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

Stephanie Kelley

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

Anton Ovchinnikov

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

Draft: October 2020

Abstract

Problem definition: We study the impact of the existing anti-discrimination laws in different countries on gender bias in the non-mortgage consumer fintech lending setting.

Academic/Practical Relevance: Building on the study of discrimination in operations, financial economics, and computer science, our paper investigates the impact and drivers of discrimination in machine learning models, trained on alternative data, and provides technically and legally permissible approaches for firms to reduce discrimination, whilst managing profitability.

Methodology: We train statistical and machine learning (ML) models on a large and realistically rich publicly available dataset and measure the impact on model *discrimination, predictive quality,* and *firm profitability*. We use ML explainability techniques to understand the drivers of ML discrimination.

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

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

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

1. Introduction

Algorithms and artificial intelligence (AI) are fundamentally transforming the way organizations make decisions. Their adoption, however, has been accompanied by reports of *bias* and *discrimination*, which in this context refer to what concerned parties, like consumers and the media, think is ethically problematic; a noncomparative wrong, a failure to treat a group of individuals the way they are entitled to be treated (Hellman 2016). Discrimination based on protected attributes, like gender or race, is considered undesirable[1], and countries around the world have adopted various anti-discrimination laws to ensure equality across protected groups. However, the rapid advances in decision-making technologies, like AI, have outpaced the changes in these laws (Barocas and Selbst 2016), creating situations where the anti-discrimination laws may paradoxically hurt, rather than help the groups they are supposed to protect.

For a recent example, consider the Apple Card, which was accused of discrimination against women in their algorithmic lending decisions[2]. The bias occurred despite the firm apparently adhering to the applicable regulation (US Equal Credit Opportunity Act, ECOA), which prohibits not just the use, but also the collection of protected attribute data. Goldman Sachs, one of the partners in the Apple Card venture confirmed: "*we have not and never will make decisions based on factors like gender. In fact, we do not know your gender or marital status...*"[3]. Insurance companies including Allstate, Geico, and Liberty Mutual were also found to be discriminating against several minority groups in their use of machine learning (ML) algorithms for car insurance pricing, despite "*...not collect[ing] any information regarding the race or ethnicity of the people they sell policies to.*"[4] These firms adhered to the anti-discrimination laws in their operations, yet their algorithmic/automated AI decisions proliferated bias.

The goal of our paper is to investigate how anti-discrimination laws across countries impact discrimination when decisions are made by automated AI systems. Specifically, we focus on gender-based discrimination in non-mortgage consumer fintech lending. We examine gender, and not race or other attributes due to its cross-country universality and focus on lending due to data availability; our choices do not in any way diminish the need to investigate discrimination in other contexts. The anti-discrimination laws appliable to non-mortgage consumer lending vary across jurisdictions, but similarities in their operational guidance with respect to gender data allow us to categorize them into three levels[5]:

- Level 1 laws (e.g., Singapore) allow for the collection and use of gender data in AI models;

[1] Per the Universal Declaration of Human Rights (https://www.ohchr.org/EN/UDHR/Documents/UDHR_Translations/eng.pdf)
[2] https://www.bbc.com/news/business-50365609
[3] https://twitter.com/gsbanksupport/status/1194022629419704320?lang=en
[4] https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-white-areas-same-risk
[5] We provide a more detailed discussion of the categorization of the laws in the supplemental material.

- Level 2 laws (e.g., European Union) allow for the collection but prohibit the use of gender data in the AI models;
- Level 3 laws (e.g., United States) prohibit the collection, and thus also the use of gender data.

We utilize a realistically-large (hundreds of features for hundreds of thousands of borrowers) publicly available dataset[6] from a real fintech lender working under Level 1 and 2 regulation to train dozens[7] of advanced[8] statistical and machine learning (ML) models that simulate how a fintech lender would assign credit to men and women[9] had it operated under Level 1, or 2, or 3 laws. We then measure model predictive quality (as it depends on what data the models can or cannot use), discrimination (as the resultant models impact men and women differently), and profitability (as the models affect lenders differently), and draw a number of insights which we summarize below.

1.1. Summary of findings

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

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

To understand our approach and findings, first recall how a non-mortgage consumer fintech lender operates: it makes loan accept/rejection decisions, followed by the interest rates pricing decisions for new applicants (Henley and Hand 1997). We focus on the loan accept/reject decision in this paper as it is more operationally relevant, and interest rates were not available in our data, or other consumer lending datasets with gender (necessary for the investigation of discrimination). The lender makes the accept/reject decision via a three-step process. First, it trains a model using data about past borrowers[10] to

[6] Our data is sourced from Kaggle's "Home Credit Default Risk" machine learning competition (https://www.kaggle.com/c/home-credit-default-risk). Disclaimer: the competition's rules prohibit using the data for research, however, Home Credit granted us explicit permission to use it for this study.

[7] We compare over 50 models in DataRobot, a commercially available automated ML platform, including extreme gradient boosting, generalized additive model, elastic-net, light gradient boosted tree, kernel SVM, random forest, Naïve Bayes, and a neural network. Access to DataRobot can be obtained through their Academic Support Program: https://www.datarobot.com/success/academic-support-program/. We trained two additional models in R: a logistic regression model (using the glm package), and an XGBoost model (using the xgboost package).

[8] Kaggle provides a practically relevant measure of model quality via the competition leaderboard, and our XGBoost model, which we utilize in §5.2, would have landed in the top 10 of 7,000+ models in the competition, illustrating that the model is highly competitive with other state-of-the-art models. We provide the R code for the LR and XGBoost models here: https://github.com/stephaniekelley/genderbias.

[9] We acknowledge that gender is non-binary but proceed to examine gender discrimination for women and men given the binary nature of our data; we are hopeful our findings could be extrapolated for non-binary gender too.

[10] Although past borrowers are not necessarily representative of all applicants, empirical research into this potential selection bias has show than there is only modest scope for model improvement when all applicants, both those accepted by the lender and those rejected are included in the training data (Banasik et al. 2003). Given this, in practice lenders have continued to predict applicant default based on accepted borrower data (Banasik et al. 2003).

predict whether a new applicant will repay or default if given a loan, a task generally referred to as binary classification. Since default is uncertain, the model predicts a numeric score, which can be intuitively interpreted as the probability of default (the distinction between this score and the probability is a technical nuance, which is not particularly relevant for this investigation, so we therefore refer to the score as the probability). Second, the lender compares the predicted probability to a classification threshold and rejects the applicant if the predicted probability is above the threshold, and issues credit otherwise (Lessmann et al. 2015). Throughout the paper we present results across a range of realistic thresholds (5 – 30%) in our figures, and for consistency, in the main discussion, report the results at a 13% threshold, the average of the realistic thresholds in Stein (2005), which is also in line with our observations from practical industry work. Third, the lender optimizes the threshold given the economics of the loan, i.e., the cost of default and the revenue from repayment.

If a fintech operates in a Level 1 jurisdiction, the accept/reject decision model has access to the gender feature and the numerous consequences of that: feature engineering, hyperparameter tuning, etc., see §1.1.2 and §5.2. If the firm operates in a Level 2 or 3 jurisdiction, then it cannot use gender in the accept/reject decision model. We refer to these two kinds of models as Model 1 (with gender) and Model 2 (without gender). We are yet to specify what the "model" is; we discuss this further in §3.1.1 and §3.1.2 below. Importantly, Model 1 differs from Model 2: their predictions vary, and consequently, the optimal thresholds differ as well. Hence, the exact same applicant could be issued a loan under one Level of law but rejected under another. We study whether this affects men and women differently.

In §4 we find that Level 2 and 3 laws, which force the firm to exclude gender in the final model (i.e., the firm must use Model 2 as opposed to Model 1) leads to a 69.43% increase[11] in gender discrimination in logistic regression (LR), and a 178.14% increase in discrimination in the top-performing ML model ("Average Blender" discussed in §3.1.2). Level 2 and 3 laws also negatively impact the firm: profitability is on average 0.46% lower for the LR model, and 0.87% lower for the ML model. Interestingly, the exclusion of gender does not significantly impact the model predictive quality, measured by the area under the curve (AUC).

The paradoxical discriminatory effects of anti-discrimination laws have been investigated before for traditional lending models trained on traditional data. Chandler and Ewert (Elliehausen and Durkin 1989) evaluated the ECOA soon after its launch in 1979, and found that the law's operational modeling guidance, which prohibits the use and collection of gender, creates a detrimental increase in the rejection

[11] Recall, that when a single number is reported, it corresponds to a 13% threshold – the average of realistic values reported in the finance literature (e.g., Stein 2005), and in line with what is observed in our practical work. A wider range of values is presented in the figures (5 to 30%); this range, again, is based on the literature and practice.

rates of women, compared to models that use gender. Andreeva and Matuszyk (2019) similarly found that the EU Gender Directive, which prohibits the use of gender in the final lending model, leads to a greater increase in rejection rates for women compared to men, versus models which include gender. Both these works use private, NDA-protected datasets, preventing investigation or replication of their results. Further, they evaluate a single statistical model without a practically relevant measure of model quality, meaning the phenomena observed could be model specific or driven by models that are "just not that good." We introduce the novel fintech setting and alternative data akin to what fintech firms use in practice, evaluate both traditional statistical models and over 50 state-of-the-art ML models trained on publicly available data, and use the Kaggle competition leaderboard[7] as a measure of model quality to address these shortcomings.

1.1.2. Drivers of statistical and machine learning discrimination

What drives discrimination, and, further, what drives the differences in discrimination between statistical and ML models? One intuitive explanation from the traditional statistics and econometrics literature is omitted variable bias (OVB) (Wooldridge 2015); indeed, Level 2 and 3 laws remove gender from the set of variables, changing the model coefficients for the remaining variables. Andreeva and Matuszyk (2019) use a traditional statistical modeling approach (which we replicate in §3.1.1) and show that the Level 2 and 3 anti-discrimination laws indeed create OVB. When trained on data with women as the minority, as is common in lending settings, the OVB leads to coefficient estimates dominated by men, the less creditworthy group, which increase the rejection rates of women compared to men. Kleinberg et al (2018), and Žliobaitė and Custer (2016) approach the topic from a computer science perspective using a generic modeling framework; and conclude that laws, akin to the Level 2 and 3 regulations, that prohibit the use of protected attributes create OVB, which can lead to discrimination. They suggest that absent legal constraints, a protected attribute should be included to reduce discrimination and improve accuracy; we support this conclusion.

However, the traditional statistics and econometrics view of OVB makes several simplifying assumptions which are not true for a fintech that uses modern ML. The removal of gender impacts numerous other stages in the ML modeling process pipeline, which are not captured by OVB; see Fig. 1:

- **Feature engineering**: having access to the gender variable allows for the creation of new variables (features), such as interactions (e.g., "gender * balance"), or binning. For an example of binning consider "age." In the absence of gender, a feature "=IF(age>65,1,0)" could mean "retiree." But if in a certain country men and women retire at different ages, then in the presence of gender the "retiree" bins could be created differently for women and men

(this observation underpins an example we use in §6). In §5.2 permutation importance shows that ~20% of the most impactful features in Model 1 are engineered using gender; they are not even considered by Model 2, and hence cannot be "omitted" by definition. Interestingly, gender is not an important feature by itself, so it does not impact Model 1 or Model 2 much.

- **Algorithm selection:** having access to gender could mean that the top-performing algorithm could change. For example, a regularized logistic regression may outperform a random forest model on the data without gender (and the resultant engineered features), but a random forest may perform better with those features.
- **Feature selection**: access to gender in model training can change the set of features that are selected to be "in the model." For example, with access to gender, the model may select "age" during feature selection, but when gender is excluded, it may exclude "age." In §5.2 we observe that when gender is excluded the algorithm ultimately excludes certain features that we refer to as gender-reliant, and selects others in their place, which we refer to as gender-redundant. Using SHAP values and SHAP interaction values (Lundberg et al. 2019, Lundberg and Lee 2017), we find that the gender-reliant features are on average 19 times more important for women compared to men, so when the gender-reliant features are not selected by the algorithm due to the exclusion of gender, discrimination against women increases.
- **Hyperparameter selection**: multiple ML models, such as the aforementioned random forest, have numerous parameters that guide learning, rather than are learned from data directly. For example, the number of trees in the forest, or the size of each tree; these are called "hyperparameters." Having access to gender and the resultant engineered features can result in a different set of hyperparameters, even if the algorithm itself is the same. Our analyses in §7 shows that selecting the hyperparameters when gender data is available can change the model predictions, ultimately reducing discrimination, even if gender is not used in the final learning model parameter estimation.

Incorporating these four elements, when gender is excluded (Model 2) our top-performing "Average Blender" ML model, hereinafter referred to as AB, is less discriminatory across all thresholds (at a 13% threshold, 4.32% less discriminatory), of significantly better predictive quality (AB AUC 78.14% [77.48 - 78.81%] vs. LR 73.19% [72.46% - 73.91%]), and on average 7.50% more profitable than the traditional statistical logistic regression model. When gender is included (Model 1), the AB model is even less discriminatory across thresholds (41.7% less discriminatory at a 13% threshold), of greater predictive quality (AB AUC 78.29% [77.63% - 78.95%] vs. LR 73.47% [72.75% - 74.20%]), and on

average 7.58% more profitable compared to the LR model. This illustrates that both firms *and* lending applicants should prefer machine learning models over traditional statistical models in the non-mortgage consumer lending setting, as the additional elements beyond OVB allow the ML algorithm to partly recover the negative impacts of excluding gender in the final estimation.

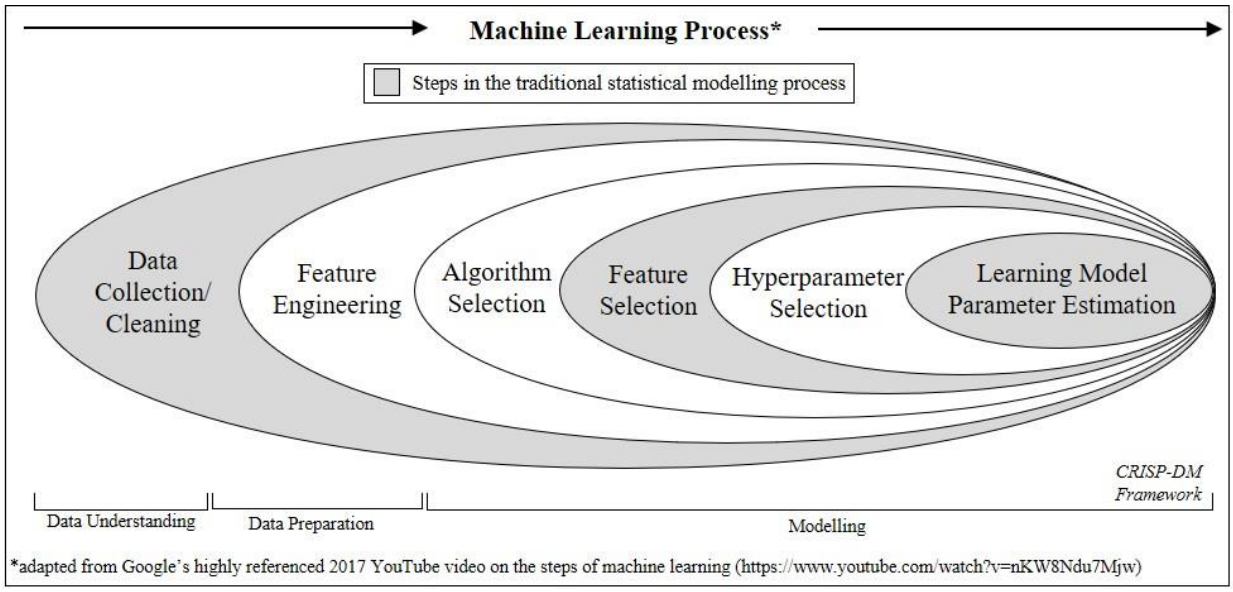

Fig. 1. Comparing the machine learning and traditional statistical modeling processes

1.1.3. Possible approaches to reduce machine learning discrimination

In §7 we evaluate several possible approaches for firms to reduce the gender discrimination given the restrictions on the use and collection of gender, imposed by anti-discrimination laws:

- **Down-sampling the training data to rebalance gender** is a form of data pre-processing (Kamiran and Calders 2012) where observations are randomly removed from the majority class (men) until the count matches that of the minority group (women). This approach is technically feasible and legal[5] for firms under Level 1 and 2 regulation and results in -9.43% discrimination, -174 bps predictive quality, - 0.06% avg profitability in our data.
- **Gender-aware hyperparameter tuning** is an approach where model hyperparameters are tuned on a dataset with gender, but gender is then not used in learning the accept/reject decision model. It is similar to the fair Bayesian optimization technique (Perrone et al. 2020) and is model agnostic and therefore quite flexible. It is technically feasible and legal[5] for firms under Level 1 and 2 regulation and results in -24.56% discrimination, -280 bps predictive quality, - 0.11% avg profitability in our data.

- **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 technically feasible and legal[5] under Level 1 and 2 regulation and results in -2.55% discrimination, no significant change to predictive quality, - 0.04% avg profitability in our data.
- **Probabilistic gender proxy, PGP, modeling** is an approach that is technically feasible for firms operating in multiple jurisdictions: a model is trained to predict the gender of existing borrowers (using data from a Level 1 or 2 jurisdiction), then the model is used to predict gender for applicants in a Level 3 jurisdiction, and finally this gender prediction is used as a feature in the final accept/reject model; see Zhang (2018) and Chen et al (2019) for examples of using PGP to predict race/ethnicity in lending settings. Although quite effective in our data (-62.08% discrimination, no significant change to predictive quality, +0.01% avg profitability) this approach is currently illegal in the US, a prime example of Level 3 regulation (Chen et al. 2019).

We reviewed several other discrimination-reducing approaches, including generating gender-specific models and using gender-specific thresholds (Lipton et al 2018). However, these approaches use gender in the final accept/reject model, and treat the two genders differently, which is in direct contradiction with Level 2 and 3 regulations; we therefore excluded them from consideration. Taken together, our findings suggest that firms under Level 2 laws can do quite a lot to reduce discrimination with data science techniques, but those under Level 3 laws cannot. Even worse, firms under Level 3 regulation have no way to even evaluate how discriminatory their models are as they are prohibited from collecting gender data, and therefore from knowing exactly which borrowers are women.

Overall, our findings add to the growing body of the operations management literature on technology discrimination (e.g., Cui et al. 2020, Lambrecht and Tucker 2019, Mejia and Parker 2020) and put our work among the first to examine the interaction of fintech operations, data science, ethics, and the law; see §2 for related literature. We further hope that not only what we find, but also how we conduct our investigation, i.e., using public verifiable data (see §3 for a description of our approach), speaks to the responsible study of operations and would be of interest to other researchers and practitioners.

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 the operations of technology-based businesses

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

Of particular relevance for our study are the investigations of discrimination in the use of machine learning algorithms. Lambrecht and Tucker (2019) find that online advertising algorithms lead to automated gender bias because of the higher economic valuation assigned to the views of women. Obermeyer et al (2019) examine a commercial healthcare prediction algorithm and find that it proliferates racial bias due to biased training data. Both works uncover unique drivers of discrimination specific to the operational setting where the technology is applied. Our work investigates the drivers of machine learning discrimination in a new operational setting, fintech lending, and we explore mechanisms to reduce the bias, which are said to be often overlooked in the literature (Mejia and Parker 2020).

2.2. Discrimination in non-mortgage consumer lending

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

For these reasons, most of the existing studies of discrimination in lending 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 in 1979 (Elliehausen and Durkin 1989) and Andreeva and Matuszyk (2019). These works focus on outdated statistical lending models with private and inaccessible data, they are each focused on a single legal jurisdiction, they have no objective measure of model quality, no formal measure of discrimination, and they do not measure the impact of the laws on firm profitability, or provide recommendations for firms to reduce discrimination. Our unique public data, and extended

modern ML approach addresses all these shortcomings, making the findings more operationally relevant for fintechs, regulators and the public across several levels of regulation. Further, we replicate the key qualitative findings from these two papers to show where our findings extend theirs or differ.

2.3. Fairness in machine learning

Third, our work is related to the computer science study of fairness in machine learning. A handful of works have investigated the impact of excluding protected attributes, like gender, on discrimination. Lipton et al (2019) and Kleinberg et al (2018) explore the impact of US anti-discrimination laws and conclude that absent legal constraints, a protected attribute should be included to improve fairness and model accuracy. Žliobaitė and Custers (2016) perform a comparable investigation, in the context of EU anti-discrimination laws, and arrive at a similar conclusion. Like Kleinberg et al (2018), they explain the drivers of algorithmic discrimination using the OVB framework. While this arm of the literature succinctly points out the discriminatory effect of excluding protected attributes, these studies lack domain-specific operational details. For instance, Žliobaitė and Custers (2016) use salary data for male and female college professors with 52 observations of 6 variables. It is hard to imagine that the Dean of Faculty at some college would use a model trained on that dataset to make salary decisions. In contrast, our study is operationally grounded: our context, data, process, and models are essentially the same as non-mortgage (fintech) lenders utilize in practice. We are also the first to provide an aggregated analysis across jurisdictions (Level 1 through 3), as opposed to an analysis of a single country's laws.

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

That said, 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 the mathematical definition of discrimination (Berk et al. 2017, Chouldechova 2017), see §3.2[12]

[12] For a more detailed review of the mathematical definitions of fairness/discrimination see Žliobaitė (2017) for a European-focused legal discussion, and Berk et al (2018) for a US-focused discussion.

3. Data, analytical approach, and key metrics

3.1 Data and analytical approach

As we were interested in analyzing gender discrimination, we needed to source data with the gender feature, and therefore acquired real data for 307,507[13] borrowers from a fintech firm, operating under Level 1 and 2 regulation. In the spirit of transparency, we sourced publicly available data from the "Home Credit Default Risk" machine learning competition on Kaggle, which at the time of writing was available at https://www.kaggle.com/c/home-credit-default-risk/data. Note that the competition's rules prohibit using the data for research, however, Home Credit granted us the explicit permission to use the data for this study. We excluded the competition test dataset as it was missing borrower default status and therefore could not be use for our research purposes, leaving us with seven data files, which we then gathered into one file with a single observation for each borrower.

3.1.1. The traditional statistical modeling process

We use logistic regression (LR) as the "traditional statistical" model throughout the paper as it is the preferred model of institutional lenders (Thomas et al. 2017), used in past lending discrimination studies. To generate a LR lending model, a fintech firm would follow the traditional statistical modeling process (visualized in gray in Fig. 1) by first *collecting and cleaning the data*. As our data was originally provided for an ML competition, we had to exclude some of the time series features in the LR feature set that did not adhere to the modeling assumptions of LR, leaving us with a subset of 122 borrower features for a firm under Level 1 or 2 regulation, and 121 features for a firm under Level 3 regulation who cannot collect gender.

Following the standard methodology for LR lending models used by Andreeva and Matuszyk (2019) we first cleaned the data exactly as they did: we coarse-classified the continuous features, first by splitting them into 10 intervals, and then manually merged the adjacent intervals with similar default rates, generating separate coarse classes for missing observations. We then grouped the small categories in the categorical variables and transformed each level into a binary dummy variable, removing the largest category to avoid identification issues. We then trained the logistic regression model to determine which features should be manually selected to remain in the final model. Per their methodology(Andreeva and Matuszyk 2019), we first *manually selected features* that were statistically significant at 0.05 level in the model with gender, and then excluded gender from this dataset to generate the genderless feature set. This final set of features was then used to train a logistic regression model, providing the *learning model parameter estimations*. We discuss the results of the model estimation in §4.

[13] The dataset has 307,511 borrowers, but 4 observations were missing gender and were therefore excluded.

3.1.2. The machine learning process

The ML process requires three additional steps: feature engineering, algorithm selection, and hyperparameter tuning (recall Fig. 1). We started with the same set of 122 features (or 121 if under Level 3 regulation) from the LR data, and did not perform any additional *data cleaning* as the data was prepared for ML. We then proceeded to *feature engineering*, a procedure whereby several additional features are generated based on interactions and/or transformations of the original feature set. We used feature engineering techniques inspired by the publicly available code of the top-ranking teams in the Kaggle competition to gather the additional features into a format that could be used by machine learning models, and generate several ratios between the original features to improve predictive performance. This resulted in 744 features for a firm under Level 1 or 2 regulation, and 743 features, excluding gender, for a firm until Level 3 regulation. We then allowed the algorithm to perform another round of automated feature engineering, specific to the training dataset. As discussed, we then *selected the algorithm* from DataRobot with the best predictive quality (measure by the five-fold cross-validated AUC reported in DataRobot, a metric we discuss in §3.2); the top-performer was an "Average Blender," an ensemble classifier that averages the predictions from multiple models, in our case, several forms of XGBoost and Light Gradient Boosting models, each of which individually had strong predictive quality. Ensemble models, like the Average Blender have been found to have stronger predictive quality compared to individual models in credit lending (Lessmann et al. 2015). The algorithm then performed automated *feature selection* (as opposed to the manual feature selection performed in the traditional statistical modeling process), extracting the explanatory features to be used in the final learning model parameter estimation. Given this final subset of features, we then let the algorithm *tune hyperparameters*, modeling values used to further improve the predictive quality of the chosen algorithm. The algorithm then performed the final *learning model parameter estimation*, results of which are also discussed in §4 onwards.

In §5.2 we introduce a single XGBoost tree ensemble model to support our investigation of the drivers of ML discrimination. We introduce this second model because the explainability techniques required for our analysis (SHAP values and SHAP interaction values), can only be calculated with access to the full model training process (not possible in DataRobot), and are designed for single-class ensembles like XGBoost, not multi-class ensembles like the Average Blender (Lundberg and Lee 2017). Our XGBoost model is trained in R using the xgboost package. The model would have landed in the top 10 of 7,000+ models in the Kaggle competition, illustrating the model is highly competitive with other state-of-the-art models. The code for this mode, and our LR model, also trained in R is available here: *https://github.com/stephaniekelley/genderbias*.

3.1.3. Data sampling

To support our investigation, we generated several samples from the original dataset. First, as is common in predictive modeling, we randomly split the original data into an 80% Training set, and a 20% Testing set. We use the testing data throughout our paper to obtain "out-of-sample" predictions for consistent comparison across models. We then generated three additional samples from the Training set:

1. **Minority data** (80% men/20% women), created by randomly down-sampling women from the Training data to reflect the global statistics on credit access (Ongena and Popov 2016), and used throughout the main analysis to investigate the impact of the three levels of law;
2. **Rebalanced data** (50% men/50% women), created by randomly down-sampling men from the minority sample; and
3. **Rebalanced collected data** (50% men/50% women), we mimic data collection by randomly sampling "extra" women observations discarded in the creation of the minority data.

The sampling procedure is summarized in Fig. 2.

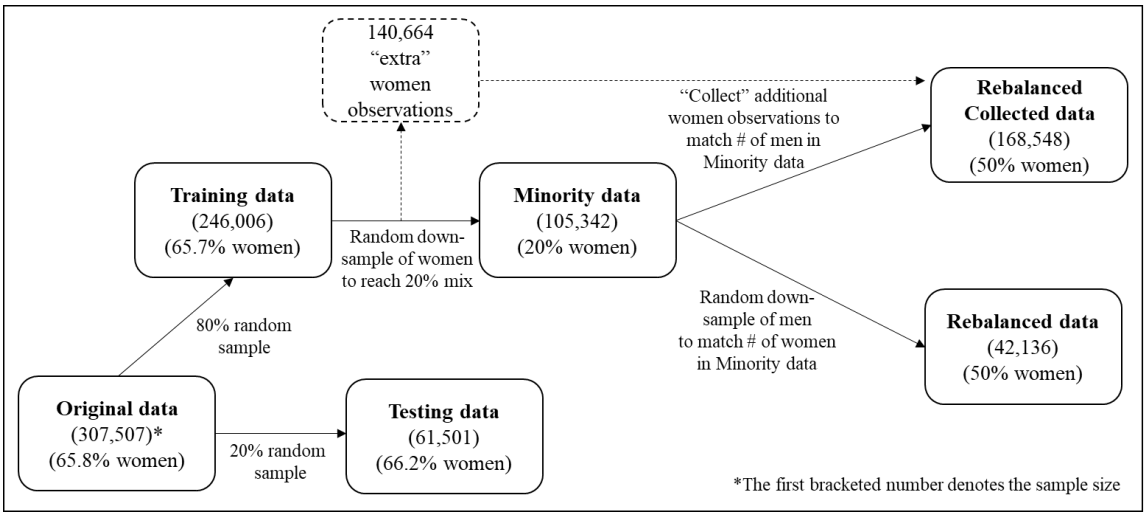

Fig. 2. Summary of the data sampling procedure

Note that the features used to train the traditional statistical and machine learning models differ due to variations in the modeling processes, but the samples and individual observations remain consistent for accurate comparisons. Across all samples we observe women to be more creditworthy than men, with fewer observed defaults (~6.7% compared to ~10.0%), in line with data from past empirical investigations and reports on gender and lending (D'Espallier et al. 2011).

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

Before proceeding to the results in §4, we briefly review the metrics used throughout our analysis: discrimination, predictive quality, and firm profitability, all calculated out-of-sample.

3.2.1. Discrimination

We measure *discrimination* with a somewhat standard metric introduced by Chouldechova (2017), and Berk et al. (2017): the difference of the positive predictive values between genders, i.e., 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, D_τ, is the number of true positive predictions (i.e., correctly predicted defaults), TP_τ, for men (M), divided by the sum of the true positive predictions and false positive predictions, FP_τ i.e., all default predictions for men (M), minus the same ratio for women (W) [note that TPs and FPs are functions of τ]:

$$D_\tau = \frac{TP\tau_M}{(TP\tau_M + FP\tau_M)} - \frac{TP\tau_W}{(TP\tau_W + FP\tau_W)} \quad (1)$$

Discrimination greater than 0 denotes bias against women, less than 0, discrimination against men, and equal to 0 indicates no discrimination (a fair model). The use of TP_τ and FP_τ values in the mathematical definition aligns with our theoretical definition of discrimination as a non-comparative wrong: a failure to treat (predict) a group of individuals (one gender) the way they are entitled to be treated (predicted correctly) (Hellman 2016). Throughout the paper we compare discrimination between models by reporting the mean and 95% confidence intervals, calculated using a 30-fold cross validation.

3.2.2. Predictive quality

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

3.2.3. Firm's profitability

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 to 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 an applicant they do not grant credit to who would default (a true positive), 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 (π_τ)

is the revenue from repayment (R) times the number of true negative predictions ($TN\tau$) less the cost of default (C) times the number of false negative predictions ($FN\tau$) (Eq. 2).

$$\pi_\tau = (R * TN\tau) - (C * FN\tau) \quad (2)$$

To examine different operating scenarios we consider 2,431 cost-to-revenue (*C:R)* pairs, covering the full range of reported ratios (up to 35x) from the literature (Altman et al. 1977, Stein 2005). To calculate the firm profitability (π^*) at each *C:R* ratio we first generate a 90% random sample of the out-of-sample predictions and calculate the $TN\tau$ and $FN\tau$ counts across 9,500 thresholds (from 0.01% to 0.95%, in increments of 0.01%). We then calculate the profit for each *C:R* ratio and every $TN\tau/FN\tau$ pair and find the maximum profit and corresponding optimal threshold for each pair. We apply those optimal thresholds to the 10% holdout and calculate the $TN\tau$ and $FN\tau$ counts, then calculate the optimal profit given the *C* and *R* for each threshold. We take the average of the optimal profit across the 30 folds to calculate firm profitability. We compare performance between models by reporting the mean difference of firm profitability across all *C:R* ratios and the number and range of the statistically significant differences, calculated using a two-sided paired t-test with a 95% confidence interval.

4. Impact of anti-discrimination laws on gender-based discrimination

In this section we compare the impact of the three levels of anti-discrimination laws on the traditional statistical model, LR, and the top performing ML model, AB. All models in §4 are trained on the Minority data, reflecting the data commonly available to a fintech lender.

4.1. Impact on logistic regression models

Given a preference for logistic regression, and access to Minority data, a fintech lender would follow the traditional statistical modeling strategy discussed in §3.1.1 to generate a predicted probability of default for each applicant. Under Level 1 regulation, the fintech firm can include gender in the data used to train the model, while those under Level 2 and 3 cannot.

Observation #1*: Compared to the LR model with gender (LR:M1), the LR model without gender (LR:M2),*

a. increases discrimination by 69.43% (0.0532 [0.0527 – 0.0536] vs. LR:M1 0.0314 [0.0308 – 0.0320], τ=0.13) (see Fig 3a. for all thresholds),

b. does not significantly impact predictive quality (AUC 73.19% [72.46% - 73.91%] vs. LR:M1 73.47% [72.75% - 74.20%]), and

c. *decreases firm profitability by an average 0.83% (31.88% of the profit differences are statistically significant, 79% [-27.20 - -0.02%], and 21% [0.007 – 3.79%]).*

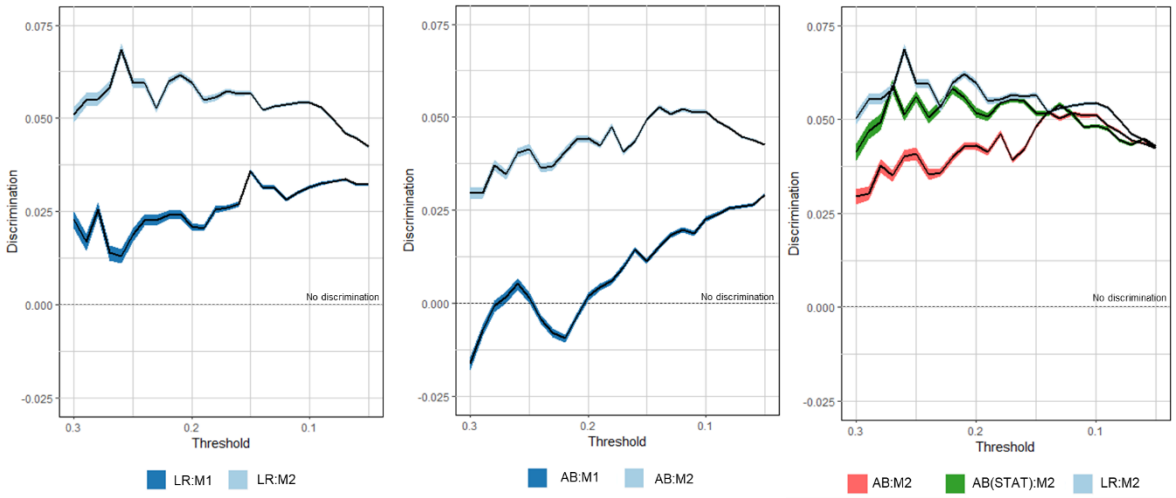

Fig. 3a. **Fig. 3b.** **Fig. 3c.**

Figs. 3a – 3c. Gender discrimination in: 3a. Average Blender models, 3b. Logistic Regression models, 3c. Average Blender & Logistic Regression models without gender

Note: figures are color-blind-friendly and printer-friendly, color recommendations are sourced from ColorBrewer (https://colorbrewer2.org/#type=qualitative&scheme=Paired&n=4)

4.2. Impact on machine learning models

With a preference for ML, a firm would follow the ML modeling process discussed in §3.1.2; firms under Level 1 law would be able to use gender, while those under Level 2 and 3 would not.

Observation #2*: Compared to the model with gender (AB:M1), when gender is excluded from the Avg Blender ML Model (AB:M2),*

a. *discrimination increases by 178.14% (0.0509 [0.0503 – 0.0515] vs. AB:M1 0.0183 [0.0175 – 0.0191], τ=0.13)* (see Fig 3b. for all thresholds),
b. *predictive quality is not significantly impacted (AUC: 78.14% [77.48% - 78.81%] vs. AB:M1 78.29% [77.63% -78.95%]), and*
c. *firm profitability decreases on average by 0.91% (31.88% of the profit differences are statistically significant, 62% [-18.93 - -0.28%], and 38% [0.007 – 0.14%]).*

4.3. Comparison of gender-based discrimination

Two insights follow from Observations 1 and 2. First, the operational guidance to not use gender in training models prescribed by the Level 2 and 3 anti-discrimination regulations, such as the EU Gender

Directive, and the US ECOA, respectively, lead to increased discrimination and decreased firm profitability compared to the Level 1 laws which allow for the use of gender. This negative impact occurs in both traditional statistical models, and machine learning models. The results are troubling as they demonstrate that the Level 2 and 3 laws create a detrimental outcome for both lending applicants (increased discrimination) and fintech lenders (decreased firm profitability), confirming the reports of automated bias that motivated our work.

5. Drivers of statistical and machine learning discrimination

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

Given the discriminatory impact of excluding gender, we investigate whether the lending applicants and fintech lending firms would be better off (i.e., observe lower levels of discrimination, higher predictive quality, and higher firm profitability) using traditional statistical or machine learning models. To do so, we compare the LR and AB models, without gender, trained on the Minority data, per Level 2 and 3 laws (LR:M2 & AB:M2).

Observation #3*: Compared to the LR model without gender (LR:M2), the AB model (AB:M2)*

a. *reduces discrimination by 4.32% (0.0509 [0.0503 – 0.0515] vs. LR:M2 0.0532 [0.0527 – 0.0536], τ=0.13)* (see Fig. 3c for all thresholds),
b. *increases predictive quality by 495 bps (AUC of 78.14% [77.48% - 78.81%] vs. LR:M2 73.19% [72.46% - 73.91%]), and*
c. *increase firm profitability on average by 7.50% (91.73% of the profit differences are statistically significant [0.1% - 57.94%])*

For robustness, we also report the results of a second AB model trained on the traditional statistical feature set used by the LR model (AB(STAT:M2). This allows us to observe the incremental improvement of changing from LR to ML (AB(STAT):M2 vs. LR:M2), in addition to the true operational benefit a fintech firm would observe by shifting from traditional statistical to the machine learning models (AB:M2 vs. LR:M2, per Observation #3).

Observation #4*: Compared to the LR model without gender (LR:M2), the AB model trained on the traditional statistical feature set (AB(STAT):M2)*

a. *is of comparable discrimination (AB(*STAT*):M2 0.542 [0.0535 – 0.0548] vs. LR:M2 0.0532 [0.0527 – 0.0536], τ=0.13)* (see Fig. 3c for all thresholds),

b. *increases predictive quality by 240 bps (AUC 75.33% [74.63% - 76.03%] vs. LR:M2 73.19% [72.46% - 73.91%]), and*
c. *increases firm profitability on average by 7.73% (90.21% of the profit differences are statistically significantly difference, 97% [0.02 – 59.00%], and 3% [-0.007- -0.005%]).*

These results illustrate that trained on the non-engineered, traditional statistical feature set, the AB model is of better predictive quality and greater profitability, but of comparable discrimination to the LR model. Trained on the full feature engineered dataset, the AB model becomes less discriminatory, and has even greater predictive quality, and profitability compared to the LR model. This demonstrates that both fintech firms and lending applicants would benefit from the use of ML models in place of traditional statistical models, like LR, but the full benefit, particularly the decrease in discrimination, relies on the comprehensive ML process (particularly the extensive feature engineering, as discussed in §3.1.2.), in addition to the selection of an ML algorithm. Qualitative findings are similar for firms under Level 1 regulation, who can include gender in their model (41.7% less discriminatory, +482bps increase in predictive quality, and average 7.58% increase in profitability); see the supplemental materials.

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

The previous results illustrate that under Level 2 and 3 regulation, even the top-performing ML model (AB:M2) still proliferates gender discrimination. In this section we seek to understand the drivers of that and do so using two ML explainability techniques. First, we use permutation importance to understand the impact of excluding gender in the AB model. Second, we use SHAP values and SHAP interaction values (Lundberg et al. 2019, Lundberg and Lee 2017) on the state-of-the-art XGBoost model (discussed in §3.1.2.) to investigate further; recall, SHAP values and SHAP interaction values cannot be created for the multi-model ensembles. Before proceeding to the discussion of ML discrimination in §5.2.2. we review OVB which drives LR discrimination.

5.2.1. Statistical discrimination: omitted variable bias

In the traditional statistical modeling process, recall Fig. 1, it is well known that when a LR model has access to the comprehensive set of causal features, it can estimate the true, unbiased learning model parameters (Wooldridge 2015). Following this, when an important causal feature is excluded in data collection (e.g., gender), the learned model parameter estimates become biased, in the statistical sense of the word (i.e., inaccurate); a phenomenon referred to as omitted variable bias, OVB (Wooldridge 2015). Andreeva and Matuszyk (2019) show that when gender is excluded from a LR lending model, it creates OVB. Men, the less creditworthy gender are a majority, and the exclusion of gender creates an upward

bias in the parameter estimates, leading to an increase in the rejection rates of women compared to the model with gender. Observation #1 confirms that, as expected, OVB occurs in our data too.

5.2.2. Machine learning discrimination: gender-blind feature selection

In the more complex ML modeling process (per Fig. 1), the exclusion of gender in the data collection stage affects four other aspects of the modeling process: feature engineering, algorithm selection, feature selection, and hyperparameter selection, which impact the final learning model parameter estimation. We focus our investigation on *feature engineering* and *feature selection* here as these stages they are most impacted by the exclusion of gender and leave the discussion of *hyperparameter selection* to §7. We first compute the permutation importance (using the feature importance tool in DataRobot) for the AB Models 1 and 2. To illustrate the result visually we add 10 manually generated gender interactions with the top 5 features (5 for women, 5 for men) to Model 1. The permutation importance of the top 25 features for Models 1 and 2 are visualized in Figs. 4a and 4b, respectively. We observe[14] the following:

Observation #5*: When gender is included in the ML model, gender interaction features account for 4 of the top 25, and 2 of the top 10 features* (Fig. 4a); *gender is also selected as a feature (outside the top 25).*

Observation #6*: When gender is excluded from the ML model, different features are selected by the algorithm (6 of the top 25 features), with different permutation importance rankings (21 of the top 25 features), compared to when gender is included* (Fig. 4b.).

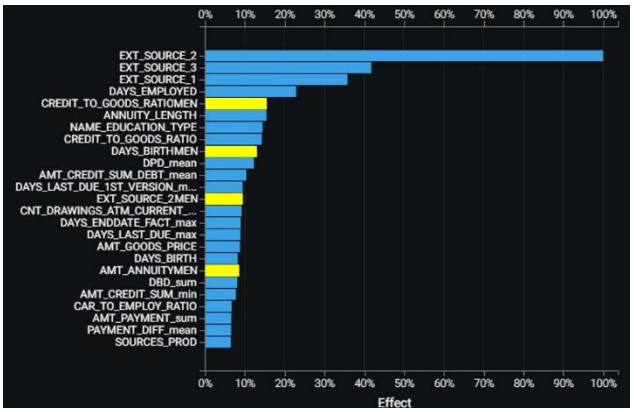

Fig. 4a. Permutation importance for the Average Blender model with gender & interactions (M1)

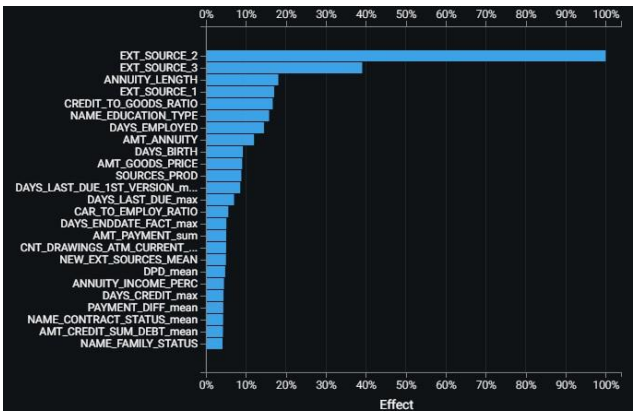

Fig. 4b. Permutation importance for the Average Blender model without gender (M2)

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

[14] These results are a lower bound on the number of affected features, as gender could have impacted engineered features that are not shown separately; e.g., the binning of "EXT_SOURCE"s may be different in M1 vs M2.

feature engineering and feature selection in more detail using the SHAP values and SHAP interaction values. As a reminder, we opted to use XGBoost for this part of the investigation given the restrictions of the SHAP interaction values (they cannot be calculated for the multi-class ensemble model).

We observe that certain features, like EXT_SOURCE_2 (an external credit score), 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 extremely 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, according to their SHAP values. Other features are selected by the algorithm when gender is present, but not when gender is excluded; we refer to them as *gender-reliant*.

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

Next, we compute the SHAP interaction values using a condensed feature set of the explanatory features, given the computation requirements of the SHAP interaction values (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 to understand the ML discrimination.

Observation #7*: The SHAP interaction values for the gender-reliant features with gender are on average 19 times greater for women compared to men; 2.5x the difference of the gender-neutral features between women and men.*

This illustrates that for the *gender-reliant* features, the predictions for women rely on the interactions with gender more than men, and therefore women are more detrimentally impacted by the gender exclusion enforced by Level 2 and 3 anti-discrimination laws. Summarizing, we find that the ML gender discrimination is driven by the *gender-blind feature selection* phenomenon; without gender, *gender-reliant* features are excluded from the algorithm in the automated feature selection, and in their place the ML algorithm sources gender information from *gender-redundant* features, despite their lower explanatory value.

The above discussion is limited to the specific dataset that we use, and while that dataset is from a real fintech firm that uses the data for the actual lending decisions, to emphasize the generalizability of

our insights, below we illustrate the interplay between the *gender-reliant* and *gender-redundant* features on a stylized synthetically-constructed example.

6. A stylized example: generalizing the properties of gender-blind feature selection

Let us assume that most borrowers default in the 5 years after they retire. A fintech firm does not have access to retirement status, so they must infer it from age to help predict default. Many applicants realize this and do not report their age, resulting in data missingness (see Table 1 for a summary). It also happens that women, who are generally better borrowers, tend to retire later in life than men to support spending across their longer lifespans, so that the Age feature is uniformly distributed between 70-74 for women who default, 65-69 for men who default, and is uniformly distributed between 18-90 for those who do not default. Age is the *gender-reliant* feature in this example; the distribution across genders differs for defaulters.

The firm also collects job Tenure, Income, and External Credit Score (Score) to help predict default. In this fictitious country, women apply for credit as soon as they start working, and work anywhere up to 20 years, taking off many more years for child rearing than men, who work for up to 40 years, and wait to apply for credit until they are 10 years into their jobs. This results in women having a Tenure uniformly distributed between 1-20, and men between 10-40, regardless of default. Tenure is this example is the *gender-redundant* feature; it has limited explanatory power but is a proxy for Gender. The other features, Income, and Score are *gender-neutral*; we constructed the example so that there are no differences in the distributions between genders for these features.

	Women Default	**Women No Default**	**Men Default**	**Men No Default**
Age (missingness)	70 – 74 (70%)	18 – 90 (70%)	65 – 69 (90%)	18 – 90 (90%)
Income (missingness)	30 – 100K (none)	30 – 800K (none)	30 – 100K (none)	30 – 800K (none)
Score (missingness)	0 – 400 (none)	200 – 800 (none)	0 – 400 (none)	200 – 800 (none)
Tenure (missingness)	1 – 20 (none)	1 – 20 (none)	10 – 40 (none)	10 – 40 (none)

Table 1. Synthetic dataset summary

We generate a synthetic dataset with 60,000 observations given these assumptions, which mirrors the gender mix (80%M/20%W) and default properties (10%M/7%W) of the Minority data from the main analysis. We then trained two stylized LR models, with Gender (M1) and without Gender (M2); we select

LR to conceptually bridge the *gender-blind feature selection* phenomenon in ML to the OVB that drives discrimination in traditional statistical models. We force the LR model to mimic the automated feature engineering in the ML process by generating an interaction feature between Gender and Age and replicate the automated feature selection from ML by allowing Model 2 to select from the entire feature set. This contrasts the process in LR (§3.1.1) where any features not significant in the model with Gender (i.e., Tenure) would be dropped in the model without Gender.

To illustrate the detrimental impact to women, we assess the predicted probability of default of a woman and a man of the same age, income, external credit score, and tenure (65 years, $100,000, 360, 20 years) between the stylized LR models with (M1) and without gender (M2). At a 13% threshold, the model with gender (M1) predicts the woman not to default, with a probability of 12.18%, whilst the man, identical in all non-protected features (Income, Score, and Tenure), is predicted to default, with a probability of 15.40%. The model without Gender (M2) however is unable to differentiate between the woman and the man, incorrectly predicting the woman to default, with a probability of 14.21%, the same probability as the man, despite the knowledge that women between 65-69 do not default (per the data summary in Table 1). These results illustrate that with access to Gender, the model can differentiate between women and men aged 65-69, who differ in their default behaviour, but without Gender, the model cannot differentiate between the two genders and incorrectly predicts the 65 year old woman to default, learning the default pattern from the majority population, men.

Next, we review the parameter estimates for these two synthetic models to understand the impact of *gender-blind feature selection* phenomenon (see columns 1 and 2 in Table 2). In both Models 1 and 2, the *gender-neutral* features Income and Score are statistically significant, and their estimates are stable across models. In Model 1 (column 1), when Gender is included, the model select the *gender-reliant* feature Age (as it is statistically significant), and the interaction of Age and Gender (the interaction feature Gender(Men):Age is statistically significant) to arrive at the prediction. In Model 2 (column 2), when Gender is excluded, the model excludes the *gender-reliant* feature Age, and cannot create the Age and Gender interaction; in their place the model selects Tenure, the *gender-redundant* feature, despite not selecting it (it is not statistically significant) when it had access to Gender (per Model 1 in column 1).

Summarizing this example and reinforcing the preceding discussion, our analysis shows that ML discrimination is driven by the *gender-blind feature selection* phenomenon, which is vastly different from the OVB that drives discrimination in traditional statistical models. On top of this, data quality also plays a role.

	Dependent variable: Default Target					
	(1)	(2)	(3)	(4)	(5)	(6)
	Stylized LR (Model 1)	**Stylized LR (Model 2)**	**Stylized LR No Missing (Model 1)**	**Stylized LR No Missing (Model 2)**	**Regular LR No Missing (Model 1)**	**Regular LR No Missing Model 2)**
Constant	6.86741***	7.15306***	1.44334***	3.15317***	3.44878***	3.77815***
	(0.14492)	(0.13834)	(0.42487)	(0.18699)	(0.18514)	(0.17254)
Gender(Men)	0.63999***	Excluded	2.35314***	Excluded	0.44510***	Excluded
	(0.09557)	by law	(0.44928)	by law	(0.09084)	by law
Income	-0.00004***	-0.00004***	-0.00004***	-0.00004***	-0.00004***	-0.00004***
	(0.000001)	(0.000001)	(0.000001)	(0.000001)	(0.000001)	(0.000001)
Score	-0.01545***	-0.01544***	-0.01567***	-0.01569***	-0.01564***	-0.01563***
	(0.00029)	(0.00029)	(0.00032)	(0.00032)	(0.00031)	(0.00031)
Age	**0.00705*****	*0.00182*	0.008750***	0.06612***	0.05877***	0.05896***
	(0.00218)	*(0.00125)*	(0.00594)	(0.00230)	(0.00225)	(0.00225)
Tenure	*-0.00330*	**0.00784*****	*-0.00346*	0.00996***	-0.00072	
	(0.00322)	**(0.00271)**	*(0.00346)*	(0.00292)	(0.00341)	
Gender(Men):Age	**-0.00566****		-0.02474***			
	(0.00269)		(0.00636)			
Observations	60,000	60,000	60,000	60,000	60,000	60,000
Log Likelihood (LL)	-4,277.46	-4,300.39	-3,700.45	-3,734.44	-3,849.10	-3,865.67
AIC.	8,568.91	8,610.78	7,414.89	7,478.87	7,710.20	7,739.35

Notes: Model parsimony is characterized by higher LL, and lower AIC; *p<0.1, **p<0.05, ***p<0.01

Table 2. Learning model parameter estimates for synthetic example

6.1. Exploring the impact of data quality (missing values) on ML discrimination

We further investigate the impact of data quality, specifically missing values on ML gender discrimination. In our real data, we observe that the *gender-reliant* features have some degree of missingness. Here we remove the missingness from Age, to explore its discriminatory impact. We estimate the two models again (M1, M2), and observe that when gender is excluded (M2, column 4), the lack of missingness changes the Tenure parameter estimate, and the constant compared to the same model with missingness (M2, column 2). Instead of being excluded, Age, the *gender-reliant* feature is selected by the model, contrary to what we observe in the *gender-blind feature selection* phenomenon (column 2), and in the main analysis. This suggests that it is the greater majority class missingness (Age missingness for men > Age missingness for women) that leads to the exclusion of the *gender-reliant* feature during feature selection. This illustrates an important property of the *gender-reliant* features: the discriminatory effect of removing the gender feature is inflated (+7.50% *(0.0889 [0.0876 – 0.0902] vs. no missingness 0.0827 [0.0814 – 0.0839], τ=0.13)*) when there is greater missingness in the majority class.

6.2. Comparing gender-blind feature selection to omitted variable bias

Lastly, to summarize the difference between the *gender-blind feature selection* phenomenon and OVB, we provide the parameter estimates for two traditional LR models (per the process in §3.1.1.), with and without gender trained on the no missingness data (columns 5 and 6 of Table 2). With gender (M1,

column 5) we see the *gender-neutral* (Income, Score), and the *gender-reliant* feature (Age) are statistically significant, but, in line with the other ML models with gender, the *gender-redundant* feature (Tenure) is not statistically significant. It is possible that *gender-redundant* features, like Tenure, would not even be gathered in the data collection stage of the traditional statistical modeling process as the use of proxies is not valid given the strict modeling assumptions of LR, but we include it in the Model 1 estimation to ensure a direct comparison. Following the standard LR modeling process, Tenure is excluded from the feature set for the estimation of the model without gender (M2, column 6). We observe that when gender is excluded, the parameter estimate for Age, as well as the constant, which captures women, both increase, which we know is the effect of OVB (Andreeva and Matuszyk 2019, Kleinberg et al. 2018). Note that without feature engineering in the process, there is no Gender and Age interaction in Model 1 (column 5).

6.3. Summary of the drivers of machine learning discrimination

Although somewhat simplified, we hope the stylized example and the accompanying discussion helps to explain the impact of the exclusion of protected attributes, like gender, on the ML modeling process. In ML, when gender is excluded from the data collection, it prevents the algorithm from feature engineering interactions with other features and gender. It also impacts algorithm, feature and hyperparameter selection; we observe that certain *gender-reliant* features are excluded, and in their place *gender-redundant* features are selected. In this setting, the exclusion of the *gender-reliant* features is significantly more detrimental to women, compared to men, creating discrimination. This gender-blind feature selection phenomenon is vastly different from the OVB that drives discrimination in traditional statistical models and leads to new understanding of discrimination for ML models.

7. Possible approaches to reduce discrimination

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

7.1. Approaches to reduce discrimination under Level 2 regulation

Fintech firms under Level 2 and 3 regulations face the reality of not being able to use gender in their accept/reject models, which we now know leads to discrimination. Those under Level 2 regulation, like the EU Gender Directive, are, however, allowed to collect applicant gender. Next, we explore several possible approaches for these firms to reduce discrimination given their ability to collect gender:

1. Down-sampling the training data to rebalance gender, i.e., under-sampling the majority class (men) to match the count of the minority class (women), leading to the Rebalanced Data (DS:M2) (see Fig. 2. for a summary of the data sampling procedure);

2. **Gender-aware hyperparameter tuning**, which involves creating a single model that has the hyperparameters tuned using gender (we use the XGBoost model, and hyperparameters inspired by the top Kaggle competition teams), which allows the model to learn from the gender feature at an aggregate level before it is trained on the Rebalanced data without individual applicant gender (HT:M2); and

3. **Up-sampling the training data to rebalance gender**, which involves a firm collecting more data from the minority class (women) to achieve a balanced sample, which we emulate by "collecting" data from the Original data resulting in the Rebalanced Collected data (US:M2).

We re-estimate the AB model, without gender (M2) from the main analysis using these three techniques and make the following observations (discrimination is visualized in Fig. 5).

Observation #8*: Compared to the AB model, without gender (AB:M2),*

1. ***Down-sampling the training data to rebalance gender*** *(DS:M2)*
    a. *decreases discrimination by 9.43% (0.0461 [0.0454 – 0.0467] vs. AB:M2 0.0509 [0.0503 – 0.0515], τ=0.13)* (see Fig. 5 for all thresholds),
    b. *decreases predictive quality by 174 bps (AUC 76.40% [75.71% - 77.09%]), and*
    c. *decreases firm profitability on average by 0.06% [-50.38 - -0.01%];*
2. ***Gender-aware hyperparameter tuning*** *(HT:M2)*
    a. *decreases discrimination by 24.56% (0.0384 [0.0374 – 0.0394], τ=0.13)* (see Fig. 5 for all thresholds),
    b. *decreases predictive quality by 280 bps (AUC 75.34% [75.71% - 77.09%]), and*
    c. *decreases firm profitability on average by 0.11% [-44.31% - -0.01%];*
3. ***Up-sampling the training data to rebalance gender*** *(US:M2)*
    a. *decreases discrimination by 2.55% (0.0496 [0.0489 – 0.0503], τ=0.13)* (see Fig. 5 for all thresholds),
    b. *does not significantly impact predictive quality (AUC 77.36% [76.68% - 78.03%]), and*
    c. *decreases firm profitability on average by 0.04% [-10.15% - -0.01%].*

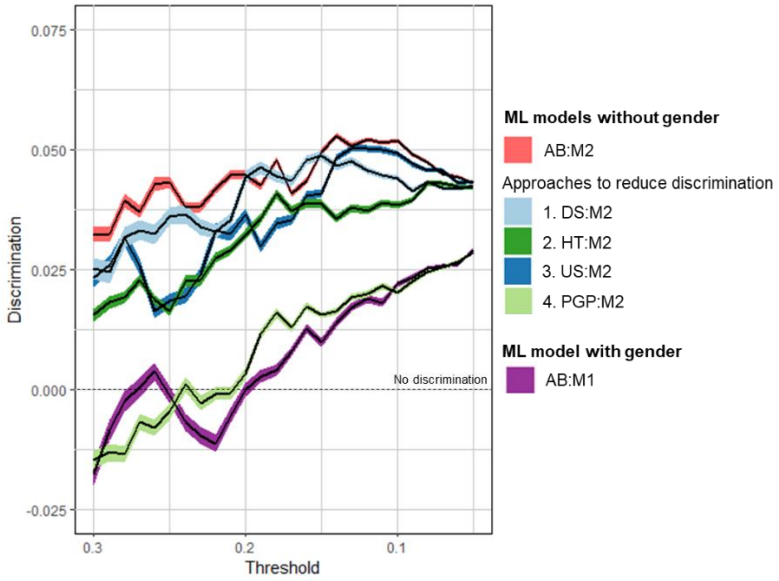

Fig. 5. Approaches to reduce gender discrimination in machine learning models

The key insight is that although fintech lenders under Level 2 regulation (i.e., countries in the EU) cannot use gender to train the final lending accept/reject model, they can use it to perform several discrimination-reducing approaches. The ultimate 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. The good news is that firms in Level 2 jurisdictions have several possibilities to reduce discrimination using data science techniques.

7.2. Discrimination reducing techniques for firms operating across jurisdictions

Here we explore an additional approach that may be technically feasible for a firm that operates in several jurisdictions: a probabilistic gender proxy model (PGP:M2). This involves first training an ML model to predict gender, or "impute" in the lingo of Zhang (2018), and then using that gender prediction as a feature in the accept/reject default prediction model. Barring distributional shift and data consistency, the lender could use data from a Level 1 or 2 jurisdiction to create a model to predict the gender of borrowers, then apply that model to predict gender for applicants in the Level 3 jurisdiction. Our gender prediction model achieved a five-fold cross-validated AUC of 91.08%, implying that gender could, in fact, be predicted with excellent accuracy from the other 700+ available features. We tuned the gender classification threshold to 20% to closely match the predictions of the model with access to gender (Model 1), which we know (per Observation #2) has lower discrimination, and higher profitability compared to the model without gender.

Observation #9*: Compared to the AB model without gender (AB:M2),*

4. ***Probabilistic gender proxy, PGP, modeling*** *(PGP:M2)*

a. *decreases discrimination by 62.08% (0.0193 [0.0187 – 0.0200] vs AB:M2 0.0509 [0.0503 – 0.515] τ=0.13),* (see Fig. 5 for all thresholds),
b. *does not significantly impact predictive quality (AUC 78.34% [77.68% - 79.00%] vs. AB:M2 78.14% [77.48% - 78.81%]), and*
c. *increases firm profitability by an average 0.01% (17.56% of the profit differences are statistically significant, 43% [-0.13 - -0.01%], and 57% [0.01- 7.64%]).*

This illustrates the benefits of probabilistic gender proxy modeling for applicants (reduced discrimination), and fintech firms (increased predictive quality and profitability). Unfortunately, we determined the methodology is currently prohibited in the US (one of the largest jurisdictions under Level 3 regulation), and has been observed to generate upward statistical bias in default predictions, albeit in the mortgage setting, not consumer credit, (Chen et al. 2019). Down-sampling, gender aware hyperparameter tuning, and up-sampling also cannot be implemented by fintech firms under Level 3 regulation (i.e., the US) as they are prohibited not just from using, but also collecting gender, which means fintech firms under this level of law, like the Apple Card, are restricted in their ability to reduce discrimination.

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

Lastly, we return to the operational modeling guidance of Level 1 regulations, which allow for both the collection *and* use of gender in the lending model. Summarizing the findings of several observations throughout the paper: the machine learning model, with gender, results in the lowest discrimination across thresholds (AB:M1, Fig. 5), the highest predictive quality, and firm profitability, compared the ML model that excludes gender, the ML approaches we suggest to reduce discrimination in the absence of gender, and the LR models. In short, our results suggest that the best way to reduce automated bias in this setting is to use machine learning models and allow for both the collection and use of gender.

8. Discussion & conclusions

We use publicly available, large, real alternative fintech data to investigate the impact of the three levels of anti-discrimination laws on gender discrimination: Level 1 laws, which allow for the collection and use of protected attributes, Level 2, which allow for only the collection but prohibit the use of gender in the accept/reject decision model, and Level 3 which prohibit both the collection and use. We find that prohibiting the use of gender in the final accept/reject model (per Level 2 and 3 laws), leads to increased discrimination, and decreased firm profitability, without significantly impacting model predictive quality, in both traditional statistical and machine learning lending models. We find that across all levels of anti-discrimination laws, ML models are less discriminatory, of better predictive quality, and higher

profitability when trained on the alternative data commonly used by fintech firm. We determine that the ML discrimination is driven by a novel phenomenon, *gender-blind feature selection*, a process we show is vastly different from the OVB that drives discrimination in traditional statistical models.

In addition, we show that the seemingly subtle difference between Level 2 and 3 regulations, allowing for the collection of gender, presents fintech firms under Level 2 regulation with four possible approaches to reduce discrimination: 1) down-sampling the training data to rebalance gender , 2) gender-aware hyperparameter tuning, 3) up-sampling the training data to rebalance gender, and 4) probabilistic gender proxy modeling, each with varying impacts to model predictive quality and firm profitability. We conclude by highlighting the importance of allowing for the collection and use of gender to reduce discrimination whilst preserving predictive quality and firm profitability.

The overarching implication of our work is that the growing adoption of algorithmic decision-making in non-mortgage consumer credit lending requires a rethink of the anti-discrimination laws and their operational guidance, specifically with respect to the collection and use of protected attributes. Our analysis points to the importance of allowing for the responsible collection and use of gender data, in line with the operational guidance of Level 1 regulations. Allowing fintech firms to collect protected attributes, like gender, would at minimum, give them the ability to assess the potential bias in their model, and potentially allow them to attempt 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. Recent work has begun to investigate alternative ways for firms to check for fairness without directly collecting sensitive attributes, like gender (see e.g., Kilbertus et al. 2018, Veale and Binns 2017), however the legality of these methods is not yet clear.

Increased data access should however, come with greater firm accountability and responsibility. For example, in Singapore, the FEAT Principles[15], which this paper's authors had the privilege to contribute in the development of, recommend that lenders should be able to collect and use protected attributes, like gender and race in their algorithmic lending models, but are responsible for discrimination in the algorithmic output. This is contrary to the situation in the US where lenders have used the existing laws as a way to elude responsibility for discriminatory outcomes, as Goldman Sachs did with their Twitter statement mentioned in the Introduction: "*we have not and never will make decisions based on*

[15] https://www.mas.gov.sg/~/media/MAS/News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf

factors like gender. In fact, we do not know your gender or marital status..." To that end, as of mid-2020 both the US and the EU have proposed regulatory guidelines for the responsible and ethical use of artificial intelligence. Both draft regulations will likely have implications for automated algorithmic decision-making in consumer credit.

The US draft regulation, titled "Maintaining American Leadership in Artificial Intelligence[16]," 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[17]" in which offers more structured guidance to firms. The Act suggests users of automated algorithms perform a bias impact assessment to mitigate potential discrimination. A consequence of our findings is that in the US, the ECOA will make it virtually impossible for lenders to adhere to the new proposed Act as they will not be able to test for discrimination without first being able to collect protected attributes like race, disability, and gender.

The European Commission's new draft regulation titled "White Paper on Artificial Intelligence: A European approach to excellence and trust,[18]" discusses that future regulation will likely include additional requirements for "high-risk" applications including stronger controls on training data, data governance and explainability, reporting, robustness and accuracy, and human oversight. It is unclear as of yet whether non-mortgage consumer lending will be deemed high-risk, but the possible approaches we discuss, specifically down-sampling, and up-sampling to rebalance gender are two methods that could ensure "a sufficiently representative training dataset," per the White Paper's recommendations.

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

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

---

[16] https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/

[17] https://www.congress.gov/bill/116th-congress/house-bill/2231/all-info

[18] https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

References

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

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

AlgorithmWatch, Bertelsmann Stiftung (2019) *Automating Society: Taking Stock of Automated Decision-Making in the EU*

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

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

Banasik J, Crook J, Thomas L (2003) Sample selection bias in credit scoring models. *J. Oper. Res. Soc.* 54(8):822–832.

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

Bartlett R, Morse A, Stanton R, Wallace N (2019) *Consumer-Lending Discrimination in the FinTech Era**

Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2017) *Fairness in Criminal Justice Risk Assessments: The State of the Art*

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

Chen IY, Johansson FD, Sontag D (2018) Why is my classifier discriminatory? Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. *Adv. Neural Inf. Process. Syst.* (Montreal, Canada), 3539–3550.

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

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

Cohen M, Fiszer MD, Ratzon A, Sasson R (2019) *Incentivizing Commuters to Carpool: A Large Field Experiment with Waze*

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

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

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

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

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

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

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

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

Fuster A, Goldsmith-Pinkham P, Ramadorai T, Walther A (2018) *Predictably Unequal? The Effects of Machine Learning on Credit Markets*

Ge Y, Knittel, Christopher R, MacKenzie D, Zoepf S (2016) *Racial and Gender Discrimination in Transportation Network Companies* (Cambridge, MA).
Hadfield G (2016) *Rules for a Flat World: Why Human Invested Law and How to Reinvent It for a Complex Global Economy* (Oxford University Press, New York).
Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. *Adv. Neural Inf. Process. Syst.*:3323–3331.
Hellman D (2016) Two Concepts of Discrimination. *Virgina Law Rev.* 102(4):895–952.
Henley WE, Hand DJ (1997) Statistical Classification Methods in Consumer Credit Scoring: a Review. *J. R. Stat. Soc. Ser. A (Statistics Soc.* 160(Part 3):523–541.
Holstein K, Vaughan JW, Daumé H, Dudík M, Wallach H (2019) Improving fairness in machine learning systems: What do industry practitioners need? *Conf. Hum. Factors Comput. Syst. - Proc.* 1–16.
Hurley M, Adebayo J (2016) Credit Scoring in the Era of Big Data. *Yale J. Law Technol.* 18(1):148–216.
Kamiran F, Calders T (2012) Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* 33(1):1–33.
Kilbertus N, Gascón A, Kusner M, Veale M, Gummadi KP, Weiler A (2018) Blind justice: Fairness with encrypted sensitive attributes. *Proc. 35th Int. Conf. Mach. Learn.* (Stockholm, Sweden), 2630–2639.
Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2018) Human Decisions and Machine Predictions. *Q. J. Econ.* 133(1):237–293.
Lambrecht A, Tucker C (2019) Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Manage. Sci.* 65(7):2966–2981.
Lessmann S, Baesens B, Seow HV, Thomas LC (2015) Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *Eur. J. Oper. Res.* 247(1):124–136.
Lundberg SM, Erion GG, Lee SI (2019) *Consistent Individualized Feature Attribution for Tree Ensembles*
Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds. *Adv. Neural Inf. Process. Syst. 30*. (Curran Associates, Inc.), 4765–4774.
Mejia J, Parker C (2020) When Transparency Fails: Bias and Financial Incentives in Ridesharing Platforms. *Manage. Sci.* (Articles in Advance 05 May 2020).
Monetary Authority of Singapore (2018) *Guide to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in the Singapore Financial Sector*
Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science (80-.).* 366(6464):447–453.
Ongena S, Popov A (2016) Gender Bias and Credit Access. *J. Money, Credit Bank.* 48(8):1691–1724.
Perrone V, Donini M, Kenthapadi K, Archambeau C (2020) Fair Bayesian Optimization. *7th ICML Work. Autom. Mach. Learn.* 1–15.
Pope DG, Sydnor JR (2011a) Implementing Anti-Discrimination Policies in Statistical Profiling Models. *Am. Econ. J. Econ. Policy* 3(3):206–231.
Pope DG, Sydnor JR (2011b) What's in a Picture? Evidence of Discrimination from Prosper.com. *J. Hum. Resour.* 46(1):53–92.
Stein RM (2005) The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing. *J. Bank. Financ.* 29(5):1213–1236.
Tang CS, Yang SA, Wu J (2018) Sourcing from suppliers with financial constraints and performance risk. *Manuf. Serv. Oper. Manag.* 20(1):70–84.
Taylor W (2011) Proving Racial Discrimination and Monitoring Fair Lending Compliance: The Missing Data Problem in Nonmortgage Credit. *Rev. Bank. Financ. Law* 31:199–264.
Thomas LC, Edelman DB, Crook JN (2017) *Credit Scoring and Its Applications* 2nd ed. (Society for Industrial and Applied Mathematics Publishing, Philadelphia).
U.S. Department of Justice (2017) The Equal Credit Opportunity Act. Retrieved (November 9, 2018), https://www.govinfo.gov/content/pkg/USCODE-2011-title15/html/USCODE-2011-title15-chap41-subchapIV.htm.
Veale M, Binns R (2017) Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data Soc.* 4(2):205395171774353.

Veale M, Edwards L (2018) Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling. *Comput. Law Secur. Rev.* 34(2):398–404.
Wooldridge JM (2015) *Introductory Econometrics A Modern Approach* 5th ed.
Younkin P, Kuppuswamy V (2018) The colorblind crowd? Founder race and performance in crowdfunding. *Manage. Sci.* 64(7):3269–3287.
Zafar MB, Valera I, Rodriguez MG, Gummadi KP (2017) Fairness constraints: Mechanisms for fair classification. *Proc. 20th Int. Conf. Artif. Intell. Stat. AISTATS 2017* 54.
Zhang Y (2018) Assessing Fair Lending Risks Using Race/Ethnicity Proxies. *Manage. Sci.* 64(1):178–197.
Zliobaite I (2015) On the Relation Between Accuracy and Fairness in Binary Classification. *Proc. 2nd Work. Fairness, Accountability, Transpar. Mach. Learn.*

Supplemental Materials

Anti-discrimination laws, specifically with respect to gender and credit

The exact anti-discrimination legislation governing non-mortgage lending decisions clearly depends on the jurisdiction, but as we discuss below, there are certain marked similarities. *Level 1* regulations allow lenders to collect and use gender throughout the lending modeling process. *Level 2* regulations are more stringent, and allow lenders to collect gender data, but prohibit its use in the accept/reject lending model used to grant credit. *Level 3* regulations are the strictest and prohibit lenders from collecting and using gender data.

In the United States (US), the Equal Credit Opportunity Act (ECOA), and its accompanying Regulation B, prohibits lenders from discriminating against credit applicants based on sex, along with several other protected attributes (U.S. Department of Justice 2017). It prohibits two types of discrimination: disparate treatment, when an individual is treated unfavourably on the basis of one or more protected attributes, like gender; and disparate impact, when a supposedly neutral lending policy results in less favorable terms for members of a protected group, compared to another group of similar applicants (Hurley and Adebayo 2016). Regulation B of the ECOA explicitly attempts to prevent lenders from making discriminatory lending decisions (specifically, disparate treatment) by prohibiting them from requesting information about an applicant's gender[19]. In practice, this means data used for nonmortgage credit decisions in the US seldom includes gender. As the regulation prevents both the collection and use of gender data, we categorize the ECOA as Level 3 regulation.

In the European Union, Directive 2004/133/EC (commonly referred to as the EU Gender Directive) prohibits direct and indirect discrimination based on gender in the pricing, access to and supply of goods and services (European Commission 2012). The Gender Directive is referenced in the EU's Consumer Credit Directive (Directive 2008/48/EC) to ensure non-discrimination in the provision of consumer credit in the EU. Article 4 of the Gender Directive (Principle of equal treatment), states there should be no direct discrimination, "where one person is treated less favourably, on grounds of sex, than another is, has been or would be treated in a comparable situation" (Council of the EU 2004). In practice this means that although there may be differential pricing offered to men and women, the difference must be based on actions or behaviours linked to the lending product, not an applicant's gender. Article 4 also states there should be no indirect discrimination, "where an apparently neutral provision, criterion or practice would put persons of one sex at a particular disadvantage compared with persons of the other sex,

[19] With the caveat that a creditor may ask for an applicant's courtesy title (Mr., Mrs., Dr., etc.) as long as it is an optional disclosure (U.S. Department of Justice 2017).

unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary" (Council of the EU 2004). The operational modeling guidance, with respect to "actuarial factors" is addressed in Article 5, however the March 2011 judgement, referred to as "the Test-Achats ruling" makes the original directive invalid from December 21, 2012 onwards (European Commission 2012). The Test-Achats ruling, specifically Article 5(1) "prohibits any results whereby differences arise in individuals' premiums and benefits due to the use of gender as a factor in the calculation of premiums and benefits" (European Commission 2012). The law explicitly addresses insurers, and providers of other related financial services such as pensions, but a 2018 court ruling in Finland, an EU member state, references the Test-Achats ruling in a multiple discrimination case (gender, language, and age) in consumer credit provision as support for their claim that "gender in the scoring system involves…discrimination…" (AlgorithmWatch and Bertelsmann Stiftung 2019). So, although the ruling does not explicitly discuss consumer credit provision, in practice it has been applied to this setting. Importantly, the Test-Achats ruling states that it "remains possible to collect, store and use gender status or gender-related information…," and provides examples including reserving and internal pricing, reinsurance pricing, and marketing and advertising. Ultimately this means that gender is prohibited from being used in the final accept/reject decision-making models as it could lead to differences in the provision of credit directly due to the use of gender but, the lender can still collect gender and use it in other ways. A review of several EU online consumer credit application system confirmed that gender is collected in practice. EU member states have the option to put in place additional regulations, but to date no members have further restricted the collection of gender. Although Article 9 (Processing of special categories of personal data) in the General Data Protection Regulation (GDPR) prohibits the collection of several protected characteristics, gender is not included (Veale and Edwards 2018), so the operational guidance in GDPR does not impact consumer non-mortgage credit provision. As the EU Gender Directive allows for the collection of gender, but not its use in the final lending accept/reject decision-making model, we categorize it as Level 2 regulation.

Singapore has specific regulatory guidance on the use of algorithms, analytics, and artificial intelligence, with its Guide to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in the Singapore Financial Sector. The guidance does not prohibit the collection or use of gender (Monetary Authority of Singapore 2018), and interviews on the topic with several data science experts at leading banks in Singapore have confirmed this; we therefore categorize it as Level 1 regulation.

Comparison of discrimination in traditional statistical and machine learning models under Level 1 regulation

Observation #3(B): *Compared to the LR model with gender (LR:M1), the AB model (AB:M1)*

a. *reduces discrimination by 41.7% (0.0183 [0.0175 – 0.0191] vs. LR:M1 0.0314 [0.0308 – 0.0320], τ=0.13), without gender* (see Fig. B for all thresholds),
b. *increases predictive quality by 482 bps (AUC 78.29% [77.63% - 78.95%] vs. LR:M1 73.47% [72.75% - 74.20%]), and*
c. *increases firm profitability on average by 7.58% (89.88% of the profit differences are statistically significant, 97% [0.04 -58.03%], and 3% [-0.007 - -0.01%]).*

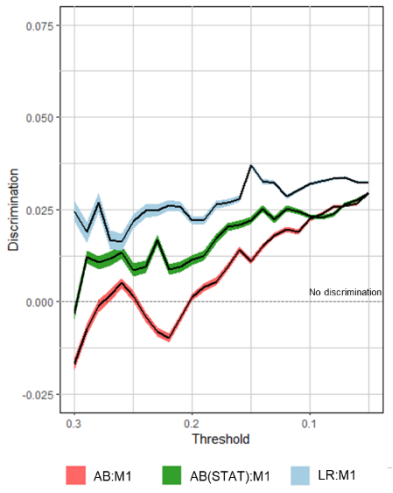

Fig. B. Average Blender & Logistic Regression models with gender

As we do in the main analysis in §5.1, for robustness we report the results of a second AB model trained on the traditional statistical feature set used by the LR model (AB(STAT):M1).

Observation #4(B): *Compared to the LR model with gender (LR:M1), the AB model trained on the traditional statistical feature set (AB(STAT):M1)*

a. *reduces discrimination by 29.62% (0.0221 [0.0212 – 0.0230] vs. LR:M1 0.0314 [0.0308 – 0.0320], τ=0.13),*
b. *increases predictive quality by 208 bps (AUC 75.55% [74.85% - 76.25%] vs. LR:M1 73.47% [72.75% - 74.20%]), and*
c. *increases firm profitability on average by 7.00% (89.63% of the profit differences are statistically significantly difference [0.02 - 57.27%]).*

These results illustrate that when gender can be used in the lending model (per Level 1 laws), the AB model is less discriminatory, of better predictive quality, and greater profitability than the LR model.

Supplemental References

AlgorithmWatch, & Bertelsmann Stiftung. (2019). *Automating Society: Taking Stock of Automated Decision-Making in the EU*. https://algorithmwatch.org/wp-content/uploads/2019/01/Automating_Society_Report_2019.pdf

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

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

Hurley, M., & Adebayo, J. (2016). Credit Scoring in the Era of Big Data. *Yale Journal of Law and Technology*, *18*(1), 148–216. https://doi.org/10.3868/s050-004-015-0003-8

Monetary Authority of Singapore. (2018). *Guide to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in the Singapore Financial Sector*.

U.S. Department of Justice. (2017). The Equal Credit Opportunity Act. https://www.govinfo.gov/content/pkg/USCODE-2011-title15/html/USCODE-2011-title15-chap41-subchapIV.htm. Accessed 9 November 2018

Veale, M., & Edwards, L. (2018). Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling. *Computer Law and Security Review*, *34*(2), 398–404. https://doi.org/10.1016/j.clsr.2017.12.002