Record Reconstruction Risk from LIME XAI Metrics

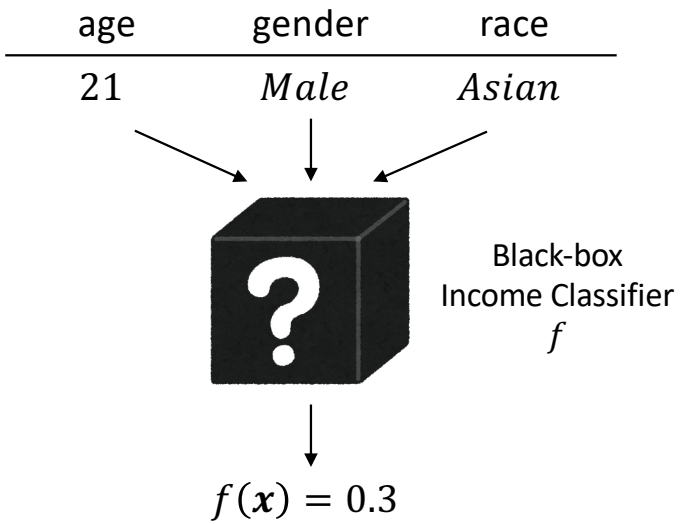
Ryotaro Toma and Hiroaki Kikuchi

Meiji University

Background

- ML models are almost black box
  - Neural Networks
  - Random Forests
  - etc.

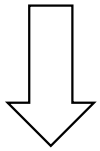
- Explainable AI (XAI) provides …
  - Transparency
  - Fairness
  - A sense of understanding



Two XAI techniques

- Shapley values [Shapley 1953]

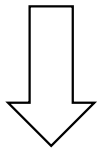
	x_1	x_2	x_3	$f(\boldsymbol{x})$
$\boldsymbol{x}$	1.5	True	A	0.8
$\boldsymbol{x}^1$	-0.4	False	B	0.6
$\boldsymbol{x}^2$	0.1	False	A	0.3
$\boldsymbol{x}^3$	0.8	True	C	0.9
$\boldsymbol{x}^4$	-1.1	True	A	0.2



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

- LIME [Ribeiro 2016]

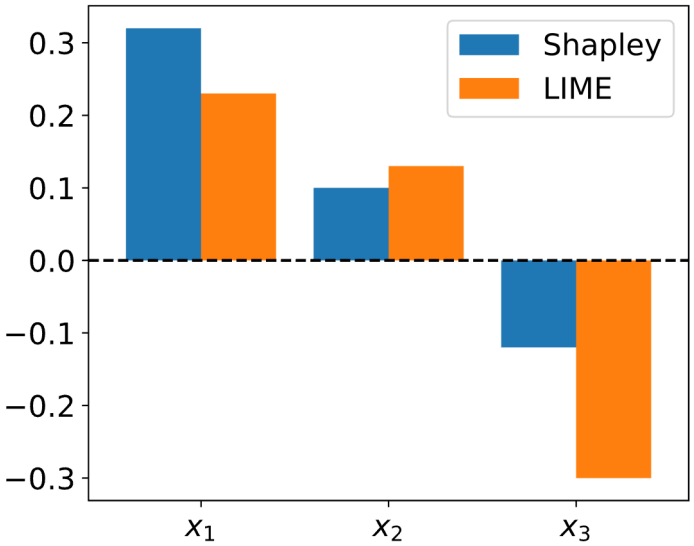
	x_1	x_2	x_3	$f(\boldsymbol{x})$
$\boldsymbol{x}$	1.5	True	A	0.8
$\boldsymbol{x}^1$	-0.4	False	B	0.6
$\boldsymbol{x}^2$	0.1	False	A	0.3
$\boldsymbol{x}^3$	0.8	True	C	0.9
$\boldsymbol{x}^4$	-1.1	True	A	0.2



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

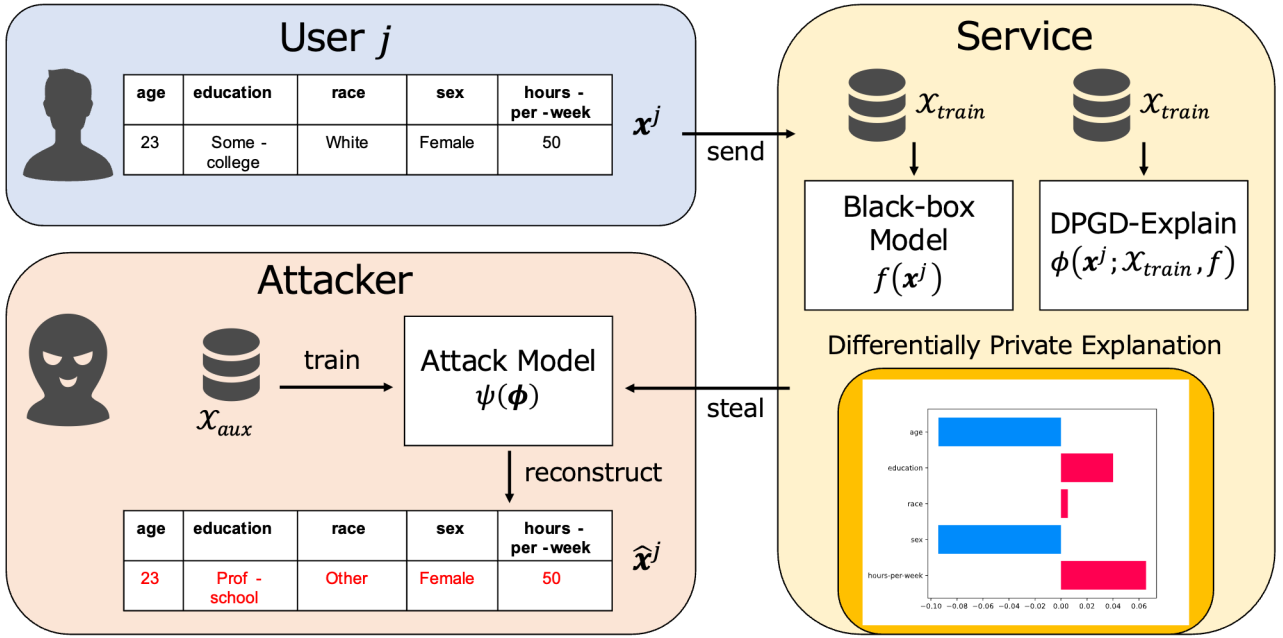
Shapley values and LIME

- The model $f(x) = \frac{1}{1+\exp(-x_1-x_2-x_3)}$
- The influences of x_2 and x_3 should be same because the input $\boldsymbol{x} = (1.5, \mathrm{True}, A) = (1.5, 1, -1)$
- In this case, the explanation by Shapley values is more appropriate



Issue of XAI: Feature Inference Attack [Luo 2022]

- The attacker can predict or infer a confidential input vector $\boldsymbol{x}^i$ from the given Shapley values $\boldsymbol{s}^i$



Research Questions

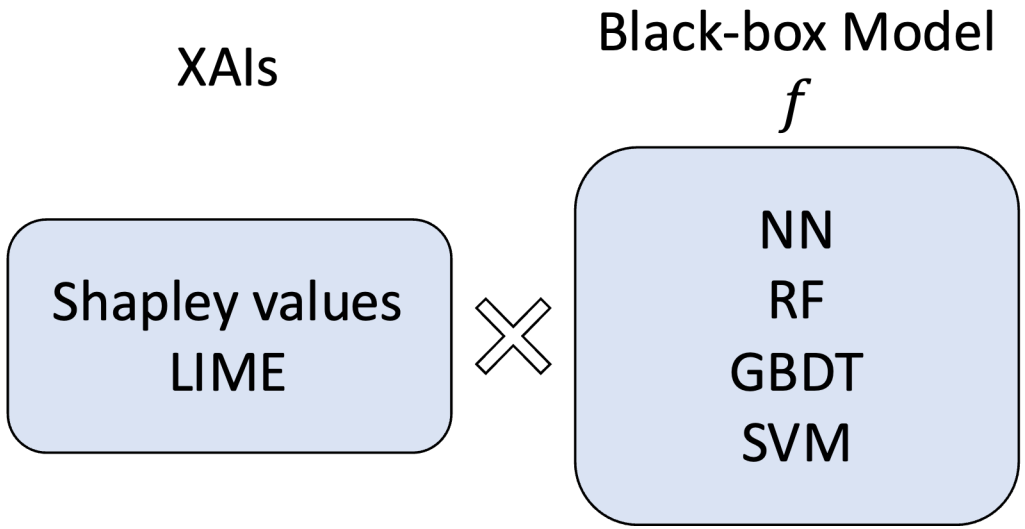
1. Is the explanation by LIME as risky as Shapley values in the record reconstruction?

2. Which machine learning models are vulnerable?

3. Which one is more vulnerable, Shapley values or LIME?

Our Proposal

- The record reconstruction risks for XAIs: Shapley values and LIME
- Q. What combinations are vulnerable?



Methodology

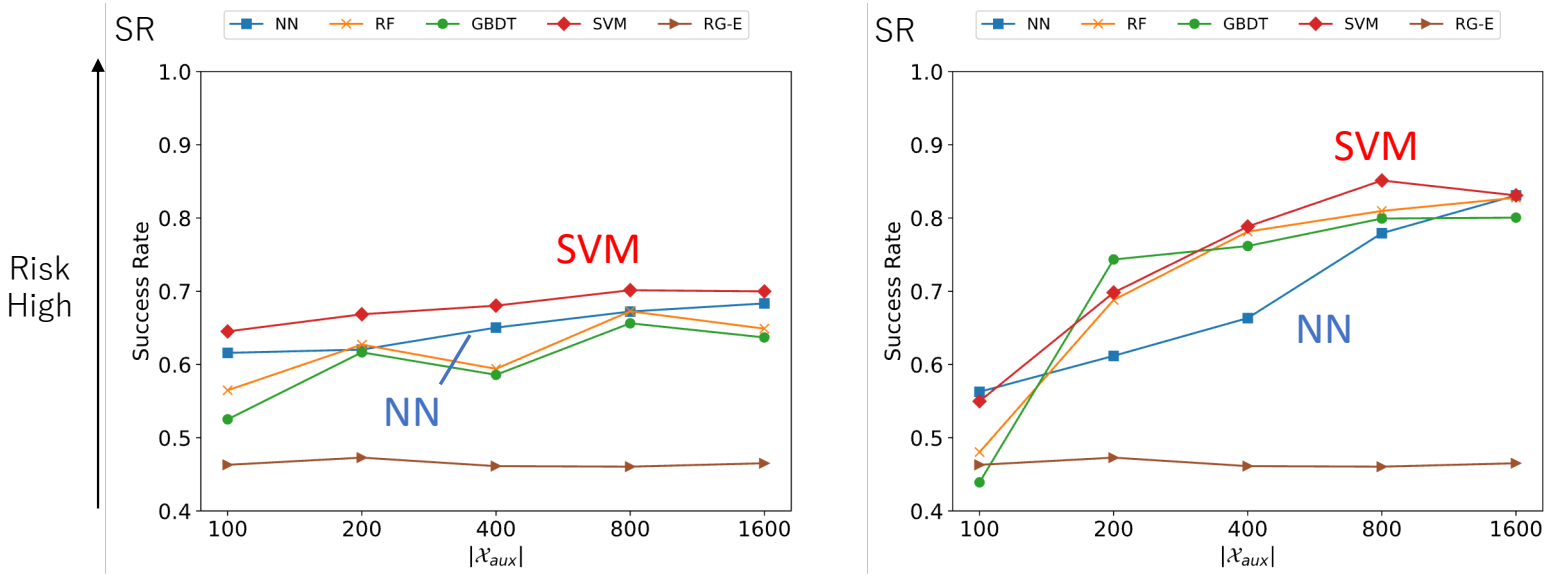
- model f （NN， RF， GBDT， SVM）
- 3 open datasets

Dataset	Records	Classes	Features
UCI Adult	48842	2	14
Bank Marketing	45211	2	16
Credit Card	30000	2	24

- metrics
  - Adversary's MAE
    - m: the number of rows, n: the number of features
    - $\ell_1(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left|\hat{x}_i^j - x_i^j\right|$
  - Success Rate
    - the fraction of the feature values that were identified successfully
    - $SR(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \frac{success(\hat{\boldsymbol{x}}, \boldsymbol{x})}{mn}$

Result 1: Reconstruction Risk

- The reconstruction risks almost increased as $|\mathcal{X}_{aux}|$ increased



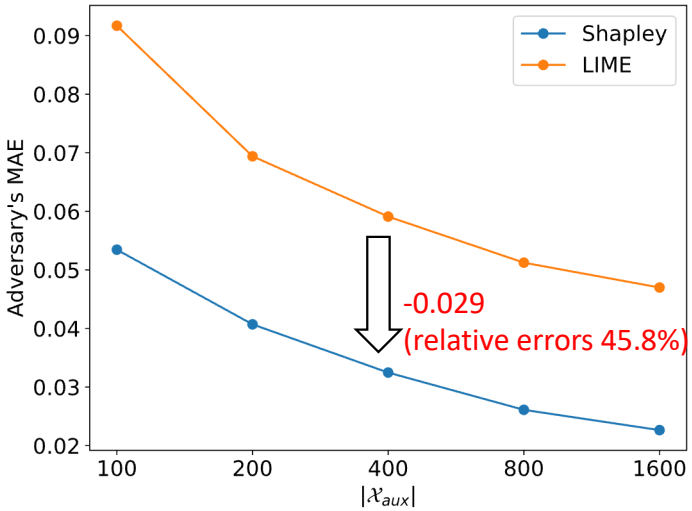
Shapley values

LIME

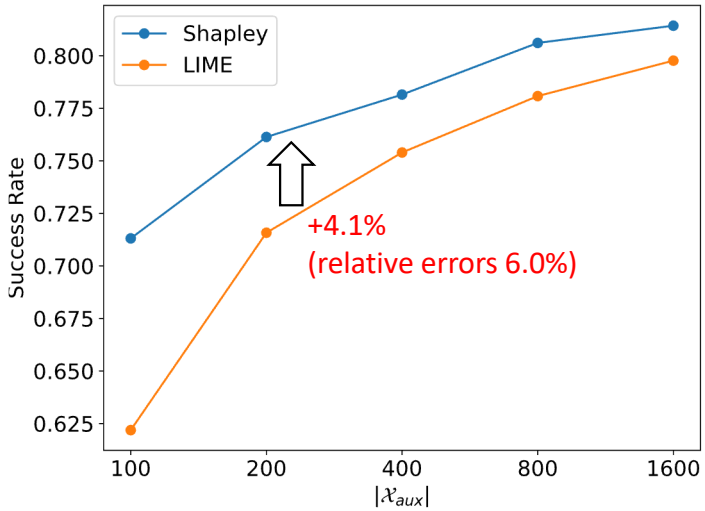
Result 2： Shapley values and LIME

- The risk of Shapley values is clearly higher than LIME

Adversary's MAE



SR



Conclusions

1. Is the explanation by LIME as risky as Shapley values in the record reconstruction?
   - ➢ LIME has the record reconstruction risk as similar as Shapley values
2. Which machine learning models are vulnerable?
   - ➢ All models have similar privacy risk
3. Which one is more vulnerable, Shapley values or LIME?
   - ➢ The explanation by Shapley values is more vulnerable

- Countermeasures against the reconstruction attack
  - Access control of XAI metrics
  - Investigation the change of risks by noise addition
  - Proposal of new robust XAI method