Algorithmic Fairness

Sanjiv Das,[1] Richard Stanton,[2] and Nancy Wallace[3]

[1]Leavey School of Business, Santa Clara University, Santa Clara, CA 95053
[2]Haas School of Business, U.C. Berkeley, Berkeley, CA 94720
[3]Haas School of Business, U.C. Berkeley, Berkeley, CA 94720; email: newallace@berkeley.edu

Annual Review of Financial Economics
2022. 14:1–32

https://doi.org/10.1146/((please add article doi))

Keywords

algorithms, machine learning, bias, fairness metrics, credit scoring
JEL C55, G21, G23, G28

Abstract

This paper reviews the recent literature on algorithmic fairness, with a particular emphasis on credit scoring. We discuss human vs. machine bias, bias measurement, group vs. individual fairness, and a collection of fairness metrics. We then apply these metrics to the U.S. mortgage market, analyzing HMDA data on mortgage applications between 2009 and 2015. We find evidence of group imbalance in the dataset for both gender and (especially) minority status, which can lead to poorer estimation/prediction for female/minority applicants. Loan applicants are handled mostly fairly across both groups and individuals, though we find that some local male (non-minority) neighbors of otherwise-similar rejected female (minority) applicants were granted loans, something that warrants further study. Finally modern machine-learning techniques substantially outperform logistic regression (the industry standard), though at the cost of being substantially harder to explain to denied applicants, regulators, or the courts.

Contents

1. INTRODUCTION

In March 2022, Aaron Braxton, a Black homeowner whose application to refinance his mortgage had been rejected by Wells Fargo, sued the bank for discrimination following the publication of a Bloomberg report showing that they denied over 50% of all refinancing applications by Black borrowers in 2020.[1] In 2017, Amazon shut down an experimental machine-learning-based recruiting tool after finding that they could not stop it from discriminating against women (see Dastin 2018). And in fall 2019, Obermeyer et al. (2019) showed that an algorithm used by health insurer UnitedHealth to allocate hospital resources was inadvertently leading to discrimination against African-American patients.

New statistical methods and machine-learning (ML) techniques, especially for predicting prospective-borrower creditworthiness, have been rapidly adopted in the financial services industry to increase efficiency, expand access to credit, and enhance profitability. The pace of this adoption has led to interest in evaluating the risks associated with the use of artificial-intelligence (AI) and ML tools and in determining whether the gains from their adoption are evenly distributed in U.S. society (see Kleinberg et al. 2018a,b).[2]

A key policy question is whether AI/ML technology leads to more gender, age-related, ethnic and racial discriminatory biases in consumer credit markets. A second potential risk with the broader application of AI/ML technology in consumer lending includes a type of

[1] This rate was higher than for other banks (see Flitter 2022, Donnan et al. 2022).

[2] ML algorithms, by design, reduce the predictive mean-squared error, so produce predictions with greater variance. There will thus always be some classes of borrower who are systematically scored as riskier, and others who are systematically scored as less risky, than under preexisting technologies.

Hirshleifer effect (Hirshleifer 1971), wherein improved prediction may reduce the opportunities to share risk across agents and may therefore be welfare decreasing for society. The significant growth in algorithmic decision-making, due in part to the availability of unprecedented data on individuals and the accompanying rise in AI/ML tools, in conjunction with examples like those above, has also heightened concern that algorithmic decision-making may not eliminate possible face-to-face biases in consumer lending. The opacity of AI/ML, and the problems that can arise with its use in consumer lending, are emergent phenomena resulting from the interaction of straightforward algorithms with complex data to produce complex predictions (see Kearns & Roth 2020, Barocas & Selbst 2016). Of course, the use of AI/ML has more than one effect. In particular, by reducing costs and increasing the speed of decision-making, it may allow access to markets by people who were previously excluded (see, for example, Fuster et al. 2019). While this paper focuses solely on fairness, it is important to keep all of these effects in mind.

The paper is organized as follows. Section 2 presents an overview of issues in algorithmic fairness, including human vs. machine bias, bias measurement and attributes, antecedents of machine bias, group vs. individual fairness, bias stages, and value systems. Section 3 discusses a selection of fairness metrics, and Section 4 presents an empirical application to mortgages. Section 5 concludes.

2. ISSUES IN ALGORITHMIC FAIRNESS

Although there are no universal definitions of artificial intelligence or machine learning, for current purposes AI can be defined as the development of computer systems to perform tasks that ordinarily require human intelligence. This definition incorporates expert systems — where humans teach machines — and also machine learning, where the machines learn from data. Interest in ML in particular has become increasingly popular, due to a combination of more digitized data, faster computers, and better algorithms to analyze data. ML is similar to statistics in that both seek to learn from the data and use many of the same tools, and the two disciplines are increasingly learning from each other. The biggest difference is that statistics has historically emphasized hypothesis testing and statistical inference, whereas ML emphasizes obtaining the best prediction. As a result, ML is not guided by economic (or other social sciences) theory (which would generate the hypotheses for statistical testing), which has the advantage that ML sometimes identifies relationships that are not (currently) predicted by theory. The disadvantage is that some of the relationships ML identifies will not be causal and, hence, cannot be usefully exploited. AI and ML are general-purpose technologies that may be used in a wide variety of areas within a financial institution. These include refinements to existing products, such as better credit and risk management, tools for uncovering asset pricing anomalies, and helping institutions comply with regulatory requirements; this is a related field called "RegTech." AI and ML are also essential inputs into the creation of a variety of new financial services.

As many domains of application of human judgment increasingly yield to automated machine-based decision-making, the question of algorithmic fairness becomes critical. Bias is rife in human decisions and is exacerbated by machines that are trained on biased datasets, in several domains such as finance, healthcare, college admissions, criminal cases, etc., as highlighted by O'Neil (2016).

We survey the literature in the domain of algorithmic fairness and develop a framework that broadly captures the scope of this field as it pertains to the financial domain. At a

narrow level, we hope to provide a checklist for algorithmic fairness that may be deployed for financial algorithms.[3]

Algorithmic fairness relates to many regulations and legal frameworks, such as the US Civil Rights Act of 1964, the European Union's General Data Protection Regulation (GDPR), the Fair Credit Reporting Act (FCRA), the Equal Credit Opportunity Act (ECOA), SR11-7 regulation, and Reg B, to name a few. These regulations have largely applied to decision-making by humans and not by algorithms.

2.1. Human vs. Machine Bias

What are the differences between bias imposed by humans in the loop versus machines? *First*, human bias is applied far less systematically than machine bias. Once a machine is instantiated with a biased algorithm, it will apply it tirelessly.

Second, bias can cascade with machine learning. If the dataset on which the algorithm is trained is biased, the machine learns the bias and displays it in its decisions, which are eventually added to the data, thereby exacerbating the bias. With humans, this effect is less prevalent, but existent nonetheless. As one example of perpetuating dataset bias, consider the gender bias in small-business lending in the US, which comes from historically small fractions of women in the borrower pool (Alesina et al. 2013, Chen et al. 2017, Brock & De Haas 2021). A similar bias is evidenced in the hiring of women (see Iris 2016).

Third, human decisions, even when biased, are often tempered within reasonable bounds, unlike machines, which may spiral out of control. For example, AIs that generate text, such as ChatGPT, are trained on very large text corpora, such as Wikipedia, Book Corpus, etc. They learn to write by effectively "reading and memorizing" much of the text that is available in huge text databases, both public and private. Hence, AIs may regurgitate hate speech, learning from datasets that contained several unacceptably biased samples of text. GPT-3 from OpenAI learned to generate text biased against Muslims.[4] It also made racist jokes, condoned terrorism, and accused people of being rapists.[5] Nurture, applied mechanistically and at scale, can simulate the darker side of our nature quite effectively!

Fourth, explaining bias in black box algorithms may appear to be hard, but is quite feasible with modern tools that explain machine decisions made by deep-learning algorithms (Samek & Müller 2019, Srinivas & Fleuret 2019, Arya et al. 2019). These papers provide explanations by identifying which inputs to the model were salient in determining the model's decision. In some cases, stronger explanations that identify cause and effect between inputs and predictions are available as in Budhathoki et al. (2022). Asking a human to explain how they arrived at a decision is often much harder. Rudin (2019) argues that while ex post explanations are a good start in assessing if bias exists, better ex ante design of machine learning algorithms for interpretability is needed.

2.2. Bias Measurement and Attributes

How is bias measured? The literature is now replete with different types of bias metrics, considered in useful overviews by Bellamy et al. (2019), Barocas et al. (2019), Pessach &

[3] https://www.oreilly.com/radar/of-oaths-and-checklists/

[4] https://www.vox.com/future-perfect/22672414/ai-artificial-intelligence-gpt-3-bias-muslim

[5] https://www.wired.com/story/efforts-make-text-ai-less-racist-terrible/

Shmueli (2020), Mehrabi et al. (2021), Das et al. (2021), Hardt et al. (2021). The main metrics in the literature will be covered in Section 3. These metrics will measure bias in data and in models.

In addition to metrics, model explanations are important in ascertaining fairness. By highlighting which aspects of the feature set (independent variables) the model focused on in making its decision, one may be able to investigate if and how bias is present in the algorithm. For instance, if the model focused primarily on age or gender in a hiring decision, the explanation would reveal bias (Iris 2016).

In this vein, counterfactual explanations may be especially helpful (Kusner et al. 2017, Barocas et al. 2020), as they may delineate causal drivers of bias (Pearl 2009, Menzies & Beebee 2019). As an example, if a senior citizen were denied a service, but an otherwise identical but younger individual was not declined, it would suggest age bias, detected by such counterfactual tests (suggested by Black et al. 2020).

In addition to bias measurement and mitigation being interwoven with the evolving literature on explaining machine models, it is also connected to other attributes of automated decision making. *First*, bias is a function of model accuracy. When models are inaccurate, they are unfair even though sometimes they may result in undeserved benefits to some people, while handing out erroneous negative outcomes to others. Both Type I and II errors treat recipients of the model's decisions unfairly. Different levels of accuracy may be observed across sub-segments of the population. For example, if a model is more accurate in its predictions for one gender versus another, then it is biased against the latter group. Diana et al. (2021) recommend the objective of minimax group fairness to manage such bias. The minimax criterion minimizes (across groups) the maximum difference in outcomes within groups.

Second, bias is interwoven with the notion of causality (Pearl 2009). Determining causal drivers of bias may be more effective in its mitigation (Galhotra et al. 2017). *Third*, a lack of privacy may lead to bias, as implicitly or explicitly revealing personally identifiable information (PII) may unlock biased algorithms (Dwork et al. 2012). As an example of bias mitigation, auditions often place musicians behind a screen to prevent decision makers being influenced by protected characteristics (e.g., age, gender, race, etc.). In the machine world, such characteristics are excluded from the feature set, and ex post checking can also reveal if the characteristics "leaked" into the model's predictions through the other features in the dataset, e.g., zip code as a proxy for ethnicity. However, such checking of models is often difficult, as today's models are "opaque, unregulated, and uncontestable" (O'Neil 2016). Getting these three elements (accuracy, causality, privacy) right is a key element for developing trust in the fairness of algorithms.

2.3. Where does Machine Bias Come From?

There are several antecedents of algorithmic bias noted in Das et al. (2021). *First*, labels may be biased because they are generated by biased humans, e.g., decisions driven by stereotyping eventually accumulate in datasets. Inconsistent labeling arises in public datasets with multiple labelers, as in police data, public opinion datasets, etc. (see Wauthier & Jordan 2011). For example, when humans are asked to label facial expressions as happy or sad, their subjective decisions may vary considerably, making the model less accurate and harder to train. *Second*, feature bias or "curation" bias occurs when the modeler is biased towards choosing some features over others, which leads to disfavoring a protected

group. In lending, choosing features such as income and education or social network data, but excluding features such as community service, engagement in sports, etc., may disfavor minorities, even though all these features reflect responsible social behavior and would improve credit scores.

Third, "objective function" bias (Menestrel & Wassenhove 2016). If the loss function is overly focused on outliers, which tend to be more from one group than another, it may result in models that treat each group differently. *Fourth*, as machines are used to make predictions, if these are biased, then they feed back into the training of future models, resulting in perpetuating the same kind of biased labels. For example, an algorithm trained to detect hit songs will then perpetuate the same kind of music as the hit song, leading to "homogenization" bias. *Fifth*, "active" bias arises deliberately, as in the case of fake news, i.e., text generation of false material, or when a racist loan officer singles out and rejects loan applicants of a certain ethnicity, overriding the recommendations from a machine algorithm. *Sixth*, "random" bias may occur if an algorithm does not have guard rails or constraints. A machine tasked with eradicating poverty may find that the optimal solution is the elimination of poor people, which would be inadmissible unless constraints on action are missing.

2.4. Bias across data modalities

An interesting aspect of algorithmic fairness arises when we consider different data modalities. Is bias measurement in tabular datasets very different from that of text or images? At higher levels of abstraction, the differences become minor. For instance, if bias is being assessed for a protected class in a dataset, e.g., redlining in mortgage lending, whether the feature set is tabular or text will not matter much, as the predictions of the model may be assessed for bias in a manner that is independent of the modality of the feature set. As we will see in Section 3, some metrics do not need the feature set, only the labels and the protected classes.

Further, bias assessment applied to tabular or text datasets can be applied also to multimodal combinations of text and tabular data. However, it is also conceivable that the range of bias in some modalities may be greater. In the case of natural language generation, the generation of racially-biased text may be quite varied in form, and also harder to anticipate. Bias in text embeddings is known and measurable, and techniques for mitigation are being developed (Papakyriakopoulos et al. 2020). Detecting and measuring such bias becomes harder, and in fact, it may be such that you know it when you see it as humans, but it is much harder to codify an algorithm to detect such bias. Similar instances arise in computer vision, where in the case of facial recognition, datasets tend to be overpopulated by white males, leading to less accurate detection of other sorts of people (Grother et al. 2011, Lunter 2020). Further, detecting biases in the captioning of images may be much harder.[6]

2.5. Group vs. Individual Fairness

An important dimension for the assessment of fairness is group fairness versus individual fairness. An algorithm may be fair across groups but unfair to individuals within groups

[6] `https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/`

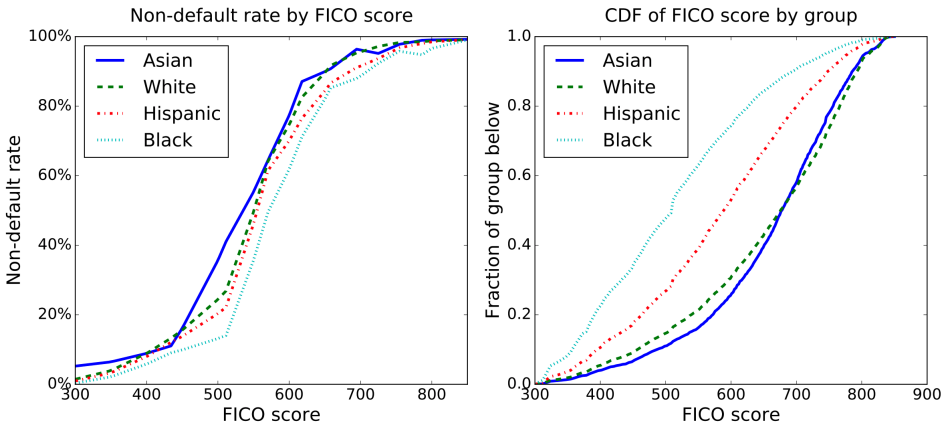


Figure 1: Bias in FICO scoring in comparison to default rates. This pair of plots, taken from Hardt et al. (2016), shows potential group unfairness in credit scoring.

(see Hutchinson & Mitchell 2019, for an excellent survey of this and other issues). For example, classification accuracy may be poor in one group.

In an interesting study of FICO scores, Hardt et al. (2016) show that individual fairness may be incentivized by shifting the cost of poor classification from disadvantaged groups to the decision maker, who may respond by improving the classification accuracy in a group that has faced less accurate classification. A classic depiction of potential group unfairness can be seen from the plots in their paper, shown in **Figure 1**. Given equal default rates by FICO score, an unconditional analysis would suggest that minorities appear to receive a larger number of low FICO scores. However, this may not be the case after conditioning on borrower features. Therefore, whether bias exists depends very much on the metrics used for algorithmic fairness and what features are conditioned upon, and we discuss various metrics in Section 3.

We may achieve group fairness, yet fail miserably on individual fairness. Fairness for individuals is based on the idea that similar people are treated similarly and protected characteristics are not used for decision-making. This may be tested using counterfactual approaches suggested in (Black et al. 2020), where the test looks at whether individuals who are otherwise similar (i.e., close in characteristics) but for their protected characteristics (such as gender, race, etc.) are treated in the same way. A useful mathematical criterion for individual fairness is suggested in Dwork et al. (2012), via a Lipschitz condition that requires that any two individuals x, y that are at distance $d(x, y) \in [0, 1]$ map to distributions $M(x)$ and $M(y)$, respectively, such that the statistical distance between $M(x)$ and $M(y)$ is at most $d(x, y)$, i.e., satisfy the (D, d)-Lipschitz property $D[M(x), M(y)] \leq d(x, y)$. Thus, the distributions over outcomes observed by the two individuals are similar within a distance of $d(x, y)$. Stated simply, people within small groups should experience very similar ranges of outcomes. A related extension to individual fairness is proposed by Diana et al. (2021), where a minimax criterion is applied to minimize (across groups) the maximum difference in outcomes within group.

2.6. Bias Stages

Bias enters at different stages in the machine learning modeling pipeline, which we categorize into three broad stages in which we may impose fairness: pre-training, in-processing, and post-training.

In the pre-training phase we assess the data itself for bias and attempt to mitigate bias therein. Vasudevan & Kenthapadi (2020) and Das et al. (2021) offer several metrics for detecting such bias, ranging from differences in sizes of advantaged versus disadvantaged groups to label imbalance across groups. For example, the existence of very few women in small business loan datasets might drive a model to learn that women should not be given loans, even when the datasets labels are equally proportionate across genders. Or, there may be equal numbers of men and women in the dataset, but there is imbalance in the numbers of accepted and rejected loan applicants across genders.

In the in-processing stage, fairness may be imposed on the algorithm by fitting the model's objective function (e.g., loss minimization) while including a fairness constraint (see Perrone et al. 2019, 2021).

Ensuring fairness in the dataset as well as constraining the model training with fairness constraints may not result in fair classification or predictions by the trained model on the test dataset (see Canetti et al. 2019). Therefore, it is important to develop fairness/bias metrics for the model's predictions as well.

2.7. Value Systems

Notions of fairness perforce reflect the value system of the society in which the algorithms make decisions. These value systems are reflected in the socio-legal-ethical mores of the population. For example, ageism is a bias that matters because age is a protected characteristic in the US, but it may not be so in other countries.

Even when a protected characteristic is used to define a dimension of possible bias, different societies may have different thresholds for bias. For example, there is a greater tolerance for religious bias in some countries versus others. The abstraction of a "golden truth" (or ground truth) is often necessitated in assessing the tolerable extent of bias, though models agnostic to this are also proposed in the literature (see Aka et al. 2021).

The connection between value systems and acceptable levels of bias may be determined through the courts, resulting in thumb rules such as the 80% rule commonly applied in the context of loans. This rule implies that a model is fair when the success rate for any group is not less than 80% of that of any other group. Nuances to rules such as this may be driven by society's interpretation of what fairness means.

3. METRICS

In this section we summarize some common metrics to assess the fairness of algorithms in finance. Biases may exist in the dataset used to train financial models. Even when the data used is unbiased, the algorithm's predictions may be biased. The specific quantification of bias depends on the metric used to measure it, which in turn relies on the socio-legal-ethical value system underpinning fairness.

Fairness may be assessed at group levels and at individual levels. To make matters concrete, we assess potential indicators of bias, because the presence of a bias metric in excess of that expected is only a marker and would need further investigation for the actual

presence of bias. Bias may exist in (i) the dataset and/or (ii) in the classifier model. One, the dataset may be biased, in terms of imbalance in the numbers in the advantaged and disadvantaged groups. It may also be biased in the proportion of labels in each class across the groups. Two, the models that are fit may be biased in their decisions.

Below, we present some of the common metrics with examples. This is far from being an exhaustive list, though the metrics below appear in many papers in the literature cited in Section 2 and therefore, we chose these to review here. After introducing the notation in Section 3.1, we present dataset bias metrics in Section 3.2 and classifier model bias in Section 3.3.

3.1. Notation and Setup

To quantify the metrics we will discuss, we use a binary classification example. To fix ideas, consider a model that decides whether or not to approve a loan application. The true labels may come from a dataset where borrowers repaid a loan or defaulted on it. We use this dataset to train a model to discriminate between borrowers who will repay the loan versus those who will not.

Bias arises when one group is favored over another. For the purposes of this exposition we define these groups as the advantaged or privileged group denoted by "A" and the disadvantaged group "D". The size of group A (D) is n_A (n_D). The true labels in the dataset are denoted Y and the predicted labels from the model are denoted as $\hat{Y}$.

When we talk of a positive (negative) outcome, we think of either Y or $\hat{Y}$ as taking the value $+1$ (-1). For example, $\hat{Y} = +1$ if a loan applicant is approved. Likewise, the model correctly predicts the label if $Pr[\hat{Y} = +1|Y = +1] \geq \tau$, where τ is the threshold probability in the model (usually defaulted to $\tau = 0.5$). A similar condition applies to negative outcomes, i.e., the model is correct if $Pr[\hat{Y} = -1|Y = -1] \geq \tau$. We also define a bias threshold β, which is the extent of allowable bias under the metric. For example, if $\beta = 0.20$, then if the metric is less than β no bias is assumed.

We use a confusion matrix to delineate the performance of our classification algorithm. All the diagonal elements denote correctly classified outcomes, while the mis-classified outcomes are represented on the off diagonals of the confusion matrix. For example, say the model produces the following confusion matrix for all groups combined:

Groups A and D		**True Labels** -1	$+1$
Predicted Labels	-1	160	50
	$+1$	25	140

Here, the true positives (TP) number 140 and the true negatives (TN) 160. Likewise, there are 25 false positives (FP) and 50 false negatives (FN). We can compute various standard metrics from this confusion matrix:

(a) Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = 300/375 = 0.800$
(b) Precision: $\frac{TP}{TP+FP} = 140/165 = 0.848$
(c) Recall: $\frac{TP}{TP+FN} = 140/190 = 0.737$
(d) F1 score: $\frac{2}{1/Precision+1/Recall} = \frac{2(0.848\times0.737)}{(0.848+0.737)} = 0.789$

Precision measures the proportion of predicted positives that are correct. Recall measures the proportion of positives that are correctly predicted. The F1 score assesses the balance

between precision and recall and is the harmonic mean of the two. These metrics are standard in machine learning for classification models.

For bias assessment, the overall confusion matrix may be decomposed by group, i.e., one for the advantaged group and one for the disadvantaged group. We show these matrices below. First, we present the disadvantaged group:

Group D only		**True Labels** -1	$+1$
Predicted Labels	-1	70	40
	$+1$	5	60

Here, the accuracy is 0.743, precision is 0.923, recall is 0.600, and the F1 score is 0.727. The confusion matrix for the advantaged group is as follows:

Group A only		**True Labels** -1	$+1$
Predicted Labels	-1	90	10
	$+1$	20	80

The accuracy is 0.850, precision is 0.800, recall is 0.889, and the F1 score is 0.842. Imbalance between precision and recall is greater for group D.

The two individual group confusion matrices sum up to the overall confusion matrix. The model performs better for Group A than for Group D, in terms of accuracy and F1 score. Hence, Group D is subject to noisier outcomes, implying that even within group, outcomes may be less fair for D than A. Precision is higher for Group D, suggesting that the model may be only approving clearly good borrowers in Group D, but allowing more borrowers with poorer credit from Group A. Recall is lower for Group D than Group A, suggesting that the model proportionately declines loans for more deserving borrowers in Group D than it does for Group A. Thus, individual group confusion matrices may be used to assess model fairness quickly using simple arithmetic.

Which of these different aspects of fairness might we care more about? What are the sources of bias? These questions can be answered using specific bias metrics that we review next.

3.2. Dataset Fairness

A classifier may become biased if the dataset is imbalanced. Imbalance may arise from differences in the size of groups or from differences in the number of positive and negative labels across groups in the dataset. There are some simple measures to assess these imbalances:

1. *Group Imbalance (GI)*: If the number of members if Group A greatly exceeds the number in Group D, then this may bias a model, both in accuracy and predictions, so it is important to check for this. For example, if a small business loan dataset has very few women borrowers, then the model may be less accurate in assessing women's loan applications or may just learn that very few women should receive loan approvals. GI is fair when it lies within threshold, i.e.,

$$GI = \frac{n_D}{n_A} \geq 1 - \beta \qquad 1.$$

Suppose $\beta = 0.20$. In our example, we see from the confusion matrices that $n_A = 200$ and $n_D = 175$, i.e., $GI = 0.875 > 1 - \beta = 0.80$. Hence, group imbalance is above the required threshold and is not a concern.

2. *Class Imbalance (CI)*: In the same loan setting, if the ratio of approved to declined applications differs considerably across groups, it may lead to a biased classifier. This may be checked by the following condition:

$$CI = \frac{\#(Y = +1|A)}{\#(Y = -1|A)} - \frac{\#(Y = +1|D)}{\#(Y = -1|D)} \leq \beta \qquad 2.$$

From the example, we see that $90/110 - 100/75 = -0.433 < \beta$. CI is not likely to be a concern.

3. *Distributional Imbalance (DI)*: CI may be generalized to more than two classes using divergence formulas. For brevity, we present one such metric proposed in Das et al. (2021), namely KL divergence (Kullback & Leibler 1951), where

$$DI = \sum_y f_A(y) \log_2 \left(\frac{f_A(y)}{f_D(y)} \right) \leq \beta \qquad 3.$$

where $f_g(y) = \frac{\#(Y=y|g)}{n_g}$, y is the category (in the binary case, class $+1/-1$), and subscript g is the group (in the binary case $g = \{A, D\}$). (The natural log may also be used.) The analysis of the example data gives: $f_A(Y = +1) = 0.45$, $f_A(Y = -1) = 0.55$, $f_D(Y = +1) = 0.571$, $f_D(Y = -1) = 0.429$. Then, $DI = 0.45 \cdot \log_2(0.45/0.571) + 0.55 \cdot \log_2(0.55/0.429) = 0.043$. Other approaches mentioned in Das et al. (2021) are Jensen-Shannon divergence, L_p-norm differences, total variation distance, and Kolmogorov-Smirnov distance. These DI metrics generalize CI.

These metrics may be viewed in relation to "base rates" (the percentage of the population in each group, or the percentage within group that have a given characteristic). Next we analyze fairness in the decisions made by trained algorithms.

3.3. Classifier Fairness

Classifier fairness may be broken down into metrics for group fairness and individual fairness. We begin with group fairness.

1. *Disparate Impact (DImp)*: Any discussion of group fairness perforce begins with the concepts of "disparate treatment" (intentional discrimination) and "disparate impact" (unintentional discrimination). Before statistical models, humans would process loan applications, apply some rules, and eventually make the decision on whether or not to grant a loan. Humans would often, consciously or unconsciously, base their decision on protected characteristics such as ethnicity, gender, age, etc. When such decisions are conscious, they would be examples of disparate treatment, and when unconscious, they represent disparate impact, which is defined as follows:

$$DImp = \frac{Pr[\hat{Y} = 1|D]}{Pr[\hat{Y} = 1|A]} \geq 1 - \beta \qquad 4.$$

Computing this metric using the data above, we get $DImp = (65/175)/(100/200) = 0.743$. If $\beta = 0.20$, then fairness may not be assumed under the metric. We often set

$\beta = 20\%$ because of the commonly used 80% rule established by various committees dealing with labor practices.[7] Note that $DImp$ does not factor in whether or not one group is more deserving than the other. It is purely based on predicted labels from the model. Ensuring that $DImp$ is fair even when groups are differentially deserving deprecates merit and is often associated with affirmative action.
Disparate impact is presented as a ratio. It can also be computed as a difference, which is commonly known as "*demographic parity*", i.e., $Pr[\hat{Y} = 1|A] - Pr[\hat{Y} = 1|D] = 0.129$.

2. *Equal Opportunity (EOpp)*: A deficiency of Disparate Impact is that it does not take into account the true labels, only the predicted labels. *EOpp* accounts for both by requiring that the two groups do not differ too much in their true positive rate (TPR).

$$\text{TPR difference: } |Pr[\hat{Y} = +1|D, Y = +1] - Pr[\hat{Y} = +1|A, Y = +1]| \leq \beta \quad 5.$$

Computing this, we get a value of $|60/100 - 80/90| = 0.289$. This would be greater than β, suggesting a lack of equal opportunity.

3. *Equalized Odds (EOdds)*: This metric extends *EOpp* by also requiring equalization of the false positive rate (FPR) across groups. This poses the additional condition:

$$\text{FPR difference: } |Pr[\hat{Y} = +1|D, Y = -1] - Pr[\hat{Y} = +1|A, Y = -1]| \leq \beta \quad 6.$$

Computing this, we get a value of $|5/75 - 20/110| = 0.115$. This is less than β, but since the TPR difference is greater than β, *EOdds* is also violated.

4. *Accuracy Difference (AD)*: Differences in accuracy reflect unfairness for the group that has less accurate predictions than the other. AD is a simple metric that quantifies this potential bias.

$$AD = Pr[\hat{Y} = Y|A] - Pr[\hat{Y} = Y|D] \leq \beta \quad 7.$$

In our example, $AD = 0.850 - 0.743 = 0.107$. We may also examine the difference in F1 scores, which in this case would be $0.842 - 0.727 = 0.115$.

5. *Treatment Equality (TEq)*: This metric assesses whether the kinds of error made by the algorithm are similar across groups, i.e., are the ratios of false negatives to false positives the same?

$$TEq = \left| \frac{\#(\hat{Y} = -1|A, Y = +1)}{\#(\hat{Y} = +1|A, Y = -1)} - \frac{\#(\hat{Y} = -1|D, Y = +1)}{\#(\hat{Y} = +1|D, Y = -1)} \right| \leq \beta \quad 8.$$

For the example, this works out to $|10/20 - 40/5| = 7.5$, which is very large, evidencing unfairness on this metric.

6. *Predictive Parity (PPP, NPP)*: This is a metric that assesses how true outcomes track predicted ones. Positive predictive parity is defined as:

$$PPP = |Pr[Y = +1|A, \hat{Y} = +1] - Pr[Y = +1|D, \hat{Y} = +1]| \leq \beta \quad 9.$$

It may suggest deliberate bias. For example, here $PPP = |80/100 - 60/65| = -0.123$, a value that favors group D. Here, the loan officer may be deliberately giving more

[7] `https://en.wikipedia.org/wiki/Disparate_impact`

loans to Group D borrowers in a form of affirmative action, which would be bias against Group A. A similar concept of negative predictive parity applies:

$$NPP = |Pr[Y = -1|A, \hat{Y} = -1] - Pr[Y = -1|D, \hat{Y} = -1]| \leq \beta \qquad 10.$$

And this computes to $NPP = |90/100 - 70/110| = 0.264$, suggesting lack of parity across groups.

Next, we present some metrics of individual fairness. These metrics have the following properties: (a) They are conditional, i.e., depend on the feature set X. (b) They are "local", i.e., by focusing on nearest neighbors, they are assessed on a partial view of the entire dataset. (c) While computed for individual observations, they may be aggregated to obtain metrics for group fairness if needed.

1. *Individual Fairness (IFair)*: This is based on the simple idea that similar individuals should receive similar outcomes.

$$IFair = |Pr[\hat{Y}_i = y|X_i] - Pr[\hat{Y}_j = y|X_j]| \leq \beta, \quad \text{if } d(X_i, X_j) \leq \epsilon \qquad 11.$$

Here X_i, X_j are the feature vectors of two individuals and $d(X_i, X_j)$ is a distance metric between the vectors. If two individuals are within ϵ of each other under the distance metric, then we would like them to receive the same decision y, i.e., in our case, loan approval or rejection. We note that the group indicator $g = \{A, D\}$ is not part of the feature set X. We may also narrow this definition to check if individual fairness is maintained within group, i.e.,

$$IGFair = |Pr[\hat{Y}_i = y|g, X_i] - Pr[\hat{Y}_j = y|g, X_j]| \leq \beta, \quad \text{if } d(X_i, X_j) \leq \epsilon \qquad 12.$$

2. *Counterfactual Fairness (CFair)*: An extension of $IFair$ leads to counterfactual fairness. We examine if, for cases where the disadvantaged individual has been rejected for a loan, whether the nearest neighbors in the advantaged group were also rejected for their loans.

$$CFair = |Pr[\hat{Y}_i = +1|A, X_i] - Pr[\hat{Y}_j = -1|D, X_j]| \leq \beta, \quad \text{if } d(X_i, X_j) \leq \epsilon \qquad 13.$$

This assesses whether flipping the group variable changes the decision (see Black et al. 2020). For example, if a member of Group D is denied a loan, we may check if the nearest neighbors in the Group A have been approved. Counterfactual fairness comes close to a causal imputation of bias.

We note here that the individual fairness metrics may also be computed for the true labels (Y), not just predicted outcomes ($\hat{Y}$). This would be symptomatic of the actual individual bias versus model bias.

4. APPLICATION TO THE U.S. MORTGAGE MARKET

4.1. Algorithmic Credit Scoring

Automated credit-scoring systems, which evaluate applications for credit, identify prospective borrowers, and manage existing credit accounts, represent some of the most successful applications of risk modeling in finance and banking. The widespread adoption of automated systems and the development of statistical credit scoring models have served as the

foundation to the phenomenal growth in consumer credit since the mid 1960s. Automated underwriting (AU) technology has provided accurate and readily scalable risk assessment tools, without which lenders of consumer credit could not have successfully reduced loan defaults, underwriting costs, and loan pricing; expanded access to credit; and leveled the competitive playing-field for small, medium-sized, and large lenders (see Thomas 2009, Avery et al. 2009, Abduo & Pointon 2011, FinRegLab 2021). For secondary-market securitizers, who purchase loans, AU systems have greatly reduced the extent of principal-agent-based adverse selection that has long been a risk of aggregating loans from numerous counter-parties (see Straka 2000).

In the U.S., early concepts of automated credit scoring systems were built on statistical methods of classification (see Fisher 1936, Durand 1941) and were introduced for commercial applications by Henry Wells of the Spiegel mail order catalogue in the 1950s. The creation of the Fair Isaac corporation (now FICO, Inc) in 1956 led to the first standardized credit scoring product, the FICO® Score.[8] Through the 1970s, automated credit scoring systems were primarily used for credit card and auto lending, however, they remained largely absent in mortgage lending until the mid 1990s (see Lewis 1992, Thomas 2009).

In 1995, Freddie Mac and Fannie Mae endorsed the use of FICO® Scores for evaluating the credit quality of residential mortgages through their proprietary automated underwriting systems, Loan Prospector® (LP) and Desktop Underwriter® (DU), and required all U.S. lenders to provide FICO® Scores for each mortgage delivered to them.[9] The importance of FICO® Scores in the mortgage market continues to this day, where as shown in the Q1:2022 10-K SEC filing for FICO Inc., its scoring products segment comprised 89% of FICO Inc's total revenue and the U.S. mortgage market segment was a "significant portion of our revenues." The market forces that continue to drive the rapid adoption of AU technology in consumer credit markets, especially the mortgage market, include competition and narrowing profit margins; the increased accuracy and efficiency of AU systems in assessing performance risk; the need for more efficient staffing management over credit cycles; and competitive and regulatory pressures to explore new markets, particularly for more affordable (often higher-risk) loan products, as a means to expand market share and enhance growth (see Avery et al. 1996, Straka 2000).

For most AU scoring systems, such as LP, DU, and FICO® Score, the statistical weights assigned to specific features and the functional forms of the predictive equations used to determine the scores are proprietary. As a consequence, these scoring systems have a "black-box" aspect to them. Nonetheless, most scoring systems share a number of elements. For example, most credit history scoring systems consider records of bankruptcy, current and historic ninety-day delinquencies, and the number of credit lines. Most mortgage scoring systems additionally consider factors such as the loan-to-value ratio, the FICO® Score, the ratio of debt payment to income, and measures of employment stability, among others. However, the risk weights assigned to these factors vary from system to system. More recently, FICO Inc. and Freddie Mac have somewhat increased visibility into the scoring features of their systems. For example, FICO® Score now reports the five primary components and associated weights of their scoring system as: loan repayment history (35%);

[8] `https://www.fico.com/25years/`

[9] See `https://www.alta.org/news/news.cfm?20020611-Freddie-Macs-Loan-Prospector-Celebrates-20-Million-Loans` for a discussion of Freddie Mac's LP mortgage underwriting system and McDonald et al. (1997) for a discussion of Fannie Mae's DU.

amounts owed (30%); length of credit history (15%); new credit accounts 10%; and types of credit used (10%).[10] Freddie Mac has also recently introduced somewhat more transparency through Loan Product Advisor®.[11]

Since the financial crises of 2007–2008, regulatory supervisors have reoriented to "data-driven" regulation, a prominent example of which is the collection and analysis of detailed contractual terms for the bank loan and trading book stress-testing protocols (Flood et al. 2016). These regulatory data requirements, the additional exponential growth and availability of consumer and loan performance microdata, and the massive increases in computer processing power have accelerated the ability of all lenders to analyze and quantify risk. The current era of automated underwriting 2.0 increasingly relies on advanced artificial intelligence (AI) technology that electronically undertakes the decision making process for scoring and managing credit applications. AI applications are now an overarching concept that span techniques and algorithms for realizing machine-based intelligence through Machine Learning (ML), and the specific subset of Deep Learning (DL), applied to big data. ML algorithms enable computer algorithms to consume large datasets, learn from them, and then make accurate predictions by compiling and analyzing new datasets. DL, meanwhile, is a particular branch of ML that uses large artificial neural networks to recognize more complex patterns in big data.

Evidence for these trends in the U.S. mortgage markets include empirical studies finding that fintech lenders appear to price risk differently (see Fuster et al. 2022, Buchak et al. 2018) and AI focused fintech lenders have more efficient application processes (see Fuster et al. 2019). Recent survey evidence indicates that industry participants believe that more sophisticated models, including the use of AI, will play an ever-increasing role in lending, including the evaluation of creditworthiness.[12]

Advanced algorithmic techniques, such as deep neural nets and tree-based models provide alternatives to conventional statistical techniques, such as discriminant analysis, probit/logit analysis and logistic regression. The point of using sophisticated techniques, such as these, is their capacity for modeling extremely complex functions, and of course, this stands in contrast to traditional linear techniques, such as linear regression and linear discriminant analysis. As noted by Dixon et al. (2020), a pivotal concern with machine learning, and especially deep learning, is the propensity for over-fitting given the number of parameters in these models and why skill is needed to fit them. In frequentist statistics, over-fitting is addressed by penalizing the likelihood function with a penalty term. A common approach is to select models based on Akaike's information criteria (Akaike 1973), which assumes that the model error is Gaussian. Machine learning methods such as least absolute shrinkage, selection operators, and ridge regression more directly optimize a loss function with a penalty term. Dixon et al. (2020) also point out that overfitting and the lack of regularization techniques largely explains why credit scoring applications, first introduced in the 1990s using neural networks, fell out of favor in the finance industry due to poor performance (for mortgage examples, see Bansal et al. 1993, Burrell & Folarin 1997, Waller & Aiken 1998, Episcopos et al. 1998, Feldman & Gross 2005).

[10] `https://www.myfico.com/credit-education/whats-in-your-credit-score`

[11] `https://sf.freddiemac.com/tools-learning/loan-advisor/our-solutions/loan-product-advisor`

[12] See, for example, ForbesInsights, "Key Takeaways on the Rise of AI in the Mortgage Industry," 2020, `https://forbesinfo.forbes.com/the-rise-of-AI-in-the-Mortgage-Industry`.

4.2. Algorithmic Fairness

A long literature has looked for evidence of discrimination between minority and non-minority borrowers in loan approval rates, interest rates, other loan costs, and service levels in U.S. mortgage markets, the largest consumer-loan market in the United States.[13]

More recently, researchers have looked to see whether the rise of Fintech lenders has made a difference to the level or type of discrimination. Using Home Mortgage Disclosure Act (HMDA) data merged with data from ATTOM Data Solutions, Black Knight McDash, and Equifax, Bartlett et al. (2022b) find that the level of pricing discrimination exhibited by Fintech lenders is somewhat lower than for all lenders, but is still significantly different from zero. Bhutta & Hizmo (2021) find no significant differences between interest rates paid by minority and non-minority borrowers, once the level of discount points is adjusted for. However, using a larger sample, Bartlett et al. (2022b) find that the differences remain significant, even after conditioning on discount points. Bhutta et al. (2022) look at recent HMDA data, which contains recommendations from Automated Underwriting Systems (AUS). They find that conditioning on AUS recommendations, lenders deny minority applicants more often than non-minority applicants. They also find that the AUS recommend higher denial rates for minority applicants. Fuster et al. (2022) study mortgage data from 2009–2019 and find that machine-learning techniques to evaluate credit quality may result in differential impact on loan provision to minority versus non-minority borrowers. Blattner & Nelson (2021) find that the credit reports of minority loan applicants are on average less informative than those of non-minority applicants, and argue that this may cause credit-scoring models to produce unequal outcomes for minority and non-minority applicants. Berg et al. (2020) analyze the information content of the digital footprint — information that people leave online simply by accessing or registering on a website — for predicting consumer default for furniture purchases from an E-commerce company in Germany. They find that the digital footprints match the information content of credit bureau scores suggesting that lender could use a combination of credit bureau and digital footprint data to make superior lending decisions compared to lenders that only access one of the two sources of information. Tantri (2021) uses data from an Indian bank and loan-officer application decision rates and finds that machine learning (ML) algorithms improved lending efficiency without compromising loan quality in a credit environment where soft information dominates. D'Acunto et al. (2022) exploit a leading FinTech peer-to-peer lending platform in India paired with an automated robo-advising lending tool and find that stereotypical discrimination is pervasive and leads to 32% higher default rates and about 11% lower returns on the loans issued to borrowers of the favored demographic. Dobbie et al. (2021) test for bias by a high-cost consumer lender in the U.K. using a quasi-random assignment of loan examiners to identify the profitability of marginal applicants. They find significant bias against immigrant and older applicants when using the firm's preferred measure of long-run profits. They calculate that the lender would have earned approximately £157 more in profit per applicant if they had used the machine learning algorithm rather than the loan examiner decisions.

Regardless of whether we are talking about human or algorithmic bias, one root cause

[13]See, for example, Black et al. (1978), Munnell et al. (1996), Courchane & Nickerson (1997), Black et al. (2003), Ghent et al. (2014), Bayer et al. (2018), Reid et al. (2017), Ambrose et al. (2021), Cheng et al. (2015), Begley & Purnanandam (2021), Hanson et al. (2016). Related studies in other consumer-debt markets include Dobbie et al. (2021) and Butler et al. (2023).

is the data available to mortgage researchers. In particular, sample-selection bias has long plagued the literature because the widely used HMDA data includes loan-level information on race and ethnicity but includes no information on the actual performance of accepted and rejected mortgages (see Munnell et al. 1996, Berkovec et al. 1996, Ladd 1998, Lee & Floridi 2021). More recently, the literature has applied statistical merges of the HMDA data with other loan-level information on the performance of accepted loans (see Ghent et al. 2014, Bhutta & Hizmo 2021, Bartlett et al. 2022b, Bhutta et al. 2022, Giacoletti et al. 2022) to at least partially address potential biases associated with false positive mortgage decisions. Blattner & Nelson (2021) quantify the importance of precision differences in borrower credit histories using TransUnion credit data and a structural model to solve for counterfactual mortgage approval decisions under alternative information structures. The only quantitative information on the association between false negative loan decisions and race and ethnicity are small samples studies of the differential treatment for racially paired "identical" loan applicants (see Galster 1996, Ross & Yinger 2002).

4.3. Legal and Regulatory Oversight

For this review article, we define U.S. fair-lending law as including the Fair Housing Act[14] and the Equal Credit Opportunity Act (ECOA)[15] together with all implementing regulations and judicial interpretations relating to them. ECOA prohibits discrimination in any aspect of a credit transaction and prohibits the consideration of nine borrower characteristics including race, color, religion, national origin, sex, marital status, age, public assistance program status, or prior exercise of rights under the Consumer Protection Act. The Fair Housing Act prohibits discrimination in all residential real-estate-related transactions and prohibits the use of seven characteristics including race, color, national origin, religion, sex (including gender identity and sexual orientation), family status, and disability. Under these laws, the courts have ruled that lenders may use proxy variables, even if they lead to worse outcomes for minorities, as long as the lender can show that these variables have a legitimate business necessity. While lenders might view many activities as being necessary for profit maximization, the courts have consistently limited the legitimate-business-necessity defense to the use of borrower characteristics and scoring practices to ascertain creditworthiness only (Bartlett et al. 2022a,b).

The passage of the Dodd-Frank Wall Street Reform and Consumer Protection Act (2010) led to significant changes in federal prudential regulation of the development, oversight, and use of all models by U.S. regulated financial services firms.[16] Although the Federal Reserve Board, the Office of the Comptroller of the Currency, and the Federal Deposit Insurance Corporation each established their own supervisory and regulatory guidelines,[17]

[14]U.S. Department of Housing and Urban Development regulations (24 C.F.R. §100).

[15]Regulation B (12 C.F.R. §202).

[16]Models are defined in SR 11-7 as "a quantitative method, system, or approach that applies statistical, economic, financial, or mathematical theories, techniques, and assumptions to process input data into quantitative estimates," see page 2, `https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm`.

[17]See Board of Governors of the Federal Reserve System, Supervisory & Regulation Letter (SR11-7), April 4, 2011, Guidance on Model Risk Management, April 4, 2011, `https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm`, Office of the Comptroller of the Currency, *Bulletin* 2011-12, Sound Practices for Model Risk Management, Supervisory Guidance on Model Risk Management, April 4, 2011, `https://www.occ.gov/news-issuances/bulletins/2011/`

the overall regulatory structure under FRB, SR 11-7 regulations requires models to include "a clear statement of purpose to ensure the model is developed in line with its intended use; sound design, theory, and logic underlying the model; robust model methodologies and processing components; rigorous assessment of data quality and relevance; and appropriate documentation."[18] For financial institutions with important business segments in retail or consumer lending, model risk management under SR 11-7 requires the application of validation methods to assure expected model performance and to account for potential model inaccuracies. Adverse notices for credit scoring are regulated by the Equal Credit Opportunity Act and the Fair Credit Reporting Act.[19] In 2011, post Dodd-Frank, an amendment to the FCRA took effect that requires adverse notices if it is found that credit terms for some borrowers were "materially less favorable" than the terms granted to a "substantial proportion" of other consumers (see 15 U.S.C. § 1681m(h); 12 C.F.R. §§ 222.70-75).

4.4. Transparency and Interpretability

Standard logistic regression, or significantly simplified and linearized machine learning models, remain the gold standard methodology in the credit scoring industry, despite the general superiority of modern machine learning techniques (see Szepannek 2017, Molinar 2022, Sudjianto et al. 2020, FinRegLab 2021). The transparency and frequent monitoring requirements of the SR 11-7 regulations and the Dodd-Frank Act Stress Testing (DFAST) methodologies have led to strong managerial preferences for the use of easily interpreted linear models in regulated financial institutions.

Unlike logistic regression models, where feature importance is a transparent byproduct, the use of black-box deep learning models for lending decisions raises important issues about explaining model decisions, especially when rejecting a loan application. Under the Equal Credit Opportunity Act (ECOA), lenders are required to disclose the reasons for why a loan application was denied. This requires an understanding of the key variables that influenced a black-box model in the lending decision and an assessment of feature importance at the *instance* (observation) level. Determining which variables are important for the model's decisions over the entire dataset is known as *global* feature importance. The key paper by Lundberg & Lee (2017) argues for feature importance evaluations at the instance level by implementing a Shapley (1952) values analysis, see the open-source SHAP library.[20] This approach determines the contribution of each feature to the model's decision by assessing prediction over all "coalitions" (subsets of the features). It has three important properties, i.e., (i) Dummy feature (if a feature never adds any marginal explanation, its Shapley value is zero). (ii) Substitutability (if two features always add the same marginal value to any subset to which they are added, their importance should be the same). (iii) Additivity (the payoff of a feature in two subsets of features should be additive to the sum of the payoffs in the combined set). This last property enables aggregation of instance-level

---

`bulletin-2011-12.html`, Department of the Treasury (2014), Federal Deposit Insurance Corporation, Financial Institution Letter 22-2017, Supervisory Guidance Model Risk Management, June 7, 2017 `https://www.fdic.gov/news/financial-institution-letters/2017/fil17022a.pdf`.

[18] See page 3, `https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm`.

[19] 15 U.S.C. §§ 1681-1681x, `https://www.ftc.gov/legal-library/browse/statutes/fair-credit-reporting-act`.

[20] `https://github.com/slundberg/shap`

explanations into global explanations. Shapley value is the only attribution method that satisfies these properties. It builds upon earlier ideas for local explanations such as LIME (Local Interpretable Model-Agnostic Explanations), Ribeiro et al. (2016). Both LIME and SHAP offer model-agnostic methods, i.e., they do not need access to the model itself to obtain instance-level explanations. However, approaches such as SHAP are computationally expensive.

Model-dependent approaches are also widely used, because access to the model enables faster and more accurate assessment of feature importance, but require revealing the model to the explainer algorithm. Common techniques here are the SHAP tree explainer[21], saliency maps (Adebayo et al. 2018), integrated gradients (Sundararajan et al. 2017), and testing with concept activation vectors (TCAV) (Kim et al. 2018). Many surveys detail the pros and cons of various explanation approaches, such as (Guidotti et al. 2018, Gilpin et al. 2018, Tjoa & Guan 2020, Burkart & Huber 2021).

4.5. Analysis

We illustrate the bias measures from the previous section with an analysis of mortgage data, using the dataset created in Bartlett et al. (2022b). We start with Home Mortgage Disclosure Act (HMDA) data for over 14 million mortgage applications between 2009 and 2015 that have complete HMDA data for the supervisory status of the lender, origination loan amount, applicant income, conventional loan status, loan rejection status, HMDA loan refinance status, property two-digit FIPS state identifier, three-digit FIPS county identifier, census tract number, a property-type indicator for mortgages on one-to-four family residences, lender name, and respondent identification.[22] Because the 2009–2015 vintages of the HMDA data do not include important underwriting variables such as the loan-level credit score and loan-to-value ratio, we proxy for these variables for each loan using the census-tract-level 25th, 50th, and 75th percentile credit score and LTV, estimated using data from a statistical merge of loan-origination data from ATTOM Data Solutions and Black Knight McDash.[23] Based on the HMDA lender name for each loan we further augment our dataset with a computed measure of each lender's census tract-level Herfindahl index by annual loan origination volume and an indicator variable for whether the lender was identified by *Inside Mortgage Finance* as among the top 25 lenders by loan origination for that year. We then construct a random sub-sample of 500,000 loans from these data. Section S1 in the Supplemental Material describes the variables in more detail and **Table 1** shows summary statistics for the data.

Our empirical illustration is based on a representative random sample of 500,000 mortgage applications. We use these to assess whether the mortgage loan business (across lending entities) has biases that may require addressing. The dataset contains a flag for

[21]Used with models such as XGBoost, this is extremely fast: `https://shap-lrjball.readthedocs.io/en/latest/generated/shap.TreeExplainer.html`

[22]The HMDA data cover 90% of mortgage originations in the U.S. (see Engel & McCoy 2011) and are the only data source with loan-level information on the identified first applicant race and ethnicity.

[23]As discussed in Bartlett et al. (2022b), 77.79% of the candidate residential single family loans in the Black Knight McDash data set (20,022,570 loans) were successfully merged to the single family residential loans in the ATTOM Data Solutions data. For more details on both the data and the merge, see the Internet Appendix to Bartlett et al. (2022b), available at `https://www.jfinec.com/internet-appendices`.

	mean	sd	min	max
25th percentile census tract × year loan-to-value ratio	.6532	.1136	.3	1
50th percentile census tract × year loan-to-value ratio	.7371	.1336	.3	1
75th percentile census-tract × year loan-to-value ratio	.8959	.0811	.3016	1
25th percentile by census tract × year credit score	705.2787	28.5041	630	829
50th percentile by census tract × year credit score	747.3291	36.6128	630	843
75th percentile by census tract × year credit score	781.9119	16.9646	630	837
HMDA rejection indicator	.4698	.4991	0	1
HMDA borrower 1 minority	.1719	.3773	0	1
HMDA borrower 1 female	.3117	.4632	0	1
HMDA origination year	2011.717	1.9771	2009	2015
HMDA origination loan amount ($000)	210.3832	107.7479	40	729
HMDA conventional conforming loan indicator	.6981	.4591	0	1
HMDA refinance loan indicator	.6419	.4794	0	1
Top 25 lender by year	.4665	.4989	0	1
HMDA origination income ($000)	98.3894	102.2259	20	9634
Lender Herfindahl index by census tract	.0535	.0220	.0169	.8233
HMDA supervisory agency indicator	5.8717	2.9371	1	9

Table 1: Summary statistics

whether the loan application has been rejected (dummy=1) or accepted (dummy=0). Note here that the positive outcome is when the dummy=0, i.e., the loan application was successful (dummy=1 is when the loan application is rejected). The proportion of rejected applications in the data is 48%. To examine bias, we construct two separate protected characteristic variables from this data: (1) Gender,[24] and (2) Minority.[25]

4.5.1. Machine-learning models. We run a large number of machine learning models on the dataset, to fit the reject/accept decision to the training data set. The models are implemented using the AutoGluon library from Amazon Web Services (AWS).[26] The library is used to implement a stack ensembled prediction model to a high level of accuracy with an objective of maximizing the F1 score (for binary classification models, this balances precision and recall, while not eschewing accuracy). The optimized F1 score is 0.847 (with accuracy of 0.845) on the hold-out test dataset. **Table 2** reports model performance metrics as well as those from logistic regression (the industry standard), showing that ensembled ML models greatly improve on logistic regression.[27] Model performance, during training, on the validation dataset matches performance on the hold-out test dataset, i.e., there is no overfitting of the model.

4.5.2. Dataset Fairness. We generate three different dataset bias metrics as described in Section 3. There is clear evidence of group imbalance in the dataset, both on gender and minority status of the first mortgage applicant. The group imbalance metric is 0.45

[24] We create a dummy variable equal to 1 if female and 0 if male.

[25] We create a dummy variable equal to 1 if the borrower is of Black/Hispanic origin, 0 otherwise.

[26] See `https://auto.gluon.ai/stable/index.html`.

[27] For details on the models, see `https://auto.gluon.ai/dev/api/autogluon.tabular.models.html`.

Model#	Model	F1 score (test data)	F1 score (validation data)
0	WeightedEnsemble	0.846569	0.868559
1	CatBoost	0.846230	0.855942
2	LightGBMXT	0.835934	0.845719
3	LightGBMLarge	0.832945	0.851866
4	NeuralNetFastAI	0.830090	0.848218
5	XGBoost	0.829700	0.848974
6	LightGBM	0.825924	0.849737
7	RandomForestEntr	0.825821	0.834900
8	NeuralNetTorch	0.825794	0.847359
9	RandomForestGini	0.825186	0.828184
10	ExtraTreesEntr	0.808492	0.809487
11	ExtraTreesGini	0.807285	0.809672
12	KNeighborsUnif	0.702274	0.723244
13	KNeighborsDist	0.701865	0.724540

(a) **F1 scores**. The F1 score is the harmonic mean of precision and recall for rejected loans.

Metric	ML Model	Logistic Regression
F1	0.847	0.654
Accuracy	0.845	0.663
Balanced accuracy	0.850	0.664
MCC	0.702	0.328
ROC/AUC	0.926	0.717
Precision	0.779	0.622
Recall	0.927	0.689

(b) **Ensembled model versus logistic regression.** Balanced recall is the average of recall for both type 0 and 1 labels. The Matthews Correlation Coefficient (mcc) is a balanced metric from the confusion matrix (`https://scikit-learn.org/stable/modules/generated/sklearn.metrics.matthews_corrcoef.html`).

Table 2: **Model results.** Panel (a) shows the F1 scores for both test and validation datasets from running several different ML models. We use 100K observations for training (balanced 50/50 for accept/reject labels) and validation and 400K for testing (imbalanced as per the original data). Panel (b) compares the preferred ensembled model with logistic regression. The best model is delivered by the weighted stack ensemble.

for gender, i.e., for every 100 men there are only 45 women in the dataset. This can lead to poorer estimation and model prediction for female applicants. The imbalance is worse for minorities who comprise only 1 of 5.8 applicants. Given the large number of applicants in mortgage datasets, it is not difficult to handle these imbalances by designing the training dataset (usually a smaller subset of the original data) to have balanced numbers of applicants, or balanced numbers of cases with accept/reject decisions.

	Importance	SD	p-value
50th percentile of the census tract × year loan-to-value ratio	0.178057	0.007499	0.000000
50th percentile of the census tract × year credit score	0.149286	0.003512	0.000000
HMDA origination year	0.023086	0.001935	0.000006
HMDA lender name	0.021093	0.005877	0.000654
25th percentile of the census-tract × year loan-to-value ratio	0.018860	0.004643	0.000407
HMDA conventional conforming loan indicator	0.012190	0.002891	0.000352
25th percentile of the census-tract × year credit score	0.009984	0.003147	0.001042
HMDA loan amount	0.008576	0.003550	0.002843
75th percentile of the census-tract × year credit score	0.006459	0.002713	0.002994
HMDA refinance loan indicator	0.006368	0.002574	0.002610
HMDA origination income	0.005661	0.002120	0.001977
75th percentile of the census-tract × year loan-to-value ratio	0.004405	0.001330	0.000888
HMDA supervisory agency indicator	0.001820	0.001641	0.034124
Top 25 lender by year indicator	0.000624	0.001741	0.233750
Lender Herfindahl Index by census tract	−0.000282	0.001061	0.707712

Table 3: **Feature importance.** The feature importance of each variable is based on column permutation analysis: when the column for each feature is shuffled, important features experience a bigger drop in the target metric, F1 score in this case (`https://auto.gluon.ai/stable/tutorials/tabular_prediction/tabular-indepth.html#interpretability-feature-importance`).

Class imbalance measures the difference in the ratio of accepts to rejects between the advantaged and disadvantaged groups. For gender, we find a disadvantage of 0.017 ($\sim$ 1.7%) for women, and a disadvantage of 0.42 ($\sim$ 42%) for minorities. Depending on societal fairness norms, these may be within reasonable limits. The results also suggest greater minority imbalance in accepts vs. rejects than for gender.

Using the distributional imbalance metric (i.e., Kullback-Leibler divergence[28]), the metrics for imbalance are weaker, about zero for gender and 0.032 for minorities. Again, minority bias is greater than gender bias as was noticed in the case of class imbalance.

4.5.3. Model Fairness (Groups). Whereas dataset fairness assesses bias in the historical record, model fairness examines biases in the models used for classifying mortgage applicants for reject/accept decisions. As discussed in Section 3, we compute six metrics to assess model fairness across groups of applicants. These are summarized in **Table 4**. We see that despite imbalances in the dataset, noted in the previous section, model fairness across groups is encouraging. The only metric that appears to suggest imbalance is Treatment Equality, where there is a difference between the ratios of false negatives to false positives, only in the case of minorities. This is outside of the bounds of $\beta = 0.20$ suggested by the 80% rule, though it is not usually applied to this metric.[29] Further, we note that the imbalance is in favor of the minority group, as the model delivers 2.73 false negatives for each false positive for minorities, whereas there are 3.83 false negatives for every false positive for non-minorities.

There are two deficiencies of these group fairness measures. *One*, these measures are

[28]This is a relative entropy measure (`https://en.wikipedia.org/wiki/Kullback-Leibler_divergence`).

[29]The rule has also been applied in labor markets to assess bias in hiring (see `https://en.wikipedia.org/wiki/Disparate_impact`).

Metric	Gender	Assessment	Minority	Assessment
Disparate Impact	0.996	OK	0.817	OK
Equal Opportunity	0.006	OK	0.020	OK
Equalized Odds	0.004	OK	0.000	OK
Accuracy Difference	−0.005	OK	−0.007	OK
Treatment Equality	0.046	OK	1.100	Imbalance
Predictive Parity:				
PPP	0.003	OK	0.008	OK
NPP	0.036	OK	0.056	OK

Table 4: **Model fairness metrics.** We report the values computed for the six metrics from Section 3. When the metrics are within the tolerance level based on $\beta = 0.20$, we signify this by an OK assessment. Only disparate impact for the protected characteristic gender violates the defined tolerance.

unconditional, as they are not based on the features of applicants other than protected-characteristic status. A disparate-impact score of less than 80% does not take into account whether the applicants in the disadvantaged group are less qualified for a mortgage than applicants in the advantaged group.

Two, even if there is no bias across groups demarcated on protected characteristics such as gender and ethnicity, individuals within each group may be discriminated against. To assess the data and models in a conditional manner, we consider using measures of individual fairness in addition to the measures of group fairness. This is the subject of the next section.

4.5.4. Model Fairness (Individual). These metrics examine whether an individual applicant has been treated fairly in comparison to other applicants. Even though group fairness may be maintained, specific individuals within a protected characteristic group may have received unfair treatment. In order to assess individual fairness, we condition on the features of an individual applicant. Our implementation of conditional analysis is undertaken in a simple manner that is easy to implement, and is also non-parametric. We examine the nearest neighbors of an applicant and determine whether the applicant in consideration obtained a different decision from the model than its nearest neighbors. This conditions on the feature set X in the same manner as undertaken with matched samples. For conditioning, we examine the 5 nearest neighbors for every applicant. While somewhat expensive computationally, the approach is general and effective.

Our analysis focuses on the 100K observations in the training dataset. After training the model to a high level of accuracy we assess how often the model gives different predictions for an applicant relative to its neighbors. Because the predictions from the model have 81% correlation with the true labels in the data, this is also an assessment of actual individual fairness decisions in the dataset made by the lending institutions. We note that the correlation of the loan rejection dummy with the gender dummy is 0.15% and with the minority dummy is 7.53%.

We compute the Euclidean distance of all 100K applicants with every other applicant, i.e., 4,999,950,000 distance pairs, computed on the feature set after min-max scaling for normalizing the features (this handles categorical data in a manner comparable to continu-

ous numerical data). For every applicant we store an ordered list of nearest neighbors from which we can always extract any number of closest applicants. We conduct three sets of individual fairness analyses.

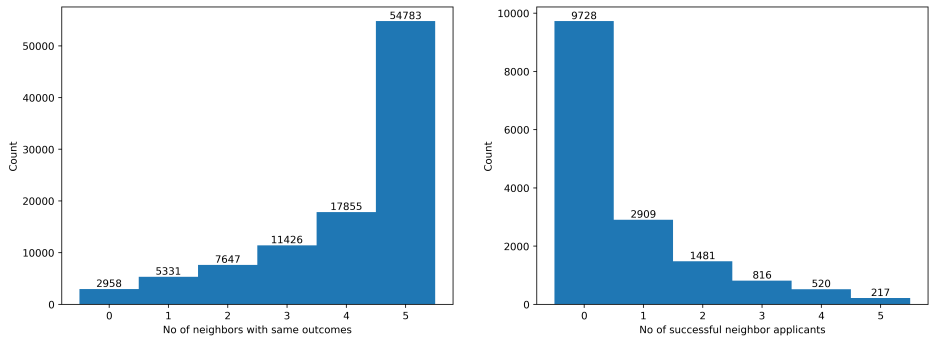
1. *Differential Treatment.* Our first assessment is whether the model generates similar decisions for similar applicants, irrespective of protected characteristic. For each applicant we count how many of the five nearest neighbors received the same decision. The histogram for the distribution of the number of similar decisions is shown in **Figure 2**, Panel (a). We see that a vast majority of individual decisions are consistent with those of the nearest neighbors. However, about 8% of applicants receive decisions that differ from all or all but one of the neighbors. Such cases would require further examination in case unfair decisions were made by the model.
2. *Gender bias.* In **Figure 2**, Panel (b), we examined nearest male neighbors for all rows in which female applicants were rejected. We report how many nearest neighbors obtain accept decisions. In 62% of these cases, all similar men also had their loan applications rejected. On the other hand, in 1.4% of the cases all nearest male applicants received approvals. Whereas there is evidence that by and large, rejected female applicants received the same decision as the majority of their nearest neighbors, there are a few cases where the majority of the closest male neighbors were given favorable outcomes. This approach enables flagging such cases for additional (possibly manual) review.
3. *Minority bias.* This is depicted in Panel (c) in **Figure 2**. We examined the nearest non-minority neighbors for all rows in which minority applicants were rejected. We report how many nearest neighbors obtain accept decisions. In 63% of these cases, all non-minorities also had their loan applications rejected. On the other hand, in just 1.4% of the cases all nearest non-minority applicants received approvals. These cases may be flagged for additional (possibly manual) review. Individual fairness appears to be equally high across both gender and minority characteristics.

Overall, we note that the sample mortgage dataset and model predictions do not seem to be riddled with gender or minority bias. Individual fairness metrics enable an assessment of bias that is conditional and local, supporting manual examination of possibly unfair cases. Understanding why local neighbors of rejected applicants have had their loan applications accepted may also support actionable explanations to unsuccessful loan applicants. This case study suggests that an overall assessment of algorithmic fairness is possible using simple, understandable metrics, and is easy to implement, predicating optimism for standardized regulation. The literature is replete with fairness metrics, some that are more complex, but the simple ones computed here provide a holistic view of the overall gestalt of fairness in data and models.

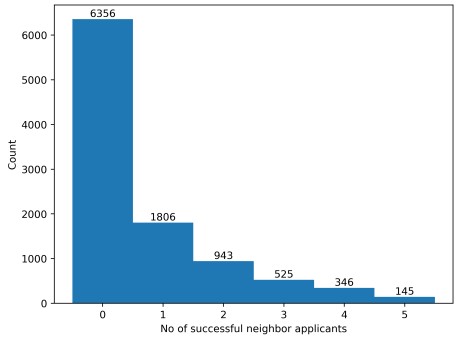
5. SUMMARY POINTS AND FUTURE ISSUES

SUMMARY POINTS

1. **Growth in algorithmic decision-making.** The last few years have seen significant growth in algorithmic decision-making in finance, healthcare, etc., driven in



(a) No. neighbors with same outcome (b) Successful male neighbors / rejected female



(c) Successful non-minority neighbors / rejected minority

Figure 2: **Individual fairness across the sample.** In Panel (a), for each applicant we count how many of the five nearest neighbors received the same decision. In Panel (b), we examined nearest male neighbors for all rows in which female applicants were rejected. We report how many nearest neighbors obtain accept decisions. In Panel (c), we examined nearest non-minority neighbors for all rows in which minority applicants were rejected. We report how many nearest neighbors obtain accept decisions. The total numbers of observations in panels (a), (b), and (c) are 100,000, 15,671, and 10,121, respectively.

part due to the availability of new data on individuals and to the development of new AI/ML tools.

2. **Bias remains.** Although algorithmic decision-making eliminates face-to-face biases (by eliminating faces), several high-profile recent examples suggest that they may still lead to biased decisions. Bias may be introduced at various stages in the system: pre-training, in-processing, and post-training.
3. **Metrics.** A range of measures have been developed to assess fairness in both data and models, at both the group and individual level.

4. **Imbalanced mortgage data.** We find evidence of group imbalance in the HMDA mortgage data for both gender and (especially) minority status, which can lead to poorer estimation/prediction for female/minority applicants.
5. **Model fairness.** Model fairness is closer across both groups and individuals, though we find that some local male (non-minority) neighbors of similar, rejected female (minority) applicants were granted loans, something that warrants further study.
6. **ML outperforms logistic regression.** Modern machine-learning techniques substantially outperform logistic regression (the industry standard), though at the cost of being substantially harder to explain to denied applicants, regulators, or the courts.

FUTURE ISSUES

1. **Additional data modalities**. This article focuses on tabular data only; however, many of the metrics are feature-agnostic and are extendable to other modalities that are increasingly used in financial analyses, such as text. Future work on bias assessment in language models may soon become widespread, with metrics such as accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency suggested in Liang et al. (2022).
2. **Regulatory harmonization for AI/ML use in consumer lending.** The dominance of logistic regression in the underwriting practices of regulated financial institutions, despite increasing evidence of the superiority of more flexible machine-learning techniques in credit scoring (see Szepannek 2017, Fuster et al. 2022, Molinar 2022, FinRegLab 2021), suggests there is a need for further research to develop policies that would achieve greater regulatory harmonization under SR 11-7 for "best practice" AI/ML consumer lending applications as well as developing well-vetted standard metrics for group and individual fairness, explainability, and auditability.
3. **De-biasing mortgage data.** Many, if not all, historical mortgage data sets exhibit significant deficiencies in the quality and availability of adequate credit reporting for ethnic and racial minorities (see Blattner & Nelson 2021), evidence of the long-term negative effects of redlining on property values and access to credit (see Aliprantis et al. 2023, Aaronson et al. 2021, Cutler et al. 1999), and significant sorting and racial disparities in property valuation assessments (see Avenancio-Leon & Howard 2022, McMillen & Singh 2020, Perry et al. 2018). The effect of these imbalances in existing datasets deserves more study as does the identification of new data sources that allow for the measurement of responsible financial behavior and lower credit risks amongst poorly served borrower populations.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We are grateful for financial support from the Fisher Center for Real Estate and Urban Economics at the Haas School of Business.

LITERATURE CITED

Aaronson D, Hartley DA, Mazumder S. 2021. The effects of the 1930s HOLC 'redlining' maps. *American Economic Journal: Economic Policy* 13:355–392

Abduo HA, Pointon J. 2011. Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management* 18:59–88

Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. 2018. Sanity checks for saliency maps, In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 9525–9536, Red Hook, NY, USA: Curran Associates Inc.

Aka O, Burke K, Bauerle A, Greer C, Mitchell M. 2021. Measuring model biases in the absence of ground truth. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery, 327–335

Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, eds. BN Petrov, F Csaki. Budapest: Akademiai Kiado, 267–281

Alesina AF, Lotti F, Mistrulli PE. 2013. Do women pay more for credit? Evidence from Italy. *Journal of the European Economic Association* 11(s1):45–66

Aliprantis D, Carroll DR, Young ER. 2023. What explains neighborhood sorting by income and race? *Journal of Urban Economics* (forthcoming)

Ambrose BW, Conklin JN, Lopez LA. 2021. Does borrower and broker race affect the cost of mortgage credit? *Review of Financial Studies* 34(2):790–826

Arya V, Bellamy RKE, Chen PY, Dhurandhar A, Hind M, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. *arXiv:1909.03012 [cs, stat]*

Avenancio-Leon CF, Howard T. 2022. The assessment gap: Racial inequality in property taxation. *Quarterly Journal of Economics* 137(3):1383–1434

Avery RB, Bostic RW, Calem PS, Canner GB. 1996. Credit risk, credit scoring and the performance of home mortgages. *Federal Reserve Bulletin* July:621–648

Avery RB, Brevoort KP, Canner GB. 2009. Credit scoring and its effects on the availability and affordability of credit. *The Journal of Consumer Affairs* 43(3):516–530

Bansal A, Kauffman RJ, Weitz RR. 1993. Comparing the modeling performance of regression and neural networks as data quality varies: A business value approach. *Journal of Management Information Systems* 10(1):11–32

Barocas S, Hardt M, Narayanan A. 2019. Fairness and machine learning. `fairmlbook.org`

Barocas S, Selbst AD. 2016. Big Data's disparate impact. *California Law Review* 104:671–732

Barocas S, Selbst AD, Raghavan M. 2020. The hidden assumptions behind counterfactual explanations and principal reasons, In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 80–89, Barcelona, Spain: Association for Computing Machinery

Bartlett R, Morse A, Stanton R, Wallace N. 2022a. Algorithmic discrimination and input accountability under the Civil Rights Acts. *Berkeley Technology Law Journal* 36:675–736

Bartlett R, Morse A, Stanton R, Wallace N. 2022b. Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics* 143(1):30–56

Bayer P, Ferreira F, Ross SL. 2018. What drives racial and ethnic differences in high-cost mortgages? The role of high-risk lenders. *Review of Financial Studies* 31(1):175–205

Begley TA, Purnanandam AK. 2021. Color and credit: Race, regulation, and the quality of financial services. *Journal of Financial Economics* 141:48–65

Bellamy RKE, Dey K, Hind M, Hoffman SC, Houde S, et al. 2019. AI fairness 360: An extensible

toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *IBM Journal of Research and Development* 63(4/5):4:1–4:15

Berg T, Burg V, Gombović A, Puri M. 2020. On the rise of fintechs: Credit scoring using digital footprints. *Review of Financial Studies* 33(7):2845–2897

Berkovec JA, Canner GB, Hannan TH, Gabriel SA. 1996. Mortgage discrimination and FHA loan performance. In *Mortgage Lending, Racial Discrimination and Federal Policy*, eds. JJ Choi, B Oskan. Taylor & Francis Group, 29–43

Bhutta N, Hizmo A. 2021. Do minorities pay more for mortgages? *Review of Financial Studies* 34:763–789

Bhutta N, Hizmo A, Ringo D. 2022. How much does racial bias affect mortgage lending? Evidence from human and algorithmic credit decisions. Working paper, Federal Reserve Board

Black E, Yeom S, Fredrikson M. 2020. FlipTest: Fairness testing via optimal transport, In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 111–121, Barcelona, Spain: Association for Computing Machinery

Black H, Schweitzer RL, Mandell L. 1978. Discrimination in mortgage lending. *American Economic Review* 68(2):186–191

Black HA, Boehm TP, DeGennaro RP. 2003. Is there discrimination in mortgage pricing? The case of overages. *Journal of Banking and Finance* 27(6):1139–1165

Blattner L, Nelson S. 2021. How costly is noise? Data and disparities in consumer credit. Working paper, Stanford University

Brock JM, De Haas R. 2021. Discriminatory lending: Evidence from bankers in the lab. Working paper, European Bank for Reconstruction and Development

Buchak G, Matvos G, Piskorski T, Seru A. 2018. Fintech, regulatory arbitrage, and the rise of shadow banks. *Journal of Financial Economics* 130(3):453–483

Budhathoki K, Minorics L, Bloebaum P, Janzing D. 2022. Causal structure-based root cause analysis of outliers, In *Proceedings of the 39th International Conference on Machine Learning*, eds. K Chaudhuri, S Jegelka, L Song, C Szepesvari, G Niu, S Sabato, vol. 162 of *Proceedings of Machine Learning Research*, pp. 2357–2369, PMLR

Burkart N, Huber MF. 2021. A survey on the explainablity of supervised machine learning. *Journal of Artificial Intelligence Research* 70:245–317

Burrell PR, Folarin BO. 1997. The impact of neural networks in finance. *Neural Computing & Applications* 6:193–200

Butler AW, Mayer EJ, Weston JP. 2023. Racial disparities in the auto loan market. *Review of Financial Studies* 36:1–41

Canetti R, Cohen A, Dikkala N, Ramnarayan G, Scheffler S, Smith A. 2019. From soft classifiers to hard decisions: How fair can we be?, In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 309–318, New York, NY, USA: Association for Computing Machinery

Chen D, Li X, Lai F. 2017. Gender discrimination in online peer-to-peer credit lending: Evidence from a lending platform in China. *Electronic Commerce Research* 17(4):553–583

Cheng P, Lin Z, Liu Y. 2015. Racial discrepancy in mortgage interest rates. *Journal of Real Estate Finance and Economics* 51(1):101–120

Courchane M, Nickerson D. 1997. Discrimination resulting from overage practices. *Journal of Financial Services Research* 11:133–151

Cutler DM, Glaeser EL, Vigdor JL. 1999. The rise and decline of the American ghetto. *Journal of Political Economy* 107:455–506

D'Acunto F, Ghosh P, Jain R, Rossi AG. 2022. How costly are cultural biases? Working paper, Georgetown University

Das S, Donini M, Gelman J, Haas K, Hardt M, et al. 2021. Fairness measures for machine learning in finance. *Journal of Financial Data Science* Fall

Dastin J. 2018. Amazon scraps secret AI recruiting tool that showed bias against women.

Reuters Oct. 10 https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Department of the Treasury. 2014. Supervisory guidance on implementing Dodd-Frank Act company-run stress tests for banking organizations with total consolidates assets of more than $10 billion but less than $50 billion. *Federal Register* 79(49)

Diana E, Gill W, Kearns M, Kenthapadi K, Roth A. 2021. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 66–76

Dixon MF, Halperin I, Bilokon P. 2020. Machine learning in finance: From theory to practice. Cham, Switzerland: Springer

Dobbie W, Liberman A, Paravisini D, Pathania V. 2021. Measuring bias in consumer lending. *Review of Economic Studies* 88:2799–2832

Donnan S, Choi A, Levitt H, Cannon C. 2022. Wells Fargo rejected half its Black applicants in mortgage refinancing boom. *Bloomberg* March 10https://www.bloomberg.com/graphics/2022-wells-fargo-black-home-loan-refinancing/

Durand D. 1941. Risk elements in consumer installment financing. Working paper, NBER

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. 214–226

Engel KC, McCoy PA. 2011. The subprime virus: Reckless credit, regulatory failure, and next steps. New York: Oxford University Press

Episcopos A, Pericli A, Hu J. 1998. Commercial mortgage default: A comparison of logit with radial basis function networks. *Journal of Real Estate Finance and Economics* 17(2):163–178

Feldman D, Gross S. 2005. Mortgage default: Classification trees analysis. *Journal of Real Estate Finance and Economics* 30(4):369–396

FinRegLab. 2021. The use of machine learning for credit underwriting: Market and data science context. Technical report, FinRegLab

Fisher RA. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* :179–188

Flitter E. 2022. A Black homeowner is suing Wells Fargo, claiming discrimination. *New York Times* March 21https://www.nytimes.com/2022/03/21/business/wells-fargo-mortgages-discrimination-suit.html

Flood MD, Jagadish HV, Raschid L. 2016. Big data challenges and opportunities in financial stability monitoring. *Financial Stability Review* April:129–142

Fuster A, Goldsmith-Pinkham P, Ramadorai T, Walther A. 2022. Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance* 77(1):5–47

Fuster A, Plosser M, Schnabl P, Vickery J. 2019. The role of technology in mortgage lending. *Review of Financial Studies* 32(5):1854–1899

Galhotra S, Brun Y, Meliou A. 2017. Fairness testing: Testing software for discrimination, In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, ESEC/FSE 2017, pp. 498–510, New York, NY, USA: Association for Computing Machinery

Galster G. 1996. Future directions in mortgage discrimination research and enforcement. In *Mortgage Lending, Racial Discrimination and Federal Policy*, eds. JJ Choi, B Oskan. Taylor & Francis Group, 38–44

Ghent AC, Hernández-Murillo R, Owyang MT. 2014. Differences in subprime loan pricing across races and neighborhoods. *Regional Science and Urban Economics* 48:199–215

Giacoletti M, Heimer R, Yu EG. 2022. Using high-frequency evaluations to estimate disparate treatment: Evidence from mortgage loan officers. Working paper, University of Southern California

Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter MA, Kagal L. 2018. Explaining explanations: An overview of interpretability of machine learning, In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 80–89

Grother PJ, Quinn GW, Phillips PJ. 2011. Report on the evaluation of 2D still-image face recognition algorithms. Tech. Rep. NIST IR 7709, National Institute of Standards and Technology, Gaithersburg, MD

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys* 51(5):93:1–93:42

Hanson A, Hawley Z, Martin H, Liu B. 2016. Discrimination in mortgage lending: Evidence from a correspondence experiment. *Journal of Urban Economics* 92:48–65

Hardt M, Chen X, Cheng X, Donini M, Gelman J, et al. 2021. Amazon SageMaker Clarify: Machine learning bias detection and explainability in the cloud, In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pp. 2974–2983, New York, NY, USA: Association for Computing Machinery

Hardt M, Price E, Srebro N. 2016. Equality of opportunity in supervised learning, In *Advances in Neural Information Processing Systems*, eds. D Lee, M Sugiyama, U Luxburg, I Guyon, R Garnett, vol. 29. Curran Associates, Inc.

Hirshleifer J. 1971. The private and social value of information and the reward to inventive activity. *American Economic Review* 61:561–574

Hutchinson B, Mitchell M. 2019. 50 years of test (un)fairness: Lessons for machine learning, In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 49–58

Iris B. 2016. What works: Gender equality by design. Harvard University Press

Kearns M, Roth A. 2020. The ethical algorithm. Oxford: Oxford University Press

Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV), In *International Conference on Machine Learning*, pp. 2668–2677, PMLR

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S. 2018a. Human decisions and machine predictions. *Quarterly Journal of Economics* 133(1):237–293

Kleinberg J, Ludwig J, Mullainathan S, Sunstein CR. 2018b. Discrimination in the age of algorithms. *Journal of Legal Analysis* 10:113–174

Kullback S, Leibler RA. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22(1):79–86

Kusner MJ, Loftus J, Russell C, Silva R. 2017. Counterfactual fairness, In *Advances in Neural Information Processing Systems*, eds. I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, R Garnett, vol. 30. Curran Associates, Inc.

Ladd H. 1998. Evidence on discrimination in mortgage lending. *Journal of Economic Perspectives* 12(2):41–62

Lee MSA, Floridi L. 2021. Algorithmic fairness in mortgage lending: From absolute conditions to relational trade-offs. *Minds and Machines* 31:165–191

Lewis E. 1992. An introduction to credit scoring. San Rafael, California: Athena Press

Liang P, Bommasani R, Lee T, Tsipras D, Soylu D, et al. 2022. Holistic Evaluation of Language Models. ArXiv:2211.09110 [cs]

Lundberg SM, Lee SI. 2017. A unified approach to interpreting model predictions, In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4768–4777, Red Hook, NY, USA: Curran Associates Inc.

Lunter J. 2020. Beating the bias in facial recognition technology. *Biometric Technology Today* 2020(9):5–7

McDonald DW, Pepe CO, Bowers HM, Dombroski EJ. 1997. Desktop Underwriter: Fannie Mae's automated mortgage underwriting expert system. Working paper, IAAI-97 Proceedings

McMillen D, Singh R. 2020. Assessment regressivity and property taxation. *Journal of Real Estate Finance and Economics* 60:155–169

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54(6):115:1–115:35

Menestrel ML, Wassenhove LN. 2016. Subjectively biased objective functions. *EURO Journal on*

Decision Processes 4(1):73–83

Menzies P, Beebee H. 2019. Counterfactual theories of causation. In *The Stanford Encyclopedia of Philosophy*, ed. EN Zalta. Metaphysics Research Lab, Stanford University, winter 2019 ed.

Molinar C. 2022. Interpretable machine learning: A guide for making black box models explainable. Leanpub

Munnell AH, Browne L, McEneaney J, Tootel G. 1996. Mortgage lending in Boston: Interpreting HMDA data. *American Economic Review* 86(1):25–54

Obermeyer Z, Powers B, Vogel C, Mullainathan S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366:447–453

O'Neil C. 2016. Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books, reprint ed.

Papakyriakopoulos O, Hegelich S, Serrano JCM, Marco F. 2020. Bias in word embeddings, In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 446–457, New York, NY, USA: Association for Computing Machinery

Pearl J. 2009. Causal inference in statistics: An overview. *Statistics Surveys* 3:96–146

Perrone V, Donini M, Zafar MB, Schmucker R, Kenthapadi K, Archambeau C. 2021. Fair Bayesian optimization, In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, p. 854–863, New York, NY, USA: Association for Computing Machinery

Perrone V, Shcherbatyi I, Jenatton R, Archambeau C, Seeger M. 2019. Constrained Bayesian optimization with max-value entropy search. *arXiv:1910.07003 [cs, stat]*

Perry A, Rothwell J, Harshbarger D. 2018. The devaluation of assests in black neighborhood: The case of residential property. Tech. rep.

Pessach D, Shmueli E. 2020. Algorithmic fairness. Working paper, Tel-Aviv University

Reid CK, Bocian D, Li W, Quercia RG. 2017. Revisiting the subprime crisis: The dual mortgage market and mortgage defaults by race and ethnicity. *Journal of Urban Affairs* 39(4):469–487

Ribeiro MT, Singh S, Guestrin C. 2016. "Why should I trust you?": Explaining the predictions of any classifier, In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*

Ross SL, Yinger J. 2002. The color of credit: Mortgage discrimination, research methodology, and fair-lending enforcement. Cambridge, MA: MIT Press

Rudin C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5):206–215

Samek W, Müller KR. 2019. Towards explainable artificial intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, eds. W Samek, G Montavon, A Vedaldi, LK Hansen, KR Müller, Lecture Notes in Computer Science. Cham: Springer International Publishing, 5–22

Shapley LS. 1952. A value for n-person games. Santa Monica, CA: RAND Corporation

Srinivas S, Fleuret F. 2019. Full-gradient representation for neural network visualization. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.

Straka JW. 2000. Shift in the mortgage landscape: The 1990s move to automated credit evaluations. *Journal of Housing Research* 11(2):207–232

Sudjianto A, Knauth W, Singh R, Yang Z, Zhang A. 2020. Unwrapping the black box of deep ReLU networks: Interpretability, diagnostics, and simplification. Tech. rep., Carnegie Mellon University

Sundararajan M, Taly A, Yan Q. 2017. Axiomatic attribution for deep networks, In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, p. 3319–3328, JMLR.org

Szepannek G. 2017. On the practical relevance of model machine learning algorithms for credit scoring applications. *WIAS Report Series* 17:88–96

Tantri P. 2021. Fintech for the poor: Financial intermediation without discrimination. *Review of Finance* 25(2):561–593

Thomas LC. 2009. Consumer credit models: Pricing, profit, and portfolios. Oxford: Oxford University Press

Tjoa E, Guan C. 2020. A survey on explainable artificial intelligence (XAI): Towards medical XAI. *IEEE Transactions on Neural Networks and Learning Systems* :1–21

Vasudevan S, Kenthapadi K. 2020. LiFT: A scalable framework for measuring fairness in ML applications. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* :2773–2780

Waller B, Aiken M. 1998. Predicting prepayment of residential mortgages: A neural network approach. *Information and Management Science* 39(4):37–44

Wauthier FL, Jordan MI. 2011. Bayesian bias mitigation for crowdsourcing. In *Advances in Neural Information Processing Systems 24*, eds. J Shawe-Taylor, RS Zemel, PL Bartlett, F Pereira, KQ Weinberger. Curran Associates, Inc., 1800–1808