Detecting and Mitigating Algorithmic Bias in Credit Models: A Tutorial Using Public Financial Data

Júlia Wotzasek Pereira
Universidade Federal de São Paulo - UNIFESP
pereira.julia@unifesp.br

***Abstract*—This tutorial presents a practical walk-through for detecting and mitigating algorithmic bias in credit scoring models using public financial datasets. We demonstrate the application of fairness metrics and mitigation techniques via open-source toolkits such as AIF360 and Fairlearn, with a focus on responsible and equitable AI in financial decision-making.**

***Index Terms*—Algorithmic Fairness, Credit Scoring, Bias Mitigation, Financial Data, Data-Centric AI, AIF360, Fairlearn**

I. Introduction

The adoption of automated decision-making systems in financial services has expanded rapidly, particularly in credit scoring and loan approval workflows. Machine learning models are increasingly used to assess creditworthiness based on historical applicant data. While these systems offer efficiency and scalability, they also inherit risks from the data on which they are trained. If sensitive or socially correlated features such as gender, age, or nationality are present—either directly or through proxies—the model may produce outcomes that disproportionately disadvantage certain groups.

This concern has brought algorithmic fairness to the forefront of AI governance and risk management. In credit decision contexts, unfair treatment can lead not only to reputational and ethical issues, but also to regulatory non-compliance. Traditional approaches to fairness in machine learning often focus on removing protected attributes from the dataset (fairness through unawareness), but this is insufficient when bias persists via correlated features. To address this, the field has developed a range of fairness-aware methods for auditing and mitigating bias throughout the model lifecycle. However, many practical implementations—especially in financial applications—lack accessible, reproducible workflows that illustrate these methods using real-world data.

II. Identified Gaps

This work aims to fill this gap by presenting a hands-on tutorial for auditing and mitigating algorithmic bias in a credit scoring scenario. Using the South German Credit dataset, we walk through the full modeling pipeline, from data preparation and sensitive attribute analysis to fairness evaluation and mitigation. Our goal is to provide a reproducible baseline for researchers and practitioners seeking to understand how fairness tools such as AIF360 can be applied in practice to financial decision models.

III. Related Work

Studies such as Barocas and Selbst (2016) and Mehrabi et al. (2021) discuss sources and implications of algorithmic bias. Huston et al. (2023) review fairness metrics in credit scoring. Toolkits like AIF360[1] and Fairlearn[2] offer technical support for evaluating and mitigating bias. Our tutorial builds on these to offer a reproducible example applied to credit data.

IV. Materials and Methods

A. Dataset

1) Origin and Description: We employ the South German Credit Data from the UCI Machine Learning Repository [3], which contains 1000 anonymized credit applications collected between 1973 and 1975. This dataset is a refined version of the classic Statlog German Credit Data [4], but fixing some inconsistencies and providing corrections about the background information [3].

2) Feature Types and Preprocessing: This dataset contains 14 categorical variables, 2 boolean variables, and 3 numerical variables. Since the columns are written in German, we first translate their names using the codetable provided in [3]. Several features are coded numerically, so we also use the codetable to translate the numbers to interpretable labels.

3) Target Variable: The binary target variable, `credit_risk`, represents the bank's assessment of the applicant: `0 = bad credit` and `1 = good credit`. It is important to note that the dataset contains an oversampled number of bad credit cases—a common approach in domains with imbalanced outcomes. However, this means that the observed class distribution does not reflect the real-world prevalence of defaults, showed in Fig. 1.

4) Sensitive Attributes and Proxy Risk: We identify three features as sensitive for the purposes of fairness analysis:

- **age** – a continuous variable that may be subject to age-based discrimination;
- **foreign_worker** – a binary variable associated with nationality and potential racial bias;
- **personal_status_sex** – a categorical feature encoding both gender and marital status.

In addition to directly sensitive attributes, we must also consider *proxy attributes*: variables that are not explicitly sensitive but may carry correlated information about protected characteristics [5]. For example, employment duration or housing status may reveal age-related patterns, while credit amount

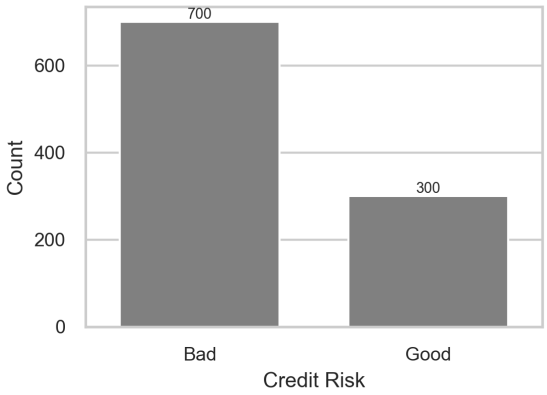

Fig. 1. Credit risk distribution (good vs bad).

may indirectly relate to gender roles in financial responsibility. In Subsection V-A2, we present an empirical analysis of such correlations to identify latent proxies.

V. Tutorial Workflow

In this section, we outline the full pipeline used to audit and mitigate bias in the South German Credit dataset. The process is divided into seven distinct stages, each aimed at identifying, quantifying, and addressing potential algorithmic unfairness.

A. Pipeline Overview

The tutorial follows this sequence:

1) Data loading and initial preprocessing;
2) Mapping sensitive and proxy attributes;
3) Exploratory Data Analysis (EDA): feature distribution, class balance, and sensitive attribute correlation;
4) Base model training: Logistic Regression and Decision Trees using Scikit-learn;
5) Fairness Evaluation: applying metrics such as Demographic Parity Difference and Equal Opportunity Difference using AIF360;
6) Bias Mitigation: using pre-processing (Reweighing) and post-processing (Reject Option Classification) techniques;
7) Result Comparison: measuring accuracy and fairness before and after mitigation.

Each step is detailed in the following subsections.

1) Data Loading and Initial Preprocessing: We read the South German Dataset downloaded from its original ASCII format and assigned meaningful English columns names in replace to the German ones. Also, we used the codetable to replace the numerical values of the categorical variables to meaningful categories.

We also analysed the existence of missing values or inconsistencies and nothing was found.

2) Mapping Sensitive and Proxy Attributes: Three sensitive variables were selected: `age`, `foreign_worker`, and `personal_status_sex`. We also analyzed potential proxy variables using correlation and mutual information to determine whether non-sensitive features encode sensitive information indirectly.

To support this analysis, we computed and visualized the top features most strongly correlated or informationally linked to each sensitive variable. These visualizations help identify potential proxy attributes that may leak sensitive information.

In Fig. 2 the most correlated features with `age` are `employment_duration`, `residence_since`, and `job`, which are clearly related to one person's stage in life. This is an indicator that those features may be proxy sensitive, indicating a risk of age-related patterns within them.

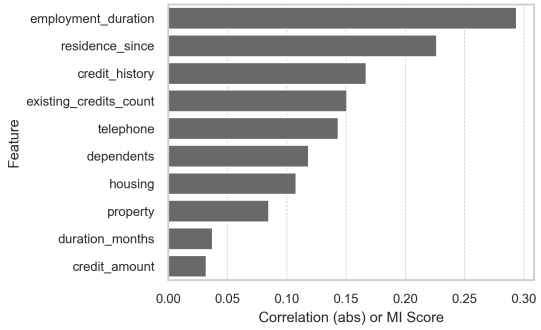

Fig. 2. Top correlated and informative features with `age`.

In Fig. 3, there is a moderate information gain from `credit_amount`, which indicates that financial variables such as previous credit provided may reflect differences in treatment for foreign applicants.

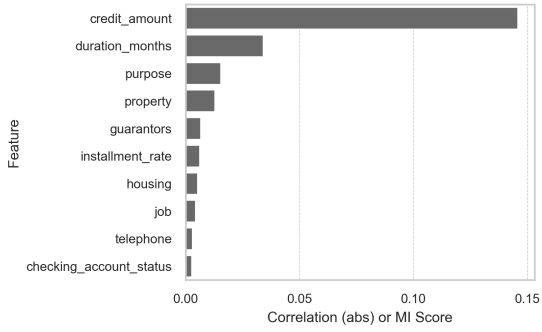

Fig. 3. Top correlated and informative features with `foreign_worker`.

In Fig. 4, there is a considerable high correlation between `personal_status_sex` and `credit_amount`, which indicates again that this financial variable may bring unfairness to the model. For now we will include `credit_amount` as a **sensitive proxy attribute**.

3) Exploratory Data Analysis (EDA): We examined the distribution of each feature, checked for class balance, and

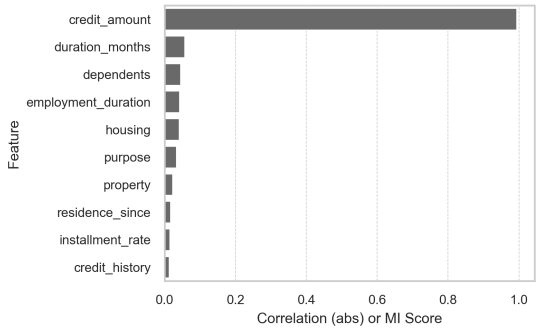

Fig. 4. Top correlated and informative features with `personal_status_sex`.

visualized relationships between the target and sensitive attributes.

In Fig. 5 we can notice that the applicants with bad credit tend to be slightly younger on average in comparison with good credit risk ones. The bad curve also has a earlier around late 20s, while the good group is more spreaded. It suggests that age plays an implicit role in credit evaluation.

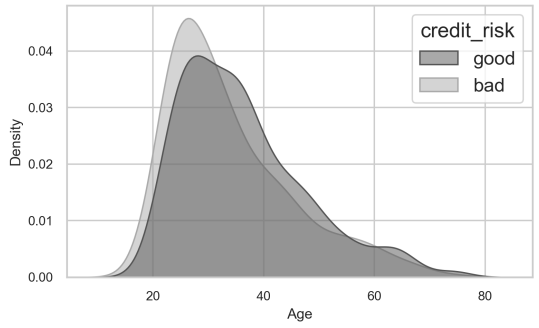

Fig. 5. Age vs. Credit Risk.

In Fig. 6 we can notice that the disparity between domestic applicants (`Foreign Worker == 'no'`) receiving **good** credit rating is high (667 good compared to 296 bad credit rating) while for foreigners it is considerably lower (33 good compared to 4 bad credit). Notice also that the foreigner group is under-represented (only 3.7% of the total data), limiting the model's ability to generalize fairly for this group, what amplifies the risk of discrimination.

For gender-status subgroups there are substantial differences. Male applicants, if married or widowed, receive the major number of approvals, followed by female applicants which are non-single or single male. In contrast, single female and divorced male have lower absolute counts and more balanced negative distribution. These differences suggest that marital status and gender may influence credit decisions. The underlying historical bias in these categories may result in unfair treatment, especially for under-represented groups such as single women or divorced men.

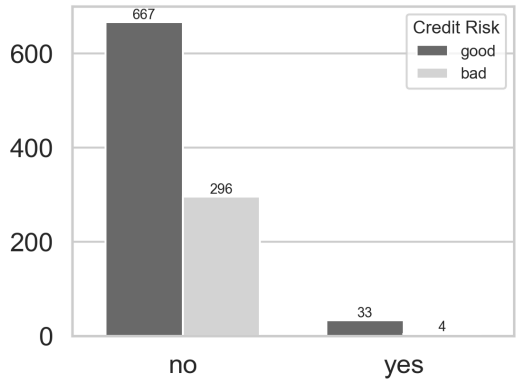

Fig. 6. Foreigner Worker vs. Credit Risk.

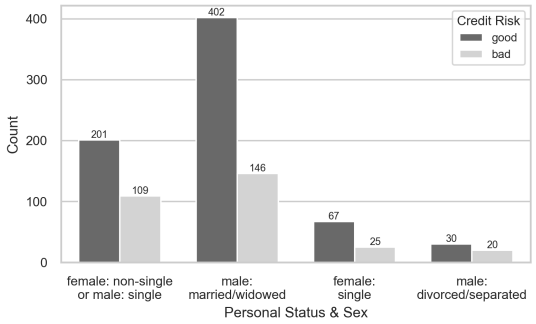

Fig. 7. Personal Status vs. Credit Risk.

4) Baseline Model: Random Forest: Following the Statlog benchmark, we implemented a Random Forest classifier using Scikit-learn with 100 estimators and default hyperparameters. The dataset was split into training and test sets with a 70/30 ratio, and one-hot encoding was applied to categorical features. The sensitive attributes were excluded from training to simulate fairness through unawareness.

The model achieved an overall accuracy of 0.76. The classification report showed higher recall and precision for the positive class (good credit), but poor performance in identifying bad credit cases:

- **Accuracy:** 0.76
- **F1-score (class 0):** 0.48
- **F1-score (class 1):** 0.84

5) Fairness Evaluation: To evaluate fairness, we computed the *Demographic Parity Difference (DPD)* across three sensitive attributes: `foreign_worker`, `personal_status_sex`, and binned `age`. The results revealed the following disparities in selection rates between groups:

- **DPD (foreign_worker):** 0.2517
- **DPD (personal_status_sex):** 0.1512
- **DPD (age):** 0.0580

These disparities indicate that the model indirectly encodes

bias through correlated features, even when sensitive variables are excluded. This motivates the need for mitigation techniques in the next stage.

6) Fairness Mitigation via Reweighing: To mitigate group-based disparities, we applied the Reweighing algorithm from the AIF360 toolkit [1]. This pre-processing method adjusts the instance weights to reduce bias in the training data, without altering the features or labels. We defined `foreign_worker` as the protected attribute, with "yes" as the privileged group.

We trained a Random Forest classifier on the reweighted dataset using the same architecture as before. The model achieved a higher overall accuracy of **0.839**, while significantly reducing the demographic disparity observed prior to mitigation.

- **Accuracy after reweighing:** 0.839
- **DPD after reweighing:** 0.0705

Compared to the baseline model (DPD = 0.2517), this result demonstrates that reweighing is effective in reducing bias while preserving or even improving predictive performance.

7) Result Comparison: Table I summarizes the performance and fairness trade-offs between the baseline model and the mitigated model using reweighing. While the reweighted model improved overall accuracy, it also significantly reduced demographic disparities across all evaluated attributes.

TABLE I
COMPARISON OF ACCURACY AND FAIRNESS BEFORE AND AFTER REWEIGHING

Model	Accuracy	DPD (FW / PS / Age)
Baseline RF	0.753	0.2517 / 0.1512 / 0.0580
Reweighed RF	0.839	0.0705 / 0.0814 / 0.0453

These results demonstrate that fairness-aware preprocessing can reduce bias across multiple dimensions while maintaining or even improving predictive performance.

VI. State of the Art

Recent advances in the literature on algorithmic fairness have refined and expanded the set of available strategies for bias mitigation in machine learning. Mitigation methods are typically classified into three categories: pre-processing, in-processing, and post-processing. Pre-processing methods focus on adjusting the dataset prior to training, aiming to neutralize unwanted correlations between sensitive attributes and the target variable. One prominent example is counterfactual data augmentation, where synthetic instances are generated with swapped sensitive attributes to reduce statistical dependence between protected variables and model outputs [6]. In-processing techniques incorporate fairness constraints directly into the learning process. A notable approach is representation neutralization, which modifies the internal representations learned by the model to obscure sensitive attributes, enhancing fairness without significant losses in predictive performance [7]. Post-processing methods operate after training, altering the final predictions to meet fairness criteria. These techniques are especially useful when access to training data is restricted or when model retraining is not feasible. Recent proposals include calibrated adjustment algorithms that re-score outputs to ensure fairness metrics such as demographic parity and equal opportunity are satisfied [8].

In the financial domain, particularly in credit scoring, the deployment of fairness-aware machine learning models is both a technical and regulatory challenge. Disparities in credit approval rates across demographic groups—such as age, gender, or nationality—may stem from historical bias encoded in training data. These disparities, when left unmitigated, not only compromise fairness but may also violate fair lending regulations [9]. Consequently, recent studies have evaluated fairness mitigation methods specifically in credit contexts, balancing predictive performance with ethical and legal considerations. For example, benchmark experiments with reweighing and reject option classification have demonstrated consistent reductions in demographic disparities without degrading model accuracy [10]. Furthermore, the literature has acknowledged the role of proxy variables, which are non-sensitive features that correlate with protected attributes and may inadvertently reintroduce bias even when direct sensitive attributes are excluded. As the use of AI models in financial decision-making expands, ensuring transparency and fairness through systematic auditing has become a central concern, especially for institutions operating under strict regulatory frameworks.

VII. Expected Contribution

This work provides a reproducible framework for fairness auditing in credit scoring models, using public data and open-source tools. It bridges fairness theory and applied machine learning by demonstrating concrete steps for bias identification and mitigation. Future work may explore alternative mitigation techniques or extend this approach to more complex financial contexts.

Acknowledgment

We acknowledge the authors and maintainers of AIF360 and Fairlearn for their tools and documentation.

Code Availability

The complete code and reproducible workflow used in this tutorial are available at:
https://github.com/JuliaWPereira/fairness.

References

[1] R. K. Bellamy *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *IBM Journal of Research and Development*, 2019.

[2] S. Bird *et al.*, "Fairlearn: A toolkit for assessing and improving fairness in ai systems," https://fairlearn.org, 2020.

[3] Unknown, "South German Credit," 2019. [Online]. Available: https://archive.ics.uci.edu/dataset/522

[4] H. Hofmann, "Statlog (German Credit Data)," 1994. [Online]. Available: https://archive.ics.uci.edu/dataset/144

[5] D. Pessach and E. Shmueli, "Algorithmic fairness: A comparative survey," *arXiv preprint arXiv:2001.09784*, 2020. [Online]. Available: https://arxiv.org/abs/2001.09784

[6] ——, "Algorithmic fairness: A comparative survey," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–44, 2023.

[7] S. Caton and C. Haas, "Fairness in machine learning: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–37, 2020.
[8] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.
[9] A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther, "Predictably unequal? the effects of machine learning on credit markets," *The Journal of Finance*, vol. 77, no. 1, pp. 5–47, 2022.
[10] R. K. Bellamy, K. R. Dey, M. Hind, S. Hoffman, S. Houde, S. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4–1, 2019.