FEDERAL RESERVE BANK
PHILADELPHIA

Working Papers
RESEARCH DEPARTMENT

One Threshold Doesn't Fit All
Tailoring Machine Learning Predictions of Consumer Default for Lower-Income Areas

Vitaly Meursault
Federal Reserve Bank of Philadelphia Research Department

Daniel Moulton
Federal Reserve Bank of Philadelphia Consumer Finance Institute

Larry Santucci
Federal Reserve Bank of Philadelphia Consumer Finance Institute

Nathan Schor
Federal Reserve Bank of Philadelphia Research Department

WP 22-39

PUBLISHED
November 2022

REVISED
October 2024

FRB CFI
PHILADELPHIA

ISSN: 1962-5361

DOI: https://doi.org/10.21799/frbp.wp.2022.39

One Threshold Doesn't Fit All: Tailoring Machine Learning Predictions of Consumer Default for Lower-Income Areas

Vitaly Meursault, Daniel Moulton, Larry Santucci, Nathan Schor

Federal Reserve Bank of Philadelphia

October 2024

Abstract

Improving fairness across policy domains often comes at a cost. However, as machine learning (ML) advances lead to more accurate predictive models in fields like lending, education, healthcare, and criminal justice, policymakers may find themselves better positioned to implement effective fairness measures. Using credit bureau data and ML, we show that setting different lending thresholds for low and moderate income (LMI) neighborhoods relative to non-LMI neighborhoods can equalize the rate at which equally creditworthy borrowers receive credit. ML models alone better identify creditworthy individuals in all groups but remain more accurate for the majority group. A policy that equalizes access via separate thresholds imposes a cost on lenders, but this cost is outweighed by the substantial gains from ML. This approach aligns with the motivation behind existing laws such as the Community Reinvestment Act, which encourages lenders to meet the credit needs of underserved communities. Targeted Special Purpose Credit Programs could provide the opportunity to prototype and test these ideas in the field.

Keywords:
fair lending policy, credit scores, group disparities, machine learning, fairness
JEL Classifications: G51, C38, C53

Author information: Vitaly Meursault (corresponding author), vitaly.meursault@phil.frb.org; Daniel Moulton, daniel.moulton@phil.frb.org; Larry Santucci, larry.santucci@phil.frb.org. Nathan Schor participated in this research in his prior position at the Federal Reserve Bank of Philadelphia.

Acknowledgments: We are grateful for the helpful comments of Ken Benton, Neil Bhutta, Julia Cheney, Fumiko Hayashi, Peter Hull, Bob Hunt, Lauren Lambie-Hanson, Scott Nelson, Jeanne Rentezelas, P-R Stark, and Jack Terruso, as well as the seminar and conference participants at the University of Delaware, Carnegie Mellon University, ETH Zurich, and the Fourth Workshop on Payments, Lending, and Innovations in Consumer Finance. We also thank Kerry Rowe for data management support and Kellen O'Connor for computing infrastructure support.

1. Introduction

Disparities in access to credit have been a persistent issue in the United States. For example, a 1996 study of the Boston mortgage market found that 28 percent of Black mortgage applicants were rejected, compared with just 10 percent of White applicants (Munnell et al. 1996). A recent study shows a narrowed but still large gap: 18 percent of Black applicants were denied a mortgage, compared with 8 percent of White applicants (Bhutta et al. 2021).[1] Policymakers have recognized the need to address disparities such as these, with the Community Reinvestment Act (CRA) of 1977 being a notable example. The CRA sought to encourage banks to meet the credit needs of low- and moderate-income neighborhoods, reflecting a recognition that fairness, in the form of adequate credit access in underserved communities, is an important objective of credit markets along with economic efficiency and lender profits.

The goal of our study is to demonstrate the potential of machine learning (ML) and, broadly, artificial intelligence (AI) models to better align the objectives of fairness and efficiency. Two key developments have made this possible: the improved predictive power of these models, driven by advances in data availability and computational capacity, and the emergence of fairness techniques in ML to address disparities in model predictions.[2] While increased predictive accuracy alone may not necessarily lead to fairer outcomes, it creates a more favorable environment for implementing ML fairness techniques, which would otherwise involve a larger trade-off in predictive performance. Importantly, our work shows that certain fair ML approaches, such as adjusting decision thresholds for protected groups, are consistent with existing laws such as the Special Purpose Credit Programs (SPCP) provision of the Equal Credit Opportunity Act (ECOA). We show that the gains in predictive power from advanced models can more than offset the costs

[1] These raw denial gaps should not be directly interpreted as evidence of race-based discrimination in the legal sense. The gap is also smaller, and in the case of Bhutta et al. (2021), falls to 2 percentage points when accounting for observable borrower characteristics. We highlight these raw denial gaps to reflect broader economic disparities that extend beyond a specific borrower-lender interaction.

[2] This is in line with ongoing questions about the implications of big data in policy research, such as those raised by Lane (2016) and Jarmin and O'Hara (2016).

associated with fairness-oriented policies, making ML/AI beneficial from both efficiency and fairness perspectives. This approach to ML/AI regulation could be fruitful across various policy domains that rely on assessing individuals' likelihood of future success, such as lending, education, employment, healthcare, and justice.[3]

In recent studies, the integration of ML techniques in consumer credit markets has been scrutinized for its potential to exacerbate disparities among borrowers, particularly affecting minority groups. Research such as Fuster et al. (2021) has shown a potential for modestly worse pricing outcomes for minorities with the adoption of ML underwriting, while Blattner and Nelson (2021) have highlighted the persistence of disparities in credit score noisiness between demographic groups. Building upon this context, we present a novel analysis, demonstrating that lenders could use group-specific credit score thresholds in conjunction with ML to address some aspects of these disparities without sacrificing profitability. Our approach to introducing fairness preferences follows the ideas of Kleinberg et al. (2018), who argue that it is optimal to use the most predictive model coupled with separate decision thresholds to satisfy the fairness preference, rather than to focus on blinding the algorithm to group membership, which is the current practice in lending.

Intuitively, lenders could use ML to create a new credit scoring model and then set lower credit score approval thresholds for disadvantaged groups. For instance, they could use a credit score approval threshold of 680 for a non-disadvantaged group and a lower approval threshold of 660 for a disadvantaged group. This would lead to more approvals for individuals in the disadvantaged group. Because scores are less accurate for disadvantaged groups (see, e.g., Blattner and Nelson 2021), this approach can reduce gaps in access to credit to creditworthy individuals.

A key empirical contribution of our study lies in examining the feasibility of implementing differentiated credit score thresholds in lending. Specifically, our analysis

[3] For some examples of predictive systems and fair ML interventions in these and other policy domains, see Lamba et al. 2021, Chouldechova et al. 2018, Mullainathan and Obermeyer 2021, and Arnold et al. 2022.

focuses on whether lowering the approval threshold for disadvantaged groups would significantly increase default rates, given the existing disparity in credit score accuracy for these groups.[4] Our empirical contribution investigates this trade-off, demonstrating that integrating ML can effectively balance the goals of increasing access to credit for the disadvantaged group with maintaining lender profitability.

We focus on inequalities between lower- and higher-income areas based on the concept of historically underserved communities found in the Community Reinvestment Act (CRA; 12 U.S.C. §2901). The CRA was passed in 1977 to encourage financial institutions to help meet the credit needs of LMI neighborhoods.[5] Because our objective is to examine how increased fairness can be achieved using group-specific thresholds, it seems appropriate to construct our thresholds in a manner consistent with this concept of LMI neighborhoods as defined by the CRA. This also serves to conceptually link fairness in the context of group-specific prediction thresholds to fairness considerations in existing law and lending practice.

We begin by confirming that the predictive power gap across population groups documented in papers such as Fuster et al. (2021) and Blattner and Nelson (2021) also exists in the CRA context. For individuals who live in LMI census tracts, credit scores based on models we estimate have lower predictive power than for non-LMI tract consumers. For lending decisions based solely on credit scores, this means that in LMI areas, consumers who should receive credit are relatively less likely to get it, and customers who won't pay back loans are more likely to receive a loan.

We proceed with a novel analysis that considers the introduction of group-specific lending thresholds within the context of technological progress in default risk assessment.

---

[4] If the credit scores of false negatives (i.e., individuals who would have repaid but were denied) in disadvantaged groups fall far below the common cutoff used in a single-threshold scenario, the necessary adjustment to alleviate disparities might be too large. If the concentration of true negatives (i.e., rejected applicants who would default if approved) is high relative to the density of false negatives for the disadvantaged group at a threshold just below the common cutoff, lowering it for this group could incur substantial costs for lenders.

[5] A census tract is defined as LMI if the tract's median income is less than 80 percent of the metropolitan statistical area/metropolitan division's (MSA/MD's) median income.

Our empirical approach focuses on a binary prediction of loan repayment that corresponds to the lending practice of approving loans for consumers with credit scores above a certain threshold. We compare the predictive performance under different rules for setting the credit score approval threshold.

The benchmark for the comparison is setting a single threshold for all applicants. This corresponds to the current regulatory framework, which prohibits lenders from considering information related to sensitive attributes such as race, ethnicity, and gender for most lending decisions. Lenders are also prohibited from using variables that are close proxies for prohibited attributes. Variables that identify an individual's geographic area or exact location are typically considered proxy variables and thus prohibited from use in lending decisions. While the intent of this policy is to reduce discrimination, a growing body of literature suggests that this approach is not optimal for reducing disparities in outcomes (Kleinberg et al. 2018; Lamba et al. 2021). We show that a single-threshold approach creates disparities in true positive rates (TPR) between groups (where repaying the loan is the positive outcome). In our main example, discussed in Section 4.5, a creditworthy LMI tract consumer is about 9 percentage points less likely to be classified as creditworthy than a creditworthy non-LMI tract consumer.

The alternative approach we consider would permit the use of specific geographic variables (e.g., residence in an LMI versus a non-LMI neighborhood) to equalize true positive rates among different groups. As a result, LMI tract groups with noisier credit scores would be assigned lower thresholds. Kleinberg et al. (2018) ground such approaches theoretically by arguing that modifying decision thresholds is an optimal way to incorporate a fairness preference. Crucially, this approach does not require lenders to build separate predictive models for LMI and non-LMI areas and can be used with any credit score, even when nothing is known about the model that generated it.

The reduction of TPR disparities comes at a cost of some eventual defaulters being misclassified as non-defaulters. This is a cost from the lender profit perspective and, potentially, to consumer welfare.[6] We show that the costs from the lender profit perspective can be mitigated if fairness constraints are paired with model improvements.

This paper makes four main contributions to the literature. First, we show that considering fairness constraints alongside model improvement can lead to different conclusions about the effects of ML on the equality of credit access. Fuster et al. (2021) and Blattner and Nelson (2021) show that the more advanced models predict default better overall, but they have only a marginal effect on the relative access to credit between groups. In contrast, we show that the more advanced models combined with fairness constraints can significantly reduce the gap in credit access for creditworthy consumers while still improving overall default prediction. Second, we focus on model improvement in the context of generic credit scoring rather than specifically mortgage default prediction. Thus, our results are relevant for a wide range of markets, including credit cards and auto loans, which have higher participation among historically underserved groups than mortgages. Third, we focus more on some of the practical aspects of ML introduction in credit scoring. On the modeling side, this includes rolling window model estimation and threshold generation that is aimed at eliminating the look-ahead bias. On the policy side, this includes a discussion of regulatory hurdles of ML and fairness constraint adoption and a potential path forward via the SPCP provision. Fourth, our results highlight the potential benefits of imposing fairness constraints by explicitly considering certain geographies that are likely to be correlated with protected attributes during the design of loan approval policies.[7]

[6] Estimating the net welfare impact of a higher probability of getting a loan with the consequences of default is beyond the scope of the paper. For some relevant considerations, see, e.g., Fedaseyeu and Hunt (2018), Fulford and Nagypal (2023), Kermani and Wong (2021), and Sodini et al. (2023)

[7] Gillis (2022) provides a legal discussion of input- and output-based fair lending scrutiny of credit scoring models. Caro and Nelson (2023) advocate for the explicit inclusion of fairness constraints in the selection and implementation of screening models and data inputs within these systems from a legal perspective.

1.1. Related literature

Our work continues the long line of economic research about discrimination in lending, but it is most related to the recent work that focuses on racial, ethnic, and income group disparities that can arise from the use of consumer default prediction models. We also build on the Fairness in Machine Learning literature that studies the gaps in predictive power across groups and introduces techniques to mitigate these gaps.

A large body of literature examines the disparate impact (or lack thereof) in independent variables incorporated into predictive models through their correlation with group membership rather their predictive power for future default (see, for example, Avery et al., 2012). Relative to this literature, we shift our focus to reducing the disparities in predictive power between the groups, regardless of the origin of these disparities, and highlight the potential benefits of using protected attributes for lending decisions in a way that makes outcomes more equitable.

The predictive power of credit scores for different population groups has been a point of interest in economics and beyond for some time. The literature started with considering the effects of a specific credit scoring approach and moved to considering an additional dimension of credit scoring model sophistication (Avery et al. 2012, Fuster et al. 2021, Blattner and Nelson 2021, Bartlett et al. 2022). We contribute to this literature by adding a new dimension, fairness constraints, and show that joint movement on the model sophistication and fairness constraint dimensions can achieve both higher profits and more equal outcomes.

In concurrent work, Blattner et al. (2023) explore the intersection of fairness and model sophistication in predictive algorithms used in lending and beyond. Their study emphasizes the importance of nuanced regulation, particularly through targeted algorithmic audits that address specific disparities such as racial biases. They find that more complex models, when appropriately regulated, can lead to both efficient and fair outcomes, a conclusion that mirrors our findings. This parallel research underscores the emerging consensus on the potential of advanced models, coupled with fairness-focused regulation, to improve decision-making processes.

This paper is also relevant to the literature that considers the racial differences in default conditional on credit score. The Federal Reserve compiled an exhaustive report to Congress in 2007,[8] which contained a discussion of higher default rates among some groups of minority individuals relative to the majority for a given credit score bin. Other work includes Bayer et al. (2016), which showed that Black and Hispanic homeowners had much higher rates of delinquency and default during the housing bust. Our work highlights that relative default rates for a given credit score are dependent on the credit scoring model and how it is applied, which suggests that the negative consequences highlighted in these papers can be mitigated with technological change and appropriate policy.

Before machine learning, the introduction of traditional credit scoring was itself a technological change that had a profound impact on the markets, studied, for example, by Edelberg (2006) and Einav et al. (2013). More recently, advancements in credit scoring have shown the potential to help a larger portion of alternative financial service users qualify for more conventional forms of credit (Servon, 2017). Our work continues this trend of studying the effects of screening technology on consumer finance outcomes.

Basing lending decisions on credit score cutoffs is a common feature in financial markets, frequently studied by researchers (e.g., Keys et al. 2010; Laufer and Paciorek 2022; Bronson et al., 2019). The focus of our work is setting these thresholds optimally to optimize for a double objective of profit and fairness.

There is a large literature in the field of ML focusing on measuring and mitigating disparities in predictive power between demographic groups. Chouldechova and Roth (2018) provide a view of the frontier of the academic literature in 2018. A variety of methods exists to make model predictions fairer, by preprocessing the data, modifying the predictive algorithms, or adjusting existing predictions (see Lamba et al. 2021 for comparison on a variety of tasks). We choose to focus on adjusting existing predictions using an approach similar to Hardt et al. (2016), because it is tractable, easy to implement

[8] See https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/creditscore.pdf. Last accessed: 10/08/2024.

and explain, isn't clearly dominated by other approaches (Lamba et al., 2021), and has theoretical grounding (Kleinberg et al., 2018). In this paper, we focus on introducing fairness constraints into the consumer lending context in a simple way to highlight the trade-off between profit and fairness and show how model improvement can soften it. We leave examining the important nuances of fairness constraints and their effect on consumers, such as subgroup fairness (Kearns et al., 2018) and dynamic fairness (Liu et al., 2019), to future work.

1.2. Policy and regulatory considerations

Both components of our analysis, the use of ML models and the use of sensitive attributes in lending decisions, are areas of policy discussion.

In the paper, we primarily discuss model improvement as the lender's decision and the imposition of fairness constraints as the policy maker's choice. However, in reality, the adoption of ML in the lending industry is determined jointly by lenders and regulators. According to a 2021 report by FinRegLab (FinRegLab, 2021), technologically enabled lenders use ML for data analysis and feature engineering across sectors and asset classes. However, ML underwriting models are still in their early stages.[9] These models offer potential benefits but also raise a variety of concerns, including model performance in unexpected conditions, fairness, inclusion, privacy, security, and transparency.[10] Regulators thus have an ability to facilitate ML adoption in the industry by providing clear guidance and reducing regulatory uncertainty. In this sense, both model improvement and the introduction of fairness constraints can be viewed as regulatory decisions.

Two federal laws explicitly prohibit discrimination in fair lending: the Equal Credit Opportunity Act (ECOA; 15 U.S.C. §1691) and the Fair Housing Act (FHA; 42 U.S.C. §§3601-3619).[11] Fair lending laws generally prohibit lenders from favoring a particular

[9] FinRegLab (2021)

[10] FinRegLab (2021)

[11] The Equal Credit Opportunity Act (ECOA), implemented by Regulation B (12 C.F.R. §202), prohibits discrimination in any aspect of a credit transaction and applies to any extension of credit. The discrimination prohibition covers nine prohibited factors: race, color, religion, national origin, sex, marital status, age, because an applicant receives income from a public assistance program, or because an

class of borrowers in any aspect of a lending decision, even if that class has been historically discriminated against. Thus, it is unclear whether the general use of group-specific thresholds would be permissible under existing federal law as currently implemented.

That said, ECOA permits lenders to design and implement tailored SPCPs with rules favoring a historically disadvantaged class of borrowers. Some of the largest US banks have recently implemented such programs, including Wells Fargo, Bank of America, JP Morgan, and TD Bank. Government Sponsored Enterprises (GSEs) are also developing SPCPs and are working to facilitate the purchase of home loans originated through lender SPCPs.[12]. We discuss SPCPs and other policy considerations in greater detail in Appendix A.

While the current version of the CRA does not explicitly address whether loans originated under a SPCP may be considered during a lender's CRA assessment[13], changes to the CRA promulgated in 2023 (effective April 1, 2024, with compliance dates beginning January 1, 2026) appear to clarify the relationship. In the final rule, SPCPs are listed as a type of credit product that could be considered responsive to the Retail Services and Products component of a lender's CRA assessment. Furthermore, some of the recent legal research argues for the explicit use of sensitive attributes and outcome-based fair lending analysis (e.g., Gillis 2022; Caro and Nelson 2023). Finally, it is important to note that a policy that combines the encouragement of model improvement with fairness constraints can be implemented in more than one way. For example, Blattner et al. (2023) arrive at

---

applicant has in good faith exercised any right under the Consumer Protection Act. The Fair Housing Act (FHA) is implemented by the U.S. Department of Housing and Urban Development regulations (24 C.F.R. §100) and prohibits discrimination in all aspects of residential real estate–related transactions. In the case of the FHA, there are seven prohibited bases: race, color, national origin, religion, sex (including gender identity and sexual orientation), familial status, and disability.

[12] See, e.g., https://freddiemac.gcs-web.com/news-releases/news-release-details/freddie-macs-2023-equitable-housing-finance-plan-builds-year-one. Last accessed: 10/08/2024.

[13] See https://www.consumercomplianceoutlook.org/2022/fourth-issue/overview-of-special-purpose-credit-programs/. Last accessed: 10/08/2024.

similar conclusions with a very different “fairness audit” procedure. We leave the economic and legal comparison of different approaches to future research.

The increasing adoption of ML in financial services, along with large lender programs aimed at historically underserved borrowers, grounds the main elements of our analysis in the existing lending landscape. By highlighting the potential for improved fairness and efficiency through ML models, our research suggests a promising direction for policy development that balances innovation with consumer protection.

2. Data

The primary dataset used in our analysis is the Federal Reserve Bank of New York Consumer Credit Panel/Equifax data (CCP). The CCP is an anonymized, consumer-level dataset comprising quarterly credit bureau records for a 5 percent random sample of individuals with a credit file.[14] We augment the CCP with the census tract-level demographic data from the U.S. Census Bureau processed by the Federal Financial Institutions Examination Council (FFIEC) to determine the CRA status of consumers based on the income of their census tract of residence.[15]

Unless otherwise noted, all plots and tables are based on authors’ calculations using CCP data with consumers in LMI census tracts identified using the dataset produced by the FFIEC, based on the U.S. Census Bureau data.

2.1. Credit information from the Consumer Credit Panel

The CCP includes quarterly snapshots of credit bureau information on credit accounts, credit inquiries, and public records (e.g., collections, bankruptcy, foreclosure, and tax

---

[14] For additional information about the CCP, see Lee and van der Klaauw, 2010; for a more general discussion of credit report data, see Avery et al., 2003.

[15] See https://www.ffiec.gov/censusapp.htm. Last accessed: 10/08/2024. The FFIEC census data files are compiled using the decennial census and American Community Survey (ACS) data, and they are updated annually. A tract is defined as LMI if the tract’s median income is less than 80 percent of the metropolitan statistical area/metropolitan division’s (MSA/MD’s) median income. For tracts outside a MSA/MD, statewide income is used. It’s important to note that the LMI cutoffs, based on MSA-relative income, inject heterogeneity into the two groups based on the MSA. In other words, a LMI tract in San Jose may not be the same as a LMI tract in Detroit.

liens) at the consumer level. Credit bureau data is the primary input to credit decisioning in the industry.

Our credit data includes several hundred variables covering outstanding and maximum available balances, payment amounts, number of trades, amounts past due, and number of days past due for a range of debt products used by the consumers, including credit cards, mortgages, auto loans, and other kinds of loans. The CCP includes each individual's year of birth as well as geographic designations, including their current census tract. It contains no additional demographic information, such as ethnicity, race, or gender, and does not contain any information about the individual's income or asset holdings. We use data for the years 2000 to 2021. For tractability, we use a random sample of 1/100 consumers in the panel in the main analysis but confirm that our results are robust to using a 1/10 random sample.

Our data-cleaning procedure includes removing duplicate consumer–quarter pairs, observations without valid census tract information, deceased consumers, and consumers who are only intermittently available in the dataset (*fragment files*). To eliminate fragment files, we only keep consumer–quarters that have eight consecutive quarters of delinquency status following the current quarter. This procedure is similar to Hunt and Wardrip (2013), Mikhed and Vogan (2018), and Blascak et al. (2018). We also remove consumer–quarters that have less than 2 quarters with at least one open account within the period between the current quarter and two years starting the next quarter (the period for which we compute our forward-looking default variable).

2.2. Exclusion of consumers already in default

Our data-cleaning procedure also excludes consumers who are currently in default. Thus, we focus on the transition from the state of non-default to the state of default. This is different from the approach taken, for example, by Albanesi and Vamossy (2019), who retain current defaulters and estimate the whole Markov transition matrix.

When building a default model, it is standard practice for credit risk modeling professionals to exclude consumers who are already in default from the model building exercise. Model builders may use those accounts to predict other things, such as the probability of curing to a current status, or the probability of moving from one state of default, say 90–120 days delinquent, to a more severe state, such as charge-off. Since the focus of this paper is on estimating the default probabilities for consumers who are currently not delinquent, for our main analysis, we choose to exclude observations that are currently more than 90 days past due on one or more of their accounts.[16]

3. Predictive models

The goal of our predictive modeling is to generate two credit scores — one based on traditional statistical models and one using newer ML models. In this section, we outline the key steps in the process and refer the interested readers to Appendix B for details.

We predict consumer repayment status using two models: logistic regression with ridge regularization (referred to simply as the *logistic model*) and eXtreme Gradient Boosting (*XGBoost*).[17] Ridge logistic regression represents the class of linear models that have been commonly used for credit scoring in the last few decades, whereas XGBoost represents non-linear models that are being increasingly used for credit scoring today. We refer to the output of our predictive models as a credit score.

We define our prediction target, *non-default*, as a binary variable that is equal to one if the consumer is not in the state of default within two years starting from the next quarter. We define the state of default as being 90 or more days past due on at least one account.[18]

[16] In Appendix E, we perform a robustness check, training a separate model for predicting future default (or recovery) for consumers currently in default and analyzing on the combined set of predictions for both consumers who are current and consumers who are currently in default. Our conclusions are unaffected.

[17] See Hastie et al. (2017) for an introduction to regularized linear models and gradient boosting.

[18] We choose non-default as the event state rather than default for stylistic reasons. In some sections of the paper, we focus on the elements of the confusion matrix such as *true positives* and *false positives.* Choosing non-default as the event aligns the meaning of *positive* as *beneficial* with the meaning of *positive* as in *positive test* facilitating the communication of results.

Table 1 breaks down the percentage of consumer–quarters in the data by the current payment status and future default rate. As discussed in Section 2.2, we only include presently current consumers in our main analysis. Rates for other groups are for reference.

We identify a total of 447 variables suitable for credit scoring in the CCP. This large number of variables presents overfitting challenges, which we address with variable selection, regularization, and hyperparameter tuning. See Appendix C for the details about the variables included in our models.

Credit scoring models are typically updated over time as the data-generating process evolves. Both the way some of the variables are reported and the relationship between RHS variables and default can change over time. To account for this, we estimate our model in a rolling window fashion. Crucially, all decisions about the model fairness policy are made using training period data, while all the results are based on model predictions on data from periods in the future relative to the training periods. This is to say that all evaluation metrics are based on performance on data separate from data used to train the models.

3.1. Model evaluation metrics

We use the receiver operating characteristic area under curve (ROC AUC) as the main overall measure of model performance. ROC AUC is a metric that goes back to the World War II-era analysis of radar receivers (Van Meter and Middleton, 1954). It evaluates the performance of a binary classifier by aggregating true positive and false positive rates at every possible threshold. It takes values from 0.5 (corresponding to random chance) to 1 (corresponding to a perfectly accurate model). Intuitively, ROC AUC represents the probability of a random non-defaulter having a higher credit score than a random defaulter (Fawcett, 2006).

In the sections of the paper that discuss fairness constraints, we focus on the true positive rate (TPR) and false positive rate (FPR) at specific decision thresholds. We also compare the percentages of population that fall into true positive (TP), false positive (FP), true negative (TN), and false negative (FN) groups.[19]

In our context, TP refers to non-defaulting consumers identified by the model as such, FP refers to defaulters identified by the model as non-defaulters, TN refers to correctly identified defaulters, and FN refers to non-defaulters mistakenly identified as defaulters.

3.1.1. Profit: a cost-sensitive classification metric

We also look at a measure of simulated profits, which we define as $Pr = TP - \lambda FP$. At a higher level, this is a cost-sensitive learning metric. The need for such metrics arises when different classification errors are associated with different costs to the decision-maker (Elkan, 2001). We call it "profit" despite not being able to measure profits directly because it reflects an important feature of the profit function — lenders lose more money on an account that defaults than they gain from the account that pays the loan back. At zero profit, λ may be viewed as the number of good accounts required to break even on a single charged-off account. The value of λ can vary widely, depending on the type of loan and the lender's pricing strategy and risk management. We calibrate λ using administrative data for consumer credit card accounts held at large bank holding companies from the Federal Reserve Board's Capital Assessments and Stress Testing (Y-14M) report for January 2014 to December 2022. Refer to Appendix D for more information on the Y-14M report.

Specifically, we calculate the λ values for seven major credit card lenders in the United States. This involves estimating the ratio of the average cost incurred from defaulting loans to the average profit earned over five years from non-defaulting credit card loans. As an illustration, consider a scenario where the average loss from a defaulted loan is \$300, and the average profit from a fully repaid loan is \$50. In this case, the λ value would be calculated as 300/50 = 6. Our approach to calculating account-level profit largely adheres to the methodology outlined in Section A of the Online Appendix of Agarwal et al. (2014). We encompass all general-purpose consumer credit card accounts in our analysis, tracking

[19] See Rodolfa et al. (2016) for a high-level introduction to these concepts from a Fairness in Machine Learning perspective.

the account activity for the initial 60 months of each account's lifespan. We have computed the lender-specific λ values annually for different cohorts spanning 2014 to 2017. The details of the calculation are presented in Appendix D.

The resulting annual average λ values are presented in Table 2. The average λ across banks is 5.14 (5.39 if weighted by total credit card assets). We use the rounded average value of 5 in our main specification and confirm the qualitative robustness of our results to λ values two standard deviations above and below the mean (approximately between 1 and 11).

3.1.2. Disparity in TPR: A fairness metric

In addition to evaluating the predictive power of the models, we also discuss their fairness properties. We choose the disparities in TPR between the LMI and non-LMI areas as our main measure of fairness because it focuses on people who have a need for regulatory intervention — creditworthy individuals in LMI areas. Many fairness definitions are available (see, for example, Hurlin et al. 2022), and they are often incompatible with each other.[20] In practice, our approach of equalizing TPR leads to increased lending in LMI areas, which is consistent with the goal of policies like the CRA. The choice of TPR for the lending case is also consistent with the "fairness tree" guidelines in Rodolfa et al. (2016) as our intervention is assistive (it increases loan access) and targets many people in need (creditworthy consumers), and we are primarily concerned with creditworthy consumers (as opposed to everyone without regard to creditworthiness). Hardt et al. (2016) also note that such a measure "improves incentives by shifting the cost of poor classification from disadvantaged groups to the decision maker, who can respond by improving the classification accuracy," which is consistent with the goals of the paper.

From the perspective of the lending practice, TPR is less commonly used for analyzing disparities than Adverse Impact Ratio (AIR) or Standardized Mean Difference (SMD) (FinRegLab, 2021), and is sometimes described as an alternative metric. We prefer to focus

[20] For example, Kleinberg et al. (2016) provide a famous impossibility theorem in the fairness space, though the degree of its empirical relevance is being debated (Bell et al., 2023).

on TPR because, unlike AIR and SMD, it takes into account eventual outcomes (defaults) and thus recognizes the differences in default probabilities between groups for the determination of the desired equal outcome. It is important to note that our proposed policy combining model improvement with fairness constraints does not necessarily require the use of TPR as a metric and that variants of the policy we propose could be implemented using standard metrics such as AIR.

4. Results

4.1. More sophisticated models improve overall default prediction, but predictive power remains unequal

We begin by reestablishing some results from the literature using our models and data. We first verify that more complex models such as XGBoost perform better in out-of-sample prediction than simpler models such as logistic regression. Afterward, we confirm that in our setting, model performance varies between non-LMI and LMI areas even though the model is unaware of a consumer's location.

Figure 1, Panel A, shows that better models improve our ability to predict default and consequently produce better credit scores. (Later, we show that it corresponds to a 2 percent profit difference under a set of assumptions; see Section 4.6.) We document the predictive power, as measured by ROC AUC, for individual years (2004–2019). The performance of both models fluctuates over the years, dipping around the 2008 financial crisis, but overall and in all periods, XGBoost performs better than a logistic model. This confirms the first important fact from the recent literature — more advanced models produce better credit scores (e.g., Albanesi and Vamossy, 2019; Fuster et al., 2021; and Blattner and Nelson, 2021).

The second important result from the literature is that credit scores do not perform equally well at predicting default for different groups. Figure 1, Panels B and C, show the out-of-sample predictive power separately for non-LMI and LMI tract consumers. We see that both models perform better for non-LMI consumers.

A crucial point is that the disparities in predictive power occur despite the fact that consumer geography or socio-demographic characteristics aren't directly considered by the models. While a credit score is neutral in the sense that it is not based on protected attributes, there is a large gap in how useful the scores are for predicting repayment between the groups. Thus, we confirm that the results of Blattner and Nelson (2021) hold in the context of general credit scoring and non-LMI/LMI tract consumer groups.

4.2. Inequality can be reduced by setting separate lending thresholds: Preliminaries

Having documented the gap in predictive power, we turn our attention to mitigating it. We choose a specific metric to equalize between groups (TPR), implement a procedure based on ML literature to reduce the disparities in that metric by choosing group-specific decision thresholds, and explore the trade-offs that arise from the point of view of the consumer, lender, and regulator.

Until now, we have examined the predictive power of continuous risk scores. However, when determining whether a loan application will be approved, what matters the most is whether a consumer is below or above a set threshold, which is a binary label. From now on, we will focus on the predictive power of the binary label, *good or bad credit,* which we define explicitly next. We focus on the credit origination decision for a generic loan, so we assign the label at the consumer–quarter level.

We consider a hypothetical lender that predicts whether the consumer will default on a loan or not using one of the credit scores (XGBoost or logistic). We normalize the credit scores into percentiles $\{0,1,2,...,100\}$ that are decreasing in probability of default. At a credit score of zero, the lender has the most certainty that the consumer will default, while at a score of 100, the lender has the most certainty that the consumer will pay back the loan. Paying back the loan (not defaulting) is the positive outcome. The lender picks a credit score threshold and lends if the consumer has a credit score above the threshold. The true default label becomes known in the future when the consumer either repays or defaults. We assume that the repayment behavior on the observed lines of credit reflects the repayment behavior on the hypothetical lines of credit. Under this assumption, even if our hypothetical lender does not lend to the consumer, we can still infer the consumer's true label.[21] We also assume that the counterfactual loans are all equal in their terms and that granting an additional loan doesn't change the probability of default (and consequently the label).[22]

Every threshold creates a confusion matrix. Figure 2 is a hypothetical example. Suppose we have 100 consumers and pick a threshold of 60 (the first matrix on Figure 2). Given this threshold, we predict that 50 consumers will repay the loan. Of this 50, 40 do repay the loan (TPs) and 10 default (FPs). For the 50 consumers who we predicted would not repay their loan, 30 of them do not repay the loan (TNs) and 20 of them do repay the loan (FNs). In the second confusion matrix in Figure 2, we show how adjusting the threshold impacts all four cells in the confusion matrix. Lowering the threshold lowers the barrier to receiving loans, so more consumers receive loans when the threshold is lowered to 40. (Instead of 50 consumers receiving loans, 70 now receive loans.) These additional 20 consumers increase the number of TPs but also the number of FPs. This also means that fewer consumers are denied loans — instead of 50 consumers being predicted to default, only 30 are now predicted to do so. The number of FNs has decreased, but so has the number of FPs. Ideally, we want to maximize the number of TPs and TNs while minimizing the number of FPs and FNs. However, in practice, non-defaulters and defaulters are not perfectly separated in the credit score space. Therefore, no threshold

[21] The issue of outcomes being only observed conditional on the selection decision is common and arises in areas like child protection (Chouldechova et al. 2018), health care (Mullainathan and Obermeyer 2021) and the judicial system (Arnold et al. 2022). Unobserved confounders affecting both selection and outcomes can induce bias and affect, among other things, the estimates of disparities. In lending, such a missing data problem is commonly referred to as reject inference. The solutions available to researchers are limited and often require some kind of quasi-experimental variation, such as the random assignment of judges to the cases (Arnold et al., 2022). Our approach is most similar to Blattner and Nelson (2021), who also use performance on observed loans as a proxy for performance on counterfactual loans. In industry practice, lenders periodically extend credit to individuals below their typical lending threshold to estimate the number of FPs at lower thresholds. Caro and Nelson (2023) discuss other ways the lenders can perform reject inference if required to assess counterfactual outcomes.

[22] This is a simplification that is typical for the generic credit scoring context. For discussion of issues in the mortgage context where it is relatively more important due to large monthly payments, see, for example, Fuster and Willen (2017).

change, whether an increase or a decrease, can increase TP without also increasing FP. This is the basis of the fairness–profit trade-off we discuss next.

The set of confusion matrices at every possible threshold provides the raw materials from which various metrics of binary classifier performance can be constructed.
We focus on true positive rates (TPR) and false positive rates (FPR) because of their importance from regulatory and business perspectives. TPR, defined as
$TPR = TP/(TP + FN)$, is the percentage of good-credit consumers who are assigned the good-credit label. Maximizing this is important for a regulator who seeks to maximize credit access to creditworthy consumers. It is also important for the lender because loans to good credit consumers are profitable. (We assume that the regulator cares about reducing the difference between the TPR of LMI and non-LMI tract groups; see next.) FPR, defined as $FPR = FP/(FP + TN)$, is the percentage of bad-credit consumers who are mistakenly assigned a good-credit label. This measure is especially important from the lenders' perspective since more money is lost when a consumer defaults on a loan than is gained when the loan is repaid. (In the main specification, we assume that the losses from default are larger than gains from a repaid loan by a factor of 5, so that $Pr = TP - \lambda \times FP$.)

4.3. Objectives of lenders and the regulator differ

As discussed in Section 3.1.1, we assume that lenders set lending thresholds to maximize profit, defined as $Pr = TP - \lambda \times FP$. This means that a successful loan gives the lender one unit of money, and a loan that defaults costs the lender λ units of money. We set λ to be 5 in our main specification based on our previously discussed calibration exercise.

We assume that lenders are regulated by a government agency ("*regulator*") that values equal credit access for consumers who have the ability to repay their loan. This is operationalized as the difference in TPR between the non-LMI and LMI tract groups, $\Delta TPR = TPR(non\text{-}LMI) - TPR(LMI)$. Because the regulator also values the ongoing viability of the lender, it might accept a partial reduction in ΔTPR that results in a smaller reduction in lending profit instead of requiring ΔTPR to be zero.

We quantify the trade-off between fairness and lender profits by considering four possible levels of fairness constraints. The benchmark for comparison is setting the thresholds in a way that is blind to non-LMI or LMI status (profits are empirically the largest in this case). The *strong fairness constraint* corresponds to setting separate thresholds for the groups in such a way that ΔTPR is 0 in-sample. *Medium* and *weak fairness constraints* involve setting thresholds that are located 66 percent and 33 percent of the distance between the strong fairness constraint threshold and the blind threshold, respectively. In all cases, the threshold for the non-LMI group remains the same as in the single-threshold scenario. The details of the procedure are described in Section 4.5.

While we assume that lenders are strictly profit driven, there are, of course, reasons why a lender might value equalizing TPRs between the non-LMI and LMI tract groups. Such reasons include being a mission-oriented organization, avoiding potential fair-lending violations, or satisfying CRA requirements. In such instances, a lender would be willing to lend to more would-be defaulters in LMI groups than would be expected otherwise. Another way of thinking of it is that lenders that value fairness will face lower hurdles to achieving fairness goals with technological improvements. Thus, our results may be interpreted as upper bounds to the fairness–profit trade-off.

4.4. One threshold doesn't fit all

Under the current policy, lenders are prohibited (with some exceptions) from using consumer demographics in most lending scenarios. This corresponds to using a group membership-blind model and picking a single credit score threshold for all consumers. On the surface, it is a neutral policy intended to reduce discrimination. However, this approach affects non-LMI and LMI consumers differently.

We simultaneously visualize the TPR and FPR for all lending thresholds and all consumers using the XGBoost credit score based on the in-sample data of the last rolling window (2014Q1–2015Q4, see Appendix B) in Figure 3, Panel A. This figure is similar to a ROC curve in that it includes TPR and FPR, but unlike a ROC curve, the figure makes threshold values explicit by plotting them on the *x* axis. The two lines represent the two rates. The y-axis gives the corresponding TPR and FPR for each threshold. We see that with a threshold of 0, every consumer receives a loan. This means that we correctly give a loan to every consumer who is indeed creditworthy. However, this means that we also give a loan to every consumer who is not creditworthy. On the other extreme, a threshold of 100 means that no consumer receives a loan. Consumers who are not creditworthy are denied a loan, as are consumers who are creditworthy. The optimal threshold is somewhere in the middle, and we pick one that maximizes lender profits. The figure shows the threshold that maximizes simulated profit, under which 86 percent of creditworthy consumers get the loan; 21 percent of defaulters do, as well.

Figure 3, Panel B, has the same *x* and *y* axes as Panel A but breaks down TPR and FPR by non-LMI and LMI tract consumers. We see that at the optimal threshold, non-LMI-tract consumers have a substantially higher TPR than LMI-tract consumers. In non-LMI tracts, 88 percent of creditworthy customers are approved for a loan, but only 79 percent of creditworthy customers in LMI tracts are approved. In the terminology of Hardt et al. (2016), the difference in TPR between groups measures "equality of opportunity."

4.5. Tailoring default predictions via separate thresholds can reduce inequality, but at a cost

In this section, we discuss the introduction of group-specific lending thresholds and the associated fairness–profit trade-offs arising in our illustrative model.

To set the separate thresholds, we modify the approach from Hardt et al. (2016). We keep the threshold for the non-LMI group the same as in the single-threshold case and set the threshold for the LMI group in a way that the difference in TPR is as close to zero as possible.[23]

---

[23] This is a modification of the equal opportunity thresholds from Hardt et al. (2016). The original approach considers all pairs of thresholds that equalize TPR and picks one that maximizes some objective (in our case lender profits). In that case, compared with using a single threshold, the separate threshold for the non-LMI group is slightly higher, and the threshold for the LMI group is lower. We choose to limit our analysis to the case in which the outcomes of the non-LMI group remain the same as under the single-threshold policy to avoid a decrease in TPR for any group of consumers.

This approach lends itself very easily to relaxation. We can adjust the thresholds depending on how much the regulator values fairness relative to the business need to maximize profits. We do so by picking thresholds for the LMI group between the single threshold and the threshold that would eliminate differences in TPR. Any objective weight can be accommodated. For simplicity, we focus on three possible levels of fairness constraint. First, we determine the threshold that eliminates the difference in TPR between non-LMI and LMI individuals. We call this threshold a *strong fairness constraint*. We also generate LMI thresholds that are 66 percent and 33 percent of the way between the single threshold and the $\Delta TPR = 0$ threshold and call them *medium* and *weak fairness constraints*. The thresholds are generated on the rolling window basis. Figure 4 shows the TPR and FPR for strong (Panel A), medium (Panel B), and weak (Panel C), as well the single threshold (Panel D), based on the last rolling window.

In the case of the strong fairness constraint, both groups have a TPR of approximately 88 percent. This way, creditworthy consumers have an 88 percent chance to get the loan regardless of which group they are in.[24]

More creditworthy LMI consumers are classified as good credit under all fairness constraints than under a single threshold. However, Figure 4 shows that lowering the threshold for the LMI group increases its FPR. This is important from the lenders' perspective since the number of FP enters the lenders' objective function with a multiplier of $\lambda > 1$. This is key for the fairness–profit trade-off.

4.6. Best of both worlds: Linking fairness constraints to model improvement

In Section 4.1, we showed that better credit scoring models improve the accuracy of default prediction for both non-LMI and LMI groups, but the gaps in model performance between the non-LMI and LMI groups remain. In Section 4.5, we showed how we can establish separate lending thresholds to reduce disparities in equality of opportunity. Now

[24] The slight difference in TPR on the plot is due to us focusing on integer thresholds and picking the LMI threshold with the smallest absolute difference between groups.

we will look at the fairness–profit trade-offs that arise from this approach and how the trade-offs interact with model improvement.

The analysis in this section is based on fairness constraints applied out-of-sample to the test sets. As described in the preceding section, we apply four different thresholds to the resulting score — profit maximizing and three degrees of fairness constraint. All thresholds are selected using training data and applied out-of-sample to obtain out-of-sample binary predictions. We then combine all eight test sets to generate the results that follow.

Figure 5 illustrates the fairness–profit trade-offs at different levels of fairness constraints and for different models. The x-axis is profit, calculated as $Pr = TP - \lambda \times FP$ and normalized so that the baseline profit of the logistic model is equal to 100. In our main specification, we set λ to 5 based on the calibration exercise described in Section 3.1.1. The y axis is the difference in TPR between the LMI group and the non-LMI group and is our measure of equality of opportunity (ΔTPR). The color is the model type, and the label is the strength of the constraint. In Figure 6, we compare results for $\lambda \in \{3,4,5\}$. Finally, Table 4 shows the robustness of our results to $\lambda \in [1,11]\}$, the range motivated by our calibration exercise discussed in Section 3.1.1.

We see that for every model in our illustrative setting, making the fairness constraint stronger reduces profits — this is the fairness–profit trade-off. For the XGBoost model, it costs about 1 percent of profits to eliminate the TPR gap. So, fairness doesn't come free and affects lenders. The degree to which increased fairness affects profit is indicated by the slope of the line.

We also observe that improvements in modeling technology shift the fairness–profit curve rightward, increasing profitability at every threshold. This is intuitive, since a better-performing model approves fewer defaulters and more creditworthy consumers. Thus, adopting a more sophisticated model can improve profitability at every threshold level.

The combined effects of fairness constraints and model improvement suggest a way forward that blends the best of both worlds. If a lender using a particular model were to adopt group-specific lending thresholds in the absence of a significant improvement in model quality, the lender would become less profitable. However, a lender that simultaneously adopts both more sophisticated modeling technology and group-specific thresholds could experience increases in both fairness and profit. An XGBoost model with the strong fairness constraint generates more profit for the lender than does a logistic model without a fairness constraint. This observation allows us to revisit a major result from the recent literature. Papers such as Fuster et al. (2021) and Blattner and Nelson (2021) argue that better models improve credit scoring accuracy but do little to reduce inequality. However, Figure 5 shows that this result crucially depends on the sensitive attribute blindness requirement for the credit score threshold. While well intentioned, blindness prevents the regulator from introducing fairness constraints that tackle disparities head-on.

If we consider an alternative policy that requires lenders to consider the sensitive attributes in a way that is designed to promote fairness, we get alternative characterizations of how model improvement affects fairness. For example, if the regulator places a high weight on fairness and requires the improvements in credit score technology to be paired with fairness constraints, then introducing a new credit scoring model can lead to a very large improvement in fairness combined with a more modest increase in our measure of profits, $TP - 5 \times FP$. In our illustrative setting, going from the logistic model with a single threshold to the XGBoost model with a strong fairness constraint decreases the TPR gap from 9 percent to 0 percent while increasing profits from 100 to 100.4. If the weight the regulator places on fairness is a bit lower, but the weight on business need to maximize profits is larger, fairness improvements are lower but profits are larger (up to the maximum profit of 101.8). Table 3 confirms the robustness of this result to a range of λ assumptions. Only at higher $\lambda > 9$, uncommon in our calibration sample, does the profit of XGBoost with a strong fairness constraint fall under 100. (The value is 99.7 in the case of $\lambda = 11$.) The XGBoost profit under the medium fairness constraint is always higher than 100 (101.2 when $\lambda = 11$).

In addition to the effect on lenders, fairness constraints also affect consumers. We base our analysis of winners and losers among the consumers on the confusion matrix. We consider the TP group to be winners. (They get a loan they can pay back.) Conservatively, we consider the FP group to be losers. (They are more likely to suffer default on the loan they might get due to a lower threshold, even though ex ante they might still prefer to get the loan.) The TN group is considered to be neutral (as they don't get a loan they can't pay back), and the FN group is considered to be losers (as they don't get a loan they can pay back).

We highlight that both winner and loser groups increase after the introduction of fairness constraints using an XGBoost credit score as an example. In Figure 7, the x-axis is the strength of the constraint and the y-axis is the percentage of the non-LMI or LMI group belonging to the TP, FP, TN, or FN category, depicted as different lines. By construction, the composition of the non-LMI group doesn't change when fairness constraints are introduced.[25] All changes are among the LMI tract individuals. In the single threshold scenario, 65.3 percent of the LMI population are TP: people with good credit correctly predicted to be good credit. Under strong fairness constraints, this number goes up to 72.7 percent. By definition, the increase comes from the decrease in the FN group, reducing the number of losers. However, the FP percent for the LMI group also rises, from 3.9 percent to 6.2 percent. This means that more consumers are more likely to get loans they might have trouble paying back. By definition, this increase comes from the decrease in the TN group who are neutral since they don't benefit from a credit score qualifying them for a loan, but they also are not hurt by the potential consequences of defaulting on more loans. While many more consumers benefit from a fairness constraint than are hurt by it, the increase in the FP group needs to be taken into consideration.

Notably, our approach to labeling consumers as winners or losers is more conservative than that in Fuster et al. (2021), who treat consumers as winners from a model change if their credit score goes up even if they are consumers with a higher default probability. We take the more conservative approach for two reasons. First, it's important for the policymaker to consider the costs of increased credit access as well as the benefits. Second,

[25] This is slightly different from the original equality of opportunity approach in Hardt et al. (2016). There, the separate thresholds are picked in a way such that the threshold for the LMI is lower (as it is in our case) and the threshold for the non-LMI group is slightly higher, resulting in a slightly higher percentage of TP and a slightly lower percentage of FP in the non-LMI group, as well as slightly higher lender profits.

and most important, we want to emphasize that the trade-off is favorable even when the increase of FP is treated as a cost to consumers, because many more creditworthy consumers get access to credit.

5. Conclusion

The technological advancements in underwriting can benefit lenders significantly. With appropriate policy guidance, these advancements can also bring substantial fair lending benefits. The gap in TPR between non-LMI and LMI areas can be reduced by adopting group-specific thresholds. However, this equality comes with a cost to lender profits. Using more complex models in conjunction with introducing separate thresholds can help to mitigate these losses. We describe a trade-off that needs to be appropriately managed rather than a first-best solution. However, we think that if the trade-off is managed appropriately, incentives can change in a way such that both fairness and profits can improve over time as lenders invest more into reducing the data disparities between the non-LMI and LMI groups underlying the predictive power gap (Blattner and Nelson, 2021).

References

Agarwal, S., Chomsisengphet, S., Mahoney, N., Stroebel, J., 2014. Regulating consumer financial products: Evidence from credit cards. *Quarterly Journal of Economics* 130, 111–164. doi:10.1093/qje/qju037.

Albanesi, S., Vamossy, D.F., 2019. Predicting consumer default: A deep learning approach. Working paper.

Arnold, D., Dobbie, W., Hull, P., 2022. Measuring racial discrimination in bail decisions. *American Economic Review* 112, 2992–3038.

Avery, R.B., Brevoort, K.P., Canner, G., 2012. Does credit scoring produce a disparate impact? *Real Estate Economics* 40, 65–114.

Avery, R.B., Calem, P.S., Canner, G.B., Bostic, R.W., 2003. An overview of consumer data and credit reporting. Federal Reserve Bulletin, 47–73.

Bartlett, R., Morse, A., Stanton, R., Wallace, N., 2022. Consumer-lending discrimination in the fintech era. *Journal of Financial Economics* 143, 30–56.

Bayer, P., Ferreira, F., Ross, S.L., 2016. The vulnerability of minority homeowners in the housing boom and bust. *American Economic Journal: Economic Policy* 8, 1–27.

Bell, A., Bynum, L., Drushchak, N., Zakharchenko, T., Rosenblatt, L., Stoyanovich, J., 2023. The possibility of fairness: Revisiting the impossibility theorem in practice, in: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery*, New York, NY, USA. p. 400–422. https://doi.org/10.1145/3593013.3594007.

Bhutta, N., Hizmo, A., Ringo, D., 2021. How much does racial bias affect mortgage lending? Evidence from human and algorithmic credit decisions. Working paper.

Blascak, N., Cheney, J., Hunt, R., Mikhed, V., Ritter, D., Vogan, M., 2018. Financial consequences of severe identity theft in the U.S. Working paper.

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

Blattner, L., Nelson, S., Spiess, J., 2023. Unpacking the black box: Regulating algorithmic decisions. arXiv:2110.03443.

Breiman, L., 2001. Random forests. *Machine Learning* 45, 5–32.

Bronson, A., Nadauld, T., Palmer, C., 2019. Real effects of search frictions in consumer credit markets. Working paper.

Caro, S., Nelson, S., 2023. The arity of disparity: Updating disparate impact for modern fair lending. Working paper.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, NY, USA. pp. 785–794.

Chouldechova, A., Benavides-Prado, D., Fialko, O., Vaithianathan, R., 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions, in: Friedler, S.A., Wilson, C. (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, PMLR. pp. 134–148. URL: https://proceedings.mlr.press/v81/chouldechova18a.html.

Chouldechova, A., Roth, A., 2018. The frontiers of fairness in machine learning. Working paper.

Edelberg, W., 2006. Risk-based pricing of interest rates for consumer loans. *Journal of Monetary Economics* 53, 2283–2298.

Einav, L., Jenkins, M., Levin, J., 2013. The impact of credit scoring on consumer lending. *RAND Journal of Economics* 44, 249–274.

Elkan, C., 2001. The foundations of cost-sensitive learning, in: *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. pp. 973–978.

Fawcett, T., 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27, 861–874. doi:10.1016/j.patrec.2005.10.010.

Fedaseyeu, V., Hunt, R., 2018. The economics of debt collection: Enforcement of consumer credit contracts. Federal Reserve Bank of Philadelphia Working Paper 18-04.

FinRegLab, 2021. The use of machine learning for credit underwriting. Report.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33, 1–22.

Fulford, S., Nagypal, E., 2023. Using the courts for private debt collection: How wage garnishment laws affect civil judgments and access to credit. Consumer Financial Protection Bureau Office of Research Working Paper 23-02. URL: https://ssrn.com/abstract=4394821.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., Walther, A., 2021. Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance* 77, 5–47.

Fuster, A., Willen, P.S., 2017. Payment size, negative equity, and mortgage default. *American Economic Journal: Economic Policy* 9, 167–91. URL: https://www.aeaweb.org/articles?id=10.1257/pol.20150007.

Gerardi, K., Lambie-Hanson, L., Willen, P., 2021. *Racial Differences in Mortgage Refinancing, Distress, and Housing Wealth Accumulation during COVID-19*. Federal Reserve Bank of Boston Current Policy Perspectives.

Gerardi, K., Willen, P.S., Zhang, D.H., 2023. Mortgage prepayment, race, and monetary policy. *Journal of Financial Economics* 147, 498–524.

Gillis, T., 2022. The input fallacy. *Minnesota Law Review*.

Hardt, M., Price, E., Srebro, N., 2016. Equality of opportunity in supervised learning, in: *Proceedings of the 30th International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA. p. 3323–3331.

Hastie, T., Tibshirani, R., Friedman, J., 2017. *The elements of statistical learning: Data mining, inference, and prediction. 12th ed.*, Springer.

Hoerl, A.E., 1962. Application of ridge analysis to regression problems. *Chemical Engineering Progress* 58, 54–59.

Hothorn, T., Zeileis, A., 2015. partykit: A modular toolkit for recursive partitioning in r. *Journal of Machine Learning Research* 16, 3905–3909. URL: http://jmlr.org/papers/v16/hothorn15a.html.

Hunt, R., Wardrip, K., 2013. Residential migration, entry, and exit as seen through the lens of credit bureau data. Federal Reserve Bank of Philadelphia Payment Cards Center discussion paper 13-04.

Hurlin, C., Perignon, C., Saurin, S., 2022. The fairness of credit scoring models. arXiv:2205.10200.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2023. An Introduction to Statistical Learning. 2nd ed., Springer.

Jarmin, R.S., O'Hara, A.B., 2016. Big data and the transformation of public policy analysis. *Journal of Policy Analysis and Management* 35, 715–721.

Kearns, M., Neel, S., Roth, A., Wu, Z.S., 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness, in: Dy, J., Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning, PMLR. pp. 2564–2572.

Kermani, A., Wong, F., 2021. Racial Disparities in Housing Returns. Working Paper 29306. National Bureau of Economic Research. URL: http://www.nber.org/papers/w29306, doi:10.3386/w29306.

Keys, B.J., Mukherjee, T., Seru, A., Vig, V., 2010. Did Securitization Lead to Lax Screening? Evidence from Subprime Loans. *Quarterly Journal of Economics* 125, 307–362.

Kleinberg, J., Ludwig, J., Mullainathan, S., Rambachan, A., 2018. Algorithmic fairness. *AEA Papers and Proceedings* 108, 22–27.

Kleinberg, J.M., Mullainathan, S., Raghavan, M., 2016. Inherent trade-offs in the fair determination of risk scores, in: *Information Technology Convergence and Services*. URL: https://api.semanticscholar.org/CorpusID:12845273.

Kovner, A., Van Tassel, P., 2021. Evaluating regulatory reform: Banks' cost of capital and lending. *Journal of Money, Credit and Banking* 54, 1313–1367. URL: http://dx.doi.org/10.1111/jmcb.12875.

Kuhn, M., Wickham, H., 2020. Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles.

Lamba, H., Rodolfa, K.T., Ghani, R., 2021. An empirical comparison of bias reduction methods on real-world problems in high-stakes policy settings. *SIGKDD Explorations Newsletter* 23, 69–85.

Lambie-Hanson, L., Reid, C., 2018. Stuck in subprime? examining the barriers to refinancing mortgage debt. *Housing Policy Debate* 28, 770–796.

Lane, J., 2016. Big data for public policy: The quadruple helix. *Journal of Policy Analysis and Management* 35, 708–715.

Laufer, S., Paciorek, A., 2022. The effects of mortgage credit availability: Evidence from minimum credit score lending rules. *American Economic Journal: Economic Policy* 14, 240–76.

Lee, D., van der Klaauw, W., 2010. An introduction to the FRBNY consumer credit panel. Staff Reports 479, Federal Reserve Bank of New York.

Liu, L.T., Dean, S., Rolf, E., Simchowitz, M., Hardt, M., 2019. Delayed impact of fair machine learning, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization. pp. 6196–6200.

Mikhed, V., Vogan, M., 2018. How data breaches affect consumer credit. *Journal of Banking and Finance* 88, 192–207.

Mullainathan, S., Obermeyer, Z., 2021. Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care. *Quarterly Journal of Economics* 137,679–727. URL: https://doi.org/10.1093/qje/qjab046.

Munnell, A.H., Tootell, G.M.B., Browne, L.E., McEneaney, J., 1996. Mortgage lending in Boston: Interpreting HMDA data. *American Economic Review* 86, 25–53. URL: http://www.jstor.org/stable/2118254.

Rodolfa, K.T., Saleiro, P., Ghani, R., 2016. Bias and fairness (in machine learning), in: *Big data and social science: A practical guide to methods and tools*. Chapman and Hall/CRC Press.

Servon, L., 2017. Are payday loans harmful to consumers? *Journal of Policy Analysis and Management* 36, 240–248.

Skanderson, D., Ritter, D., 2014. Fair lending analysis of credit cards. Working paper.

Sodini, P., Van Nieuwerburgh, S., Vestman, R., von Lilienfeld-Toal, U., 2023. Identifying the benefits from homeownership: A Swedish experiment. *American Economic Review*

113, 3173–3212. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20171449, doi:10.1257/aer.20171449.

Tibshirani, R., 1994. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Series B 58, 267–288.

Van Meter, D., Middleton, D., 1954. Modern statistical approaches to reception in communication theory. *Transactions of the IRE Professional Group on Information Theory* 4, 119–145.

Tables

Note: Unless otherwise stated, all tables are based on authors' calculations using CCP data, with consumers in low- and moderate-income (LMI) census tracts identified using the dataset produced by the FFIEC, based on U.S. Census Bureau data.

Table 1: Observations by current delinquency status, the CRA status of their census tract, and repayment outcome. The calculations are based on a 1 percent sample of the CCP. Individuals are observed at the quarterly frequency; % Default column indicates the stock of defaulters in the given group. We provide the overall default statistic and split the results by CRA status, current delinquency status, and both. (The % observations in each of these splits sum to 100 percent.) Individuals in low and moderate-income (LMI) census tracts are identified using a dataset produced by the FFIEC, based on U.S. Census Bureau data. *Default* is defined as being more than 90 days past due on any credit account within a two-year period starting with the next quarter.

Consumer Group	% of Observations	% Default (Next 2 Y.)
Overall	100.0	22.3
Census Tract Status Non-LMI	79.9	19.4
LMI	20.1	33.8
Current Delinquency Status		
Current	84.4	8.5
<90 Days Past Due	2.8	59.9
≥90 Days Past Due	12.8	93.4
Census Tract Status & Current Delinquency Status		
Non-LMI & Current	68.1	7.8
Non-LMI & <90 Days Past Due	2.3	58.0
Non-LMI & ≥90 Days Past Due	9.5	93.3
LMI & Current	14.9	14.0
LMI & <90 Days Past Due	0.7	66.6
LMI & ≥90 Days Past Due	4.5	94.6

Table 2: Average λ values for seven large financial institutions. The λ value is the ratio of the average cost of default to the average five-year profit earned on non-defaulting loans. The weighted mean is the average λ weighted by total credit card assets. The calculations are based on the account- and portfolio-level Y-14M administrative data for January 2014 to December 2022.

	2014	2015	2016	2017	All
Unweighted mean	5.00	4.76	4.96	5.85	5.14
Weighted Mean	5.53	5.15	5.05	5.81	5.39
Std Dev	3.16	2.02	1.84	2.43	2.32
Num Obs	7	7	7	7	28

Table 3: Fairness–profit trade-offs at different levels of fairness constraints and λ (profit trade-off of good to bad loans). Profit is calculated as $Pr = TP - \lambda \times FN$, where TP is the number of true positives, FP is the number false positives, and λ corresponds to the monetary loss associated with a loan that is not repaid relative to the gain from a loan that is repaid. We normalize profit by the profit of the logistic model with no fairness constraints. The positive outcome is non-default within the next two years. Fairness, or ΔTPR, is the difference in TPR between the LMI and non-LMI groups. TPR is defined as $\frac{TP}{TP+FN}$, where TP is the true positives and FN is the false negatives. Individuals in low- and moderate-income (LMI) census tracts are identified using a dataset produced by the FFIEC, based on U.S. Census Bureau data. The labels Strong, Medium, Weak, and Blind represent different strengths of the fairness constraint. (See Section 4.5 for definitions.)

			λ					
		Policy	1	3	5	7	9	11
Fairness								
	Log	Blind	-1.9	-6.5	-8.9	-10.6	-11.6	-12.5
		Weak	-1.3	-4.2	-5.8	-6.9	-7.7	-8.2
		Medium	-0.6	-1.9	-2.9	-3.3	-3.6	-4
		Strong	0	0.3	0.2	0.3	0.4	0.3
	XGB	Blind	-1.6	-6	-8.7	-10.4	-11.4	-12.3
		Weak	-1.1	-4	-5.6	-6.8	-7.5	-8
		Medium	-0.5	-1.8	-2.8	-3.4	-3.5	-4
		Strong	0	0.1	0.2	0.1	0.4	0.2
Profit								
	Log	Blind	100	100	100	100	100	100
		Weak	100	99.9	99.8	99.8	99.7	99.7
		Medium	100	99.7	99.4	99.2	99	98.9
		Strong	99.9	99.3	98.8	98.3	98	97.6
	XGB	Blind	100.5	101.1	101.8	102.2	102.4	102.6
		Weak	100.5	101	101.5	101.9	102	102.1
		Medium	100.4	100.8	101.1	101.3	101.2	101.2
		Strong	100.4	100.4	100.4	100.3	99.9	99.7

Figures

Note: Unless otherwise stated, all figures are based on authors' calculations using CCP data with consumers in low- and moderate-income (LMI) census tracts identified using the dataset produced by the FFIEC, based on the U.S. Census Bureau data.

Figure 1: Performance of repayment status predictions on the combined evaluation set, 2004Q1–2019Q2. ROC AUC is a measure of binary classifier performance running between 0.5 and 1 (the more, the better) that accounts for the true positive rate (how many non-defaulters are correctly identified as such) and the false positive rate (how many defaulters are incorrectly identified as non-defaulters) at every possible decision threshold. The two lines correspond to the logistic and XGBoost models. Individuals in low- and moderate-income (LMI) census tracts are identified using a dataset produced by the FFIEC, based on U.S. Census Bureau data. *Default* is defined as being more than 90 days past due on any credit account within a two-year period starting with the next quarter. For example, observations in 2019 use data up to 2021 to compute the default variable. All metrics are calculated fully out of sample from each model's training period.

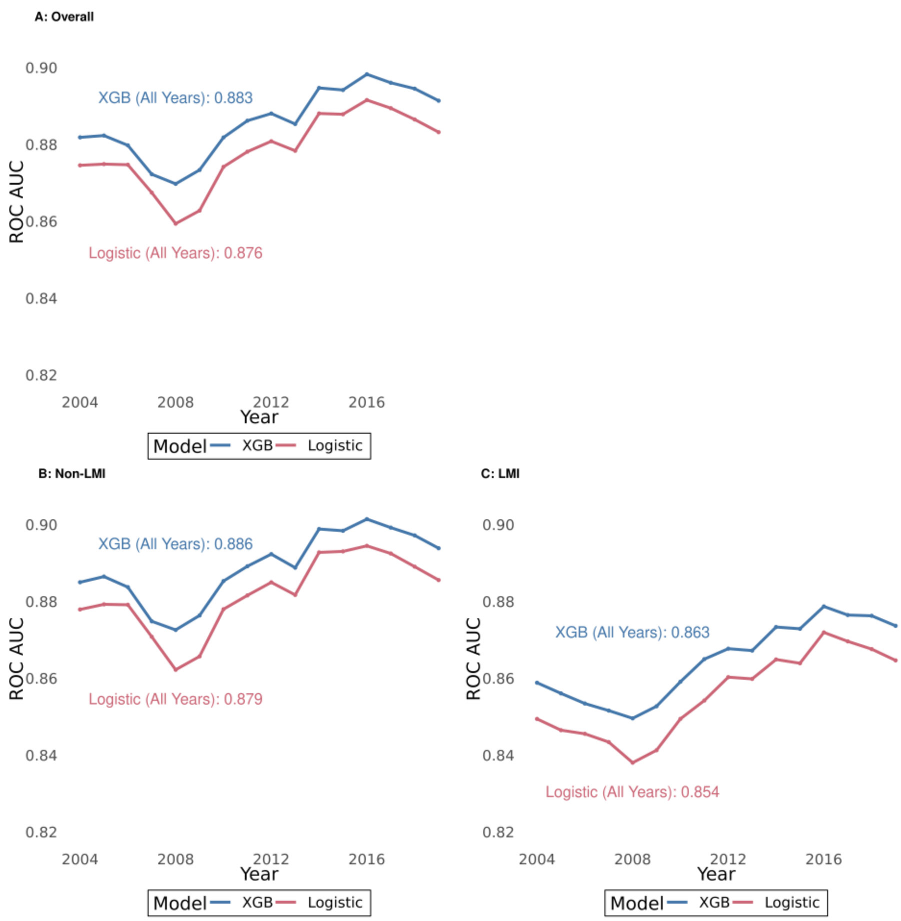


Figure 2: A hypothetical confusion matrix of 100 individuals. Thresholds can take values between 0 and 100. At the 0 threshold, everyone is predicted to be positive (not in default within two years); at the 100 threshold, everyone is predicted to be negative (in default within two years). The cells correspond to (left to right, top to bottom): true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). Changing the threshold values changes values in all four cells.

Threshold: 60	Actual Positive	Actual Negative
Predicted Positive	40	10
Predicted Negative	20	30

Threshold: 40↓	Actual Positive	Actual Negative
Predicted Positive	50↑	20↑
Predicted Negative	10↓	20↓

Figure 3: Comparison of true positive rates (TPR) and false positive rates (FPR), single-threshold approach. Threshold values 0 to 100 correspond to percentiles of model outputs. The vertical line represents the single profit-maximizing threshold. *TPR* is defined as $\frac{TP}{TP+FN}$, where TP is the true positives and FN is the false negatives. *FPR* is defined as $\frac{FP}{FP+TN}$, where FP is the false positives and TN is the true negatives. Individuals in low- and moderate-income (LMI) census tracts are identified using a dataset produced by the FFIEC, based on U.S. Census Bureau data. *Default* is defined as being more than 90 days past due on any credit account within a two-year period starting with the next quarter. This plot is based on the XGBoost model and the training set of the last rolling window, 2014Q1–2015Q4.

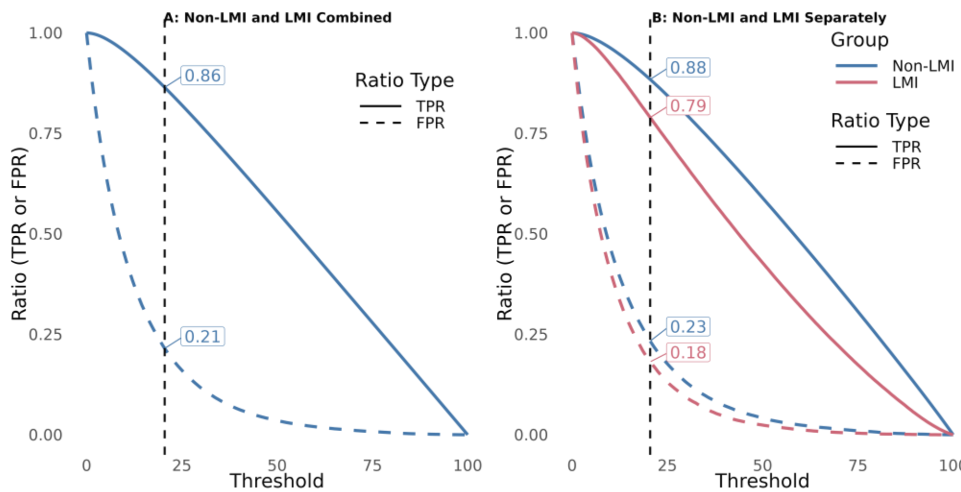


Figure 4: Comparison of true positive rates (TPR) and false positive rates (FPR), group-specific threshold approach. Threshold values 0 to 100 correspond to percentiles of model outputs. The vertical dotted lines represent the group-specific profit-maximizing thresholds. (Panels A, B and C, respectively, depict strong, medium and weak fairness constraints; see Section 4.5 for definitions.) The vertical line on Panel D represents the group-unaware single profit-maximizing threshold. *TPR* is defined as $\frac{TP}{TP+FN}$, where TP is the true positives and FN is the false negatives. *FPR* is defined as $\frac{FP}{FP+TN}$, where FP is the false positives and TN is the true negatives. Individuals in low- and moderate-income (LMI) census tracts are identified using a dataset produced by the FFIEC, based on U.S. Census Bureau data. *Default* is defined as being more than 90 days past due on any credit account within a two-year period starting with the next quarter. This plot is based on the XGBoost model and the training set of the last rolling window, 2014Q1–2015Q4.

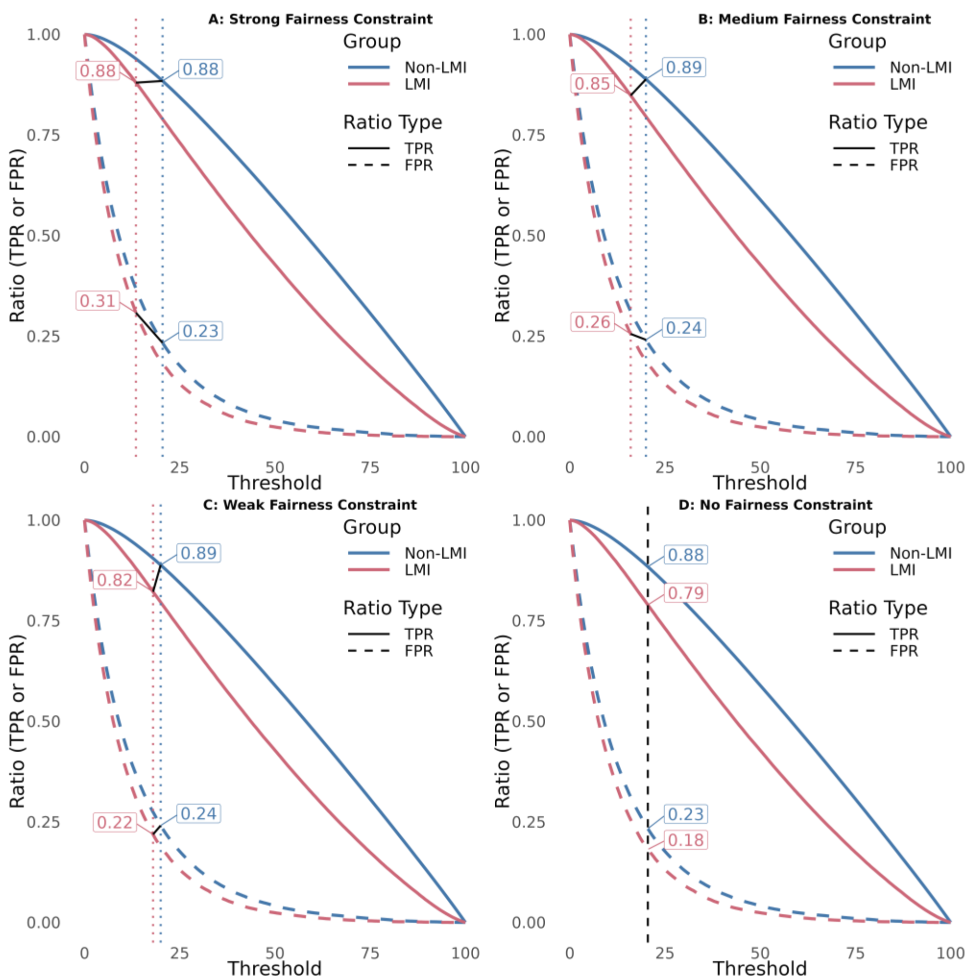


Figure 5: Fairness–profit trade-offs at different levels of fairness constraints and for the logistic and XGBoost models. Profit is calculated as $Pr = TP - 5 \times FN$,where TP is the number of true positives and FP is the number of false positives. Profit has been normalized to a 0 to 1000 scale. The positive outcome is non-default within the next two years. ΔTPR is the difference in TPR between the LMI and non-LMI groups, which is the measure of fairness we adopt in this paper. TPR is defined as $\frac{TP}{TP+FN}$, where TP is the true positives and FN is the false negatives. Individuals in low and moderate-income (LMI) census tracts are identified using a dataset produced by the FFIEC, based on U.S. Census Bureau data. Labeled points on lines represent different strengths of the fairness constraint: Strong, Medium, Weak, and Blind. (See Section 4.5 for definitions.)

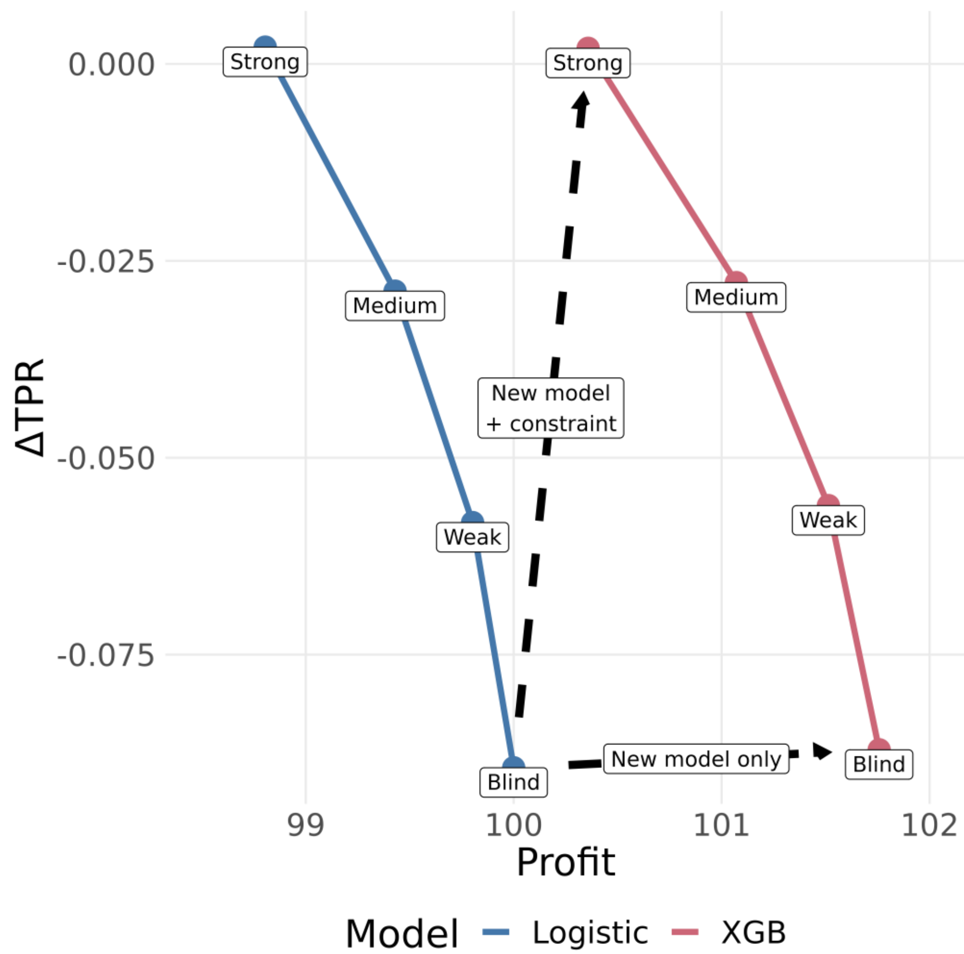


Figure 6: Fairness–profit trade-offs at different levels of fairness constraints and λ (profit trade-off of good to bad loans). Profit is calculated as $Pr = TP - \lambda \times FN$, where TP is the number of true positives, FP is the number false positives, and λ corresponds to the monetary loss associated with a loan that is not repaid relative to the gain from a loan that is repaid. We normalize profit by the profit of the logistic model with no fairness constraints. The positive outcome is non-default within the next two years. ΔTPR is the difference in TPR between the LMI and non-LMI groups, which is the measure of fairness we adopt in this paper. TPR is defined as $\frac{TP}{TP+FN}$, where TP is the true positives and FN is the false negatives. Individuals in low and moderate-income (LMI) census tracts are identified using a dataset produced by the FFIEC, based on U.S. Census Bureau data. Labeled points on lines represent different strengths of the fairness constraint: Strong, Medium, Weak, and Blind. (See Section 4.5 for definitions.)

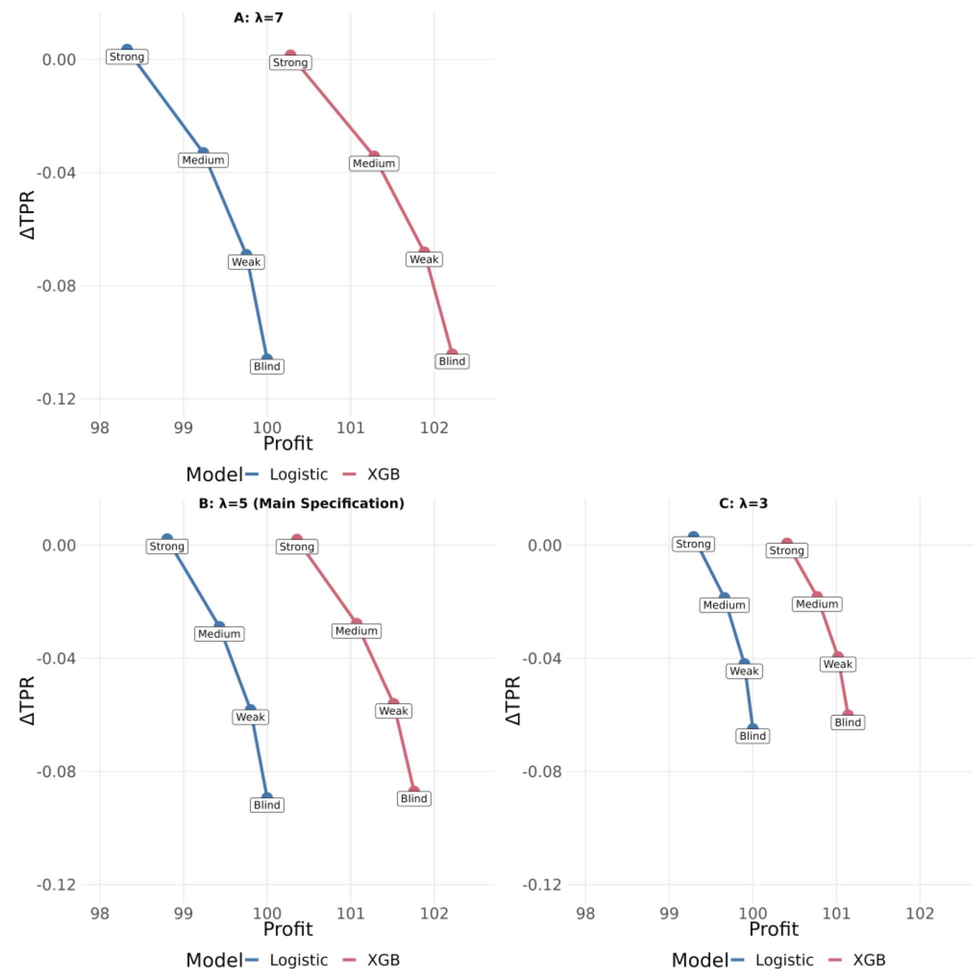
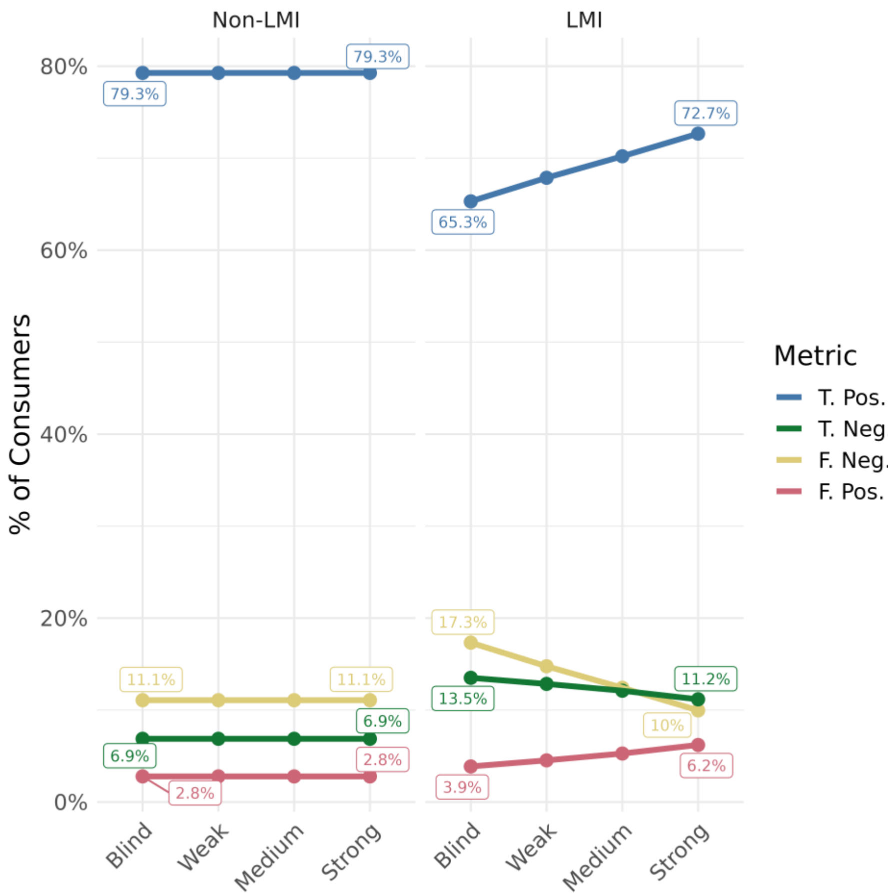


Figure 7: Winners and losers under different fairness constraints, XGBoost model. Points on the x-axis represent different strengths of the fairness constraint, with Blind corresponding to a single threshold without fairness constraints, and Strong representing the most stringent constraint aimed at eliminating ΔTPR. The lines correspond to the fractions of population in the group (non-LMI or LMI) belonging to the true positive, true negative, false negative, or false positive category. The positive outcome is non-default within the next two years. Individuals in low- and moderate-income (LMI) census tracts are identified using a dataset produced by the FFIEC, based on U.S. Census Bureau data. Labeled points on lines represent different strengths of the fairness constraint: Strong, Medium, Weak, and Blind. (See Section 4.5 for definitions.)



Appendix A. Policy considerations

Our illustrative examples suggest that applying group-specific lending thresholds in default prediction can reduce disparities in TPR between non-LMI and LMI neighborhoods. Disparities are reduced mechanically as new group-specific thresholds are introduced and are not dependent on whether the lender bases its lending decisions on a logistic regression model or a more sophisticated machine learning model. However, in either case, increased fairness comes at the cost of higher default rates in LMI neighborhoods.

We also provide evidence that lenders that simultaneously adopt both ML models and group-specific lending thresholds may experience increases in fairness as well as profit. This can occur when the ML model identifies sufficiently many creditworthy loan applicants in both the LMI and non-LMI neighborhoods (relative to the baseline model) such that it more than offsets the defaults incurred from lowering the lending threshold for LMI neighborhoods.

Under certain circumstances, borrowers benefit from the adoption of group-specific lending thresholds and more sophisticated credit risk assessment technology.[26] Group-specific lending thresholds ensure that the approval rates across neighborhood types achieve a level of parity that would not be achievable using a single lending threshold. Moreover, group-specific lending thresholds can achieve greater fairness without diminishing the approval rates enjoyed by residents of non-LMI neighborhoods under a single-threshold lending model.

The possibility of dual adoption, whereby lenders simultaneously adopt both ML credit risk models and group-specific lending thresholds, has the opportunity to establish a new trajectory for fair lending without the loss of profit that would arise from the imposition of

[26] Throughout the paper, we focus on the outcome of a lending decision in which an applicant is either approved or rejected for a loan and may subsequently terminate the loan in good standing or default. We assume that higher approval rates in non-defaulting populations make the individuals better off. Of course, one could argue that an LMI individual who is approved for a loan at significantly more onerous terms has not been made better off. Appendix F and Fuster et al. (2021) suggest that pricing implications are likely small. However, a full accounting of individuals' welfare is beyond the scope of this paper.

group-specific thresholds in isolation. However, there are some important challenges and limitations that would affect the likelihood, scale, and nature of adoption. In particular, it is unclear whether group-specific thresholds corresponding to LMI neighborhoods are broadly implementable under current fair lending law. Although the lender's objective in assigning group-specific thresholds is to increase fairness, lenders may put themselves at risk for further regulatory scrutiny or disparate impact litigation by including variables in the lending decision that are correlated with race or other protected characteristics. In addition, widespread adoption of ML models in credit underwriting has been impeded by the newness and sophistication of the technology, which has created operational, legal, and regulatory uncertainty (FinRegLab, 2021).

In the following sections, we discuss some of the legal and regulatory hurdles that a hypothetical lender might encounter when initially adopting ML credit risk models and group-specific lending thresholds for LMI and non-LMI neighborhoods. There are several reasons why we believe this discussion should be considered a hypothetical — rather than a practical — exercise. First, it is important to note that our credit scoring exercise combines data from multiple lenders. Our results are the product of market-level aggregations of consumer and lender behavior and do not necessarily reflect the experience that any one particular lender might have when implementing credit scoring or group-specific lending thresholds. While we have no reason to believe that the fairness-accuracy trade-off exhibited by our models would not be present at the lender level, our analysis does not provide sufficient evidence to rule out this possibility. Thus, we caution the reader that the following discussion rests upon the assumption that our market-level outcomes and trade-offs are representative of what lenders might observe in their own data. Second, although fairness and model accuracy in lending are two closely related topics, each falls under distinct regulatory umbrellas that may, at times, conflict with each other. An exhaustive analysis of related regulatory issues is beyond the scope of this paper. In the discussion that follows, our purpose is to highlight some of the key challenges to implementation and to shed light on a particular aspect of existing fair lending law that may prove useful for lenders and policymakers seeking to explore the topic further.

Appendix A.1. Group-specific lending thresholds

By introducing group-specific lending thresholds corresponding to LMI and non-LMI neighborhoods, our intent is to demonstrate that greater fairness can be achieved when the lender explicitly considers sensitive borrower characteristics. Residents of neighborhoods who have historically experienced barriers to credit can achieve TPRs equal to borrowers living in high-income neighborhoods who are less likely to have experienced lending discrimination and reduced access to credit. Moreover, we show that group-specific lending thresholds can achieve greater fairness without diminishing the approval rates enjoyed by residents of non-LMI neighborhoods under a single-threshold lending model.

From a fair lending perspective, it is not clear whether a lender could implement neighborhood-based lending thresholds in the credit underwriting process of a typical loan program. Fair lending laws generally prohibit lenders from favoring a particular class of borrowers in any aspect of a lending decision, even if that class has been historically discriminated against. The ECOA and its implementing Regulation B make it illegal for covered lenders to discriminate against certain classes of loan applicants and prohibit lenders from using certain personal characteristics, including race and national origin, in any aspect of a credit transaction (Skanderson and Ritter, 2014). While residence in an LMI neighborhood is not explicitly protected under the ECOA, it can be correlated with characteristics that are explicitly prohibited, such as race or ethnicity. However, as noted previously, the ECOA and Regulation B do permit lenders to design and implement SPCPs in order to extend credit to a class of persons who would otherwise be denied credit or would receive it on less favorable terms (12 C.F.R. §1002.8(a)(3)(ii)).

We discuss SPCPs in more detail below. We also discuss recent changes to the Community Reinvestment Act (CRA) that, once effective, should provide lenders with additional incentives to design and implement SPCPs.

Appendix A.1.1. Special Purpose Credit Programs

While fair lending laws may prohibit lenders from using group-specific lending thresholds corresponding to neighborhood types, the ECOA does permit lenders to establish Special Purpose Credit Programs (12 C.F.R. §1002.8), in which prohibited factors such as race or ethnicity receive favorable consideration. These programs are intended to extend credit to those who would be unlikely to receive it under the lender's customary lending standards or would receive credit on less favorable terms (12 C.F.R. §1002.8(a)(3)(ii)).[27]

As an example of such a program, Fannie Mae recently published its Equitable Housing Finance Plan. Its goal is to reduce racial disparities in access to mortgage financing. Part of the plan involves the creation and deployment of SPCPs with the objective of "enabling access to credit and encouraging sustainable homeownership for Black consumers." Fannie Mae's SPCPs are focused on "people residing in formerly redlined and other underserved areas with majority Black populations."[28]

While SPCPs have historically been underutilized,[29] several recent actions by federal regulatory agencies have encouraged their use. A 2020 Advisory Opinion by the Consumer Financial Protection Bureau (CFPB) noted that a lender can initiate a SPCP without CFPB approval, provided the lender's program meets the compliance standards and general rules set forth in Regulation B (Official Interpretations, 12 C.F.R. pt. §1002(supp. I), sec. §1002.8, 8(a)-1).[30] The lender must first demonstrate a need for the program, either by analyzing its own lending data or reviewing research or data from an outside source. In addition, the lender must have a written plan that identifies the program's intended beneficiaries and establishes the procedures and standards the lender will use for extending

[27] It is unclear whether, by providing credit to persons living in historically underserved neighborhoods at more favorable terms than persons living outside these neighborhoods, SPCPs pose a fair lending risk, since those living outside historically underserved neighborhoods would not qualify for the program. As noted previously, we show in our paper that group-specific lending thresholds can achieve greater fairness without diminishing the approval rates of the residents of non-LMI neighborhoods under a single-threshold lending model. Thus, in our framework, neither group is worse off under group-specific thresholds and lenders mitigate the risk of discriminating against any borrower.

[28] See https://www.fanniemae.com/media/43636/display. Last accessed: 10/08/2024.

[29]See https://www.hud.gov/sites/dfiles/FHEO/documents/FHEO_Statement_on_Fair_Housing_and_Special_Purpose_Programs_FINAL.pdf. Last accessed: 10/08/2024.

[30] See https://files.consumerfinance.gov/f/documents/cfpb_advisory-opinion_special-purpose-credit-program_2020-12.pdf. Last accessed: 10/08/2024.

credit under the program.[31] The plan must also state the expected duration of the program and the criteria by which its continuing need will be evaluated.

In December 2021, the Department of Housing and Urban Development issued guidance clarifying that SPCPs that conform to the ECOA and Regulation B would generally not violate the FHA.[32] This opinion was followed by interagency guidance from the Federal Reserve Board, the Federal Deposit Insurance Corporation, the National Credit Union Administration, the Office of the Comptroller of the Currency, the CFPB, the Department of Housing and Urban Development, the Department of Justice, and the Federal Housing Finance Agency. The February 2022 guidance encourages lenders to explore opportunities to develop SPCPs.[33] The interagency guidance notes that lenders are permitted to consider the use of SPCPs across all types of credit covered by the ECOA and Regulation B.

Some of the largest lenders in the US have introduced SPCPs, including Wells Fargo, Bank of America, JP Morgan, and TD Bank.[34] In late 2022, JP Morgan announced that, after a successful pilot phase, it would be expanding nationally a SPCP designed to increase lending to small businesses located in majority-minority neighborhoods.[35]

In addition to the guidance and opinions listed above, forthcoming changes to the CRA — promulgated in 2023 and effective April 1, 2024, with staggered compliance dates of January 1, 2026, and January 1, 2027 — may also spur the creation of SPCPs in the coming

[31] See 12 C.F.R. §1002.8(a)(3)(i).

[32] See https://www.hud.gov/sites/dfiles/GC/documents/Special_Purpose_Credit_Program_OGC_guidance_12-6-2021.pdf. Last accessed: 10/08/2024.

[33] See Interagency Statement on Special Purpose Credit Programs Under the Equal Credit Opportunity Act and Regulation B. https://www.fdic.gov/news/financial-institution-letters/2022/fil22008a.pdf. Last accessed: 10/08/2024.

[34] See https://www.americanbanker.com/news/banks-expanding-special-purpose-credit-programs and https://www.jchs.harvard.edu/blog/designing-new-programs-narrow-racial-homeownership-gaps. Last accessed: 10/08/2024.

[35] See https://www.americanbanker.com/news/jpmorgan-chase-takes-special-purpose-credit-program-national. Last accessed: 10/08/2024.

years.[36] The final rule, issued by the Federal Reserve Board, the Federal Deposit Insurance Corporation, and the Office of the Comptroller of the Currency, is the first update to CRA regulations since 1995. Once effective, the final rule indicates that 10 percent of a large bank's CRA grade will be determined by a Retail Services and Products Test, and that the use of alternative credit scores, SPCPs, and other credit products that assist low- or moderate- income individuals with purchasing a home could be considered responsive credit products under that test.[37]

Appendix A.2. Machine learning adoption

The second major challenge to dual adoption of ML and group-specific lending thresholds is the risk, expense, and uncertainty surrounding the use of ML models in credit underwriting. While sophisticated ML models are pervasive in fintech lending, banks and other traditional lenders have proceeded more cautiously when considering the use of ML models for credit risk assessment. The use of ML models has made significant inroads into certain credit products, such as credit cards and unsecured consumer loans, and are also used in automotive and small business lending (FinRegLab, 2021). Overall, banks appear to be in the early stages of adopting ML in credit underwriting. This is partly due to the number of ways in which ML models complicate internal model development and governance processes, as well as lenders' ability to satisfy their legal and regulatory requirements. ML models require technical expertise that may not exist at a traditional lender, as well as the ability to absorb implementation costs to purchase and build computing infrastructure (FinRegLab, 2021).

A 2021 report by FinRegLab indicated that broader acceptance and use of ML models is also hindered by a variety of risk and trustworthiness concerns, including model risk management, fair lending, model transparency and explainability, and the ability to

---

[36] See https://www.federalreserve.gov/aboutthefed/boardmeetings/frn-cra-20231024.pdf. Last accessed: 10/08/2024.

[37] According to the final rule, a bank with more than $2 billion in assets would be classified as a large bank and be evaluated under four performance tests, including the Retail Services and Products Test.

generate adverse action notices, as required by law (FinRegLab, 2021).[38] While a discussion of these challenges is outside the scope of this paper, it is important to recognize that some lenders — both nonbank fintechs and traditional lenders — are currently using ML models in a variety of lending decisions that potentially affect millions of individuals, but that uncertainty remains, in no small part because of the complexity of ML models and the uncertainty as to how these models fit into existing legal and regulatory frameworks.

Appendix A.3. Dual adoption strategy

The possibility of dual adoption of group-specific lending thresholds and ML models seems unlikely to occur overnight, given the challenges of fair lending law and ML adoption. However, there may be an opportunity for lenders and regulators to leverage the provision of the SPCP to learn more about the effects of ML-based credit decisions on lending fairness in a well-defined space in which fairness is a primary objective. Such an arrangement would almost certainly require additional regulatory guidance from the CFPB and perhaps an interagency group of regulators to ensure that lenders would be undertaking no additional risk by participating in a compliant dual adoption program. Under such an arrangement, lenders might also be encouraged to refine their group-specific lending thresholds, examining classifications based not only on LMI neighborhoods but also on LMI income cut-offs, census tract-based racial and ethnic concentrations, and minority and women-owned businesses.

The adoption of group-specific lending thresholds needn't be limited to lenders that have yet to adopt ML underwriting models. Within the group of lenders that have already made the transition from regression-based models to ML credit underwriting models — particularly lenders with a mission of reaching underserved populations — lenders could be encouraged to design and adopt their own SPCP. To understand and fully characterize the gains from dual adoption, these lenders would need to establish a benchmark. For example, lenders could score credit applicants with both a machine learning model and a

[38] The ECOA requires lenders to disclose up to four reasons why an individual was denied credit or received less favorable credit terms on an existing loan or credit arrangement.

legacy regression model. Likewise, fintech lenders that have been using ML models since their inception might benchmark against a previous version of their model, a regression-based model, or a model without alternative data.

Appendix B. Predictive model details

This section describes our modeling pipeline in more detail relative to Section 3. First, we select the relevant class of models. Second, we set up the inputs: selecting the LHS variable, setting up rolling windows, and performing RHS variable selection (using lasso regression) and preprocessing via binning. Finally, we train our models, tune their hyperparameters, and evaluate their performance. We perform lasso variable selection and tune and train the logistic model using the R library *glmnet* (Friedman et al., 2010). We tune and train the XGBoost model using the R library *tidymodels* (Kuhn and Wickham, 2020) with *xgboost* backend (Chen and Guestrin, 2016). The details of the pipeline are provided below.

Appendix B.1. Model selection

We predict the consumer repayment status using two models: logistic regression with ridge regularization (referred to simply as *logistic model*) and eXtreme Gradient Boosting (*XGBoost*).[39] Ridge logistic regression represents the class of linear models that have been commonly used for credit scoring in the last few decades, whereas XGBoost represents non-linear models that are being increasingly used for credit scoring today. We refer to the output of our predictive models as a credit score. Logistic regression with ridge regularization includes a penalty on the sum of squared coefficients, shrinking them closer to zero (but not to zero) in order to avoid overfitting, as in Hoerl (1962). Ridge produces familiar coefficients with readily interpretable coefficients — while allowing for large numbers of input variables and non-linear relationships with appropriate preprocessing (subsection Appendix B.4). Because we perform variable selection (subsection Appendix B.4) before fitting our model, ensuring that nearly all RHS inputs are relevant to predicting the LHS, Ridge regression is likely to outperform other linear regularization methods (James et al., 2023).[40]

[39] See Hastie et al. (2017) for an introduction to regularized linear models and gradient boosting.

[40] In untabulated results, we also evaluated the performance of lasso and elastic-net regularization and found that ridge regularization is empirically superior.

XGBoost is a tree-based algorithm, related to Random Forest and decision trees (Chen and Guestrin, 2016). Decision trees split the dataset based on RHS variables and predict a specific LHS variable value for each split. For example, consumers with two or three credit card accounts but without a mortgage are assigned one probability of default, while consumers with a single credit card and a mortgage are assigned another probability, and so on. Random forests average many decision trees (Breiman, 2001). XGBoost instead estimates trees sequentially, with each new tree focusing on examples that the previous one fit poorly. XGBoost naturally incorporates a wide range of non-linear relationships. Interpretability is not as readily obtainable as in regularized linear models, but various measures of variable importance can be calculated. XGBoost is prone to overfitting training data — a pitfall we address in Appendix B.3 and Appendix B.5.

Appendix B.2. LHS variable: Definition of non-default and default

We define our prediction target, *non-default*, as a binary variable that is equal to 1 if the consumer is not in the state of default within two years starting next quarter. We define the state of default as being 90 or more days past due on at least one of the accounts.[41]

Table 1 breaks down the percentage of consumer-quarters in the data by their current payment status and future default rate. As discussed previously, we only include presently current consumers in our main analysis. Rates for other groups are for reference.

Appendix B.3. Rolling window setup

Credit-scoring models are typically updated over time as the data-generating process evolves. Both the way some of the variables are reported and the relationship between RHS variables and default can change over time. To account for this, we estimate our model in a rolling window fashion. Each model is trained on eight quarters of data and used to obtain out-of-sample predictions on another eight quarters of data. The window is

[41] We choose non-default as the event state rather than default for stylistic reasons. In some sections of the paper, we focus on the elements of the confusion matrix such as *true positives* and *false positives.* Choosing non-default as the event aligns the meaning of *positive* as *beneficial* as in *positive test*, facilitating the communication of results.

rolled for eight quarters at a time. The first model is trained on data from 2000Q1 to 2001Q4 and evaluated on the period from 2004Q1 to 2005Q4. The gap between the training and evaluation samples is needed to obtain the payment status for all training sample consumers. The gap period becomes the training period for the next window. Table B4 shows the training, gap, and evaluation quarters for all eight windows. All results reported below are based on fully out-of-sample results — that is, only in the evaluation windows.

Appendix B.4. Variable selection and processing

To reduce the computational cost of training the models and to combat overfitting, we perform a variable selection procedure that is separate from model training.

We begin by manually inspecting all variable definitions in the CCP and select 457 variables that conceptually can be used in a credit scoring model. These variables cover things such as balances, utilization, performance, inquiries, age of accounts, and number of trades. There are variables aggregating these metrics across all account types and for specific trade types (mortgages, auto loans, credit cards, personal loans, etc.). There are both contemporaneous measures and those that look back across the prior three to 24 months. See Appendix C for more detail on the set of variables used in our models after the variable selection step described below.

We then bin the variables because many have a large fraction of missing values and/or are mixed-type variables (for example, the variable *% of bankcard accounts always paid as agreed* can include either a continuous percentage value or a special value corresponding to *no relevant accounts*). We use supervised discretization via recursive partitioning (Hothorn and Zeileis, 2015).

Next, we perform variable selection using lasso regularized regression. Lasso applies a penalty on the sum of absolute values of the coefficients (Tibshirani, 1994). Such a penalty assigns zero to a large number of coefficients, making lasso useful for variable selection (James et al., 2023). We choose the smallest regularization value that results in 100 variables having at least one bin with a non-zero coefficient.

Appendix B.5. Hyperparameter tuning and training

For the logistic model, the only hyperparameter is regularization strength. We choose it by performing a five-fold cross-validation over the training sample (Hastie et al., 2017). The final model is then estimated on the full training set.

For XGBoost, we tune nine different hyperparameters that affect the complexity of the model.[42] Tuning these parameters is crucial for regularization and performance. We perform a grid search across a sample of potential combinations of parameters, with bounds based on Blattner and Nelson (2021). To help prevent the XGB model from overfitting, we employ early stopping for both hyperparameter tuning and model training.[43]

Because of its nonlinear nature, XGBoost is much more sensitive to the way the validation set is chosen than the logistic model. We use the first quarter of the training set corresponding to a given rolling window to estimate models with all sets of candidate hyperparameters and pick the optimal set using the performance on the last quarter of the training set. This way, the chosen set of hyperparameters are more robust to changes in the underlying data-generating process over time. The final model is estimated on the full training set.

[42] Specifically, we tune the number of trees, tree depth, number of predictors for individual trees, minimum number of observations in a node, minimum loss reduction required to make a split, sample size for individual trees, and learning rate.

[43] Early stopping causes the model to cease training when the log-loss on a holdout set doesn't improve for a specific number (in our case, 10) of iterations (additions of a new tree) in a row.

Table B.4: Rolling window setup. These subsets of quarters are used to train and evaluate the logistic and XGBoost credit-scoring models and to compute lending thresholds. Training quarters are used for variable selection, hyperparameter tuning, and model training. Gap quarters are needed to compute default for the training quarters. The trained model is used to produce out-of-sample predictions of consumer default on evaluation quarters. Default is defined as being more than 90 days past due on any credit account within a two-year period starting with the next quarter. A training set is also used to compute lending thresholds that are applied to the evaluation quarters out-of-sample.

Window	Training quarters	Gap	Evaluation quarters
1	2000Q1–2001Q4	2002Q1–2003Q4	2004Q1–2005Q4
2	2002Q1–2003Q4	2004Q1–2005Q4	2006Q1–2007Q4
3	2004Q1–2005Q4	2006Q1–2007Q4	2008Q1–2009Q4
4	2006Q1–2007Q4	2008Q1–2009Q4	2010Q1–2011Q4
5	2008Q1–2009Q4	2010Q1–2011Q4	2012Q1–2013Q4
6	2010Q1–2011Q4	2012Q1–2013Q4	2014Q1–2015Q4
7	2012Q1–2013Q4	2014Q1–2015Q4	2016Q1–2017Q4
8	2014Q1–2015Q4	2016Q1–2017Q4	2018Q1–2019Q2

Appendix C. Variables

As described in Section 3, for each training window, we start with 457 variables from the Federal Reserve Bank of New York Consumer Credit Panel/Equifax data (CCP) and then perform variable selection using lasso regression. This process results in the 100 variables used in each window's models. In this section we provide summaries of these variables — in terms of both specific fields and categories of variables.

For this exercise, we have classified input variables in two ways: by variable type — what credit behavior or attribute does the variable measure — and by loan type. Variable categories are as follows.

- Delinquency — contemporaneous or lagged payment history relative to required payments
- Number or share of accounts — count or share of accounts for a given account type
- Balance — outstanding balance
- Account age — number of months since account opening
- Utilization — percent of total available credit in use
- Credit inquiries — number of times lenders have pulled the customer's credit file, a measure of credit demand
- Collections — number or amount of loans in collection
- Payments — amount customer paid back on a given loan type
- Bankruptcy — has customer recently been in (any type of) bankruptcy
- Derogatory events — enumerates derogatory events, such as foreclosure or charge-off
- Credit limit — credit limit for given loan type

Loan types are as defined by the FRBNY CCP staff report by Lee and van der Klaauw, 2010.

- All — all accounts
- Revolving — all revolving accounts
- Installment — all installment accounts
- Bankcard — credit card accounts for banks, bankcard companies, national credit card companies, credit unions, and savings and loan holding associations
- Retail — credit card accounts for clothing, groceries, department stores, home furnishings, gas, etc.
- Department Store — subset of retail
- Mortgage — close-ended loans secured on property
- Student loans — loans to finance educational expenses
- Auto — loans taken out to purchase a car, including auto bank loans (provided by banking institutions) and auto finance loans (provided by automobile dealers and automobile financing companies)
- Consumer Finance — sales financing and personal loans
- Home Equity Revolving — home equity loans with a revolving line of credit with a credit limit

Given that we perform variable selection independently for each training window, our credit-scoring models consider slightly different variables depending on the window. That said, there are 28 variables shared across all training windows. We describe these variables in table C.7. Measures capturing contemporaneous and recent delinquency, utilization, and credit demand (via inquiries) account for more than half of these variables. Variables that span across all account types also cover more than half the list, with aggregates across revolving accounts making up another 20 percent.

We also look across all eight training windows and create an aggregate share by variable and loan type — with weights determined by the total frequency across windows. In table C.5, we aggregate by variable type. We find that 40 percent of all variables track contemporaneous and recent delinquency, followed by another 14 percent that track the total number of accounts. Table C.6 shows that 42 percent of all the variables considered aggregate across all of an individual's loans. In total, revolving credit, as measured by revolving, bankcard, department store, and retail trades, cover an additional 38 percent.

Table C.5: Share of Variables by Type of Variable

Variable Type	Share of Variables
Delinquency	40%
Number/% of Accounts	14%
Balance	9%
Account Age	8%
Utilization	7%
Credit Inquiries	6%
Collections	5%
Payments	5%
Bankruptcy	3%
Derogatory Events	3%
Credit Limit	1%

Table C.6: Share of Variables by Account Type

Account Type	Share of Variables
All	42%
Revolving	19%
Bankcard	13%
Installment	7%
Mortgage	4%
Department Store	3%
Student Loan	3%
Retail	3%
Auto	2%
Consumer Finance	2%
Home Equity Revolving	1%

Table C.7: Variables Selected via Lasso in All 8 Training Windows

Description	Variable Type	Account Type
Num of inquiries w/in 3m	Credit Inquiries	All
Num of inquiries w/in 12m	Credit Inquiries	All
Num of inquiries w/in 24m	Credit Inquiries	All
Age oldest acct	Account Age	All
Age oldest bankcard acct	Account Age	Bankcard
Age oldest revolving acct	Account Age	Revolving
Bal open finance/student loan w/updt w/in 3m	Balance	Student Loan
Credit limit, revolving acct w/updt w/in 3m	Credit Limit	Revolving
Total past due amount	Delinquency	All
Num 30 DPD occur w/in 12m, revolving acct	Delinquency	Revolving
Num 30 DPD occur w/in 24m, installment acct	Delinquency	Installment
Num 120-180+ DPD occur w/in 24m	Delinquency	All
Num open retail acct w/updt w/in 3m $\geq$ 50% util	Utilization	Retail
Num open revolving acct w/updt w/in 3m $\geq$ 50% util	Utilization	Revolving
Num open bankcard acct w/updt w/in 3m $\geq$ 75% util	Utilization	Bankcard
% acct opened w/in 6m to all acct	Num/% of Acct	All
% acct opened w/in 12m to all acct	Num/% of Acct	All
% revolving acct opened w/in 12m to all rev acct	Num/% of Acct	Revolving
Utilization, open bankcard acct w/updt w/in 3m	Utilization	Bankcard
Utilization, open revolving acct w/updt w/in 3m	Utilization	Revolving
% bal to total loan, installment acct w/updt w/in 3m	Balance	Installment
% acct always satisfactory	Delinquency	All
Bankruptcy w/in 24m	Bankruptcy	All
3rd party collection amount, w/in 12m	Collections	All
3rd party collection amount, w/in 24m	Collections	All
3rd party collection amount, total	Collections	All
% acct always satisfactory, w/in 6m	Delinquency	All
% inquiries 3m to inquires 12m	Credit Inquiries	All

Appendix D. Details of λ calibration

As discussed in Section 3.1.1, we calibrate the λ value, which represents the ratio of the average cost of default to the average five-year profit earned on non-defaulting loans, using account- and portfolio-level administrative data from the Federal Reserve Board's Capital Assessments and Stress Testing report (Y-14M).[44] Our calculations closely follow the methodology of Agarwal et al. (2014).

We use both account- and portfolio-level credit card data because some revenue and cost measures are not observed at the account level. These include all expenses as well as interchange income. In addition, because the Y-14M data does not include marketing and acquisitions expenses, we augment it with portfolio-level data from the Office of the Comptroller of the Currency (OCC) running from 2008 to 2013. To scale portfolio-level expenses down to the account level, we follow Agarwal et al. (2014) and assume that revenues and expenses we don't observe at the account level broadly scale with either balances or purchase volume. We first compute monthly lender-level ratios and then apply them to individual accounts by multiplying by the average daily balance (ADB) or the total purchase volume.

The five-year profit calculation (approximating the lifetime profit) is a present value discounted version of the account-level profit calculation. The formula for profit is defined as:

$$\text{Profit}(\mathit{Loss}) = \text{Total Income} - \text{Total Expenses} - \text{Net Charge-Off Amount}$$

Total Income is calculated as follows:

[44] For more information about the Y-14M, refer to https://www.federalreserve.gov/apps/reportingforms/Report/Index/FR_Y-14M. Last accessed: 10/08/2024.

$$\text{Total Income} = \text{Finance Charges} + \text{Total Fees} + \text{Net Interchange Income},$$

where:

- Finance Charges are observed directly in the account-level data.
- Total Fees include late, over limit, insufficient funds, cash advance, balance transfer, annual or monthly membership, and debt suspension fees, and are also from the account-level data.
- Net Interchange Income is computed based on a Net Interchange Income ratio. This ratio is the sum of interchange income, interchange expense, rewards expense, and fraud expense, all divided by month-end managed receivables. However, since this factor should scale with purchases rather than balances, which can include fees and finance charges, we multiply it by the ratio of balances to purchases before scaling it by account-level monthly purchases. The balances-to-purchases ratio is computed using the monthly account-level Y-14M data and is the ratio of total ADB to total purchases at the lender level.

Total Expenses are computed as follows:

$$\begin{aligned}\text{Total Expense} =\, &\text{Interest Expense} + \text{Collections Expense} + \\ &\text{Marketing and Acquisition Expense} + \text{Other Expenses},\end{aligned}$$

where each component is calculated as follows:

- Interest Expense is computed using an Interest Expense ratio, the ratio of interest expense to month-end managed receivables. The monthly account-level interest expense is the product of ADB and the Interest Expense ratio.
- Collections Expense is computed using a Collections Expense ratio, the ratio of Collections Expense to month-end managed receivables. The monthly account-level Collections Expense is the product of ADB and the Collections Expense ratio.

- Marketing and Acquisition Expense is computed using a factor of 0.000723, the long-run mean of the bank-level ratios of the expense to total ADB, computed using the OCC data. This factor is applied to the monthly ADB at the account level. For example, an account with a $1,500 ADB would be associated with $1.08 in monthly marketing and account acquisition expenses.

- Other Expenses is computed using the ratio of Other Expenses to month-end managed receivables, applied similarly to the previous expense categories. According to Y-14M documentation, the Other Expenses category includes servicing, billing, processing interchange and payments, issuing cards, authorizations, and outside services. The Other Expenses ratio is the ratio of Other Expenses to month-end managed receivables. Monthly account-level Other Expense is the product of ADB and the Other Expenses ratio.

Net Charge-Off amount is computed as follows:

$$\text{Net Charge-Off Amount} = \text{Gross Charge-Off Amount} - \text{Recovery Amount}.$$

Actual gross charge-off amounts are included in account-level profit calculations. The charge-off amount is reduced by any recoveries the lender receives during the 12 months following the charge-off. Net Charge-Off Amount equals 0 for all non-defaulting accounts.

Future cash flows are discounted back to the year of account origination using a discount factor of 10 percent (Kovner and Van Tassel, 2021).

Finally, λ is computed as follows:

$$\lambda = \frac{Average\ Loss\ (charge{-}off\ accounts)}{Average\ Profit\ (non\ charge{-}off\ accounts)}$$

Appendix E. Robustness to including borrowers currently in default

As discussed in Section 2.2, our main sample excludes consumers currently in default. Here we present a robustness check that tests sensitivity of our main results to this sample selection strategy.

To accommodate this change, we trained separate logistic and XGB models specifically for individuals in this category.[45] After training the models, we map the scores into probabilities using isotonic regression and combine the outputs of the models trained on currently current and currently in default samples. The rest of the analyses proceed in the same way as before.

Table E.8 presents the fairness–profit trade-offs for λ ranging between 1 and 11. The comparison with Table 4, which is based on the main sample selection strategy that excludes individuals currently in default, suggests that including them leads to qualitatively and quantitatively similar results.

[45] Training a single model for currently current and currently defaulted individuals leads to extreme out-performance of the logistic by XGB on the currently in-default sub-sample, skewing the overall performance statistics.

Table E.8: Fairness–profit trade-offs at different levels of fairness constraints and λ (profit trade-off of good to bad loans), sample including individuals currently in default. Profit is calculated as $Pr = TP - \lambda \times FN$, where TP is the number of true positives, FP is the number of false positives, and λ corresponds to the monetary loss associated with a loan that is not repaid relative to a gain from a loan that is repaid. We normalize profit by the profit of the logistic model with no fairness constraints. The positive outcome is non-default within the next two years. Fairness, or ΔTPR, is the difference in TPR between the LMI and non-LMI groups. *TPR* is defined as *TP/(TP+FN),* where TP are the true positives and FN are the false negatives. Individuals in low- and moderate-income (LMI) census tracts are identified using a dataset produced by the FFIEC, based on U.S. Census Bureau data. The labels Strong, Medium, Weak, and Blind represent different strengths of the fairness constraint. (See Section 4.5 for definitions.)

		λ					
	Policy	1	3	5	7	9	11
Fairness							
Log	Blind	-3.4	-7.7	-10.1	-11.4	-12.5	-13.1
	Weak	-2.3	-4.9	-6.6	-7.4	-7.8	-8.1
	Medium	-1.0	-2.4	-2.9	-3.1	-3.6	-3.9
	Strong	0.1	0.1	0.2	0.2	0.4	0.2
XGB	Blind	-3.1	-7.4	-9.7	-11.2	-12.2	-12.9
	Weak	-2.1	-4.8	-6.3	-7.2	-7.7	-8.0
	Medium	-1.0	-2.1	-3.1	-3.6	-3.8	-3.9
	Strong	0.1	0.2	0.2	0.3	0.1	0.2
Profit							
Log	Blind	100.0	100.0	100.0	100.0	100.0	100.0
	Weak	100.0	100.0	99.9	99.8	99.8	99.7
	Medium	100.0	99.8	99.5	99.3	99.2	99.0
	Strong	100.0	99.5	99.0	98.5	98.3	97.9
XGB	Blind	100.7	101.3	102.1	102.5	102.7	102.9
	Weak	100.7	101.2	101.9	102.2	102.4	102.4
	Medium	100.7	101.0	101.5	101.7	101.7	101.6
	Strong	100.6	100.7	100.8	100.7	100.6	100.2

Appendix F. Potential pricing implications of model improvement and fairness constraints

In this section, we discuss the potential implications of credit scoring model improvement and fairness constraints for loan pricing. While the pricing decision is as equally important as the loan-granting decision, the pricing information is generally not available in the CCP. To address this limitation, we introduce an additional dataset that allows us to estimate a mapping from the probability of default to the price of a first mortgage. We are then able to simulate a scenario in which all consumers applied for a mortgage in a nationally representative market. The HMDA-McDash-CRISM dataset[46] contains data on both mortgage pricing and borrower delinquency and default. Using these data, we are able to assign interest rates to individuals based on their probability of default implied by our credit-scoring models.

We first argue that changes in loan pricing resulting from improvements in credit scoring models are likely to be small, since the changes in default probabilities due to such improvements are generally modest. Then we turn to examining the potential loan prices faced by consumers who only receive loans after the introduction of fairness constraints. We contend that fairness constraints are compatible with only a limited effect on loan pricing, as the probability of default among consumers who benefit from them does not warrant sharp price increases.

Appendix F.1. Data processing

Our goal is to estimate the mortgage interest rates that consumers might receive from a lender using the Logistic and XGB credit scores estimated in this paper.[47] There are two

[46] HMDA-McDash-CRISM is a combination of several sources of anonymized data: Home Mortgage Disclosure Act (HMDA) data, Black Knight McDash data, and a credit bureau dataset, Equifax Credit Risk Insight Servicing data, that is linked to the McDash data (known as CRISM). Both CRISM and the merged HMDA-McDash-CRISM datasets are anonymized. The combined dataset covers more than 60 percent of the U.S. mortgage market (in some years as much as 80 percent).

For more information see Gerardi et al. (2021).

[47] We use the exact same models as described in section 3, which use the data (a lasso-selected subset of 100 variables out of 457 available variables) as described in section 2. We do no new training and add no

key intermediate steps. First, we convert the VantageScore credit score available in the HMDA-McDash-CRISM dataset that only covers matched consumers with mortgages to the ex-post default probability. This allows us to put the credit score available in the HMDA-McDash-CRISM dataset on the same scale as the credit scores we computed using the CCP, since we also can convert them to default probabilities. Second, we map the default probability to the mortgage rates.

We process the HMDA-McDash-CRISM matched dataset in the following way. First, we apply a series of filters to isolate loans meeting specific criteria. Loans are filtered to first lien only, with no more than six months of seasoning and a confident match. We exclude borrowers in default on any loan at the time of origination. Then we compute a forward-looking variable of default (on any product, not only the mortgage) within two years of the mortgage origination.[48] Then, we map the VantageScore credit score to the probability of default using isotonic regression.[49] This allows us to map interest rates to the ex-post probability of default to the Vantage score and our model scores. We apply that mapping to the probabilities of default associated with the logistic and XGB models obtained, again, using isotonic regression. That gives us our main object of interest: estimated mortgage rates likely to be offered to the borrowers based on their estimated credit score.

We choose to combine different types of loans in our interest rate estimation rather than focusing on a single product, such as a 30-year fixed-rate mortgage. Thus, our estimated rates reflect the interest rate on the mortgage loans that consumers with a given probability of default are most likely to receive when access to credit is expanded.[50]

new data. We score the relevant customers using the pre-trained models and run the pricing analysis forward from there.

[48] We define default the same as we do for CCP, being more than 90 days past due on any loan product.

[49] All isotonic regressions are run separately for each rolling window used in the estimation of our main models.

[50] In untabulated analyses, we confirm that all the results discussed here also hold if we focus on any fixed-rate loans or exclusively on 30-year fixed-rate mortgages. The interest rate spreads in these alternative analyses are smaller and the price impact is even more muted than in our main analysis.

Appendix F.2. Analysis

To obtain insights about potential pricing implications based on the default probabilities of consumers granted credit, we assume that pricing decisions are closely linked to the default probability.[51] Our proposed policy links model improvement and fairness constraints, and both parts of the policy impact the probability of default of consumers granted credit: Model improvement does so by directly re-estimating the probabilities, and fairness constraints do so by expanding the number of people who receive the loans in LMI areas and granting loans to people with a higher probability of default than before. We examine the pricing effects of both in turn.

We begin by analyzing the impact of model improvement on loan pricing. Table F.9 provides insights into the creditworthiness of individuals in different income groups, as measured by their probability of default. The table shows that a higher percentage of people in the non-LMI group benefit from model improvement by being assigned a lower probability of default than before: 68 percent of non-LMI consumers benefit compared with 61 percent of LMI consumers. Therefore, the non-LMI group is expected to benefit more in terms of loan pricing.

The table also indicates that less than half of the individuals in the lowest quartile of creditworthiness are likely to benefit from model improvement. Moreover, those belonging to the LMI group are more likely to fall into this category. However, this is not necessarily a negative outcome, as the Δ%TP and Δ%FP columns show that more people belong to the TP group under the XGB model, while fewer people belong to the FP group. This means that the decreases in the probability of default are concentrated among the defaulters, for whom the negative effects of higher perceived default probability are

[51] More generally, mortgage pricing decisions are a function of the default risk, the prepayment risk, and the opportunity cost. While we don't take the prepayment risk into account directly, differences in prepayment rates are unlikely to affect pricing below the threshold. Recent research shows that subprime and minority groups are less likely to refinance mortgage debt (see Lambie-Hanson and Reid 2018; Gerardi et al. 2021, Gerardi et al. 2023). If prepayment risk decreases with credit score, lenders are less likely to demand higher prepayment premiums below the non-LMI threshold than just above it.

partially offset because they are likely to face default penalties if granted the loan, resulting in even lower access to credit in the future.[52]

Table F.10 illustrates how changes in the probability of default translate to estimated interest rate spreads. The results show that, congruent with the improved accuracy of the XGB model, interest rate spreads for the XGB model are better than those for the logistic model for both groups. The improvement is larger for the non-LMI group, but the differences in improvement are small: The non-LMI group improves from -0.010 percent to -0.041 percent, while the LMI group improves from 0.042 percent to 0.016 percent. This finding contrasts with the results of Fuster et al. (2021) for the mortgage market, where they find negative effects on pricing among minorities. However, both their results and ours are measured in basis points, so they are very small in magnitude.

Next, we move on to examine the potential pricing implications of fairness constraints for consumers who only receive loans after the constraints are introduced. We show that the probability of default of the new consumers justifies only relatively modest price increases. To conduct this analysis, we focus on the predictions of the XGB model for the LMI group, and we assume that consumers who receive the loans under the single-threshold policy receive them with the same interest rate as in the blind case.

Figure F.8 illustrates the distribution of the estimated interest rate spreads faced by consumers who always receive the loan and consumers who only receive loans under the strong fairness constraint. The results show no sharp discontinuities in rates for the newly granted loans. The median new borrower pays only 5 basis points more than the maximum rate charged to the borrowers in the always-granted group, and 20 basis points more than the borrowers in the 75th percentile. The whole spread between the 100th percentile of the

[52] This does not hold in the expected value for every individual with a high probability of default. Very risky borrowers who are granted credit owing to an inaccurate prediction from a model benefit if they do not end up defaulting. This is especially true in the mortgage market, where the benefits of homeownership can be substantial.

always-granted group and the 100th percentile of the newly granted group is smaller than the spread between the 75th and 100th percentiles of the always-granted groups.[53]

Overall, we find suggestive evidence that the impact on pricing of combining model improvement with fairness constraints for new borrowers can be small. This is because the probability of default for borrowers who are granted credit under different scenarios is relatively modest, leading to only minor differences in spreads in the observable mortgage market.

[53] In addition to our main specification, which uses all HMDA-McDash-CRISM loans to estimate interest rates, these result hold qualitatively in untabulated analyses that use only fixed-rate loans and only 30-year fixed-rate loans. The alternative specifications lead to tighter spreads and reduced price impact.

Figure F.8: Comparison of the quartiles of the estimated Interest Rate Spread for consumers in LMI areas who are granted loans without fairness constraints and consumers who are only granted loans when strong fairness constraints are introduced. This analysis is for mortgages only. The spread between the 100th percentile of the always-granted group and the 100th percentile of the newly granted group is smaller than the spread between the 75th and 100th percentiles of the always-granted group. The Interest Rate Spread is defined as the estimated mortgage interest rate for a given level of probability of default minus the average estimated mortgage interest in a given quarter. The mapping between the probability of default and the interest rate is estimated using the HMDA-McDash-CRISM matched dataset. The spread is based on the probability of default estimated by the XGB model. Individuals in low- and moderate-income (LMI) census tracts are identified using a dataset produced by the FFIEC, based on U.S. Census Bureau data. *Default* is defined as being more than 90 days past due on any credit account within a two-year period starting with the next quarter.

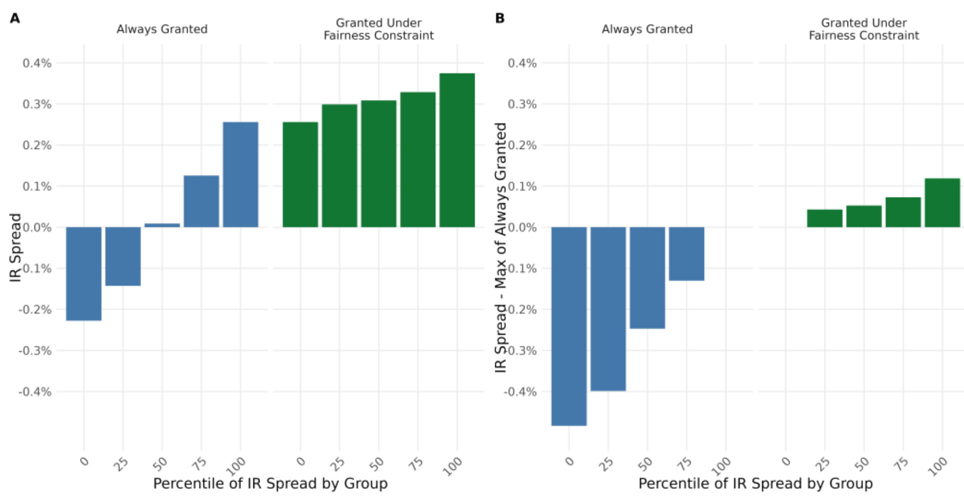


Table F.9: Creditworthiness by model and income group. The table compares creditworthiness between the XGB and logistic models, showing the percentage of the group with lower estimated probability of default. Non-LMI consumers benefit more from the model improvement (68 percent) than LMI consumers (61 percent). The lowest creditworthiness quartile has less than 50 percent of members with increased creditworthiness. However, more belong to the true positive group and fewer to the false positive group under the XGB model, which offsets negative effects for defaulters. Individuals in low- and moderate-income (LMI) census tracts are identified using a dataset produced by the FFIEC, based on U.S. Census Bureau data. *Default* is defined as being more than 90 days past due on any credit account within a two-year period starting with the next quarter.

	% Better		% of the Group		Δ%TP		Δ%FP	
Q of C/w	LMI	non-LMI	LMI	non-LMI	LMI	non-LMI	LMI	non-LMI
All	61	68	100	100	0.449	0.290	-0.203	-0.092
1Q	47	49	36	22	1.823	2.652	-0.377	-0.073
2Q	51	54	27	25	-0.735	-1.209	-0.248	-0.304
3Q	77	79	20	26	-0.008	-0.006	-0.002	-0.001
4Q	89	88	17	27	0.000	-0.001	-0.002	-0.001

Table F.10: Comparison of the Means and Standard Deviations of the Interest Rate Spread by area income level and model for creditworthy consumers. The creditworthiness is determined using our main specification of the lender decision function under the blind threshold policy. The Interest Rate Spread is defined as the estimated mortgage interest rate for a given level of probability of default minus the average estimated mortgage interest in a given quarter. The mapping between the probability of default and the interest rate is estimated using the HMDA-McDash-CRISM matched dataset. Individuals in low- and moderate-income (LMI) census tracts are identified using a dataset produced by the FFIEC, based on U.S. Census Bureau data. *Default* is defined as being more than 90 days past due on any credit account within a two-year period starting with the next quarter.

	LMI	non-LMI
Mean IR Spread Logistic (%)	0.042	-0.010
Mean IR Spread XGB (%)	0.016	-0.041
St. Dev. IR Spread Logistic (%)	0.329	0.289
St. Dev. IR Spread XGB (%)	0.350	0.310