Anti-discrimination Laws, AI, and Gender Bias in Non-mortgage Fintech Lending

Stephanie Kelley

Smith School of Business, Queen's University, Kingston, ON, Canada, K7L 3N6,
stephanie.kelley@queensu.ca

Anton Ovchinnikov

Smith School of Business, Queen's University, Kingston, ON, Canada, K7L 3N6,
anton.ovchinnikova@queensu.ca
INSEAD, Fontainebleau, 77300, France, anton.ovchinnikov@insead.edu

Draft: November 2021

Abstract

Problem definition: We use a realistically large, publicly available dataset from a global fintech lender to simulate the impact of different anti-discrimination laws, and their corresponding data management and model building regimes, on gender-based discrimination in the non-mortgage fintech lending setting.

Academic/Practical Relevance: Our paper extends the conceptual understanding of model-based discrimination from computer science to a realistic context that simulates the situations faced by fintech lenders in practice, where advanced machine learning (ML) techniques are used with high-dimensional, feature rich, highly multi-collinear data. We provide technically and legally permissible approaches for firms to reduce discrimination across different anti-discrimination regimes, whilst managing profitability.

Methodology: We train statistical and ML models on a large and realistically rich publicly available dataset to simulate different anti-discrimination regimes, and measure their impact on model *discrimination*, *predictive quality*, and *firm profitability*. We use ML explainability techniques to understand the drivers of ML discrimination.

Results: We find that regimes which prohibit the use of gender (like those in the United States) substantially increase discrimination, and slightly decrease firm profitability. We observe ML models are less discriminatory, of better predictive quality, and more profitable compared to traditional statistical models like logistic regression. Unlike omitted variable bias which drives discrimination in statistical models, ML discrimination is driven by changes in the model training procedure, including feature engineering and feature selection, when gender is excluded. We observe that down-sampling the training data to rebalance gender, gender-aware hyperparameter tuning, and up-sampling the training data to rebalance gender, all reduce discrimination, with varying trade-offs in predictive quality and firm profitability. Probabilistic gender proxy modeling (imputing applicant gender) further reduces discrimination, with negligible impact on predictive quality, and a slight increase in firm profitability.

Managerial Implications: A rethink is required of the anti-discrimination laws, specifically with respect to the collection and use of protected attributes for machine learning models. Firms should be able to collect protected attributes to, at minimum, measure discrimination, and ideally, take steps to reduce it. Increased data access should come with greater accountability for firms.

Keywords: discrimination, bias, ethics, law, fintech, artificial intelligence, machine learning, gender

1. Introduction

Algorithms and artificial intelligence (AI) are fundamentally transforming the way organizations make decisions. Their adoption, however, has been accompanied by reports from consumers and the media of discrimination. The reports refer to discrimination as what is ethically problematic, as opposed to what is illegal, essentially a noncomparative wrong whereby an algorithm fails to treat a group of individuals the way they are entitled to be treated (Hellman 2016). As such, in this paper we define and measure discrimination not by what is legal, but rather by how a model treats a group of individuals. While an ethically centered definition is salient, most countries have adopted anti-discrimination laws to increase equality for protected groups, each with their own unique legal definition of discrimination. Rapid advances in AI have however outpaced changes in these laws (Barocas and Selbst 2016), resulting in regulations that may paradoxically hurt, rather than help the groups they are supposed to protect.

For example, consider the Apple Card, which was accused by consumers and the media of discrimination against women (Vigdor 2019): it declined a woman's application for a credit line increase while granting one to her husband, resulting in a 20 times difference between them, even though she "*had a better credit score and other factors in her favor*." In response to the accusation, Goldman Sachs, a partner in the Apple Card venture, stated: "*we have not and never will make decisions based on factors like gender…we do not know your gender or marital status*..." (Franck 2019). This statement is not surprising: the US Equal Credit Opportunity Act (ECOA) prohibits the use (and even collection) of protected attributes, like gender, for non-mortgage lending decisions, and a New York State Department of Financial Services (2021) report indeed found no violations of fair lending by the Apple Card venture. The outcome however is paradoxical: adhering to the existing anti-discrimination laws and the data management and model building practices they imply could produce ethically problematic outcomes.

The anti-discrimination laws with respect to gender and credit differ across countries; while their exact language varies greatly and is not a subject of this investigation, their data management and model building guidance imply three main regimes (discussed in detail in the Supplemental Materials, S1):

1. Regime 1 (e.g., Singapore) allows for the collection and use of gender data in AI models;
2. Regime 2 (e.g., European Union) allows for the collection of gender but prohibits the use of gender as a feature in the training and screening models used for individual lending decisions;
3. Regime 3 (e.g., United States) prohibits the collection, and thus also the use of gender data.

In this paper we use a realistically large, publicly available dataset from a global fintech lender, Home Credit, to simulate the impact of these regimes (and the corresponding implications for data management and model building) on gender discrimination. We examine gender, the characteristics of women and men that are socially constructed, as opposed to sex, the characteristics that are biologically

determined (WHO/Europe | Gender: definitions), given that the majority of global anti-discrimination efforts focus on "gender equality". We use the terms "women" and "men" throughout the paper, which are consistent with our available data, gender. Further, we focus on gender, and not race or other attributes due to its universality; our choices do not in any way diminish the need to investigate discrimination in other contexts and with other datasets.

Our study follows how a consumer lending fintech firm makes loan accept/reject decisions; see Figure 1 for a schematic. First, a lender uses data about past *borrowers* to train a model that predicts whether a customer will repay or default if given a loan, a task generally referred to as binary classification (Henley and Hand 1997). Since default is uncertain, the model predicts a numeric score, which can be intuitively interpreted as the predicted probability of default. Note that the predicted default scores from different model classes (e.g., logistic regression (LR) vs. tree-based ensembles) may not necessarily be calibrated (i.e., a predicted score of 0.3 from one model may not be equivalent to a 0.3 from a different model), and therefore are not directly comparable. Second, the firm uses the trained model (Kleinberg et al. 2020 refer to it as a "screening model") to predict defaults for the new *applicants*, resulting in a set of applicant default probability predictions. Third, the predicted probabilities are compared to a classification threshold (τ) to arrive at a binary default classification: applicants are rejected if the predicted probability is above the threshold, and issued credit otherwise (Lessmann et al. 2015). This threshold is optimized given the economics of the loan, i.e., the cost of default and revenue from repayment.

Ideally, a firm would measure discrimination by comparing the probability that the screening model rejects someone who will not pay back the loan, across groups, but because a firm cannot possibly know the payment outcomes for a rejected applicant, they must use an alternative measure using available data. We use two such measures of discrimination, positive predictive value, PPV (Chouldechova 2017), which measures a model's ability to correctly predict outcomes for one group, conditional on the known outcome, compared to another group; and within group mean difference, WMGD (Žliobaitė 2017), the difference in the mean predicted default rate for the protected class, between models. See §3.1.

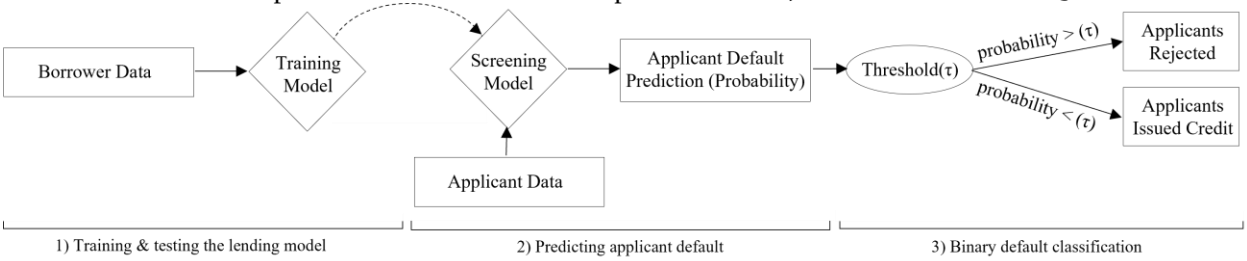


Fig. 1. Schematic view of the 3-step fintech lending modeling process

A key observation that underlies our investigation is that under Regime 1, a firm's training and screening models have access to gender (and the numerous machine learning (ML) consequences of that: feature engineering and hyperparameter tuning, etc., see §1.1.2 and §6.2) while under Regime 2, the use of gender is restricted, and under Regime 3, prohibited entirely. This leads to two model types, corresponding to the different regimes: Model 1 (with gender, per Regime 1) and Model 2 (without gender, per Regimes 2 and 3). We specify what the "model" is in §3.1, but, importantly, Model 1 differs from Model 2: their applicant default probability predictions vary, and consequently, their optimal thresholds differ as well. Hence, the exact same applicant could be issued credit under one regime but rejected under another. We study whether this affects men and women differently, dissect why the differences occur, and show what firms under different regimes can do to reduce the discrimination, while quantifying the associated impact on profitability.

1.1. Summary of findings

The insights from our paper can be classified into three categories: 1) impact of anti-discrimination regimes on gender-based discrimination, 2) drivers of statistical and machine learning discrimination, and 3) possible approaches to reduce machine learning discrimination.

1.1.1. Impact of anti-discrimination regimes on gender-based discrimination

In §5 we find that Regimes 2 and 3, which force the exclusion of gender in the screening models (i.e., the firm must use Model 2) do not significantly impact predictive quality, measured by area under the curve (AUC). Despite not impacting predictive quality, the exclusion of gender negatively impacts firm profitability, which is on average 0.25% lower for the model without gender. Most shockingly, the gender exclusion leads to on average a 285.04% increase in gender discrimination (measured by positive predictive value, PPV, §3.1) in the top-performing ML model ("Average Blender" discussed in §4.3.2), trained on our data.

The paradoxical discriminatory effects of anti-discrimination regimes have been investigated before in both computer science and financial economics. In computer science, most notably, Kleinberg et al. (2018) use a conceptual framework and a regression-based empirical example to show that algorithmic decision-makers should prefer a model that includes protected attributes, such as race or gender, given they are useful for predicting the outcome. Other works in computer science have used a combination of conceptual frameworks and small-scale empirical examples (<10 features, low multi-collinear datasets) to show that absent legal constraints, protected attributes should be included to reduce discrimination and improve predictive quality (Lipton et al. 2018, Žliobaitė and Custers 2016). We support this conceptual conclusion, with results from a realistic dataset and modeling process that mimics a fintech's operations.

In the financial economics literature: Chandler and Ewert (Elliehausen and Durkin 1989) evaluate the ECOA, and find the operational modeling guidance, which prohibits the use and collection of gender, creates a detrimental increase in the rejection rates of women, compared to regression models that use gender. Andreeva and Matuszyk (2019) use classical statistical techniques and find that the EU Gender Directive, which prohibits the use of gender as a feature in the training and screening models, leads to a greater increase in rejection rates for women compared to men, versus models which include gender. §5 extends these analyses into the modern ML setting, exploring multiple nuances stemming from the use of ML methods. For robustness, we also replicate the statistical analyses in the Supplemental Materials (S2).

1.1.2. Drivers of statistical and machine learning discrimination

What drives discrimination, and, further, what drives the differences in discrimination between statistical and ML models? One intuitive explanation from the traditional statistics and econometrics literature is omitted variable bias (OVB) (Wooldridge 2015); indeed, Regimes 2 and 3 remove gender from the variable set, changing the model coefficients for the remaining variables. Andreeva and Matuszyk (2019) use a traditional statistical modeling approach (discussed in §4.3.1) and empirically show that the data collection and modeling guidance of Regimes 2 and 3 indeed create OVB: when trained on data with women as the minority, as is common in lending, the exclusion of gender leads to coefficient estimates dominated by men, the less creditworthy group, which disproportionately increases the rejection rates of women. Kleinberg et al. (2018) and Žliobaitė and Custers (2016) similarly show using a conceptual modeling framework that Regime 2 and 3 regulations, which prohibit the use of protected attributes, create OVB if those features have explanatory power, leading to discrimination.

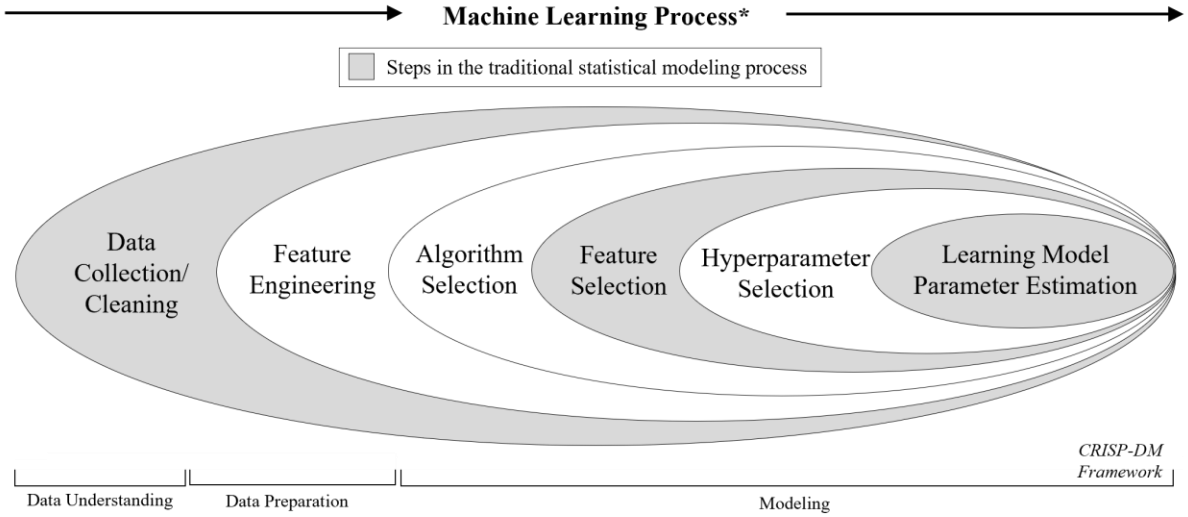
This traditional statistics view of OVB makes several simplifying assumptions which are not true for a fintech that uses modern ML. Kleinberg et al. (2020) extends beyond OVB and suggests that algorithmic bias can come from three places in the modern ML process: the choice of outcome measure, the choice of input variables, and the construction of the model training procedure. We show that the modern ML used by fintechs introduces differences in the construction of the model training procedure, suggesting that OVB may not the main driver of discrimination in the modern ML setting.

1. **Choice of outcome measure:** the choice of default probability as the outcome measure is well-founded in the lending space (Henley and Hand 1997) and does not change in ML; therefore the use of modern ML should not introduce additional bias through the choice of outcome measure.
2. **Choice of input variables:** the various anti-discrimination regimes do not differ in their guidance for traditional statistical and machine learning models, so the choice of input variables does not change between the two modeling processes: gender is either included as a feature in the training and screening models (per Regime 1) or excluded (per Regimes 2 and 3).

3. **Construction of the model training procedure:** recall from Fig. 1 that borrower data is used to train the training model; a process which can be summarized at a more granular level by CRISP-DM (the CRoss Industry Standard Process for Data Mining) (Wirth and Hipp 2000), the most commonly used model training procedure in modern ML, used by several fintech industry partners. Fig. 2. visualizes the details of CRISP-DM and the additional modeling steps introduced into the training procedure (white in Fig. 2). The removal of gender, per Regimes 2 and 3, affects each of these additional steps, the impact of which is not captured by OVB, specifically:
    a. **Feature engineering**: having access to the gender variable allows for the creation of new variables or features including interactions (e.g., "gender * income") and binning (e.g., "=IF(age>65,1,0)"). In §6.2.2, permutation importance analysis shows that ~20% of the most impactful features in Model 1 (with gender) are engineered using gender; they are not even considered by Model 2 (without gender), and hence cannot be "omitted" by definition.
    b. **Algorithm selection:** having access to gender could change the top-performing algorithm. For example, a regularized logistic regression may outperform a random forest model on data without gender (and the resultant engineered features) but may perform worse with gender.
    c. **Feature selection**: access to gender in model training can change the set of features that are selected to be "in the model." For example, with access to gender, the model may select "age" during feature selection, but when gender is excluded, it may exclude "age." In §6.2.2 we observe that when gender is excluded the algorithm also excludes certain features that we refer to as *gender-reliant*, and selects others in their place, which we refer to as *gender-redundant*. Using SHAP values and SHAP interaction values (Lundberg et al. 2019, Lundberg and Lee 2017), we find that the gender-reliant features are on average 19 times more important for women compared to men, so when gender is excludes, so too are the gender-reliant features, increasing discrimination against women.
    d. **Hyperparameter selection**: many ML models, such as random forest, have numerous parameters that guide learning, rather than are learned from data directly. For example, the number of trees, or the size of each tree; these are called "hyperparameters." Having access to gender and the resultant engineered features can result in a different set of hyperparameters, even if the algorithm itself is the same. Our analyses in §7 shows that selecting the hyperparameters when gender is available can change model predictions, ultimately reducing discrimination, even if gender is excluded in the learning model parameter estimation.

Incorporating the differences in the construction of the model training procedure, when gender is included (Model 1) our top performing "Average Blender" ML model, hereinafter referred to as AB, is

44.06% less discriminatory on average (measured by PPV), of significantly better predictive quality (+472 bps AUC), and on average 7.86% more profitable than the traditional statistical logistic regression model. When gender is excluded (Model 2), the AB model remains less discriminatory (on average by 9.56%, per PPV) across thresholds, with greater predictive quality (+487 bps AUC), and is on average 7.60% more profitable compared to the LR model. This illustrates that both firms *and* applicants should prefer ML models over traditional statistical models in the non-mortgage consumer fintech lending setting, as ML allows the model to partially recover the negative impact of excluding gender.



*adapted from Google's highly referenced 2017 YouTube video on the steps of machine learning (https://www.youtube.com/watch?v=nKW8Ndu7Mjw)

Fig. 2. The 6-step machine learning modeling process compared to the 3-step traditional statistical modeling process, arranged according to the CRISP-DM Framework

1.1.3. Possible approaches to reduce machine learning discrimination

In §7 we evaluate several possible approaches for firms to reduce the gender discrimination given the restrictions on the use and collection of gender, imposed under various regimes:

1. **Down-sampling the training data to rebalance gender** is a form of pre-processing (Kamiran and Calders 2012): observations are randomly removed from the majority class (men) until counts are equal with the minority group (women). It is feasible under Regimes 1 and 2 and results on average in -4.54% discrimination (PPV), -175 bps predictive quality, and -4.47% average profitability in our data.
2. **Gender-aware hyperparameter tuning** is an approach where model hyperparameters of the training model are tuned on data with gender of past borrowers, but applicant gender is not used in the screening model. It is model agnostic, and similar to the fair Bayesian optimization technique (Perrone et al. 2020). It is feasible for firms under Regimes 1 and 2, and results on average in -37.73% discrimination (PPV), -278 bps predictive quality, and -4.42% average profitability in our data.

3. **Up-sampling the training data to rebalance gender** is an approach that involves collecting additional observations from the minority class (women) to match the count of the majority class (men); see Chen et al. (2018). As gender must be collected, this approach is feasible under Regimes 1 and 2 and results on average in -24.47% discrimination (PPV), no significant change to predictive quality, -1.46 % average profitability in our data.
4. **Probabilistic gender proxy, PGP, modeling** is an approach that is feasible for firms operating across multiple jurisdictions: a training model is created (using data from a Regime 1 or 2 country) to predict gender for applicants in a Regime 3 jurisdiction (where gender is not permitted to be collected), and this gender prediction is used as a feature in the screening model; see Zhang (2018) and Chen et al. (2019) for PGP used to predict race/ethnicity in lending. Although quite effective in our data (on average -71.09% discrimination (PPV), no significant change to predictive quality, +0.13% average profitability) this approach is prohibited in the US, an example of Regime 3 (Chen et al. 2019).

We reviewed several other discrimination-reducing approaches, including generating gender-specific models and using gender-specific thresholds (Lipton et al. 2018); however, these approaches use gender in both the training and screening models, and treat the two genders differently, in direct contradiction with Regime 2 and 3 guidance; we therefore excluded them from consideration.

2. Related literature

Our work is related to the study of discrimination in three areas: operations management, financial economics, and computer science.

2.1. Discrimination in technology-based business operations

First, our work is related the empirical study of technology-based business operations (e.g., Cohen and Harsha 2020, Cui et al. 2018), and the discrimination they proliferate in: crowdfunding (Pope and Sydnor 2011b, a, Younkin and Kuppuswamy 2018), online auctions (Doleac and Stein 2013), social networks (Acquisti and Fong 2020), ride sharing (Ge et al. 2016, Mejia and Parker 2021), online labor markets (Chan and Wang 2018), online advertising (Lambrecht and Tucker 2019), online vacation rental marketplaces (Cui et al. 2020), and healthcare treatment (Obermeyer et al. 2019).

Of particular relevance for our study are the investigations of discrimination in the use of machine learning algorithms. Lambrecht and Tucker (2019) find that online advertising algorithms lead to automated gender bias because of the higher economic valuation assigned to the views of women. Obermeyer et al. (2019) examine a commercial healthcare prediction algorithm and find that it proliferates racial bias due to biased training data. Our work investigates the drivers of machine learning discrimination in a new operational setting, fintech lending, and we contribute to the literature by: 1)

studying the model training procedures and resulting implications for firms, as opposed to only the model outputs; 2) exploring techniques to reduce the bias, which are said to be overlooked (Mejia and Parker 2021); and 3) studying the impact of these discrimination reducing techniques on firm profitability. Lastly, we believe it is important to bring forward the topics of gender equality and discrimination for operations management researchers because they are aligned with the Vision of the Responsible Research in Business and Management (RRBM 2017), and the United Nation's Sustainable Development Goals; see Supplemental Materials (S1).

2.2. Discrimination in non-mortgage consumer lending

Second, our work is related to the financial economics literature on discrimination in consumer lending. The vast majority of this empirical literature considers mortgage lending (Bartlett et al. 2019, Fuster et al. 2018) due to data availability (Taylor 2011) which differs from our non-mortgage context in three ways. First, the lenders' operating models are different: most mortgage fintech firms are intermediaries who connect borrowers and lenders by structuring the loan applications and leaving them on the platform to be funded by individual or institutional lenders; they do not make the loan accept/reject decisions. Second, lenders in several major markets who make such decisions use variations of the Fair Isaac score (FICO) in logistic regressions, where discrimination is driven by OVB. Third, the collection of gender data is not prohibited for mortgage lenders in most jurisdictions.

For these reasons, most of the existing studies of lending discrimination are not directly relevant to our work, with, to the best of our knowledge, only two studies that are similar to ours: Chandler and Ewert (Elliehausen and Durkin 1989) evaluate the impact of the ECOA in 1979, and Andreeva and Matuszyk (2019) evaluate the impact of the EU Gender Directive in the 2000s. They both find that the operational modeling guidance of the laws, which restrict the use of gender in the training and screening models, creates a detrimental increase in the rejection rates of women, compared to models that use gender. Both these works use private, NDA-protected datasets, preventing investigation or replication of their results. They focus on outdated statistical regression models in a single legal jurisdiction, do not produce a formal measure of discrimination, and do not measure the impact on firm profitability, or provide recommendations for firms to reduce discrimination. Our unique public data, and modern ML approach addresses these shortcomings, making the findings more operationally relevant for fintech firms, regulators, and the public across several regulatory regimes. We find similar qualitative findings as these papers (see Supplemental Materials, S2), then extend the analysis to the ML setting (§5, §6).

2.3. Fairness in machine learning

Third, our work is related to the computer science study of fairness in machine learning. A handful of works have investigated the impact of excluding protected attributes, like gender, on

discrimination. Lipton et al (2019) and Kleinberg et al (2018) explore the impact of US anti-discrimination laws and conclude that absent legal constraints, a protected attribute should be included to decrease discrimination and improve model accuracy. Žliobaitė and Custers (2016) perform a comparable investigation, in the context of EU anti-discrimination laws, and arrive at a similar conclusion. Like Kleinberg et al (2018), they conceptually explain the drivers of algorithmic discrimination using the OVB framework. While this arm of the literature succinctly points out the discriminatory effect of excluding protected attributes, these studies lack domain-specific, realistic operational details. For instance, Žliobaitė and Custers (2016) use a small-scale salary dataset with 52 observations and 6 variables; while they admit it is small, it is difficult to extend their OVB findings to a true operational high-dimensional, feature rich, and highly multi-collinear dataset used by fintechs to train ML models. In contrast, our study is operationally grounded: the data, process, and models are selected to simulate those used by fintech lenders, which allowed us to uncover the *gender-blind feature selection* phenomenon; this mechanism has not been discussed in the more generalized computer science investigations. Further, we are the first to provide an aggregated analysis across regimes.

Kleinberg et al. (2020) extends beyond traditional statistical models and OVB to suggest that algorithmic bias can stem from three aspects of the ML modeling process: the choice of outcome measure, the choice of input variables, and the construction of the model training procedure. We extend this conceptual framework and empirically measure the algorithmic bias introduced through a change in the construction of the model training procedure (from LR to ML). Note the choice of outcome measure (default probability), and input variables (specifically, with respect to gender) do not change.

Further, while several other works suggest a range of solutions to reduce discrimination through pre-processing (e.g., Kamiran and Calders 2012, Chen et al. 2018, Chen et al. 2019), in-processing (e.g., Perrone et al. 2020, Zafar et al. 2017), and post-processing (e.g., Hardt et al. 2016), practitioners reported "*struggling to apply existing auditing and de-biasing methods in their contexts*," and found there were limited "*domain-specific education resources, metrics, processes, and tools…*", as the majority of computer science studies focus on non-business contexts, such as recidivism (Holstein et al. 2019). Our business-oriented and operationally relevant approach, again, directly addresses these concerns.

Finally, the theoretical progress in this literature provides several useful concepts that we utilize in our work, such as the fairness-accuracy trade-off (see Žliobaitė 2015 for a summary discussion) and several measures of discrimination (Berk et al. 2017, Chouldechova 2017, Žliobaitė 2017), see §3.1.

Summarizing our results vis-à-vis the existing literature, we build on a large body of work that established a conceptual understanding of model-based discrimination (Kleinberg et al. 2018, 2020, Lipton et al. 2018, Žliobaitė and Custers 2016). We extend this understanding to a realistic context that

mimics situations faced by fintech lenders in practice, where advanced non-regression techniques are used with high-dimensional, feature rich, highly multi-collinear data in conjunction with sophisticated feature engineering. Further, we measure the economic impact on firms, which has not previously been explored. To what extent these practical elements alter the conceptual findings about model-based discrimination is unclear from prior research; our paper presents an investigation that is relevant for firms, consumers, and regulators around the world. In fact, the insights from our work have already impacted the policies and guidelines adopted by multiple financial institutions and regulatory bodies.

3. Key metrics: discrimination, predictive quality, and firm profitability

3.1. Discrimination measure selection

Prior literature proposes three main classes of discrimination measures: classification parity, calibration, and anti-classification (Berk et al. 2017), and it is well-known (Chouldechova 2017, Kleinberg et al. 2017) that multiple measures cannot be simultaneously satisfied unless model accuracy is perfect or base rates are equal across groups, which are unrealistic assumptions. From our knowledge of fintech lending and discussions with industry partners we determined that to be practically relevant for a fintech firm, a discrimination measure should achieve three conditions, in order of importance:

1. adjust for unequal base default rates between protected groups (e.g., a difference in default rates between genders should be preserved as it ensures groups are treated the way they are entitled to be treated, per the ethically centered definition of discrimination from consumers and the media, per §1);
2. be calculated without an external risk score (as there are no such risk scores available: credit scores are not a valid external score as they are used as a feature in the model);
3. be calculated with a known default outcome label (helps to preserve absolute default rates to ensure the model predictions achieve comparable default rates to historical values).

Table 1 provides a summary of measures and highlights two: positive predictive value (PPV) and within group mean difference (WGMD). PPV satisfies all three conditions, and WGMD is the only other measure that adjusts for unequal based default rates (the most important condition) and can be calculated without an external risk score.

One potential weakness of the above conditions is that requiring a known default outcome label (condition #3) means that a lender cannot measure discrimination in the applicant group that was rejected; without the known outcome label, they are faced with the selective, or missing label problem (Lakkaraju et al. 2017). We address the missing labels problem with a sampling procedure called augmentation (Hsia 1978) as discussed in §4.2, but given that WGMD does not require a known default outcome label it

provides a useful alternative measure of discrimination that avoids the selective labels problem. We discuss both PPV and WGMD below.

Discrimination Measures	Reference	Adjusts for unequal base default rates	Calculated without external risk score	Calculated with a known default outcome label
Classification Parity				
Statistical Parity	Berk et al. (2017)		X	X
Equalized Odds	Hardt et al. (2016)		X	X
Treatment Equality	Berk et al. (2017)		X	X
Balance for the Positive Class	Kleinberg et al. (2017)		X	X
Positive Predictive Value	Choudlechova (2017)	X	X	X
Mean Difference	Žliobaitė (2017)		X	
Within Group Mean Difference	Adapted from Žliobaitė (2017)	X	X	
Calibration	Corbett-Davies and Goel (2018)	X		
Anti-classification	Grgic-Hlaca et al. (2016)		X	

Table 1. Summary of the discrimination measure selection rationale

PPV (Chouldechova 2017) represents the difference in the model's ability to correctly predict default, conditional on actual default between men and women. Given a classification threshold, τ, discrimination measured by $PPV(\tau)$ is the number of true positive predictions (i.e., correctly predicted defaults for men [M]), $TP_M(\tau)$, divided by all default predictions for men (i.e., the sum of $TP_M(\tau)$ and the number of false positive predictions for men, $FP_M(\tau),$), minus the same ratio for women (W):

$$PPV(\tau) = \frac{TP_M(\tau)}{[TP_M(\tau)+FP_M(\tau)]} - \frac{TP_W(\tau)}{[TP_W(\tau)+FP_W(\tau)]} \quad (1)$$

A $PPV(\tau)$ greater than 0 denotes bias against women, less than 0, bias against men, and equal to 0 indicates no discrimination. The use of $TP_M(\tau)$ and $FP_M(\tau)$ values (e.g., for men) in the mathematical definition aligns with our theoretical definition of discrimination as a non-comparative wrong: a failure to treat (predict) a group of individuals (one gender) the way they are entitled to be treated (predicted correctly) (Hellman 2016). For brevity, we refer to $PPV(\tau)$ as simply PPV throughout the paper.

WGMD (Žliobaitė 2017) measures the difference in the mean predicted default rate for the protected class, women, between models. That is, if $\hat{Y}_{iW}(\tau) = 1$ denotes a default prediction for woman i in our dataset (as made by some model at some threshold τ), and N_W denotes the total number of women in the data, then:

$$WGMD(\tau) = \left[\frac{\sum_{i=1}^{n_W} \hat{Y}_{iW}(\tau)=1}{N_W}\right]_{Model\ A} - \left[\frac{\sum_{i=1}^{n_W} \hat{Y}_{iW}(\tau)=1}{N_W}\right]_{Model\ B} \quad (2)$$

A $WGMD(\tau)$ value greater than zero denotes an increase in discrimination against the protected class in model A versus model B. A value less than zero denotes a decrease in discrimination, and a value

of zero indicates no change in discrimination between the models. For brevity, we refer to $WGMD(\tau)$ as simply WGMD throughout the paper. Note that a known default outcome is not required to calculate the positive default predictions, increasing the generalizability of the measure by avoiding the selective labels problem.

Throughout the paper we compare discrimination between models using PPV as our main measure as it achieves all three conditions (per Table 1), and WGMD for robustness, and report the mean (the mean of the values across the 30-fold cross validation) and 95% confidence intervals (CIs, the upper and lower bounds are calculated using the sample mean and sample standard deviation across the 30-fold cross validation). Both values were calculated in R using the `ci` function from the `gmodels` package. For additional robustness, we test the significance of the differences of the discrimination from each model across a range of thresholds (5-30%). We use the Shapiro-Wilk normality test to determine if the differences are statistically significantly different from the normal distribution, then proceed with a paired t-test if the differences are normally distributed, and a paired samples Wilcoxon test if they are not. An alpha of 0.05 was used for all tests.

3.2. Model predictive quality measure

We measure *model predictive quality* using area under the curve (AUC), a percentage, calculated using predicted and known outcomes, with higher numbers denoting better quality. It is commonly used to measure lending model quality as it performs well with the imbalanced datasets typical in the credit setting (Akkoç 2012, Lessmann et al. 2015). We compare predictive quality between models by reporting AUC and the 95% confidence intervals, computed using the DeLong method (DeLong et al. 1988), with 2000 stratified bootstraps.

3.3. Firm profitability measure

We measure *firm profitability* as the optimal profit across classification thresholds (Akkoç 2012, Lessmann et al. 2015). A firm receives revenue for each applicant they grant credit who does not default (a true negative prediction) and incurs a cost when they grant credit to someone who defaults (a false negative prediction). We assume a firm is not impacted by an applicant they do not grant credit to who would default (a true positive prediction), and for simplicity, assume they incur no opportunity cost for rejecting an applicant who would not default (a false positive prediction). Profit at a given threshold $\pi(\tau)$ is the revenue from repayment (R) times the number of true negative predictions at that threshold, $TN(\tau)$, less the cost of default (C) times the number of false negative predictions at that threshold, $FN(\tau)$:

$$\pi(\tau) = R * TN(\tau) - C * FN(\tau) \quad (3)$$

To examine different operating scenarios we consider 2,431 cost-to-revenue (*C:R)* pairs, covering the full range of reported ratios (up to 35x) from the literature (Altman et al. 1977, Stein 2005). To

calculate the firm optimal profitability $\pi(\tau^*)$ at each *C:R* ratio we first generate a 90% random sample of the out-of-sample predictions and calculate the $TN(\tau)$ and $FN(\tau)$ counts across 9,500 thresholds from 0.01% to 0.95%, in increments of 0.01%. We then calculate the profit for each *C:R* ratio and every $TN(\tau)/FN(\tau)$ pair and find the maximum profit and corresponding optimal threshold for each pair. We apply those optimal thresholds to the 10% holdout and calculate the $TN(\tau)$ and $FN(\tau)$ counts, then calculate the optimal profit given the *C* and *R* for each threshold. We take the average of the optimal profit across the 30 folds to calculate firm profitability. We compare performance between models by reporting the mean difference of firm profitability across all *C:R* ratios and the number and range of the statistically significant differences, calculated using a two-sided paired t-test with a 95% confidence interval.

4. Data, sampling, and analytical approach

We needed to source data that had both gender and known default outcomes for all observations given the choice of key metrics: discrimination, predictive quality, and firm profitability. Real-world applicant data does not however have complete known default outcomes, as some applicants are rejected by the lender and therefore default versus repayment is not observed; a challenge referred to as the *selective labels problem* (Lakkaraju et al. 2017). A naïve approach would be to use the borrower data, which has complete default outcome information; however, this could introduce a selection bias as the borrower population does not necessarily represent the applicant population we are interested in measuring (Lakkaraju et al. 2017). To overcome this problem in practice, lenders use reject inference techniques (Hand and Adams 2014) to incorporate data from rejected loan applicants into the lending modeling process. There are four reject inference techniques (Hand and Adams 2014):

1. augmentation, which adjusts the distribution of borrowers to match the applicant population;
2. extrapolation, which estimates the outcome labels of rejected cases using known features;
3. conducting experiments, in which lenders purposely provide credit to individuals who they believe will default, to gather the missing label of these applicants; and
4. gathering outcomes of the rejected applicants who manage to obtain credit from another lender.

Our industry contacts confirmed that experiments are too costly and therefore rarely used. Privacy regulations restrict us from gathering the outcomes of rejected applicants as the data is anonymized. We select augmentation as it is model agnostic, relying only on resampling of the testing data, and therefore prevents additional noise from being introduced to the outcomes, which could occur if extrapolation were used. We follow the well developed augmentation methodology from Hsia (1978).

4.1. Data sourcing

We acquired real data for 307,507 borrowers from Home Credit, a global fintech that operates in 9 countries under Regime 1 and 2 jurisdictions, each allowing for the collection of gender. Machine learning is an integral part of how Home Credit manages their strategy, risk, products, funding, and customer lifecycles (https://www.homecredit.net/about-us/our-vision-and-business-model.aspx); its importance led them to create the "Home Credit Default Risk" competition on Kaggle from which we source of our data. At the time of writing this paper the data was available at https://www.kaggle.com/c/home-credit-default-risk/data. We used the competition training datasets (gathered into one file with observations for each borrower) but excluded the testing dataset as it was missing default outcomes, and therefore could not be used for our study. Note that the competition's rules prohibit using the data for published research; however, Home Credit granted us permission to use the data for this study. Additional exploration of the data, and replication of our results is possible through participation in the Kaggle competition.

4.2. Sampling: reject inference – augmentation

It is critical to note, that since our data comes from a Kaggle competition, it excludes the records for the rejected applicants, i.e., those who were denied credit in the past, leading to the aforementioned selective labels problem. To overcome this we utilized a reject inference technique called *augmentation* (Hsia 1978), a common method used in research (Hand and Adams 2014) and industry. Augmentation adjusts the borrower data (which importantly, includes default outcomes) to better reflect the applicant population. Technically, we used the applicant data (provided by our industry partner and not available on Kaggle) to create two joint distributions (one of borrowers, one of applicants) across the five most important predictive features from the lending application form: gender, age, income, occupation, and marital status. Then, we measured the proportional differences between borrowers and applicants across each of the five key characteristics. Borrowers and applicants were found to be statistically significantly different (according to a two-proportion z-test, at an alpha of 0.05) in their distribution across income and occupation; while gender, age, and marital status were comparable. Using this proportional difference information, we down-sampled the borrower data accordingly to create a "look-a-like" applicant dataset, with complete default outcomes, which helps to address the selective labels problem. See schematic in Figure 3. The technical details and accompanying code for the augmentation sampling procedure can be found at the author's github: https://github.com/stephaniekelley/genderbias.

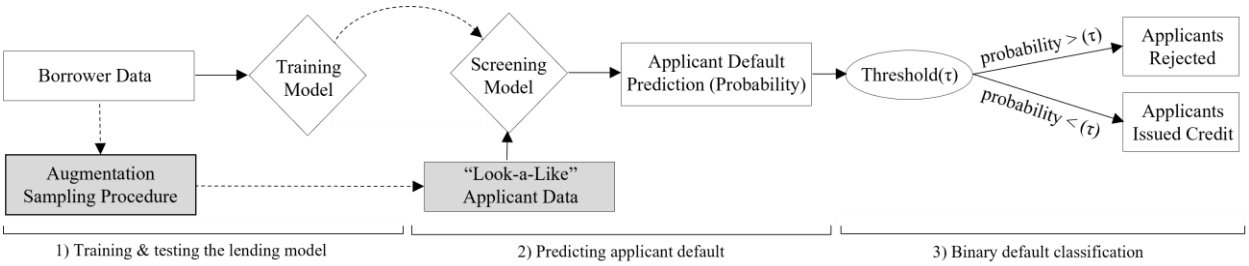


Fig. 3. Schematic view of the integration of the augmentation sampling procedure into the 3-step fintech lending modeling process

Because it is not practically feasible to perform the augmentation sampling procedure using every variable (we select five key variables), it is likely that some selection bias remains after the augmentation procedure. As we use the same testing dataset throughout the paper, any selection bias would be present across all models and would therefore not impact our qualitative findings. Additional selection biases could remain in our results due to other selection steps in the credit lending process, including: who lenders advertise to, who applies for credit, and which applicants accept loans (Henley and Hand 1997). While these selection steps are operationally relevant to the firm, they remain out of scope for this investigation as they require exceedingly rare data, which may not even be collected by firms, to analyze.

To replicate the training and testing step from the lending modeling process (Fig. 3) throughout the paper we created two samples from the Home Credit Borrower data (the sampling procedure is summarized in Fig. 4). First, per predictive modeling best practices, we randomly split the Home Credit Borrower Data into an 80% Borrower Training Data set, and a 20% Borrower Testing Data set. Then:

1. From the Borrower Training Data set we generated **Minority Training Data** (20% women), created by randomly down-sampling women to better reflect the gender imbalances present in borrower datasets available to fintech lenders globally (Ongena and Popov 2016). This data is used throughout the paper for training models in the main analysis (§5 & §6).
2. From the Borrower Testing Data, we generated **Look-a-Like Applicant Testing Data** (66.2% women), created using the augmentation sampling procedure (discussed in §4.2) to reduce the selective labels problem. This data is used throughout the paper as input for the screening model to obtain "out-of-sample" predictions, providing a consistent comparison across key metrics.

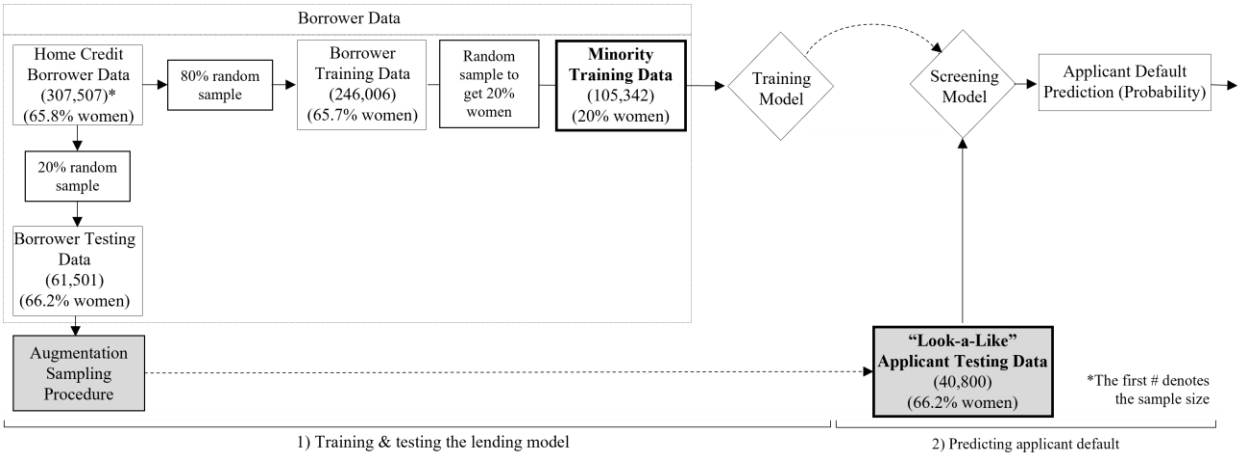


Fig. 4. Schematic view of the integration of the augmentation sampling procedure into the 3-step fintech lending modeling process

Across all samples we observe women to be more creditworthy than men, with fewer observed defaults (~6.8% and ~10.4% for training data, ~6.9% and ~10.5% for testing data), in line with data from past empirical investigations and reports on gender and lending (D'Espallier et al. 2011).

4.3. Analytical approach

Prior to presenting the results in §5, we review two modeling processes that we use in the study to simulate the model building of fintech lenders: traditional statistical modeling, and machine learning.

4.3.1. The traditional statistical modeling process

We use logistic regression (LR) as the "traditional statistical" model as it is the preferred model of lenders (Thomas et al. 2017), and used in past lending discrimination studies. To generate a LR model, a fintech firm would follow the traditional statistical modeling process (introduced in gray in Fig. 2) by first *collecting and cleaning the data*. The data was originally used for a ML competition, so we had to exclude some of the time series features that did not adhere to the modeling assumptions of LR, leaving us with a subset of 122 features (under Regimes 1 or 2, where gender can be collected), or 121 features (under Regime 3, where gender cannot be collected).

Following the standard methodology for *data cleaning* of LR credit models used by Andreeva and Matuszyk (2019) we coarse-classified the continuous features: features were first split into 10 intervals, adjacent intervals with similar default rates were manually merged, and separate coarse classes were generated for missing observations. Small categories were grouped into categorical variables and transformed into binary dummy variables, with the largest category removed to avoid identification issues. Per Andreeva and Matuszyk (2019) we trained a LR model and *manually selected features* that were significant at an alpha of 0.05 in the model with gender, and then excluded gender from this dataset

to generate the genderless feature set. These features were then used to train the LR training model, resulting in the *learning model parameter estimation*, results of which are discussed in §5 onwards.

4.3.2. The machine learning process

The ML process introduces three additional steps: feature engineering, algorithm selection, and hyperparameter tuning (recall Fig. 2, and §1.1.2). We started with the same 122 features (or 121 in Regime 3) from the LR data, with no further *data cleaning*. We then proceeded to *feature engineering*, whereby more features were generated based on interactions and/or transformations of the original feature set. We used techniques inspired by the publicly available code of the top-ranking teams in the Kaggle competition to gather features into a format that could be used by ML models and generate several ratios between the features to improve predictive performance. This resulted in 744 features (Regimes 1 or 2), or 743 features, excluding gender (Regime 3).

We then compared over 50 ML models (including extreme gradient boosting, generalized additive model, elastic-net, light gradient boosted tree, kernel SVM, random forest, Naïve Bayes, and a neural network) in DataRobot, a commercially available automated ML platform. We *selected the algorithm* with the best predictive quality (measure by the five-fold cross-validated AUC reported in DataRobot, a metric we discuss in §3.2). Access to DataRobot can be obtained through their Academic Support Program: https://www.datarobot.com/success/academic-support-program/.

The top performing DataRobot algorithm was "Average Blender" (AB) an ensemble classifier that averages the predictions from multiple models, in our case, several forms of XGBoost and Light Gradient Boosting models, each with strong predictive quality. Ensemble models, like the AB often have stronger predictive quality compared to individual models in credit lending (Lessmann et al. 2015). The algorithm then performed automated *feature selection* (as opposed to the manual feature selection performed in the traditional statistical modeling process), explanatory features were extracted for use in the final learning model parameter estimation. The algorithm then *tuned hyperparameters*, modeling values used to further improve the predictive quality of the chosen algorithm, and then the final *learning model parameter estimation* occurred, results of which are discussed in §5 onwards.

In §6.2 we introduce a single XGBoost tree ensemble model to support our investigation of the drivers of ML discrimination. We introduce this second ML model because the explainability techniques required for our analysis (SHAP values and SHAP interaction values), can only be calculated with access to the full model training process (not possible in DataRobot), and are designed for single-class ensembles like XGBoost, not multi-class ensembles like the AB (Lundberg and Lee 2017). Our XGBoost model is trained in R using the xgboost package. The Home Credit competition on Kaggle provides a practically relevant external measure of model quality via the competition leaderboard; our model would have landed

in the top 10 of 7,000+ models in the Kaggle competition, illustrating the model is highly competitive with other state-of-the-art models. As Kleinberg et al (2020) note, a firm can never know the true prediction function, but the closer the algorithm is to the true function (i.e., the better the predictive quality), the lower the bias will be. So, our top ranked model is likely one of the least bias models possible. The code for this model, and our LR model, also trained in R is available here on github: https://github.com/stephaniekelley/genderbias.

5. Impact of anti-discrimination regimes on gender-based discrimination

In this section we compare the impact of the three regimes on the top performing ML model, AB.

5.1. Impact of gender-exclusion on machine learning models

In line with what a fintech lender would do in practice, we follow the ML modeling process discussed in §4.3.2 and train a model on the Minority Training Data (discussed in §4.2) to generate a default prediction score for each new applicant. Under Regime 1, the firm can use gender as a feature in training and screening models (resulting in Model 1), while those under Regimes 2 and 3 are restricted in the use of gender as a feature (resulting in Model 2).

Observation #1 (see Table 2) reports the impact of removing gender from the Average Blender model (AB:M2) compared to the same model with gender (AB:M1); in summary, discrimination, measured by PPV, increases by 285.04%, predictive quality is not impacted, and firm profitability decreases by 0.25%. Discrimination is also visualized in Figures 5a, for PPV, and 5b, for WGMD.

Obs.	Model	Discrimination (PPV)	Discrimination (WGMD)	Predictive Quality (AUC)	Firm Profitability (Average Optimal Profit)
Compared to the Average Blender model with gender (AB:M1)					
1	AB:M2	+285.04% [51.72%- 1011.46%][1] (See Fig. 5a.)	+34.75% [17.37% - 42.93%][1] (See Fig. 5b.)	Not impacted 77.97% [77.16 – 78.79%] vs. AB:M1 78.06% [77.24 – 78.88%]	-0.25%[2]
Compared to the Logistic Regression model with gender (LR:M1)					
2	AB:M1	-44.06% [-127.53% – +2.60%][3] (See Fig. 6a.)	Inapplicable[4]	+472 bps 78.06% [77.24 – 78.88%] vs. LR:M1 73.34% [72.46 – 74.22%]	+7.86%[5]
3	AB(STAT):M1	-16.32% [-54.66% - +56.75%][1] (See Fig. 6a)	Inapplicable[4]	+192 bps 75.26% [74.41 – 76.12%] vs. LR:M1 73.34% [72.46 – 74.22%]	+3.92%[6]
Compared to the Logistic Regression model without gender (LR:M2)					
4	AB:M2	-9.56% [-36.87% – +13.15%][1] (See Fig. 6b)	Inapplicable[4]	+487 bps 77.97% [77.16 – 78.79%] vs. LR:M2 73.10% [72.22 – 73.98%]	+7.60%[7]

5	AB(STAT):M2	Is of comparable discrimination (See Fig. 6b)	Inapplicable[4]	+200bps 75.10% [74.24 – 75.59%] vs. LR:M2 73.10% [72.22 – 73.98%]	+4.19%[8]

Notes. [1]Across thresholds (5-30%) the differences are statistically significant ($p<0.01$, paired t-test; see §3.1 for details). [2]11.11% of the profit differences are statistically significant, 23% are negative differences [-2.43 - -1.19%], and 77% are positive [0.02 – 23%]. [3]Across thresholds (5-30%) the differences are statistically significant ($p<0.01$, paired samples Wilcoxon test). [4]Discrimination (WGMD) is inapplicable when comparing AB and LR models as it does not adjust for the lack of calibration between the model predictions, which are from two different model families (see §1 for details). [5]89.88% of the profit differences are statistically significant, 98% are positive differences [0.04 -68.20%], and 2% are negative [-0.006 - -0.005%]. [6]82.06% of the profit differences are statistically significant, 93% are positive differences [0.06 - 51.10%], and 7% are negative [-0.006 - -0.005%]. [7]90.58% of the profit differences are statistically significant [0.02 - 63.50%]. [8]77.87% of the profit differences are statistically significantly different [0.03 – 51.30%].

Table 2. Observations #1 - 5: A comparison of lending models across discrimination (PPV & WGMD), predictive quality (AUC), and firm profitability (average optimal profit)

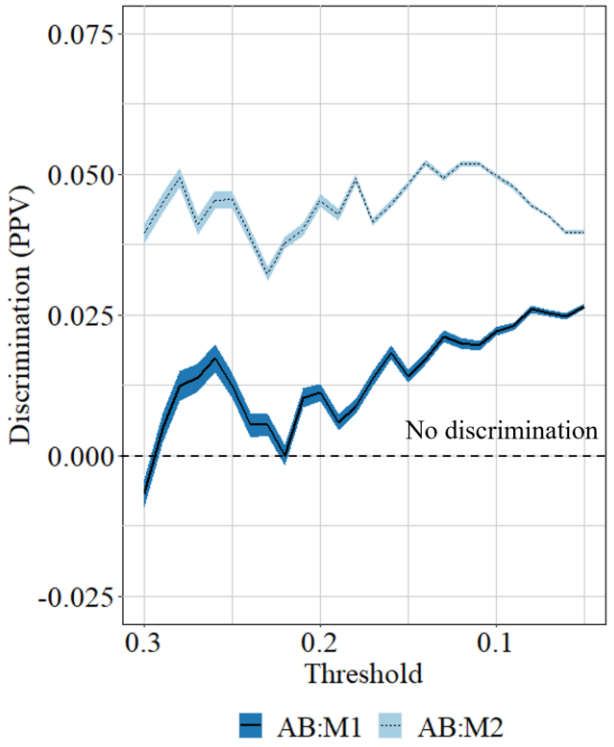


Fig. 5a. Discrimination, measured by PPV, of AB models with gender (AB:M1) and without gender (AB:M2)

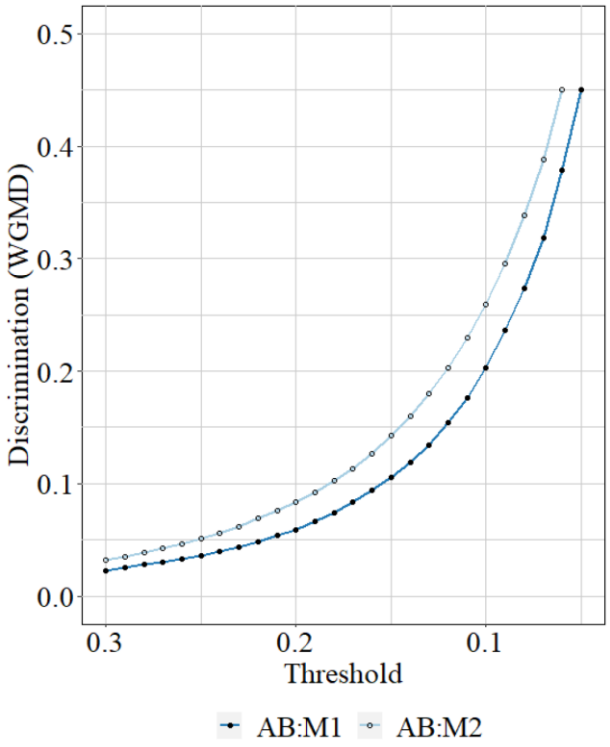


Fig. 5b. Discrimination, measured by WGMD, of AB models with gender (AB:M1) and without gender (AB:M2)

A key insight follows from Observation #1: the operational guidance to exclude gender as a feature in training and screening models prescribed by Regimes 2 and 3, leads to increased discrimination and decreased firm profitability compared to Regime 1 which allows for the use of gender. This negative impact occurs in both ML, and LR models (the results for which are included in the Supplemental Materials, S2 for brevity). The results are troubling as they demonstrate that Regimes 2 and 3 create a

detrimental outcome for both lending applicants (increased discrimination) and fintech lenders (decreased firm profitability), confirming the reports of automated bias in fintech lending that motivated our work.

6. Drivers of statistical and machine learning discrimination

6.1. Comparison of discrimination in traditional statistical and machine learning models

Given the discriminatory impact of excluding gender, we examine whether the lending applicants and fintech firms would be better off (i.e., observe lower levels of discrimination, higher predictive quality, and higher firm profitability) using LR or ML models. To do so, we compare the LR and AB models, with gender (Regime 1); and without gender (Regimes 2 and 3), trained on the Minority Training data.

Observation #2 (Table 2) reports the impact of using an Average Blender model with gender (AB:M1) versus a Logistic Regression model with gender (LR:M1); in summary, discrimination, measured by PPV, decreases by 44.06%, predictive quality increases by 472 bps, and firm profitability increases by 7.86%. Discrimination (PPV) is visualized in Figure 6a.

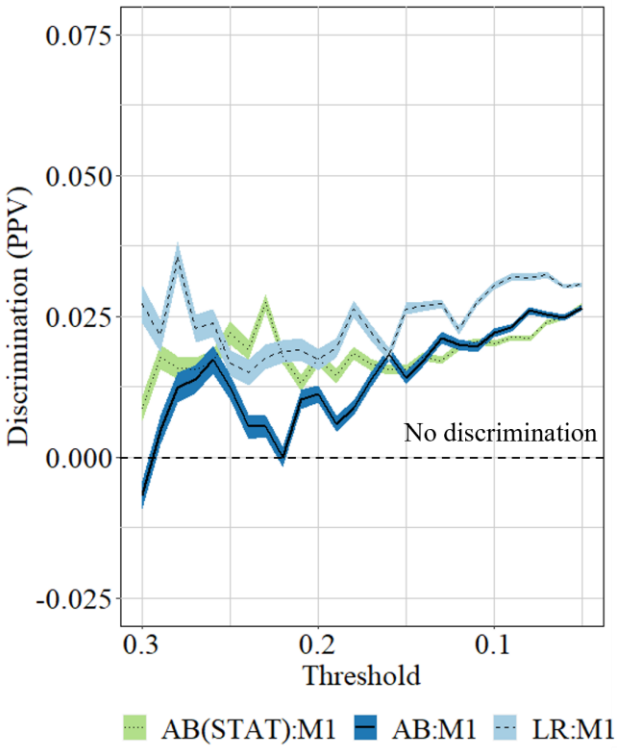


Fig. 6a. Discrimination, measured by PPV, of LR and AB models with gender (LR:M1, AB:M1) and the AB model trained on the traditional statistical dataset with gender (AB(STAT):M1)

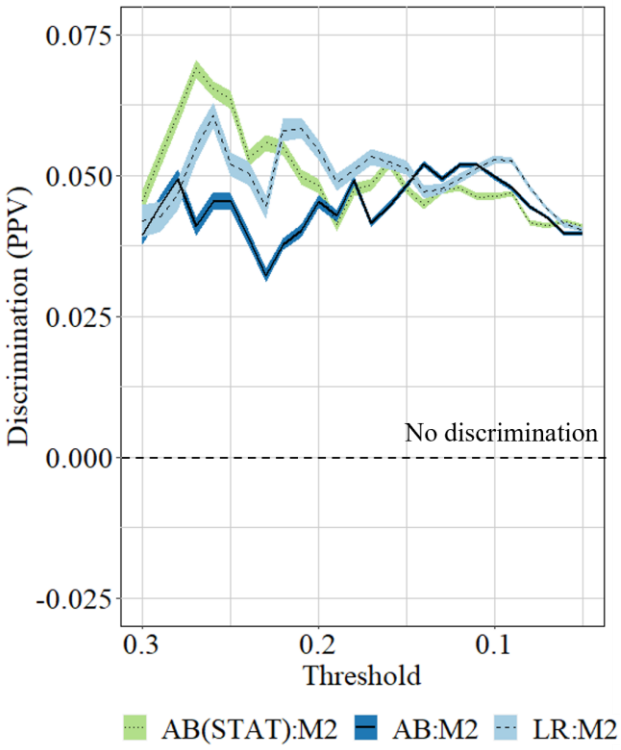


Fig. 6b. Discrimination, measured by PPV, of LR and AB models without gender (LR:M2, AB:M2) and the AB model trained on the traditional statistical dataset without gender (AB(STAT):M2)

These results illustrate that when gender can be used in the training and screening models (per Regime 1), the AB model is less discriminatory, of better predictive quality, and greater profitability than the LR model. From our understanding of the traditional statistical modeling process (§4.3.1) and the ML process (§4.3.2), we know that the results from Observation #2 occur due to two things: 1) a change in the model (LR to AB), and 2) access to a larger ML feature set. For robustness we report the results of a second AB model trained on the traditional statistical feature set used by the LR model (AB(STAT):M1), which allows us to observe the impact of 1) a change in the model (LR to AB), by itself.

Observation #3 (Table 2) reports the impact of using an Average Blender model trained on the traditional statistical feature set with gender (AB(STAT):M1) versus a Logistic Regression model with gender (LR:M1); in summary, discrimination, measured by PPV, decreases by 16.32%, predictive quality increases by 192 bps, and firm profitability increases by 3.92%. Discrimination (PPV) is visualized in Figure 6a.

Observation #3 illustrates that even on the traditional statistical feature set, with fewer engineered features, the AB model is less discriminatory, of better predictive quality and greater profitability than the LR model. This suggests that lenders and applicants should both prefer ML models over LR models under Regime 1. Next, we investigate whether lenders and applicants would be better off with ML models under Regimes 2 and 3, when gender is not included as a feature in the training and screening models.

Observation #4 (Table 2) reports the impact of using an Average Blender model without gender (AB:M2) versus a Logistic Regression model without gender (LR:M2); in summary, discrimination, measured by PPV, decreases by 9.56%, predictive quality increases by 487 bps, and firm profitability increases by 7.60%. Discrimination (PPV) is also visualized in Figure 6b.

The above findings show that without access to gender, the AB model (AB:M2) is less discriminatory (PPV), of better predictive quality, and greater firm profitability compared to the LR model. For robustness, as we did with Model 1, we also report the results of a second AB model trained on the traditional statistical feature set used by the LR model (AB(STAT:M2).

Observation #5 (Table 2) reports the impact of using an Average Blender model trained on the traditional statistical feature set without gender (AB(STAT):M2) versus a Logistic Regression model with gender (LR:M2); in summary, discrimination, measured by PPV, is comparable, predictive quality increases by 200 bps, and firm profitability increases by 4.19%. The impact on discrimination (PPV) is visualized in Figure 6b.

Observation #5 suggests that without gender, the AB model without the engineered features (AB(STAT):M2) is of better predictive quality and greater profitability, but of comparable discrimination

to the LR model; it is not able to reduce discrimination to the degree it can with access to the full feature-engineered dataset (AB:M2).

In aggregate, Observations #2 - #5 illustrate that trained on the full feature engineered dataset, the AB model is less discriminatory, of greater predictive quality and profitability compared to the LR model; both when gender is included and excluded. Trained on the non-engineered, traditional statistical feature set, the AB model is of better predictive quality and greater profitability compared to the LR model, regardless of whether gender is included or not; however, it is of comparable discrimination, suggesting that feature engineering has a significant impact on reducing discrimination. This demonstrates that both fintech firms and lending applicants would benefit from the use of ML models in place of traditional statistical models, like LR, but the greatest decrease in discrimination relies on the complete ML process (particularly the feature engineering, as discussed in §4.3.2), and the inclusion of gender.

6.2. Using ML explainability techniques to uncover the drivers of ML discrimination

The previous results illustrate that under Regimes 2 and 3, even the top-performing ML model (AB:M2) still proliferates gender discrimination. Next, we seek to understand the drivers of that discrimination using two ML explainability techniques: permutation importance, to understand the impact of excluding gender in the AB model; and SHAP values and SHAP interaction values (Lundberg et al. 2019, Lundberg and Lee 2017) of the state-of-the-art XGBoost model (discussed in §4.3.2). Recall, SHAP values and SHAP interaction values cannot be created for model ensembles. Before proceeding to the discussion of ML discrimination in §6.2.2 we review OVB which drives LR discrimination.

6.2.1. Statistical discrimination: omitted variable bias

In the traditional statistical modeling process (per Fig. 2) it is well known that when a LR model has access to a comprehensive set of causal features, it can estimate the true, unbiased learning model parameters (Wooldridge 2015). Following this, when an important causal feature is excluded in data collection (e.g., gender), the learned model parameter estimates become biased, in the statistical sense of the word (i.e., inaccurate); a phenomenon referred to as omitted variable bias, OVB (Wooldridge 2015). Andreeva and Matuszyk (2019) show that when gender is excluded from a LR model, it creates OVB. Men, the less creditworthy gender are a majority, and the exclusion of gender creates an upward bias in the parameter estimates, leading to an increase in the rejection rates of women compared to the model with gender. We confirm in the Supplemental Materials, S2, that as expected, OVB occurs in our data too.

6.2.2. Machine learning discrimination: gender-blind feature selection

Recall (from §1.1.2) that the ML modeling process (Fig. 2) alters the construction of the model training procedure from that of LR, and the exclusion of gender affects each additional step in the ML

modeling process: feature engineering, algorithm selection, feature selection, and hyperparameter selection, which then change the final learning model parameter estimation. We focus our investigation on *feature engineering* and *feature selection* as these steps are most impacted by the exclusion of gender and leave the discussion of *hyperparameter selection* to §7. We first compute the permutation importance (using the feature importance tool in DataRobot) for the AB Models 1 and 2. To illustrate the result visually we add 10 manually generated gender interactions with the top 5 features (5 for women, 5 for men) to Model 1. The permutation importance of the top 25 features for Models 1 and 2 are visualized in Figs. 7a. and 7b., respectively. We observe the following:

Observation #6*: When gender is included in the ML model, gender interaction features account for 4 of the top 25, and 2 of the top 10 features (Fig. 7a); gender is also selected as a feature (outside the top 25).*

Observation #7*: When gender is excluded from the ML model, different features are selected by the algorithm (6 of the top 25 features), with different permutation importance rankings (21 of the top 25 features), compared to when gender is included (Fig. 7b).*

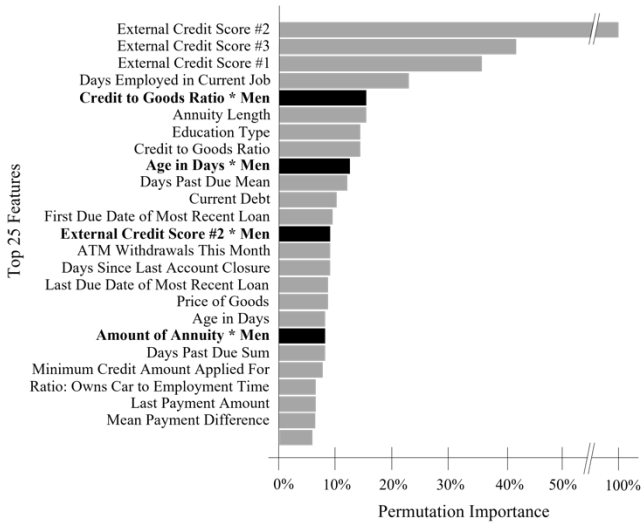


Fig. 7a. Permutation importance of the top 25 features of the AB model with gender & gender interactions (M1)

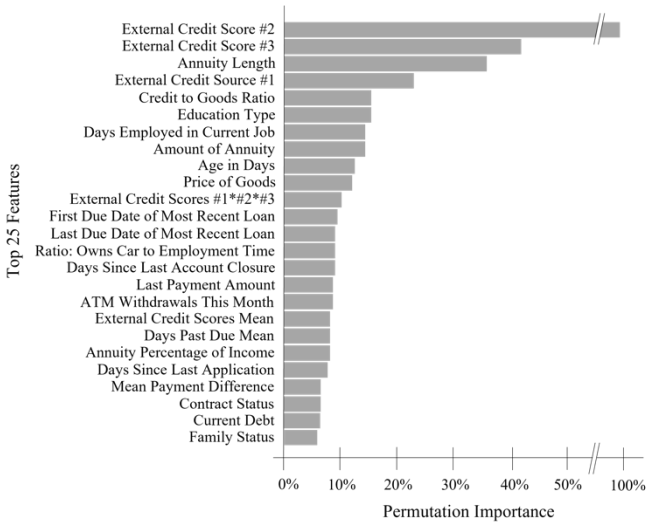


Fig. 7b. Permutation importance of the top 25 features of the AB model without gender (M2)

Results are a lower bound on the number of affected features; gender may have impacted engineered features not shown (e.g., binning of External Credit Scores may be different in M1 and M2). We refer to this phenomenon, where different features are selected by the algorithm when gender is not present, as *gender-blind feature selection*. To better understand the phenomenon, we investigate the feature engineering and feature selection in more detail using SHAP values and SHAP interaction values. As a reminder, we opted to use the XGBoost model for this part of the investigation given the restrictions of the SHAP interaction values (they cannot be calculated for the AB multi-class ensemble model).

We observe that certain features, like External Credit Score #2, and Annuity Length are always selected by the algorithm, regardless of whether gender is included; we refer to these as *gender-neutral* features. These features consistently have the highest SHAP values (feature importance), and are very important for the final prediction, accounting for 97.1% of the final 655 features selected in Model 1, and 97.7% of the final 659 features in Model 2. Other features are selected by the algorithm when gender is present, but not when it is excluded; we refer to them as *gender-reliant*.

The final set of features are *gender-redundant;* when gender is present, they are "redundant" and are excluded by the algorithm in the automated feature selection step, but when gender is excluded, they are selected by the algorithm. We observe these features have very low SHAP values compared to *gender-neutral* features, and higher gender inference information compared to *gender-reliant* features.

Next, we compute the SHAP interaction values using a condensed feature set of the explanatory features, given the computation requirements (Lundberg et al. 2019). These values tell us the feature importance for every feature engineered pairwise interaction and help us to understand why the *gender-blind feature selection* phenomenon leads to discrimination. We look specifically at the SHAP interaction values between gender and the *gender-reliant* features, and gender and the *gender-redundant* features, visualized in Figure 8.

Observation #8*: The mean SHAP interaction values for the gender-reliant features with gender are 19x greater for women versus men; 2.5x the difference of gender-neutral features.*

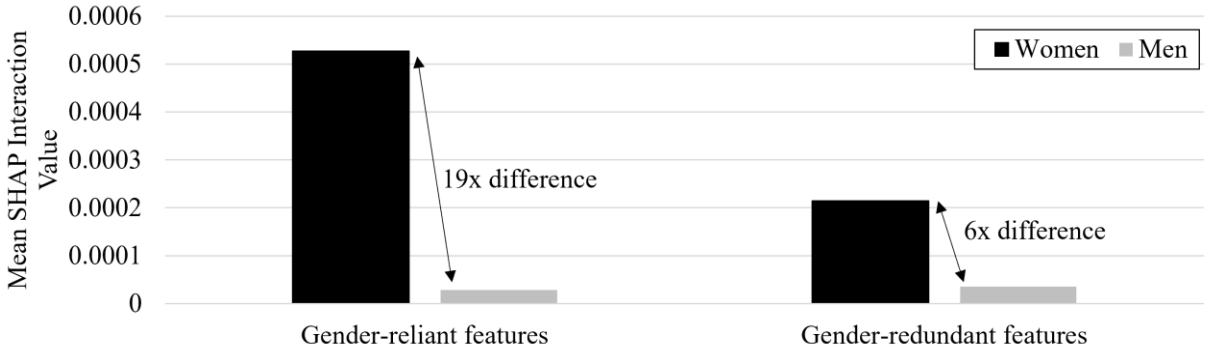


Fig. 8. Mean SHAP interaction values for gender-reliant features (those selected by the algorithm when gender is present, but not when it is excluded), and gender-neutral features (those always selected by the algorithm) across women and men

This illustrates that predictions for women rely more on the *gender-reliant* features than men, and therefore women are more detrimentally impacted by the gender exclusion enforced by Regimes 2 and 3.

6.3. Summary of the drivers of machine learning discrimination

Summarizing, in ML, when gender is excluded, it prevents the algorithm from feature engineering, e.g., from creating interactions with other features and gender. It also impacts algorithm,

feature and hyperparameter selection; we observe that certain *gender-reliant* features are excluded, and in their place *gender-redundant* features are selected. In this setting, the exclusion of the *gender-reliant* features is significantly more detrimental to women, compared to men, increasing discrimination. This gender-blind feature selection phenomenon is vastly different from the OVB that drives discrimination in traditional statistical models; we show the discrimination is linked to changes in the construction of the model training procedure, leading to a new understanding of ML discrimination. The above discussion is limited to the dataset that we use, and while that dataset is from a real fintech firm, to emphasize the generalizability of our insights, we provide a stylized example in the Supplemental Materials (S3).

7. Possible approaches to reduce discrimination

Finally, we consider what ethically minded fintech firms can do to reduce the gender discrimination given the restrictions of the anti-discrimination regimes.

7.1. Approaches to reduce discrimination under Regime 2

Fintech firms under Regime 3 are not able to collect and therefore use gender as a feature in their training and screening models, which we now know leads to discrimination (§5). Those under Regime 2 are prohibited from using gender as a feature in the training and screening models used for individual lending decisions but are allowed to collect gender and use it in other ways during the modeling process. Next, we explore several possible approaches for firms to reduce discrimination given these permissions:

1. **Down-sampling the training data to rebalance gender**, i.e., under-sampling the majority class (men) to match the count of the minority class (women), leading to the Rebalanced Down-sampled Training data (50% women/50% men, n=42,136) (DS:M2);
2. **Gender-aware hyperparameter tuning**, which involves creating a training model that tunes the hyperparameters using borrower gender data (we use the XGBoost model, and hyperparameters inspired by the top Kaggle teams), which allows the training model to learn about gender at an aggregate level before it is retrained on the Rebalanced Down-sampled Training data without gender (HT:M2); and
3. **Up-sampling the training data to rebalance gender**, which involves a firm collecting more data from the minority class (women) to achieve a balanced sample, which we emulate by "collecting" data from the Borrower Training data excluded during the creation of the Minority Training data, resulting in the Rebalanced Collected Training data (50% women/50% men, n=168,548) (US:M2).

We re-estimate the AB model, without gender (M2) from the main analysis using these three techniques and make the following observations (#9-11), which are summarized and expanded upon in Table 3. Accompanying discrimination results are visualized in Figures 9a, for PPV, and 9b, for WGMD.

Observation #9 reports the impact of down-sampling the training data to rebalance gender (DS:M2) compared to the Average Blender model without gender (AB:M2); in summary, discrimination, measured by PPV, decreases by 4.54%, predictive quality decreases by 175 bps, and firm profitability decreases by 4.47%.

Observation #10 reports the impact of gender-aware hyperparameter tuning (HT:M2) compared to AB:M2; in summary, discrimination, measured by PPV, decreases by 37.73%, predictive quality decreases by 278 bps, and firm profitability decreases by 4.42%.

Observation #11 reports the impact of up-sampling the training data to rebalance gender (US:M2) compared to AB:M2; in summary, discrimination, measured by PPV, decreases by 24.47%, predictive quality is not impacted, and firm profitability decreases by 1.46%.

Obs.	Model	Discrimination (PPV)	Discrimination (WGMD)	Predictive Quality (AUC)	Firm Profitability (Average Optimal Profit)
Compared to the Average Blender model with gender (AB:M1)					
9	DS:M2	-4.54% [-19.38% - +14.03%][1] (See Fig. 9a.)	-21.5% [-37.15% - -1.51%][1] (See Fig. 9b.)	-175 bps 76.22% [75.37 – 77.07%] vs. AB:M2 77.97% [77.16 – 78.79%]	-4.47%[2]
10	HT:M2	-37.73% [-75.84% – +9.67%][1] (See Fig. 9a.)	-30.37% [-47.01% - -6.71%][3] (See Fig. 9b.)	-278 bps 75.19% [74.33 – 76.04%] vs. AB:M2 77.97% [77.16 – 78.79%]	-4.42%[4]
11	US:M2	-24.47% [-59.76% – +2.79%][1] (See Fig. 9a)	-40.08% [-43.35% - -44.85%][3] (See Fig. 9b)	Not impacted 77.05% [76.22 – 77.88%] vs. AB:M2 77.97% [77.16 – 78.79%]	-1.46%[5]
12	PGP:M2	- 71.09% [-112.05% - -35.15%][1] (See Fig. 9a.)	-25.74% [-31.20% - -13.89%][3] (See Fig 9b.)	Not impacted 78.10% [77.28 – 78.92%] vs. AB:M2 77.97% [77.16 – 78.79%]	+0.13%[6]

Notes. [1]Across thresholds (5-30%) the differences are statistically significant ($p<0.01$, paired t-test; see §3.1 for details). [2][-50.40 - -0.01%] [3]Across thresholds (5-30%) the differences are statistically significant ($p<0.01$, paired samples Wilcoxon test). [4][-59.30% - -0.01%] [5]All of the profit differences are statistically significant, 88% are negative differences [-35.90% - -0.1%], and 12% are positive [0.01- 0.03%]. [6][0.01% - 2.75%]

Table 3. Observations #9 - 12: A comparison of possible approaches to reduce discrimination (PPV & WGMD), accompanied by predictive quality (AUC), and firm profitability (average optimal profit) changes

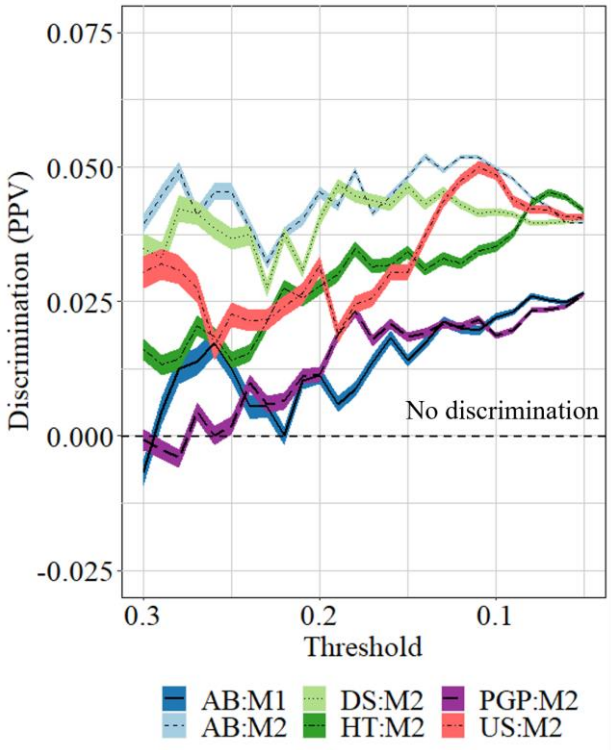


Fig. 9a. Discrimination, measured by PPV, of AB models with gender (AB:M1), without gender (AB:M2), and 4 possible approaches to reduce discrimination (DS:M2, PGP:M2, HT:M2, US:M2)

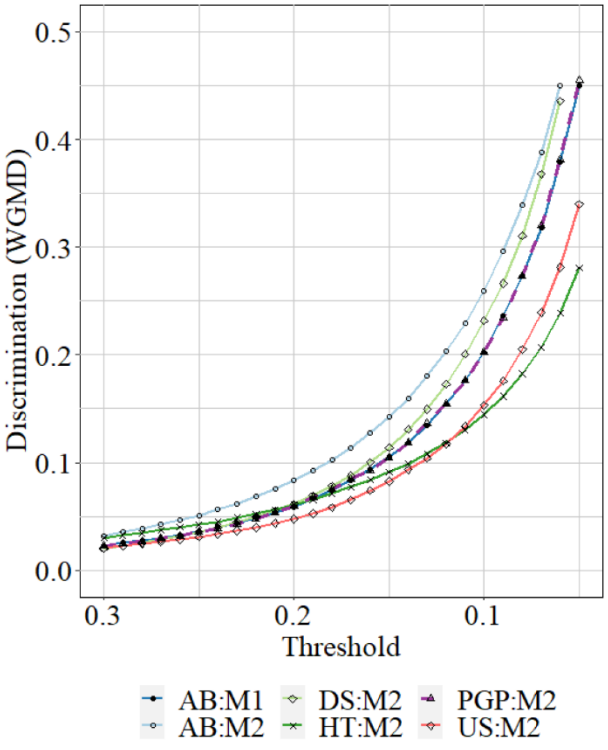


Fig. 9b. Discrimination, measured by WGMD, of AB models with gender (AB:M1), without gender (AB:M2), and 4 possible approaches to reduce discrimination (DS:M2, PGP:M2, HT:M2, US:M2)

The key insight is that although fintech lenders under Regime 2 (e.g., countries in the EU) cannot use gender as a feature in the training and screening models, they can use it to perform several alternative discrimination-reducing approaches. The approach selected by a fintech firm will depend on their threshold selection, and their acceptance of the potential fairness accuracy trade-off between reducing discrimination and the reduced predictive quality and firm profitability. Fortunately, firms in Regime 2 jurisdictions have several possibilities to reduce discrimination in models using data science techniques.

7.2. Discrimination reducing techniques for firms operating across jurisdictions

Here we explore an additional approach that may be technically feasible for a firm that operates in several jurisdictions: a probabilistic gender proxy model (PGP:M2). This involves first training an ML model to predict gender, or "impute," per Zhang (2018), and then using that gender prediction as a feature in the screening model to predict default. Barring distributional shift and data consistency, the lender could use data from a Regime 1 or 2 jurisdiction to create a model to predict the gender of borrowers, then apply that model to predict gender for applicants in the Regime 3 jurisdiction. Our gender prediction

model achieved a five-fold cross-validated AUC of 91.08%, implying that gender could, in fact, be predicted with excellent accuracy from the 700+ other available features in our data. We tuned the gender classification threshold to 20% to closely match the predictions of the model with gender (Model 1), which we know (per Observation #1) has lower discrimination, and higher profitability.

Observation #12 (summarized and expanded up in Table 3) reports the impact of the probabilistic gender proxy, PGP, model (PGP:M2) compared to the Average Blender model without gender (AB:M2); in summary, discrimination, measured by PPV, decreases by 71.08%, predictive quality is not impacted, and firm profitability increases by 0.13%.

This illustrates the benefits of probabilistic gender proxy modeling for applicants (reduced discrimination), and fintech firms (no change to predictive quality and increased profitability). Unfortunately, we determined the methodology is currently prohibited in the US (the largest jurisdiction under Regime 3), and has been observed to generate upward statistical bias in default predictions, albeit in the mortgage setting, not consumer credit (Chen et al. 2019). Down-sampling, gender aware hyperparameter tuning, and up-sampling also cannot be implemented by fintech firms under Regime 3 (e.g., the US) as they are prohibited not just from using, but also collecting gender, which means fintech firms under this regime, like the Apple Card, are restricted in their ability to measure and reduce discrimination.

7.3. Allowing for the collection and use of gender to reduce discrimination

Lastly, we return to the operational modeling guidance of Regime 1 regulations, which allow for both the collection *and* use of gender in the training and screening models. Summarizing the findings of several observations throughout the paper: the ML model, with gender, results in the lowest discrimination (PPV, WGMD) across thresholds (5-30%), the highest predictive quality, and the greatest firm profitability, compared to: the ML model that excludes gender, the ML approaches that reduce discrimination in the absence of gender, and the LR models. In short, our results suggest that the best way to reduce bias in this setting is to use ML models and allow for both the collection and use of gender.

8. Discussion & conclusions

We use publicly available, real, feature-rich and highly-multicollinear fintech data to investigate the impact of three anti-discrimination legal regimes on gender discrimination: Regime 1, which allows for the collection and use of protected attributes in both training and screening models; Regime 2, which allows for the collection of gender but prohibits its use as a feature in the training and screening model; and Regime 3, which prohibits both the collection and use of gender in any model. We find that prohibiting the use of gender as a feature in the screening model (per Regimes 2 and 3), leads to increased

discrimination, and decreased firm profitability, without significantly impacting model predictive quality, in both traditional statistical and machine learning models. We find that across all anti-discrimination regimes, ML models are less discriminatory, of better predictive quality, and higher profitability when trained on the data commonly used by fintech firms, due to differences in the construction of the training procedure: feature engineering, feature selection, and hyperparameter tuning. We determine that ML discrimination is driven by a novel phenomenon, *gender-blind feature selection*, a process that is vastly different from OVB which drives discrimination in traditional statistical models.

In addition, we show that the seemingly subtle difference between Regimes 2 and 3, allowing for the collection of gender, presents fintech firms under Regime 2 with four possible approaches to reduce discrimination: 1) down-sampling the training data to rebalance gender, 2) gender-aware hyperparameter tuning, 3) up-sampling the training data to rebalance gender, and 4) probabilistic gender proxy modeling, each with varying impacts to model predictive quality and firm profitability. While these approaches reduce discrimination in our applicant dataset, it is unclear how they might impact applicant self-selection, either persuading or dissuading individuals from applying for credit from the fintech lender; this could be an important avenue for future research.

The overarching implication of our work is that the growing adoption of algorithmic decision-making in non-mortgage consumer credit lending requires a rethink of the anti-discrimination laws and their operational guidance, specifically with respect to the collection and use of protected attributes. Our analysis points to the importance of allowing for the responsible collection and use of gender data, in line with the operational guidance of Regime 1 regulations. Allowing fintech firms to collect protected attributes, like gender, would at minimum, give them the ability to assess the potential bias in their model, and could allow them to reduce discrimination through approaches such as down-sampling to rebalance gender, gender-aware hyperparameter tuning, up-sampling to rebalance gender, and probabilistic gender proxy modeling. These approaches could also in theory be leveraged to support affirmative action (also referred to as positive discrimination) initiatives, notwithstanding the critiques of the practice.

From a lender's perspective, the findings can serve as guidelines and influence lenders to revisit their existing data usage and algorithm design processes. For an industry partner involved in this work, residing in Regime 1, the findings are particularly interesting as the use of ML and AI becomes more widespread for decision-making in the financial sector. While there are disputes about the reduced explainability in AI models and potential reduction of fairness driven by model complexity (e.g., deep neural networks), the findings of this work are supportive of pursuing sophisticated AI model design and setting intra-organization data collection and usage requirements, which include responsible usage of personal attributes like gender, as part of an organization's AI ethics guidelines.

Our work also paves the way for fair economic welfare of both the financial institutions and the individual customers by approving loans for customers who deserve the financial support but are currently discriminated against when traditional modeling approaches or regulatory-binding guidelines are applied. The customers' chances for economic wellbeing are improved and likewise the profitability of the lending company increases because of a lower default risk. The collection and use of gender should be supported by a strong customer communication strategy; the benefits of using personal attributes should be well described and a suitable level of AI education carried out to increase customer confidence in the suggested approach.

Increased data access should however, come with greater firm accountability and responsibility. For example, in Singapore, the FEAT Principles (Monetary Authority of Singapore 2018), which this paper's authors had the privilege to contribute to the development of, recommend that lenders should be able to collect and use protected attributes, like gender and race, in their training and screening models, but are responsible for discrimination in the algorithmic output. This is contrary to the situation in the US where lenders have used the existing laws to elude responsibility for discriminatory outcomes, as Goldman Sachs did with their Twitter statement mentioned in the Introduction: "*we have not and never will make decisions based on factors like gender. In fact, we do not know your gender or marital status...*" To that end, as of early-2021 both the US and the EU have proposed regulatory guidelines for the responsible and ethical use of artificial intelligence. Both draft regulations will likely have implications for automated algorithmic decision-making in non-mortgage consumer fintech lending.

The US draft regulation, titled "Maintaining American Leadership in Artificial Intelligence," highlights automated bias as a potential risk, but does not suggest specific actions to mitigate it; however, members of the House of Representatives previously proposed an "Algorithmic Accountability Act" which offers more structured guidance to firms. The Act suggests users of automated algorithms perform a bias impact assessment to mitigate potential discrimination. A consequence of our findings is that in the US, the ECOA will make it virtually impossible for lenders to adhere to the new proposed Act as they will not be able to test for discrimination without first being able to collect protected attributes like race, disability, and gender.

The European Commission's new draft regulation titled "Artificial Intelligence Act" categorizes AI systems used for credit lending as high-risk and specifies certain mandatory requirements with regards to training data, data governance and explainability, reporting, robustness and accuracy, and human oversight. For example, this includes ensuring a sufficiently representative training dataset; two of the possible approaches to reduce discrimination we discuss, specifically down-sampling, and up-sampling to rebalance gender are methods that could meet these requirements.

Alternatively, organizations could take a self-regulation approach, as proposed by some legal scholars (Hadfield 2016), by developing fairness certification programs or voluntary AI ethics guidelines. To date, we have worked with several large multi-national banks and fintech firms who have developed these kinds of voluntary AI ethics guidelines in the absence of formal regulation.

Clearly, our findings show that there are inconsistencies between the objectives of the existing anti-discrimination regimes and their detrimental impact when decisions impacting minorities are made by algorithms. We considered one setting, consumer fintech lending, and urge other researchers to continue investigating the implications and drivers of other forms of discrimination, and potential solutions in additional contexts and operational settings.

References

Acquisti A, Fong C (2020) An Experiment in Hiring Discrimination via Online Social Networks. *Manage. Sci.* 66(3):1005–1024.

Akkoç S (2012) An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. *Eur. J. Oper. Res.* 222(1):168–178.

AlgorithmWatch, Bertelsmann Stiftung (2019) Automating Society: Taking Stock of Automated Decision-Making in the EU. Report, Algorithm Watch and Bertelsmann Stiftung, Berlin.

Altman EI, Haldeman RG, Narayanan P (1977) ZETA Analysis: A new model to identify bankruptcy risk of corporations. *J. Bank. Financ.* (1):29–54.

Andreeva G, Matuszyk A (2019) The law of equal opportunities or unintended consequences?: The effect of unisex risk assessment in consumer credit. *J. R. Stat. Soc. Ser. A Stat. Soc.* 182(Part 4):1287–1311.

Anon WHO/Europe | Gender: definitions. Retrieved (July 15, 2021), https://www.euro.who.int/en/health-topics/health-determinants/gender/gender-definitions.

Barocas S, Selbst AD (2016) Big Data's Disparate Impact. *Calif. Law Rev.* 104:671–732.

Bartlett R, Morse A, Stanton R, Wallace N (2019) Consumer-Lending Discrimination in the FinTech Era, Working Paper, University of California, Berkeley, Berkeley.

Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2017) Fairness in Criminal Justice Risk Assessments: The State of the Art, Working Paper, University of Pennsylvania, Philadelphia.

Chan J, Wang J (2018) Hiring preferences in online labor markets: Evidence of a female hiring bias. *Manage. Sci.* 64(7):2973–2994.

Chen IY, Johansson FD, Sontag D (2018) Why is my classifier discriminatory? *Adv. Neural Inf. Process. Syst.* (Montreal, Canada), 3543–3554.

Chen J, Kallus N, Mao X, Svacha G, Udell M (2019) Fairness Under Unawareness: Assessing Disparity When Protected Class is Unobserved. *FAT* '19 Proc. Conf. Fairness, Accountability, Transpar.* (ACM New York, Atlanta, GA), 339–348.

Chouldechova A (2017) Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5(2):153–163.

Cohen M, Fiszer MD, Ratzon A, Sasson R (2019) Incentivizing Commuters to Carpool: A Large Field Experiment with Waze. Working Paper, McGill University, Montréal.

Cohen MC, Harsha P (2020) Designing Price Incentives in a Network with Social Interactions. *Manuf. Serv. Oper. Manag.* 22(2):292–309.

Council of the EU (2004) Council Directive 2004/113/EC Implementing the principle of equal treatment between men and women in the access to and supply of goods and services. *Off. J. Eur. Communities* L373(December 2000):37–43.

Cui R, Gallino S, Moreno A, Zhang DJ (2018) The Operational Value of Social Media Information. *Prod. Oper. Manag.* 27(10):1749–1769.

Cui R, Li J, Zhang D (2020) Reducing Discrimination with Review in the Sharing Economy: Evidence from Field Experiments on Airbnb. *Manage. Sci.* 66(3):1071–1094.

DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* 44:837–845.

Doleac JL, Stein LCD (2013) The Visible Hand: Race and Online Market Outcomes. *Econ. J.* 123(572):F469–F492.

Elliehausen GE, Durkin TA (1989) Theory and evidence of the impact of Equal Credit Opportunity: An agnostic review of the literature. *J. Financ. Serv. Res.* 2(2):89–114.

European Commission (2012) Guidelines on the Application of Council Directive 2004/113/EC to Insurance, in the Light of the Court of Justice of the European Union in Case C-236/09 (Test-Achats). *Off. J. Eur. Union* C11(March 2011):1–11.

Franck T (2019) Warren rips Goldman over its Apple card, “discriminatory” algorithms. *CNBC*. Retrieved (April 18, 2021), https://www.cnbc.com/2019/11/14/warren-rips-goldman-over-its-apple-card-discriminatory-algorithms.html.
Fuster A, Goldsmith-Pinkham P, Ramadorai T, Walther A (2020) Predictably Unequal? The Effects of Machine Learning on Credit Markets. Working Paper.
Ge Y, Knittel, Christopher R, MacKenzie D, Zoepf S (2016) Racial and Gender Discrimination in Transportation Network Companies. Working Paper, National Bureau of Economics, Cambridge, MA.
Hadfield G (2016) *Rules for a Flat World: Why Human Invested Law and How to Reinvent It for a Complex Global Economy* (Oxford University Press, New York).
Hand DJ, Adams NM (2014) Selection bias in credit scorecard evaluation. *J. Oper. Res. Soc.* 65(3):408–415.
Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. *Adv. Neural Inf. Process. Syst.* (NIPS 2016, Barcelona, Spain),3323–3331.
Hellman D (2016) Two Concepts of Discrimination. *Virgina Law Rev.* 102(4):895–952.
Henley WE, Hand DJ (1997) Statistical Classification Methods in Consumer Credit Scoring: a Review. *J. R. Stat. Soc. Ser. A (Statistics Soc.* 160(Part 3):523–541.
Holstein K, Vaughan JW, Daumé H, Dudík M, Wallach H (2019) Improving fairness in machine learning systems: What do industry practitioners need? *Conf. Hum. Factors Comput. Syst. - Proc.* (CHI 2019, Glasgow, UK), 1–16.
Hsia D (1978) Credit Scoring and the Equal Credit Opportunity Act. *Hastings Law J.* 30(2):371.
Hurley M, Adebayo J (2016) Credit Scoring in the Era of Big Data. *Yale J. Law Technol.* 18(1):148–216.
Kamiran F, Calders T (2012) Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* 33(1):1–33.
Kleinberg J, Ludwig J, Mullainathan S, Rambachan A (2018) Algorithmic Fairness. *AEA Pap. Proc.* 108:22–27.
Kleinberg J, Ludwig J, Mullainathan S, Sunstein CR (2018) Discrimination in the Age of Algorithms. *J. Leg. Anal.* 10:113–174.
Kleinberg J, Ludwig J, Mullainathan S, Sunstein CR (2020) Algorithms as discrimination detectors. *Proc. Natl. Acad. Sci.* 117(48):30096–30100.
Lakkaraju H, Kleinberg J, Leskovec J, Ludwig J, Mullainathan S (2017) The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. (KDD 2017, Halifax, Canada), 275–284.
Lambrecht A, Tucker C (2019) Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Manage. Sci.* 65(7):2966–2981.
Lessmann S, Baesens B, Seow HV, Thomas LC (2015) Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *Eur. J. Oper. Res.* 247(1):124–136.
Lipton ZC, Chouldechova A, McAuley J (2018) Does mitigating ML’s impact disparity require treatment disparity? *32nd Conf. Neural Inf. Process. Syst. (NeurIPS 2018)*. 1–11.
Lundberg SM, Erion GG, Lee SI (2019) Consistent Individualized Feature Attribution for Tree Ensembles. Working Paper, University of Washington, Seattle.
Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst. 30*. (NIPS 2017, Long Breach, CA), 4768–4777.
Mejia J, Parker C (2021) When Transparency Fails: Bias and Financial Incentives in Ridesharing Platforms. *Manage. Sci.* 67(1):166-184.
Monetary Authority of Singapore (2018) Guide to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in the Singapore Financial Sector. Report, Monetary Authority of Singapore, Singapore.
Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 366(6464):447–453.
Ongena S, Popov A (2016) Gender Bias and Credit Access. *J. Money, Credit Bank.* 48(8):1691–1724.

Perrone V, Donini M, Kenthapadi K, Archambeau C (2020) Fair Bayesian Optimization. *7th ICML Work. Autom. Mach. Learn.* (AutoML 2020, Virtual), 1–15.
Pope DG, Sydnor JR (2011a) Implementing Anti-Discrimination Policies in Statistical Profiling Models. *Am. Econ. J. Econ. Policy* 3(3):206–231.
Pope DG, Sydnor JR (2011b) What's in a Picture? Evidence of Discrimination from Prosper.com. *J. Hum. Resour.* 46(1):53–92.
RRBM (2017) A Vision of Responsible Research in Business and Management: Striving for Useful and Credible Knowledge. Retrieved (July 17, 2021), https://rrbm.network/wp-content/uploads/2020/04/Position-Paper_revised_8April2020.pdf
Stein RM (2005) The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing. *J. Bank. Financ.* 29(5):1213–1236.
Tang CS, Yang SA, Wu J (2018) Sourcing from suppliers with financial constraints and performance risk. *Manuf. Serv. Oper. Manag.* 20(1):70–84.
Taylor W (2011) Proving Racial Discrimination and Monitoring Fair Lending Compliance: The Missing Data Problem in Nonmortgage Credit. *Rev. Bank. Financ. Law* 31:199–264.
Thomas LC, Edelman DB, Crook JN (2017) *Credit Scoring and Its Applications* 2nd ed. (Society for Industrial and Applied Mathematics Publishing, Philadelphia).
U.S. Department of Justice (2017) The Equal Credit Opportunity Act. Retrieved (November 9, 2018), https://www.govinfo.gov/content/pkg/USCODE-2011-title15/html/USCODE-2011-title15-chap41-subchapIV.htm.
Veale M, Edwards L (2018) Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling. *Comput. Law Secur. Rev.* 34(2):398–404.
Vigdor N (2019) Apple Card Investigated After Gender Discrimination Complaints - The New York Times. *New York Times*. Retrieved (April 18, 2021), https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html.
Wirth R, Hipp J (2000) CRISP-DM : Towards a Standard Process Model for Data Mining. *Proc. Fourth Int. Conf. Pract. Appl. Knowl. Discov. Data Min.* (24959):29–39.
Wooldridge JM (2015) *Introductory Econometrics A Modern Approach* 5th ed. (Cengage Learning, Boston)
Younkin P, Kuppuswamy V (2018) The colorblind crowd? Founder race and performance in crowdfunding. *Manage. Sci.* 64(7):3269–3287.
Zafar MB, Valera I, Rodriguez MG, Gummadi KP (2017) Fairness constraints: Mechanisms for fair classification. *Proc. 20th Int. Conf. Artif. Intell. Stat. AISTATS 2017* 54.
Zhang Y (2018) Assessing Fair Lending Risks Using Race/Ethnicity Proxies. *Manage. Sci.* 64(1):178–197.
Zliobaite I (2015) On the Relation Between Accuracy and Fairness in Binary Classification. *Proc. 2nd Work. Fairness, Accountability, Transpar. Mach. Learn.* (FATML 2017, Lille, France)
Žliobaitė I (2017) Measuring discrimination in algorithmic decision making. *Data Min. Knowl. Discov.* 31(4):1060–1089.
Žliobaitė I, Custers B (2016) Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artif. Intell. Law* 24(2):183–201.