Evaluating Racial Fairness in Credit Granting Models

Vinicius Martins Goulart[1] **and Lilian Berton**[1]

[1]Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo (UNIFESP)
São José dos Campos – SP – Brasil

{vmgoulart07,lberton}@unifesp.br

Abstract. *Automated credit models risk propagating socioeconomic disparities. This paper evaluates the trade-off between predictive accuracy and algorithmic fairness in Decision Trees, Random Forests, and XGBoost using four frameworks: (i) Baseline, (ii) Fairness through Blindness, (iii) Counterfactual Fairness, and (iv) In-processing Mitigation (Exponentiated Gradient). Results reveal that tree-based models easily exploit proxy variables under blind conditions, rendering simple race removal ineffective. Conversely, in-processing mitigation strictly enforces demographic parity, actively mapping the critical trade-offs between financial risk discrimination and algorithmic equity.*

1. Introduction

Access to credit is a fundamental pillar of economic mobility, influencing an individual's ability to secure housing, pursue higher education, and build generational wealth. Consequently, the integration of Machine Learning (ML) algorithms by financial institutions to automate credit scoring has sparked critical discussions regarding algorithmic fairness and systemic bias [Barocas and Selbst 2016, Rabonato and Berton 2025, Wecchi and Berton 2026]. While predictive models promise increased efficiency and objectivity, they risk learning, perpetuating, and amplifying historical socioeconomic disparities embedded within training datasets [Mehrabi et al. 2021, dos Santos Porphirio et al. 2025].

Empirical investigations into public financial data consistently reveal that raw financial indicators like income and loan amounts do not fully account for the systematic disparities in credit approval rates across different racial groups. These persistent gaps underscore a need to evaluate how different classifier architectures and feature-handling strategies impact both model performance and algorithmic fairness [Kozodoi et al. 2022].

To evaluate these dynamics, this paper investigates the behavior of three prominent tree-based classification algorithms widely utilized in the financial sector due to their robustness and handling of non-linear relationships: Decision Trees, Random Forests, and Extreme Gradient Boosting (XGBoost). We evaluate these models across four distinct operational frameworks designed to measure, isolate, and mitigate racial bias:

- **Baseline Approach:** Models are trained using all available financial and demographic features, serving as the benchmark for unmitigated predictive performance and demographic disparity [Barocas and Selbst 2016].
- **Fairness through Blindness (Race Removal):** The explicit racial identifiers are completely omitted from the training phase, testing the common but often flawed assumption that ignoring a protected attribute eliminates bias despite the presence of proxy variables [Mehrabi et al. 2021].

- **Counterfactual Fairness Testing:** Models are trained on all features, but at inference time, the protected racial attribute of an individual is synthetically flipped to observe if the classifier alters its final credit decision [Kusner et al. 2017]. This approach directly measures the causal dependency of the model's predictions on race alone.
- **In-processing Mitigation (Exponentiated Gradient):** The models are trained as a constrained optimization problem, explicitly penalizing group-level algorithmic disparities during the learning phase. This evaluates the trade-off between strict predictive accuracy and structural demographic parity [Agarwal et al. 2018].

By contrasting these strategies, this study contributes to the growing body of literature on algorithmic auditing in finance. Our empirical setup aims to identify not only which tree-based architectures are inherently more prone to exploiting racial proxies, but also whether standard mitigation techniques successfully disrupt the residual racial disparities that statistical and baseline models consistently detect in credit data [Kozodoi et al. 2022].

2. Related work

The deployment of tree-based machine learning ensemble, such as Random Forests, and Extreme Gradient Boosting (XGBoost), has drastically improved risk discrimination and predictive performance in consumer credit scoring platforms [Hlongwane et al. 2024]. However, optimizing solely for statistical metrics like overall accuracy or F1-Score poses a substantial risk of perpetuating or amplifying historical inequalities embedded within lending datasets [Barocas and Selbst 2016, Das et al. 2023]. Consequently, auditing credit scoring systems through the lens of algorithmic fairness has evolved into a prominent subfield of financial machine learning [Das et al. 2023, Kozodoi et al. 2022].

Algorithmic auditing frameworks traditionally differentiate between group-level statistical fairness and individual causal fairness [Cornacchia et al. 2023]. Statistical metrics focus on outcome frequencies across demographic segments, with demographic parity serving as a foundational benchmark [Pessach and Shmueli 2023]. Demographic parity dictates that the probability of receiving a positive credit decision should be equal across all protected groups [Pessach and Shmueli 2023].

A pervasive and historical strategy to achieve group fairness has been "fairness through blindness," wherein explicit protected attributes (e.g., race) are entirely removed from the training data [SD 2026]. However, literature consistently demonstrates that omitting sensitive features fails to mitigate bias because complex non-linear models, particularly tree-based ensembles, excel at reconstructing the protected attributes using correlated socioeconomic proxies such as income, loan amount, and geographic data [SD 2026, Pessach and Shmueli 2023].

To address the mathematical limitations of explicit feature blindness, recent algorithmic auditing literature has shifted toward in-processing mitigation frameworks. A foundational method in this domain is the reduction approach proposed by Agarwal et al. [Agarwal et al. 2018], which reformulates fair classification as a constrained optimization problem. By applying an exponentiated gradient descent algorithm, this methodology treats fairness as a sequence of cost-sensitive learning problems. It dynamically updates sample weights during training to balance the predictive loss function against strict struc-

tural fairness constraints, such as demographic parity or equalized odds, generating a randomized ensemble that provides provable theoretical bounds on algorithmic discrimination [Agarwal et al. 2018].

To overcome the limitations of group-level metrics, individual fairness paradigms demand that "like cases be treated alike" [Cornacchia et al. 2023]. A major advancement in this domain is counterfactual fairness, formalized through structural causal models [Kusner et al. 2017]. Under this definition, a predictive model is considered counterfactually fair toward an individual if its prediction remains invariant in both the actual world and a counterfactual world where the individual's protected attribute is synthetically altered while all non-descendant features are held constant [Kusner et al. 2017, Cornacchia et al. 2023].

In practical credit applications, full counterfactual testing presents unique challenges due to the strict and often untestable assumptions required to map entire causal systems or latent background variables [Ivanov 2026]. To bridge this gap, recent studies employ localized counterfactual perturbations or post-hoc auditing [Ivanov 2026, Cornacchia et al. 2023]. These methods simulate counterfactual profiles at inference time by keeping factual creditworthiness indicators constant and flipping only the protected attribute to directly isolate explicit bias and evaluate local prediction consistency [Kusner et al. 2017, Cornacchia et al. 2023]. While tree-based classifiers exhibit varying degrees of robustness when exposed to demographic perturbations [Hlongwane et al. 2024], our study builds upon this foundation by explicitly comparing how different operational feature frameworks—ranging from full baseline inclusion and explicit feature blindness to counterfactual demographic flipping—manifest across Decision Tree, Random Forest, and XGBoost architectures within credit-granting contexts.

3. Methodology

This section outlines the empirical framework used to evaluate racial fairness across tree-based classifiers in automated credit granting. Our methodology is structured into three main phases: (i) pipeline definition and predictive modeling, (ii) operational feature configuration design, and (iii) algorithmic bias and counterfactual auditing.

3.1. Classification Models

To capture complex, non-linear interactions within credit scoring datasets, we employ three distinct tree-based classification paradigms:

- **Decision Tree (DT):** Serves as our structural baseline model, splits the feature space orthogonally based on impurity criteria (e.g., Gini index or entropy) to establish deterministic, human-interpretable routing rules.
- **Random Forest (RF):** An ensemble bagging approach that constructs a large pool of uncorrelated decision trees trained on parallel bootstrapped samples of the dataset [Breiman 2001]. It averages individual tree outputs to control variance and prevent overfitting to proxy features.
- **Extreme Gradient Boosting (XGBoost):** An optimized, sequential boosting framework that minimizes regularized objective functions [Chen and Guestrin 2016]. It fits successive trees to the residual errors of preceding iterations, providing high predictive performance in tabular risk models.

For each classifier, the target variable $Y \in \{0, 1\}$ represents the credit decision, where $Y = 1$ signifies credit approval and $Y = 0$ indicates rejection.

3.2. Experimental Design and Feature Configuration Approaches

To isolate the mechanisms through which racial bias manifests in the models, we evaluate each of the three tree-based classifiers across three operational feature setups with an additional fourth step for the XGBoost model, related to the In-process mitigation method exhibit in Figure 1:

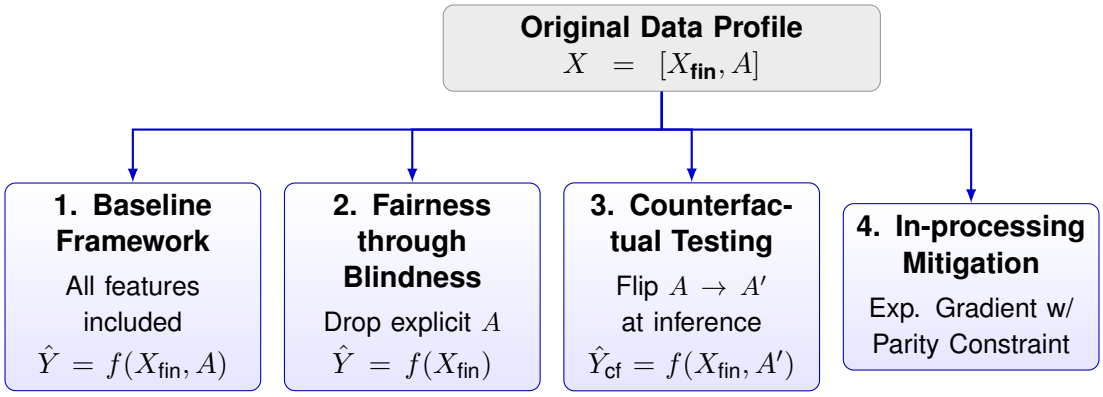


Figure 1. System Architecture of Feature Handling Operational Configurations.

3.2.1. Baseline Framework (All Features)

In this setup, the models are trained on the complete feature matrix $X = [X_{\text{fin}}, A]$, where X_{fin} denotes structural financial and transactional attributes (e.g., income, loan amount) and A denotes the protected sensitive attribute representing race ($A \in \{\text{Group}_1, \text{Group}_2, \ldots, \text{Group}_n\}$). This configuration serves as our performance benchmark:

$$\hat{Y}_{\text{base}} = f(X_{\text{fin}}, A) \tag{1}$$

3.2.2. Fairness through Blindness (Race Removal)

This configuration evaluates the efficacy of explicit attribute suppression, a historical standard for compliance. The sensitive attribute A is completely purged from the dataset prior to training, restricting the classifier's input matrix exclusively to X_{fin}:

$$\hat{Y}_{\text{blind}} = f(X_{\text{fin}}) \tag{2}$$

This setup tests the resilience of tree-based models against proxy variable exploitation, analyzing whether the classifiers reconstruct the missing race attribute A using highly correlated socioeconomic indicators residing within X_{fin}.

3.2.3. Counterfactual Robustness Testing (Demographic Attribute Flipping)

To explicitly capture individual causal dependence on race, we establish a post-hoc counterfactual testing strategy. The model is originally trained using the complete baseline matrix $X = [X_{\text{fin}}, A]$. However, during the inference phase, the financial attributes X_{fin} are held strictly constant while the demographic label A is synthetically perturbed to an alternative racial category A' ($A' \neq A$):

$$\hat{Y}_{\text{cf}} = f(X_{\text{fin}}, A') \tag{3}$$

By evaluating whether a synthetic change in race flips the credit prediction ($\hat{Y}_{\text{base}} \neq \hat{Y}_{\text{cf}}$), this metric measures explicit model dependence on demographic indicators without altering the applicant's underlying financial profile.

3.2.4. In-processing Mitigation (Exponentiated Gradient)

To evaluate the impact of explicitly incorporating algorithmic fairness into model training, we implement the Exponentiated Gradient (EG) reduction algorithm [Agarwal et al. 2018]. As an in-processing mitigation method, EG solves fair classification through a constrained optimization procedure, where predictive risk is minimized subject to a fairness constraint. In this study, we target Demographic Parity. Due to the high computational complexity of iteratively fitting multiple base estimators over a large-scale dataset, EG was applied specifically to XGBoost. This architectural choice is motivated by XGBoost's status as the top-performing predictive baseline and the prevailing industry standard for tabular credit risk modeling, making it the most critical candidate for evaluating the frontier between optimal predictive power and strict fairness constraints.

A convenient way to express the objective is through a Lagrangian formulation:

$$\min_{h \in \mathcal{H}} \max_{\lambda \geq 0} \left[L(h(X), Y) + \lambda \, \Delta(h, A) \right], \tag{4}$$

where L denotes the predictive loss, $\Delta(h, A)$ measures violation of the fairness constraint, and λ is the corresponding Lagrange multiplier. In the EG reduction, these dual variables are updated multiplicatively across iterations, while the base learner is repeatedly trained under reweighted costs induced by the current constraint violations.

Rather than yielding a single deterministic classifier, EG typically produces a randomized mixture over classifiers that balances predictive performance and fairness. This makes it well suited for empirically studying the trade-off between classification performance and demographic parity, although F1-Score is used here only as an evaluation metric rather than as part of the optimization objective itself.

3.3. Evaluation Metrics

The interaction between predictive performance and fairness is measured across all configurations using a joint validation suite. Based on the specific dynamics of our target dataset, the models are evaluated using the following taxonomy show in Table 1:

- **Predictive Performance (Precision, Recall, F1-Score):** To assess the classification quality, we evaluate Precision and Recall. However, the primary predictive

Table 1. Taxonomy of Evaluation Metrics

Category	Specific Metrics Used
Predictive Performance	Precision, Recall, & F1-Score
Group Fairness	Demographic Parity Ratio (DPR) Demographic Parity Difference (DP Diff)
Individual Fairness	Prediction Flip Rate

benchmark is the F1-Score. As the harmonic mean of Precision and Recall, the F1-Score provides a robust evaluation of model performance that is highly sensitive to the structural class imbalances commonly found in credit approval datasets.

- **Demographic Parity Ratio (DPR) and Difference (DP Diff):** These metrics quantify group-level statistical fairness by examining the allocation of positive outcomes. DPR computes the ratio of acceptance rates between the unprivileged and privileged groups. DP Diff measures the absolute margin between these approval probabilities:

$$\text{DP Diff} = P(\hat{Y} = 1 \mid A = \text{privileged}) - P(\hat{Y} = 1 \mid A = \text{unprivileged}) \quad (5)$$

  A perfectly fair model under this statistical paradigm achieves a DPR of 1.0 and a DP Diff of 0.0.
- **Prediction Flip Rate:** Evaluates the individual causal vulnerability of the baseline models. This metric measures the proportion of errors whose predictions change when only the protected demographic attribute is altered:

$$\text{PFR} = \frac{\text{Flips}}{\text{Total Errors}} \times 100\% \quad (6)$$

  where Flips is the number of test instances whose predictions change after flipping race, and Total Errors is the total number of model errors.

3.4. Dataset and Feature Engineering

The empirical analysis in this study is conducted using the publicly available Home Mortgage Disclosure Act (HMDA) dataset [Barocas and Selbst 2016,Das et al. 2023]. Enacted by the United States Congress in 1975, the HMDA mandates that financial institutions report detailed, loan-level registers regarding housing-related credit applications to identify potentially discriminatory lending patterns and measure structural public access to mortgage systems [Das et al. 2023].

The dataset contains a rich set of demographic, financial, and transactional attributes. Our experimental pipeline targets five primary core variables to model individual credit risk and algorithmic bias profiles:

- **`derived_race` / `applicant_race-1`:** The explicit categorical variable indicating the self-reported race of the primary credit applicant, serving as the target protected sensitive attribute (A). In accordance with the HMDA data dictionary, the privileged group is defined as White applicants (encoded as 5), while the unprivileged minority group analyzed is defined as Black or African American applicants (encoded as 3).

- **action_taken:** The binary target label (Y), where originations and accepted pre-approvals are mapped to a positive credit outcome ($Y = 1$), and explicitly denied applications are mapped to a negative outcome ($Y = 0$).
- **income:** Continuous indicator representing the gross annual income of the applicant (in thousands of dollars).
- **loan_amount:** Continuous metric specifying the exact monetary value of the mortgage amount requested.
- **interest_rate:** The numerical pricing index denoting the contractual interest rate assigned to the financial transaction.

3.5. Data Preprocessing and Exploratory Transformations

Raw credit registers frequently exhibit data quality constraints and administrative noise that can distort statistical learning algorithms. To ensure robust model training and computational efficiency, our data preparation framework implements the following pipeline:

1. **Target Engineering and Data Reduction:** The dataset is randomly downsampled to 30% using a fixed random seed (SEED = 42) to optimize computational overhead while preserving statistical representativeness, resulting in a massive final dataset of 3,668,789 instances. The raw credit decision (`action_taken`) is mapped into a binary target, where originations and approved pre-approvals are classified as positive outcomes (1), and all others as negative (0). Redundant administrative metadata and secondary demographic identifiers are systematically removed.
2. **Missing Value Imputation:** Financial columns containing textual artifacts (e.g., "Exempt") are coerced into missing values. Numerical features are subsequently imputed using the median of their respective distributions, while missing categorical variables are explicitly isolated using a constant "Missing" label.
3. **Feature Scaling and Encoding:** To normalize distributions for the baseline algorithms, numerical variables are standardized using zero-mean and unit-variance scaling. Categorical features are converted into discrete numerical representations via ordinal encoding.

To establish a baseline for algorithmic fairness, we analyzed the historical approval rates within this final dataset. White applicants (the privileged group) represent the majority at 57.74% (2,118,277 instances) and historically experience an approval rate of 62.43%. In contrast, Black applicants (the unprivileged group) constitute 8.57% of the dataset (314,572 instances) and face a significantly lower historical approval rate of 48.43%. This initial discrepancy establishes a clear structural disparity in the raw data prior to any machine learning intervention. Preliminary exploratory data analysis reveals significant distributional variations across racial segments. A non-parametric Kruskal-Wallis test applied to compare income distributions across distinct demographic lines yielded a test statistic corresponding to a p-value of $p = 0.0$, indicating statistically significant differences among group distributions.

Furthermore, stratification experiments across fixed income intervals show that cross-demographic disparities in approval rates remain constant even when conditioning on identical income brackets, just as shown in Figure 2. This finding strongly implies that income alone does not fully explain the approval disparities observed across groups, providing a clear justification for evaluating tree-based classifiers using counterfactual and blindness frameworks.

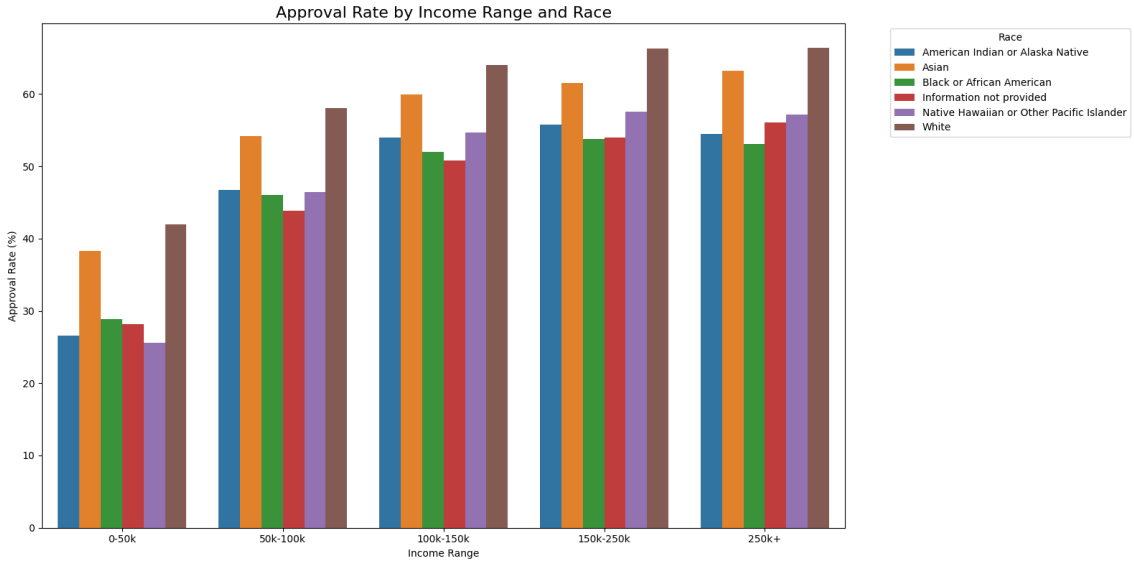


Figure 2. Approval rate for each race among different income brackets.

3.6. Experimental Setup and Model Tuning

To ensure rigorous evaluation and reproducibility, the dataset was partitioned into three subsets using a stratified holdout approach: Training (64%), Validation (16%), and Test (20%). Stratification was applied based on the target variable (`loan_approved`) to maintain consistent class distributions across all partitions, explicitly handling class imbalance without relying on synthetic data generation, which could distort underlying demographic distributions.

Hyperparameter optimization was conducted using the Optuna framework. For each model, we ran 5 optimization trials targeting the maximization of the F1-Score on the Validation set. The hyperparameter search spaces were defined as follows:

- **Decision Tree:** `max_depth` $\in [3, 10]$, `min_samples_split` $\in [2, 10]$.
- **Random Forest:** `n_estimators` $\in \{50, 100, 200\}$, `max_depth` $\in [3, 10]$.
- **XGBoost:** `n_estimators` $\in \{50, 100, 200\}$, `learning_rate` $\in [0.01, 0.20]$, `max_depth` $\in [3, 7]$.

Once optimized, the models were evaluated on the isolated Test set. To ensure spatial uniformity during the Counterfactual Twin generation, a standard scaler (zero-mean, unit-variance) was applied to all numerical features.

4. Results

Table 2 presents the predictive performance of the evaluated models. The baseline tree-based classifiers exhibit near-perfect F1-scores, indicating a strong capacity to map the training data. The application of fairness through blindness (Unaware - UA models) results in negligible performance degradation, maintaining F1-scores above 0.99. This demonstrates that removing the explicit racial attribute does not hinder the models' ability to predict default risk. However, the introduction of the Exponentiated Gradient mitigation framework (XGBoost + ExpGrad-DP) induces a noticeable predictive trade-off, reducing Precision to 0.8662 and the overall F1-score to 0.9266.

Table 2. Predictive performance of the evaluated models.

Model	Precision	Recall	F1-Score
Decision Tree	0.9977	0.9941	0.9959
Decision Tree (UA)	0.9985	0.9954	0.9970
Random Forest	0.9956	0.9927	0.9941
Random Forest (UA)	0.9924	0.9969	0.9947
XGBoost	0.9984	0.9965	0.9974
XGBoost (UA)	0.9984	0.9964	0.9974
XGBoost + ExpGrad-DP	0.8662	0.9961	0.9266

The cost of this predictive trade-off becomes evident when analyzing the algorithmic equity metrics detailed in Table 3. Both the baseline and unaware (UA) configurations fail to achieve group fairness, exhibiting Demographic Parity Differences around 0.21 and DP Ratios roughly at 0.66. This confirms that explicitly ignoring race does not mitigate systemic disparities, as the models successfully reconstruct demographic profiles via financial proxies. Furthermore, the prediction flip rate reveals that the Random Forest and XGBoost baseline models are vulnerable to demographic perturbations, with 0.58% and 1.20% of their errors changing after a synthetic change in race, respectively. Interestingly, the Decision Tree ended up with a 0%, which indicates a solution that is not based on the race category itself, however, its DP diff and DP ratio suggests that the solution found used proxy categories that reproduced the bias. Conversely, the XGBoost model penalized with Exponentiated Gradient effectively closes the fairness gap, achieving a DP Difference of 0.0163 and a DP Ratio of 0.9738, thereby satisfying strict demographic parity constraints.

Table 3. Fairness metrics of the evaluated models.

Model	DP Diff	DP Ratio	Prediction Flip Rate
Decision Tree	0.2123	0.6589	0.00%
Decision Tree (UA)	0.2117	0.6599	–
Random Forest	0.2114	0.6603	0.58%
Random Forest (UA)	0.2147	0.6579	–
XGBoost	0.2119	0.6599	1.20%
XGBoost (UA)	0.2117	0.6601	–
XGBoost + ExpGrad-DP	**0.0163**	**0.9738**	–

5. Discussion

The experimental evaluation of tree-based models under different feature-handling configurations highlights a critical tension between optimization-driven credit assessment and algorithmic equity. Our results demonstrate that while the complete suppression of explicit protected attributes ("fairness through blindness") remains a standard corporate practice to comply with non-discrimination legislation, it fails to eliminate racial disparities. This failure stems from proxy variable leakage, wherein high-capacity ensemble models such as Random Forests and XGBoost efficiently reconstruct demographic profiles using highly correlated financial features like income, interest rates, and loan amounts. When scrutinized under individual counterfactual auditing—synthetically flipping racial indicators at inference time—the baseline models demonstrate fragile consis-

tency, showing that minor demographic perturbations can explicitly alter credit classification outcomes even when the underlying economic risk remains unchanged.

To fully understand these mathematical vulnerabilities, algorithmic auditing cannot be decoupled from the socioeconomic structures that shape credit registries. Although our empirical experiments rely on the US HMDA dataset—owing to its regulatory transparency and granular demographic attributes, which provide an ideal and necessary auditing benchmark rarely available in open data regimes elsewhere—the mathematical mechanisms of proxy leakage observed here transcend geographical borders. In emerging markets such as Brazil, credit ecosystems similarly reflect a historical accumulation of systemic socioeconomic inequalities rather than neutral risk indicators. Consequently, the predictive dependencies exhibited by our audited classifiers provide critical structural parallels to the dynamics of financialized consumption and credit precarity described in recent Brazilian literature.

Specifically, Santos (2026) challenges the traditional, optimistic narrative of "financial inclusion" in his study *Parcelado: dinâmicas de consumo na periferia* [Santos 2026]. Santos observes that the modern integration of low-income, predominantly Black populations into formal credit systems coexists with severe structural precarity [Santos 2026]. Within the Brazilian retail and banking sectors, access to credit occurs under highly unfavorable operational terms characterized by exorbitant interest rates and compounding financial costs [Santos 2026]. Rather than functioning as an instrument of economic mobility or wealth generation, credit acts as an expensive mechanism for managing daily scarcity, transforming historical exclusion into a process of predatory subordination to financial capital [Santos 2026].

Our machine learning auditing directly mirrors the structural mechanics of this economic extraction. Empirical data from the Brazilian credit sector confirms that Black applicants face structurally higher interest rates and lower approval limits, exposing a cumulative cycle of racial disadvantage where systemic bias is continuously reinforced. Credit architectures in Brazil are built around high-rate consumption financing (such as credit card revolving lines and direct retailer consumer credit), which prioritizes the extraction of resources from low-income groups over sustainable financial inclusion.

When tree-based machine learning algorithms are trained on datasets emerging from this highly financialized and asymmetrical ecosystem, they do not just learn to predict risk; they formalize and optimize the extraction of resources. If an algorithmic model determines that a specific racial group exhibits a higher default probability or must be subjected to higher interest rates to mitigate risk, it is structurally codifying the historical economic extraction detailed by Santos [Santos 2026]. Consequently, removing the race variable does not prevent the model from identifying and penalizing these marginalized profiles via proxy metrics. True algorithmic fairness in financial modeling requires moving away from feature blindness and adopting rigorous counterfactual regularizations that actively prevent classifiers from reproducing the compounding, asymmetric cycles of the financial market.

6. Conclusion

This study evaluates the intersection of predictive performance and algorithmic fairness in credit scoring utilizing tree-based classification frameworks. Our empirical analysis

conclusively demonstrates that the prevalent industry practice of "fairness through blindness"—the explicit omission of racial identifiers—fundamentally fails to mitigate demographic disparities. High-capacity ensemble models, particularly Random Forests and XGBoost, efficiently reconstruct protected attributes through socioeconomic proxy variables, thereby maintaining significant demographic parity gaps. Furthermore, localized counterfactual testing exposes the fragile causal consistency of these models, revealing that identical financial profiles can yield divergent credit outcomes when only the explicit racial attribute is synthetically altered.

To effectively break the cycle of proxy exploitation, proactive in-processing interventions are essential. Our results show that implementing the Exponentiated Gradient reduction method successfully enforces strict demographic parity, reducing the statistical disparity difference to near zero. While this constraint-driven approach inevitably introduces a trade-off with predictive precision, it provides a mathematical guarantee that the compounding, asymmetrical cycles of historical economic exclusion are not mechanically optimized and reproduced by the algorithmic system. Future research should prioritize the integration of structural causal models to map the deeper socio-economic variables driving these disparities, ultimately guiding the design of more equitable financial technologies.

Acknowledgments

The authors would like to thank Serasa Experian, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) grant 2021/14725-3, for supporting the development of this study.

Declaration of Generative AI and AI-Assisted Technologies

During the preparation of this work, the authors used AI-assisted tools and generative artificial intelligence technologies (specifically, Gemini) to polish the phrasing, enhance grammatical readability, and translate draft text into structured English prose. After using these services, the authors thoroughly reviewed, edited, and verified all generated content, and take full responsibility for the final scientific accuracy.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104:671.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Cornacchia, G., Anelli, V. W., Narducci, F., Ragone, A., and Di Sciascio, E. (2023). Counterfactual fair opportunity: Measuring decision model fairness with counterfactual reasoning. Technical Report arXiv:2302.08158, arXiv.

Das, S., Stanton, R., and Wallace, N. (2023). Algorithmic fairness. *Annual Review of Financial Economics*, 15:565–593.

dos Santos Porphirio, J., dos Santos, D. J., Azevedo, S., and Berton, L. (2025). Analyzing data bias through data augmentation: A case study in financial data. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 950–961. SBC.

Hlongwane, R., Ramaboa, K. K. K. M., and Mongwe, W. (2024). Enhancing credit scoring accuracy with a comprehensive evaluation of alternative data. *PLOS ONE*, 19(5):e0303566.

Ivanov, A. (2026). Fairness calibration in credit scoring via counterfactual perturbation and group-wise regularization. Technical report, Preprints.org.

Kozodoi, N., Jacob, J., and Lessmann, S. (2022). Fairness in credit scoring: An assessment of optimization algorithms and resource allocation. *Expert Systems with Applications*, 207:117951.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems (Rules of NeurIPS)*, volume 30.

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.

Pessach, D. and Shmueli, E. (2023). Algorithmic fairness. In *Machine Learning for Data Science Handbook*, pages 867–886. Springer International Publishing.

Rabonato, R. T. and Berton, L. (2025). A systematic review of fairness in machine learning. *AI and Ethics*, 5(3):1943–1954.

Santos, K. L. d. (2026). *Parcelado: dinâmicas de consumo na periferia*. Editora Fósforo, São Paulo.

SD, N. (2026). Structural gender bias in credit scoring: Proxy leakage. Technical Report arXiv:2601.18342, arXiv.

Wecchi, M. and Berton, L. (2026). Algorithmic fairness and bias mitigation in financial artificial intelligence: scoping review. *AI and Ethics*, 6(4):409.