Analyzing Data Bias through Data Augmentation: A Case Study in Financial Data

Julia dos Santos Porphirio[1], Diogo José dos Santos[1], Sérgio Azevedo[2], Lilian Berton[1]

[1] Instituto de Ciência e Tecnologia, Universidade Federal de São Paulo (UNIFESP)
São José dos Campos, SP, Brasil
[2] Serasa Experian
Avenida das Nações Unidas, 14.401, São Paulo, SP, Brasil

{julia.porphirio, diogo.santos06, lberton}@unifesp.br

Abstract. *This study examines the impact of introducing group unbalances through oversampling strategies on model fairness and feature importance, especially concerning sensitive attributes like sex, marital status, and education in a financial dataset. We hypothesize that linear models are more vulnerable to fairness distortions introduced by oversampling generated by synthetic data generation than more complex models, such as gradient-boosted decision trees and support vector machines with an RBF kernel. To test this, we evaluate oversampling approaches like SMOTE and RandomOverSampler within a consistent framework, comparing linear classifiers against XGBoost and SVM (RBF). Our assessment includes predictive performance, the stability of fairness metrics, and changes in feature importance rankings before and after oversampling.*

1. Introduction

As machine learning (ML) systems become increasingly embedded in financial decision-making processes, concerns around algorithmic fairness and bias have gained attention [Bajracharya et al. 2022, Agu et al. 2024]. Bias in predictive models can stem not only from model architecture or training procedures but often originates in the structure and distribution of the input data itself. In particular, imbalances or hidden correlations involving protected attributes, such as gender, race, or age, can subtly influence outcomes, leading to systemic unfairness.

Class imbalance is a prevalent issue in real-world datasets, especially in domains involving health, finance, and social decision-making. To mitigate the adverse effects of imbalance, oversampling techniques such as SMOTE (Synthetic Minority Oversampling Technique) [Chawla et al. 2002] and its variants are widely used to increase the representation of the minority class. While effective in improving predictive performance, recent studies have raised concerns about the impact of these methods on fairness and the interpretability of machine learning models [Mehrabi et al. 2021].

This study investigates the effects of different oversampling strategies on fairness and feature importance, particularly in the context of models that include sensitive attributes such as *sex*, *marital status*, and *education*. A key hypothesis explored here is that linear models may be more susceptible to fairness distortion introduced by synthetic data generation compared to more complex models such as gradient-boosted decision trees and support vector machines with an RBF kernel.

To this end, we evaluate a variety of oversampling approaches, including SMOTE and RandomOverSampler [LemaÃŽtre et al. 2017], using a consistent modeling framework. Several linear classifiers are tested alongside XGBoost and SVM (RBF) as comparative non-linear baselines. We assess not only predictive performance but also the stability of fairness metrics and feature importance rankings before and after oversampling.

The findings contribute to a better understanding of how oversampling affects model behavior in the presence of categorical sensitive features, offering practical insights for model selection and fairness-aware data preprocessing.

2. Related Works

Some reviews explore the ethical challenges and fairness issues in AI-driven financial services. [Bajracharya et al. 2022] surveys the presence of algorithmic bias and fairness challenges in financial services, where AI and machine learning systems are increasingly used for tasks such as loan approvals, credit scoring, and credit limit decisions. Authors examine sources and examples of bias across key financial sectors and review detection and mitigation techniques designed to improve transparency and accountability in automated decision-making. [Agu et al. 2024] identifies key sources of algorithmic unfairness, such as biased training data and opaque decision-making, and proposes solutions including algorithmic audits, inclusive data practices, ethical design principles, and regulatory oversight. It also emphasizes the need for stakeholder collaboration and outlines future directions like explainable AI, improved bias detection, and ethical governance frameworks to ensure more equitable outcomes in financial technologies. [de Castro Vieira et al. 2025] summarize techniques used in fair credit decision systems, emphasizing how proxies like location and credit history often embed bias.

Recent research has addressed the ethical and practical challenges of integrating ML into financial services, with a focus on fairness and bias mitigation. Studies such as [Kim et al. 2023] and [Hurlin et al. 2024] highlight how protected attributes (like gender, age, marital status) and model evaluation strategies impact equitable outcomes in credit systems. [Christensen 2021] emphasizes the need for end-to-end bias interventions, while [Choi et al. 2025] explore fairness-aware modeling techniques using real-world datasets. Collectively, these works underscore the importance of rigorous fairness evaluation frameworks and contribute to a growing body of knowledge aimed at building more transparent, accountable, and inclusive AI systems in finance.

Previous works also explored fairness through data augmentation, [Zhou et al. 2023] provide theoretical and empirical support for SMOTE's effectiveness in enhancing fairness without compromising performance. [Welfert et al. 2024] investigate how augmenting latent representations can boost performance for worst-off subpopulations in classification settings. While most studies report positive outcomes from data augmentation, our findings reveal that, in certain cases, it can inadvertently amplify bias. Therefore, its use should be approached with caution and thoughtful evaluation.

3. Methodology

In this study, we analyzed the impact of various oversampling techniques on the fairness and variable importance of machine learning models. All experiments were conducted using a dataset in which the variables *age*, *sex*, *marriage*, and *education* were treated as categorical features.

3.1. Dataset

This study uses the *Default of Credit Card Clients* dataset [Yeh 2009], published by Yeh (2009) via the UCI Repository and accessed through Kaggle[1]. The dataset contains 30,000 observations, each representing a credit card client. The binary target variable (target) indicates whether the client defaulted in the following month (1 for default, 0 otherwise), with the positive class being the minority, representing approximately 22% of the records.

The dataset includes demographic attributes (such as sex, age, marital status, and education) as well as financial and behavioral features (such as credit limit, billing amount, and payment history). Some variables were preprocessed, including the discretization of age into intervals and the mapping of ordinal attributes. Table 1 summarizes the main variables considered in this work and used as inputs for model training and fairness analysis.

Table 1. Variables used from the Credit Card Default dataset after preprocessing.

Variable(s)	Description	Type
target	Default in the following month (1 for default, 0 otherwise)	Binary
LIMIT_BAL	Amount of credit limit	Continuous
sensitive_sexo	Gender	Categorical
EDUCATION	Education level	Categorical
MARRIAGE	Marital status	Categorical
sensitive_faixa_idade	Age group (categorized)	Categorical
ordinal_PAY_0 to PAY_6	Past monthly payment status (6 months)	Ordinal
BILL_AMT1 to BILL_AMT6	Amount of bill statement in each of the last 6 months	Continuous
PAY_AMT1 to PAY_AMT6	Amount paid in each of the last 6 months	Continuous

3.2. Oversampling Strategies

In this stage, we aimed to investigate the effects of targeted oversampling on both the performance and fairness of classification models, with a particular focus on addressing vulnerable subgroups. For this purpose, we selected a specific group of interest composed of defaulting clients who are single women with a high school education, referred to as the focus group. This profile was chosen for representing a critical intersection of demographic characteristics often associated with challenges in accessing credit.

Unlike traditional approaches that aim to balance the target variable as a whole, our study applies oversampling specifically to this subgroup. To that end, we intentionally increased its representation in the training set using different oversampling techniques,

[1] https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset

in order to assess whether such intervention could impact both the overall model performance and fairness-related disparities.

The following strategies were evaluated:

- **Original (No Oversampling)**: Represents the baseline scenario, with no intervention applied. The distribution of the focus group reflects its natural occurrence in the dataset and serves as a benchmark for comparison.
- **RandomOverSampler**: A simple strategy based on randomly replicating existing records from the focus group. Although it does not generate synthetic examples, it is useful as a baseline to compare against more sophisticated techniques.
- **SMOTE_default**: A technique based on SMOTENC (SMOTE for datasets with both categorical and continuous variables), which creates new synthetic samples for the focus group by interpolating between existing records and their nearest neighbors, preserving the mixed structure of the data.
- **SMOTE_expand_5x**: A variant configured to expand the focus group to approximately 5,000 instances. It uses a reduced number of neighbors (k=3) to generate new samples in an attempt to preserve realistic characteristics and minimize noise.

The target variable is binary, where 1 indicates a client who defaulted on payment in the following month, and 0 indicates no default. In the training set, the focus group—defined as defaulting clients who are female, single, and have completed high school—represented only 154 individuals out of 21,000, which corresponds to approximately 0.73%. The absolute counts of this group before and after oversampling are shown in Figure 1.

Figures 2, 3, and 4 illustrate how oversampling strategies affected the distribution of the focus group across the categorical variables *education*, *marriage*, and *sex*. It is evident that the oversampling approaches led to a substantial increase in the number of minority class (defaulting) individuals within the focus group, highlighting the direct impact of these techniques on the training data composition.

Although this approach deviates from traditional class balancing practices, the intentional distortion was designed as a controlled experiment to evaluate how different classifiers behave when exposed to a demographically underrepresented group with artificially increased presence. This setup does not aim to propose a general-purpose data preprocessing pipeline, but rather to expose model vulnerabilities under targeted augmentation.

These strategies were tested in combination with different classification models to analyze how each technique affects not only predictive performance but also the distribution of errors across sensitive groups. The models considered in this study will be presented in the next section.

3.3. Evaluated Models

In the modeling stage, our primary objective was to investigate whether linear classification models are more sensitive to the potential biases introduced by targeted oversampling techniques. To this end, we implemented and evaluated a diverse set of algorithms, encompassing both linear and non-linear approaches, and trained them on the datasets generated through each oversampling strategy.

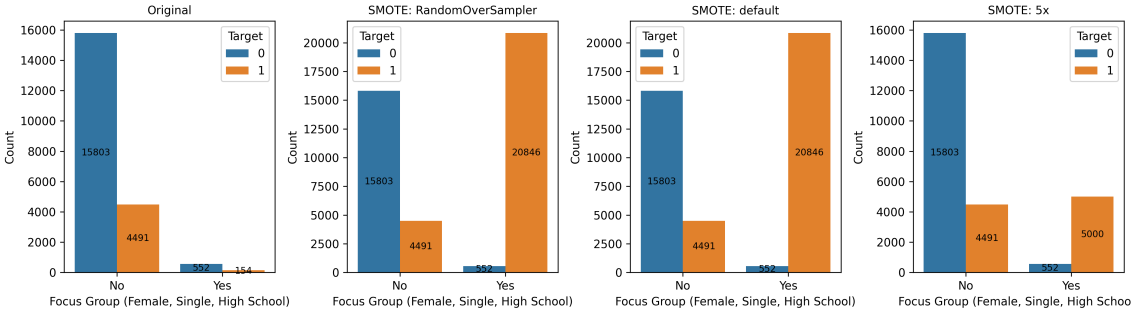

Figure 1. Absolute count of focus group individuals (female, single, high school, and defaulting) before and after oversampling.

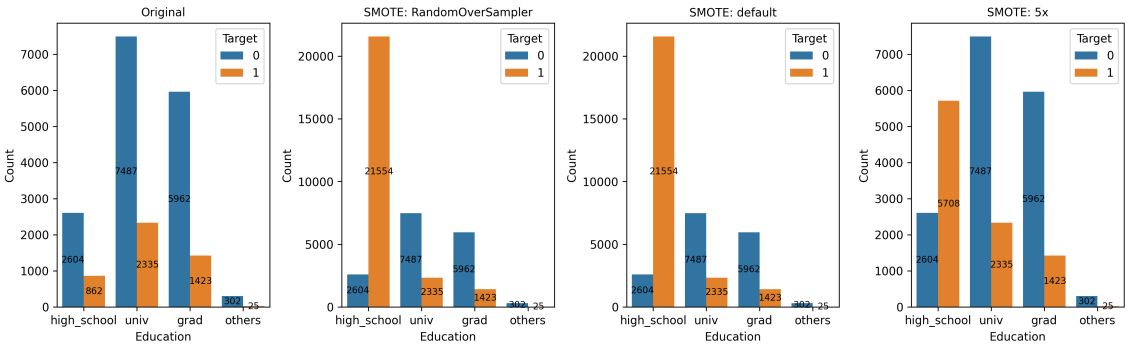

Figure 2. Distribution of the target variable by education level across different oversampling strategies. The synthetic data generation clearly increases the representation of defaulted individuals with a high school degree.

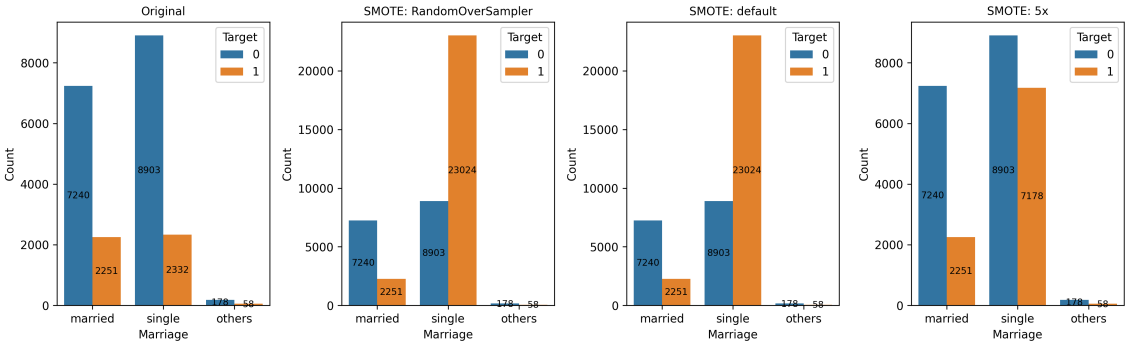

Figure 3. Effect of oversampling strategies on the distribution of the target variable by marital status. The focus group (single individuals in default) becomes more prominent in augmented datasets.

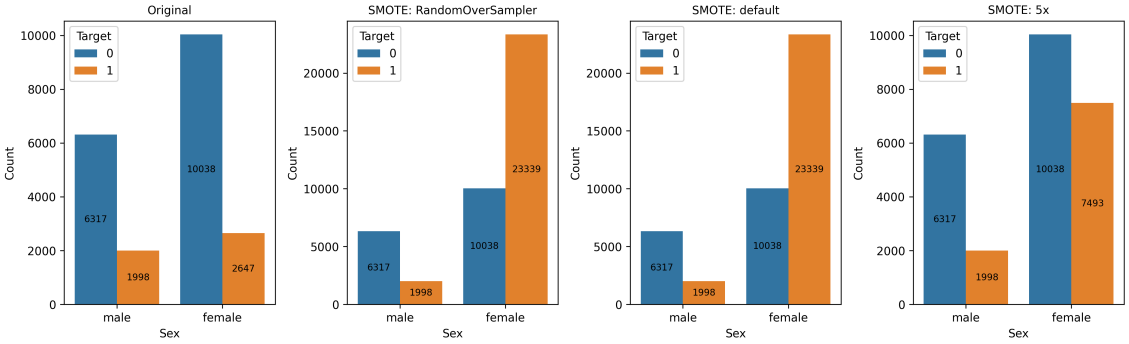

Figure 4. Sex-based target distribution before and after oversampling. A noticeable expansion is seen in the number of defaulted female individuals, in alignment with the focus group definition.

The following models were tested:

- **Linear Models:**
  - **Logistic Regression**: A classical linear model for binary classification that estimates probabilities using the logistic function. It is often used as a baseline due to its simplicity and interpretability.
  - **Ridge Classifier (L2 regularization)**: A linear model with L2 regularization that penalizes large coefficients to reduce overfitting. It tends to handle multicollinearity better than standard logistic regression.
  - **Lasso Classifier (L1 regularization)**: A linear model with L1 regularization, which encourages sparsity by driving some coefficients to zero. It is suitable for feature selection and interpretable modeling.
  - **ElasticNet**: Combines L1 and L2 regularization in a single linear model. It balances the benefits of both ridge and lasso regularization, especially in scenarios with many correlated features.
  - **SGDClassifier (Logistic Loss)**: A linear model trained using stochastic gradient descent (SGD) with logistic loss. It is efficient for large datasets and allows flexible control over learning parameters.
  - **Single-Layer Perceptron**: A basic neural network with a single hidden layer. Although still linear in nature when no hidden activation is applied, it serves as a baseline for shallow learning.
- **Comparative Non-Linear Models**:
  - **XGBoost (Extreme Gradient Boosting)**: A powerful ensemble method based on decision trees and gradient boosting. It is used as a non-linear comparative baseline to assess the robustness of linear models under the same oversampling conditions.
  - **SVM (Support Vector Machine with RBF kernel)**: A non-linear model that maps data into higher-dimensional space using a radial basis function kernel. It is effective at separating complex class boundaries and helps assess how synthetic data affects margin-based classifiers.

All models were trained using 5-fold stratified cross-validation, ensuring that the distribution of the target variable was preserved across folds, and their performance and fairness metrics were recorded across different oversampling scenarios. To ensure a fair

comparison across models, hyperparameter tuning was performed using Grid Search within each fold, optimizing for the macro-averaged F1-score. The objective was to compare not only their predictive capabilities but also how the data manipulation impacted the behavior of each model with respect to fairness and sensitivity to the focus group. The full experimental setup is available in the project repository [Porphirio 2025].

4. Results

The results show significant differences in performance metrics across models before and after applying oversampling strategies. Figure 5 illustrates these variations, revealing that the original scenario—without any oversampling—consistently achieved the highest accuracy and precision values across all models. In contrast, recall improved after oversampling was applied, indicating better identification of the minority class, although at the expense of precision. Across all oversampling scenarios, non-linear models — XGBoost and SVM(RBF) — achieved the highest F1-Scores and precision, consistently outperforming linear models by a wide margin. These models demonstrated greater stability and robustness to changes in class distribution, maintaining a good balance between precision and recall even after oversampling interventions.

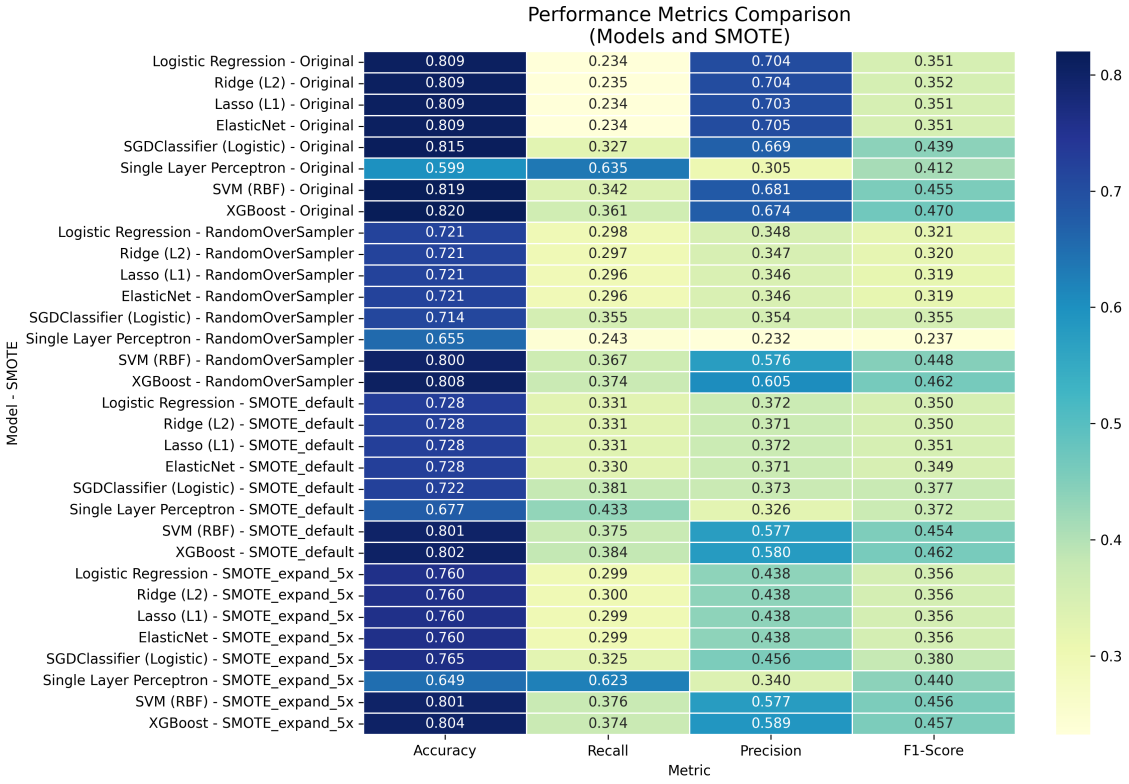

Figure 5. Performance metrics across models and SMOTE strategies.

Figures 6 and 7 compare fairness metrics between gender-sensitive groups (male and female) under different oversampling strategies. For clarity and space limitations, only the charts for Logistic Regression and XGBoost are presented, as they are representative of the patterns observed among linear and non-linear models, respectively. Although fairness metrics vary across sensitive groups, models within the

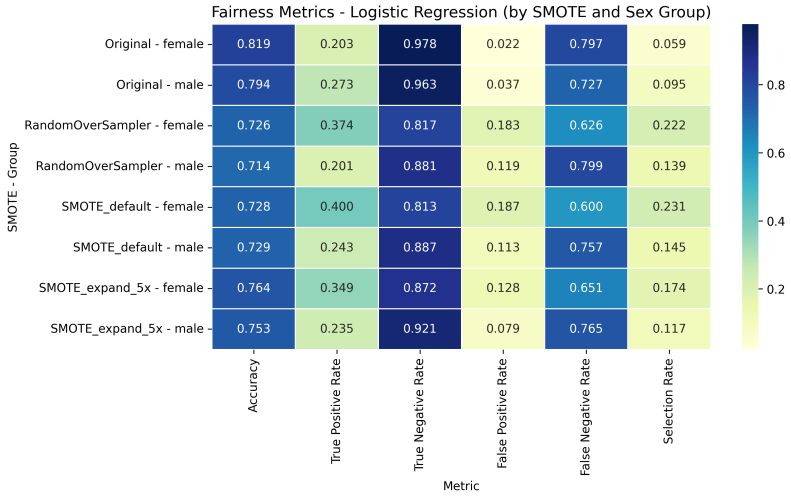

Figure 6. Fairness metrics for Logistic Regression across SMOTE strategies and sex subgroups.

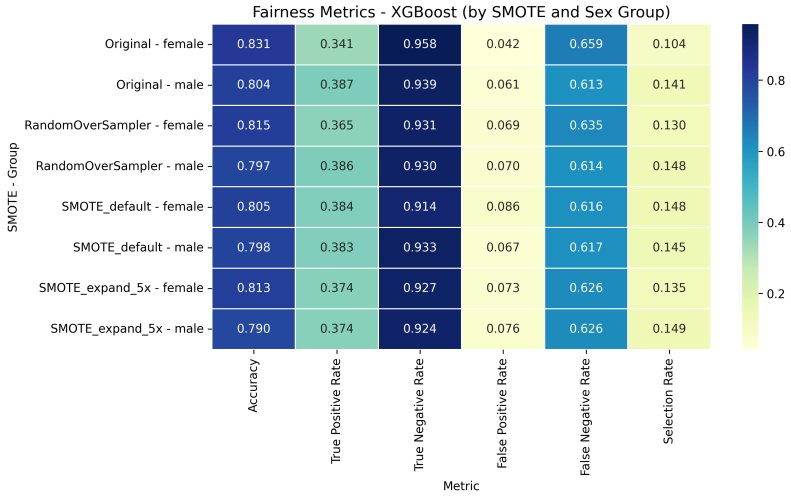

Figure 7. Fairness metrics for XGBoost across SMOTE strategies and sex subgroups.

same class (linear or non-linear) exhibited highly similar behavior, justifying a more concise visualization.

In linear models, represented here by Logistic Regression, the original scenario shows moderate disparities between groups: women tend to have a lower True Positive Rate (TPR) and a higher False Negative Rate (FNR), although still close to male metrics. After applying oversampling strategies, the artificial increase in female defaulters introduces a structural shift in the training data and injects bias into the learning process. As a result, the models begin to classify women more frequently as positives, which leads to a substantial increase in TPR and a decrease in FNR for the female group. However, this forced adjustment also increases the False Positive Rate (FPR) for both groups and reduces overall accuracy, indicating a degradation in predictive quality. This same pattern was observed across all linear models evaluated.

In contrast, non-linear models demonstrated greater robustness to the changes

introduced by oversampling. XGBoost, for example, already exhibits lower group disparity in metrics such as TPR, FNR, and Selection Rate in the original scenario. Even after the application of oversampling techniques, the model remains stable, with minimal variation in both performance and fairness metrics. This behavior was also observed in the SVM (RBF), which displayed patterns very similar to those of XGBoost, reinforcing the resilience of non-linear models to both class imbalance and the demographic distortions introduced by resampling techniques.

Figure 8 presents the differences in fairness metrics computed by sex for each combination of model and SMOTE strategy. These differences quantify disparities between male and female subgroups across key fairness dimensions, including demographic parity—measured by the difference in selection rates—and equal opportunity—measured by the difference in true positive rates between groups. In practice, these metrics assess whether individuals from different groups have the same chance of receiving a favorable outcome (selection) and whether the model is equally capable of correctly identifying positive cases across groups.

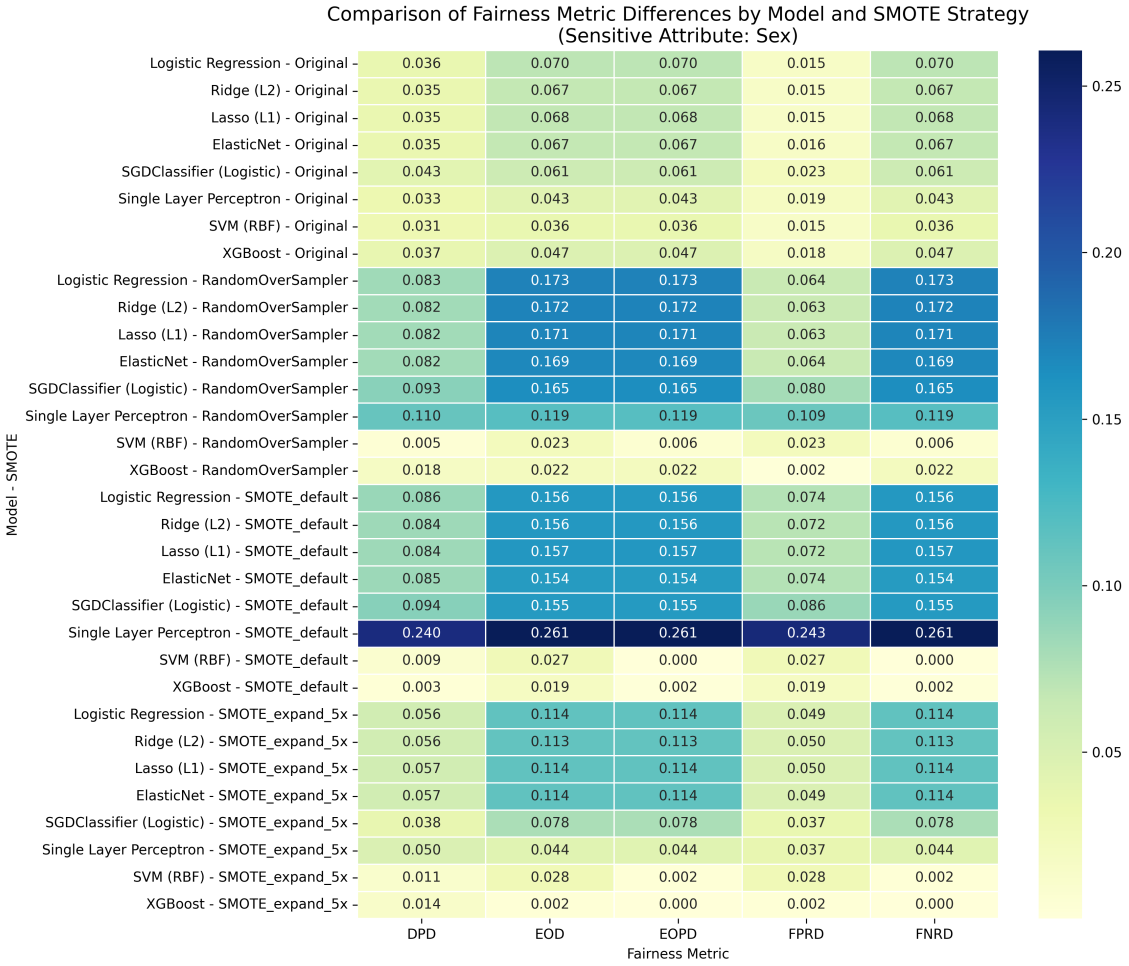

Figure 8. Differences in fairness metrics across models and SMOTE strategies for the sensitive attribute *sex*. Metric abbreviations: DPD = Demographic Parity Difference, EOD = Equalized Odds Difference, EOPD = Equal Opportunity Difference, FPRD = False Positive Rate Difference, FNRD = False Negative Rate Difference.

In the original scenario (without oversampling), the SVM (RBF) achieved the lowest disparities across all fairness metrics, indicating balanced treatment between male and female subgroups. Although linear models performed slightly worse than non-linear models in this scenario, the differences were still relatively modest. However, after applying oversampling strategies, a clear separation emerges: non-linear models—especially XGBoost and SVM—consistently maintain low disparity levels across all metrics, while most linear models begin to show substantial increases in group differences. For example, under the SMOTE strategy, the Single Layer Perceptron model produces the worst result, with a demographic parity difference of 0.24 and an equal opportunity difference of 0.26, highlighting the sensitivity of linear models to artificially induced class imbalance.

The permutation-based feature importance analysis reinforces this pattern. After applying oversampling techniques, variables related to the focus group, such as *marriage = single* and *education = high school*, more frequently appeared among the top ten most important features in most linear models. In contrast, these features were not prominent in the original scenario.

Specifically, the *education = high school* variable consistently ranked among the top three features in ElasticNet, Lasso, Logistic Regression, and Ridge models after oversampling, as shown in Table 2. Similarly, *marriage = single* gained prominence in models such as SGDClassifier and Perceptron with SMOTE. These results suggest that the alteration in class distribution, particularly with the sharp increase in positive examples from the focus group, directly affected the patterns learned by linear models, making these variables more predictive in this new context.

Another important finding concerns the *sex* variable. In all post-oversampling scenarios, its importance score was negative, meaning that permuting this variable improved model performance. This behavior suggests that the variable was being used counterproductively or with bias, especially after SMOTE artificially increased the presence of women in the positive class. This pattern reflects greater sensitivity of these models to the bias introduced by class rebalancing.

Moreover, XGBoost and SVM (RBF) stood out for maintaining a stable distribution of important features, even after oversampling. Although *education = high school* appeared among the top 10 features in some scenarios, *marriage = single* and *sex* remained of low relevance or showed negative importance in both models. This behavior reinforces the robustness of non-linear models to demographic imbalance and artificial alterations introduced in the training data.

4.1. Conclusion

In conclusion, the findings indicate that using imbalanced data can introduce bias in linear models, particularly when oversampling strategies are applied without caution. Linear models seek direct and proportional relationships between inputs and outputs. Distortion in these synthetic data can lead the model to establish incorrect linear relationships with sensitive attributes, affecting fairness.In this context, non-linear models emerge as safer and more stable alternatives, offering greater reliability for applications where algorithmic fairness is critical.

Table 2. Ranking and importance of sensitive variables based on permutation analysis, by model and SMOTE strategy. Bold values indicate that the variable ranked among the top 10 most important features.

Model - SMOTE	**EDUCATION High School**	**MARRIAGE Single**	**SEX**
ElasticNet - Original	27 (-0.001)	26 (-0.001)	15 (0.001)
ElasticNet - RandomOverSampler	**2 (0.020)**	27 (-0.006)	29 (-0.007)
ElasticNet - SMOTE_default	**2 (0.014)**	12 (0.001)	30 (-0.006)
ElasticNet - SMOTE_expand_5x	**2 (0.013)**	24 (-0.001)	30 (-0.005)
Lasso (L1) - Original	28 (-0.002)	26 (-0.001)	15 (0.001)
Lasso (L1) - RandomOverSampler	**2 (0.019)**	27 (-0.006)	29 (-0.007)
Lasso (L1) - SMOTE_default	**2 (0.014)**	13 (0.001)	30 (-0.006)
Lasso (L1) - SMOTE_expand_5x	**2 (0.013)**	24 (-0.001)	30 (-0.005)
Logistic Regression - Original	27 (-0.001)	26 (-0.001)	17 (0.001)
Logistic Regression - RandomOverSampler	**2 (0.020)**	27 (-0.005)	29 (-0.007)
Logistic Regression - SMOTE_default	**2 (0.014)**	12 (0.001)	30 (-0.005)
Logistic Regression - SMOTE_expand_5x	**2 (0.013)**	23 (-0.001)	30 (-0.005)
Ridge (L2) - Original	27 (-0.001)	26 (-0.001)	13 (0.002)
Ridge (L2) - RandomOverSampler	**2 (0.019)**	27 (-0.006)	29 (-0.007)
Ridge (L2) - SMOTE_default	**2 (0.014)**	13 (0.001)	29 (-0.006)
Ridge (L2) - SMOTE_expand_5x	**2 (0.013)**	24 (-0.001)	30 (-0.005)
SGDClassifier (Logistic) - Original	25 (-0.001)	21 (-0.000)	16 (0.001)
SGDClassifier (Logistic) - RandomOverSampler	**2 (0.017)**	**9 (0.002)**	28 (-0.007)
SGDClassifier (Logistic) - SMOTE_default	**4 (0.010)**	**7 (0.004)**	29 (-0.005)
SGDClassifier (Logistic) - SMOTE_expand_5x	**5 (0.011)**	11 (0.003)	25 (-0.000)
Single Layer Perceptron - Original	16 (0.000)	20 (0.000)	13 (0.000)
Single Layer Perceptron - RandomOverSampler	**2 (0.014)**	22 (-0.000)	24 (-0.003)
Single Layer Perceptron - SMOTE_default	**3 (0.007)**	**8 (0.002)**	30 (-0.004)
Single Layer Perceptron - SMOTE_expand_5x	**6 (0.000)**	29 (-0.002)	16 (0.000)
SVM (RBF) - Original	19 (0.0)	30 (-0.0)	29 (-0.0)
SVM (RBF) - RandomOverSampler	**5 (0.008)**	15 (0.002)	19 (0.001)
SVM (RBF) - SMOTE_default	**6 (0.009)**	15 (0.002)	28 (-0.0)
SVM (RBF) - SMOTE_expand_5x	**6 (0.009)**	14 (0.003)	30 (-0.001)
XGBoost - Original	21 (0.000)	17 (0.000)	20 (0.000)
XGBoost - RandomOverSampler	**8 (0.007)**	18 (0.003)	30 (-0.002)
XGBoost - SMOTE_default	**10 (0.004)**	15 (0.002)	30 (-0.001)
XGBoost - SMOTE_expand_5x	11 (0.001)	12 (0.001)	22 (-0.000)

5. Ackowledgement

We thanks Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) and Serasa Experian.

References

Agu, E. E., Abhulimen, A. O., Obiki-Osafiele, A. N., Osundare, O. S., Adeniran, I. A., and Efunniyi, C. P. (2024). Discussing ethical considerations and solutions for ensuring fairness in ai-driven financial services. *International Journal of Frontier Research in Science*, 3(2):001–009.

Bajracharya, A., Khakurel, U., Harvey, B., and Rawat, D. B. (2022). Recent advances in algorithmic biases and fairness in financial services: a survey. In *Proceedings of the Future Technologies Conference*, pages 809–822. Springer.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. 16:321–357.

Choi, Y., Hong, J., Lee, E., Kim, J., and Kim, S. (2025). Enhancing fairness in financial ai models through constraint-based bias mitigation. *Journal of Information Processing Systems*, 21(1):89–101.

Christensen, J. (2021). Ai in financial services. In *Demystifying AI for the Enterprise*, pages 149–192. Productivity Press.

de Castro Vieira, J. R., Barboza, F., Cajueiro, D., and Kimura, H. (2025). Towards fair ai: Mitigating bias in credit decisions—a systematic literature review. *Journal of Risk and Financial Management*, 18(5):228.

Hurlin, C., Pérignon, C., and Saurin, S. (2024). The fairness of credit scoring models. *Management Science*.

Kim, S., Lessmann, S., Andreeva, G., and Rovatsos, M. (2023). Fair models in credit: Intersectional discrimination and the amplification of inequity. *arXiv preprint arXiv:2308.02680*.

LemaÃŽtre, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of machine learning research*, 18(17):1–5.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Porphirio, J. (2025). Credit risk fairness. `https://github.com/jporphirio/credit-risk-fairness`. Accessed: August 6, 2025.

Welfert, M., Stromberg, N., and Sankar, L. (2024). Fairness-enhancing data augmentation methods for worst-group accuracy. *Proceedings of Machine Learning Research*, 279:156–172.

Yeh, I.-C. (2009). Default of Credit Card Clients. UCI Machine Learning Repository.

Zhou, Y., Kantarcioglu, M., and Clifton, C. (2023). On improving fairness of ai models with synthetic minority oversampling techniques. In *Proceedings of the 2023 SIAM international conference on data mining (SDM)*, pages 874–882. SIAM.