When Machine Learning Models Leak
An Exploration of Synthetic Training Data

Slokom, Manel; de Wolf, Peter Paul; Larson, Martha

DOI
10.1007/978-3-031-13945-1_20

Publication date
2022

Document Version
Final published version

Published in
Privacy in Statistical Databases - International Conference, PSD 2022, Proceedings

Citation (APA)
Slokom, M., de Wolf, P. P., & Larson, M. (2022). When Machine Learning Models Leak: An Exploration of Synthetic Training Data. In J. Domingo-Ferrer, & M. Laurent (Eds.), *Privacy in Statistical Databases - International Conference, PSD 2022, Proceedings* (pp. 283-296). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 13463 LNCS). Springer. https://doi.org/10.1007/978-3-031-13945-1_20

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

When Machine Learning Models Leak: An Exploration of Synthetic Training Data

Manel Slokom[1,2,3(✉)], Peter-Paul de Wolf[2], and Martha Larson[3]

[1] Delft University of Technology, Delft, The Netherlands
m.slokom@tudelft.nl
[2] Statistics Netherlands, The Hague, The Netherlands
pp.dewolf@cbs.nl
[3] Radboud University, Nijmegen, The Netherlands
m.larson@cs.ru.nl

Abstract. We investigate an attack on a machine learning classifier that predicts the propensity of a person or household to move (i.e., relocate) in the next two years. The attack assumes that the classifier has been made publically available and that the attacker has access to information about a certain number of target individuals. That attacker might also have information about another set of people to train an auxiliary classifier. We show that the attack is possible for target individuals independently of whether they were contained in the original training set of the classifier. However, the attack is somewhat less successful for individuals that were not contained in the original data. Based on this observation, we investigate whether training the classifier on a data set that is synthesized from the original training data, rather than using the original training data directly, would help to mitigate the effectiveness of the attack. Our experimental results show that it does not, leading us to conclude that new approaches to data synthesis must be developed if synthesized data is to resemble "unseen" individuals to an extent great enough to help to block machine learning model attacks.

Keywords: Synthetic data · Propensity to move · Attribute inference · Machine learning

1 Introduction

Governmental institutions charged with collecting and disseminating information may use machine learning models to produce estimates, such as imputing missing values or inferring variables that cannot be directly observed. When such estimates are published, it is also useful to publish the machine learning model itself, so that researchers using the estimates can evaluate it closely, or

The views expressed in this paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands.

J. Domingo-Ferrer and M. Laurent (Eds.): PSD 2022, LNCS 13463, pp. 283–296, 2022.
https://doi.org/10.1007/978-3-031-13945-1_20

even produce their own estimates. Moreover, society also asks for more insight into the models that are used, e.g., to address possible discrimination caused by decisions based on machine learning models.

Unfortunately, machine learning models can be attacked in a way that allows an attacker to recover information about the data set that they were trained on [19]. For this reason, publishing machine learning models can lead to a risk that information in the training set is leaked. In this paper, we carry out a case study of an attribute inference attack on a machine learning classifier to better understand the nature of the risk. The classifier that we study predicts *propensity to move*, i.e., whether an individual or household will relocate their home within the next two years. The attack scenario assumes that the classifier has been released to the public, and that an attacker wishes to learn a sensitive attribute for a group of victims, i.e., target individuals. The attacker has non-sensitive information about these target individuals that is used for the attack and has scraped information about other people from the Web.

Our experimental investigation first confirms that a machine learning classifier is able to predict propensity to move for individuals in its training data set as well as for previously "unseen" individuals, reproducing [3]. We then attack this classifier and demonstrate that an attacker can learn sensitive attributes both for individuals in the training data as well as for previously "unseen" individuals. However, for "unseen" individuals the attack is somewhat less successful. We reason that data synthesis might potentially allow us to create data that we could use for training and that would be far enough from the original data, than any real individual would have the somewhat higher resistance to attack of an "unseen" individual. Based on this idea, we create a synthetic training set, train a machine learning classifier on that set, and repeat the attacks. Interestingly, the resulting classifier is just as susceptible to attack as the original classifier, which was trained on the original data. We relate this finding to the success of an attack that infers sensitive information from individuals using priors and not the machine learning model. Our findings point to the direction that future research must pursue in order to create synthetic data that could reduce the risk of attack when used to train machine learning models.

2 Threat Model

Our goal is to test whether a machine learning model trained on synthetic data can replace a machine learning model trained on original data. The idea is to release a machine learning model trained on synthetic data such that there is no leak of original data. The synthetic data serves as a replacement of the original data. In this section, we specify our goal more formally in the form of a threat model.

Inspired by [23], a threat model follows three main dimensions. First, the threat model describes the adversary by looking at the resources at the adversary's disposal and the adversary's objective. In other words, it specifies what the attacker is capable of and what the attacker's goal is. Second, it describes

the vulnerability, including the opportunity that makes an attack possible. Then, the threat model specifies the nature of the countermeasures that can be taken to prevent the attack.

Table 1 provides the specifications of our threat model for each of the dimensions. As resources, we assume that the attacker has access to our released machine learning classifier. In addition to the ML model, the attacker has a subset of the data that is used to train an attacker model. The adversary's objective is to infer sensitive information about individuals. In our experiments, the attack model is trained using subset of data in addition to the released machine learning model that predicts propensity-to-move. The opportunity for attack is the possession of original data including sensitive attributes. Finally, the countermeasure that we are investigating is data synthesis.

Table 1. Threat model addressed by our approach

Component	Description
Adversary: Objective	Specific attributes about individuals
Adversary: Resources	The attacker has access to the released classifier and has a subset of data
Vulnerability: Opportunity	Possession of original data and inference of individuals' sensitive data
Countermeasure	Make access to original data and model unreliable

3 Background and Related Work

In this section, we give a brief overview on basic concepts and related work on predicting the propensity to move, on privacy in machine learning, and model inversion attribute inference attack.

3.1 Propensity to Move

The propensity to move is defined as desires, expectations, or plans to move to another dwelling [5]. Multiple factors come to play to understand and estimate the propensity to move in a population. In [5], the authors have grouped those factors into two categories: (1) *Residential satisfaction* which is defined as the satisfaction with the dwelling and its location or surroundings. Residential satisfaction is divided into housing satisfaction and neighborhood satisfaction. (2) *Household characteristics* which is related to demographic and socioeconomic characteristics of the household. The gender and age are indicators of a household are important demographic attributes. For instance, a male household has different mobility patterns than a female household. Also, education and income of the household are important socioeconomic attributes.

In [10], authors investigated the possible relationship between involuntary job loss and regional mobility. In a survey, the German socio-economic panel [10]

looked at whether job loss increases the probability to relocate to a different region and whether displaced workers who relocate to another region after job loss have better labor market outcomes than those staying in the same area. They found that job loss has a strong positive effect on the propensity to relocate. In [17], the authors examined the residential moving behavior of older adults in the Netherlands. [17] used a data collected from the Housing Research Netherlands (HRN) to provide insights into the housing situation of the Dutch population and their living needs. A logistic regression model was used to assess the likelihood that respondents would report that they are willing to move in the upcoming two years. Among their key findings, they showed that older adults with a propensity to move are more often motivated by unsatisfactory conditions in the current neighborhood. Further results revealed that older adults are more likely to have moved to areas with little deprivation, little nuisance, and a high level of cohesion.

In [3], the authors studied the possibility of replacing a survey question about moving desires by a model-based prediction. To do so, they used machine learning algorithms to predict moving behavior from register data. The results showed that the models are able to predict the moving behavior about equally well as the respondents of the survey. In [4], the authors used data collected by the British Household Panel Survey. The data is conducted using a face to face interviews. They examined the reasons why people desire to move and how these desires affect their moving behavior. The results show that the reasons people report for desiring to move vary considerably over the life course. People are more likely to relocate if they desire to move for targeted reasons like job opportunities than if they desire to move for more diffuse reasons relating to area characteristics. In [18], the authors studied the social capital and propensity to move of four different resident categories in two Dutch restructured neighborhoods. They defined social capital as the benefit of cursory interactions, trust, shared norms, and collective action. Using a logistic regression model, they showed that (1) age, length of residency, employment, income, dwelling satisfaction, dwelling type and perceived neighborhood quality significantly predict residents' propensity to move and (2) social capital is of less importance than suggested by previous research.

3.2 Privacy in Machine Learning

In this section, we will discuss challenges and possible solutions in privacy preserving techniques. Existing works can be divided into three categories according to the roles of machine learning (ML) in privacy [19]: First, *making the ML model private*. This category includes making ML model (its parameters) and data private. Second, *using ML to enhance privacy protection*. In this category, the ML is used as a tool to enhance privacy protection of the data. Third, *ML based privacy attack*. The ML model is used as an attack tool of the attacker.

Based on the threat model, both data and the prediction model are important. Predicting and estimating the propensity to move requires access to models as well as to data. However, since the propensity to move data contains sensitive data such as income, gender, age, education level, the data is treated as sensitive and once collected from individuals it cannot be shared with third parties.

One possible solution is to generate synthetic data that captures the distribution of the original data and generates artificial, but yet realistic data. The synthetic data offers a replacement for the original data to enable model training, model validation and model explanation. In order to attempt to protect the machine learning model before release or sharing, we propose to train our model on the synthetic data instead of the original data. The goal is to test whether it is possible to release a machine learning model trained on synthetic data without leaking sensitive information.

Synthetic data generation is based on two main steps: First, we train a model to learn the joint probability distribution in the original data. Second, we generate a new artificial data set from the same learned distribution. In recent years, advances in machine learning and deep learning models have offered us the possibility to learn a wide range of data types.

Synthetic data was first proposed for Statistical Disclosure Control (SDC) [8]. The SDC literature distinguishes between two types of synthetic data [8]. First, *fully synthetic data sets* create an entirely synthetic data based on the original data set. Second, *partially synthetic data sets* contain a mix of original and synthetic values. It replaces only observed values for variables that bear a high risk of disclosure with synthetic values. In this paper, we are interested in fully synthetic data. For data synthesize, we used an open source and widely used R toolkit: *Synthpop*. We used a CART model for synthesize since it has been shown to perform well for other type of data [9]. Data synthesis is based on sequential modeling by decomposing a multidimensional joint distribution into conditional and univariate distributions. In other words, the synthesis procedure models and generates one variable at a time, conditionally to previous variables:

$$f_{x_1,x_2,..,x_n} = f_{x_1} \times f_{x_2|x_1} \times .. \times f_{x_n|x_1,x_2,..x_{n-1}} \tag{1}$$

Synthesis using CART model has two important parameters. First, the order in which variables are synthesized called *visiting.sequence*. This parameter has an important impact on the quality of the synthetic data since it specifies the order in which the conditional synthesize will be applied. Second, the *stopping rules* that dictate the number of observations that are assigned to a node in the tree.

3.3 Attribute Inference Attack

Privacy attacks in machine learning [6,22] include membership inference attacks [24], model reconstruction attacks such as attribute inference [29], model inversion attacks [11,12], and model extraction attacks [28]. Here, we focus on a form of model inversion attacks, namely, attribute inference attack.

Model inversion attacks try to recover sensitive features or the full data sample based on output labels and partial knowledge (subset of data) of some features [1,22]. [1] provided a summary of possible assumptions about adversary capabilities and resources for different model inversion attribute inference attacks. In [11,12], the authors introduced two types of model inversion attacks: Black-box

attack and white-box attack. The difference between black-box attack and white-box attack lies in the amount of resources that are available for the adversary. In [1], the authors proposed two types of model inversion attacks: (1) confidence score-based model inversion attack and (2) label-only model inversion attack. The first attack assumed that the adversary has access to the target model's confidence scores, whereas the second assumed that the adversary has access to the target model's label predictions only. Other attacks such as [14] assumed that the attacker does not have access to target individuals non-sensitive features.

Attribute Inference Attack. An attribute inference attack or attribute disclosure occurs if an attacker is able to learn new information about a specific individual, i.e., the values of certain attributes. Examples from the Statistical Disclosure Control (SDC) literature include [8,16].

Here, we study attribute inference attack as prediction. An attacker trains a model to predict the value of an unknown sensitive attribute from a set of known attributes given access to raw or synthetic data [15,25]. We implemented our attribute inference attack using *adversarial robustness toolbox*[1]. In order to perform an attribute inference attack, we assume that the attacker has access to a subset of data, a marginal prior distribution representing possible values for the sensitive features in the training data, and the released ML model's predictions. Using this resources, an attacker is able to train a model to learn sensitive information. This attack is called black-box attack because the predictions of the model, but not the architecture or the weights are available to the attacker. Further details about our black-box attack will be discussed in Sect. 4.3.

In addition to black-box attack, we use two other attack models as baselines for comparison, namely, *random attack* and *baseline attack*. Both attacks assume that the attacker does not have access to the released ML model. First, the random attack has only access to the marginal prior distribution of the sensitive feature that is being targeted. Our random attack uses random classifier with a stratified strategy, i.e., it generates random predictions that respect the class distribution of the training data. Second, the baseline attack also access to the prior distribution of the sensitive feature. However, in addition it also uses a ML model, i.e., a random forest classifier, to infer sensitive attributes. Recall that only the black-box attack is related to our threat model defined in Sect. 2. The random and baseline attacks provide comparative conditions, which the black-box attack must outperform.

Measuring Success of Inference. Prior work on synthetic data disclosure risk [26] looked at either matching probability by comparing perceived match risk, expected match risk, and true match risk [20], or Bayesian estimation approach by assuming that an attacker seeks a Bayesian posteriori distribution [21]. In this paper, our black-box attack is considered successful if its accuracy outperforms the accuracy of a random attack. In other words, we assume that going beyond a random guess, can reveal sensitive information about individuals. This type of

[1] https://github.com/Trusted-AI/adversarial-robustness-toolbox.

measurement is similar to previous work on model inversion attribute inference attacks [11,12,14], which measure the difference between the adversary's predictive accuracy given the model and the best, i.e., ideal, accuracy that could be achieved without the model [29]. Methods for measurements of success are discussed in [2], who also covers the precise or probabilistic measures conventionally used in the SDC community, i.e., using matching or Bayesian estimate.

4 Experimental Setup

In this section, we describe our data sets, utility measures measured by applying different machine learning algorithms, and adversary resources.

4.1 Data Set

For our experiments, we used an existing data about someone's propensity-to-move. The data was collected by [3]. [3] linked several registers from the Dutch System of Social Statistical Datasets (SSD). The data set has around 150K individuals including 100K individuals drawn randomly from register data and 50K individuals are sampled from the Housing Survey 2015 (HS2015) respondents. The resulting data set has used in [3] has 700 variables containing for each individual: (1) "y01" the binary target variable indicating whether (=1) or not (=0) a person moved in year j where j = 2013, 2015. The target attribute "y01" is imbalanced and dominated by class 0. (2) time independent personal variables, (3) time dependent personal, household, and housing variables, (4) information about regional variables.

Feature Selection. Different from [3], we applied feature selection to reduce the number of features. Some features can be noise and potentially reduce the performance of the models. Also, reducing number of feature helps to reduce the complexity of synthesize and to better understand the output of the ML model. To do so, we applied *SelectKBest* from *Sklearn*[2]. We use chi2 method as a scoring function. We selected top $K = 30$ features with the highest scores. Our final data set contains 30 best features for a total of 150K individuals[3]. In addition to the 30 features, we added gender (binary), income (categorical with five categories), and age (categorical with seven categories) as sensitive features that will be used in our attribute inference attack later (Sect. 5.2). Gender, age, and income have balanced classes. Similar to [3], we found that the most important features are age (lft), time since latest change in household composition (inhehalgr3), and time since latest move or number of moves (rinobjectnummer).

[2] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html.

[3] We note that reducing the number of features does not have an impact on the success rate of the attack because there is a redundancy in some variables since they go until 17 years back [3].

Data Splits. As mentioned earlier, our propensity to move data was collected in 2013 and 2015. Following [3], we use the 2013 data to train our classifier and the 2015 data to test the classifier and to carry out the attacks. The 2015 data contains individuals who were present in the 2013 data set, and also new individuals. We split the 2015 data set into two parts "original individuals" (inclusive) and "new in 2015 individuals" (exclusive) in order to test our classifier and our attacks on individuals who were in the training set but also in the also on "unseen individuals".

4.2 Utility Measures

Machine Learning Algorithms. We selected a number of machine learning algorithms to predict propensity-to-move. The chosen machine learning techniques provide insight into the importance of the features and are easy to interpret and understand [3].

In our experiments in Sect. 5.1, we used: *decision tree* where a tree is created/learned by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner. Extra trees and random forest are part of ensemble methods. In *random forest*, each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. *Extra trees* fits a number of randomized decision trees on various sub-samples of the data set and uses averaging to improve the predictive accuracy and control overfitting. *Naive Bayes* is a probabilistic machine learning algorithm based on applying Bayes' theorem with strong (naive) independence assumptions between the features. *KNN*, K-nearest neighbors, is a non-parametric machine learning algorithm. KNN uses proximity to make predictions about the grouping of an individual data point.

Metrics for Evaluating Performance of ML Models. Similar to [3] and since our target propensity-to-move attribute is imbalanced, we used: F1-score, as a harmonic mean of precision and recall score. Matthews Correlation Coefficient (MCC), and Area Under the Curve (AUC) that measures the ability of a classifier to distinguish between classes.

4.3 Adversary Resources

In Sect. 3.3, we provided description of our attack models. The attacker is interested to infer target individual sensitive features. Below, we briefly discuss different attack models used in our experiments along with different resources that are available for the attacker.

- *Random attack*: uses a subset of data and marginal prior distribution.
- *Baseline attack*: uses a subset of data, marginal prior distribution, and random forest classifier.
- *Black-box attack*: uses a subset of data, marginal prior distribution, released ML model, and random forest classifier.

In random attack model, a random classifier[4] randomly infers target individual's sensitive features i.e., gender, age, income. In baseline attack model, a random forest classifier[5] is trained on a subset of data and marginal prior distribution to predict sensitive features. Last but not least, a black-box attack model has access to the released ML model's predictions, in addition to having access to subset of data and marginal prior distribution. Then, a random forest classifier is trained to infer target individual's sensitive features.

Understanding the vulnerability of a model to attribute inference attack requires using right metric to evaluate different attack models. Since our sensitive target features (gender, age, income) are balanced [11], we used precision, recall to measure the effectiveness of the attacks. Precision measures the ability of the classifier not to label as positive a sample that is negative. Precision is the ratio of $tp/(tp + fp)$ where tp is the number of true positives and fp the number of false positives. Recall measures the ability of the classifier to find all the positive samples. Recall is the ratio of $tp/(tp+fn)$ where tp is the number of true positives and fn the number of false negatives. We also measure accuracy which is defined as the fraction of predictions that our classifier got right.

5 Experimental Results

Now, that we have defined our threat model including the adversary resources and capabilities, and utility measures to evaluate the quality of synthetic data and machine learning algorithms, we turn to discuss our experimental results.

5.1 Evaluation of Machine Learning Algorithms

Table 2 shows our results of classification performance of propensity to move, and confirms the results of [3]. As expected, all classifiers outperform the random baseline, with classifiers using trees generally the stronger performers. We also see that when the test set includes only individuals already present in the training set (*inclusive*), the performance is better than when it includes only "unseen" individuals (*exclusive*). Note that if the data for the *inclusive* individuals were identical in the training and test set, we would have expected very high classification scores. However, the data is not identical because it was collected on two different occasions with two years intervening, and individuals' situations would presumably have changed.

Reproducing Burger et al.,'s [3] *results* In Table 2, results show that all machine learning classifiers outperform random classifier. Overall we observe that our results are in line with [3] across different metrics. This confirms that we can still predict individuals moving behavior in the same level as in [3] even after reducing number of features.

[4] Random Classifier using Stratified strategy from https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html.

[5] Random Forest Classifier: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

Table 2. Classification performance of propensity-to-move measured in terms of AUC, MCC, and F1-score on **original data** and **synthetic data**. (Right) the data splitting is similar to [3]. The training set individuals and test set individuals are inclusive. (Left) A different data splitting where we train the model on individuals data from 2013, then, we test the model on different individuals from 2015.

Machine learning algorithms		Training and test individuals are exclusive			Training and test individuals are inclusive		
		AUC	MCC	F1-score	AUC	MCC	F1-score
Original data	Random	0.4962	−0.0105	0.2139	0.5014	0.0029	0.1633
	NaiveBayes	0.5656	−0.0328	0.5491	0.6815	0.2204	0.2992
	RandomForest	0.7061	0.3210	0.6322	0.7532	0.3121	0.4460
	DecisionTree	0.6372	0.2692	0.5376	0.6568	0.2292	0.3057
	ExtraTrees	0.7226	0.3197	0.6325	0.7597	0.3212	0.4525
	KNN	0.6304	0.2074	0.4104	0.6717	0.1744	0.2235
Synthetic data	Random	0.4991	−0.025	0.2261	0.5011	0.0022	0.1657
	NaiveBayes	0.5658	0.045	0.5451	0.6822	0.2029	0.2578
	RandomForest	0.7053	0.3282	0.6343	0.7467	0.3133	0.4471
	DecisionTree	0.6489	0.2598	0.4878	0.6618	0.2125	0.3078
	ExtraTrees	0.7188	0.3185	0.6321	0.7557	0.3138	0.4464
	KNN	0.6067	0.1152	0.1857	0.6542	0.1637	0.2070

In addition to reproducing [3], we looked at another prediction model where train and test individuals are exclusive/different. We found that it is also possible to predict moving behavior of new individuals from 2015 based on a classifier trained on different individuals from 2013.

Measuring the Utility of Synthetic Data. In order to evaluate the quality of synthetic data, we run machine learning algorithms on synthesized training set (2013 data). we used $TSTR$ [13] evaluation strategy where we train classifiers on 2013 synthetically generated data and we test on 2015 original data. Results in Table 2 show that the performance of machine learning algorithms trained on synthetic data is very close and comparable to the performance of machine learning algorithms trained on original data. This confirms that the synthetic training set can replace the original training set. In the remainder of the paper, we will focus on decision tree model. We will assume that we are releasing a decision tree model.

5.2 Model Inversion Attribute Inference Attack

In this section, we present the results of our experiments on attribute inference attack using the three attack models: (1) random attack, (2) baseline attack, (3) black-box attack (Sect. 4.3). Recall that we assume that the adversary can have access to three different subsets of data (Sect. 2).

1. `Inclusive individuals (2013)`: the attacker has access to a subset of the data that is used from 2013 to train the released machine learning algorithm.
2. `Inclusive individuals (2015)`: the attacker has access to a more recent subset of data from 2015, but for the same set of individuals that are used to train the released machine learning algorithm.
3. `Exclusive individuals (2015)`: the attacker has access to a recent subset of data from 2015, but the individuals are different from individuals that are used to train the released machine learning algorithm.

Table 3 shows results of different attribute inference attacks for three type of sensitive features gender, age and income. We notice that attack always achieves better than random scores, which demonstrates the viability of the attack.

Table 3. Results of model inversion attribute inference attacks. Adversary resources can be either: `Inclusive individuals (2013)`, `Inclusive individuals (2015)`, or `Exclusive individuals (2015)`. ± represents the standard deviation over ten times of running the experiments. Numbers in gray represent the best inference results across conditions. Note that only black-box attack is related to threat model described in Sect. 2. An attack is considered successful if its score is higher than a score of random attack.

Adversary Resources	*Released ML*	*Attack Models*	*Gender*			*Age*			*Income*		
			Accuracy	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
Inclusive individuals (2013)	Original	*Random*	0.500 ±0.00	0.500 ±0.00	0.500 ±0.00	0.1095 ±0.00	0.1129 ±0.00	0.1112 ±0.00	0.2086 ±0.00	0.2060 ±0.00	0.2077 ±0.00
		Baseline	0.6107 ±0.007	0.6103 ±0.007	0.6104 ±0.007	0.1472 ±0.003	0.1566 ±0.003	0.1407 ±0.001	0.1483 ±0.005	0.1590 ±0.005	0.2323 ±0.006
		Black-Box	0.6187 ±0.005	0.6181 ±0.005	0.6183 ±0.005	0.1482 ±0.004	0.1577 ±0.004	0.1412 ±0.001	0.1469 ±0.004	0.1576 ±0.005	0.2302 ±0.006
	Synthetic	*Random*	0.500 ±0.00	0.500 ±0.00	0.500 ±0.00	0.1164 ±0.00	0.1223 ±0.00	0.1213 ±0.00	0.1838 ±0.00	0.1889 ±0.00	0.1983 ±0.00
		Baseline	0.6262 ±0.00	0.6263 ±0.006	0.6264 ±0.006	0.1562 ±0.004	0.1561 ±0.004	0.1412 ±0.001	0.1509 ±0.003	0.1575 ±0.003	0.2189 ±0.004
		Black-Box	0.6298 ±0.005	0.6299 ±0.005	0.6300 ±0.005	0.1562 ±0.003	0.1561 ±0.003	0.1412 ±0.001	0.1492 ±0.003	0.1553 ±0.004	0.2182 ±0.006
Inclusive individuals (2015)	Original	*Random*	0.500 ±0.00	0.500 ±0.00	0.500 ±0.00	0.1095 ±0.00	0.1129 ±0.00	0.1112 ±0.00	0.2086 ±0.00	0.2060 ±0.00	0.2077 ±0.00
		Baseline	0.6240 ±0.006	0.6228 ±0.006	0.6227 ±0.00	0.1552 ±0.003	0.1590 ±0.003	0.1467 ±0.001	0.1502 ±0.004	0.1552 ±0.004	0.2327 ±0.007
		Black-Box	0.6235 ±0.009	0.6226 ±0.009	0.6223 ±0.009	0.1547 ±0.003	0.1585 ±0.003	0.1463 ±0.001	0.1545 ±0.003	0.1599 ±0.003	0.2428 ±0.005
	Synthetic	*Random*	0.500 ±0.00	0.500 ±0.00	0.500 ±0.00	0.1164 ±0.00	0.1223 ±0.00	0.1213 ±0.00	0.1838 ±0.00	0.1889 ±0.00	0.1983 ±0.00
		Baseline	0.6186 ±0.006	0.6188 ±0.006	0.6186 ±0.006	0.1657 ±0.003	0.1606 ±0.003	0.1465 ±0.001	0.1620 ±0.003	0.1592 ±0.003	0.2169 ±0.005
		Black-Box	0.6236 ±0.006	0.6237 ±0.006	0.6236 ±0.006	0.1646 ±0.003	0.1595 ±0.003	0.1456 ±0.001	0.1626 ±0.003	0.1596 ±0.003	0.2259 ±0.006
Exclusive individuals (2015)	Original	*Random*	0.500 ±0.00	0.500 ±0.00	0.500 ±0.00	0.1095 ±0.00	0.1129 ±0.00	0.1112 ±0.00	0.2086 ±0.00	0.2060 ±0.00	0.2077 ±0.00
		Baseline	0.5269 ±0.009	0.5198 ±0.009	0.5201 ±0.009	0.0830 ±0.001	0.2116 ±0.005	0.1279 ±0.001	0.0829 ±0.003	0.1779 ±0.008	0.2182 ±0.02
		Black-Box	0.5272 ±0.005	0.5195 ±0.005	0.5199 ±0.005	0.0817 ±0.001	0.2100 ±0.005	0.1280 ±0.001	0.0804 ±0.003	0.1693 ±0.008	0.2283 ±0.02
	Synthetic	*Random*	0.500 ±0.00	0.500 ±0.00	0.500 ±0.00	0.1164 ±0.00	0.1223 ±0.00	0.1213 ±0.00	0.1838 ±0.00	0.1889 ±0.00	0.1983 ±0.00
		Baseline	0.5268 ±0.009	0.5198 ±0.009	0.5201 ±0.009	0.0825 ±0.001	0.2116 ±0.005	0.1279 ±0.001	0.0829 ±0.003	0.1779 ±0.008	0.2182 ±0.02
		Black-Box	0.5272 ±0.005	0.5195 ±0.005	0.5198 ±0.005	0.0817 ±0.001	0.2100 ±0.005	0.1280 ±0.001	0.0804 ±0.003	0.1693 ±0.008	0.2283 ±0.02

Comparing the row "Original" for the three individuals sets and across all three sets of sensitive attributes (columns), we see that the attack is less successful for the "Exclusive" individuals who were unseen in the training data of the classifier. This fact might lead us to wonder whether training the classifier on synthetic data might lead to less successful attacks, since the individuals in the training data would be in some way "different" with the target individuals. This, however, turns out not to be the case. Comparing the row "Synthetic" for the three individuals sets and across all three sets of sensitive attributes (columns), we see that if the training data is synthesized using the original training data, the model is just as susceptible to attack as when trained on the original data. This point is less surprising when we take into account the high success of the "Random" attack. This attack recovers sensitive attributes of individuals without access to the trained machine learning model. Instead, priors are used. We assume that the information of the priors is also retained in the trained model. These results demonstrate the magnitude of the challenge that we face, if we wish to release a trained machine learning model publically.

6 Conclusion and Future Work

In this paper, we have investigated an attack on a machine learning model trained to predict individual's propensity-to-move i.e., in the next two years. for individuals in the training data as well as for "unseen" individuals. However, we observed that for "unseen" individuals, the attribute inference attack is somewhat less successful. This result is consistent with the training data used to train ML model having a different distribution than the "unseen" individuals.

To explore the ability of synthetic data to protect against attribute inference attack, we created fully synthetic data using CART model. The ML model trained on synthetic data maintained prediction performance, but was found to leak in the same way as the original classifier. This result is not particularly surprising. Synthetic data mimics properties of the original data including overall structure, correlation between features, and the joint distributions [25].

Our results is interesting because until now The SDC community working with synthetic data has mainly focused on measuring the risk of identity disclosure rather than attribute disclosure [26]. In the identity disclosure literature, synthetic data has been shown to provide protection [7,27].

Our work draws attention to the fact a lot of work is still needed to protect against attribute disclosure [2]. A potential solution to protect against attribute inference attack is to apply privacy-preserving techniques during synthesis, e.g., data perturbation or masking sensitive attributes. Also, it would be interesting to explore different combinations of ML and conventional models to synthesize and carry out attribute attacks. From an evaluation perspective, future work should look at other metrics [15] (e.g., from SDC and/or ML perspective) to evaluate and quantify the success of attribute inference attack for a given target individual. Finally, future research should expand the threat model that we have adopted in this research (Sect. 2) and other attack scenarios in which the attacker

has access to more limited resources, e.g., assuming that attacker does not have access to all attributes in data.

References

1. Mehnaz, S., Dibbo, S.V., Kabir, E., Li, N., Bertino, E.: Are your sensitive attributes private? Novel model inversion attribute inference attacks on classification models. In: 31st USENIX Security Symposium (USENIX Security), Boston, MA. USENIX Association (2022)
2. Andreou, A., Goga, O., Loiseau, P.: Identity vs. attribute disclosure risks for users with multiple social profiles. In: Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 163–170. ASONAM (2017)
3. Burger, J., Buelens, B., de Jong, T., Gootzen, Y.: Replacing a survey question by predictive modeling using register data. In: ISI World Statistics Congress, pp. 1–6 (2019)
4. Coulter, R., Scott, J.: What motivates residential mobility? Re-examining self-reported reasons for desiring and making residential moves. Popul. Space Place **21**(4), 354–371 (2015)
5. Crull, S.R.: Residential satisfaction, propensity to move, and residential mobility: a causal model. In: Digital Repository at Iowa State University (1979). http://lib.dr.iastate.edu/
6. De Cristofaro, E.: A critical overview of privacy in machine learning. IEEE Secur. Priv. **19**(4), 19–27 (2021)
7. Domingo-Ferrer, J.: A survey of inference control methods for privacy-preserving data mining. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy-Preserving Data Mining, pp. 53–80. Springer, Boston (2008). https://doi.org/10.1007/978-0-387-70992-5_3
8. Drechsler, J.: Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation, vol. 201. Springer, New York (2011). https://doi.org/10.1007/978-1-4614-0326-5
9. Drechsler, J., Reiter, J.P.: An empirical evaluation of easily implemented, non-parametric methods for generating synthetic datasets. Comput. Stat. Data Anal. **55**(12), 3232–3243 (2011)
10. Fackler, D., Rippe, L.: Losing work, moving away? Regional mobility after job loss. Labour **31**(4), 457–479 (2017)
11. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM Conference on Computer and Communications Security (SIGSAC), CCS 2015, pp. 1322–1333 (2015)
12. Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., Ristenpart, T.: Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. In: 23rd USENIX Security Symposium (USENIX Security), San Diego, CA, pp. 17–32. USENIX Association (2014)
13. Heyburn, R., et al.: Machine learning using synthetic and real data: similarity of evaluation metrics for different healthcare datasets and for different algorithms. In: Data Science and Knowledge Engineering for Sensing Decision Support: Proceedings of the 13th International FLINS Conference, pp. 1281–1291. World Scientific (2018)

14. Hidano, S., Murakami, T., Katsumata, S., Kiyomoto, S., Hanaoka, G.: Model inversion attacks for prediction systems: without knowledge of non-sensitive attributes. In: 15th Annual Conference on Privacy, Security and Trust (PST), pp. 115–11509. IEEE (2017)
15. Hittmeir, M., Mayer, R., Ekelhart, A.: A baseline for attribute disclosure risk in synthetic data. In: Proceedings of the 10th ACM Conference on Data and Application Security and Privacy, pp. 133–143 (2020)
16. Hundepool, A., et al.: Statistical Disclosure Control. Wiley, Hoboken (2012)
17. de Jong, P.A.: Later-life migration in the Netherlands: propensity to move and residential mobility. J. Aging Environ. **36**, 1–10 (2020)
18. Kleinhans, R.: Does social capital affect residents' propensity to move from restructured neighbourhoods? Hous. Stud. **24**(5), 629–651 (2009)
19. Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., Lin, Z.: When machine learning meets privacy: a survey and outlook. ACM Comput. Surv. **54**(2), 1–36 (2021)
20. Reiter, J.P., Mitra, R.: Estimating risks of identification disclosure in partially synthetic data. J. Priv. Confid. **1**(1) (2009)
21. Reiter, J.P., Wang, Q., Zhang, B.: Bayesian estimation of disclosure risks for multiply imputed, synthetic data. J. Priv. Confid. **6**(1) (2014)
22. Rigaki, M., Garcia, S.: A survey of privacy attacks in machine learning. arXiv preprint arXiv:2007.07646 (2020)
23. Salter, C., Saydjari, O.S., Schneier, B., Wallner, J.: Toward a secure system engineering methodology. In: Proceedings of the 1998 Workshop on New Security Paradigms, pp. 2–10. NSPW (1998)
24. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: Symposium on Security and Privacy (SP), pp. 3–18. IEEE (2017)
25. Stadler, T., Oprisanu, B., Troncoso, C.: Synthetic data-anonymisation groundhog day. In: 29th USENIX Security Symposium (USENIX Security). USENIX Association (2020)
26. Taub, J., Elliot, M., Pampaka, M., Smith, D.: Differential correct attribution probability for synthetic data: an exploration. In: Domingo-Ferrer, J., Montes, F. (eds.) PSD 2018. LNCS, vol. 11126, pp. 122–137. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99771-1_9
27. Templ, M.: Statistical Disclosure Control for Microdata: Methods and Applications in R. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-50272-4
28. Tramèr, F., Zhang, F., Juels, A., Reiter, M.K., Ristenpart, T.: Stealing machine learning models via prediction APIs. In: 25th USENIX Security Symposium (USENIX Security 2016), Austin, TX, pp. 601–618. USENIX Association (2016)
29. Yeom, S., Giacomelli, I., Fredrikson, M., Jha, S.: Privacy risk in machine learning: analyzing the connection to overfitting. In: 31st Computer Security Foundations Symposium (CSF), pp. 268–282. IEEE (2018)