Black-box Model Inversion Attribute Inference Attacks on Classification Models

Shagufta Mehnaz
Dartmouth College
United States
shagufta.mehnaz@dartmouth.edu

Ninghui Li
Purdue University
United States
ninghui@purdue.edu

Elisa Bertino
Purdue University
United States
bertino@purdue.edu

ABSTRACT

Increasing use of ML technologies in privacy-sensitive domains such as medical diagnoses, lifestyle predictions, and business decisions highlights the need to better understand if these ML technologies are introducing leakages of sensitive and proprietary training data. In this paper, we focus on one kind of model inversion attacks, where the adversary knows non-sensitive attributes about instances in the training data and aims to infer the value of a sensitive attribute unknown to the adversary, using oracle access to the target classification model. We devise two novel model inversion attribute inference attacks– confidence modeling-based attack and confidence score-based attack, and also extend our attack to the case where some of the other (non-sensitive) attributes are unknown to the adversary. Furthermore, while previous work uses accuracy as the metric to evaluate the effectiveness of attribute inference attacks, we find that accuracy is not informative when the sensitive attribute distribution is unbalanced. We identify two metrics that are better for evaluating attribute inference attacks, namely G-mean and Matthews correlation coefficient (MCC). We evaluate our attacks on two types of machine learning models, decision tree and deep neural network, trained with two real datasets. Experimental results show that our newly proposed attacks significantly outperform the state-of-the-art attacks. Moreover, we empirically show that specific groups in the training dataset (grouped by attributes, e.g., gender, race) could be more vulnerable to model inversion attacks. We also demonstrate that our attacks' performances are not impacted significantly when some of the other (non-sensitive) attributes are also unknown to the adversary.

1 INTRODUCTION

Across numerous sectors, a variety of institutions are streamlining their processes by adopting machine learning (ML) technologies and leveraging commercial ML-as-a-service APIs. In many cases, these ML technologies are trained on proprietary and sensitive datasets, e.g., in the domains of personalized medicine [1–4], product recommendation [5–7], finance and law [8–10], social media [11–13], etc. Moreover, with the increasing use of ML technologies in personal data, we have seen a recent surge of serious privacy concerns that were previously ignored. Therefore, it is very important to also understand whether public access to such trained models introduces new attack vectors against the privacy of these proprietary and sensitive datasets used for training ML models, such as lifestyle surveys, genetic data, purchase history of sensitive items, etc.

A *model inversion attack* is the one that turns the one-way journey from training data to model into a two-way one, i.e., this attack allows an adversary to infer part of the training data even when given only oracle access to the target classification model. This attack can take the most catastrophic form when the datasets used to train such machine learning models are privacy sensitive or proprietary. Recently, Fredrikson et al. [14, 15] proposed two formulations of model inversion attacks. In the first one, which we call **model inversion attribute inference (MIAI)** attack, the adversary aims to learn some sensitive attribute of an individual whose data are used to train the target model, and whose other attributes are known to the adversary. This can be applied, e.g., when each instance gives information of one individual. In the second formulation, which we call **typical instance reconstruction (TIR)** attack, the adversary is given oracle access to a classification model and a class, and aims to come up with a typical instance for that class. For example, the adversary, when given access to a model that recognizes different users' faces, tries to reconstruct an image that is similar to a target individual's actual facial image.

Several recent works study TIR attacks [16–18]. Note that for TIR attacks to be considered successful, it is not necessary for a reconstructed instance to be quantitatively close to any specific training instance. Evaluation is typically done by having humans assessing similarity of the reconstructed instances (e.g., reconstructed facial images) to training instances. Thus a model that is able to learn the essence of each class and generalizes well (as opposed to relying on remembering information specific to training instances) will likely remain vulnerable to such an attack. Indeed, it is proven [18] that a model's predictive power and its vulnerability to such TIR attacks are two sides of the same coin. This is because highly predictive models are able to establish a strong correlation between features and labels and this is the property that an adversary exploits to mount the TIR attacks [18]. In other words, the existence of TIR attacks is a feature of good classification models, although the feature may be undesirable in some settings. We point out that such is not the case for MIAI attacks, which is evaluated by the ability to predict exact attribute values of individual instances.

In this paper, we focus only on MIAI attacks on classification models where data about individuals are used. More specifically, we consider the attribute inference attacks where the adversary leverages black-box access to an ML model to infer the sensitive attributes of a target individual. While attribute inference in other contexts have been studied extensively in the privacy literature, there exists little work studying to what extent model inversion introduces new AI vulnerabilities. In the rest of the paper, we refer to MIAI attacks whenever we use the term model inversion attack.

Proposed new model inversion attacks: We design two new black-box model inversion attacks: (1) confidence modeling-based attack, and (2) confidence score-based attack. These attacks are different in terms of the adversary's capability assumptions, and

therefore, represent a variety of adversaries that can pose different levels of threats to training data privacy. We define a threat model to clearly state the adversary assumptions not only for our proposed attacks but also for the existing attacks in the literature. Our confidence score-based attack assumes an adversary who can only access the predicted label and the confidence scores returned by the model, whereas the adversary assumed in our confidence modeling-based attack has access to a dataset collected from the same population the target model training dataset has been obtained from. However, there is no overlap between these two datasets. While all the existing attacks [14, 15] assume that the adversary has full knowledge of other non-sensitive attributes of the target individual, it is not clear how the adversary would perform in a setting where it has only partial knowledge of those attributes. To understand the vulnerability of model inversion attacks in such practical scenarios, we also propose an attack that works even when some non-sensitive attributes are unknown to the adversary. Moreover, we also investigate if there are scenarios when model inversion attacks do not threaten the privacy of the overall dataset but are effective on some specific groups of instances (e.g., individuals grouped by race, gender, education level, etc.). We empirically show that there exists such discrimination across different groups of the training dataset where a group is more vulnerable than the others.

While the existing MIAI attacks [14] have been evaluated only on decision tree models, we evaluate our attacks also on *deep neural network* models. Therefore, we train two models– a *decision tree* and a *deep neural network* with each of the two real datasets in our experiments, General Social Survey (GSS) [19] and Adult dataset [20], to evaluate our proposed attacks.

Effective evaluation of model inversion attacks: To understand if model inversion attacks pose a broader risk, it is important that we use appropriate metrics to evaluate our proposed attacks as well as the existing attacks. Although the Fredrikson et al. attack [14] primarily uses accuracy, in this paper, we argue that accuracy is not the best measure. This is because simply predicting the majority class for all the instances can achieve very high accuracy which certainly misrepresents the performances of model inversion attacks. Moreover, we argue that the F1 score, a widely used metric, is also not sufficient by itself since it emphasizes only the positive class, and simply predicting the positive class for all the instances can achieve significant F1 score. Hence, we propose to use G-mean [21] and Matthews correlation coefficient (MCC) [22] as metrics to design a framework that can effectively evaluate any model inversion attack.

Understanding model inversion attacks in-depth: We also propose to compare the performances of various model inversion attacks with those from attacks that do not query the target model, e.g., randomly guessing the sensitive attribute according to some distribution. Such random guessing attacks do not even access the model and thus certainly cannot invert the model. When a particular model inversion attack deployed against a target model performs similarly to such attacks, we can conclude that the target model in not vulnerable to that particular model inversion attack. Hence, in this paper, we address the following general research question- is it possible to identify when a model should be classified as vulnerable to such model inversion attacks? *More specifically, does black-box access to a particular model really help the adversary to estimate the sensitive attributes in the training dataset which is otherwise impossible for the adversary to estimate (i.e., without access to the black-box model)?*

Summary of contributions: In summary, this paper makes the following contributions:

(1) We define the various capabilities of the adversary and provide a detailed threat model. Based on the threat model we design two new black-box model inversion attacks: (1) confidence modeling-based attack and (2) confidence score-based attack. We propose to use the G-mean and Matthews correlation coefficient (MCC) metrics along with accuracy and F1 score to compare our proposed attacks with existing [14] and other baseline attacks.
(2) We conduct extensive evaluation of our attacks using two types of ML models, decision tree and deep neural network, trained with two real datasets. Evaluation results show that our proposed attacks significantly outperform the existing attacks.
(3) We also empirically show that a particular subset of the training dataset (grouped by attributes, such as, gender, education level, etc.) could be more vulnerable than others to the model inversion attacks.
(4) We evaluate the risks incurred by model inversion attacks when the adversary does not have knowledge of all other non-sensitive attributes of the target individual and demonstrate that our attack's performance is not impacted significantly in those circumstances.

2 PROBLEM DEFINITION AND EXISTING ATTACK STRATEGIES

2.1 Model Inversion Attribute Inference Attack

An ML model can be represented using a deterministic function $f : \mathcal{R}^d \longrightarrow \mathcal{Y}$. The input of this function is a d-dimensional vector $\mathbf{x} = [x_1, x_2, ..., x_d] \in \mathcal{R}^d$ that represents d attributes and $y' \in \mathcal{Y}$ is the output. In the case of a regression problem, $\mathcal{Y} = \mathcal{R}$. However, in this work, we focus on classification problems. Therefore, more specifically, $f : \mathcal{R}^d \longrightarrow \mathcal{R}^m$ where m is the number of unique class labels $(y_1, y_2, ..., y_m)$ and $\mathcal{R}^m$ represents the confidence scores returned by the classification model for these m class labels. Finally, the class label with the highest confidence score is considered as the output of the prediction model. We denote the dataset on which the f model is trained as DS_T. From now on, we use the term y to represent the actual value in the training dataset DS_T whereas y' is used to represent the model output $f(\mathbf{x})$. The values of y and y' are the same in the case of a correct prediction or different in the case of an incorrect prediction by f.

Now, some of the attributes in $\mathbf{x}$ introduced above could be privacy sensitive. Without loss of generality, let's assume that $x_1 \in \mathbf{x}$ is a sensitive attribute that the individual corresponding to a data record in the training dataset does not want to reveal to the public. However, a model inversion attack may allow an adversary to infer this x_1 attribute value of a target individual given some specific capabilities, such as, access to the black-box model (i.e., target model), background knowledge about the target individual, etc.

Table 1: Assumption of adversary capabilities/knowledge for different attack strategies.

Attack strategy	Predicted label	Confidence score along with predicted label	Target individuals' all other (non-sensitive) attributes	All possible values of the sensitive attribute	Marginal prior of the sensitive attribute	Marginal prior of all other (non-sensitive) attributes	Confusion matrix of the model	Attacker dataset DS_A
Naive attack				✓	✓			
Random guessing attack				✓	✓(optional)			
Fredrikson et al. Attack [14]	✓		✓	✓	✓	✓	✓	
Confidence modeling-based attack	✓	✓	✓	✓				✓
Confidence score-based attack	✓	✓	✓	✓				

2.2 Threat Model

The adversary is assumed to have all or a subset of the following capabilities/knowledge (see Table 1):

- Access to the black-box target model, i.e., the adversary can query the model with $\mathbf{x} = [x_1, x_2, ..., x_d]$ and obtain a class label y' as the output.
- The confidence scores returned by the target model for m class labels, i.e., $\mathcal{R}^m$.
- Full/partial knowledge of the non-sensitive attributes of the target individual except his/her sensitive attribute.
- All possible (k) values of the sensitive attribute x_1.
- Knowledge of marginal prior of the sensitive attribute x_1, i.e., $\mathbf{p_1} = \{p_{1,1}, p_{1,2}, ..., p_{1,k}\}$ where k is the number of all possible values of x_1 and $p_{1,k}$ is the probability of the $k-$th unique possible value.
- Knowledge of confusion matrix (C) of the model where $C[y, y'] = Pr[f(x) = y' | y$ is the true label].
- Access to a dataset collected from the same population the target model's training dataset has been obtained from. However, there is no overlap between the target model's training dataset (DS_T) and the dataset that the adversary has access to (DS_A).

Note that, for the attacks designed in this paper, the adversary does not need the knowledge of marginal priors of any attributes (sensitive or non-sensitive) or the confusion matrix. Also, we only consider a passive adversary that does not aim to corrupt the machine learning model or influence its output in any way.

2.3 Baseline Attack Strategies

2.3.1 Naive Attack. A naive model inversion attack assumes that the adversary has knowledge about the probability distribution (i.e., marginal prior) of the sensitive attribute and always predicts the sensitive attribute to be the value with the highest marginal prior. Therefore, this attack does not require access to the target model. Note that this attack can still achieve significant accuracy if the sensitive attribute is highly unbalanced, e.g., if the sensitive attribute can take only two values and there is an 80%-20% probability distribution, predicting the value with higher probability would result in 80% accuracy.

2.3.2 Random Guessing Attack. The adversary in this attack also does not require access to the target model. The adversary randomly predicts the sensitive attribute by setting a probability for each possible value. The adversary may or may not have access to the marginal priors of the sensitive attribute. Fig. 5(a) in Appendix A.1 shows the optimal performance of random guessing attack in terms of different metrics when the adversary sets different probabilities for predicting the positive class sensitive attribute independent of its knowledge of marginal prior 0.3. Note that, predicting the positive class for all the instances with this attack (i.e., setting a probability 1 for the positive class) would result in a significantly high F1 score, mainly due to a recall of 100% (Fig. 5(a) in Appendix).

2.4 Fredrikson et al. Attack [14]

The Fredrikson et al. [14] black-box model inversion attack assumes that the adversary can obtain the model's predicted label, has knowledge of all the attributes of a targeted individual (including the true y value) except the sensitive attribute, has access to the marginal priors of all the attributes, and also to the confusion matrix of the target model. The adversary can query the target model multiple times by varying the sensitive attribute (x_1) and obtain the predicted y' values. After querying the model k times with k different x_1 values ($x_{1,0}, x_{1,1}, \ldots, x_{1,k-1}$) while keeping the other known attributes unchanged, the adversary computes $C[y, y'] * p_{1,i}$ for each possible sensitive attribute value, where

$$C[y, y'] = Pr[f(x) = y' |\ y \textit{ is the true label}]$$

and $p_{1,i}$ is the marginal prior of the i-th possible sensitive attribute value. Finally, the attack predicts the sensitive attribute value for which the computed $C[y, y'] * p_{1,i}$ value is the maximum.

3 METRICS FOR EVALUATING THE VULNERABILITY OF A MODEL TO INVERSION ATTACKS

Though the impact of model inversion attacks can be overwhelming, in this section, we aim to take a deep dive to understand if it is possible to determine when a model should be classified as vulnerable and if the metrics considered in the existing model inversion attack research are sufficient. More specifically, we investigate the following general research question- *does black-box access to a particular model really help the adversary to estimate the sensitive attributes in the training dataset which is otherwise impossible for the adversary to estimate (i.e., without access to that black-box model)?*

Understanding a model's vulnerability to inversion attacks requires a meaningful metric to evaluate and compare different model inversion attacks. The Fredrikson et al. attack [14] primarily uses accuracy. However, if we care only about accuracy, the naive attack of simply guessing the majority class for all the instances can achieve very high accuracy. Another widely used metric is the F1 score. However, the F1 score of the positive class emphasizes only on that specific class and thus, as a one-sided evaluation, cannot be considered as the only metric to evaluate the attacks. Otherwise, always guessing the positive class may achieve similar or even better F1 score (mainly due to a recall of 100%) than any sophisticated

model inversions attack that identifies the positive class instances strategically.

To understand whether access to the black-box model considerably contributes to attack performance and also to compare the baseline attack strategies (that do not require access to the model, i.e., naive attack and random guessing attack) to our proposed attacks, we use the following two metrics in addition to accuracy and F1 scores: G-mean [21] and Matthews correlation coefficient (MCC) [22], as described below.

G-mean: G-mean is the geometric mean of sensitivity and specificity [21]. Thus it takes all of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) into account. With this metric, the random guessing attack can achieve a maximum performance of 50%. Note that, even if the adversary has knowledge of marginal priors of the sensitive attribute, it is not able to achieve a G-mean value of more than 50% by setting different probabilities for predicting the positive class sensitive attribute (Fig. 5(a) in Appendix). For random guessing attack, the optimal G-mean value can be achieved by setting the probability to 0.5. The G-mean for the naive attack is always 0%.

$$\text{G} - \text{mean} = \sqrt{\frac{\text{TP}}{\text{TP} + \text{FN}} * \frac{\text{TN}}{\text{TN} + \text{FP}}} \tag{1}$$

Matthews correlation coefficient (MCC): This MCC metric also takes into account all of TP, TN, FP, and FN, and is a balanced measure which can be used even if the classes of the sensitive attribute are of very different sizes [22]. It returns a value between -1 and +1. A coefficient of +1 represents a perfect prediction, 0 represents a prediction no better than the random one, and -1 represents a prediction that is always incorrect. Note that, even if the adversary has the knowledge of marginal priors of the sensitive attribute, it is not able to achieve an MCC value of more than 0 with the random guessing attack strategy (details in Appendix A.1). Also, the naive attack always results in an MCC of 0, independent of the marginal prior knowledge.

$$\text{MCC} = \frac{(\text{TP} * \text{TN}) - (\text{FP} * \text{FN})}{\sqrt{(\text{TP} + \text{FP}) * (\text{TP} + \text{FN}) * (\text{TN} + \text{FP}) * (\text{TN} + \text{FN})}} \tag{2}$$

4 NEW MODEL INVERSION ATTACKS

We design two new attack strategies: (1) confidence modeling-based model inversion attack and (2) confidence score-based model inversion attack. Table 1 shows the different adversary capabilities/knowledge assumptions for these attacks in contrast to the existing attacks.

4.1 Confidence Modeling-based Model Inversion Attack (CMMIA)

In confidence modeling-based model inversion attack, the adversary models the predictions and confidence values returned by the target model f to *predict the sensitive attribute value*. We assume that the adversary has access to a dataset, DS_A, collected from the same population the DS_T dataset has been obtained, where $DS_A \cap DS_T = \emptyset$. The adversary first queries the target model f using records in DS_A dataset and collects predictions and confidence scores returned by the target model to train the attack models. The attack steps are described in the following.

4.1.1 Collecting data for training attack models. The adversary first queries the target model f using the records in dataset DS_A (see Fig. 1). Note that, the adversary has knowledge of the actual sensitive attributes (x_1) in the DS_A dataset along with non-sensitive attributes, $x_2, ...x_d$ and actual y. The goal of this attack step is to collect data on how the target model responds to the queries performed with the DS_A dataset, i.e., the predictions and confidence scores returned by the target model, when the adversary varies the sensitive attribute x_1. The number of queries performed for this data collection (attack models' training data) step is: $n_A * k$, where n_A is the number of records in DS_A and k is the number of unique possible values of the sensitive attribute x_1. For instance, if there are two possible values of a sensitive attribute (i.e., $k = 2$, well depicted by a yes/no answer from an individual in response to a survey question), the adversary queries the model by setting the sensitive attribute value x_1 to both *yes* and *no* while all other known input attributes of the target individual remain the same. Let y'_0 and $conf_0$ be the returned model prediction and confidence score when the sensitive attribute is set to *no*. Similarly, y'_1 and $conf_1$ are the model prediction and confidence score when the sensitive attribute is set to *yes*. After querying the target model, there are three possible outcomes:

Case (1) The target model f predicts the correct y *only for a single sensitive attribute value*, e.g., $y = y'_0 \wedge y \mathrel{!}= y'_1$ or $y \mathrel{!}= y'_0 \wedge y = y'_1$, in the case of a binary sensitive attribute. The adversary collects the records $[\{y'_0, conf_0\}, ..., \{y'_{k-1}, conf_{k-1}\}, x_1]$, i.e., the predictions with k possible sensitive attribute values and the associated confidence scores, to train the case 1 attack models (see Fig. 1).

Case (2) The target model f predicts the correct y for multiple sensitive attribute values, i.e., multiple y's equal y. The adversary collects the records $[\forall i \{y'_i, conf_i\}$ *s.t.* $y'_i = y, x_1]$ to train the case 2 attack models.

Case (3) The target model f predicts incorrect y's for all possible sensitive attribute values. The adversary collects the records $[\{y'_0, conf_0\}, ..., \{y'_{k-1}, conf_{k-1}\}, x_1]$ to train the case 3 attack models, where for any i in $\{0, ..., k-1\}$, $y'_i \neq y$.

4.1.2 Training the attack models. The adversary trains m attack models for each of the 3 cases mentioned above where m is the number of target model class labels (mentioned in Section 2.1). The adversary first separates the records under each case into m subsets based on the original target class label (y) of the records in DS_A. The reason behind training separate attack models for each target model class label is that the target model's performance (returned predictions and confidence scores) can be different for different class labels which may be due to various reasons, e.g., underlying unbalanced training dataset. For example, a vehicle image classification model may perform very good in identifying images of trucks but demonstrate poor performance in identifying images of vans.

In summary, we train $3 * m$ attack models that take as input the predictions and confidence scores returned by the target model and outputs a prediction for the sensitive attribute:

$$\mathcal{A}^q_l : [\{y'_0, conf_0\}, ..., \{y'_{k-1}, conf_{k-1}\}] \longrightarrow x_1 \tag{3}$$

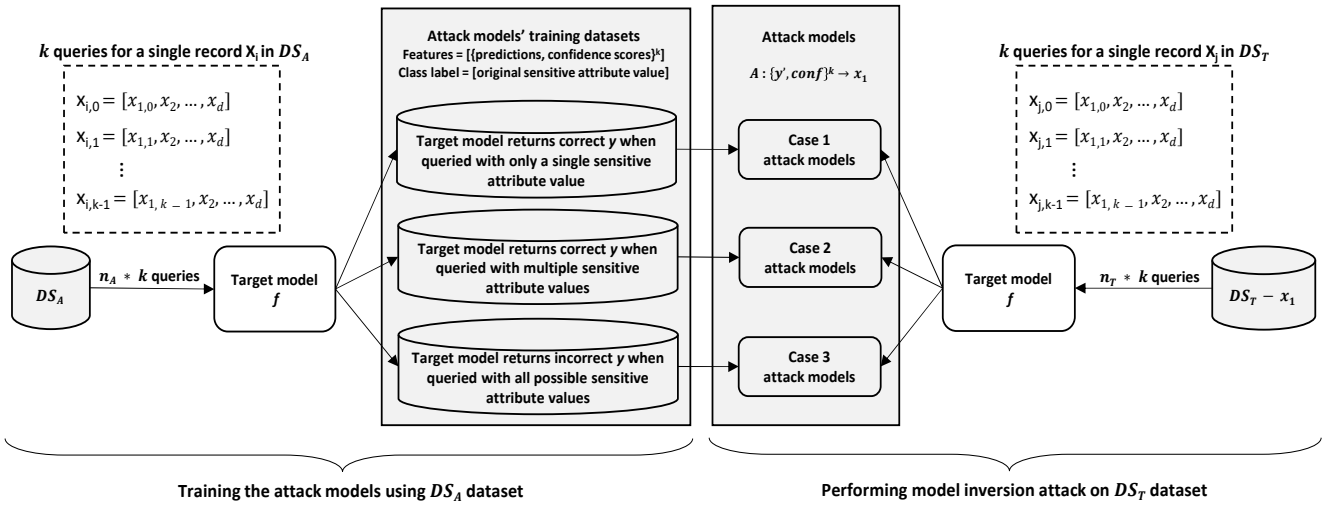


Figure 1: Confidence modeling-based model inversion attack (CMMIA). First, the adversary collects the attack models' training datasets by querying the target model f using DS_A dataset to train the attack models. The adversary then leverages the trained attack models to predict the sensitive attribute values of the records in DS_T dataset.

where $1 \leq l \leq m$ represents original class label index and $1 \leq q \leq 3$ represents case number. For instance, $\mathcal{A}_2^3$ represents the attack model that is trained to predict the sensitive attribute x_1 for a record when the original class label of that record is y_2 and the target model returns incorrect predictions for all k queries, i.e., for any i in $\{0, ..., k-1\}$, $y_i' \neq y_2$. Note that, in case (2), the input of attack models can have less features since the adversary collects the target model's right predictions only, $[\forall i\{y_i', conf_i\}\ s.t.\ y_i' = y, x_1]$, as described in Section 4.1.1.

4.1.3 Performing model inversion attack on DS_T. The adversary now performs $n_T * k$ queries to the target model f, where n_T is the number of records in DS_T and k is the number of unique possible values of the sensitive attribute x_1. Unlike DS_A, the adversary does not know the x_1 attribute of the records in DS_T and the goal here is to estimate this attribute.

When querying the target model f with DS_T dataset, the adversary classifies the outcomes into three cases in the same way as described in Section 4.1.1. Whereas in Section 4.1.1 the adversary collects data for training the attack models, here the adversary leverages the trained attack models to estimate the sensitive attribute x_1. For instance, to estimate the sensitive attribute x_1 of a record with *original class label* y_2, the adversary first queries the target model f with varying sensitive attribute values and obtains the corresponding predictions and confidence score pairs $\{y_0', conf_0\}$, ..., $\{y_{k-1}', conf_{k-1}\}$. The adversary then inputs these values to an attack model based on the outcomes of the target model. If the outcomes of the target model fall in–

- case (1), the adversary queries attack model $\mathcal{A}_2^1$.
- case (2), the adversary queries attack model $\mathcal{A}_2^2$.
- case (3), the adversary queries attack model $\mathcal{A}_2^3$.

Finally, the attack model outputs a prediction for the sensitive attribute x_1.

4.2 Confidence Score-based Model Inversion Attack (CSMIA)

We propose another attack that exploits the confidence scores returned by the model but in this attack the adversary does not have access to the dataset DS_A. Therefore, this attack is suitable for an adversary that is not able to obtain such a dataset. Unlike Fredrikson et al. [14] attack, the adversary assumed in this attack does not have access to the marginal priors or the confusion matrix. Similar to the steps described in Section 4.1, the adversary queries the model by setting the sensitive attribute value x_1 to all possible k values while all other known input attributes of the target individual remain the same. *The key idea of this attack is that the target model's returned prediction is more likely to be correct and the confidence score is more likely to be higher when it is queried with a record containing the real sensitive attribute value (since the target model encountered the record in the training dataset with the real sensitive attribute value).* In order to determine the value of x_1, this attack considers the following cases:

Case (1) If the target model predicts the correct y *only for a single sensitive attribute value*, e.g., $y = y_0' \wedge y != y_1'$ or $y != y_0' \wedge y = y_1'$, in the case of a binary sensitive attribute, the attack selects the sensitive attribute to be the one for which the prediction y' matches y. For instance, if $y = y_1' \wedge y != y_0'$, the attack predicts *yes* for the sensitive attribute and vice versa.

Case (2) If the model predicts the correct y for multiple sensitive attribute values, i.e., $y = y_0' \wedge y = y_1'$, the attack selects the sensitive attribute to be the one for which the prediction confidence score is the maximum. In the above example, if the model predicts the y value with a higher confidence when *yes* value is set for the sensitive attribute, the attack outputs the *yes* value for the x_1 prediction and vice versa.

Case (3) If the model outputs incorrect predictions for all possible sensitive attribute values, i.e., y != $y'_0 \wedge y$!= y'_1, the attack selects the sensitive attribute to be the one for which the prediction confidence is the minimum. In the above example, if the model outputs the incorrect prediction with a higher confidence when *yes* value is set for the sensitive attribute, the attack outputs the *no* value for the x_1 prediction and vice versa.

5 ATTACK WITH PARTIAL KNOWLEDGE OF TARGET INDIVIDUAL'S NON-SENSITIVE ATTRIBUTES

All of our proposed attacks mentioned in Section 4 as well as the Fredrikson et al. [14] attack assume that the adversary has full knowledge of the target individual's non-sensitive attributes. Although these attacks raise serious privacy concerns against a model trained on sensitive dataset, it is not clear how much risk is incurred by these model inversion attacks if the adversary has only partial access to the other (non-sensitive) attributes. In many cases, it may be difficult or even impossible for an adversary to obtain all of the non-sensitive attributes of a target individual. Therefore, the goal of this section is to quantify the risk of model inversion attacks in the case where all non-sensitive attributes of a target individual are not available to the adversary.

In the following, we describe our *confidence score-based model inversion attack* for this special case. Therefore, this model inversion attack with partial knowledge of target individual's non-sensitive attributes does not require the adversary to have access to the dataset DS_A.

For simplicity, we assume that there is only one non-sensitive attribute that is unknown to the adversary. Extending our attack steps to more than one unknown attribute is straightforward. In Section 6.7, we conduct experiments when a single as well as multiple non-sensitive attributes are unknown to the adversary. Without loss of generality, let $x_2 \in \mathbf{x}$ be the non-sensitive attribute unknown to the adversary. Also, let u be the number of unique possible values of x_2. We query the model by varying the unknown non-sensitive attribute with its different unique possible values (in the same way we vary the sensitive attribute x_1 in the attacks described in Section 4) while all other known non-sensitive attributes $\{x_3, ..., x_d\}$ remain the same. Hence, in this attack, we query the model u times for each possible value of the sensitive attribute. As a result, the complexity of the attacks described in this section is u times the complexity of the attacks in Section 4.

According to the notations used in Section 4, let C_0=$\sum_{i=1}^{u}(y = y'_{0\_i})$ be the number of times the predictions are correct with the sensitive attribute *no* and C_1=$\sum_{i=1}^{u}(y = y'_{1\_i})$ be the number of times the predictions are correct with the sensitive attribute *yes*.

In order to determine the value of x_1, this attack considers the following cases:

Case (1) If C_0 != C_1, i.e., the number of correct target model predictions are different for different sensitive attribute values, the attack selects the sensitive attribute to be the one for which the number of correct predictions is higher. For instance, if $C_1 > C_0$, the attack predicts *yes* for the sensitive attribute and vice versa.

Table 2: Distribution of sensitive attributes in datasets.

Dataset	Sensitive attribute	Positive class label	Negative class label	Positive class count	Positive class %
GSS	Watched x-rated movie	Yes	No	4002 (3017)	19.70% (19.80%)
Adult	Marital status	Married	Single	21639 (16893)	47.85% (47.96%)

Case (2) If $C_0 = C_1$ and both are non-zero, we compute the sum of the confidence scores (only for the correct predictions) for each sensitive attribute and the attack selects the sensitive attribute to be the one for which the sum of the confidence scores is the maximum.

Case (3) If $C_0 = 0 \wedge C_1 = 0$, we compute the sum of the confidence scores for each sensitive attribute and the attack selects the sensitive attribute to be the one for which the sum of the confidence scores is the minimum.

If there is a second non-sensitive attribute that is unknown to the adversary (let that unknown attribute be x_3) and v is the number of unique possible values for that unknown non-sensitive attribute, we query the model by varying both x_2 and x_3 while all other known non-sensitive attributes $\{x_4, ..., x_d\}$ remain the same. Hence, in this attack, we query the model $u * v$ times for each possible value of the sensitive attribute. As a result, the complexity of the attack becomes $u * v$ times the complexity of the attacks in Section 4.

6 EVALUATIONS

In this section, we discuss our experiment setup (i.e., datasets, machine learning models, and performance metrics) and evaluate our proposed attacks. *To facilitate reproducibility, we will publicly release the code-base, all the datasets, and machine learning models used in the following experiments.*

6.1 Datasets

General Social Survey (GSS) [19]: Fredrikson et al. [14] attack, denoted as *FJR* attack in the rest of this paper, uses the *General Social Survey (GSS)* dataset to demonstrate their attack effectiveness. This dataset has 51,020 records with 11 attributes and is used to train a model that predicts how happy an individual is in his/her marriage. However, the training dataset for this model contains sensitive attribute about the individuals: e.g., responses to the question '*Have you watched X-rated movies in the last year?*'. Removing the data records that do not have either the sensitive attribute or the attribute that is being predicted by the target model (i.e., happiness in marriage) results in 20,314 records that we use in our experiments. For CMMIA, DS_A consists of randomly chosen 5,079 records and the rest 15,235 records belong to DS_T (i.e., a 25%-75% split). To ensure consistency, we evaluate our CSMIA attack and other baseline attack strategies including FJR attack [14] on the target models trained on the DS_T dataset (15,235 records). Among the 15,235 records in the DS_T dataset, 3,017 individuals answered *yes* (i.e., sensitive attribute x_1 = *yes*) to the survey question on whether they watched X-rated movies in the last year and the rest 12218 individuals answered *no* (i.e., x_1 = *no*).

Adult Dataset [20]: This dataset, also known as *Census Income* dataset, is used to predict whether an individual earns over $50K a

year. The number of instances in this dataset is 48, 842 and it has 14 attributes. We merge the 'marital status' attribute into two distinct clusters, Married: {Married-civ-spouse, Married-spouse-absent, Married-AF-spouse} and Single: {Divorced, Never-married, Separated, Widowed}. We then consider this attribute (Married/Single) as the sensitive attribute that the adversary aims to learn. After removing the data records with missing values, the final dataset consists of 45222 records. For CMMIA, DS_A consists of randomly chosen 10, 000 records and the rest 35, 222 records belong to DS_T. To ensure consistency, we evaluate all the model inversion attacks in comparison (proposed and baseline) against the target models trained on the DS_T dataset (35, 222 records). Among the 35222 records, 16893 individuals are *married* (i.e., sensitive attribute x_1 = *married*) and the rest 18329 individuals are *single* (i.e., x_1 = *single*). For the experiments in this paper, we have removed the '*relationship*' attribute from this dataset since this attribute's values (e.g., husband, wife, unmarried) are directly related to the *marital status* attribute that the model inversion attacks aim to learn.

6.2 Machine Learning Models

In order to evaluate our proposed attacks and other existing attacks, we train target models on each of the two datasets mentioned in Section 6.1 using two different machine learning techniques, decision tree and deep neural network. The confusion matrices of all the trained models are given in Appendix (Tables 9, 10, 11, and 12). We leverage BigML [23], an ML-as-a-service system, to train these target models. Users can leverage such a service by uploading their datasets, selecting an attribute as the objective field, and training a model to predict that objective field when the other attributes are given as input. BigML allows users to publish their models in black-box mode, i.e., the models can be queried by other users and they can obtain the model predictions along with the confidence scores. CMMIA strategy's attack models are also trained using BigML's decision tree machine learning technique.

6.3 Attack Performance Metrics

As mentioned earlier, the *accuracy* metric may fail to evaluate an attack or even misrepresent the attack performance if the dataset is unbalanced. Table 2 shows the distribution of sensitive attribute values in the GSS and Adult datasets (positive class count and % in DS_T are shown in parenthesis). Since the sensitive attribute in the GSS dataset is unbalanced, a naive attack always predicting the negative class would result in $\sim$ 80% accuracy, which is a misleading evaluation of attack performance. Moreover, the *F1 score* alone is not a meaningful metric to evaluate the attacks since it emphasizes only on the positive class. Therefore, we also use *G-mean* and *MCC* metrics as described in Section 3 to evaluate our attacks as well as to compare their performances with that of the FJR attack [14] and the baseline attacks (naive and random guessing).

6.4 New Model Inversion Attacks' Results

6.4.1 GSS Dataset. Tables 3 and 4 show the performance of the baseline attacks and our proposed attacks against the decision tree and deep neural network target models trained on the GSS dataset, respectively. As shown in the results, the FJR attack [14] achieves a very low recall and thus low F1 score. This is due to the fact that the FJR attack [14] relies on the marginal prior of the sensitive attribute while performing the attack (described in Section 2.4). Since the sensitive attribute in the GSS dataset is unbalanced, the FJR attack [14] mostly predicts the negative sensitive attribute (i.e., the individual didn't watch x-rated movie, marginal prior $\sim$ 0.8) and rarely predicts the positive sensitive attribute (i.e., the individual watched x-rated movie, marginal prior $\sim$ 0.2). The FJR attack [14] against the deep neural network target model performs almost like a naive attack with only 1 true positive and 5 false positives as demonstrated in Table 4. In contrast, our proposed CMMIA and CSMIA strategies achieve significantly high recall, F1 score, G-mean, and MCC while also improving precision. The FJR attack [14] performs better only in terms of accuracy. However, note that the naive attack also achieves an accuracy of 80.2%, the highest among all attacks, but there is no attack efficacy (0 true positive).

In CMMIA, for each of the three cases, we train three decision tree models, one for each possible y value of the target model (i.e., happiness in marriage– 'not too happy', 'pretty happy', and 'very happy'), i.e., nine decision tree models in total. In most of the experiments, the CMMIA strategy performs better than CSMIA since it leverages the attack models. Note that, attack models take the predictions and confidence scores returned by the target model when queried with varying sensitive attributes as input and outputs a prediction for the sensitive attribute. As a result, for instance, in case (2), the CMMIA strategy learns that even if the target model returns higher confidence for the negative sensitive attribute (i.e., the individual didn't watch x-rated movie), the actual sensitive attribute value could be positive (i.e., the individual watched x-rated movie) and thus reports more true positives (see case (2) results in Tables 13 and 14 in Appendix that show the contrast between the confidence modeling-based and confidence score-based attacks in details).

Note that, all our attacks perform consistently across different machine learning models. In contrast, the FJR attack [14] shows notably different performance in terms of identifying the positive cases against the deepnet target model.

6.4.2 Adult Dataset. Tables 5 and 6 show the performance of the baseline attacks and our proposed attacks against the decision tree and deep neural network target models trained on the Adult dataset, respectively. While our CMMIA strategy results in the highest recall, the CSMIA attack performs better in terms of precision. For the CMMIA strategy, for each of the three cases, we train two decision tree models, one for each possible y value of the target model (i.e., income– '<=50K' and '>50K'), i.e., six decision tree models in total. Tables 15 and 16 in Appendix show the contrast between CMMIA and CSMIA in details.

Overall, the attacks against the target models trained on Adult dataset demonstrate more effectiveness than that of against the target models trained on GSS dataset. However, the correlations between the sensitive attributes and the corresponding target models trained on these datasets (in other words, importance of the sensitive attributes in the target models) do not differ significantly. For instance, the importance of the 'x-rated-movie' and 'marital-status' sensitive attributes in their corresponding decision tree target models are 7.3% and 9.6%, respectively. Fig. 7 in Appendix shows the importance of all attributes in these models.

Table 3: Attack performance against the decision tree target model trained on GSS dataset.

Attack Strategy	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC
Naive attack	0	12218	0	3017	0%	0%	80.2%	0%	0%	0%
FJR attack [14]	131	11709	509	2886	20.47%	4.34%	77.72%	7.16%	20.39%	0.3%
Confidence modeling-based attack	1766	7605	4610	1254	27.7%	58.48%	61.51%	37.59%	60.34%	16.8%
Confidence score-based attack	1490	7844	4373	1528	25.41%	49.37%	61.27%	33.55%	56.3%	11.1%

Table 4: Attack performance against the deep neural network target model trained on GSS dataset.

Attack Strategy	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC
FJR attack [14]	1	12213	5	3016	16.67%	0.03%	80.17%	0.07%	1.82%	−0.2%
Confidence modeling-based attack	1100	8133	4085	1917	21.22%	36.46%	60.61%	26.82%	49.26%	2.2%
Confidence score-based attack	1212	8058	4160	1805	22.56%	40.17%	60.85%	28.89%	51.47%	5.1%

Table 5: Attack performance against the decision tree target model trained on Adult dataset.

Attack Strategy	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC
Naive attack	0	18329	0	16893	0%	0%	52.04%	0%	0%	0%
FJR attack [14]	3788	17818	511	13105	88.11%	22.42%	61.34%	35.75%	46.69%	29.9%
Confidence modeling-based attack	12311	11619	6710	4582	64.72%	72.88%	67.94%	68.56%	67.97%	36.4%
Confidence score-based attack	7664	17085	1244	9229	86.04%	45.37%	70.27%	59.41%	65.03%	44.3%

Table 6: Attack performance against the deep neural network target model trained on Adult dataset.

Attack Strategy	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC
FJR attack [14]	3592	17717	612	13301	85.44%	21.26%	60.5%	34.05%	45.34%	27.62%
Confidence modeling-based attack	11907	11328	7001	4986	62.97%	70.48%	65.97%	66.52%	66.01%	32.35%
Confidence score-based attack	7490	17139	1190	9403	86.29%	44.34%	69.93%	58.58%	64.39%	43.87%

6.5 Understanding Model Inversion Attacks In-depth

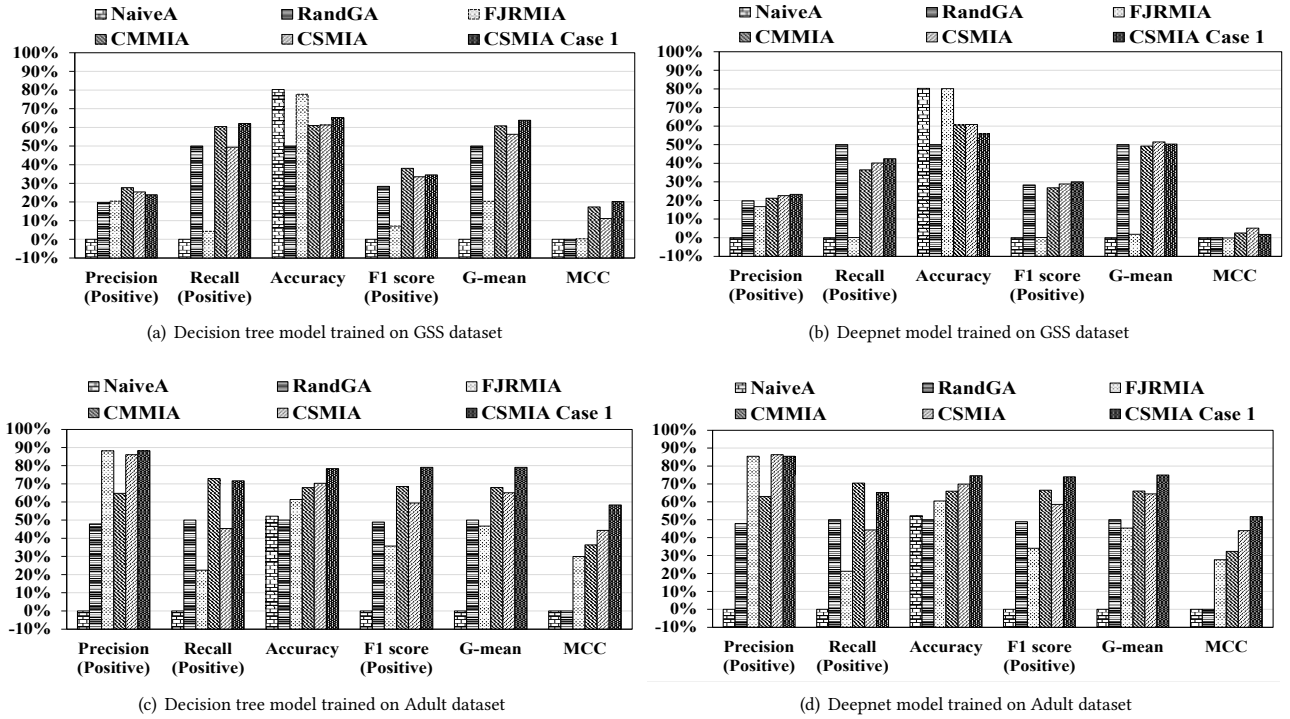
As described in Section 3, the goal of the following experiments is to understand whether releasing the black-box model really adds more advantage to the adversary to learn the sensitive attributes in the training dataset. Therefore, we compare all model inversion attacks (both existing and our proposed ones), i.e., Fredrikson et al. attack (FJRMIA [14]), confidence-modeling based attack (CMMIA), and confidence score-based attack (CSMIA) with baseline attack strategies that do not require access to the target model, i.e., naive attack (NaiveA) and random guessing attack (RandGA). Note that, the case (1) of CSMIA does not even require the knowledge of confidence scores. The adversary can perform the case (1) CSMIA attack with the knowledge of the predicted labels (y) only. Therefore, we pay special attention to the case (1) of CSMIA and analyze its performance along with the other attacks.

In random guessing attack (RandGA), always predicting the positive class would result in a 100% recall and thus a high F1 score but a G-mean of 0%. Therefore, for all the comparisons in the following, the RandGA strategy predicts the positive class with a probability of 0.5, thus maximizing G-mean at 50% and also ensuring a recall of 50%. Figures 5(b) and 5(c) in Appendix show the optimal performance of random guessing attack on GSS and Adult datasets, respectively.

6.5.1 GSS Dataset. Since the sensitive attribute in this dataset has an unbalanced distribution, the NaiveA strategy (described in Section 2.3), also mentioned in [14], predicts the sensitive attribute as *no* for all the individuals and achieves an accuracy of 80.3%. However, the precision, recall, F1 score, G-mean, and MCC would all be 0% as shown in Figures 2(a) and 2(b).

Fig. 2(a) shows how the existing and our proposed model inversion attacks compare with NaiveA and RandGA against the decision tree model trained on the GSS dataset. As demonstrated in the figure, the FJRMIA [14] strategy achieves high accuracy, similar to NaiveA, but does not perform well in terms of any other metrics. Our attacks consistently perform better than RandGA in terms of all metrics. We emphasize that the individuals who belong to the case (1) are more vulnerable to CSMIA.

Fig. 2(b) shows how the existing and our proposed model inversion attacks compare with NaiveA and RandGA against the deep neural network model trained on the GSS dataset. As demonstrated in the figure, the FJRMIA [14] strategy achieves a high accuracy but an extremely low recall. The RandGA strategy has the same results as Fig. 2(a) since this strategy is independent of the machine learning model. Our attacks' performances against this model are not significantly better than RandGA, even the CSMIA case (1) does not show significant results. Therefore, it may seem that according to the overall performance, the deep neural network model trained on the GSS dataset may not be vulnerable to model inversion attacks since another adversary even without access to the model may achieve comparable performances with RandGA. *However, it is very important to note that the RandGA strategy predicts the sensitive attribute randomly whereas the model inversion attacks rely on the outputs of a model that is trained on the dataset containing the actual sensitive attributes. Even if the overall performance of a model inversion attack on the entire dataset does not seem to be a threat, some specific groups of individuals in the dataset could still*



(a) Decision tree model trained on GSS dataset

(b) Deepnet model trained on GSS dataset

(c) Decision tree model trained on Adult dataset

(d) Deepnet model trained on Adult dataset

Figure 2: Comparison of model inversion attacks: Fredrikson et al. attack (FJRMIA [14]), confidence-modeling based attack (CMMIA), confidence score-based attack (CSMIA), and case (1) of CSMIA with baseline attack strategies: naive attack (NaiveA) and random guessing attack (RandGA).

be vulnerable. We discuss such discrimination in performances of model inversion attacks later in Section 6.6.

6.5.2 Adult Dataset. Since the sensitive attribute is more balanced in this dataset, the NaiveA strategy has an accuracy of only 52.1%, and the other metrics are at 0%, similar to that of previous results of this attack.

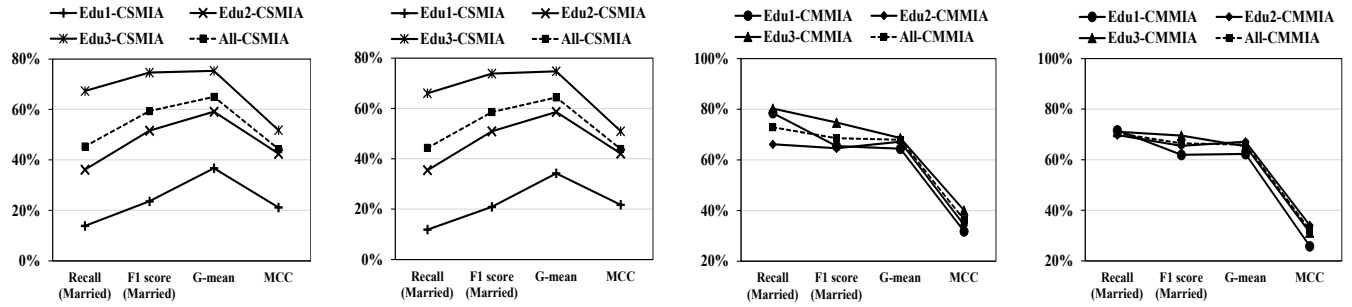
As demonstrated in Fig. 2(c), FJRMIA [14] results in a precision comparable to our attacks but achieves much less in terms of the other metrics. Our attacks significantly outperform RandGA in terms of all metrics except the recall of CSMIA. Fig. 2(d) shows results very similar to that of Fig. 2(c). Observing the results of our proposed attacks and also the CSMIA case (1) performance, we conclude that *releasing the models trained on the Adult dataset* would add significant advantage to the adversary in terms of learning the 'marital status' sensitive attribute. This is because all our proposed attacks that query the models for sensitive attribute inference perform significantly better when compared to the RandGA adversary that does not need access to the model.

6.6 Differences in Vulnerability Among Different Groups

In this section, we further investigate the vulnerability of model inversion attacks by analyzing the attack performances on different groups in the dataset. If a particular group in a dataset is more vulnerable to model inversion attacks than others, it raises serious privacy concerns for that particular group. For instance, we studied the vulnerability of individuals in the Adult dataset by grouping them according to their increasing education levels. We clustered them into three groups, Edu1: {Preschool-12th}, Edu2: {HS-grad, Some-college}, and Edu3: {Assoc-voc, Assoc-acdm, Bachelors, Masters, Prof-school, Doctorate}. The clustered data is then more balanced with number of instances 4397, 19169, and 11656, respectively. The percentages of married individuals in these groups are 43.39%, 45.01%, and 54.55%, respectively. Fig. 3 shows the contrast of attack performances against these three groups.

Figures 3(a) and 3(b) show how the CSMIA attack performs on these three groups of individuals where the target models are decision tree and deep neural network, respectively. As demonstrated in the figures, the Edu3 group is the most vulnerable one with the highest recall, F1 score, G-mean, and MCC, independent of the machine learning algorithm used to train the target model. We also demonstrate the performance of CSMIA on the entire dataset (all of Edu1, Edu2, and Edu3 combined) with dotted lines in Figures 3(a) and 3(b).

Note that the performance of an adversary with RandGA strategy would not differ significantly for these different groups because of their random prediction. Due to the differences in the underlying distributions of the married individuals in these groups, the RandGA strategy would only show slightly different performance in terms of precision and thus in the F1 score. While this result shows only one instance of such inequity in the model inversion attack performances on different groups, this is a potentially serious issue and needs to be further investigated. Otherwise, while it



(a) CSMIA on decision tree model (b) CSMIA on deepnet model (c) CMMIA on decision tree model (d) CMMIA on deepnet model

Figure 3: Differences in attack performances against individuals with different education levels in the Adult dataset. Edu1: {Preschool-12th}, Edu2: {HS-grad, Some-college}, Edu3: {Assoc-voc, Assoc-acdm, Bachelors, Masters, Prof-school, Doctorate}.

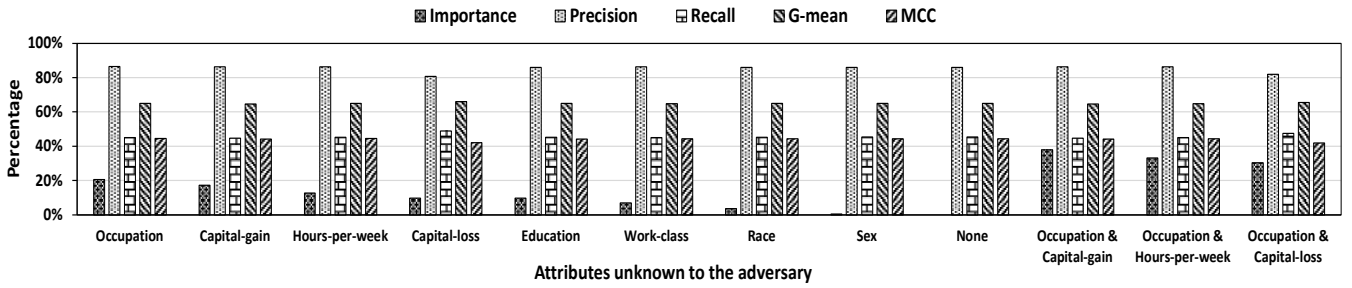


Figure 4: CSMIA performance against the decision tree model trained on Adult dataset when some of the other (non-sensitive) attributes of a target individual are also unknown to the adversary.

may seem that the attack performance on the overall dataset is not a significant threat, some specific groups in the dataset could be significantly more vulnerable to model inversion attacks.

The CMMIA strategy's performance on these three groups of individuals does not vary significantly as shown in Figures 3(c) and 3(d). This confirms that the CMMIA strategy is capable of better handling the sensitivity of the target model in terms of the returned confidence scores when the attributes, both sensitive and non-sensitive, are varied (also discussed in Sections 6.4.1 and 6.8).

6.7 Model Inversion Attack Results With Partial Knowledge of Target Individual's Non-sensitive Attributes

In Section 6.5, we have observed that the models trained on the Adult dataset are notably more vulnerable to model inversion attacks when compared to the models trained on the GSS dataset. Therefore, in this section, we focus on only the Adult dataset to evaluate how our proposed attacks would perform when the adversary has partial knowledge of the target individual's non-sensitive attributes.

As mentioned earlier, we have removed the '*relationship*' attribute from the Adult dataset due to its very high correlation with the *marital status* sensitive attribute. Excluding the sensitive attribute ('marital status') and the output of the target model ('income'), we consider each of the remaining (non-sensitive) attributes to be unknown to the adversary once at a time, i.e., denoting those as x_2. Fig. 4 shows the performance of our confidence score-based attack on the decision tree target model trained on the Adult dataset when some of the non-sensitive attributes are unknown to the adversary. The x-axis shows the non-sensitive attributes that are unknown. The attributes are sorted (from left to right) according to their *importance* in the model, a parameter computed by BigML. We also present the original results (i.e., when *none* of the non-sensitive attributes is unknown to the adversary) to compare how the partial knowledge of the target individual's non-sensitive attributes impacts our attacks' performances. As demonstrated in Fig. 4, we observe that the performance of our attack does not deteriorate and remain almost same when some of the non-sensitive attributes are unknown to the adversary, independent of the importance of the attributes in the target model. We observe only slightly lower precision (and slightly higher recall) when the 'capital-loss' attribute is unknown to the adversary. We also perform experiments where a combination of non-sensitive attributes are unknown to the adversary– 'occupation and capital-gain' (combined importance 37.8%), 'occupation and hours-per-week' (combined importance 33.3%), and 'occupation and capital-loss' (combined importance 30.4%). As demonstrated in Fig. 4, our attack does not show any significant deterioration. *These results not only show an increased vulnerability of model inversion attacks but also escalate the practicability of such attacks in the real world where the adversary may not know all other attributes of a target individual.*

Table 7: Attack performance against the deep neural network target model trained on Adult dataset.

Target model class label	Attack Strategy	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score
<=50K	FJR attack [14]	13	17108	13	9315	50%	00.14%	64.73%	00.28%
	Confidence score-based attack	127	17018	103	9201	55.22%	1.36%	64.82%	2.66%
	Confidence modeling-based attack	5643	11399	5722	3685	49.65%	60.5%	64.43%	54.54%
>50K	FJR attack [14]	3775	710	498	3790	88.34%	49.9%	51.12%	63.78%
	Confidence score-based attack	7537	67	1141	28	86.85%	99.63%	86.68%	92.8%
	Confidence modeling-based attack	6668	220	988	897	88%	88.14%	78.51%	87.62%

Due to space constraints and also due to similarities with the results in Fig. 4, the performance details against the deepnet target model in this setting have been discussed in Appendix A.2.

Table 8: Confusion matrix of decision tree target model trained on Adult dataset.

Actual \ Predicted	<=50K	>50K	Total	Recall
<=50K	24912	1537	26449	94.19%
>50K	3343	5430	8773	61.89%
Total	28255	6967	35222	Avg. recall 78.04%
Precision	88.17%	77.94%	Avg. precision 83.05%	Accuracy 86.15%

6.8 Model Inversion Attacks' Efficacy on Different Class Labels of Target Model

In this section, we aim to understand the efficacy of model inversion attacks for different class labels of the target model and focus on the decision tree target model trained on Adult dataset (Table 7).

Table 8 shows the confusion matrix of this target model. From the confusion matrix, it is evident that the model's performance is better for class label <= $50K$ than that of class label > $50K$.

Table 7 shows a comparison among FJR attack [14], CMMIA, and CSMIA performances for different class labels of the target model. For both FJR attack [14] and CSMIA, the performance of the attacks are significantly different for the two class labels, e.g., the recall of identifying 'married' individuals in class <=50K is very low compared to the recall of identifying 'married' individuals in class >50K. In contrast, the performance of CMMIA strategy in identifying 'married' individuals is more consistent across the two class labels.

The reason behind the low recall of FJR attack [14] for class <=50K is that when the adversary queries the model with records from class <=50K by varying sensitive attribute values, i.e., single and married, most of the time both the queries predict the correct class <=50K. In other words, both the query outcomes fall into the [<=50K, <=50K] cell of the confusion matrix. Therefore, when the FJR attack [14] computes $C[y, y'] * p_{1,i}$ for each possible sensitive attribute value, the $C[y, y']$ term for both the sensitive attribute values become same. This attack then relies on the $p_{1,i}$ term for attack prediction and thus mostly predicts 'single' because of its higher marginal prior which *simply boils the FJR attack [14] down to a naive attack*. However, for the records with class label >50K, the FJR attack [14] performs better since queries with varying sensitive attribute values produce target model outcomes that fall into different cells of the confusion matrix. The $C[y, y']$ term dominates $p_{1,i}$ and thus the attack returns more true positives.

The CSMIA strategy also shows similar trends based on the confidence scores returned by the target model, i.e., the performance of the attack is significantly different for the two class labels. However, CSMIA outperforms FJR attack [14] significantly in terms of recall for class >50K.

Unlike FJR attack [14] and CSMIA, the CMMIA strategy with trained attack models performs well on <=50K class instances with a recall of 60.5% while achieving a precision similar to the other attacks. The CMMIA strategy also performs well on >50K class instances with significantly high recall when compared to FJR attack [14]. These results show the advantage of the CMMIA strategy over the other attacks which is clearly due to the fact that the CMMIA strategy focuses on learning the correlation between the outcome (both predictions and confidence scores) of the target model and the sensitive attribute by training the attack models.

6.9 Discussion

As mentioned in Section 1, the TIR attacks have strong correlations with the model's predictive power. This is because highly predictive models are able to establish a strong correlation between features and labels, and this is the property that an adversary exploits to mount the TIR attacks [18]. However, we argue that such is not the case for MIAI attacks. In Section 6.8, where it is evident that the target model's performance is better for class label <= $50K$ than that of class label > $50K$, we demonstrate that the MIAI attacks perform better against the records of class label > $50K$. Moreover, in Section 6.4, we show that the correlations between the sensitive attributes and the corresponding target models trained on GSS and Adult datasets do not differ significantly whereas the proposed MIAI attacks on these target models demonstrate significantly different results (attacks against target models trained on Adult dataset are more effective than that of against the target models trained on GSS dataset). This indicates that only controlling the *importance* of the sensitive attributes in the target model may not be always sufficient to reduce the risk of model inversion attacks. Finally, we investigate if black-box access to a particular model really helps the adversary to estimate the sensitive attributes in the training dataset which is otherwise impossible for the adversary to estimate, i.e., without access to the black-box model, by performing extensive experiment with baseline attacks (random guessing attack and naive attack).

Hence, *ours is the first work that studies the MIAI attacks in such details* and it is evident that further investigation is required to better understand the potentially serious threats of model inversion attacks such as its unequal impact on different groups of individuals as described in Section 6.6.

7 RELATED WORK

In [15], Fredrikson et al. introduced the concept of model inversion attacks and applied their attack to linear regression models. In [14], Fredrikson et al. extended their attack so that it could also be applicable to non-linear models, such as decision trees. The later work presents two types of applications of the model inversion attack. The first one assumes an adversary who has access to a model (for querying) and aims to learn the sensitive attributes in the dataset that has been used to train that model (also known as attribute inference attack). In the second setting, the adversary aims to reconstruct instances similar to ones in the training dataset using gradient descent. Particularly, their attack generates images similar to faces used to train a facial recognition model. As mentioned earlier, we focus on the first one, i.e., attribute inference attack. Subsequently, Wu et al. [24] presented a methodology to formalize the model inversion attack.

A number of attribute inference attacks have been shown to be effective in different domains, such as social media [25–31] and recommender systems [32, 33]. In the case of social media, the adversary infers the private attributes of a user (e.g., gender, political views, locations visited) by leveraging the knowledge of other attributes of that same user that are shared publicly (e.g., list of pages liked by the user, etc). The adversary first trains a machine learning classifier that takes as input the public attributes and then outputs the private attributes. In order to build such a classifier, these attacks [25–31] rely on social media users who also make their private attributes public. Also, for the attacks shown in the recommender systems [32], at first, the adversary collects data of the users who also share their private attributes (e.g., gender) publicly along with their public rating scores (e.g., movie ratings). While our CMMIA strategy assumes the adversary to be able to obtain a dataset from the same population the target model training dataset has been obtained from, in our CSMIA attack we do not make such an assumption. This is because in some special scenarios such an assumption (adversary having access to a similar dataset) may not be valid and our goal is to also incorporate these scenarios in our attack surface so that our attack could be applied more widely.

Shokri et al. [34] investigate whether transparency of machine learning models conflicts with privacy and demonstrate that record-based explanations of machine learning models can be effectively exploited by an adversary to reconstruct the training dataset. In their setting, the adversary can generate unlimited transparency queries and for each query, the adversary is assumed to get in return some of the original training dataset records (that are related to the queries) as part of the transparency report. He et al. [35] devise a new set of model inversion attacks against collaborative inference where a deep neural network and the corresponding inference task are distributed among different participants. The adversary, as a malicious participant, can accurately recover an arbitrary input fed into the model, even if it has no access to other participants' data or computations, or to prediction APIs to query the model.

Most of the work mentioned above assume that the attributes of a target individual, except the sensitive attribute, are known to the adversary. Hidano et al. [36] proposed a method to infer the sensitive attributes without the knowledge of non-sensitive attributes. However, they consider an online machine learning model and assume that the adversary has the capability to poison the model with malicious training data. In contrast, our model inversion attack with partial knowledge of target individual's non-sensitive attributes does not require poisoning and performs similar to scenarios where the adversary has full knowledge of target individual's non-sensitive attributes.

Zhang et al. [18] present a generative model-inversion attack to invert deep neural networks. They demonstrate the effectiveness of their attack by reconstructing face images from a state-of-the-art face recognition classifier. They also prove that a model's predictive power and its vulnerability to inversion attacks are closely related, i.e., highly predictive models are more vulnerable to inversion attacks. Aïvodji et al. [16] introduce a new black-box model inversion attack framework, GAMIN (Generative Adversarial Model INversion), based on the continuous training of a surrogate model for the target model and evaluate their attacks on convolutional neural networks. In [17], Yang et al. train a second neural network that acts as the inverse of the target model while assuming partial knowledge about the target model's training data. The objective of the works mentioned above is typical instance reconstruction (TIR), i.e., similar to the second attack mentioned in [14].

8 CONCLUSION AND FUTURE WORK

In this paper, we demonstrate two new black-box model inversion attribute inference (MIAI) attacks: (1) confidence modeling-based attack (CMMIA) and (2) confidence score-based attack (CSMIA). The CMMIA strategy assumes the adversary has access to a dataset from the same population the target model training dataset has been obtained from (similar to other state-of-the-art attacks [17, 25–32]) whereas the CSMIA strategy does not make such an assumption. Along with accuracy and F1 score, we propose to use the G-mean and Matthews correlation coefficient (MCC) metrics in order to ensure effective evaluation of our attacks as well as the state-of-the-art attacks. We perform an extensive evaluation of our attacks using two types of machine learning models, decision tree and deep neural network, that are trained with two real datasets [19, 20]. Our evaluation results show that the proposed attacks significantly outperform the existing ones. Moreover, we empirically show that model inversion attacks have inequity property and consequently, a particular subset of the training dataset (grouped by attributes, such as gender, education level, etc.) could be more vulnerable than others to the model inversion attacks. We also evaluate the risks incurred by model inversion attacks when the adversary does not have knowledge of all other non-sensitive attributes of the target individual and demonstrate that our attack's performance is not impacted significantly in those scenarios.

Further investigating the potentially serious threats of model inversion attacks such as its inequity property on different groups of individuals and also, extending the model inversion attacks to dynamic environments where the model and/or the non-sensitive attributes of the target individuals could change over time are left as future work. Since the defense methods designed to mitigate reconstruction of instances resembling those used in the training dataset (TIR attacks) [37, 38] do not directly apply to our MIAI attack setting, it would also be an interesting direction for future work to explore new defense methods.

REFERENCES

[1] International Warfarin Pharmacogenetics Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8):753–764, 2009.

[2] Jeremy C Weiss, Sriraam Natarajan, Peggy L Peissig, Catherine A McCarty, and David Page. Machine learning for personalized medicine: Predicting primary myocardial infarction from electronic health records. *Ai Magazine*, 33(4):33–33, 2012.

[3] Davide Cirillo and Alfonso Valencia. Big data analytics for personalized medicine. *Current opinion in biotechnology*, 58:161–167, 2019.

[4] Marinka Zitnik, Francis Nguyen, Bo Wang, Jure Leskovec, Anna Goldenberg, and Michael M Hoffman. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 50:71–91, 2019.

[5] Xiaoyuan Su and Taghi M Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009, 2009.

[6] G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[7] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[8] Christian Dunis, Peter W Middleton, A Karathanasopolous, and K Theofilatos. *Artificial intelligence in financial markets*. Springer, 2016.

[9] Robert R Trippi and Efraim Turban. *Neural networks in finance and investing: Using artificial intelligence to improve real world performance*. McGraw-Hill, Inc., 1992.

[10] Mireille Hildebrandt. Law as computation in the era of artificial legal intelligence: Speaking law to the power of statistics. *University of Toronto Law Journal*, 68(supplement 1):12–35, 2018.

[11] Daniel Gayo-Avello, Panagiotis Takis Metaxas, Eni Mustafaraj, Markus Strohmaier, Harald Schoen, and Peter Gloor. The power of prediction with social media. *Internet Research*, 2013.

[12] Golnoosh Farnadi, Geetha Sitaraman, Shanu Sushmita, Fabio Celli, Michal Kosinski, David Stillwell, Sergio Davalos, Marie-Francine Moens, and Martine De Cock. Computational personality recognition in social media. *User modeling and user-adapted interaction*, 26(2-3):109–142, 2016.

[13] Marcin Skowron, Marko Tkalčič, Bruce Ferwerda, and Markus Schedl. Fusing social media cues: personality prediction from twitter and instagram. In *Proceedings of the 25th international conference companion on world wide web*, pages 107–108, 2016.

[14] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pages 1322–1333, New York, NY, USA, 2015. ACM.

[15] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 17–32, San Diego, CA, August 2014. USENIX Association.

[16] Ulrich Aïvodji, Sébastien Gambs, and Timon Ther. Gamin: An adversarial approach to black-box model inversion. 2019.

[17] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, page 225–240, New York, NY, USA, 2019. Association for Computing Machinery.

[18] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[19] The general social survey. https://gss.norc.org/.

[20] Adult dataset. http://archive.ics.uci.edu/ml/datasets/Adult.

[21] Yanmin Sun, Andrew K. C. Wong, and Mohamed S. Kamel. Classification of imbalanced data: a review. *Int. J. Pattern Recognit. Artif. Intell.*, 23:687–719, 2009.

[22] B.W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442 – 451, 1975.

[23] Bigml. https://bigml.com/.

[24] Xi Wu, Matthew Fredrikson, Somesh Jha, and Jeffrey F Naughton. A methodology for formalizing model-inversion attacks. In *2016 IEEE 29th Computer Security Foundations Symposium (CSF)*, pages 355–370. IEEE, 2016.

[25] Neil Zhenqiang Gong and Bin Liu. Attribute inference attacks in online social networks. *ACM Trans. Priv. Secur.*, 21(1), January 2018.

[26] Jinyuan Jia, Binghui Wang, Le Zhang, and Neil Zhenqiang Gong. Attriinfer: Inferring user attributes in online social networks using markov random fields. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1561–1569, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

[27] Neil Zhenqiang Gong and Bin Liu. You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 979–995, Austin, TX, August 2016. USENIX Association.

[28] Neil Zhenqiang Gong, Ameet Talwalkar, Lester Mackey, Ling Huang, Eui Chul Richard Shin, Emil Stefanov, Elaine (Runting) Shi, and Dawn Song. Joint link prediction and attribute inference using a social-attribute network. *ACM Trans. Intell. Syst. Technol.*, 5(2), April 2014.

[29] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[30] Abdelberi Chaabane, Gergely Acs, and Mohamed Ali Kaafar. You are what you like! information leakage through users' interests. In *In NDSS*, 2012.

[31] Elena Zheleva and Lise Getoor. To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, page 531–540, New York, NY, USA, 2009. Association for Computing Machinery.

[32] Udi Weinsberg, Smriti Bhagat, Stratis Ioannidis, and Nina Taft. Blurme: Inferring and obfuscating user gender based on ratings. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 195–202, 2012.

[33] Le Wu, Yonghui Yang, Kun Zhang, Richang Hong, Yanjie Fu, and Meng Wang. Joint item recommendation and attribute inference: An adaptive graph convolutional network approach. 2020.

[34] Reza Shokri, Martin Strobel, and Yair Zick. Privacy risks of explaining machine learning models. *arXiv preprint arXiv:1907.00164*, 2019.

[35] Zecheng He, Tianwei Zhang, and Ruby B. Lee. Model inversion attacks against collaborative inference. In *Proceedings of the 35th Annual Computer Security Applications Conference*, ACSAC '19, page 148–162, New York, NY, USA, 2019. Association for Computing Machinery.

[36] S. Hidano, T. Murakami, S. Katsumata, S. Kiyomoto, and G. Hanaoka. Model inversion attacks for prediction systems: Without knowledge of non-sensitive attributes. In *2017 15th Annual Conference on Privacy, Security and Trust (PST)*, pages 115–11509, 2017.

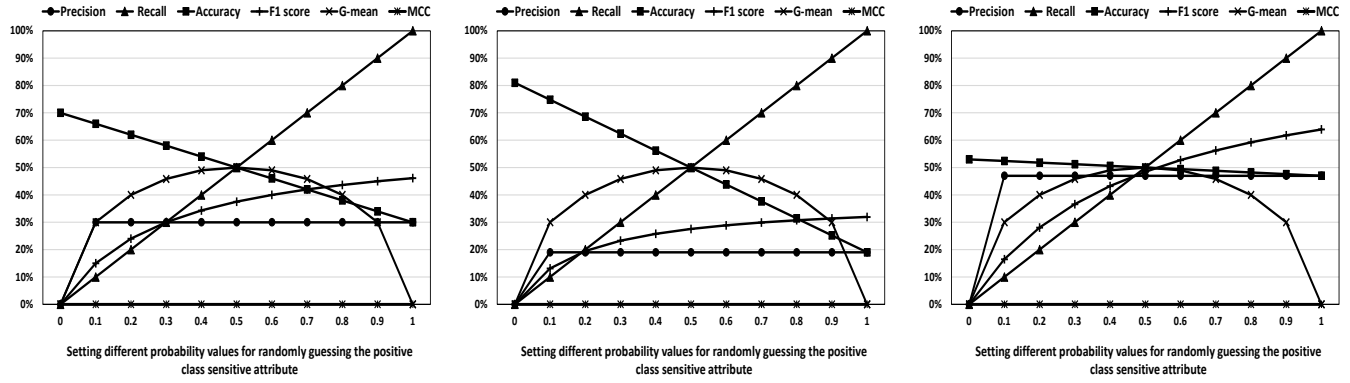
[37] Tiago A. O. Alves, Felipe M. G. França, and Sandip Kundu. Mlprivacyguard: Defeating confidence information based model inversion attacks on machine learning systems. In *Proceedings of the 2019 on Great Lakes Symposium on VLSI*, GLSVLSI '19, page 411–415, New York, NY, USA, 2019. Association for Computing Machinery.

[38] Ziqi Yang, Bin Shao, Bohan Xuan, Ee-Chien Chang, and Fan Zhang. Defending model inversion and membership inference attacks via prediction purification, 2020.

A APPENDIX

A.1 Random Guessing Attack Performances

In this attack, the adversary randomly predicts the sensitive attribute by setting a probability for the positive class sensitive attribute value. Fig. 5(a) shows the optimal performance of random guessing attack when the marginal prior of the positive class sensitive attribute is 0.3 and the adversary sets different probabilities to predict the positive class sensitive attribute value (probabilities in x-axis). As shown in the figure, the maximum G-mean a random guessing attack can achieve is 50%, independent of the knowledge of marginal prior. The precision for predicting the positive class sensitive attribute is constant and equals the marginal prior of that class as long as the set probability is > 0. This is because when the attack randomly assigns positive class label to the records, approximately 30% of those records' sensitive attributes would turn out to be originally positive according to the marginal prior of the positive class sensitive attribute which is 0.3. The recall of random guessing attack increases with the probability set to predict the positive class sensitive attribute. For example, if the adversary reports all the records' sensitive attributes as positive, there is no false negative left and thus recall reaches 100%. Figures 5(b) and 5(c) show the performance of random guessing attack on the GSS and Adult datasets, respectively, when the adversary sets different probability values to predict the positive class sensitive attribute. As shown in



(a) Marginal prior of the positive class attribute is 0.3.

(b) GSS dataset where the marginal prior of the positive class attribute is 0.197.

(c) Adult dataset where the marginal prior of the positive class attribute is 0.479.

Figure 5: Random guessing attack performances for different marginal priors of the positive class sensitive attribute value.

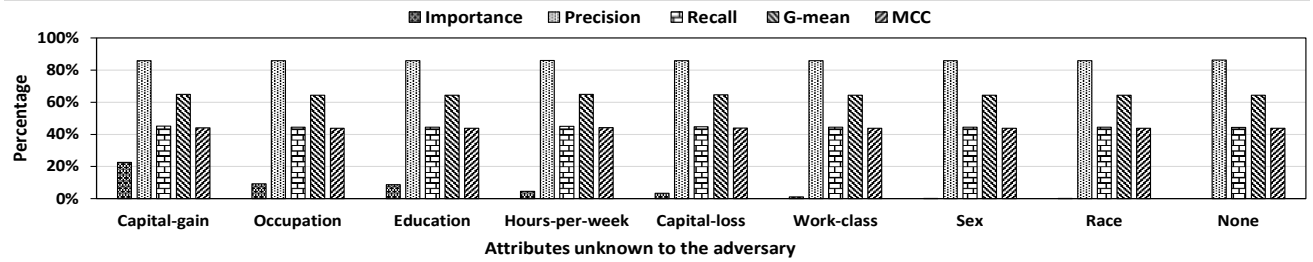
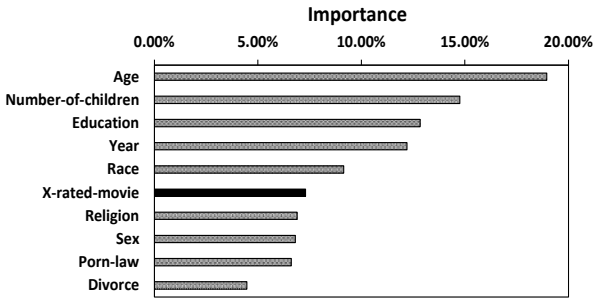


Figure 6: Confidence score-based attack performance against the deep neural network model trained on Adult dataset when some of the other (non-sensitive) attributes of a target individual are also unknown to the adversary.

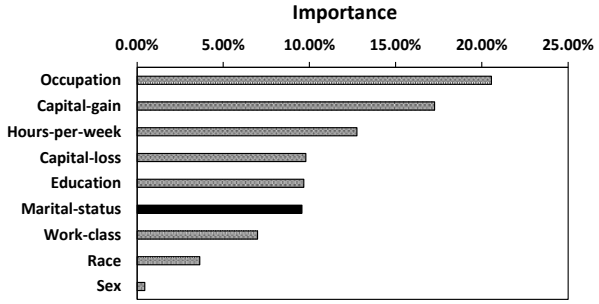
Figures 5(a), 5(b), and 5(c), the MCC of the random guessing attacks is always 0.

A.2 Model Inversion Attack Results With Partial Knowledge of Target Individual's Non-sensitive Attributes

Fig. 6 shows the performance of our confidence score-based attack on the deep neural network target model trained on the Adult dataset when some of the non-sensitive attributes are unknown to the adversary. The x-axis shows the non-sensitive attribute that is unknown. The attributes are sorted (from left to right) according to their *importance* in the model. We also present the original results (i.e., when *none* of the non-sensitive attributes is unknown to the adversary) to compare how the partial knowledge of the target individual's non-sensitive attributes impacts our attacks' performances. As demonstrated in the figure, we observe that the performances of our attack do not deteriorate and remain almost the same when some of the non-sensitive attributes are unknown to the adversary, independent of the importance of the attributes in the target model.



(a) GSS dataset attributes' importance



(b) Adult dataset attributes' importance

Figure 7: Importance of GSS and Adult dataset attributes on their corresponding decision tree target models.

Table 9: Confusion matrix of decision tree target model trained on GSS dataset.

Actual \ Predicted	Not too happy	Pretty happy	Very happy	Total	Recall
Not too happy	5	63	370	438	1.14%
Pretty happy	0	813	4178	4991	16.29%
Very happy	0	526	9280	9806	94.64%
Total	5	1402	13828	15235	Avg. recall 37.36%
Precision	100%	57.99%	67.11%	Avg. precision 75.03%	Accuracy 66.28%

Table 10: Confusion matrix of deepnet target model trained on GSS dataset.

Actual \ Predicted	Not too happy	Pretty happy	Very happy	Total	Recall
Not too happy	1	102	335	438	0.23%
Pretty happy	0	565	4426	4991	11.32%
Very happy	0	598	9208	9806	93.90%
Total	1	1265	13969	15235	Avg. recall 35.15%
Precision	100%	44.66%	65.92%	Avg. precision 70.19%	Accuracy 64.16%

Table 11: Confusion matrix of decision tree target model trained on Adult dataset.

Actual \ Predicted	<=50K	>50K	Total	Recall
<=50K	24912	1537	26449	94.19%
>50K	3343	5430	8773	61.89%
Total	28255	6967	35222	Avg. recall 78.04%
Precision	88.17%	77.94%	Avg. precision 83.05%	Accuracy 86.15%

Table 12: Confusion matrix of deepnet target model trained on Adult dataset.

Actual \ Predicted	<=50K	>50K	Total	Recall
<=50K	24433	2016	26449	92.38%
>50K	3276	5497	8773	62.66%
Total	27709	7513	35222	Avg. recall 77.52%
Precision	88.18%	73.17%	Avg. precision 80.67%	Accuracy 84.97%

Table 13: Our proposed attacks' performance details against the decision tree target model trained on GSS dataset.

Attack	Case	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC
Confidence modeling-based attack	(1)	160	1477	554	196	22.41%	44.94%	68.58%	29.91%	57.17%	13.7%
Confidence score-based attack		219	1336	698	134	23.88%	62.04%	65.14%	34.49%	63.83%	20.2%
Confidence modeling-based attack	(2)	1050	4035	2840	618	26.99%	62.95%	59.52%	37.78%	60.78%	17.2%
Confidence score-based attack		661	4466	2409	1007	21.53%	39.63%	60.01%	27.91%	50.74%	3.8%
Confidence modeling-based attack	(3)	556	2093	1216	440	31.38%	55.82%	61.53%	40.17%	59.42%	16.3%
Confidence score-based attack		610	2042	1266	387	32.52%	61.18%	61.61%	42.46%	61.46%	19.5%

Table 14: Our proposed attacks' performance details against the deep neural network target model trained on GSS dataset.

Attack	Case	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC
Confidence modeling-based attack	(1)	66	592	193	160	25.48%	29.21%	65.08%	27.22%	46.93%	4.4%
Confidence score-based attack		96	468	317	130	23.24%	42.48%	55.79%	30.05%	50.32%	1.8%
Confidence modeling-based attack	(2)	636	4863	2681	1030	19.17%	38.18%	59.71%	25.53%	49.61%	2.1%
Confidence score-based attack		55	7339	205	1611	21.15%	3.3%	80.28%	5.71%	17.92%	1.4%
Confidence modeling-based attack	(3)	398	2678	1211	727	24.74%	35.38%	61.35%	29.11%	49.36%	3.8%
Confidence score-based attack		1061	251	3638	64	22.58%	94.31%	26.17%	36.44%	24.67%	1.3%

Table 15: Our proposed attacks' performance details against the decision tree target model trained on Adult dataset.

Attack	Case	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC
Confidence modeling-based attack	(1)	4520	2150	1827	766	71.21%	85.51%	72.01%	77.71%	67.99%	42.2%
Confidence score-based attack		3788	3466	511	1498	88.11%	71.66%	78.31%	79.04%	79.03%	58.4%
Confidence modeling-based attack	(2)	5991	9323	4693	3081	56.07%	66.04%	66.33%	60.65%	66.28%	31.9%
Confidence score-based attack		1375	13560	456	7697	75.09%	15.16%	64.68%	25.22%	38.29%	21.5%
Confidence modeling-based attack	(3)	1800	146	190	735	90.45%	71.01%	67.78%	79.56%	55.55%	10.1%
Confidence score-based attack		2501	59	277	34	90.03%	98.66%	89.17%	94.15%	41.62%	29.5%

Table 16: Our proposed attacks' performance details against the deep neural network target model trained on Adult dataset.

Attack	Case	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC
Confidence modeling-based attack	(1)	3332	2982	1468	2178	69.42%	60.47%	63.39%	64.64%	63.66%	27.3%
Confidence score-based attack		3592	3838	612	1918	85.44%	65.19%	74.6%	73.96%	74.98%	51.8%
Confidence modeling-based attack	(2)	6129	8304	5275	2792	53.74%	68.7%	64.15%	60.31%	64.82%	29.2%
Confidence score-based attack		1467	13235	344	7454	81.01%	16.44%	65.34%	27.34%	40.03%	25%
Confidence modeling-based attack	(3)	2446	42	258	16	90.46%	99.35%	90.08%	94.7%	37.29%	29%
Confidence score-based attack		2431	66	234	31	91.22%	98.74%	90.41%	94.83%	46.61%	35.1%