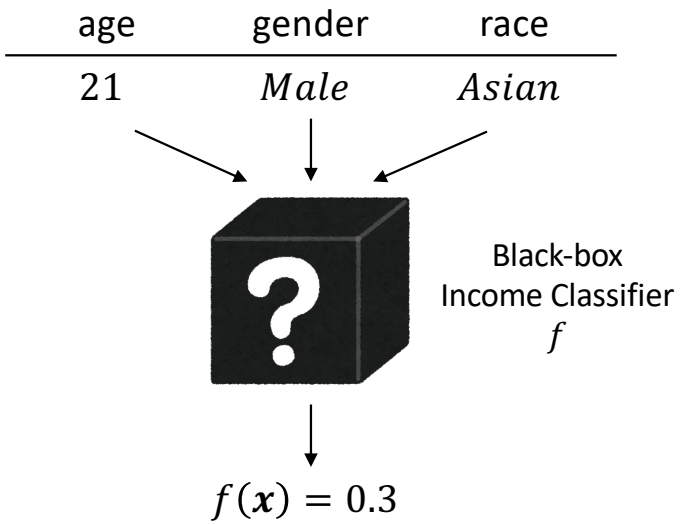
Combinations of AI Models and XAI Metrics Vulnerable to Record Reconstruction Attack

Ryotaro Toma and Hiroaki Kikuchi
Meiji University

Background

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

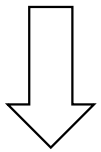
- Explainable AI (XAI) provides …
  - Transparency
  - Fairness
  - A sense of understanding



Shapley values [Shapley 1953]

- Many studies re-visited the Shapley values
- There are approximation methods
  - Monte Carlo sampling [Štrumbelj et al. 2016]
  - SHAP [Lundberg et al. 2017]
- Shapley values can explain the contributions of each feature

	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

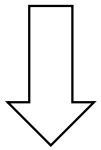


	s_1	s_2	s_3
$\boldsymbol{s}$	0.32	0.10	-0.12

LIME [Ribeiro et al. 2016]

- A method for providing an explanation for each input feature like Shapley values
- LIME approximates the behavior of f around the specific input vector $\boldsymbol{x}$
- Both XAI metrics are consistent in the sense

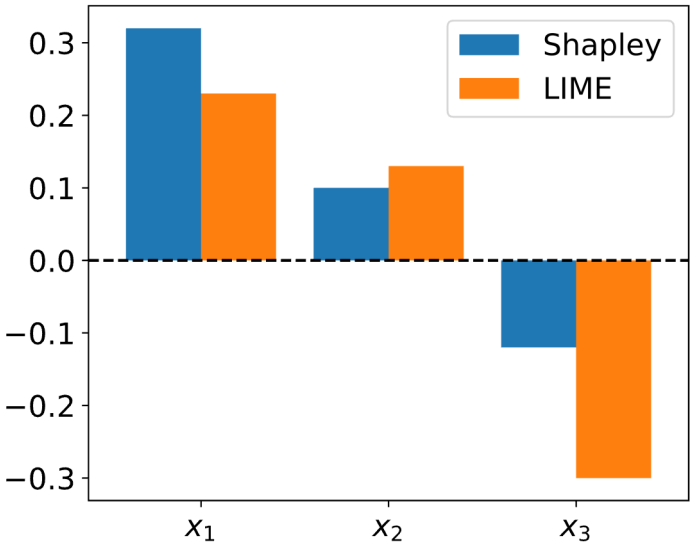
	x_1	x_2	x_3	$f(\boldsymbol{x})$
$\boldsymbol{x}$	1.5	True	A	0.8
$\boldsymbol{x}^1$	-0.4	False	B	0.6
$\boldsymbol{x}^2$	0.1	False	A	0.3
$\boldsymbol{x}^3$	0.8	True	C	0.9
$\boldsymbol{x}^4$	-1.1	True	A	0.2



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

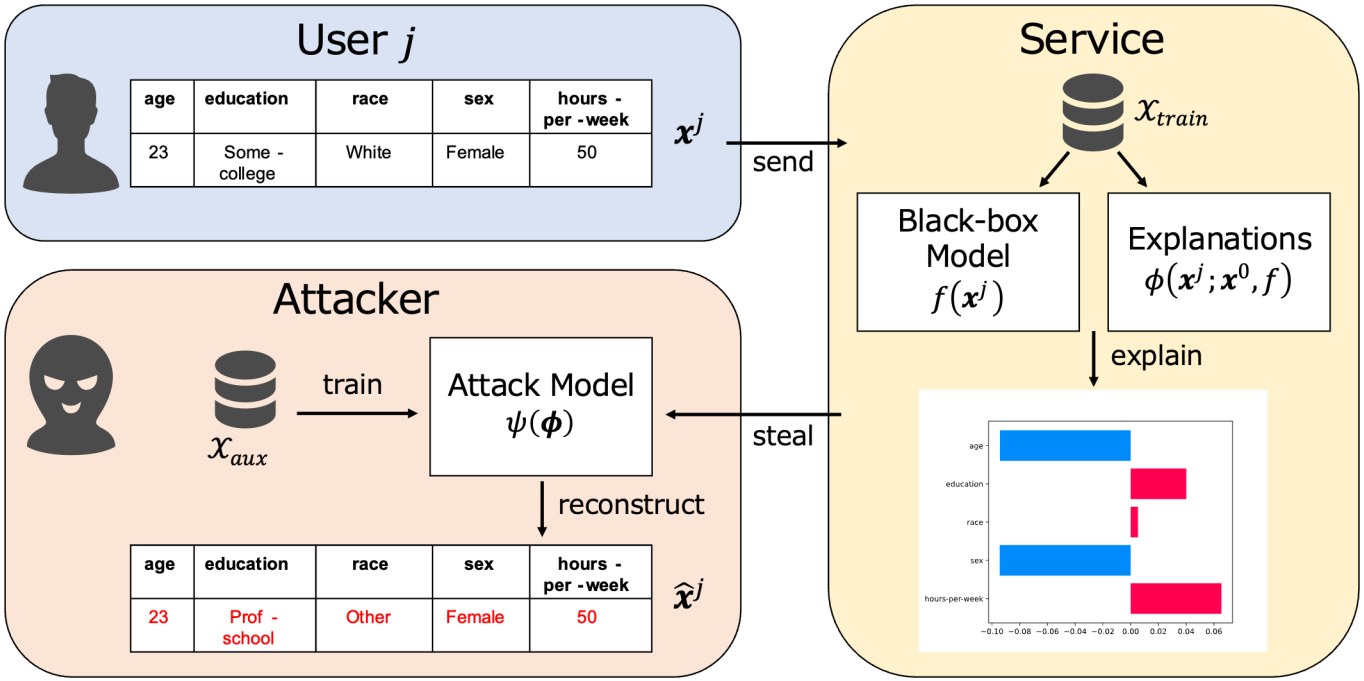
Shapley values and LIME

- 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



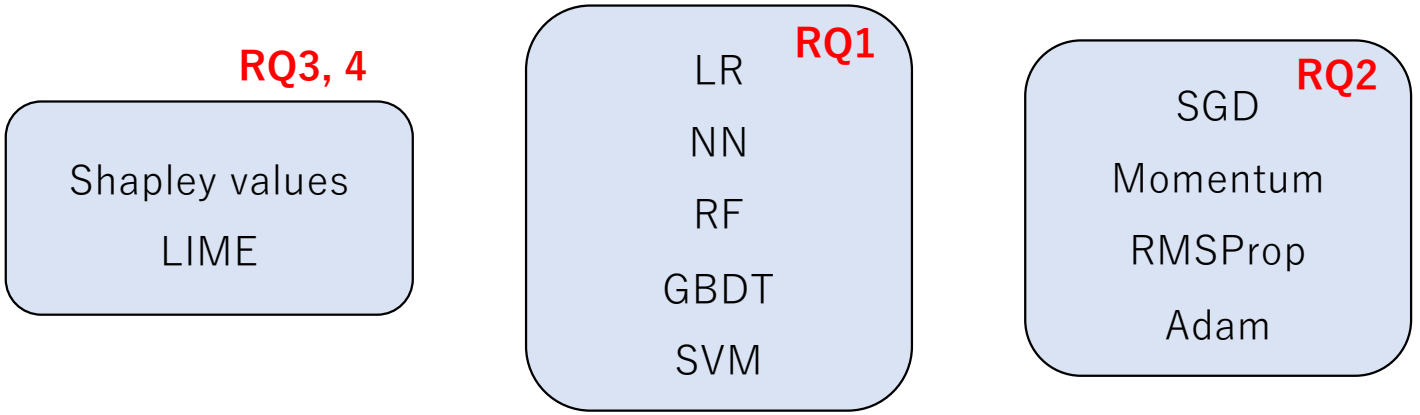
Issue of XAI: Feature Inference Attack [Luo et al. 2022]

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



Research Questions

1. Which black-box models are vulnerable?
2. Is there any difference depending on the optimization algorithms?
3. Is the explanation by LIME as risky as Shapley values in the record reconstruction?
4. Which one is more vulnerable, Shapley values or LIME?



The Answer to RQ1

Proposition 1. *Let f and ψ be a linear black-box model and a linear attack model, respectively. When $n < |\mathcal{X}_{aux}|$, the attacker's MAE with ψ is 0.*

- A record reconstruction could be perfect if the Shapley values from the linear model were to be used

- The attacker's MAE is zero when enough number of auxiliary data are given

Proof of Proposition 1

- Shapley value s_i is defined as the weighted average

$$
\begin{aligned}
s_i &= \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} f(\boldsymbol{x}_{[S \cup \{i\}]}) - f(\boldsymbol{x}_{[S]}) \\
&= \lambda_i (x_i - x_i^0) \qquad \text{where } \lambda_i = \textstyle\sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} \beta_i
\end{aligned}
$$

- Then, a linear attack model ψ can correctly estimate

$$
\begin{aligned}
\hat{x}_i &= \alpha_0 + \alpha_1 s_1 + \cdots + \alpha_n s_n \\
&= \alpha_0 + \alpha_1 (\lambda_1 (x_1 - x_1^0)) + \cdots + \alpha_n (\lambda_n (x_n - x_n^0)) \\
&= \gamma_0 + \gamma_1 x_1 + \cdots + \gamma_n x_n \\
&\text{where } \gamma_i = \alpha_i \lambda_i \text{ and } \gamma_0 = \alpha_0 - \textstyle\sum_{k=1}^{n} \alpha_k \lambda_k x_k^0
\end{aligned}
$$

Methodology

- model f （LR, NN, RF, GBDT, SVM）

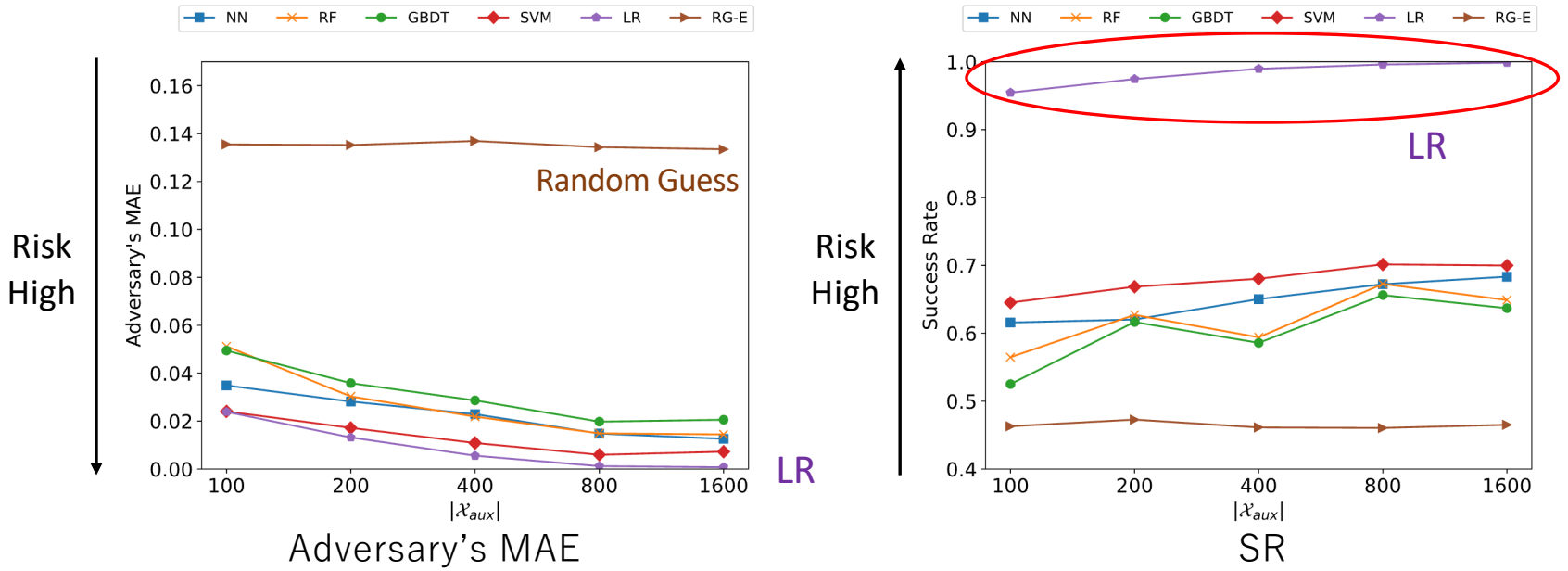
- 3 open datasets

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
    - $SR(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \frac{success(\hat{\boldsymbol{x}}, \boldsymbol{x})}{mn}$

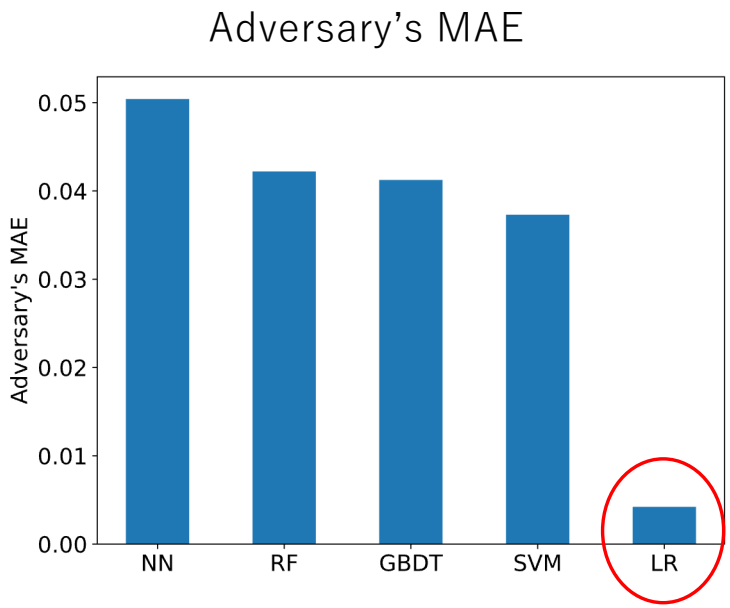
Result 1: Reconstruction Risk (RQ1)

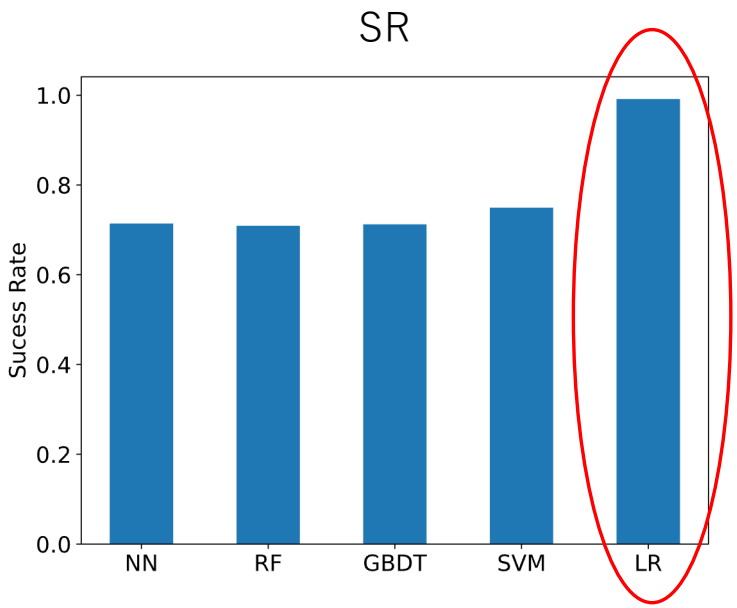
- The reconstruction risks almost increased as $|\mathcal{X}_{aux}|$ increased
- The attacker succeeded exactly in estimation



Result 2： Vulnerability by models (RQ1)

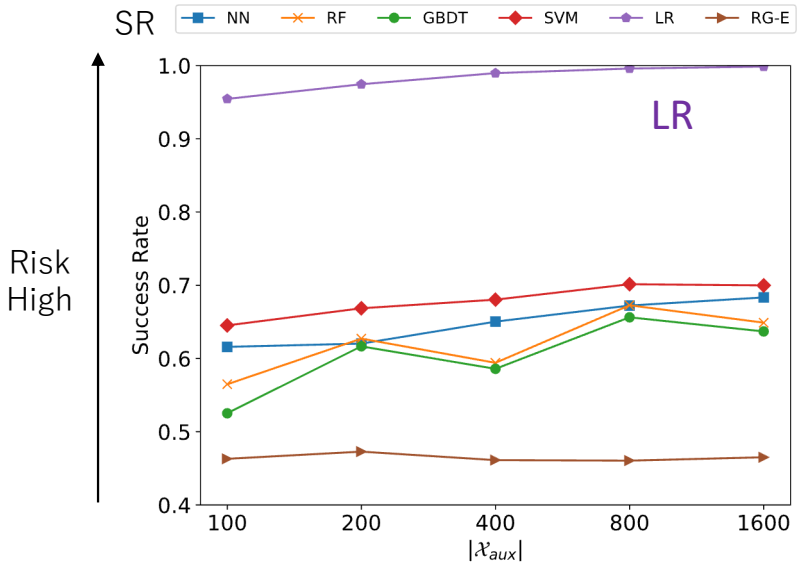
- We show the aggregated MAEs and Success Rates
- The risk of Linear Regression is the most vulnerable for both metrics



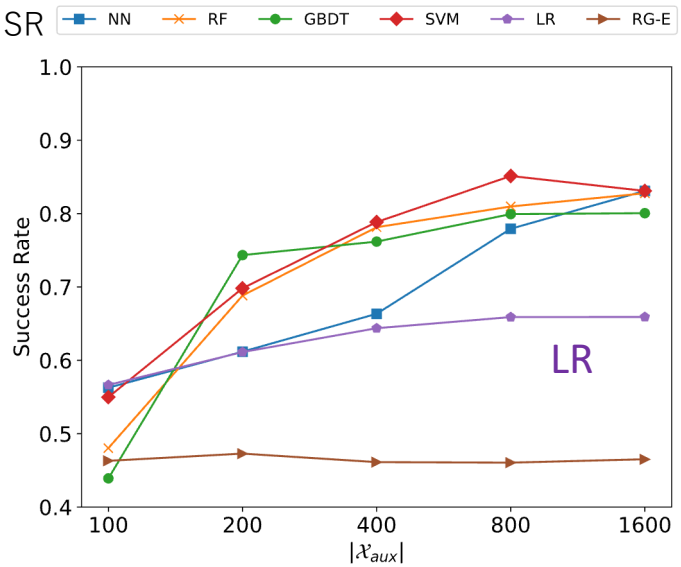


Result 3: Reconstruction Risk of LIME (RQ3)

- The risk of LIME increased as $|\mathcal{X}_{aux}|$ increased as the risk of Shapley values



Shapley values

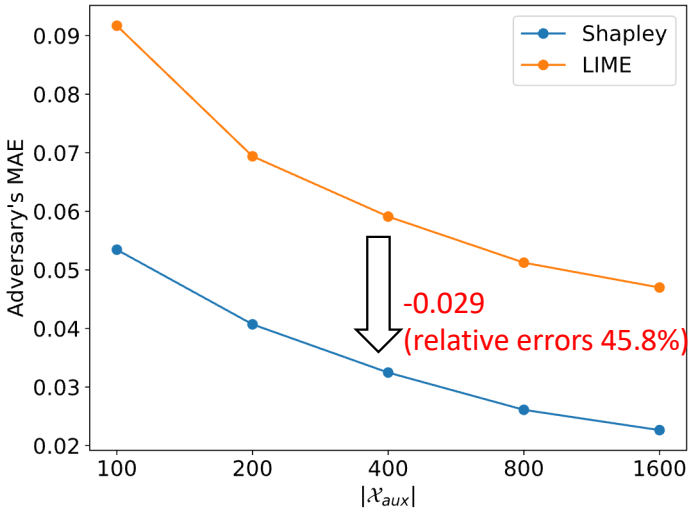


LIME

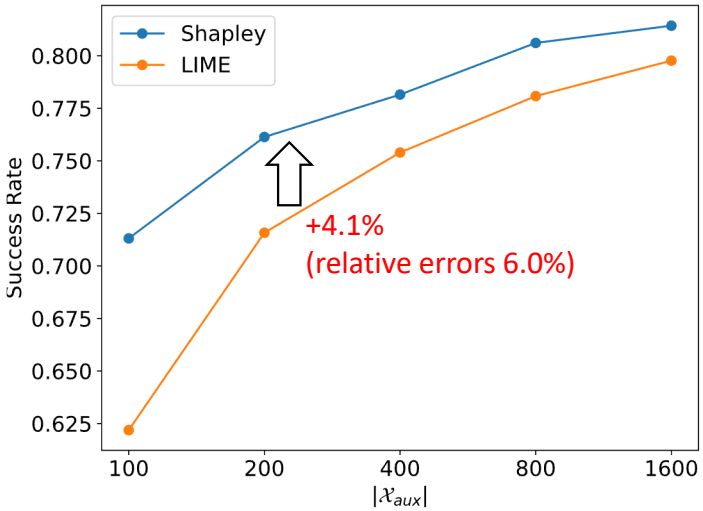
Result 4： Shapley values and LIME (RQ4)

- The risk of Shapley values is clearly higher than that of LIME

Adversary's MAE

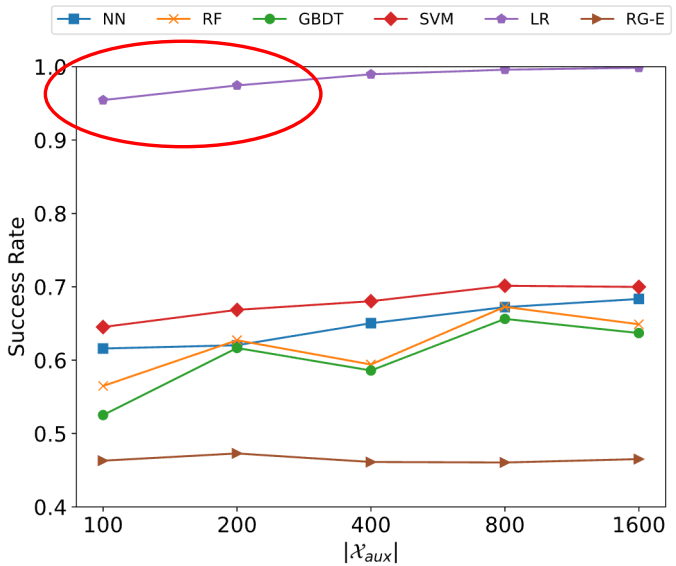


SR



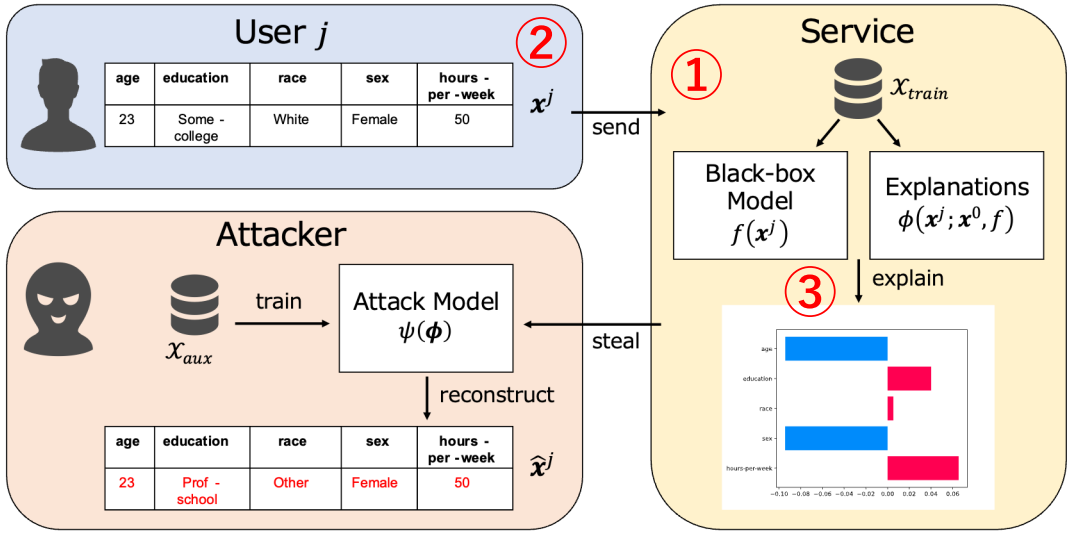
Differences between the Theorem and the Experiments

- Our proposition says the attacker can reconstruct correctly if the model f is linear
- When $|\mathcal{X}_{aux}|$ is lower, the Success Rate are NOT always exactly 1.0
- This error was caused due to the lack of a sufficient number of instances



Mitigation

1. Access control
   - Limitation of the number of requests
2. Privacy-enhancing technologies
   - Synthetic data
   - Differential privacy
3. Processing XAI values
   - Quantization
   - Masking



Conclusions

1. Which machine learning models are vulnerable?
   - ➢ The linear model causes more higher vulnerability than the other models

2. Is the explanation by LIME as risky as Shapley values in the record reconstruction?
   - ➢ LIME has the similar record reconstruction risk to Shapley values
   - ➢ The combination of linear model and LIME is not so vulnerable.

3. Which one is more vulnerable, Shapley values or LIME?
   - ➢ The explanation by Shapley values is more vulnerable