bank-account-fraud: Tabular Dataset(s) for Fraud Detection under Multiple Bias Conditions

Sérgio Jesus
Feedzai / Universidade do Porto
sergio.jesus@feedzai.com

José Pombal
Feedzai

Duarte Alves
Feedzai

André F. Cruz
Feedzai

Pedro Saleiro
Feedzai

Rita P. Ribeiro
Universidade do Porto

João Gama
Universidade do Porto

Pedro Bizarro
Feedzai

Abstract

Evaluating new ML techniques on realistic datasets plays a crucial role in the development of ML research and broader adoption by practitioners. Furthermore, with the growing ethical concerns around the potential of bias in algorithmic decision-making, fairness evaluation is becoming a standard practice in ML. However, while there has been a significant increase of publicly available unstructured datasets for computer vision and NLP tasks, there is still a lack of large-scale domain-specific tabular datasets, which hinders potential research applied to this particular domain. Additionally, many high-stakes decision-making applications, where an accurate evaluation of algorithmic fairness is of paramount importance, rely on this type of data. Ultimately, this affects the quality of the deployed models in critical applications, and consequently, automated decisions applied to people. To tackle this issue, we present `bank-account-fraud`, the first publicly available, large-scale, and privacy-preserving suite of tabular datasets for fraud detection. The suite was generated by applying state-of-the-art tabular dataset generation techniques on an anonymized, real-world bank account opening application dataset. This setting carries a set of challenges that are commonplace in real world applications, including distribution shifts in time and significant class imbalance. Additionally, to allow practitioners to *stress test* both performance and fairness of ML methods in dynamic environments, each dataset variant of the presented suite depicts a specific predetermined and controlled type of data bias, including time-related patterns. With this dataset, we hope to potentiate ML research, through a more realistic, complete and robust test bed for novel and existing ML methods.

1 Introduction

The ability to collect and handle large-scale data has laid the foundations for the widespread adoption of Machine Learning (ML) [1, 2]. Regardless of the application, evaluating new ML techniques on realistic datasets plays a crucial role in the development of ML research, and subsequent adoption by practitioners [3] [4]. Additionally, with the growing ethical concerns around the potential of bias in algorithmic decision-making [5, 6, 7], fairness evaluation is becoming a standard practice in ML [8, 9, 10]. However, while the vast majority of publicly available datasets are directed to computer vision and NLP tasks, there is a scarcity of large-scale domain-specific tabular datasets. The latter are the centerpiece of most high-stakes decision-making applications, where fairness testing is of

Submitted to the 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks. Do not distribute.

paramount importance. As it stands, the most relevant tabular datasets in the fair ML literature suffer from a series of limitations [11, 12, 13], which we will detail in Section 2. Furthermore, most real-world settings are dynamic, featuring temporal distribution shifts, class imbalance, and other phenomena that are not reflected in most of the datasets in fair ML literature [14]. We will discuss how the `bank-account-fraud` suite of datasets tackles these limitations, and outline its utility as a general-purpose tool for the evaluation of performance and fairness in dynamic environments.

What is a good dataset for ML practitioners?

In general, good datasets for ML benchmarks are ones that are representative of the distribution and dynamics of some target population, and that, symbiotically, are useful to train ML models for a given task. Large-scale datasets based on real-world use cases achieve both goals, as they contain a wide variety of observations, and findings from benchmarks conducted on them are considered to be sufficiently generalizable to real tasks [15].

Adding to these characteristics, a key aspect of a dataset for fair ML is the context of the task: high-impact domains, where decisions produced by an ML system have substantial consequences on the lives of the decision subjects, are strongly preferred [9, 10]. Applications of this nature may be found in the criminal justice, hiring, and financial services domains, among others. Another important aspect for the community is the fidelity of the setting. That is, datasets originating from real-world scenarios are favoured, especially if ML methods were employed. In these cases, the impact of a new method can be measured and compared to other alternatives, or even the original decision-making solution. These measurements can then be translated into real-world solutions, namely making models fairer with respect to a historically discriminated group, for example. Other important components for these datasets include the available protected attributes, privacy, representation, scale, and how recent they are.

What is the current landscape of datasets for fair ML research?

Only a limited amount of datasets are consistently used for validating and benchmarking fairness methods, following a trend of less datasets being used for more often for experimental observations (*i.e.*, *funneling*) observed in the ML community in general [16]. Common issues regarding these datasets are expanded in section 2.1. The relative age of the majority of the datasets used in fair ML, combined with the saturation of tests performed on them, makes the observed results stagnate. These constitute technical considerations for deprecating the dataset [17], and limit any possible validation of novel solutions extracted on these. The lack of quality datasets for fair ML — identified in the 2021 Stanford University's AI Index Report [18] — has prompted the appearance of several initiatives advocating the public sharing of private datasets for decision-making containing protected attributes. Symbiotically, many tools have been recently developed which facilitate the sharing of the data, namely on best practices in documentation and privacy-preserving methods. However, there is still no observable shift between the usage of the more commonplace and older datasets, and the comparatively less explored and updated datasets.

What are the characteristics of the introduced dataset?

The `bank-account-fraud` suite of datasets was generated from a real-world online bank account opening fraud detection dataset. This is a relevant application for fair ML, as model predictions result in either granting or denying financial services to individuals. Each dataset variant in the suite features predetermined and controlled types of data bias over multiple time-steps, obtained by a) sampling the generator with different rates depending on given criteria, and b) appending columns with a distribution depending on properties of the instance. Section 3.3 contains more details about each bias pattern observable in the data. The aforementioned variants, combined with the temporal distribution shifts inherent to the underlying data distribution, amount to a mean for *stress testing* the performance and fairness of ML models meant to operate in dynamic environments.

The datasets on the suite was generated by leveraging state-of-the-art Generative Adversarial Network (GAN) models. One important reason for choosing these methods was to preserve the privacy of the applicants — an ever-growing concern in today's societal and legislative landscape [19]. Each dataset is comprised of a total of one million instances of individual applications, using a total of thirty features. The latter represent observed properties of the applications, either obtained directly from the applicant (*e.g.*, employment status), or derived from the provided information (*e.g.*, whether the provided phone number is valid), and aggregations of the data (*e.g.*, frequency of applications on a given zip code). The data spans eight months of applications, which can be identified in the column ``month''. Regarding protected attributes, the dataset provides the age, personal income, and employment status of the applicant. More details on the dataset are included in Sections 3 and 4, and in the dataset's datasheet[20], provided as supplementary material.

2 Background

In this section, we will first go through the most popular datasets used by the fair ML community, and their shortcomings (Section 2.1). We will also describe privacy-preserving techniques, which play a pivotal role in the development and publication of private datasets, such as our own.

2.1 Shortcomings on Popular Tabular Datasets

Among the datasets used for the benchmark of fairness methods, the UCI Adult dataset [21, 22] stands out as the most popular dataset in the field [11, 23]. Despite its popularity, the dataset has recently been criticized [11, 12], mainly due to three aspects: a) the sampling strategy, based on the poorly documented variable `fnlwgt`, b) the arbitrary choice of task — predicting individuals whose income is above 50,000 dollars — which is not connected to any real census task, and c) the age of the data itself (it is based on 1994 US census data).

The same and other issues are found on a variety of datasets. As an example, the second most popular dataset for fairness benchmarks, COMPAS [6], is afflicted with measurement biases [13], missing values, label leakage [11] and sampling incongruities [24]. Most importantly for the context of the application: the decision-making process where the RAI (Risk Assessment Instrument) is inserted in usually has multiple points of discretion (*i.e.*, different agents, such as juries and ML models, conveying a decision, score or recommendation on the subject) [13], which ultimately render the measurement of fairness of the system based only on a single model's predictions unrealistic. Additionally, one major concern is regarding the privacy of the data, as it is possible to identify accused individuals based on criminal record and other Personal Identifiable Information (PII) [11]. The third most popular dataset is the German Credit dataset [22], which has several documentation issues, including the information regarding what is used as sensitive attribute. Here, the sex of the individual is not retrievable by the ``Personal status and sex'' attribute, as there are overlaps between the possible values. A posterior release of the dataset addresses some documentation errors, but also clarifies that retrieving the applicant's sex through the aforementioned attribute is not possible [25]. This limits the utility of the dataset in the context of algorithmic fairness. Additionally, the dataset is composed of applicants from 1973 to 1975, which hinders the generalization of any insights to today's world. Recently, a study on the datasets used in Machine Learning Research (MLR) identified a funneling tendency in the field, whereby increasingly fewer datasets are being used for benchmarking [16, 11]. These datasets are generally also being used in different tasks than originally intended [16]. Such a trend is also observed in the fairness community, where the previously mentioned UCI Adult dataset [21, 22] was repurposed from its original task [21]. This highlights the necessity of renewing the currently available datasets for fair ML.

2.2 Privacy-Preserving Approaches and Generative Models

A major concern regarding the publication of datasets is the rise of potentially dangerous privacy-breaching applications for the data [26]. This is especially important when considering the field of Responsible AI, where evaluation takes into consideration sensitive attributes of individuals, such as gender, sexual orientation, or religion. To avoid these issues, it is required to either remove, transform, or obfuscate any information that leads to the identification of a particular individual.

One of the more consensual means of evaluation of methods for the purpose of privacy-preservation, is the measurement of differential privacy [27]. This metric determines the maximum difference in an arbitrary measurement or transformation applied to a dataset induced by any individual instance. Lower values of this metric correspond to higher preservation of privacy. Upper-bound levels of this metric are met on several generative models [28, 29]. However, the default implementations of generative models do not take into consideration common problems faced in the tabular data domain. These are mostly caused by having categorical and non-normally distributed continuous variables. One particular architecture that tackles these problems is the CTGAN [30]. This architecture, however, does not have differential privacy guarantees, which is observed in models adapted to the image domain, and constitutes a gap in generative models for tabular data. There is still no consensus on the evaluation of generative models, however [31]. In the computer vision domain, most approaches present a measurement of distance between the original and generated data distributions, such as the Inception Score (IS) and Fréchet Inception Distance (FID) [31]. For tabular data, the practice revolves mostly around validating the generated data through training models on the combination of the generated and original datasets [30, 32], analyzing statistics derived from distance between individual feature distributions, and computing paired correlations [32].

3 A Suite of Datasets for Fraud Detection and Fairness Measurement

In this section we will go into detail on the main characteristics of the original dataset, how the generated sample was obtained, the decisions made regarding the generative process, and the different variants presented in the suite of datasets this work introduces.

3.1 Dataset Overview

The introduced dataset regards the detection of fraudulent online bank account opening applications in a large European bank. In this scenario, fraudsters usually attempt to either impersonate someone via identity theft, or create a fictional individual in order to gain access to the banking services. After the services being granted, the fraudster wishes to max out the credit available for the account, which then proceeds to default. The costs of default are sustained by the banking company. Our use case is considered a high-stakes domain of application for ML. A positive prediction is usually followed by a punitive action, in this case, denying the application. As mentioned in Section 1, holding a bank account is a basic right in the European Union, making fraud detection an extremely pertinent application from a societal perspective. Following the recent awareness of the risk of unfair decision-making using ML systems, banks and merchants are in a front position to become early adopters of Fair ML methods. Nonetheless, a few percentage points in recall may represent millions of fraud losses, which makes the requirements for Fair ML particularly strict.

Each instance (row) of the dataset represents an individual application. All of the applications were made in an online platform, where explicit consent to store and process the gathered data was granted by the applicant. The label of each instance is stored in the `"is_fraud"` column. A positive instance represents a fraudulent attempt, while a negative instance represents a legitimate application. The dataset comprises eight months of information ranging from the February to September (including). The prevalence of fraud varies between 0.85% of the instances and 1.5% of the instances over the months. We observe that these values are higher for the later months. Additionally, the distribution of applications is unbalanced by month. Some months have a higher number of applications (15% of the total applications) and some have lower number (9.5% of the total applications). These distributions are reference in order to define the approximate number of legitimate and fraudulent instances that should be sampled each month for each variant of the dataset in the suite.

During the process of training a generative model, as well as obtaining the empirical observations, several choices were made. These are listed and justified bellow.

Splitting Strategy: We follow the original strategy for the evaluation of models in the dataset, by training on the first six months of data and validating the models on the last two months.

Protected Attributes: The dataset includes three relevant features that are possible to use as protected attributes for the data: `"customer_age"`, `"income"` and `"employment_status"`. The original and generated distributions of each of these attributes are available in Appendix. In this study, we focus in customer age. Since this is a continuous variable, and to be able to compute group fairness metrics, we create a categorical version by separating applicants with age >50 in one group and ≤ 50 in the other group.

Performance Metric: Due to the low prevalence figures in the data, it is important to define a relevant threshold and metric for the application. This is done mainly through defining a specific operating point in the ROC space of the model. In this case, we select the threshold in order to obtain 5% false positive rate (FPR), and measure the true positive rate (TPR) at that point. This metric is typically imposed by clients in the fraud detection domain, since it strikes as a balance between detecting fraud (recall), and keeping customer attrition low — each false positive is a not satisfied customer that may wish to change the banking company after being falsely flagged as fraudulent.

Fairness Metric: In this scenario, a penalizing effect for an individual would be a wrongful classification for a legitimate applicant, *i.e.*, a False Positive. Because of this, for the context of fairness, we want to guarantee that the probability of being wrongly classified as a fraudulent application is independent of the sensitive attribute value of the individual, hence we measure the ratio between FPRs, *i.e.*, *predictive equality*.

3.2 Training and Validating a Generative Model

In this section we will describe the process of obtaining the generated dataset, as well as a generative model that was most capable of approximating the original data.

The first step of this process was to reduce the number of original features in the dataset. This has two main consequences; firstly, we improve the convergence time and results of the generative model. Secondly, we also improve the privacy of the resulting dataset, since there is less available information for tracing applicants. To this end, we started by selecting the five best performing LightGBM [33] models obtained through random search in the original dataset. Then, we selected the junction of the top thirty most important features for these five models, according to the default feature importance method of LightGBM (number of splits per feature in the model). This resulted in a total of forty three features. This selection was reduced further to thirty features, by selecting more expressive, interpretable, and less redundant features manually.

Afterwards, we trained the CTGAN models on the original dataset with the selected features. Since there are no generative models architectures capable of modeling temporal data out-of-the-box, we add this functionality by creating a column representing the month where the application was made. We found this segmentation to be a good trade-off between sample size and granularity. The selection of hyperparameters for the generative model was done through random search, resulting in a total of 70 trained models. The tested hyperparameters are available in Appendix A.1. Generative models were trained in parallel, in four Nvidia GeForce RTX 2080 Ti models. The average (non-paralelized) time to train a single generative model was of 4.53 hours, totaling in close to 13 days of computation time. Some of the patterns of the data would not be not accurately modeled by the generative model by default. For example, personal income is rounded to have two significant figures, while the modeled results are arbitrary float values. Because of this, we manually applied transformations to some fields in the data. Moreover, for each instance, we encoded a single unique identifier depending on the feature values, so that there could be no repetitions between the original data and the generated datasets, or among the generated datasets.

With the aforementioned setup, we would create samples from the generative model of 2.5M instances. From these samples, we would reduce to candidate datasets with 1M instances by further sampling observations, such that the observed month distribution and prevalence by month corresponded approximately to the original dataset's. These datasets would then be evaluated to assess the quality of the generated model (for more details on the results see the Appendix, Section A.3). The first group of metrics regards the predictive performance of ML models on combinations of data. This extends previous works [30, 32]. In these, the trained models use generated data in train and are tested on real data. In our study, we also train with real data and test on generated data, and train and test using exclusively generated data. The second group of metrics regards the statistical similarity between the real and generated data. We calculate the average absolute difference in Pearson correlation between the real data and the generated data [32], as well as the average distance between the empirical cumulative distribution functions of each feature for the datasets. The goals of leveraging both sets of metrics are to make sure that models trained on the generated datasets are effective at the task at hand, and to guarantee that the generated distribution is realistic and faithful to the original data.

3.3 Bias Patterns

To further enhance its generalization capabilities, the generated suite contains variants of the base dataset with pre-determined and controllable bias patterns. The data biases we considered were the following:

Group size disparity is present if $P\left[A=a\right] \neq \frac{1}{N}$, where $a \in A$ represents a single group from a given protected attribute A, and N the number of possible groups. This represents different group-wise frequencies in the dataset, and might be caused by numerous reasons, such as an original population with imbalanced groups, or uneven adoption of an application by demographic segments. Considering the example of the presented dataset, where age is the protected attributed, group size disparity would imply that age groups have different sizes. This pattern is observed in the original dataset, with a higher proportion of applications being made by the younger age group.

Prevalence disparity occurs when $P\left[Y\right] \neq P\left[Y|A=a\right]$, *i.e.*, the class probability depends on the protected group. We leverage this property to generate datasets whose probability of the label is conditioned by the different groups of the protected attribute. Similarly to the original dataset, the proposed dataset shows higher fraud rates for older age groups. The reason for this might be because fraudsters have an incentive to impersonate older people: banks provide older applicants with larger lines of credit once an account is opened, which fraudsters try to max out before being caught.

Separability disparity extends the previous definition by including the joint distribution of input features X and label Y, $P\left[X,Y\right] \neq P\left[X,Y|A=a\right]$. An example of this, consider an ATM withdrawal scenario, where we have a binary feature (illumination) indicating if the ATM has external light close by, and age. Also, suppose that the age group 20-40 has a higher probability of using ATMs in dark places. This leads to a greater likelihood of having their card cloned by a fraudster. The illumination feature will help identify fraud instances for records within that group, but not for the remaining instances.

The first and second disparities are obtained through undersampling or oversampling the instances, depending on the group and label, respectively. The third is obtained through appending two columns, with different multivariate normal distributions, whose means depend on the group and label, with different controllable linear separability, similar to previous approaches to creation of synthetic datasets [34].

3.4 Dataset Variants

It is important to stress that each dataset variant follows the same underlying distribution as the base dataset, but with additional controlled data bias patterns. This implies that, save for prevalence and group disparities in some cases, whatever biases were present in the base dataset are also present in the variants. The goal is to offer a diverse set of additional algorithmic fairness challenges. A summary of the generated variants can be found in Table 1.

Variant I. Contrary to the Base and Original datasets, the groups in the protected attribute of this variant do not have disparate fraud rates. Instead, the group size disparity is aggravated, reducing the size of the minority group from approximately 20% of the dataset to 10%. As such, while models trained on this dataset will not face the challenge of group-wise prevalence imbalance, they still have

Table 1: Summary table of the generated variants in the study. Approximate values for the original dataset. Values in parentheses are applied to the test set.

Dataset	Group	Group Size	Prevalence	Separability
Original	Majority	80%	1%	-
	Minority	20%	2%	-
Base	Majority	77%	0.9%	-
	Minority	23%	1.8%	-
Variant I	Majority	90%	1.1%	-
	Minority	10%	1.1%	-
Variant II	Majority	50%	0.4%	-
	Minority	50%	1.9%	-
Variant III	Majority	50%	1.1%	Increased
	Minority	50%	1.1%	Equal
Variant IV	Majority	50%	0.3% (1.5%)	-
	Minority	50%	1.7% (1.5%)	-
Variant V	Majority	50%	1.1%	Increased (Equal)
	Minority	50%	1.1%	Equal (Equal)
Global	-	-	1.8%	-

to be robust to the fact that there is an even smaller minority group, which may be left under-explored and under-represented.

Variant II. Instead of exhibiting group size disparities, like the Base and Variant I datasets, this variant features steeper prevalence disparities — the minority group has five times the fraud rate of the majority, instead of approximately two times. Thus, this variant serves as a *stress test* for the prevalence disparity bias.

Variant III. This dataset features the Separability disparity presented in Section 3.3, whereby the classification task is made relatively simpler for the majority group by manipulating the correlations between the protected attribute, appended features, and the target. This type of bias calls for more nuanced interventions; for instance, re-sampling the data to balance prevalence and group size is ineffective, as they are already balanced. Thus, for models to be fair and stay performant under this variant, it is important to reach an equilibrium between countering the relations among some features and the protected attribute, while still learning useful patterns.

Variant IV. This variant introduces a temporal aspect to the presented data biases. In particular, similar to Variant II, it features prevalence disparities over the first six months, but no disparity for the remainder. Considering the first six months as a training set, and the rest as validation data, the observed disparity can be caused by a biased training data collection process, for example. Taking such aspects into account is fundamental to model realistic dataset variants, since real-world use cases are susceptible to biases outside of the practitioner's immediate control, and that change across time.

Variant V. Similarly as the previous variant, this dataset features changes in data bias patterns over time. However, we keep group-size and prevalence balanced. Instead, we add a separability bias component on the first six months, and remove it on the remainder. This is essentially a feature distribution shift across time, where we make sure that the features that change are related to both the protected attribute and the target. Most models in the real-world operate in highly dynamic environments, which makes them highly susceptible to temporal distribution shifts. In fact, this variant is analogous to a very common phenomenon in fraud detection: fraudsters adapting to the outcomes. That is, fraud detection is an adversarial classification setting [35] (a subset of performative prediction [36]), where fraudsters may adapt their behaviour over time to evade avoid detection. This means that features that were useful to detect fraud for a time, may become obsolete afterwards, as fraudsters learn to escape the system. In Variant V, these features are related to the protected attribute and the target, creating a potential for drastic change in the landscape of algorithmic fairness.

4 Empirical Observations

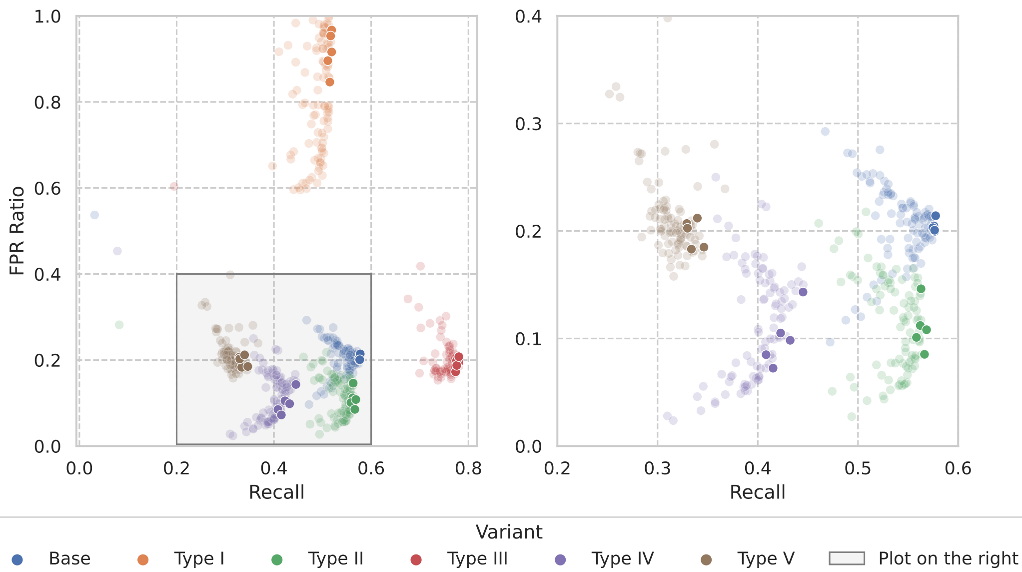

Figure 1: On the left, fairness and performance of 100 LightGBM models across all datasets in the suite. On the right, a zoom-in that focuses on the base dataset and the Type II variant, compared with the variants that feature temporal bias (Types IV and V). Opaque points represent the top 5 models in terms of performance in the Base dataset, across all variants. The top performing models on the Base dataset are not necessarily the best ones on the other variants.

To paint a teaser picture of the performance and fairness challenges that practitioners would face using our suite, we assessed how fairness-blind models fared on each dataset. To this end, we sampled 100 hyperparameter configurations of LightGBM — a popular algorithm for tabular data — and trained them on each dataset. We measure performance as recall at 5% FPR, as explained in Section 3. Our fairness metric is *predictive equality* (ratio of group FPRs), which ensures no sub-group is being disproportionately denied access to banking services. This metric is appropriate for our *punitive* setting [37], as a positive classification translates into denial of banking services. That said, we strongly encourage practitioners to explore other fairness and performance metrics, as well as fairness-aware models on these datasets.

Figure 1 shows the fairness and predictive performance of all the models evaluated on the test set, using the first 6 months for training and the rest for testing. One pattern that stands out is that models are distributed in significantly different areas of the fairness-accuracy space, depending on the dataset they were used on. This is promising in terms of our goal of providing the community with a diverse suite of datasets. Additionally, the base dataset alone provides a demanding fairness challenge, with the top performing models lying around 0.2 FPR ratio. This implies that legitimate applications from individuals in the group of higher ages are five times more likely to be flagged as fraudulent, when compared to the group of lower ages.

Focusing on the variants, many models produced fairer results under Type 1, when compared to the baseline. Still, there is significant variance in the fairness axis, leaving room for improvement. Fairness of models decreases under the Type II variant, compared with the baseline. This is justified by the exacerbated prevalence of fraud imbalance, as the rest of the distribution is similar. With the appended features to induce the separability bias, models under Type III were able to increase performance, at a comparable level of fairness of the base dataset and Type II.

As for the variants with biases that change across time, there are some interesting findings. Looking at Figure 1, model performance deteriorated under the Type IV variant, relative to its counterpart Type II. The fact that the learned patterns in the training set do not carry over to the test set (like in Type II) explains this gap in performance. The same reasoning applies to models under Type V, which, compared to those under Type III, show a similar, yet much more pronounced performance degradation phenomenon, and no gains in fairness. The plot on the right in Figure 1 shows how the best performing models under the baseline dataset were not necessarily the best ones, especially after introducing temporal biases (Type IV and Type V datasets). In fact, several models achieved better fairness-accuracy trade-offs under these datasets. This shows how performant models in static environments may fall short in more realistic, dynamic ones.

All in all, the proposed suite seems to be an adequate tool to benchmark the fairness and performance of ML models meant for static and dynamic environments. We limited our analysis to fairness-blind models hoping that this encourages practitioners to experiment with other alternatives, including fairness-aware methods.

5 Limitations and Intended Uses

We identify two main limitations regarding the suite dataset. The first is regards theoretical guarantees of privacy. Although using aggregation features and generative models to further anonymize the data, provide some privacy guarantees, there are still no applicable methods to measure or to create an upper bound limit for the metric of differential privacy, especially when taking into consideration the tabular setting for the generative model. The identification of individuals in the data should be, in practice, impossible due to the number and nature of the features, allied to the stochastic nature of samples obtained from GANs. In future works, this may be guaranteed once methods for the generation of tabular data with an upper bound for differential privacy are introduced.

The other limitation is related to the method of obtaining information. Many of the fields in applications were filled by the applicant. This might lead to wrongful information, either provided intentionally by fraudsters to boost their chances of success, or accidentally by legitimate applicants. To the best of our knowledge, there is no solution to this problem.

There are several possible uses for this suite of datasets. We note, however, that this dataset should only be used for the purpose of evaluating ML methods and fair ML interventions, as the patterns and behaviours of banking fraud are highly dynamic and context-dependant. Models trained on this data should not be directly employed in real-world fraud detection scenarios, with the potential risk of under-performing or outputting biased decisions.

In this study, we limited our analysis to the original data split, *i.e.* training models with the initial 6 months of data, and testing on the remainder. These, however, can and should be adapted to other scenarios, which would confer more realistic and robust results *e.g.*, having part of the data for validation of the hyperparameters or threshold definition, or having a sliding window approach to train and validate models. Additionally, we defined a threshold for the studied protected attribute (age), at the value of 50. We selected this value as it represents a decent compromise between group size (approximately an 80/20 split) and prevalence (approximately 2 times larger for the older group). This threshold, however, is not intended to be mandatory; other thresholds or group definitions should be taken into consideration. Another interesting approach would be to leverage the continuous nature of this variable for fairness studies.

We encourage other authors and practitioners to experiment with different ML or fair ML algorithms on this suite of datasets. We expect that with this work, the quality of evaluation of novel ML methods increases, potentiating the development of the area. Additionally, we hope it encourages other similar relevant datasets to be published from other authors and institutions.

References

[1] Alon Y. Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intell. Syst.*, 24(2):8–12, 2009.

[2] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 843–852. IEEE Computer Society, 2017.

[3] Ravit Dotan and Smitha Milli. Value-laden disciplinary shifts in machine learning. In Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna, editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, page 294. ACM, 2020.

[4] Moritz Hardt and Benjamin Recht. *Patterns, predictions, and actions: Foundations of machine learning*, chapter 8. Princeton University Press, 2022.

[5] Ayanna Howard and Jason Borenstein. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Science and engineering ethics*, 24(5):1521–1536, 2018.

[6] Julia Angwin, Jeff Larson, Lauren Kirchner, and Surya Mattu. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`, May 2016. Accessed: 2022-05-16.

[7] Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2016.

[8] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *NIPS tutorial*, 1:2017, 2017.

[9] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338, 2019.

[10] Hemank Lamba, Kit T Rodolfa, and Rayid Ghani. An empirical comparison of bias reduction methods on real-world problems in high-stakes policy settings. *ACM SIGKDD Explorations Newsletter*, 23(1):69–85, 2021.

[11] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. Algorithmic fairness datasets: the story so far. *arXiv preprint arXiv:2202.01711*, 2022.

[12] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, 2021.

[13] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. It's compaslicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

[14] Andrey Malinin, Neil Band, Yarin Gal, Mark J. F. Gales, Alexander Ganshin, German Chesnokov, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, Vyas Raina, Denis Roginskiy, Mariya Shmatova, Panagiotis Tigas, and Boris Yangel.

Shifts: A dataset of real distributional shift across multiple large-scale tasks. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

[15] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

[16] Bernard Koch, Emily Denton, Alex Hanna, and Jacob G. Foster. Reduced, reused and recycled: The life of a dataset in machine learning research. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

[17] Alexandra Sasha Luccioni, Frances Corry, Hamsini Sridharan, Mike Ananny, Jason Schultz, and Kate Crawford. A framework for deprecating datasets: Standardizing documentation, identification, and communication. May 2022.

[18] Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Niebles, Michael Sellitto, Yoav Shoham, Jack Clark, and Raymond Perrault. The ai index 2021 annual report. In *AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA*, March 2021.

[19] European Parliament and Council. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. *Official Journal of the European Union (OJ)*, 59:1, 2016.

[20] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets, March 2018.

[21] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 202–207. AAAI Press, 1996.

[22] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[23] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 03 2022.

[24] Matias Barenstein. ProPublica's COMPAS Data Revisited. page arXiv:1906.04711, Jun 2019.

[25] Ulrike Grömping. South german credit data: Correcting a widely used data set. In *Reports in Mathematics, Physics and Chemistry*, 2019.

[26] Jacob Leon Kröger, Milagros Miceli, and Florian Müller. How data can be used against people: A classification of personal data misuses. *SSRN Electronic Journal*, 2021.

[27] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[28] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018.

[29] Qingrong Chen, Chong Xiang, Minhui Xue, Bo Li, Nikita Borisov, Dali Kaafar, and Haojin Zhu. Differentially private data generative models. *CoRR*, abs/1812.02274, 2018.

[30] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems*, 2019.

[31] Ali Borji. Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022.

[32] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y. Chen. Ctab-gan: Effective table data synthesizing. In Vineeth N. Balasubramanian and Ivor Tsang, editors, *Proceedings of The 13th Asian Conference on Machine Learning*, volume 157 of *Proceedings of Machine Learning Research*, pages 97–112. PMLR, 17–19 Nov 2021.

[33] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[34] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, 2017.

[35] Nilesh N. Dalvi, Pedro M. Domingos, Mausam, Sumit K. Sanghai, and Deepak Verma. Adversarial classification. In Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel, editors, *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 99–108. ACM, 2004.

[36] Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 7599–7609. PMLR, 2020.

[37] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic Decision Making and the Cost of Fairness. In *Proc. of the 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining - KDD '17*, pages 797–806, New York, New York, USA, jan 2017. ACM Press.

Checklist

1. For all authors...
    (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The main contribution is a suite of datasets for the evaluation of ML methods. This is described in Section 3.
    (b) Did you describe the limitations of your work? [Yes] This is discussed in Section 5.
    (c) Did you discuss any potential negative societal impacts of your work? [Yes] This is discussed in Section 5.
    (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
    (a) Did you state the full set of assumptions of all theoretical results? [N/A]
    (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
    (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] This information is included both in the description of the dataset, Section 3.1 and Appendix.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We do not experiment by varying random seeds in the process. We apply hyperparameter optimization algorithms, and show the results in Section 4 and Appendix.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] This information is included in Section 3.2 for the generative models. The calculation regarding the training of models in the datasets was omitted, as it was negligible when compared to the computation time of the generative models.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Discussed in Section 3.1.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] This is discussed throughout the paper, and one of the main reasons to do feature engineering and use a generative model.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

A.1 Hyperparameter spaces for trained CTGANs

The tested hyperparameters were:

- Batch Size (100 to 5000);
- Epochs (50 to 1000);
- Generator embedding layer dimension (8 to 256 neurons);
- Number of layers and neurons per layer in the generator (1 to 3 layers, 128 to 512 neurons per layer);
- Number of layers and neurons per layer in the critic (1 to 2 layers, 64 to 256 neurons per layer);
- Learning rates of the generator and critic.

Default values were used for omitted hyperparameters available in CTGAN's [30] implementation [1].

Additionally, undersampling the dataset was included as hyperparameter, where the target prevalence was increased to either 5%, 10%, or 20%. Not performing undersampling was also one possible value for this hyperparameter.

[1] `https://github.com/sdv-dev/CTGAN/tree/v0.4.3`

Include extra information in the appendix. This section will often be part of the supplemental material. Please see the call on the NeurIPS website for links to additional guides on dataset publication.

A.2 Hyperparameter spaces for trained LightGBM models

The tested hyperparameters for trained LightGBM models were:

- Number of estimators (20 to 10000);
- Maximum tree depth (3 to 30 splits);
- Learning rate (0.02 to 0.1);
- Maximum tree leaves (10 to 100);
- Boosting algorithm (GBDT, GOSS);
- Minimum instances in leaf (5 to 200);
- Maximum number of buckets for numerical features (100 to 500);
- Exclusive feature bundling (True or False).

Default values were used for omitted hyperparameters available in LGBM's [33] official implementation [2].

A.3 Results of Generative Models

In this section, we present the evaluation results of the 70 trained generative models. Out of these 70 models, 20 (≈28%) were not able to produce a candidate sample that followed the observed distribution of month and prevalence in the original datasets. These were excluded from the analysis, as they were incapable of learning the distribution of the data over time to an acceptable extent. We present a table with the best performing generative models, when testing with the generated train and test sets.

Table 2: Results of the evaluation on trained generative models (Top 5 Models).

ID	Train & Test ↓	Train Set	Test Set	KS Metric	Correlation Diff.
1 (Selected)	54.8%	63.1%	44.2%	0.074	0.018
2	51.2%	63.6%	41.3%	0.077	0.025
3	50.6%	65.3%	39.6%	0.078	0.017
4	49.4%	53.3%	32.0%	0.071	0.027
5	48.4%	62.5%	40.7%	0.086	0.024
Mean (Std.)	26.5% (16.3%)	30.9% (23.3%)	16.7% (13.7%)	0.127 (0.061)	0.031 (0.012)

The first three columns of metrics represent the obtained predictive performance (TPR with thresholding at 5% FPR) with the possible combinations of datasets. Here the column **Train & Test** represents training and testing on the generated dataset; the column **Train Set** represents training on the generated train set and testing on the original test set, and; the column **Test Set** represents training on the original training set and testing on the generated test set. The selection criterion for the generative model was the highest performance when training and testing on generated data. No model was able to achieve performance similar to training and testing on the original data, which was of 75.4%TPR. Observing the table results, we notice a larger degradation in performance when using the generated test set only. The selected model, in fact, obtained the best performance with the generated test set, while other models produced slightly better results with the generated training sets. In this regard, a part of the models was not capable of converging, with performances close to a random estimator in the ROC space (TPR=FPR). Regarding the statistical similarity metrics, we observe that these values are not correlated with the ML performance of the datasets.

[2] https://lightgbm.readthedocs.io/en/v3.2.1/Parameters.html

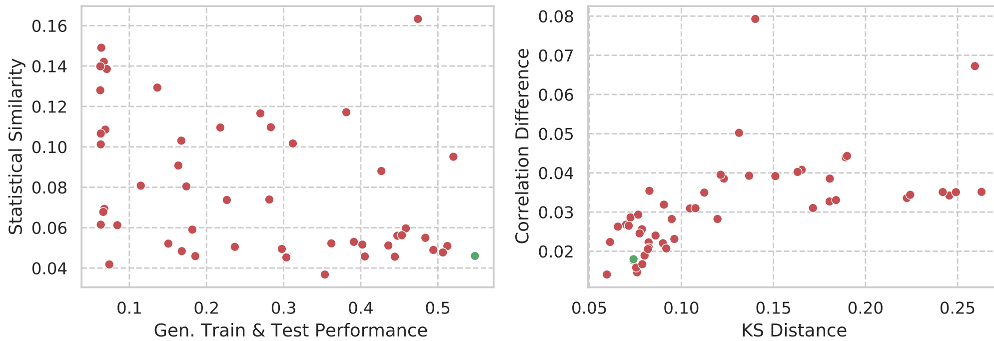

Figure 2: Generative models metrics. The left plot represents performance (with generated train and test sets) versus statistical similarity. The right plot represents the two metrics of statistical similarity. The selected generative model is represented in green.

In these plots, the main conclusion that we can obtain is that there is no clear correlation between ML performance and statistical similarity. The better performing models, however, have better than average results in the statistical similarity metrics.

A.4 Distributions of protected attributes

Customer Age

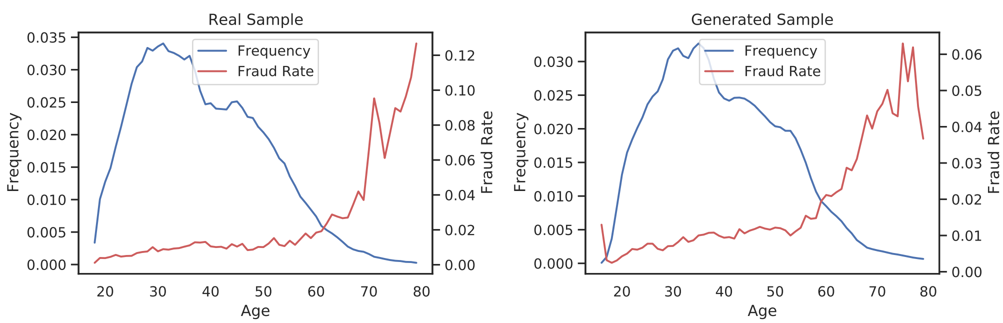

Figure 3: Distribution of age and prevalence of fraud by age in real (left) and generated (right) datasets. Ages truncated to 80, due to the lower frequencies and higher noise in higher values.

Personal Income

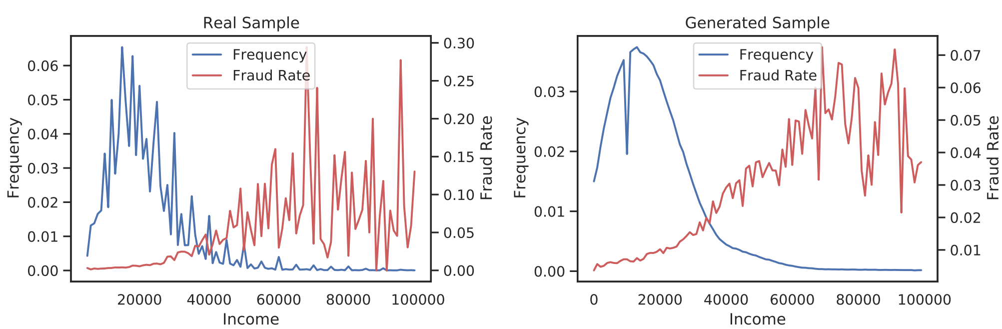

Figure 4: Distribution of income and prevalence of fraud by income in real (left) and generated (right) datasets. Truncated to 100k, due to high variance and low frequencies for higher income values.