Are Your Sensitive Attributes Private? Novel Model Inversion Attribute Inference Attacks on Classification Models

Shagufta Mehnaz, *The Pennsylvania State University;* Sayanton V. Dibbo and
Ehsanul Kabir, *Dartmouth College;* Ninghui Li and Elisa Bertino, *Purdue University*

This paper is included in the Proceedings of the 31st USENIX Security Symposium.

August 10–12, 2022 • Boston, MA, USA

Are Your Sensitive Attributes Private?
Novel Model Inversion Attribute Inference Attacks on Classification Models

Shagufta Mehnaz[1], Sayanton V. Dibbo[2], Ehsanul Kabir[2], Ninghui Li[3], and Elisa Bertino[3]

[1]Department of Computer Science and Engineering, The Pennsylvania State University
[2]Department of Computer Science, Dartmouth College
[3]Department of Computer Science, Purdue University
smehnaz@psu.edu, {sayanton.v.dibbo.gr, ehsanul.kabir.gr}@dartmouth.edu,
{ninghui, bertino}@purdue.edu

Abstract

Increasing use of machine learning (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 leakage of sensitive and proprietary training data. In this paper, we focus on model inversion attacks where the adversary knows non-sensitive attributes about records in the training data and aims to infer the value of a sensitive attribute unknown to the adversary, using only black-box access to the target classification model. We first devise a novel confidence score-based model inversion attribute inference attack that significantly outperforms the state-of-the-art. We then introduce a label-only model inversion attack that relies only on the model's predicted labels but still matches our confidence score-based attack in terms of attack effectiveness. We also extend our attacks to the scenario where some of the other (non-sensitive) attributes of a target record are unknown to the adversary. We evaluate our attacks on two types of machine learning models, decision tree and deep neural network, trained on three real datasets. Moreover, we empirically demonstrate the disparate vulnerability of model inversion attacks, i.e., specific groups in the training dataset (grouped by gender, race, etc.) could be more vulnerable to model inversion attacks.

1 Introduction

Across numerous sectors, the use of ML technologies trained on proprietary and sensitive datasets has increased significantly, e.g., in the domains of personalized medicine [1, 2], product recommendation [3–5], finance and law [6, 7], social media [8, 9], etc. Companies provide access to such trained ML models through APIs [10–14] whereas users querying these models are charged on a pay-per-query basis. With the increasing use of ML technologies in personal data, we have seen a recent surge of serious privacy concerns that were previously ignored [15–18]. Therefore, it is important to investigate 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. A *model inversion attack* is one of such attacks on ML 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 when it is given access to the target ML model.

Fredrikson et al. [18, 19] 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 a 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 about one individual. In the second formulation, which we call **typical instance reconstruction (TIR)** attack, the adversary is given access to a classification model and a particular 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 individuals' faces, tries to reconstruct an image that is similar to a target individual's actual facial image. For TIR attacks [20–22] to be considered successful, it is not necessary for a reconstructed instance to be quantitatively close to any specific training instance. In contrast, MIAI attacks are evaluated by the ability to predict *exact* attribute values of individual instances. Evaluation of TIR attacks is typically done by having humans assess the 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 has been proven [22] 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 [22]. 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 investigate whether the root cause of TIR attacks (high predictive power) also applies to MIAI attacks. According to our observation, we

point out that such is not the case.

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 has been studied extensively in the privacy literature (e.g., user attribute inference in social networks [23, 24]), there exists little work studying to what extent model inversion introduces new attribute inference 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: In this paper, we devise two new black-box MIAI attacks: (1) confidence score-based model inversion attack (CSMIA) and (2) label-only model inversion attack (LOMIA). The confidence score-based attack assumes that the adversary has access to the target model's confidence scores whereas the label-only attack assumes the adversary's access to the target model's label predictions only. To the best of our knowledge, ours is the first work to propose a label-only MIAI attack. We empirically show that despite having access to only the predicted labels, our label-only attack performs on par with the proposed confidence score-based attack. Also, both of our proposed attacks outperform state-of-the-art attacks significantly. Furthermore, we note that defense mechanisms [18] that reduce the precision of confidence scores or introduce noise in the confidence scores to thwart model inversion attacks are ineffective against our label-only attack.

While the existing attacks [18, 19] assume that the adversary has full knowledge of other non-sensitive attributes of the target record, we also propose extensions of our attacks that work even when some non-sensitive attributes are unknown to the adversary. We also evaluate cases where an adversary aims to estimate multiple sensitive attributes of a target record which also has not been explored in the existing MIAI attacks [18, 19]. Moreover, we 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., records grouped by race, gender, occupation, 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. We use the term *disparate vulnerability* to represent such discrimination. We further investigate if model inversion attribute inference attacks are able to infer the sensitive attributes in data records that do not belong to the training dataset of the target model but are drawn from the same distribution. A model inversion attack with such capability compromises the privacy of not only the target model's training dataset but also breaches its *distributional privacy*.

We train two models– a *decision tree* and a *deep neural network* with each of the three real datasets in our experiments, General Social Survey (GSS) [25], Adult dataset [26], and

FiveThirtyEight dataset [27], to evaluate our proposed attacks. To the best of our knowledge, ours is the first work that studies MIAI attacks in such details on *tabular* datasets which is the most common data type used in real-world ML [28].

Effective evaluation of model inversion attacks: Although the Fredrikson et al. attack [18] primarily uses accuracy to evaluate model inversion attacks, 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 a significant F1 score. Hence, we propose to also use G-mean [29] and Matthews correlation coefficient (MCC) [30] as metrics in addition to precision, recall, accuracy, false positive rate (FPR), and F1 score to design a framework that can effectively evaluate any model inversion attack. While the existing MIAI attacks [18, 19] evaluate their performance on binary sensitive attributes only, we evaluate our attacks on multi-valued sensitive attributes as well. We use attack confusion matrices to evaluate the attack performances in estimating multi-valued sensitive attributes. Finally, we evaluate the required number of queries to the black-box target models to perform the proposed attacks.

Comparison with baseline attribute inference attacks: We also compare the performances of our MIAI attacks with those from attacks that do not query the target model, e.g., randomly guessing the sensitive attribute according to some distribution. When a particular model inversion attack deployed against a target model performs similarly to such attacks, we can conclude that the target model is 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 which is otherwise impossible for the adversary? We demonstrate that our proposed attacks significantly outperform baseline attribute inference attacks that do not require access to the target model.

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

1. We design two new black-box MIAI attacks: (1) confidence score-based MIAI attack and (2) label-only MIAI attack. We define the various capabilities of the adversary and provide a detailed threat model.

2. We conduct an extensive evaluation of our attacks using two types of ML models, decision tree and deep neural network, trained with three real datasets. Evaluation results show that our proposed attacks significantly outperform the existing attacks in inferring both binary and

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

Attack strategy	Predicted label	Confidence score along with predicted label	Target individuals' all non-sensitive attributes including true label	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
NaiveA				✓	✓		
RandGA				✓	✓ (optional)		
FJRMIA [18]	✓		✓	✓	✓	✓	✓
CSMIA	✓	✓	✓	✓			
LOMIA	✓		✓	✓			✓

multi-valued sensitive attributes. Moreover, our label-only attack performs on par with the proposed confidence score-based MIAI attack despite having access to only the predicted labels of the target model.

3. We extend both of our proposed attacks to the scenario where some of the other (non-sensitive) attributes of a target record are unknown to the adversary and demonstrate that the performance of our attacks is not impacted significantly in those circumstances.

4. We uncover that a particular subset of the training dataset (grouped by attributes, such as gender, race, etc.) could be more vulnerable than others to the model inversion attacks, a phenomenon we call *disparate vulnerability*.

2 Problem Definition and Existing Attacks

2.1 Model Inversion Attribute Inference

An ML model can be represented using a deterministic function f where the input of this function is a d-dimensional vector $\mathbf{x} = [x_1, x_2, ..., x_d]$ 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, f outputs y' if it returns only the predicted label and outputs $\mathcal{R}^m$ if it returns the confidence scores as well, where m is the number of unique class labels $(y_1, y_2, ..., y_m)$ and $\mathcal{R}^m$ represents the corresponding confidence scores returned for these m class labels. Finally, the class label with the highest confidence score is considered as the model's prediction label. 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 and vice versa.

Now, some of the attributes in $\mathbf{x}$ 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, e.g., access to the black-box target model, background knowledge about the target individual, etc.

2.2 Threat Model

The adversary is assumed to have all or a subset of the following capabilities in different attacks (see Table 1 for details):

- Access to the black-box target model, i.e., the adversary can query the model with $\mathbf{x}$ and obtain y'.

- 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 and also knowledge of the true label of the target record.
- 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 $(\mathcal{C})$ of the model where $\mathcal{C}[y, y'] = Pr[f(\mathbf{x}) = y' | y$ is the true label$]$. Here, confusion matrix represents the performance of an ML model when queried on the entire training dataset [18].

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). While our CSMIA strategy does not require the knowledge of the target model confusion matrix, the LOMIA strategy indirectly assumes this knowledge. The adversary has only black-box access to the model, i.e., it has no knowledge of the model details (e.g., architecture or parameters). Finally, 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 (NaiveA)

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 (RandGA)

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. Figure 7(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 is

independent of its knowledge of marginal prior (0.3 in this example). 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% (Figure 7(a) in Appendix).

2.4 Fredrikson et al. Attack [18] (FJRMIA)

This black-box model inversion attack [18] assumes that the adversary can obtain the model's predicted label, has knowledge of all the attributes of a targeted record (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 (see Table 1). The adversary queries the target model by varying the sensitive attribute (x_1) and obtains 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 $C[y, y'] = Pr[f(\mathbf{x}) = y' \mid y \text{ is the true label}]$ and $p_{1,i}$ is the marginal prior of i-th possible sensitive attribute value. Finally, the attack predicts the sensitive attribute value for which the $C[y, y'] * p_{1,i}$ value is the maximum.

3 Metrics for Evaluating MIAI Vulnerability

Understanding a model's vulnerability to inversion attacks requires a meaningful metric to evaluate and compare different model inversion attacks. The FJRMIA [18] primarily uses accuracy. However, if we care only about the 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 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 a 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 more strategically. Therefore, 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 precision, recall, accuracy, FPR, and F1 score: G-mean [29] and Matthews correlation coefficient (MCC) [30], as described below.

G-mean: G-mean is the geometric mean of sensitivity and specificity [29]. Thus it takes all of the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) into account. With this metric, the random guessing attack can achieve 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 (Figure 7(a) in Appendix). For the 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%.

$$G - mean = \sqrt{\frac{TP}{TP+FN} * \frac{TN}{TN+FP}} \qquad (1)$$

Matthews correlation coefficient (MCC): This MCC metric also takes into account all of TP, TN, FP, and FN, and is a balanced measure that can be used even if the classes of the sensitive attribute are of very different sizes [30]. 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 (either TP=FP=0 or TN=FN=0).

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

4 New Model Inversion Attacks

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

4.1 Confidence Score-based Attack (CSMIA)

This attack exploits the confidence scores returned by the target model. Unlike FJRMIA [18], the adversary assumed in this attack does not have access to the marginal priors or the confusion matrix. The adversary knows the true labels for the records it is attacking (Table 1). In many privacy attacks, the attacker compromises the privacy of just a small subset of the records. For such a small set of target records (e.g., AOL-dataset [31]), it is common to assume that the attacker knows all information (including the true label) except the attribute it is attacking. This assumption is consistent with prior work [18]. The *key idea* of our CSMIA 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 original sensitive attribute value (since the target model encountered the target record with original sensitive attribute value during training). In contrast, the target model's returned prediction is more likely to be incorrect when it is queried with a record containing the wrong sensitive attribute value.

The adversary first 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 record remain the same. If the sensitive attribute is continuous, we can use binning
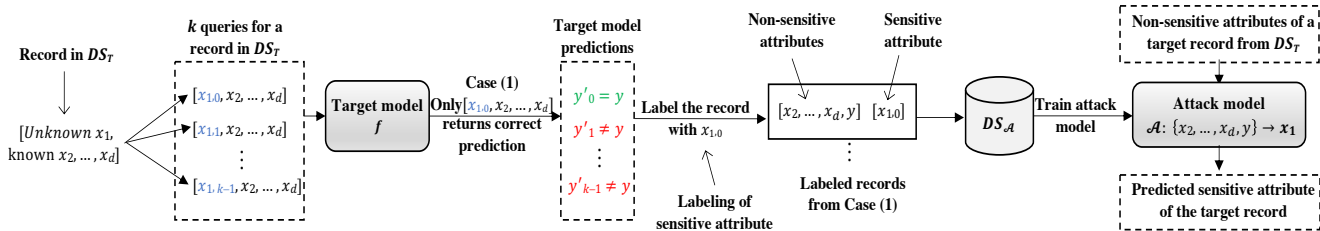
Figure 1: LOMIA: The adversary collects the case (1) records by querying the target model f, obtains the $DS_\mathcal{A}$ dataset, and trains the attack model $\mathcal{A}$. The adversary then leverages the attack model to predict the sensitive attribute values of the target records.

to turn it into a categorical attribute and recover an approximate value. 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 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 record 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*. In order to determine the value of x_1, this attack considers the following three cases:

Case (1) If the target model's prediction is correct *only for a single sensitive attribute value*, e.g., $y = y_0' \land y \mathrel{!=} y_1'$ or $y \mathrel{!=} y_0' \land y = y_1'$ in the event of a binary sensitive attribute, the attack selects the sensitive attribute to be the one for which the prediction is correct. For instance, if $y = y_1' \land y \mathrel{!=} y_0'$, the attack predicts *yes* for the sensitive attribute and vice versa. *Note that, for this case, the adversary only requires the predicted labels and does not require the confidence scores.* We leverage the records that fall into this case in our label-only attack as described later in Section 4.2.

Case (2) If the model's prediction is correct for multiple sensitive attribute values, i.e., $y = y_0' \land y = y_1'$, 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's prediction is correct with 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., $y \mathrel{!=} y_0' \land y \mathrel{!=} y_1'$, 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 higher confidence when *yes* value is set for the sensitive attribute, the attack predicts the *no* value for x_1 and vice versa.

4.2 Label-Only Attack (LOMIA)

This advanced attack assumes the adversary's access to the target model's predicted labels only. Therefore, defense mechanisms [18] that reduce the precision of confidence scores or introduce noise in the confidence scores in order to thwart

model inversion attacks are ineffective against our label-only attack. The attack has the following steps as shown in Figure 1: (1) obtaining an attack dataset ($DS_\mathcal{A}$), (2) training an attack model $\mathcal{A}$ from $DS_\mathcal{A}$, and (3) leveraging $\mathcal{A}$ to infer the sensitive attributes of target records.

4.2.1 Obtaining Attack Dataset $DS_\mathcal{A}$

The *key intuition* of this attack step is that if the target model f returns the correct prediction (y) for only one possible value of the sensitive attribute, it is highly likely that this particular value represents the original sensitive attribute value, e.g., sensitive attribute value $x_{1,0}$ in Figure 1. Hence, the adversary then labels the record in this example with $x_{1,0}$. The adversary collects all such labeled records that fall into Case (1) as described in Section 4.1 and obtains the $DS_\mathcal{A}$ dataset. Note that, the labeling of sensitive attributes might have some errors, e.g., $x_{1,0}$ in Figure 1 might not be the original sensitive attribute of the record even though only with this value the target model returns the correct prediction. Table 3 in Section 5.2 shows the sizes of the $DS_\mathcal{A}$ datasets obtained from different target models in our experiments and their corresponding accuracy. However, since the LOMIA attacker does not know the original sensitive attribute values, it uses the entire $DS_\mathcal{A}$ datasets to train the attack models.

Note that, while building the attack model dataset $DS_\mathcal{A}$, we assume that the adversary knows the real y attribute of all the instances in the training dataset. In other words, unlike CSMIA, the adversary in the LOMIA strategy assumes the knowledge of the target model confusion matrix (Table 1).

4.2.2 Training Attack Model $\mathcal{A}$

The next step is to train an attack model $\mathcal{A}$ where the input would be the set of non-sensitive attributes from a target record, i.e. a d-dimensional vector $[x_2, ..., x_d, y]$ and the output would be a prediction for the sensitive attribute x_1. The adversary trains this attack model using the $DS_\mathcal{A}$ dataset. The *key goal* of this attack step is to learn how the target model correlates the sensitive attribute with the other non-sensitive attributes including the target model's prediction label. Note that, the dataset used to train the attack model ($DS_\mathcal{A}$) represents a strong correlation of the sensitive attribute values with other non-sensitive ones ($[x_2, ..., x_d, y]$) since it considers only the Case (1) records.

4.2.3 Performing Sensitive Attribute Inference using $\mathcal{A}$

Once the attack model $\mathcal{A}$ is trained, the adversary can simply query $\mathcal{A}$ with the non-sensitive attributes of a target record and obtain a prediction for the sensitive attribute. It is important to note that the adversary could also query the model with the non-sensitive attributes of a record that is not in the training dataset (DS_T), i.e., the record is not used while training the target model. In Section 5.6, we demonstrate the effectiveness of our attacks not only in compromising the privacy of the training dataset but also their performance in breaching the distributional privacy.

4.3 Estimating Multiple Sensitive Attributes

Our LOMIA and CSMIA strategies can be easily extended to cases where the adversary aims to estimate multiple sensitive attributes of a target record. Let, x_1, x_2 be the sensitive attributes the adversary aims to estimate. Our strategies first perform two instances of the attacks and then stitch them together. In other words, while trying to infer x_1, the adversary queries the target model without setting any value for x_2 and vice versa. Making predictions with missing values is common in APIs provided by companies like BigML, Amazon, Microsoft, etc. [17]. For example, in a BigML decision tree model, if an input attribute is missing, there are two strategies to handle such a situation: (1) *last prediction* strategy (default), where the model returns the prediction given by the last node (i.e., parent node) after reaching missing value, and (2) *proportional* strategy, where all branches from the missing value node are considered in majority voting [32]. We adopt the *last prediction* strategy in our experiments. In the case of CSMIA, we estimate the values of x_1 and x_2 independently by executing the CSMIA strategy for each of these two attributes as described in Section 4.1. In the case of LOMIA, we execute the LOMIA strategy independently for each of these two attributes as described in Section 4.2.1 and train two separate attack models to estimate the values of x_1 and x_2. The attack model to estimate x_1 does not take x_2 as an input (since the adversary does not know x_2) and vice versa. Once the multiple sensitive attributes are estimated, we evaluate the performance of the attacks on these two attributes independently.

4.4 Attacks With Partial Knowledge of Target Record's Non-sensitive Attributes

Our attacks proposed in this section as well as the FJR-MIA [18] strategy assume that the adversary has full knowledge of the target record's non-sensitive attributes. However, in many cases, it may be difficult or even impossible for an adversary to obtain all of the non-sensitive attributes of a target record. Therefore, the goal of this section is to quantify the risk of MIAI attacks in the cases where all non-sensitive attributes of a target record are not known to the adversary.

4.4.1 CSMIA With Partial Non-sensitive Attributes

For simplicity, we assume that there is only one non-sensitive attribute ($x_2 \in \mathbf{x}$) that is unknown to the adversary. Extending our attack steps to more than one unknown attribute is straightforward. Let u be the number of unique possible values of x_2. The attacker queries 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 $C_0 \mathrel{!=} C_1$, 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.

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 (e.g., x_3 with v unique possible values), the complexity of the attack becomes $u * v$ times the complexity of the attacks in Section 4. While CSMIA is prone to combinatorial explosion with increasing number of unknown non-sensitive attributes, LOMIA is not, as we describe below.

4.4.2 LOMIA With Partial Non-sensitive Attributes

With partial knowledge of non-sensitive attributes, the attacker first queries the target model omitting the unknown non-sensitive attributes which is handled using the *last prediction* strategy explained in section 4.3 and thus the attacker obtains the $DS_{\mathcal{A}}$ dataset. Note that, the attack dataset $DS_{\mathcal{A}}$ for LOMIA is obtained from Case (1) instances, i.e., the instances where only one sensitive attribute value yields the correct model prediction y while all other known non-sensitive attributes remain unchanged, see Figure 1. Hence, the attack models in LOMIA are highly dependent on the y attribute and are less dependent on other non-sensitive attributes. Therefore, even if multiple non-sensitive attributes, except the y

attribute, are not considered in the attack model, the LOMIA strategy's performance with partial non-sensitive attributes does not degrade significantly.

5 Evaluation

In this section, we discuss our experiment setup and evaluate our proposed attacks. The links to all the datasets and models are available here: https://github.com/smehnaz/black-boxMIAI.

Table 2: Distribution of sensitive attributes in datasets.

Dataset	Sensitive attribute	Positive class label	Negative class label	Positive class count	Positive class %
GSS	X-movie	Yes	No	4002 (3017)	19.7% (19.8%)
Adult	Marital status	Married	Single	21639 (16893)	47.8% (47.9%)
Fivethirtyeight	Alcohol	Yes	No	266	80.3%

5.1 Datasets

General Social Survey (GSS) [25]: FJRMIA [18] uses the *General Social Survey (GSS)* dataset to demonstrate their attack effectiveness. This dataset has 51020 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 attributes 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 20314 records that we use in our experiments. Among these 20314 records, 4002 individuals answered *yes* (sensitive attribute $x_1 = yes$) to the survey question on whether they watched X-rated movies in the last year, i.e., 19.7% positive class (see Table 2). In order to understand if our proposed model inversion attribute inference attacks also breach the privacy of data that is not in the training dataset of the target model but is drawn from the same distribution, we split the dataset and use 75% data to train the target models (15235 records in DS_T) and use the rest 25% data to evaluate attacks on other data from the same distribution (5079 records in DS_D). To ensure consistency, we evaluate other baseline attack strategies including FJRMIA [18] on the target models trained on the DS_T dataset. Among the 15235 records in the DS_T dataset, 3017 individuals answered *yes* to the question on x-rated movies, i.e., 19.8% positive class (see Table 2).

Adult [26]: 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 48842 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

Table 3: $DS_{\mathcal{A}}$ datasets' details obtained from target models.

Dataset	Target Model	Number of instances in $DS_{\mathcal{A}}$	Number of instances with correctly labeled sensitive attribute in $DS_{\mathcal{A}}$
GSS	DT	2387	1555
	DNN	1011	564
Adult	DT	9263	7254
	DNN	9960	7430
Fivethirtyeight	DT	49 (alcohol)	48
	DT	75 (age-group)	72

of 45222 records. Similar to the GSS dataset, we also split the Adult dataset and use 35222 records to train the target models (DS_T) and use the rest 10000 records to evaluate attacks on data from the same distribution (DS_D) but not in DS_T. Among the 45222 (35222) records, 21639 (16893) individuals are *married* (i.e., sensitive attribute $x_1 = married$), i.e., 47.8% (47.9%) positive class (Table 2). To ensure consistency, we evaluate all attacks against the target models trained on the DS_T dataset. The '*relationship*' attribute in this dataset (values: husband, wife, unmarried) is directly related to the *marital status* sensitive attribute. Hence, for the attack setup practicality, we have removed the '*relationship*' attribute from this dataset since otherwise the adversary could perform a straightforward attack: if relationship is husband or wife, the individual is married, otherwise, the individual is single.

Fivethirtyeight [27]: This dataset is from a survey conducted by the Fivethirtyeight Datalab, also used in FJRMIA [18]. 553 individuals were surveyed on a variety of questions. This dataset is used to train a model that predicts how an individual would like their steak prepared. In order to evaluate the cases of estimating multivalued and multiple sensitive attributes, we consider two sensitive attributes in this dataset: which age-group an individual belongs to (multivalued, {18-29, 30-44, 45-60, > 60}) and whether an individual drinks alcohol (binary, {yes,no}). Removing the data records missing either the sensitive attributes or the model output results in 331 data records. Due to its small size, we do not split this dataset further. Among 331 individuals, 266 answered yes to the question on drinking alcohol, i.e., 80.3% positive class (Table 2). The age-group marginal prior distribution is {21.1%, 28.1%, 26%, 24.8%}, respectively.

5.2 Machine Learning Models

To ensure a fair comparison with [18] which uses decision tree models, we first trained decision tree (DT) target models on the three datasets mentioned in Section 5.1. To further demonstrate the generalizability of our attacks, we also trained deep neural network (DNN) target models. However, we do not use the DNN model trained on the Fivethirtyeight dataset as the model's performance is very poor due to the small training set size. The confusion matrices of all the trained models are given in Tables 4, 5, 6, 7, 8, and 9. Since our attacks are black-box, the underlying model architecture does not make any difference in our attacks' algorithms. We leverage BigML [33], an ML-as-a-service system, and use its default configurations (1-click supervised training feature) to train these target mod-

els. The DT target models use BigML's memory tree optimization algorithm and smart pruning technique. Each DNN target model has 3 hidden layers and uses ADAM [34] as the optimization algorithm with a learning rate of 0.005. The attack models of LOMIA are trained using BigML's *ensemble* training algorithm with default configurations, i.e., decision forest algorithm and smart pruning technique. Table 3 shows the sizes of the $DS_\mathcal{A}$ datasets obtained from different target models along with the number of instances with the correctly labeled sensitive attribute in $DS_\mathcal{A}$.

Table 4: Confusion matrix of DT target model trained on GSS dataset.

Predicted / Actual	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 5: Confusion matrix of DNN target model trained on GSS dataset.

Predicted / Actual	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 6: Confusion matrix of DT target model trained on Adult dataset.

Predicted / Actual	<=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 7: Confusion matrix of DNN target model trained on Adult dataset.

Predicted / Actual	<=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 8: Confusion matrix of DT target model trained on FiveThirtyEight dataset.

Predicted / Actual	Medium	Medium Well	Medium Rare	Rare	Well	Total	Recall
Medium	105	0	3	0	1	109	96.33%
Medium Well	0	55	1	0	0	56	98.21%
Medium Rare	3	1	122	1	1	128	95.31%
Rare	0	1	0	17	0	18	94.44%
Well	0	0	0	0	20	20	100%
Total	108	57	126	18	22	331	Avg. Rec. 96.9%
Precision	97.2%	96.5%	96.8%	94.4%	90.9%	Avg. Prec. 95.2%	Acc. 96.4%

Table 9: Confusion matrix of DNN target model trained on FiveThirtyEight dataset.

Predicted / Actual	Medium	Medium Well	Medium Rare	Rare	Well	Total	Recall
Medium	9	0	95	5	0	109	8.26%
Medium Well	10	0	42	4	0	56	0.00%
Medium Rare	12	0	104	11	1	128	81.25%
Rare	2	0	15	1	0	18	5.56%
Well	3	0	13	4	0	20	0.00%
Total	36	0	269	25	1	331	Avg. Rec. 19%
Precision	25%	0%	38.7%	4%	0%	Avg. Prec. 13.5%	Acc. 34.4%

5.3 Attack Performance Metrics

As described in Section 3, along with precision, recall, accuracy, and F1 score, we also use *G-mean* and *MCC* metrics to evaluate our attacks on binary sensitive attributes as well as to compare their performances with that of the FJRMIA [18] and the baseline attacks (NaiveA and RandGA). We discuss the false positive rates (FPR) of the attacks in Section 5.5. In order to evaluate the proposed and existing attacks on multivalued sensitive attributes, we compute and compare the confusion matrices of the attacks as shown in Section 5.4.3.

We also evaluate the number of queries performed to the target model by the FJRMIA, CSMIA, and LOMIA strategies. Section A.2 in the Appendix presents the details of this comparison. Note that, while the CSMIA extension for partial knowledge of non-sensitive attributes suffers from a combinatorial explosion and make significantly more queries to the target model (Appendix A.3), the LOMIA strategy in the cases of partial knowledge of non-sensitive attributes does not require any extra query to the target model (see Section 4.4.2).

5.4 New Model Inversion Attacks' Results and Comparison with Baseline Attacks

In this section, we compare CSMIA and LOMIA with existing FJRMIA [18], and also with baseline attack strategies that do not require access to the target model, i.e., NaiveA and RandGA. As described in Section 3, the goal behind comparing with NaiveA and RandGA 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. We pay special attention to the Case (1) instances and analyze the LOMIA performance on them separately.

In RandGA, always predicting the positive class would result in 100% recall and thus a high F1 score but a G-mean of 0%. Therefore, for all the experiments in the following, RandGA predicts the positive class with a 0.5 probability, thus maximizing G-mean at 50% and ensuring a recall of 50%.

5.4.1 GSS Dataset

Figures 2(a) and 2(b) show the performances of the proposed attacks against the DT and DNN target models trained on the GSS dataset, respectively, and present a comparison with FJRMIA, NaiveA, and RandGA. Table 12 in Appendix shows the details of the metrics along with the TP, TN, FP, and FN values. Since the sensitive attribute in this dataset has an unbalanced distribution, the NaiveA strategy, also mentioned in [18], predicts the sensitive attribute as *no* for all the individuals and achieves an accuracy of 80.2%. However, the precision, recall, F1 score, G-mean, and MCC are all 0% as shown in Figures 2(a) and 2(b). Note that, NaiveA performance is independent of the target ML model type.

As demonstrated in Figure 2(a), the FJRMIA [18] achieves a very low recall and thus low F1 score. This is due to the fact

(a) DT model, GSS dataset (b) DNN model, GSS dataset (c) DT model, Adult dataset (d) DNN model, Adult dataset
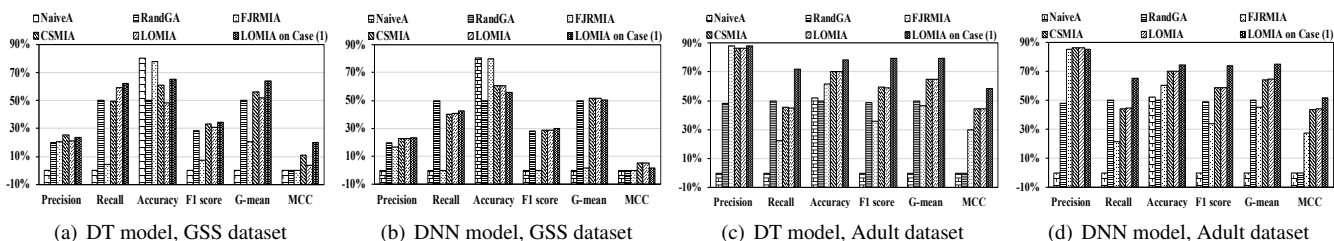
Figure 2: Comparison of attacks: FJRMIA [18], CSMIA, and LOMIA with baseline attack strategies NaiveA and RandGA.

Table 10: Attacks against DT target model trained on FiveThirtyEight dataset to infer multivalued 'age' sensitive attribute

(a) Attack confusion matrix of FJRMIA

Actual \ Predicted	18-29	30-44	45-60	>60	Total	Recall
18-29	0	70	0	0	70	0%
30-44	0	93	0	0	93	100%
45-60	0	86	0	0	86	0%
>60	0	82	0	0	82	0%
Total	0	331	0	0	331	Avg. rec. 25%
Precision	0%	28.1%	0%	0%	Avg. prec. 7.02%	Accuracy 28.1%

(b) Attack confusion matrix of CSMIA

Actual \ Predicted	18-29	30-44	45-60	>60	Total	Recall
18-29	40	9	8	13	70	57.14%
30-44	13	49	12	19	93	52.69%
45-60	15	17	36	18	86	41.86%
>60	11	19	21	31	82	37.8%
Total	79	94	77	81	331	Avg. rec. 47.37%
Precision	50.63%	52.13%	46.75%	38.27%	Avg. prec. 46.95%	Accuracy 47.13%

(c) Attack confusion matrix of LOMIA

Actual \ Predicted	18-29	30-44	45-60	>60	Total	Recall
18-29	41	20	9	0	70	58.57%
30-44	21	50	18	4	93	53.76%
45-60	28	24	32	2	86	37.21%
>60	30	30	12	10	82	12.2%
Total	120	124	71	16	331	Avg. rec. 40.43%
Precision	34.17%	40.32%	45.07%	62.5%	Avg. prec. 45.51%	Accuracy 40.18%

(d) Attack confusion matrix of LOMIA (Case 1)

Actual \ Predicted	18-29	30-44	45-60	>60	Total	Recall
18-29	21	0	0	0	21	100%
30-44	0	23	0	0	23	100%
45-60	1	0	19	1	21	90.48%
>60	1	0	0	9	10	90%
Total	23	23	19	10	75	Avg. rec. 95.12%
Precision	91.3%	100%	100%	90%	Avg. prec. 95.33%	Accuracy 96%

that the FJRMIA [18] relies on the marginal prior of the sensitive attribute while performing the attack. Since the sensitive attribute in the GSS dataset is unbalanced, the FJRMIA [18] mostly predicts the negative sensitive attribute (i.e., the individual didn't watch any x-rated movie, marginal prior ~ 0.8) and rarely predicts the positive sensitive attribute (i.e., the individual watched x-rated movies, marginal prior ~ 0.2). In contrast, our proposed CSMIA and LOMIA strategies achieve significantly high recall, F1 score, G-mean, and MCC while also improving precision. The FJRMIA [18] performs better only in terms of accuracy. However, note that the NaiveA also achieves an accuracy of 80.2%, the highest among all attacks, but there is no attack efficacy (0 true positive, see Table 12). Our attacks also consistently outperform RandGA in terms of all metrics. We emphasize that the records that belong to Case (1) are more vulnerable to model inversion attacks.

It is noteworthy that the LOMIA strategy performs similarly to CSMIA despite having access to only the predicted labels. Unlike CSMIA, the LOMIA strategy does not have cases and uses a single attack model for all the target records.

As shown in Figure 2(b), the FJRMIA [18] strategy again achieves a high accuracy but an extremely low recall. It performs almost like NaiveA with only 1 true positive and 5 false positives (see Table 12). The RandGA strategy has the same results as Figure 2(a) since this strategy is independent of the target model (similar to NaiveA). Our attacks' performances against this model are not significantly better than RandGA, even the LOMIA results on Case (1) are not significant. There-

fore, it may seem that according to the overall performance, the DNN model trained on the GSS dataset may not be vulnerable to model inversion attacks since the RandGA attack even without access to the model may achieve comparable performances. 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 records (e.g., individuals grouped by race, gender) in the dataset could still be vulnerable. We discuss such discrimination in performances of model inversion attacks later in Section 5.7.

5.4.2 Adult Dataset

Figure 2(c) shows the performances of the attacks against the DT target model trained on the Adult dataset. The results for the DNN target model as shown in Figure 2(d) are very similar to that of Figure 2(c). Table 13 in Appendix shows the details along with the TP, TN, FP, and FN values. 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%. FJRMIA [18] results in a precision comparable to our attacks but achieves much less in terms of the other metrics. Our attacks also significantly outperform RandGA in terms of all metrics except the recall.

Observing the results of the proposed attacks and also the

performance against Case (1) instances, we conclude that *releasing the models trained on the Adult dataset* would add a significant advantage to the adversary in terms of learning the 'marital status' sensitive attribute. This is because all our proposed attacks that query the target models perform significantly better when compared to the NaiveA and RandGA adversaries that do not need any access to the model.

Overall, the attacks demonstrate more effectiveness against the target models trained on the Adult dataset than against the target models trained on the GSS dataset. Therefore, we investigated if the correlations between the sensitive attributes and the corresponding target models trained on these datasets (in other words, the importance of the sensitive attributes in the target models) differ significantly. According to our observation, this is not the case. For instance, the importance of the 'x-rated-movie' and 'marital-status' sensitive attributes in their corresponding DT target models are 7.3% and 9.6%, respectively. Figures 7(b) and 7(c) in Appendix show the importance of all attributes in these models.

5.4.3 FiveThirtyEight Dataset

In this section, we perform two sets of attack experiments against the DT target model trained on the FiveThirtyEight dataset: (i) inferring multivalued sensitive attribute *age-group*, when all other non-sensitive attributes are known to the adversary, and (ii) inferring both *alcohol* and *age-group*, i.e., the case of estimating multiple sensitive attributes.

(i) Estimating Multivalued Sensitive Attributes: Tables 10 (a), (b), and (c) show the performances of the FJRMIA, CSMIA, and LOMIA strategies, respectively, in terms of estimating a multivalued sensitive attribute, i.e., *age* in the FiveThirtyEight dataset. FJRMIA [18] predicts the age-group $30 - 44$ for all the target records (i.e., it boils down to NaiveA, age-group $30 - 44$ has the highest marginal prior among all, 28.1%). Also, the RandGA strategy would achieve maximum accuracy of 25% in estimating this multivalued sensitive attribute (not shown in tables). In contrast, our proposed CSMIA and LOMIA strategies achieve significantly better results. The results in Table 10 (d) show the performance of LOMIA on Case (1) instances which has an accuracy of 96%. Hence, we emphasize that the records in Case (1) are significantly more vulnerable to model inversion attacks.

(ii) Estimating Multiple Sensitive Attributes: In this attack setting, the adversary estimates both the age-group and alcohol sensitive attributes of a target individual. The attack results for estimating the multivalued age-group attribute, in this case, are similar to that of Table 10. Due to space constrains, the attack results for estimating the binary attribute alcohol are given in Table 15 in Appendix. Also, we demonstrate the results of the FJRMIA, CSMIA, LOMIA, and LOMIA Case (1) in terms of estimating the age-group attribute in Table 16.
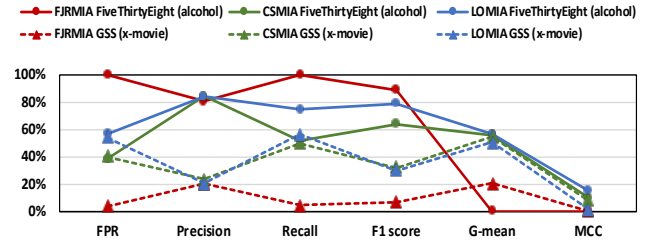

Figure 3: Comparison among different attack strategies in terms of FPR and other metrics. Marginal priors: alcohol positive class 80.3%, x-movie positive class 19.8%.

5.5 False Positive Rates and Attack Stability

In order to demonstrate the false positive rate (FPR) comparison between our proposed attacks and the existing FJRMIA [18] strategy, we perform experiments with two scenarios: (1) estimating the 'alcohol' sensitive attribute in the FiveThirtyEight dataset which has 80.3% positive class marginal prior (i.e., alcohol=yes), and (2) estimating the 'x-movie' sensitive attribute in the GSS dataset which has only 19.8% positive class marginal prior (i.e., x-movie=yes), see Table 2. Figure 3 shows the comparison among FJRMIA, CSMIA, and LOMIA in terms of FPR and other metrics. The solid lines represent the attack performances of estimating alcohol in the FiveThirtyEight dataset whereas the dashed lines represent the attack performances of estimating x-movie in the GSS dataset. Since FJRMIA is heavily dependent on the marginal priors of the sensitive attributes, it achieves extreme FPRs in these two scenarios: 100% FPR in estimating alcohol and 4.17% FPR is estimating x-movie. In contrast, our proposed attacks are more stable and their superior performances in both scenarios are evident by the G-mean and MCC metrics in Figure 3. The comparison of these attacks' FPRs for the Adult dataset where the sensitive attribute is more balanced is given in Table 13. The FPRs of our proposed attacks are comparable to that of FJRMIA ($\sim 6\%$ vs. $\sim 3\%$). However, our attacks outperform FJRMIA in terms of other metrics as shown in Figures 2(c) and 2(d). Also, note that, lower FPR may not always indicate a better attack, e.g., NaiveA has an FPR of 0% but the attack has no efficacy.

In the following, we characterize the datasets for which our proposed attacks may have higher FPR and vice versa. In a dataset, if the sensitive attribute's positive class has a high marginal prior, our proposed attacks achieve lower FPRs than FJRMIA (e.g., FiveThirtEight dataset in Figure 3). On the other hand, if the sensitive attribute has lower positive class marginal prior, our attacks result in higher FPRs than FJRMIA (e.g., GSS dataset in Figure 3). Finally, if the sensitive attribute has a more balanced marginal prior, our attacks achieve FPRs comparable to FJRMIA (e.g., Adult dataset in Table 13). However, irrespective of the marginal priors, for different datasets, our attacks consistently outperform FJRMIA in terms of G-mean and MCC metrics.
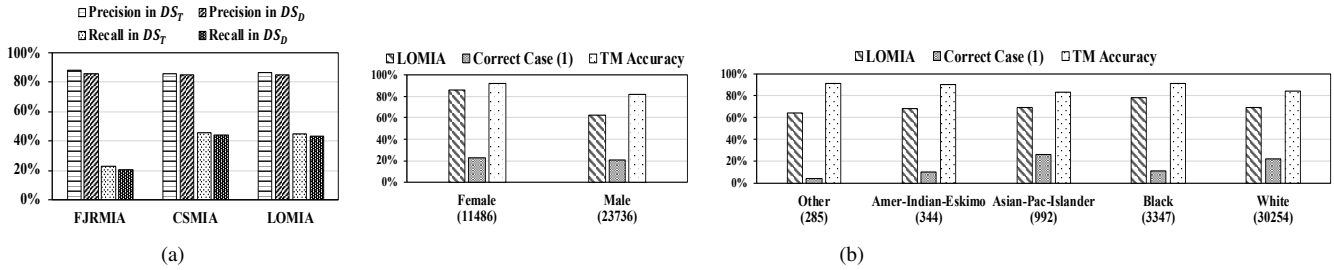
Figure 4: (a) Privacy leakage for DS_T and DS_D, (b) disparate vulnerability of LOMIA for different gender and race groups.

5.6 Distributional Privacy Leakage

In order to investigate if our MIAI attacks also breach the privacy of data that is not in the training dataset of the target model but is drawn from the same distribution, we evaluate our attacks on the corresponding DS_D datasets as described in Section 5.1. Figure 4(a) compares the performance of our attacks as well as the performance of FJRMIA on the DT model trained on the Adult dataset. Our observation shows that our attacks are equally effective against the records in the training dataset (DS_T) and the records outside of the training dataset but drawn from the same distribution (DS_D). We observe similar trends in the proposed attacks against other target models as shown in Figure 8 in Appendix.

5.7 Disparate Vulnerability of MIAI Attacks

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 these attacks than others, it raises serious privacy concerns for that particular group.

Figure 4(b) shows the contrast in the performances of LOMIA against different gender and race populations. The attack is performed against the DNN model trained on the Adult dataset. The x-axis represents gender/race identities along with the number of records in the training dataset that belong to the particular subgroups. For instance, the numbers of female and male individuals in the Adult dataset are $11,486$ and $23,736$, respectively. According to our observation, LOMIA predicts correct marital status for 85.9% of the female population whereas it predicts correct marital status for only 62.4% of the male population. LOMIA also shows disparate attack performance against different race groups and is most successful against the Black race subgroup with 78.2% accuracy. Since the attack model of LOMIA is trained on $DS_{\mathcal{A}}$ dataset obtained from the Case (1) instances, we investigated what percentage of records of each of the female and male subgroups are labeled with correct sensitive attributes in $DS_{\mathcal{A}}$ dataset and if that has any impact on such disparate vulnerability. However, we observe that around a similar percentage ($\sim 21\%$) of both female and male records, i.e., 2593 and 4837, respectively, are labeled with the correct sensitive attribute

(single/married) in the $DS_{\mathcal{A}}$ dataset, which is shown using Correct Case (1) bar in Figure 4(b). We also investigated if the accuracy of the target model for different subgroups plays a role in disparate vulnerability, shown using the TM Accuracy bar in Figure 4(b). We observe that the target model is 92.4% accurate for the female population and only 81.4% accurate for the male population in predicting their income, which correlates with the disparate vulnerability. However, we have not observed this correlation consistently, e.g., in the case of disparate vulnerability for race subgroups. LOMIA shows disparate vulnerability against other subgroups, such as religions (DT model trained on GSS dataset) and occupations (DNN model trained on Adult dataset). The results are demonstrated in Appendix (see Figures 9 and 10, respectively). Note that, we have observed disparate vulnerability across all datasets and models but reported the most interesting results only.

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 our findings here show only a few instances of such disparity in the model inversion attack performances on different groups, this is a potentially serious issue and needs to be further investigated. Otherwise, while it may seem that the attack performance on the overall dataset is not a significant threat, some specific groups in the dataset could still remain significantly vulnerable to MIAI attacks.

5.8 Attack Results With Partial Knowledge of Target Record's Non-sensitive Attributes

Figure 5(a) shows the performance details of LOMIA against the DT model trained on the Adult dataset when 1-9 non-sensitive attributes (NSA) increasingly become unknown (u) to the adversary in the following order: work-class, sex, race, fnlwgt, occupation, education, hours-per-week, capital-gain, and capital-loss. This order reflects the importance of the Adult dataset attributes in the LOMIA attack model trained against the DT target model (see Figure 6). Since the 'income' attribute occupies 90.4% importance in the LOMIA attack model, the unavailability of 9 other non-sensitive attributes does not degrade the performance of LOMIA. Figure 5(b)
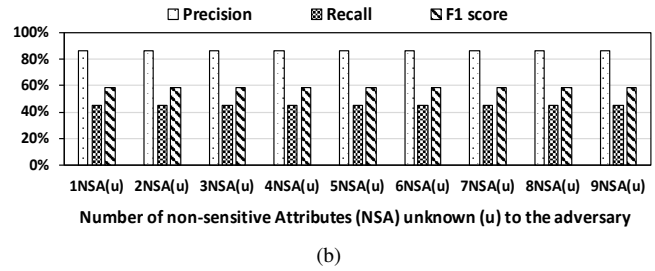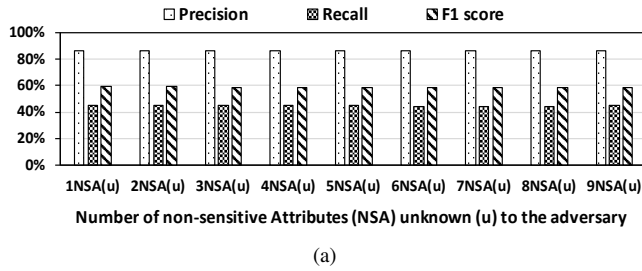
Figure 5: LOMIA performance against the (a) DT and (b) DNN target models trained on Adult dataset when 1-9 non-sensitive attributes (NSA) are unknown (u) to the adversary.
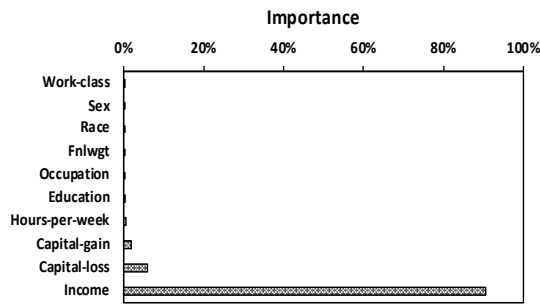

Figure 6: Adult dataset attributes' importance in the LOMIA attack model trained against the DT target model.

Table 11: Attack performance against the DT target model trained on Adult dataset.

Target model class label	Attack Strategy	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score
<=50K	FJRMIA [18]	13	17108	13	9315	50%	0.14%	64.73%	0.28%
	CSMIA	127	17018	103	9201	55.22%	1.36%	64.82%	2.66%
	LOMIA	26	17085	36	9302	41.94%	0.28%	64.69%	0.55%
>50K	FJRMIA [18]	3775	710	498	3790	88.34%	49.9%	51.12%	63.78%
	CSMIA	7537	67	1141	28	86.85%	99.63%	86.68%	92.8%
	LOMIA	7548	47	1161	17	86.67%	99.78%	86.57%	92.76%

shows a similar performance of LOMIA against the DNN model trained on the Adult dataset. Note that, the income attribute occupies 100% importance in the LOMIA attack model trained against the DNN target model. Hence, we use the same order shown in Figure 6 to make 1-9 non-sensitive attributes (NSA) increasingly unknown (u) to the adversary. We have observed similar LOMIA results against the target models trained on GSS datasets. Figure 11 in Appendix shows the importance of the GSS dataset attributes in the LOMIA attack model against the DT target model and the corresponding performance details of LOMIA when 9 non-sensitive attributes (NSA) increasingly become unknown. Figure 12 shows the same for the DNN target model trained on the GSS dataset.

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 record. Due to space constraints, the performance details of the CSMIA partial knowledge attack have been discussed in Appendix A.3.

5.9 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 DT model trained on the Adult dataset.

Table 11 shows a comparison among FJRMIA [18], CSMIA, and LOMIA performances for different class labels of the target model. Note that, the attack performances are significantly different for the two class labels, e.g., the recall values of identifying 'married' individuals in class <=50K are significantly low when compared to the recall values of identifying 'married' individuals in class >50K. The precision values also demonstrate disparate attack performances on these two target model class labels.

5.10 Discussion and Limitations

5.10.1 Discussion

To our knowledge, ours is the first work that studies MIAI attacks in such details on *tabular* datasets which is the most common data type used in real-world machine learning [28]. We discuss some of our notable findings in the following:
TIR vs. MIAI: 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 [22]. However, we argue that such is *not* the case for MIAI attacks. Table 6 shows the confusion matrix for the DT model trained on Adult dataset. From the matrix, it is evident that the target model's performance (both precision and recall) is better for class label $<= 50K$ than that of for class label $> 50K$. If the root causes of MIAI attacks were similar to that of TIR attacks, the attacks would be more effective against the records of class label $<= 50K$. On the contrary, in Section 5.9, we demonstrate that the MIAI attacks (both existing and proposed) perform better against the records of class label $> 50K$.
Importance of sensitive attribute in target model: As discussed in Section 5.4.2, the importance of sensitive attributes

in the corresponding target models trained on GSS and Adult datasets do not differ significantly whereas the proposed MIAI attacks against target models trained on the Adult dataset are significantly 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. We identify the difference in the distribution of sensitive attributes in these datasets (Adult dataset 47.9% positive class vs. GSS dataset 19.8% positive class) as a factor that has contributed to this attack performance difference. We leave investigating this and other factors to future work.

Disparate vulnerability: We have investigated correct Case (1) percentage and target model accuracy for different subgroups as possible factors behind disparate vulnerability. It is evident that further investigation is required to better understand the disparate impact on different groups of records which is a serious threat of model inversion attacks.

Distributional privacy breach: Existing research [19, 22] shows that differential privacy (DP)-based defense mechanisms against model inversion attacks suffer from significant loss of model utility. Moreover, DP mechanisms provide privacy guarantees to only the training data records. In contrast, our experiments show that model inversion attacks not only breach the privacy of sensitive training datasets but also leak distributional privacy. Therefore, the effectiveness of DP mechanisms against model inversion attacks needs further investigation.

5.10.2 Limitations

Attribute inference attack is not a real threat when a dataset has a lot of attributes, since the model prediction is likely to depend very little on each individual attribute. Therefore, in this paper, we study the MIAI attacks only on datasets with fewer attributes. Also, in certain datasets, where the sensitive attribute has lower positive class marginal prior, our attacks result in high FPRs. We leave the investigation of designing improved model inversion attacks that ensure marginal prior agnostic lower FPRs to future work.

6 Related Work

In [19], Fredrikson et al. introduced the concept of model inversion attacks and applied their attack to linear regression models. In [18], 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. In the second setting, the adversary aims to reconstruct instances similar to ones in the training dataset using gradient descent. As mentioned earlier, we focus on the first one, i.e., attribute inference attack. Subsequently, Wu

et al. [35] 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 [23, 24, 36–40] and recommender systems [41, 42]. In these attacks, the adversary first trains a machine learning classifier that takes as input the public attributes and then outputs the private attributes. However, in order to build such a classifier, these attacks [23, 24, 36–41] have to rely on the users who also share their private attributes (e.g., gender, political views, locations visited) along with their public attributes (e.g., pages liked, movie ratings). Therefore, the adversary's machine learning classifier can be built only in those scenarios where it can collect the private-public attribute pairs of real users. In contrast to the adversaries assumed in these attacks [23, 24, 36–41], the adversaries assumed in our attacks are *not* assumed to be able to obtain a dataset from the same population the DS_T dataset has been obtained from. This is because in many scenarios such an assumption (adversary having access to a similar dataset) may not be valid. Therefore, while designing our attacks, it has been part of our goal to incorporate these practical scenarios in our attack surface so that our proposed attacks could be applied more widely.

Most of the work mentioned above assumes that the attributes of a target individual, except the sensitive attribute, are known to the adversary. Hidano et al. [43] proposed a method to infer the sensitive attributes without the knowledge of non-sensitive attributes. However, unlike our model inversion attacks, 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 the target individual's non-sensitive attributes does not require poisoning and performs similar to scenarios where the adversary has full knowledge of the target individual's non-sensitive attributes.

7 Conclusion and Future Work

In this paper, we demonstrate two new black-box model inversion attacks: (1) confidence score-based attack (CSMIA) and (2) label-only attack (LOMIA). We perform an extensive evaluation of our attacks using two types of ML models, decision tree and deep neural network, that are trained with three real datasets [25–27]. Our evaluation results show that the proposed attacks significantly outperform the existing ones. Moreover, we empirically show that model inversion attacks have disparate vulnerability property. We also evaluate the risks incurred by model inversion attacks when the adversary does not have the knowledge of all other non-sensitive attributes of a target record and demonstrate that our attacks' performance is not impacted significantly in those scenarios. Designing effective defense methods that protect privacy against our label-only MIAI attack without degrading the target model's performance is left as future work.

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

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

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

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

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

[8] 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.

[9] 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.

[10] Amazon geteventprediction api. https://docs.aws. amazon.com/frauddetector/latest/ug/real-time-fraud-prediction.html.

[11] Google text-to-speech api. https://cloud.google.com /text-to-speech/docs/reference/rest/.

[12] Face++ emotion api. https://www.faceplusplus.com /emotion-recognition/.

[13] Microsoft speech api. https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/.

[14] Google nlp api. https://cloud.google.com/natural-language.

[15] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P. Wellman. Sok: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS P)*, pages 399–414, 2018.

[16] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017.

[17] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *25th USENIX Security Symposium*, pages 601–618, 2016.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

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

[24] 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.

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

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

[27] Walt Hickey. Fivethirtyeight.com data-lab: How americans like their steak. https://fivethirtyeight.com/features/how-americans-like-their-steak/.

[28] Yuanfei Luo, Hao Zhou, Wei-Wei Tu, Yuqiang Chen, Wenyuan Dai, and Qiang Yang. Network on network for tabular data classification in real-world applications. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2317–2326, 2020.

[29] 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.

[30] 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.

[31] Aol dataset. https://www.kaggle.com/dineshydv/aol-user-session-collection-500k.

[32] How does bigml handle missing values to predict with your models and ensembles? https://support.bigml.com/hc/en-us/articles/206616349-How-does-BigML-handle-missing-values-to-predict-with-your-models-and-ensembles-.

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

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

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

[40] 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.

[41] 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.

[42] 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.

[43] 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.
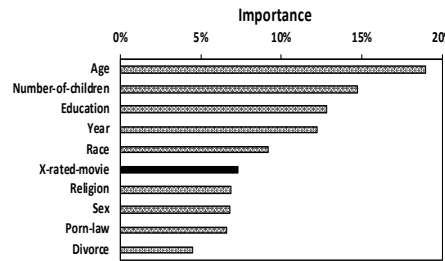
A Appendix

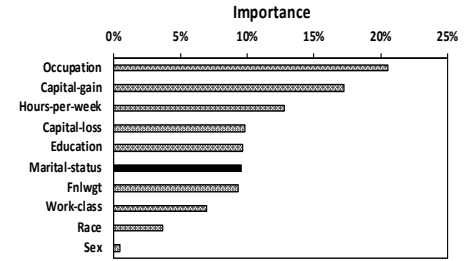## A.1 Random Guessing Attack Performances

In this attack, the adversary randomly predicts the sensitive attribute by setting a probability for the positive class sensitive attribute value. Fig. 7(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
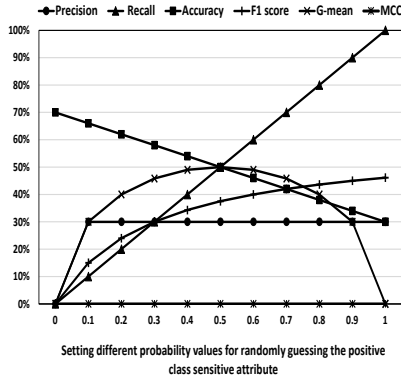
(a) Positive class marginal prior 0.3. (b) GSS dataset attributes' importance (c) Adult dataset attributes' importance

Figure 7: (a) RandGA performance. Importance of (b) GSS and (c) Adult dataset attributes in their respective DT target models.

when the attack randomly assigns the 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 the 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%. The MCC of the random guessing attacks is always 0.

A.2 Comparison of Attacks' Query Numbers

In FJRMIA, the adversary predicts sensitive attribute value from all possible values based on the score $C[y,y'] * p_{1,i}$, i.e., the attack selects the sensitive attribute value that maximizes this score (see section 2.4). Since for each instance in the training dataset the adversary queries k times (k is the number of possible values of the sensitive attribute), it requires $k*n$ queries in total, where n is the number of instances in the training dataset. Our CSMIA strategy performs queries similarly, i.e., during the attack, but instead of comparing $C[y,y'] * p_{1,i}$ for different sensitive attribute values, it compares the confidence scores returned by queries (see section 4.1). Therefore,

both FJRMIA and CSMIA require $k*n$ queries while attacking the entire training dataset. For LOMIA, the number of queries to the target model is the same as CSMIA and FJRMIA to perform the attack on the entire training dataset. However, these queries are performed to generate the Case (1) attack dataset and to train the attack model rather than while performing the attack as in CSMIA and FJRMIA. Once the attack model is built, the attacker in LOMIA does not need to query the target model anymore, it queries the attack model instead. We present the target model query numbers for different attacks on different datasets in Table 14.

For example, GSS DS_T dataset has 15235 instances ($n = 15235$) and sensitive attribute x_movie has two possible values ($k = 2$). Therefore, the total number of queries for all attacks in this dataset is $k*n = 15235 \times 2 = 30470$. The total number of queries for estimating single sensitive attributes in Adult (i.e., *marital-status*) and FiveThirtyEight datasets (i.e., *alcohol* or *age-group*) are calculated similarly. For multiple sensitive attribute inference, i.e., estimating age-group and alcohol in the FiveThirtyEight dataset, we consider one sensitive attribute to be missing [32] and query the target model with all possible values of the other sensitive attribute. Therefore, the total number of queries while simultaneously estimating age-group

Table 12: Attack performance against the DT and DNN target models trained on GSS dataset.

Target Model	Attack Strategy	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC	FPR
DT/DNN	NaiveA	0	12218	0	3017	0%	0%	80.2%	0%	0%	0%	0%
DT	FJRMIA [18]	131	11709	509	2886	20.47%	4.34%	77.72%	7.16%	20.39%	0.3%	4.17%
DT	CSMIA	1490	7844	4373	1528	25.41%	49.37%	61.27%	33.55%	56.3%	11.1%	35.79%
DT	LOMIA	1782	5565	6653	1235	21.13%	59.07%	48.22%	31.12%	51.87%	3.7%	54.45%
DNN	FJRMIA [18]	1	12213	5	3016	16.67%	0.03%	80.17%	0.07%	1.82%	−0.2%	0.04%
DNN	CSMIA	1212	8058	4160	1805	22.56%	40.17%	60.85%	28.89%	51.47%	5.1%	34.05%
DNN	LOMIA	1225	8015	4203	1792	22.57%	40.6%	60.65%	29.01%	51.61%	5.16%	34.4%

Table 13: Attack performance against the DT and DNN target models trained on Adult dataset.

| Target Model	Attack Strategy	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC	FPR
DT/DNN	NaiveA	0	18329	0	16893	0%	0%	52.04%	0%	0%	0%	0%
DT	FJRMIA [18]	3788	17818	511	13105	88.11%	22.42%	61.34%	35.75%	46.69%	29.9%	2.79%
DT	CSMIA	7664	17085	1244	9229	86.04%	45.37%	70.27%	59.41%	65.03%	44.3%	6.79%
DT	LOMIA	7574	17132	1197	9319	86.35%	44.84%	70.14%	59.02%	64.74%	44.3%	6.53%
DNN	FJRMIA [18]	3592	17717	612	13301	85.44%	21.26%	60.5%	34.05%	45.34%	27.6%	3.34%
DNN	CSMIA	7490	17139	1190	9403	86.29%	44.34%	69.93%	58.58%	64.39%	43.9%	6.49%
DNN	LOMIA	7565	17121	1208	9328	86.23%	44.78%	70.09%	58.95%	64.68%	44.2%	6.59%

(a)   (b)   (c)
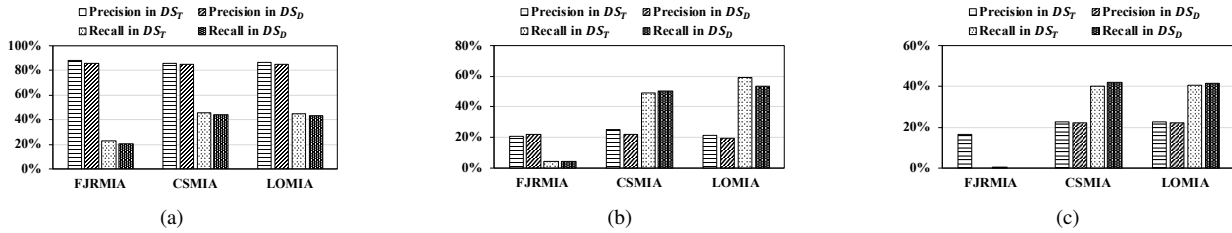
Figure 8: Privacy leakage for DS_T and DS_D: against (a) DNN target model trained on Adult dataset, (b) DT target model trained on GSS dataset, and (c) DNN target model trained on GSS dataset.
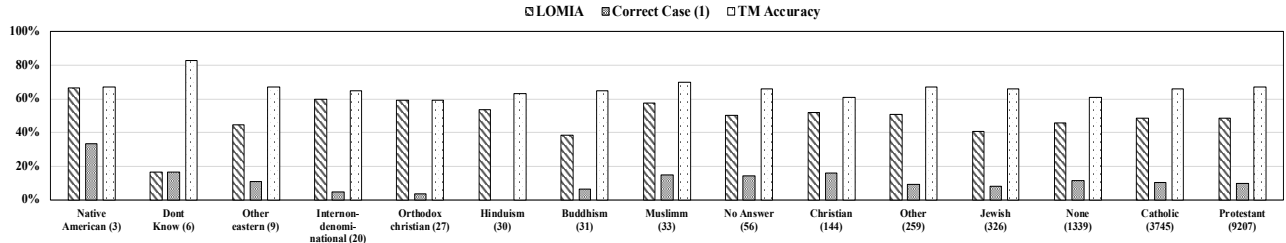

Figure 9: Disparate vulnerability of LOMIA for different religion groups (attack on the DT model trained on GSS dataset)
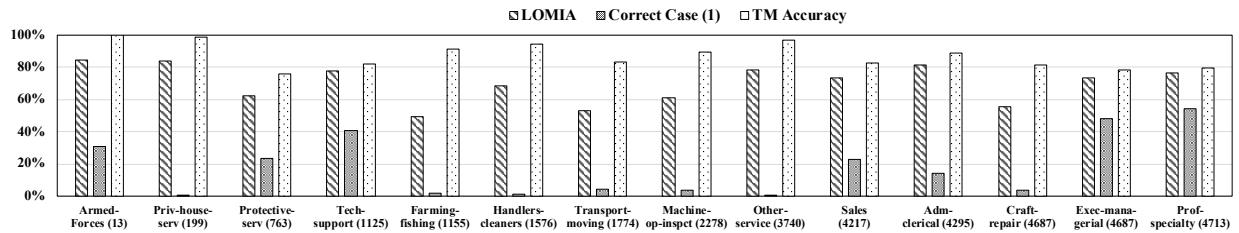

Figure 10: Disparate vulnerability of LOMIA for different occupation groups (attack on the DNN model trained on Adult dataset)

Table 14: Query Numbers for Different Attacks

Attack Strategy	GSS, (x-movie) Section 5.4.1	Adult, (marital-status) Section 5.4.2	Fivethirtyeight , (alcohol) Section 5.5	Fivethirtyeight, (age-group) Section 5.4.3 (i)	Fivethirtyeight (age-group & alcohol) Section 5.4.3 (ii)
FJRMIA	30470	70444	662	1324	1986
CSMIA	30470	70444	662	1324	1986
LOMIA	30470	70444	662	1324	1986

and alcohol sensitive attributes is 331x(4+2)=1986.

A.3 CSMIA Results With Partial Knowledge of Non-sensitive Attributes

Excluding the sensitive attribute ('marital status') and the output of the target model ('income'), we first 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. Figure 13 shows the performance of CSMIA on the DT 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 along with the number of queries to the target model. 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 Figure 13, we observe that the performance of our attack does not deteriorate and remains 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. 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 Figure 13, our attack does not show any significant deterioration. Due to the combinatorial complexity of our CSMIA partial knowledge attack and increasing number of target model queries, we limit the number of unknown non-sensitive attributes to two for these experiments.

Table 15: Inferring the sensitive attribute alcohol (alcohol=yes 80.3% marginal prior), attack performances against the DT target model trained on FiveThirtyEight dataset (adversary also estimates the age-group sensitive attribute).

Attack Strategy	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC
FJRMIA [18]	256	5	60	10	81.01%	96.24%	78.85%	87.97%	27.21%	7.51%
CSMIA	151	34	31	115	82.97%	56.77%	55.89%	67.41%	54.49%	7.25%
LOMIA	192	19	46	74	80.67%	72.18%	63.75%	76.19%	45.93%	1.25%

Importance

Divorce
Race
Religion
Sex
Education
Age
Year
Number-of-children
Porn-law
Hap-marriage

☐ Precision ⊠ Recall ▨ F1 score

1NSA(u) 2NSA(u) 3NSA(u) 4NSA(u) 5NSA(u) 6NSA(u) 7NSA(u) 8NSA(u) 9NSA(u)

Number of non-sensitive Attributes (NSA) unknown (u) to the adversary

(a)　　　(b)

Figure 11: (a) GSS dataset attributes' importance in the LOMIA attack model trained against the DT target model. (b) LOMIA performance against the DT model trained on GSS dataset when 1-9 non-sensitive attributes (NSA) are unknown (u) to the adversary in the following order: divorce, race, religion, sex, education, age, year, number-of-children, and porn-law.

Importance

Divorce
Year
Sex
Age
Number-of-children
Race
Religion
Porn-law
Education
Hap-marriage

☐ Precision ⊠ Recall ▨ F1 score

1NSA(u) 2NSA(u) 3NSA(u) 4NSA(u) 5NSA(u) 6NSA(u) 7NSA(u) 8NSA(u) 9NSA(u)

Number of non-sensitive Attributes (NSA) unknown (u) to the adversary
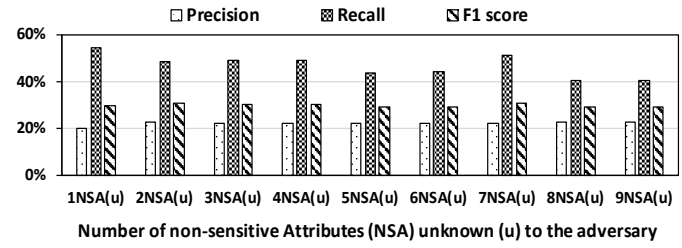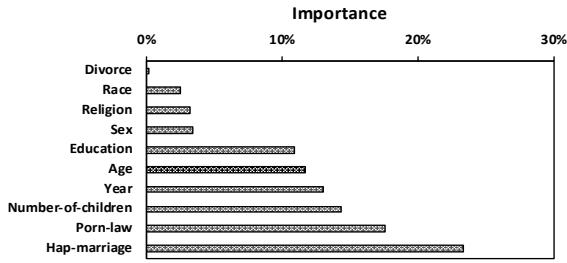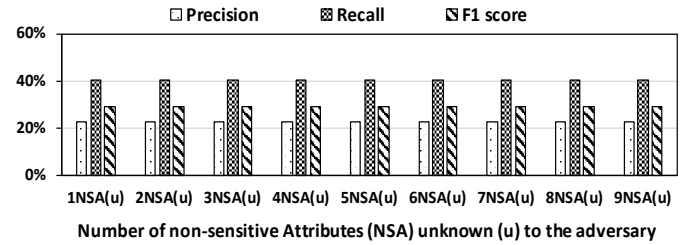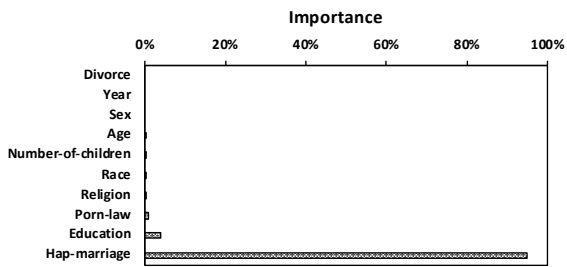
(a)　　　(b)

Figure 12: (a) GSS dataset attributes' importance in the LOMIA attack model trained against the DNN target model. (b) LOMIA performance against the DNN model trained on GSS dataset when 1-9 non-sensitive attributes (NSA) are unknown (u) to the adversary in the following order: divorce, year, sex, age, number-of-children, race, religion, porn-law, and education.
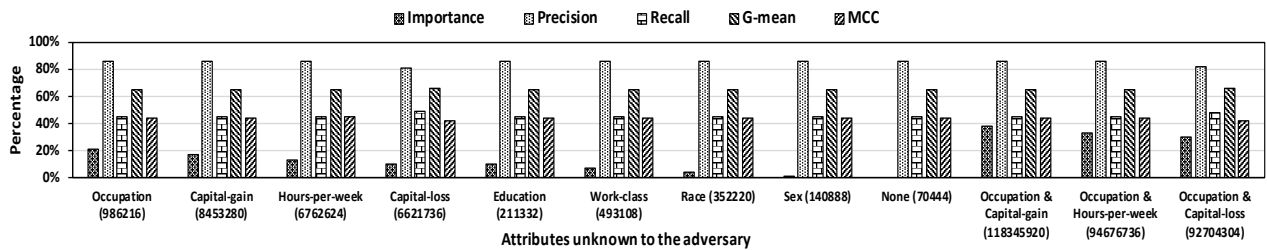
⊠ Importance ▤ Precision ☐ Recall ▨ G-mean ▧ MCC

Occupation (986216), Capital-gain (8453280), Hours-per-week (6762624), Capital-loss (6621736), Education (211332), Work-class (493108), Race (352220), Sex (140888), None (70444), Occupation & Capital-gain (118345920), Occupation & Hours-per-week (94676736), Occupation & Capital-loss (92704304)

Attributes unknown to the adversary

Figure 13: CSMIA performance against the DT model trained on Adult dataset when some of the other (non-sensitive) attributes of a target individual are also unknown to the adversary.

Table 16: Attacks against DT target model trained on FiveThirtyEight dataset to infer multivalued age-group sensitive attribute (along with inferring binary sensitive attribute alcohol)

(a) Attack confusion matrix of FJRMIA

Actual \ Predicted	18-29	30-44	45-60	>60	Total	Recall
18-29	0	64	0	6	70	0%
30-44	0	88	0	5	93	94.62%
45-60	0	84	0	2	86	0%
>60	0	77	0	5	82	6.1%
Total	0	313	0	18	331	Avg. recall 25.18%
Precision	0%	28.12%	0%	27.78%	Avg. precision 13.97%	Accuracy 28.1%

(b) Attack confusion matrix of CSMIA

Actual \ Predicted	18-29	30-44	45-60	>60	Total	Recall
18-29	35	12	7	16	70	0.5%
30-44	14	52	12	15	93	55.91%
45-60	16	14	36	20	86	41.86%
>60	16	24	17	25	82	30.49%
Total	81	102	72	76	331	Avg. recall 44.57%
Precision	43.21%	50.98%	50%	32.89%	Avg. precision 44.27%	Accuracy 44.71%

(c) Attack confusion matrix of LOMIA

Actual \ Predicted	18-29	30-44	45-60	>60	Total	Recall
18-29	33	23	13	1	70	47.14%
30-44	24	48	15	6	93	51.61%
45-60	19	29	33	5	86	38.37%
>60	21	34	15	12	82	14.63%
Total	97	134	76	24	331	Avg. recall 37.94%
Precision	34.02%	35.82%	43.42%	50%	Avg. precision 40.82%	Accuracy 38.07%

(d) Attack confusion matrix of LOMIA (Case 1)

Actual \ Predicted	18-29	30-44	45-60	>60	Total	Recall
18-29	15	1	0	0	16	93.75%
30-44	0	18	0	1	19	94.74%
45-60	1	0	16	2	19	84.21%
>60	1	1	0	6	8	75%
Total	17	20	16	9	62	Avg. recall 86.92%
Precision	88.24%	90%	100%	66.67%	Avg. precision 86.23%	Accuracy 88.71%