Predictably Unequal?
The Effects of Machine Learning on Credit Markets

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

This draft: March 2018

[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, Krisk Gerardi, Ralph Koijen, Karthik Muralidharan, Jonathan Roth, Johannes Stroebel, and Stijn van Nieuwerburgh for useful conversations and seminar participants at Imperial College Business School, NYU Stern, University of Rochester, Queen Mary University of London, the Office for Financial Research, and the Southern Finance Association 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, new predictive statistical methods and machine learning techniques have been rapidly adopted by businesses seeking efficiency gains in a broad range of industries.[2] The pace of adoption of these technologies has prompted concerns that society has not carefully evaluated the risks associated with their use, including the possibility that any efficiency gains may not be evenly distributed.[3] In this paper, we study the distributional consequences of the adoption of machine learning techniques in the important domain of household credit markets. We do so by developing simple theoretical frameworks to analyze these issues, and by using structural estimation to evaluate counterfactuals using a large administrative dataset of loans in the US mortgage market.

The essential insight of our paper is that a more sophisticated statistical technology (in the sense of reducing predictive mean squared error) will, by definition, produce predictions with greater variance. Put differently, improvements in predictive technology act as mean-preserving spreads for predicted outcomes—in our application, predicted default propensities on loans.[4] This means that there will always be some borrowers considered less risky by the new technology ("winners"), while other borrowers will be deemed riskier ("losers"), relative to their position in equilibrium under the pre-existing technology. The key question is then how these winners and losers are distributed across societally important categories such as race, age, income, or gender.

We attempt to provide clearer guidance to identify the specific groups most likely to win or lose from the change in technology. To do so, we first solve a simple model in closed form for a lender who uses a single exogenous variable (e.g., a borrower characteristic such as income) to predict default. We then provide graphical intuition to help assess distributional

[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 applying machine learning to credit risk modeling includes Khandani, Kim, and Lo (2010), and Sirignano, Sadhwani, and Giesecke (2017).

outcomes in the case of a lender who uses two exogenous input variables to predict default. In both of these cases, we find 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 point is to consider an economy endowed with a primitive prediction technology which simply 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 uses 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.[5] We therefore provide evidence on these issues by going to the data. We estimate a set of increasingly sophisticated statistical models, beginning with a simple logistic regression of default outcomes on borrower and loan characteristics, and culminating in a random forest machine learning model (Ho, 1998; Breiman, 2001). We use these models to predict default in a large dataset of close to 10 million US mortgages originated between 2009 and 2013.

[5]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.

Using these data, we find that changes in predicted default propensities across race and ethnic groups differ significantly as statistical technology improves. In particular, while a large fraction of borrowers belonging to the majority group (e.g., White non-Hispanic) "win," that is, experience lower estimated default propensities under the machine learning technology than the less sophisticated logit technology, these benefits do not accrue to the same degree to members of minority race and ethnic groups (e.g., Black and Hispanic borrowers).

We investigate this issue further, by comparing the performance of the naïve and sophisticated statistical models when race and ethnicity are included and withheld from the information set used to predict default. We find that the logistic regression models benefit more from the inclusion of this information in the sense that it improves their predictive accuracy, while the machine learning model is barely affected by the inclusion of race and ethnic identifiers. Moreover, the machine learning models are far better than the logistic models at predicting race using borrower information such as FICO score and income. These findings are interesting since the *spirit* of the law suggests that models assessing borrower credit risk should be colorblind.[6] While this is the case by construction for the less sophisticated models in our analysis, the omission of these variables from the machine learning models barely affects their performance, as they are able to "triangulate" the predictive information contained in race and ethnicity for default probabilities, and to use it in credit risk assessments. This is reminiscent of recent work in the computer science literature which shows that anonymizing data is ineffective if sufficiently granular data on characteristics about individual entities is available (e.g., Narayanan and Shmatikov, 2008).

Our analysis finds that predicted default propensities across race and ethnic groups experience different changes as technology improves from the simple logistic approach to the more sophisticated machine learning technology. We go on to evaluate how these changes might translate into actual outcomes, i.e., whether different groups of borrowers will be granted mortgages and the interest rates that they will be asked to pay. To do so, we embed these

[6] In practice, compliance with the letter of the law 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).

statistical models in a simple equilibrium model of credit provision in a competitive credit market. When evaluating counterfactual equilibrium outcomes and performing comparative statics with respect to underlying technologies, we face a number of obvious challenges to identification. These arise from the fact that the data that we use to estimate the default models were not randomly generated, but rather, a consequence of the interactions between borrowers and lenders who may have had access to additional information whilst making their decisions.

We confront these challenges in a number of ways. First, we focus on a loan origination period which is well after the financial crisis. Post-crisis, mortgage underwriting operates on fairly tight observable criteria that are set by the government-sponsored enterprises (GSEs) Fannie Mae and Freddie Mac, as well as the Federal Housing Administration (FHA), which jointly insure most loans. Second, we restrict our analysis to securitized mortgages backed by Fannie Mae and Freddie Mac, as they are less likely to suffer from selection by lenders on unobservable borrower characteristics; instead, lenders mainly focus on whether a borrower fulfills the underwriting criteria set by the GSEs.[7] And finally, we undertake a bias adjustment of our estimated sensitivities of default to changes in interest rates, by computing an adjustment factor based on credibly causal estimates of these sensitivities estimated by Fuster and Willen (2017).

We compute 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 another to evaluate comparative statics on outcomes across groups. We find that the machine learning model appears to provide a slightly larger number of borrowers access to credit, and marginally reduces disparity in acceptance rates (i.e., the extensive margin) across race and ethnic groups in the borrower population. However, the story is different on the intensive margin. Here, the cross-group disparity of

[7] In influential work, Keys, Mukherjee, Seru, and Vig (2010) argue that there are discontinuities in lender screening at FICO cutoffs that determine the ease of securitization, but only for low-documentation loans (where soft information is likely more important), not for full-documentation loans such as the ones we consider.

equilibrium rates increases significantly (by 23%) under the machine learning model relative to the less sophisticated logistic regression models. This is also accompanied by a substantial increase in within-group dispersion in equilibrium interest rates as technology improves—it rises significantly more for Black and Hispanic borrowers under the machine learning model than it does for White non-Hispanic borrowers, i.e., Black and Hispanic borrowers get very different rates from one another under the machine learning technology.

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 effects of the improved technology are the substantial rise in the dispersion of rates across race groups, as well as the significant rise in the dispersion of rates within the group of Black and Hispanic borrowers.

Our focus in this paper is on the distributional impacts of changes in technology rather than on explicit taste-based discrimination (Becker, 1971) or "redlining," which seeks to use geographical information to indirectly differentiate on the basis of excluded characteristics, and which is also explicitly prohibited.[8] However, similarly in spirit to this work, we also seek a clearer understanding of the sources of inequality in household financial markets.[9] Our work is also connected more broadly to theories of statistical discrimination,[10] though we do not model lenders as explicitly having access to racial and ethnic information when estimating borrowers' default propensities. In future versions of this draft, we intend to

[8] Bartlett, Morse, Stanton, and Wallace (2017) study empirically whether "FinTech" mortgage lenders in the US appear to discriminate more across racial groups. Buchak, Matvos, Piskorski, and Seru (2017) and Fuster, Plosser, Schnabl, and Vickery (2018) study other aspects of FinTech lending in the US mortgage market.

[9] 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), 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).

[10] See Fang and Moro (2010) for an excellent survey, and the classic references on the topic, including Phelps (1972) and Arrow (1973).

clarify the connection between our work and statistical discrimination models, as well as to provide greater insight into how to evaluate tradeoffs between efficiency and disparity 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 that we use in our work. Section 4 introduces the default forecasting models that we employ on these data. Sections 5 sets up our equilibrium model of credit provision under different technologies, and Section 6 discusses how changes in technology affect measures of disparity in the US mortgage data. Section 7 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 a vector of observable characteristics x. We start by assuming that the lender takes as given a mortgage contract (interest rate, loan-to-value ratio, etc.) when drawing inferences, and study how these inferences are affected by changes in the statistical technology that they are able to apply. In a later section, we allow interest rates to be determined in competitive equilibrium, and also consider how changes in technology affect equilibrium rates.

The lender wishes to to find a function $\hat{y} = \hat{P}(x) \in \mathcal{M}$ which maps the observable characteristics x into a predicted y. We represent the statistical technology that the lender can use to find this function as $\mathcal{M}$, which comprises a class of possible functions that can be chosen.[11] We say that a statistical technology $\mathcal{M}_2$ is *better than* $\mathcal{M}_1$ if it gives the lender a larger set of functional options, i.e., $\mathcal{M}_1 \subset \mathcal{M}_2$.

[11] For example, if linear regression technology is all that the lender has available, then $\mathcal{M}$ is the space of linear functions of x.

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:

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

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

Our first step is to consider the impact of improvements in technology on predictions, and find that such improvements necessarily leads to predictions that are more disperse:

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

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

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

Proof: See Appendix.

This result is intuitive: by definition, improvements in technology will yield predictions with a mean-square error that is less than or equal to the pre-existing predictions. These new 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 is very simple, but makes it clear that there will be both winners and losers when better technology becomes available in credit markets, motivating the distributional concerns at the heart of our analysis. 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.

However, Lemma 1 is not specific about the identities of those who gain and lose in credit markets when statistical technology improves. This is a complex problem, and so to build intuition, we analyze two simple special cases in the remainder of this section. These examples employ one- and two-dimensional borrower characteristics and continue to assume that contract characteristics are given. In both cases, we consider the potential impacts of introducing a more sophisticated statistical technology on subgroups (g) of borrowers in the broader population. In what follows, we characterize these subgroups by the conditional distributions of their characteristics, i.e., $x|g$.[12]

2.1 Case 1: One-Dimensional Borrower Characteristics

We assume here that lenders predict default as a function of a scalar x. We further 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 $\hat{P}(x|\mathcal{M}_2)$, 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:

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

where a is the value of the characteristic of a "representative" borrower such that $\frac{\partial^j \hat{P}(a|\mathcal{M}_2)}{\partial x^j} =$

[12] 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.

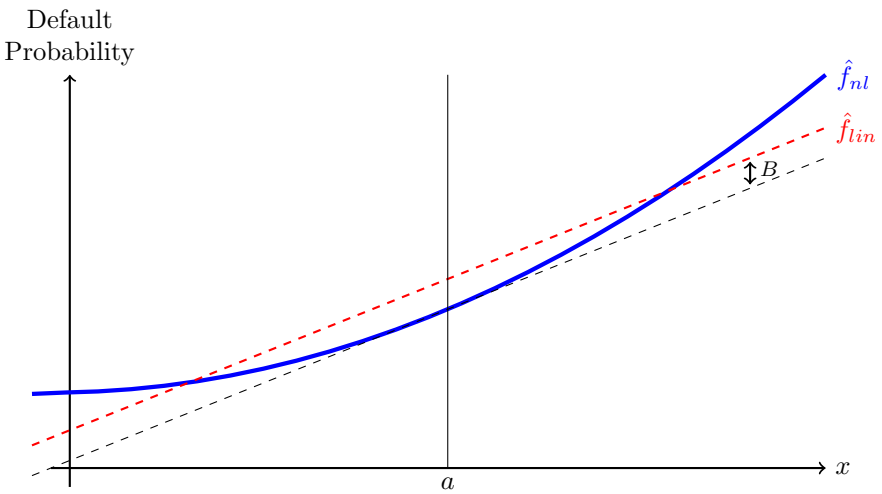
$\frac{\partial^j \hat{P}(a|\mathcal{M}_1)}{\partial x^j}$, and $B = \hat{P}(a|\mathcal{M}_1) - \hat{P}(a|\mathcal{M}_2)$ is a constant.

Proof: See Appendix.

Lemma 2 shows that in this case, 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 $\frac{\partial^j \hat{P}(a|\mathcal{M}_2)}{\partial x^j}$, evaluated at a.

Figure 1 illustrates what happens when the prediction using the new statistical technology, denoted $\hat{P}(x|\mathcal{M}_2) = \hat{P}_{quad}$, is a convex quadratic function of x. As in Lemma 2, the linear prediction $\hat{P}(x|\mathcal{M}_1) = \hat{P}_{lin}$ can be expressed as a shifted approximation of $\hat{P}_{quad}$ around the representative point $x = a$. In this case, the leading term in equation (2) indicates that a subgroup 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 far from the representative borrower's value.

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



In the special case when $\hat{P}(x|\mathcal{M}_2) = \hat{P}_{quad}$, borrowers belonging to minority subgroups of the population are likely to lose under the new technology. To see this more clearly, 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. Of course, this is a special case, and if the superior technology were concave rather than convex in x, this result would be reversed and minority subgroups would benefit under the new technology.

More generally, Lemma 2 implies that a subgroup g of borrowers is likely to lose under the superior statistical technology 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 $\hat{P}(x|\mathcal{M})$.[13]

2.2 Case 2: Two-Dimensional Borrower Characteristics

To develop further intuition, we now consider the case of two-dimensional borrower characteristics, i.e., $x = (x_1, x_2)$. For concreteness, let x_1 be the borrower's income, and x_2 her FICO credit score.

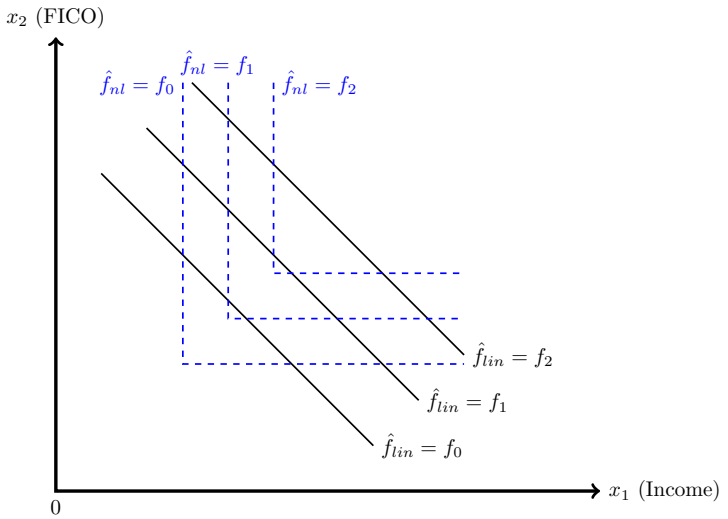
As an illustration, Panel (a) of Figure 2 plots the level sets of the predicted default probabilities $\hat{P}(x|\mathcal{M}_1) = \hat{P}_{lin}$ from a linear probability model, alongside predictions $\hat{P}(x|\mathcal{M}_2) = \hat{P}_{nl}$ from a superior, Nonlinear model which takes the Leontief shape, i.e., $\hat{P}_{nl} = \min\{ax_1, bx_2\}$. These choices of functional forms are in anticipation of our empirical analysis, where we consider mainly Logit models of default, which have linear or stepped level sets (depending on whether underlying characteristics enter linearly or in binned levels), and machine learning models based on decision trees, which tend to yield predicted default probabilities that can pick up more complex interactions of underlying characteristics. In this example, for both technologies, we assume that predicted default probabilities are decreasing in both FICO and income.

[13]For example, if the distribution of $x|g$ is right-skewed, and the third derivative of $\hat{P}(x|\mathcal{M})$ is positive, then the introduction of $\hat{P}(x|\mathcal{M})$ relative to the previously available technology will penalize the right tail of x, causing members of subgroup 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 $\hat{P}(x|\mathcal{M})$ 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.

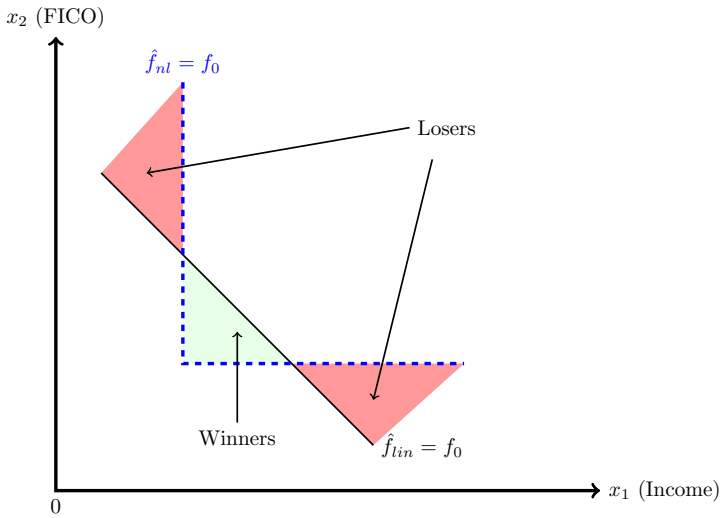
Panel (b) of Figure 2 focuses on comparing one level set across the two technologies, and shows those who will be predicted to have lower credit risks ("winners") and higher credit risks ("losers") upon the introduction of the new technology. The specific assumption that the new technology is Leontief means that income and FICO act as complements, where under the linear technology they acted as substitutes. Losers under the new technology are therefore borrowers who fall short on one of these criteria, while doing well on the other.

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

(a) Level sets of predicted PD



(b) Winners and losers from improved technology



2.3 Discussion

The main insights from our simple theoretical analysis are as follows. First, Lemma 1 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 specific differences between the new and old predictions.

It is worth re-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 true patterns that exist in the data. For example, consider the case in which the new technology allows a lender to more efficiently use 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 on average, though dispersion of outcomes within the group would rise as a result.[14]

As a result, while we have a better understanding of the underlying forces at work, uncovering the identities of the winners and losers will require moving to the data. In the next section, therefore, we discuss how predicted default probabilities estimated in the data vary with statistical technology, and concentrate on the distributional impacts of these technologies across race and ethnicity-based subgroups of the population.

Another shortcoming of our discussion thus far is that it has not touched upon the more realistic scenario of endogenously assigned contract characteristics, meaning that we cannot

[14]The case of the monline credit card company, CapitalOne, more efficiently using demographic information during the decade from 1994 to 2004 is evocative in this context. See, for example, Wheatley, Malcolm (November 1, 2001). "Capital One Builds Entire Business on Savvy Use of IT," CIO magazine.

at this stage predict how changing probabilities of default translate into changes in interest rates or exclusion. We return to this issue in some detail after the next section.

3 US Mortgage Data

To study how these issues play out in reality, we use high-quality administrative data on the US mortgage market, which results from merging two loan-level datasets: (i) data collected under the Home Mortgage Disclosure Act (HMDA), and (ii) the McDash^{TM} mortgage servicing dataset which is owned and licensed by 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.[15] 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.

However, HMDA lacks a number of key pieces of information that we need for our analysis. Loans in this dataset are only observed at origination, so it is impossible to know whether a borrower in the HMDA dataset ultimately defaulted on an originated loan. Moreover, a number of borrower characteristics useful for predicting default are also missing from the HMDA data, such as the credit score (FICO), loan-to-value ratio (LTV), the term of the issued loan, and information on the cost of a loan (this is only reported for "high cost" loans).[16]

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

[16] 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

The McDash^{TM} dataset from Black Knight contains much more information on the contract and borrower characteristics of loans, including mortgage interest rates. Of course, these data are only available for originated loans, which the dataset follows over time. The dataset also contains a monthly indicator of a loan's delinquency status, which has made it one of the primary datasets that researchers have 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 zipcode, lien type, loan purpose (i.e., 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, and we discuss how this affects our sample size below.

Our entire dataset extends from 2009-2016, and we use these data to estimate three-year probabilities of delinquency (i.e., three or more missed payments, also known as "90-day delinquency") on all loans originated between 2009 and 2013.[17] 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 data experienced positive house price growth throughout the sample period. This means that delinquency is likely driven to a large extent by idiosyncratic borrower shocks rather than macro shocks, mapping more closely to our theoretical discussion.

For the origination vintages from 2009-2013, our HMDA-McDash dataset corresponds to 45% of all loans in HMDA. This fraction is driven by the coverage of McDash (corresponding to 73% 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 issued) fixed-

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.

[17] We do so in order to ensure that censoring of defaults affects all vintages similarly for comparability.

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 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 42.2 million originated mortgages on 1-4 family properties in the 2009-2013 HMDA data. The matched HMDA-McDash sample imposing only the non-excessive-seasoning restriction contains 16.84 million loans, of which 72% are conventional loans. After imposing all of our remaining data filters on this sample, we end up with 9.37 million loans. For all of these loans, we observe whether they ever enter serious delinquency over the first three years of their life—this occurs for 0.74% 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.[18] 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, known as "SATO") for the Black and White Hispanic borrowers. Intuitively, such differences in characteristics make minority populations look

[18]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.

different from the "representative" borrower discussed in the single-characteristic model of default probabilities in the theory section. Depending on the shape of the functions under the new statistical technology, these 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, 2009-2013 Originations.**

Group		FICO	Income	LoanAmt	Rate (%)	SATO (%)	Default (%)
	Mean	739	122	277	4.24	-0.07	0.42
Asian	Median	773	105	251	4.25	-0.05	0.00
(N=574,812)	SD	140	74	149	0.71	0.45	6.49
	Mean	717	91	173	4.42	0.11	1.88
Black	Median	742	76	146	4.50	0.12	0.00
(N=235,673)	SD	127	61	109	0.71	0.48	13.57
	Mean	723	90	187	4.36	0.07	0.99
White hispanic	Median	755	73	159	4.38	0.07	0.00
(N= 381,702)	SD	138	63	115	0.71	0.47	9.91
	Mean	736	110	208	4.33	-0.00	0.71
White non-hispanic	Median	772	92	178	4.38	0.02	0.00
(N=7,134,038)	SD	144	73	126	0.69	0.44	8.37
Native Am, Alaska,	Mean	721	97	204	4.39	0.04	1.12
Hawaii/Pac Isl	Median	759	82	175	4.45	0.04	0.00
(N=59,450)	SD	151	65	123	0.70	0.46	10.52
	Mean	731	119	229	4.38	0.00	0.79
Unknown	Median	770	100	197	4.50	0.02	0.00
(N=984,310)	SD	151	78	141	0.68	0.44	8.85

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. Default is defined as being 90 or more days delinquent at some point over the first three years after origination. Data source: HMDA-McDash matched dataset of fixed-rate mortgages originated over 2009-2013.

It is worth noting one point regarding our data and the US mortgage market more broadly. The vast majority of loans in the sample (over 90%) end up 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 loans.[19] As a result, the

[19]For instance, in addition to their flat "guarantee fee" (i.e., insurance premium), the GSEs charge so-

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 we discuss later in the paper, when we discuss counterfactual equilibria associated with new statistical technologies, this feature of the market makes it less likely that there is selection on unobservables by lenders originating GSE securitized loans, a key factor that we require for identification. Nevertheless, in this section of the paper, we estimate default probabilities using both GSE-securitized and portfolio loans, as we would like to learn about default probabilities using as much data as possible—as we believe a profit maximizing lender would also seek to do.

In the next section we estimate increasingly sophisticated statistical models to predict default in the mortgage dataset. We then evaluate how the predicted probabilities of default from these models vary across race- and ethnicity-based groups in the population of mortgage borrowers.

4 Estimating Probabilities of Default Using Different Statistical Technologies

In this section, we use different prediction methods to estimate $\hat{p}(x, R)$, the three-year probability of default for originated mortgages in the US mortgage dataset, which we will later use to understand the impact of different statistical technologies on mortgage lending.[20]

First, we implement two Logit models to approximate the "standard" prediction tech-

called "loan-level price adjustments" that depend on borrower FICO score, LTV ratio, and some other loan characteristics.

[20]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 lifetime probability of default as $P(x, R) = Pr(\text{Default}|x, R)$. In practice, we do not estimate lifetime probabilities of default, but rather, three-year probabilities of default. We denote these shorter-horizon estimates as $\hat{p}(x, R)$. In the appendix, we discuss the assumptions needed to convert estimated $\hat{p}(.)$ into estimates of $\hat{P}(.)$, which we need for our equilibrium computations later in the paper.

nology typically used by both researchers and industry practitioners (e.g. Demyanyk and Van Hemert, 2011; Elul, Souleles, Chomsisengphet, Glennon, and Hunt, 2010). Second, to provide insights into how more sophisticated prediction technologies will affect outcomes across groups, we estimate a tree-based model and augment it using a number of techniques commonly employed in machine learning applications. More specifically, as we describe below, we implement a Random Forest model (Breiman, 2001), and use 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:

$$\log\left(\frac{g(x)}{1-g(x)}\right) = x'\beta. \tag{3}$$

We estimate two models using this framework, by varying the way in which the covariates in x enter the model. In the first model, all of the variables in x (listed in Table 2) enter 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.[21]

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 the variables in x entering the model for the log-odds ratio linearly, we bin them to allow for the possibility of a nonlinear relationship between x and the log-odds ratio. In particular, we bin LTV, into 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

[21] The Random Forest model, which we describe next, uses the same set of variables as the Logit model.

Table 2: **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.

maximum). Finally, we bin income, using US $25,000 intervals from 0 to US $500,000. We henceforth refer to this model as the Nonlinear Logit.

4.2 Tree-Based Models

We then turn to using machine learning models to estimate $\hat{p}(x, R)$. The term is quite broad, but essentially refers to the use of a range of 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.

Several features differentiate machine learning approaches from more standard approaches to these sorts of problems. For one, the models tend to be non-parametric. Another difference is that these approaches generally use computationally intensive techniques 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, we focus here 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 only a brief overview of the technique here—for a more in-depth discussion of tree-based models applied to a default forecasting problem, see, 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 cross-entropy or 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 Nonlinear 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. An even more 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 Nonlinear 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 needed in a leaf and the minimum number of samples required on a leaf.

4.2.1 Translating Classifications into Probabilities

An important difference between the Random Forest model and the Logit models is that the latter class of models naturally produce an estimate of the probability of default given x. In contrast, the Random Forest model (and indeed, many machine learning models focused on generating "class labels") 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. However, for many purposes, including credit evaluation, the default probability is more useful than the class label alone. In order to use the predictions of the machine learning model as inputs into a model of lending decisions, we need to convert these outputs into probabilities that particular loans will default.

In tree-based models such as 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 known default outcomes in the training dataset ("sigmoid calibration"), and searching across the space of monotonic functions to

find the best fit function connecting the noisy estimates with the true values ("isotonic regression calibration").[22]

We employ isotonic regression calibration to translate the predicted classifications into probability estimates. In the online appendix, we provide more details of this procedure, and discuss how this translation affects the raw estimates in the Random Forest model.

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 performance of the models on a *test* set, which the models have not seen before. In particular, we use 70% of the sample to estimate and train the models, and 30% to test the models. When we sample, we randomly select across all loans, such that the training and test sample are chosen independent of any characteristics, including year of origination. An alternative sampling procedure could sample within year, but given that there are a massive number of loans within each year, the differences between the two procedures should be small.[23]

The training sample is also split into two subcomponents. We use 70% of the training sample as a *model* sample, which we use to estimate the Logit and Nonlinear Logit models, and to train the Random Forest model. 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 as described above. This ensures that both the Random Forest and Logit models have the same amount of data used to estimate their default probabilities.

[22] 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 use this approach in our empirical application.

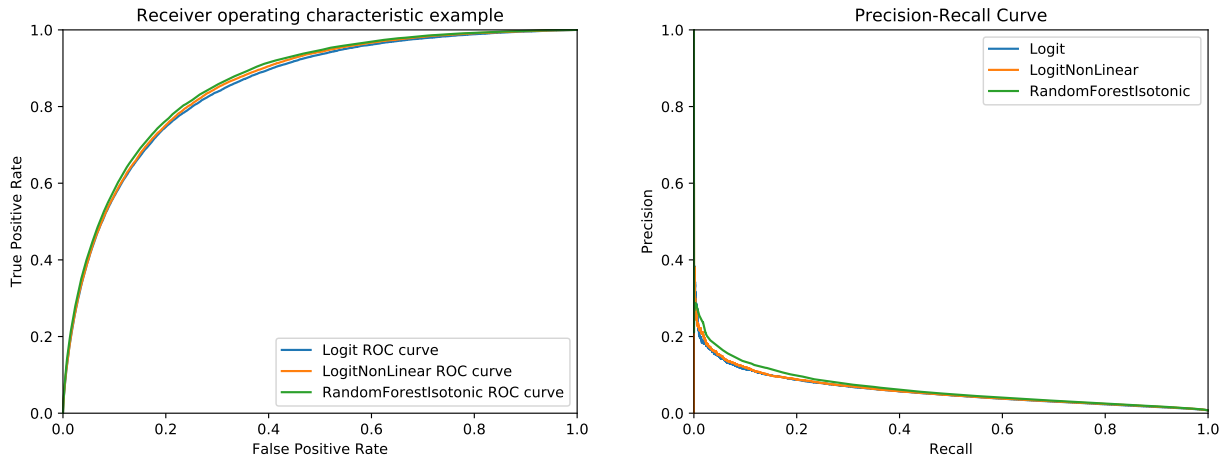
[23] We estimate the Random Forest model using Python's scikit-learn package, and the Logit models using Python's statsmodels package.

4.3 Model Performance

We evaluate the performance of the different models on the test set in several ways. We plot Receiver Operating Characteristics (ROC) curves, which show the variation in the true positive rate (TPR) and 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). A popular metric used to summarize the information in the ROC curve is the Area Under the Curve (AUC; see, for example, Bradley, 1997). 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).[24]

One drawback of the AUC is that it is less informative in datasets which are sparse in defaulters, since FPRs are naturally low in datasets of this nature (see, for example, Davis and Goadrich, 2006). We therefore also compute the *Precision* of each classifier, calculated as $P(y = 1|\hat{y} = 1)$, and the *Recall*, as $P(\hat{y} = 1|y = 1)$, and draw Precision-Recall curves which plot Precision against Recall for different probability thresholds.

Figure 3: **ROC and Precision-Recall Curves**



Panel A Panel B

[24]The TPR is the proportion of defaults in the test set that are correctly identified as such, and the FPR is the fraction of observations in the test set incorrectly identified as defaulters. An intuitive explanation of the AUC is that it captures the probability that a randomly picked defaulter will have been ranked more likely to default by the model than a randomly picked non-defaulter.

Panels A and B of Figure 3 shows the ROC and Precision-Recall curves on the test dataset for the three models that we consider. These figures do not include the race of the borrower as a covariate. They show that the Random Forest model performs better than both versions of the Logit model. In Panel A, the TPR appears to be weakly greater for the Random Forest model than the others for every level of FPR. In Panel B, the Precision-Recall curves, which are better suited for evaluating models on the kind of dataset we consider (sparse in defaulters) show stronger gains for the Random Forest model over the Logit models.

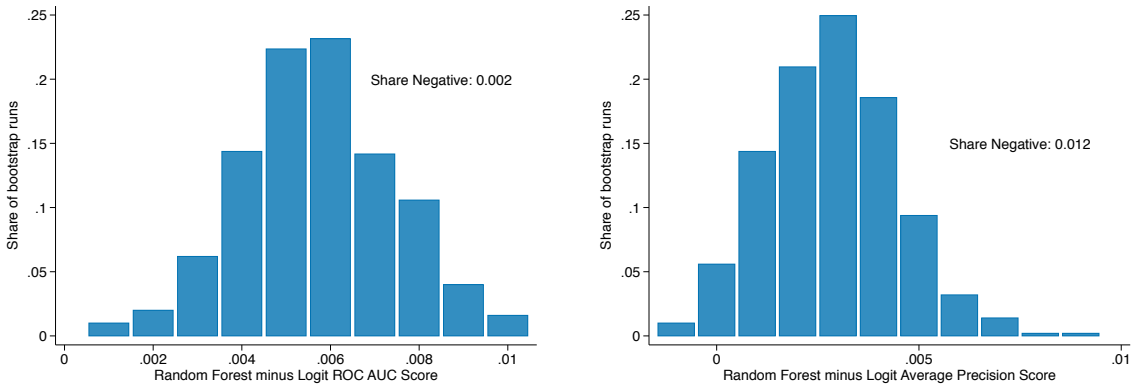
The first and third columns of Table 3 confirm that the AUC and Precision are indeed greater for the Random Forest model than for the other two, suggesting that the machine learning model more efficiently uses the information in the training dataset in order to generate more accurate predictions out of sample.

Table 3: **AUC and Precision for Different Statistical Technologies Predicting Default**

	ROC AUC Score		Precision Score	
Model	No Race	Race	No Race	Race
Logit	0.8517	0.8522	0.0589	0.0592
Logit Non-Linear	0.8565	0.8569	0.0600	0.0603
Random Forest	0.8626	0.8626	0.0633	0.0635

In order to verify that these differences are indeed statistically significant, we use bootstrapping. We randomly resample with replacement from the original dataset to create 500 bootstrapped sample test datasets. Holding fixed our estimated models, we re-estimate the average Precision and AUC scores for all of the models on each bootstrapped sample. Panels A and B of Figure 4 plot the histogram across boostrapped datasets of the difference in these scores between the Random Forest and the Nonlinear Logit models. The figure shows that the Random Forest AUC is greater than that of the Nonlinear Logit 99.8% of the time, with an average improvement of 0.7 percent, and the corresponding Precision score increases 98.8% of the time, with an average improvement of 5.5 percent.

Figure 4: **Bootstrap Estimates of Differences in AUC and Average Precision**



Panel A: Difference in ROC AUC

Panel B: Difference in Average Precision

4.3.1 Model Performance With and Without Race

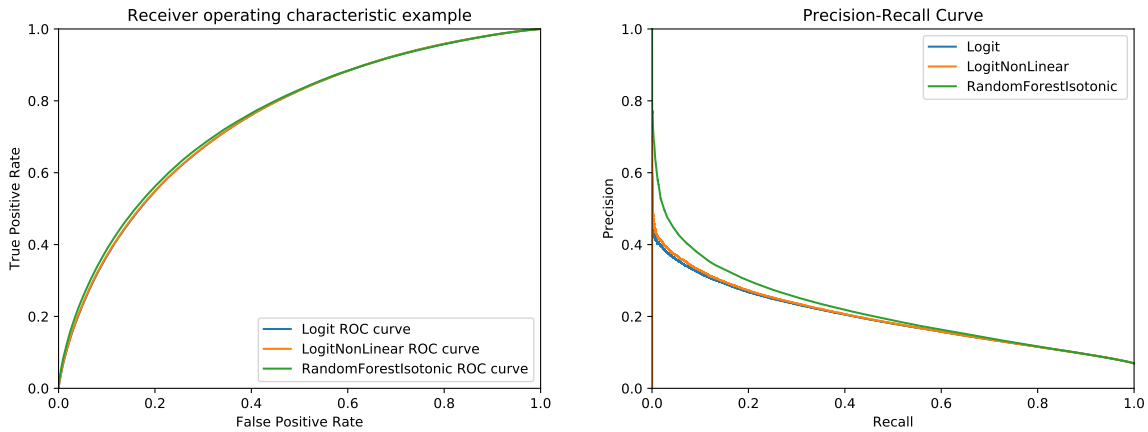
The second and fourth columns of Table 3 show that the inclusion of race has different effects on the three models. Both of the Logit models benefit from the inclusion of this excluded variable—allowing the AUC of these models to reduce the gap versus the AUC of the Random Forest model. In contrast, there is virtually no change in the AUC of the Random Forest model.

Evaluating changes in the relative predictive ability of the models as a result of the inclusion of race is interesting. In keeping with the spirit of the law prohibiting differentiation between borrowers on the basis of excluded characteristics, assessments of borrower risk should be colorblind. This seems to be the case for the two Logit models, in the sense that race appears to marginally augment their performance. The Random Forest model seems less affected by the elimination of information about race, suggesting that the model is able to more efficiently triangulate the association between race and default using the remaining borrower characteristics.

To explore this issue further, we employ the three models to predict whether a borrower is Hispanic or Black using the same set of variables used to predict default. This exercise reveals striking differences between the models, especially in Panel B of Figure 5. Table 4

confirms that the Random Forest outperforms the other two models, which have very similar scores, by 7.8% in terms of average precision and 0.7% in terms of AUC.

Figure 5: **ROC and Precision-Recall Curves of Predicting Race**



Panel A Panel B

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

Model	ROC AUC Score	Precision Score
Logit	0.7478	0.1948
Logit Non-Linear	0.7484	0.1974
Random Forest Isotonic	0.7537	0.2128

Next, we document how estimated probabilities of default from these models vary across race-based groups in US mortgage data.

4.4 Differences in Predicted Default Propensities

Having estimated the different models, we can inspect how they 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 compare with those estimated using the Nonlinear Logit model. The figure focuses on understanding which race and ethnic groups "win" and "lose" under the new technology, in keeping with our central motivation.

Panel A of the figure shows the cumulative distribution function of the increase in the estimated default probability and moving from Nonlinear Logit to Random Forest, holding constant the interest rate at $R = 4.5\%$ for all borrowers. Each line in this plot represents a different race group. Borrowers for whom this difference is negative are "winners" from the new technology (in the sense of having a lower estimated default probability), and those with a positive difference are "losers". Panel B plots the log difference in default probabilities to highlight the proportional benefit for each group.[25]

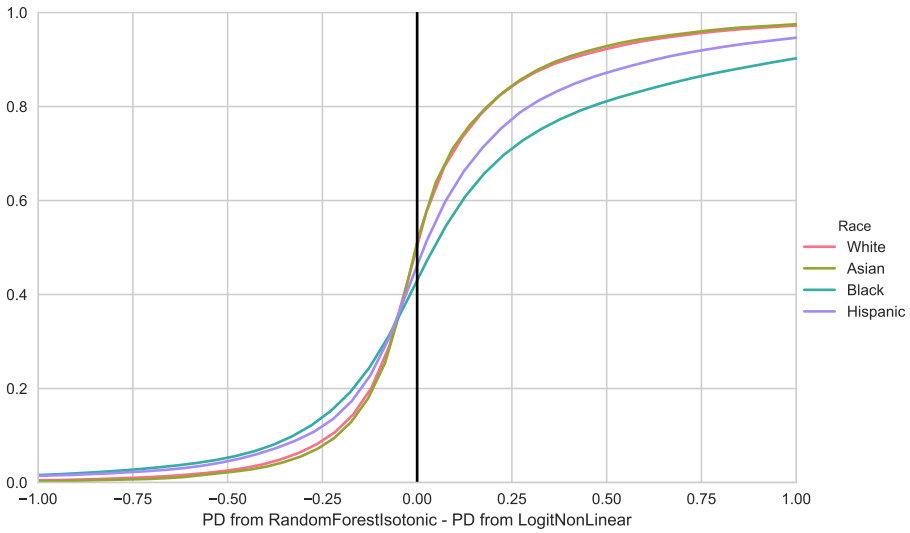
Panel B shows that there is a proportional reduction in default risk under the Random Forest model for the population as a whole. The y-axis of the plot shows that across all groups, the share of borrowers for whom the estimated probability of default falls under the new technology is either marginally above 50%, or at worst, slightly less than 50%. In this sense, aggregated across the population, and weighting by the representation of different race groups, the new technology seems to perform similarly to the old technology in terms of assigning lower probabilities of default to the borrower population.

However, what is also evident are important differences between different race groups. Panel B shows that the winners from the new technology are disproportionately White non-Hispanic and Asian – the share of the borrowers in these groups that benefit from the new technology is above the 50% mark. 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. In particular, the cdfs of the differences evaluated at 0 for the White non-Hispanic and Asian populations are clearly above the corresponding cdfs for the Black and White Hispanic groups. We also 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 Nonlinear Logit model, and return to this finding later.[26]

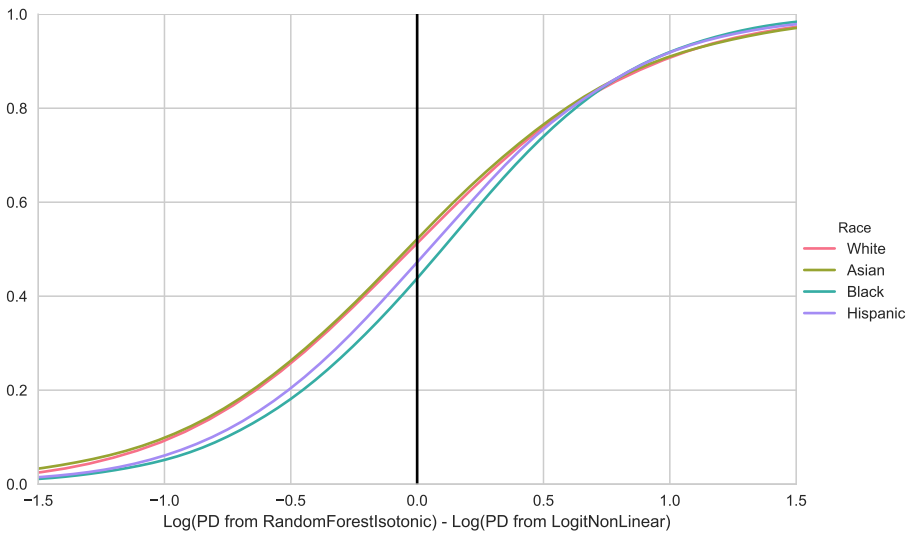
[25] For ease of visual representation, we have truncated the x-axes on these plots, as there are a small share of cases in which the estimated differences in the pds are substantial.

[26] It is also true that the distributions are right-skewed, i.e., the Random Forest model has a tendency to predict far higher probabilities of default for some of the borrowers in all groups than the Logit model.

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



Panel A



Panel B

The figure provides useful insights into the questions that motivate our analysis, and suggest that there may indeed be variations in the fraction of winners and losers across race groups engendered by technology. However, to make further progress, we need to better understand how changing probabilities of default translate into changes in interest rates or exclusion. The next section discusses how we model equilibrium when contract characteristics are endogenous, to facilitate more meaningful statements about possible changes to these ultimate outcomes as technology varies.

5 Equilibrium and Statistical Technology

Thus far, our discussion has concentrated on the case in which lenders evaluate default probabilities 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.

5.1 A Simple Model of Equilibrium

We 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) at origination are pre-determined for each borrower.[27] We therefore think of L and LTV at origination 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.[28]

[27] In reality, of course, these parameters are often dictated, or at least confined to a narrow range, by local property prices and liquidity constraints faced by the borrower.

[28] In the online 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 ($\hat{P}(x, R|\mathcal{M})$) of default on the interest rate, where $\mathcal{M}$ continues to denote a given statistical technology.

We begin with a general NPV formula of the mortgage to a risk-neutral lender for a loan of size L, at interest rate R:

$$NPV = \frac{1}{1+\rho}\left[(1-P)(1+R)L + P\hat{L}\right] - L. \tag{4}$$

In equation (4), lenders' net cost of capital between dates 0 and 1 is denoted by $\rho > 0$. In the event of default, the lender receives $\hat{L}$, the recovery value in default. P is the lifetime probability of default. As we have discussed in the previous sections, we estimate P using borrowers' observable characteristics and offered rates in the data. We therefore denote P as $P(x, R)$ and NPV as $N(x, R)$ to make this dependency explicit.

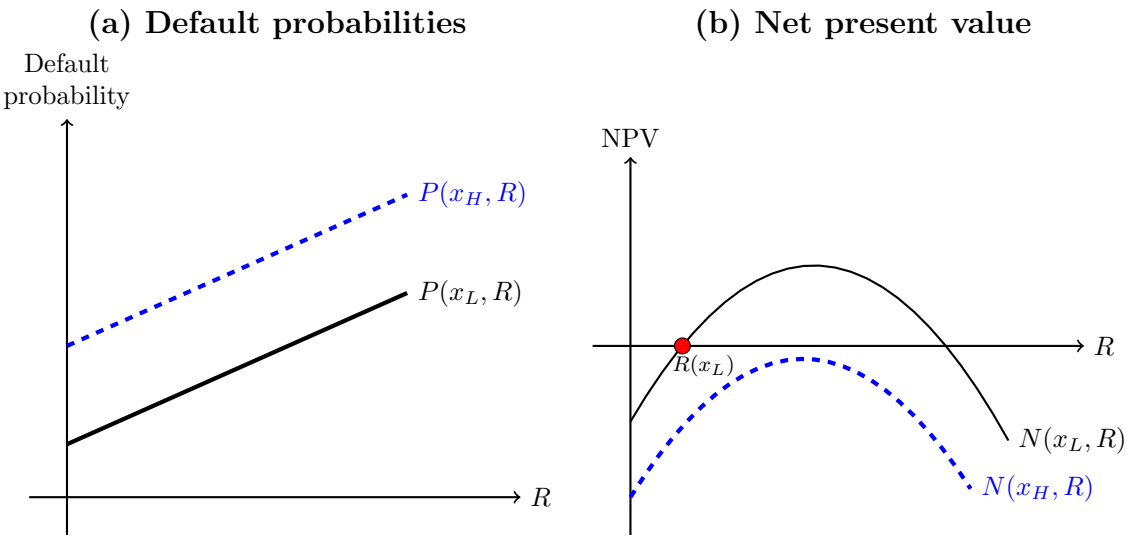
Note that $N(x, 0) < 0$ for all x; intuitively, a positive interest rate is required to allow lenders to break even. In general, the NPV need not be a monotonic function of R, since higher interest rates increase the yield on the mortgage, but a greater interest burden may generate a strong temptation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

$$R(x) = \min\{R|N(x, R) = 0\} \tag{5}$$

Figure 7 illustrates the determination of equilibrium in this model using a simple example where predicted default probabilities $\hat{P}(x, R|\mathcal{M})$ 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 online appendix, we formally derive the above equilibrium conditions in a canonical model of lender and borrower behavior.

Figure 7: **Equilibrium determination.**



To facilitate mapping the model to the data, we assume that $\hat{L} = \gamma\delta V$. Here, V is the house price at origination, the expected house value at default is δV, where $\delta < 1$ reflects the fact that default correlates with low house prices, and we assume that the lender can recapture γ of the value at default in the event of foreclosure, where the remainder $1 - \gamma$ captures deadweight costs of foreclosure. Using the additional identity that LTV$= \frac{L}{V}$ at loan origination, we get (see online appendix for more details):

$$NPV = \frac{L}{1+\rho}\left[(1-P)(1+R) + P\frac{\gamma\delta}{\text{LTV}} - (1+\rho)\right] \tag{6}$$

5.2 Identification

In equation (6), lenders base their decisions on a reduced-form prediction $\hat{P}(x, R|\mathcal{M})$ of default probabilities, given statistical technology $\mathcal{M}$. 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. However, when estimating counterfactual equilibria under alternative statistical technologies, we note several potential identification issues that arise from our approach of relying on this reduced form approach, and take steps to account for these issues.

The essential identification problem that we face is that our calculation of the lender's expected NPV is valid if and only if $\hat{P}(x, R|\mathcal{M})$ that we estimate in reduced form is an *unbiased* predictor of the true likelihood of default once the mortgage is originated.

Several selection issues arise in this context. First, we only observe one potential default response for each borrower in the data, namely the one associated with the interest rate actually assigned to the borrower in the data. Second, this issue is further complicated by the fact that if a borrower is not granted a mortgage by lenders in the data, we do not observe her at all. Third, 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.

In order to deal with the first issue above, we make the standard assumption permitting identification in the face of selection issues, namely, conditional independence, i.e., given observable borrower characteristics x_i, the treatment (interest rate R_i) is drawn independently

of potential default outcomes $y_i(R)$, for all potential R. Since this is a strong assumption, as we explain in more detail later, we further correct our estimates for the bias introduced by selection on unobservables, adjusting our estimated default sensitivities to interest rates to be in line with plausibly causal estimates from the mortgage literature.

To deal with the second issue above (we cannot observe borrowers not granted mortgages), we restrict our counterfactual statements to populations with distributions of borrower characteristics identical to the one we observe in the data. That is to say, when reporting population averages, we will implicitly weight borrower characteristics by the observed density of characteristics in the HMDA-McDash merged dataset. Under the assumption that borrowers denied a mortgage are high credit risks, we will therefore potentially understate (overstate) the population averages of extensive margin credit expansions (contractions) when evaluating equilibrium under a counterfactual technology.[29]

The third issue is mitigated by the fact that the population object we are estimating, $\hat{P}(x, R|\mathcal{M})$, is the probability of default conditional on the borrower's decision to accept the contract. This is the relevant probability for a lender calculating expected profit. In our dataset (and indeed in any such dataset), we estimate default propensities using borrowers who accepted contract offers. However, we show in the online appendix that under conditional independence, estimates from such a dataset are unbiased for $\hat{P}(x, R|\mathcal{M})$.[30]

5.2.1 GSEs and No Selection on Unobservables

As we discuss in more detail in the online appendix, a natural sufficient condition for identification using the conditional independence assumption is that there is no selection on unobservables. If lenders have no information that correlates with determinants of borrower

[29]We are unable, of course, to draw inferences about counterfactual acceptances in regions of borrower characteristics that we do not observe in the data. However, it is worth noting that we can still make statements about increased counterfactual densities of borrower acceptances in regions of the characteristic distribution that we do observe.

[30]However, it is still the case that unobservable changes in the borrower population's propensity to accept offers will generate selection issues in our estimates.

behavior other than the "hard" information that we observe, x_i, then default predictions are identifiable, even when counterfactual lending and pricing decisions are not observed.

In our empirical work, the sample period that we focus on occurs after the 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.

However, as discussed earlier, to be conservative, we restrict our analysis to GSE-insured mortgages (i.e. those securitized through Fannie Mae or Freddie Mac), as they are far less likely to suffer from selection by lenders on unobservable borrower characteristics; instead, lenders mainly focus on whether a borrower fulfills the underwriting criteria set by the GSEs.[31] For estimating default propensities which feed in to equilibrium computations, we therefore only include loans securitized by the GSEs, and which are marked as having been originated with full documentation of borrower income and asset. This leaves us with 5.16 million loans, of which 0.60% enter serious delinquency over the first three years of their life.[32]

Once we estimate the functions $\hat{P}(x, R|\mathcal{M})$ under the different technologies using all

[31] As mentioned earlier, Keys, Mukherjee, Seru, and Vig (2010) argue that there are discontinuities in lender screening at FICO cutoffs that determine the ease of securitization, but only for low-documentation loans (where soft information is likely more important), not for full-documentation loans such as the ones we consider.

[32] Restricting the estimation sample to loans for which the GSEs, and not the originating lender, bear the credit risk may appear at odds with the model we consider, where loans are held in lender portfolios. However, even a lender that only makes portfolio loans would wish to learn about default probabilities using as much data as they can acquire, and GSE loans account for the vast majority of loans in our sample of conventional loans. Furthermore, 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). 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.

GSE mortgages in the data, we apply a further correction to them, described below. In this version of the paper, for computational purposes, we then use the corrected $\hat{P}$ functions to evaluate counterfactual equilibria using a subset of these loans. The "equilibrium sample" comprises 100,000 randomly selected GSE, full documentation, 30-year purchase loans for owner-occupied homes in 2011.[33]

5.2.2 Interest Rate Sensitivity Adjustment

Even after restricting the sample to the set of GSEs, there may still be remaining selection concerns about using our estimated sensitivities of probabilities of default to mortgage rates (or SATO) in the equilibrium calculation.[34] Concretely, such concerns are that as we change counterfactual mortgage rates in our equilibrium calculation, we could be overstating the importance of interest rates for default probabilities, and may reach mistaken conclusions on equilibrium rates under alternative technologies.[35]

To further correct for this source of bias, therefore, we rely on and extend existing work that estimates the *causal* effect of interest rate changes on mortgage default. Specifically, Fuster and Willen (2017) use downward rate resets of hybrid adjustable-rate mortgages to estimate the sensitivity of default probabilities to changes in rates. These resets occur three years or more after origination of the mortgages and are determined by the evolution of benchmark interest rates (such as LIBOR). Using the same dataset as Fuster and Willen (2017) (non-agency hybrid ARMs), we estimate a (non-causal) cross-sectional sensitivity of

[33]The online appendix shows the summary statistics for the 2011 sample and the cumulative distributions of differences in default probabilities between the models for the equilibrium sample. It shows that the patterns are very similar to those in the full sample. For computational purposes, in this draft we focus on purchase loans, to focus on what lenders might do when originating new loans, and also because these loans are more relevant when thinking about issues of access to credit. We have checked that 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.

[34]For instance, it is possible that some borrowers were charged a higher rate precisely because they were at higher risk of default in ways not observable in our data. Alternatively, it is possible that borrowers with higher financial literacy shop around for low mortgage rates more, and are also at lower risk of default (even if they did not have a lower-rate mortgage).

[35]Of course, we also note here that it is not clear that lenders armed with the same data as us would recognize that estimated rate sensitivities are not structural when deciding on their rate offerings.

3-year default probabilities to a 50 basis point change in the interest rate spread at origination (SATO), using the same hazard model used for the Fuster and Willen (2017) causal estimates. When we compare the resulting non-causal estimate to their causal estimates, we find that it is 1.7 times as large. The online appendix describes how we use this factor to adjust our empirical estimates before plugging them into the NPV calculations. We have reason to believe that this adjustment is quite conservative, since the non-causal estimate comes from defaults occurring in the first-three years—this is more likely to comprise the segment of interest-rate sensitive borrowers.

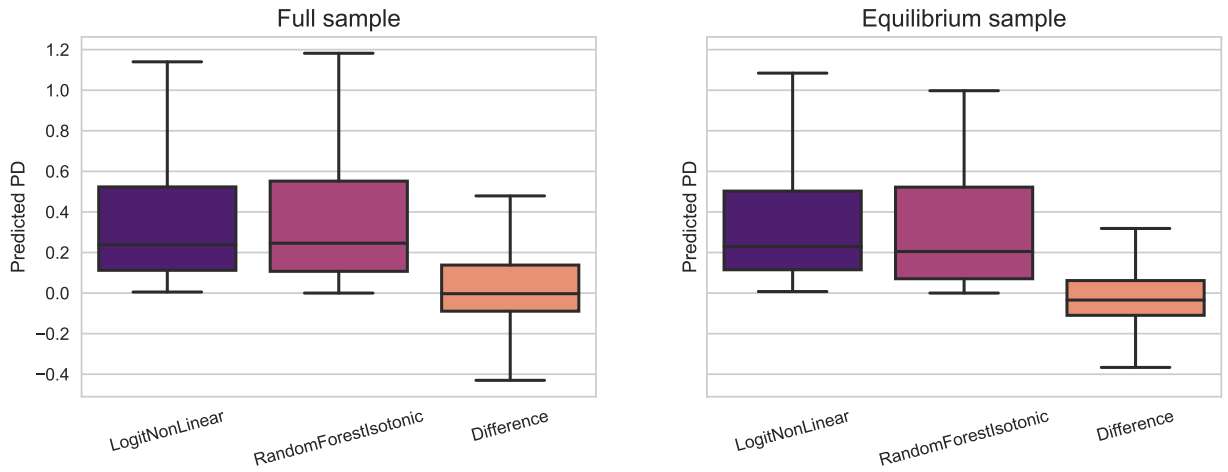
5.3 Parameter Choices and Estimation Details

When taking equation (6), we make choices for the parameters entering the equation: First, we set the WACC $\rho = 4.5$ percent to match the average observed interest rate of 4.62 percent. This minimizes the amount of extrapolation beyond frequently observed combinations of borrower characteristics and interest rates in our data. Second, we set the parameter combination $1 - \gamma\delta$ to 0.25, roughly in line with the loss severities that An and Cordell (2017) document for Freddie Mac insured loans originated post 2008.

When computing equilibrium, for every borrower i, we evaluate $NPV(x_i, R)$ at a grid of 10 interest rates between 1.5 and 6 percent. We then use linear interpolation to solve for the equilibrium interest rate $R_i^\star$, i.e. the smallest root of $NPV(x_i, R) = 0$. If no such solution exists within the grid of interest rates considered, we conclude that borrower i is not accepted for a loan.

The estimated default propensities that we estimate using the different statistical technologies are predictions of default in the first 36 months of each loan's life, meaning that all our default data are censored 36 months after origination for all cohorts. We denote these estimated 36 month default rates by $\hat{p}(x, R)$. However, equation (6) takes as an input lifetime default rates $P(x, R)$. We therefore convert our estimates of $\hat{p}(x, R)$ into estimates of $\hat{P}(x, R)$ using a procedure based on the Standard Default Assumptions (SDA) used in the

Figure 8: **Predicted PD, comparing Full and Equilibrium samples.**



mortgage investor community, as described in the online appendix.

5.4 Residual Interest Rate Variation

Figure 8 shows how the estimated probabilities of default from the different models differ between the full sample and the equilibrium sample. The figure shows that the variance, and indeed, the right tail, of estimated default probabilities is smaller in the equilibrium sample. The reduction in the variance of the estimated default probabilities is consistent with less unobservable information used in the selection and pricing of the loans in the equilibrium sample.

Table 5 below shows results from a more direct way to check for the prevalence of soft information. It shows that the residual variation in interest rate spreads at origination (SATO), when regressed on the observable variables in our model, is clearly smaller in the equilibrium sample.

Finally we check if, when computing equilibrium, we are predicting default rates for combinations of borrower characteristics and interest rates that are scarcely observed in the data. This would place a great burden of extrapolation on our estimated models, and we

Table 5: **Residual Variation in SATO, comparing Full and Equilibrium samples.**

Mortgage type	SATO residual	SATO
Equilibrium sample	0.259	0.291
Other	0.287	0.412

Note: In the full sample, we regress observed SATO on characteristics (i.e. the RHS variables in the linear Logit). This table shows the standard deviations of the residual from this regression (left column) and of the raw SATO series (right column) conditional on loan type. The first row shows standard deviations among loans that satisfy the restrictions imposed on the equilibrium sample (GSE, full documentation, 30-year purchase loans for owner-occupied homes in 2011). The second row shows standard deviations for remaining loans in the full sample. 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 fixed-rate mortgages.

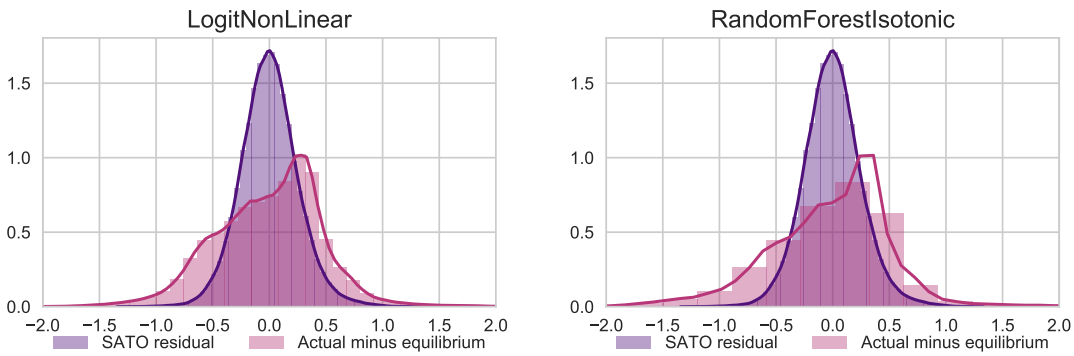
would like to avoid this (although one might argue that a profit-maximizing lender would also use some extrapolation if the data was sparse). We also therefore compare the residual SATO to the difference between actual interest rates and model-implied equilibrium rates for all borrowers in our sample. Figure 9 shows histograms and kernel density estimates for the SATO residual and the difference between actual and equilibrium rates.

The figure shows that the counterfactual equilibrium rates that we predict differ from actual rates, but for the most part, these changes to the predictions lie within the region covered by residual variation, or the "noise" in observed interest rates. It is true that a small fraction of our predictions is driven by extrapolation outside the noise in rates that we observe (the area under the actual rates minus equilibrium rates curve that does not overlap measures this fraction), but the patterns in the plot are broadly reassuring about the fairly limited extent of this extrapolation.[36]

We next turn to understanding how rates and exclusion outcomes for different groups in the population are likely to change as technology varies.

[36]Counterfactual differences lying precisely within the range of the residuals, are "supported" by the noise in the residuals, and counterfactual differences lying outside the range of residuals, are outside the space of fitted rates, meaning that we may be venturing into ranges of the data that may have been generated by selection on unobservables. The plot shows that the latter case occurs relatively infrequently.

Figure 9: **Residual interest rate variation**.



6 Technology and Disparity in the Data

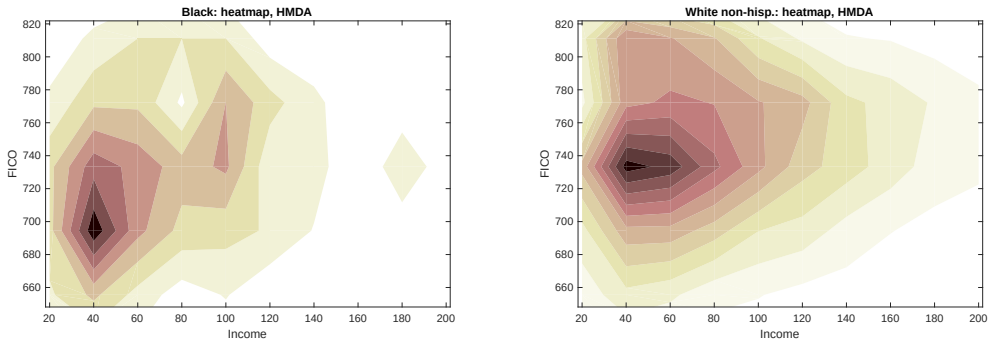
In this section we discuss how the issues highlighted in the initial theory section actually play out in the real US mortgage market data that we analyze, once we input estimated default probabilities into our equilibrium model. To build intuition, we begin in two-dimensional space, as in our theoretical 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 $P(x, R)$. 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 focus on our equilibrium sample in these plots, plotted for an interest rate $R = 4.5\% = \rho$.

While these plots are helpful in continuing to build intuition, they are not representative of the patterns in the entire data, and we therefore return to tabulating more aggregate measures for the entire equilibrium sample at the end of the graphical presentation to illustrate how different groups win and lose on both the extensive margin (exclusion) and the intensive margin (rates) as statistical technology improves.

Group-Conditional Distributions of Borrower Characteristics

Figure 10 shows the empirical frequency of borrower FICO and income by racial group, for both Black (left panel) and White (right panel) borrowers in the equilibrium sample.[37]

Figure 10: **Distribution of Borrower Characteristics.**



The 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 substantially 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 $g \in \{\text{Black,White}\}$.[38]

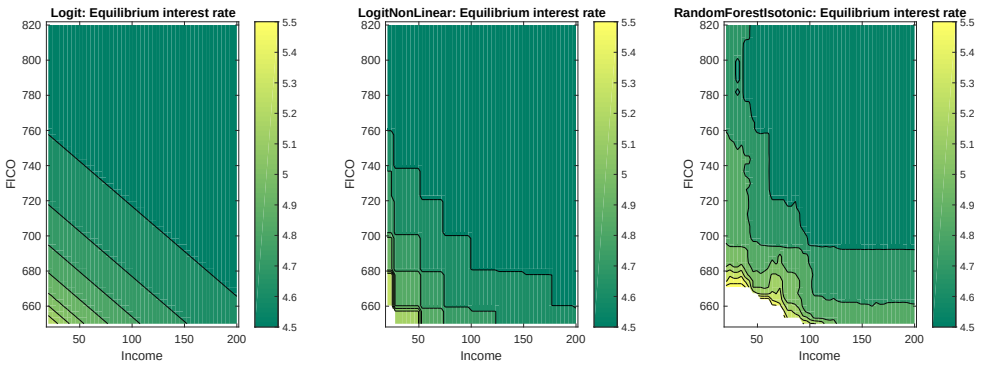
[37] We plot the distribution here for all borrowers with loan amount $L \in [250000, 350000]$ and $LTV \in [75, 85]$ to correspond with our analysis of rates and predicted default later.

[38] If we were only to use the sample of mortgage borrowers that were acccepted once they applied for a mortgage, it would understate exclusion since rejected applicants do not 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 online appendix. Figure 10 shows the imputed joint distribution of FICO and income for Black and White borrowers. The imputation procedure simply lowers the means of both variables in both distributions, and increases the variances.

6.1 Equilibrium Outcomes

Figure 11 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.[39] From this graphical analysis, it appears as if 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.

Figure 11: **Equilibrium Interest Rates.**



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 12 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

[39]For the graphical analysis of equilibrium rates, we vary FICO and income while holding all other observable borrower characteristics constant. We therefore restrict attention 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 4.5% (comprised of a mortgage base rate of 4.4%, and SATO of 10 bp).

Figure 12: **Equilibrium Interest Rates and Distribution of Characteristics.**

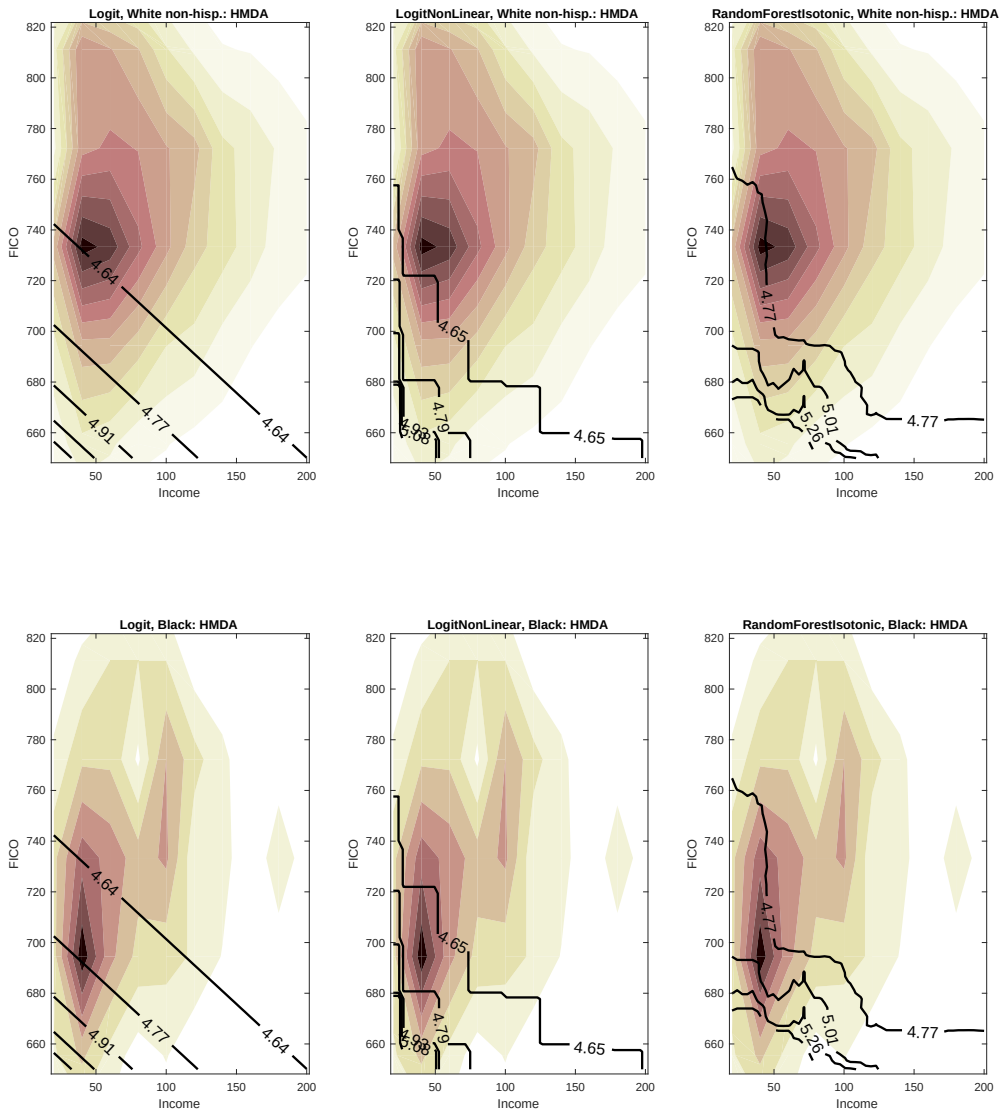
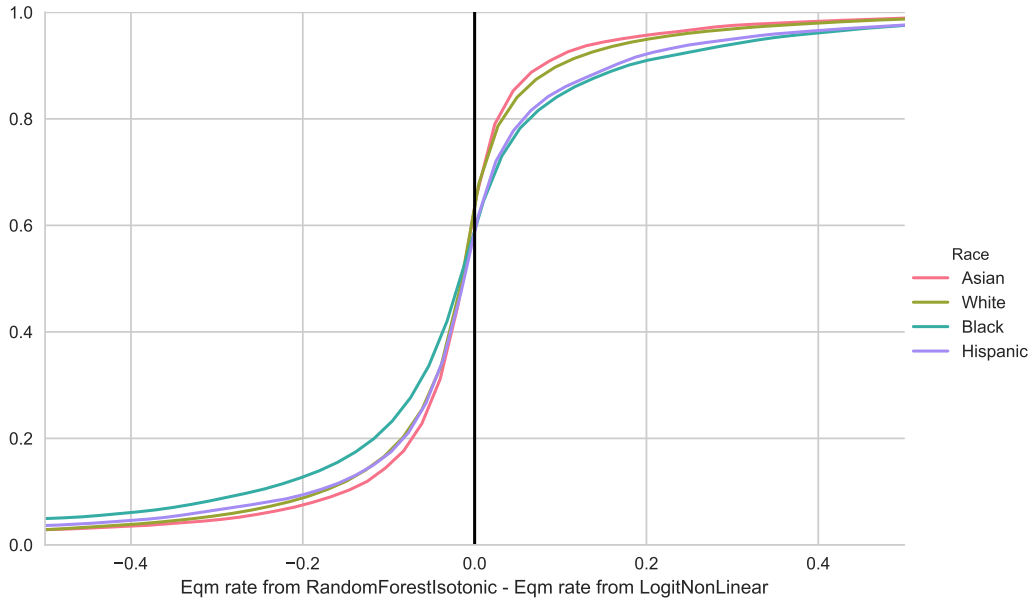


Figure 13: **Comparison of Equilibrium Interest Rates.**



Black borrowers on to the equilibrium rates and exclusion regions associated with the different underlying statistical technologies. The bottom panels of Figure 12 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.

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 13 plots the difference of offered rates in equilibrium under the Random Forest model and those under the Nonlinear Logit model, for the borrowers approved for a loan under both technologies.

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

Table 6: **Equilibrium Outcomes**.

	Proportion accepted	Mean eq. rate	SD eq. rate
LogitNonLinear	0.898	4.613	0.306
RandomForestIsotonic	0.92	4.591	0.524
Actual		4.624	0.398

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.

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.

6.2 Measuring Disparity

To make further progress, we first turn to Table 6, which looks at selected summary statistics from equilibrium computed using the different technologies.

The first and second columns of the table show that the proportion of borrowers accepted and the average rates for borrowers across the Logit and Random Forest models are very similar. However, the third column shows that the dispersion of rates is very different across the models, with the more sophisticated technology producing predictions with a far higher spread. These facts are reminiscent of our Lemma 1, in which the new technology generates pds which are a mean-preserving spread of the older technology.

Who wins and who loses in the new equilibrium associated with the more sophisticated technology? The first column of Table 7 shows mean equilibrium acceptance rates. The second column shows the mean interest rate for the group in equilibrium, and the final column shows population frequencies of each racial group. The first five rows of the table show these statistics for each of the racial groups in the data, and the sixth, averaged

across the entire population. The panels show these statistics for the underlying statistical technologies.

Table 7: **Cross-Group Disparity.**

LogitNonLinear			
	Acceptance rate	Av. interest	Frequency
Asian	0.929	4.578	0.068
White	0.902	4.614	0.775
Hispanic	0.827	4.643	0.045
Black	0.761	4.661	0.02
Other	0.908	4.61	0.092
Population	0.898	4.614	
Cross-group st.dev	0.026	0.013	
RandomForestIsotonic			
	Acceptance rate	Av. interest	Frequency
Asian	0.946	4.55	0.068
White	0.925	4.592	0.775
Hispanic	0.85	4.635	0.045
Black	0.802	4.638	0.02
Other	0.923	4.587	0.092
Population	0.92	4.592	
Cross-group st.dev	0.024	0.016	

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., acceptance 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:

$$\delta_\tau = \sqrt{\sum_g \phi_g \left(\gamma_{g,\tau} - \bar{\gamma}_\tau\right)^2} \tag{7}$$

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. The first finding here

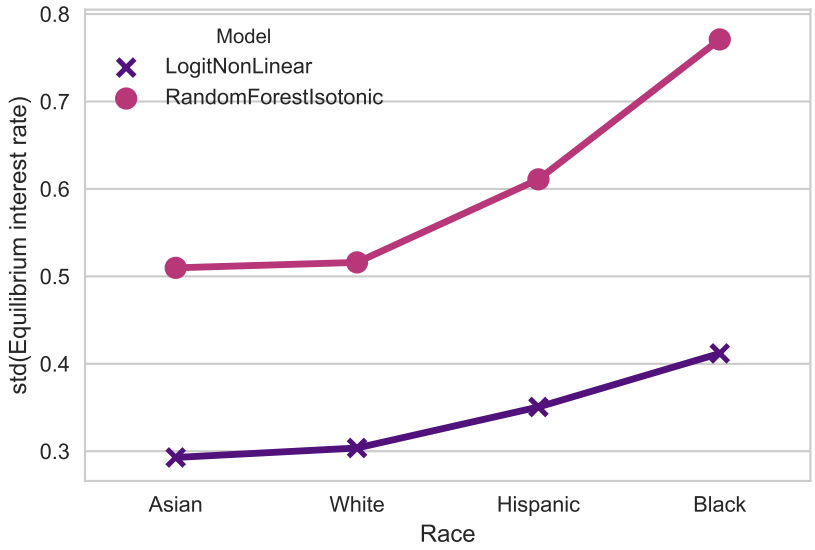
is that the Random Forest model has a slightly lower δ_τ for acceptance – 7.7% lower than for the Nonlinear Logit model. This is also accompanied by a lower average rejection rate for this model across all groups (i.e., an average rejection rate of 8% under the Random Forest model as opposed to 10.2% 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. The equilibrium rates are very similar under the two technologies (around 2.2 basis points higher on average under Nonlinear Logit). However, the disparity of rates across groups is significantly higher under the new technology. The point estimate of $\delta_\tau = 0.016$ is 23% higher than the comparable point estimate for the Nonlinear Logit model. This reflects the differential changes in the average rate across groups.

Overall the picture that Table 7 paints is an interesting one. As we have seen earlier, the Random Forest model is a more accurate predictor of defaults. Moreover, it generates higher acceptance rates and slightly lower interest rates on average. However, it penalizes some minority race groups significantly more than the previous technology in the process, by giving them higher interest rates.

Figure 14 shows the *within* group dispersion of predicted equilibrium rates associated with the different statistical technologies. The table shows that this dispersion goes up by a factor of two - from 40 to roughly 80 basis points for the group of Black borrowers under the Random Forest model, while there is a smaller increase in this dispersion for White non-Hispanic and Asian borrowers - roughly a factor of 1.6, with an increase from 30 to roughly 50 basis points. Overall, these patterns in within group dispersion suggest that the Random Forest model screens within minority groups more extensively than the Nonlinear Logit model, leading to changes in both exclusion and rate patterns associated with the new technology.

Figure 14: **Within-group dispersion of equilibrium rates**.



7 Conclusion

In this paper, we find that changes in statistical technology used to identify creditworthiness can generate disparity in credit outcomes across different groups of borrowers in the economy. We present simple theoretical frameworks to provide insights about the underlying forces driving towards such changes in outcomes, and verify that the issue manifests itself in US mortgage data.

The essential insight is that a more sophisticated statistical technology, virtually by definition, generates more disperse predictions as it better fits the predicted outcome variable (in the case that we consider, this is the probability of mortgage default). It immediately follows that such dispersion will generate both "winners" and "losers" relative to their position in equilibrium under the pre-existing technology.

It is of course clear that efficiency gains can arise from the improved use of underlying information by new technologies. However our work highlights that at least one reason to more carefully study the impact of introducing such technologies is that the winners

and losers from their widespread adoption can be unequally distributed across societally important categories such as race, age, income, or gender.

In our empirical application, we find that even though the new statistical technology is not explicitly allowed to use information about race group membership during default prediction, it is better able to triangulate the information connecting default propensity with these memberships using legitimately included variables. We also find that minority groups appear to lose, in terms of the distribution of predicted default propensities, and in our counterfactual evalution, 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 likely winners and losers in society from the inevitable adoption of machine learning and artificial intelligence.

8 Appendix

8.1 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 $\hat{f}_1$.

8.2 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

$$\hat{f}(x|\mathcal{M}) - \hat{f}(x|\ell) = \sum_{j=2}^{\infty} \frac{1}{j!} \frac{\partial^j \hat{f}(a|\mathcal{M})}{\partial x^j}(x - a)^j - B \tag{8}$$

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

References

An, X., and L. Cordell (2017): "Regime Shift and the Post-Crisis World of Mortgage Loss Severities," Working Paper No. 17-08, Federal Reserve Bank of Philadelphia.

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.

Bartlett, R., A. Morse, R. Stanton, and N. Wallace (2017): "Consumer Lending Discrimination in the FinTech Era," Working paper, UC Berkeley.

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).

Bradley, A. P. (1997): "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, 30(7), 1145 – 1159.

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.

Buchak, G., G. Matvos, T. Piskorski, and A. Seru (2017): "Fintech, Regulatory Arbitrage, and the Rise of Shadow Banks," Working Paper 23288, National Bureau of Economic Research.

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.

Davis, J., and M. Goadrich (2006): "The Relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240. ACM.

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.

Fuster, A., M. Plosser, P. Schnabl, and J. Vickery (2018): "The Role of Technology in Mortgage Lending," Staff Report 836, Federal Reserve Bank of New York.

Fuster, A., and P. Willen (2017): "Payment Size, Negative Equity, and Mortgage Default," *American Economic Journal: Economic Policy*, 9(4), 167–191.

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.

Keys, B. J., T. Mukherjee, A. Seru, and V. Vig (2010): "Did Securitization Lead to Lax Screening? Evidence from Subprime Loans," *Quarterly Journal of Economics*, 125(1), 307–362.

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.

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.

Online Appendix to "Predictably Unequal? The Effect of Machine Learning on Credit Markets"

A.1 Isotonic regressions and calibration

Denote by y_i 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:

$$\hat{z} = \arg\min_{z} \sum (y_i - z(l_i))^2. \tag{9}$$

Figure A-1: **Calibration Curve.**

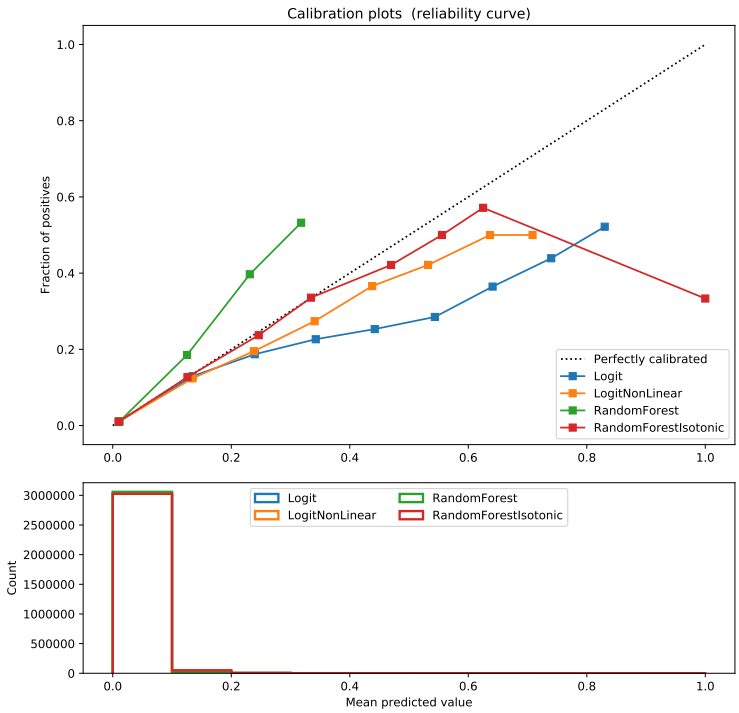


Figure A-1 plots the number of defaults within each bin shown on the y-axis against the binned predictions from each of the models on the x-axis. A well-calibrated model would lie along the 45° line. The Non-Linear Logit model looks relatively well-calibrated, but in comparison, the Random Forest model (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.

A.2 LTV and LGD

We begin with a general NPV formula:

$$NPV = \frac{1}{1+\rho}\left[(1-P)(1+R)L + P\hat{L}\right] - L$$

where $\hat{L}$ is the recovery value in default, and P is the lifetime probability of default.

At loan origination,

$$\text{LTV} = \frac{L}{V}$$

where V is the house price at origination. Now suppose that the expected house value at default is δV, where $\delta < 1$ reflects the fact that default correlates with low house prices. In the event of default, assume that the lender seizes the house, and is able to recover γ of its current value, where the remainder $1-\gamma$ captures deadweight costs of foreclosure.

Then the total recovery amount is:

$$\begin{aligned}\hat{L} &= \gamma\delta V \\ &= \gamma\delta\frac{L}{\text{LTV}}\end{aligned}$$

and the NPV is therefore:

$$NPV = \frac{L}{1+\rho}\left[(1-P)(1+R) + P\frac{\gamma\delta}{\text{LTV}} - (1+\rho)\right] \tag{10}$$

In what follows, we discuss how we estimate P using borrowers' observable characteristics and offered rates in the data. We therefore denote P as $P(x, R)$ to make this dependency explicit.

A.3 Derivation of equilibrium prices

An industry of $N \geq 2$ mortgage lenders faces a population of potential borrowers. Each borrower has a vector $x \in \mathcal{X}$ of observable characteristics, which lenders observe. As in the text, we treat the loan amount L and the house price V as exogenously given, and subsume them into the vector x of observables.

The timing of the game is then as follows: First, each lender offers mortgage rates R to borrowers at date 0, the terms of which can be made contingent on x. We write $R = \emptyset$ if a lender is unwilling to make any offer to x-borrowers.

Borrowers then decide which lender's offer to accept, if any, based on the selection of offers they receive, and potentially also based on private information about their own circumstances. Without explicitly modelling borrowers' preferences, we define $g(x, R) \in [0, 1]$ as the proportion of x-borrowers who prefer a mortgage at rate R to remaining without a mortgage. We assume that all borrowers have a preference for lower interest rates. Therefore, $g(x, R)$ is decreasing in R. When indifferent between several offers, borrowers select a lender randomly to break ties.

Lenders are risk-neutral. Their cost of capital is ρ and the repayment they can recover in default is $\hat{L} = \gamma\delta V$ as discussed above. Lenders have a common belief that the probability of default by a borrower with characteristics x, who accepts a mortgage at interest rate R, is $P(x, R)$. As we discuss in greater detail in Appendix (A.6), $P(x, R)$ is therefore the structural probability of default conditional on acceptance by borrowers, which accounts for selection effects when borrowers have private information.

A.3.1 Lenders' profits

As in the text, the expected Net Present Value of a mortgage with rate R for the lender is

$$NPV = \frac{1}{1+\rho}\left[(1 - P(x, R))(1 + R)L + P(x, R)\hat{L}\right] - L \equiv N(x, R),$$

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:*

$$R_0 = \inf\{R | N(x, R) = 0\} \tag{11}$$

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

We can fully characterize equilibrium as follows:

Lemma 2 *If $N(x,R) < 0$ for all R such that $g(x,R) > 0$, then no x-borrowers obtain a mortgage with positive probability in equilibrium. Conversely, if $N(x,R) \geq 0$ and $g(x,R) > 0$ for some R, then all x-borrowers are offered credit and the unique accepted equilibrium rate is $R(x) = R_0$, defined as in Equation (5).*

Proof. Consider first the case where $N(x,R) < 0$ for all R such that $g(x,R) > 0$. Suppose that x-borrowers accept a mortgage with positive probability. Then an individual lender whose offer is accepted with positive probability can profitably deviate by rejecting, meaning equilibrium cannot be sustained. Thus, x-borrowers do not obtain credit (one equilibrium strategy which sustains this is for all lenders to offer $R = \emptyset$ to x-borrowers).

Suppose next that $N(x,R) \geq 0$ and $g(x,R) > 0$ for some R. If all lenders reject x-borrowers in equilibrium, then an individual lender can profitably deviate by offering $R_0 + \epsilon$ and capturing the entire market. Thus, x-borrowers must be offered credit in equilibrium, and will accept only the lowest offer. If the lowest offer is $R < R_0$, then the lender offering it makes a loss and has a profitable deviation by offering $R = \emptyset$. If the lowest offer is $R > R_0$ in equilibrium, then an individual lender can deviate by offering $R_0 + \epsilon$, poach the entire market, and strictly increase her profits. Hence, the unique equilibrium rate is R_0 as required. ■

A.4 Discussion of endogenous contracting terms

In our model, lenders' Net Present Value depends on contracting terms beyond the interest rate. In particular, equation (10) makes clear that the NPV depends on the loan-to-value ratio (LTV) at origination. Under different assumptions about recovery rates in default, NPV could further depend on loan size (L) or other details of the mortgage contracts.

We have so far assumed that all contract characteristics except for the mortgage interest rate are pre-determined. In this section of the 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, which can be characterized as one interest rate $R(h, x)$ (or possibly rejection) for each possible contract $h = \{L,\text{LTV}\}$, given observable characteristics x.

Given a menu $R(h, x)$, let $\pi_h(h|x)$ 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 $\pi_x(x)$ be the population distribution of x.

In any equilibrium, the proportion of borrowers obtaining a mortgage across the population is

$$C = \int \int 1\{R(h, x) \neq \emptyset\}\pi_h(h|x)\pi_x(x)dhdx$$

and the average mortgage rate conditional on obtaining credit is

$$\bar{R} = C^{-1} \int \int 1\{R(h, x) \neq \emptyset\}R(h, x)\pi_h(h|x)\pi_x(x)dhdx$$

From the population of potential borrowers, we can obtain an estimate $\hat{\pi}_x(x)$ of the distribution of exogenous characteristics x. We also obtain an estimate $\hat{\pi}_h(h|x)$ of the conditional empirical distribution of contract characteristics given exogenous characteristics. We then assume that this is an unbiased estimate of the choice function $\pi_h(h|x)$ specified above:

$$\hat{\pi}_h(h|x) = \pi_h(h|x) + \varepsilon$$

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.

A.5 Estimating Lifetime Default Rates

In our empirical work, we estimate the cumulative probability of default up to a time period post-loan issuance of 36 months. We denote this estimate as $\hat{p}(.)$. We do so using both standard as well as machine learning models over our sample period, and do so in order to maintain comparability in modelling across cohorts of issued loans, with a view to using data up until the present.

This generates the need for further modelling, as the appropriate input into the NPV computations is the lifetime cumulative default probability on the loan. This section of

the appendix discusses how we use the Standard Default Assumption (SDA) curve[40] in combination with our estimated three year cumulative probabilities of default to estimate the lifetime cumulative probability of default.

Let $h(t)$ represent the default hazard on a loan. The SDA curve has a piecewise linear hazard rate, which linearly increases to a peak value h_{max} at t_1, stays there until t_2, then decreases linearly to a floor value h_{min} at t_3, staying at that level until the terminal date of the loan T.

Formally:

$$h(t) = \begin{cases} \frac{h_{max}}{t_1} t, & 0 \leq t \leq t_1 \\ h_{max}, & t_1 < t \leq t_2 \\ h_{max} - (t - t_2)\frac{h_{max} - h_{min}}{t_3 - t_2}, & t_2 < t \leq t_3 \\ h_{min} & t_3 < t < T \end{cases}$$

SDA sets $t_1 = 30$, $t_2 = 60$, $t_3 = 120$ months, $h_{max} = 0.6\%$, $h_{min} = 0.03\%$.

We assume that the hazard rates of the mortgages in our sample can be expressed as multiples M of $h(t)$, i.e., as a scaled version of the same basic SDA shape. Using this assumption, we back out M from our empirically estimated 3-year cumulative default rates $\hat{f}$, and then the resulting lifetime hazard profile to calculate the cumulative default probability over the life of the mortgage. In particular, we can map scaled hazard rates to a cumulative default probability $P(t)$ as:

$$P(t) = 1 - \exp\left[-MH(t)\right]$$

where

$$H(t) = \int_0^t h(t)dt$$

The $\hat{p}(\hat{t})$ that we measure is the cumulative probability of default up to $\hat{t} = 36$, i.e. up to just past the peak of hazard rates. We therefore assume that $\hat{t} \in (t_1, t_2)$, meaning that:

$$\begin{aligned} \hat{p} = P(\hat{t}) &= 1 - \exp\left[-M\left(\int_0^{t_1} \frac{h_{max}}{t_1} t dt + \int_{t_1}^{\hat{t}} h_{max} dt\right)\right] \\ &= 1 - \exp\left[-M\left(h_{max}\left(\hat{t} - \frac{t_1}{2}\right)\right)\right] \end{aligned}$$

[40] This was originally introduced by the Public Securities Association – see Andrew K. Feigenberg and Adam S. Lechner, "A New Default Benchmark for Pricing Nonagency Securities," Salomon Brothers, July 1993.

Rearranging, we can therefore express M as:

$$M = -\frac{1}{h_{max}} \frac{\log(1-\hat{p})}{\hat{t} - \frac{t_1}{2}}.$$

Having found M, we then find the lifetime cumulative default probability as:

$$\begin{aligned} P(T) &= 1 - \exp[MH(T)] \\ &= 1 - \exp\left[\frac{1}{h_{max}} \frac{\log(1-\hat{p})}{\hat{t} - \frac{t_1}{2}} H(T)\right] \\ &\equiv P_T(\hat{f}) \end{aligned} \tag{12}$$

where $H(T)$ is just a constant determined by T and the SDA:

$$\begin{aligned} H(T) &= \int_0^{t_1} \frac{h_{max}}{t_1} t dt + \int_{t_1}^{t_2} h_{max} dt + \int_{t_2}^{t_3} \left(h_{max} - (t - t_2)\frac{h_{max} - h_{min}}{t_3 - t_2}\right) dt + \int_{t_3}^{T} h_{min} dt \\ &= \frac{h_{min}}{2}(2T - t_2 - t_3) + \frac{h_{max}}{2}(t_2 + t_3 - t_1). \end{aligned}$$

We then simply substitute equation (12) into equation (10) and proceed.

A.6 Identification and Estimation

Structural Relationship

Assume that each borrower i has a *potential* default response $y_i(R)$ for every potential environment R. $y_i(R)$ is the *structural* relationship between the environment R and behavior – more concretely, we can think of $y_i(R)$ as the probability of default, given interest rate R, in an optimizing model of borrower behavior (for example, Campbell and Cocco, 2015).

When estimating probabilities of default, a competitive lender facing a new borrower with characteristics x needs to know the sufficient statistic:

$$E[y_i(R)|x_i = x] \equiv p(x, R)$$

for every R on a grid. When lenders know $p(x, R)$, they can mechanically translate it into

NPV values on the grid using procedures of the sort that we have outlined above, and into an equilibrium price. This $p(x, R)$ is the structural mapping from x and R to behavior that must be identified in order for us to make progress on evaluating counterfactuals.

Identification Problem

We do not observe counterfactual pricing and acceptance decisions. This leads to (at least) two selection problems. First, for each i we only observe one potential response $y_i(R_i)$, the one associated with the interest rate R_i that was actually assigned to borrower i in the data, but we cannot observe $y_i(R)$ when $R \neq R_i$. This is what we term the "intensive margin" problem. Second, if a borrower is not granted a mortgage by lenders in the data, we do not observe her at all, leading to what we term the "extensive margin" problem.[41]

Because of these issues, we cannot measure $p(x, R)$. We only observe its empirical counterpart:

$$E[y_i(R)|x_i = x, R_i = R] \equiv \tilde{p}(x, R),$$

which differs from $p(x, R)$ whenever the assignment of R_i to borrowers is not random, so that there is information about potential outcomes in the conditioning event $R_i = R$.[42]

For any given statistical technology, the econometrician (in order to approximate a counterfactual lender) must therefore solve two problems:

1. *Identification*: Find a situation in which $\tilde{p}(x, R) = p(x, R)$.
2. *Estimation*: Guess the (potentially nonlinear) population function $\tilde{p}(x, R)$ from finite data using both standard and machine learning techniques.

No Selection on Unobservables Permits Identification

The standard assumption permitting identification is conditional independence, i.e., given observable borrower characteristics x_i, the treatment (interest rate) R_i is drawn independently of potential outcomes $y_i(R)$, for all potential R:

$$R_i \perp y_i(R)|x_i, \forall R$$

[41] Indeed, we face another selection problem. We do not observe borrowers that were granted a mortgage but turned down the offer. We begin by assuming that every offer that is made is accepted, focusing initially on selection by lenders. We then return to selection by borrowers at the end of our discussion.

[42] The event $R_i = R$ is a double condition meaning "borrower is accepted, and offered R", reflecting the two counterfactuals we do not observe.

Under this strong assumption, identification follows as:

$$\begin{aligned} p(x,R) &= E[y_i(R)|x_i = x] \\ &= \sum_{R'} Pr[R_i = R']E[y_i(R)|x_i = x, R_i = R'] \\ &= \sum_{R'} Pr[R_i = R']E[y_i(R)|x_i = x, R_i = R] \\ &= E[y_i(R)|x_i = x, R_i = R] = \tilde{p}(x,R) \end{aligned}$$

In the third line, we use $E[y_i(R)|x_i = x, R_i = R'] = E[y_i(R)|x_i = x, R_i = R]$, since by conditional independence, $E[y_i(R)|x_i = x, R_i = R'] = E[y_i(R)|x_i = x] = E[y_i(R)|x_i = x, R_i = R]$.

To operationalize this assumption, suppose that we can find segments of the credit market under consideration in which lenders base their credit acceptance and rate-setting decisions based only on observables x_i. We can also allow lenders to differ in their preferences using a (potentially random) parameter η_i. We can then fully characterize lenders' behavior as $Pr[R_i = R|x_i] = g(x_i, \eta_i)$, for some *deterministic* function $g(.)$. Then, if we can assume that lender preferences η_i are independent of borrower behavior, conditional independence holds. More formally, conditional independence holds when, for all possible variables z_i that affect borrower behavior, we have $Pr[R_i = R|x_i, z_i] = Pr[R_i = R|x_i]$. But this is trivially true when $Pr[R_i = R|x_i] = g(x_i, \eta_i)$, as long as η_i is independent of z_i.

A natural sufficient condition for identification is therefore selection on observables: If lenders have no information that correlates with determinants of borrower behavior other than x_i, then default predictions are identifiable, even when counterfactual lending and pricing decisions are not observed. In our empirical work, we restrict our analysis to government sponsored enterprise (GSE) securitized mortgages, as they are far less likely to suffer from selection on unobservable borrower characteristics.

A.6.1 Selection by borrowers

The discussion can be made more general in a world with borrowers that can accept or reject offers that are made to them. We let $a_i(R) \in \{0, 1\}$ be a dummy for whether borrower i accepts an offer with mortgage rate R. Now the object of interest for the competitive lender is

$$E[y_i(R)|x_i = x, a_i(R) = 1] \equiv p_a(x, R).$$

Again, the observable counterpart is

$$E[y_i(R)|x_i = x, R_i = R, a_i(R) = 1] \equiv \tilde{p}_a(x, R).$$

To get identification in this context, we must slightly modify the conditional independence assumption. Assume that conditional on x_i, the treatment R_i is independent of both the borrower's default decision $y_i(R)$ and her acceptance decision $a_i(R)$, for every potential R. Then identification is achieved because:

$$\begin{aligned} p_a(x,R) &= E[y_i(R)|x_i = x, a_i(R) = 1] \\ &= \sum_{R'} Pr[R_i = R']E[y_i(R)|x_i = x, R_i = R', a_i(R) = 1] \\ &= \sum_{R'} Pr[R_i = R']E[y_i(R)|x_i = x, R_i = R, a_i(R) = 1] \\ &= E[y_i(R)|x_i = x, R_i = R] = \tilde{p}(x,R) \end{aligned}$$

Again the proof hinges on the third line, which uses conditional independence to argue that $E[y_i(R)|x_i = x, R_i = R', a_i(R) = 1] = E[y_i(R)|x_i = x, a_i(R) = 1]$.

Estimation

If conditional independence, and therefore identification holds, we might still face challenges in estimating counterfactuals. One obvious potential issue is sparse data. For example, suppose that borrowers with $FICO < 500$ are *always* rejected in the data. Then, even though estimation for $FICO > 500$ is unbiased, we cannot meaningfully make predictions or simulate equilibrium for $FICO < 500$, unless we permit extrapolation this group from predictions for higher-FICO borrowers. In an ideal world, we would have "full support", i.e., that the density

$$Pr[x_i = x, R_i = R] > 0$$

for all values of x and R that are used in equilibrium computation. We use an ex-post method here, i.e., we estimate $f(x,R)$, simulate equilibrium, and check whether the data is dense over the range of x and R that lenders consider in simulated equilibrium.

A.7 Adjusting Empirical Estimates to Match Causal Estimates

As we discuss above, if there is no selection on unobservables, this is sufficient for identification. We therefore restrict our analysis to the segment of GSE loans, which are less likely to suffer from selection on unobservables. However, it is still possible that the GSE loans in the sample are not completely immune to concerns about selection on unobservables. We therefore implement an additional adjustment to our estimates to account for this possibility.

Our approach is to use a recently proposed causal estimate of the sensitivity of default rates to interest rates R due to Fuster and Willen (2017), who use downward rate resets of hybrid adjustable-rate mortgages to estimate the sensitivity of default probabilities to changes in rates. Using the same dataset as they do (non-agency hybrid ARMs), we estimate a (non-causal) cross-sectional sensitivity of 3-year default probabilities to a 50 basis point change in the interest rate spread at origination (SATO), using the same hazard model as they use for their causal estimates. When we compare the resulting non-causal estimate to their causal estimates, we find that it is 1.7 times as large. We therefore adopt the factor $b = \frac{1}{1.7}$ as a measure of bias in our non-causal estimates estimated using GSE loans, assuming that the bias on 3-year default sensitivities estimated for the FRMs in our sample is the same as the one estimated using the non-agency hybrid ARMs. We have reason to believe that this adjustment is quite conservative, since the non-causal estimate comes from defaults occurring in the first-three years – this is more likely to comprise the segment of interest-rate sensitive borrowers.

How do we implement the bias adjustment on our estimates? First, as is standard in the literature, let us consider default intensities as a Cox proportional hazard model, with hazard rate:

$$h(t|R) = h_0(t)\exp(\phi R)$$

abstracting from other determinants of default for clarity. Here, $h_0(t)$ is the baseline hazard, and $\exp(\phi R)$ is the dependence of the hazard on the loan interest rate.

We can integrate the hazard function to get the cumulative hazard over the lifetime of the mortgage:

$$H(T|R) = H_0(T)\exp(\phi R).$$

The survival function (the cumulative probability of no default) is therefore:

$$\begin{aligned} S(R) &= e^{-H(T|R)} \\ &= (S_0)^{\exp(\phi R)} \end{aligned}$$

where $S_0 = e^{-H_0(T)}$, and therefore:

$$\phi = \frac{\partial \log\left(-\log\left(S(R)\right)\right)}{\partial R}$$

The cumulative probability of default is $P(R) = 1 - S(R)$, which is what we input into our NPV calculations. Now suppose that we have estimates of the lifetime cumulative probability of default on a grid of interest rates $\{R^{(0)}, ..., R^{(n)}\}$. Let the predicted probability at $R^{(j)}$ be $\hat{P}^{(j)}$. We define the transformation:

$$\Lambda(P) = \log\left(-\log(1-P)\right)$$

Note that this transformation is invertible with $P = \Lambda^{-1}(\Lambda) = 1 - e^{-e^{\Lambda}}$. Computationally, we will avoid taking the log of numbers close to zero by using:

$$\Lambda(P) = \log\left(-\log(1-P) + \epsilon\right)$$

where ϵ is a small number, and the inverse $\Lambda^{-1}(\Lambda) = 1 - e^{\epsilon - e^{\Lambda}}$.

We know that our estimates imply a sensitivity $\hat{\phi}$ which is biased, i.e., we can assume that the true parameter is $b\hat{\phi}$, where b measures the bias as discussed above. We need to debias the estimates to arrive at the appropriate cumulative probabilities of default. Our goal is therefore to adjust the $\hat{P}^{(j)}$ into corrected estimates $P^{(j)}$ by accounting for this bias.

The procedure below can be run separately for each borrower, with different implied $\hat{\phi}$ for each borrower and each interest rate level. Therefore, this method can preserve nonlinearities and interactions between interest rates and borrower characteristics in the estimated $\hat{P}^{(j)}$. We only assume that the proportional bias in estimated sensitivities b is constant. Under that assumption, we can derive a simple adjustment algorithm.

To build intuition, consider increasing interest rates from $R^{(j)}$ by moving up one notch on the grid, for a single borrower. Our estimates give an implicit $\hat{\phi}$ for this step

$$\hat{\phi}^{(j)} = \frac{\Delta \log\left(\log\left(S(R)\right)\right)}{\Delta R} = \frac{\Lambda\left(\hat{P}^{(j+1)}\right) - \Lambda\left(\hat{P}^{(j)}\right)}{R^{(j+1)} - R^{(j)}} \tag{13}$$

The true probabilities, on the other hand, satisfy

$$\phi = \frac{\Lambda\left(P^{(j+1)}\right) - \Lambda\left(P^{(j)}\right)}{R^{(j+1)} - R^{(j)}} = b\hat{\phi}^{(j)}$$

so that the true sensitivites are described by

$$\begin{aligned}\Lambda\left(P^{(j+1)}\right) - \Lambda\left(P^{(j)}\right) &= b\left(R^{(j+1)} - R^{(j)}\right)\hat{\phi}^{(j)} \\ &= b\left[\Lambda\left(\hat{P}^{(j+1)}\right) - \Lambda\left(\hat{P}^{(j)}\right)\right],\end{aligned}$$

where the last line uses equation (13).

Now suppose we know that the non-causal estimate at some $R^{(j)}$ is equal to the causal estimate. Then $P^{(j)} = \hat{P}^{(j)}$ and we can solve for a corrected estimate at $R^{(j+1)}$:

$$\Lambda\left(P^{(j+1)}\right) = b\Lambda\left(\hat{P}^{(j+1)}\right) + (1-b)\Lambda\left(P^{(j)}\right). \tag{14}$$

Equation (14) shows that, given the starting value, the corrected Λ's are an exponentially weighted moving average of the estimated Λ's. The smoothing parameter is a simple function of the bias, $(1-b)$.

Similarly, consider moving down one notch to $R^{(j-1)}$. Then we have

$$\Lambda\left(P^{(j)}\right) - \Lambda\left(P^{(j-1)}\right) = b\left[\Lambda\left(\hat{P}^{(j)}\right) - \Lambda\left(\hat{P}^{(j-1)}\right)\right]$$

and assuming that our estimate at $R^{(j)}$ is correct, we get $P^{(j)} = \hat{P}^{(j)}$ and the corrected estimate for $R^{(j-1)}$ is

$$\Lambda\left(P^{(j-1)}\right) = b\Lambda\left(\hat{P}^{(j-1)}\right) + (1-b)\Lambda\left(P^{(j)}\right)$$

We can therefore implement the bias adjustment using a simple recursive algorithm, namely:

- Assume that for one level of the interest rate, say $R^{(B)}$, our non-causal estimates equal the causal estimates (in our empirical implementation, we assume that this is true at the mean rate in the population, i.e., at a SATO of 0).
- Transform all $\hat{P}^{(j)}$ into $\hat{\Lambda}^{(j)} = \hat{\Lambda}(P^{(j)})$. Initialize the corrected $\Lambda^{(B)} = \hat{\Lambda}^{(B)}$ at the base rate.
- Then execute two loops:
  1. Forward adjustment: For $j = B, B+1, ..., n-1$
     - Calculate corrected estimate at interest rate $R^{(j)}$ as

$$\Lambda^{(j+1)} = b\hat{\Lambda}^{(j+1)} + (1-b)\Lambda^{(j)}$$

  2. Backward adjustment: For $j = B, B-1, ..., 1$
     - Calculate corrected estimate at interest rate $R^{(j)}$ as

$$\Lambda^{(j-1)} = b\hat{\Lambda}^{(j-1)} + (1-b)\Lambda^{(j)}$$

- Finally, transform corrected estimates back to probabilities $P^{(j)} = \Lambda^{-1}\left(\Lambda^{(j)}\right)$.

A.8 Descriptive Statistics, Equilibrium Sample

We show descriptive statistics for the 2011 sample in Table A-1. The table simply confirms that the patterns that are evident in the broader set of summary statistics are also evident

for this subsample.

Table A-1: **Descriptive Statistics, 2011 Originations.**

Group		FICO	Income	LoanAmt	Rate (%)	SATO (%)	Default (%)
	Mean	738	124	266	4.32	-0.09	0.43
Asian	Median	775	107	240	4.38	-0.05	0.00
(N=101,369)	SD	146	76	148	0.58	0.50	6.52
	Mean	720	93	167	4.58	0.13	1.85
Black	Median	743	77	139	4.62	0.20	0.00
(N=43,204)	SD	122	63	108	0.56	0.50	13.46
	Mean	724	91	179	4.54	0.09	0.91
White hispanic	Median	757	74	150	4.50	0.11	0.00
(N=68,567)	SD	139	65	113	0.56	0.49	9.50
	Mean	737	111	199	4.43	-0.00	0.68
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	8.23
Native Am, Alaska,	Mean	724	99	195	4.50	0.05	0.94
Hawaii/Pac Isl	Median	760	83	166	4.50	0.11	0.00
(N=9,890)	SD	150	68	122	0.56	0.49	9.65
	Mean	736	120	221	4.46	0.00	0.76
Unknown	Median	772	100	185	4.50	0.07	0.00
(N=172,970)	SD	142	79	141	0.56	0.49	8.68

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. Default is defined as being 90 or more days delinquent at some point over the first three years after origination. Data source: HMDA-McDash matched dataset of fixed-rate mortgages originated in 2011.

Figure A-2 shows the cumulative distribution functions of the differences between the pds produced by the different models, restricted to the loans in the equilibrium sample. It shows that the patterns are very similar to those evident in the full sample.

A.9 Imputation procedure for FICO in HMDA data

We calculate the population frequency $\eta(FICO, Y|L, LTV)$, 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:

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

We can calculate the weights conditional on acceptance, $\eta(FICO, Y|L, LTV, \mathcal{A} = 1)$, directly from the merged HMDA-McDash sample. We then obtain the frequency of acceptance $\eta(\mathcal{A} = 1)$ 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 $\eta(\mathcal{A} = 0)$ is the proportion of borrowers with flag 3 (Application denied by financial institution). We normalize these frequencies so that $\eta(\mathcal{A} = 1) + \eta(\mathcal{A} = 0) = 1$.

We impute the weights conditional on rejection, $\eta(FICO, Y|L, LTV, \mathcal{A} = 0)$, 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:

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

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

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

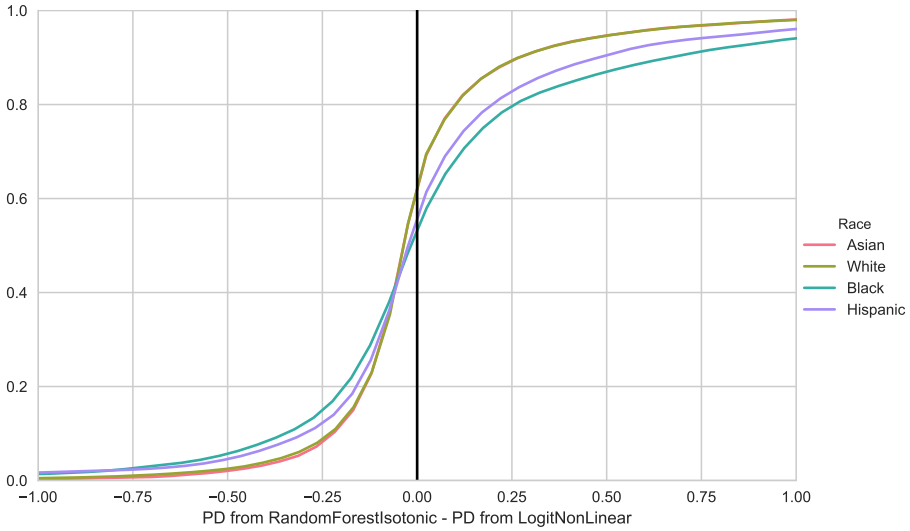
   where $Q_F \in (0, 1)$ is a parameter measuring the degree of adjustment. Our baseline figures are based on $Q_F = 0.3$.

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:

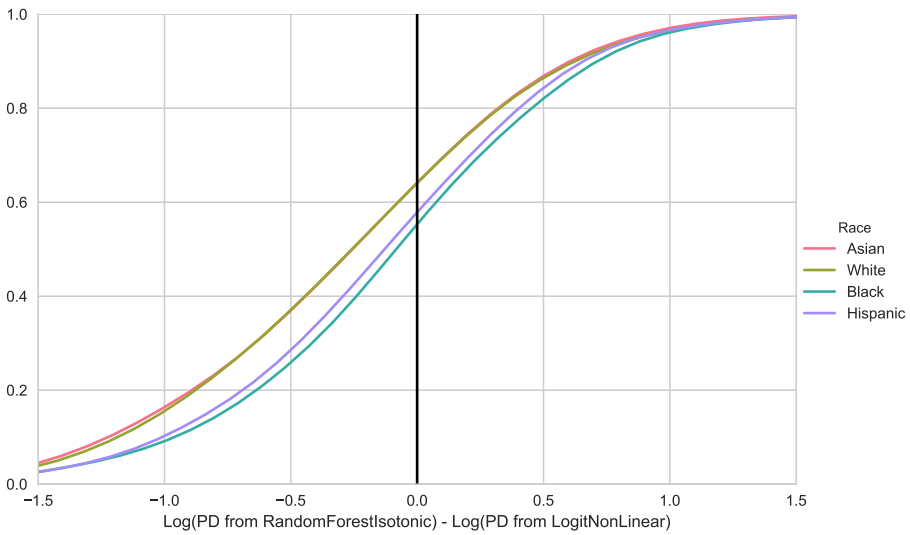
$$\eta(FICO, Y|L, LTV, \mathcal{A} = 1) = \eta(Y|L, \mathcal{A} = 1)\eta(\hat{F}, Y|L, LTV, \mathcal{A} = 0) \tag{15}$$

Given these assumptions, we obtain the imputed frequencies conditional on rejection according to equation (15), where we get the first factor $\eta(Y|L, \mathcal{A} = 1)$ from the HMDA (sub)sample of rejected borrowers, and the second factor $\eta(\hat{F}, Y|L, LTV, \mathcal{A} = 0)$ from the HMDA-McDash sample with adjusted FICO scores.

Figure A-2: **Comparison of Predicted Default Probabilities, Equilibrium Sample.**



Panel A



Panel B