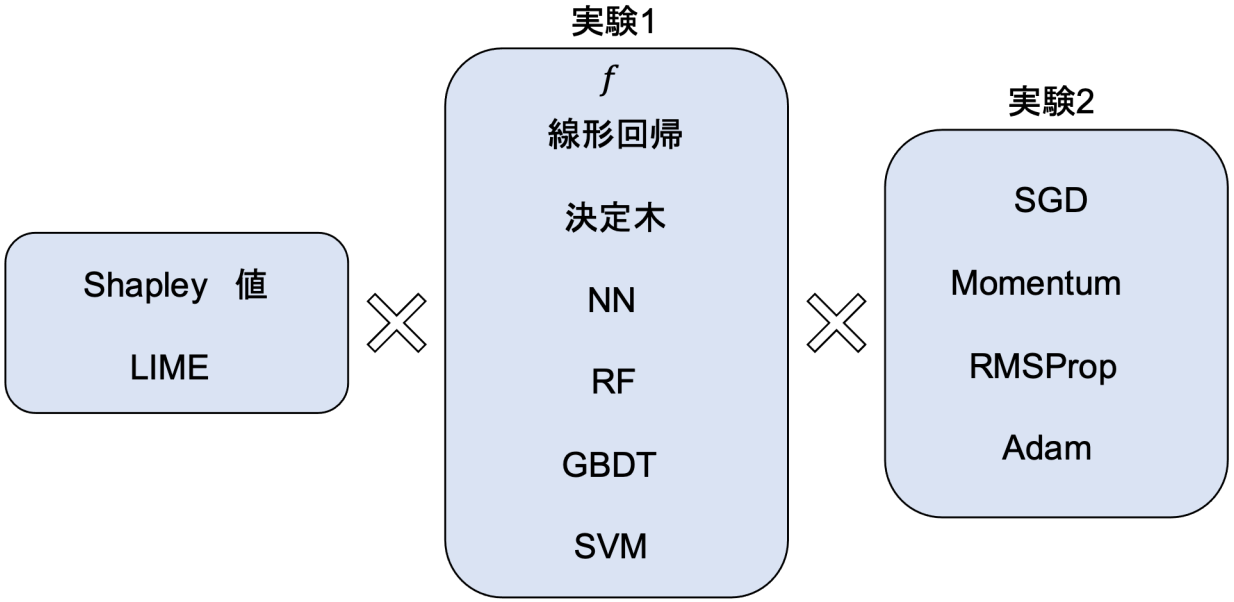
1. 入力属性間の相関
2. モデルの構造
3. Shapley値以外のXAI
4. ψ の学習方式



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

1. モデル f （線形回帰，NN，RF，GBDT，SVM）
2. 学習方法（SGD，Ｍｏｍｅｎｔｕｍ，ＲＭＳProp，Ａｄａｍ）

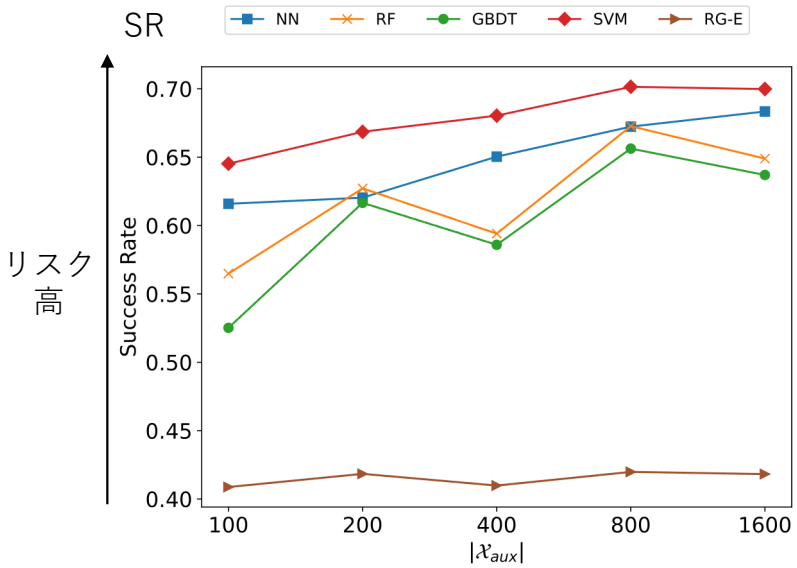
- 評価指標
  - Mean Absolute Error （MAE）
    - $\ell_1(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left|\hat{x}_i^j - x_i^j\right|$
  - Success Rate (SR)
    - 推定に成功した入力特徴量の割合
    - $SR(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \frac{success(\hat{\boldsymbol{x}}, \boldsymbol{x})}{mn}$

表 4.1: 使用データセット

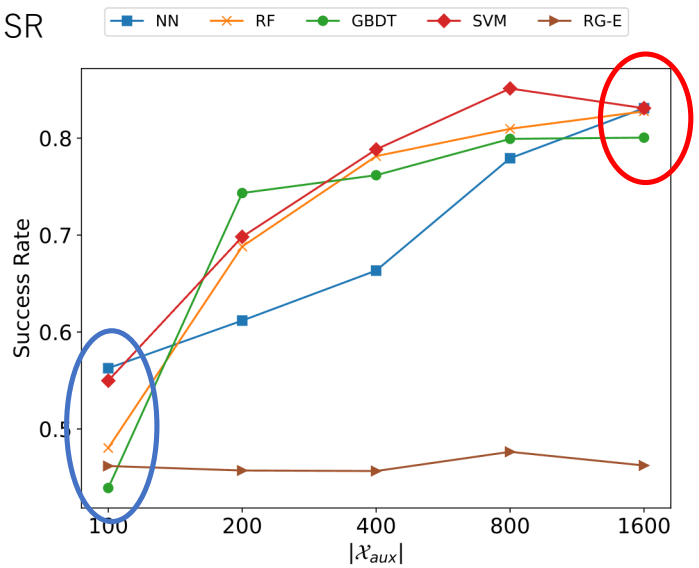
データセット	レコード数	クラス	特徴量
Adult [10]	48842	2	14
Bank Marketing [11]	45211	2	16
Credit Card [12]	30000	2	24

結果1 (Adultデータセット)

- 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



Shapley値



LIME

表 4.3: Shapley 値と LIME の属性推定リスク比較

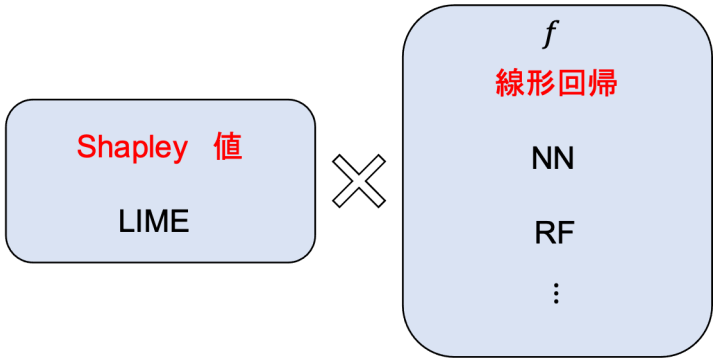
	平均 SR	
	$\|\mathcal{X}_{aux}\| = 100$	$\|\mathcal{X}_{aux}\| = 1600$
Shapley 値	**0.65**	0.77
LIME	0.62	**0.83**

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

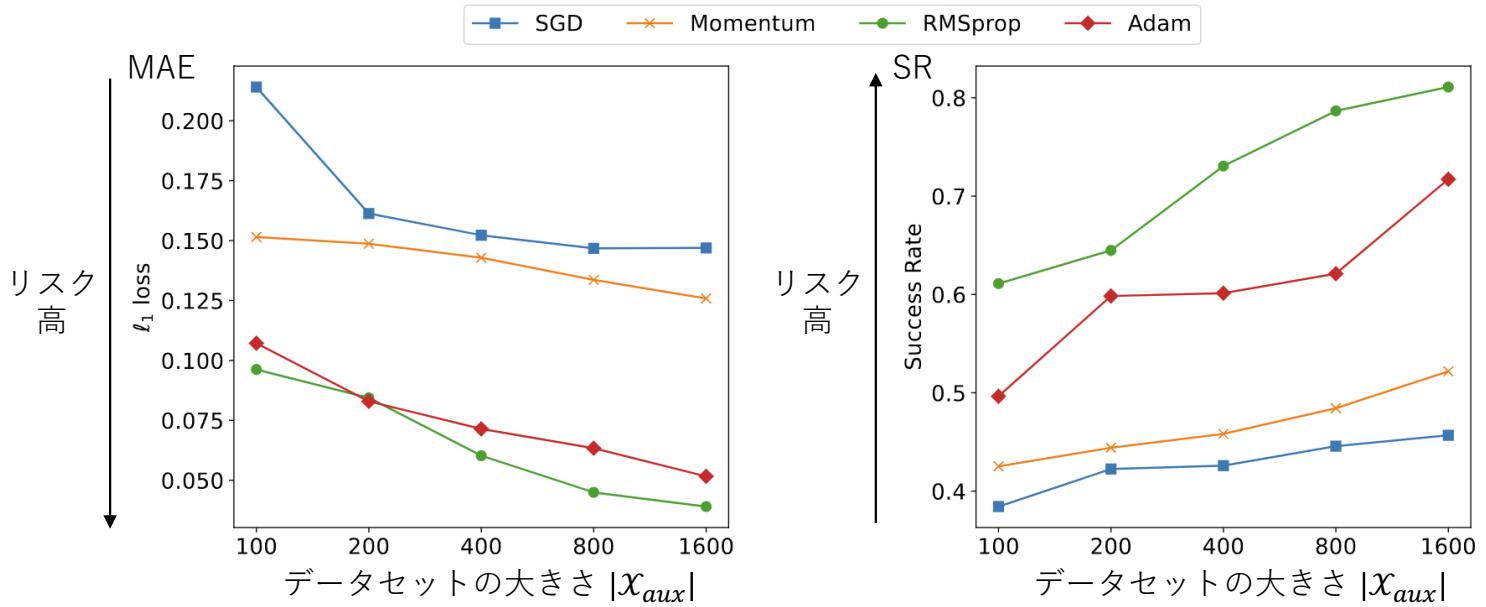
		x_1	x_2	x_3	x_4	y
x_{test}	$\boldsymbol{x}$	1.8	0.1	0.3	−0.4	1.9
x_{aux}	$\boldsymbol{x}^0$	−1	0.3	−0.3	0.5	
Shapley値	$\boldsymbol{s}$	1.30	0.02	0.06	−0.04	
推定	$\hat{\boldsymbol{x}}$	1.8	0.1	0.3	−0.4	
MAE		0.0	0.0	0.0	0.0	

ψ



結果2

- 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



結論

- モデル f の構造によって属性推定リスクは変化しないが，f が線形モデルのとき属性推定が誤差なく出来る
- 補助データが少ない時はShapley値の方が推定リスクが高いが多い時はLIMEの方が推定リスクが高い
- 攻撃アルゴリズム ψ の学習方法によって推定リスクが変わる
- 今後の課題
  - データや説明ベクトルにノイズを加えたときの推定リスクの調査
  - Shapley値とLIME以外のXAIに対する推定リスクの調査