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

Ryotaro Toma[1] and Hiroaki Kikuchi[1][0000−0002−0903−8430]

Graduate School of Advanced Mathematical Sciences, Meiji University
4–21–1 Nakano, Tokyo 164–8525, Japan
cs242022@meiji.ac.jp
kikn@meiji.ac.jp

Abstract. Explainable AI (XAI) metrics have gained attention because of a need to ensure fairness and transparency in machine learning models by providing users with some understanding of the models' internal processes. Many services, including Amazon Web Services, the Google Cloud Platform, and Microsoft Azure run machine-learning-as-a-service platforms, which provide several indices, including Shapley values, that explain the relationship between the output of the black-box model and its private input features. However, in 2022, it was demonstrated that a Shapley-value-based explanation could lead to the reconstruction of private attributes, posing a privacy risk of information leakage from the model. It was shown that the leaked value would depend on the AI black-box model used. However, it was not clear which combinations of black-box model and XAI metric would be vulnerable to a reconstruction attack. The present study shows, both theoretically and experimentally, that Shapley values are indeed vulnerable to a reconstruction attack. We prove that Shapley values for a linear model can lead to a perfect reconstruction of records, that is, they can enable an accurate estimation of private values. In addition, we investigate the impact of various optimization algorithms used in attack models on the reconstruction risk.

Keywords: Shapley Values · Explainabilities · XAI · Reconstruction Attack.

1 Introduction

Machine learning (ML) models have recently been used in important use cases, including finance, healthcare, E-commerce, and employment [1–3]. One well-known issue with artificial intelligence (AI) models is a lack of transparency. Many kinds of models, particularly deep neural networks and ensemble models, have complex internal structures regarded as "black boxes," which prevent internal analysis of their operation and therefore how the models arrive at their decisions. For an application to be trustworthy, we need some transparency about the mapping between the input features and the outputs of the model used.

Explainable AI (XAI) technologies are therefore necessary to guarantee the transparency of models and to explain the relationship between their input features and their outputs [1, 4]. XAI helps to build trust by providing AI users

with insights into how systems work and why they have made specific decisions. In addition to gaining transparency in AI models, XAI can mitigate biases in AI models and can address social concerns about their use.

Currently, XAI indices are available for most machine-learning-as-a-service (MLaaS) platforms, which provide ML models with some XAI indices that explain how the input features affect the outputs of the models. In particular, Shapley-value-based XAI methods [15, 16] are state-of-the-art and are offered for the major commercial MLaaS platforms, including Amazon Web Services [5], Google Cloud [18], and Microsoft Azure [6].

Unfortunately, XAI raises serious privacy concerns. Luo et al. [7] showed that private input record values can be inferred from an explanation based on Shapley values. They proposed an algorithm that estimates attribute values using a gradient-descent method with a sampled auxiliary dataset. This enables private records to be reconstructed with a significant probability of success. They demonstrated the feasibility of attacks using six open datasets plus three synthetic datasets and four black-box AI models.

However, reconstruction accuracy depends on the black-box models used, such as support vector machines (SVMs) or decision trees. The risk must also depend on the XAI metric used because there are many XAI technologies, including local interpretable model-agnostic explanations (LIME) [8], Shapley additive explanations (SHAP) [15], and Anchors [17]. Luo et al. investigated only the Shapley-value approach. It remains unclear if particular combinations of black-box models and XAI metrics may be particularly vulnerable to attack. This is the motivation for our investigation of the privacy risks in vulnerable combinations of black-box models and XAI metrics.

In this work, we claim that *a particular combination of black-box AI model and XAI metric is vulnerable in the context of record reconstruction risks*. We prove that a linear regression model using a Shapley-value XAI metric enables an attacker to reconstruct private records completely (no estimation error). Note that this is an inevitable vulnerability, obtained via theoretical analysis, which will hold for any kind of dataset when the exact Shapley values are given. However, in most platforms, an approximation to or a variation of the Shapley values is used because the exact computation of Shapley values suffers from an exponential computation cost with respect to the number of attributes and becomes infeasible for large-scale systems with more than 50 attributes.

To address these additional questions, we conducted experiments to investigate the reconstruction risks with respect to the differences in the algorithms used by potential attackers. We used three open datasets: Adult [9], Bank Marketing [10], and Credit Card Client [11], with Shapley values. We studied not only the black-box AI models used for training the baseline estimation but also the attack models used to estimate the private record values and the optimization algorithms used by the attack models, including SGD, Momentum, RMSProp, and Adam. The experiments demonstrated that our theoretical result holds for common open datasets.

Our work makes the following contributions:

– We proved that the combination of a linear regression method and using Shapley values is vulnerable against a reconstruction attack, in that the attacker can reconstruct private input attributes exactly.
– We evaluated the record reconstruction risk on Shapley values with respect to the various optimization algorithms used in attacker models. Based on our experimental results, we suggest a possible mitigation approach that adjusts the learning rate with the aim of minimizing the estimation risk.

The remainder of the paper is organized as follows. In Section 2, we introduce some background definitions, Shapley values, and some related work. In Section 3, we present the problem statement and the threat model assumed in our study. We show some examples of record reconstruction attacks on a toy example, aiming to provide insight into attack methods. Our main theorem is given in Section 4, where we prove that using Shapley values with a linear regression model allows an attacker to identify the private inputs without error. In Section 5, we describe our experiments using open data with various parameter settings. We discuss the generalization of our methodology and its limitations in Section 6. We also suggest a possible approach to mitigating the effects of reconstruction attacks. We conclude our work in Section 7.

2 Preliminaries

2.1 Shapley Values

Shapley values [12], proposed by Shapley in 1953, are indices that quantify the contributions of each player in cooperative game theory. In this work, let $\boldsymbol{s} = (s_1, \ldots, s_n)$ be the Shapley values representing the local explanation for the output of model $f(\boldsymbol{x})$ for n input features $\boldsymbol{x} = (x_1, \ldots, x_n)$.

Let $N = \{1, 2, \ldots, n\}$ be the index set of features, S be a subset of N, $\boldsymbol{x}^0$ be a reference sample for calculating Shapley values, and $\phi(\boldsymbol{x}; \boldsymbol{x}^0, f) = (s_1, \ldots, s_n)$ be a mapping to calculate Shapley values. Then, the Shapley value s_i is:

$$s_i = \phi_i(\boldsymbol{x}; \boldsymbol{x}^0, f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \left(f(\boldsymbol{x}_{[S \cup \{i\}]}) - f(\boldsymbol{x}_{[S]}) \right), \quad (1)$$

where $\boldsymbol{x}_{[S]} = ((x_{[S]})_1, \ldots, (x_{[S]})_n)$ denotes an input vector corresponding to S and is defined for $i = 1, \ldots, n$ as:

$$(x_{[S]})_i = \begin{cases} x_i & \text{if } i \in S, \\ x_i^0 & \text{otherwise.} \end{cases} \quad (2)$$

For example, given the input vector $\boldsymbol{x} = (\mathbf{2}, \mathbf{5}, \mathbf{1}, \mathbf{3})$, the reference sample $\boldsymbol{x}^0 = (0, 3, 2, 1)$, and the subset $S = \{2, 3\}$, the vector $\boldsymbol{x}_{[S]}$ is $\boldsymbol{x}_{[S]} = (0, \mathbf{5}, \mathbf{1}, 1)$.

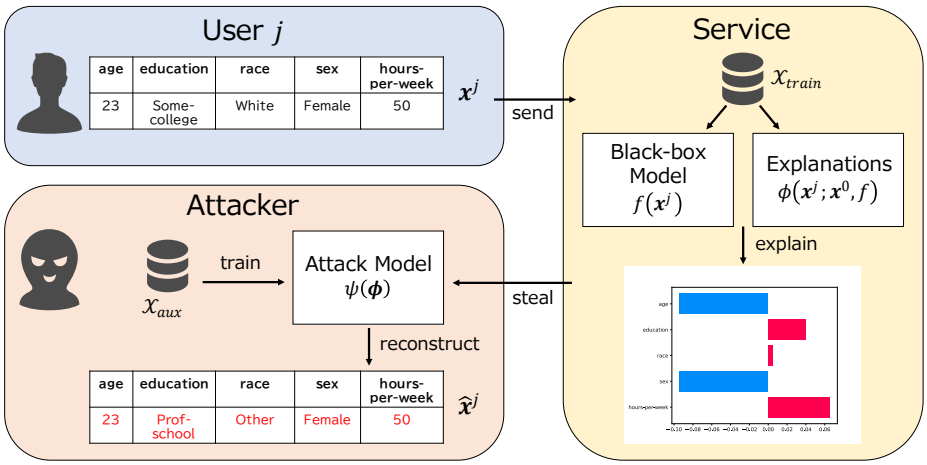


Fig. 1: Overview of the system model

2.2 Feature Inference Attack on Shapley Values

System Model Luo et al. [7] assumed a system model by which the service provider trains their black-box models f based on a private training dataset $\mathcal{X}_{train}$ and uses them for their service provided on MLaaS platforms. Fig. 1 shows the overview of their system model. The **Attacker** has access to the **Service** with its store of explanation for target **User** j. Given the explanation $\phi(\boldsymbol{x}^j; \boldsymbol{x}^0, f)$, the **Attacker** attempts to reconstruct the original record $\boldsymbol{x}^j$ using the attack model $\phi(\boldsymbol{\phi})$.

Threat Model An attacker can send data to the service and receive the Shapley values as an explanation of the data. In addition, the attacker can steal the Shapley values of other users from the service. Under this assumption, the attacker performs the feature inference.

2.3 Related Work

Explainable AI Metrics There are many ways to explain black-box and white-box models. Explainabilities can be partitioned into global and local methods. Global methods explain overall model behavior and compute feature importance values [13, 14], whereas local methods explain the feature importance for each input [8, 15–17]. Major MLaaS platforms including Amazon SageMaker [5], Microsoft Azure [6], and Google Cloud Platform [18] offer SHAP [15], an approximation to Shapley values [12, 16] using local methods such as LIME [8].

Our study therefore focuses on Shapley values [12, 15, 16] as local explainabilities.

Privacy Risk with Explainability Many previous studies have investigated the privacy risk with explainabilities against a variety of attacks, including membership inference [19–21], model extraction [19, 22, 23], feature inference [7, 24], and adversarial attack [19, 25]. In particular, Luo et al. [7] identified a feature inference attack with Shapley values and investigated the privacy risk from the explanations.

Defenses against Attack Some defensive approaches have been proposed, including differential privacy [29–31] and using synthetic data [29, 32, 33]. In addition, several studies have proposed privacy-preserving XAI methods [26–28]. For example, Patel et al. [26] proposed using a local explainability method such as LIME with differential privacy. Nevertheless, the privacy risk in the context of feature inference attacks remains unclear for explainabilities other than using Shapley values.

3 Problem Statement

3.1 Record Reconstruction Attack

We can define a record reconstruction attack (or a feature inference attack) [7] as follows.

Let f and ψ be black-box and attack models, respectively. Let $\mathcal{X}_{train}$, $\mathcal{X}_{aux}$, and $\mathcal{X}_{test}$ be a training dataset for training f, an auxiliary dataset for training ψ, and a test dataset, respectively. Let $\boldsymbol{x}^j = (x_1^j, ..., x_n^j)$ be an input vector for the user $j = 1, \ldots, m$. Let $\boldsymbol{x}^0$ and $\boldsymbol{s}^j = \phi(\boldsymbol{x}^j; \boldsymbol{x}^0, f)$ be the reference sample and the Shapley values of the input vector x^j, respectively. Given the explanation dataset $\mathcal{S}_{aux}$ for all $\boldsymbol{x}_{aux} \in \mathcal{X}_{aux}$ and the black-box model f, the attacker trains the attack model $\psi : \mathcal{S}_{aux} \to \mathcal{X}_{aux}$.

3.2 Evaluation Metrics

We use two metrics for evaluating the record reconstruction risk, the attacker's mean absolute error (MAE) and the success rate (SR).

Attacker's MAE The MAE is the average of a set of absolute errors. The attacker's MAE of the estimated data $\hat{\boldsymbol{x}}$ for the dataset $\boldsymbol{x}$ with m rows and n columns is given as:

$$MAE(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} |\hat{x}_i^j - x_i^j|. \tag{3}$$

Attacker's SR The SR represents the ratio of correctly estimated features to all input features. We say that a feature estimation is a success if the estimated value is identical to the original values for discrete variables. For continuous variables, the estimation is a success if the absolute error is below a particular threshold value. The SR of the estimated data $\hat{\boldsymbol{x}}$ for the dataset $\boldsymbol{x}$ is given as:

$$SR(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \frac{success(\hat{\boldsymbol{x}}, \boldsymbol{x})}{mn}, \tag{4}$$

where $success(\hat{\boldsymbol{x}}, \boldsymbol{x})$ is the number of correctly estimated features.

4 The Record Reconstruction Risk with Linear Regression

In this section, we show the vulnerability of using the combination of a linear model and the Shapley values. First, we provide a simple example to give insight into the vulnerability.

4.1 Example of a Record Reconstruction Attack

Consider an example dataset of 10 rows and 5 columns. We synthesize the dataset $x_1 = n_1$, $x_2 = n_2$, $x_3 = n_1 n_2$, $x_4 = n_2 n_3$, and $y = x_1 - x_3 x_4$ using three independent identically distributed sequences n_1, n_2, and n_3, as shown in Table 1.

Let $\mathcal{X}_{test}$ and $\mathcal{X}_{train}$ comprise rows 1 to 5 and rows 6 to 10 of the dataset, respectively. The Shapley value is the average of the values with respect to each row of $\mathcal{X}_{train}$ in the reference sample, i.e., $\boldsymbol{s} = \frac{1}{5}\sum_{j=6}^{10} \phi(\boldsymbol{x}; \boldsymbol{x}^j, f)$.

Table 2 shows the Shapley values $\mathcal{S}_{test}$ for the input dataset $\mathcal{X}_{test}$ with the linear black-box model f trained for the dataset $\mathcal{X}_{train}$ in the example.

The attacker's MAEs for the model f and the estimation algorithm ψ are then shown in Table 3. f and ψ are either a linear regression or a decision tree. They are implemented via scikit-learn and trained on $\mathcal{X}_{train}$. Note that the input features are correctly estimated without error when f and ψ are both linear models.

4.2 Theoretical Analysis of MAE

According to Eq. (1), the Shapley values are computed as the weighted average of the difference of two outputs $f(\boldsymbol{x}_{[S\cup\{i\}]}) - f(\boldsymbol{x}_{[S]})$. First, we show the linearity of this difference.

Lemma 1. *Let f be a linear black-box model. For any $i \in N$, $S \subseteq N \setminus \{i\}$, and reference sample $(x_1^0, x_2^0, \ldots, x_n^0)$, the following holds:*

$$f(\boldsymbol{x}_{[S\cup\{i\}]}) - f(\boldsymbol{x}_{[S]}) = \beta_i(x_i - x_i^0) \tag{5}$$

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.

This result implies that a record reconstruction could be perfect if the Shapley values from the linear model were to be used.

Table 1: Example dataset

	x_1	x_2	x_3	x_4	y
$\mathcal{X}_{test}$	1.8	0.1	0.3	−0.4	1.9
	0.4	1.5	0.6	1.0	−0.2
	1.0	0.8	0.7	0.7	0.5
	2.2	0.1	0.3	−0.1	2.2
	1.9	0.4	0.8	1.0	1.1
$\mathcal{X}_{train}$ ($\mathcal{X}_{aux}$)	−1.0	0.3	−0.3	−0.5	−1.2
	1.0	1.5	1.4	0.1	0.9
	−0.2	−0.2	0.0	0.0	−0.2
	−0.1	0.3	0.0	0.5	−0.1
	0.4	−0.9	−0.4	−1.3	−0.1

Table 2: Shapley values $s_i \in \mathcal{S}_{test}$ when $\boldsymbol{x}^1$, $\boldsymbol{x}^2$, $\boldsymbol{x}^3$, $\boldsymbol{x}^4$, and $\boldsymbol{x}^5$ represent the reference sample

	s_1	s_2	s_3	s_4
$\boldsymbol{x}^1$	1.30	0.02	0.06	−0.04
$\boldsymbol{x}^2$	0.28	−0.29	0.18	0.34
$\boldsymbol{x}^3$	0.72	−0.13	0.21	0.26
$\boldsymbol{x}^4$	1.59	0.02	0.06	0.04
$\boldsymbol{x}^5$	1.37	−0.04	0.25	0.34

Table 3: The attacker's MAEs for the combination of black-box model f and attack model ψ

Black-box	Attack	Attacker's MAE				
f	ψ	x_1	x_2	x_3	x_4	average
Linear Regression	Linear Regression	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Linear Regression	Decision Tree	0.82	1.24	0.74	1.18	1.00
Decision Tree	Linear Regression	0.69	0.52	0.41	0.53	0.54
Decision Tree	Decision Tree	0.68	1.16	0.82	0.54	0.80
average		0.55	0.73	0.49	0.59	

5 Experiments

5.1 Dataset

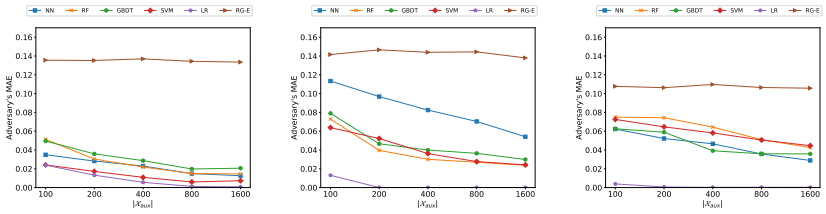
Table 4 shows the dataset used for the experiments. For these settings, we aim to clarify the record reconstruction risk of XAI metrics and the differences between the optimizers used in the attack algorithm.

5.2 Experiment 1: Risk with Respect to Black-box Models and XAI Methods

We evaluated the record reconstruction risk for the explanation of black-box model f given by Shapley values. The black-box model f was one of five models: neural networks (NN), random forest (RF), gradient-boosting decision tree (GBDT), kernel SVM (kSVM), and linear regression (LR).

Table 4: Three open datasets used in the experiments

Dataset	No. of Records	Classes	No. of Features
Adult [9]	48,842	2	14
Bank Marketing [10]	45,211	2	16
Credit Card [11]	30,000	2	24



(a) Adult (b) Bank Marketing (c) Credit Card

Fig. 2: The attacker's MAEs for a record reconstruction attack using Shapley values with respect to the number of rows in the auxiliary dataset $|\mathcal{X}_{aux}|$ and the black-box model f

We implemented NN using PyTorch [34] using an n-dimensional input layer, a c-dimensional output layer, and two hidden layers of $2n$ neurons each. The activation function was softmax for the output layer and rectifier linear unit (ReLU) for the remainder. The other models, RF, SVM, GBDT, and LR, were implemented via scikit-learn. The number of trees and the maximal depth were 100 and 5, respectively, for RF, with those for GBDT being 100 and 3, respectively. We used default values for other parameters if not specifically mentioned. By "RG-E", we denote a random guess prediction from the empirical distribution based on $\mathcal{X}_{aux}$ for comparison purposes.

5.3 Experiment 2: Risk with Respect to Optimization Algorithms

We investigated the record reconstruction risk of the various optimization algorithms that could be used by a potential attacker. An optimizer would be used to update the parameter θ_ψ of the attack model ψ as:

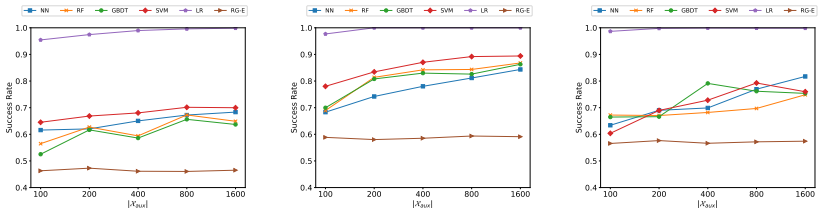
$$\theta_\psi \leftarrow \theta_\psi - \alpha \nabla_{\theta_\psi} \text{loss} \tag{6}$$

In this experiment, we investigated four types of optimization algorithms including SGD [35], Momentum [36], RMSProp [37], and Adam [38]. Using the settings in Luo et al. [7], we examined the attack model ψ deployed in neural networks with $4n$ neurons in the hidden layer and n neurons in the output layer for n features, with the sigmoid function being used for all layers. We implemented model ψ using PyTorch, using default values for all parameters except $\eta = 0.01$ for the learning rates of SGD and Momentum and $momentum = 0.9$ for Momentum. (Momentum refers to an SGD for which $momentum \neq 0$).

5.4 Results

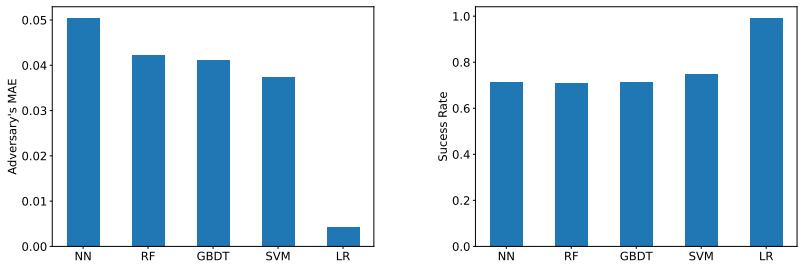
Figs. 2 and 3 show the record reconstruction risk for Shapley values with respect to the number of rows of the auxiliary dataset $\mathcal{X}_{aux}$.

We found that the attacker's MAE decreases and the SR increases as the number of rows $|\mathcal{X}_{aux}|$ increases. Note that the record reconstruction risks for



(a) Adult (b) Bank Marketing (c) Credit Card

Fig. 3: The attacker's SRs for a record reconstruction attack using Shapley values with respect to the number of rows in the auxiliary dataset $|\mathcal{X}_{aux}|$ and the black-box model f



(a) The attacker's MAE (b) The attacker's SR

Fig. 4: The evaluation averages for each model f

Shapley values are small but not zero when $|\mathcal{X}|$ is small for (a) Adult and (b) Bank scenarios. These results appear to be in conflict with Proposition 1, which states that there should be no errors for a linear black-box model. However, these small errors may remain because the conditions in Proposition 1 do not hold here. Another possible reason is that the Shapley values were not calculated exactly as defined but approximated [15, 16]. Note also that the small MAE increases in proportion to the size of the auxiliary dataset.

Fig. 4 shows the average of the attacker's MAE and SR for each model f and dataset size $|\mathcal{X}_{aux}|$ for an attack based on Shapley values. The linear model had the highest record reconstruction risk in terms of both the attacker's MAE and SR.

Fig. 5 shows the distributions of the attacker's MAE and SR for the various optimization algorithms used in an attack. In common with the attacker's MAE and SR, the accuracy of record reconstruction using SGD is the least and, with RMSProp, it is is greatest. For all optimization algorithms, the attacker's MAE decreased and SR increased with $|\mathcal{X}_{aux}|$. Adam and RMSProp tended to increase the record reconstruction risk, whereas SGD and Momentum tended to decrease it. It is known that Adam and RMSProp are methods that tune the learning

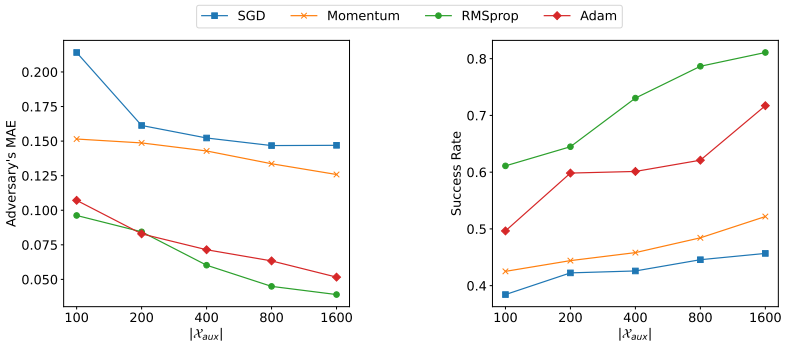


Fig. 5: The attacker's MAE and SR with respect to the size of dataset $|\mathcal{X}_{aux}|$ for the four types of optimizer available to an attacker

rate when training the model. Therefore, it was tuning the learning rate that improved the accuracy of the record reconstruction.

6 Discussion

6.1 Differences between the Theorem and the Experiments

In Proposition 1, we proved that an attacker can reconstruct the original input features correctly when both the black-box model f and the attack model ψ are linear. However, our experimental results showed small errors remaining. We consider that the lack of a sufficient number of instances in some smaller datasets may be the source of the error. The premise of the proposition would then not hold and the linear regression would fail to estimate the exact values. We also note that using Shapley values involves approximations and a reconstruction attack using XAI values might therefore include few mistakes. We plan to explore these conjectures in future work.

6.2 Linearity of the Black-box Models

The proof of Proposition 1 was based on the assumption of linearity in the black-box model f. The reconstruction risk could therefore be nonzero if the black-box model f is not linear. However, we believe that there is a positive correlation between the reconstruction risk and the additivity of the explanations. In 2017, Lundberg et al. [15] proposed a class of additive feature attribution methods that included methods based on Shapley values. There are several methods belonging to the class, except for Shapley values, but a property called "local accuracy" is not fully satisfied in these cases. We therefore believe that the property of local accuracy may be a key feature of the vulnerability when using Shapley values.

6.3 Effects of Encoding Methods for Qualitative Variables

In our experiments, we encoded the qualitative variables into discrete values using one-hot encoding. This could lead to a dimensionality issue. It is possible that, as the number of features increases and the accuracy of the approximations in Shapley values decreases, this could result in a higher reconstruction risk than the original risk. We are also concerned that an attacker could gain more information from the explanations because the dimensionality of the explanation vector would be higher than that for the original. Again, we plan to conduct further experiments using other encoding methods.

6.4 Optimization for Black-box Models

In our experiments, we investigated the privacy risk with respect to optimization algorithms used in the attack model. It is possible that the optimizer in the black-box model affects the reconstruction risk when the black-box model is a neural network. We might consider that the privacy risk would be higher if the accuracy of the black-box model were increased.

6.5 Mitigation

To decrease the record reconstruction risk, we suggest three defensive methods.

First, we could use one of several privacy-enhancing technologies, such as using synthetic data to train the black-box model or using differential privacy for the explanation values. In 2022, Patel et al. [26] proposed explainability via differential privacy. Using a privacy-preserving method should reduce the reconstruction risk.

Second, the access control on MLaaS platforms could be made more efficient. For example, a limitation on the number of requests could reduce the amount of information available to a potential attacker.

Finally, the quantization and masking of Shapley values could be useful ways of decreasing the record reconstruction risk.

7 Conclusion

We have examined the record reconstruction risk of XAI methods based on Shapley values using the attack algorithm of Luo et al. [7]. We also found that learning-rate tuning in the optimization algorithms used by an attacker increases the privacy risk, particularly for Adam and RMSProp optimizers. Using Shapley values can enable the exact reconstruction of private inputs when both black-box and attack models are linear.

To mitigate these risks, we recommend using synthetic data for training black-box models and applying differential privacy to explanation values. Limiting requests and access on MLaaS platforms could also enhance privacy protection.

In future work, we aim to explore the reconstruction risk with explanation vectors that use additive noise and develop new XAI methods to further reduce privacy risks.

Acknowledgments. Part of this work was supported by JSPS KAKENHI Grant Number 23K11110 and JST, CREST Grant Number JPMJCR21M1, Japan.

References

1. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019)
2. Zest AI Insights, `https://www.zest.ai/insights/why-zaml-makes-your-ml-platform-better`, last accessed 2024/04/19
3. Sakai, A., Komatsu, M., Komatsu, R., Matsuoka, R., Yasutomi, S., Dozen, A., Shozu, K., Arakaki, T., Machino, H., Asada, K., Kaneko, S., Sekizawa, A., Hamamoto, R.: Medical Professional Enhancement Using Explainable Artificial Intelligence in Fetal Cardiac Ultrasound Screening. Biomedicines 2022 **10**(3), 551 (2022)
4. Chen, J., Song, L., Wainwright, M., Jordan, M.: Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. In: 35th International Conference on Machine Learning, pp. 882–891. PMLR 80, Stockholm, Sweden (2018)
5. Amazon SageMaker Documentation. `https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-model-explainability.html`, last accessed 2024/04/19
6. Azure Machine Learning Documentation. `https://learn.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability`, last accessed 2024/04/19
7. Luo, X., Jiang, Y., Xiao, X.: Feature Inference Attack on Shapley Values. In: 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22), pp. 2233–2247. Association for Computing Machinery, Los Angeles, CA, USA (2022)
8. Ribeiro, M., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), pp. 1135–1144. Association for Computing Machinery, San Francisco, California, USA (2016)
9. Becker, B., Kohavi, R.: Adult. UCI Machine Learning Repository. (1996). https://doi.org/10.24432/C5XW20
10. Moro, S., Cortez, P., Rita, P.: A data-driven approach to predict the success of bank telemarketing. Decision Support Systems **62**, 22–31 (2014)
11. Yeh, I., Lien, C.: The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert systems with applications **36**(2), 2473–2480 (2009)
12. Shapley, L.: 17. A Value for n-Person Games. Contributions to the Theory of Games (AM-28) **II**, 307–318 (1953)
13. Covert, I., Lundberg, S., Lee, S.: Understanding Global Feature Contributions With Additive Importance Measures. In: 34th International Conference on Neural Information Processing Systems (NIPS '20), pp. 17212–17223. Curran Associates Inc., Vancouver, BC, Canada (2020)

14. Fisher, A., Rudin, C., Dominici, F.: All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. Journal of Machine Learning Research **20**(177), 1–81 (2019)
15. Lundberg, S. Lee, S.: A Unified Approach to Interpreting Model Predictions. In: 31st International Conference on Neural Information Processing Systems (NIPS'17), pp. 4768–4777. Curran Associates Inc., Long Beach, California, USA (2017)
16. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. **41**, 647–665 (2016). https://doi.org/10.1007/s10115-013-0679-x
17. Ribeiro, M., Singh, S., Guestrin, C.: Anchors: High-Precision Model-Agnostic Explanations. In: Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18), pp. 1527–1535. AAAI Press, New Orleans, Louisiana, USA (2018)
18. Introduction to Vertex Explainable AI, https://cloud.google.com/vertex-ai/docs/explainable-ai/overview, last accessed 2024/04/19
19. Kuppa, A., Le-Khac, N.: Adversarial XAI Methods in Cybersecurity. IEEE Transactions on Information Forensics and Security **16**, 4924–4938 (2021)
20. Shokri, R., Strobel, M., Zick, Y.: On the Privacy Risks of Model Explanations. In: 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), pp. 231–241. Association for Computing Machinery, Virtual Event, USA (2021)
21. Liu, H., Wu, Y., Yu, Z., Zhang, N.: Please Tell Me More: Privacy Impact of Explainability through the Lens of Membership Inference Attack. In: 2024 IEEE Symposium on Security and Privacy (SP), pp. 119–138. IEEE Computer Society, San Francisco, CA, USA (2024)
22. Yan, A., Hou, R., Liu, X., Yan, H., Huang, T., Wang, X.: Towards explainable model extraction attacks. International Journal of Intelligent Systems **37**(11), 9936–9956 (2022)
23. Yan, A., Huang, T., Ke, L., Liu, X., Chen, Q., Dong, C.: Explanation leaks: Explanation-guided model extraction attacks. Information Sciences: an International Journal **632**(C), 269–284 (2023)
24. Fredrikson, M., Jha, S., Ristenpart, T.: Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In: 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS '15), 1322–1333. Association for Computing Machinery, Denver, Colorado, USA (2015)
25. Baniecki, H., Biecek, P.: Adversarial attacks and defenses in explainable artificial intelligence: A survey. Information Fusion **107**, 102303 (2024)
26. Patel, N., Shokri, R., Zick, Y.: Model Explanations with Differential Privacy. In: 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), pp. 1895–1904. Association for Computing Machinery, Seoul, Republic of Korea (2022)
27. Nguyen, T., Lai, P., Phan, H., Thai, M.: XRand: Differentially Private Defense against Explanation-Guided Attacks. Proceedings of the AAAI Conference on Artificial Intelligence **37**(10), 11873–11881 (2023)
28. Bozorgpanah, A., Torra, V., Aliahmadipour, L.: Privacy and Explainability: The Effects of Data Protection on Shapley Values. Technologies **10**(6), 125 (2022)
29. Shlomo, N.: Integrating Differential Privacy in the Statistical Disclosure Control Tool-Kit for Synthetic Data Production. In: Domingo-Ferrer, J., Muralidhar, K. (eds) Privacy in Statistical Databases (PSD 2020), LNCS, vol. 12276, pp. 271–280. Springer, Cham. (2020). https://doi.org/10.1007/978-3-030-57521-2_19

30. Wang, G., Gehrke, J. Xiao, X.: Differential Privacy via Wavelet Transforms. IEEE Transactions on Knowledge & Data Engineering **23**(8), 1200–1214 (2011)
31. Ito, S., Miura, T., Akatsuka, H., Terada, M.: Differential Privacy and Its Applicability for Official Statistics in Japan - A Comparative Study Using Small Area Data from the Japanese Population Census. In: Domingo-Ferrer, J., Muralidhar, K. (eds) Privacy in Statistical Databases (PSD 2020), LNCS, vol. 12276, pp. 337–352. Springer, Cham. (2020). https://doi.org/10.1007/978-3-030-57521-2_24
32. Slokom, M., Wolf, P., Larson, M.: When Machine Learning Models Leak: An Exploration of Synthetic Training Data. In: Domingo-Ferrer, J., Laurent, M. (eds.) Privacy in Statistical Databases (PSD 2022), LNCS, vol. 13463, pp. 283–296. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-13945-1_20
33. Tritscher, J., Ring, M., Schlr, D., Hettinger, L., Hotho, A.: Evaluation of Post-hoc XAI Approaches Through Synthetic Tabular Data. In: Helic, D., Leitner, G., Stettinger, M., Felfernig, A., Raś, Z.W. (eds) Foundations of Intelligent Systems (ISMIS 2020), LNCS, vol 12117, pp. 422–430. Springer, Cham. (2020). https://doi.org/10.1007/978-3-030-59491-6_40
34. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: an imperative style, high-performance deep learning library. In: 33rd International Conference on Neural Information Processing Systems (NeurIPS 2019), pp. 8026–8037. Curran Associates Inc., Vancouver, Canada (2019)
35. Bottou, L.: On-line Learning and Stochastic Approximations. On-Line Learning in Neural Networks, 9–42 (1999)
36. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: 30th international conference on machine learning (ICML-13), pp. 1139–1147. JMLR.org, Atlanta, GA, USA (2013)
37. Hinton, G.: Coursera Neural Networks for Machine Learning Lecture 6. `https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf`, last accessed 2024/04/19
38. Kingma, D., Ba, J.: Adam: A Method for Stochastic Optimization. In: International Conference on Learning Representations (ICLR) 2015. (2015)

A Algorithm in Luo et al. [7]

In the model investigated by Luo et al. [7], a user j sends a private input vector $\boldsymbol{x}^j$ to a service and receives the output of the model $f(\boldsymbol{x}^j)$ and explanations about n features $\boldsymbol{s} = \phi(\boldsymbol{x}^j; \boldsymbol{x}^0, f) = (s_1, ..., s_n)$. The attacker has an auxiliary dataset $\mathcal{X}_{aux}$, which is distributed similarly to the training dataset $\mathcal{X}_{train}$. The attacker sends $\boldsymbol{x}_{aux} \in \mathcal{X}_{aux}$ to the model f and receives the explanation dataset $\mathcal{S}_{aux} = \{\phi(\boldsymbol{x}_{aux}; \boldsymbol{x}^0, f) | \boldsymbol{x}_{aux} \in \mathcal{X}_{aux}\}$. The attacker then trains the attack model $\psi : \mathcal{S}_{aux} \rightarrow \mathcal{X}_{aux}$ to minimize the loss $L(\psi(\mathcal{S}_{aux}), \mathcal{X}_{aux})$. Finally, the attacker estimates the original private input features $\boldsymbol{x}^j$ as $\hat{\boldsymbol{x}}^j = \psi(\boldsymbol{s})$ with the given Shapley values $\boldsymbol{s}$. This is described formally in Algorithm 1.

Algorithm 1 Estimating algorithm using the auxiliary dataset [7]

Input: Black-box model f, auxiliary dataset $\mathcal{X}_{aux}$, learning rate α, attacked Shapley vector $\boldsymbol{s}$
Output: Estimated private input features $\hat{\boldsymbol{x}}$

$\mathcal{S}_{aux} \leftarrow \phi(\mathcal{X}_{aux}; f)$
$\theta_\psi \leftarrow \mathcal{N}(0, 1)$
for each epoch **do**
  for each batch **do**
    $loss \leftarrow 0$
    $B \leftarrow$ randomly select a batch of samples
    for $j \in 1, \ldots, |B|$ **do**
      $(\hat{\boldsymbol{x}}_{aux})^j \leftarrow \psi((\boldsymbol{s}_{aux})^j; \theta_\psi)$
      $loss \leftarrow loss + L((\hat{\boldsymbol{x}}_{aux})^j, (\boldsymbol{x}_{aux})^j)$
    end for
    ${\theta_\psi}' \leftarrow \theta_\psi - \alpha \nabla_{\theta_\psi} loss$
  end for
end for
$\hat{\boldsymbol{x}} \leftarrow \psi(\boldsymbol{s}; \theta_\psi)$
return $\hat{\boldsymbol{x}}$

B Proofs

B.1 Proof of Lemma 1

Proof. Denoting the model f as $f(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$, for any subset S and element i,

$$\begin{aligned} f(\boldsymbol{x}_{[S\cup\{i\}]}) - f(\boldsymbol{x}_{[S]}) &= \beta_0 + \sum_{k\in S\cup\{i\}} \beta_k x_k + \sum_{k\in N\setminus(S\cup\{i\})} \beta_k x_k^0 \\ &\quad - \beta_0 + \sum_{k\in S} \beta_k x_k + \sum_{k\in N\setminus S} \beta_k x_k^0 \\ &= \beta_i (x_i - x_i^0). \end{aligned}$$

B.2 Proof of Proposition 1

Proof. From Lemma 1, the i-th Shapley value s_i is calculated as follows:

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

where $\lambda_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} \beta_i$. Therefore, the attack model ψ is given as the linear equations of $x_1, \ldots, x_n$ as follows:

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

where $\gamma_i = \alpha_i \lambda_i$ and $\gamma_0 = \alpha_0 - \sum_{k=1}^{n} \alpha_k \lambda_k x_k^0$. Note that this is a linear polynomial of $x_0, \ldots, x_n$. If $\mathcal{X}_{aux}$ is large enough and the number of rows exceeds $n+1$, the coefficients $\gamma_1, \ldots, \gamma_n$ are correctly estimated by the least squares method. A linear regression will give the exact input variables, proving the proposition.