Operationalizing the Search for Less Discriminatory Alternatives in Fair Lending

Talia Gillis
tbg2117@columbia.com
Columbia University
New York, New York, USA

Vitaly Meursault
Federal Reserve Bank of Philadelphia
Philadelphia, PA , USA

Berk Ustun
berk@ucsd.edu
UC San Diego
La Jolla, CA, USA

ABSTRACT

The *Less Discriminatory Alternative* is a key provision of the disparate impact doctrine in the United States. In fair lending, this provision mandates that lenders must adopt models that reduce discrimination when they do not compromise their business interests. In this paper, we develop practical methods to audit for less discriminatory alternatives. Our approach is designed to verify the existence of less discriminatory machine learning models – by returning an alternative model that can reduce discrimination without compromising performance (*discovery*) or by certifying that an alternative model does not exist (*refutation*). We develop a method to fit the *least* discriminatory linear classification model in a specific lending task – by minimizing an exact measure of disparity (e.g., the maximum gap in group FNR) and enforcing hard performance constraints for *business necessity* (e.g., on FNR and FPR). We apply our method to study the prevalence of less discriminatory alternatives on real-world datasets from consumer finance applications. Our results highlight how models may inadvertently lead to unnecessary discrimination across common deployment regimes, and demonstrate how our approach can support lenders, regulators, and plaintiffs by reliably detecting less discriminatory alternatives in such instances.

ACM Reference Format:
Talia Gillis, Vitaly Meursault, and Berk Ustun. 2024. Operationalizing the Search for Less Discriminatory Alternatives in Fair Lending. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3630106.3658912

1 INTRODUCTION

Disparate impact doctrine is the focal point for discussions of algorithmic fairness [see e.g., 12, 38, 40]. Much of the literature on this topic examines how the legal requirements surrounding disparate impact can be translated into methods to measure or mitigate algorithmic discrimination across protected classes [see, e.g., 6, 31, 81]. To date, these efforts have overlooked a crucial component of disparate impact doctrine – namely that policies that lead to discrimination over protected groups should only be used when there is no less discriminatory alternative policy to achieve the same objectives. This requirement, known as the *less discriminatory alternative*, mandates that entities who support or automate decisions with a discriminatory model must demonstrate that a less discriminatory model does not exist.

The LDA provision is a powerful opportunity to promote algorithmic fairness in domains ranging from housing to hiring. In consumer finance – a "high-risk" domain [30] where models have long been subject to regulatory oversight [see e.g., 67] – the LDA provision represents an avenue to safeguard consumers against discrimination at scale through audits. On one hand, a plaintiff may conduct an LDA search to challenge discriminatory practices [56]. On the other, a lender may use an LDA search to show regulators that their lending models do not discriminate, or that the disparities would inevitably compromise their business interests. Such internal and external audits are increasingly relevant for fair lending. Regulators have already expressed the position that lenders have an affirmative duty to search for an LDA [see e.g., 15, 59]. Likewise, third-party auditors now provide services to conduct an LDA search in lending applications – see e.g., FairPlay.AI [1] who provide a service "[t]une... models to be fairer while preserving or enhancing accuracy."

Despite the potential of an LDA search to establish that a lending practice is not discriminatory, there is no standardized method to satisfy this requirement. Traditional approaches to establishing the existence of an LDA focused on examining a model's dependence on specific features and their influence on disparities [11, 41]. In practice, these input-based and ad hoc methods fall short of facilitating substantial LDA searches, especially when lending decisions are automated by complex models with numerous features. The considerable latitude in conducting an LDA search, coupled with the absence of regulatory guidelines, fails to harness the potential of the LDA provision for consumer protection. The result is that consumers remain vulnerable, lenders face compliance challenges [58], and regulators lack effective oversight.

In this paper, we turn the search for LDA into a formal audit that promotes transparency – by specifying measures of disparity and business interest and returning information that minimizes reliance on human discretion. The key technical contribution of our approach is that it allows for both *discovery*, when we are able to demonstrate the existence of an LDA, and *refutation*, when the problem is infeasible and we are unable to find an LDA – meaning there is no linear model that can reduce discrimination on a population of interest. While our framework sets the measures of disparity and

Disclaimer: The views in this paper are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. No statements here should be treated as legal advice. Any errors or omissions are the responsibility of the authors.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658912

business interest, we also provide flexibility within the structured search where there is currently significant legal ambiguity. Our main contributions include:

(1) We formalize the LDA problem by considering a setting in which a lender uses a probabilistic classifier to predict the probability of repayment and then uses a threshold for lending decisions.
(2) We develop a method to search for the *least* discriminatory linear classification model. Our method solves a combinatorial optimization problem that can capture exact measures of disparity and performance. By directly solving this problem, we can exhaustively search over all models that obey hard constraints on business necessity, thus reliably returning less discriminatory models when they exist and refuting their existence when they do not
(3) We present results from a comprehensive empirical study of least discriminatory alternatives for real-world classification tasks from consumer finance. Our results highlight how models may inadvertently lead to unnecessary discrimination in deployment, and demonstrate how our methods can detect and mitigate such instances.
(4) We provide a Python implementation of our method, available on Github https://github.com/ustunb/ldasearch.

2 BACKGROUND AND RELATED WORK

Discrimination Law. In the context of the disparate impact doctrine, practices that negatively affect a protected group are deemed impermissible, even when they are justified as fulfilling a legitimate business need, if there is a "less discriminatory alternative" (LDA) to achieve the same goal. This LDA requirement represents the final stage in the three-step process of a disparate impact claim. Within this burden-shifting framework, particularly in litigation, the plaintiff bears the initial burden of demonstrating that a policy creates disparities. Subsequently, the defendant must justify the challenged policy by showing that it was intended for a legitimate goal ("business necessity"). However, even with this justification, the defendant may still face liability if "those interests could not be served by another practice that has a less discriminatory effect" [71].

Our work focuses on the disparate impact doctrine under fair lending laws. Discriminatory lending is prohibited under the *Fair Housing Act* (FHA)[1] and the *Equal Credit Opportunity Act* (ECOA)[2]. The FHA prohibits housing discrimination, including mortgage lending, on the basis of race, color, religion, sex, disability, familial status, and national origin, and ECOA bans discrimination in all types of credit transactions on the basis of sex, marital status, race, color and religion. ECOA and FHA cover intentional or direct discrimination (*disparate treatment*) and facially neutral conduct that has a discriminatory effect (*disparate impact*). [3]

While the FHA and ECOA adopt the LDA requirement—that a policy not be adopted unless there is no less discriminatory alternative to achieve a legitimate policy goal—there is little guidance on how to conduct an LDA search. In practice, the LDA stage of a disparate impact claim is rarely litigated and "[c]ourts applying the disparate impact standard under the FHA or the ECOA rarely have discussed, much less reached, the 'third prong' of less discriminatory alternatives analysis" [48]. Regulatory enforcement action typically settles before any meaningful discussion of the standard is developed resulting in a lack of case law to formulate the less discriminatory alternative standard [9, 48].

Existing approaches to establish the existence of an LDA primarily focus on a evaluating the inputs to a model, rather than how they contribute to disparity. For example, when lenders use a borrower's gross income in their underwriting decisions, it can negatively impact elderly applicants because the measure overlooks differences in tax rates. A less discriminatory alternative to measuring income would therefore take into account the increased value of nontaxable income. [41]. As a separate example, a 2012 study on the disparate impact of credit scores considered whether including features that encode credit characteristics would exacerbate discrimination in a model [11]. Such exploratory input-based and ad hoc approaches are unlikely to spot discrimination or refute the existence of alternative models – especially with more complex models.

A formal approach to audit for an LDA may provide guidance to multiple stakeholders. These include: lenders and regulators who may audit lending decisions as part of their lender oversight [see e.g., 7, 55]; and plaintiffs to challenge discriminatory lending practices. Under the disparate impact burden-shifting framework, the LDA provision is often considered the burden of the plaintiff[4]. In the context of fair lending, this requires that a plaintiff must show that a lender who successfully defended a practice as meeting a legitimate business should still face liability. Assuming a plaintiff has access to a lender's model and dataset, they would be able to use our method to test if a LDA exists. In the context of fair lending, regulators have already expressed the position that lenders have an affirmative duty to search for a LDA [59]. Likewise, scholars have argued that the burden to search for a LDA should be on the entity deploying the algorithmic models, given their superior position in conducting a LDA search [15].

Algorithmic Fairness. Our work is broadly related to a stream of methods on fairness in machine learning, including: methods to learn fair models [28, 57, 82]; methods to reduce performance disparities through post-processing [see e.g., 20, 76, 79]; and methods to characterize the trade-off between fairness and accuracy [14, 50, 75].

Our approach differs from methods for learn fair models or reducing discrimination via post-processing as it outputs a model that optimizes for accuracy and fairness directly – by minimizing exact measures such as the 0-1 loss rather than convex surrogates

[1] The FHA (42 U.S.C. §§ 3601-3619, 3631), also known as Title VIII of the Civil Rights Act of 1968, which protects renters and buyers from discrimination by sellers or landlords and covers a range of housing-related conduct, including the setting of credit terms.

[2] 15 U.S.C. §§ 1691-1691f

[3] The Supreme Court affirmed that disparate impact claims could be made under FHA in Texas *Dep't of Hous. & Cmty. Affs.* v. *Inclusive Communities Project, Inc.*, 576 U.S. 519 (2015). There is not an equivalent Supreme Court case with respect to ECOA, but the Consumer Financial Protection Bureau (CFPB), the agency primarily responsible for enforcing the ECOA, and lower courts have found that the statute allows for a claim of disparate impact [?]. In recent years, there have been some challenges to the recognition of disparate impact under ECOA [54].

[4] See U.S. Department of Housing and Urban Development (HUD) [71], who write: "If the respondent or defendant satisfies the burden of proof... the charging party or plaintiff may still prevail upon proving that the substantial, legitimate, nondiscriminatory interests supporting the challenged practice could be served by another practice that has a less discriminatory effect."

like the logistic loss. This is an algorithm design decision that sacrifices computation to achieve more reliable *discovery* and *refutation* (see Section 4). In the context of an LDA search, methods using approximations return models that achieve a lower trade-off between fairness and accuracy – which would compromise an audit by suggesting that we cannot reduce discrimination without impacting performance. In such cases, we may also not be able to rule out the existence of an LDA – as we cannot determine if we failed to find an alternative model because it is not viable, or because of the misalignment in our problem specification.

Our approach returns an estimate of the maximum reduction in disparity with respect to a baseline model that the lender could deploy. This reduction could be determined using methods to bound or characterize trade-offs between fairness and accuracy [see e.g., 14, 50, 75]. These techniques could be adapted to bound the best performance achievable under constraints on group fairness, which may be valuable in refuting the existence of an LDA but would not be able to produce a tangible model that could be shown to a regulator as evidence of an LDA or to a lender as a starting point of a search.

Our results highlight the existence of LDAs across prediction tasks. This result broadly reflects the fact that datasets admit competing models that perform almost equally well – a notion that is known as *model multiplicity* [17]. Existing work highlights how competing models can differ in terms of salient properties – such as their predictions [26, 49, 77, 78], explanations [19], interpretability [62, 63], and fairness [24]. These results suggest that many lending tasks may admit alternative models that can reduce discrimination without affecting a lender's bottom line [15].

The closest work to ours is that of Coston et al. [24], who propose an algorithm to train a model that bounds minimum performance disparity in a prediction task with selective labels. This work considers a formulation that is similar to ours – i.e., minimize disparity over models that achieve near-optimal loss – but solved it using a reductions approach that outputs a *stochastic classifier* – i.e., a collection of models that are randomly chosen to assign predictions at test time. In the context of an LDA search, this model would not stand as a viable alternative model that a lender could deploy as it would assign predictions that change across multiple applications [see e.g., 25, for a list]. For instance, a stochastic classifier would allow an applicant who is denied to be approved by simply applying multiple times.

3 PROBLEM STATEMENT

Preliminaries. We consider a classification task where a lender trains a model from a dataset of n examples $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. Each example contains of a vector of $d+1$ features $\boldsymbol{x}_i = [1, x_{i1}, \ldots, x_{id}] \in \mathcal{X} \subseteq \mathbb{R}^{d+1}$ and a label $y_i \in \mathcal{Y} = \{0, 1\}$, where $y_i = 1$ indicates an outcome of interest – e.g., applicant i will repay a loan within 2 years.

We assume that the lender uses the dataset to train a *classification model* that takes as input the features of each applicant $\boldsymbol{x}_i \in \mathcal{X}$ and returns as output a prediction that will determine their loan decision $\hat{y}_i \in \{0, 1\}$.

- If the model outputs a *predicted probability of repayment* $f : \mathcal{X} \to [0, 1]$ (e.g., a logistic regression model), the lender approves applicants whose predicted probability exceeds a threshold value $\alpha \in [0, 1]$. In this case, the loan decision for an applicant with features $\boldsymbol{x}_i$ is $\hat{y}_i := \mathbb{I}[f(\boldsymbol{x}_i) \geq \alpha]$, where $\mathbb{I}[\cdot]$ denotes the indicator function.
- If the model outputs a *predicted label* $f : \mathcal{X} \to \mathcal{Y}$ (e.g., a random forest), the lender approves all applicants who are predicted to repay. In this case, the loan decision for an applicant with features $\boldsymbol{x}_i$ is $\hat{y}_i := f(\boldsymbol{x}_i)$.

In both cases, we can map the features of an individual applicant $\boldsymbol{x}_i$ to their loan decision, and evaluate model performance in terms of standard metrics such as those listed in Table 1.

We assume lenders train and deploy their model to maximize profits. In Remark 1, we show how a lender who uses a probabilistic classifier can maximize their expected profits in a simple lending task by adjusting their thresholds.

Remark 1 (Profit-Maximizing Threshold). *Consider a lending task where the cost of each instance is denoted as $C(\hat{y}, y)$ where $\hat{y} \in \{0, 1\}$ denotes a lender's decision and $y \in \{0, 1\}$ denotes an applicant's repayment. In a task where a lender issues loans for $L > 0$ and expects an interest payment of $R > 0$, we have that $C(1, 0) = L$, $C(1, 1) = -R$, $C(0, 1) = C(0, 0) = 0$. Thus, the lender maximizes profit by approving applicants using the threshold rule:*

$$f(\boldsymbol{x}) \geq \frac{L/R}{1 + L/R} \tag{1}$$

Given that each threshold maps to a specific combination of FNR and FPR, Remark 1 implies that a lender could adjust their profits by adjusting its FNR and FPR. In what follows, we therefore restrict our attention to a setting where a lender will use a model that outputs hard label predictions $f_0 : \mathcal{X} \to \mathcal{Y}$ for clarity. This is without loss of generality since we can measure the performance and profits for a probabilistic classifier in the same way as a model that outputs hard label predictions.

3.1 Operationalizing the Search for Least Discriminatory Alternatives

We consider a setting where a lender issues loans with a *baseline classifier* f_0 that exhibits *disparate impact* over *groups* defined by *protected characteristics* such as age, sex, ethnicity, gender identity or disability status [70]. In practice, disparate impact can be quantified by comparing model performance or approval rates across protected groups. The disparities metric is externally defined rather than being determined by the lender. As a result, any disparities that arise based on this metric can give rise to a presumptive claim of disparate impact, regardless of the lender's intent. In effect, disparities may arise inadvertently when lenders issue loans with a baseline model trained using data from consumers in a different market [13, 65] or at a different period in time [13], or without data on protected groups [e.g., due to practical challenges or legal barriers 10].

Characterizing Business Necessity. Our goal is to determine if there exists an *alternative model* f' that reduces discrimination across a set of protected groups $g \in \mathcal{G}$. In practice, lending policies that lead to disparities are defensible when their goals are considered a *business necessity.*[5] Thus, we start by specifying the

[5] Although there is some disagreement on the types of profit-making policies are justifiable under the "business necessity" defense [37], it is widely accepted that

Metric	Definition	Estimator	Interpretation
False Negative Rate	$\mathsf{FNR}(f) := \mathbb{E}[\hat{y} \neq y \mid y = 1]$	$\widehat{\mathsf{FNR}}(f) := \frac{1}{n^{+}} \sum_{i \in I^{+}} \mathbb{I}[\hat{y}_i = 0]$	Denial rate of applicants who would repay
False Positive Rate	$\mathsf{FPR}(f) := \mathbb{E}[\hat{y} \neq y \mid y = 0]$	$\widehat{\mathsf{FPR}}(f) := \frac{1}{n^{-}} \sum_{i \in I^{-}} \mathbb{I}[\hat{y}_i = 1]$	Approval rate of applicants who would default
Error Rate	$\mathsf{ERR}(f) := \mathbb{E}[\hat{y} \neq y]$	$\widehat{\mathsf{ERR}}(f) := \sum_{i=1}^{n} \mathbb{I}[\hat{y}_i \neq y_i]$	Proportion of applicants with incorrect loan decisions

Table 1: Performance measures for binary classification models in loan approval. We assume that applicant with features $\boldsymbol{x}_i$ is approved when $\hat{y}_i = \mathbb{I}[f(\boldsymbol{x}_i) \geq \alpha] = 1$ and that they repay their loan when $y_i = 1$.

requirements for an alternative model to comply with business necessity.

Definition 2 (Business Necessity). *Consider a lender who issues loans using a model f_0, we say that a model f' is an* alternative model *so long as that $FNR(f') \leq FNR(f_0)$ and $FPR(f') \leq FPR(f_0)$.*

These conditions require that alternative models achieve comparable FNR *and* FPR to the baseline model on the population of interest. These requirements follow from Remark 1, which implies that the profits in a lending task are determined by a threshold rule written in terms of the FPR and FNR of the baseline model. In other words, if one can find an alternative model that obeys Definition 2, then a lender could use it to issue loans without affecting profits.

The conditions in (2) reflect those in a simple lending task and could be generalized to more complex lending tasks. Even in this simple case, however, business necessity conditions are stronger than the classical conditions in the model multiplicity literature, which would require alternative models that achieve comparable accuracy [49] or loss [24, 35]. In this case, the conditions on FNR and FPR rule out alternative models that achieve comparable accuracy but adversely impact business interests by trading off FNR for FPR.

Measuring Disparity. We measure the discrimination of all model f over a set of protected groups $g \in \mathcal{G}$ in terms of *group disparity metric* denoted $\Delta(f, g, g')$. We allow $\Delta(f, g, g')$ to represent any metric that captures the difference in performance or predictions between two groups g and g', and list examples in Table 2. Given a group disparity metric, we measure discrimination at the population level through an aggregate disparity measure denoted $\Delta(f)$. Given a model f, a set of group attributes $\mathcal{G}$, and group disparity metric, the *aggregate disparity* $\Delta(f)$ is a population-level statistic that summarizes the group disparity $\Delta(f, g, g')$ over all pairs of groups $g, g' \in \mathcal{G}$. In practice, we would consider the *worst-case group disparity* or the *mean group disparity*:

$$\overset{\text{Worst-Case Disparity}}{\Delta(f) := \max_{g,g' \in \mathcal{G}} \Delta(f, g, g')} \qquad \overset{\text{Mean Disparity}}{\Delta(f) := \sum_{g,g' \in \mathcal{G}} \frac{1}{n_g} \frac{1}{n_{g'}} \Delta(f; g, g')}$$

Training the Least Discriminatory Model. We test for the existence of *less discriminatory alternative* (LDA) model by fitting the *least discriminatory model* in a specific class of models. Our procedure requires two inputs:

(1) *Auditing Dataset* $\mathcal{D}^{\text{LDA}} := (\boldsymbol{x}_i, y_i, g_i)_{i=1}^{n}$ containing features, labels, and group attributes from the target population.
(2) *Performance Metrics* of the baseline model f_0 on $\mathcal{D}^{\text{LDA}}$ to enforce business necessity constraints in Definition 2, namely the false negative rate $\mathsf{FNR}_0 := \mathsf{FNR}(f_0; \mathcal{D}^{\text{LDA}})$ and false positive rate $\mathsf{FPR}_0 := \mathsf{FPR}(f_0; \mathcal{D}^{\text{LDA}})$.

Given these inputs, we fit a model that minimizes discrimination while adhering to constraints on business necessity by solving an empirical risk minimization problem of the form:

$$\begin{aligned} \min_{f \in \mathcal{F}} \quad & \Delta(f; \mathcal{D}^{\text{LDA}}) \\ \text{s.t.} \quad & \mathsf{FNR}(f; \mathcal{D}^{\text{LDA}}) \leq \mathsf{FNR}_0 + \varepsilon^{\text{FNR}} \\ & \mathsf{FPR}(f; \mathcal{D}^{\text{LDA}}) \leq \mathsf{FPR}_0 + \varepsilon^{\text{FPR}}. \end{aligned} \tag{2}$$

We refer to the optimization problem in (2) as the *LDA Problem* and denote an optimal solution as f_{LDA}. Here, the objective ensures that the alternative model f_{LDA} minimizes aggregate disparity as measured in terms of $\Delta(f)$. The constraints ensure compliance with business necessity constraints by controlling the difference in false negative rates and false positive rates between any alternative model and the baseline model. Our formulation cap the difference in $\mathsf{FNR}(f)$ and FNR_0 and $\mathsf{FPR}(f)$ and FPR_0 in terms of user-specified *slack paramters* $\varepsilon^{\text{FNR}} \in [0, 1]$ and $\varepsilon^{\text{FPR}} \in [0, 1]$, respectively. By default, we set $\varepsilon^{\text{FNR}} = 0$ and $\varepsilon^{\text{FPR}} = 0$ so that f_{LDA} will guarantee performance. In general, however, we can set these parameters to positive values to show we can achieve a meaningful reduction in discrimination at a negligible cost to performance, or find least discriminatory alternatives for other model classes.

3.2 Auditing with a Least Discriminatory Model

Solving the LDA problem returns the least discriminatory model among a set of models in a given lending task when it exists. This procedure can support a number of use cases – either by returning an alternative model (discovery) or by refuting its existence (refutation).

Discovery. Say we were to solve the LDA problem and recover an alternative model f_{LDA}. In this case, we can claim that there exists an alternative model that can reduce discrimination by up to $\text{Gain}(f_{\text{LDA}}; \mathcal{D}^{\text{LDA}})$.

$$\text{Gain}(f_{\text{LDA}}; \mathcal{D}^{\text{LDA}}) := \Delta(f_0; \mathcal{D}^{\text{LDA}}) - \Delta(f_{\text{LDA}}; \mathcal{D}^{\text{LDA}}) \tag{3}$$

When using the worst-case disparity metrics in (3.1), this would reflect the worst-case performance disparity over all groups. In cases where the LDA search is undertaken by auditors, such as regulators or private plaintiffs challenging lending practices, the least f_{LDA} provides evidence that there exists an alternative model. The alternative model f_{LDA} and the auditing dataset $\mathcal{D}$ could be shared with the lender so that the revised model development takes into consideration the existence of the LDA in addition to other constraints or considerations the lender faces.

lending decisions based on an empirical prediction of creditworthiness are likely to constitute a business necessity [see e.g., 61].

Metric	Definition	Estimate	Reference
FNR Gap	$\mathbb{E}[\,\hat{y}=0 \mid y=1, \mathcal{G}=g]$ - $\mathbb{E}[\,\hat{y}=0 \mid y=1, \mathcal{G}=g']$	$\frac{1}{n_g^+}\sum_{i\in I_g^+}\mathbb{I}[\,\hat{y}_i=0] - \frac{1}{n_{g'}^+}\sum_{i\in I_{g'}^+}\mathbb{I}[\,\hat{y}_i=0]$	Meursault et al. [51]
FPR Gap	$\mathbb{E}[\,\hat{y}=1 \mid y=0, \mathcal{G}=g]$ - $\mathbb{E}[\,\hat{y}=1 \mid y=0, \mathcal{G}=g']$	$\frac{1}{n_g^-}\sum_{i\in I_g^-}\mathbb{I}[\,\hat{y}_i=1] - \frac{1}{n_g^-}\sum_{i\in I_{g'}^-}\mathbb{I}[\,\hat{y}_i=1]$	Hurlin et al. [42]
Approval Gap	$\mathbb{E}[\,\hat{y}=1 \mid \mathcal{G}=g]$ - $\mathbb{E}[\,\hat{y}=1 \mid \mathcal{G}=g']$	$\frac{1}{n_g}\sum_{i\in I_g}\mathbb{I}[\,\hat{y}_i=1] - \frac{1}{n_{g'}}\sum_{i\in I_{g'}}\mathbb{I}[\,\hat{y}_i=1]$	FinRegLab [34]

Table 2: Group disparity metrics that can be used with our approach. Each gap measures the difference in performance or predictions between group g and group g'. We adopt the convention that smaller gaps $\Delta(f, g, g')$ are desirable. Thus, $\Delta(f, g, g') \geq 0 \implies$ that a model f performs “worse" on group g than on group g' – i.e., f attains a larger error rate, FPR, or FNR on group g than group g'.

Refutation. This may arise when the problem is infeasible – meaning that it is impossible to find an alternative model that meets the baseline FNR and FPR required for business necessity. Alternatively, we may find that the optimal model only achieves a negligible and insufficient reduction in discrimination [see e.g., 68, discussing the magnitude of decrease in disparities in LDA]. Refutation provides the lender with evidence to support a claim that a lender could not find a LDA to the current model [see e.g., 15, 59, who suggest that lenders should proactively engage in internal LDA audits]. Localized refutation-i.e. evidence that there is no LDA with respect to a particular model class and auditing dataset–can be generalized by additional LDA searches over other model classes and a richer set of features by lenders.

Applicability. Auditing for discrimination is challenging in practice – as lenders view their models as trade secrets [8] and datasets with group attributes may be hard to obtain or share [66]. Given these challenges, our procedure could be used to support audits in a variety of settings – e.g., by regulators or private plaintiffs – because as it can learn an LDA model using only data from the target population and aggregate performance statistics from the baseline model. This information can either be required to be disclosed to regulators or obtained by private parties through a discovery process. Ideally, an auditor would access both the baseline model and the dataset with group attributes – as this would allow them to determine the reduction in disparity with respect to the baseline model shown in (3). Lenders can use this auditing procedure inhouse when proactively demonstrating the refutation of an LDA or as part of their required fair lending compliance [59].

4 METHODOLOGY

In this section, we present an integer programming method to search for the least discriminatory model over the class of linear classifiers.

4.1 MIP Formulation

We consider a version of the LDA problem (2) to search for LDA models over the family of linear classification models. We consider the set of all linear classifiers with parameters $\boldsymbol{w} \in \mathbb{R}^{d+1}$

$$\hat{y} = f(\boldsymbol{x}) = 1 \text{ if and only if } \langle \boldsymbol{w}, \boldsymbol{x} \rangle \geq 0 \tag{4}$$

We fit the parameters for the LDA model by solving the following mixed-integer program:

$$
\begin{array}{llllr}
\min\limits_{\boldsymbol{w}} & \Delta & & & \\
\text{s.t.} & \Delta \geq \delta^+_{g,g'} & g, g' \in \mathcal{G} & \Delta \textit{ must exceed FN Gap between } g \textit{ and } g' & \text{(5a)} \\
& \Delta \geq \delta^-_{g,g'} & g, g' \in \mathcal{G} & \Delta \textit{ must exceed FP Gap between } g \textit{ and } g' & \text{(5b)} \\
& FN \leq FN_0 + \varepsilon_{\text{FN}} & & \textit{Business Necessity for FN} & \text{(5c)} \\
& FP \leq FP_0 + \varepsilon_{\text{FP}} & & \textit{Business Necessity for FP} & \text{(5d)} \\
& \delta^+_{g,g'} = \frac{FN_g}{n^+_g} - \frac{FN_{g'}}{n^+_{g'}} & g, g' \in \mathcal{G} & \textit{FN Gap between } g \textit{ and } g' & \text{(5e)} \\
& \delta^-_{g,g'} = \frac{FP_g}{n^-_g} - \frac{FP_{g'}}{n^-_{g'}} & g, g' \in \mathcal{G} & \textit{FP Gap between } g \textit{ and } g' & \text{(5f)} \\
& FN = \frac{1}{n^+} \sum\limits_{g \in \mathcal{G}} n^+_g FN_g & & \textit{Total FN Count} & \text{(5g)} \\
& FP = \frac{1}{n^-} \sum\limits_{g \in \mathcal{G}} n^-_g FP_g & & \textit{Total FP Count} & \text{(5h)} \\
& FN_g = \sum\limits_{i \in I^+_g} l_i & g \in \mathcal{G} & \textit{FN for Group } g & \text{(5i)} \\
& FP_g = \sum\limits_{i \in I^-_g} l_i & g \in \mathcal{G} & \textit{FP for Group } g & \text{(5j)} \\
& M_i l_i \geq y_i(\gamma - \sum\limits_{j=0}^{d} w_j x_{ij}) & i = 1, \ldots, n & \textit{Mistake for Point } i & \text{(5k)} \\
& w_j = w^+_j + w^-_j & j = 0, \ldots, d & \textit{Set Coefficients} & \\
& 1 = \sum\limits_{j=0}^{d} (w^+_j - w^-_j) & & \textit{Fix } \| \boldsymbol{w} \|_1 = 1 & \text{(5l)} \\
& l_i \in \{0, 1\} & i = 1, \ldots, n & \textit{Mistake Indicators} & \\
& w_j \in [-1, 1] & j = 0, \ldots, d & \textit{Coefficient Values} & \\
& w^+_j \in [0, 1] & j = 0, \ldots, d & \textit{Positive Components of } w_j & \\
& w^-_j \in [-1, 0] & j = 0, \ldots, d & \textit{Negative Components of } w_j &
\end{array}
$$

Here, each l_i is a binary variable set as $l_i \leftarrow \mathbb{I}[f(\boldsymbol{x}_i) \neq y_i]$ if a linear classifier with weights $\boldsymbol{w}$ makes a mistake on point i. The indicator behavior is enforced through constraints (5k). These constraints depend on a margin parameter γ, which should be set to a small positive number (e.g., 10^{-4}), and “Big-M" parameters M_i, which can be set as $M_i = \gamma + \max_i \|\boldsymbol{x}_i\|_\infty$ since we fix $\|\boldsymbol{w}\|_1 = 1$ in constraint (5l).

Auditing with a Solver. We formulate the ERM problem (2) as a mixed-integer program and solve it with a MIP solver such as CPLEX, Gurobi, and CBC. MIP solvers find the global optimum of a discrete optimization problem using exhaustive search algorithms like branch-and-bound [80]. In our setting, this returns three pieces of information:

1. Best Alternative Model f_{LDA}, i.e., the best solution found by the solver (2).
2. Upper Bound on Disparity $\Delta(f_{\text{LDA}}; \mathcal{D})$, i.e. disparity of f_{LDA} on the auditing dataset.
3. Lower Bound on Disparity, i.e. a bound on the lowest possible disparity that one could achieve using model in $\mathcal{F}$.

When the upper bound matches the lower bound, the alternative model that we obtain is *certifiably optimal.* If the solver fails to return a certifiably optimal solution within a user-specified time limit, the upper bound reflects the achievable reduction in disparity, which may be informative for discovery. Likewise, the lower reflects the minimal possible disparity one could hope to achieve, which could be useful for refutation.

Design Considerations. Solving the LDA Problem is a challenging computational task. Many measures of performance and disparity are discrete quantities that can optimized or constrained by solving hard combinatorial optimization problems. In effect, each measure requires that we count the number of mistakes over a subset of examples, which can be seen as special cases of 0-1 loss minimization. Modern approaches to train fair classification models through empirical risk minimization will replace such measures with surrogate measures that can be optimized efficiently [see e.g., 81]. Our approach can optimize these quantities directly because it leads to two major practical benefits in the context of an LDA Audit:

(1) *Reliability*: Our approach can recover the least discriminatory model when it exists and refute its existence when it does not. In contrast, consider an alternative approach where we fit a model by optimizing or constraining approximate measures of disparity or performance. Such a method would return models that are optimal or feasible with respect to approximate measures, which would compromise our ability to certify or refute the existence of an LDA in terms of the measures that we care about. For example, consider a method that optimizes approximate measures to return an LDA model that only achieves a small reduction in disparity. In practice, this may suggest the lender could not reduce disparity without compromising performance. Such a claim may be incorrect – as the method may have underestimated the reduction in disparity one could achieve in terms of the measures that we care about. Such behavior affects a large class of methods in fair machine learning [see e.g., 46, for further evidence on the effect of surrogate measures].

(2) *Versatility*: Our approach can measure disparity and performance in terms of a broad class of functions that capture predictions and performance at the group level. These include all measures of Table 2, as well as other variants currently used in fair lending, such as the Adverse Impact Ratio (AIR), measuring the differences in loan approvals for white and racial minority applicants [also see e.g., 21, who propose a combined measure of demographic parity and error disparity]. This degree of versatility is valuable given diversity of metrics that are used in practice by lenders [33] as well as the lack of consensus on which error rates should be used for discrimination [see e.g., 40].

Variants and Extensions. The MIP Formulation in (5) is a general formulation that minimizes worst-case disparity as measured through either the FNR gap or the FPR gap. In practice, this formulation can be adjusted to minimize other measures of aggregate disparity – e.g., the mean disparity gap by adjusting constraints (5a) and (5c). Likewise, we can adapt the formulation to minimize aggregate disparity over FNR gaps by dropping constraints (5f) and (5c), or over FPR gaps (5e) and (5a). Our formulation could also be extended to search for a least discriminatory model that complies with business necessity constraints at multiple operating points. Such a model could be useful in situations where a lender stipulates that an alternative model must be adaptive. This formulation differs from the variants described above in that it would be more challenging to solve – specifically, we would require n additional variables and constraints to count mistakes for each additional operation point.

Practical Considerations for Non-Linear Models. Our method is bound to achieve a reduction in disparity whenever we fit the LDA model from a class of models that contains the baseline model so that $f_0 \in \mathcal{F}$. This condition is bound to be satisfied whenever f_{LDA} and f_0 belong to the same model class so that $\mathcal{F} = \mathcal{F}_0$ (e.g., in an internal audit), or when we search over a simpler model class $\mathcal{F}_0 \subset \mathcal{F}$ (e.g., when the baseline model belongs to a class of linearly separable models). In cases where we have no guarantees on realizability, we can ensure the feasibility of the LDA Problem by adjusting the slack parameters ε^{FNR} and ε^{FPR} to capture potential discrepancies in performance a priori. This can be achieved by setting these values to capture the difference in FNR/FPR rates between f_0 and f_{LDA} on the auditing dataset, or through a model distillation approach where we fit a linear classifier to predict the output of the baseline model on the auditing dataset. As we show in Section 5, our approach may still discover an LDA model that leads to a meaningful reduction in disparity in prediction tasks where we cannot guarantee the realizability of a baseline model $f_0 \neq \mathcal{F}$.

5 EXPERIMENTS

In this section, we apply our method to search for less discriminatory alternatives in classification tasks in consumer finance. We have three goals: (i) to validate our method; (ii) to demonstrate how it can be used in practice; (iii) to characterize the prevalence of less discriminatory models in common deployment regimes.

5.1 Setup

Datasets. We work with three datasets from consumer finance applications in which lenders, regulators, or plaintiffs may wish to conduct an LDA search. Each dataset is publically available, used in prior work, and pertains to an application where models are used to support lending decisions, either directly by predicting repayment (`fico`, `german`), or indirectly by predicting income [6] (`income`). We process each dataset by imputing or dropping missing values and binarizing features. We present a list of summary statistics for each processed dataset in Table 3.

We split each dataset into three parts: *training* (60%), used to train a baseline model; *auditing* (20%), used to search for an LDA; and *test* (20%), used to compute unbiased estimates of disparity and performance.

Baseline Model. We use each dataset to fit a *baseline model* that a lender would deploy in a given application. We fit baseline models using *logistic regression* (LR) and *random forests* (RF). We choose

[6] Income prediction is a major component of alternative credit scores in the United States, which use predictions of income level as a feature in loan approval models for customers without credit history. Most companies develop these models internally, credit bureaus offer package solutions for income prediction https://www.transunion.com/content/dam/transunion/global/business/documents/product-creditvision-income-estimator-as.pdf [69]

these methods because they are widely used by industry practitioners [see e.g., 36, 61, 64] and cover model classes that we can and cannot search over;

LDA Model. We consider a simple audit in which we search for less discriminatory models with respect to groups defined by a binary attribute. We use $\mathcal{G}$ = sex = {male, female} for german and income, which represents a protected attribute as per FHA and ECOA. We use $\mathcal{G}$ = thin-file = {yes, no} for fico, which reflects the availability of data in an applicant's credit file. Even as *thin-file* is not a protected attribute, it is used as a proxy in internal audits as it leads to disparate impact across protected groups [see e.g., 18].

Given the baseline model, we search for the least discriminatory model by solving the LDA Problem Eq. (2). We consider instances that minimize the worst-case disparity and that measure group disparity in terms of the FNR Gap. We set the slack parameters in our formulation to $\varepsilon^{\text{FNR}} = 0.0\%$ and $\varepsilon^{\text{FPR}} = 0.0\%$ for LR, and to $\varepsilon^{\text{FNR}} = 0.5\%$ and $\varepsilon^{\text{FPR}} = 0.5\%$ for RF to adjust for possible changes in operating points due to changes in the underlying model class. We formulate each instance as a MIP using the formulation in (5), and solve it using CPLEX v22.1 [43] on a 3.6 GHz CPU with 32 GB RAM, setting a maximum time limit of one hour.

Deployment Regimes. We perform our LDA audit for each dataset and each baseline model in three *deployment regimes* that capture common distributional shifts in lending applications:

- Standard: This regime captures how an LDA search would perform if it were conducted by the lender as part of model development. We train a baseline model and set its operating point to impose a 3:1 cost between false negatives and true positives. This reflects a standard rule of thumb in which a lender needs 3 good accounts to break even on 1 defaulter [51].
- ModelShift: We consider a setting where a lender updates their lending policy after a model is deployed. This represents a credit-tightening regime in which lenders react to changing economic conditions by increasing credit approval thresholds. At higher approval thresholds, models typically exhibit greater disparities [51]. We simulate this shift by adjusting the baseline model to reflect a 5:1 cost between false positives and false negatives by adjusting the approval threshold (LR) or retraining the model (RF).
- LabelShift: We consider a setting where a baseline model is used to issue loans to consumers with a different repayment rate $\Pr(y = 1)$. This regime captures a common class of distribution shift when a lending model is used to issue loans in a new market, or under new economic conditions. We simulate this regime by undersampling positive points from the audit and test samples (specifically, we remove 10% of the positive examples). See [52] for a discussion of repayment change over time in the mortgage market.

5.2 Results

We summarize the results of our audit across all datasets, baseline model classes, and deployment regimes in Table 3. In what follows, we discuss these results.

On Generalization. Our results in Table 3 shows that we can find a least discriminatory alternative in all of our audits. In practice, these models correspond to models that can often lead to meaningful reductions in disparity while adhering to constraints on FNR and FPR to avoid compromising business necessity. In practice, these results reflect the performance we observe using the audit dataset, and that should be confirmed by evaluating their disparity and performance on a test dataset as shown in Table 4. In this case, we find that the gains we observe generalize across 15/18 instances. This result reflects one of the benefits of fitting the least discriminatory model over linear classifiers – as simple models are more likely to generalize.

On the Prevalence of LDAs. Our results highlight diverse ways in which an LDA can arise in practice – e.g., as a result of suboptimality that arises when training models without directly optimizing for performance or fairness [see e.g., 46], or as a result of model shifts and distributional shifts that arise in lending applications. Our audits show that the reduction in disparity changes across datasets, model classes, and deployment regimes. In general, the reduction that we can hope to achieve using the least discriminatory model is capped by the disparity of the baseline model – see e.g., income in the LabelShift where the baseline models achieve disparities of 0.9%. Beyond this, the relationships may be unpredictable and counterintuitive. In the fico dataset, for example, an audit of the LR in the Standard deployment regime, the least discriminatory model reduces disparity by 10.7%. In a LabelShift regime, however, the least discriminatory model may only reduce disparity by 4.4%. In practice, such results arise over datasets and model classes, highlighting the need to audit consistently.

On Searching across Model Classes. Although our method is designed to search over the class of linear models, our results highlight its ability to return valuable information when evaluating baseline models from other model classes. Considering the results for RF in Table 3, we see that we can discover a linear LDA that achieves meaningful reductions in disparity without compromising performance on the fico and income datasets. In the german dataset, however, these models can only achieve comparable reductions in disparity by alternating their operating point. In this case, our results for RF reflect instances where we have set the slack parameter in the business necessity constraints to small values to ensure feasibility. Given these results, we could claim that an LDA exists for fico and income (even without slack), and refute its existence for german. In general, such results should not be surprising – as many methods will return models that perform well on tabular datasets [see e.g., and the results for RF in Table 3 83].

On Computation. Our method returns certifiably optimal models in 14 of 18 audits often within minutes. In general, we can improve this behavior by initializing the search with a feasible model (e.g., the baseline model or linear models trained on the auditing dataset). In cases where we fail to find a certifiably optimal solution, we can use the outputs from the model to inform our audit. For example, in fico, Standard regime and LR, where in allotted time we find an LDA model that can reduce the disparity from 13.9% to 3.3%. In practice, this model may be sufficient to support a claim of discovery. In settings where the resulting reduction in disparity is insufficient,

		Standard		ModelShift		LabelShift	
Dataset	**Metrics**	LR	RF	LR	RF	LR	RF
`fico` $n = 10,459$ $d = 147$ $\mathcal{G}$ = `thin-file` FICO [32]	Gain(f_{LDA})	**10.6%**	**1.4%**	**9.7%**	**2.4%**	**4.4%**	**0.9%**
	LDA Disparity	3.3%	0.0%	0.0%	0.0%	9.5%	0.0%
	Baseline Disparity	13.9%	1.4%	9.7%	2.4%	13.9%	1.0%
	LDA FNR/FPR	50.9%/8.6%	66.1%/3.2%	66.1%/3.3%	85.5%/0.8%	50.6%/7.2%	64.7%/4.3%
	Baseline FNR/FPR	52.3%/9.6%	70.9%/4.5%	71.6%/4.0%	87.2%/1.6%	52.3%/9.6%	71.7%/4.5%
`german` $n = 1,000$ $d = 56$ $\mathcal{G}$ = `sex` Dua and Graff [29]	Gain(f_{LDA})	**12.3%**	**10.2%**	**15.0%**	**4.7%**	**8.3%**	**9.8%**
	LDA Disparity	0.0%	0.0%	0.0%	0.0%	4.0%	0.0%
	Baseline Disparity	12.3%	10.1%	15.0%	4.7%	12.3%	9.8%
	LDA FNR/FPR	0.0%/28.8%	0.0%/49.2%	0.0%/22.0%	0.0%/39.0%	2.1%/30.5%	0.0%/49.2%
	Baseline FNR/FPR	36.2%/30.5%	17.7%/47.5%	51.8%/22.0%	28.4%/37.3%	36.2%/30.5%	15.7%/47.5%
`income` $n = 32,561$ $d = 30$ $\mathcal{G}$ = sex Kohavi [44]	Gain(f_{LDA})	**0.9%**	**0.7%**	**2.5%**	**0.7%**	**0.1%**	**0.8%**
	LDA Disparity	0.0%	0.0%	0.0%	0.0%	0.8%	0.0%
	Baseline Disparity	0.9%	0.7%	2.5%	0.7%	0.9%	0.8%
	LDA FNR/FPR	81.8%/1.1%	83.1%/1.0%	90.3%/0.4%	89.0%/0.4%	81.1%/1.2%	80.8%/1.6%
	Baseline FNR/FPR	84.0%/1.2%	85.0%/1.2%	90.7%/0.4%	89.2%/0.7%	84.0%/1.2%	84.8%/1.2%

Table 3: Overview of LDA audits for all datasets, baseline model classes, and deployment regimes. We search for an LDA using the worst-case disparity metric and measure group disparity in terms of the FNR Gap. We report the following measures for each audit: Gain($f_{\text{LDA}}; \mathcal{D}^{\text{LDA}}$); the reduction in worst-case disparity between the LDA f_{LDA} and f_0; the aggregate disparity of the LDA model f_{LDA}; the aggregate disparity of the baseline model f_0; and the FNR/FPR of the LDA model f_{LDA} and the baseline model f_0. These values are on the audit dataset. We show analogous results on test data in generalization in (4).

		Standard		ModelShift		LabelShift	
Dataset	**Metrics**	LR	RF	LR	RF	LR	RF
`fico` $n = 10,459$ $d = 147$ $\mathcal{G}$ = `thin-file` FICO [32]	Gain(f_{LDA})	**7.8%**	**0.2%**	**5.8%**	**1.9%**	**3.6%**	**3.6%**
	LDA Disparity	2.3%	3.9%	1.3%	0.2%	6.6%	0.1%
	Baseline Disparity	10.2%	4.1%	7.1%	2.1%	10.2%	3.7%
	LDA FNR/FPR	52.1%/12.6%	66.1%/10.4%	66.6%/8.2%	84.5%/4.2%	51.6%/12.5%	65.4%/9.8%
	Baseline FNR/FPR	51.2%/11.6%	69.9%/6.0%	72.5%/3.9%	86.8%/1.7%	51.2%/11.6%	69.3%/6.0%
`german` $n = 1,000$ $d = 56$ $\mathcal{G}$ = `sex` Dua and Graff [29]	Gain(f_{LDA})	**2.2%**	**6.5%**	**-0.7%**	**1.1%**	**5.7%**	**-2.9%**
	LDA Disparity	4.7%	1.7%	3.3%	4.1%	1.3%	13.4%
	Baseline Disparity	6.9%	8.2%	2.6%	5.3%	6.9%	10.5%
	LDA FNR/FPR	20.9%/62.3%	18.7%/72.1%	27.3%/62.3%	12.9%/63.9%	21.6%/67.2%	18.3%/72.1%
	Baseline FNR/FPR	42.4%/34.4%	20.9%/57.4%	56.8%/21.3%	23.7%/45.9%	42.4%/34.4%	19.0%/57.4%
`income` $n = 32,561$ $d = 30$ $\mathcal{G}$ = sex Kohavi [44]	Gain(f_{LDA})	**0.4%**	**1.1%**	**-0.3%**	**1.4%**	**4.6%**	**0.7%**
	LDA Disparity	4.3%	1.8%	1.6%	2.3%	0.2%	3.1%
	Baseline Disparity	4.7%	2.9%	1.3%	3.6%	4.8%	3.8%
	LDA FNR/FPR	82.8%/1.1%	84.6%/1.2%	90.2%/0.6%	89.2%/0.6%	83.4%/1.4%	81.5%/1.7%
	Baseline FNR/FPR	84.1%/1.2%	84.9%/1.2%	90.9%/0.4%	88.1%/0.7%	84.1%/1.2%	84.6%/1.2%

Table 4: Overview of LDA audits for all datasets, baseline model classes, and deployment regimes on the test dataset. This table should be used to evaluate the generalization of results from Table 3.

a lower bound may be used to refute the existence of an LDA or to continue the search.

6 CONCLUDING REMARKS

The legal mandate to demonstrate the absence of an LDA to avoid liability under the disparate impact doctrine has been in place for decades. Traditionally, this inquiry has been conducted in an ad hoc manner, lacking standardized and formal guidelines, with analyses of alternatives heavily reliant on heuristic-based human intuition. With regulatory bodies increasingly mandating that lenders proactively seek LDAs to comply with fair lending laws [59], it becomes imperative to provide methods and best practices to support their efforts.

Our work seeks to operationalize the search for LDA in a way that promotes transparency – by specifying measures of disparity and business interest and returning information that minimizes reliance on human discretion. The key technical contribution of our approach is that it allows for both *discovery*, when we can demonstrate the existence of an LDA, and *refutation*, when the problem is infeasible and we have shown that there is no linear model that can reduce discrimination on a population of interest.

On the Value and Interpretation of Discovery and Refutation. Our procedure can support the legal implementation of the LDA search, both for lenders and external auditors, like regulators and private plaintiffs. Using our tools, lenders could refute an LDA within a specific class of models to potentially demonstrate to regulators that disparities are unavoidable, thereby meeting the lender's LDA burden and directing stakeholders to explore other alternatives that may reduce discrimination.

In both cases, these outcomes should be viewed as evidence to guide efforts to reduce discrimination. For example, refutation does not mean the absolute non-existence of an LDA – alternatives may exist, for example, by searching for alternatives over non-linear models or by training models to use additional features or predict alternative outcomes. Likewise, the discovery of a less discriminatory model should not viewed as a model that a lender should adopt – but rather as evidence that the lender should revisit the model development and build a model accurate model for the population at hand.

On the Benefits and Limitations of Problem Specification. One of the key benefits of operationalizing an LDA search is that it requires stakeholders to explicitly specify key measures to evaluate discrimination and business necessity – from measures of performance, disparity, and protected groups [53]. Without a formal specification of disparate impact and business necessity, we are unable to discover or refute the existence of less discriminatory alternatives.

Our framework provides stakeholders with substantial flexibility to choose the elements of this specification – allowing them to search for alternative models that minimize disparity with respect to different metrics and across different protected groups, or models that reduce discrimination by relaxing business necessity constraints. Even as this versatility is valuable to promote adoption across use cases, it may lead to misleading results or facilitate manipulation [e.g., by "fairwashing" where for example, a malicious lender could claim the lack of an LDA using specific disparity metrics 4, 5]

Such degrees of freedom in choosing a specification point to gaps in policy and legislation. Overall, we see a need for more specific guidance from financial regulators such as the Consumer Financial Protection Bureau in choosing suitable measures of disparity for a specific audit [7, 55]. Likewise, we see a need for clarity on the legal front with regards to the degree to which the business necessity constraints can be relaxed for an LDA. [see e.g., discussions on whether to require the adoption of an LDA even when it is not "equally effective" in 72].

In closing, we caution that the search for alternative models to support fair lending will exhibit many of the limitations that stem from imposing group fairness – ranging from the potential to arbitrarily alter individual loan decisions [23, 47] to their potential to exhibit detrimental feedback loops time [2, 22, 27, 84]. Given these limitations, an LDA audit should be paired in conjunction with other individual protections such as a right to recourse [45, 73, 74].

Future Work. Our proposed method to search for the least discriminatory model over linear models could be applied to search over a salient model class that can be encoded in combinatorial optimization problems, such as decision trees [3, 39] or rule lists [5]. Another promising approach to improve the discovery of alternative models is to develop a wrapper algorithm that repeatedly solves the LDA problem and generates interactions to improve feasibility. Such an algorithm would broadly allow for the discovery of LDAs in any classification dataset – but require a technique to propose feature interactions to limit computation and avoid overfitting.

In closing, our work restricts the scope of an LDA search to changes in the training procedure – assuming that the dataset and prediction task are fixed. In general, however, an aternative model could refer to any alternative model that can be produced by altering components in a machine learning pipeline, including changes in features, outcome, model class, or model training procedure [16]. In turn, an LDA search may require a broader search over other elements of the pipeline and consideration of alternatives to algorithmic decision-making altogether.

REFERENCES

[1] [n. d.]. Fair Play Fairness Optimizer. https://fairplay.ai/fairness-tools/#fairness-optimizer. Accessed: 2024-01-21.

[2] 2021. Fairness in learning-based sequential decision algorithms: A survey. In *Handbook of Reinforcement Learning and Control*. Springer, 525–555.

[3] Sina Aghaei, Andrés Gómez, and Phebe Vayanos. 2021. Strong optimal classification trees. *arXiv preprint arXiv:2103.15965* (2021).

[4] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*. PMLR, 161–170.

[5] Ulrich Aïvodji, Julien Ferry, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. 2022. Leveraging integer linear programming to learn optimal fair rule lists. In *International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research*. Springer, 103–119.

[6] Ifeoma Ajunwa, Sorelle Friedler, Carlos E Scheidegger, and Suresh Venkatasubramanian. 2016. Hiring by algorithm: predicting and preventing disparate impact. *Available at SSRN* (2016).

[7] Michael Akinwumi, John Merrill, Lisa Rice, Kareem Saleh, and Maureen Yap. 2021. An AI fair lending policy agenda for the federal financial regulators. https://www.brookings.edu/articles/an-ai-fair-lending-policy-agenda-for-the-federal-financial-regulators/.

[8] Gabriele Spina Alì and Ronald Yu. 2021. Artificial intelligence between transparency and secrecy: from the EC whitepaper to the AIA and beyond. *European*

Journal of Law and Technology 12, 3 (2021).

[9] Michael G Allen, Jamie L Crook, and John P Relman. 2014. Assessing HUD's disparate impact rule: A practitioner's perspective. *Harv. CR-CLL Rev.* 49 (2014), 155.

[10] McKane Andrus and Sarah Villeneuve. 2022. Demographic-reliant algorithmic fairness: Characterizing the risks of demographic data collection in the pursuit of fairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1709–1721.

[11] Robert B Avery, Kenneth P Brevoort, and Glenn Canner. 2012. Does Credit Scoring Produce a Disparate Impact? *Real Estate Economics* 40 (2012), S65–S114.

[12] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California law review* (2016), 671–732.

[13] Ainhize Barrainkua, Paula Gordaliza, Jose A Lozano, and Novi Quadrianto. 2023. Preserving the Fairness Guarantees of Classifiers in Changing Environments: a Survey. *Comput. Surveys* (2023).

[14] Andrew Bell, Lucius Bynum, Nazarii Drushchak, Tetiana Zakharchenko, Lucas Rosenblatt, and Julia Stoyanovich. 2023. The Possibility of Fairness: Revisiting the Impossibility Theorem in Practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 400–422.

[15] Emily Black, John Logan Koepke, Pauline Kim, Solon Barocas, and Mingwei Hsu. 2024. Less Discriminatory Algorithms. *Available at SSRN* (2024).

[16] Emily Black, Rakshit Naidu, Rayid Ghani, Kit Rodolfa, Daniel Ho, and Hoda Heidari. 2023. Toward Operationalizing Pipeline-aware ML Fairness: A Research Agenda for Developing Practical Guidelines and Tools. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–11.

[17] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 850–863. https://doi.org/10.1145/3531146.3533149

[18] Laura Blattner and Scott Nelson. 2021. How costly is noise? Data and disparities in consumer credit. *arXiv preprint arXiv:2105.07554* (2021).

[19] Marc-Etienne Brunet, Ashton Anderson, and Richard Zemel. 2022. Implications of Model Indeterminacy for Explanations of Automated Decisions. *Advances in Neural Information Processing Systems* 35 (2022), 7810–7823.

[20] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems* 30 (2017).

[21] Spencer Caro and Scott Nelson. 2023. The Arity of Disparity: Updating Disparate Impact for Modern Fair Lending. (December 2023). https://faculty.chicagobooth.edu/-/media/faculty/scott-nelson/research/arityofdisparity.pdf

[22] Jennifer Chien, Margaret Roberts, and Berk Ustun. 2023. Algorithmic Censoring in Dynamic Learning Systems. *arXiv preprint arXiv:2305.09035* (2023).

[23] A Feder Cooper, Solon Barocas, Christopher De Sa, and Siddhartha Sen. 2023. Variance, Self-Consistency, and Arbitrariness in Fair Classification. *arXiv preprint arXiv:2301.11562* (2023).

[24] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. 2021. Characterizing Fairness Over the Set of Good Models Under Selective Labels. *CoRR* abs/2101.00352 (2021). arXiv:2101.00352 https://arxiv.org/abs/2101.00352

[25] Andrew Cotter, Maya Gupta, and Harikrishna Narasimhan. 2019. On making stochastic classifiers deterministic. *Advances in Neural Information Processing Systems* 32 (2019).

[26] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395* (2020).

[27] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 525–534.

[28] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical risk minimization under fairness constraints. *Advances in neural information processing systems* 31 (2018).

[29] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[30] European Commission. 2021. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206. COM(2021) 206 final 2021/0106(COD).

[31] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[32] FICO. 2018. FICO HELOC. https://community.fico.com/s/explainable-machine-learning-challenge

[33] FinRegLab. 2023. Explainability and Fairness in Machine Learning for Credit Underwriting. https://finreglab.org/wp-content/uploads/2023/12/FinRegLab_2023-12-07_Research-Report_Explainability-and-Fairness-in-Machine-Learning-for-Credit-Underwriting_Policy-Analysis.pdf. Accessed: January 20, 2024.

[34] FinRegLab. 2023. Machine Learning & Financial Inclusion: Findings Overview. https://finreglab.org/wp-content/uploads/2023/07/FRL_ML-FindingsOverview_Final.pdf.

[35] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research* 20, Vi (2019).

[36] ANDREAS FUSTER, PAUL GOLDSMITH-PINKHAM, TARUN RAMADORAI, and ANSGAR WALTHER. 2022. Predictably Unequal? The Effects of Machine Learning on Credit Markets. *The Journal of Finance* 77, 1 (2022), 5–47. https://doi.org/10.1111/jofi.13090 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.13090

[37] Talia Gillis. 2024. "Price Discrimination" Discrimination. (May 2024). Working Paper.

[38] Talia B Gillis and Jann L Spiess. 2019. Big data and discrimination. *The University of Chicago Law Review* 86, 2 (2019), 459–488.

[39] Oktay Gunluk, Jayant Kalagnanam, Minhan Li, Matt Menickelly, and Katya Scheinberg. 2016. Optimal generalized decision trees via integer programming. *arXiv preprint arXiv:1612.03225* (2016).

[40] Deborah Hellman. 2020. Measuring algorithmic fairness. *Virginia Law Review* 106, 4 (2020), 811–866.

[41] HUD and OFHEO and DOJ and Treasury and OCC and OTS and FRB and FDIC and FHFB and FTC and NCUA. 1994. Policy Statement on Discrimination in Lending. Federal Register. Issued by various U.S. government agencies.

[42] Christophe Hurlin, Christophe Pérignon, and Sébastien Saurin. 2022. The Fairness of Credit Scoring Models. arXiv:2205.10200 [stat.ML]

[43] IBM Corporation. 2023. IBM ILOG CPLEX Optimization Studio V22.1. https://www.ibm.com/products/ilog-cplex-optimization-studio. Accessed: 2024-01-19.

[44] Ron Kohavi. 1996. Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In *KDD*. 202–207. http://www.aaai.org/Library/KDD/1996/kdd96-033.php

[45] Avni Kothari, Bogdan Kulynych, Tsui-Wei Weng, and Berk Ustun. 2024. Prediction without Preclusion: Recourse Verification with Reachable Sets. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=SCQfYpdoGE

[46] Michael Lohaus, Michael Perrot, and Ulrike Von Luxburg. 2020. Too Relaxed to Be Fair. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 6360–6369. https://proceedings.mlr.press/v119/lohaus20a.html

[47] Carol Xuan Long, Hsiang Hsu, Wael Alghamdi, and Flavio Calmon. 2023. Individual Arbitrariness and Group Fairness. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[48] Peter E Mahoney. 1998. The End (s) of Disparate Impact: Doctrinal Reconstruction, Fair Housing and Lending Law, and the Anti-Discrimination Principle. *Emory LJ* 47 (1998), 409.

[49] Charles Marx, Flavio P. Calmon, and Berk Ustun. 2019. Predictive multiplicity in classification.

[50] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*. PMLR, 107–118.

[51] Vitaly Meursault, Daniel Moulton, Larry Santucci, and Nathan Schor. 2022. The Time Is Now: Advancing Fairness in Lending Through Machine Learning. *FRB of Philadelphia Working Paper* 22, 39 (2022). https://doi.org/10.21799/frbp.wp.2022.39

[52] Atif Mian and Amir Sufi. 2009. The Consequences of Mortgage Credit Expansion: Evidence from the U.S. Mortgage Default Crisis*. *The Quarterly Journal of Economics* 124, 4 (11 2009), 1449–1496. https://doi.org/10.1162/qjec.2009.124.4.1449 arXiv:https://academic.oup.com/qje/article-pdf/124/4/1449/5407278/124-4-1449.pdf

[53] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application* 8 (2021), 141–163.

[54] Mick Mulvaney. 2018. Statement of the Bureau of Consumer Financial Protection on Enactment of S.J. Res. 57. Press Release.

[55] National Community Reinvestment Coalition and Upturn and Zest AI. 2022. CFPB Should Encourage Lenders To Look For Less Discriminatory Models. https://ncrc.org/cfpb-should-encourage-lenders-to-look-for-less-discriminatory-models/. Accessed: 2024-01-21.

[56] OCC, FDIC, FRB, OTS and NCUA. 2009. Interagency Fair Lending Examination Procedures.

[57] Luca Oneto, Michele Donini, and Massimiliano Pontil. 2020. General fair empirical risk minimization. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[58] Richard Pace. 2023. Fool's Gold 3: Do LDA Credit Models Really Improve Fairness? https://www.paceanalyticsllc.com/post/fools-gold-3. Accessed: 2024-01-21.

[59] Practical Law Finance. 2023. CFPB Clarifies Duty to Perform Fairness Testing on Lending Models. https://content.next.westlaw.com/practical-law/document/I7770d7c1da5611ed8921fbef1a541940/CFPB-Clarifies-Duty-to-Perform-Fairness-Testing-on-Lending-Models. Published on Practical Law website, jurisdiction: USA (National/Federal).

[60]]RegulationB Regulation B [n. d.]. 12 C.F.R. § 1002 et seq.

[61] Relman Colfax PLLC. 2021. Fair Lending Monitorship of Upstart Network's Lending Model, Second Report of the Independent Monitor. https://www.relmanlaw.com/cases-406. Accessed: 2024-01-21.

[62] Lesia Semenova, Harry Chen, Ronald Parr, and Cynthia Rudin. 2023. A Path to Simpler Models Starts With Noise. *arXiv preprint arXiv:2310.19726* (2023).

[63] Lesia Semenova, Cynthia Rudin, and Ronald Parr. 2019. A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. *ArXiv* (2019), 1–64. http://arxiv.org/abs/1908.01755

[64] Naeem Siddiqi. 2012. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. Vol. 3. John Wiley & Sons.

[65] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. 2021. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 3–13.

[66] Winnie Taylor. 2011. Proving racial discrimination and monitoring fair lending compliance: the missing data problem in nonmortgage credit. *Rev. Banking & Fin. L.* 31 (2011), 199.

[67] Winnie F Taylor. 1980. Meeting the Equal Credit Opportunity Act's Specificity Requirement: Judgmental and Statistical Scoring Systems. *Buff. L. Rev.* 29 (1980), 73.

[68] Kevin Tobia. 2017. Disparate statistics. *The Yale Law Journal* (2017), 2382–2420.

[69] TransUnion. 2014. CreditVision Income Estimator. https://www.transunion.com/content/dam/transunion/global/business/documents/product-creditvision-income-estimator-as.pdf Product information sheet.

[70] U.S. Congress. 2016. Fair Credit Reporting Act. https://www.ftc.gov/system/files/fcra_2016.pdf

[71] U.S. Department of Housing and Urban Development (HUD). 2013. Implementation of the Fair Housing Act's Discriminatory Effects Standard. Federal Register. , 11460-11461 pages. Final rule.

[72] U.S. Department of Housing and Urban Development (HUD). 2023. Reinstatement of HUD's Discriminatory Effects Standard. Federal Register. , 19450-19500 pages. Final rule, Document Number: 2023-05836.

[73] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*. 10–19.

[74] Suresh Venkatasubramanian and Mark Alfano. 2020. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 284–293.

[75] Hao Wang, Luxi He, Rui Gao, and Flavio P Calmon. 2023. Aleatoric and Epistemic Discrimination in Classification. *arXiv preprint arXiv:2301.11781* (2023).

[76] Hao Wang, Berk Ustun, and Flavio Calmon. 2019. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *International Conference on Machine Learning*. PMLR, 6618–6627.

[77] Jamelle Watson-Daniels, Solon Barocas, Jake M Hofman, and Alexandra Chouldechova. 2023. Multi-Target Multiplicity: Flexibility and Fairness in Target Specification under Resource Constraints. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 297–311.

[78] Jamelle Watson-Daniels, David C Parkes, and Berk Ustun. 2023. Predictive multiplicity in probabilistic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 10306–10314.

[79] Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio P Calmon. 2020. Optimized score transformation for fair classification. *Proceedings of Machine Learning Research* 108 (2020).

[80] Laurence A Wolsey. 1998. *Integer Programming*. Vol. 42. Wiley New York.

[81] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. 2019. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research* 20, 1 (2019), 2737–2778.

[82] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*. PMLR, 962–970.

[83] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180, 3 (2017), 689–722.

[84] Zhaowei Zhu, Tianyi Luo, and Yang Liu. 2021. The rich get richer: Disparate impact of semi-supervised learning. *arXiv preprint arXiv:2110.06282* (2021).