Scientific Initiation Internship Report (MC040) Measuring fairness of synthetic minority oversampling on credit datasets

Author: Decio Miranda Filho[1]
Associated: Thalita Biazzuz Veronese[1] Marcos Medeiros Raimundo[1*]
[1]Instituto de Computação - Universidade Estadual de Campinas (UNICAMP)
d236087@dac.unicamp.br {veronese, mrai}@unicamp.br

Abstract

Machine Learning models often face performance issues due to class imbalance, a common problem characterized by datasets biased towards a majority class. Oversampling the minority class through synthetic generators has become a popular solution for balancing data, giving rise to many rebalancing techniques, like ADASYN and SMOTE. Practitioners usually lean on performance metrics to either refute or advocate for adopting some resampling method. However, considering the increasing ethical and legal demands for fair machine learning models, testing the neutrality of these methods concerning fairness is essential. We investigated the effects of oversampling on gender bias by analyzing statistical parity difference (SPD) and equal opportunity difference (EOD) obtained from four credit datasets. Similarly to performance, the fairness impact caused by synthetic minority oversampling was shown to be more significant for weak classifiers. Our results suggest that synthetic oversampling should be used cautiously to avoid amplifying or creating biased data.

1 Introduction

Class imbalance is a common issue in many real-world problems, especially when collecting more data on the minority class is either unwieldy or even not feasible [3, 30, 2, 32, 27, 25]. Combined with the increasing use of machine learning in almost every decision-making process [12], the disparate class distribution of available data has driven the development of a myriad of techniques aimed at tackling unbalanced learning [41]. However, a recent widespread of works analyzing the benefits of SMOTE [16, 35, 37] is showing a reduced to no gain in performance, which jeopardizes the need for applying this technique. Idrissi et al. (2022) [24] show that oversampling leads to overfitting, especially for big datasets. At the same time, simple balancing of classes and groups by either subsampling or reweighting data is faster to train and achieves state-of-the-art accuracy. Subsampling is also pointed by Chaudhuri et al. (2023) [10] as a better strategy for learning on heavy-tailed data.

On the other hand, various recent works are questioning the ethical aspects of machine learning ranging through a series of real-life examples. O'Neil (2016) [29] describes how the teacher assessment tool IMPACT led many teachers to lose their jobs in 2010. In 2018, an investigation by Angwin et al. (2018) [1] scrutinized the COMPAS recidivism algorithm and reported the ubiquitous racial disparity responsible for countless unfair decisions about defendants' freedom. Buolamwini et al. (2018) [7] published another exposure unveiling accuracy discrepancies with considerably higher error rates for dark-skinned subjects, especially females. More recently, Bender et al. (2021) [4] discuss the possible risks associated with Large Language Models, like reinforcement of hegemonic biases and the consequential harms to marginalized populations.

We believe that oversampling methods, beyond having small to no performance gain, might also reinforce discrimination contained in data and propagate it in models. We add this concern to the statements published in recent works discouraging oversampling and incorporate a new justification to oppose oversampling methods. From an ethical point of view, those methods may have unwanted effects on fairness metrics beyond the already surveyed performance ones. This work experimentally evaluates the impact of performance-guided use of oversampling methods to deal with class imbalance under the most commonly adopted performance evaluation approach, that is, fixing the decision threshold, and the most recommended one, optimizing the decision threshold. The results obtained on four credit datasets suggest that augmented synthetic samples noisily replicate already discriminating samples, disseminating social bias.

2 Fairness metrics

Anyone seeking a consensual definition of algorithmic fairness will face a vast complexity of the theme, a challenge that social and economic scientists face. In a recent paper, Chohlas et al. (2023) [12] synthesize the contributions raised so far and call researchers out to a more robust discussion on the topic. According to the authors, algorithmic fairness constraints can be classified among three notions of fairness. The first is blinding, reflecting the principle *fairness through unawareness*. This principle states that an algorithm cannot have access to sensitive features. However, many otherwise non-sensitive features can be correlated and predictive of the sensitive attribute [20]. Furthermore, this blind approach is still subject to erroneous and potentially harmful decisions by miscalibrating predictions, even without proxy features.

In an alternative approach, an algorithm is considered fair if its decision rates are equal across demographic groups. Chohlas et al. (2023) [12] argue that, in this approach, the algorithm can harm members of all groups due to infra-marginality problem [14]. A third approach defends that fairness is achieved by equalizing error rates across demographic groups, in which members of all groups could also be harmed.

Nevertheless, fairness is relegated to a second-class metric by resampling approaches, which usually focus only on prediction accuracy improvement [26]. In this paper, we evaluate synthetic minority oversampling techniques on four credit datasets concerning both accuracy and fairness. We selected two fairness metrics: Statistical Parity Difference (SPD) and Equal Opportunity Difference (EOD). Statistical Parity Difference quantifies the independence between the decision $\hat{y}(X)$ and the protected attribute Z, and is given by:

$$SPD = P(\hat{y}(X) = gain \mid Z = unpr) - P(\hat{y}(X) = gain \mid Z = priv).$$

In the equation above, $\hat{y}(X)$ represents the forecasted target, and X represents the features employed for prediction. Like being granted a loan, the aspired benefit is designated by $gain$. The protected characteristic is given by Z and can take one of two values: $priv$ (privileged) and $unpr$ (unprivileged). Statistical Parity Difference encompasses an evaluation of autonomy and equal classification [38]. SPD can also be calculated on the dataset by measuring the independence between the actual target (ground truth), denoted by Y, and the protected attribute.

Equal Opportunity Difference (EOD) [20] assesses the disparity in access to a favorable outcome $\hat{y}(X) = gain$ between the unprivileged group $Z = unpr$ and the privileged group $Z = priv$, when an individual rightfully deserves that outcome, that is, $Y = gain$ (true positive rate):

$$EOD = P(\hat{y}(X) = gain \mid Z = unpr,\ Y = gain) - P(\hat{y}(X) = gain \mid Z = priv,\ Y = gain).$$

This metric encompasses an evaluation of distinction and equal classification [38]. In both metrics, a value near zero would indicate a fair model.

Under a normative analysis of fairness metrics, Statistical Parity Difference can be interpreted as a *bias transforming* metric, reflecting the non-neutrality of the status quo represented by training data and aiming to address social structural inequality [39]. According to this interpretation, Equalized Opportunity Difference can be classified as *bias transforming*, in which matched error rates across groups reflect the preservation of target labels distribution.

3 Synthetic Minority Oversampling Methods

In our experiments, we selected two popular oversampling methods, ADASYN and SMOTE, and two variations of SMOTE, namely SVM-SMOTE and Borderline-SMOTE. In this section, we briefly describe the overall operation behind each of the selected oversampling methods and discuss its limitations on statistically translating the missing ground truth data into representative synthetic samples. We also discuss the idea of joining class and group data rebalancing that underpins some attempts to develop fair resampling approaches.

3.1 SMOTE and its variations

SMOTE SMOTE stands for Synthetic Minority Oversampling Technique and denominates an algorithm proposed by Chawla et al. (2002) [11] as "an approach to the construction of classifiers from imbalanced datasets", that has been adopted and explored in several works since it was published [36][40][6]. SMOTE generates synthetic minority class examples by selecting a minority class instance and interpolating new examples between the original instance and its k-nearest neighbors. The difference between the feature vector (sample) currently under consideration and its closest neighbor is scaled by a random number ranging from 0 to 1 and added back to the original feature vector. According to the authors, this approach makes the decision region of the minority class less specific and more expansive. In essence, SMOTE selects a random point along the line segment connecting these two specific features. This creates new instances that are similar to the existing minority class instances but are not exact copies.

Borderline-SMOTE Based on the SMOTE method, Borderline-SMOTE1 and Borderline-SMOTE2 techniques exclusively target and strengthen the borderline minority instances [19]. In Borderline-SMOTE1, only minority instances misclassified by their nearest neighbors undergo oversampling. Conversely, in Borderline-SMOTE2, both the minority instances misclassified by their nearest neighbors and the near-decision boundary instances (correctly classified but still in proximity of majority instances) are oversampled. The effectiveness of these methods stems from their focus on the most informative minority instances, namely those near the decision boundary. By oversampling these instances, the techniques enhance the classification performance of the minority class without introducing excessive noise. Experimental findings performed by the authors demonstrate that Borderline-SMOTE1 and Borderline-SMOTE2 yield superior True Positive rates and F-values compared to SMOTE and random oversampling methods.

SVM-SMOTE SVM-SMOTE is a variation of SMOTE that uses Support Vector Machines (SVMs) to classify and generate new samples at the borderline between classes. SVMs are trained to find the best possible hyperplane that maximizes the separation between the minority class and the majority class of the dataset [8]. The importance of support vector machines (SVMs) in addressing the problem of class imbalance in machine learning is highlighted by Tang et al. (2009) [34]. Through a different rebalancing heuristic in SVM modeling, including cost-sensitive learning, oversampling, undersampling, and combinations of both, it is possible to improve classification performance by extracting informative samples that are essential for classification, and eliminating redundant or noisy samples [33].

3.2 ADASYN

The Adaptive Synthetic (ADASYN) algorithm is an oversampling method similar to SMOTE that produces a variable quantity of samples determined by estimating the local distribution of the minority class. The primary distinction about SMOTE is that ADASYN adopts an adaptive approach instead of a random one. In ADASYN, the generation of synthetic samples is influenced by the density of the data distribution, deciding how many synthetic samples will be created for each instance of the minority class. This process creates instances that are harder to learn [21].

3.3 Fair Oversampling Methods

Motivated by the increasing claims for fair machine learning models, an effort to create fair oversampling methods seems to be emerging, characterized by techniques that try to consider fairness con-

straints by addressing both minority class samples generation and group bias mitigation [43, 26, 31]. Below, we describe some oversampling methods that pursue fairness found in the literature.

Fair-SMOTE is an algorithm designed by Chakraborty et al. (2021) [9] to remove biased labels and provide an equal proportion of examples for positive and negative classes. In the work of Lavalle et al. (2022) [26], bias and unfairness in the protected attributes are treated as a rebalancing problem. Extensions of undersampling, oversampling, and SMOTE have applied to the COMPAS dataset to achieve a proportion of 25% African-American non-recidivists, 25% African-American recidivists, 25% Caucasian non-recidivists and 25% Caucasian recidivists. The authors claim that applying the proposed methodology makes it possible to identify the most appropriate data rebalancing techniques that maximize fairness. In another work, Sonoda et al. (2023) [31] state that fair oversampling techniques may cause classifier overfitting and propose to solve this problem by generating synthetic data with class-mix features or group-mix features and enhancing its validity by considering the original cluster distribution and data noise.

Despite the existence of such methods, using oversampling without any discussion about its possible interference with fairness is the most commonly adopted approach, pointing out the need for researchers to consider and explore the effects of rebalancing techniques from ethical points of view. In this paper, we look at this problem by shedding some light on the effects of some popular synthetic minority oversamplers on gender bias in credit models.

4 Experimental Setup

4.1 Datasets

Our experiments were performed on four publicly available financial datasets commonly used in credit scoring research: German, Taiwan, Home Credit, and PAKDD. A brief description of them is presented below. We excluded the categorical features from training data for all datasets, except for gender, the protected characteristic defining groups in our analysis.

German German Credit Data [23] comprises 1,000 instances representing people who once took credit from a bank in Germany, together with a binary class attribute rating the applicant as either a good or a bad borrower, split 70% to 30%, respectively. Each instance is represented by a set of 20 attributes, including seven numerical and 13 categorical, discrete-valued attributes, the latter having 4 to 10 possible values.

Taiwan Taiwan dataset [42] comprises 30,000 instances representing Taiwanese bank customers' payment data, with a binary class attribute rating the applicant as either a credible or a non-credible client, split 78% to 22%, respectively. Each instance contains a set of 24 attributes, including 16 numerical and eight categorical, discrete-valued attributes, the latter having 2 to 10 possible values.

Home Credit The Home Credit dataset [28] comprises 307,511 instances representing borrowers' historical data, together with a binary class attribute rating the applicant as either a non-default or a default (after two years), split 93.3% to 6.7%, respectively. Each instance contains a set of 121 attributes, including 106 numerical and 15 categorical, discrete-valued attributes, the latter having 2 to 58 possible values.

PAKDD PAKDD 2009 dataset [13] comprises 39,988 instances representing credit card clients from a retail chain in Brazil, together with a binary class attribute rating the applicant as either a good or a bad payer, split 80.2% to 19.8%, respectively. Each instance contains a set of 28 attributes, including 13 numerical and 15 categorical, discrete-valued attributes, the latter having 2 to 5 possible values.

4.2 Evaluation metrics

Metrics like Balanced Accuracy and Area Under the Curve (AUC) are valuable to address class imbalance. A higher AUC value signifies a more effective classifier [15]. Balanced Accuracy is given by:

$$BalancedAccuracy = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right).$$

where TP (True Positives) represents the number of correctly predicted positive instances; TN (True Negatives) represents the number of correctly predicted negative instances; FN (False Negatives) represents the number of actual positive instances incorrectly predicted as negative; FP (False Positives) represents the number of actual negative instances incorrectly predicted as positive.

AUC quantifies a classifier's ability to distinguish between positive and negative instances, and is given by:

$$AUC = P[p(y = 1|X_i) > p(y = 1|X_j)|y_i = 1, y_j = 0].$$

A greater AUC value indicates a more effective classifier [15]. In more straightforward language, it signifies the probability that the score of a sample i belonging to class 1 ($p_i = 1$) is greater than the score $p(y|X_j)$ of a sample j belonging to class 0.

4.3 Algorithms

In this experimental design, we covered the three main classifiers in credit scoring [22] - Logistic Regression, Random Forest, and Gradient Boosting (XGBoost) - to evaluate the impact of employing the following oversampling methods: ADASYN, SMOTE, and two variations of SMOTE, namely SVM-SMOTE and Borderline-SMOTE. Those algorithms were optimized considering a utility measure (loss, entropy, balanced accuracy, or AUC) to emulate a performance-driven use of imbalanced methods. All baselines were trained to their standard procedure (loss for logistic regression and entropy for the tree-based classifiers), and hyperparameter tuning used balanced accuracy and a fixed threshold of 0.5 in the first experiment and used AUC and a threshold defined with the maximal difference between TNR and FPR in the second experiment. All the experiments were performed on the same test set, which encompasses 20% of total data.

The hyperparameter tuning for the oversampling methods entailed experimenting with a range of parameter values, encompassing a spectrum of possible and significant choices. For the base classifiers, it involved adjusting their main parameters: the regularization strength and optimization solver of Logistic Regression, the number of estimators of Random Forest, and the parameters of the decision tree (depth, number of splits, and number of leaves); and the decision trees (as mentioned earlier) and other model optimization hyperparameters (gamma, alpha, evaluation metric, learning rate, min child weight) from XGBoost.

In our analysis, we do not use any undersampling or reweighing technique. With this choice, we intended to measure the isolated effect of oversampling on each machine-learning model's performance. We also explored two strategies for dealing with the lack of calibration in unbalanced-trained classifiers: (1) We fixed a threshold of 0.5 for every classifier and optimized the hyperparameters using balanced accuracy; (2) We optimized the hyperparameters using AUC and selected the threshold that corresponds to the largest difference in True Negative and False Positive rates (KS statistic) [5]. Both approaches are used because adopting a predefined threshold of 0.5 is the most common choice in rebalancing problems [17], and using an optimized threshold showed great results [16].

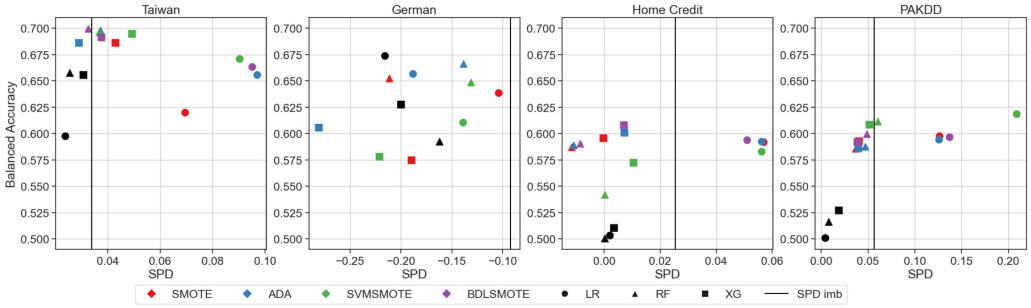
5 Experimental Results

In our experiments, we applied four distinct oversampling methods - ADASYN, SMOTE, SVM-SMOTE, and Borderline-SMOTE - to mitigate class imbalance and improve the representation of minority class. Table 1 presents the group distribution of minority class samples in the imbalanced dataset, as well as the group distributions of synthetic (minority class) samples generated by each of the selected oversampling methods.

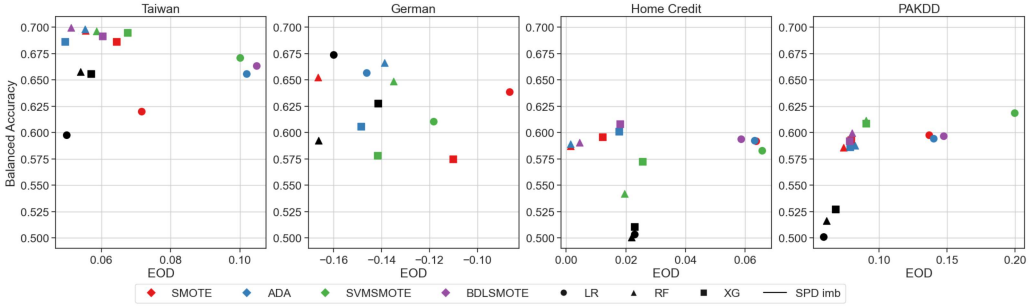
In Taiwan, Home Credit (HC), and PAKDD datasets, we observe that the proportions of synthetic samples closely match the imbalanced dataset ones, with a slight variation for Taiwan and PAKDD (± 0.1). The differences become more noticeable in the German dataset, with a variation of almost 6% more female samples (unprivileged group) generated by Borderline-SMOTE in comparison with the

Sampling	Taiwan		German		HC		PAKDD	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**
Imbalanced	0.43	0.57	0.64	0.36	0.45	0.55	0.36	0.64
SMOTE	0.43	0.57	0.61	0.39	0.45	0.55	0.35	0.65
ADASYN	0.42	0.58	0.62	0.38	0.45	0.55	0.36	0.64
SVMSMOTE	0.44	0.56	0.60	0.40	0.45	0.55	0.38	0.62
BorderlineSMOTE	0.43	0.57	0.58	0.42	0.45	0.55	0.36	0.64

Table 1: Proportion of Minority Class Samples by Gender.



(a) SPD (Statistical Parity Difference) vs. balanced accuracy. The black line represents the SPD from the dataset.



(b) EOD (Equal Opportunity Difference) vs. balanced accuracy.

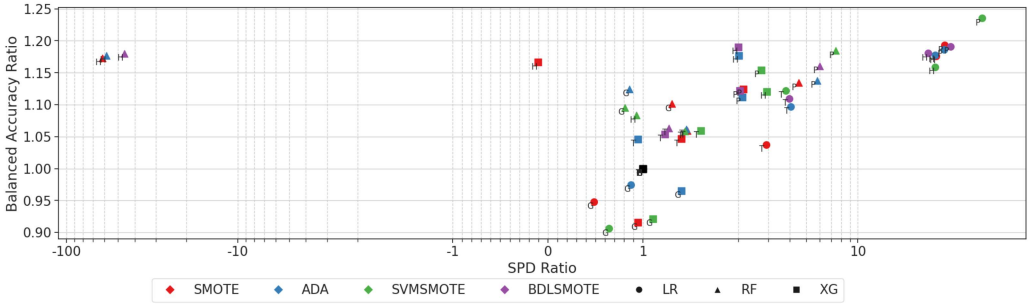
Figure 1: Relationship between Balanced Accuracy and fairness metrics obtained in different classifiers (identified by geometric shape) including Logistic Regression (LR), Random Forest (RF), and XGBoost (XG) with different oversampling methods (identified by color). Models' hyperparameters were tuned to maximize Balanced Accuracy on validation with a fixed decision threshold of 0.5.

original proportion obtained from the imbalanced dataset, which may induce an increase in data-level discrimination.

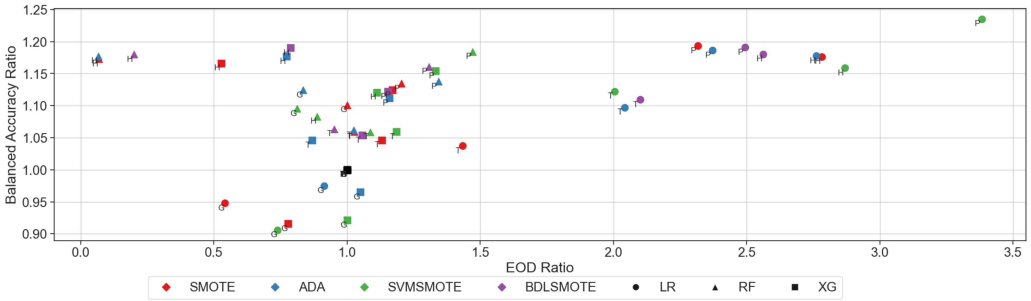
5.1 Fixed decision threshold

In this section, we analyze how classifiers behave for each oversampling method when we fix the threshold and tune the hyperparameters to maximize Balanced Accuracy. Figure 1a encompasses each dataset in a separate subplot and presents the position of balanced accuracy in relation to SPD. In Taiwan and PAKDD datasets, we can notice that, for most models (with or without oversampling), fairness metrics orbit near (either to the left or to the right) the imbalanced dataset SPD value. Logistic Regression (circles) is the only exception, with a considerably higher SPD. In terms of balanced accuracy, the oversampling methods showed some improvement. In the German dataset, the impact of oversampling methods is inconclusive, both SPD and Balanced Accuracy showed improvements and losses depending on the classifier and oversampling method. The Home Credit dataset, on the other hand, presents a more significant improvement in performance for almost all oversampling methods. However, when compared with no oversampling (black shapes) most models presented a larger SPD, except for Random Forest and SMOTE+XGBoost. Figure 1b represents the bias-preserving

fairness metric EOD vs. accuracy. The results and insights are very similar to the ones described for bias-transforming SPD metrics.



(a) SPD (Statistical Parity Difference) vs. balanced accuracy.



(b) EOD (Equal Opportunity Difference) vs. balanced accuracy.

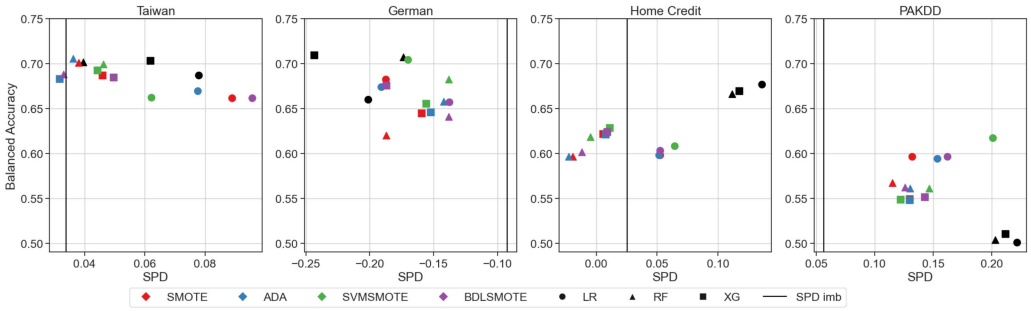
Figure 2: Relationship between Balanced Accuracy and fairness metrics gain obtained in different classifiers (identified by geometric shape) including Logistic Regression (LR), Random Forest (RF), and XGBoost (XG) with different oversampling methods (identified by color). This gain is measured w.r.t. the base model with no oversampling. Models' hyperparameters were tuned to maximize Balanced Accuracy on validation with a fixed decision threshold of 0.5.

The results for all the datasets are presented together in Figure 2a and 2b, with Balanced Accuracy vs. SPD or EOD Ratio representing the ratio between the respective fairness metric value obtained from the classifier with and without oversampling methods. For example, the x-axis value for the green square in Figure 2a is the ratio between SPD calculated on the outcomes of XGBoost classifier on Taiwan after SVM-SMOTE oversampling and SPD calculated on the outcomes of XGBoost classifier on original (imbalanced) Taiwan dataset. In this context, any value below one (and minus 1) indicates that the method has better fairness metrics compared with the baseline classifier trained on the original (imbalanced dataset). Concerning the vertical axis, values larger than one indicate performance improvement, while values lower than one indicate performance degradation. Overall, it is evident that most algorithms improved balanced accuracy with a fixed threshold decision. However, SPD variation raises a warning to the cautionary use of oversampling methods, which may lead to unpredictable discriminator outcomes. Figure 2b shows the variation in Balanced Accuracy and Equal Opportunity Difference amongst oversampling methods for each machine learning model. From this visualization, the higher fairness impact suffered by the Logistic Regression classifier (circles), whose models fill the rightmost portion of the graphic, becomes clear.

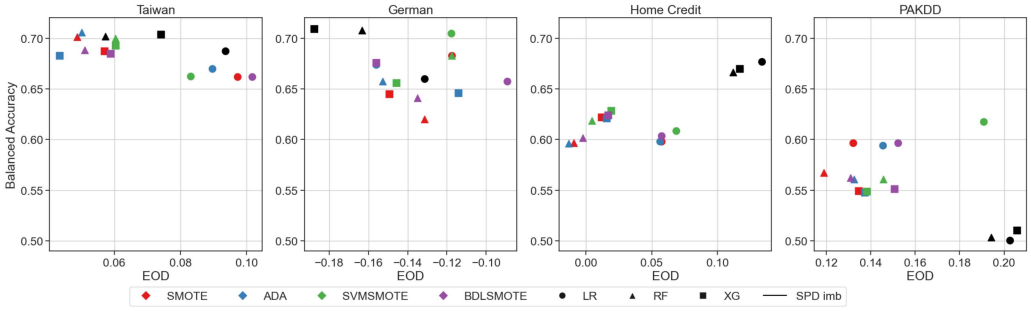
In Appendix A.1, we show the same results for AUC metrics vs. SPD and EOD in which models' hyperparameters were tuned to maximize Balanced Accuracy on validation with a fixed decision threshold of 0.5. Overall, it is evident that most algorithms have no to little improvement in AUC while most of them have a decrease in performance. SPD and EOD have no pattern, usually with fairness degradation but sometimes with improvement. In conclusion, oversampling methods might improve Balanced Accuracy when the threshold is fixed in 0.5 but it does not present any improvement in AUC and has a worrisome impact in fairness.

5.2 Optimized threshold

In this section, we analyze how classifiers behave for each oversampling method in the optimized threshold approach, obtained by tuning the hyperparameters to maximize AUC on validation and defining the threshold with the maximal difference between TNR and FPR. We can observe from Figures 3a and 3b that, for most classifiers, the use of oversampling represented little to no performance gain, except for Logistic Regression in the case of German dataset, and PAKDD overall, thus reinforcing the recent caveats in the literature against oversampling, especially for strong classifiers. However, our contribution is given by measuring the fairness impact of oversampling methods, represented by SPD and EOD variations in our analysis. In Figures 4a and 4b, we can see that, except for the Taiwan dataset, even when performance is unaffected by oversampling, there might be a considerable shift in fairness metrics. These results might suggest the emergence of apparently unpredictable ethical implications of generating synthetic samples, a behavior that we aim to investigate in a further analysis.



(a) SPD (Statistical Parity Difference) vs. balanced accuracy. The black line represents the SPD from the dataset.



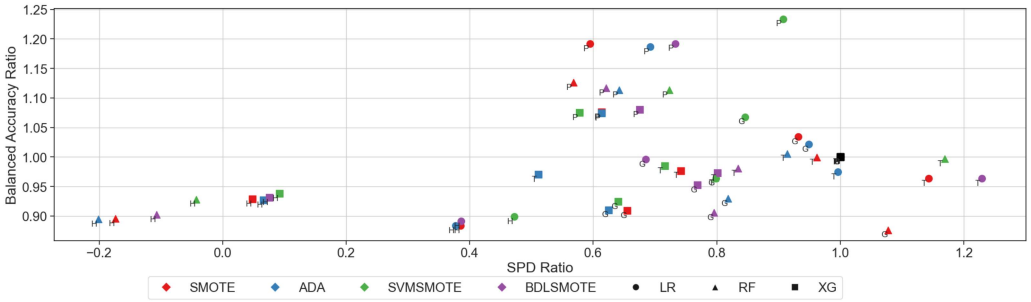
(b) EOD (Equal Opportunity Difference) vs. balanced accuracy.

Figure 3: Relationship between Balanced Accuracy and fairness metrics obtained in different classifiers (identified by geometric shape) including Logistic Regression (LR), Random Forest (RF), and XGBoost (XG) with different oversampling methods (identified by color). Models' hyperparameters were tuned to maximize AUC on validation with a threshold defined with the maximal difference between TNR and FPR.

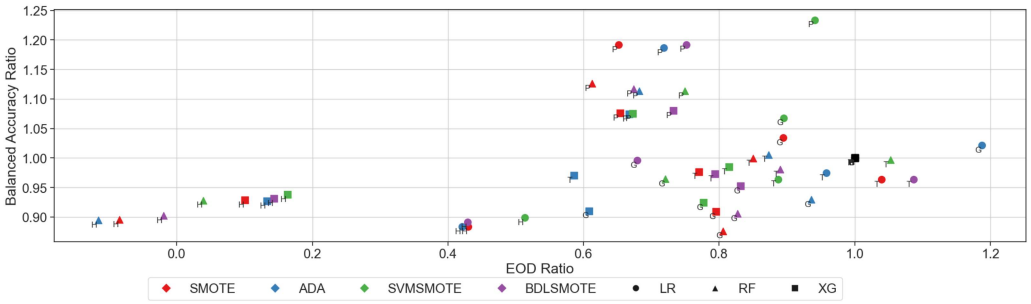
In Appendix A.2 we show the same results for AUC metrics vs. SPD and EOD in which models' hyperparameters were tuned to maximize AUC on validation with a threshold defined with the maximal difference between TNR and FPR. Overall, it is evident that most algorithms have no to little improvement in AUC while most of them have a decrease in performance. SPD and EOD have no pattern. Usually, oversampling methods improve fairness, but it is impossible to find an improvement pattern (e.g., an oversampling method that always improves fairness for a classifier).

6 Conclusion

Oversampling methods are considered the most common choice when dealing with class imbalance. In the case of SMOTE, this popularity is justified by Fernandez et al. (2018) [18] for its simple design and robustness when dealing with different problems. The success of oversamplers, though,



(a) SPD (Statistical Parity Difference) vs. balanced accuracy.



(b) EOD (Equal Opportunity Difference) vs. balanced accuracy.

Figure 4: Relationship between Balanced Accuracy and fairness metrics gain obtained in different classifiers (identified by geometric shape), including Logistic Regression (LR), Random Forest (RF), and XGBoost (XG), with different oversampling methods (identified by color). This gain is measured w.r.t. the base model with no oversampling. All models were trained to maximize AUC with a threshold defined with the maximal difference between TNR and FPR.

is measured only by performance metrics, without any reflection on the ethical implications of generating synthetic examples for high-stakes decision-making models that directly impact the distribution of social resources and sanctions.

In this work, we evaluated the impact of performance-guided use of oversampling methods to deal with class imbalance on three classifiers trained on four credit datasets. We can conclude from our results that augmented synthetic samples noisily replicate already discriminating samples, bringing out the potentially harmful effect of disseminating social bias that can arise from the unfettered use of oversampling. This concern is added to the recent works highlighting the small to no performance gain obtained by oversampling methods and may represent an ethical red flag to be heeded by those aiming to apply such methods to real-world problems.

It is worth noting that new experiments considering both oversampling and undersampling will be conducted to explore this issue further.

Acknowledgements

This project was supported by the Ministry of Science, Technology, and Innovation of Brazil, with resources granted by the Federal Law 8.248 of October 23, 1991, under the PPI-Softex. The project was coordinated by Softex and published as Intelligent agents for mobile platforms based on Cognitive Architecture technology [01245.013778/2020-21].

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 2016.

[2] John O Awoyemi, Adebayo O Adetunmbi, and Samuel A Oluwadare. Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 international conference*

on computing networking and informatics (ICCNI), pages 1–9. IEEE, 2017.

[3] Nur Athirah Azhar, Muhammad Syafiq Mohd Pozi, Aniza Mohamed Din, and Adam Jatowt. An investigation of SMOTE based methods for imbalanced datasets with data complexity analysis. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[4] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[5] Daniel Berrar. An empirical evaluation of ranking measures with respect to robustness to noise. *Journal of Artificial Intelligence Research*, 49:241–267, 2014.

[6] R. Blagus and L. Lusa. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 14:106 – 106, 2013.

[7] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[8] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[9] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. Bias in machine learning software: Why? how? what to do? In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 429–440, 2021.

[10] Kamalika Chaudhuri, Kartik Ahuja, Martin Arjovsky, and David Lopez-Paz. Why does throwing away data improve worst-group error? In *Fortieth International Conference on Machine Learning*, 2023.

[11] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[12] Alex Chohlas-Wood, Madison Coots, Sharad Goel, and Julian Nyarko. Designing equitable algorithms. *Nature Computational Science*, 3(7):601–610, 2023.

[13] PAKDD 2009 Data Mining Competition. Credit risk assessment on a private label credit card application. GitHub, 2009. https://github.com/JLZml/Credit-Scoring-Data-Sets/tree/master/2.

[14] Sam Corbett-Davies, J Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *J. Mach. Learn. Res*, 2023.

[15] Xolani Dastile and Turgay Celik. Making Deep Learning-Based Predictions for Credit Scoring Explainable. *IEEE Access*, 9:50426–50440, 2021.

[16] Yotam Elor and Hadar Averbuch-Elor. To SMOTE, or not to SMOTE? *arXiv preprint arXiv:2201.08528*, 2022.

[17] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from imbalanced data sets*, volume 10. Springer, 2018.

[18] Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018.

[19] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.

[20] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[21] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. Ieee, 2008.

[22] Hongliang He, Wenyu Zhang, and Shuai Zhang. A novel ensemble method for credit scoring: Adaption of different imbalance ratios. *Expert Systems with Applications*, 98:105–117, May 2018.

[23] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

[24] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pages 336–351. PMLR, 2022.

[25] Harsurinder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys (CSUR)*, 52(4):1–36, 2019.

[26] Ana Lavalle, Alejandro Maté, Juan Trujillo, Jorge García Carrasco, et al. A methodology based on rebalancing techniques to measure and improve fairness in artificial intelligence algorithms. In *Proceedings of the 22nd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data, DOLAP@EDBT/ICDT 2020*, pages 84–88, 2020.

[27] Menghua Luo, Ke Wang, Zhiping Cai, Anfeng Liu, Yangyang Li, and Chak Fong Cheang. Using imbalanced triangle synthetic data for machine learning anomaly detection. *Computers, Materials & Continua*, 58(1), 2019.

[28] Anna Montoya and Martin Kotek KirillOdintsov. Home Credit Default Risk. Kaggle, 2018. https://kaggle.com/competitions/home-credit-default-risk.

[29] Cathy O'Neil. *Weapons of math destruction: how big data increases inequality and threatens democracy*. Crown Publishing Group, 2016.

[30] Manisha Saini and Seba Susan. Deep transfer with minority data augmentation for imbalanced breast cancer dataset. *Applied Soft Computing*, 97:106759, 2020.

[31] Ryosuke Sonoda. Fair oversampling technique using heterogeneous clusters. *Information Sciences*, 640:119059, 2023.

[32] Kazi Abu Taher, Billal Mohammed Yasin Jisan, and Md Mahbubur Rahman. Network intrusion detection using supervised machine learning technique with feature selection. In *2019 International conference on robotics, electrical and signal processing techniques (ICREST)*, pages 643–646. IEEE, 2019.

[33] Yuchun Tang, Yan-Qing Zhang, Nitesh V Chawla, and Sven Krasser. Svms modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):281–288, 2008.

[34] Yuchun Tang, Yan-Qing Zhang, Nitesh V. Chawla, and Sven Krasser. Effectiveness of classification. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, 39(1):286–293, 2009.

[35] Ahmad S Tarawneh, Ahmad B Hassanat, Ghada Awad Altarawneh, and Abdullah Almuhaimeed. Stop oversampling for class imbalance learning: A review. *IEEE Access*, 10:47643–47660, 2022.

[36] Fredy Rodríguez Torres, Jesús A Carrasco-Ochoa, and José Fco Martínez-Trinidad. Smote-d a deterministic version of smote. In *Pattern Recognition: 8th Mexican Conference, MCPR 2016, Guanajuato, Mexico, June 22-25, 2016. Proceedings 8*, pages 177–188. Springer, 2016.

[37] Ruben van den Goorbergh, Maarten van Smeden, Dirk Timmerman, and Ben Van Calster. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association*, 29(9):1525–1534, 2022.

[38] Kush R. Varshney. *Trustworthy Machine Learning*. Independently Published, Chappaqua, NY, USA, 2022.

[39] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Bias preservation in machine learning: the legality of fairness metrics under eu non-discrimination law. *W. Va. L. Rev.*, 123:735, 2020.

[40] Juanjuan Wang, Mantao Xu, Hui Wang, and Jiwu Zhang. Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding. *2006 8th international Conference on Signal Processing*, 3:–, 2006.

[41] Le Wang, Meng Han, Xiaojuan Li, Ni Zhang, and Haodong Cheng. Review of classification methods on unbalanced data sets. *IEEE Access*, 9:64606–64628, 2021.

[42] I-Cheng Yeh. default of credit card clients. UCI Machine Learning Repository, 2016. DOI: https://doi.org/10.24432/C55S3H.

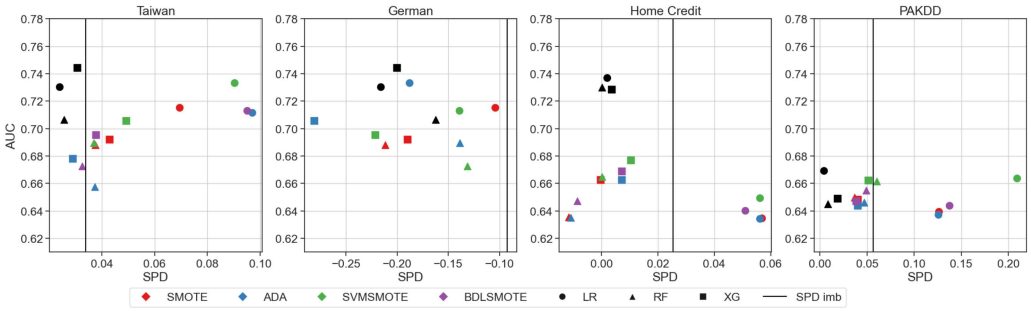
[43] Vladimiro Zelaya, Paolo Missier, and Dennis Prangle. Parametrised data sampling for fairness optimisation. *KDD XAI*, 2019.

A Experimental results comparing fairness metrics and AUC

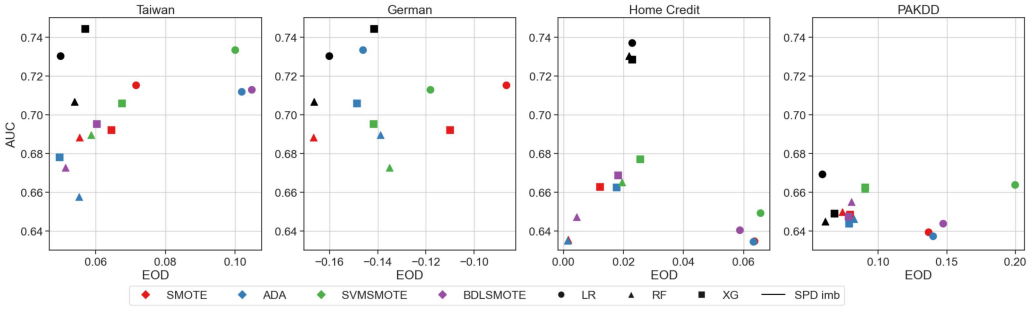
This appendix presents the results considering AUC as the performance metric.

A.1 Fixed decision threshold

This section analyzes how classifiers behave for each oversampling method when we fix the threshold and tune the hyperparameters to maximize Balanced Accuracy. Figure 5a encompasses each dataset in a separate subplot and presents the position of AUC concerning SPD. In Taiwan and PAKDD datasets, we can notice that, for most models (with or without oversampling), fairness metrics orbit near (either to the left or to the right) the imbalanced dataset SPD value. Logistic Regression (circles) is the only exception, with a considerably higher SPD. In terms of AUC, the oversampling methods showed little to no improvement.



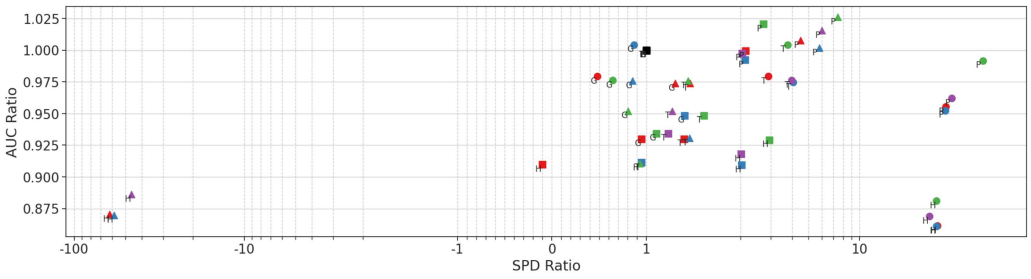
(a) SPD (Statistical Parity) vs. AUC. The black line represents the SPD from the dataset.



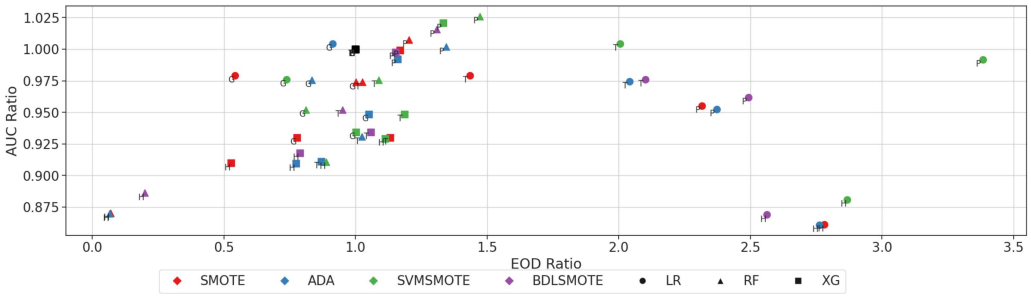
(b) EOD (Equal Opportunity Difference) vs. AUC.

Figure 5: Relationship between AUC and fairness metrics obtained in different classifiers (identified by geometric shape) including Logistic Regression (LR), Random Forest (RF), and XGBoost (XG) with different oversampling methods (identified by color). Models' hyperparameters were tuned to maximize Balanced Accuracy on validation with a fixed decision threshold of 0.5.

The results for all the datasets are presented together in Figure 6a and 6b, with Balanced Accuracy vs. SPD or EOD Ratio representing the ratio between the respective fairness metric value obtained from the classifier with and without oversampling methods. For example, the x-axis value for the red square with the letter H (right in the middle of the plot) in Figure 6a is the ratio between SPD calculated on the outcomes of XGBoost classifier on Home Credit after SMOTE oversampling and SPD calculated on the outcomes of XGBoost classifier on original (imbalanced) Taiwan dataset. In this context, any value different below one (and minus 1) indicates that the method has better fairness metrics than the baseline classifier trained on the original (imbalanced dataset). Concerning the vertical axis, values larger than one indicates performance improvement, while values lower than one indicate performance degradation. Overall, it is evident that most algorithms have no to little improvement in AUC while most have a decrease in performance. Moreover, SPD variation raises a warning to the cautionary use of oversampling methods, which may lead to unpredictable performance in fairness, with a common decrease in fairness metrics. Figure 2b shows the variation in AUC and Equal Opportunity Difference amongst oversampling methods for each machine learning model.



(a) SPD (Statistical Parity) vs. AUC.



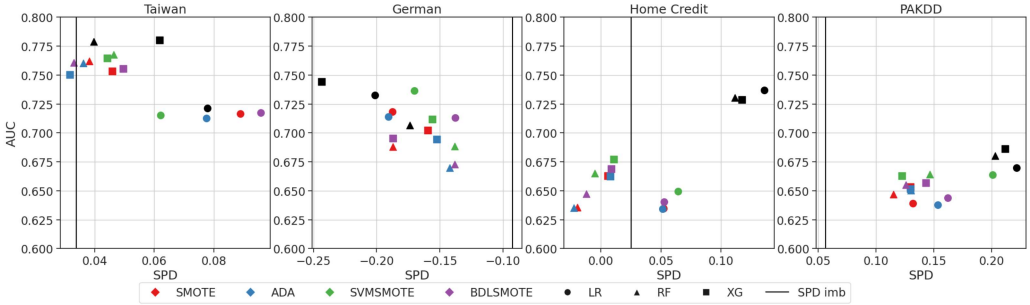
(b) EOD (Equal Opportunity Difference) vs. AUC.

Figure 6: Relationship between AUC and fairness metrics gain obtained in different classifiers (identified by geometric shape) including Logistic Regression (LR), Random Forest (RF), and XGBoost (XG) with different oversampling methods (identified by color). This gain is measured w.r.t. the base model with no oversampling. Models' hyperparameters were tuned to maximize Balanced Accuracy on validation with a fixed decision threshold of 0.5.

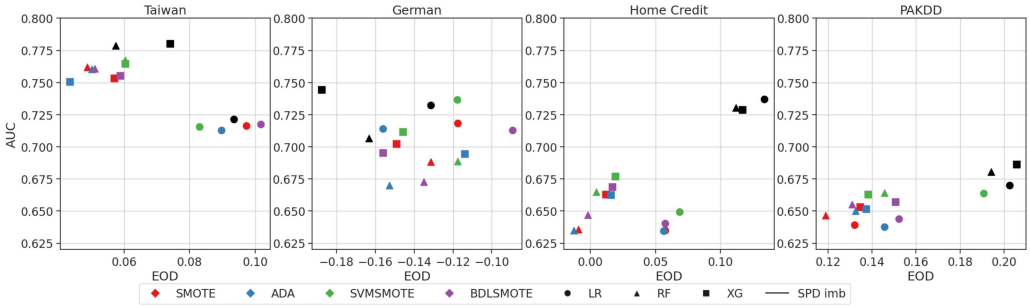
From this visualization, the higher fairness impact suffered by the Logistic Regression classifier (circles), whose models fill the rightmost portion of the graphic, becomes clear.

A.2 Optimized threshold

In this section, we analyze how classifiers behave for each oversampling method in the optimized threshold approach, obtained by tuning the hyperparameters to maximize AUC on validation and defining the threshold with the maximal difference between TNR and FPR. We can observe from Figures 7a and 7b that, for most classifiers, oversampling rarely shows performance gain. However, our contribution is given by measuring the fairness impact of oversampling methods, represented by SPD and EOD variations in our analysis. In Figures 8a and 8b, we can see that all methods showed a gain in fairness, however, in a completely unpredictable manner. We can see that oversampling methods improve fairness for the Random Forest for Home Credit dataset (the triangles in the bottom-left of the plots); however, for the dataset Taiwan (the triangles in the upper-right of the plots), it degraded the fairness metric. Generally, we can see no pattern of fairness improvement (e.g., an oversampling method always improves fairness for a classifier).

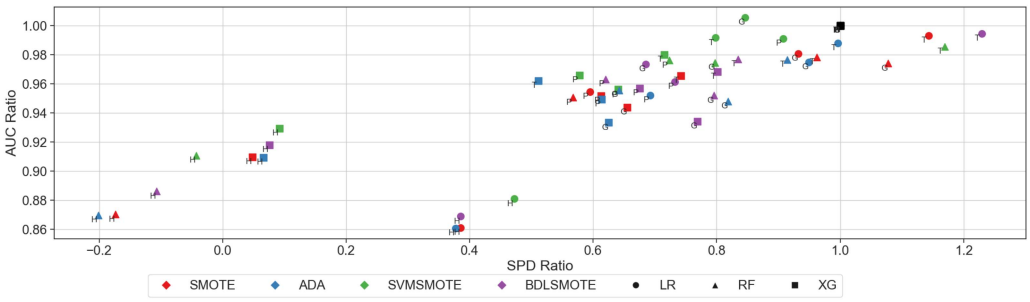


(a) SPD (Statistical Parity) vs. AUC. The black line represents the SPD from the dataset.

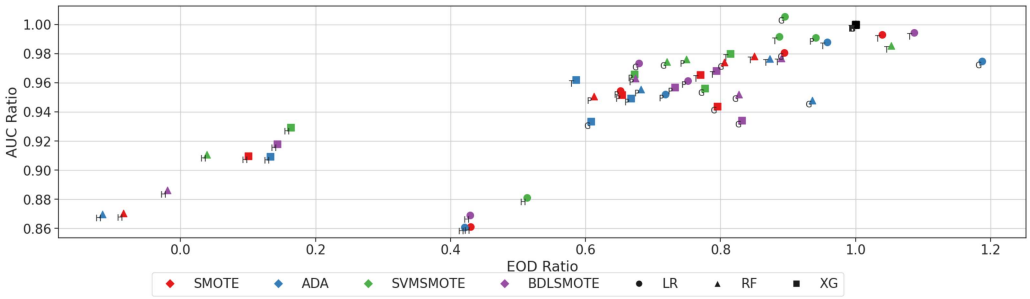


(b) EOD (Equal Opportunity Difference) vs. AUC.

Figure 7: Relationship between AUC and fairness metrics obtained in different classifiers (identified by geometric shape) including Logistic Regression (LR), Random Forest (RF), and XGBoost (XG) with different oversampling methods (identified by color). Models' hyperparameters were tuned to maximize AUC on validation with a threshold defined with the maximal difference between TNR and FPR.



(a) SPD (Statistical Parity) vs. AUC.



(b) EOD (Equal Opportunity Difference) vs. AUC.

Figure 8: Relationship between AUC and fairness metrics gain obtained in different classifiers (identified by geometric shape) including Logistic Regression (LR), Random Forest (RF), and XGBoost (XG) with different oversampling methods (identified by color). This gain is measured w.r.t. the base model with no oversampling. All models were trained to maximize AUC with a threshold defined with the maximal difference between TNR and FPR.