FEDERAL RESERVE BANK
PHILADELPHIA

Working Papers
RESEARCH DEPARTMENT

Measuring Fairness in the U.S. Mortgage Market

Hadi Elzayn
Stanford University

Simon Freyaldenhoven
Federal Reserve Bank of Philadelphia

Ryan Kobler
Federal Reserve Bank of Philadelphia

Minchul Shin
Federal Reserve Bank of Philadelphia

WP 25-04

PUBLISHED
February 2025

FRB CFI
PHILADELPHIA

ISSN: 1962-5361

DOI: https://doi.org/10.21799/frbp.wp.2025.04

Measuring Fairness in the U.S. Mortgage Market

Hadi Elzayn
Stanford University

Simon Freyaldenhoven
Federal Reserve Bank of Philadelphia

Ryan Kobler
Federal Reserve Bank of Philadelphia

Minchul Shin
Federal Reserve Bank of Philadelphia[*]

January 10, 2025

Abstract

Black Americans are both substantially more likely to have their mortgage application rejected and substantially more likely to default on their mortgages than White Americans. We take these stark inequalities as a starting point to ask the question: How fair or unfair is the U.S. mortgage market? We show that the answer to this question crucially depends on the definition of fairness. We consider six competing and widely used definitions of fairness and find that they lead to markedly different conclusions. We then combine these six definitions into a series of stylized facts that offer a more comprehensive view of fairness in this market. To facilitate further exploration, an interactive Online Appendix allows the user to examine our fairness measurements further across both time and space.

JEL-Classification: D63, G21, G28, J15, R21
Keywords: fairness, discrimination, inequality, measurement, algorithmic decisions, HMDA

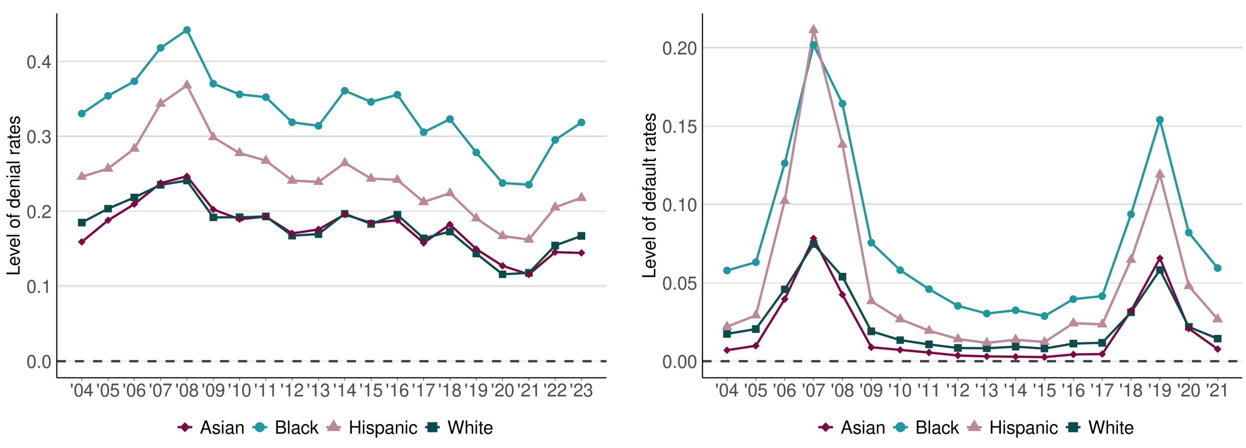
[*]We are grateful to Andrew Gross, Kellen O'Connor, and Eliana Sena Sarmiento for excellent research assistance. We thank Neil Bhutta, Bob Hunt, Lauren Lambie-Hanson, and Edison Yu for helpful comments and suggestions. The views expressed herein are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia, or the Federal Reserve System. No statements here should be treated as legal advice. Emails: hselzayn@law.stanford.edu, simon.freyaldenhoven@phil.frb.org, ryan.kobler@phil.frb.org, minchul.shin@phil.frb.org.

1 Introduction

Is the $2 trillion U.S. mortgage market *fair*? Figure 1 depicts the default rates and denial rates for mortgage applications filed in the U.S. in recent years for different demographic groups. We see that Black and Hispanic borrowers are substantially more likely to have their loan application rejected than White and Asian borrowers. At the same time, Black and Hispanic borrowers also exhibit a substantially higher default rate than White and Asian borrowers.

We take these stark inequalities as a starting point to ask the question: How fair or unfair is the U.S. mortgage market with respect to the race of an applicant? To answer this question, we first need a formal definition of fairness. However, many competing definitions of fairness exist. For example, Narayanan [2018] reviews a total of 21 fairness definitions. Similarly, as we discuss in Section 3, different parts of existing laws and regulatory guidance use different definitions of fairness. A second question we address in this paper is therefore: Does it matter what definition of fairness we use? Are the conclusions and policy implications the same across fairness definitions or do they differ?

Understanding fairness in the U.S. mortgage market is important for several reasons. Mortgage balances are the largest source of debt for most Americans: by the end of 2022, out of a total household debt in the U.S. of $17 trillion, mortgage balances made up $12 trillion (Federal Reserve Bank of New York [2023]). Additionally, disparities in this market can translate into inequality in home ownership and thus wealth accumulation; indeed, it is well known that the mortgage market plays a prominent role in the persistence of wealth gaps across generations (Charles and Hurst [2003], Kuhn et al. [2020]), and that, historically,



(a) Fraction of mortgage applications denied by year of application.

(b) Fraction of mortgages that default by year of origination.

Figure 1: Summary statistics by demographic group and year for mortgage applications filed in the U.S.

racial discrimination through practices such as redlining and racial covenants contributed to today's racial wealth gaps. We further note that mortgage lending has seen a shift towards algorithmic underwriting over the past two decades. As we argue in Section 3, this makes some of the traditional ways to measure (and regulate) fairness challenging. We thus also consider recent definitions of fairness developed in the algorithmic fairness literature.

We stylize our setting as one in which a sequence of mortgage applicants appears before a decision-maker, who reaches a decision about whether to approve or deny each application based on a set of features. The decision-maker may be a human or an algorithm. We observe some application features including the race of an applicant, the decisions made, and the resulting outcomes of originated mortgages (default/non-default). We do not assume access to the algorithm or process by which decisions are made, and explicitly allow for the possibility that it includes additional, unobserved applicant characteristics. This formulation corresponds to our empirical setting, for which we combine two anonymized sources of data. First, we utilize a confidential version of the Home Mortgage Disclosure Act (HMDA) data. This includes the vast majority of all mortgage applications filed in the U.S. and contains applicant features including protected characteristics, as well as the loan decisions. Second, we leverage the ICE, McDash (hereinafter referred to as McDash) dataset, which contains the servicing portfolios of the largest residential mortgage servicers in the U.S. and covers approximately two-thirds of installment-type loans in the residential mortgage servicing market. Matching the approved loans in HMDA to their servicing records in McDash allows us to track the performance of approved mortgages over time.

With this setting in mind, we construct six different fairness definitions: Statistical Parity, Predictive Parity, the Marginal Outcome Test, Equalized Odds, Conditional Statistical Parity, and Representativeness (we formally define these in Section 3). Some of these definitions are straightforward to compute and have been studied extensively in the literature. Others are more complex and, to the best of our knowledge, have not been applied in this setting. For example, leveraging machine learning techniques, our paper utilizes a novel strategy to detect applicants that submit multiple applications ("cross-applicants"). We then use these cross-applicants to construct two of our measures of fairness. Our paper also makes progress on empirically evaluating definitions that have thus far primarily been considered in more theoretical papers. We then contrast each of these fairness measures and find that any single fairness measure comes with significant drawbacks. Our findings are consistent with existing theoretical results showing that satisfying all or even multiple definitions at once is generally impossible (Kleinberg et al. [2017], Chouldechova [2017]).

Our paper contributes to an extensive literature on fairness and discrimination in the mortgage market, building upon works such as those by Black et al. [1978], Munnell et al. [1996], Berkovec et al. [1998], Ross and Yinger [2002], and Cheng et al. [2015]. Recent advancements in data availability have spurred a new wave of studies incorporating default outcomes, as seen in works by Bhutta and Hizmo [2021], Giacoletti et al. [2022], and Bartlett et al. [2022]. Further, our work is related to a recent literature that explores how machine learning algorithms can mitigate racial differences in lending (Tantri [2021], Bartlett et al. [2022], Fuster et al. [2022], and Bhutta et al. [2022]). Existing studies usually focus on observing differences in a single outcome of interest (e.g., denial rates, pricing, or mortgage performance) across race. For example, Munnell et al. [1996] focuses on denial rates and finds that minorities in Boston had denial rates up to twice as high as White applicants. Berkovec et al. [1998] focuses on loan performance and finds that Black borrowers are more than twice as likely to default. In contrast, we consider a comprehensive list of fairness definitions. This allows us to empirically show that different definitions of fairness can lead to very different conclusions and policy implications.

In Section 4, we describe five stylized facts based on a combination of the six fairness measures above. While this allows us to highlight recent trends and geographic patterns in fairness and inequality in the U.S. mortgage market, we encourage the reader to further explore our results interactively through our online interactive appendix. For example, throughout this paper, we focus on the race and ethnicity of an applicant, but our online appendix also includes all fairness measures for the gender of the applicant. It will be updated periodically and is available at `https://www.philadelphiafed.org/surveys-and-data/consumer-finance-data/mortgage-fairness-explorer`.

Before we proceed, we caveat two important limitations to our setup. First, we do not consider other aspects of the decision process beyond approval or denial. There is evidence suggesting that minority applicants may receive less assistance during the mortgage application process (Frame et al. [2021], Kim and Squires [1995] and may even be discouraged from applying in the first place (Ladd [1998], Yinger [1991], Lubin [2008], Ross et al. [2008]). We also do not study fairness with respect to pricing. For recent papers on discrimination in pricing in the context of mortgage applications, see Ambrose et al. [2021], Bhutta and Hizmo [2021], Bartlett et al. [2022], Willen and Zhang [2020]. Second, we restrict ourselves to measures of group fairness and do not consider notions of procedural (Grgić-Hlača et al. [2018]), individual (Dwork et al. [2012]), or compositional fairness (Dwork and Ilvento [2019]).

2 The Data

We draw our data from two high-quality administrative data sources. All figures and tables are based on authors' calculations using these two datasets:

1. **Home Mortgage Disclosure Act (HMDA).** This dataset contains anonymized data on mortgage applications. With few exceptions, all mortgage applications filed in the U.S. are subject to HMDA reporting and thus included in this database.[1] HMDA data has been one of the primary datasets in the literature to study inequality in mortgage finance across protected classes (e.g. Munnell et al. [1996], Berkovec et al. [1998], Ross and Yinger [2002], Bayer et al. [2018], Bhutta et al. [2022] to mention but a few). In fact, "identifying possible discriminatory lending patterns" was one of

[1] Previous studies have estimated HMDA captures between 80 to 92 percent of mortgages across the U.S., with coverage being higher in metropolitan relative to non-metropolitan areas (Bhutta et al. [2017]).

the stated purposes of establishing HMDA in 1975.[2] Unusual for financial datasets, HMDA contains protected attributes of the applicants, such as an applicant's race, ethnicity, and gender. While a publicly available version of this dataset exists, we work directly with a confidential version that is available to users within the Federal Reserve System and includes more detailed information for each loan application (e.g., the exact date an application was filed, applicant and coapplicant age, credit score, automated underwriting system results, debt-to-income ratio (DTI) and the loan-to-value ratio (LTV) among others).

Throughout this paper we construct and focus our attention on four non-overlapping major demographic groups. These are Asian applicants, Black applicants, Hispanic White applicants, and non-Hispanic White applicants.[3] For ease of notation, we will refer to these four groups simply as Asian, Black, Hispanic and White, respectively. We denote the variable that encodes the four groups by G_i. Since alternate race and ethnicity categories were introduced in the HMDA data in 2004 (in accordance with changes made to the U.S. Census), our analysis includes mortgage applications filed between 2004 and 2023 to avoid ambiguity in the definition of G_i. We further retain only first-lien mortgages and applications that are either approved or denied, dropping applications that are withdrawn by the applicant before a decision was made, applications closed for incompleteness, loan purchases, and applications that only went through the preapproval process. Finally, we drop applications filed outside the 50 states and Washington D.C.

On its own, HMDA has some notable shortcomings. Since the dataset consists of mortgage applications, it does not include any information on the subsequent performance of the originated loans, such as whether a loan becomes delinquent. The set of available observed features associated with an application also changes throughout the sample period, and a number of important borrower and loan characteristics, such as the credit score and the LTV, are available only starting in 2018. Our second dataset, McDash, addresses some of these shortcomings.

2. **ICE, McDash (McDash).** This dataset comprises the de-identified servicing portfolios of the largest residential mortgage servicers in the U.S., covering approximately two-thirds of originated loans in the residential mortgage servicing market. It allows us to track the performance of originated mortgages over time. In particular, it includes a monthly variable indicating loan delinquency status. McDash also includes a richer set of borrower and loan characteristics across a longer time span relative to HMDA, such as the credit score, DTI, and LTV. On the other hand, McDash does not include a number of borrower characteristics available in HMDA, such as a borrower's race and ethnicity, and by definition includes only approved (and originated) applications.

[2]For more background, see https://www.ffiec.gov/hmda/history.htm.

[3]Applicants are classified into mutually exclusive groups based on the first race and ethnicity reported for the primary applicant in HMDA.

3. **Matched Sample (HMDA-McDash).** While we can compute some of our fairness measures using only the HMDA dataset (e.g., Statistical Parity), other fairness measures are infeasible to compute with HMDA alone because we need access to the subsequent performance of originated loans. We therefore match the loan applications in HMDA with the servicing records in McDash. Individual observations in HMDA and McDash are matched using origination date, loan amount, property ZIP code, lien type, loan purpose (e.g., purchase or refinance), loan type (e.g., conventional or FHA), and occupancy type (e.g., owner-occupied, absentee or investment property) using the same matching logic that is standard in the literature (e.g., Fuster et al. [2022]).[4]

   For the matched sample that links the loans in HMDA to their servicing records, we further restrict our analysis to mortgage applications filed until 2021. This allows us to observe the servicing record of each originated loan for at least two years using servicing records in McDash through 2023. Accordingly, we use two-year probabilities of delinquency throughout to ensure comparability of different vintages. Specifically, we consider a loan delinquent if it is ever 90 or more days past due (i.e., three or more missed payments) within 24 months of origination.[5]

   We further restrict the matched sample to loans with a term of 10, 15, 20, or 30 years, a loan-to-value ratio (LTV) between 0% and 200%, a credit score between 500 and 820, debt-to-income ratio (DTI) within 0% to 250%, income between 0 and $1 million, and a loan amount up to $1.5 million. These restrictions both reduce the amount of data errors and facilitate the estimation of our default model by restricting the heterogeneity of the matched sample.

We conclude the discussion of our data with a number of summary statistics. The HMDA data contains 415.3 million mortgage applications filed between 2004 and 2023. After filtering for first-lien mortgage applications that are either approved or denied in which the applicant identifies as Asian, Black, Hispanic, or White, and imposing our geographic restriction, we retain about 211.8 million applications in our sample. The full matched sample of originated mortgages and their subsequent servicing record contains 74.0 million loans originated between 2004 and 2021. After imposing our additional data filters, we end up with about 39.3 million loans.

Table 1 presents summary statistics for a subset of observable features. It includes the median loan amount, income, credit score, and LTV ratio, broken down by demographic group. We calculate these separately for our two datasets. Table 1 further includes the default probability of originated loans in the matched HMDA-McDash data. Perhaps unsurprisingly,

[4]Overall, among McDash mortgages originated over our sample period, each year 61-80% of the servicing records are matched to at least one HMDA loan. We limit our analysis to loans that can be uniquely matched, meaning that each McDash loan is matched to just one HMDA loan and that each HMDA loan has only one McDash match candidate. Using this conservative sample reduces our match rate to between 47% and 67% depending on the year, although it tends to improve over time.

[5]Note that this coincides with the default outcome targeted by the most widely used credit scoring models.

Group	Dataset	Obs	Amount (1000s)	Income (1000s)	Credit Score	LTV	P(Default) (%)
Total	HMDA	211,814,666	180	73	732	75	
	HMDA-McDash	39,315,246	193	79	747	79	2.9
Asian	HMDA	12,785,703	294	104	761	74	
	HMDA-McDash	2,452,690	283	105	765	75	1.8
Black	HMDA	19,136,690	154	58	678	80	
	HMDA-McDash	2,589,287	177	64	694	89	8.3
Hispanic	HMDA	19,670,586	191	64	707	80	
	HMDA-McDash	3,209,512	191	65	717	82	5.5
White	HMDA	160,221,687	175	74	739	75	
	HMDA-McDash	31,063,757	190	80	751	79	2.3

Table 1: Descriptive Statistics across demographic groups. Depicted across the columns are the median of loan amount, income, credit score, and LTV, as well as the empirical frequency of default. The HMDA sample contains mortgage applications filed 2004-2023; HMDA-McDash contains mortgages originated 2004-2021. In HMDA, credit score and LTV are available only starting in 2018. The corresponding columns for both HMDA and HMDA-McDash are therefore based only on applications filed in 2018 and later.

we see that there are large differences across demographic groups. For example, White borrowers tend to have substantially higher income and credit scores than Black and Hispanic borrowers.

3 Fairness Measures

To define the fairness measures, we first introduce some notation. An applicant i applies for a mortgage at a potential lender l. Applicants have a vector of features partitioned into Z_i, η_i, and G_i, and there is a joint distribution $\mathcal{D}$ over which these features are drawn. G_i is a discrete variable indicating membership in a protected class (e.g. a binary indicator for applicant gender or a categorical variable for applicant race). Z_i and η_i are covariates that may be related to the default probability of individual i. An applicant may apply for a mortgage at multiple lenders, and we denote the number of applications applicant i submits by N_i.

The lender l decides whether to approve or deny an application submitted by an applicant i. Then, L_{il} is a binary indicator variable that takes a value of one if the application is approved and zero otherwise. D_{il} is a binary indicator variable that takes a value of one if a person defaults on his loan that was approved by lender l and zero otherwise.

Throughout, we maintain that $E(D_{il}|Z_i, \eta_i, G_i, l) = E(D_{il}|Z_i, \eta_i, G_i)$, implying that the default probability does not depend on the lender, l, conditional on i. Note that D_{il} is realized (and thus observed) only when an application is both approved and originated.

While the lender observes Z_i, η_i, and G_i, the econometrician (or regulator) who wants to measure the fairness of the decision process observes only Z_i and G_i, but not η_i. Thus, Z_i and G_i are visible to both the lender and to the regulator, while η_i is visible only to the lender.

We maintain the following Assumption, which states that the group membership G_i has no *direct* impact on an individual's default probability.

Assumption 1.

$$E(D_{il}|Z_i, \eta_i) = E(D_{il}|Z_i, \eta_i, G_i) \tag{1}$$

However, since G_i may be correlated with both Z_i and η_i, we allow for differential default probabilities across groups, both unconditionally and conditional on the observed covariates Z_i (i.e. we do not assume that $E(D_{il}|Z_i) = E(D_{il}|Z_i, G_i)$).

Furthermore, we denote the geographic state and time of a loan application by S_{il} and T_{il}, with realizations s and t, respectively. We additionally denote by $\mathcal{I}_{\mathcal{A}} := \{i, l : a_{il} \in \mathcal{A}\}$ the set of applications for which the condition $a_{il} \in \mathcal{A}$ holds. For example, $\mathcal{I}_{s,t,g}$ denotes the set of applications in state s during year t and of demographic group g. Finally, we define $N_{s,t,g} = |\mathcal{I}_{s,t,g}|$, which is the cardinality of the set $\mathcal{I}_{s,t,g}$.

We consider six classes of fairness definitions, which we will define more formally below:

1. Statistical Parity (difference in denial rates)
2. Predictive Parity (difference in default rates)
3. Marginal Outcome Test (difference in lending standards)
4. Equalized Odds (difference in denial rates for creditworthy borrowers)
5. Conditional Statistical Parity (conditional difference in denial rates)
6. Representativeness (amount of under-representation among approved)

As we explain below in more detail, these correspond to common notions of fairness frequently alluded to in either current regulations or public debate.

3.1 Statistical Parity

We begin with the notion of Statistical Parity ("SP"), which is arguably the most elementary notion of fairness. Statistical Parity requires that borrowers have equal approval (equivalently, denial) probabilities regardless of their group.

Thus, we say that Statistical Parity holds if $P(L_{il} = 0|G_i = g) = P(L_{il} = 0|G_i = g')$ and define any violation of Statistical Parity, δ_{SP}, as the difference in denial rates between groups

g and g':

$$\delta_{SP}(g, g') := P(L_{il} = 0 | G_i = g) - P(L_{il} = 0 | G_i = g'). \quad (2)$$

Throughout, g' represents a "reference group." Throughout the paper, this will always be non-Hispanic White applicants and all fairness violations are defined such that larger numbers correspond to worse outcomes for minority applicants relative to non-Hispanic White applicants. In practice, we then simply replace the population quantities on the right by their empirical counterpart.[6] Statistical Parity is easy to compute and captures an intuitive notion about inequality across groups. It also corresponds with language used in current federal regulations. The Uniform Guidelines for Employee Selection Procedures of 1978 (also see the Civil Rights Act of 1964 and the Equal Employment Opportunity Act of 1972) state in §1607.4D:

> "Adverse impact and the 'four-fifths rule.' A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact."

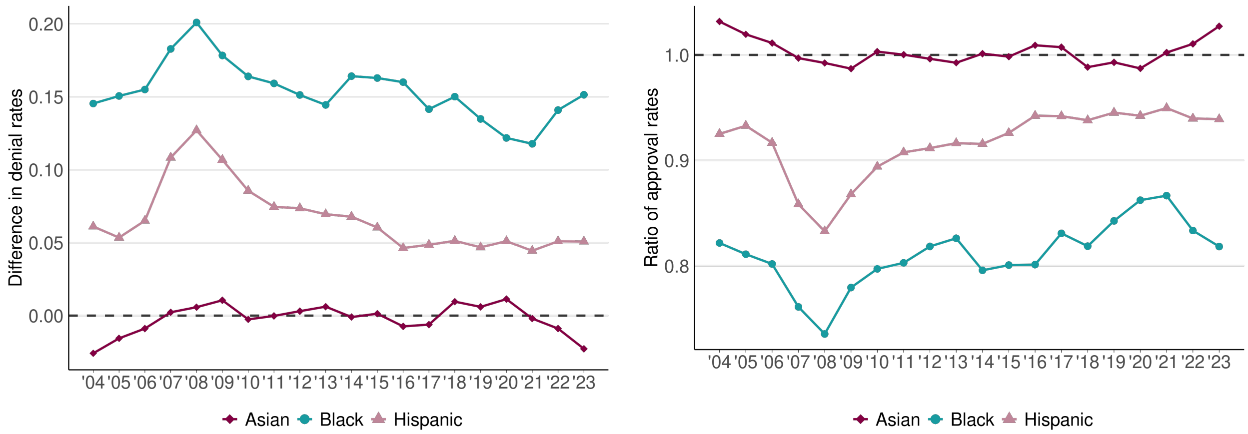
In 2015, in "Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc." the Supreme Court ruled that adverse impact claims are also cognizable under the Fair Housing Act (Civil Rights Act of 1968, Titles VIII-IX), extending the notion of disparate impact to housing cases.

Figure 2 illustrates the inequality in mortgage decisions between minority and White applicants for applications filed between 2004 and 2023. Figure 2a depicts violations of Statistical Parity, defined in (2), with respect to White applicants. Figure 2b depicts an alternative measure of Statistical Parity violations that uses the ratio of approvals rather than the difference in denials between minority and White borrowers, in line with the "four-fifths" rule we mention above. Both plots point to a large and persistent gap between denial rates for Black and White applicants. Focusing on the gap between Black and White applicants, we note that the gap appears relatively stable until 2016, but then has improved over the last six years of our data. In fact, following the aforementioned Supreme Court ruling, the approval ratio has consistently been above 80% for the first time in recent years.

[6] We sometimes construct our measures of fairness on a subset of the data, based on a set of discrete variables, for example the year and state of an application. The violation of Statistical Parity for a state-year combination s, t is then computed as:

$$\delta_{\text{SP}_{s,t}}(g, g') := \frac{1}{N_{s,t,g}} \sum_{i,l \in \mathcal{I}_{s,t,g}} \mathbf{1}[L_{il} = 0] - \frac{1}{N_{s,t,g'}} \sum_{i,l \in \mathcal{I}_{s,t,g'}} \mathbf{1}[L_{il} = 0]$$

To simplify notation, we will omit the corresponding subscripts to denote a subset of the data in the remainder of the paper when it is clear from context. For example, any regression-based measures at the state-year level will be based on a separate regression for each state-year combination.



(a) Difference in denial rates between minority applicants and White applicants.

(b) Ratio of approval rates relative to White applicants.

Figure 2: Figure based on mortgage applications filed in the U.S. between 2004 and 2023.

3.2 Predictive Parity

We now turn to Predictive Parity ("PP"). Instead of comparing the probability of loan approval across groups (cf. Figure 1a), Predictive Parity asks that the default rate be similar among approved borrowers regardless of their group (cf. Figure 1b). Like Statistical Parity, Predictive Parity is easy to compute and communicate, and captures an intuitive and meaningful notion of inequality across groups.[7]

Thus, we say that Predictive Parity holds if $P(D_{il} = 1 | L_{il} = 1, G_i = g) = P(D_{il} = 1 | L_{il} = 1, G_i = g')$ and define violation of Predictive Parity, δ_{PP}, as the difference in default rates between groups g and g':

$$\delta_{PP}(g, g') := P(D_{il} = 1 | L_{il} = 1, G_i = g) - P(D_{il} = 1 | L_{il} = 1, G_i = g'). \tag{3}$$

Figure 3 illustrates the inequality in default rates between minority and White applicants for applications filed between 2004 and 2021. We note a persistent disparity in default rates across demographic groups. For example, the default rate for Black borrowers is 2-13 percentage points higher than that of White borrowers during our sample period. Figure 3 further highlights a substantial increase in the *differences* in default rates between demographic groups during the 2007/2008 Financial Crisis and the 2020 COVID-19 Pandemic: Minority homebuyers suffered disproportionately from these crises (also see Bayer et al. [2016]).

The fact that Black borrowers default at substantially higher rates reflects the systematically worse financial outcomes Black Americans face, and may further compound existing economic

[7] Also see Gerardi and Willen [2009], Bayer et al. [2016], or Li and Mayock [2019] for recent papers that focus on differences in default rates across demographic groups.

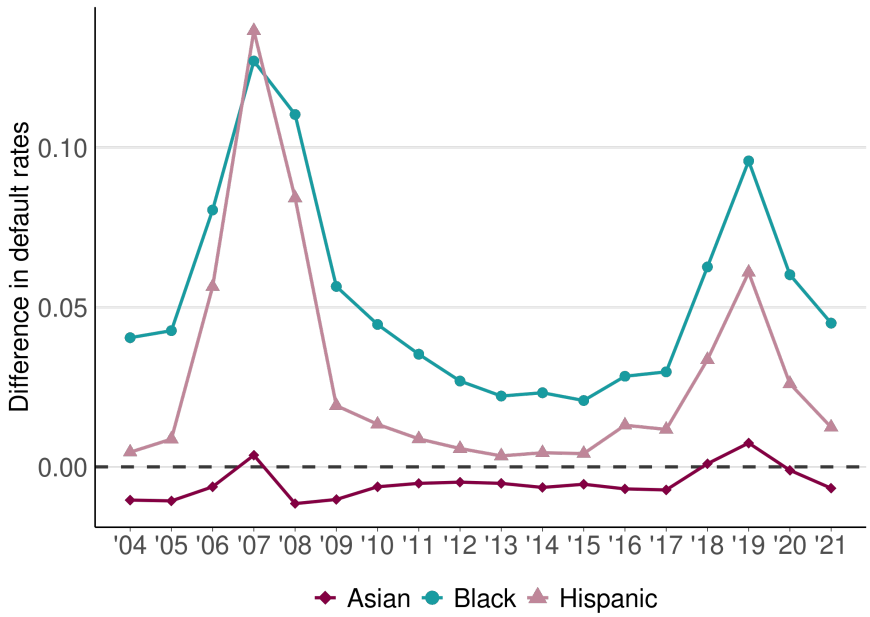


Figure 3: Difference in default rates between minority borrowers and White borrowers. Figure based on mortgage originations between 2004 and 2021.

and financial disparities. However, we note that different default rates among approved applicants are not necessarily suggestive of different lending standards. This is because, as Simoiu et al. [2017] notes, such outcome-based tests suffer from the problem of "infra-marginality"–in other words, even without discrimination, default rates may differ between groups if they have different underlying risk distributions (also see Yinger [1996], Ross [1997]).

We therefore look directly at candidates "on the margin" next.

3.3 Marginal Outcome Test

One intuitive notion of fairness widely used in economics and current law requires that the same credit standards are applied across protected classes. In order to assess whether this is indeed the case, much of the economics literature, going back to the classic theory of Becker [1957], has focused on marginal candidates, both in lending decisions and other contexts (see, e.g., Anwar and Fang [2015] and Arnold et al. [2018] in the context of the criminal justice system; Dobbie et al. [2021] in consumer lending; and Berkovec et al. [1998] in mortgage lending). The idea is that if the same threshold of creditworthiness is used for Black and White applicants, people at this threshold (or "on the margin") should default at the same rate. We call this the "Marginal Outcome Test"("MOT").

This focus on marginal applications and/or applicants is also reflected in the Interagency

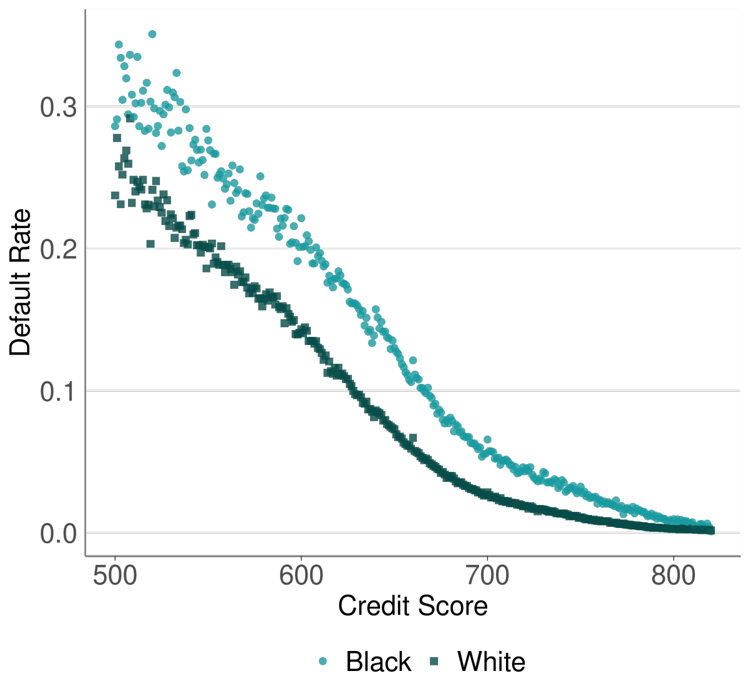


Figure 4: Default rates as a function of credit scores for mortgage applications filed in 2004-2021.

Fair Lending Examination Procedures,[8] for example:

> "The examiner-in-charge should, during the following steps, judgmentally select from the initial sample only those denied and approved applications which constitute marginal transactions."

Empirically, the key step to a fairness measure based on the Marginal Outcome Test is therefore to identify a set of marginal candidates: candidates who were just barely offered a loan.[9] However, obtaining marginal candidates or transactions is often difficult, as evident by the fact that examiners are instructed to "judgmentally" select them in the guidelines above.

One approach to identifying marginal candidates is via observables. For example, a major input into the loan decision of a lender is an individual's credit score, which is explicitly used in the underwriting decisions of the government-sponsored enterprises (GSEs) Fannie Mae and Freddie Mac (Bubb and Kaufman [2014]). One could thus construct marginal candidates as those applicants within a particular (intermediate) credit score range.

However, in Figure 4, which depicts the empirical default rate as a function of an applicant's credit score, we observe that Black applicants have higher default rates than White applicants

[8]These are publicly available online at `https://www.ffiec.gov/pdf/fairlend.pdf`.

[9]Note that identifying such candidates empirically (i.e., detecting which applications were just barely offered a loan) does not re-identify individuals from our anonymized data.

at all credit scores. This suggests that credit score by itself is not a well-calibrated default model in this context – Black and White applicants with the same credit score are associated with different default risks (cf. Kermani and Wong [2021]) – and that additional variables are needed to determine creditworthiness. In Appendix B, we show a strikingly similar pattern for more sophisticated default models that involve a rich set of covariates. Training a machine learning model[10] to predict default on a large set of observable features (not including race), we obtain a similar degree of miscalibration, again systematically underestimating the default risk of Black applicants. Such miscalibration of the underlying default risk model reflects an inherent concern with any approach that uses estimated creditworthiness to identify marginal candidates. That is, identifying marginal candidates based on this type of approach is imprecise at best.

Therefore, we propose an alternative way to construct marginal applicants by identifying candidates who apply for multiple mortgages. An applicant is then marginal if she submits two (nearly) identical applications, and receives one approval and one denial. We consider such an applicant marginal by "revealed preference": The fact that one lender approved the loan application, while another lender rejected the same application, reveals that the corresponding applicant is on the threshold between acceptance and denial.

While we describe the algorithm used to identify pairs of applications that correspond to applicants who submit multiple (nearly identical) applications ("cross-applicants") in more detail in a companion paper, we briefly outline the main idea here. The HMDA dataset is at the application level and does not include the person identifier i. To identify pairs of applications from the same applicant, we apply a state-of-the-art agglomerative hierarchical clustering algorithm to find clusters (usually pairs) of applications that are nearly identical. We then use the rate at which clusters contain multiple originations to both fine-tune our algorithm and to estimate the frequency with which our clusters correctly correspond to single individuals. The idea is the following: if all clusters are pairs of applications from two applicants, most clusters with two approvals would have two originations. On the other hand, if all clusters are pairs of applications from one applicant, no clusters with two approvals can have two originations, since it is impossible to take out two first-lien loans on the same property. Using this logic, we estimate that more than 92% of our estimated cross-applicants indeed represent single individuals at our preferred specification (see Elzayn et al. [2024] for more detail).[11]

We can then select from these (estimated) individuals that submitted two nearly identical

[10] In particular, we use a histogram-based gradient-boosted classification tree (HGBC) to capture the rich nonlinear relationships between observable features such as credit score, loan-to-value ratio, applicant income, loan type, among others and default. See Appendix B for more detail.

[11] On the other hand, since our algorithm requires applications to be nearly identical to be in the same cluster, we may categorize only a small fraction of applicants that submit multiple applications into the same cluster. That is, we estimate that almost all clusters correspond to cross-applicants, but the reverse is not true: we do not know which fraction of cross-applicants correspond to clusters. For further discussion, see Elzayn et al. [2024].

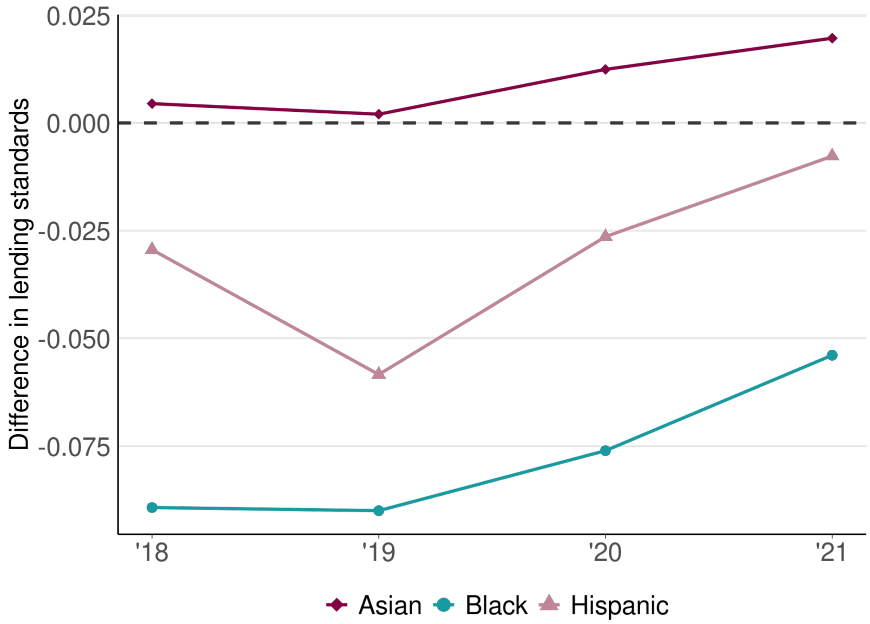


Figure 5: Implied difference in lending standards (i.e., difference in default rates) between minority and White marginal borrowers.

applications those who received one approval and one denial, and calculate any violations of our Marginal Outcome Test in our dataset as follows:

$$\delta_{MOT}(g, g') := P(D_i = 1 | i \in \mathcal{M}_{g'}) - P(D_i = 1 | i \in \mathcal{M}_g), \tag{4}$$

where $\mathcal{M}_g$ denotes the set of applicants (clusters) in group g that submitted multiple applications and experienced both an approval and a denial before originating a (single) loan. Since each individual can originate only one loan, we may suppress the subscript l and simply write D_i. Note that $\delta_{MOT}(g, g') > 0$ if the default rate of minority applicants is *lower* than that of White applicants. This is because lower default rates at the margin for members of group g imply higher lending standards for this group. Contrasted with Predictive Parity, which captures a broader notion of inequality and systematically worse financial outcomes, the Marginal Outcome Test addresses the problem of "infra-marginality" and hence may be able to measure such differences in lending standards.

Figure 5 depicts the implied difference in lending standards between minority marginal applicants and White marginal applicants based on (4). We find that marginal Black and Hispanic applicants default more frequently. This is not consistent with higher lending standards for Black applicants. In fact, since higher default rates for minority applicants imply slightly *lower* lending standard for minority applicants, our fairness violation is "negative" for Black and Hispanic borrowers, following our definition in (4). This result would be consistent with industry efforts or existing public programs aimed at approving more minority borrowers and to reduce inequality (e.g., Agarwal et al. [2012]).

However, there are multiple alternative explanations for why the default rates of Black marginal applicants are higher than those of White marginal applicants. For example, this pattern may also arise as the result of miscalibrated default models. As discussed above, we find that default models excluding the applicant's race tend to underestimate the default risk of Black applicants (e.g. due to further unobserved factors correlated with both race and creditworthiness). If lenders base their underwriting decisions on similarly miscalibrated models, marginal Black applicants will tend to be riskier than White applicants. Another explanation may be that lenders base their decision on other factors beyond the default risk of an applicant. That is, we implicitly equate creditworthiness and default risk (also see Stylized Fact 3 in Section 4). This will generally not be true if profitability of a loan is determined by more than its default risk. We plan to examine the drivers of our documented difference in default rates among marginal candidates further in future work.

3.4 Equalized Odds

While the Interagency Fair Lending Examination Procedures explicitly invoke *marginal applicants* to measure fairness (see our discussion in the previous section), the Equal Credit Opportunity Act instead explicitly invokes *creditworthy applicants*. In particular, the Equal Credit Opportunity Act states in Regulation B, 12 CFR § 1002.1(b):

> "The purpose of this part is to promote the availability of credit to all creditworthy applicants without regard to race, color, religion, [...]"

Defining a "creditworthy applicant" ex post as a borrower that did not default (also see Meursault et al. [forthcoming]) allows the construction of the following two measures of fairness.

1. Equality of Opportunity ("EOP"): Consider cross-applicants with an originated loan that did not default. We can ask the question: How likely was such a "creditworthy applicant" denied at least once at the time of application? Intuitively this captures the notion of being *unfairly denied*, since the borrower repaid her loan. We can then ask whether the rate of these *unfair denials* varies by group membership.

2. Equality of Goodwill ("EGO"): Consider cross-applicants with an originated loan that defaulted. We can ask the question: How likely was such an "uncreditworthy applicant" to be approved across all her applications? Intuitively this captures the notion of being *unfairly approved*, since the borrower did not repay her loan. We can then ask whether the rate of these *unfair approvals* varies by group membership.

These two measures (Equality of Opportunity; Equality of Goodwill) are often combined into the term "Equalized Odds"(e.g. Hardt et al. [2016]). Both capture the idea of equal treatment to groups of individuals based on their "true type": defaulters and non-defaulters.[12]

[12] Also see Angwin et al. [2016] for an application of this fairness measure in a criminal justice context, where algorithms are widely used to rate a defendant's risk of future crime. Angwin et al. [2016] find that, among defendants that re-offended, risk scores for White defendants are substantially lower than those of

In particular, Equality of Opportunity holds if the denial rates are identical for Black and White borrowers who would not default. Equality of Goodwill holds if the approval rates are identical for Black and White borrowers who would default.

Unfortunately, computing Equality of Opportunity and Equality of Goodwill is generally infeasible. This is due to the standard selective labels issue (e.g. Lakkaraju et al. [2017]): we know whether the applicant defaulted only if the loan is extended to (and originated by) the applicant. However, leveraging applicants who submit multiple (nearly identical) applications, we can calculate a version of EOP and EGO. For cross-applicants who apply for multiple mortgages with nearly identical applications and are approved and originate at least once, we can observe the rate at which their applications were approved, conditional on their ultimate default outcome. We therefore compute violations of Equality of Opportunity and Equality of Goodwill in our dataset using the estimated cross-applicants we obtained using our agglomerative hierarchical clustering algorithm explained in the previous section (and derived in more detail in Elzayn et al. [2024]). Specifically, we base our measure on applicants who i) applied to multiple mortgages ($N_i \geq 2$); ii) had at least one of their applications approved ($\sum_l L_{il} \geq 1$); and iii) originated at least one of their approved loans ($O_i = 1$):

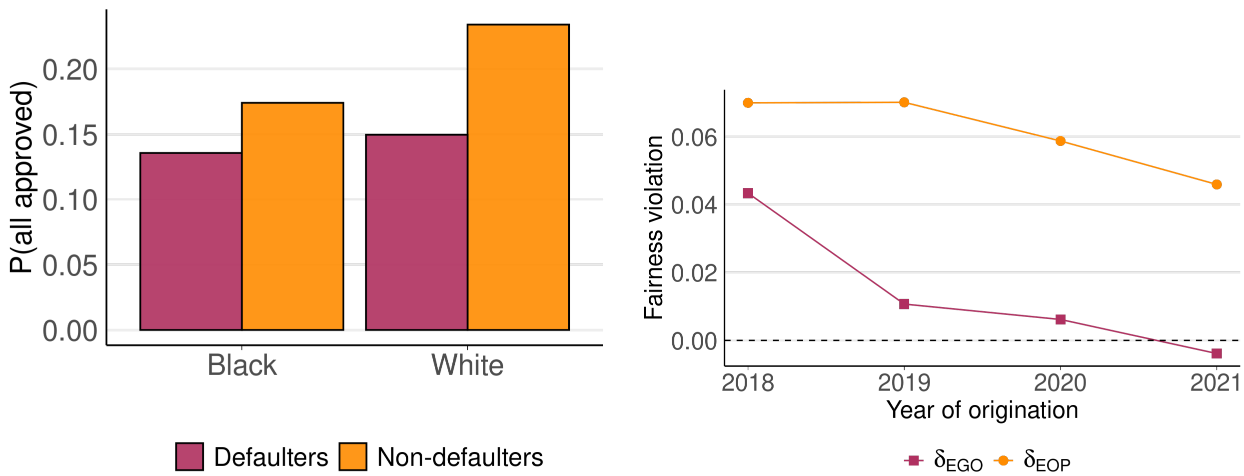
$$\begin{aligned}
\delta_{EOP}(g,g') = P&\left(\sum_l (1-L_{il}) \geq 1 \middle| N_i \geq 2, O_i = 1, D_i = 0, G_i = g\right) \\
&- P\left(\sum_l (1-L_{il}) \geq 1 \middle| N_i \geq 2, O_i = 1, D_i = 0, G_i = g'\right), \qquad (5)
\end{aligned}$$

$$\begin{aligned}
\delta_{EGO}(g,g') = P&\left(\sum_l L_{il} = N_i \middle| N_i \geq 2, O_i = 1, D_i = 1, G_i = g'\right) \\
&- P\left(\sum_l L_{il} = N_i \middle| N_i \geq 2, O_i = 1, D_i = 1, G_i = g\right). \qquad (6)
\end{aligned}$$

Thus, we say that Equality of Opportunity, defined by $\delta_{EOP}(g,g') = 0$, holds in our data if the frequency with which at least one application was denied for cross-applicants that 1) have at least one application originated and 2) did not default, is equal for members of group g and g'. We say that Equality of Goodwill, defined by $\delta_{EGO}(g,g') = 0$, holds in our data if the frequency with which all applications were approved for cross-applicants that 1) have at least one application originated and 2) defaulted, is equal for members of group g and g'.

Figure 6a depicts the frequency with which Black and White cross-applicants have all their applications approved. We see that, among cross-applicants who default on their loan,

Black defendants, and therefore conclude that the algorithm is thus racially biased.



(a) Proportion of cross-applicants (with an origination) who had all of their applications approved.

(b) EGO and EOP fairness violations over time.

Figure 6: Equalized Odds for mortgages originated in 2018-2021.

the difference is small: While Black defaulters are slightly less likely to have all of their applications approved, around 15% of defaulters from each demographic group are (unfairly) approved. On the other hand, among cross-applicants who did not default on their loan, Black applicants are 6 percentage points less likely to have all of their applications approved. Remarkably, the rate at which Black creditworthy applicants (non-defaulters) are approved is much closer to the rate at which White un-creditworthy applicants (defaulters) are approved than to that of White creditworthy applicants.

Consequently, we find larger violations for Equality of Opportunity than for Equality of Goodwill (cf. Figure 6b). We further observe that both violations become smaller during our sample and that Equality of Goodwill is approximately satisfied starting in 2019.

3.5 Conditional Statistical Parity

As mentioned above, conditioning on the true outcome (default/non-default) requires both multiple applications and an approved loan, limiting the sample for which we can compute the previous measures. To sidestep these difficulties, we next turn to Conditional Statistical Parity. Instead of conditioning on the true outcome, this involves conditioning on other features of the applicant or application.

Formally, we define violations of Conditional Statistical Parity ("CSP") for a given y as

$$\delta_{CSP}(g, g', y) := P(L_{il} = 0 | Y_i = y, G_i = g) - P(L_{il} = 0 | Y_i = y, G_i = g'), \tag{7}$$

where Y_i is the conditioning set, which denotes a subset of features in $(Z_i \cup \eta_i)$.[13] This can

[13] Note that one can think of EGO/EOP as (an infeasible version of) CSP if $D_i = Y_i$.

then be translated into a scalar value by, for example, aggregating over the distribution of Y_i.

Intuitively, Conditional Statistical Parity states that a decision-maker's decision is "race-blind" after taking into account all other relevant characteristics collected in Y_i: the denial rates are identical for Black and White borrowers with the same characteristics. Several papers on measuring discrimination in the U.S. mortgage market are based on variations of Conditional Statistical Parity (for example, Avery et al. [1997], Black et al. [2001], Bhutta et al. [2022], and Hurtado and Sakong [forthcoming]).

Conditional Statistical Parity is also frequently used in legal settings. How any violations are interpreted depends on the choice of the conditioning set Y_i. For example, as Ayres [2010] argues, "Disparate treatment tests [...] control for any and all variables that plausibly had a causal impact on a defendant's decision making." On the other hand, "disparate impact tests should only include controls for attributes that are plausibly business justified" (Ayres [2010]).

As the legal setting suggests, balancing these requirements makes implementing Conditional Statistical Parity challenging in practice. The first obstacle is to define the right set of controls, or conditioning set, Y_i. Clearly, different choices of Y_i will lead to different measures. In our experience, the "correct" choice of Y_i is often far from obvious; a too narrow choice of Y_i can omit economically relevant dimensions of the decision, while too wide a choice can mask discrimination.[14]

We thus present results for a variety of conditioning sets Y_i:[15]

1. $Y_i = \emptyset$ ("Unconditional", corresponding to Statistical Parity)
2. $Y_i = \{Z_i^{small}\}$ ("few covariates")
3. $Y_i = \{Z_i\}$ ("many covariates")
4. $Y_i = \{AUS_i, Z_i\}$ ("including AUS")

where $Z_i^{small} \subset Z_i$ includes loan purpose, loan amount, applicant income, and an indicator for co-applicant, and Z_i additionally includes the applicant's credit score, loan term, debt-to-income ratio, and loan-to-value ratio (which are available in HMDA starting in 2018). AUS_i denotes the recommendation from an Automated Underwriting System (AUS), typically provided by one of the government-sponsored enterprises (GSEs). The AUS provides a

[14] For instance, a lender could include a proxy that is irrelevant for the default probability of an applicant, yet correlates with race. Including this proxy in the conditioning set would result in no CSP violation. See Prince and Schwarcz [2019] for a comprehensive discussion of proxy discrimination. To give a specific example, the CFPB Examination Procedures (`https://files.consumerfinance.gov/f/201306_cfpb_laws-and-regulations_ecoa-combined-june-2013.pdf`) explicitly caution against using underwriting models that use ZIP codes, postulating a negative disparate impact on protected classes.

[15] Also see Bohren et al. [2022] for a discussion on what makes a "good" covariate in the context of measuring discrimination.

recommendation to underwriters based on a statistical default model.[16]

Second, while Conditional Statistical Parity may be easy to calculate for a regulator in simple models, it can become increasingly difficult under more sophisticated models: If the lender's decision rule is known to be linear in the elements of Y_i, and if Y_i is observed by the regulator, it simply states that the coefficient on G_i is equal to zero in a multivariate regression of L_{il} on Y_i and G_i. On the other hand, if models are highly nonlinear, e.g. derived via machine learning algorithms, and may access rich feature sets, testing Conditional Statistical Parity becomes challenging. This is exacerbated by the fact that more sophisticated models will be better at triangulating the protected class of an applicant. Intuitively, more flexible models are better able to proxy for the relationship between G_i and the default probability of an applicant using nonlinear functions of the variables in Y_i (also see Prince and Schwarcz [2019] and Fuster et al. [2022]).

We thus consider a range of model types and first define our measure of Conditional Statistical Parity maintaining additive separability in G_i and Y_i for $P(L_{il}|G_i, Y_i)$. That is, we base these measures of Conditional Statistical Parity on estimates that take the form $P(L_{il}|G_i, Y_i) = f(G_i) + h(Y_i)$. While restrictive, this reduces our measure of Conditional Statistical Parity for group g to a single number which allows for convenient visualization and interpretation of our results. In particular, our measure of Conditional Statistical Parity (with respect to reference group g'), $\delta_{CSP}(g, g', y)$, is then equal to β_g, estimated from the following regression model:

$$(1 - L_{il}) = \beta_0 + \sum_{g \in \mathcal{G} \setminus g'} \beta_g \mathbf{1}[G_i = g] + h(Y_i).$$

Our first version ("linear") imposes linearity on the variables in Y_i (i.e. $(h(Y_i) = \sum_j \alpha_j \iota_{ij})$ for $\iota_{ij} \in Y_i$), and is thus based on a simple linear regression of L_i on a group dummy and the variables in Y_i. Our second version ("nonlinear") creates 20 bins with roughly the same number of observations for each variable in Y_i and controls for dummies indicating membership in each of those bins.

Further, we add two additional specifications. First, we add the county of the property to the conditioning set of our largest nonlinear model ("including county"). While this suggests an even smaller difference between Black and White applicants, we may be concerned that including location fixed effects induces included-variable bias (Ayres [2010], Jung et al. [2018]). In fact, denying an applicant a loan for housing based on a certain neighborhood is,

[16]The most commonly used AUS in our dataset is Fannie Mae's Desktop Underwriter. While the statistical model used in the Desktop Underwriter is unknown to us, Fannie Mae publishes a list of risk factors it considers in its AUS (Fannie Mae [2023]). These include both credit report variables, such as an applicant's credit history, as well as non-credit risk factors, such as an applicant's liquid reserves and housing expense ratio. Also see Bhutta et al. [2022] for a more detailed description of Automated Underwriting Systems.

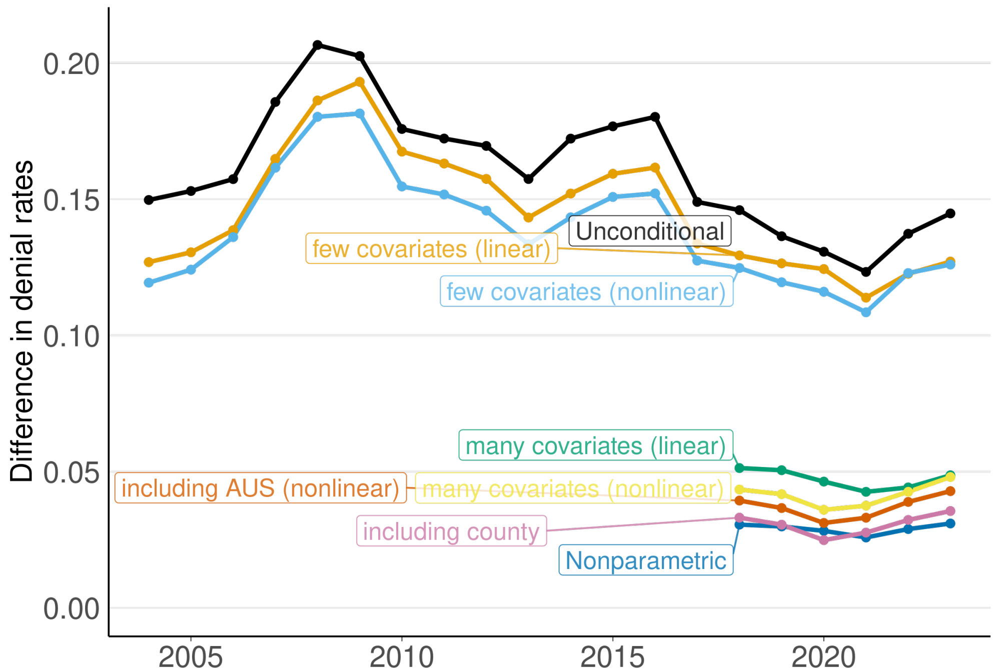


Figure 7: Conditional Statistical Parity over time for various implementations.

in general, considered *redlining* and is illegal.[17]

Second, since additive separability in G and Y_i for $P(L_{il} = 0|G_i, Y_i)$ may be overly restrictive, we also consider the following fully nonparametric specification. We first use machine learning (an HGBC model, similar to the default model introduced in Appendix B) to estimate the conditional expectation function $E(L_{il} = 0|Y_i, G_i) = g(Y_i, G_i)$.[18] This attempts to capture the underlying relationship between application outcomes and applicant features, *including race*. Then, we can define the individual effect $b_i \equiv g(Y_i, 1) - g(Y_i, 0)$ to capture the individual-specific impact of being a minority applicant on the denial probability. In other words, it captures the change in probability of denial for applicant i if she belonged to a different group. Once we have the individual specific coefficients b_i, we average them across the entire population to yield a nonparametric estimate of β_g ("nonparametric").

[17]However, a bank may under certain conditions consider such economic factors as the condition, use, or design of nearby properties, the availability of neighborhood amenities or city services, and the need of the lender to hold a balanced real estate loan portfolio, with a reasonable distribution of loans among various neighborhoods (see the Federal Fair Lending Regulations and Statutes available at `https://www.federalreserve.gov/boarddocs/supmanual/cch/fair_lend_fhact.pdf`).

[18]Specifically, we train a HGBC model with monotonicity constraints identical to those in the default model (see Appendix B) using HMDA data to predict denial with the following covariates: state of the property for which the application is filed, loan purpose, applicant income, DTI, LTV, credit score, loan amount, whether a coapplicant is present on the application, the loan term in months, and an indicator for whether an applicant is Black.

Figure 7 depicts the difference in denial rates between Black and White applicants after conditioning on the aforementioned sets and across our models of varying complexity. Each line corresponds to a different specification in terms of both which covariates are included and how flexible a model we allow. While all measures point to a disproportionate rate of denials for Black applicants, the difference in denial rate between Black and White applicants tends to decline as we increase the size of the conditioning set Y_i. Similarly, more flexible functional forms tend to reduce the size of our estimate, with our nonparametric specification suggesting the smallest deviation from Conditional Statistical Parity between White and Black applicants.

We conclude that the choice of both the conditioning set and the functional form assumptions are extremely important and can lead to vastly different results when calculating violations of Conditional Statistical Parity (in our case, the measures based on alternative specifications differ by a factor of more than six). While we find that all of our specifications maintain a racial gap in denial rates, this gap becomes smaller for increasingly rich conditioning sets. On the one hand, this could be consistent with a "race-blind" decision maker that has access to additional attributes in η_i that are correlated with group membership (e.g. following similar arguments as in Altonji et al. [2005] and Oster [2019]). On the other hand, richer conditioning sets may be more likely to include covariates that are themselves racially biased, illegal to use, decision-irrelevant proxies, or otherwise inappropriate controls. Thus, the ultimate significance of the declining gap is not clear. Taken together, our estimates of Conditional Statistical Parity are suggestive of some disparity. However, their range makes it difficult to assess magnitude and economic significance, and the large degree of sensitivity of this measure to its particulars is a weakness of this approach.

3.6 Representativeness

Finally, we consider a notion of fairness called Representativeness ("RP", Ross and Yinger [2002]). Intuitively, it corresponds to the idea that the population of approved candidates should be "representative" of the qualified candidates. Conceptually, a decision process would be representative if:

$$P(G_i = g|L_{il} = 1) = P(G_i = g|E(D_{il} = 1|Z_i, \eta_i) < c). \tag{8}$$

In practice, since η is unobserved, it is of course infeasible to compute the right-hand-side term in (8). We thus replace $E(D_{il} = 1|Z_i, \eta_i) < c$ with a feasible counterpart, $E(D_{il}|Y_i) < c$, and implement our measure of Representativeness as follows.

First, for a chosen conditioning set Y_i, we rank all applicants according to their default risk, $E(D_{il}|Y_i)$.[19] Denote by n_a the number of approved loans in the data. We then compare the

[19] To construct $E(D_{il}|Y_i)$ in practice, we estimate default probabilities of applications submitted in year t from an HGBC model trained on data from year $t-3$. This training data comprises mortgages originated in year $t-3$ along with the subsequent two years of loan performance in which we count defaults. The covariates include credit score, LTV, DTI, original loan amount, loan type (conventional, VA, etc.), an indicator for

group composition of the n_a applicants with the lowest estimated default risk (the "qualified" applicants) to the group composition of the approved applicants. This means we calculate any deviations from Representativeness as

$$\delta_{RP}(g) := P(G_i = g|E(D_{il}|Y_i) < \hat{c}) - P(G_i = g|L_{il} = 1), \tag{9}$$

where Y_i denotes a conditioning set in the estimation of default risk and $\hat{c}$ is defined as

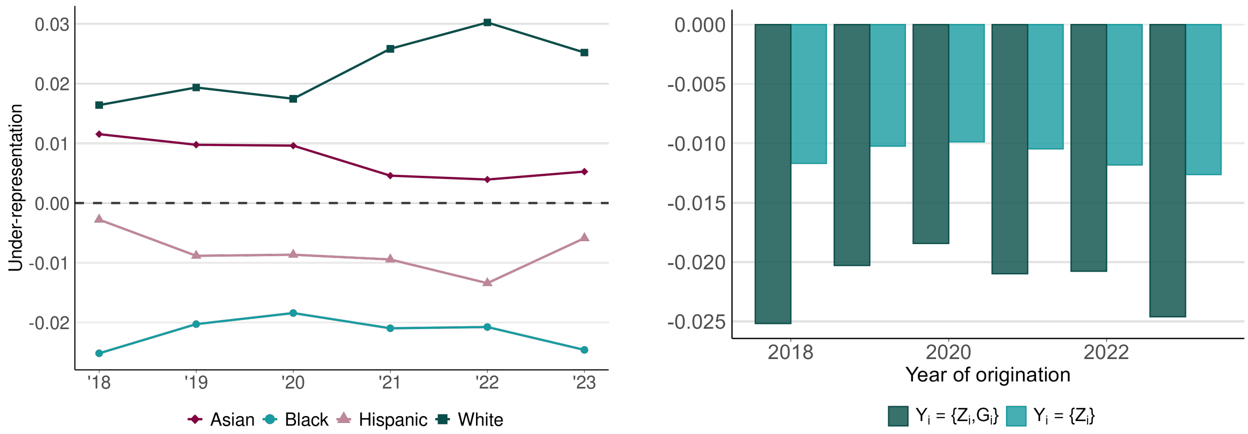
$$\hat{c} = \min c \qquad \text{such that} \sum_i 1[E(D_{il}|Y_i) \leq c] \geq \sum_i 1[L_{il} = 1] \equiv n_a.$$

This is the minimum c, or default risk threshold, such that the number of applicants predicted to be less risky than c is greater than or equal to the number of approved applicants, n_a. Representativeness has the following intuitive interpretation. If $\delta_{RP}(g) > 0$ for $G_i = g$, this means that individuals in group g are accepted less frequently in the data than they "deserve based on their merit" (i.e., riskier individuals from one group are accepted at the expense of less risky applicants from another group). Thus, a positive value indicates that group g is *under-represented* among approved borrowers, while a negative value indicates over-representation of group g.

We note that the choice of the conditioning set Y_i is again crucial. We start by following Ross and Yinger [2002] and use $Y_i = \{Z_i, G_i\}$, in which case $E(D_{il}|Z_i, G_i)$ becomes the expected default probability conditional on all observed variables, *including group membership* G. Figure 8a depicts our measure of Representativeness for the four demographic groups using mortgage applications submitted across the United States between 2018 and 2023. We observe persistent differences in Representativeness across groups. In particular, this measure suggests that minorities (in particular Black applicants) are slightly over-represented among approved applicants relative to their model-implied riskiness.

This is again suggestive that approved Black borrowers on the margin are slightly riskier than approved White borrowers on the margin, consistent with our finding based on the Marginal Outcome Test. In Figure 8b we illustrate our measure of Representativeness for Black applicants with and without including G_i in Y_i. Dropping G_i, the difference between the actual approval rate and a hypothetical approval rate based on $E(D_{il}|Y_i)$ becomes smaller. This suggests that being Black is associated with a higher default probability in the default model that includes G_i. In other words, a default model that does not have access to group membership will result in lower predicted default probabilities for minority applicants. In turn, this implies that Black borrowers are less overrepresented when using $Y_i = \{Z_i\}$. Note that, since in both specifications we find that Black applicants are overrepresented, our fairness violation is "negative" using this definition.

whether a coapplicant is present, the state of the property's location, and loan term in months, among others. See Appendix B for the full specification and more details.



(a) Across demographic groups, using $Y_i = \{Z_i, G_i\}$ as the conditioning set.

(b) For Black applicants, with and without G_i in the conditioning set.

Figure 8: Representativeness for mortgages originated in 2018-2023.

4 Stylized Facts

We next summarize and combine the six measures we introduced in the previous section to describe five stylized facts that highlight recent trends and geographic patterns in fairness and inequality in the U.S. mortgage market.

Stylized Fact 1: Broad measures suggest stark systemic inequality. We find strong evidence of systematically worse outcomes for Black Americans. Black applicants are denied at a significantly higher rate, restricting their access to home ownership and thus wealth accumulation. At the same time, Black borrowers default at higher rates, thus disproportinally bearing the associated costs (for a discussion of the cost of bankruptcy, see Argyle et al. [2023]). This paints a picture of a society with large racial disparities in financial well-being and economic opportunity.

This holds true across the United States. Figure 9 presents the violations of Statistical Parity and Predictive Parity between Black and White Americans across states. Each state corresponds to one of the 51 observations depicted (50 states plus DC). We have also labeled the most populous state in each of the four census regions. The dashed lines correspond to equal outcomes in both dimensions. It is evident that Black applicants in all states face higher denial rates than White applicants. Similarly, Black borrowers have higher default rates in all states. Furthermore, there is a pronounced positive correlation across states between racial gaps in denial rates and default rates.

While this paints a stark picture of inequality in the mortgage sector, which may perpetuate existing inequalities in wealth and economic well-being, these differences are not necessarily reflective of explicit discrimination in the mortgage sector. They may also reflect existing

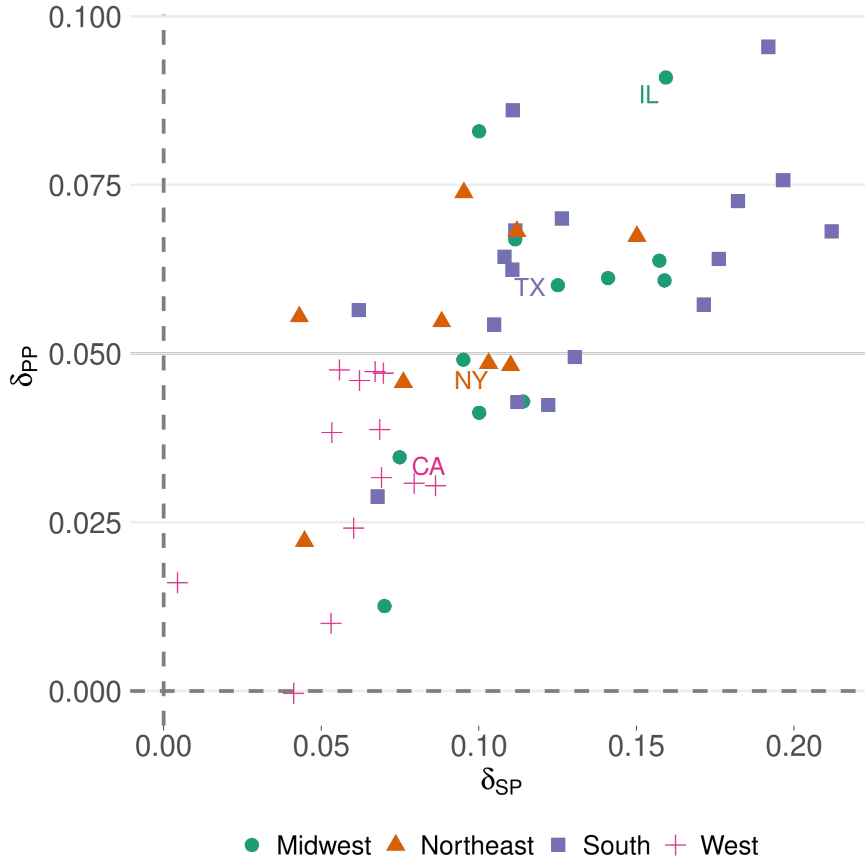
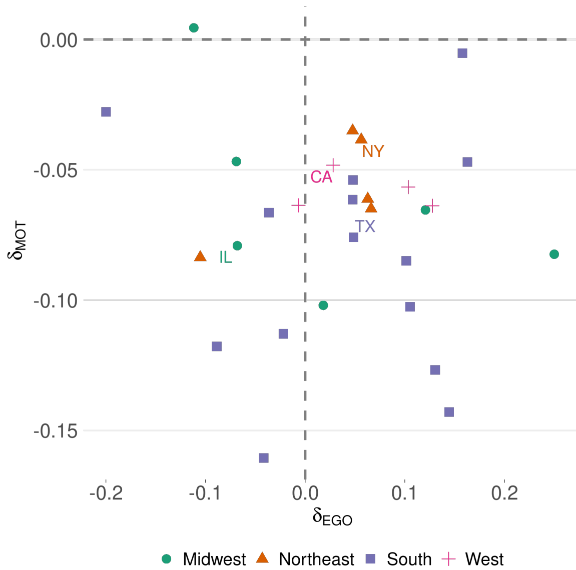


Figure 9: Fairness violations at the state level as measured by Statistical Parity (δ_{SP}) and Predictive Parity (δ_{PP}) based on applications filed in 2018-2021.

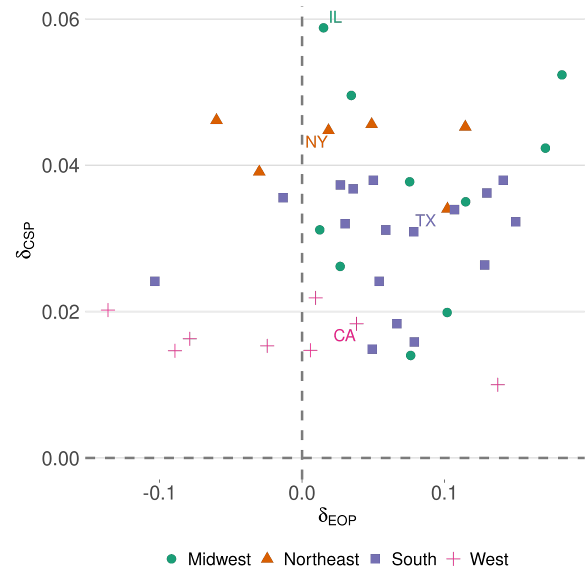
systemic and historical inequalities in the United States between Black and White Americans.

Stylized Fact 2: Decision-driven fairness measures provide mixed evidence. More narrow definitions of fairness that try to isolate the decision of the lender on mortgage applications yield a nuanced picture. We illustrate this at the state-level in Figure 10.

First, to isolate the decision of the lender, we compare default rates at the decision boundary by restricting to a sample of marginal borrowers. We find that marginal Black borrowers are more likely to default than marginal White borrowers. Since the higher default rate for minority applicants implies a slightly *lower* lending standard for minority applicants, our fairness violation based on the Marginal Outcome Test is negative for Black borrowers in all but one state. This is reflected by the fact that all but one state in Figure 10a are below the dashed line indicating equality. We note that this is further in line with our results for Representativeness (cf. Section 3.6), which suggest that minority applicants are slightly *overrepresented* relative to their default risk.



(a) Marginal Outcome Test (δ_{MOT}) and Equality of Goodwill (δ_{EGO})



(b) Conditional Statistical Parity (δ_{CSP}) and Equality of Opportunity (δ_{EOP})

Figure 10: Fairness violations at the state level comparing "more narrow" fairness measures based on applications filed in 2018-2021.

Second, we find violations of Equality of Goodwill are approximately equal to zero nationally, and distributed around zero at the state level, suggesting that there is no systematic fairness violation in aggregate using this definition of fairness.[20] In other words, among applicants who submitted multiple applications and ultimately defaulted, Black applicants were about as likely as White applicants to obtain approvals on all their loans.

Third, in Figure 10b we observe mostly positive violations of Equality of Opportunity across the depicted states (restricting our sample to those 41 states with at least five Black cross-applicants between 2018-2021 who did not default), indicating that Black applicants who did not default were denied more frequently than White applicants who did not default. This suggests that for creditworthy Black applicants it is still harder to obtain credit than for creditworthy White applicants in most states. Similarly, in Figure 10b we observe positive violations of Conditional Statistical Parity in all depicted states: Black applicants are more likely to be denied in their loan applications even conditional on a rich set of covariates.[21]

These seemingly contradictory patterns are perhaps not surprising, given existing impossibility results (e.g. Kleinberg et al. [2017]). On balance, they provide mixed evidence of unfairness. They do, however, emphasize that the choice of fairness definition matters, and

[20]Note that Figure 10a includes only 29 states since we restrict our sample to the 29 states with at least five Black crossapplicants between 2018-2021 who defaulted.

[21]Here, we depict the specification "including AUS (nonlinear)". The results are qualitatively similar for our alternative specifications.

show how different definitions can lead to contrasting results. It is also suggestive that a regulator indeed faces inherent trade-offs when deciding which notion of fairness to measure, aim for, or enforce (Kleinberg et al. [2017], Kleinberg et al. [2020]).

Stylized Fact 3: HMDA is inconsistent with a decision rule based on default risk. It is instructive to consider a hypothetical lender who makes *merit-based* lending decisions.[22] To do so, we consider a profit-maximizing risk-neutral lender that does not exhibit any taste-based discrimination. Specifically, a merit-based decision maker (or lender l) applies the following decision rule:

$$P(L_{il} = 1) = \begin{cases} 1 & \text{if } E(D_{il}|Z_i, \eta_i) < c \\ 0.5 & \text{if } E(D_{il}|Z_i, \eta_i) = c \\ 0 & \text{if } E(D_{il}|Z_i, \eta_i) > c, \end{cases} \tag{10}$$

such that applicants with a default probability lower than c are approved, applicants with a default probability higher than c are rejected, and approval is random right at the cutoff. In particular, this decision rule does not take into account an applicant's group membership, but is solely based on an individual's "merit."[23]

We again stress that, by allowing for the presence of η_i in the decision rule, we do not assume we observe all the loan and applicant characteristics that the lender considers. Under this theoretical framework, in Appendix A we derive whether each measure is satisfied in general under a merit-based decision maker. Table 2 contrasts these theoretical predictions with our empirical findings. The left column of Table 2 summarizes how data generated by a merit-based decision maker would manifest itself in terms of violation of our fairness measures. The right column presents the empirical violations of our fairness measures in our data. Our main conclusion is that the data is not consistent with a merit-based decision maker.

In particular, the Marginal Outcome Test is expected to be satisfied under a merit-based decision maker while it is not satisfied empirically in our data. In fact, as presented in the previous section, minorities tend to have higher default rates at the decision boundary in the data. One possible explanation is that Assumption 1 may be violated. In other words, the default model that lenders use (which may include η_i) may also be miscalibrated, which means that lenders may underestimate the default risk of minority borrowers. On the other hand, Equality of Goodwill is (approximately) satisfied in our data, which will generally not be the case under a merit-based decision maker.

[22] We use the term *merit-based* narrowly to refer to a decision-making process based solely on expected default risk. This definition does not encompass broader notions of merit that might incorporate other attributes (e.g., terms of the contract and the applicant's ability to repay beyond default risk.)

[23] See Kasy and Abebe [2021] for a discussion of why a merit-based decision maker may be deemed normatively undesirable.

	Merit-based decision maker	Empirical results
Statistical Parity	Violated	Violated
Predictive Parity	Violated	Violated
Marginal Outcome Test	**Satisfied**	Violated
Equalized Odds: Equality of Opportunity	Violated	Violated
Equalized Odds: Equality of Goodwill	Violated	**Satisfied**
Conditional Statistical Parity	Violated	Violated
Representativeness	Violated	Violated

Table 2: Fairness violations implied by a merit-based decision maker, and as observed in our data. See Appendix A for more detail.

Stylized Fact 4: Common trends. We next ask to what extent our six measures are correlated. For this exercise, we consider Statistical Parity, Predictive Parity, Conditional Statistical Parity, the Marginal Outcome Test, Representativeness, and Equality of Opportunity.[24] Since we observe some of our measures starting only in 2018, we look at state-level correlation and calculate each measure for the pooled sample across 2018-2021.

First, columns 1-6 of Table 3 present the correlation among fairness violation measures. Intuitively, Table 3 suggests that we can classify our measures into two distinct groups. The first group is formed by Statistical Parity, Predictive Parity, Equality of Opportunity, and Conditional Statistical Parity. The second group consists of the Marginal Outcome Test and Representativeness. Measures within a group are positively correlated, while measures between groups are negatively correlated. This suggests that there is no clear ordering of the states in terms of their overall fairness. Another way to see this: There is no state that is ranked in the top 10 by all six measures, and there is no state that is ranked in the bottom 10 by all six measures.

We also see this clustering pattern reflected in the results from principal component analysis. The first principal component explains 52% of the variation in fairness violations across states, is strongly negatively correlated with Statistical Parity, Predictive Parity, Equality of Opportunity, and Conditional Statistical Parity, and is strongly positively correlated with the Marginal Outcome Test and Representativeness - in line with the group structure a visual inspection of the correlation matrix suggests. On the other hand, the second principal component (which explains another 19% of the total variation) is positively correlated with five of the six fairness violations (and almost uncorrelated with Predictive Parity), and thus captures a tendency that states with high violations in one measure tend to also have high violations in other measures.

[24] Equality of Goodwill is excluded because of sample limitations at the state level. Restricting our sample to those states with at least five Black marginal applicants and at least five Black cross-applicants who did not default, we are left with 40 states to analyze in Stylized Facts 4 and 5.

	SP	PP	MOT	EOP	CSP	RP	PC1	PC2
SP	1						-0.85	0.24
PP	0.66	1					-0.88	-0.09
MOT	-0.41	-0.55	1				0.66	0.6
EOP	0.47	0.24	0.09	1			-0.4	0.81
CSP	0.52	0.58	-0.28	0.21	1		-0.66	0.17
RP	-0.58	-0.64	0.57	-0.23	-0.24	1	0.78	0.21

Table 3: Correlation among *violations* of fairness measures and first two principal components. PC1 and PC2 denote the first two principal components. Recall that all fairness violations are defined such that larger numbers correspond to worse outcomes for Black applicants or borrowers.

Stylized Fact 5: Geographic patterns. Finally, we illustrate the geographic heterogeneity across the United States in more detail. First, we project the six measures into the space spanned by the first two principal components. Figure 11 suggests a clustering of these states by region. All Western states (pink crosses) cluster on the right-hand side of the figure, reflecting positive loadings on the first principal component. This suggests that states in the West tend to have smaller violations of Statistical Parity, Predictive Parity, Conditional Statistical Parity, and Equality of Opportunity, and larger violations of Representativeness and the Marginal Outcome Test. With the exception of New Mexico, the Western states also tend to have negative loadings on the second principal component, which suggests smaller fairness violations across the board. On the other hand, the Midwestern states (green circles) have mostly positive loadings on the second principal component, which suggests larger fairness violations overall.

Figure 12 attempts a partial overall ranking of states. We say that state A "strictly dominates" state B if state A has lower fairness violations than state B for all six measures. In other words, no matter what fairness measure (or combination of measures) one chooses, state A is more fair than state B according to this measure. In Figure 12, we depict for each state the number of states it strictly dominates. For example, California (with a value of four) has a smaller fairness violation according to all six measures than four states: Kansas, Missouri, Nebraska and Oklahoma. The map suggests that the Midwestern states in particular tend to be strictly dominated by other states, with Illinois, Nebraska, Missouri, Kansas and Wisconsin being strictly dominated by at least one other state. In fact, Missouri has worse outcomes in all six measures than 10 other states.

This geographical pattern is further underlined if we rank all states according to each of the fairness measures, and then average these ranks. While the top six states (those with the lowest overall fairness violations according to this metric) fall into the Western region, eight of the nine Midwestern states included are ranked in the bottom 15 states.

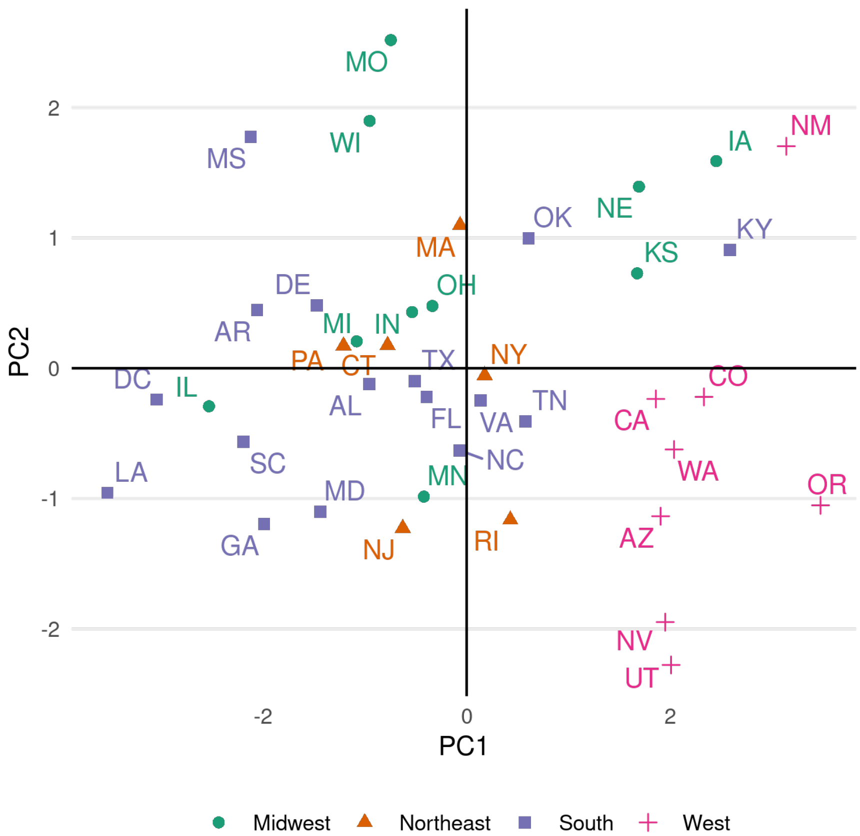


Figure 11: Fairness violations of 40 states over 2018-2021, projected into the space of the first two principal components.

Interactive data visualization. While we have presented five Stylized Facts that highlight recent trends in fairness and inequality in the mortgage market, we are able to highlight only a small subset of interesting results in this section. An online appendix in the form of an interactive data visualization at `https://www.philadelphiafed.org/surveys-and-data/consumer-finance-data/mortgage-fairness-explorer` allows the reader to explore our results further across both time and space. Not only does it include all of our measures for Hispanic and Asian borrowers, it also extends our analysis to gender as a second protected group. A reader can visualize time series trends at the state level, geographic maps based on individual measures, and comparisons of multiple measures in one figure. It will be updated periodically.

5 Conclusion

The first question we set out to answer was whether it matters what definition of fairness one uses when assessing the fairness of the U.S. mortgage market. To answer this question, we

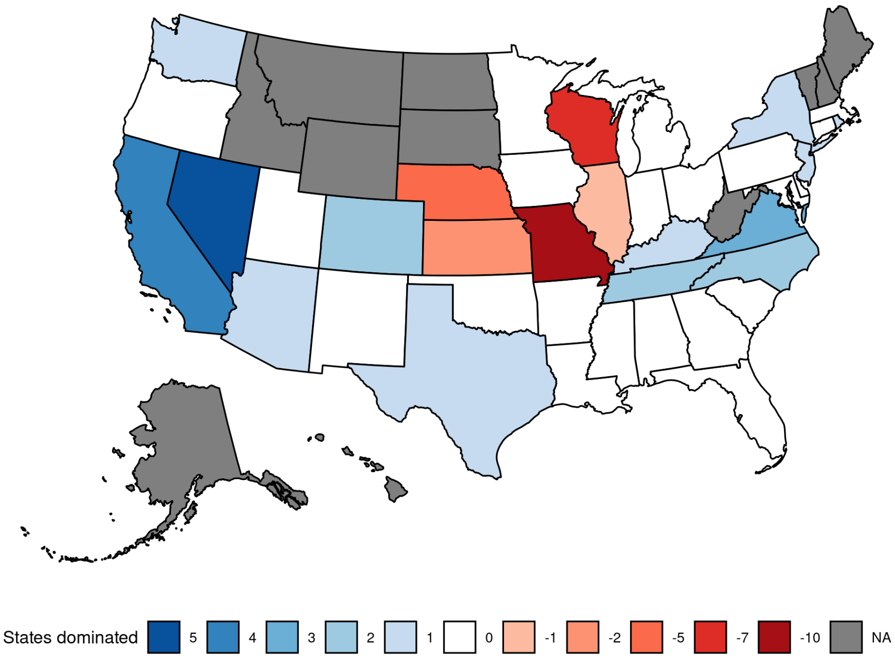


Figure 12: Geographic visualization of how many other states a given geographic state strictly dominates across all measures. For a given state, positive numbers indicate the number of states with higher fairness violations across all six measures. Negative values indicate the number of states with lower fairness violations across all six measures. All numbers exclude the eleven grey regions with insufficient data.

considered a wide range of fairness definitions stemming from the economics and computer science literature, law and regulatory guidance, as well as public debate. We find strong evidence that the definition of fairness indeed matters. Different fairness definitions will lead to very different conclusions. This has important policy implications: depending on the context, policymakers and regulators should carefully decide on the appropriate definition of fairness to be used, as this choice will be important in shaping policy decisions.

The second question we set to answer was how fair or unfair the outcome of the U.S. mortgage market looks. In light of the answer to our first question above, we do not find a conclusive answer. We find strong evidence of systematically worse outcomes for Black Americans. We are, however, unable to conclude whether these disparities are due to unfair decisions by the lenders, as different fairness definitions lead to different conclusions. One implication from this is that any one definition (or study) may be misleading, and a comprehensive analysis might include several competing definitions.

We hope that our work contributes to a more nuanced understanding of how we measure and regulate fairness, both in the mortgage market and more generally.

Our findings open up a number of promising avenues for future research. We document large differences in our fairness measures across both time and space in the United States. Exploring the drivers of these differences could provide valuable insights into the causal factors that influence fairness in this market. It would also be interesting to study how technological advancements, such as AI-driven lending decisions, interact with our various fairness measures. Finally, closely related to the question of how to measure fairness is the question of how to regulate fairness, and investigating the long-term economic and social impacts of different fairness regulations would be a compelling topic for future work.

References

Uniform guidelines on employee selection procedures. 43 FR 38290, 1978. Federal Register.

Sumit Agarwal, Efraim Benmelech, Nittai Bergman, and Amit Seru. Did the Community Reinvestment Act (CRA) lead to risky lending? Working paper, National Bureau of Economic Research, 2012.

Stefania Albanesi and Domonkos F Vamossy. Predicting consumer default: A deep learning approach. Working paper, National Bureau of Economic Research, 2019.

Joseph G. Altonji, Todd E. Elder, and Christopher R. Taber. Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy*, 113(1):151–184, 2005.

Brent W Ambrose, James N Conklin, and Luis A Lopez. Does borrower and broker race affect the cost of mortgage credit? *The Review of Financial Studies*, 34(2):790–826, 2021.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. And it's biased against Blacks. *ProPublica*, 2016.

Shamena Anwar and Hanming Fang. Testing for racial prejudice in the parole board release process: Theory and evidence. *The Journal of Legal Studies*, 44(1):1–37, 2015.

Bronson Argyle, Sasha Indarte, Benjamin Iverson, and Christopher Palmer. Explaining racial disparities in personal bankruptcy outcomes. Working paper, 2023.

David Arnold, Will Dobbie, and Crystal S Yang. Racial bias in bail decisions. *The Quarterly Journal of Economics*, 133(4):1885–1932, 2018.

R.B. Avery, P.E. Beeson, and P.S. Calem. Using HMDA data as a regulatory screen for fair lending compliance. *Journal of Financial Services Research*, 11:9–42, 1997.

Ian Ayres. Testing for discrimination and the problem of "included variable bias". *Yale Law School Mimeo*, 2010.

Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. Consumer-lending discrimination in the fintech era. *Journal of Financial Economics*, 143(1):30–56, 2022.

Patrick Bayer, Fernando Ferreira, and Stephen L Ross. The vulnerability of minority homeowners in the housing boom and bust. *American Economic Journal: Economic Policy*, 8 (1):1–27, 2016.

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

Gary S Becker. *The economics of discrimination*. University of Chicago Press, 1957.

James A Berkovec, Glenn B Canner, Stuart A Gabriel, and Timothy H Hannan. Discrimination, competition, and loan performance in FHA mortgage lending. *Review of Economics and Statistics*, 80(2):241–250, 1998.

Neil Bhutta and Aurel Hizmo. Do minorities pay more for mortgages? *The Review of Financial Studies*, 34(2):763–789, 2021.

Neil Bhutta, Steven Laufer, and Daniel Ringo. Residential mortgage lending in 2016: Evidence from the Home Mortgage Disclosure Act data. Technical report, Board of Governors of the Federal Reserve System, 2017.

Neil Bhutta, Aurel Hizmo, and Daniel Ringo. How much does racial bias affect mortgage lending? Evidence from human and algorithmic credit decisions. FEDS working paper, 2022.

H.A. Black, B.L. Robinson, and R.L. Schweitzer. Comparing lending decisions of minority-owned and White-owned banks: Is there discrimination in mortgage lending? *Review of Financial Economics*, 10:23–39, 2001.

Harold Black, Robert L Schweitzer, and Lewis Mandell. Discrimination in mortgage lending. *The American Economic Review*, 68(2):186–191, 1978.

Interagency Fair Lending Examination Procedures. Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency, Federal Deposit Insurance Corporation, Consumer Financial Protection Bureau, 2009. URL `https://www.ffiec.gov/pdf/fairlend.pdf`.

J Aislinn Bohren, Peter Hull, and Alex Imas. Systemic discrimination: Theory and measurement. Working paper, National Bureau of Economic Research, 2022.

Ryan Bubb and Alex Kaufman. Securitization and moral hazard: Evidence from credit score cutoff rules. *Journal of Monetary Economics*, 63:1–18, 2014.

Kerwin Kofi Charles and Erik Hurst. The correlation of wealth across generations. *Journal of Political Economy*, 111(6):1155–1182, 2003.

Ping Cheng, Zhenguo Lin, and Yingchun Liu. Racial discrepancy in mortgage interest rates. *The Journal of Real Estate Finance and Economics*, 51:101–120, 2015.

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

Yuliya Demyanyk. Your credit score is a ranking, not a score. *Economic Commentary*, (2010-16), 2010.

Will Dobbie, Andres Liberman, Daniel Paravisini, and Vikram Pathania. Measuring bias in consumer lending. *The Review of Economic Studies*, 88(6):2799–2832, 2021.

Cynthia Dwork and Christina Ilvento. Fairness under composition. *10th Innovations in Theoretical Computer Science*, 2019.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

Hadi Elzayn, Simon Freyaldenhoven, and Minchul Shin. Constructing applicants from loan-level data: A case study of mortgage applications. Working paper, 2024.

Fannie Mae. Fannie Mae Selling Guide B3-2-03, Risk Factors Evaluated by DU. `https://selling-guide.fanniemae.com/sel/b3-2-03/risk-factors-evaluated-du`, 2023. Accessed: 2023-02-22.

Federal Reserve Bank of New York. Quarterly Report on Household Debt and Credit 2022:Q4. `https://www.newyorkfed.org/medialibrary/interactives/householdcredit/data/pdf/HHDC_2022Q4`, 2023. Accessed: 2023-02-28.

W Scott Frame, Ruidi Huang, Erik J Mayer, and Adi Sunderam. The impact of minority representation at mortgage lenders. *SMU Cox School of Business Research Paper*, (22-08), 2021.

Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance*, 77(1):5–47, 2022.

Kristopher Gerardi and Paul Willen. Subprime mortgages, foreclosures, and urban neighborhoods. *The BE Journal of Economic Analysis & Policy*, 9(3), 2009.

Marco Giacoletti, Rawley Heimer, and Edison G Yu. Using high-frequency evaluations to estimate discrimination: Evidence from mortgage loan officers. Working paper, Philadelphia Federal Reserve, 2022.

Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Agustin Hurtado and Jung Sakong. Racial disparities in the US mortgage market. In *AEA Papers and Proceedings*. American Economic Association, forthcoming.

Jongbin Jung, Sam Corbett-Davies, Ravi Shroff, and Sharad Goel. Omitted and included variable bias in tests for disparate impact. *arXiv preprint arXiv:1809.05651*, 2018.

Maximilian Kasy and Rediet Abebe. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 576–586, 2021.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3149–3157, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Amir Kermani and Francis Wong. Racial disparities in housing returns. Technical report, National Bureau of Economic Research, 2021.

Sunwoong Kim and Gregory D Squires. Lender characteristics and racial disparities in mortgage lending. *Journal of Housing Research*, pages 99–113, 1995.

J. Kleinberg, J. Ludwig, S. Mullainathan, and Ashesh Rambachan. An economic approach to regulating algorithms. Working paper, National Bureau of Economic Research, 2020.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference*, volume 67, pages 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017.

Moritz Kuhn, Moritz Schularick, and Ulrike I Steins. Income and wealth inequality in America, 1949–2016. *Journal of Political Economy*, 128(9):3469–3519, 2020.

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

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284, 2017.

Phillip Li and Tom Mayock. Mortgage characteristics and the racial incidence of default. *Journal of Housing Economics*, 46:101655, 2019.

Paul C Lubin. Fair lending testing: Best practices, trends and training. *Joint Centre for Housing Studies, Harvard University*, 2008.

Vitaly Meursault, Daniel Moulton, Larry Santucci, and Nathan Schor. One threshold doesn't fit all: Tailoring machine learning predictions of consumer default for lower-income areas. *Journal of Policy Analysis and Management*, forthcoming.

Alicia H Munnell, Geoffrey MB Tootell, Lynn E Browne, and James McEneaney. Mortgage lending in Boston: interpreting HMDA data. *American Economic Review*, pages 25–53, 1996.

Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proceedings Conference on Fairness, Accountability, and Transparency*, volume 1170, page 3, New York, NY, 2018.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632, 2005.

Supreme Court of the United States. Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc., 2015. URL `https://www.justice.gov/sites/default/files/crt/legacy/2015/06/25/tdhcainclusiveopinion.pdf`. 135 S. Ct. 2507.

Emily Oster. Unobservable selection and coefficient stability: Theory and validation. *Journal of Business Economics and Statistics*, 37(2):187–204, 2019.

Anya ER Prince and Daniel Schwarcz. Proxy discrimination in the age of artificial intelligence and big data. *Iowa Law Review*, 105:1257, 2019.

Stephen L Ross. Mortgage lending discrimination and racial differences in loan default: A simulation approach. *Journal of Housing Research*, pages 277–297, 1997.

Stephen L Ross and John Yinger. *The color of credit: Mortgage discrimination, research methodology, and fair-lending enforcement*. MIT Press, 2002.

Stephen L Ross, Margery Austin Turner, Erin Godfrey, and Robin R Smith. Mortgage lending in Chicago and Los Angeles: A paired testing study of the pre-application process. *Journal of Urban Economics*, 63(3):902–919, 2008.

Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

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

United States Congress. Civil Rights Act of 1964. U.S. Code Congressional and Administrative News, 1964. Public Law 88-352, 78 Stat. 241.

United States Congress. Civil Rights Act of 1968. U.S. Statutes at Large, 1968. URL https://www.govinfo.gov/content/pkg/STATUTE-82/pdf/STATUTE-82-Pg73.pdf. Public Law 90-284.

United States Congress. Equal Employment Opportunity Act of 1972. U.S. Code Congressional and Administrative News, 1972. Public Law 92-261, 86 Stat. 103.

United States Congress. Fair Housing Act (FHAct), title VIII of the Civil Rights Act of 1968, as amended (42 USC 3601 et seq.). Federal Fair Lending Regulations and Statutes, Consumer Compliance Handbook, 2006. URL https://www.federalreserve.gov/boarddocs/supmanual/cch/fair_lend_fhact.pdf.

Paul Willen and David Hao Zhang. Do lenders still discriminate? A robust approach for assessing differences in menus. Working paper, National Bureau of Economic Research, 2020.

John Yinger. Acts of discrimination: Evidence from the 1989 housing discrimination study. *Journal of Housing Economics*, 1(4):318–346, 1991.

John Yinger. Why default rates cannot shed light on mortgage discrimination. *Cityscape*, pages 25–31, 1996.

A Merit-Based Decision Maker

Definition 1. We call a decision maker (lender) l a **merit-based decision maker** if she applies the following decision rule:

$$P(L_{il}=1)=\begin{cases}1 & \text{if } E(D_{il}|Z_i,\eta_i)<c\\ 0.5 & \text{if } E(D_{il}|Z_i,\eta_i)=c\\ 0 & \text{if } E(D_{il}|Z_i,\eta_i)>c,\end{cases} \tag{11}$$

such that applicants with a default probability lower than c are approved, applicants with a default probability higher than c are rejected, and there is some randomness right at the cutoff. In particular, this decision rule does not take into account an applicant's group membership, but is solely based on an individual's "merit."

Under such a merit-based decision maker, and in the presence of characteristics η_i that are observed to the decision maker but unobserved to the regulator, we then ask whether our varying measures of fairness are satisfied or violated.

We will illustrate any violations by means of a simple counterexample. In this example there exist two potential realizations for Z_i and η_i. Table 4 displays the conditional probabilities of (Z_i,η_i) conditional on G_i and the conditional probabilities of default. Throughout this section, we will omit the subscripts i and l whenever they are not crucial to the main argument.

Table 4: A simple example

		Conditional probability		Conditional default probability
Z	η	$P(Z,\eta\|G=0)$	$P(Z,\eta\|G=1)$	$P(D=1\|Z,\eta)$
1	1	0	1/3	2/3
1	2	1/3	1/3	1/3
2	1	1/3	1/6	1/3
2	2	1/3	1/6	0

Let us introduce an auxiliary random variable ϕ which takes 1 with probability 1/2 otherwise 0. This random variable is independent to all other variables. When $\phi=1$, the lender accepts the application with $E(D|Z,\eta)=c$.

A.1 Statistical Parity

Statistical Parity will generally not be satisfied under a merit-based decision maker. To see this, we first define the set $\Phi^c=\{(z,\eta*) : E(D|Z=z,\eta=\eta^*)<c\}$ and $\Psi^c=\{(z,\eta*) :$

$E(D|Z = z, \eta = \eta^*) = c\}$. Then, we have

$$P(L = 1|G = g) = P((Z, \eta) \in \Phi^c|G = g) + \frac{1}{2}P((Z, \eta) \in \Psi^c|G = g)$$

But if the joint distribution of (Z, η) depends on G, $P((Z, \eta) \in \Phi^c|G = g)$ and $P((Z, \eta) \in \Psi^c|G = g]$ may not be equal to $P((Z, \eta) \in \Phi^c|G = g')$ and $P((Z, \eta) \in \Psi^c|G = g')$, respectively. Hence, $P(L = 0|G = g)$ and $P(L = 0|G = g')$ may differ, and Unconditional Statistical Parity will not hold in general.

As a specific counterexample, take the data generating process from Table 4. Let $c = 1/3$, then

$$\begin{aligned}
P(L = 1|G = 0) = &\ P((Z, \eta) = (1, 2), \phi = 1|G = 0) \\
&+ P((Z, \eta) = (2, 1), \phi = 1|G = 0) \\
&+ P((Z, \eta) = (2, 2)|G = 0) \\
= &\ 1/3 \cdot 1/2 + 1/3 \cdot 1/2 + 1/3 = 2/3.
\end{aligned}$$

while

$$\begin{aligned}
P(L = 1|G = 1) = &\ P((Z, \eta) = (1, 2), \phi = 1|G = 1) \\
&+ P((Z, \eta) = (2, 1), \phi = 1|G = 1) \\
&+ P((Z, \eta) = (2, 2)|G = 1) \\
= &\ 1/3 \cdot 1/2 + 1/6 \cdot 1/2 + 1/6 = 5/12.
\end{aligned}$$

A.2 Predictive Parity

Predictive Parity will generally not be satisfied under a merit-based decision maker. The general argument closely follows the one above, and we thus illustrate this using the example introduced at the beginning of the section. We again set c to $1/3$.

Then,

$$\begin{aligned}
P(D = 1|L = 1, G = 0) &= \frac{P(D = 1, L = 1|G = 0)}{P(L = 1|G = 0)} \\
&= \frac{1/9}{2/3} = 1/6,
\end{aligned}$$

because

$$\begin{aligned}
&P(D=1, L=1|G=0)\\
&= P(D=1, (Z,\eta)=(1,2), \phi=1|G=0)\\
&\quad + P(D=1, (Z,\eta)=(2,1), \phi=1|G=0)\\
&\quad + P(D=1, (Z,\eta)=(2,2)|G=0)\\
&= P(D=1|(Z,\eta)=(1,2)|G=0)P((Z,\eta)=(1,2), G=0, \phi=1)P(\phi=1)\\
&\quad + P(D=1|(Z,\eta)=(2,1), G=0)P((Z,\eta)=(2,1)|G=0)P(\phi=1)\\
&\quad + P(D=1|(Z,\eta)=(2,2), G=0)P((Z,\eta)=(2,2)|G=0)\\
&= 1/3 \cdot 1/3 \cdot 1/2 + 1/3 \cdot 1/3 \cdot 1/2 + 0 \cdot 1/3 = 1/9
\end{aligned}$$

and the denominator is from the previous derivation for statistical parity.

Similarly,

$$\begin{aligned}
P(D=1|L=1, G=1) &= \frac{P(D=1, L=1|G=1)}{P(L=1|G=1)}\\
&= \frac{1/12}{5/12} = 1/5,
\end{aligned}$$

because

$$\begin{aligned}
&P(D=1, L=1|G=1)\\
&= P(D=1, (Z,\eta)=(1,2), \phi=1|G=1)\\
&\quad + P(D=1, (Z,\eta)=(2,1), \phi=1|G=1)\\
&\quad + P(D=1, (Z,\eta)=(2,2)|G=1)\\
&= P(D=1|(Z,\eta)=(1,2), G=1)P((Z,\eta)=(1,2)|G=1)P(\phi=1)\\
&\quad + P(D=1|(Z,\eta)=(2,1), G=1)P((Z,\eta)=(2,1)|G=1)P(\phi=1)\\
&\quad + P(D=1|(Z,\eta)=(2,2), G=1)P((Z,\eta)=(2,2)|G=1)\\
&= 1/3 \cdot 1/3 \cdot 1/2 + 1/6 \cdot 1/3 \cdot 1/2 + 0 = 1/12.
\end{aligned}$$

A.3 Marginal Outcome Test

The Marginal Outcome Test (based on loan decisions) will be satisfied under a merit-based decision maker. We define a set of marginal candidates as approved applicants who originated their loans but have been denied by some lenders.

$$\mathcal{M}_g = \{i : N_i > 1, \sum_l L_{il} < N_i\}$$

where $N_i = \sum_l (L_{il} + |1 - L_{il}|)$ is the total number of applications submitted by i because $|1 - L_{il}|$ is a denial indicator. Note that under a merit-based decision maker, we have that

$$\mathcal{M}_g = \{i : E(D_{il}|Z_i, \eta_i) = c, G_i = g\}$$

Because c does not depend on g, the desired result follows,

$$P(D_{il} = 1|i \in \mathcal{M}_g) = P(D_{il} = 1|i \in \mathcal{M}_{g'}).$$

A.4 Equalized Odds

Equality of Opportunity will not be satisfied under a merit-based decision maker. Again, let us take the example from Table 4 with $c = 1/3$. First note that

$$P(L=0|D=0,G=0) = \frac{P(L=0,D=0|G=0)}{P(D=0|G=0)}.$$

Note that the numerator is

$$\begin{aligned}
&P(L=0,D=0|G=0)\\
&= P(D=0,(Z,\eta)=(1,1)|G=0)\\
&\quad + P(D=0,(Z,\eta)=(1,2),\phi=0|G=0)\\
&\quad + P(D=0,(Z,\eta)=(2,1),\phi=0|G=0)\\
&= P(D=0|(Z,\eta)=(1,1),G=0)P((Z,\eta)=(1,1)|G=0)\\
&\quad + P(D=0|(Z,\eta)=(1,2),\phi=0,G=0)P((Z,\eta)=(1,2)|G=0)P(\phi=0)\\
&\quad + P(D=0|(Z,\eta)=(2,1),\phi=0,G=0)P((Z,\eta)=(2,1)|G=0)P(\phi=0)\\
&= 1/3\cdot 0 + 1/3\cdot 1/3\cdot 1/2 + 1/3\cdot 1/3\cdot 1/2 = 1/9.
\end{aligned}$$

And the denominator is $1 - P(D=1|G=0)$. We get that

$$\begin{aligned}
P(D=1|G=0) &= P(D=1|G=0,(Z,\eta)=(1,1))P((Z,\eta)=(1,1)|G=0)\\
&\quad + P(D=1|G=0,(Z,\eta)=(1,2))P((Z,\eta)=(1,2)|G=0)\\
&\quad + P(D=1|G=0,(Z,\eta)=(2,1))P((Z,\eta)=(2,1)|G=0)\\
&\quad + P(D=1|G=0,(Z,\eta)=(2,2))P((Z,\eta)=(2,2)|G=0)\\
&= 2/3\cdot 0 + 1/3\cdot 1/3 + 1/3\cdot 1/3 + 0\cdot 1/3 = 2/9.
\end{aligned}$$

So

$$P(L=0|D=0,G=0) = \frac{P(L=0,D=0|G=0)}{P(D=0|G=0)} = \frac{1/9}{1-2/9} = 1/7.$$

For $G = 1$,

$$\begin{aligned}
&P(L=0,D=0|G=1)\\
&= P(D=0,(Z,\eta)=(1,1)|G=1)\\
&\quad + P(D=0,(Z,\eta)=(1,2),\phi=0|G=1)\\
&\quad + P(D=0,(Z,\eta)=(2,1),\phi=0|G=1)\\
&= P(D=0|(Z,\eta)=(1,1),G=1)P((Z,\eta)=(1,1)|G=1)\\
&\quad + P(D=0|(Z,\eta)=(1,2),\phi=0,G=1)P((Z,\eta)=(1,2)|G=1)P(\phi=0)\\
&\quad + P(D=0|(Z,\eta)=(2,1),\phi=0,G=1)P((Z,\eta)=(2,1)|G=1)P(\phi=0)\\
&= 1/3\cdot 1/3 + 2/3\cdot 1/3\cdot 1/2 + 2/3\cdot 1/6\cdot 1/2 = 5/18.
\end{aligned}$$

And the denominator calculated similarly as above is $1 - P(D = 1|G = 1) = 1 - (2/3 \cdot 1/3 + 1/3 \cdot 1/3 + 1/3 \cdot 1/6 + 0 \cdot 1/6) = 7/18$, leaving us with a value of 5/11 in total.

Therefore, $P(L = 0|D = 0, G = 0) \neq P(L = 0|D = 0, G = 1)$.

We note that Equality of Goodwill will not be satisfied under a merit-based decision maker following similar arguments as above.

A.5 Conditional Statistical Parity

Conditional Statistical Parity with respect to Z will generally not be satisfied under a merit-based decision maker. We use the same example as above with $c = 1/3$. Note that

$$\begin{aligned} P(L = 0|Z = 1, G = 0) &= \frac{P(L = 0, Z = 1|G = 0)}{P(Z = 1|G = 0)} \\ &= \frac{P(Z = 1, \eta = 1|G = 0) + P(Z = 1, \eta = 2, \phi = 0|G = 0)}{P(Z = 1|G = 0)} \\ &= \frac{0 + 1/3 \cdot 1/2}{1/3} = 1/2, \end{aligned}$$

and

$$\begin{aligned} P(L = 0|Z = 1, G = 1) &= \frac{P(L = 0, Z = 1|G = 1)}{P(Z = 1|G = 1)} \\ &= \frac{P(Z = 1, \eta = 1|G = 1) + P(Z = 1, \eta = 2, \phi = 0|G = 1)}{P(Z = 1|G = 1)} \\ &= \frac{1/3 + 1/3 \cdot 1/2}{2/3} = 3/4. \end{aligned}$$

A.6 Representativeness

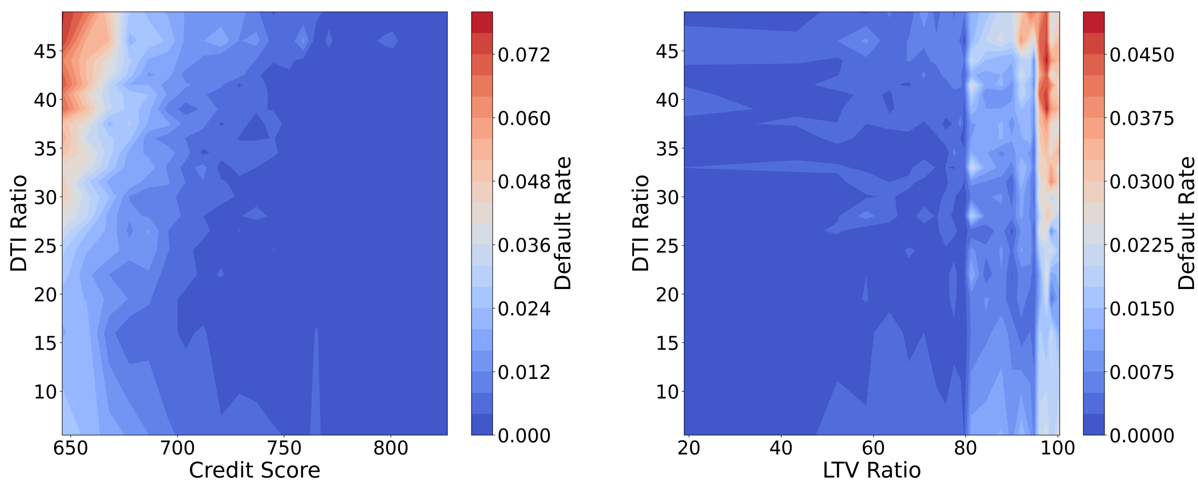
Representativeness with respect to Z and G will generally not be satisfied under a merit-based decision maker. This is because there is no $\hat{c}$ such that $L = 1\{E(D|Z, \eta) < c\} = 1\{E(D|Z, G) < \hat{c}\}$ unless the creditworthiness ordering by $E(D|Z, \eta)$ and $E(D|Z, G)$ is identical.

In our simple example from Table 4, it becomes evident that the creditworthiness orderings based on $E(D|Z, \eta)$ and $E(D|Z, G)$ can diverge. First note that for any applicant with $Z = z$ and $G = g$, we can calculate the conditional probability of D given Z and G, $E(D|Z = z, G = g)$, from Table 4. When we rank applicants by $E(D|Z = z, G = g)$, we observe that more applicants with $(Z = 1, \eta = 1, G = 0)$ are included while applicants with $(Z = 1, \eta = 2, G = 1)$ are excluded than when we rank them by $E(D|Z, \eta)$. This discrepancy arises because η and G exhibit a negative correlation with each other when conditioned on $Z = 1$.

B Modeling Default

Estimating Representativeness requires predicted default probabilities. Using the matched HMDA-McDash sample described in Section 2, we can construct a model for these probabilities. That is, we use this sample to estimate $P(D = 1|Z)$, with the usual caveat that we observe default behavior conditional only on loan origination (and thus acceptance). In this section, we briefly describe our model and its performance.[25]

Model Family. Figure 13 presents contour plots of the relation between the empirical default probability for mortgages originated in 2014 and a number of observed covariates.



(a) As a function of credit score and debt-to-income ratio.

(b) As a function of loan-to-value and debt-to-income ratio.

Figure 13: Empirical default probability (defined as 90+ days past due within two years of origination) across a number of key observables.

The observed pattern of nonlinear interactions motivates the use of a flexible machine learning algorithm to develop a default prediction. In contrast with traditional econometric models, such as logistic regressions, we want to allow for highly nonlinear relationships and rich interactions between the covariates.

Specifically, our main prediction model is a *histogram-based gradient-boosted classification tree* (HGBC) with monotonicity constraints. An HGBC tree is an ensemble method combining multiple decision trees. Unlike other ensemble methods, where combined elements are formally independent of one another, gradient-boosted trees instead proceed iteratively.

[25]In addition to the machine learning model described in this section, we also considered linear and nonlinear logistic regression models. Across all years, we found that the area under the receiver operator characteristic (ROC) curve is larger for our machine learning model, and we thus focus on this model.

That is, each subsequent iteration of the model is obtained by adding a new weak learner that is fit to the *gradient* of the loss function at the total predictions so far.

Formally, given a loss function $L(y, \hat{y})$, a learning rate γ_m, and a class of weak learners $\mathcal{F}$ (in the HGBC case shallow decision trees), the learning process for gradient-boosted trees can be described as the following algorithm:

First, obtain f_1 such that

$$f_1 \in \text{argmin}_{f \in \mathcal{F}} \sum_i L(y_i, f(X_i))$$

and set $F_1 = f_1$. Then, iterate the following steps for $m \geq 2$.

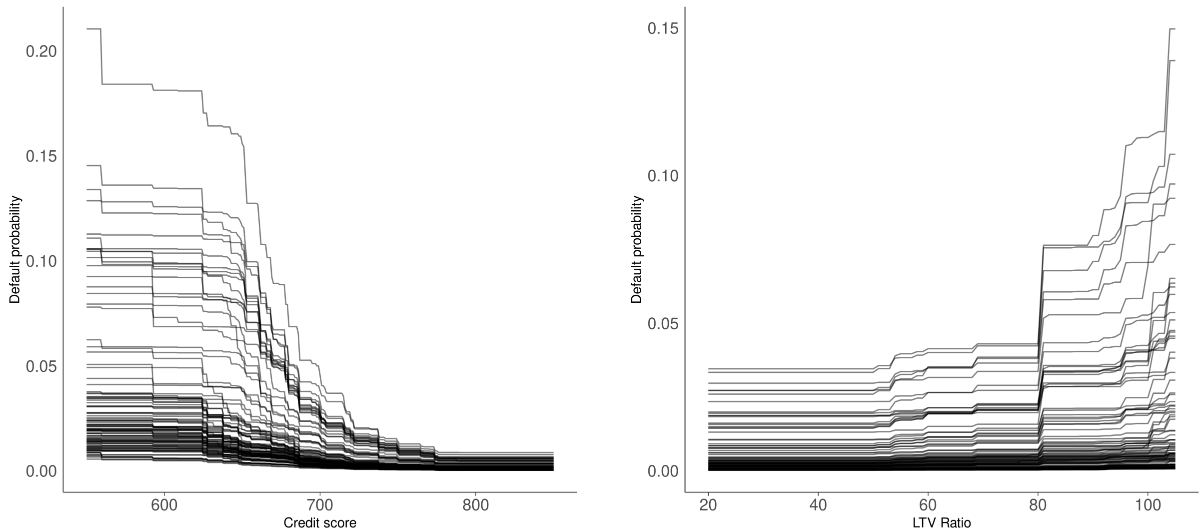
1. Get f_m from
$$f_m \in \text{argmin}_{f \in \mathcal{F}} \sum_i L\left(\frac{\partial L}{\partial \hat{y}}\bigg|_{F_{m-1}(X_i)},\ f(X_i) \right).$$
We use mean-squared error loss, such that $\frac{\partial L}{\partial \hat{y}} = y - \hat{y}$, which can be viewed as a residual.

2. Update F_m based on
$$F_m(X) = \sum_{i=1}^{m} \gamma_i f_i(X) = F_{m-1}(X) + \gamma_m f_m(X)$$
and stop when the max number of iterations, M, has been reached, i.e., $m = M$.

What makes our models a histogram-based gradient-boosted tree is the process of the data-driven selection of binning scheme to consolidate features into lower-cardinality categorical features, as described further in Ke et al. [2017].

We use sklearn in our implementation and, within a model we fix γ_m as a constant, γ. Additionally, because it is non-standard, we briefly discuss the monotonicity constraints that we impose on the algorithm. Such constraints a priori impose relationships based on economic arguments and serve as regularization. To visualize the effects of the monotonicity constraints, we depict individual conditional expectations (Goldstein et al. [2015]) in Figure 14.

Each black line represents a mortgage application filed in 2014 from a random sample of 100 applications. For example, in panel (a), we construct each black line by varying the credit score from its actual value reported on the application and fixing all other features. The resulting line traces out predicted default probability at each value of the credit score and is called an Individual Conditional Expectation (Goldstein et al. [2015]). Our monotonicity constraints enforce that at an individual level, the relationship between the covariate and the default probability is always monotone.



(a) Constrained HGBC, credit score **(b)** Constrained HGBC, LTV ratio

Figure 14: Individual conditional expectation plots for credit score and LTV ratio under the constrained HGBCs. Each black line represents a mortgage application filed in 2014 from a random sample of 100 applications.

Direct estimates of probability that come from tree-based models can sometimes be noisy and require an additional calibration step (Niculescu-Mizil and Caruana [2005]). However, in our application, our monotonicity constraint models lead to well-calibrated predictions on the test set, and so we do not recalibrate with an additional model.

Features. Our baseline features are the following covariates observed at time of origination: credit score (-), Loan-to-Value (LTV) ratio (+), Debt-to-Income (DTI) ratio (+), original loan amount (+), applicant income (-), a dummy for whether a coapplicant is present on the application (-), a code for the geographic state of the property's location, loan term (in months), funding source (the type of purchaser of the loan), a dummy indicating private mortgage insurance (PMI), and dummies for loan type (conventional, government-insured, VA, farm) and loan purpose (investment, refinancing). A plus or minus sign in brackets indicates that we impose a monotonicity constraint in the indicated direction in our model.

Train-test split. We start by applying a random 70%-30% split to create our training and testing sets in a given year. Within the training set, we use 3-fold cross validation for hyperparameter tuning. In particular, we use a parameter grid of $\gamma \in \{0.02, 0.025, 0.03\}$, $M \in \{300, 350, 400\}$, and maximum leaf nodes from $\{16, 20, 24, 28\}$. The models trained separately by year are then allowed to vary over all combinations of this set of hyperparameters and select the combination resulting in the lowest average mean squared error across folds.

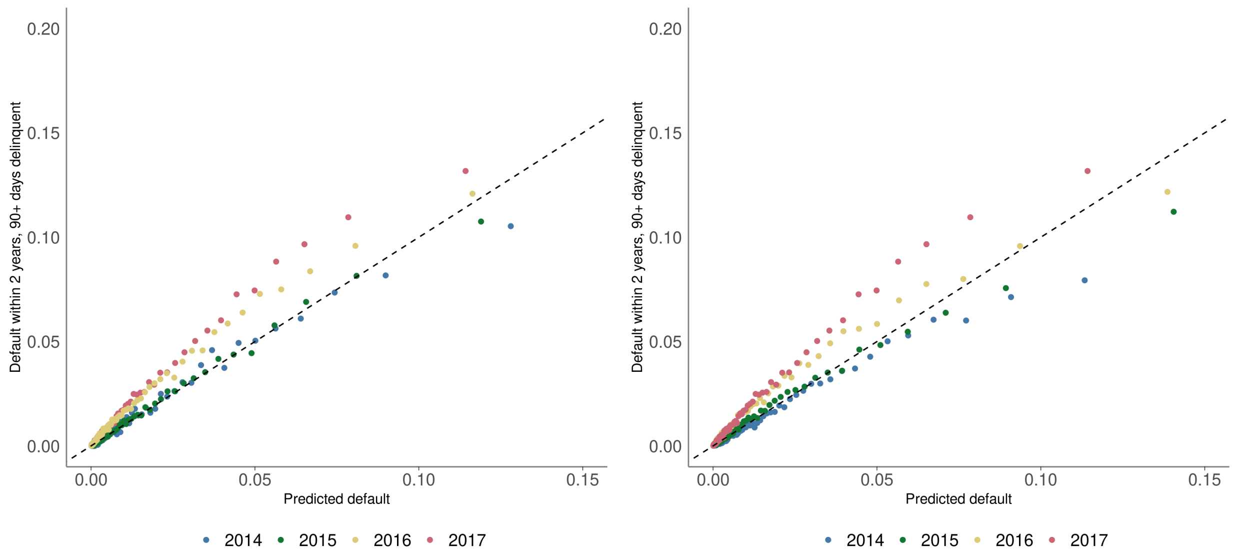
However, evaluating a model against mortgages from the same year as the training set may be misleading because the model may incorporate future information during the span in which loan performance is measured (in our case, 24 months from origination). Thus, to account for this look-ahead bias, a test set that avoids any information leakage must contain mortgages originated at least three years after the training set. For instance, we first split mortgage applications filed in 2014 into two subsamples and use the 30% withheld as a first test set to evaluate the performance of our model. We then also evaluate the performance of our model, trained on 2014 data, on the following years. For mortgages originating in 2015 and 2016 we still have (decreasing) amounts of information leakage, while mortgages originating in 2017 and beyond are free of information leakage. For our fairness measures, we therefore use the model trained on data in year $t-3$ to estimate the default probabilities of mortgages originated in year t.

Model Performance. We illustrate our model performance visually in Figure 15. Here, we use a binscatter (a binned scatter plot) as a flexible, yet parsimonious way of summarizing the relationship between our predicted default probability and the empirical default rates. We use percentiles as our bins, such that each bin represents one percentage point. Visually assessing whether our prediction model is effective in risk-ordering the applicants amounts to checking whether the empirical default rates are monotonically increasing. Visually assessing whether our prediction model is well-calibrated amounts to checking whether the empirical default rates are close to the 45-degree line (indicated by the dashed lines).

In Figure 15a, we illustrate the performance of the model trained on 2014 data across subsequent years. In Figure 15b, we depict the performance of our model on test data from 2014-2017 (t), where the model is trained in 2011-2014 ($t-3$).

In both panels we observe monotonically increasing empirical default rates, indicating that our prediction model performs well at risk-ordering the applicants. In other words, our default model seems to perform well at predicting the relative risk of applicants. On the other hand, as the temporal gap between train and test data increases, our models are not always well calibrated: For example, in the years depicted here, there is a general tendency to underestimate the default risk for applications filed in 2016 and 2017. We note that this is in line with the performance of traditional credit scores, which do well in their relative ranking of consumers but are not designed to be time consistent (see, e.g. Demyanyk [2010], Albanesi and Vamossy [2019]).

Figure 16 depicts a binscatter for the model trained on 2014 data, with the predicted and observed default probabilities plotted for each demographic group. This lets us assess whether there are any clear discrepancies in model performance across demographic groups. We note that in the 2017 test data, our algorithm significantly underpredicts the default risk of Black applicants relative to White applicants, with the blue points lying farther above the 45-degree line relative to the red points, particularly at higher predicted probabilities of default. This pattern is qualitatively similar for other years. We note that the amount of miscalibration is remarkably similar to what we observed in Figure 4, which illustrated how



(a) Using 2014 as training data

(b) Using $t - 3$ as training data

Figure 15: Binscatter depicting empirical default rates in year t as a function of predicted default probabilities. Each bin represents one percentage point.

credit score alone appears to underpredict the risk of Black applicants.

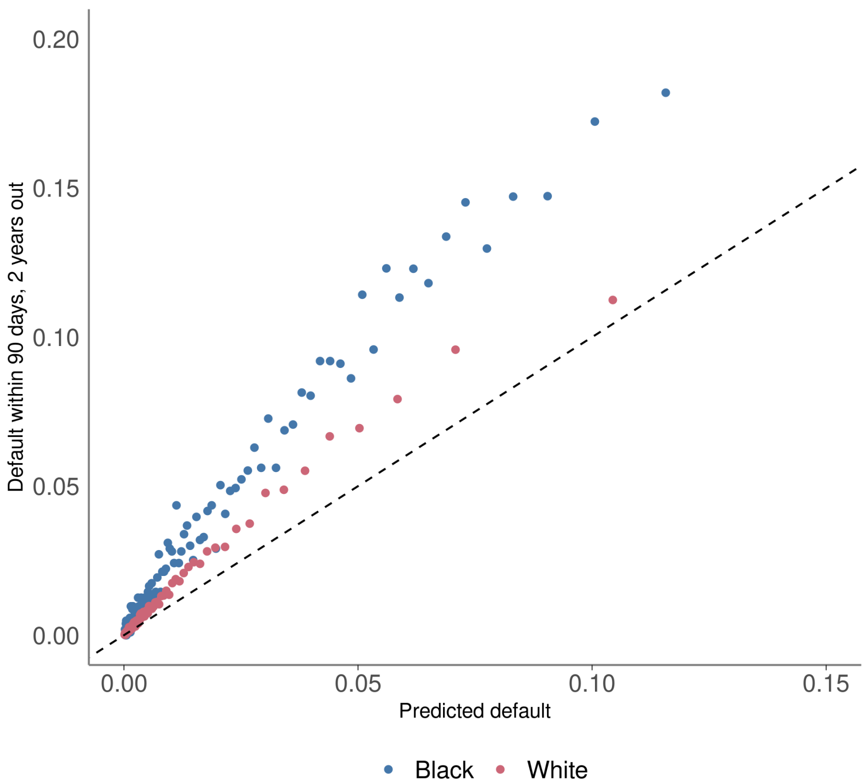


Figure 16: Binscatter depicting empirical 24-month-out default rates for mortgages originated in 2017 as a function of predicted default probabilities based on a 2014 race-blind default model, separated by demographic groups: Black and White. Each bin represents one percentage point.