Subgroup Preservation in Financial Data Anonymized by a Variational Autoencoder

Samuel C. Maina
IBM Research – Africa
Nairobi, Kenya
samuelm@ke.ibm.com

Reginald E. Bryant
IBM Research – Africa
Nairobi, Kenya
bryantre@ke.ibm.com

William Ogallo
IBM Research – Africa
Nairobi, Kenya
william.ogallo@ibm.com

Kush R. Varshney
IBM Research – Africa
Nairobi, Kenya
krvarshn@us.ibm.com

Skyler Speakman
IBM Research – Africa
Nairobi, Kenya
skyler@ke.ibm.com

Celia Cintas
IBM Research – Africa
Nairobi, Kenya
celia.cintas@ibm.com

Aisha Walcott-Bryant
IBM Research – Africa
Nairobi, Kenya
awalcott@ke.ibm.com

Robert-Florian Samoilescu
IBM Research – Africa
Nairobi, Kenya
robert-florian.samoilescu@ibm.com

Abstract

Many data-driven decision-making tasks performed by financial service institutions rely on market segmentation and the discovery of anomalous subgroups in individual-level data. Given that privacy is a critical consideration, we investigate the effect of anonymization on anomalous subgroup discovery. In particular, we train a binary classifier on a real world financial dataset to model a binary outcome (product uptake). We then identify the most anomalous subgroup in the original data by maximizing the bias between the group's predicted odds ratio from the model and the observed odds ratio from the data. We then perform anonymization using a variational autoencoder to synthesize an entirely new data set but, ideally, drawn from the distribution of the original, real-world data. Finally, we repeat the anomalous subgroup detection task on the new data and compare to what was identified pre-anonymization.

1 Introduction

Banks, insurance companies, and other financial services providers along with their partners, vendors and contractors are increasingly relying on data-driven and machine learning-based approaches to perform a variety of functions. These functions include shaping marketing strategies [1], designing bespoke portfolio and insurance products and services [2], managing risk [3], detecting fraud [4], and complying with anti-discrimination or fairness regulations [5]. A common machine learning objective in these tasks is to discover customer segments or subgroups in individual-level data, defined by demographic, psychographic, behavioural or socioeconomic variables, that are interesting or anomalous according to some criterion [6].

While working with individual-level data, privacy is an important consideration and is protected by laws such as the Gramm-Leach-Bliley Act in the United States [7] and the General Data Protection Regulation [8] in Europe. Failure to appropriately protect customer data exposes companies to significant reputational and legal risks. Data can be misused by bad actors within a company, stolen

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

by cyber-criminals or inappropriately shared with third parties. This can lead to direct financial losses from fraud claims or as a result of regulatory fines.

Anonymization seeks to protect private and sensitive information of a client and or their activities. This allows the data to be used for a variety of functions such as the ones mentioned in the first paragraph by the institution itself or by its partners. When sharing data with privileged partners, stakeholders or regulators, most institutions use popular anonymization techniques (e.g. removal, redaction, encryption, data masking, perturbation, and generalization) to safeguard the privacy and secure the sanctity of the data.

Different anonymization techniques distort the dataset in various ways that may be unwanted depending on the downstream task the dataset is to be used for [9]; after all, anonymization without preserving the information content and utility of the data is not desired. In this paper, our interest is in examining the extent to which anonymization techniques preserve interesting subgroups of individuals. Therefore, our utility or figure of merit for an anonymization technique (besides its ability to prevent deanonymization attacks) is its ability to yield the same or similar most-anomalous subgroup.

Recent advances in generative machine learning such as variational autoencoders [10] (VAEs) and generative adversarial networks [11] (GANs) are starting to be applied to the anonymization problem [12, 13, 14, 15, 16, 17, 18]. The main idea of the generative approach is to learn the salient characteristics of the data distribution and sample new (synthetic) individuals from the distribution. Synthetic data from the generative models retains the properties of the original data and can therefore be used as a proxy and be shared without risks of re-identification or information leakage [14].

We investigate how VAE-based synthesis techniques fare with respect to preserving subgroup properties. We do not pre-specify the subgroups but instead we use a scanning method that examines various combinations of input attributes to identify the specific subgroup which is consistently misclassified at a rate greater than that of the entire dataset. In particular, we use the *Bias-Scan* method [19] for anomalous subgroup discovery. It is a linear-time approach that can be used with large-scale data. We apply Bias-Scan on the original data as well as the synthesized data coming from a VAE and investigate the proportion of the subgroup that was preserved.

In the empirical section, we use the *Bank of Portugal Dataset*, a labeled marketing-campaign dataset consisting of 10 customer-attributes (a portion of the original 150) with the binary outcome of the acceptance of an offered long-term bank deposit product [20]. Our results indicate two things of note. First, from the perspective of privacy, the data synthesis procedure does not preserve identities of the most anomalous individual records as shown by the high Jaccard distance index values indicating that the VAE performs quite well in anonymizing the original dataset. Second, from the perspective of utility, the transformation process yields mid-value Jaccard distance values (about 0.5) for attribute-value overlap indicating that to a good extent, the overall statistical properties of the data are preserved.

The outline of the remainder of the paper is as follows. In Section 2, we give a brief introduction to VAEs and the Bias-Scan methods as used in the paper. In Section 3, we present details about the data and the experimental setup. In Section 4, we present the empirical results and discuss the key findings. Section 5 concludes the paper and offers some thoughts on future work.

2 Background

In this section, we briefly describe the theoretical and algorithmic frameworks that we use for the purposes of this experiment; namely the subset scan method for identification of systemic bias in the dataset and the VAE technique for generating synthetic data.

2.1 Variational Autoencoders

Autoencoders are generative models which are designed to capture the underlying distribution of input data and reproducing it from its essential underlying attributes. These essential features are lower order representations of the data determined within the internal structure of the autoencoder. This lower-order or compressed representation is termed the *latent space*. Note that as information about the original data is compressed, the output of autoencoders are not fundamentally the same as the input–the output only captures certain aspects of the original data. VAEs take this notion of inexact replication of the original data to another level. VAEs are designed to produce variations of the input data, thus creating data that didn't exist, but could have existed based on the underline statistics of that input data. Typically used for image data reproduction, we are focused on using VAEs to reproduce tabular data as a way to represent the underlying statistics of the input data, thus establishing a data protection mechanism to ensure privacy.

2.2 Bias-Scan Algorithm

This work takes the "subset scanning" approach to detecting bias in binary classifiers [19]. This view treats the task as a search problem with the goal of finding the subpopulation that is the most systematically over- (or under-) risked by the classifier. This is done by efficiently maximizing a score function $F(S)$ (a likelihood ratio from the Bernoulli distribution) over the exponentially-many subgroups. The efficient maximization is enabled by the Linear Time Subset Scanning property (LTSS) of the scoring function.

The categorical features of the dataset create a multi-dimensional tensor with each feature being a mode in the tensor and each record falling into one of the "cells". Bias-Scan [19] identifies an "axis-aligned" subset of these cells such that the records in the cells maximize the scoring function. An example of axis-aligned subsets that span three modes of a tensor would be: Low or Middle income, Black or Hispanic, Females. Bias-Scan iteratively optimizes over each mode of the tensor until convergence to a local maximum is found. Exploiting the LTSS property of the scoring function guarantees that each optimization step over a mode in the tensor is done efficiently and exactly. However, the joint optimization over all modes depends on the order in which the modes are optimized, and therefore multiple restarts are used to help explore the space and reach a global maximum. Consider a feature with 6 possible attribute values. An exhaustive search over this mode would require 2^6 possible combinations. However, this maximization can be done by only scoring 6 combinations while still guaranteeing that the optimal subset will be found.

The scoring function efficiently maximized in Bias-Scan is:

$$score_{bias}(S) = max_q log(q) \sum_{i \in S} y_i - \sum_{i \in S} log(1 - \hat{p_i} + q\hat{p_i})$$

This function is derived from a likelihood ratio of the Bernoulli distribution that operates on a subset of records in the data. The null hypothesis assumes the binary outcomes (y_i) are drawn correctly from the binary classifier's predicted probability for the records in the subset. The alternative hypothesis assumes that the binary outcomes of the records in the subset are generated by a different probability than predicted by the classifier. In particular, the alternative hypothesis assumes that the odds $\frac{p}{1-p}$ have been increased by a multiplicative factor $q > 1$. The end result of efficiently maximizing this scoring function over all possible "axis aligned" subsets of the data is identifying the subgroup that has the most evidence of being generated by the alternative hypothesis stated above. That is, those records have the largest divergence between what the binary classifier predicted for their outcomes and what was actually observed.

The information contained in such a subgroup can provide unique insights into the dataset. A bespoke financial product can be designed and marketed that takes into consideration the unique traits of the sub-population. For example; [21] provide a review on the application of AI on healthcare data and suggested how this can be used to identify groups that will benefit from tailored premium rates.

3 Experimental Setup

We use data from a direct marketing campaign of a Portuguese banking institution. This data was collected from May 2008 to November 2010 and consists of 41,188 records with 20 labeled attributes. This data describes whether a customer will accept a long-term bank deposit account (binary target label) [20]. Our preliminary experiment with this dataset takes a subset of the categorical attributes with both nominal and ordinal values. Specifically, we performed experimental evaluation on a filtered dataset that included the following 10 discrete attributes (listed with and their cardinality): *Job* (12); *Martial status* (4); *Education* (8); *Default* (2); *Housing* (3); *Loan* (3); *Method of Contact* (2); *Month* (10); *Day of Week* (5); and *Outcome of Previous Marketing Campaign* (2). The final dataset consists of $37,069$ records, 10 attributes, and 1 binary target label.

The procedure of the experiment is as follows. First, we use the original data to train a Random Forest (RF) predictive classifier and an XGBoost predictive classifier. Each classifier's parameters were optimized by cross-validation, with model performance measured using Area Under the ROC Curve (AUC). The classifiers were used to generate the predicted (calibrated) probabilities of a respondent taking up the proposed term deposit product. Second, we apply the Bias-Scan algorithm to identify the single subgroup of attribute space with the highest bias score. This is the subgroup of data that is most anomalous with regard to the predicted and the observed outcome. The bias score, therefore, reflects the extent to which the classifier makes the prediction error for the observations in this subgroup.

Third, following [22], we train a standard categorical variational autoencoder (VAE) to generate new samples of the of the original dataset. In our case, we use Adaptive Moment Estimation (ADAM) [23] as the optimization method, which computes adaptive learning rates for each parameter with a $LR = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$. Here, the loss function is the reconstruction loss added to the K-L divergence. The input shape of the vectors varies depending on the dataset; all variables were encoded using one-hot encoding. We ran several simulations of the VAE on the original data to obtain 6 different sets of the synthetic data samples. For each of the 6 synthetic data samples, we trained a RF and an XGBoost model that was then used to obtain the respective predictive probabilities for each new dataset using the binary classifiers. Fourth, we applied the Bias-Scan algorithm to each of the 6 synthetic dataset samples to identify the most anomalous subgroup (subgroup with highest bias score).

Lastly, for each classifier, we evaluated the pairwise dissimilarity between the individual records belonging to the most anomalous subgroup in the original dataset and the individual records belonging to the most anomalous subgroup in the 6 VAE synthetic samples. We also evaluated the pairwise dissimilarity between the most anomalous subset of attribute values in the original dataset and the most anomalous subset of attribute values in the 6 VAE synthetic samples for each classifier. We quantified pairwise dissimilarity between sets of subgroups using the Jaccard distance, $d_{X,Y}$, defined as a complement of the Jaccard coefficient and obtained by subtracting the Jaccard coefficient from 1:

$$d_{X,Y} = 1 - \frac{X \cap Y}{X \cup Y}$$

where X and Y are sets of discrete elements and $0 \leq d_{X,Y} \leq 1$ with a higher values implying greater dissimilarity.

4 Empirical Results and Discussion

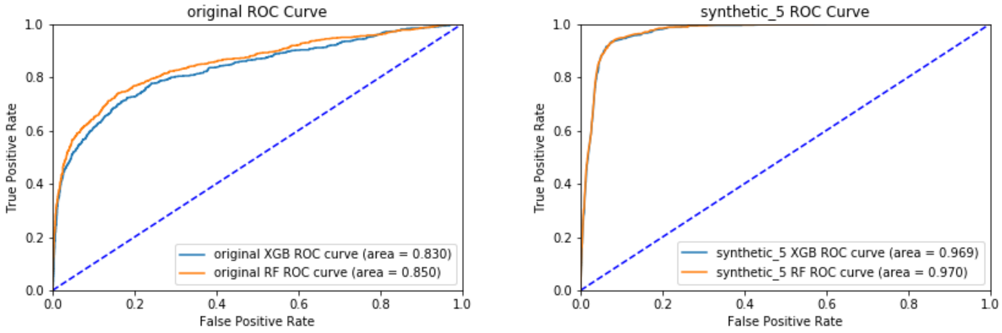
As shown in Table 1, we observe that the two predictive classifiers have relatively high AUC scores both the original data and the 6 synthetic samples. For example, Figure 1 illustrates the ROC curves for both for the original dataset and a sample VAE generated dataset under both classifiers and suggests, comparable performance of the random forest and XGBoost models for each dataset.

We also investigate the overlaps between the anomalous individual observations between the original data and the VAE synthetic datasets to infer the impact of the transformation. The results are shown in Table 2. The Jaccard distances for individual records indicate little to no overlap between individual records of the subgroup most affected by the predictive bias of Random Forest and XGBoost classifiers. The Jaccard distances for the attribute values suggest considerable overlap between subsets of the attribute space that are most affected by predictive bias.

This observation is exemplified in Figure 2 which compares the overlap between the most anomalous subgroups in the original dataset and a sample VAE generated synthetic dataset under the random forest classifier. As shown in Figure 2A, there is no overlap (Jaccard distance = 1) between the individual records of the subgroup most affected by predictive bias of the classifier as measured by the Jaccard distance metric which shows values very close to one. This suggests that the VAE performs relatively well in data anonymization and re-identification of individuals may be hard. However, we observe that for the attribute (feature) values, there is a significant overlap (Jaccard distance = 0.462) between the most anomalous subgroups of the model as shown in in Figure 2B. This implies that the characteristics of subgroups of interest in a dataset would not be lost when the data is transformed using VAEs.

model	original	synth_0	synth_1	synth_2	synth_3	synth_4	synth_5
RF	0.85	0.779	0.951	0.946	0.915	0.900	0.970
XGBoost	0.83	0.789	0.952	0.949	0.914	0.903	0.969

Table 1: Area under curve (AUC) scores for the Random Forest and XGBoost predictive models on the original Bank of Portugal dataset and six synthetic VAE generated datasets.

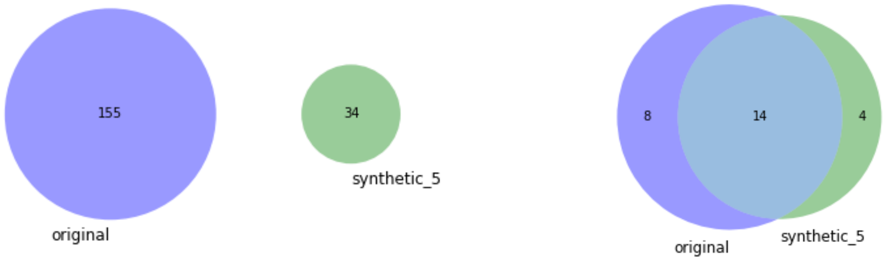


(A) (B)

Figure 1: Receive Operating Characteristic (ROC) curves for Random Forest and XGBoost classifiers trained on the original Bank of Portugal dataset (A) and a sample VAE synthetic dataset (B).

category	model	synth_0	synth_1	synth_2	synth_3	synth_4	synth_5
Individual Records	RF	0.965	0.949	0.959	0.949	0.947	1.000
	XGB	0.953	0.958	0.958	0.975	0.951	1.000
Attribute Values	RF	0.571	0.481	0.516	0.469	0.438	0.462
	XGB	0.556	0.536	0.543	0.545	0.441	0.581

Table 2: Dissimilarity, quantified by the Jaccard distance, between subgroups in the original datatet vs 6 VAE synthetic samples that are most affected by bias under different setups. Higher values imply greater dissimilarity



(A) (B)

Figure 2: (A) shows that for a Random Forest classifier trained on the original Bank of Portugal data and on a synthetic data generated by a variational autoencoder (VAE) there is zero overlap (Jaccard distance = 1) between individual records of the subgroup most affected by predictive bias of the classifier. (B) shows that for the same setup, there is considerable overlap (Jaccard distance = 0.462) between attributes values of the subsets of the original and synthetic data that are most affected by predictive bias.

	Random Forest		XGBoost	
attribute	**original**	**synthetic_5**	**original**	**synthetic_5**
contact	cellular	cellular	cellular	cellular
day_of_week	tue wed thu fri	tue fri	tue wed thu fri	tue fri
default	no	no	no	
education	basic.9y high.school university.degree	basic.9y high.school university.degree professional.course	high.school university.degree unknown	basic.9y high.school university.degree professional.course unknown
housing	no	no yes	no	no yes
job	services management admin. unknown	admin. blue-collar	services management admin. Retired housemaid unknown	admin. blue-collar
loan	no	no	no	no
marital	married single	married single	married single	married single divorced
month	apr jul aug nov	apr aug	apr jul aug nov dec	apr aug
poutcome	nonexistent	failure	nonexistent	failure

Table 3: Subsets of attribute values of original and a VAE synthetic sample that are most affected by predictive bias of Random Forest and XGBoost models.

Lastly, shown in Table 3 are the subsets of the original data and a VAE synthetic sample that are most affected by predictive bias of the two classifiers used as identified by the Bias Scan method. These findings illustrate the ability of the Bias Scan methodology to efficiently identify nuanced subgroups that are most adversely affected by the predictive bias of a model. As would be expected, the results suggest that the same classifier trained on original data and VAE synthetic data result in identifying different subgroups that are most affected by bias.Similarly, given the same dataset, different classifiers yield different subgroups that are most anomalous.

5 Conclusion

In this paper we have presented the standard variational autoencoder (VAE) as a tool for anonymization and utility preservation. In particular, we have demonstrated that this model achieves anonymization on individual records which allows for data reuse and sharing amongst the key stakeholders and clients. This is confirmed by the high Jaccard distance values on the overlaps for the anomalous subsets of individual records for the original and synthetic datasets after the data has been run through a classifier and the Bias-Scan algorithm. We have also shown that the VAE synthesizer preserves some of the ‘global’ statistical distributional properties of the data as demonstrated by the mid-level values of the Jaccard distance metric for the feature values. We therefore conclude that the data transformation preserves (to an extent) some key subgroup properties of the data.

A potentially fruitful direction for further extensions of this work could be on applying a methodology that attains higher subgroup overlaps between the two datasets. An initial recommendation is to modify the standard VAE for data generation to a case where we incorporate constraints on how each attribute can vary with respect to one another when producing synthetic data from the VAE. One can also pursue other generative approaches such as Generative Adversarial Networks [11, 14, 24] to synthesize tabular data and check for consistency values between different attributes while preserving the privacy of individual records.

References

[1] Ning Sun, Jacqueline G. Morris, Jian Xu, Xiufang Zhu, and Ming Xie. iCARE: A framework for big data-based banking customer analytics. *IBM Journal of Research and Development*, 58(5/6):4, September/November 2014.

[2] Chris Lamberton, Damiano Brigo, and Dave Hoy. Impact of robotics, RPA and AI on the insurance industry: Challenges and opportunities. *Journal of Financial Perspectives*, 4(1):8–20, May 2017.

[3] Bart van Liebergen. Machine learning: A revolution in risk management and compliance? *Journal of Financial Transformation*, 45:60–67, 2017.

[4] Anuj Sharma and Prabin Kumar Panigrahi. A review of financial accounting fraud detection based on data mining techniques. *International Journal of Computer Applications*, 39(1):37–47, 2012.

[5] Matthew Adam Bruckner. The promise and perils of algorithmic lenders' use of big data. *Chicago-Kent Law Review*, 93(1), 2018.

[6] Stefan Rueping. Ranking interesting subgroups. In *Proceedings of the International Conference on Machine Learning*, pages 913–920, June 2009.

[7] Gramm-Leach-Bliley Act (GLBA). `shorturl.at/rIOV4`.

[8] General Data Protection Regulation (GDPR). `https://eugdpr.org`.

[9] Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Distribution-preserving k-anonymity. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 11(6):253–270, 2018.

[10] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[12] Yujia Li Max Welling Richard Zemel Christos Louizos, Kevin Swersky. The variational fair autoencoder. In *Proceedings of the International Conference on Learning Representations*, 2016.

[13] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *Proceedings of the International Conference on Learning Representations*, 2016.

[14] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment*, 11(10):1071–1083, 2018.

[15] Reginald Bryant, Celia Cintas, Isaac Wambugu, Andrew Kinai, Abdigani Diriye, and Komminist Weldemariam. Evaluation of bias in sensitive personal information used to train financial models. In *Proceedings of the 7th IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2019.

[16] Ho Bae, Dahuin Jung, and Sungroh Yoon. Anomigan: Generative adversarial networks for anonymizing private medical data. *CoRR*, abs/1901.11313, 2019.

[17] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. DeepPrivacy: A Generative Adversarial Network for Face Anonymization. *arXiv e-prints*, page arXiv:1909.04538, Sep 2019.

[18] Samuel Ainsworth, Nicholas J. Foti, Adrian K. C. Lee, and Emily B. Fox. Interpretable vaes for nonlinear group factor analysis. *CoRR*, abs/1802.06765, 2018.

[19] Zhe Zhang and David Neill. Identifying significant predictive bias in classifiers. *2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017)*, 2017.

[20] Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 2014.

[21] David Thesmar, David Sraer, Lisa Pinheiro, Nick Dadson, Razvan Veliche, and Paul Greenberg. Combining the power of artificial intelligence with the richness of healthcare claims data: Opportunities and challenges. *PharmacoEconomics*, 37(6), Jun 2019.

[22] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

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

[24] Ashutosh Kumar, Arijit Biswas, and Subhajit Sanyal. ecommercegan: A generative adversarial network for e-commerce. *arXiv preprint arXiv:1801.03244*, 2018.