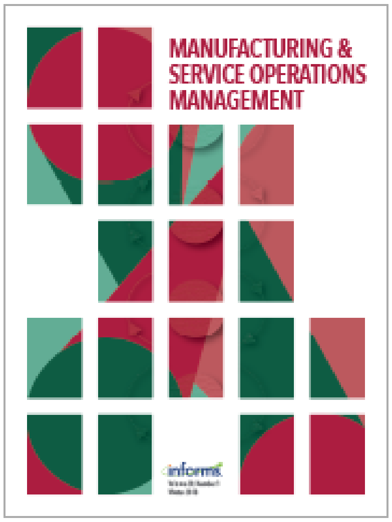


Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Antidiscrimination Laws, Artificial Intelligence, and Gender Bias: A Case Study in Nonmortgage Fintech Lending

Stephanie Kelley, Anton Ovchinnikov, David R. Hardoon, Adrienne Heinrich

informs
https://pubsonline.informs.org/journal/msom
MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 24, No. 6, November–December 2022, pp. 3039–3059
ISSN 1523-4614 (print), ISSN 1526-5498 (online)

Antidiscrimination Laws, Artificial Intelligence, and Gender Bias: A Case Study in Nonmortgage Fintech Lending

Stephanie Kelley,[a,*] Anton Ovchinnikov,[a,b] David R. Hardoon,[c] Adrienne Heinrich[c]

[a] Smith School of Business, Queen's University, Kingston, Ontario K7L 3N6, Canada; [b] INSEAD, 77300 Fontainebleau, France; [c] Artificial Intelligence and Innovation Center of Excellence, Union Bank of the Philippines, and Aboitiz Data Innovation, Pasig City 1605, Philippines
*Corresponding author

Contact: stephanie.kelley@queensu.ca, https://orcid.org/0000-0002-2184-4507 (SK); anton.ovchinnikov@queensu.ca, https://orcid.org/0000-0001-5972-2217 (AO); drhardoon@unionbankph.com (DRH); aheinrich@unionbankph.com (AH)

Received: October 1, 2020
Revised: April 30, 2021; September 24, 2021; December 9, 2021
Accepted: January 11, 2022
Published Online in Articles in Advance: May 18, 2022

https://doi.org/10.1287/msom.2022.1108

Abstract. ***Problem definition*:** We use a realistically large, publicly available data set from a global fintech lender to simulate the impact of different antidiscrimination laws and their corresponding data management and model-building regimes on gender-based discrimination in the nonmortgage 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 multicollinear data. We provide technically and legally permissible approaches for firms to reduce discrimination across different antidiscrimination regimes whilst managing profitability. ***Methodology*:** We train statistical and ML models on a large and realistically rich publicly available data set to simulate different antidiscrimination regimes and measure their impact on model quality and firm profitability. We use ML explainability techniques to understand the drivers of ML discrimination. ***Results*:** We find that regimes that prohibit the use of gender (like those in the United States) substantially increase discrimination and slightly decrease firm profitability. We observe that ML models are less discriminatory, of better predictive quality, and more profitable compared with 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 selection, 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 antidiscrimination laws, specifically with respect to the collection and use of protected attributes for ML 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.

History: This paper has been accepted for the *Manufacturing & Service Operations Management* Special Section on Responsible Research in Operations Management.

Supplemental Material: The online appendix is available at https://doi.org/10.1287/msom.2022.1108.

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 of discrimination from consumers and the media. The reports refer to discrimination as being 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 feel they are entitled to be treated

Manufacturing & Service Operations Management, 2022, vol. 24, no. 6, pp. 3039–3059, © 2022 The Author(s)

(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. Although an ethically centered definition is salient, most countries have adopted antidiscrimination 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; 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" (Vigdor 2019). 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 United States (U.S.) Equal Credit Opportunity Act (ECOA) prohibits the use (and even collection) of protected attributes, like gender, for nonmortgage lending decisions, and a report by the New York State Department of Financial Services (2021) found no violations of fair lending by the Apple Card venture. The outcome, however, is paradoxical; adhering to the existing antidiscrimination laws and the data management and model-building practices they imply could produce ethically problematic outcomes.

The antidiscrimination laws with respect to gender and credit differ across countries. Their exact language varies greatly and is not a subject of this investigation, but their data management and model-building guidance imply three main regimes (discussed in detail in Online Appendix S1).

1. Regime 1 (e.g., Singapore) allows for the collection and use of gender data in AI models.

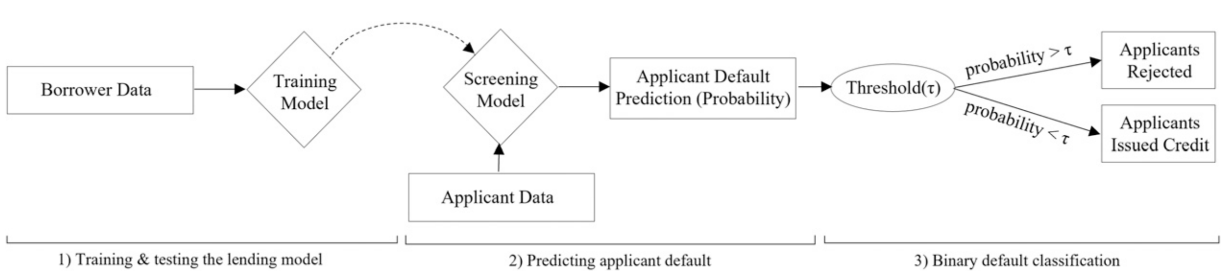
2. Regime 2 (e.g., the European Union (EU)) allows for the collection of gender but prohibits the use of gender as a feature in the training and screening models used for individual lending decisions.

3. Regime 3 (e.g., the United States (US)) prohibits the collection and thus, also the use of gender data.

In this paper, we use a realistically large publicly available data set 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 2022), given that the majority of global antidiscrimination 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 because of its universality; our choices do not in any way diminish the need to investigate discrimination in other contexts and with other data sets.

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). Because 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) versus tree-based ensembles) may not necessarily be calibrated (i.e., a predicted score of 0.3 from one model may not be equivalent to 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 with 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).

Figure 1. Schematic View of the Three-Step Fintech Lending Modeling Process



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. However, because a firm cannot possibly know the payment outcomes for a rejected applicant, we must use an alternative measure using available data. We use two such measures of discrimination between models: 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 with another group, and within-group mean difference (WMGD) (Žliobaitė 2017), the difference in the mean predicted default rate for the protected class between models.

A key observation that underlies our investigation is as follows. Under regime 1, a firm's training and screening models have access to gender and the numerous machine learning (ML) consequences associated with it, such as feature engineering, hyperparameter tuning, etc. (see Sections 1.2 and 6.2), whereas under regime 2, the use of gender is restricted; under regime 3, it is prohibited entirely. These regimes, therefore, lead to two model types: model 1 (with gender per regime 1) and model 2 (without gender per regimes 2 and 3). We specify what the "model" is in Section 3.1, but importantly, models 1 and 2 differ in that their applicant default probability predictions vary; consequently, their optimal thresholds differ as well. Hence, the exact same applicant could be issued credit under one regime and 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.

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

1.1. Impact of Antidiscrimination Regimes on Gender-Based Discrimination

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, however, 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 PPV) (Section 3.1) in the top-performing ML model ("average blender" (AB), which is discussed in Section 4.3) trained on our data. This finding is summarized in Observation 1 in Section 5.

The paradoxical discriminatory effects of antidiscrimination regimes have been investigated before in both computer science and financial economics. In the computer science literature, 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 multicollinear data sets) to show that, absent legal constraints, protected attributes should be included to reduce discrimination and improve predictive quality (Žliobaitė and Custers 2016, Lipton et al. 2018). We support this conceptual conclusion with results from a realistic data set and a modeling process that mimics a fintech's operations.

In the financial economics literature, Chandler and Ewert (1976) evaluate the ECOA and find that the operational modeling guidance, which prohibits the use and collection of gender, creates a detrimental increase in the rejection rates of women compared with 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 with men versus models that include gender. In Section 5, we extend these analyses into the modern ML setting, exploring multiple nuances stemming from the use of ML methods. For robustness, we also replicate the statistical approach of Andreeva and Matuszyk (2019) on our data in Online Appendix S2.

1.2. Drivers of Statistical and Machine Learning Discrimination

What drives discrimination? Moreover, 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 Section 4.3) 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 credit worthy group, which in turn, disproportionately increases the rejection rates of women. Using a conceptual modeling framework, Žliobaitė and Custers (2016) and Kleinberg et al. (2018) similarly show 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, however, makes several simplifying assumptions that are not true for a fintech that uses modern ML. Kleinberg et al. (2020) extend beyond OVB and suggest that algorithmic discrimination 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. Utilizing this framework, we detail how the specifics of a fintech's operations could introduce discrimination within these three elements.

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 does not introduce discrimination through the choice of outcome measure.

2. Choice of input variables. The various antidiscrimination 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).

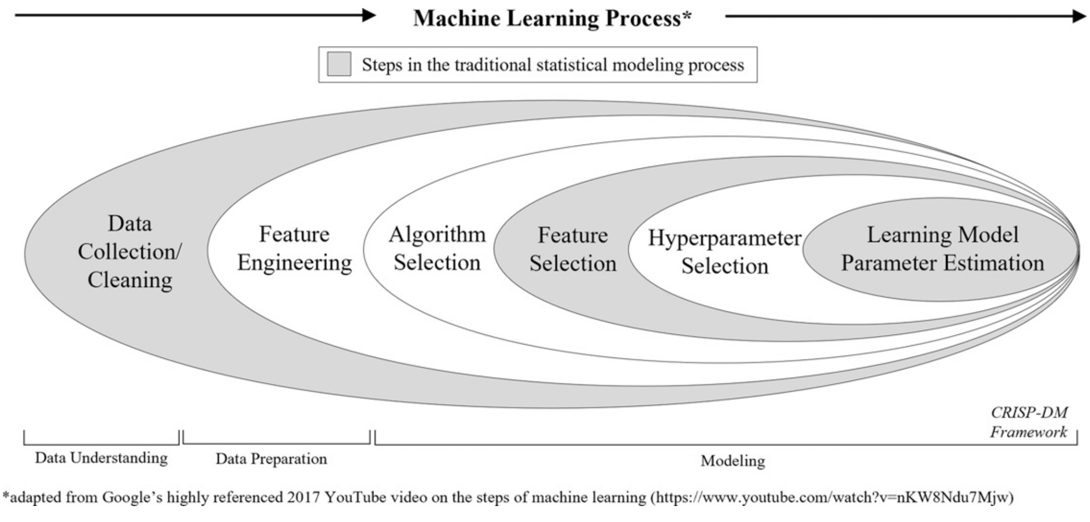
3. Construction of the model training procedure. Recall from Figure 1 that borrower data are used to train the training model, a process that can be summarized at a more granular level by the crossindustry standard process for data mining (CRISP-DM) (Wirth and Hipp 2000), the most commonly used model training procedure in modern ML that is employed by our fintech industry partners. Figure 2 depicts the details of CRISP-DM and the additional modeling steps introduced into the training procedure (white in Figure 2). The removal of gender, per regimes 2 and 3, affects each of these additional steps and introduces discrimination that is not captured by OVB.

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 Section 6.2, permutation importance analysis shows that approximately 20% of the most impactful features in model 1 (with gender) are engineered using gender. In contrast, these features 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, leading to differences in discrimination between algorithms.

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; when gender is excluded, however, it may exclude "age." Indeed, in Section 6.2, we observe that, when gender is excluded, the algorithm also excludes certain features, which we refer to as gender reliant. In their place, the algorithm selects other features, which we refer to as gender redundant. Using Shapley additive explanations (SHAP) values and SHAP interaction values (Lundberg and Lee 2017, Lundberg et al. 2019), we find that the gender-reliant features are on average 19

Figure 2. The Six-Step Machine Learning Modeling Process Compared with the Three-Step Traditional Statistical Modeling Process Arranged According to the CRISP-DM Framework



times more important for women compared with men. As such, when gender is excluded, so too are the gender-reliant features, thereby 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. Such parameters 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 Section 7 show that selecting the hyperparameters when gender is available can change model predictions and reduce 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 is 44.06% less discriminatory on average (measured by PPV), of significantly better predictive quality (+472 basis points (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 it is on average 7.60% more profitable compared with the LR model. These findings are summarized in Observations 2–5 in Section 5 and Observations 6–8 in Section 6. Together, they illustrate that both firms and applicants should prefer ML models over traditional statistical models in the nonmortgage consumer fintech lending setting, as ML allows the model to partially recover the negative impact of excluding gender.

1.3. Possible Approaches to Reduce Machine Learning Discrimination

In Section 7, we evaluate four 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 preprocessing (Kamiran and Calders 2012). Observations are randomly removed from the majority class (men) until counts are equal with the minority group (women). Doing so 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 the gender of past borrowers. However, applicant gender is not used in the screening model. This approach is model agnostic and similar to the fair Bayesian optimization technique (Perrone et al. 2021). 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, and −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, and +0.13% average profitability), this approach is prohibited in the United States, an example of regime 3 (Chen et al. 2019).

These findings are summarized in Observations 9–12 in Section 7. Note that 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 the 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 to empirical studies on technology-based business operations (e.g., Cui et al. 2018, Cohen and Harsha 2020) and the discrimination in which they proliferate: crowdfunding (Pope and Sydnor 2011a, b; Younkin and Kuppuswamy 2018), online auctions (Doleac and Stein 2013), social networks (Acquisti and Fong 2020), ride-sharing (Ge et al. 2020, 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 because of 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 that it is important to bring forward the topics of gender equality and discrimination for operations management researchers because they are aligned with both the Vision of the Responsible Research in Business and Management (Cofounders for RRBM 2017) and the United Nation's Sustainable Development Goals (Online Appendix S1).

2.2. Discrimination in Nonmortgage 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. 2022, Fuster et al. 2022) because of data availability (Taylor 2011), which differs from our nonmortgage context in three ways. First, the lenders' operating models are different. Most mortgage fintech firms are intermediaries that 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 that make such decisions use variations of the Fair Issac score and logistic regression models, 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 (1976) evaluate the impact of the ECOA in 1979 and Andreeva and Matuszyk (2019) evaluate the impact of the EU Gender Directive from 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 when compared with models that use gender. Both of these works use proprietary 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 neither measure the impact on firm profitability nor provide recommendations for firms to reduce discrimination. Our unique public data and modern ML approach addresses these shortcomings, making our findings more operationally relevant for fintech firms, regulators, and the public across several regulatory regimes.

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. Kleinberg et al. (2018) and Lipton et al. (2018) explore the impact of U.S. antidiscrimination 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 antidiscrimination laws and arrive at a similar conclusion. Like Kleinberg et al. (2018), they conceptually explain the drivers of algorithmic discrimination using the OVB framework. Although 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 data set with 52 observations and six variables. Although they admit it is small, it is difficult to extend their OVB findings to a true operational, high-dimensional, feature-rich, and highly multicollinear data set 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 that neither the choice of outcome measure (default probability) nor the input variables (specifically with respect to gender) change.

Further, although several other works suggest a range of solutions to reduce discrimination through preprocessing (e.g., Kamiran and Calders 2012; Chen et al. 2018, 2019), in processing (e.g., Zafar et al. 2019, Perrone et al. 2021), and postprocessing (e.g., Hardt et al. 2016), practitioners reported "struggling to apply existing auditing and de-biasing methods in their contexts" (Holstein et al. 2019, p. 2) and found there were limited "domain-specific education resources, metrics, processes, and tools," (Holstein et al. 2019, p. 9) as the majority of computer science studies focus on nonbusiness contexts, such as recidivism. 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 Section 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 (Žliobaitė and Custers 2016; Kleinberg et al. 2018, 2020; Lipton et al. 2018). We extend this understanding to a realistic context that mimics situations faced by fintech lenders in practice, where advanced nonregression techniques are used with high-dimensional, feature-rich, highly multicollinear 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 anticlassification (Berk et al. 2017). Furthermore, it is well known (Chouldechova 2017, Kleinberg et al. 2018) 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 these 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; see Section 1);

2. Be calculated without an external risk score (as there are no such risk scores available; credit scores are not valid external scores as they are used as a feature in the model); and

3. Be calculated with a known default outcome label (which 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: PPV and WGMD. PPV satisfies all three conditions, whereas 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 these conditions is that requiring a known default outcome label (condition (3))

Table 1. Summary of the Discrimination Measure Selection Rationale

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. (2018)		X	X
Positive predictive value	Chouldechova (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		
Anticlassification	Grgic-Hlaca et al. (2016)		X	

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 label problem with a sampling procedure called augmentation (Hsia 1978), as discussed in Section 4.2. However, 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.

Positive predictive value (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, $TP_M(\tau)$) divided by all default predictions for men (i.e., the sum of $TP_M(\tau)$ and $FP_M(\tau)$, the number of false-positive predictions for men) minus the same ratio for women:

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

A $PPV(\tau)$ greater than zero denotes bias against women, less than zero denotes bias against men, and equal to zero 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 noncomparative 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.

Within-group mean difference (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 data set (as made by some model at some threshold τ) and N_W denotes the total number of women in the data, then

$$WGMD(\tau) = \left[\frac{\sum_{i=1}^{N_W} \hat{Y}_{iW}(\tau) = 1}{N_W}\right]_{ModelA} - \left[\frac{\sum_{i=1}^{N_W} \hat{Y}_{iW}(\tau) = 1}{N_W}\right]_{ModelB}. \tag{2}$$

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. We report the means and 95% confidence intervals (95% CIs) across the 30-fold crossvalidation. Both values are 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. We then proceed with a paired t test if the differences are normally distributed and a paired samples Wilcoxon test if they are not. An α of 0.05 is used for all tests.

3.2. Model Predictive Quality Measure

We measure *model predictive quality* using the 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 data sets 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 2,000 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 does default (a false-negative prediction). We assume a firm is not impacted by applicants 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, $FN(\tau)$:

$$\pi(\tau) = R \times TN(\tau) - C \times FN(\tau). \tag{3}$$

To examine different operating scenarios, we consider 2,431 cost-to-revenue (C:R) pairs, covering the full range of reported ratios (up to 35×) 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, we 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

Given the choice of key metrics (discrimination, predictive quality, and firm profitability), we needed to source data that had both gender and known default outcomes for all observations. Real-world applicant data do not, however, have complete known default outcomes. 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 have complete default outcome information. However, doing so could introduce a bias as the borrower population may 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. Furthermore, privacy regulations restricted us from gathering the outcomes of rejected applicants as the data are anonymized. We, therefore, selected augmentation as it is model agnostic and relies only on resampling of the testing data. This approach prevented additional noise from being introduced to the outcomes, which could have occurred if extrapolation was used. We followed the well-defined 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 nine 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 life cycles (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 our data. At the time of writing, the data were available at https://www.kaggle.com/c/home-credit-default-risk/data. We used the competition training data sets (gathered into one file with observations for each borrower) but excluded the testing data set, 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; Home Credit, however, granted us permission to use the data for this study. Additional exploration of the data and replication of our results are possible through participation in the Kaggle competition.

4.2. Sampling: Reject Inference—Augmentation

It is critical to note that, because our data come from a Kaggle competition, they exclude 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 problem, we applied a reject inference technique called *augmentation* (Hsia 1978), a common method used in both research (Hand and Adams 2014) and industry. Augmentation adjusts the borrower data (which importantly, include 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.

Next, 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 α of 0.05) in their distribution across income and occupation, whereas gender, age, and marital status were comparable. Using this proportional difference information, we down sampled the borrower data accordingly to create a "look-alike" applicant data set with complete default outcomes, which helps to address the selective labels problem (see the schematic in Figure 3). The technical details and accompanying code for the augmentation sampling procedure can be found at the authors' GitHub: https://github.com/stephaniekelley/genderbias.

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 data set 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 because of 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). Although these selection steps are operationally relevant to the firm, they remain beyond the scope of this investigation as they require exceedingly rare data—data that may not even be collected by firms—to analyze.

To replicate the training and testing steps from the lending modeling process (Figure 3) throughout the paper, we created two samples from the Home Credit Borrower data (the sampling procedure is summarized in Figure 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.

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 data sets available to fintech lenders globally (Ongena and Popov 2016). These data are used throughout the paper for training models in the main analysis (Sections 5 and 6).

2. From the borrower testing data, we generated look-alike applicant testing data (66.2% women), created using the augmentation sampling procedure (discussed in Section 4.2) to reduce the selective labels problem. These data are used throughout the paper as input for the screening model to obtain "out-of-sample" predictions, providing a consistent comparison across key metrics.

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 Section 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.

We use LR as the "traditional statistical" model as it is the preferred model of lenders (Thomas et al. 2017) and was used in past lending discrimination studies. To generate an LR model, a fintech firm would follow the traditional statistical modeling process (introduced in gray in Figure 2) by first collecting and cleaning the data. The data were originally used for an 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 as follows. 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 then 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 an LR model and manually selected features that were significant at an α of 0.05 in the model with gender; then, we excluded gender from this data set to generate the genderless feature set. These features were then used to train the LR training model, resulting in the learning model parameter estimation, the results of which are discussed in Section 5 onward.

The ML process introduces three additional steps: feature engineering, algorithm selection, and hyperparameter tuning (recall Figure 2 and Section 1.2). We started with the same 122 features (or 121 in regime 3) from the LR

Figure 3. Schematic View of the Integration of the Augmentation Sampling Procedure into the Three-Step Fintech Lending Modeling Process

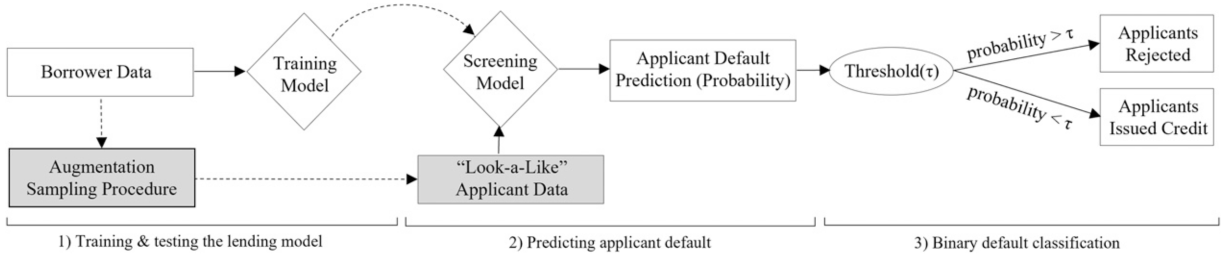
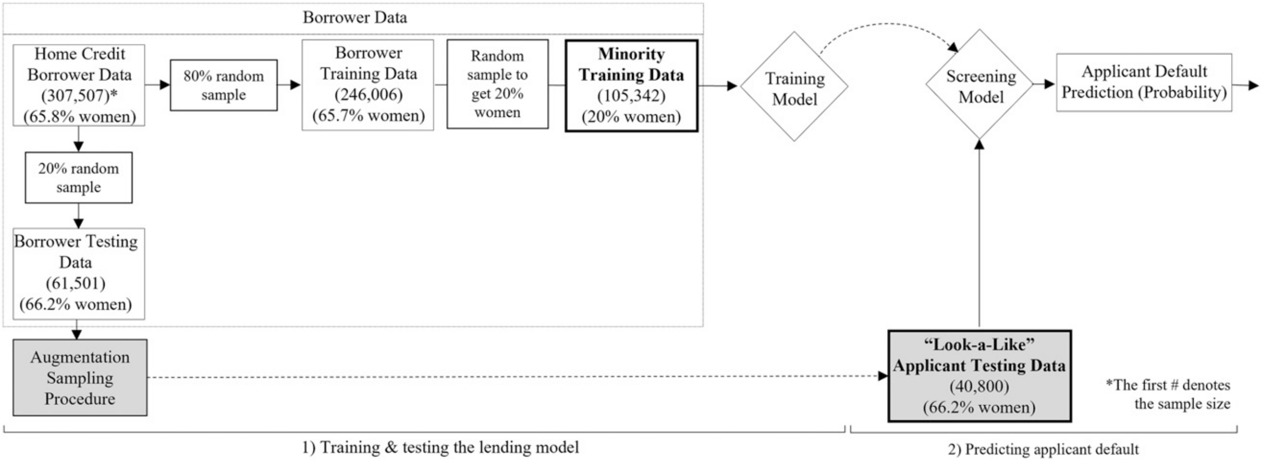


Figure 4. Schematic View of the Sampling Procedure Incorporated into Steps 1 and 2 of the Three-Step Fintech Lending Modeling Process



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 generated 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 support vector machine, 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 (measured by the five-fold crossvalidated AUC reported in DataRobot, a metric we discuss in Section 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 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 with 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 next tuned hyperparameters, modeling values used to further improve the predictive quality of the chosen algorithm, and then, the final learning model parameter estimation occurred, the results of which are discussed in Section 5 onward.

In Section 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, rather than multiclass 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 that 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 biased models possible. The codes for this model and for our LR model also trained in R are available on GitHub: https://github.com/stephaniekelley/genderbias.

5. Impact of Antidiscrimination Regimes on Gender-Based Discrimination

In this section, we compare the impact of the three regimes on both ML and LR models and make five observations (Observations 1–5), which are summarized in Table 2.

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

Table 2. Observations 1–5: A Comparison of Lending Models Across Discrimination (PPV and WGMD), Predictive Quality (AUC), and Firm Profitability (Average Optimal Profit)

Observation	Model	Discrimination (PPV)	Discrimination (WGMD)	Predictive quality (AUC)	Firm profitability (average optimal profit)
Compared with the average blender model with gender (AB:M1)					
Observation 1	AB:M2	+285.04% [51.72%–1,011.46%][a] (see Figure 5(a))	+34.75% [17.37%–42.93%][a] (see Figure 5(b))	Not impacted: 77.97% [77.16%–78.79%] vs. AB:M1: 78.06% [77.24%–78.88%]	−0.25%[b]
Compared with the logistic regression model with gender (LR:M1)					
Observation 2	AB:M1	−44.06% [−127.53%–+2.60%][c] (see Figure 6(a))	Inapplicable[d]	+472 bps: 78.06% [77.24%–78.88%] vs. LR:M1: 73.34% [72.46%–74.22%]	+7.86%[e]
Observation 3	AB(STAT):M1	−16.32% [−54.66%–+56.75%][a] (see Figure 6(a))	Inapplicable[d]	+192 bps: 75.26% [74.41%–76.12%] vs. LR:M1: 73.34% [72.46%–74.22%]	+3.92%[f]
Compared with the logistic regression model without gender (LR:M2)					
Observation 4	AB:M2	−9.56% [−36.87%–+13.15%][a] (see Figure 6(b))	Inapplicable[d]	+487 bps: 77.97% [77.16%–78.79%] vs. LR:M2 73.10% [72.22%–73.98%]	+7.60%[g]
Observation 5	AB(STAT):M2	Is of comparable discrimination (see Figure 6(b))	Inapplicable[d]	+200 bps: 75.10% [74.24%–75.59%] vs. LR:M2: 73.10% [72.22%–73.98%]	+4.19%[h]

[a]Across thresholds (5%–30%), the differences are statistically significant ($p < 0.01$, paired t test; see Section 3.1 for details).
[b]11.11% of the profit differences are statistically significant, 23% are negative differences (−2.43% to −1.19%), and 77% are positive (0.02%–23%).
[c]Across thresholds (5%–30%), the differences are statistically significant ($p < 0.01$, paired samples Wilcoxon test).
[d]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 Section 1 for details).
[e]89.88% of the profit differences are statistically significant, 98% are positive differences (0.04%–68.20%), and 2% are negative (−0.006% to −0.005%).
[f]82.06% of the profit differences are statistically significant, 93% are positive differences (0.06%–51.10%), and 7% are negative (−0.006% to −0.005%).
[g]90.58% of the profit differences are statistically significant (0.02%–63.50%).
[h]77.87% of the profit differences are statistically statistically significant (0.03%–51.30%).

Section 4.3 and train a model on the minority training data (discussed in Section 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), whereas those under regimes 2 and 3 are restricted in the use of gender as a feature (resulting in model 2).

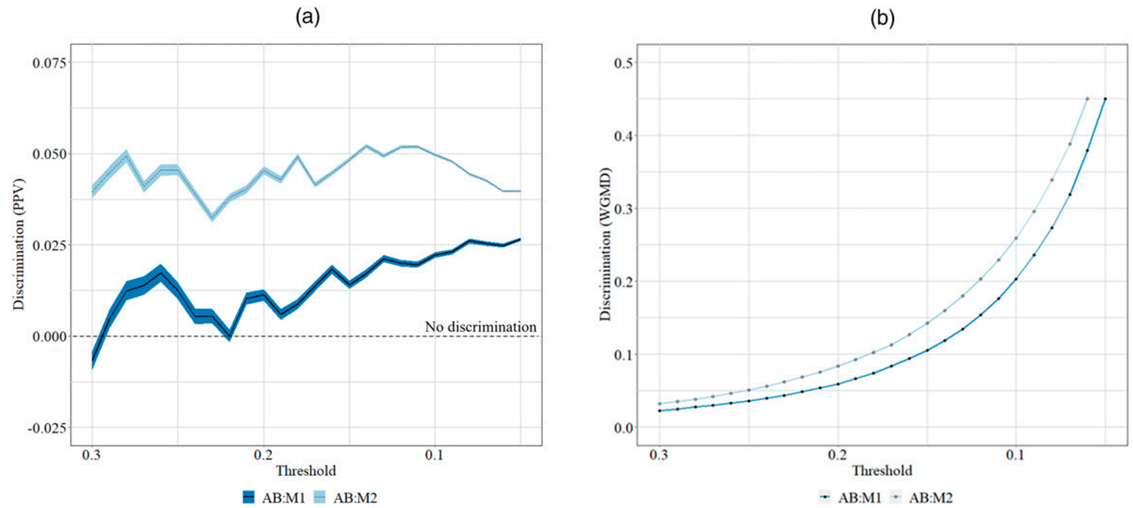
Observation 1. We start by examining the impact of removing gender from the average blender model (AB:M2) compared with the same model with gender (AB:M1). We find that discrimination measured by PPV increases by 285.04%, predictive quality is not impacted, and firm profitability decreases by 0.25% (see Observation 1 and Table 2). Discrimination is visualized in Figure 5(a) for PPV and Figure 5(b) for WGMD. The key implication is that 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 with 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 Online Appendix 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

Figure 5. (Color online) (a) Discrimination Measured by PPV of AB Models with Gender (AB:M1) and Without Gender (AB:M2). (b) Discrimination Measured by WGMD of AB Models with Gender (AB:M1) and Without Gender (AB:M2)



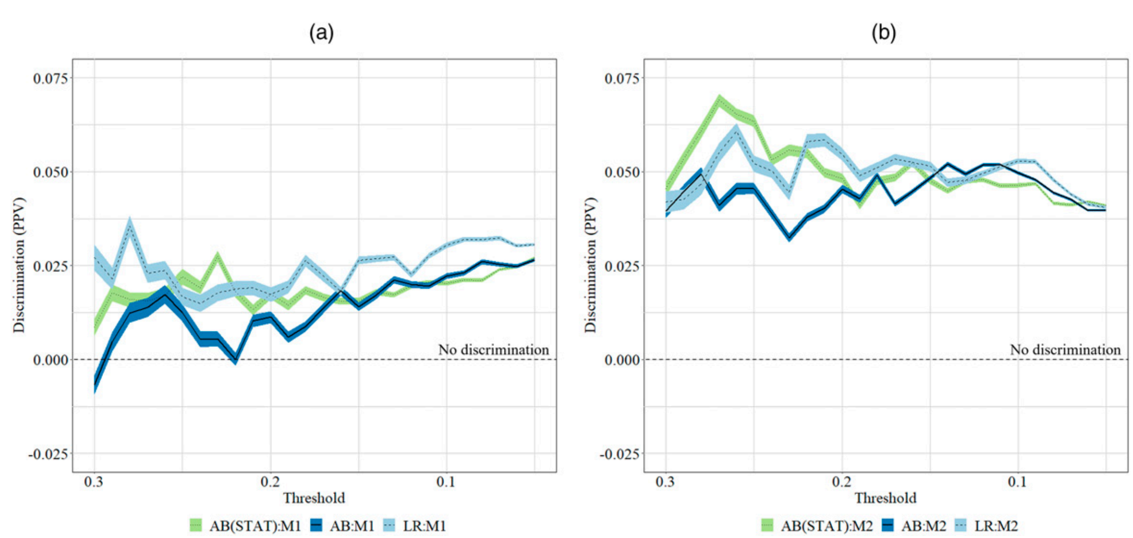
firms would be better off (i.e., observe lower levels of discrimination, higher predictive quality, and higher firm profitability) using LR or ML models. To do so, we compare the LR and AB models, with gender (regime 1) and without gender (regimes 2 and 3), trained on the minority training data.

Observation 2. We compare the average blender model with gender (AB:M1) to the logistic regression model with gender (LR:M1) and find that discrimination measured by PPV decreases by 44.06%, predictive quality increases by 472 bps, and firm profitability increases by 7.86% (see Observation 2 and Table 2). Discrimination (PPV) is visualized in Figure 6(a). The key implication is that when gender is used in the training and screening models (per regime 1), the AB model is less discriminatory, is of better predictive quality, and produces greater profitability than the LR model.

From our understanding of the traditional statistical modeling and ML modeling processes (Section 4.3), we know that the results from Observation 2 occur for two reasons: 1) a change in the model (LR to AB) and 2) access to a larger ML feature set.

Observation 3. For robustness, we compare the results of a second AB model trained on the smaller traditional statistical feature set used by the LR model (AB(STAT):M1) with the LR model with gender

Figure 6. (Color online) (a) Discrimination Measured by PPV of LR and AB Models with Gender (LR:M1, AB:M1) and the AB Model Trained on the Traditional Statistical Data Set with Gender (AB(STAT):M1). (b) Discrimination Measured by PPV of LR and AB Models Without Gender (LR:M2, AB:M2) and the AB Model Trained on the Traditional Statistical Data Set Without Gender (AB(STAT):M2)



(LR:M1). This comparison allows us to observe the impact of a change in the model (LR to AB) by itself. We find that this change from LR to AB decreases discrimination measured by PPV by 16.32%, increases predictive quality by 192 bps, and increases firm profitability by 3.92% (Observation 3 and Table 2). Discrimination (PPV) is visualized in Figure 6(a). The key implication is that even on the traditional statistical feature set with fewer engineered features, the AB model is less discriminatory, is of better predictive quality, and produces greater profitability than the LR model. This finding suggests that under regime 1, both lenders and applicants should prefer ML over LR models.

Observation 4. 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. To do so, we compare the impact of using an average blender model without gender (AB:M2) versus a logistic regression model without gender (LR:M2). We find that discrimination measured by PPV decreases by 9.56%, predictive quality increases by 487 bps, and firm profitability increases by 7.60% (Observation 4 and Table 2). Discrimination (PPV) is also visualized in Figure 6(b). The key implication is that without access to gender (per regimes 2 and 3), both firms and customers should prefer the AB model as it is less discriminatory, is of better predictive quality, and produces greater firm profitability compared with LR.

Observation 5. For robustness, as we did with model 1, we also compare an AB model trained on the traditional statistical feature set used by the LR model without gender (AB(STAT):M2) with an LR model without gender (LR:M2). We find that discrimination measured by PPV is comparable, predictive quality increases by 200 bps, and firm profitability increases by 4.19% (Observation 5 and Table 2). The impact on discrimination (PPV) is illustrated in Figure 6(b). The key implication is that although the AB model trained on the traditional statistical feature set used by the LR model is of better predictive quality and greater profitability, it is of comparable discrimination to the LR model without gender, as it does not have access to the full feature engineered data set.

In aggregate, Observations 2–5 illustrate that, trained on the full feature engineered data set, the AB model is less discriminatory and of greater predictive quality and profitability compared with the LR model both when gender is included and when gender is excluded. Trained on the nonengineered, traditional statistical feature set, the AB model is of better predictive quality and greater profitability compared with the LR model, regardless of whether gender is included; however, it is of comparable discrimination, suggesting that feature engineering has a significant impact on reducing discrimination. This finding 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 Section 4.3), and the inclusion of gender.

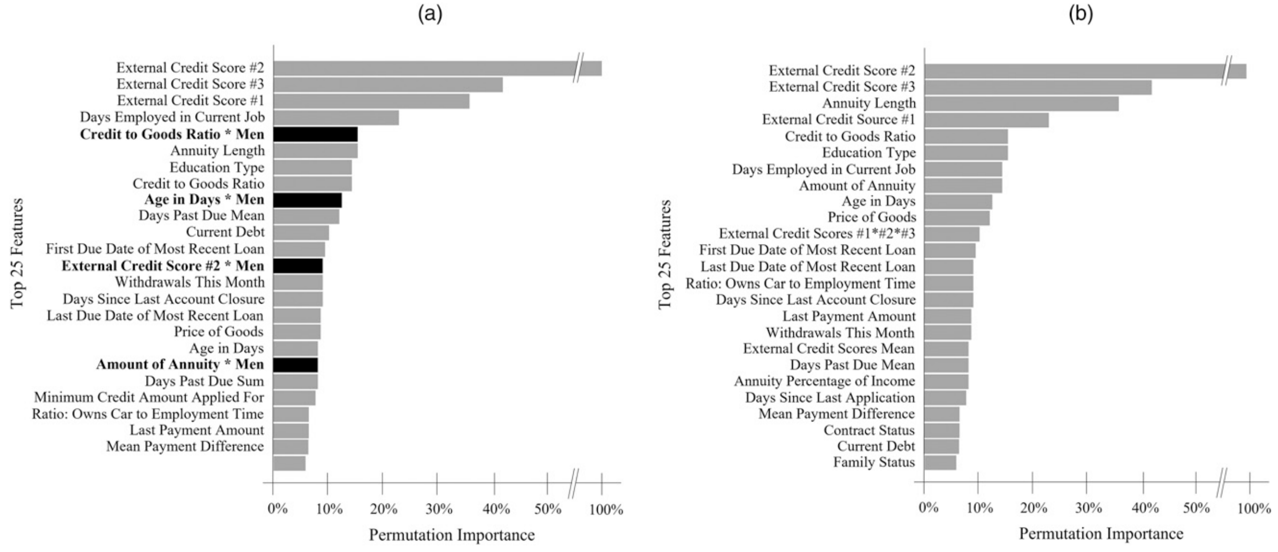
6.2. Using ML Explainability Techniques to Uncover the Drivers of ML Discrimination

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: (1) permutation importance to understand the impact of excluding gender in the AB model and (2) SHAP values and SHAP interaction values (Lundberg and Lee 2017, Lundberg et al. 2019) in our state-of-the-art XGBoost model (Section 4.3). This is because SHAP values and SHAP interaction values cannot be created for model ensembles. Before proceeding to the discussion of ML discrimination, we review OVB, which drives LR discrimination.

In the traditional statistical modeling process (per Figure 2), it is well known that, when an LR model has access to a comprehensive set of causal features, it can estimate the true, unbiased learning model parameters (Wooldridge 2015). Therefore, 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 (Wooldridge 2015). Andreeva and Matuszyk (2019) show that when gender is excluded from an LR model, it creates OVB. Specifically, because men, the less creditworthy gender, are a majority, the exclusion of gender creates an upward bias in the parameter estimates, leading to an increase in the rejection rates of women compared with the model with gender. As expected, OVB occurs in our data too (Online Appendix S2), where we replicate the approach from Andreeva and Matuszyk (2019).

Recall (from Section 1.2) that the ML modeling process (Figure 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 in turn, changes 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 Section 7. We first compute the

Figure 7. (a) Permutation Importance of the Top 25 Features of the AB Model with Gender and Gender Interactions (M1). (b) Permutation Importance of the Top 25 Features of the AB Model Without Gender (M2)



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 is illustrated in Figure 7, (a) and (b), 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 features and 2 of the top 10 features (Figure 7(a)); 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 with when gender is included (Figure 7(b)).

Results from Observations 6 and 7 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 ensemble model).

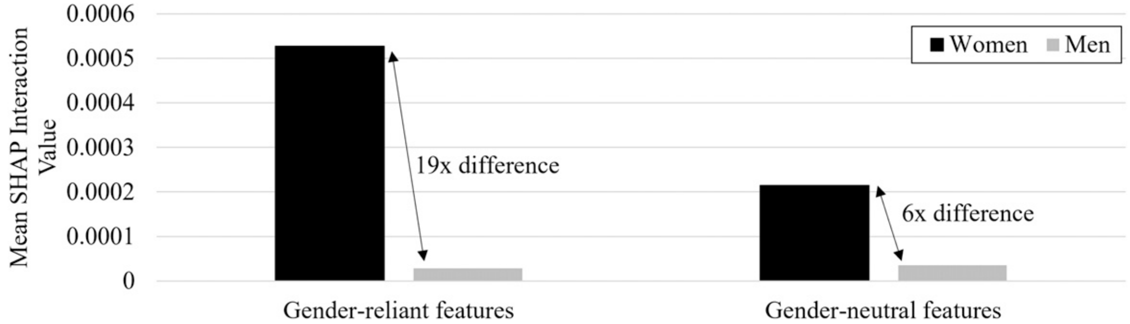
We observe that certain features, like external credit score 2 and annuity length, are always selected by the algorithm, regardless of whether gender is included; we refer to these as *gender-neutral* features. These features consistently have the highest SHAP values (feature importance) and are very 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 these features as *gender reliant*.

The final set of features is *gender redundant*. When gender is present, these features are "redundant" and excluded by the algorithm in the automated feature selection step; when gender is excluded, however, they are selected by the algorithm. We observe that these features have very low SHAP values compared with gender-neutral features and higher gender inference information compared with 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 between gender and the gender-neutral features visualized in Figure 8.

Observation 8. The mean SHAP interaction values for the gender-reliant features with gender are 19 times

Figure 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



greater for women than for men; this finding is 2.5 times the difference of gender-neutral features.

This observation illustrates that predictions for women rely more on the *gender-reliant* features than predictions for men, and therefore, women are more detrimentally impacted by the gender exclusion enforced by regimes 2 and 3. Summarizing, in ML, the exclusion of gender 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 with men, thereby increasing discrimination. This gender-blind feature selection phenomenon is vastly different from the OVB that drives discrimination in traditional statistical models; we show the ML discrimination is linked to changes in the construction of the model training procedure. This discussion is limited to the data set that we use, and although that data set is from a real fintech firm, to emphasize the generalizability of our insights, we provide a stylized example in Online Appendix S3.

7. Possible Approaches to Reduce Discrimination

Finally, we consider what ethically minded fintech firms can do to reduce gender discrimination given the restrictions of the antidiscrimination 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 (Section 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 to reduce discrimination for firms under regime 2:

1. down sampling the training data to rebalance gender (i.e., undersampling 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) and 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 (we emulate this 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 retrain the AB model using these three techniques and compare the results with the AB model without gender (AB:M2).

Observation 9. We find that down sampling the training data to rebalance gender (DS:M2) decreases discrimination measured by PPV by 4.54%, decreases predictive quality by 175 bps, and decreases firm profitability by 4.47% (Observation 9 and Table 3).

Observation 10. We observe that gender-aware hyperparameter tuning (HT:M2) decreases discrimination measured by PPV by 37.73%, decreases predictive quality by 278 bps, and decrease firm profitability by 4.42% (Observation 10 and Table 3).

Observation 11. Lastly, we find that up sampling the training data to rebalance gender (US:M2) decreases

Table 3. Observations 9–12: A Comparison of Possible Approaches to Reduce Discrimination (PPV and WGMD) Accompanied by Predictive Quality (AUC) and Firm Profitability (Average Optimal Profit) Changes

Observation	Model	Discrimination (PPV)	Discrimination (WGMD)	Predictive quality (AUC)	Firm profitability (average optimal profit)
Compared with the average blender model with gender (AB:M1)					
Observation 9	DS:M2	−4.54% [−19.38%–+14.03%][a] (see Figure 9(a))	−21.50% [−37.15% to −1.51%][a] (see Figure 9(b))	−175 bps: 76.22% [75.37%–77.07%] vs. AB:M2: 77.97% [77.16%–78.79%]	−4.47%[b]
Observation 10	HT:M2	−37.73% [−75.84%–+9.67%][a] (see Figure 9(a))	−30.37% [−47.01% to −6.71%][c] (see Figure 9(b))	−278 bps: 75.19% [74.33%–76.04%] vs. AB:M2: 77.97% [77.16%–78.79%]	−4.42%[d]
Observation 11	US:M2	−24.47% [−59.76%–+2.79%][a] (see Figure 9(a))	−40.08% [−43.35% to −44.85%][c] (see Figure 9(b))	Not impacted: 77.05% [76.22%–77.88%] vs. AB:M2: 77.97% [77.16%–78.79%]	−1.46%[e]
Observation 12	PGP:M2	−71.09% [−112.05% to −35.15%][a] (see Figure 9(a))	−25.74% [−31.20% to −13.89%][c] (see Figure 9(b))	Not impacted: 78.10% [77.28%–78.92%] vs. AB:M2: 77.97% [77.16%–78.79%]	+0.13%[f]

[a]Across thresholds (5%–30%), the differences are statistically significant ($p < 0.01$, paired t test; see Section 3.1 for details).
[b]−50.40% to −0.01%.
[c]Across thresholds (5%–30%), the differences are statistically significant ($p < 0.01$, paired samples Wilcoxon test).
[d]−59.30% to −0.01%.
[e]All of the profit differences are statistically significant: 88% are negative differences (−35.90% to −0.1%), and 12% are positive differences (0.01%–0.03%).
[f]0.01%–2.75%.

discrimination measured by PPV by 24.47%, does not impact predictive quality, and decreases firm profitability by 1.46% (Observation 11 and Table 3). Accompanying discrimination results are depicted in Figure 9(a) for PPV and Figure 9(b) for WGMD.

Taken together, the key insight is that although fintech lenders under regime 2 (e.g., countries in the European Union) 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, such as those discussed.

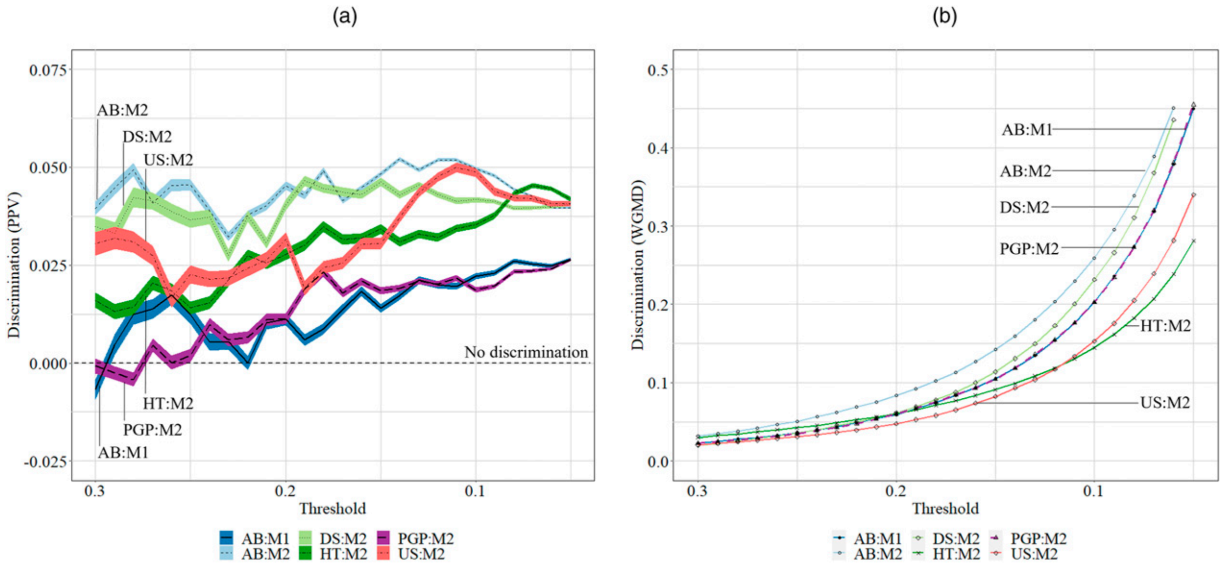
7.2. Discrimination-Reducing Techniques for Firms Operating Across Jurisdictions

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 approach 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 and then, apply that model to predict gender for applicants in the regime 3 jurisdiction. Our gender prediction model achieved a fivefold crossvalidated 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. We compare the PGP model (PGP:M2) with the AB model without gender (AB:M2) and observe that discrimination measured by PPV decreases by 71.08%, predictive quality is not impacted, and firm profitability increases by 0.13% (Observation 12 and Table 3).

This finding illustrates the benefits of probabilistic gender proxy modeling for applicants (reduced

Figure 9. (Color online) (a) Discrimination Measured by PPV of AB Models with Gender (AB:M1), Without Gender (AB:M2), and Four Possible Approaches to Reduce Discrimination (DS:M2, PGP:M2, HT:M2, US:M2). (b) Discrimination Measured by WGMD of AB Models with Gender (AB:M1), Without Gender (AB:M2), and Four Possible Approaches to Reduce Discrimination (DS:M2, PGP:M2, HT:M2, US:M2)



discrimination) and fintech firms (no change to predictive quality and increased profitability). Unfortunately, we determined that the methodology is currently prohibited in the United States (the largest jurisdiction under regime 3) and has been observed to generate upward statistical bias in default predictions, albeit in the mortgage setting and not in 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 United States) as they are prohibited from not using but also, collecting gender, which means that 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 the use of gender in the training and screening models. Summarizing the findings of several observations throughout the paper, we find that 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 when compared with 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 the use of gender.

8. Discussion and Conclusions

We use publicly available, real, feature-rich, and highly multicollinear fintech data to investigate the impact of three antidiscrimination 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 models; and regime 3, which prohibits both the collection and the 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 antidiscrimination regimes, ML models are less discriminatory, of better predictive quality, and of higher profitability when trained on the data commonly used by fintech firms because of 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 the omitted variable bias that 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, each with varying impacts to model predictive quality and firm profitability: 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. Although these approaches reduce discrimination in our applicant data set, it is unclear how they might impact applicant self-selection, either persuading or dissuading individuals from applying for credit from the fintech lender. This uncertainty could be an important avenue for future research.

The overarching implication of our work is that the growing adoption of algorithmic decision making in nonmortgage consumer credit lending requires a rethink of the antidiscrimination 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. Furthermore, doing so 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 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. Although there are disputes about the reduced explainability in AI models and the 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 intraorganization data collection and usage requirements, which include the responsible use of personal attributes, like gender, as part of an organization's AI ethics guidelines.

Our work also paves the way for the fair economic welfare of both financial institutions and 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 well-being 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 should be carried out to increase customer confidence in the suggested approach.

Increased data access should, however, come with greater firm accountability and responsibility. For example, the Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector (Monetary Authority of Singapore 2018)—to the development of which this paper's authors had the privilege to contribute—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 recommendation is contrary to the situation in the United States, where lenders have used the existing laws to elude responsibility for discriminatory outcomes. For instance, Goldman Sachs did so with their Twitter statement mentioned in Section 1: "we have not and never will make decisions based on factors like gender. In fact, we do not know your gender or marital status" (Franck 2019). To that end, as of late 2021, both the United States and the European Union 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 nonmortgage consumer fintech lending.

The U.S. 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 that users of automated algorithms perform a bias impact assessment to mitigate potential discrimination. A consequence of our findings is that, in the United States, 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 the "Artificial Intelligence Act" categorizes AI systems used for credit lending as high risk and specifies certain mandatory requirements with regard to training data, data governance and explainability, reporting, robustness and accuracy, and human oversight. These requirements include, for example, ensuring a sufficiently representative training data set. 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, multinational banks and fintech firms that 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 antidiscrimination regimes and their detrimental impact when decisions impacting minorities are made by algorithms. We consider one setting—consumer fintech lending—and urge other researchers to continue investigating the implications and drivers of other forms of discrimination as well as potential solutions in additional contexts and operational settings.

Acknowledgments

The authors are grateful to department editors Jérémie Gallien and Serguei Netessine, an associate editor, and two referees for their constructive comments that helped improve the paper.

References

Acquisti A, Fong C (2020) An experiment in hiring discrimination via online social networks. *Management 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.

Altman EI, Haldeman RG, Narayanan P (1977) ZETA analysis: A new model to identify bankruptcy risk of corporations. *J. Banking Finance* 1(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. Roy. Statist. Soc. Ser. A* 182(Part 4): 1287–1311.

Barocas S, Selbst AD (2016) Big data's disparate impact. *Calif. Law Rev.* 104:671–732.

Bartlett R, Morse A, Stanton R, Wallace N (2022) Consumer-lending discrimination in the FinTech Era. *J. Financial Econom.* 143(1):30–56.

Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2017) Fairness in criminal justice risk assessments: The state of the art. *Sociol. Methods Res.* 50(1):3–44.

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

Chandler GG, Ewert DC (1976) Discrimination on the basis of sex under the equal credit opportunity act. Working Paper No. 8, Credit Research Center, Purdue University, West Lafayette, IN.

Chen IY, Johansson FD, Sontag D (2018) Why is my classifier discriminatory? Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. *Adv. Neural Inform. Processing Systems* (Red Hook, NY), 31:3543–3554.

Chen J, Kallus N, Mao X, Svacha G, Udell M (2019) Fairness under unawareness: Assessing disparity when protected class is unobserved. Boyd D, Morgenstern J, eds. *FAT* '19 Proc. Conf. Fairness Accountability Transparency* (New York, NY), 339–348.

Chouldechova A (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5(2):153–163.

Cofounders for RRBM (2017) A vision of responsible research in business and management: Striving for useful and credible knowledge. Accessed November 23, 2021, https://www.rrbm.network/position-paper.

Cohen MC, Harsha P (2020) Designing price incentives in a network with social interactions. *Manufacturing Service Oper. Management* 22(2):292–309.

Corbett-Davies S, Goel S (2018) The measure and mismeasure of fairness: A critical review of fair machine learning. Working paper, Stanford University, Stanford, CA.

Cui R, Li J, Zhang D (2020) Reducing discrimination with review in the sharing economy: Evidence from field experiments on Airbnb. *Management Sci.* 66(3):1071–1094.

Cui R, Gallino S, Moreno A, Zhang DJ (2018) The operational value of social media information. *Production Oper. Management* 27(10):1749–1769.

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(3):837–845.

D'Espallier B, Guérin I, Mersland R (2011) Women and repayment in microfinance: A global analysis. *World Development* 39(5):758–772.

Doleac JL, Stein LCD (2013) The visible hand: Race and online market outcomes. *Econom. J.* 123(572):F469–F492.

Franck T (2019) Warren rips Goldman over its Apple card, "discriminatory" algorithms. *CNBC* (November 14), 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 (2022) Predictably unequal? The effects of machine learning on credit markets. *J. Finance* 77(1):5–47.

Ge Y, Knittel CR, MacKenzie D, Zoepf S (2020) Racial discrimination in transportation network companies. *J. Public Econom.* 190:104205.

Grgic-Hlaca N, Bilal Zafar M, Gummandi KP, Weller A (2016) The case for process fairness in learning: Feature selection for fair decision making. Working paper, Max Planck Institute for Software Systems, Saarbrücken, Germany.

Hadfield G (2016) *Rules for a Flat World: Why Human Invented 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. Lee DD, von Luxburg U, Garnett R, Sugiyama M, Guyon I, eds. *Adv. Neural Inform. Processing Systems* (Red Hook, NY), 29:3323–3331.

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. Roy. Statist. Soc. Ser. A* 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? Brewster S, Fitzpatrick G, eds. *Proc. Conf. Human Factors Comput. Systems* (New York, NY), 1–16.

Hsia D (1978) Credit scoring and the Equal Credit Opportunity Act. *Hastings Law J.* 30(2):371–448.

Kamiran F, Calders T (2012) Data preprocessing techniques for classification without discrimination. *Knowledge Inform. Systems* 33(1):1–33.

Kleinberg J, Ludwig J, Mullainathan S, Rambachan A (2018) Algorithmic fairness. Johnson WR, Markel K, eds. *AEA Paper Proc.* (Pittsburgh, PA), 108:22–27.

Kleinberg J, Ludwig J, Mullainathan S, Sunstein CR (2020) Algorithms as discrimination detectors. *Proc. Natl. Acad. Sci. USA* 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. Matwin S, Yu S, Farooq F, eds. *KDD '17: Proc. 23rd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (New York, NY), 275–284.

Lambrecht A, Tucker C (2019) Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads. *Management 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? Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. *32nd Conf. Neural Inform. Processing Systems (NeurIPS 2018)* (Red Hook, NY), 1–11.

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds. *Adv. Neural Inform. Processing Systems* (Red Hook, NY), 4765–4774.

Lundberg SM, Erion GG, Lee SI (2019) Consistent individualized feature attribution for tree ensembles. Working paper, University of Washington, Seattle.

Mejia J, Parker C (2021) When transparency fails: Bias and financial incentives in ridesharing platforms. *Management Sci.* 67(1):166–184.

Monetary Authority of Singapore (2018) Principles to promote fairness, ethics, accountability and transparency (FEAT) in the use of artificial intelligence and data analytics in Singapore' financial sector. Accessed November 23, 2021, https://www.mas.gov.sg/publications/monographs-or-information-paper/2018/FEAT.

New York State Department of Financial Services (2021) Report on Apple Card Investigations. Accessed November 24, 2021, https://www.dfs.ny.gov/system/files/documents/2021/03/rpt_202103_apple_card_investigation.pdf.

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 Banking* 48(8):1691–1724.

Perrone V, Donini M, Kenthapadi K, Archambeau C (2021) Fair Bayesian optimization. Fourcade M, Kuipers B, Lazar S, Mulligan D, eds. *AIES '21 Proc. 2021 AAAI/ACM Conf. AI Ethics Soc.* (New York, NY), 854–863.

Pope DG, Sydnor JR (2011a) Implementing anti-discrimination policies in statistical profiling models. *Amer. Econom. J. Econom. Policy* 3(3):206–231.

Pope DG, Sydnor JR (2011b) What's in a picture? Evidence of discrimination from Prosper.com. *J. Human Resources* 46(1):53–92.

Stein RM (2005) The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing. *J. Banking Finance* 29(5):1213–1236.

Taylor W (2011) Proving racial discrimination and monitoring fair lending compliance: The missing data problem in nonmortgage credit. *Rev. Banking Financial 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).

Vigdor N (2019) Apple Card investigated after gender discrimination complaints. *New York Times* (November 11), https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html.

WHO/Europe (2020) Gender: Definitions. Accessed July 15, 2021, https://www.euro.who.int/en/health-topics/health-determinants/gender/gender-definitions.

Wirth R, Hipp J (2000) CRISP-DM: Toward a standard process model for data mining. Mackin N, ed. *Proc. Fourth Internat. Conf. Practical Appl. Knowledge Discovery Data Mining* (Blackpool, LA), 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. *Management Sci.* 64(7):3269–3287.

Zafar MB, Valera I, Rodriguez MG, Gummadi KP (2019) Fairness constraints: Mechanisms for fair classification. *J. Machine Learn. Res.* 20:1–42.

Zhang Y (2018) Assessing fair lending risks using race/ethnicity proxies. *Management Sci.* 64(1):178–197.

Žliobaitė I (2015) On the relation between accuracy and fairness in binary classification. *2nd Work Fairness Accountability Transparency Machine Learn*, Lille, France.

Žliobaitė I (2017) Measuring discrimination in algorithmic decision making. *Data Mining Knowledge Discovery* 31(4):1060–1089.

Žliobaitė I, Custers B (2016) Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence Law* 24(2):183–201.