European Journal of Operational Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

Fairness in credit scoring: Assessment, implementation and profit implications

Nikita Kozodoi[a,*], Johannes Jacob[a], Stefan Lessmann[a]

School of Business and Economics, Humboldt University of Berlin, Unter den Linden 6, Berlin 10099, Germany

ARTICLE INFO

Article history:
Received 3 December 2020
Accepted 11 June 2021
Available online xxx

Keywords:
OR in banking
Machine learning
Algorithmic fairness
Credit scoring

ABSTRACT

The rise of algorithmic decision-making has spawned much research on fair machine learning (ML). Financial institutions use ML for building risk scorecards that support a range of credit-related decisions. Yet, the literature on fair ML in credit scoring is scarce. The paper makes three contributions. First, we revisit statistical fairness criteria and examine their adequacy for credit scoring. Second, we catalog algorithmic options for incorporating fairness goals in the ML model development pipeline. Last, we empirically compare different fairness processors in a profit-oriented credit scoring context using real-world data. The empirical results substantiate the evaluation of fairness measures, identify suitable options to implement fair credit scoring, and clarify the profit-fairness trade-off in lending decisions. We find that multiple fairness criteria can be approximately satisfied at once and recommend separation as a proper criterion for measuring the fairness of a scorecard. We also find fair in-processors to deliver a good balance between profit and fairness and show that algorithmic discrimination can be reduced to a reasonable level at a relatively low cost. The codes corresponding to the paper are available on GitHub.

1. Introduction

Financial institutions increasingly rely on machine learning (ML) to support decision-making (Crook, Edelman, & Thomas, 2007). The paper considers ML applications in the retail credit market, which is a large and economically important segment of the credit industry. For example, the total outstanding amount of retail credit in the US exceeded $4,161 billion in 2020[1]. ML-based scoring models, also called scorecards, have played a major role in the approval of the corresponding loans.

In 2016, the Executive Office of the President of the US published a report on algorithmic systems, opportunity, and civil rights (Executive Office of the President, 2016), which highlights the dangers of automated decision-making to the detriment of historically disadvantaged groups. It emphasizes credit scoring as a critical sector with a large societal impact, calling practitioners for using the principle of "equal opportunity by design" across different demographic groups. Similar actions were taken by the EU when they supplemented their General Data Protection Regulation with a guideline that stresses the need for regular and systemic monitoring of the credit scoring sector (European Commission, 2017). The guidelines issued by the EU and the US evidence political concern that potential violations of anti-discrimination law in credit scoring might affect debt and wealth distributions and have undesired economic effects on the society (Liu, Dean, Rolf, Simchowitz, & Hardt, 2018).

A growing literature on fair ML echos these concerns and proposes a range of statistical fairness measures and approaches for their optimization. It is common practice to discuss algorithmic fairness through the lens of differences between groups of individuals. The groups emerge from one or multiple categorical attributes that are considered sensitive. Examples include gender, religious denomination or ethnic group. The goal of fair ML is then to ensure that model predictions meet statistical fairness criteria. Narayanan (2018) distinguishes 21 such criteria while Barocas, Hardt, & Narayanan (2019) show that most criteria can be derived from one of three main fairness measures: independence, separation, and sufficiency. Beyond quantifying fairness in model-based predictions, fairness criteria also serve as constraints or objectives in the optimization problem that underlines the training of an ML model. Approaches to adjust model training to optimize fairness criteria next to common indicators of model fit are known as fairness processors.

Surprisingly, the literature on fair ML and credit scoring share few touching points. As we detail in Section 3.1, only three studies (Fuster, Goldsmith-Pinkham, Ramadorai, & Walther, 2017; Hardt, Price, & Srebro, 2016; Liu et al., 2018) have considered the interface between the two disciplines. None of them focuses on operational

* Corresponding author.
E-mail address: nikita.kozodoi@hu-berlin.de (N. Kozodoi).
[1] Source: https://www.federalreserve.gov/releases/g19/current

https://doi.org/10.1016/j.ejor.2021.06.023
0377-2217/

decisions in the loan approval process and the potential trade-off between fairness and profit. Therefore, the goal of the paper is to i) provide a broad overview and systematization of recently developed fairness criteria and fairness processors, and to ii) empirically test their adequacy for credit scoring. While the fairness enhancing procedures that we consider are not new and have been developed in the fair ML literature, we suggest that our holistic and integrative perspective is useful to help risk analysts stay abreast of recent developments in that literature, judge their impact on credit scoring practices, and focus future research initiatives concerning fair credit scoring.

In pursuing its objective, the paper makes the following contributions: First, we revisit statistical criteria for measuring fairness and examine whether these criteria and their underlying understanding of distributional equality are appropriate for credit scoring. Given that different fairness criteria typically conflict with one another (Chouldechova, 2017), our analysis is useful to inform the selection of a suitable fairness criterion (or set of criteria). Considering the relative costs of classification errors for banks and retail clients, we identify separation as a preferable criterion to appraise fairness in a lending context. More generally, our analysis may raise awareness for the risk of algorithmic discrimination in credit scoring, which, given the sparsity of prior work on the topic, may be seen as a valuable contribution to the credit risk community.

Second, we review and catalog state-of-the-art fairness processors across multiple important dimensions, including the target fairness criterion, the implementation method, and requirements for the classification problem. The catalog provides a systematic overview of fairness processors and clarifies whether and when these meet requirements associated with loan approval processes and the application context of credit scoring. The catalog also addresses the critique of Mitchell, Potash, Barocas, DAmour, & Lum (2021), who demand a more uniform fairness terminology among scholars.

Last, we empirically compare a range of different fairness processors along several performance criteria using seven real-world credit scoring data sets. Unlike prior studies on fair ML, our analysis recognizes prediction performance indicators that are established in credit scoring and, importantly, the profitability of a scoring model. Furthermore, to extend the conceptual discussion on the suitability of the fairness criteria for credit scoring, we measure fairness not only with the criterion optimized by a processor but a range of different fairness criteria. The corresponding results provide original insights concerning the agreement among fairness criteria in credit scoring and their compatibility with profit. More specifically, our comparative analysis contributes to the empirical credit scoring literature by identifying fairness processors that best serve the interests and requirements of risk analysts and by elucidating the trade-off between profitability and fairness of a credit scoring system. A deeper understanding of this trade-off is crucial for managers and policy-makers to decide on the deployment of fairness enhancing procedures in financial institutions and regulatory directives to enforce certain levels of fairness, respectively.

2. Theoretical background

This section covers relevant background on fair ML. We first examine methods to integrate fairness constraints into the model development pipeline and then review established fairness criteria. We focus on independence, separation and sufficiency because these criteria encompass a variety of other fairness concepts (Barocas et al., 2019; Mitchell et al., 2021). Table A1 in the online Appendix details how independence, separation and sufficiency have synonymously been referred to in the literature and how they relate to the other formulations of fairness.

2.1. Fairness optimization in the modeling pipeline

Research on fair ML has recently emerged from the continuous integration of automated decision-making into important areas of social life and fairness concerns arising during this process (Barocas & Selbst, 2016). Much fair ML literature focuses on classification settings in which an unprivileged demographic group experiences discrimination through a classification model (Mitchell et al., 2021). Several attempts have been made to formalize the concept of fairness. Incorporating the corresponding fairness criteria in the ML pipeline facilitates measuring the degree to which class predictions discriminate against minorities (Barocas et al., 2019).

Algorithmic interventions designed to implement statistical fairness constraints are denoted as fairness processors. A processor can alter different stages in the ML pipeline. The literature distinguishes three methods of intervention: pre-processing, in-processing and post-processing (Berk, Heidari, Jabbari, Kearns, & Roth, 2021). Their application generally depends on the conceptual and technical feasibility of a given prediction task. Fig. 1 illustrates the fairness processors within an ML pipeline. We describe selected approaches from each group in Section 4.

Integrating a fairness processor into the pre-processing stage transforms the training data such that the input to a model is fair with respect to one or more sensitive features. Typically, fair pre-processing involves decorrelating the feature space with the sensitive attribute (e.g., Calmon, Wei, Vinzamuri, Natesan Ramamurthy, & Varshney, 2017). Even though modifying the training data is sometimes not possible or practical, the advantage of fair pre-processing is that if fairness is ensured before ML model training, it will also be ensured during the next model development steps (Barocas et al., 2019).

In-processing methods introduce auxiliary fairness constraints during ML model training. Then, training involves minimizing the empirical risk of the model while also optimizing a fairness criterion. In-processing renders a learned classifier (approximately) fair for the training data (Zafar, Valera, Gomez Rodriguez, & Gummadi, 2017a). Optimizing fairness during training has the potential to generate the highest utility as the tuning process also considers the fairness constraint. At the same time, in-processors are typically developed for settings with specific requirements (e.g., supporting only a single sensitive attribute), which limits their generality (Barocas et al., 2019). Another disadvantage is that implementing a fair in-processor requires full access to the training process and the input data. This is especially problematic in heavily regulated domains such as credit scoring, where changes to a risk model might require regulatory approval and are associated with high costs.

After an ML model is trained, post-processing can be applied to adjust the learned classifier or change its predictions according to the requirements of a particular fairness criterion (Hardt et al., 2016). The standard procedures include modifying the predicted scores or labels for specific observations. Unlike pre- or in-processing, post-processors need no information about the input data or the base model. This has the advantage that post-processors can be applied to any set of predictions. However, generality has a price. Post-processing is often less effective than alternative approaches and may substantially decrease classification accuracy (Barocas et al., 2019).

2.2. Fairness criteria

This subsection introduces three established fairness criteria from a credit scoring perspective. Consider a setting in which a financial institution uses data on previous customers to predict whether a loan applicant will default. Let $X \in \mathbb{R}^k$ denote the k fea-

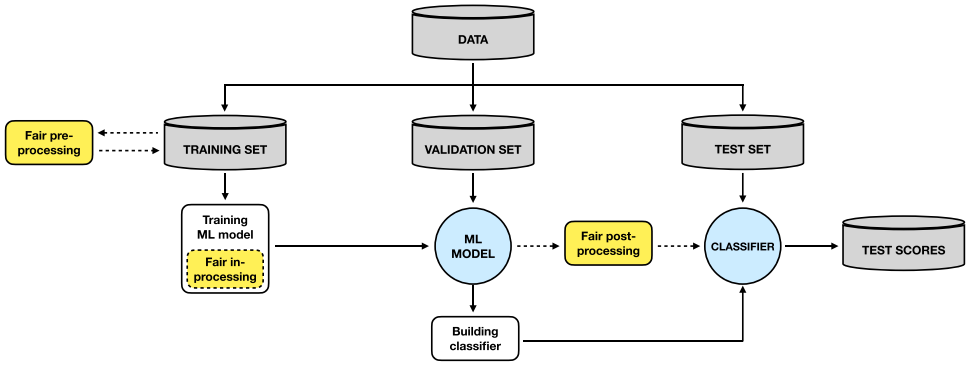

Fig. 1. Fairness Integration in the ML Pipeline: In-processing, Pre-processing and Post-processing.

tures of a loan applicant and $y \in \{0, 1\}$ a random variable indicating if the applicant repays the loan ($y = 1$) or defaults ($y = 0$). The institution approves applications using a scoring model that predicts risk scores $s(X) = \mathbf{P}(y = 1|X)$. The score function can be turned into a classifier by accepting customers with scores above a cutoff τ. Let $x_a \in \{0, 1\}$ denote a protected attribute associated with certain characteristics of an applicant. For example, x_a could indicate whether she has a disability ($x_a = 1$) or not ($x_a = 0$). Clearly, the value of x_a must not impact the decision of the credit institution.

In the following, we consider a binary protected attribute to simplify the exposition. The discussed fairness criteria generalize to multinomial protected attributes (i.e., protected attributes with more than two unique values). Also, note that the fair ML literature often uses the terms protected attribute and sensitive attribute interchangeably. From a methodological perspective, it is less important whether the use of an attribute is socially undesirable or regulated by law. We use the term sensitive attribute throughout the paper while acknowledging that our example attribute disability is not only sensitive but protected. The groups created when splitting individuals by a sensitive attribute are referred to as sensitive groups.

2.2.1. Independence

The score $s(X)$ satisfies independence at a cutoff τ if the fraction of customers classified as good risks ($y = 1$) is the same in each sensitive group. Formally, this condition can be written as:

$$\mathbf{P}[s(X \,|\, x_a = 0) > \tau] = \mathbf{P}[s(X \,|\, x_a = 1) > \tau] \tag{1}$$

Eq. (1) states that $s(X)$ is statistically independent of the sensitive attribute x_a (Barocas et al., 2019). Classifier predictions are not affected by the sensitive attribute, and the probability to be classified as a good risk is the same in both groups (Pleiss, Raghavan, Wu, Kleinberg, & Weinberger, 2017). In the prior work, the independence condition is also known as demographic or statistical parity (Chouldechova, 2017).

This strict constraint is usually not feasible for real-world applications like credit scoring, as the resulting loss in model performance can make a business unsustainable. Therefore, it is a common practice in anti-discrimination law to allow the score and the sensitive attribute to share at least some mutual information and introduce a relaxation of the independence criterion (Barocas & Selbst, 2016). The Equal Opportunity Credit Act has a regulation that is referred to as the "80 percent rule" (Feldman, Friedler, Moeller, Scheidegger, & Venkatasubramanian, 2015). The rule requires that $\mathbf{P}(s(X \,|\, x_a = 1) > \tau) \leq 0.8 \cdot \mathbf{P}(s(X \,|\, x_a = 0) > \tau)$, where $\{x_a = 0\}$ is the privileged group (Kleinberg, Mullainathan, & Raghavan, 2017).

Following the relaxation of the independence condition suggested in the prior work (Barocas et al., 2019), we measure independence using a metric denoted as IND, which we define as:

$$\text{IND} = |\mathbf{P}[s(X \,|\, x_a = 0) > \tau] - \mathbf{P}[s(X \,|\, x_a = 1) > \tau]| \tag{2}$$

A positive difference between the two terms implies that the group $\{x_a = 0\}$ is considered the privileged group and vice versa. The closer IND is to zero, the lower is the discrimination.

2.2.2. Separation

The separation condition, also known as the equalized odds condition, is satisfied if the classification based on the predicted score $s(X)$ and the cutoff τ is independent on x_a conditional on the true outcome y (Barocas et al., 2019). Formally, the score $s(X)$ satisfies separation at a cutoff τ if:

$$\begin{cases} \mathbf{P}[s(X \,|\, y = 0, x_a = 0) > \tau] & = \mathbf{P}[s(X \,|\, y = 0, x_a = 1) > \tau] \\ \mathbf{P}[s(X \,|\, y = 1, x_a = 0) \leq \tau] & = \mathbf{P}[s(X \,|\, y = 1, x_a = 1) \leq \tau] \end{cases} \tag{3}$$

The expression in the first line compares the false positive rate (FPR) across the sensitive groups, whereas the second line compares the false negative rate (FNR) per group. The separation criterion, therefore, requires that the FNR and the FPR are the same for the sensitive groups.

Separation acknowledges that x_a may be correlated with y (e.g., applicants with a disability might have a higher default rate). However, the criterion prohibits the use of x_a as a direct predictor for y. When the difference between group sizes is large, the criterion will punish models that perform well only on the majority group (Hardt et al., 2016). To measure the degree to which the separation condition is satisfied, we suggest using a criterion denoted as SP, which we define as:

$$\text{SP} = \frac{1}{2}\left|(\text{FPR}_{\{x_a=1\}} - \text{FPR}_{\{x_a=0\}}) + (\text{FNR}_{\{x_a=1\}} - \text{FNR}_{\{x_a=0\}})\right| \tag{4}$$

SP calculates the average absolute difference between the group-wise FPR and FNR. A positive difference between each of the two group-wise error rates indicates that the $\{x_a = 0\}$ group has a lower misclassification rate and is, therefore, the privileged group. Perfect separation (i.e., $\text{SP} = 0$) is observed when the group-wise FPR and FNR are equal. Higher values of SP indicate stronger discrimination through a larger difference in model performance across the sensitive groups.

2.2.3. Sufficiency

The score $s(X)$ is sufficient at a cutoff τ if the likelihood that an individual belonging to a positive class is classified as positive is the same for both sensitive groups (Barocas et al., 2019). This implies that for all values of $s(X)$ the following condition holds:

$$\mathbf{P}(y = 1 \,|\, s(X) > \tau, x_a = 0) = \mathbf{P}(y = 1 \,|\, s(X) > \tau, x_a = 1) \tag{5}$$

Eq. (5) requires that the positive predictive value (PPV) is the same for the sensitive groups (Chouldechova, 2017). This paper defines

the sufficiency metric SF as the absolute difference between the group-wise PPV:

$$\text{SF} = \left|\text{PPV}_{\{x_a=0\}} - \text{PPV}_{\{x_a=1\}}\right| \tag{6}$$

A large difference between the group-wise PPV indicates inconsistent model performance across the sensitive groups. The closer SF is to zero, the higher is the achieved sufficiency.

3. Fairness and credit scoring

The section discusses the interplay between fair ML and credit scoring. We summarize previous work in the field and examine the adequacy of fairness criteria for credit scoring.

3.1. Prior work on fair credit scoring

Prior literature on fair ML for credit scoring is surprisingly sparse. To our best knowledge, only three studies address algorithmic discrimination in credit scoring, and their focus differs substantially from that of this study. A first study by Fuster et al. (2017) considers the credit market. The authors formalize the introduction of ML as a market intervention and examine the corresponding effect on interest rates in demographically different groups. Liu et al. (2018) take a similar perspective. Referring to the sample-selection bias, which arises from training scorecards on previously accepted cases (Banasik & Crook, 2007), they argue that selection bias leads to scorecards overestimating the creditworthiness of some groups of applicants and perpetuates existing unfairness. To remedy this effect, Liu et al. (2018) call for mathematical constraints that optimize fairness as a long-term societal goal. However, the formulation of these constraints is still subject to further research. More generally, the long-term perspective of Fuster et al. (2017) and Liu et al. (2018) emphasizes regulatory questions and is orthogonal to the focus on static fairness interventions, which prevails in the fair ML literature. These interventions address operational loan approval decisions and provide concrete approaches to remedy algorithmic bias.

Focusing on fairness interventions, a third study of Hardt et al. (2016) is related to this paper more closely. Hardt et al. (2016) propose the equalized odds fairness criterion and develop an algorithm that adjusts classifier predictions to raise fairness according to this criterion. The authors report enhanced fairness compared to a maximum profit benchmark using a credit scoring example based on FICO scores. In comparison to the focal paper, Hardt et al. (2016) focus on the specific combination of one fairness processor and one fairness criterion. Their study does not examine the trade-off between profit and fairness and provides limited empirical evidence on how equalized odds compare to other fairness criteria or how fairness is best ensured in an ML pipeline.

In summary, the main distinction between the focal paper and previous studies on fairness in credit scoring is that we undertake a comprehensive empirical analysis of alternative fairness criteria and fairness processors, which optimize these criteria. Prior work fails to account for the breadth of approaches that have been proposed in the scope of fair ML. Also, no previous study examines the interplay between fairness criteria and processors. Therefore, we aim at consolidating different advancements in fair ML, discussing their suitability for credit scoring, and providing rich empirical results that clarify the degree to which fairness constraints affect the predictive ability of credit scorecards and the corresponding profit implications, and how the trade-off between fairness and profit develops across fairness criteria and processors. We hope that our results offer actionable insights on how to set and pursue fairness objectives in credit scoring.

3.2. Fairness criteria for credit scoring

The choice of the fairness criterion has severe consequences for the social impact of lending decisions (Liu et al., 2018). An unconstrained scoring model will take full advantage of the available (sensitive) information and discriminate between protected groups if this enhances predictive performance. The purpose of introducing fairness is, therefore, to adjust decision-making (i.e., scoring) practices for a better, discrimination-free outcome. According to the U.S. anti-discrimination law, for example, the demographic properties of a loan applicant should not influence lending decisions (Equal Credit Opportunity Act, 1974). Arguably, the societal goal behind such law is an equal opportunity for financial well-being across demographically different groups. Achieving this goal in credit scoring is difficult as clients face unequal misclassification costs. Applicants that are denied a loan they could have repaid face the cost of a missed opportunity to enhance their social and economic position. However, if applicants receive a loan they cannot repay, they are confronted with financial debt and a long-term worsening of their financial situation as future access to financing will be more difficult. With these characteristics of credit scoring in mind, the following considerations elaborate on the extent to which independence, separation and sufficiency fulfill the goal of equal opportunity for financial well-being in society.

Forcing independence on a scoring model's results in the same rate of accepted customers within sensitive groups. The problem with this approach is that the ability to repay a loan can have a different distribution in each group (Barocas et al., 2019). If this is the case, but members of both groups have the same probability of receiving a loan, one group will experience more actual defaults. For a client, the consequences of defaulting can be more severe than the opportunity costs associated with a rejected application. Typically, the historically unprivileged group has a higher rate of non-solvent customers. Handing out loans to such individuals might worsen their financial situation in the long term (Hardt et al., 2016). Instead of achieving fairness, this can lead to further perpetuating existing unfairness. The goal of better financial equality would not be met, and the financial gap in society could become even wider.

The separation criterion addresses this dilemma and acknowledges that a sensitive attribute might correlate with default rates. Requiring the same error rates between groups but allowing different positive classification rates, separation achieves a fair result that is closer to the reality of credit allocation decisions and more desirable from a customer's perspective. More precisely, separation accounts for different misclassification costs between groups. On the contrary, separation would be inadequate if credit scoring had a strictly preferred outcome for a customer, as is the case in domains like college admission (Mitchell et al., 2021). Interestingly, the first formulation of the separation criterion in the context of ML by (Hardt et al., 2016) is based on the example of the credit scoring domain and the limitations of the independence criterion to meet its requirements.

Sufficiency requires the ratio of true positive classifications over all positive classifications to be the same for the sensitive groups. This concept has two disadvantages for credit scoring. First, it allows for substantial discrimination in separation. For both groups, the proportion of correctly labeled non-default clients can be the same, satisfying sufficiency. In contrast, the likelihood of a potential non-default customer being classified as a bad risk can still differ between groups, violating the separation constraint. Second, most ML algorithms are designed to achieve sufficiency without integrating a fairness constraint if the model can predict the sensitive attribute from the other features (Barocas et al., 2019). In credit scoring, the question would, therefore, be if the current procedure for assessing a customer's default risk and the associated

Table 1
Fairness Processors.

Fairness processor	Reference	Method	Criterion	MT	MS	MA	PE	This paper
Reweighing	Calders et al. (2009)	PRE	IND					✓
Massaging	Calders et al. (2009)	PRE	IND					
Classification without discrimination	Kamiran & Calders (2009)	PRE	IND					
Discrimination discovery K-NN	Luong, Ruggieri, & Turini (2011)	PRE	IND	✓				
Fair representation learning	Zemel, Wu, Swersky, Pitassi, & Dwork (2013)	PRE	IND	✓				
Disparate impact remover	Feldman et al. (2015)	PRE	IND		✓	✓		✓
Variational fair autoencoder	Louizos, Swersky, Li, Welling, & Zemel (2016)	PRE	IND	✓	✓	✓		
Feature adjustment	Johndrow, Lum et al. (2019)	PRE	IND	✓	✓	✓		
Discrimination-free pre-processing	Calmon et al. (2017)	PRE	IND		✓	✓		
Prejudice remover regularizer	Kamishima et al. (2012)	IN	IND	✓				✓
Fair accuracy maximizer	Zafar, Valera, Gomez Rodriguez, & Gummadi (2017b)	IN	IND	✓	✓	✓		
Non-discriminatory Learner	Woodworth, Gunasekar, Ohannessian, & Srebro (2017)	IN	SP					
Adversarial debiasing	Zhang et al. (2018)	IN	SP	✓	✓	✓		✓
Meta-fairness algorithm	Celis et al. (2019)	IN	IND, SP, SF		✓	✓		✓
Group-wise Platt scaling	Platt (1999), Barocas et al. (2019)	POST	SF	✓	✓	✓		✓
Group-wise histogram binning	Zadrozny & Elkan (2001)	POST	SF	✓	✓	✓		
Group-wise isotonic regression	Niculescu-Mizil & Caruana (2005)	POST	SF	✓	✓	✓		
Fairness-aware classifier	Calders & Verwer (2010)	POST	IND					
Reject option classification	Kamiran et al. (2012)	POST	IND, SP		✓	✓		✓
Fairness constraint optimizer	Goh, Cotter, Gupta, & Friedlander (2016)	POST	IND	✓	✓	✓		
Equalized odds processor	Hardt et al. (2016)	POST	SP		✓		✓	✓
Calibrated equalized odds	Pleiss et al. (2017)	POST	SP					

Abbreviations: IND = Independence, SP = separation, SF = sufficiency; PRE = pre-processor, IN = in-processor, POST = post-processor;
MT = multinomial target, MS = multinomial sensitive attribute, MA = multiple sensitive attributes, PE = profit-driven evaluation.

distribution of loans is fair. The literature suggests a negative answer to this question (Fuster et al., 2017; Hardt et al., 2016; Liu et al., 2018). Hence, sufficiency appears less suitable for credit scoring.

Based on these considerations, the separation criterion appears most suitable to achieve a desirable form of fairness in credit scoring. Separation accounts for the imbalanced misclassification costs of the customer, and, as these imbalanced costs also exist for the financial institution, separation is also able to consider the interests of the loan market.

The considerations provided in this section suggest that the question of which fairness constraint is most adequate for credit scoring should be a part of a wider academic and societal debate. Such a democratic process should also acknowledge the importance of studying the long-term effects of implementing different fairness constraints to judge whether the societal goal of better financial equality between demographic groups can be achieved with specific interventions (Liu et al., 2018).

4. Methodology

This section systematically reviews and catalogs fairness processors suggested in the prior work across different dimensions and discusses their applicability in credit scoring. Using the constructed catalog, we select and describe eight fairness processors that are part of the empirical study.

4.1. Cataloging fairness processors

The fair ML literature has developed a variety of fairness processors to implement independence, separation and sufficiency constraints. The complexity between these processors varies considerably, from simply relabeling the prediction outcomes (e.g., Kamiran, Karim, & Zhang, 2012) to complex deep learning approaches for training a discrimination-free classifier (e.g., Zhang, Lemoine, & Mitchell, 2018). Furthermore, some processors are limited to specific problem setups. This motivates us to develop a structured overview of fairness processors with respect to their characteristics and applicability. Specifically, we catalog existing fairness processors in Table 1 using six dimensions: (i) point of intervention into the ML pipeline; (ii) optimized fairness criterion; (iii) classification problem type supported by a processor (binary or multinomial); (iv) possible number of sensitive attributes (one or multiple); and (vi) supported types of sensitive attributes (binary or multinomial).

Three main conclusions emerge from Table 1. First, the majority of processors implement the independence criterion. This may come from the other criteria being invented only recently (see Table A1 in the online Appendix for comparison). Furthermore, independence allows implementation via pre-processing, which provides an additional point of intervention in the ML pipeline. In many scenarios, however, fairness through independence may not be a suitable choice. This calls for additional processors that implement the other two criteria.

Second, the choice of a suitable fairness processor is limited by the application and implementation context of a scorecard. The application context determines the type of target variable and sensitive attribute(s) to be handled by a processor. For instance, in a setup with multiple sensitive attributes optimizing separation is only possible via the adversarial debiasing or reject option classification. This is a severe limitation for credit scoring because financial institutions commonly face several protected attributes: the U.S. anti-discrimination law distinguishes nine bases that must not influence lending decisions, including race, color, religion and other customer attributes (Equal Credit Opportunity Act, 1974). The implementation context can also limit possible points of intervention in the ML pipeline. Replacing a scorecard with a fair in-processor might require regulatory approval and incur additional costs. Post-processors are easier to implement since they are agnostic of the input data and the scorecard and only require access to the predicted scores.

Third, it is a standard procedure to embed the fairness processor into an accuracy-optimizing framework. The loss in predictive accuracy is commonly used as a performance measure to judge the cost of integrating a fairness constraint. In line with this framework, Friedler et al. (2019) conducted a comparative study to examine the achieved fairness and accuracy of four fairness processors. However, recent credit scoring literature criticizes the practice of using standard performance measures for evaluating scor-

ing models and calls for profit-driven evaluation (Verbraken, Bravo, Weber, & Baesens, 2014). In such a setup, evaluation of fairness processors should be performed with a profit maximization objective instead of standard statistical performance measures such as accuracy.

To conclude, the catalog suggests that a comparative analysis of fairness processors under profit maximization is needed to clarify the "cost of fairness". We argue that the profitability aspect is underrepresented in the fair ML literature, while it is highly relevant for real-world applications. A better understanding of the (dis)agreement of profitability and different fairness criteria is also useful for policy making as it sheds some light on the thorny question of which criterion lending institutions should emphasize. Which fairness processor to use for optimizing the desired criterion is yet another question with high relevance for practice. Prior literature offers limited guidance due to assessing processors typically only in terms of the single criterion that this processor implements. Contributing toward answering these pressing questions is the overall goal of the paper.

4.2. Selected fairness processors

This subsection overviews eight fairness processors from the catalog presented in Table 1. The selection of processors covers all combinations of fairness interventions. Following the setup introduced in Section 2, we consider a credit scoring setup with a binary target variable $y \in \{0, 1\}$ and a binary sensitive attribute $x_a \in \{0, 1\}$ to introduce the processors. Some of the considered processors also generalize to multinomial target and sensitive attributes (see Table 1 for details).

4.2.1. Pre-Processors

Fairness pre-processors transform the input data to achieve fairness. Reweighing is a pre-processor that assigns weights to each observation in the training set based on the overall probabilities of the group-class combinations (Calders, Kamiran, & Pechenizkiy, 2009). Thus, weights for observations with $(x_a = 1, y = 1)$ are greater than weights for observations with $(x_a = 0, y = 1)$ if members of the group $\{x_a = 1\}$ have a lower probability to belong to a positive class than those of the group $\{x_a = 0\}$:

$$W(X \mid x_a = 1,\ y = 1) = \frac{\mathbf{P}_{exp}(x_a = 1 \mid y = 1)}{\mathbf{P}_{obs}(x_a = 1 \mid y = 1)}, \tag{7}$$

where $\mathbf{P}_{\text{exp}}$ is the expected probability and $\mathbf{P}_{obs}$ is the observed probability. For instance, assume that 90% of all individuals belong to the positive class and 20% percent belong to the group $\{x_a = 1\}$. Then, $\mathbf{P}_{\text{exp}}(x_a = 1 \mid y = 1) = 0.9 \cdot 0.2 = 0.18$. If, in fact, only 12% of all cases in $\{x_a = 1\}$ belong to the positive class, then $W(X \mid x_a = 1, y = 1) = \frac{0.18}{0.12} = 0.9$.

Based on the computed weights, a fair training set is resampled with replacement such that combinations with a higher weight reappear more often. This procedure helps to fulfill the independence criterion. A discrimination-free classifier can then be trained on the resampled data.

Another pre-processing technique is the disparate impact remover proposed by Feldman et al. (2015). The intuition behind this processor is to ensure independence by prohibiting the possibility of predicting the sensitive attribute x_a with the other features in X and the outcome y. This is achieved by transforming X into $\overline{X}$ while preserving the rank of X within sensitive groups defined by x_a. By preserving the rank of X given x_a, the classification model $f(\overline{X})$ will still learn to choose higher-ranked credit applications over lower-ranked ones based on the other features.

The transformation is performed using an interpolation based on a quantile function and the cumulative distribution of F: $\mathbf{P}(X \mid x_a = a)$. This ensures that given the transformed $\overline{X}$ at some rank, the probability of drawing an observation given $x_a = a$ is the same as for the entire data set. Hence, x_a cannot be predicted with the other attributes, and the independence criterion is fulfilled. Since ensuring perfect independence can have a strong negative impact on a classifier utility, the transformation can be modified to only partially remove disparate impact. The meta-parameter $\lambda \in [0, 1]$ allows controlling the desired level of fairness-utility trade-off during transformation.

4.2.2. In-Processors

In-processors achieve fairness when building a classifier. One of such methods, prejudice remover, introduces a fairness-driven regularization term to the classification model (Kamishima, Akaho, Asoh, & Sakuma, 2012). Regularization is a standard statistical approach to penalize a model for some undesired behavior. This is typically done by adding a regularizer term to the loss function.

The fairness-driven regularization introduced by Kamishima et al. (2012) is based on the prejudice index PI, which quantifies the degree of unfairness based on the independence criterion:

$$\text{PI} = \sum_{(y, x_a) \in D} \mathbf{P}(y, x_a) \ln \frac{\mathbf{P}(y, x_a)}{\mathbf{P}(x_a)\mathbf{P}(y)}, \tag{8}$$

where $\mathbf{P}(y, x_a)$, $\mathbf{P}(y)$ and $\mathbf{P}(x_a)$ are empirical distributions of y and x_a over the sample D. PI measures the amount of mutual information between y and x_a. High values of PI indicate that a sensitive attribute x_a is a good predictor for y. The optimization problem extends to:

$$\min_f L[f(X), y] + \eta \text{PI}, \tag{9}$$

where $L(\cdot)$ is the underlying loss function of the model $f(X)$, and η controls the importance of the term PI. In this study, we tune η to maximize the profitability of a scorecard. The regularization term ensures that the sensitive attribute x_a becomes less influential in the final prediction.

Adversarial debiasing is another in-processor that stacks two neural networks with contrary objectives on top of each other (Zhang et al., 2018). The first network (predictor) is trying to learn a function to predict y given X, while also minimizing the success of the second network. The second network (adversary) takes the output layer of the first model $\hat{y}$ and the true labels y as input and tries to predict the sensitive attribute x_a. Both models have objective-specific loss functions and weights that can be optimized using standard gradient-based optimization methods such as stochastic gradient descent or Adam (Kingma & Ba, 2014).

The adversary is assumed to have weights U and loss function $L_A(\hat{x}_a, x_a)$. The weights U are updated according to the gradient $\nabla_U L_A$ to minimize L_A. The weights of the predictor denoted as W are modified based on a gradient that minimizes its loss function $L_P(\hat{y}, y)$ but also maximizes the loss function of the adversary: $\nabla_W L_P(\hat{y}, y) - \alpha \nabla_W L_A(\hat{x}_a, x_a)$, where α is a meta-parameter.

Since the adversary takes the output of the predictor $\hat{y}$ as input, the predictor aims to hold back any additional information about the sensitive attribute x_a in its output $\hat{y}$ as it would improve the adversary's loss. In other words, the predictor will try to deceive the adversary and not share any additional information in $\hat{y}$. As y is known to the adversary, the algorithm acknowledges that the sensitive attribute might correlate with y, and only unnecessary information will be avoided. Hence, the adversarially debiased model will converge towards the separation criterion.

The meta fair classification algorithm is yet another in-processor designed to achieve fairness according to one of the different fairness criteria. For a given criterion, Celis, Huang, Keswani, & Vishnoi (2019) suggest using a corresponding group-wise fairness metric denoted as FM, where similar values of FM across sensitive groups indicate a higher level of fairness. Given a classifier

$f(X)$ with a loss function $L(f(X), y)$, they add a fairness constraint to the loss optimization problem during training:

$$\min_f L(f(X), y) \quad \text{s.t.} \quad \frac{\min[\text{FM}(f(X \mid x_a = 0)),\ \text{FM}(f(X \mid x_a = 1))]}{\max[\text{FM}(f(X \mid x_a = 0)),\ \text{FM}(f(X \mid x_a = 1))]} \geq \sigma, \tag{10}$$

where $\sigma \in [0, 1]$ is a desired fairness bound. Higher values of the fraction in Eq. (10) indicate a higher similarity of FM across sensitive groups, and $\sigma = 1$ implies perfect fairness.

For example, in case of sufficiency, FM is set to positive predictive value (PPV) such that $\text{FM}(f) = \text{PPV}(f) = \frac{\mathbf{P}(f=1 \mid x_a=a, y=1)}{\mathbf{P}(f=1 \mid x_a=a)}$. If the group $\{x_a = 1\}$ has a low PPV and the group $\{x_a = 0\}$ has a high PPV, the fraction in the optimization condition is close to zero. A high σ will, therefore, bound the classifier to a high degree of fairness. During training, the value for σ can be tuned such that it maximizes profit while minimizing the loss in fairness, i.e., the loss in sufficiency.

4.3. Post-processors

As a post-processing method, reject option classification is based on the output of a learned classifier (Kamiran et al., 2012). In a credit scoring setup, the classifier output is a credit score that reflects the posterior probability to not default for each customer $s(X) = \mathbf{P}(\hat{y} = 1|X)$. The closer the score is to 1 or 0, the higher is the certainty with which the classifier assigns the corresponding labels, whereas a score close to 0.5 implies a high degree of uncertainty.

Reject option classification defines a critical region of high uncertainty and reassigns labels for customers that have predicted scores within this region, such that members of the unprivileged group receive a positive label ($y = 1$) and vice versa. Formally, the critical region is defined as:

$$\max\left[\mathbf{P}(\hat{y} = 1|X) \ \ 1 - \mathbf{P}(\hat{y} = 1|X)\right] \leq \theta, \tag{11}$$

where $0.5 < \theta < 1$. Given a set of predicted scores and the true outcomes, a suitable value of θ and the number of required posterior reclassifications can be tuned to optimize a fairness criterion (e.g., independence) within a specified interval restricted by the lower and the upper bound of the fairness metric denoted as $[\sigma_l, \sigma_u]$.

Equalized odds processor uses a different logic to post-process classifier predictions. It finds a cutoff value τ that optimizes the predictive performance while satisfying the separation criterion, i.e., ensuring the same false negative and false positive rate per group (Hardt et al., 2016).

Consider the receiver operating characteristic (ROC) curves that depict the trade-off between true and false positive rates for two sensitive groups. In an unfair scenario, the group-wise ROC curves have different slopes, which implies that not all trade-offs are achievable in each group. In the accuracy optimization setting, the optimal cutoff that satisfies sufficiency lies at the intersection of group-wise ROC curves. When optimizing for profit, the misclassification costs are not the same for both error rates. Thus, the optimal cutoff could lie somewhere else. Given a loss function $L(\cdot)$, Hardt et al. (2016) suggest to derive a suitable cutoff τ by optimizing the following objective:

$$\begin{aligned}&\min \mathbf{P}(s(X|x_a = a, y = 0) \leq \tau) \cdot L(\hat{y} = 1, y = 0) \\ &+ [1 - \mathbf{P}(s(X|x_a = a, y = 1) > \tau)] \cdot L(\hat{y} = 0, y = 1)\end{aligned} \tag{12}$$

Platt scaling is a post-processing method that stems from the notion of calibration (Platt, 1999). Calibration addresses the problem that some classification algorithms are not able to make a statement about the certainty of their prediction, i.e., the probability with which an instance belongs to a certain class. In credit scoring, the predicted score could be an indicator of default risk but not the actual probability of default. A score $s(X)$ is calibrated if $\mathbf{P}(y = 1 \mid s(X) = \tau) = \tau$.

When extending the calibration condition to the group level, it becomes apparent that it implements the sufficiency criterion (see Barocas et al. (2019) for proof):

$$\mathbf{P}[y = 1 \mid s(X) = \tau,\ x_a = 1] = \ \mathbf{P}[y = 1 \mid s(X) = \tau,\ x_a = 0] = \tau \tag{13}$$

To achieve calibration per group, Platt scaling is applied separately to each sensitive group. The method uses the output of a possibly uncalibrated score $s(X)$ as input for logistic regression fitted against the target variable y. Based on the loss function of the logistic regression, the result is a new calibrated score that represents the probability that an instance belongs to the positive class. Formally, Platt scaling minimizes the log-loss $-\mathbb{E}[y \log(\sigma) + (1 - y)\log(1 - \sigma)]$ by finding the optimal parameters a and b of the sigmoid function $\sigma = \frac{1}{1+\exp(aS+b)}$.

5. Experimental setup

5.1. Data

The empirical experiment is based on seven credit scoring data sets. Data sets *german* and *taiwan* stem from the UCI Machine Learning Repository[2]. *Pakdd*, *gmsc* and *homecredit* were provided by different companies for the data mining competitions on PAKDD[3] and Kaggle[4]. *Bene* and *uk* were collected from financial institutions in the Benelux and UK (Lessmann, Baesens, Seow, & Thomas, 2015).

Each data set has a unique set of features describing a loan applicant and loan characteristics. The target variable y is a binary indicator of whether the applicant has repaid the loan ($y = 1$) or not ($y = 0$). Each data set also contains a sensitive demographic attribute x_a indicating the applicant's age group. The Equal Credit Opportunity Act prohibits that demographic characteristics such as the applicants' age impact credit approval decisions. We distinguish two groups of applicants: $\{x_a = 1\}$ contains applications where the applicant's age is below ψ years, and $\{x_a = 0\}$ refers to the applications from customers older than ψ. We set $\psi = 25$, following the findings of Kamiran & Calders (2009), who used one of the consumer credit scoring data sets to discover that applicants from different age groups exhibit the greatest disparate impact (i.e., difference in $\mathbf{P}[y = 1 \mid x_a = a]$) at a threshold of 25 years. Table 2 summarizes the main characteristics of the data sets.

5.2. Experimental setup

On each data set, we implement the eight fairness processors introduced in Section 4, following the model development pipeline depicted in Fig. 1[5]. First, we partition the data into training (60%) and test (40%) sets. We then perform five-fold cross-validation on the training set. Each of the five combinations of training folds is used to train a scoring model and implement fairness processors. An unconstrained scoring model (i.e., a model that does not include any fairness-optimizing procedures) serves as a benchmark and represents the profit maximization scenario. Next, we consider

[2] Source: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data), https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

[3] Source: https://www.kdnuggets.com/2010/03/f-pakdd-2010-data-mining-competition.html.

[4] Source: https://kaggle.com/c/home-credit-default-risk, https://kaggle.com/c/givemesomecredit.

[5] The code reproducing the experiments is available at https://github.com/kozodoi/Fair_Credit_Scoring.

Table 2
Credit Scoring Data Sets.

Data set	Sample size	No. features	Default rate	Sensitive group rate
german	1000	61	0.30	0.19
bene	3123	82	0.33	0.12
taiwan	23,531	76	0.23	0.14
uk	30,000	51	0.04	0.20
pakdd	50,000	185	0.26	0.11
gmsc	150,000	68	0.07	0.02
homecredit	307,511	92	0.08	0.04

Table 3
Cost Matrix for Profit Computation.

	Predicted label	
Actual label	Bad risk	Good risk
Bad risk	$\pi_0 F_0(\tau)$ benefit: 0	$\pi_0(1-F_0(\tau))$ cost: B
Good risk	$\pi_1 F_1(\tau)$ cost: C	$\pi_1(1-F_1(\tau))$ benefit: C

in-processors in the form of the prejudice remover, adversarial debiasing and the meta fair algorithm. Relying on an in-processor implies that the trained in-processor serves as a scorecard. This contrasts pre- and post-processors, in which the actual scorecard is still based on a conventional ML algorithm. We consider reweighing and the disparate impact remover to pre-process (i.e., transform) the training data before developing a scoring model. Reject option classification, the equalized odds processor and Platt scaling represent the post-processors in our study. To learn a post-processing model, we apply each of them to the validation fold predictions of the unconstrained scorecard.

Fairness pre- and post-processors, as well as an unconstrained scorecard, use four base classifiers: logistic regression, artificial neural network and the tree-based ensemble learners random forest and extreme gradient boosting (XGB). Using multiple base learners allows us to check the robustness of processors across different classifiers. The base learners are established in credit scoring (e.g., Kozodoi, Lessmann, Papakonstantinou, Gatsoulis, & Baesens, 2019; Lessmann et al., 2015), whereby XGB (Chen & Guestrin, 2016) is maybe less known in the community. We include XGB due to its reputation as a highly powerful learning algorithm in Kaggle competitions and its strong performance in a recent credit scoring study by Gunnarsson, Vanden Broucke, Baesens, Óskarsdóttir, & Lemahieu (2021), who find XGB outperforming challenging deep learning benchmarks. Meta-parameters of the base classifiers are tuned in a nested four-fold cross-validation on the training data. The meta-parameters of fairness processors are also tuned using grid search. The details on the meta-parameter values and the tuning procedure are provided in the online Appendix.

Fairness processors and benchmarks are evaluated on the test set using multiple performance metrics. First, we measure the profitability of a scorecard by computing profit per EUR issued by a financial institution. To estimate profit, we start from the Expected Maximum Profit (EMP) criterion (Verbraken et al., 2014). The EMP measures the incremental profit compared to a base scenario in which loan applications are accepted without screening. This often leads to a small magnitude of EMP differences across classifiers (Kozodoi et al., 2019) and complicates the interpretation of the metric. To enable a more direct interpretation, we normalize misclassification costs such that the base scenario represents rejecting all applications.

Table 3 provides the confusion matrix of a scoring model, where π_i are prior probabilities of good and bad risks, and $F_i(\tau)$ are predicted cumulative density functions of the scores of class i given a cutoff value τ. If an applicant is predicted to be a good risk, a financial institution faces cost B in case of an incorrect prediction and earns C from an accurate prediction. In contrast, if an applicant is predicted to be a bad risk, a company faces an opportunity cost C in case of an incorrect prediction. Parameters B and C are defined according to Verbraken et al. (2014).

The parameter B reflects the cost associated with misclassifying a bad risk. Providing credit to a defaulter, the company faces a loss; specifically, the expected loss in case of default:

$$B = \frac{\text{LGD} \cdot \text{EAD}}{A}, \tag{14}$$

where LGD refers to the loss given default, EAD is the exposure at default, and A is the principal. B varies between 0 and 1 and several distributions may arise (Somers & Whittaker, 2007). We follow Verbraken et al. (2014) and treat B as a random variable with probability distribution:

- $B = 0$ with probability p_0 (a customer repays the entire loan after default);
- $B = 1$ with probability p_1 (the bank loses the entire loan);
- B follows a uniform distribution in (0, 1) with $F(B) = 1 - p_0 - p_1$.

The parameter C reflects the opportunity cost or earned benefit associated with good risks. By accepting a good customer, the company earns the equivalent to the return on investment ROI:

$$C = \text{ROI} = \frac{I}{A}, \tag{15}$$

where I is the total interest payments. Given these parameters, we compute profit as:

$$\text{Profit} = \int_0^1 \Big[C \cdot \big(\pi_1(1-F_1(\tau)) - \pi_1 F_1(\tau)\big) - B \cdot \pi_0(1-F_0(\tau))\Big] f(B) d(B) \tag{16}$$

This paper follows the empirical findings of Verbraken et al. (2014) and assumes a constant ROI of 0.2664 and the point masses $p_0 = 0.55$ for no loss and $p_1 = 0.1$ for full loss to compute B.

Apart from estimating the profitability of each fairness processor, we also compute the area under the ROC curve (AUC), which is a widely used indicator of the discriminatory ability of a scoring model. In addition, we evaluate fairness by measuring independence, separation and sufficiency. We aggregate the performance of pre- and post-processors over seven credit scoring data sets, five training fold combinations and four base classifiers, obtaining 140 performance estimates per processor. Since in-processors do not require a base classifier, their performance is aggregated over 35 values obtained from seven data sets and five training fold combinations.

6. Empirical results

This section presents the empirical results. We first examine the correlation between the scorecard performance, profitability,

Table 4
Rank Correlation between Evaluation Metrics.

Metric	AUC	Profit	IND	SP	SF
AUC	1				
Profit	0.8014	1			
IND	−0.4707	−0.3774	1		
SP	−0.3326	−0.2994	0.9477	1	
SF	0.3489	0.1636	−0.2156	−0.1311	1

Abbreviations: AUC = area under the ROC curve, IND = independence, SP = separation, SF = sufficiency.

and fairness. Next, we compare the performance of different fairness processors. Last, drawing on the findings that suggest a strong negative correlation between profit and fairness, we examine the profit-fairness trade-off to appraise the monetary cost of fairness.

6.1. Correlation analysis

Table 4 depicts the mean Spearman correlation between the evaluation metrics. The correlation coefficients are computed on the performance estimates obtained from different variants of fairness processors and averaged over the seven credit scoring data sets. The results suggest that the AUC and profit often produce similar model rankings (correlation is 0.80). Still, there is some disagreement between the two measures, which indicates that optimizing profit is important to identify potentially more profitable scorecards. Therefore, we emphasize profit in the following.

Comparing profit and fairness, we observe a moderate negative correlation between independence, separation, and profitability[6] As expected, integrating fairness constraints to reduce discrimination prevents a scorecard from taking full advantage of the available information, which decreases profit. At the same time, a weak positive correlation between sufficiency and profit suggests that optimizing profitability without implementing additional fairness constraints could also improve sufficiency. This result confirms the observation that most ML algorithms are designed to automatically achieve sufficiency and implies that directly optimizing sufficiency with a fairness processor is not essential.

A different conclusion emerges from examining the agreement of the other two fairness criteria. As indicated by Table 4, independence and separation have a strong positive correlation of 0.95. Optimizing either of these two criteria will, therefore, favor models that fulfill both independence and separation. In other words, reducing the mutual information between a sensitive attribute and model predictions also helps to align the parity of error rates across the sensitive groups. This is an interesting finding, given that the former constraint targeted by independence is stricter compared to the one targeted by separation. For a risk analyst, the observed result implies that it is ample to rely on a single fairness criterion. Since separation has a better ability to capture the cost asymmetry (see Section 3 for details), we conclude that optimizing and measuring the separation criterion is the most suitable way to integrate and evaluate the fairness of a credit scoring model.

6.2. Benchmarking fairness processors

Table 5 provides average performance gains from fairness processors compared to the unconstrained scoring model across the seven credit scoring data sets. A positive gain indicates a better performance of a processor relative to the unconstrained model in terms of a particular evaluation measure. Individual results for each of the data sets are provided in the online Appendix.

Table 5 confirms that using a processor to enhance fairness decreases profit compared to the unconstrained model. Results in terms of the AUC mirror this finding, whereby two processors show marginally higher AUC values than the unconstrained model. Table 5 also evidences that the unconstrained model suffers from discrimination. Six out of eight processors achieve better independence and five processors attain better separation. However, sufficiency is consistently higher in the unconstrained model, which confirms that this metric differs fundamentally from independence and separation. High agreement between the sufficiency and profit, expressed by strict dominance of the unconstrained model in Table 5, also indicates that the goal of profit maximization is compatible with maximizing sufficiency, which questions the fairness perspective that the latter embodies.

Considering individual processors, the reject option classification post-processor demonstrates the best fairness in independence and separation. This is achieved by sacrificing more than 30% profit compared to the unconstrained model. On the other hand, we observe the least profit decrease of less than 5% for the prejudice remover, which also attains a similar AUC as the unconstrained model. At the same time, the prejudice remover provides a smaller fairness improvement than other processors. These results emphasize the trade-off between profit and fairness.

Comparing processors within the implementation methods, we can identify promising techniques. Considering post-processors, the equalized odds processor is dominated by reject option classification in all evaluation measures. Platt scaling achieves higher profit and sufficiency than the latter but gives the by far worst results in independence and separation. In sum, Table 5 clearly identifies reject option classification as the most suitable post-processor. Concerning pre-processors, no clear result emerges. Reweighing achieves the best fairness but decreases profitability by 23%. The disparate impact remover retains a higher share of profit but offers substantially smaller improvements in independence and separation.

Among the in-processors, we observe the unconstrained model to dominate the meta fair algorithm, which displays negative results for all metrics of Table 5. Therefore, the meta fair algorithm does not warrant further consideration. Comparing the prejudice remover to adversarial debiasing, we find the former to deliver better results in all metrics but sufficiency. Given reservations against the fairness concept of the sufficiency metric, the results of Table 5 suggest that the prejudice remover is the best performing in-processor.

The results of Table 5 have several implications. First, we identify two fairness processors, Platt scaling and the meta-fair algorithm, inadequate for credit scoring since they decrease profit and predictive performance while not improving fairness compared to the unconstrained model. Second, we find that the equalized odds processor is dominated by another post-processor in all considered evaluation metrics and should, therefore, be avoided.

The remaining processors arrive at different solutions in the space between sacrificing profit and reducing discrimination, leaving decision-makers with the difficult task to balance these conflicting goals according to their preferences, business requirements, and regulation. In general, in-processors offer more flexibility in prioritizing fairness or profit through meta-parameters. For example, the prejudice remover incorporates a regularizer to penalize fairness violations and exposes the weight of that penalty as a meta-parameter. However, the benefit of higher flexibility carries a cost. Compared to alternative options, in-processors replace existing scorecards and impact the scoring process the most. Post-processors largely retain an existing scoring pipeline, which simpli-

[6] Higher values of the AUC and profit indicate better performance, whereas lower values of independence, separation, and sufficiency indicate higher fairness. Therefore, we invert correlation signs between the two former performance metrics and the three fairness criteria to facilitate the consistent interpretation of the results.

Table 5
Average Performance Gains from Fairness Processors Relative to the Unconstrained Model.

Method	Fairness processor	AUC	Profit	IND	SP	SF
Pre-processing	Reweighting	-3.19%	-23.04%	66.00%	61.24%	-38.18%
	Disparate impact remover	**0.82**%	-10.60%	5.33%	4.50%	**-19.99**%
In-processing	Prejudice remover	0.37%	**-4.28**%	11.51%	9.41%	-202.36%
	Adversarial debiasing	-0.21%	-13.90%	9.38%	2.98%	-148.36%
	Meta fair algorithm	-2.98%	-7.25%	-7.49%	-20.88%	-108.17%
Post-processing	Reject option classification	-8.64%	-30.71%	**74.80**%	**74.55**%	-263.51%
	Equalized odds processor	-16.22%	-59.73%	25.83%	-11.08%	-407.82%
	Platt scaling	-0.45%	-26.98%	-85.28%	-108.45%	-85.02%
Average change across fairness processors		-3.81%	-22.06%	12.51%	1.53%	-159.18%

Abbreviations: AUC = area under the ROC curve, IND = independence, SP = separation, SF = sufficiency. Values represent percentage differences relative to an unconstrained scorecard averaged over seven data sets × five folds × four base models; positive values indicate improvement.

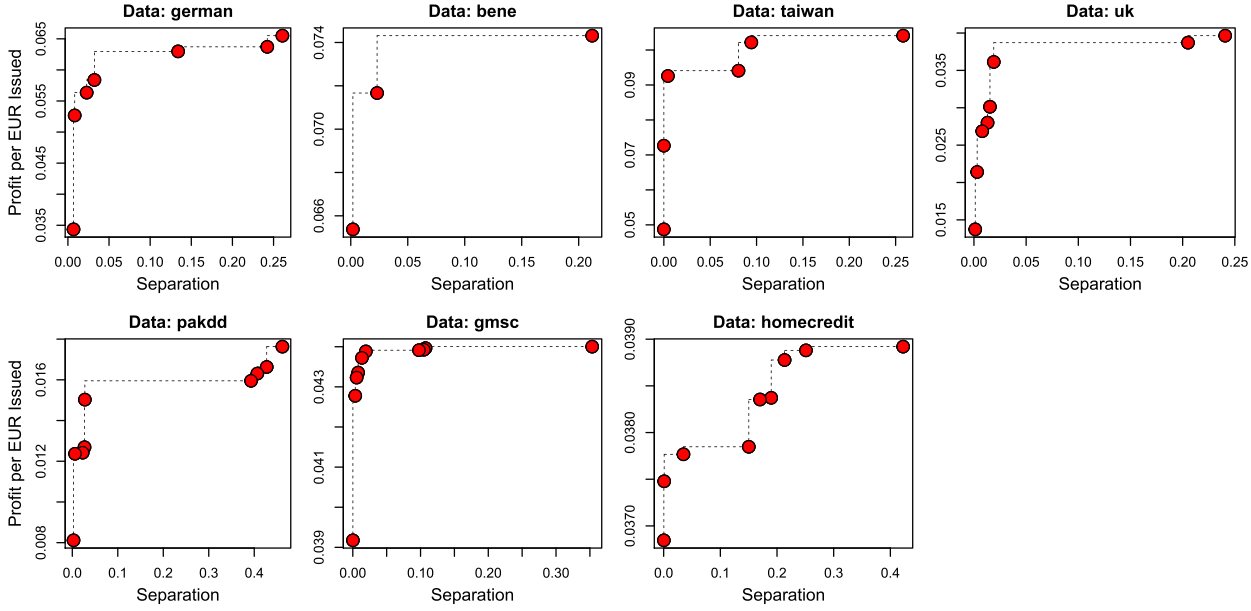

Fig. 2. Profit-Fairness Trade-Off: Frontiers with Non-Dominated Solutions.

fies their deployment. Pre-processors address fairness at the data level, which represents a more invasive change of the scoring process compared to post-processing but seems less difficult to implement than in-processing. Together with the results of Table 5, in which the best in-processor (i.e., the prejudice remover) finds a better trade-off between profit and fairness than the disparate impact remover while the best post-processor (i.e., reject option classification) increases fairness to a larger extent than reweighting, considerations related to the complexity of deploying fairness processors and revising loan approval processes suggest two options for addressing fairness in credit scoring. Decision-makers can choose between a flexible but invasive in-processor and a post-processor, which is easier to deploy but might substantially decrease profitability. Table 5 represents the corresponding options by the prejudice remover and reject option classification.

6.3. The cost of fairness

Previous results indicate that it is possible to improve fairness by sacrificing profit. Fig. 2 provides a more detailed examination of the profit-fairness trade-off on each of the seven data sets using the concept of Pareto frontiers. The points on the frontiers refer to the test set performance of fairness processors trained with different base classifiers and on different combinations of the training folds. The frontiers only contain the non-dominated solutions, i.e., the points where it is impossible to improve on one objective (i.e., profit) without harming the other objective (i.e., fairness). Based on the previous results, we use the separation criterion to measure fairness.

Fig. 2 reveals that discrimination can be substantially reduced at a relatively low cost. Recall that separation indicates the difference between the false positive and false negative rates across the sensitive groups. According to Fig. 2, reducing the difference in error rates below 0.2 is possible while sacrificing less than € 0.01 profit per EUR issued. Across the data sets, this translates to an average profit reduction of 4.91% compared to the most profitable scorecard with stronger discrimination. At the same time, completely eliminating unfairness is more costly: achieving separation of 0 is only possible when sacrificing more than 35% of the profit. However, since perfect fairness is not required by regulation, we conclude that a financial institution can reduce discrimination to a reasonable extent while maintaining a relatively high profit margin.

7. Conclusion

The paper sets out to consolidate recent advancements in fair ML from a credit scoring perspective. Cataloging approaches for

quantifying fairness and the ML pipeline interventions for fairness maximization, we have examined the adequacy of these fairness measures and processors for credit scoring. To substantiate our conceptual analysis, we have undertaken a systematic empirical comparison of several fairness processors from different families to identify preferable approaches and clarify the degree to which increasing fairness in loan approval processes harms profitability.

The conceptual comparison of different fairness criteria reveals separation to be the most appropriate metric for credit scoring. Separation acknowledges the imbalanced misclassification costs, which are instrumental to the lending business. The presented catalog of fairness processors offers practitioners a starting point for deciding which processors to consider for a given problem setting. The catalog also indicates that most processors have been evaluated based on their accuracy and that some relevant credit scoring scenarios are not well covered by the available processors. For example, in a setting with multiple sensitive attributes (e.g., race and religion), only two processors, adversarial debiasing and reject option classification, facilitate optimizing the separation criterion.

The empirical study benchmarks fairness processors in a profit-oriented credit scoring setup. Several implications emerge from the results. First, examining the agreement between the fairness criteria under study reveals that separation and independence are strongly correlated. While other empirical studies support this finding (Friedler et al., 2019), it contradicts the intuition from theoretical considerations that fairness criteria are mutually exclusive (Mitchell et al., 2021). We also find that sufficiency has a property to be achievable by any well-trained classifier that can predict the sensitive attribute from the other features (Barocas et al., 2019). This calls into question the overall suitability of sufficiency for credit scoring and further emphasizes separation as a proper criterion for measuring the fairness of credit scorecards.

Second, we find that the choice of an appropriate fairness processor depends on the implementation feasibility and preferences of a decision-maker regarding the conflicting objectives of profit and fairness. Post-processing methods such as reject option classification are the easiest to implement in production but improve fairness at a high monetary cost. In-processors such as the prejudice remover perform best in finding the profit-fairness trade-off and offer the most flexibility in calibrating the importance of the conflicting objectives. However, using in-processors requires replacing a deployed scoring model with a new algorithm, which might require regulatory approval and is, more generally, associated with considerable efforts.

Third, while achieving perfect fairness is costly, we find that reducing discrimination to a reasonable extent is possible while maintaining a relatively high profit. These results support the current anti-discrimination regulation that allows unfairness to exist up to a certain limited extent. The analysis of fairness processors from the perspective of the Pareto frontiers offers decision-makers a tool to analyze the profit-fairness trade-off specific to their context and identify techniques that reduce discrimination to a required level at the smallest monetary cost.

Our study may also have implications for customer scoring models beyond the credit industry. Fairness concerns arise from the increasing use of ML to automate decisions in many domains, such as hiring (Barocas et al., 2019), college admission (Mitchell et al., 2021) or criminal risk assessment (Berk et al., 2021). The catalog of fairness processors and the results of their empirical analysis can aid these domains in identifying suitable techniques for integrating fairness in decision support systems. Future work on fair ML may also draw value from the empirical comparison in that it highlights effective approaches that set a benchmark for new fairness processors.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2021.06.023

References

Banasik, J., & Crook, J. (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research, 183*(3), 1582–1594.
Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning*. fairmlbook.org.
Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*, 671–732.
Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research, 50*(1), 3–44.
Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *IEEE international conference on data mining workshops* (pp. 13–18).
Calders, T., & Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery, 21*(2), 277–292.
Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Advances in neural information processing systems* (pp. 3992–4001).
Celis, L. E., Huang, L., Keswani, V., & Vishnoi, N. K. (2019). Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Conference on fairness, accountability, and transparency* (pp. 319–328).
Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).
Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163.
Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research, 183*(3), 1447–1465.
Equal Credit Opportunity Act (1974). Art. 9 & 15 U.S. code §1691. https://www.law.cornell.edu/uscode/text/15/1691c.
European Commission (2017). Guidelines on data protection officers. https://ec.europa.eu/newsroom/article29/items/612048.
Executive Office of the President (2016). Big data: A report on algorithmic systems, opportunity,and civil rights. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.
Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 259–268).
Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Conference on fairness, accountability, and transparency* (pp. 329–338).
Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2017). Predictably unequal? The effects of machine learning on credit markets. *Technical Report*. National Bureau of Economic Research.
Goh, G., Cotter, A., Gupta, M., & Friedlander, M. P. (2016). Satisfying real-world goals with dataset constraints. In *Advances in neural information processing systems* (pp. 2415–2423).
Gunnarsson, B. R., Vanden Broucke, S., Baesens, B., Óskarsdóttir, M., & Lemahieu, W. (2021). Deep learning for credit scoring: Do or dont? *European Journal of Operational Research*. https://doi.org/10.1016/j.ejor.2021.03.006.
Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315–3323).
Johndrow, J. E., Lum, K., et al. (2019). An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics, 13*(1), 189–220.
Kamiran, F., & Calders, T. (2009). Classifying without discriminating. In *International conference on computer, control and communication* (pp. 1–6).
Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. In *International conference on data mining* (pp. 924–929).
Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 35–50).
Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv:1412.6980,.
Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *8th innovations in theoretical computer science conference* (pp. 43:1–43:23).
Kozodoi, N., Lessmann, S., Papakonstantinou, K., Gatsoulis, Y., & Baesens, B. (2019). A multi-objective approach for profit-driven feature selection in credit scoring. *Decision Support Systems, 120*, 106–117.
Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136.
Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed impact of fair machine learning. In *International conference on machine learning* (pp. 3150–3158).

Louizos, C., Swersky, K., Li, Y., Welling, M., & Zemel, R. (2016). The variational fair autoencoder. In *International conference on learning representations*.

Luong, B. T., Ruggieri, S., & Turini, F. (2011). K-NN as an implementation of situation testing for discrimination discovery and prevention. In *ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 502–510).

Mitchell, S., Potash, E., Barocas, S., DAmour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application, 8*, 141–164.

Narayanan, A. (2018). Translation tutorial: 21 fairness definitions and their politics. In *Conference on fairness, accountability, and transparency*.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *International conference on machine learning* (pp. 625–632).

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers, 10*, 61–74.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In *Advances in neural information processing systems* (pp. 5680–5689).

Somers, M., & Whittaker, J. (2007). Quantile regression for modelling distributions of profit and loss. *European Journal of Operational Research, 183*(3), 1477–1487.

Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research, 238*(2), 505–513.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., & Srebro, N. (2017). Learning non-discriminatory predictors. In *Conference on learning theory* (pp. 1920–1953).

Zadrozny, B., & Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *International conference on machine learning* (pp. 609–616).

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017a). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International conference on world wide web* (pp. 1171–1180).

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017b). Fairness constraints: Mechanisms for fair classification. In *International conference on artificial intelligence and statistics* (pp. 962–970).

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *International conference on machine learning* (pp. 325–333).

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *AAAI/ACM conference on ai, ethics, and society* (pp. 335–340).