IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 December 2023, accepted 14 January 2024, date of publication 12 February 2024, date of current version 29 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3365190

RESEARCH ARTICLE

You Are What You Buy: Personal Information Extraction From Anonymized Data

THOMAS CILLONI[1], (Member, IEEE), CHARLES FLEMING[2], (Member, IEEE), AND CHARLES WALTER[1], (Member, IEEE)

[1]Department of Computer and Information Science, University of Mississippi, University, MS 38677, USA
[2]Cisco, San Jose, CA 95134, USA

Corresponding author: Charles Walter (cwwalter@olemiss.edu)

ABSTRACT The exponential growth of data in the information age poses several threats to the privacy and safety of digital service users. Existing legislation, such as the GDPR in Europe and the CCPA in California, defines frameworks and guidelines to promote personal privacy but leaves freedom in the choice of means to achieve privacy. Data anonymization techniques remove information that can be used to identify individuals from the dataset, either through suppression, generalization, anatomization, permutation, or perturbation. Information suppression remains the most common, safe, and widely applicable anonymization method, though at a high data utility cost. In this paper, we argue that even information suppression may not be sufficient in some cases. We study the case of a dataset that describes the shopping habits of a grocery store's customers. All identifiers and quasi-identifiers are removed from the dataset by suppression. However, by augmenting the data in a novel multi-step, iterative process, and building a neural network enriched with prior knowledge, we show that most sensitive information can be retrieved with an accuracy of 80%.

INDEX TERMS Data anonymization, privacy, machine learning, responsible AI.

I. INTRODUCTION

In recent years there has been increasingly fast adoption of web-based services [1], generating a stream of data of size never seen before [2] and giving rise to the Information Age. Data in the Information Age is a heterogeneous mixture of raw data points that can be collected from virtually any service or platform with an internet connection. The heterogeneity of such data is the very reason why mining it can reveal extremely useful information, and at the same time, it makes it impossible to clearly define what attributes may become quasi-identifiers in the dataset when mined. Data collection, therefore, requires some form of control to ensure, with varying degrees of confidence, that such data cannot be used for malicious scopes, such as retrieving personal information and crafting cyber-attacks.

Two of the most widely used privacy preservation guidelines are the CCPA and the GDPR. While the United States does not regulate data privacy at a federal level, a number of entities follow the California Consumer Privacy Act (CCPA) [3]. This legislation defines how businesses must be transparent with their clients about the collection, usage, and sale of their data. Additionally, it gives users the right to request access to their data, forbid its sale, or erase it. On a similar note is the General Data Protection Regulation (GDPR) [4], which must be followed by all businesses operating with European consumers' data. It defines users' right to be informed about, access, rectify and delete their data, similar to the CCPA. Moreover, the GDPR gives people the right to confute any algorithmic decision made with their data and request that it be revised by a human. These guidelines are mostly concerned with the transparency of business decisions and data usage and do not specify to what extent data can be used, nor how much personal or sensitive information can be extracted from available sources.

Nowadays, most techniques to prevent data from revealing personally identifying information rely on masking, permuting, and tampering with data points. These modifications can be applied either partially, usually to render specific attribute values more generic, such as for quasi-identifiers, or totally

The associate editor coordinating the review of this manuscript and approving it for publication was Giacomo Fiumara.

in the case of identifying attributes. The goal that all such techniques share is to render it virtually impossible to infer personal information or identify an individual given a dataset containing their data. Some of the most popular techniques for data anonymization are t-closeness, k-anonymization, and l-diversity [5]. However, regardless of the use of these measures, privacy is still at risk.

In this paper, we show how anonymized records of customers of a grocery store are susceptible to attribute inference attacks, which is the disclosure of purposely cloaked information from a dataset. While certain kinds of data are intuitively quite telling of personal information associated with an individual (e.g. medications for inferring age), we only use aggregated data about the purchasing statistics of some customers and some of their responses to non-targeted advertisements.

The anonymization process consists of the suppression of all identifier and quasi-identifier attributes from the dataset, which is comprised of education level, income, age, marital status, and the number of children in the household. The remaining data points are only statistics about shopping habits, such as membership age, money spent in each department, advertisement engagement, etc. No information about individual purchases is included, such as products, dates, or cart contents.

Throughout the paper, we argue that the semantics of data attributes can be exploited to extract information beyond what the raw data alone can provide. Our contributions are as follows: 1) We provide a theoretical framework of semantic data augmentation that allows machine learning models to extract significantly more information from data; 2) We apply our data augmentation model to a dataset of customer purchase statistics and show that it allows attribute inference attacks to reach 80% accuracy; and 3) We propose guidelines on how to harness the power of this semantic data augmentation model further and give suggestions on anonymization goals to consider for future research.

II. BACKGROUND

A recent comparative study [6] tackles the problem of deciding which technique to use depending on the contents of a dataset. The authors compare five techniques, namely *generalization, suppression, distortion, swapping*, and *masking*, and discuss which kinds of attributes are more effectively anonymized by which technique (where categorical, numerical, etc. are attribute types0). Their results show that efficacy-wise, suppression is the most effective methodology while swapping is the least. As for the efficiency, or resource requirements of the techniques, swapping is the most resource-intensive and suppression the least. However, suppressing attribute values entirely by dropping the related columns in a database is sometimes still not enough, and attackers are able to retrieve the suppressed data, as we show.

Personal information is important to businesses: it allows them to tailor advertisements, predict demand for products and services, and understand the behavioral and spending habits of customers. One would assume that by not disclosing personal information when sharing a dataset, such as by suppressing it, such information can never be available. In reality, however, such information can be inferred. Lu et al. [7] created *GenderPredictor* to show that it is possible to predict the gender of a customer based on product viewing logs. *GenderPredictor* is a Gradient Boosting Decision Tree classifier that, taking as inputs the data about a customer session's duration, number of products viewed, time of login, and category of products viewed. Their results show that over 90% of female customers and almost 70% of male customers are correctly labeled.

Similar to *GenderPredictor*, Merler, Cao, and Smith [8] propose a gender prediction model that uses the profile and feed pictures of Twitter users to predict their gender. While this may seem trivial using profile pictures, not all images contain a single face to predict the gender, with many not having a face at all. Of 10K user profiles, less than 55% present a unique face in their profile or feed images. What is therefore proposed is to extract the content and context from these images using a trained network that predicts the actions, objects, and environments in a given picture, and computing statistics from images such as color pallets and background colors. Combining all this information, the authors can predict Twitter users' gender with an 88% accuracy.

A more comprehensive study, proposed by Chaabane, Acs, and Kaafar [9], takes on the challenge of determining a wider range of personal information of some Facebook users from their musical interests. The authors scraped over 100M public Facebook profiles and tried to extract personal information and musical interests, resulting in 100K usable records. They then built models to perform attribute inference attacks. This work uses a semantic augmentation procedure by building a hierarchical Wikipedia ontology and applying it to augment individual musical interests. We refer to this kind of augmentation in our framework as *Value Semantics Injection* (see III-C1). Their model is able to determine the *gender* of users with about 70% accuracy. Their results highlight the danger of disclosing seemingly harmless information online as they can be used to infer sensitive information accurately.

But how does the semantics of data exactly improve a model's accuracy? The intuition to formulate an answer can be found in the work of Baran et al. [10]. Through an empirical study, they assess the correlation between products purchased in a store and the perception that society has of individuals. Though seemingly far-fetched, the two are actually closely related. Advertisements and social media shape our opinion of brands and products. Opinionated views on the same products from different brands lead to opinions on the people who purchase said products and brands. Practically, the study tries to correlate the brands in people's shopping carts with the perception that others have on them: results show that there is a strong correlation. We believe this is indicative of the nature of data that can be extracted from the semantics of data points. As an analogy,

the items in a person's shopping cart are the true data points and their brands represent their semantic context. If humans can make judgments based on brands (regardless of their truthfulness or ethics), then so should machines if explicitly given such data. We, therefore, use this as the basis for our data augmentation model, which becomes incorporation of the semantics associated with attributes or attribute values (we thoroughly explore the first) in a dataset.

Introducing semantics-based augmented data in existing datasets is an already widely researched practice [11], [12]. Most existing techniques, however, are applicable only to specific tasks (such as context-aware web queries [13]), specific dataset types (see tabular data in [5]), or specific models (neural networks in [11] and [12]). Instead, we propose a general method that can improve the extraction of information for methods ranging from statistical models to deep neural networks.

III. SEMANTIC DATA AUGMENTATION

This section briefly describes the dataset we use and contains the design of a general-purpose, semantics-based data augmentation framework to render anonymized datasets more prone to attribute inference attacks. The framework is composed of three main parts: *attribute semantic augmentation* uses human intuition to combine existing attributes or add new ones, *value semantics augmentation* makes data values more easily understandable from a machine perspective, and *a priori knowledge injection* uses domain knowledge to add information to a dataset. We also show how to use each component in the application to the customer statistics dataset, described below.

A. DATASET DESCRIPTION

The Customer Personality Analysis (CPA) dataset[1] is a collection of 2240 records of customers who have a loyalty membership with an unnamed grocery store. For each record, the information available includes when the membership was started, the date of the last trip to the store, the amount of money spent in each of 6 categories in the supermarket (wine, fruit, meat, fish, sweets, jewelry) since the membership started, and how many purchases were made on the phone, in-store or online. All directly identifying information has been removed from the dataset (such as full names, phone, and ID numbers), leaving quasi-identifiers and other attributes. The quasi-identifiers are age, education level, marital status, income, and the number of children and teenagers in the same household. These latter attributes are what we assume the store's data managers would hide before selling the data, but we show that most of it can be recovered after semantically augmenting the data.

B. ATTRIBUTE SEMANTIC AUGMENTATION

Attributes in datasets and databases usually represent tangible entities or characteristics of data. Consequently, they have a meaning that humans can comprehend. The meaning of two certain attributes can sometimes be similar, rendering one of the two superfluous, or too complex, requiring simpler descriptions. By merging, splitting, and combining attributes, data can be represented more systematically.

[1]https://www.kaggle.com/imakash3011/customer-personality-analysis

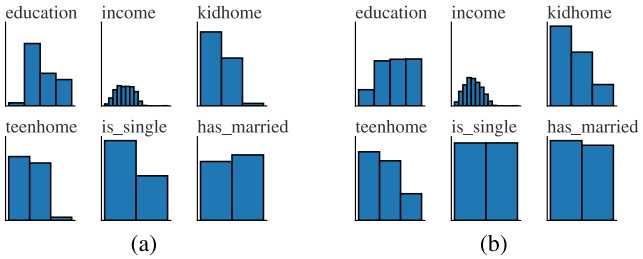

FIGURE 1. Effect of re-balancing the under-represented classes in the train set. In (a) some attribute values are barely visible, being significantly under-represented; in (b), after augmentation, attribute values are more evenly balanced and the *income* (the only non-categorical attribute) more closely follows a normal distribution.

For example, in the CPA dataset the *amount of money* spent in each store category is highly correlated with the length of time a customer has had the membership, and how often visits to the store are taken. The data can therefore be semantically augmented by introducing the *share of money* spent into each category since becoming a member of the store. Values are to be computed such that the sum of the shares per each of the five categories is 1. The same concept can be applied to the number of in-store, on-the-phone, and online orders placed.

1) REDUNDANCY REMOVAL

Two (or more) attributes can refer to the same characteristic of a data sample, thus being redundant. For instance, in a *car* database the attributes *color* and *color_code* may take values *gray* and *light gray*, rendering the first virtually useless. In this case, the first attribute can be dropped from the dataset.

2) SIMPLIFICATION

An attribute may be too complex to be processed or be a combination of multiple attributes. In the CPA dataset, the *marital status* attribute is categorical and can take one of eight values, some of which have ambiguous meanings. Given the size of the dataset, there are not enough records for a model to learn the relationship between the various attribute values, such as that *married* and *together* both mean that a person is with someone, but in the first case after marriage and in the latter before. The attribute can therefore be augmented by splitting it into two semantically equivalent binary attributes: *is_single* and *has_married*. Figure 2 shows how the distribution of records can be improved by re-arranging this attribute.

3) DISTRIBUTION

Attributes may be re-distributed to be more easily interpretable, from both a human's and a machine's perspectives. Following the *car* database example above, the two mentioned attributes can be rearranged into a *color* and a *shade*

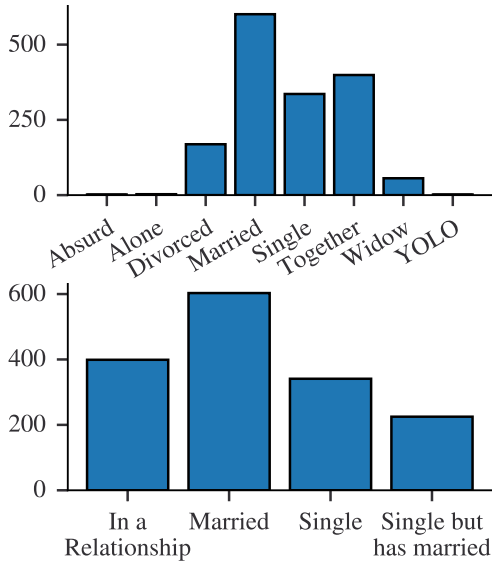

FIGURE 2. Change in the sample distribution before (figure above) and after (figure below) augmenting the marital status attribute.

attributes with values *gray* and *light*. Additionally, even data distribution is crucial to reduce bias in learned models. The sensitive information classes in the CPA dataset are unevenly distributed, as shown in Figure 1 (a). *Income* is the only attribute that follows a normal distribution, with all the others being irregular in shape. Particularly concerning are the *education, # of kids at home* and *# of teens at home* attributes, which have particularly skewed distributions and risk becoming underrepresented in a fit model. To overcome this obstacle, classes can be balanced by re-sampling the dataset with a sampling probability skewed towards records from underrepresented classes, and the result is shown in Figure 1 (b).

The main idea behind rearranging attributes is to have only a few possible values per attributed and have them be semantically interpretable. This can ensure a more uniform distribution of samples across attribute values, thus increasing the generalization of models that fit such data.

C. VALUES SEMANTIC AUGMENTATION

The meat of a dataset is the data it contains. If attribute definitions can be semantically augmented, so can attribute values (and to a larger extent). A dataset's values that can be semantically augmented are those belonging to one of two categories: categorical and cyclical. We define the process of incorporating external knowledge to the values of an attribute as *value semantics injection*. This is the potentially most powerful semantic augmentation task as it allows very large amounts of external data to be encoded in a dataset. We describe each of the two semantically augmented categories below.

1) CATEGORICAL ATTRIBUTES

These are the ones with the most potential to be semantically augmented. The simplest form of augmentation, in this case, is merging attribute values that are semantically equivalent into one. The intuition behind this augmentation is similar to the splitting of attributes described above: if an attribute has a lot of values, its distribution is more likely to be non-uniform than if the same attribute had fewer possible values. This can improve a model's generalization and reduce its susceptibility to underrepresented samples. In the CPA dataset the *education* attribute can take on one of 5 values: *Diploma, Bachelor's, Master,* 2^{nd} *Cycle, PhD*. However, a 2^{nd} cycle degree is the equivalent to a Master's degree in Europe according to the Bologna proceedings [14]. Therefore, we merge the two attribute values onto a single *Master* value.

The second and most potent augmentation for categorical attributes is the injection of the actual semantics for its values. Some Natural Language Processing models, such as Word2Vec [15], [16], learn vectorized representations of words in a semantic space, where the relative location between two-word vectors is telling of how they are related, highlighting synonyms, changes in gender, and semantically close terms. This information can be used to replace categorical attribute values with their corresponding semantic vectors, greatly reducing the training cost and highlighting correlations that would otherwise require much larger datasets. For instance, a neural network model may need thousands of training samples to learn that *chocolate spread* and *Nutella* are the same things, but if these were instead represented as their corresponding semantic vectors, their relationship would be embedded in the data. This task can be seen as the integration of a feature map learned separately (such as Word2Vec) into the data to be fed to another neural network for a different task.

Finally, some categorical attributes carry an implicit ordering and they can therefore be represented accordingly. For example, *education* is progressive in nature: a Bachelor's comes after a Diploma, a Ph.D. after a Master's and a Bachelor's, and so on. Knowledge about this progression can be injected into the data by converting the categorical values into numerical, with numbers from 0 to 4 for Diploma to Ph.D.

2) CYCLICAL ATTRIBUTES

Numerical or categorical attributes that represent cyclical entities can and should be semantically augmented. These attributes often do not carry the information about their cyclic nature, thus not representing the actual distance between their values. For instance, the month of December is very close to January, but if they were encoded as a numerical attribute their values would be on the opposite ends of the attribute range (12 and 1), signifying a large distance that is actually not true. Similarly, if they were encoded as categorical attributes, the distance between any two months would be the same, thus also not representing the semantics of months.

Semantically augmenting these attributes can be done with existing techniques, such as sine-cosine curve descriptors [17]. A cyclical attribute's values can be intuitively

distributed uniformly on a circle, and the x and y coordinates of each value be used as the encoding of said value, which can be calculated as the cosine and sine of the point on the circle. The result is a combination of two attributes that describes the distance properties between any two values to an acceptable level of truth. We apply this technique to the CPA dataset by splitting the *join date* attribute into a *join year* and a *join month* attributes, and make the month cyclical.

D. A PRIORI KNOWLEDGE INJECTION

Tangible entities have such a vast array of properties that datasets cannot possibly describe them. Human beings learn the properties of the objects around them through years of experience and data collection. Many of these properties are part of our implicit understanding of the world, while a few others are explicitly defined rules and properties. Datasets, by nature, cannot contain this kind of information. It is possible, however, to augment data in such a way as to infuse it with our understanding of things.

1) IN-DOMAIN PROPERTIES

When we think of a cat, we know what it looks like: has four legs, likely has fur, runs fast and jumps high, and makes a buzz when feeling comfortable. We also know that a cat cannot be blue, or purple, but it can be red, white, or black and that it cannot have two tails, but it can have one or none. These are tangible properties that can be implicitly embedded in a dataset, similar to how humans implicitly know them. A property can be defined as a set of class-invariant transformations for some (or all) the values of an attribute. Then, a dataset can be semantically augmented by generating new examples from existing ones by applying the transformations in a property. Referring to the previous example, a cat image can be flipped, rotated, changed the fur color to a different one, or removed the tail, without the sample ever being shifted to a different class.

A special kind of class-invariant transformation is the addition of noise. Knowing that minor noise in a record is acceptable and does not change its label is itself a form of a priori knowledge. Augmenting a dataset by introducing noisy samples is essentially equivalent to instructing a model to ignore minor variances in samples' values. In relation to the CPA dataset, there is likely a small error in the amounts of money spent in each category due to price changes throughout the year, seasonality of products, and discounts used. We incorporate this a priori knowledge by generating synthetic samples from the real dataset distribution with a small, normally distributed, additive random noise applied to all the *amounts* and *shares* categories (for both purchase locations and purchase categories).

2) DOMAIN-LIMITING PROPERTIES

There are attributes whose properties limit the values they can take, and this information can be embedded in data. A data analyst can define domain-limiting rules by setting the boundaries within which attributes can take values and within which variance is significant. With regards to values, a property that defines the numerical range of an attribute's values can be applied by removing all records in the dataset that do not meet the property, and then 0-1 normalizing the values according to the range in the property. What distinguishes this task from traditional normalization is the definition of properties independently from the recorded data: in traditional 0-1 normalization, the range values are taken from the actual min and max values registered in the dataset, whereas in our semantic augmentation framework they must be manually defined, and can ''override'' registered values by dropping the records that do not satisfy set properties.

On a similar note to class-invariant transformations, a property can be defined as the (in)precision to which variance in the values of a sample is insignificant. This is another form of knowledge injection because it allows a data analyst to instruct a model to consider two close values to be the same (which is sometimes not the case, especially with deep learning models). For a practical example, the velocity of a car can likely be expressed in miles per hour without any decimals, even if it was recorded with decimals. Removing the decimal part is equivalent to giving a model the knowledge that the decimals in the speed of a car are insignificant.

IV. ATTRIBUTE INFERENCE MODELS

For attribute inference attacks we train four machine learning models to predict sensitive information from the shopping statistics of customers. The attributes to predict are age, education level, income, number of children at home, and marital status.

70% of the available samples (1568 records) have been randomly selected for training, leaving the remaining 30% for testing. All input features for the models are 0-1 normalized based on the min and max values found in the training set. For each parameter map for a given model, training is performed following a 5-fold cross-validation procedure on the train set, and the average of the 5 runs is used for deciding which configuration is best. All models are evaluated using the Mean Squared Error (MSE) evaluation metric, computed on the left out set from the 5-fold split in each run, whereas the training loss function is dependent on the hyperparameter for each run.

Following is a brief description of the different models used for the inference attacks, with an intuitive explanation for the choice of their configuration parameters. Table 1 shows a summary of the models and parameters used.

A. GRADIENT BOOSTING

Gradient boosting is a machine learning model used for both classification and regression tasks, a technique to convert many weak learners into a single strong one. It is built as an ensemble of many weak prediction models, such as decision trees (in which case, the model is referred to as a gradient-boosting tree), in a staged fashion. At each iteration, a new weak classifier is fit to a subset of the training set.

TABLE 1. Summary of the parameters used for the four proposed ML-based attribute inference models. Bold values are those that yield the best average accuracy across all attributes.

Gradient Boosting		Random Forest	
Loss	**MSE**, MAE	Criterion	**MSE**, Poisson
Estimators	20, 100, **200**	Estimators	20, **100**
Subsample	1, **0.8**	Min Samples	**3**, 0.1, 0.01
Deep Neural Network		**Support Vector Regression**	
Solver	**Adam**, lbfgs	Loss	**L1 (MAE)**
Iterations	150, **500**	C	0.1, 1, **10**
Topology	[40,20], **[40,40,40]**	ϵ	0.2, **0.1**, 0
Activation	tanh, **ReLU**		
Learning Rate	10^{-4}, $\mathbf{10^{-3}}$		

In our experiments we explore three hyperparameters for the gradient boosting algorithm, as follows:

- *criterion* is the loss function used to evaluate the model's fit to the train data. We use Mean Absolute Error (MAE) and MSE as criteria. The first takes an average of all the differences between the predicted and true target values, while the second takes the mean of their squares. MAE is more robust to outliers, and MSE includes information on both the variance and the bias of an estimator.
- *n_estimators* is the number of weak classifiers used in the sequential tree structure that makes the gradient boost machine.
- *subsample* is the relative size of the subsets of train samples used for training each tree.

B. RANDOM FOREST

Random Forests are ensemble models similar to Gradient Boosting, but with a few key differences. As gradient boosting, a forest uses many weak tree models to make one strong classifier (or regressor). The uncorrelation between the individual weak classifiers is what makes them useful as a whole, as each weight features differently and the errors of the few are corrected by the accuracy of the many. Contrarily to Gradient boosting, which is built and used sequentially, random forests train and use trees concurrently, and the mean of the predictions of the weak classifiers is used for the overall output.

For random forests we experiment with changing the following hyperparameters:

- *criterion* is the fitness function that the model uses to determine the best split at a node. Squared error tries to minimize the errors between predictions and outputs, whereas with Poisson the model looks for the split that most reduces the Poisson deviance.
- *n_estimators* is the number of trees in the forest.
- *min_samples_split* is the minimum number of samples that must be in a node to make that node eligible to be split; decimal values indicate a percentage of the dataset, while integer values indicate exactly how many records are necessary for a node to be considered for splitting.

C. SUPPORT VECTOR REGRESSION

Support Vector Machines (SVMs) are supervised machine learning models for classification and regression tasks. For the simplest task of binary classification, an SVM learns a mapping to a feature space in which it finds a plane that separates the mapped samples in two groups. Finding wide planar gaps between the two categories is how the model better fits the training data. SMVs can be used for regression tasks too and maintain all features that characterize them, such as the maximal margin method: in this case they take the name Support Vector Regressor (SVR). SVRs use a parameter ϵ to modulate how closely the model fits to the training data by ignoring all errors equal to or smaller than ϵ.

We explore three hyperparamters for the support vector machine regression model:

- *epsilon* controls how much error can be ignored for each training sample.
- *kernel* is the type of kernel used by the SVM.
- *loss* is the loss function used for calculating the training error.
- *C* is the regularization parameter, which shapes how punitive is the penalty for the training error. A large value of *C* makes the margin separating values on a plane smaller, so that more training samples are classified correctly, but may not perform well in testing; the opposite holds true for smaller *C* values.

D. DEEP NEURAL NETWORK

A kind of artificial neural network (ANN), deep neural networks (DNNs) have multiple network layers between the input and output. All DNNs are built with the same set of components: neurons, synapses, weights, biases, and loss functions. These components make DNNs similar to the human brain and allow them to be trained as a traditional machine learning model.

We explore the following hyperparameters:

- *solver* is the method used to update the weights. Adam is a stochastic gradient descent optimizer that computes an exponential moving average of the gradients (with decay rates) with respect to the parameters separately and updates them accordingly. With limited-memory BFGS (L-BFGS) a matrix of second partial derivatives (an inverse Hessian matrix) is computed to describe the loss function around a point in the model's parameters space. This memory-limited version only approximates the matrix, which is particularly expensive to compute, by calculating only small sections of it in a sequential manner.
- *max_iter* is the maximum number of training iterations/epochs allowed to the model.
- *hidden_layer_size* controls the internal structure of the neural network. An example of network topology with two hidden layers of size 10 and 5 is [10, 5].
- *activation* is the activation function used in all the layers (but the last) the model. *tanh* maps any value to a $[-1, 1]$

range, whereas *ReLU* sets all negative values to 0 (and so has oututs in range $[0, \infty]$).
- *learning_rate_init* is the learning rate used by the solver.

V. EVALUATION

Attribute inference attacks are executed to discover the quasi-identifiers of customers that we suppose a data manager would suppress. This section provides the accuracy of the models described in the previous section, and we compare them to two baselines, showing that even a dumb classifier performs better on augmented data. Evaluation is carried out on the 30% of records left out from the training set.

The accuracy of a model for an attribute is computed differently depending on the nature of the attribute. For *education*, it is important that a prediction is as close to the true label as possible. We, therefore, use the following equation to measure the accuracy with respect to a single sample, where y is the true label of a sample and $\phi(x)$ is the model's prediction for that same sample:

$$A_{education} = 1 - |y - \phi(x)| \tag{1}$$

For the *income* and *age* attributes we use the relative error because the same absolute prediction error is less important for higher true values than it is for lower ones, and the attribute is continuous in nature. The accuracy is therefore defined as:

$$A_{income} = 1 - \frac{|y - \phi(x)|}{y} \tag{2}$$

The accuracy for the two binary attributes and the two remaining numerical attributes is computed by checking whether the predicted values match the true values or not. Before making this check, numerical values are de-normalized and rounded to integers.

$$A_{others} = \begin{cases} 1 & \text{if } y = \phi(x) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Following is a description of the baseline methods used, and then the results are discussed.

Random Baseline: The first baseline proposed makes predictions at random. In the case of binary attributes, its accuracy is equal to 50%. For continuous-valued attributes, namely income, its accuracy depends on the distributions of target values themselves.

Statistical Baseline: A more informed baseline is also discussed. Instead of guessing targets at random, the mode or mean of each attribute is used for all outputs (depending on which one is applicable). The accuracy then becomes dependent on the distribution of the targets, and in binary attributes, it is equal to the percentage of samples in the majority class.

A. RESULTS

This subsection explores the performance of the various proposed methods, including baselines and ML-based ones, to understand the usefulness of our semantic augmentation method.

TABLE 2. Accuracy of inferring target attributes with the *statistical model* on data before and after augmentation.

	Original	Augmented
Education	52%	**74%**
Income	**56%**	55%
Age	21%	**22%**
Children at home	56%	**57%**
Is single	35%	**66%**
Has married		**50%**

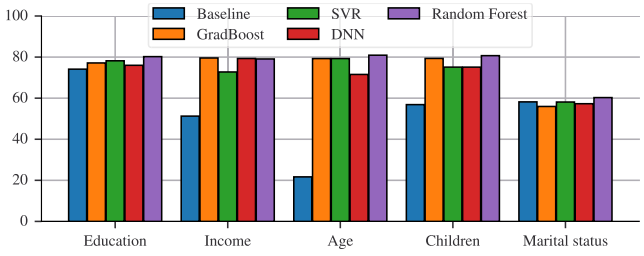

FIGURE 3. Comparison of the accuracy of the proposed models in predicting the sensitive attributes suppressed from the store customers dataset.

1) SEMANTIC AUGMENTATION EFFECT

We first show how semantically augmenting samples has an effect on the accuracy of a baseline classifier. Table 2 shows the prediction accuracy for the various target attributes of the statistical baseline model on the original and on the augmented test set. Significant improvements can be seen for *education* and *marital status/single + married* attributes, which are the ones that have been augmented the most. On average, the accuracy of the statistical model improved from 43% to 53% after augmentation.

2) ATTRIBUTE INFERENCE ATTACKS

The graph in Figure 3 compares the prediction accuracy of the proposed machine learning models against the statistical baseline model introduced earlier, highlighting the power of our semantic augmentation framework when exploited with machine learning. For each learned model type, the graph shows the results of the one configuration which performed best on average in training with 5-fold cross-validation (see Table 1 to check the best configurations).

Overall, all ML-based models perform better than the two baselines. Depending on the attribute to predict, different models perform best, with a general bias towards gradient boosting and random forests. Random forests, in fact, predict best the *education, children*, and *marital status* attributes, with an accuracy of 80% for the first two and 60% for the latter. *Income* is best predicted by deep neural networks, also with an 80% accuracy, and a support vector regression model is best at predicting the *age* with an 81% accuracy.

Of the four proposed ML-based attribute inference models two are explainable. The graphs in Figure 5 and Figure 4 show how important each feature was in the prediction of the six sensitive attributes for the gradient boosting and

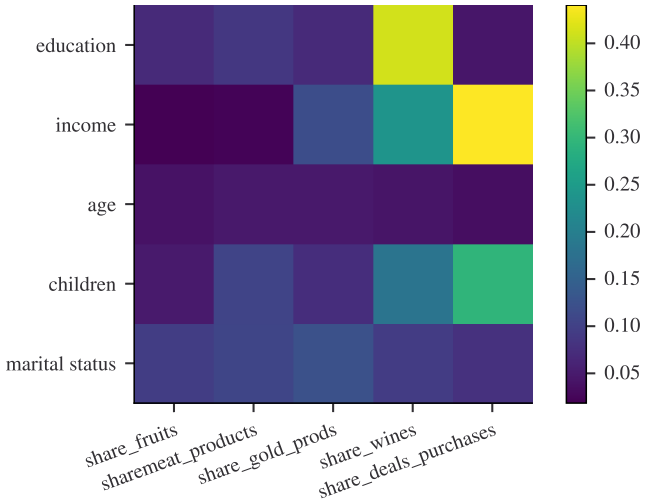

FIGURE 4. **Importance of the input features in determining the sensitive attributes for the best *Random Forest* models.**

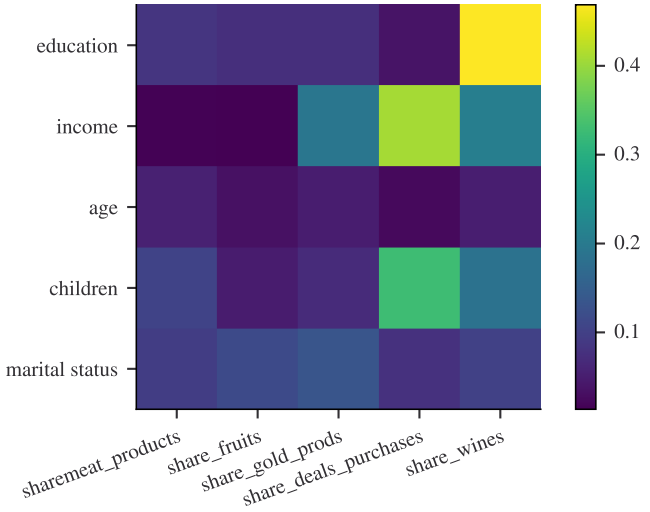

FIGURE 5. **Importance of the input features in determining the sensitive attributes for the best *Gradient Boosting* models.**

random forest models. The attributes resulting from the *value semantics injection* augmentation are consistently the most important in making predictions, further proving how semantically augmenting the data with our model has a positive impact on the amount of information that can be extracted from it.

VI. CONCLUSION

In general, information about the marital status of the store's customers proved to be more difficult to infer, while the number of children in the household, the education level, and the income were the easiest to predict, with an accuracy of around 80%. This level of precision in determining the personal, sensitive information of users from an anonymized dataset calls for action in studying new anonymization techniques that can more robustly ensure privacy.

Potential mitigation methods that could promote higher privacy levels can be looked for in the fields of adversarial machine learning and statistics. A possible statistical method could be to approximate which attributes, values or records contribute more to the inference accuracy of ML models (similarly to how we show in Figures 5 and 4), and reduce the variance for those attributes or values. A similar approach may be taken with Adversarial ML. A model can be trained to learn a feature mapping from the dataset attributes to a new manifold where clustering algorithms can be used to perform inference. Then, such a model could be used to generate adversarial examples to insert in the dataset before publishing it, with the potential result to make it harder for new models to be trained on it. We leave this exploration to future works.

Security and privacy in machine learning are two-sided coins: for every improvement in ML, new defense mechanisms need to be studied to promote privacy, and so for every new potential security threat new ML techniques need development. An additional future work that we are therefore interested in is applying our model to set-valued datasets, which have the potential to be semantically augmented even further.

REFERENCES

[1] S. Kemp. *Digital 2020: Global Digital Overview*. Accessed: Aug. 28, 2023. [Online]. Available: https://datareportal.com/reports/digital-2020-global-digital-overview

[2] J. Wiener and N. Bronson. (Oct. 2014). *Facebook's Top Open Data Problems*. [Online]. Available: https://research.fb.com/blog/2014/10/facebook-s-top-open-data-problems/

[3] E. L. Harding, J. J. Vanto, R. Clark, L. H. Ji, and S. C. Ainsworth, "Understanding the scope and impact of the California consumer privacy act of 2018," *J. Data Protection Privacy*, vol. 2, no. 3, pp. 234–253, 2019.

[4] P. Voigt and A. Von dem Bussche, *The EU general data protection regulation (GDPR): A Practical Guide*, 1st Ed. Cham, Switzerland: Springer, 2017.

[5] A. Praveena and S. Smys, "Anonymization in social networks: A survey on the issues of data privacy in social network sites," *J. Int. J. Eng. Comput. Sci.*, vol. 5, no. 3, pp. 15912–15918, 2016.

[6] S. Murthy, A. A. Bakar, F. A. Rahim, and R. Ramli, "A comparative study of data anonymization techniques," in *Proc. IEEE 5th Int. Conf. Big Data Secur. Cloud (BigDataSecurity), IEEE Int. Conf. High Perform. Smart Comput., (HPSC) IEEE Int. Conf. Intell. Data Secur. (IDS)*, May 2019, pp. 306–309, doi: 10.1109/BigDataSecurity-HPSC-IDS.2019.00063.

[7] S. Lu, M. Zhao, H. Zhang, C. Zhang, W. Wang, and H. Wang, "GenderPredictor: A method to predict gender of customers from e-commerce website," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol. (WI-IAT)*, vol. 3, Dec. 2015, pp. 13–16, doi: 10.1109/WI-IAT.2015.106.

[8] M. Merler, L. Cao, and J. R. Smith, "You are what you tweet...pic! Gender prediction based on semantic analysis of social media images," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jun. 2015, pp. 1–6, doi: 10.1109/ICME.2015.7177499.

[9] A. Chaabane, G. Acs, and M. A. Kaafar, "You are what you like! information leakage through users' interests," in *Proc. 19th Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, San Diego, CA, USA, 2012, pp. 1–14. [Online]. Available: https://hal.inria.fr/hal-00748162

[10] S. J. Baran, J. J. Mok, M. Land, and T. Y. Kang, "You are what you buy: Mass-mediated judgments of people's worth," *J. Commun.*, vol. 39, no. 2, pp. 46–54, 1989.

[11] Y. Wang, X. Pan, S. Song, H. Zhang, G. Huang, and C. Wu, "Implicit semantic data augmentation for deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/15f99f2165aa8c86c9dface16fefd281-Paper.pdf

[12] Y. Wang, G. Huang, S. Song, X. Pan, Y. Xia, and C. Wu, "Regularizing deep networks with semantic data augmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3733–3748, Jul. 2022, doi: 10.1109/TPAMI.2021.3052951.
[13] A. Burton-Jones, V. C. Storey, V. Sugumaran, and S. Purao, "A heuristic-based methodology for semantic augmentation of user queries on the web," in *Proc. Int. Conf. Conceptual Modeling*. Berlin, Germany: Springer, 2003, pp. 476–489.
[14] M. Wende, "The Bologna declaration: Enhancing the transparency and competitiveness of higher education," *Higher Educ. Eur.*, vol. 25, pp. 305–310, Jan. 2000, doi: 10.1080/713669277.
[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[16] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," 2013, *arXiv:1310.4546*.
[17] A. Wyk. (Apr. 2018). *Encoding Cyclical Features for Deep Learning*. [Online]. Available: https://www.avanwyk.com/encoding-cyclical-features-for-deep-learning/

THOMAS CILLONI (Member, IEEE) received the first bachelor's degree in computer science from Xi'an Jiaotong–Liverpool University, the second bachelor's degree in computer science from the University of Liverpool, and the Ph.D. degree in computer science from the University of Mississippi. He is currently an Applied Scientist at Amazon. His research interest includes security and privacy in machine learning, focusing on data privacy, adversarial ML, and computer vision security.

CHARLES FLEMING (Member, IEEE) received the Ph.D. degree in computer science from the University of California at Los Angeles (UCLA), in 2013. He worked in academia with Xi'an Jiaotong-Liverpool University and the University of Mississippi, University, MS, USA, before joining Cisco, in 2022. He works broadly in security and privacy, with a special interest in machine learning applications and adversarial machine learning. He also dabbles in usable security, particularly for VR/AR applications.

CHARLES WALTER (Member, IEEE) received the bachelor's, master's, and Ph.D. degrees from The University of Tulsa. Before coming to Ole Miss, he was a Postdoctoral Researcher at his alma mater. He is currently an Assistant Professor of computer and information science with the University of Mississippi. His research interests include wireless security, wearable devices, human–computer interaction, computer science education, software engineering, and self-adaptive systems.