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
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

November 14, 2022

Abstract

Modeling advances create credit scores that predict default better overall, but raise concerns about their effect on protected groups. Focusing on low- and moderate-income (LMI) areas, we use an approach from the Fairness in Machine Learning literature — fairness constraints via group-specific prediction thresholds — and show that gaps in true positive rates (% of non-defaulters identified by the model as such) can be significantly reduced if separate thresholds can be chosen for non-LMI and LMI tracts. However, the reduction isn’t free as more defaulters are classified as good risks, potentially affecting both consumers’ welfare and lenders’ profits. The trade-offs become more favorable if the introduction of fairness constraints is paired with the introduction of more sophisticated models, suggesting a way forward. Overall, our results highlight the potential benefits of explicitly considering sensitive attributes in the design of loan approval policies and the potential benefits of output-based approaches to fairness in lending.

Keywords:
Credit Scores, Group Disparities, Machine Learning, Fairness

JEL: 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, nathan.schor@phil.frb.org

Acknowledgments:
We are grateful for the helpful comments of Bob Hunt, Julia Cheney, Jeanne Rentezelas, Ken Benton, and Jack Terruso, as well as the seminar and conference participants at University of Delaware 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 the computing infrastructure support.

Disclaimer: This Philadelphia Fed working paper represents preliminary research that is being circulated for discussion purposes. The views expressed in these papers are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. No statements here should be treated as legal advice. Any errors or omissions are the responsibility of the authors. Philadelphia Fed working papers are free to download at https://philadelphiafed.org/research-and-data/publications/working-papers.
1. Introduction

The growing sophistication of consumer credit default models has raised fair lending concerns among consumer groups and financial regulators. While improvements in default prediction are generally viewed as a good thing, there is some concern that the gains from machine learning models could accrue unequally, and that disparities in credit allocation could be further ingrained in a system that is more opaque and thus more difficult to regulate. The increased availability of data from outside of the credit reporting system (e.g., social media and shopping habits) may increase the risk that more sophisticated and opaque machine learning models indirectly and inadvertently incorporate prohibited consumer sociodemographic variables into the lending decision. Recent research finds that Black and Hispanic mortgage borrowers benefit less from more sophisticated models (Fuster et al., 2021) and that credit scores are more noisy for minority borrowers (Blattner and Nelson, 2021).

We claim that combining model improvement with fairness constraints via group-specific prediction thresholds, like the ones suggested by Hardt et al. (2016), can be part of the solution to disparities in predictive power. Our group-specific thresholds are based on the concept of historically underserved communities that is 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 historically underserved segments of the communities in which they operate. Prior to the CRA, some lenders had engaged in unfair lending practices such as redlining, in which lenders avoided lending to would-be homeowners in low- and moderate-income (LMI) areas and areas with high concentrations of minority residents. 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 the concept of LMI neighborhoods as defined in the regulations that implement 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.

Two federal laws govern 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). 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. Thus, it is unclear whether group-specific thresholds would be permissible under existing federal law. To implement a lending program in which lower score cutoffs are assigned to loan applicants residing in LMI neighborhoods, lenders would need to consider whether such a policy was permissible under federal, state, and local fair lending law, a determination that is beyond the scope of this paper.

We show that fairness constraints via group-specific thresholds can dramatically reduce gaps in the true positive rates, ensuring that creditworthy consumers in different groups are about as likely to be classified as good risks. However, this comes at a

---

1The 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 applicant has in good faith exercised any right under the Consumer Protection Act. The 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 FHA, there are seven prohibited bases: race, color, national origin, religion, sex (including gender identity and sexual orientation), familial status, and disability.

2In Section 5, we discuss one way in which lenders have historically been able to implement lending programs with rules favoring a historically disadvantaged class of borrowers.
cost of some eventual defaulters being misclassified as non-defaulters, which is both a lender- and consumer-borne cost that needs to be managed. One way to manage the trade-off is to link the introduction of separate thresholds to model improvement, which softens the costs.

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 CRA context. For consumers who live in CRA-eligible low- and moderate-income (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, while other consumers end up with loans they might not be able to pay back.

Our contribution is to consider the introduction of group-specific lending thresholds within the context of technological progress in default risk assessment. 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 certain threshold. We compare the performance of potential binary predictions 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 individ-
ual’s geographic area or exact location are typically considered proxy variables and thus prohibited from use in lending decisions. We show that a single-threshold approach results in significantly different true positive rates between groups (where repaying the loan is the positive outcome). For example, a creditworthy LMI tract consumer is about 6 percentage points less likely to be classified as creditworthy than a creditworthy non-LMI tract consumer.

The single-threshold approach is predicated on the requirement to exclude certain demographics and geography from consideration when a lending decision is made. While the intent of this policy is to reduce discrimination (especially, taste-based 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).

The alternative approach we consider in this paper would permit the use of specific geographic variables (e.g., LMI and non-LMI neighborhoods) to equalize true positive rates among different groups. As a result, LMI tract groups with noisier credit scores would be assigned lower thresholds. This framework is rooted in Fairness in Machine Learning literature (e.g., Hardt et al., 2016) and has various policy applications, for example, in education, mental health, criminal justice and housing safety settings (Rodolfa et al., 2021). 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.

Fairness constraints via group-specific thresholds can dramatically reduce gaps in the true positive rates, ensuring that creditworthy consumers in different groups
are about as likely to be classified as good risks. However, this comes at a cost of some eventual defaulters being misclassified as non-defaulters. This is a cost both from a consumer welfare perspective and from a lender profit perspective. The costs from the lender profit perspective can be mitigated if the introduction of fairness constraints is paired with model improvements.

Overall, our results highlight the potential benefits of imposing fairness constraints by explicitly considering certain geographies (e.g., LMI and non-LMI neighborhoods) that are likely correlated with protected attributes during the design of loan approval policies.\[^3\]

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 than 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

\[^3\]See Gillis (2021) for a discussion of output-based fairness scrutiny of credit scoring models from a legal perspective.
these disparities and highlight the potential benefits of using protected categories for lending decisions in a way that makes outcomes more equitable.

The recent rise of interest in machine learning in lending resulted in a set of papers that either introduce new default prediction models (for example, Sadhwani et al., 2020 and Albanesi and Vamossy, 2019) or examine the impact of such models on different groups of borrowers (Fuster et al., 2021, Blattner and Nelson, 2021, Gillis et al., 2021). These papers provide the most direct basis for our work. Our contribution relative to them is to introduce fairness constraints, and to examine how fairness constraints interact with model change.

2. Data

The primary data set 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 (for additional information about the CCP, see Lee and van der Klaauw, 2010; for more general discussion credit report data, see Avery et al., 2003). 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.

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

We construct a model of consumer credit default using the Consumer Credit Panel (CCP), an anonymized, nationally representative random sample of individuals with a credit file. The dataset includes quarterly snapshots of credit bureau information on credit accounts, credit inquiries, and public records (e.g., collections, bankruptcy, foreclosure, and tax liens) at the consumer level. The credit information includes several hundred variables covering the 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 the 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 the data for years 2000 to 2021. For tractability, we use a random sample of 1/100 of consumers in the panel.

Our data-cleaning procedure includes removing duplicate consumer–quarter pairs, observations without a valid census tract information, deceased consumers, and consumers that are only intermittently available in the dataset (fragment files). Once a consumer is identified as deceased in quarter $t$, they are removed from quarter $t$ and all quarters thereafter. To eliminate intermittent consumers, we only keep consumer–quarters that have eight consecutive quarters of current delinquency status following the current quarter. This procedure is similar to Hunt and Wardrip.
of consumers already in default

Our data-cleaning procedure also excludes consumers who are currently in default. Thus, we focus solely on transition from the state of non-default to the state of default. This is notably 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 at 180 days. Thus, since the focus of this paper is on estimating the default probabilities for the consumers who are currently not delinquent, we choose to exclude observations that are currently more than 90 days past due on one or more of their accounts.

2.3. Identifying LMI census tracts

Our main analyses focus on the differences between CRA-eligible LMI census tracts, and non-CRA-eligible non-LMI tracts. We obtain information about the status of census tracts from the dataset managed by the FFIEC. 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 tract median

---

income is less than 80% of the metropolitan statistical area/metropolitan division (MSA/MD) income. For tracts outside the MSA/MD, statewide income is used. The designations are updated annually.

3. Predictive models

We predict consumer repayment status using two models: logistic regression with ridge regularization (referred to simply as *logistic model*) and eXtreme Gradient Boosting (*XGBoost*).[6] The 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 to avoid overfitting [Hoerl (1962)]. It is a conceptually simple model closely resembling the OLS regression. However, it can be very flexible in modeling non-linear relationships if the input variables are preprocessed in specific ways. Here, we introduce non-linearity by binning the RHS variables (see below). Since logistic regression produces familiar regression coefficients, the results are readily interpretable (although confidence intervals are not easily available).

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 of the splits. For example, con-

---

sumers with two or three credit card accounts but no mortgage are assigned one probability default, consumers with a single credit and a mortgage are assigned another probability, and so on. Random forests average many decision trees [Breiman, 2001]. Instead of averaging the trees, XGBoost provides a scalable way to estimate trees sequentially, with each new tree focusing on examples that the previous one fit poorly. Since XGBoost assigns predictions to groups of variables based on the splits, it can incorporate a wide range of non-linear relationships. Interpretability is not as readily obtainable as in the case of logistic and other linear models, but various measures of variable importance can be calculated.

Our modeling pipeline proceeds as follows: We perform variable selection using lasso, tune the hyperparameters, train the models, and evaluate them out-of-sample in the rolling window fashion. The models are estimated with log-loss function\(^7\) and optimal hyperparameters are picked using ROC AUC criterion. 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.

3.1. LHS variable: Definition of non-default and default

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 next quarter. We define the state of default as being 90 or more days past due on at least

\[^7\text{LogLoss}_i = -[y_i \ln p_i + (1-y_i) \ln(1-p_i)], \text{ where } y_i \text{ is the outcome variable and } p_i \text{ is the predicted probability of the positive outcome.}\]
one of the accounts.

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 analysis rates for other groups are for reference only.

3.2. 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 that, 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 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 2 shows the training, gap, and evaluation quarters for all eight windows.

---

8We 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.
3.3. Variable selection and processing

To reduce the computational cost (both CPU time and RAM usage) of training the models, we perform a variable selection procedure that is separate from model training.

We begin by manually inspecting all variable definitions in CCP and select 457 variables that conceptually can be used in a credit scoring model.

We then bin the variables because many of them 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 binning using XGBoost, which creates splits in individual variables that explain the most variation in the outcome variable in the training set. We explicitly recode missing observations as their own level called missing. We use step_discretize_xgb implementation in the embed library.

Next, we train lasso models for a sequence of regularization strength hyperparameters. Lasso is a regularized linear model with a penalty on the sum of absolute values of the coefficients (Tibshirani, 1994). Such penalty assigns zero to a large number of coefficients making lasso useful for variable selection. We choose the smallest regularization strength value that results in at least 100 variables having at least one bin with a non-zero coefficient. The variable selection is performed for each rolling window using 50% of the training set observations.

The choice to use 50% of observations for variable selection and the choice to pick the regularization strength corresponding to approximately 100 variables are driven by RAM constraints.
The variables selected by the lasso are then used for model training. For the logistic model, they are binned in the way described previously. XGBoost is designed to work with missing values and mixed type variables by default, so no binning is required.

3.4. 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. The final model is then estimated on the full training set.

For XGBoost, the hyperparameters we tune are the number of trees (trees in tidy-models), tree depth (tree_depth), number of predictors for individual trees (mtry), minimum number of observations in a node (min_n), minimum loss reduction required to make a split (loss_reduction), sample size for individual trees (sample_size), and learning rate (learn_rate). Together, these parameters affect the complexity of the model (and so the regularization strength), the performance of individual trees and the model optimization dynamics. We sample 100 potential hyperparameter combinations. The bounds for sampled values are based on [Blattner and Nelson (2021)](#).

Due to its non-linear nature, XGBoost is more sensitive to the way the validation set is chosen than the logistic model. Because of that, 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 is more robust to changes in underlying data generating process over time. The final model
is estimated on the full training set.

To help prevent the XGB model from overfitting, we employ early stopping for both hyperparameter tuning and model training. 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. We use a 20% random sample of the training set as the early stopping holdout set.

3.5. 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 takes values from 0.5 to 1 (the higher, the better) and evaluates the performance of a binary classifier by aggregating true positive and false positive rates at every possible threshold.

In the sections of the paper that discuss fairness constraints, we focus directly on the true positive rate (TPR) and false positive rate (FPR). We also compare the percentages of population that fall into true positive (TP), false positive (FP), true negative (TN), and false negative (FN) groups.[10]

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

[10] See Rodolfa et al. (2016) for a high-level introduction to these concepts from a Fairness in Machine Learning perspective.
Since the losses from loan defaults tend to outweigh the gains from repaid loans from the lenders’ perspective, we also look at weighted measure of simulated profits, which we define as \( Pr = TP - \lambda FP \). At zero profit, \( \lambda \) may be viewed as the number of good accounts required to break even on a single charged-off account. The value of \( \lambda \) can vary widely, depending on the type of loan and the lender’s pricing strategy and risk management. For the typical prime credit card portfolio, the lender may need as many as six profitable accounts to make up for a single charged-off account. For a subprime portfolio, the ratio of good to bad accounts may be closer to 3:1. In our main specification, we assume a generic loan product and set \( \lambda \) to 4. Importantly, our measure of profits doesn’t reflect actual profits by the lenders and is not measured in dollars. It is, fundamentally, a measure of model performance. We call this measure profit because it reflects the fact that lender profits are negatively affected by false negatives to a larger degree than they are affected positively by true positives.

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 the 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 1% 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, XGBoost performs better than the logistic regression.
This confirms the first important fact from the recent literature - more advanced
models do produce better credit scores (e.g., Albanesi and Vamossy 2019, Fuster

The second important result from the literature is that credit scores do not per-
form equally well at predicting default for different groups. Figure 1, Panels B and
C, show the out-of-sample predictions separately for the 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 sociodemographics are not directly taken into account
by these models. While 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) and other papers in non-LMI/LMI tract context.

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

Having documented the gap in predictive power, we turn our attention to mit-
igating it. We choose a specific performance metric to equalize (TPR), implement
a procedure based on machine learning literature to reduce the disparities in that
metric by choosing group-specific decision thresholds, and explore the trade-offs that
Differences in the predictive power across demographic groups is a common issue in machine learning, with a large literature, commonly called *Fairness in Machine Learning* focused on measuring and mitigating the disparities. Chouldechova and Roth (2018) provide a view of the frontier of the academic literature in 2018. Rodolfa et al. (2016) provides an introduction with a list of references that focus on practical implementations of fairness interventions. A variety of methods exist to make model predictions more fair, based on some criterion by preprocessing the data, by modifying the predictive algorithms, or by adjusting existing predictions. We choose to focus on adjusting existing predictions because of the tractability and ease of implementation. Similar approaches have been used in education, mental health, criminal justice, and housing safety settings (Rodolfa et al., 2021). We leave the comparison of alternative approaches to future work.

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 the credit score of zero, the lender is certain that the consumer
will default and at the credit score of 100, the lender is certain the consumer will not default. Paying the loan back (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.

Every threshold creates a confusion matrix. Figure 2 is a hypothetical example. Suppose we have 100 consumers and pick a threshold of 60 (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 (true positives) and 10 default (false positives). For the 50 consumers who we predicted would not repay their loan, 30 of them do not repay the loan (true negatives) and 20 of them do repay the loan (false negatives). In the second confusion matrix on 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 true positives, but also the number of false positives. 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 false negatives has decreased, but so has the number of false positives. Ideally, we want to maximize the number of true positives and true negatives while
minimizing the number of false positives and false negatives. However, in practice non-defaulters and defaulters are not perfectly separated in the credit score space. Therefore, no threshold change from $t$ to $t^*$ (either an increase or decrease) will have $TP_{t^*} > TP_t$ and $TN_{t^*} > TN_t$ at the same time as $FP_{t^*} < FP_t$ and $FN_{t^*} < FN_t$. 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 rate (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. 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 four, so that $Pr = TP - 4 \times FP$, see next).

4.3. Objectives of lenders and the regulator differ

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 $\lambda$ units of money. We set $\lambda$ to be four in our main specification.

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. That is operationalized as the difference in TPR between the non-LMI and LMI tract groups, $\Delta TPR = TPR(non-LMI) - TPR(LMI)$. Because the regulator also values the ongoing viability of the lender, it might accept a partial reduction in $\Delta TPR$ that results in a smaller reduction in lending profit instead of requiring $\Delta TPR$ to be zero.

We operationalize 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). *Strong fairness constraints* correspond to setting separate thresholds for the groups in such a way that $\Delta TPR$ is 0 in-sample. *Medium and weak fairness constraints* involve setting thresholds that are located 66% and 33%, respectively, of the way between the strong fairness constraint threshold and blind threshold. 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
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 together using the XGBoost credit score based on the in-sample data of last rolling window in Figure 3 Panel A. This figure is similar to a ROC curve in that it includes TPR and FPR, but unlike an 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 90% of creditworthy consumers get the loan and 29% 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 for thresholds toward the middle of the plot, for a given threshold, non-LMI consumers have a substantially higher TPR. While non-LMI tract consumers have a smaller FPR than LMI tract consumers at a given threshold, the gap between TPR for non-LMI and LMI consumers is far greater than the gap between FPR for the two groups. A creditworthy non-LMI tract consumer has a 91% chance to get the loan, but a LMI tract consumer who is equally creditworthy has only an 85% chance to get the loan. In 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.

It is possible to completely eliminate the gap in TPR and satisfy the regulator’s fairness objective by keeping the threshold for the non-LMI tract group the same as the single threshold, and reducing the threshold for the LMI group up to the point where TPR become equal. In Figure 4, Panel A, we select a threshold for LMI consumers, and a threshold for non-LMI consumers such that both groups have a TPR of roughly 91%. This way, creditworthy consumers have a 91% chance to get the loan regardless of which group they are in.11

Under separate thresholds, more creditworthy LMI consumers are classified as

---

11The 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.
good credit than under a single threshold. While this approach increases their TPR, it also increases their FPR. In Figure 4, we see how lowering the threshold for the LMI group increases their FPR. This is important from the lenders’ perspective since the number of FP enters the lenders’ objective function with a multiplier of $\lambda$, which we set to four. This is key for the fairness–profit trade-off.

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.\footnote{This is a modification of the equal opportunity thresholds from Hardt et al. (2016). The original approach considers all pairs of thresholds which 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 the 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 weights can be accommodated. For simplicity, we focus on three possible levels of fairness constraints. 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% and 33% 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 (see Section 3.2) based on the training set (in-sample) model outputs. 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.

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 non-LMI and LMI groups are large for every model we consider. In Section 4.5, we showed how we can establish separate lending thresholds to reduce disparities in equality of opportunity. Now we will look at the fairness-profit trade-offs that arise from this approach and how the trade-offs interact with model improvement.

The analyses in this section are based on fairness constraints applied out-of-sample to the test sets. We apply the thresholds generated on the training set as previously discussed to model predictions on the test set 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 1,000 is the maximum possible profit. As discussed previously in Section 3.5, our measure of profits does not reflect actual profits by lender. Rather, it is a measure of model performance that weights FP higher than TP, which is a feature it shares with lenders’ profit function. In our main specification, we set \( \lambda \) to four. In Figure 6, we also report results for \( \lambda \in \{2, 4, 6\} \). The y axis is
the difference in TPR between the protected group and the unprotected group and is our measure of equality of opportunity ($\Delta TPR$). The color is the model type, and the label is strength of the constraint.

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 0.6% of profits to eliminate the TPR gap. So fairness does not come for free and affects the 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 tend to shift the fairness-profit curve rightward, increasing profitability at every threshold. This is intuitive, since a better performing model approves fewer defaulters at every threshold level. 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 fairness constraints. 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 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, $TP - \lambda FP$, with $\lambda = 4$ in this case. In our illustrative setting, going from the logistic model with single threshold to XGBoost model with strong fairness constraint decreases gap in TPR from 7% to 0%, while increasing profits from just below 990 to just under 995. 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 1,000).

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), the FP group to be losers (they are more likely to suffer default on the loan they might get due to the credit score increase), the TN group to be neutral (as they don’t get a loan they can’t pay back), and the FN group to be losers (they don’t get
a loan they can pay back). Note that this approach to labeling consumers as winners or losers is different from Fuster et al. (2021) who treat consumers as winners from a model change if their credit score goes up. We do not consider defaulters with increases in credit score as winners since, even if they are more likely to get a loan, eventual default means negative financial consequences and reduced ability to gain loans in the future.

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.

13 All changes are among the LMI tract individuals. In the single threshold scenario, 58.6% 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 68.2%. By definition, the increase comes from the decrease in the FN group, reducing the number of losers. However, the FP % for the LMI group also rises, from 2.3% to 4.1%. 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 a from credit score qualifying them for a loan, but they also are not hurt by

\footnote{This is slightly different from the original equality of opportunity approach in Hardt et al. (2016). There, the separate thresholds are picked in way that the threshold for the protected is lower (as it is in our case), and the threshold for the unprotected group is slightly higher, resulting in a slightly higher percentage of TP and a slightly lower percentage of FP in the unprotected group, as well as slightly higher lender profits.}
the potential consequences of defaulting on more loans. While many more consumers benefit from a fairness constraint as are hurt by it, the increase in the FP group needs to be taken into consideration as a policy cost.

5. Discussion

A growing body of research shows that credit scores do not perform equally well at predicting default for different groups of individuals, whether the groups be defined according to race or ethnicity, income, or depth of credit file. As a result, lending decisions based on a single, group-unaware threshold can be unfair, in the sense that different groups can have different shares of creditworthy borrowers identified by the model as such (TPR).

Our illustrative examples suggest that applying group-specific lending thresholds in default prediction can reduce disparities in true positive rates 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 who simultaneously adopt both machine learning models and group-specific lending thresholds may experience increases in fairness as well as profit. This can occur when the machine learning model identifies sufficiently many creditworthy loan applicants in both the LMI and non-LMI neighborhoods (relative to the baseline model) 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. 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 machine learning 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 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 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,

\[\text{\footnote{Throughout the paper, we focus on the outcome of a lending decision in which an applicant is either approved or rejected for a loan, and may subsequently terminate the loan in good standing or default. We assume that higher approval rates in non-defaulting populations make the individuals better off. Of course, one could argue that an LMI individual who is approved for a loan at significantly more onerous terms has not been made better off, however, a full accounting of individuals’ welfare is beyond the scope of this paper.}}\]
widespread adoption of machine learning 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 machine learning 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 tradeoff exhibited by our models would not be present at the lender level, our analysis does not provide sufficient evidence to rule this possibility out. Thus, we caution the reader that the following discussion rests upon the assumption that our market-level outcomes and tradeoffs 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 fall under distinct regulatory umbrellas that may, at time, 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.

5.1. 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.

We leave open the question of whether group-specific thresholds are permissible.
under the existing fair lending law. Instead, in the following section, we describe how group-specific thresholds might fit under the intuition of the special purpose credit programs permitted under the ECOA.

5.1.1. Special purpose credit programs

While the fair lending law 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 those who would be unlikely to receive credit under the lender’s customary lending standards or would receive it on less favorable terms (12 C.F.R. §1002.8(a)(3)(ii)).

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 several special purpose credit programs with the objective of “enabling access to credit and encouraging sustainable homeownership for Black consumers.” Fannie Mae’s special purpose credit programs are focused on “people residing in formerly redlined and other underserved areas with majority Black populations.”

A 2020 Advisory Opinion by the Consumer Financial Protection Bureau (CFPB) noted that a lender can initiate a special purpose credit program without CFPB approval, provided the lender’s program meets the compliance standards and general

\footnote{See https://www.fanniemae.com/media/43636/display Last accessed: 08/11/2022.}
rules set forth in Regulation B (Official Interpretations, 12 C.F.R. pt. 1002(supp. I), sec. 1002.8, 8(a)-1).[16] 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 credit under the program.[17] The plan must also state the expected duration of the program and the criteria by which its continuing need will be evaluated.

The ECOA also states that, although the lender may use protected class status to determine eligibility for a special purpose credit program, the lender cannot discriminate against an applicant on a prohibited basis.[18]

It is unclear whether, by providing credit to persons in a protected class on more favorable terms than persons belonging to another class, special purpose credit programs pose a fair lending risk due to so-called reverse discrimination, since members of a non-protected class who have historically received favorable treatment would not qualify for the program. As noted previously, in our paper we show 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.

---


Recently, several regulatory agencies including the CFPB distributed interagency guidance that encourages lenders to explore opportunities to develop special purpose credit programs. The guidance notes that lenders are permitted to consider the use of special purpose credit programs across all types of credit covered by the ECOA and Regulation B. Thus, the ECOA’s exception for special purpose credit programs may be a viable starting point for future discussions into the implementability of group-specific lending thresholds for LMI and non-LMI neighborhoods.

5.2. Machine learning adoption

The second major challenge to dual adoption of machine learning and group-specific lending thresholds is the risk, expense, and uncertainty surrounding the use of machine learning models in credit underwriting. While sophisticated machine learning models are pervasive in fintech lending, banks and other traditional lenders have proceeded more cautiously when considering the use of machine learning models for credit risk assessment. The use of machine learning 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 machine learning in credit underwriting. This is partly due to the number of ways in which machine learning models complicate internal model development and governance processes, as well as lenders’ ability to satisfy their legal and regulatory requirements. Machine learning

---

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 machine learning 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 generate adverse action notices as required by law (FinRegLab, 2021). 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 machine learning models in a variety of lending decisions that potentially affect millions of individuals, but that uncertainty remains, in no small part, due to the complexity of machine learning models and the uncertainty as to how these models fit into existing legal and regulatory frameworks.

5.3. Dual adoption strategy

The possibility of dual adoption of group-specific lending thresholds and machine learning models seems unlikely to occur overnight, given the challenges of fair lending law and machine learning adoption. However, there may be an opportunity for lenders and regulators to leverage the provision of the special purpose credit program to learn more about the effects of machine learning-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

---

20The 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.
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 business ownership.

The adoption of group-specific lending thresholds needn’t be limited to lenders that have yet to adopt machine learning underwriting models. Within the group of lenders that have already made the transition from regression-based models to machine learning credit underwriting models — particularly lenders with a mission to reach underserved populations — lenders could be encouraged to design and adopt their own special purpose credit program. 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 as well as a legacy regression model. Likewise, fintech lenders that have been using machine learning models since their inception might benchmark against a previous version of their model, a regression-based model, or a model without alternative data.

6. Conclusion

The gap in true positive rates between non-LMI and LMI individuals can be reduced by adopting group-specific thresholds. However, this equality comes with costs: Lender profits are affected, and there are winners and losers among different
groups, including the protected ones. 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 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 gap (Blattner and Nelson, 2021).
References


Hunt, R., Wardrip, K., 2013. Residential migration, entry, and exit as seen ythrough the lens of credit bureau data. Working paper.


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 the U.S. Census Bureau data.
Table 1: Observations by current delinquency status, CRA status of their census tract, and repayment outcome. The calculations are based on a 1% sample of CCP. Individuals are observed at the quarterly frequency. % Default column indicates the stock of defaulters in the given group. We provide overall default statistic and split the results by CRA status, current delinquency status, and both (% observations in each of these splits sums to 100%). 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>
<thead>
<tr>
<th>Consumer Group</th>
<th>% of Observations</th>
<th>% Default (Next 2 Y.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>100.0</td>
<td>22.3</td>
</tr>
<tr>
<td><strong>Census Tract Status</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-LMI</td>
<td>79.9</td>
<td>19.4</td>
</tr>
<tr>
<td>LMI</td>
<td>20.1</td>
<td>33.8</td>
</tr>
<tr>
<td><strong>Current Delinquency Status</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current</td>
<td>84.4</td>
<td>8.5</td>
</tr>
<tr>
<td>&lt;90 Days Past Due</td>
<td>2.8</td>
<td>59.9</td>
</tr>
<tr>
<td>≥90 Days Past Due</td>
<td>12.8</td>
<td>93.4</td>
</tr>
<tr>
<td><strong>Census Tract Status &amp; Current Delinquency Status</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-LMI &amp; Current</td>
<td>68.1</td>
<td>7.8</td>
</tr>
<tr>
<td>Non-LMI &amp; &lt;90 Days Past Due</td>
<td>2.3</td>
<td>58.0</td>
</tr>
<tr>
<td>Non-LMI &amp; ≥90 Days Past Due</td>
<td>9.5</td>
<td>93.3</td>
</tr>
<tr>
<td>LMI &amp; Current</td>
<td>14.9</td>
<td>14.0</td>
</tr>
<tr>
<td>LMI &amp; &lt;90 Days Past Due</td>
<td>0.7</td>
<td>66.6</td>
</tr>
<tr>
<td>LMI &amp; ≥90 Days Past Due</td>
<td>4.5</td>
<td>94.6</td>
</tr>
</tbody>
</table>
**Table 2:** 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. Training set is also used to compute lending thresholds that are applied to the evaluation quarters out-of-sample.

<table>
<thead>
<tr>
<th>Window</th>
<th>Training quarters</th>
<th>Gap</th>
<th>Evaluation quarters</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2008Q1–2009Q4</td>
<td>2010Q1–2011Q4</td>
<td>2012Q1–2013Q4</td>
</tr>
<tr>
<td>7</td>
<td>2012Q1–2013Q4</td>
<td>2014Q1–2015Q4</td>
<td>2016Q1–2017Q4</td>
</tr>
</tbody>
</table>
Figures

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 the U.S. Census Bureau data.
Figure 1: Performance of repayment status predictions on the combined evaluation set, 2004Q1–2019Q4. ROC AUC is a measure of binary classifier performance running between 0.5 and 1 (the more, the better) that accounts for true positive rate (how many non-defaulters are correctly identified as such) and false positive rate (how many defaults are incorrectly identified as non-defaulters) at every possible decision threshold. The two lines correspond to 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.
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 2 years), at the 100 threshold, everyone is predicted to be negative (in default within 2 years). The cells correspond to (left to right, top to bottom): true positives (TP), false positives (FP), false negatives (FN), true negatives (TN). Changing the threshold values changes values in all four cells.
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 $TP$ are the true positives and $FN$ are the false negatives. $FPR$ is defined as $\frac{FP}{FP+TN}$, where $FP$ are the false positives and $TN$ are 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 are the true positives and FN are the false negatives. FPR is defined as $\frac{FP}{FP+TN}$, where FP are the false positives and TN are 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 logistic and XGBoost models. Profit is calculated as $Pr = TP - 4 \times FN$, where TP is the number of true positives, FP is the number false positives. The positive outcome is non-default within the next two years $\Delta TPR$ is the difference in TPR between the non-LMI and LMI groups, which is the measure of fairness we adopt in this paper. $TPR$ is defined as $\frac{TP}{TP+FN}$, 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. Labeled points on lines represent different strength 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 for logistic and XGBoost models. Profit is calculated as $Pr = TP - \lambda \times FN$, where TP is the number of true positives, FP is the number false positives, and $\lambda$ 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 so that largest possible value is 1,000. The positive outcome is non-default within the next two years $\Delta TPR$ is the difference in TPR between the non-LMI and LMI groups, which is the measure of fairness we adopt in this paper. TPR is defined as $\frac{TP}{TP+FN}$, 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. Labeled points on lines represent different strength 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 strength of the fairness constraint, with Blind corresponding to single threshold without fairness constraints and Strong represents the most stringent constraint aimed at eliminating $\Delta TPR$. 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 strength of the fairness constraint: Strong, Medium, Weak, and Blind (see Section 4.5 for definitions).