

Can generative AI-based data balancing mitigate unfairness issues in Machine Learning?

Benoît Ronval^{1,*}, Siegfried Nijssen¹ and Ludwig Bothmann^{2,3}

¹ICTEAM, UCLouvain, Belgium

²Department of Statistics, LMU Munich, Germany

³Munich Center for Machine Learning (MCML)

Abstract

Data imbalance in the protected attributes can lead to machine learning models that perform better on the majority than on the minority group, giving rise to unfairness issues. While a baseline method like SMOTE can balance datasets, we investigate how methods of generative artificial intelligence compare concerning classical fairness metrics. Using generated fake data, we propose different balancing methods and investigate the behavior of classification models in thorough benchmark studies using German credit and Berkeley admission data. While our experiments suggest that such methods may improve fairness metrics, further investigations are necessary to derive clear practical recommendations.

Keywords

Fairness, Generative AI, Imbalanced Data, Large Language Models, Machine Learning

1. Imbalanced data and fairness

Fairness issues in automated decision-making (ADM) systems may be attributed to different sources: the data, the machine learning (ML) algorithm (learner) and the user interactions [1]. We focus on algorithmic bias [2] where an ML model introduces a bias to previously unbiased data and tackle the subproblem of data imbalance in the protected attribute (PA). Imbalance in the PAs can give the learner a wrong incentive [3, 4]: Let us assume that a target Y shall be predicted based on features X and that the PA A is a binary feature Gender.¹ A learner that uses empirical risk minimization could gain more from fitting the majority group very closely than from spending model complexity on the minority group. This would lead to a model that has substantially better performance on the majority group, giving rise to unfairness issues [5].

A natural countermeasure would be to obtain more data from the minority group. In most applications, however, it is not possible to sample new data from the data-generating process (DGP), which calls for artificial data augmentation. The challenge of imbalanced data is not new to the ML literature and many proposals have been made to counteract this, however, mostly focusing on imbalance in the target variable Y rather than in the PA A [see, e.g., 5,

EWAF'24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany

✉ benoit.ronval@uclouvain.be (B. Ronval); siegfried.nijssen@uclouvain.be (S. Nijssen);

ludwig.bothmann@lmu.de (L. Bothmann)

🌐 <https://www.slds.stat.uni-muenchen.de/people/bothmann/> (L. Bothmann)

🆔 0000-0002-1471-6582 (L. Bothmann)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

¹While an extension for multi-categorical Gender is straight-forward, we focus on the binary version for simplicity.

6, 7]. These proposals include baseline methods such as random under/oversampling, and more sophisticated methods such as SMOTE [8]. In this work, we investigate how generative artificial intelligence (genAI) could mitigate unfairness issues related to ML models in ADM systems and compare these with SMOTE.

2. GenAI models

GenAI models learn the distribution of the original data to create fake (or synthetic) observations from that distribution. The Generative Adversarial Network (GAN) [9] for creating images comprises a generator and a discriminator, which compete against each other during training. An alternative for tabular data is the Conditional Tabular GAN (CTGAN) [10], where tabular data is represented with a *mode-specific normalization* that the GAN can use.

Large Language Models are important members of genAI models [11]. With adequate processing of the inputs and outputs, using, e.g., a format similar to "[feature] is [feature value]", LLMs can be fine-tuned to generate tabular data, as demonstrated by the model GReaT [12]. Using a relatively old LLM like GPT2 (or distil-GPT2) [13, 14], it can obtain fake data that are closer to the original distribution than with other approaches.

There exist genAI models that try to produce fair data [15, 16]. Their focus is to ensure that fake data are representative of the populations. In particular, TabFairGAN [17] is specialized for tabular data. We intend to use it in future work. Notice that we use genAI to learn better classification models from balanced data, we do not improve genAI directly.

3. Dataset corrections

We first need to train the genAI models. The CTGAN model has been trained for 200 epochs. GReaT, using distil-GPT2, has been fine-tuned for 200 epochs with a batch size of 16 and a learning rate of 0.00005. We trained the genAI models in two different setups: (1) using all original data and (2) using only 80%, leaving the remaining 20% as an unseen test set.

We consider different dataset correction methods that use the generated observations to balance the number of male and female observations. The baseline method is called **real-only** which randomly downsamples the male group. Given m and f , the original number of male and female observations, this correction reduces m to f , resulting in $2f$ original observations. The **fake-only** datasets are reduced to $2f$ observations, with equal representation of male and female and using only fake data. The **mixture-full** and **mixture** corrections are intermediate approaches. We add $m - f$ fake female observations to the original dataset, reaching $2m$ observations. The mixture approach downsamples this augmented dataset to $2f$ observations.

To compare the generative models (GReaT, CTGAN, SMOTE), corrections are done with each model excepted: (1) real-only correction is done once as it is independent of the generative models; (2) the combination fake-only and SMOTE is not possible as it creates data only for the minority group. In the end, we have 9 corrected versions for a given dataset: 3 with CTGAN, 3 with GReaT, 2 with SMOTE, and 1 real-only. We checked the fake data followed the distribution of the original with a visual comparison (not shown due to page limit). We use the German credit dataset [18] and Berkeley admission [19] and plan to add other datasets in future work.

4. Experiments

We use the above-generated 9 datasets to carry out thorough experiments to compare 3 different learners: logistic regression, classification trees, and random forests (RF). We tune their hyperparameters with 100 iterations of a random search using 3-fold cross-validation (CV).

We evaluate model performance with accuracy (ACC) and area under the curve (AUC). For a description of performance differences in the subgroups of the PA *Gender*, we use three confusion matrix-based metrics. (There has been some criticism of using these “classical fairness metrics” due to a lack of philosophical justification [see 20] – we therefore use these rather as descriptive tools than for proving or disproving fairness.) For assessing demographic parity, we compute the difference of the ratios of positive predictions for males and females (DP [21]). For assessing equalized odds, we compute the mean of the absolute values of (i) the difference between the true positive rates for males and females and (ii) the difference of the true negative rates for males and females (EO [22]). Similarly, we condition on the predicted classes and compute the mean of the absolute differences between males and females regarding positive and negative predictive values, aiming at conditional use accuracy equality (CUAE [23]).

The experiments have two settings: (1) A resampling study carries out – on each corrected dataset – nested resampling with 3-fold CV as inner resampling and 100 iterations of subsampling with a train-test ratio of 80/20 as outer resampling. The goal is to analyze the behavior of the tuned learners on unseen test data from the same distribution, including estimates of the standard deviation for statistical significance. (2) Since a generation method that generates data that are far from the original data distribution but rather simple to classify would also obtain good results in the resampling, we additionally test all methods on the same 20% test sample of the original data. In this setting, the genAI models are trained on the remaining 80% of the original data, and the learners are tuned with 3-fold CV on the resulting datasets (which are hence smaller as in the benchmark study). We use the R packages *mlr3* [24] for performing experiments and *mlr3fairness* [25] for computing fairness metrics. Results are presented in Table 1 and Table 2, respectively.

Resampling For space constraints, we selected some of the 27 combinations of learner and correction methods: Since RF performed consistently best regarding mean ACC, we limited the presentation of the results to RF. For the generation methods, we excluded the mixture method since the results are between fake-only and mixture-full (mf). For the German credit data, the generation methods outperform the results obtained by real-only (see Table 1), however, most differences are not statistically significant. Between the generation methods, a clear winner cannot be distinguished, since SMOTE leads to better predictive performances, whereas CTGAN and GReaT have better values in the fairness metrics. In-line with prior findings [26], fake-only versions are outperformed by mixture-full versions. As one example, we included GReaT-fake, which later (see Table 2) suffers from a drastic drop in performance when evaluated on original data (same holds for CTGAN-fake). For the Berkeley data (fake-versions omitted for space constraints), the generation methods also perform better, but again, differences are not statistically significant. Notably, SMOTE performs rather bad in the fairness metrics. In total, it appears that balancing the number of males and females in the datasets tends to improve the fairness metrics of the models learned on these data.

Table 1

Results of resampling. Standard deviations in parentheses. Best value per column and dataset in bold.

Data	Correction	ACC $\uparrow$	AUC $\uparrow$	DP $\downarrow$	EO $\downarrow$	CUAE $\downarrow$
German	real-only	0.737 (0.038)	0.766 (0.039)	0.060 (0.049)	0.095 (0.057)	0.115 (0.067)
German	ctgan_mf	0.739 (0.024)	0.702 (0.027)	0.043 (0.031)	0.075 (0.044)	0.095 (0.054)
German	great_mf	0.747 (0.021)	0.777 (0.025)	0.056 (0.039)	0.074 (0.039)	0.085 (0.052)
German	great_fake	0.772 (0.036)	0.799 (0.041)	0.099 (0.053)	0.116 (0.065)	0.155 (0.092)
German	smote_mf	0.779 (0.024)	0.818 (0.025)	0.062 (0.040)	0.102 (0.047)	0.105 (0.056)
Berkeley	real-only	0.653 (0.010)	0.628 (0.013)	0.177 (0.015)	0.179 (0.016)	0.079 (0.036)
Berkeley	ctgan_mf	0.658 (0.008)	0.630 (0.010)	0.159 (0.017)	0.160 (0.020)	0.055 (0.020)
Berkeley	great_mf	0.665 (0.007)	0.632 (0.009)	0.177 (0.009)	0.179 (0.010)	0.083 (0.028)
Berkeley	smote_mf	0.615 (0.033)	0.670 (0.008)	0.398 (0.151)	0.395 (0.141)	0.107 (0.023)

Table 2

Results on 20% real test data. Best value per column in bold.

Data	Correction	ACC $\uparrow$	AUC $\uparrow$	DP $\downarrow$	EO $\downarrow$	CUAE $\downarrow$
German	real_only	0.730	0.761	0.013	0.084	0.163
German	ctgan_mf	0.760	0.771	0.036	0.054	0.134
German	great_mf	0.760	0.766	0.072	0.045	0.048
German	great_fake	0.695	0.732	0.026	0.033	0.133
German	smote_mf	0.775	0.770	0.005	0.039	0.137

Evaluation on real test data Table 2 summarizes the results on the 20% test sample of the original data for the same combinations. Again, the generation methods have better values in all metrics, where a clear winner cannot be distinguished. Consistently, SMOTE leads to a well-performing model and is competitive in the fairness metrics. As mentioned above, GReaT-fake suffers from a drop in predictive performance. For Berkeley, the results of the different methods are comparable to those of the resampling (omitted in Table 2).

5. Discussion and Conclusion

The presented results indicate that generative methods can help to improve fairness metrics when facing data imbalance in the PA, where fake-only corrections do not seem to generalize well. There is, however, no clear sign that more complex, resource-intensive genAI methods like CTGAN and GReaT outperform more basic methods such as SMOTE, even if the distribution of the fake data is closer to the original data than the one observed with SMOTE. Further experiments are necessary to investigate this: beyond experiments on other real-world datasets, we plan to do thorough simulation studies with differing degrees of imbalance and complexity of the DGP. Other generative models such as TabFairGAN [17], specialized in the generation of fair tabular data, will also be studied in this context.

Acknowledgments

Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region

References

- [1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A Survey on Bias and Fairness in Machine Learning, *ACM Computing Surveys* 54 (2021) 1–35. doi:10.1145/3457607.
- [2] D. Danks, A. J. London, Algorithmic Bias in Autonomous Systems, in: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, Melbourne, Australia, 2017*, pp. 4691–4697. doi:10.24963/ijcai.2017/654.
- [3] A. Roy, V. Iosifidis, E. Ntoutsi, Multi-fairness Under Class-Imbalance, in: P. Pascal, D. Ienco (Eds.), *Discovery Science, Lecture Notes in Computer Science*, Springer Nature Switzerland, Cham, 2022, pp. 286–301. doi:10.1007/978-3-031-18840-4_21.
- [4] V. Iosifidis, B. Fethu, E. Ntoutsi, FAE: A Fairness-Aware Ensemble Framework, in: *2019 IEEE International Conference on Big Data (Big Data)*, IEEE, Los Angeles, CA, USA, 2019, pp. 1375–1380. doi:10.1109/BigData47090.2019.9006487.
- [5] P. Vuttipittayamongkol, E. Elyan, A. Petrovski, On the class overlap problem in imbalanced data classification, *Knowledge-Based Systems* 212 (2021) 106631. doi:10.1016/j.knosys.2020.106631.
- [6] V. López, A. Fernández, S. García, V. Palade, F. Herrera, An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics, *Information Sciences* 250 (2013) 113–141. doi:10.1016/j.ins.2013.07.007.
- [7] N. Japkowicz, The class imbalance problem: Significance and strategies, in: *Proceedings of the International Conference on Artificial Intelligence*, volume 56, 2000, pp. 111–117.
- [8] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, *Journal of Artificial Intelligence Research* 16 (2002) 321–357. doi:10.1613/jair.953.
- [9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, *Communications of the ACM* 63 (2020) 139–144. doi:10.1145/3422622.
- [10] L. Xu, M. Skoularidou, A. Cuesta-Infante, K. Veeramachaneni, Modeling Tabular data using Conditional GAN, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, volume 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf.
- [11] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A Survey of Large Language Models, 2023. doi:10.48550/arXiv.2303.18223.

- [12] V. Borisov, K. Seßler, T. Leemann, M. Pawelczyk, G. Kasneci, Language Models are Realistic Tabular Data Generators, 2023. doi:10.48550/arXiv.2210.06280.
- [13] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.
- [14] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, in: NeurIPS EMC2 Workshop, 2019. doi:10.48550/arXiv.1910.01108.
- [15] D. Xu, S. Yuan, L. Zhang, X. Wu, FairGAN: Fairness-aware Generative Adversarial Networks, in: 2018 IEEE International Conference on Big Data (Big Data), IEEE, Seattle, WA, USA, 2018, pp. 570–575. doi:10.1109/BigData.2018.8622525.
- [16] F. Friedrich, M. Brack, L. Struppek, D. Hintersdorf, P. Schramowski, S. Luccioni, K. Kersting, Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness, 2023. doi:10.48550/arXiv.2302.10893.
- [17] A. Rajabi, O. O. Garibay, TabFairGAN: Fair Tabular Data Generation with Generative Adversarial Networks, Machine Learning and Knowledge Extraction 4 (2022) 488–501. doi:10.3390/make4020022.
- [18] H. Hofmann, Statlog (German Credit Data), 1994. doi:10.24432/C5NC77.
- [19] P. J. Bickel, E. A. Hammel, J. W. O’Connell, Sex Bias in Graduate Admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation., Science 187 (1975) 398–404. doi:10.1126/science.187.4175.398.
- [20] L. Bothmann, K. Peters, B. Bischl, What Is Fairness? On the Role of Protected Attributes and Fictitious Worlds, arXiv, 2024. doi:10.48550/arXiv.2205.09622.
- [21] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Association for Computing Machinery, New York, NY, USA, 2012, pp. 214–226. doi:10.1145/2090236.2090255.
- [22] M. Hardt, E. Price, N. Srebro, Equality of Opportunity in Supervised Learning, in: Advances in Neural Information Processing Systems, volume 29, Curran Associates, Inc., 2016. URL: <https://papers.nips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html>.
- [23] R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, Fairness in Criminal Justice Risk Assessments: The State of the Art, Sociological Methods & Research 50 (2021) 3–44. doi:10.1177/0049124118782533, publisher: SAGE Publications Inc.
- [24] M. Lang, M. Binder, J. Richter, P. Schratz, F. Pfisterer, S. Coors, Q. Au, G. Casalicchio, L. Kotthoff, B. Bischl, mlr3: A modern object-oriented machine learning framework in R, Journal of Open Source Software (2019). doi:10.21105/joss.01903.
- [25] F. Pfisterer, W. Siyi, M. Lang, mlr3fairness: Fairness Auditing and Debiasing for ‘mlr3’, 2024. URL: <https://mlr3fairness.mlr-org.com>.
- [26] D. Manousakas, S. Aydöre, On the Usefulness of Synthetic Tabular Data Generation, 2023. doi:10.48550/arXiv.2306.15636.