BAF Dataset Suite Datasheet

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

Motivation

Q1: **For what purpose was the dataset created?**

A1: The target of this suite of datasets is to contribute to the evaluation employed in ML Research with a large-scale, realistic and up-to-date suite of tabular datasets. We focus particularly in testing fairness and ML performance in dynamic conditions and extreme scenarios, in order to *stress-test* ML methods and fair ML interventions. We introduce distinct biased patterns segmented in the different presented datasets (*i.e.*, Bias types) to register the fairness-performance trade-offs under different conditions.

Q2: **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

A2: The dataset was created by Sérgio Jesus, José Pombal, Duarte Alves, André F. Cruz, Pedro Saleiro, and Pedro Bizarro on behalf of Feedzai.

Q3: **Who funded the creation of the dataset?**

A3: The dataset creation was funded by Feedzai; there was no additional third-party funding involved in the creation of this dataset.

Composition

Q4: **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?**

A4: Each instance in the suite of datasets represents a synthetic, feature-engineered bank account opening application in tabular format. These were generated using a CTGAN trained on a real-world anonymized dataset for bank account opening fraud. There are no different types of instances.

Q5: **How many instances are there in total (of each type, if appropriate)?**

A5: The suite is composed of 6 datasets, each one with 1M instances, for a total of 6M instances. There are 5 dataset variants, each representing one or more specific bias types, and a "base" dataset, where no bias was induced in the sampling process.

Q6: **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**

A6: The presented datasets are sampled from a generative model (CTGAN). We first generate a large sample (2.5 million instances) from the best performing generative model. Then, to create the different variants of the suite, we sample instances from the large sample with different probability rates, depending on the protected attribute group and the label

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Datasets and Benchmarks track.

of the instance, totaling to 1 million instances per variant. Here, we define the sampling rates to obtain the desired group size and prevalence per month of the dataset. With this, we can create the variants of the dataset, with group size and prevalence disparities (two types of bias). Additionally, for other variants, we sample two features from multivariate normal distributions with different expected values, conditioned by the group and label of the instance, to create separability disparities (another type of bias). Representativeness in the generated data was validated through the comparison of the generated distribution of the base dataset (no disparity is artificially injected, only the naturally occurring biases of the dataset are present) and the original dataset distribution.

Q7: **What data does each instance consist of?**

A7: Each instance is a synthetic feature-engineered bank account application with the following fields:

- **income** (numeric): Annual income of the applicant (in decile form). Ranges between $[0.1, 0.9]$.
- **name_email_similarity** (numeric): Metric of similarity between email and applicant's name. Higher values represent higher similarity. Ranges between $[0, 1]$.
- **prev_address_months_count** (numeric): Number of months in previous registered address of the applicant, *i.e.* the applicant's previous residence, if applicable. Ranges between $[-1, 380]$ months (-1 is a missing value).
- **current_address_months_count** (numeric): Months in currently registered address of the applicant. Ranges between $[-1, 429]$ months (-1 is a missing value).
- **customer_age** (numeric): Applicant's age in years, rounded to the decade. Ranges between $[10, 90]$ years.
- **days_since_request** (numeric): Number of days passed since application was done. Ranges between $[0, 79]$ days.
- **intended_balcon_amount** (numeric): Initial transferred amount for application. Ranges between $[-16, 114]$ (negatives are missing values).
- **payment_type** (categorical): Credit payment plan type. 5 possible (annonymized) values.
- **zip_count_4w** (numeric): Number of applications within same zip code in last 4 weeks. Ranges between $[1, 6830]$.
- **velocity_6h** (numeric): Velocity of total applications made in last 6 hours *i.e.*, average number of applications per hour in the last 6 hours. Ranges between $[-175, 16818]$.
- **velocity_24h** (numeric): Velocity of total applications made in last 24 hours *i.e.*, average number of applications per hour in the last 24 hours. Ranges between $[1297, 9586]$
- **velocity_4w** (numeric): Velocity of total applications made in last 4 weeks, *i.e.*, average number of applications per hour in the last 4 weeks. Ranges between $[2825, 7020]$.
- **bank_branch_count_8w** (numeric): Number of total applications in the selected bank branch in last 8 weeks. Ranges between $[0, 2404]$.
- **date_of_birth_distinct_emails_4w** (numeric): Number of emails for applicants with same date of birth in last 4 weeks. Ranges between $[0, 39]$.
- **employment_status** (categorical): Employment status of the applicant. 7 possible (annonymized) values.
- **credit_risk_score** (numeric): Internal score of application risk. Ranges between $[-191, 389]$.
- **email_is_free** (binary): Domain of application email (either free or paid).
- **housing_status** (categorical): Current residential status for applicant. 7 possible (annonymized) values.
- **phone_home_valid** (binary): Validity of provided home phone.
- **phone_mobile_valid** (binary): Validity of provided mobile phone.
- **bank_months_count** (numeric): How old is previous account (if held) in months. Ranges between $[-1, 32]$ months (-1 is a missing value).
- **has_other_cards** (binary): If applicant has other cards from the same banking company.

- **proposed_credit_limit** (numeric): Applicant's proposed credit limit. Ranges between $[200, 2000]$.
- **foreign_request** (binary): If origin country of request is different from bank's country.
- **source** (categorical): Online source of application. Either browser (INTERNET) or app (TELEAPP).
- **session_length_in_minutes** (numeric): Length of user session in banking website in minutes. Ranges between $[-1, 107]$ minutes (-1 is a missing value).
- **device_os** (categorical): Operative system of device that made request. Possible values are: Windows, macOS, Linux, X11, or other.
- **keep_alive_session** (binary): User option on session logout.
- **device_distinct_emails** (numeric): Number of distinct emails in banking website from the used device in last 8 weeks. Ranges between $[-1, 2]$ emails (-1 is a missing value).
- **device_fraud_count** (numeric): Number of fraudulent applications with used device. Ranges between $[0, 1]$.
- **month** (numeric): Month where the application was made. Ranges between $[0, 7]$.
- **fraud_bool** (binary): If the application is fraudulent or not.

Q8: **Is there a label or target associated with each instance?**

A8: Yes, the label is contained in the **fraud_bool** field. A positive value (fraud_bool=1) represents a fraudulent bank account application. A negative value (fraud_bool=0) represents a legitimate bank account application. When accepted, all accounts are opened with access to credit.

Additional information on how the labels were obtained:

There is a slight selection bias in the way the labels were gathered, as there is a wide range of regulatory and business constraints that restrict the type of customers some banks may accept. For instance, anti money laundering (AML) regulation dictates that some customers may not be allowed to open a bank account (or credit line). Other applicants may have been pre-rejected due to business constraints (e.g., targeting only customers with 18+ years of age living in country XYZ). Strictly speaking, this does mean that our ground-truth label pool does not correspond to the entire universe of potential customers. It is, however, a sizable and representative pool of customers that would be expected to be able to open a bank account on any bank, as all banks are subject to the same regulatory requirements and similar business constraints. Gathering labels for pre-rejected customers would be either unrealistic or go against bank regulations - e.g., if some customer is signaled as having a very high chance of conducting money laundering, then the bank cannot legally accept their application, and therefore will never know the true label for that customer.

Moreover, no part of the pre-screening process was due to a previously implemented fraud detection ML model, which means that all customers that passed pre-screening were given a chance to prove to be legitimate or fraudulent. This is admittedly rare in real-world scenarios nowadays, as this kind of lax acceptance leads to very high losses for the bank, until some ML model can be put into production. As such, this dataset arguably covers as wide of a customer pool as possible.

Furthermore, all the labels in the data are trustworthy. Fraudulent observations are unequivocal — fraud is detected, confirmed, and labeled as such. As for non-fraudulent observations, the probability of them being mislabeled is exceedingly low, for two main reasons: first, the dataset is mature, and all accounts had a significant amount of time to prove whether they were fraudulent or legitimate (labels were gathered several months after the account application). Second, there is a strong incentive for fraudsters to conduct fraud as soon as their account is accepted; there is nothing to be gained from delaying fraudulent activities from the moment they have access to a credit line, and there is something to be lost, as the account can be closed under suspicion of fraud.

In summary, despite the aforementioned label selection issue, we argue it is unrealistic to provide data (ground-truth or not) on all applicants. This weakness is inherent to all real-world bank account opening datasets (which are seldom publicly available), and other financial services' domains, such as consumer lending applications. However, we maintain that the proposed datasets are representative of the types of financial decisions taken by ML models everyday, and, as such, are a valuable contribution to the research community.

Q9: **Is any information missing from individual instances?**

A9: There is no missing information from individual instances.

Q10: **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?**

A10: There are no relationships between individual instances. Each individual instance was generated using a CTGAN independently of each other, and each instance represents features to detect fraud, with no links to other instances.

Q11: **Are there recommended data splits (e.g., training, development/validation, testing)?**

A11: The performed data splits are based on the temporal information of the dataset. To this end, we use the column **month**, similarly to the original data.We made the column available so practitioners can test different temporal cross validation strategies. The results presented in the paper use the first six months of data for training and the last two months for validation. This was the split used for the original dataset.

Q12: **Are there any errors, sources of noise, or redundancies in the dataset?**

A12: The only known source of error in the pipeline is the random process of GAN generation. This might generate instances that might not be similar to real-world instances (although we test the generated dataset in both ML model performance and statistical similarity). Additionally, another source of noise is the sampling to produce the different variants of the dataset.

Q13: **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**

A13: The dataset is self-contained. No links to external resources.

Q14: **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?**

A14: There is no confidential data in this dataset. The selected features to be included in the dataset are not traceable to the individuals that applied to a bank account.

The original data had already been anonymized, and most features were aggregations that do not constitute a significant privacy threat to specific applicants. Still, it was considered important to improve the privacy of the proposed resource. Thus, we performed a series of privacy-promoting interventions in the process of making the dataset (see Figure 1). In particular, prior to training the GAN, we pre-selected a subset of the best and most intuitive features to improve convergence of the GAN. Second, we added Laplacian noise to the original data, with varying privacy budgets (inverse of the noise scale), and added noise to its categorical features as well. Third, continuous columns that contained personal information were categorized (customer age and income). The GAN was then trained on noisy data, with blocked access to information on real applicants. Finally, a filter was applied after generation to guarantee that no generated instance matched an original data point. To further improve privacy, we label encoded key categorical variables.

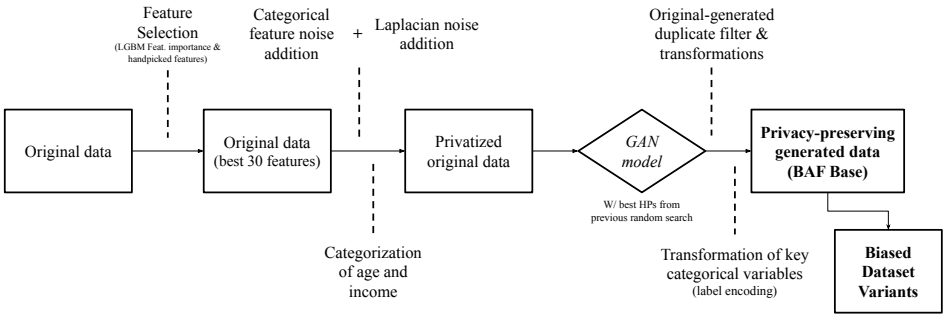

Figure 1: Illustration of the privacy-promoting interventions conducted.

Q15: **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**

A15: No.

Q16: **Does the dataset relate to people?**

A16: The dataset is composed of synthetic instances generated using a CTGAN, trained on a bank account opening dataset, where each application is related to an individual. In that sense, it is related to people, but the presented individuals are not real.

Q17: **Does the dataset identify any subpopulations (e.g., by age, gender)?**

A17: Yes, the datasets identify subpopulations, as one of the main objectives is to study of fairness under different scenarios. Age is included as one of the features of the dataset, as a continuous variable. In the dataset, the number of applications decreases with age. Inversely, the prevalence of fraud increases with age. In the empirical study, we binarize this feature based on a threshold (< 50, ≥ 50), to create groups for calculating fairness metrics. With the selected threshold we obtain approximately 80% of the applications in the former group and 20% of the applications in the latter. The latter group has a prevalence of fraud approximately two times superior to the former. These values vary across datasets in the suite. Additionally, the dataset includes annual income and professional status, which can also potentially be used for the study of fairness.

Q18: **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?**

A18: No, there is no information that allows the identification of individuals. This is expanded in answer A14.

Q19: **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**

A19: Each instance in the datasets of the suite are synthetic, and each instance represents features for fraud detection in bank account applications. Some of the learned distributions regard sensitive information, such as age and annual income, however, they are not related to a specific individual, due to the synthetic nature of the datasets.

Q20: **Any other comments?**

A20: Although the use-case is considered sensitive (banking/financial data), and there are protected attributes (age, employment status, income) one of the major concerns in this work was to guarantee that no instance in the suite represents a real identifiable individual.

Collection Process

Q21: **How was the data associated with each instance acquired?**

A21: The suite was generated by sampling the best performing CTGAN model, trained on an anonymized feature engineered dataset created to train fraud detection models. The samples were generated according to the predefined bias type. We guarantee in the sampling that no instance is repeated within the generated dataset and when compared to the original data. The original dataset was obtained directly during the application process, with user consent.

Q22: **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?**

A22: The generated dataset was collected using a CTGAN. These were trained with a random search algorithm over 70 hyperparameter configurations in parallel in 4 Nvidia GeForce RTX 2080 Ti graphics cards. We applied transformations to the data to better model patterns that are not learnable by generative models, such as the rounding in amount variables. Instances that shared the same values as real entries, or entries already existent in the sample dataset were filtered out. With this sampling, we generated a larger sample than the presented datasets, in order to sample with different probability rates, depending on the type of bias to emulate.

Q23: **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

A23: The sample present in the replica dataset was sampled from the generative model. We sampled the instances in order to obtain a similar monthly distribution to the original data, as well as to obtain the predefined bias types (depending on the dataset of the suite). In this step, we sample the instances for the dataset with different probabilities, depending on the group and label of the instance, to attain group size and prevalence disparities (Type I and Type II bias). Additional columns were appended with different multivariate normal distribution means, depending on the group and label, to attain separability disparities (Type III bias). The sampling was also made dependant on the month of the instance, to obtain prevalence and separability disparities only on the training set (Type IV and Type IV bias).

Q24: **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

A24: Only the authors were involved in the data collection process.

Q25: **Over what timeframe was the data collected?**

A25: The data was collected over a period of eight months.

Q26: **Were any ethical review processes conducted (e.g., by an institutional review board)?**

A26: No.

Q27: **Does the dataset relate to people?**

A27: Yes. Refer to answer A16.

Q28: **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

A28: The data was obtained through the sampling of a CTGAN model. The original data used to train the model was collected through a banking online form and feature engineered to obtain anonymized features.

Q29: **Were the individuals in question notified about the data collection?**

A29: For the original dataset, yes, individuals were informed about the data collection process.

Q30: **Did the individuals in question consent to the collection and use of their data?**

A30: For the original dataset, yes, the collection and use of data was consented.

Q31: **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**

A31: Yes, the consent can be revoked, regarding the original data.

Q32: **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?**

A32: No. The datasets are synthetic and should not be used to train fraud detection models to be used in real-world fraud applications. The use of these datasets should be self-contained for ML experimentation.

Preprocessing/cleaning/labeling

Q33: **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**

A33: Yes, this dataset featured a preprocessing step. This consisted in creating features 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). These features are anonymized and used to train ML models for the detection of fraud. The obtained features were then used to train a set of generative models, where one was chosen, based on predictive performance of the generated data. The process of sampling the generative model is expanded in Answers A22 and A23.

Q34: **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?**
A34: The raw data was not saved due to privacy reasons.

Q35: **Is the software used to preprocess/clean/label the instances available?**
A35: The code to sample the generative model is available on the GitHub repository of the project. The software and code for feature extraction and pre-process of the original dataset is proprietary to Feedzai.

Uses

Q36: **Has the dataset been used for any tasks already?**
A36: The suite of datasets was used only to generate the results available in the paper.

Q37: **Is there a repository that links to any or all papers or systems that use the dataset?**
A37: There are still no applications of the presented datasets. We intend to keep track of its uses in the project GitHub repo [1].

Q38: **What (other) tasks could the dataset be used for?**
A38: These datasets should be used for the context of evaluation of ML methods and fair ML interventions. Other ML evaluation tasks include the temporal evaluation of ML Models and different processing of the protected attribute.

Q39: **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**
A39: Assuming the dataset is used exclusively for the evaluation of ML methods / fair ML interventions, the composition of the dataset should not impact future uses.

Q40: **Are there tasks for which the dataset should not be used?**
A40: Using models trained in these datasets for real-world bank account opening fraud detection (or any other related application) directly should be avoided. The patterns and behaviours observed in these applications are highly dynamic and context-dependant, and using these models can result in unexpected low performances and biased decisions.

Distribution

Q41: **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**
A41: Yes, the suite is publicly accessible.

Q42: **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?**
A42: It will be distributed as a tarball on GitHub and Kaggle.

Q43: **When will the dataset be distributed?**
A43: The suite is publicly available as of today as a tarball on GitHub. It will be uploaded to Kaggle in the following weeks. There are no plans in removing the dataset from public usage.

Q44: **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**
A44: The suite is licensed under the Creative Commons CC BY-NC-ND 4.0 license.

Q45: **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**
A45: No.

Q46: **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**
A46: No.

[1] `https://github.com/feedzai/bank-account-fraud`

Maintenance

Q47: **Who is supporting/hosting/maintaining the dataset?**

A47: The suite is supported and maintained by the Feedzai Research team.

Q48: **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

A48: The authors can be contacted via email (sergio.jesus@feedzai.com or pedro.saleiro@feedzai.com). The public GitHub repository issues page can also be used as a point of contact.

Q49: **Is there an erratum?**

A49: No, there is no erratum as of yet. If necessary in the future, an erratum will be developed for the suite, as well as for this document.

Q50: **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**

A50: There are no current plans on updating the current datasets. This can change in the future, either to introduce new variants to the suite, or to correct any undetected bug in the generated datasets.

Q51: **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?**

A51: There are no applicable retention limits of the data.

Q52: **Will older versions of the dataset continue to be supported/hosted/maintained?**

A52: Currently, there is only the initial version of the suite. If any updates are published, previous versions will be available.

Q53: **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**

A53: There are no current mechanisms to contribute to the suite of datasets. Novel ideas and variants of the dataset should be submitted via email to the authors or as an issue on GitHub.

Author Statement