Predictably Unequal?
The Effects of Machine Learning on Credit Markets

Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther[1]

This draft: November 2017

[1] Fuster and Goldsmith-Pinkham: Federal Reserve Bank of New York. Email: andreas.fuster@ny.frb.org, paul.goldsmith-pinkham@ny.frb.org. Ramadorai: Imperial College, London SW7 2AZ, UK, and CEPR. Email: t.ramadorai@imperial.ac.uk. Walther: Warwick Business School. Email: Ansgar.Walther@wbs.ac.uk. We thank John Campbell, Jediphi Cabal, Ralph Koijen, Karthik Muralidharan, Johannes Stroebel, and Stijn van Nieuwerburgh for useful conversations and seminar participants at Imperial College Business School, NYU-Stern, and University of Rochester for comments. We also thank Kevin Lai, Lu Liu, and Qing Yao for research assistance. The views expressed are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of New York or the Federal Reserve System.

Abstract

Recent innovations in statistical technology, including in evaluating creditworthiness, have sparked concerns about impacts on the fairness of outcomes across categories such as race and gender. We build a simple equilibrium model of credit provision in which to evaluate such impacts. We find that as statistical technology changes, the effects on disparity depend on a combination of the changes in the functional form used to evaluate creditworthiness using underlying borrower characteristics and the cross-category distribution of these characteristics. Employing detailed data on US mortgages and applications, we predict default using a number of popular machine learning techniques, and embed these techniques in our equilibrium model to analyze both extensive margin (exclusion) and intensive margin (rates) impacts on disparity. We propose a basic measure of cross-category disparity, and find that the machine learning models perform worse on this measure than logit models, especially on the intensive margin. We discuss the implications of our findings for mortgage policy.

1 Introduction

In recent years, there has been significant innovation and improvement in predictive modelling, aided by an explosion of data availability and rapid advances in computing power. These new developments in machine learning and statistical technology have been rapidly adopted by businesses in a broad range of industries, and are also increasingly being utilized in academic economics.[2]

Alongside the undoubted efficiency benefits arising from widespread adoption and use of these new prediction technologies, concerns have also arisen about whether their use leads to gains for all groups in the economy. Perhaps the most widely voiced of these concerns is that there may be the potential for bias against certain groups in the population arising from decision-support systems which rely on sophisticated algorithms trained on past historical datasets.[3]

Household financial markets are an area of the economy in which there is vast potential for efficiency gains from the use of machine learning. This is evident from the rapid uptake of such technology in consumer finance from both incumbents and challengers, especially in the area of credit provision.[4] It is also the case that a significant focus of policy in this sphere has been to ensure equality of opportunity and access in household financial markets.[5] In practice, this has usually been interpreted to mean that differentiation between households using "excluded" characteristics such as race or gender is prohibited (see, e.g., Ladd, 1998). A related issue arises with "redlining," which seeks to use geographical information to indirectly

[2]See, for example, Belloni, Chernozhukov, and Hansen (2014), Varian (2014), Mullainathan and Spiess (2017), Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey (2017), and Athey and Imbens (2017).

[3]See, for example, O'Neil (2016), Hardt, Price, and Srebro (2016), Kleinberg, Mullainathan, and Raghavan (2016), and Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan (2017).

[4]Academic work in the area includes Khandani, Kim, and Lo (2010), and Sirignano, Sadhwani, and Giesecke (2017).

[5]A partial list of laws in the US focused on ensuring equal treatment in financial markets to all households regardless of race or gender includes Civil Rights Act (1966), Fair Housing Act (1968), Employee Retirement Income Security Act (1974), Equal Credit Opportunity Act (1974), and the FDIC Policy Statement on Lending (2004).

differentiate on the basis of such characteristics, and which is also prohibited.[6]

In this paper, we attempt to better understand the distribution of gains and losses from improvements in statistical technology. Our focus is on how such technological change affects outcomes for different groups (e.g., racial groups, or different genders), and we study these issues in household credit markets.

We begin by building a simple theoretical framework to identify the likely groups of winners and losers in credit markets as the technology used to identify creditworthiness changes. We first note that a more sophisticated technology (in the sense of reducing predictive mean squared error) will, by definition, produce predictions with greater variance. This means that better technology shifts weight from average predicted default probabilities to more extreme values. As a result, there will always be some borrowers with characteristics that are treated as less risky under the new technology, and therefore experience better credit market outcomes, while borrowers with other characteristics will be considered to be riskier. The question is then how these winners and losers are distributed across different groups or categories of borrowers.

We therefore attempt to provide guidance to identify the specific groups most likely to win or lose from the change in technology. We solve this in closed form for a lender who uses a single exogenous variable (e.g., a borrower characteristic such as income) to predict default, and then provide graphical intuition for this result in the case of a lender using two input variables to predict default. The essential insight is that winning or losing depends on both the functional form of the new technology, and the differences in the distribution of the characteristics across groups. Perhaps the simplest way to understand this insight is to consider an economy endowed with a primitive prediction technology which simply

[6]These issues have been a major focus on work in household financial markets. In mortgages and housing, see, e.g., Berkovec, Canner, Gabriel, and Hannan (1994, 1998), Ladd (1998), Ross and Yinger (2002), Pope and Sydnor (2011), Ghent, Hernández-Murillo, and Owyang (2014), and Bayer, Ferreira, and Ross (2017). In insurance markets, see, e.g., Einav and Finkelstein (2011), Chetty and Finkelstein (2013), Bundorf, Levin, and Mahoney (2012), and Geruso (2016). This work is also connected to broader theories of discrimination. See Fang and Moro (2010) for an excellent survey, and the classic references on the topic, including Becker (1971), Phelps (1972), and Arrow (1973).

uses the mean level of a single characteristic to predict default. In this case, the predicted default rate will just be the same for all borrowers, regardless of their particular value of the characteristic. If a more sophisticated linear technology which identifies that default rates are linearly increasing in the characteristic becomes available to this economy, groups with higher values of the characteristic than the mean will clearly be penalized following the adoption of the new technology, while those with lower values will benefit from the change. Similarly, a convex quadratic function of the underlying characteristic will penalize groups with higher variance of the characteristic, and so forth.

Credit default forecasting generally utilizes large numbers of variables, and machine learning involves highly nonlinear functions. This means that it is not easy to identify general propositions about the cross-group joint distribution of characteristics and the functional form predicting default. Indeed, we note that the impact of new technology could be either negative or positive for any given group of households – there are numerous real-world examples of new entrants with more sophisticated technology more efficiently screening and providing credit to members of groups that were simply eschewed by those using more primitive technologies.[7]

We therefore provide evidence on these issues by utilizing increasingly sophisticated statistical models, beginning with a simple logistic regression of default outcomes on borrower and loan characteristics, and culminating in a random forest model (Ho, 1998; Breiman, 2001) to predict default in a large dataset of over 10 million US mortgages originated between 2009 and 2014. We then embed these reduced form default forecasting models in a simple equilibrium model of credit provision in a competitive credit market, in order to be able to assess outcomes on both the extensive margin (access to credit) and the intensive margin (rates conditional on obtaining credit). We compute the counterfactual equilibria associated with each statistical technology on a subset of our data (loans originated in 2011, in this version of the paper), and then compare the resulting equilibrium outcomes with one

[7]The monoline credit card company CapitalOne is one such example of a firm that experienced remarkable growth in the nineties by more efficiently using demographic information on borrowers.

another. We use this approach to evaluate comparative statics on outcomes across groups as the underlying statistical technology available to lenders changes.

When implementing these models, we follow the *letter* of the law, and behave as lenders would, eliminating race and gender from the set of borrower characteristics. However we are able to contrast the performance of the models when race is included and withheld. We find that the logistic regression models benefit from the inclusion of race in the sense that it improves their predictive accuracy, while the machine learning model is barely affected by this inclusion. This is interesting since the *spirit* of the law suggests that the models assessing borrower risk should be colorblind. While this does seem to be the case for the more primitive models, it does not for the machine learning model, suggesting that this model is able to more efficiently triangulate the association between race and default using the remaining borrower characteristics. This is reminiscent of recent work in the computer science literature which shows that anonymizing data is ineffective if sufficiently granular data on characteristics is available on the individual entities (e.g., Narayanan and Shmatikov, 2008).

When we embed these statistical models in competitive equilibrium, we find that the machine learning model appears to provide a slightly larger number of borrowers access to credit. We also create a simple measure of cross-group disparity in outcomes (for both the probability of acceptance and rates conditional on acceptance), which is simply the standard deviation of group outcomes, weighted by group representation in the population, around the population mean outcomes. We find that on this metric, the machine learning model also does marginally better on the extensive margin, i.e., the probability of acceptance also varies slightly less across groups using this more sophisticated technology.

However, the story is different on the intensive margin. While the machine learning model accepts a few more borrowers, it has a significantly higher average rate for these borrowers across the board, by roughly 20 basis points, or roughly 4% of the average rate in the sample. More importantly, the cross-group disparity of these rates is substantially (three times) higher than for the less sophisticated logistic regression models. This reflects

the differential changes in the average rate across groups. For White borrowers, the average rate under the machine learning technology rises by a little less than 20 basis points, which is naturally close to the population average, but for Black borrowers, the comparable rise is more than double this number, at 40 basis points.

Overall, the picture is mixed. On the one hand, the machine learning model is a more effective model, predicting default more accurately than the more primitive technologies. What's more, it does appear to provide credit to a slightly larger fraction of mortgage borrowers, and slightly reduce cross-group dispersion in acceptance rates. However, the main effect of the improved technology is the substantial rise in the cross-group dispersion of rates across race groups. In future versions of this draft, we intend to provide more insight into how to evaluate these tradeoffs from a social welfare perspective.

The organization of the paper is as follows. Section 2 sets up a simple theory framework to understand how improvements in statistical technology can affect different groups of households in credit markets. Section 3 discusses the US mortgage data while Section 4 turns to the default forecasting models that we employ on these data. Section 5 discusses how changes in technology affect measures of disparity in the US mortgage data, and Section 6 concludes.

2 A Simple Theory Framework

Consider a mortgage lender who wishes to predict the probability of default, $y \in [0, 1]$, by a borrower with observable characteristics x. For now, we study the lender's inference problem given a mortgage contract (interest rate, loan-to-value ratio, etc.).[8]

The lender's inference problem is to find a function $\hat{y} = \hat{f}(x) \in \mathcal{M}$ which maps the observable vector x into a predicted y. The statistical technology that is available to the lender to find this function is represented by $\mathcal{M}$, a class of possible functions that can be

[8]We subsequently allow interest rates to be determined in competitive equilibrium.

chosen. For example, if technology only permits linear regression, then $\mathcal{M}$ is the space of linear functions of x. We say that a statistical technology $\mathcal{M}_2$ is *better than* $\mathcal{M}_1$ if it gives the lender a larger set of options, i.e., $\mathcal{M}_1 \subset \mathcal{M}_2$.

We assume that the lender chooses the best predictor in a mean-square error sense, subject to the constraint imposed by the available statistical technology:

$$\hat{f}(x|\mathcal{M}) = \arg\min_f E[(f(x) - y)^2] \text{ subject to } f \in \mathcal{M}. \tag{1}$$

Note that the prediction $\hat{f}(x|\mathcal{M})$ is itself a random variable, since it depends on the realization of characteristics x.

We first show that better technology necessarily leads to predictions that are more disperse:

Lemma 1. If $\mathcal{M}_2$ is a better statistical technology than $\mathcal{M}_1$, then $\hat{f}(x|\mathcal{M}_2)$ is a mean-preserving spread of $\hat{f}(x|\mathcal{M}_2)$, that is:

$$\hat{f}(x|\mathcal{M}_2) = \hat{f}(x|\mathcal{M}_1) + u,$$

where $E[u] = 0$ and $Cov(u, \hat{f}(x|\mathcal{M}_1)) = 0$.

Proof: See Appendix.

This result is intuitive. To begin with, almost by definition, better technology will yield predictions with a lower mean-square error. The resulting predictions $\hat{y}$ will track the true y more closely, and will therefore be more disperse on average. Moreover, this spread is mean-preserving, because optimal predictors are unbiased and will match the true y *on average* regardless of technology.

Lemma 1 motivates a concern that there will be both winners and losers when better technology becomes available in credit markets. Better technology shifts weight from average

predicted default probabilities to more extreme values. As a result, there will be borrowers with characteristics x that are treated as less risky under the new technology, and therefore experience better credit market outcomes, while borrowers with other characteristics will be considered to be riskier.

Who gains in credit markets, and who loses, when statistical technology improves? In the remainder of this Section, we analyze two simple cases that serve to build intuition. In both of these cases, we attempt to evaluate the impacts of introducing a more sophisticated statistical technology on subgroups of borrowers in the broader population. In what follows, we characterize these subgroups by the conditional distributions of their characteristics, $x|g$.[9]

2.1 The Case of One-Dimensional Borrower Characteristics

We begin by considering the case in which lenders use only one dimension of borrowers' characteristics to predict default (i.e., default is a function of a scalar x). Furthermore, we assume that the inferior technology $\mathcal{M}_1$ is the class of linear functions of x, and that the better technology $\mathcal{M}_2$ is a more general class of nonlinear, but smooth (i.e., continuous and differentiable), functions of x. Using a Taylor series representation of the improved estimate $\hat{f}(x|\mathcal{M}_2)$, we can then characterize the impact of new technology on group g in terms of the conditional moments $x|g$:

Lemma 2. Let $\mathcal{M}_1$ be the class of linear functions of x, and suppose that borrower characteristics $x \in [\underline{x}, \bar{x}] \subset \mathbf{R}$ are one-dimensional. Then the impact of the new statistical technology on the predicted default rates of borrower group g is:

$$E[\hat{f}(x|\mathcal{M}_2) - \hat{f}(x|\mathcal{M}_1)|g] = \sum_{j=2}^{\infty} \frac{1}{j!} \frac{\partial^j \hat{f}(a|\mathcal{M}_2)}{\partial x^j} E\left[(x-a)^j|g\right] - B \tag{2}$$

[9] This nests the case in which we consider borrowers individually, i.e., in groups of size 1. In this case the distribution of borrower characteristics is degenerate and places probability 1 on one particular realization of characteristics.

where a is the value of the characteristic of a "representative" borrower such that $\frac{\partial^j \hat{f}(a|\mathcal{M}_2)}{\partial x^j} = \frac{\partial^j \hat{f}(a|\mathcal{M}_1)}{\partial x^j}$, and $B = \hat{f}(a|\mathcal{M}_1) - \hat{f}(a|\mathcal{M}_2)$ is a constant.

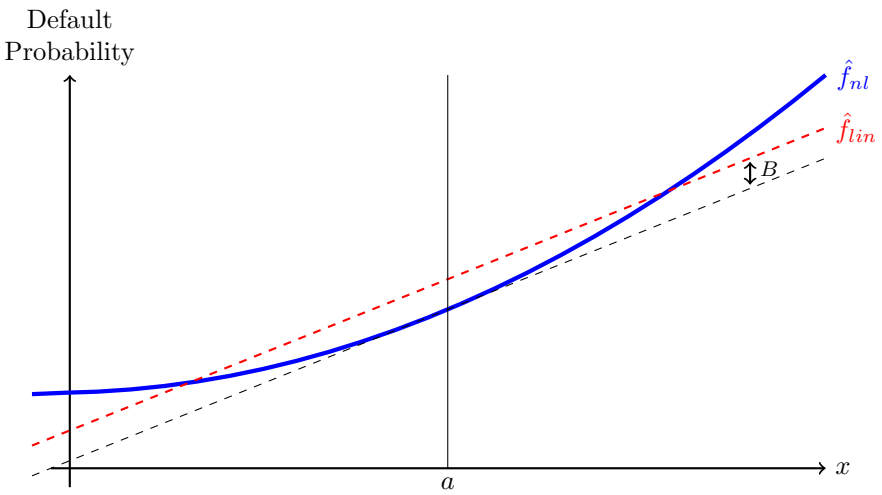
Lemma 2 shows that the impact of new technology across groups depends on two factors, namely, (i) the higher-order moments $E\left[(x-a)^j|g\right]$ of characteristics, centered around the value a of the characteristic of a representative borrower, and (ii) the higher-order derivatives of the nonlinear prediction $\frac{\partial^j \hat{f}(a|\mathcal{M}_2)}{\partial x^j}$, evaluated at a.

Figure 1 illustrates a special case, in which the prediction using the superior nonlinear statistical technology, denoted $\hat{f}(x|\mathcal{M}_2) = \hat{f}_{nl}$, is a convex quadratic function of x. The linear prediction $\hat{f}(x|\mathcal{M}_1) = \hat{f}_{lin}$ is a shifted approximation of $\hat{f}_{nl}$ around the representative point $x = a$. In this case, the leading term in equation (2) indicates that a group g will be treated as having higher default risk under this particular new technology if $E[(x-a)^2|g]$ is large, i.e., if the distribution of x given g is dispersed away from the representative borrower's value.

Under these conditions, minority borrowers, whose attributes are not representative, are likely to lose under the new technology. (It is important to note here that if the superior technology were concave rather than convex in x, this result would be reversed, of course). To see more clearly why the new technology is likely to negatively impact minority borrowers, suppose that a fraction $\mu > 1/2$ of borrowers (the majority group g_0) have attributes x_0, while the remaining $1-\mu$ (the minority group g_1) have attributes x_1. It is then easy to show that $E[(x-a)^2|g_1]$ increases to its upper bound $(x_1 - x_0)^2$ as μ approaches 1.

More generally, Lemma 2 implies that a group g of borrowers is likely to lose once superior statistical technology becomes available, if there is a positive association between the higher-order moments of the distribution of $x|g$, and the higher-order derivatives of the improved prediction $\hat{f}(x|\mathcal{M})$.

Figure 1: **One-Dimensional Example.**



For example, if the distribution of $x|g$ is right-skewed, and the third derivative of $\hat{f}(x|\mathcal{M})$ is positive, then the introduction of $\hat{f}(x|\mathcal{M})$ relative to the previously available technology will penalize the right tail of x, causing members of group g to have higher predicted default rates. Members of g would therefore lose out under the new technology. To take another example, if the distribution of $x|g$ is fat-tailed, and the fourth derivative of $\hat{f}(x|\mathcal{M})$ is negative, then the new predictions reward both tails of the conditional distribution, and members of g will be relatively better off, and so forth.

2.2 The Case of Two-Dimensional Borrower Characteristics

Now let us consider the more complex case in which borrower characteristics are two-dimensional, i.e., $x = (x_1, x_2)$. For concreteness, let x_1 be the borrower's income, and x_2 her FICO credit score.

Panel (a) of Figure 2 plots the level sets of the predicted default probabilities $\hat{f}(x|\mathcal{M}_1) = \hat{f}_{lin}$ from a linear probability model, alongside predictions $\hat{f}(x|\mathcal{M}_2) = \hat{f}_{nl}$ from a superior, non-linear model. In this particular example, we assume that this superior technology takes the Leontief shape, i.e., $\hat{f}_{nl} = \min\{ax_1, bx_2\}$. These choices of functional forms are in

anticipation of our empirical analysis, where we consider mainly Logit models of default, which have linear level sets, and machine learning models based on decision trees, which tend to yield predicted default probabilities that are kinked or step functions of underlying characteristics.[10] In this example, for both technologies, we assume that predicted default probabilities are decreasing in both FICO and income.

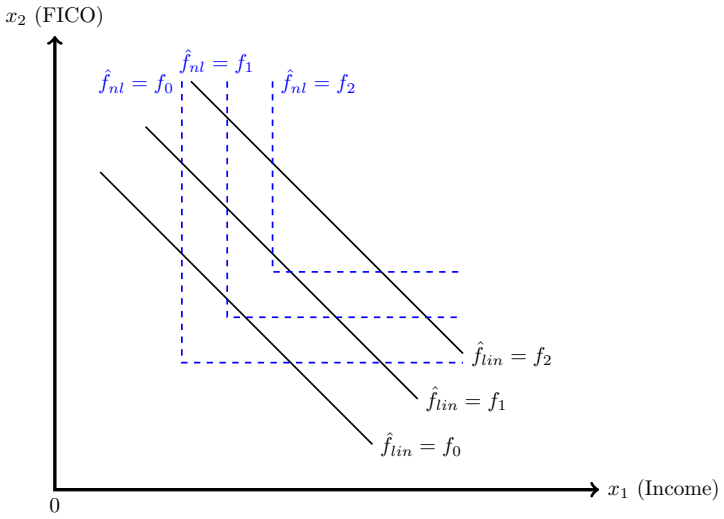
Panel (b) of Figure 2 focuses on comparing one level set across the two technologies, and shows the location of winners and losers upon the introduction of the new technology. We use the terms winners and losers to capture those who will be considered as lower or higher credit risks, respectively, under the new technology. In this example, the new technology considers income and FICO to be complementary, while the linear technology considers them to be substitutes. The losers under the new technology are therefore those borrowers who fall short on one of these criteria, while doing well on the other.

To summarize, the main insights from this analysis are as follows. First, theory clearly predicts that there will generally be both winners and losers from an improvement in statistical technology. Second, while we have studied a number of specific examples to build intuition for the potential impacts on specific groups of better technology, it is clear that these impacts are jointly determined by the shape of the underlying distribution of $x|g$ and the particular differences between the new and old predictions. Indeed, it is worth emphasizing that the intuition that we have developed using the specific functional forms (convex quadratic in the single variable case, and Leontief in the two-variable case) could well be misleading in terms of the patterns that exist in the real data. For example, consider the case in which the new technology allows a lender to more efficiently utilize demographic information in order to make better predictions, and that this technology delivers more accurate predictions by identifying the good credit risks within a minority group which was previously assigned high predicted default rates under the old technology. In this case, we might see that the introduction of new technology benefits the minority group considerably

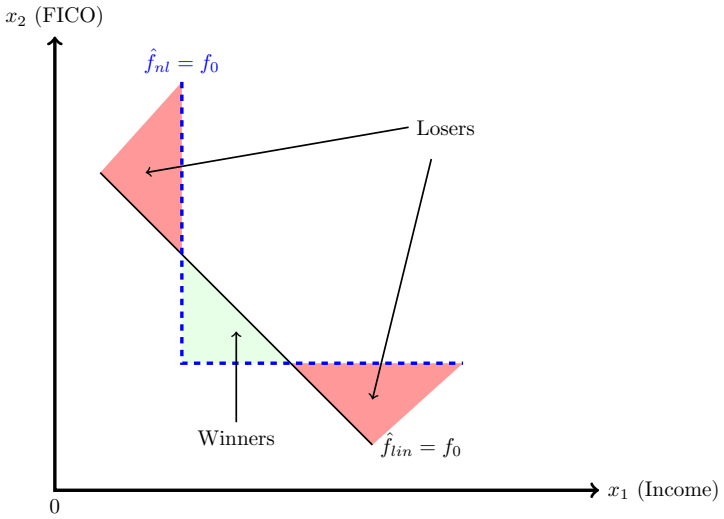
[10] We also consider logit models with binned underlying variables, in which the level sets are also step functions.

Figure 2: **Two-dimensional examples.**

(a) Level sets of predicted PD



(b) Winners and losers from improved technology



on average, though dispersion of outcomes within the group would rise as a result.[11]

As this discussion reveals, it is difficult to develop strong theoretical priors on the shape of either the functional forms likely to be selected under the new technology, or the specifics of the higher-order moments of the multivariate distributions of characteristics across groups

[11]The case of the monline credit card company, CapitalOne, more efficiently using demographic information during the decade from 1994 to 2004 is particularly evocative in this context.

of interest. As a result, at least in theory, we note that the impact of new technology could be either positive or negative for any given g. To better understand these objects, therefore, we take our analysis to the data in Section 3. Before describing our empirical results, we consider the determination of mortgage interest rates in competitive equilibrium.

2.3 Equilibrium interest rates

Thus far, our discussion has concentrated on the case in which lenders evaluate default probabilities which are based purely on borrower characteristics x, and we have assumed that mortgage contract terms are exogenously specified. We now turn to thinking about the effects on outcomes of interest when we embed the lender's prediction problem in a setting in which mortgage terms are endogenously determined in competitive equilibrium.

We therefore consider a simple two-period model, in which each lender can offer mortgages to borrowers at date 0, the terms of which can be made contingent on borrower characteristics x. A mortgage contract consists of a loan L against a house worth V, and a promised repayment $(1+R) \times L$ at date 1, where R is the mortgage interest rate. For now, we assume that the loan size L and the loan-to-value ratio $LTV = L/V$ are pre-determined for each borrower. In reality, these parameters are often dictated, or at least confined to a narrow range, by local property prices and liquidity constraints faced by the borrower. We therefore think of L and LTV as elements of the borrowers' exogenous observable characteristics x. Thus, the mortgage rate R is the only variable that can be adjusted by lenders as part of a mortgage offer. In the Appendix, we discuss the extent to which this assumption biases our calculations.

In most optimizing models of borrower behavior, a change in interest rates affects the probability of default. Therefore, when allowing the interest rate to adjust to its equilibrium value, we now make explicit the dependence of the predicted probability $\hat{f}(x, R|\mathcal{M})$ of default on the interest rate, where $\mathcal{M}$ continues to denote a given statistical technology. In practice, as we will show below, the best prediction typically results from a model that includes R as

an explanatory variable for the probability of default.

The Net Present Value of the mortgage to a risk-neutral lender, at interest rate R, is

$$N(x,R) = L \times \left[\frac{1+R}{1+\rho} \left(1 - \lambda \hat{f}(x, R|\mathcal{M}) \right) - 1 \right] \tag{3}$$

In equation (3), lenders' net cost of capital between dates 0 and 1 is denoted by $\rho > 0$. In the event of default, the lender receives only a fraction $1 - \lambda$ of the outstanding payment at date 1, where λ denotes the proportional loss given default on the loan.

Note that $N(x, 0) < 0$ for all x; intuitively, a positive interest rate is required to allow lenders to break even. In general, the NPV need not be a monotonic function of R, since higher interest rates increase the yield on the mortgage, but a greater interest burden may generate a strong temptation to default or lead to adverse selection among borrowers, thus raising the probability of default.

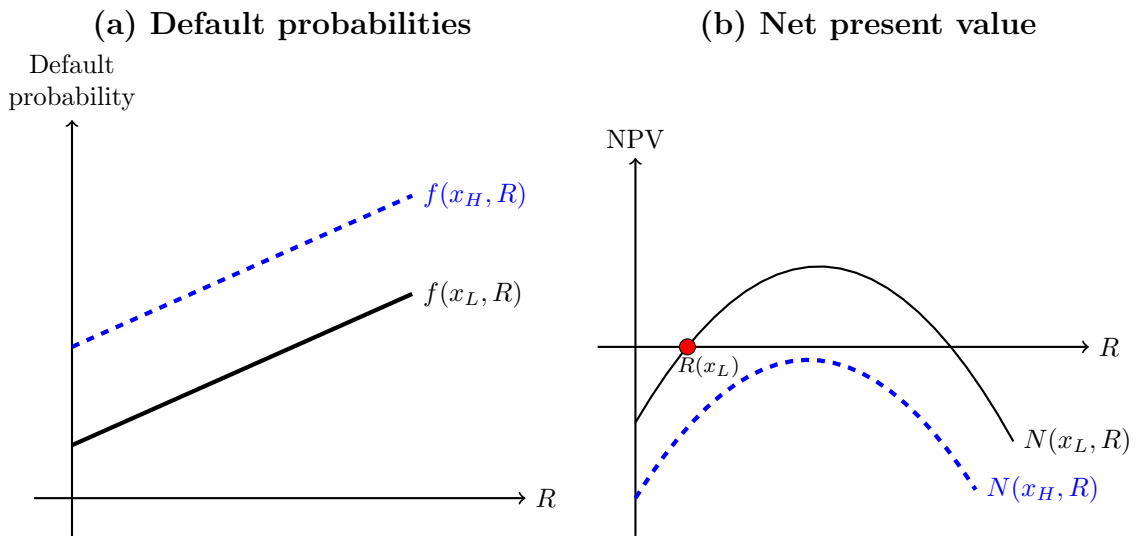
We assume that lenders are in Bertrand competition, that is, each lender simultaneously posts a schedule $R(x)$ of mortgage rates conditional on observable characteristics. We write $R(x) = \emptyset$ if a lender is unwilling to make any offer to x-borrowers. In this case, the lender rejects borrowers with characteristics x. The unique equilibrium outcome can be characterized as follows: All lenders reject borrowers with characteristics x such that $N(x, R) < 0$ for all R. For other borrowers, the equilibrium mortgage rate is the smallest rate that allows lenders to break even:

$$R(x) = \min \{R | N(x, R) = 0\} \tag{4}$$

Figure 3 illustrates the determination of equilibrium in this model using a simple example where predicted default probabilities $\hat{f}(x, R|\mathcal{M})$ are linear in interest rates R. The left panel shows predicted default rates for a borrower with high-risk characteristics x_H (dashed) and low-risk characteristics x_L (solid). The right panel shows the resulting NPV for the high-risk borrower, who is rejected in equilibrium, and the low-risk borrower, who is accepted and

receives interest rate $R(x_L)$. In the Appendix, we formally derive the above equilibrium conditions in a canonical model of lender and borrower behavior.

Figure 3: **Equilibrium determination.**



In equation (3), we assume that lenders base their decisions on a reduced-form prediction $\hat{f}(x, R|\mathcal{M})$ of default probabilities, given the available statistical technology. An alternative approach is to estimate a full structural model of borrower characteristics and behavior, and then to map these parameters into predicted default rates. We note here that in mortgage prepayment modeling, practitioners usually rely on reduced form models (see, e.g., Richard and Roll, 1989; Fabozzi, 2016). Similarly, empirical work on corporate defaults tends to suggest that a reduced form approach achieves better predictive outcomes than structural modeling (e.g., Bharath and Shumway, 2008; Campbell, Hilscher, and Szilagyi, 2008). We therefore posit that lenders take this approach.

Before moving to the empirical analysis, we note some potential identification problems in the lender's inference. Our calculation of the lender's expected NPV is valid if and only if $\hat{f}(x, R|\mathcal{M})$ that we estimate in reduced form is an *unbiased* predictor of the true likelihood of default once the mortgage is originated.

Two sample selection issues arise in this context. First, the mortgage is originated only if the borrower is willing to accept the contract with interest rate R. This gives rise to the possibility that the subset of borrowers who are willing to accept such offers have different default propensities from the population. While this issue is mitigated by the fact that in our dataset (and indeed in any such dataset), we estimate default propensities using borrowers who accepted contract offers, it is still the case that unobservable changes in the borrower population's propensity to accept offers will generate selection issues in our estimates.

Second, mortgages in past data are originated only if past lenders were willing to grant them. To the extent that past lenders had access to "soft information" beyond what is recorded, past data may therefore contain a select sample of borrowers that had favorable soft characteristics. Here, a mitigating factor is that we mainly focus in our empirical work on a period after the (low doc/no doc loan-intensive) lending boom preceding the financial crisis. Post crisis, soft information does not appear to play a large role in the US mortgage market, since mortgage underwriting operates on fairly tight criteria that are set by the government sponsored enterprises (GSEs) and the Federal Housing Administration (FHA) for all insured loans. Similarly, for jumbo loans that are held on balance sheet, banks usually have centralized criteria and automatic underwriting software for most loans. A more restrictive interpretation of our work could be that we shed light on how such centralized criteria might change with the introduction of machine learning and other sophisticated statistical technologies, and how this development would affect outcomes for different groups of borrowers.

We discuss these issues in more detail when we present the US mortgage market data that we study, as well as when we discuss our empirical results.

3 US Mortgage Data

We use high-quality administrative data on the US mortgage market to study these issues using real data. In particular, we primarily rely on a merge of two loan-level datasets: (i) data collected under the Home Mortgage Disclosure Act (HMDA), and (ii) the McDash™ mortgage servicing dataset from Black Knight.

HMDA data has traditionally been the primary dataset used to study unequal access to mortgage finance by loan applicants of different races, ethnicities, or genders; indeed "identifying possible discriminatory lending patterns" was one of the main purposes in establishing HMDA in 1975.[12] HMDA reporting is required of all lenders above a certain size threshold that are active in metropolitan areas, and the HMDA data are thought to cover 90% or more of all first-lien mortgage originations in the US (e.g., National Mortgage Database, 2017; Dell'Ariccia, Igan, and Laeven, 2012). These data also contain information on acceptances and rejections for loan applications, and are therefore useful to gauge how rejection rates might vary across different groups of borrowers. That having been said, the data lack a number of key pieces of information that are required for studying unequal access to or differential costs of credit across different groups. For instance, credit score (FICO), loan-to-value ratio (LTV), or the term of a loan are all not collected under HMDA; information on the cost of a loan is also very limited (as it is only reported for "high cost" loans).[13] HMDA also only observes loans at origination, and does not allow one to know whether a borrower ultimately defaulted on an originated loan.

The McDash™ dataset (owned and licensed by Black Knight) contains much more information on the contract and borrower characteristics of *originated* loans, including mortgage interest rates. The dataset follows these loans over time, and also contains a monthly indicator of a loan's delinquency status; it is thus one of the primary datasets that researchers have

[12] See https://www.ffiec.gov/hmda/history.htm.

[13] Bhutta and Ringo (2014) and Bayer, Ferreira, and Ross (2017) merge HMDA data with information from credit reports and deeds records in their studies of racial and ethnic disparities in the incidence of high-cost mortgages. Starting with the 2018 reporting year, additional information will be collected under HMDA; see http://files.consumerfinance.gov/f/201510_cfpb_hmda-summary-of-reportable-data.pdf for details.

used to study mortgage default (e.g., Elul, Souleles, Chomsisengphet, Glennon, and Hunt, 2010; Foote, Gerardi, Goette, and Willen, 2010; Ghent and Kudlyak, 2011). A matched dataset of HMDA and McDash loans is made centrally available to users within the Federal Reserve System. The match is done by origination date, origination amount, property ZIP code, lien type, loan purpose (purchase or refinance), loan type (e.g., conventional or FHA), and occupancy type. We only retain loans which can be uniquely matched between HMDA and McDash; we discuss how this affects our sample size below.

We estimate default probabilities using different statistical technologies on loans originated over the years 2009-2014, and in this version of the paper, study the effects of these different statistical technologies on equilibrium outcomes (credit allocation and rates) using the data for the year 2011.[14] We thus focus on loans originated after the end of the housing boom, which (unlike earlier vintages) did not experience severe declines in house prices. Indeed, most borrowers in our full sample experienced positive house price growth throughout the sample period. This means that serious delinquency of three or more missed payments (also known as "90-day delinquency"), which our analysis assumes is the outcome that lenders try to predict (and prevent), is likely driven to a large extent by idiosyncratic borrower shocks rather than macro shocks, which maps more closely to our theoretical analysis.

For loans originated in 2009-2014, our HMDA-McDash dataset corresponds to 43% of loans originated in HMDA. This fraction is driven by the coverage of McDash (70% of HMDA originations over this period) and the share of these McDash loans that can be uniquely matched to the HMDA loans (just over 60%).

For our analysis, we impose some additional sample restrictions. We only retain conventional (non-government) fixed-rate first-lien mortgages on single-family and condo units, with original loan term of 10, 15, 20, or 30 years. We furthermore only keep loans with original LTV between 20 and 100, a loan amount of US$ 1 million or less, and borrower income

[14] The inferences that we draw are very similar when we compute equilibrium using the data for the other years in the sample. These results are untabulated in the current version of the paper.

of US$ 500,000 or less. We also drop observations where the occupancy type is marked as unknown, and finally, we require that the loans reported in McDash have data beginning no less than 6 months after origination, which is the case for the majority (about 83%) of the loans in McDash originated over our sample period. This requirement that loans are not excessively "seasoned" before data reporting begins is an attempt to mitigate any selection bias associated with late reporting.

There are 48.1 million originated mortgages on 1-4 family properties in the 2009-2014 HMDA data. The matched HMDA-McDash sample imposing only the non-excessive-seasoning restriction contains 18.45 million loans, of which 72% are conventional loans. After imposing all of our remaining data filters on this sample, we end up with 10.3 million loans. For all of these loans, we observe whether they ever enter serious delinquency between origination and March 2017 – this occurs for 1.26% of these loans.

HMDA contains separate identifiers for race and ethnicity; we focus primarily on race, with one important exception. For white borrowers, we additionally distinguish between hispanic/latino white borrowers and non-hispanic white borrowers.[15] The number of borrowers in each group, along with descriptive statistics of key observable variables are shown in Table 1. The table shows that there are clear differences between the (higher) average and median FICO scores, income levels, and loan amounts for White non-hispanic and Asian borrowers relative to the Black and White hispanic borrowers. Moreover, the table shows that there are higher average default rates (and indeed interest rates and spreads at origination over average interest rates) for the Black and White hispanic borrowers. Heuristically, such differences in characteristics make minority populations look different from the "representative" borrower we discussed in the single-characteristic model of default probabilities in the theory section. Depending on the shape of the functions under the new machine learning technology, these

[15]The different race codes in HMDA are: 1) American Indian or Alaska Native; 2) Asian; 3) Black or African American; 4) Native Hawaiian or Other Pacific Islander; 5) White; 6) Information not provided by applicant in mail, Internet, or telephone application; 7) Not applicable. We combine 1) and 4) due to the low number of borrowers in each of these categories; we also combine 6) and 7) and refer to it as "unknown". Ethnicity codes are: Hispanic or Latino; Not Hispanic or Latino; Information not provided by applicant in mail, Internet, or telephone application; Not applicable. We only classify a borrower as hispanic in the first case, and only make the distinction for white borrowers.

differences will either be penalized or rewarded (in terms of estimated default probabilities) under the new technology relative to the old.

Table 1: **Descriptive Statistics, 2009-2014 Originations.**

Group		FICO	Income	LoanAmt	Rate (%)	SATO (%)	Default (%)
	Mean	738	122	280	4.23	-0.08	0.63
Asian	Median	773	105	252	4.25	-0.05	0.00
(N=631,764)	SD	141	75	153	0.69	0.45	7.93
	Mean	716	91	174	4.41	0.11	3.28
Black	Median	740	76	146	4.50	0.13	0.00
(N=269,178)	SD	125	61	111	0.68	0.48	17.82
	Mean	721	89	188	4.36	0.07	1.55
White hispanic	Median	753	73	160	4.38	0.10	0.00
(N=440,662)	SD	140	63	117	0.68	0.47	12.35
	Mean	735	110	209	4.33	-0.00	1.19
White non-hispanic	Median	771	91	178	4.38	0.02	0.00
(N=7,798,479)	SD	142	73	127	0.67	0.44	10.84
Native Am, Alaska,	Mean	720	97	204	4.38	0.04	1.72
Hawaii/Pac Isl	Median	757	81	175	4.38	0.04	0.00
(N=66,451)	SD	150	65	125	0.68	0.46	13.00
	Mean	731	119	230	4.37	-0.00	1.30
Unknown	Median	770	100	197	4.38	0.02	0.00
(N=1,066,078)	SD	150	78	143	0.67	0.44	11.32

Note: Income and loan amount are measured in thousands of USD. SATO stands for "spread at origination" and is defined as the difference between a loan's interest rate and the average interest rate of loans originated in the same calendar quarter. Data source: HMDA-McDash matched dataset of conventional fixed-rate mortgages originated over 2009-2014.

We also show descriptive statistics for the 2011 sample on which we compute equilibrium in Table 2. The table simply confirms that the patterns that are evident in the broader set of summary statistics are also evident for this subsample.

It is worth emphasizing one important point regarding our data and the US mortgage market more broadly. The vast majority of loans in our sample (over 90%) end up being securitized by the government-sponsored enterprises (GSEs) Fannie Mae or Freddie Mac, which insure investors in the resulting mortgage-backed securities against the credit risk on the loans. Furthermore, these firms provide lenders with underwriting criteria that dictate whether a loan is eligible for securitization, and (at least partly) influence the pricing of the

Table 2: **Descriptive Statistics, 2011 Originations.**

Group		FICO	Income	LoanAmt	Rate (%)	SATO (%)	Default (%)
	Mean	738	124	266	4.32	-0.09	0.62
Asian	Median	775	107	240	4.38	-0.05	0.00
(N=101,369)	SD	146	76	148	0.58	0.50	7.87
	Mean	720	93	167	4.58	0.13	3.47
Black	Median	743	77	139	4.62	0.20	0.00
(N=43,204)	SD	122	63	108	0.56	0.50	18.31
	Mean	724	91	179	4.54	0.09	1.56
White hispanic	Median	757	74	150	4.50	0.11	0.00
(N=68,567)	SD	139	65	113	0.56	0.49	12.38
	Mean	737	111	199	4.43	-0.00	1.18
White non-hispanic	Median	773	93	168	4.38	0.07	0.00
(N=1,289,050)	SD	144	75	125	0.56	0.48	10.80
Native Am, Alaska,	Mean	724	99	195	4.50	0.05	1.65
Hawaii/Pac Isl	Median	760	83	166	4.50	0.11	0.00
(N=9890)	SD	150	68	122	0.56	0.49	12.73
	Mean	736	120	221	4.46	0.00	1.27
Unknown	Median	772	100	185	4.50	0.07	0.00
(N=172,970)	SD	142	79	141	0.56	0.49	11.21

Note: Income and loan amount are measured in thousands of USD. SATO stands for "spread at origination" and is defined as the difference between a loan's interest rate and the average interest rate of loans originated in the same calendar quarter. Data source: HMDA-McDash matched dataset of conventional fixed-rate mortgages originated in 2011.

loans.[16] As a result, the lenders retain originated loans in portfolio (i.e., on balance sheet) and thus directly bear the risk of default for less than 10% of the loans in our sample. As this is the case that we consider in the model, a natural question that arises here is whether we should restrict our sample to those loans. We believe that the answer is no, since even a lender that only makes portfolio loans would wish to learn about default probabilities using as much data as they can acquire. In fact, the GSE underwriting criteria and pricing may be such that more loans are originated than in a purely private market, and this is helpful in the estimation of default probabilities (since those can only be reliably estimated for loan types actually available in the data).

We now turn to the next section, in which we estimate a set of different models to predict

[16]For instance, the GSEs charge so-called "loan-level price adjustments" that depend on borrower FICO score, LTV ratio, and some other loan characteristics.

default in the mortgage dataset. These models use increasingly sophisticated statistical technology, and allow us to evaluate the effects on the minority and majority groups in the borrower population.

4 Default Forecasting Using Different Statistical Technologies

Using the mortgage dataset, we approximate different mortgage lending technologies by using different prediction methods to estimate $\hat{f}(x, R)$, the probability of default.[17]

First, we implement two Logit models to approximate the "standard" prediction technology, estimating typical models used by both researchers and practitioners in the industry (e.g. Demyanyk and Van Hemert, 2011; Elul, Souleles, Chomsisengphet, Glennon, and Hunt, 2010). Second, to provide insights into how more sophisticated prediction technology will affect outcomes across groups, we estimate a tree-based model and augment it with a number of techniques that are commonly used in machine learning applications. More specifically, as we describe below, we implement a Random Forest model (Breiman, 2001), and utilize cross-validation and calibration to augment the performance of this model.

4.1 Logit Models

We begin by estimating two simple implementations of a standard Logit model. These models find widespread use in default forecasting applications, with a link function such that:

$$\log\left(\frac{g(x)}{1-g(x)}\right) = x'\beta. \tag{5}$$

[17] In our description of the estimation techniques, we maintain the notation in the previous sections, referring to observable characteristics as x, the loan interest rate as R, and the conditional probability of default as $f(x, R) = Pr(\text{Default}|x, R)$.

We estimate two models using this framework, by varying the way in which the covariates in x enter the model.

In the first model, x includes income, LTV, FICO score, the interest rate spread at origination (i.e., SATO, which is the interest rate on the mortgage relative to the average rate on all mortgages originated in the same calendar quarter), origination amount, and log of origination amount, with all terms entering linearly. Additionally, we include dummies for origination year, document type, occupancy type, product type, investor type, loan purpose, coapplicant status, and a flag for whether the mortgage is a jumbo. In addition, we include the term of the mortgage, and state fixed effects. We refer to this model simply as Logit.

In our second model, we allow for a more flexible use of the information in the covariates in x, in keeping with standard industry practice. In particular, we keep the same fixed effects as in the first model, but instead of income, LTV, and FICO entering the model linearly, we bin them to allow for the possibility of different coefficients to be estimated at different levels of these variables. In particular, for LTV, we use bins of size 5% from 20 to 100 percent, along with an indicator for LTV equal to 80, as this is a frequently chosen value in the data. For FICO, we use bins of 20 point width from 300 (the minimum) to 850 (the maximum). Finally, we bin income, using US $25,000 intervals from 0 to US $500,000. We refer to this model as the Non-linear Logit henceforth. Table 3 shows the list of variables employed in both models; the Random Forest model, which we describe next, uses the same set of variables as the Logit model.

4.2 Tree-Based Models

For our second approach to estimating $\hat{f}(x, R)$, we turn to machine learning models. The term is quite broad, but essentially refers to techniques to "learn" the function f that can best predict a generic outcome variable y using underlying attributes x. Within the broad area of machine learning, settings such as ours, in which the outcome variable is discrete (here, binary, as we are predicting default) are known as "classification" problems.

Table 3: **Variable List**

Logit	*Non-linear Logit*
Applicant Income (linear)	Applicant Income (25k bins, from 0-500k)
LTV Ratio (linear)	LTV Ratio (5-point bins, from 20 to 100%; separate dummy for LTV=80%)
FICO (linear)	FICO (20-point bins, from 300 to 850)
(with dummy variables for missing values)	
Common Covariates	
Spread at Origination (linear)	
Origination Amount (linear and log)	
Documentation Type (dummies for full/low/no/unknown documentation)	
Occupancy Type (dummies for vacation/investment property)	
Jumbo Loan (dummy)	
Coapplicant Present (dummy)	
Loan Purpose (dummies for purchase, refinance, home improvement)	
Loan Term (dummies for 10, 15, 20, 30 year terms)	
Funding Source (dummies for portfolio, Fannie Mae, Freddie Mac, other)	
Mortgage Insurance (dummy)	
State (dummies)	
Year of Origination (dummies)	

Note: Variables used in the models. Data source: HMDA-McDash matched dataset of conventional fixed-rate mortgages originated in 2011.

Several features differentiate machine learning approaches from more standard approaches to these sorts of problems. To list just one, the models tend to be non-parametric. Another difference is the use of computationally intensive approaches such as bootstrapping and cross-validation, which have experienced substantial growth in applied settings as computing power and the availability of large datasets have both increased.

While many statistical techniques and approaches can be characterized as machine learning, in our application here, we focus on a set of models that have been both successful and popular in prediction problems, which are based on the use of simple decision trees. In particular, we employ the Random Forest technique (Breiman, 2001).

In essence, the Random Forest is a non-parametric and non-linear estimator that flexibly

bins the covariates x in a manner that best predicts the outcome variable of interest. As this technique has been fairly widely used, we provide a brief overview of the technique here – for a more in-depth treatment of tree-based models applied in a default forecasting problem, see, for e.g., Khandani, Kim, and Lo (2010).

The Random Forest approach can best be understood in two parts. First, a simple decision tree is estimated by recursively splitting single covariates from a set x to best identify regions of default y. To fix ideas, assume that there is a single covariate under consideration, namely loan-to-value (LTV). To build a (primitive) tree, we would begin by searching for the single LTV value which best separates defaulters from non-defaulters, i.e., maximizes a criterion such as the Gini coefficient in the outcome variable between the two resulting bins on either side of the selected value, thus ensuring default prediction purity of each bin (or "leaf" of the tree). The process then proceeds recursively within each such selected leaf.

When applied to a broad set of covariates, the process allows for the possibility of bins in each covariate as in the Non-Linear Logit model described earlier, but rather than the lender pre-specifying the bin-ends, the process is fully data-driven as the algorithm learns the best function on a training subset of the dataset. Another important differentiating factor is that the process can identify *interactions* between covariates, i.e., bins that identify regions defined by multiple variables simultaneously, rather than restricting the covariates to enter additively into the link function, as is the case in the Non-Linear Logit model.

The simple decision tree model is intuitive, and fits the data extremely well in-sample, i.e., has low bias in the language of machine learning. However, it is typically quite bad at predicting out of sample, with extremely high variance on datasets that it has not been trained on, as a result of overfitting on the training sample. To address this issue, the second step in the Random Forest model is to implement (b)ootstrap (ag)gregation or "bagging" techniques. This approach attempts to reduce the variance of the out-of-sample prediction without introducing additional bias. It does so in two ways: first, rather than fit a single

decision tree, it fits many (500 in our application), with each tree fitted to a bootstrapped sample (i.e., sampling with replacement) of the original dataset. Second, at each point at which a new split on a covariate is required, the covariate in question must be from a randomly selected subset of covariates. The final step when applying the model is to take the modal prediction across all trees when applied to a new observation of covariates x.

The two approaches, i.e., bootstrapping the data and randomly selecting a subset of covariates at each split, effectively decorrelate the predictions of the individual trees, providing greater independence across predictions. This reduces the variance in the predictions without much increase in bias.

A final note on cross-validation is in order here. Several parameters must be chosen in the estimation of the Random Forest model, and can have an impact on the precision of the accuracy of the model. These include things like the maximum number of leaves, the minimum number of data points needed in a leaf in order to proceed with another split, and so on. In order to ensure the best possible fit, the common approach is to cross-validate the choice of parameters. This involves taking the training sample, and randomly splitting it into K-samples (in our case, we use $K = 3$). For each of the K samples, we fit the model (using a given set of tuning parameters) on the combined remaining samples ($K-1$ of them) to estimate the model, and then compare the out-of-sample predicted values of the model on the held-out sample. This is done K times, and the performance of those tuning parameters is averaged. This validation is done over a grid of potential tuning parameter values, and the set of parameters that maximize the out-of-sample fit in the cross-validation are chosen.

In our application, we cross-validate over the minimum number of data points need in a leaf and the minimum number of samples required on a leaf, and we use a 3-fold cross-validation.

4.2.1 Calibration

An important difference between the Random Forest model and the Logit models (both simple and non-linear) is that Logit estimation naturally produces an estimate of the probability of default given x. In contrast, the Random Forest model is geared towards providing a binary classification, i.e., given a set of covariates, the model will output either that the borrower is predicted to default, or to not default. As we require the default probability as an input into equation (3), we need to find a way to convert the output of the tree model into a probability that any given observation will be in default.

In the Random Forest model, one way to estimate this probability is to count the fraction of predicted defaults associated with the leaf into which a new borrower is classified. This fraction is generally estimated in the training dataset. However, this estimated probability tends to be very noisy, as leaves are optimized for purity, and there are often sparse observations in any given leaf.

A frequently used approach in machine learning is to "calibrate" these noisy estimated probabilities by fitting a monotonic function to smooth/transform them (see, for example, Niculescu-Mizil and Caruana, 2005). Common transformations include running a logistic regression on these probabilities to connect them to the true outcome in the training dataset, and searching across the space of monotonic functions to find the best fit function connecting the noisy estimates with the true values.[18]

This approach is known as calibration by isotonic regression, and we employ this approach to translate the predicted classifications into probability estimates. The Appendix provides more details, and discusses how this translation affects the raw estimates in the Random Forest model.

[18] In practice, the best results are obtained by estimating the calibration function on a second "calibration training set" which is separate from the training dataset on which the model is trained. The test dataset is then the full dataset less the two training datasets. See, for example, Niculescu-Mizil and Caruana (2005). We utilize this approach in our empirical application.

4.2.2 Estimation

As mentioned earlier, we first estimate our two sets of models on a subset of our full sample, which we refer to as the *training* set. We then evaluate the model on a *test* set, which the models have not seen before. In particular, we use 70% of the sample to estimate and train the model, and 30% to test.

The training sample is also split into two subcomponents. We utilize 70% of the training sample as a *model* sample, which we use to estimate the Logit, Nonlinear Logit, and Random Forest models. The remaining 30% of the training sample we dub the *calibration* sample, and use this subsample to estimate the isotonic regression to construct probabilities from the estimated Random Forest model.

Finally, we estimate the Random Forest model using Python's scikit-learn package, and the Logit models using Python's statsmodels package.

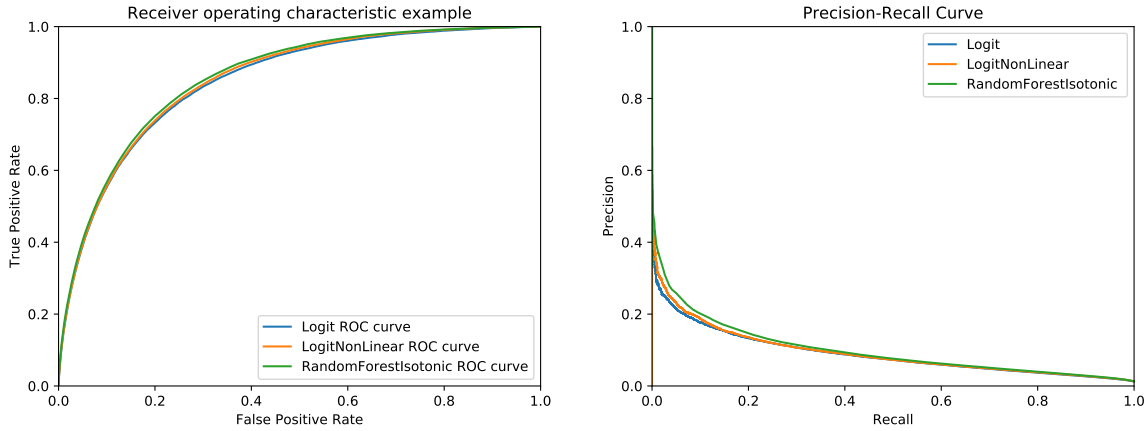
4.2.3 Evaluation

We evaluate the performance of the different models in several ways. To begin with, the Receiver Operating Characteristics (ROC) curve plots the true positive rate (TPR) against the false positive rate (FPR) as the probability threshold for declaring an observation to be a default varies (e.g., >50% is customary in Logit), but this number is a parameter that is varied in ROC analysis to see how it affects predictive accuracy and misclassification. A popular metric used to summarize the information in the ROC curve is the Area Under the Curve (AUC). Models for which AUC is higher are preferred, as these are models for which the ROC curve is closer to the northwest (higher TPR for any given level of FPR).

We also compute the *Precision* of each classifier, calculated as $P(y = 1|\hat{y} = 1)$, and the *Recall*, as $P(\hat{y} = 1|y = 1)$. These scores can each be thought of as average measures of the relative fit of the models, and as with the ROC, we draw Precision-Recall curves which plot

Precision against Recall for different probability thresholds.

Figure 4: **ROC and Precision-Recall Curves**



Panel A Panel B

Panel A of Figure 4 shows the ROC curves on the test dataset for the three models that we consider. The figure does not include race as a covariate, and shows that the Random Forest model performs better than both versions of the Logit model, in the sense that the TPR appears to be (weakly) greater for the Random Forest model than the others for every level of FPR. In Panel B, we see much stronger gains for the Random Forest model over the Logit models, which delivers higher levels of Precision for a given level of Recall.

Table 4 confirms that the AUC and Precision are indeed greater for the Random Forest model than for the other two, suggesting that the machine learning model more efficiently utilizes the information in the training dataset in order to generate more accurate predictions out of sample.

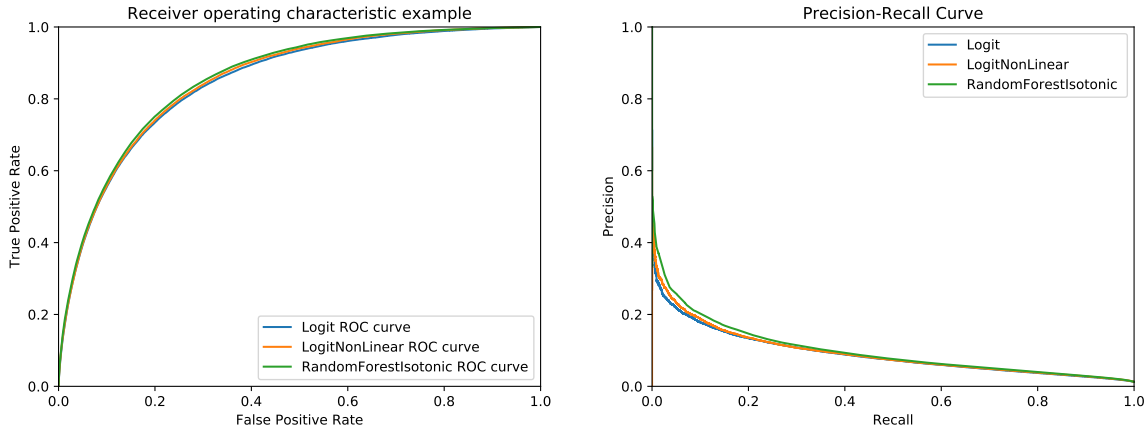
Table 4: **AUC and Precision for Different Statistical Technologies**

	ROC AUC Score		Precision Score	
Model	No Race	Race	No Race	Race
Logit	0.8477	0.8484	0.0896	0.0902
Logit Non-Linear	0.8518	0.8524	0.0923	0.0928
Random Forest Isotonic	0.8577	0.8577	0.0968	0.0968

Table 4 also shows that the inclusion of race has different effects on the three models.

Both of the Logit models benefit from the inclusion of this excluded variable, which allows these models to close the gap in predictive ability relative to the Random Forest model. While the change from 0.8518 to 0.8524 may seem small, this is roughly a 1% increase in average precision. The magnitude of this movement should be interpreted taking into account the fact that White non-hispanics make up the overwhelming majority of loans in our sample. In contrast, there is virtually no change in the predictive ability of the Random Forest model following the inclusion of race. Figure 5 shows the difference between the ROC curves from the three models once race is included in all three models.

Figure 5: **ROC and Precision-Recall Curves with Race Included.**



These facts about the change in the relative predictive ability of the models as a result of the inclusion of race are interesting, because the usual interpretation of the law is that differentiation based on excluded characteristics is prohibited. In keeping with the spirit of the law, the models assessing borrower risk should therefore be colorblind, which seems to be the case for the two Logit models (in the sense that race appears to add usefully to their performance). However, given that the Random Forest model seems relatively unaffected by the elimination of information about race, it suggests that this model is able to more efficiently triangulate the association between race and default using the remaining borrower characteristics.

We next turn to evaluating how probabilities of default, access to mortgage credit, and interest rate offers vary across excluded groups when we apply our model to the US mortgage

data.

5 Technology and Disparity in the Data

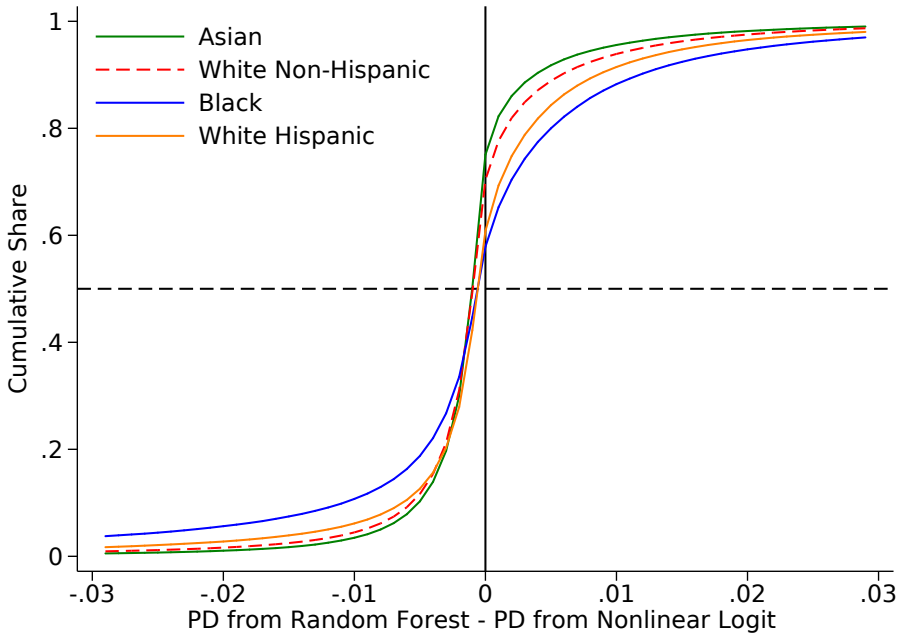
Following on from the models estimated on the data in the previous section, we begin by presenting some preliminary observations about how these models differ in their evaluation of the default risk of borrowers from different race groups.

Figure 6 provides insight into how the estimated probabilities of default in the data from the Random Forest model vary relative to those estimated using the Non-linear Logit model. Panel A of the figure shows the cumulative distribution function of the difference of the two probabilities, by race group. The difference corresponds to a relative benefit; borrowers for whom this difference is less than zero benefit from the new technology (in the sense of having a lower estimated default probability), and vice versa. In Panel B, we plot the log difference in default probabilities, which highlights the proportional benefit for each group.

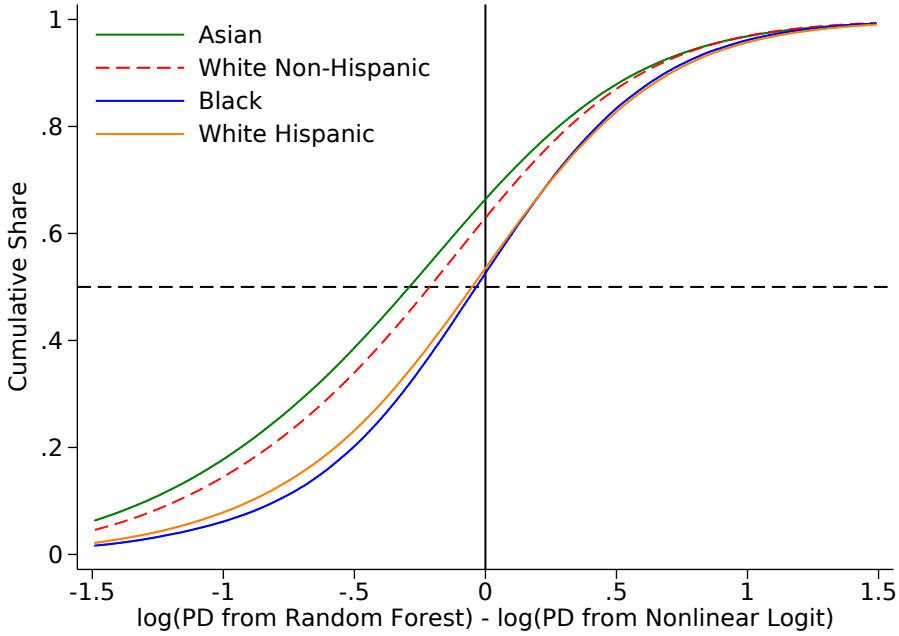
Panel B shows for all groups, there is a proportional reduction in default risk under the Random Forest model – the y-axis of the plot shows that the share of borrowers for whom the estimated probability of default falls under the new technology is above 50% for all groups (though this increase is smaller than some than for other groups). This benefit disproportionately accrues to White non-hispanic and Asian borrowers as the share of the borrowers in these groups that benefit from the new technology is over 60% (roughly 70% for Asian borrowers). In contrast, a roughly equal share of borrowers in the Black and White hispanic populations are on either side of zero, meaning that there are roughly equal fractions of winners and losers within these groups. Strikingly, we see that for both of these minority groups, the distribution of predicted default probabilities from the Random Forest model has larger variance than under the Logit model. We return to this finding later.

The figure provides useful insights into the questions that motivate our analysis, and

Figure 6: **Comparison of Predicted Default Probabilities.**



Panel A



Panel B

suggest that there may indeed be variations in the fraction of winners and losers across race groups engendered by technology. The figure shows that, at least in terms of estimated

probabilities of default, overall, White non-hispanic and Asian borrowers appear to benefit from the new technology in the sense that a larger fraction of them have lower estimated default probabilities using the new technology.

In the remainder of this section we continue to discuss how the issues highlighted in the theory section actually play out in the real US mortgage market data that we analyze, eventually moving to putting these estimated default probabilities into our equilibrium model. To build intuition, we begin in two-dimensional space, as in our theoretical presentation. We begin by plotting the distributions of FICO and income for different race groups to provide insights into variations in the distributions of $x|g$. We focus on the Black and White race groups in this initial analysis.

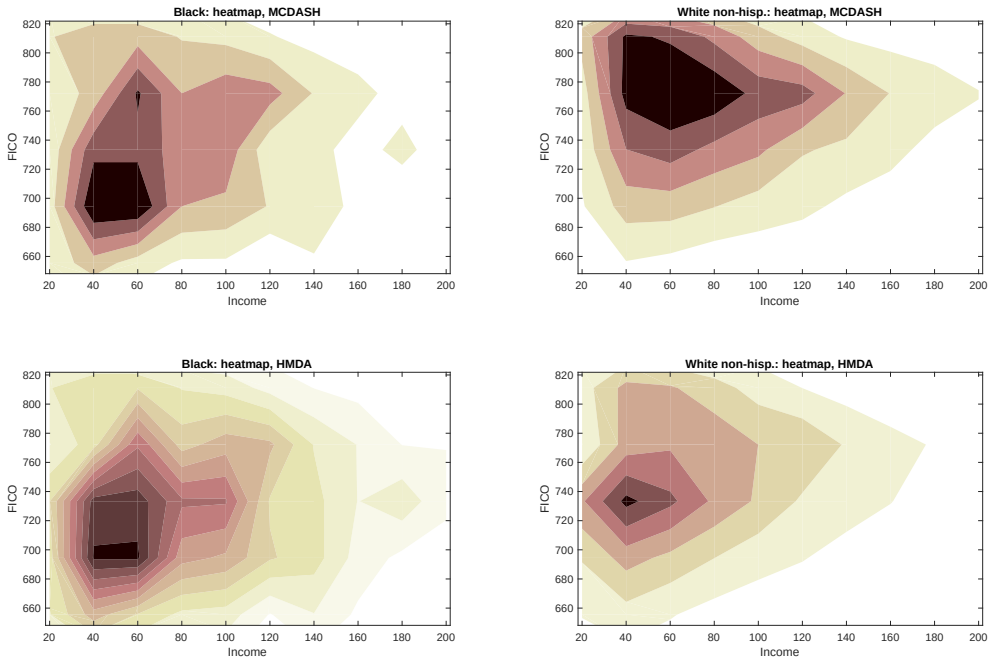
We then overlay these race-group conditional distributions on the exclusion regions and interest rate bands that arise from the use of different statistical technologies to estimate $f(x, R)$. In order to show this in FICO-income space, we must fix other borrower and contract characteristics, which also simultaneously vary with FICO and income. We therefore focus on a subgroup of loans in these plots.[19] It is worth emphasizing that these choices mean that the plots are not representative of the patterns in the entire data, and we therefore return to tabulating more aggregate measures at the end of the graphical presentation (for the particular contracts to which we restrict this analysis) to illustrate how different groups win and lose on both the extensive margin (exclusion) and the intensive margin (rates) as statistical technology improves.

[19] We restrict this graphical analysis to portfolio loans originated in California in 2011, with a loan amount of US$ 300,000, LTV 80%, and 30 year term, for the purpose of buying a home. The loans are issued to owner- occupants with full documentation, and bought by FNMA as the end investor. We drop all applicants with missing FICO or income. We also compute these probabilities of default under the assumption that the interest rate is 5% (comprised of a mortgage base rate of 4.4%, and SATO of 60 bp).

Group-Conditional Distributions of Borrower Characteristics

Figure 7 shows the empirical frequency of borrower FICO and income by racial group, for both Black (left panel) and White (right panel) borrowers in the data.[20]

Figure 7: **Distribution of Borrower Characteristics.**



The top two plots show the distribution of FICO and income computed using the HMDA-McDash merged dataset, and presented as a heatmap. The figure shows that the joint distribution of the two variables looks very different for Black and White borrowers. Clearly, the mean of both income and FICO are lower for Black borrowers. In addition, the variances of both income and FICO appear higher, and the two variables appear to be positively correlated for Black borrowers, whereas at high levels of FICO, income and FICO appear virtually uncorrelated for White borrowers. At least along the dimension of these two characteristics in the total set x, the distributions of $x|g$ look very different for $g \in \{\text{Black,White}\}$.

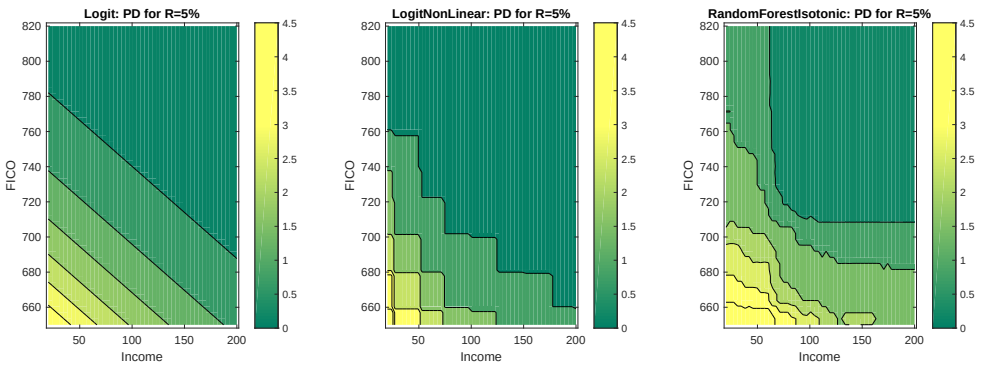
[20] We plot the distribution here for all borrowers with loan amount $L \in [250000, 350000]$ and $LTV \in [75, 85]$ to correspond with our analysis of rates and predicted default later.

While these plots are revealing about the race-group specific differences in the conditional distributions, they are drawn from a selected sample of mortgage borrowers, i.e., those who were accepted once they applied for a mortgage. Using this sample will likely understate exclusion since rejected applicants don't show up here. To address this issue, we construct a distribution of FICO and income using the entire HMDA dataset, which includes both accepted and rejected borrowers. This requires an imputation procedure, as FICO is not available for rejected borrowers in the HMDA data. We describe this procedure in the Appendix. Using this procedure, the bottom panel of Figure 7 shows the imputed joint distribution of FICO and income for Black and White borrowers. The plots show similar patterns as before, but the means of both variables in both distributions are now lower, and the variances are greater. We use these plots for the remainder of our analysis, given their greater representativeness for the broader distribution of mortgage applicants.

5.1 Probabilities of Default

We next show the probabilities of default associated with each of the statistical models for fixed exogenously specified interest rates. Figure 8 is the real-world (i.e., estimated using the actual mortgage data) equivalent of Figure 2, which drew hypothetical shapes for the level sets of the nonlinear technology.

Figure 8: **Predicted PD.**



The figure shows contours of the predicted probabilities of default as a function of borrower FICO score and income.[21] From each estimated statistical model (Logit, Non-linear Logit, and Random Forest), we calculate predicted default probabilities on a 50×50 grid of income-FICO pairs. We simply set $R = 5\%$ to draw this figure.[22] As expected, Logit has linear level sets, which, for the grid which we consider here, move in a fairly predictable fashion, with default probabilities decreasing linearly in both FICO and income, and the two serving as substitutes for one another in terms of their contribution to default risk.

Once we include narrowly-sized bins for the continuous variables in the Logit, the resulting Non-linear Logit specification is more flexible, and captures interesting non-linearities in estimated default probability as a function of income and FICO. For example, at a FICO of 760, there is an apparent downward jump in the estimated probability of default at roughly US$ 50,000 of income.

Finally, turning to the Random Forest model, it is clear that the level sets are nonlinear, though they appear to broadly follow the contours of the Non-linear Logit. The addition of high-dimensional interactions in the tree model clearly generates several differences in the estimated regions of high and low default probabilities. For example, under the new statistical technology, at a FICO of 760, there is now a region of low income (up to roughly US$ 75,000) where the Random Forest model predicts higher probabilities of default than the Non-linear Logit.

Earlier, we built intuition in the theoretical single-variable case that group outcomes depend on both the higher-order derivatives of the default-prediction function and the cross-group distribution of characteristics. We have already seen that the distributions of underlying characteristics vary by race group, and we have now seen that machine learning technologies have highly nonlinear level sets. To see which groups win and lose under the new technology, we must go further, and superimpose these two graphs on one another.

[21] As mentioned in a previous footnote, it is worth reiterating that these are plotted for a particular set of loan characteristics.

[22] The empirical mean of R is 4.62, with a standard deviation of 0.42.

Before doing so, we need to estimate intensive margin effects, i.e., rates, as well as extensive margin effects, i.e., the possibility of exclusion. We therefore turn to describing how we compute equilibrium in the data.

Computing Equilibrium

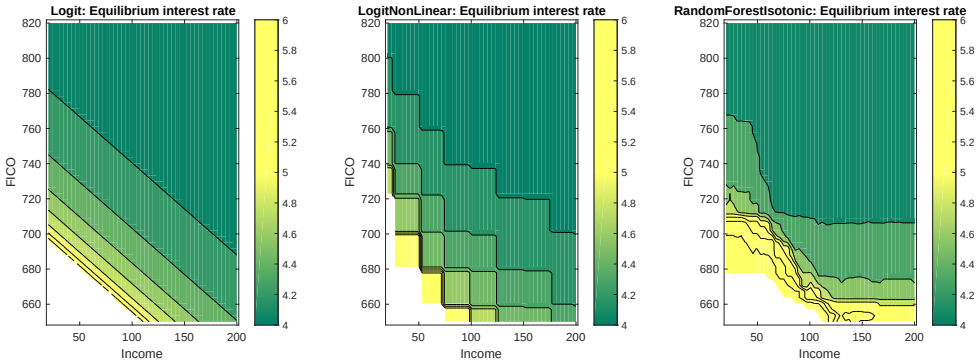
We now turn to computing equilibria in the data associated with the different statistical technologies $\hat{f}(x, R|\mathcal{M})$, using equation (3). In order to do so, we need assumptions on ρ, the cost of capital, and λ, the loss given default.

We set these parameters with reference to market yields and related literature. In particular, mortgage-backed security (MBS) yields in 2011 plus the primary-secondary spread (servicing costs plus guarantee fee) are approximately 4%, which we set as our value of ρ. Moreover, Andersson and Mayock (2014) estimate mean gross loss severity of roughly 45%, without accounting for discounting or workout costs. We therefore set loss given default, $\lambda = 50\%$.

Armed with these assumptions, we generate Figure 9, which shows equilibrium outcomes as a function of borrower FICO score and income. In this figure, we restrict the sample to Portfolio and GSE loans, 30-year purchase mortgages, owner-occupants and full documentation, and drop all applicants with missing FICO or income. From each estimated model, we compute equilibrium acceptance decisions and interest rates for each borrower in this subsample, allowing for interest rates on a 10-point grid in the range $R \in [1.5\%, 6\%]$, and obtain market-clearing rates by linear interpolation. We also compute equilibrium rejection decisions, and plot the contours of equilibrium interest rates R. In the figure, white space outside the contours denotes the rejection region.

The figure shows that there are significant differences between the rates generated by the three models, as well as the sizes of the areas of exclusion from the mortgage market. From this graphical analysis, it appears as if a) the size of the exclusion region is similarly sized for

Figure 9: **Equilibrium Interest Rates.**

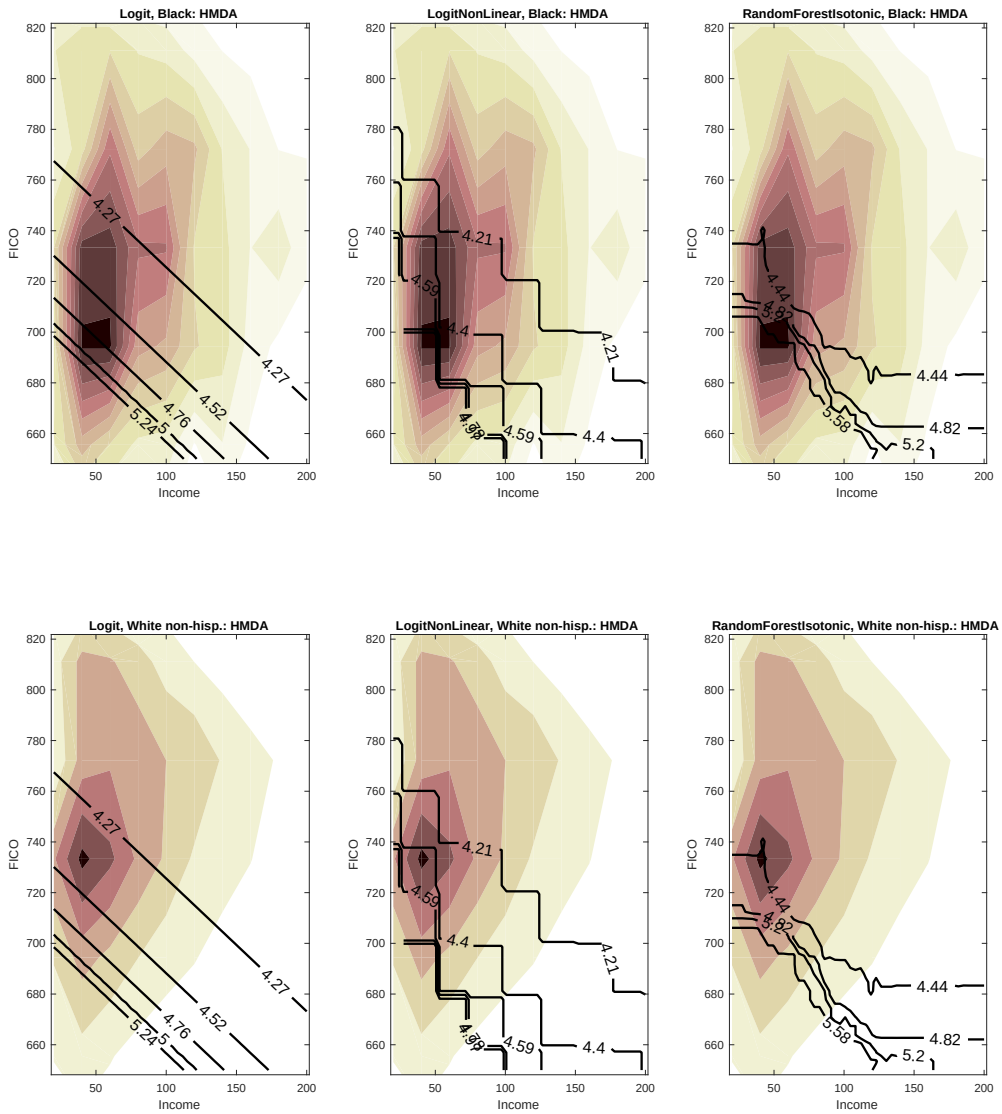


the machine learning model and the Logit model, and that b) the spread of the rates offered in the machine learning model is greater than that in the other two models, especially at low levels of FICO.

We explore this issue further in the next section, but we first proceed with the graphical analysis by overlaying the race-group-specific FICO-income joint distributions on these plots. Figure 10 does this for the White non-hispanic as well as Black borrowers in the population, and shows that there are significant differences between the treatment of these borrowers across the three models. The Nonlinear Logit model appears to treat the majority of White borrowers in this particular grid more favorably than the Logit model, though the Random Forest model appears to penalize this particular group of borrowers with higher average rates.

An interesting contrast is offered by overlaying the FICO-income joint distribution of Black borrowers on to the equilibrium rates and exclusion regions associated with the different underlying statistical technologies. The bottom panels of Figure 10 show that this joint distribution shifts both down and to the left relative to that of White borrowers, showing that there is significantly more exclusion for the subset of borrowers whose mortgages we consider in this set of plots, across all models. On average, the rates also appear to be higher for these borrowers, conditional on obtaining credit, under the machine learning model.

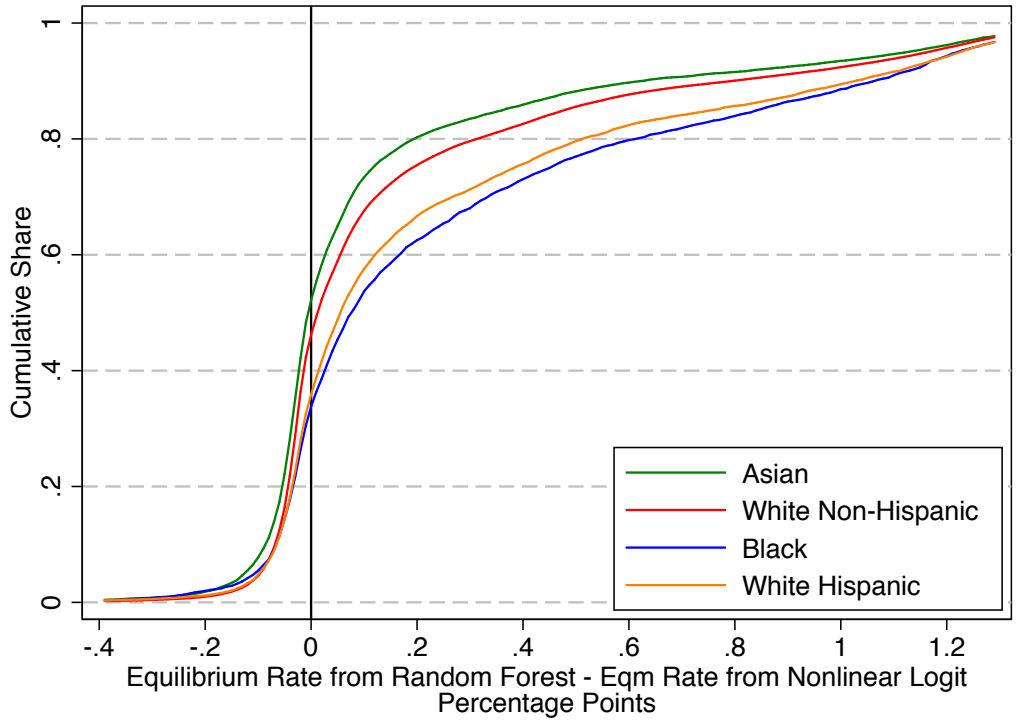
Figure 10: **Equilibrium Interest Rates and Distribution of Characteristics.**



While these patterns are revealing, they are plotted in FICO-income space holding constant a particular set of contract and borrower characteristics. To better understand the effect of the machine learning technology on offered mortgage interest rates, Figure 11 plots the difference of offered rates under the Random Forest model and those under the Nonlinear

Logit model, for those borrowers that are approved for a loan under both technologies.

Figure 11: **Comparison of Equilibrium Interest Rates.**



As before, the plot shows the cumulative distribution function of this difference by race group. Borrowers for whom this difference is negative benefit (in the sense of having a lower equilibrium rate) from the introduction of the new machine learning technology, and vice versa. Once again, the machine learning model appears to generate disparate impacts on different race groups. A larger fraction of White and especially Asian borrowers appear to benefit from the introduction of the technology, being offered lower rates under the new technology, while the reverse is true for the Black and hispanic borrowers. We also see that there is substantial skew in the rate difference: while few borrowers see a decrease of more than 10 basis points, up to a quarter of borrowers (depending on the group) see increases of 50 basis points or more.

To more rigorously assess the cross-group effects on both intensive and extensive margins, we next propose a simple approach to computing the disparate impacts of different

technologies.

5.2 A Simple Measure of Disparity

To make further progress, we need to go beyond analysis in the two-dimensional space. We therefore turn to Table 5, which evaluates the entire output of the models in the out-of-sample test dataset, once equilibrium outcomes have been computed.

Table 5: **Cross-Group Disparity.**

Nonlinear Logit				
	Predicted PD	Reject proportion	Av. interest	Frequency
Black	0.031	0.288	4.329	0.025
Native Am, Alaska, Hawaii	0.012	0.125	4.256	0.006
Unknown	0.009	0.075	4.212	0.09
White hisp	0.017	0.192	4.289	0.049
White non-hisp	0.01	0.086	4.222	0.761
Asian	0.007	0.063	4.211	0.069
Population	0.01	0.094	4.226	1.0
Cross-group st.dev.	0.002	0.016	0.009	
Random Forest				
	Predicted PD	Reject proportion	Av. interest	Frequency
Black	0.042	0.224	4.726	0.025
Native Am, Alaska, Hawaii	0.012	0.093	4.538	0.006
Unknown	0.009	0.046	4.394	0.09
White hisp	0.017	0.125	4.65	0.049
White non-hisp	0.011	0.062	4.45	0.761
Asian	0.008	0.041	4.384	0.069
Population	0.011	0.066	4.456	1.0
Cross-group st.dev.	0.002	0.012	0.027	

The first column of the table shows the mean predicted default probability by race group, evaluated at the empirically observed interest rate. The second and third columns show mean equilibrium rejection rates and equilibrium rates conditional on acceptance. The fourth column shows population frequencies of each racial group. The first six rows of the table show these statistics for each of the racial groups in the data, and the seventh, averaged

across the entire population. The panels show these statistics for each of the underlying statistical technologies, concluding with the Random Forest model.

In the final row of each panel, we compute a simple measure of cross-group disparity δ_τ under each technology τ. We denote the per-group mean of the desired measure by $\gamma_{g,\tau}$ (e.g., rejection rate, probability of default, or interest rate) under each technology τ. We then denote the measure for the entire population by $\bar{\gamma}_\tau$ under each technology τ. Finally, let ϕ_g be the frequency of each group in the population. Then:

$$\delta_\tau = \sqrt{\sum_g \phi_g \left(\gamma_{g,\tau} - \bar{\gamma}_\tau\right)^2} \tag{6}$$

The measure essentially computes the cross-group standard deviation of outcome variables, weighted by the groups' incidence in the population.

The table shows that δ_τ varies interestingly across technologies τ. While the Nonlinear Logit and the Random Forest model barely differ according to this measure when it comes to predicted default probabilities, it does seem that the Random Forest model has a lower δ_τ for exclusion. This is also accompanied by a lower average rejection rate for this model across all groups (i.e., an average rejection rate of 6.6% under the Random Forest model as opposed to 9.4% under Nonlinear Logit. Perhaps intuitively, the superior technology is better at screening, and is therefore more inclusive on average, and inclusive in a manner that cuts across race groups. However, the magnitude of these differences across models are relatively small.

The more substantial difference arises along the intensive margin. While it is true that under the machine learning technology more borrowers are allocated credit, the equilibrium rate is roughly 20 basis points higher for borrowers under this technology on average (4.456% vs. 4.226% under Nonlinear Logit). What's more, the disparity of rates across groups is substantially higher under the new technology. The point estimate of $\delta_\tau = 0.027$ is three times higher than the comparable point estimate for the Nonlinear Logit model. This reflects

the differential changes in the average rate across groups. For White borrowers, the average rate under the machine learning technology rises by a little under 20 basis points, but for Black borrowers, the comparable rise is more than double this number, at 40 basis points.

Overall the picture that Table 5 paints is an interesting one. The Random Forest model is a more accurate predictor of defaults, and generates higher acceptance rates and interest rates on average. However, it penalizes some minority race groups significantly more than the previous technology in the process, by giving them significantly higher interest rates.

Table 6: **Within-Group Disparity.**

Logit		
	St. dev. of predicted PD	St.dev. of interest rate
Black	0.027	0.271
Native Am, Alaska, Hawaii	0.012	0.255
Unknown	0.012	0.234
White hisp	0.014	0.264
White non-hisp	0.01	0.243
Asian	0.007	0.234
Population	0.011	0.243
Nonlinear Logit		
	St. dev. of predicted PD	St.dev. of interest rate
Black	0.037	0.275
Native Am, Alaska, Hawaii	0.018	0.242
Unknown	0.015	0.216
White hisp	0.021	0.266
White non-hisp	0.014	0.22
Asian	0.01	0.208
Population	0.015	0.223
Random Forest		
	St. dev. of predicted PD	St.dev. of interest rate
Black	0.041	0.677
Native Am, Alaska, Hawaii	0.014	0.614
Unknown	0.013	0.545
White hisp	0.016	0.664
White non-hisp	0.015	0.571
Asian	0.011	0.538
Population	0.016	0.576

Table 6 shows the *within* group dispersion of predicted default probabilities and rates associated with the different statistical technologies. The table shows that the predicted probabilities of default have higher dispersion within the group of Black borrowers under the Random Forest model, while there is hardly any increase in this dispersion for White non-hispanic borrowers. This increased dispersion also shows up in rates, but for both Black and White borrowers (though the dispersion in rates is higher for Black borrowers). Overall, these patterns in within group dispersion suggest that the Random Forest model screens within minority groups more efficiently than the Nonlinear Logit model, leading to changes in both exclusion and rate patterns associated with the new technology.

6 Conclusion

In this paper, we find that changes in statistical technology used to identify creditworthiness can generate significant disparity in credit outcomes across different categories of borrowers. We present simple theoretical frameworks to provide insights about the sources of these changes in outcomes, and verify that the issue manifests itself in US mortgage data.

Importantly, the disparity manifests itself across "restricted" categories such as race. Even though the statistical technologies do not use information about race group membership during default prediction, we find that machine learning models are more effectively able to triangulate the information connecting default propensity with these memberships using legitimately included variables.

While it is clear that there is an efficiency gain arising from the improved use of underlying information by the new technology, our work highlights that there are, inevitably, winners and losers from this change in technology. We also find that minority groups appear to lose, at least in terms of equilibrium rates, from the change in technology in the specific setting of the US mortgage market.

We propose in future versions of this paper to attempt to quantify the tradeoffs between lending efficiency, inclusion in credit markets, and rates conditional on inclusion arising with each underlying statistical technology. In so doing, we hope to provide a set of tools that will be useful to analyze the benefits and costs accruing to different groups in the economy of the inevitable use of machine learning and artificial intelligence.

7 Appendix

7.1 Proof of Lemma 1

We write $\mathcal{L}^2$ for the space of random variables z such that $E[z^2] < \infty$. Assume that the true default probability $f(x, R) \in \mathcal{L}^2$. On $\mathcal{L}^2$ we define the inner product $< x, y >= E[xy]$. Let $\hat{f}_j$ denote the projection of f onto a closed subspace $\mathcal{M}_j \subset \mathcal{L}^2$. The space of linear functions of x for given R, and the space of all functions of x, which we consider in the text, are both closed subspaces of $\mathcal{L}^2$. The projection $\hat{f}_j$ minimizes the mean square error $E[(f - \hat{f})^2]$, and the projection theorem (e.g. chapter 2 of Brockwell and Davis (2006)) implies that for any $m \in \mathcal{M}_j$,

$$E(m, f - \hat{f}_j) = 0$$

Letting $m \equiv 1$, we obtain $E[\hat{f}_j] = E[f]$. Now defining $u = \hat{f}_2 - \hat{f}_1$, we immediately get the required decomposition with $E[u] = E[f] - E[f] = 0$. We still need to show that $Cov(u, \hat{f}_1) = 0$. We have $u = \hat{f}_2 - f + f - \hat{f}_1$. Therefore,

$$Cov(u, \hat{f}_1) = Cov(\hat{f}_2 - f, \hat{f}_1) + Cov(f - \hat{f}_1, \hat{f}_1)$$

The first term is zero by an application of the projection theorem to $\hat{f}_2$, noting that $\hat{f}_1 \in \mathcal{M}_1 \subset \mathcal{M}_2$. The second term is zero by a direct application of the projection theorem to $\hat{f}_1$.

7.2 Proof of Lemma 2

The linear prediction can be written as $\hat{f}(x|\ell) = \alpha + \beta x$. For the nonlinear technology, let $\underline{\beta} = \min_{x \in [\underline{x}, \bar{x}]} \frac{\partial \hat{f}(x|\mathcal{M})}{\partial x}$ and $\bar{\beta} = \max_{x \in [\underline{x}, \bar{x}]} \frac{\partial \hat{f}(x|\mathcal{M})}{\partial x}$. It is easy to see that $\beta \in \left(\underline{\beta}, \bar{\beta}\right)$: If $\beta > \bar{\beta}$, for example, then it is possible to obtain an linear prediction that is everywhere closer to the nonlinear one, and therefore achieves lower mean-square error, by reducing β by a marginal unit.

By the intermediate value theorem, we can now find a representative borrower type $x = a$ such that the linear regression coefficient $\beta = \frac{\partial \hat{f}(a|\mathcal{M})}{\partial x}$. Then, we can write the linear prediction as a shifted first-order Taylor approximation of the nonlinear prediction around a:

$$\hat{f}(x|\ell) = \hat{f}(a|\mathcal{M}) + \frac{\partial \hat{f}(a|\mathcal{M})}{\partial x}(x - a) + B$$

where $B = \hat{f}(a|\ell) - \hat{f}(a|\mathcal{M})$. Now using a Taylor series expansion around a, we have

$$\hat{f}(x|\mathcal{M}) - \hat{f}(x|\ell) = \sum_{j=2}^{\infty} \frac{1}{j!} \frac{\partial^j \hat{f}(a|\mathcal{M})}{\partial x^j}(x - a)^j - B \tag{7}$$

and taking expectations conditional on group g yield the desired result.

7.3 Derivation of equilibrium prices

An industry of $N \geq 2$ mortgage lenders faces a population of potential. borrowers over two periods $t \in \{0, 1\}$. Each borrower has a vector $x \in \mathcal{X}$ of observable characteristics, which are known by lenders. In addition, each borrower has a vector $\theta \in \Theta$ of unobservable characteristics, which may be privately known to the borrower but are not seen by lenders.

Let $U(R|x, \theta)$ denote the lifetime utility enjoyed by a borrower with observable and unobservable characteristics $\{x, \theta\}$, when she accepts a mortgage with repayment R, and assume that this utility is strictly decreasing in R. Let $\bar{U}(x, \theta)$ denote the utility enjoyed without a mortgage. The borrower accepts a mortgage with rate R if and only if $U(R|x, \theta) \geq \bar{U}(x, \theta)$. If a borrower accepts a mortgage offer with rate R at date 0, she defaults at date 1 with probability $\tilde{p}(R, x, \theta)$. This general notation can represent most micro-founded models of optimal borrower behavior. Note, in particular, that $U(R|x, \theta)$ incorporates the potential value to the borrower of the option to default on her mortgage. Each lender can offer mortgage rates R to borrowers at date 0, the terms of which can be made contingent on x.

As in the text, we treat the loan amount L and V as exogenously given. The lender's cost

of capital is ρ and the fraction of the outstanding repayment recovered in default is $1 - \lambda$.

Lenders' profits

From the perspective of lenders, the default probability $\tilde{p}$ is a random variable, because it depends on unobservable characteristics. Lenders calculate their expected profits using the *ex ante* probability of default:

$$f(x, R) = E_\theta \left[\tilde{p}(R, x, \theta) | x, U(R|x, \theta) \geq \bar{U}(x, \theta) \right] \tag{8}$$

This conditional expectation allows for (adverse or advantageous) selection among borrowers. The lender's information set consists of the offered rate R and observable characteristics x. He further realizes that he will only end up doing business with borrowers who prefer this offer to their outside option. Therefore, he takes the conditional expectation of default probabilities across unobserved borrower types θ who would optimally choose to accept the offer.

This expression makes precise our discussion about identification. A statistical prediction $\hat{f}(x, R|\mathcal{M})$ can be used to calculate lenders' expected profits, given available technology, if and only if it is an unbiased predictor of the Bayesian description $f(x, R)$ of borrower behavior. The Net Present Value of offering a mortgage rate R to borrowers with observables x is then $N(x, R)$, defined as in Equation (3) in the paper.

We impose the following regularity condition:

Condition 1 *If* $\exists\ R = 0$ *such that* $N(x, R) = 0$*, then* $N(x, R)$ *is strictly increasing in* R *in a neighborhood of its smallest root* R_0*, defined as:*

$$R_0 = \inf\{R | N(x, R) = 0\} \tag{9}$$

Moreover, at any point of discontinuity in R*,* $N(x, R)$ *jumps downwards.*

This assumption rules out pathological cases. It is likely to hold under empirically realistic conditions, for two reasons. First, noting that $N(x, 0) < 0$, the NPV must cross zero from below at its smallest root R_0, so unless it is tangent (a knife-edge case), it must be strictly increasing. Second, an upward jump in $N(x, R)$ implies a downward jump in predicted default rates as the interest rate increases. This can be ruled out in most micro-founded models of borrower behavior, where default options are more likely to be exercised for high interest rates, and we consistently find that empirical default probabilities are increasing in interest rates.

Equilibrium

Lenders are in Bertrand competition: Each lender simultaneously posts a schedule $R(x)$ of mortgage rates conditional on observable characteristics. We write $R(x) = \emptyset$ if a lender is unwilling to make any offer to x-borrowers. In this case, the lender rejects borrowers with characteristics x. When borrowers are indifferent, they choose a lender at random according to a fair coin toss.

Lemma. If there is no rate $R \geq 0$ such that $N(x, R) \geq 0$, then the unique equilibrium is for all lenders to offer $R(x) = \emptyset$: If any lender accepted x-borrowers, they would make negative profits, and have a profitable deviation by rejecting instead. Otherwise, if there is an $R \geq 0$ such that $N(x, R) > 0$, then all x-borrowers obtain credit and the unique equilibrium rate for x-borrowers is $R(x) = R_0$, defined in analogy to Equation (4) as the smallest root of $N(x, R) = 0$.

Proof. We can establish this by contradiction. First, suppose that x-borrowers do not obtain credit. Then, letting R_1 denote any rate such that $N(R, x) > 0$, each lender has a profitable deviation to offering R_1, which yields strictly positive profits. Second, suppose that lenders offer $R' < R(x)$ in equilibrium. Then, the make a loss and have a profitable deviation by rejecting. Finally, suppose that lenders offer $R' > R(x)$. Then, if $N(x, R) > 0$, a lender can deviate by offering $R' - \epsilon$, poach the entire market, and strictly increase her

profits. Otherwise, if $N(x, R) \leq 0$, a lender can deviate by offering $R_0 + \epsilon$, again obtaining the entire market and strictly positive profits. Hence, the unique equilibrium rate is R_0 as required.

7.4 Discussion of endogenous contracting terms

We have assumed that loan size L and LTV are pre-determined. In this Appendix, we discuss whether this assumption biases our calculation of the proportion of borrowers accepted for credit, and of the average mortgage rate conditional on acceptance, across the population.

Suppose that lenders offer a menu: One interest rate $R(h, x)$ (or possibly rejection) for each possible contract $h = \{L, LTV\}$, given observable characteristics x.

Given a menu $R(h, x)$, let $p_h(h|x)$ be the proportion of x-borrowers whose preferred contract on the menu is h, conditional on accepting any of these offers at all (some borrowers may choose to remain without a mortgage in equilibrium). Let $p_x(x)$ be the population distribution of x.

In any equilibrium, the rate of credit across the population is

$$C = \int \int 1\{R(h, x) \neq \emptyset\} p_h(h|x) p_x(x) dh dx$$

and the average mortgage rate conditional on obtaining credit is

$$\bar{R} = C^{-1} \int \int 1\{R(h, x) \neq \emptyset\} R(h, x) p_h(h|x) p_x(x) dh dx$$

From the population of potential borrowers, we can obtain an estimate $\hat{p}_x(x)$ of the distribution of exogenous characteristics x. We also obtain an estimate $\hat{p}_h(h|x)$ of the conditional empirical distribution of contract characteristics given exogenous characteristics. We then

assume that this is an unbiased estimate of the choice function $p_h(h|x)$ specified above:

$$\hat{p}_h(h|x) = p_h(h|x) + \varepsilon$$

where ε is independent of borrower and contract characteristics. Under this condition, the average outcomes that we calculate in the paper continue to be an unbiased estimate of the integrals above, even when contract characteristics are chosen endogenously.

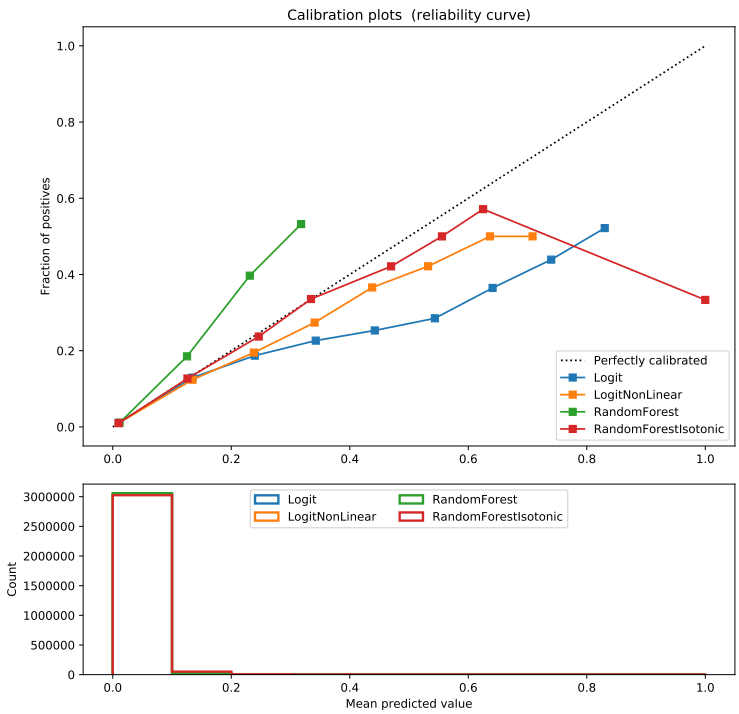
7.4.1 Isotonic regressions and calibration

Denote by y_i is the true outcome for a borrower i in the training dataset, and by l_i the ratio of predicted default to non-default observations associated with the leaf in the decision tree to which the borrower's characteristics have been classified. Then, the isotonic regression approach is to find $\hat{z}$ in the space of monotonic functions such that:

$$\hat{z} = \arg\min_{z} \sum (y_i - z(l_i))^2. \tag{10}$$

Figure 12 plots the number of defaults within each bin shown on the y-axis against the binned predictions from each of the models on the x-axis. A well-calibrated model would lie along the 45° line. The Non-Linear Logit model looks relatively well-calibrated, but in comparison, the Random Forest model (without the application of the isotonic regression model) and Simple Logit models look relatively poorly calibrated. This is because of the noisy measure of probability obtained from the leaf nodes which are optimized for purity. Following the isotonic regression, we see that the Random Forest model seems better calibrated, lying close to the 45° line, at least at lower predicted probabilities of default.

Figure 12: **Calibration Curve.**



7.5 Imputation procedure for FICO in HMDA data

We calculate the population frequency $p(FICO, Y|L, LTV)$, where Y is borrower income, and L is the loan amount. Let $\mathcal{A}$ be a dummy variable denoting acceptance for a mortgage. We can then write:

$$p(FICO, Y|L, LTV) = \sum_{\mathcal{A}\in\{0,1\}} p(\mathcal{A})p(FICO, Y|L, LTV, \mathcal{A})$$

We can calculate the weights conditional on acceptance, $p(FICO, Y|L, LTV, \mathcal{A} = 1)$, directly from the merged HMDA-McDash sample. We then obtain the frequency of acceptance $p(\mathcal{A} = 1)$ as the proportion of borrowers with action flags 1 (Loan originated) or 3 (Application approved but not accepted) in the HMDA sample. The frequency of rejection $p(\mathcal{A} = 0)$ is the proportion of borrowers with flag 3 (Application denied by financial institution). We normalize these frequencies so that $p(\mathcal{A} = 1) + p(\mathcal{A} = 0) = 1$.

We impute the weights conditional on rejection, $p(FICO, Y|L, LTV, \mathcal{A} = 0)$, since rejections are only observed in the HMDA sample, where FICO and LTV are not recorded. Our imputation is based on the following assumptions:

1. The conditional distribution of FICO among rejected borrowers is equivalent to the distribution of an adjusted FICO score, denoted $\hat{F}$, among accepted borrowers:

$$p(FICO, Y|L, LTV, \mathcal{A} = 1) = p(\hat{F}, Y|L, LTV, \mathcal{A} = 0)$$

2. Let m_Y be the ratio of median income of rejected to accepted borrowers, which is 0.756 in the HMDA sample. Then the adjusted FICO score $\hat{F}$ is

$$\hat{F} = (1 - Q_F) \times FICO + Q_F \times FICO \times m_Y$$

   where $Q_F \in (0, 1)$ is a parameter measuring the degree of adjustment. Our baseline figures are based on $Q_F = 0.3$.

3. The conditional distribution of FICO is independent of income Y conditional on L and LTV. Further, Y is independent of LTV conditional on the L. We can now write:

$$p(FICO, Y|L, LTV, \mathcal{A} = 1) = p(Y|L, \mathcal{A} = 1) \times p(\hat{F}, Y|L, LTV, \mathcal{A} = 0) \tag{11}$$

Given these assumptions, we obtain the imputed frequencies conditional on rejection according to equation (11), where we get the first factor $p(Y|L, \mathcal{A} = 1)$ from the HMDA (sub)sample of rejected borrowers, and the second factor $p(\hat{F}, Y|L, LTV, \mathcal{A} = 0)$ from the HMDA-McDash sample with adjusted FICO scores.

References

Andersson, F., and T. Mayock (2014): "Loss severities on residential real estate debt during the Great Recession," *Journal of Banking & Finance*, 46, 266 – 284.

Arrow, K. J. (1973): "The Theory of Discrimination," in *Discrimination in Labor Markets*, ed. by O. Ashenfelter, and A. Rees. Princeton University Press.

Athey, S., and G. W. Imbens (2017): "The State of Applied Econometrics: Causality and Policy Evaluation," *Journal of Economic Perspectives*, 31(2), 3–32.

Bayer, P., F. Ferreira, and S. L. Ross (2017): "What Drives Racial and Ethnic Differences in High-Cost Mortgages? The Role of High-Risk Lenders," *Review of Financial Studies*, forthcoming.

Becker, G. S. (1971): *The Economics of Discrimination*. University of Chicago Press.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, 28(2), 29–50.

Berkovec, J. A., G. B. Canner, S. A. Gabriel, and T. H. Hannan (1994): "Race, redlining, and residential mortgage loan performance," *The Journal of Real Estate Finance and Economics*, 9(3), 263–294.

——— (1998): "Discrimination, competition, and loan performance in FHA mortgage lending," *The Review of Economics and Statistics*, 80(2), 241–250.

Bharath, S. T., and T. Shumway (2008): "Forecasting Default with the Merton Distance to Default Model," *Review of Financial Studies*, 21(3), 1339–1369.

Bhutta, N., and D. R. Ringo (2014): "The 2013 Home Mortgage Disclosure Act Data," *Federal Reserve Bulletin*, 100(6).

Breiman, L. (2001): "Random forests," *Machine learning*, 45(1), 5–32.

Brockwell, P. J., and R. A. Davis (2006): *Time Series: Theory and Methods*. Springer.

Bundorf, M. K., J. Levin, and N. Mahoney (2012): "Pricing and Welfare in Health Plan Choice," *American Economic Review*, 102(7), 3214–48.

Campbell, J. Y., J. Hilscher, and J. Szilagyi (2008): "In Search of Distress Risk," *Journal of Finance*, 63(6), 2899–2939.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2017): "Double/Debiased/Neyman Machine Learning of Treatment Effects," *American Economic Review*, 107(5), 261–65.

Chetty, R., and A. Finkelstein (2013): "Social Insurance: Connecting Theory to Data," in *Handbook of Public Economics*, ed. by A. J. Auerbach, R. Chetty, M. Feldstein, and E. Saez, vol. 5 of *Handbook of Public Economics*, chap. 3, pp. 111 – 193. Elsevier.

Dell'Ariccia, G., D. Igan, and L. Laeven (2012): "Credit booms and lending standards: Evidence from the subprime mortgage market," *Journal of Money, Credit and Banking*, 44(2-3).

Demyanyk, Y., and O. Van Hemert (2011): "Understanding the Subprime Mortgage Crisis," *Review of Financial Studies*, 24(6), 1848–1880.

Einav, L., and A. Finkelstein (2011): "Selection in Insurance Markets: Theory and Empirics in Pictures," *Journal of Economic Perspectives*, 25(1), 115–38.

Elul, R., N. S. Souleles, S. Chomsisengphet, D. Glennon, and R. Hunt (2010): "What 'Triggers' Mortgage Default?," *American Economic Review*, 100(2), 490–494.

Fabozzi, F. J. (ed.) (2016): *The Handbook of Mortgage-Backed Securities*. Oxford University Press, 7th edn.

Fang, H., and A. Moro (2010): "Theories of Statistical Discrimination and Affirmative Action: A Survey," Working Paper 15860, National Bureau of Economic Research.

Foote, C. L., K. S. Gerardi, L. Goette, and P. S. Willen (2010): "Reducing Foreclosures: No Easy Answers," *NBER Macroeconomics Annual*, 24, 89–183.

Geruso, M. (2016): "Demand Heterogeneity in Insurance Markets: Implications for Equity and Efficiency," Working Paper 22440, National Bureau of Economic Research.

Ghent, A. C., R. Hernández-Murillo, and M. T. Owyang (2014): "Differences in subprime loan pricing across races and neighborhoods," *Regional Science and Urban Economics*, 48, 199–215.

Ghent, A. C., and M. Kudlyak (2011): "Recourse and Residential Mortgage Default: Evidence from US States," *Review of Financial Studies*, 24(9), 3139–3186.

Hardt, M., E. Price, and N. Srebro (2016): "Equality of Opportunity in Supervised Learning," *CoRR*, abs/1610.02413.

Ho, T. K. (1998): "The random subspace method for constructing decision forests," *IEEE transactions on pattern analysis and machine intelligence*, 20(8), 832–844.

Khandani, A. E., A. J. Kim, and A. W. Lo (2010): "Consumer credit-risk models via machine-learning algorithms," *Journal of Banking & Finance*, 34(11), 2767–2787.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2017): "Human Decisions and Machine Predictions," *Quarterly Journal of Economics*, forthcoming.

Kleinberg, J. M., S. Mullainathan, and M. Raghavan (2016): "Inherent Trade-Offs in the Fair Determination of Risk Scores," *CoRR*, abs/1609.05807.

Ladd, H. F. (1998): "Evidence on Discrimination in Mortgage Lending," *Journal of Economic Perspectives*, 12(2), 41–62.

Mullainathan, S., and J. Spiess (2017): "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31(2), 87–106.

Narayanan, A., and V. Shmatikov (2008): "Robust De-anonymization of Large Sparse Datasets," in *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, pp. 111–125. IEEE Computer Society.

National Mortgage Database (2017): "A Profile of 2013 Mortgage Borrowers: Statistics from the National Survey of Mortgage Originations," Technical Report 3.1, CFPB/FHFA, `https://s3.amazonaws.com/files.consumerfinance.gov/f/documents/201703_cfpb_NMDB-technical-report_3.1.pdf`.

Niculescu-Mizil, A., and R. Caruana (2005): "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632. ACM.

O'Neil, C. (2016): *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.

Phelps, E. S. (1972): "The Statistical Theory of Racism and Sexism," *American Economic Review*, 62(4), 659–661.

Pope, D. G., and J. R. Sydnor (2011): "Implementing Anti-Discrimination Policies in Statistical Profiling Models," *American Economic Journal: Economic Policy*, 3(3), 206–231.

Richard, S. F., and R. Roll (1989): "Prepayments on fixed-rate mortgage-backed securities," *Journal of Portfolio Management*, 15(3), 73–82.

Ross, S., and J. Yinger (2002): *The Color of Credit: Mortgage Discrimination, Research Methodology, and Fair-Lending Enforcement*. The MIT Press.

Sirignano, J., A. Sadhwani, and K. Giesecke (2017): "Deep Learning for Mortgage Risk," Discussion paper, Stanford University.

Varian, H. R. (2014): "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, 28(2), 3–28.