Predictably Unequal?
The Effects of Machine Learning on Credit Markets

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

This draft: November 2018

[1] Fuster: Swiss National Bank. Email: andreas.fuster@gmail.com. Goldsmith-Pinkham: Yale School of Management. Email: paulgp@gmail.com. Ramadorai: Imperial College London and CEPR. Email: t.ramadorai@imperial.ac.uk. Walther: Imperial College London. Email: ansgar.walther@gmail.com. We thank Philippe Bracke, Jediphi Cabal, John Campbell, Andrew Ellul, Kris Gerardi, Andra Ghent, Johan Hombert, Ralph Koijen, Andres Liberman, Gonzalo Maturana, Karthik Muralidharan, Daniel Paravisini, Jonathan Roth, Jann Spiess, Jeremy Stein, Johannes Stroebel, and Stijn Van Nieuwerburgh for useful conversations and discussions, and seminar participants at the NBER Household Finance Summer Institute, CEPR European HHF conference, Adam Smith Corporate Finance Workshop, Pre-WFA Summer Real Estate Research Symposium, SFS Cavalcade, FIRS, SITE (Financial Regulation), Banking and Financial Regulation conference at Bocconi University, Southern Finance Association conference, Boston College, Imperial College Business School, NYU Stern, University of Rochester, Queen Mary University of London, USI Lugano, University of Southampton, Office for Financial Research, Bank of England, ECB, Federal Reserve Bank of Boston, Riksbank, UK Financial Conduct Authority, Engineer's Gate, and Quantum Black for comments. We also thank Kevin Lai, Lu Liu, and Qing Yao for research assistance. Fuster and Goldsmith-Pinkham were employed at the Federal Reserve Bank of New York while much of this work was completed. The views expressed are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of New York, the Federal Reserve System, or the Swiss National Bank.

Abstract

Innovations in statistical technology, including in predicting creditworthiness, have sparked concerns about differential impacts across categories such as race. Theoretically, distributional consequences from better statistical technology can come from greater flexibility to uncover structural relationships, or from triangulation of otherwise excluded characteristics. Using data on US mortgages, we predict default using traditional and machine learning models. We find that Black and Hispanic borrowers are disproportionately less likely to gain from the introduction of machine learning. In a simple equilibrium credit market model, machine learning increases disparity in rates between and within groups; these changes are primarily attributable to greater flexibility.

1 Introduction

In recent years, new predictive statistical methods and machine learning techniques have been rapidly adopted by businesses seeking profitability 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 gains arising from better statistical modeling 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 a structural model 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, or "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, 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 consider the decision of a lender who uses a single exogenous variable (e.g., a borrower characteristic such as income) to predict default. 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.

[2] See, for example, Agrawal et al. (2018). Academic economists also increasingly rely on such techniques (e.g., Belloni et al., 2014; Varian, 2014; Kleinberg et al., 2017; Mullainathan and Spiess, 2017; Chernozhukov et al., 2017; Athey and Imbens, 2017).

[3] See, for example, O'Neil (2016), Hardt et al. (2016), Kleinberg et al. (2016), and Kleinberg et al. (2018).

[4] Academic work applying machine learning to credit risk modeling includes Khandani et al. (2010) and Sirignano et al. (2017).

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 decreasing in the characteristic becomes available to this economy, groups with lower values of the characteristic than the mean will clearly be penalized following the adoption of the new technology, while those with higher 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.

We then extend this simple theoretical intuition, noting two important mechanisms through which such unequal effects could arise. To begin with, we note that default outcomes can generically depend on both "permissible" observable variables such as income or credit scores, as well as on "restricted" variables such as race or gender. As the descriptors indicate, we consider the case in which lenders are prohibited from using the latter set of variables to predict default, but can freely apply their available technology to the permissible variables.

One possibility is that the additional *flexibility* available to the more sophisticated technology allows it to more easily recover the structural relationships connecting permissible variables to default outcomes. Another possibility is that the structural relationship between permissible variables and default is perfectly estimated by the primitive technology, but the more sophisticated technology can more effectively *triangulate* the unobserved restricted variables using the observed permissible variables. In the latter case, particular groups are penalized or rewarded based on realizations of the permissible variables, which are more accurately combined by the more sophisticated technology to estimate the influence of the restricted variables.

Our theoretical work is helpful to build intuition, but 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, 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] Armed with the intuition from our simple models, we therefore go to the data to understand the potential effects of machine learning on an important credit market, namely, the US mortgage market. We rely on a large administrative dataset of close to 10 million US mortgages originated between 2009 and 2013, in which we observe borrowers' race, ethnicity, and gender, as well as mortgage characteristics and default outcomes.[6] We estimate a set of increasingly sophisticated statistical models to predict default using these data, 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).[7]

We confirm that the machine learning technology delivers significantly higher out-of-sample predictive accuracy for default than the simpler logistic models. However, we find that predicted default propensities across race and ethnic groups look very different under the more sophisticated technology than under the simple technology. In particular, while a large fraction of borrowers belonging to the majority group (e.g., White non-Hispanic) gain, 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 some minority race and ethnic groups (e.g., Black and Hispanic borrowers).

We propose simple empirical measures to try to bound the extent to which flexibility or triangulation is responsible for these results, by comparing the performance of the naïve and

[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.

[6]We track default outcomes for all originated loans for up to three years following origination, meaning that we follow the 2013 cohort up to 2016.

[7]We also employ the eXtreme Gradient Boosting (XGBoost) model (Chen and Guestrin, 2016), which delivers very similar results to the Random Forest. We therefore focus on describing the results from the Random Forest model, and provide details on XGBoost in the online appendix.

sophisticated statistical models when race and ethnicity are included and withheld from the information set used to predict default. We find that the majority of the predictive accuracy gains from the more sophisticated machine learning model are attributable to the increased flexibility of the model, with 8% or less attributable to pure triangulation. This finding suggests that simply prohibiting certain variables as predictors of default propensity will likely become increasingly ineffective as technology improves.[8] While in some measure this is due to the ability of nonlinear methods to triangulate racial identity,[9] the main effect seems to arise from the fact that such regulations cannot protect minorities against the additional flexibility conferred by the new technology.

How might these changes in predicted default propensities across race and ethnic groups 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 when granted mortgages? To evaluate these questions, we embed the statistical models in a simple equilibrium model of credit provision in a competitive credit market.[10]

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.

[8]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).

[9]We also find that the machine learning models are far better than the logistic models at predicting race using borrower information such as FICO score and income. 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).

[10]We consider a model in which lenders bear the credit risk on mortgage loans (which is the key driver of their accept/reject and pricing decisions) and are in Bertrand competition with one another. The US mortgage market over the period covered by our sample is one in which the vast majority of loans are insured by government-backed entities that also set underwriting criteria and influence pricing. This introduces some variance between our model and the current state of the market. That said, our equilibrium exercise can be viewed as evaluating the effects of the changes in default probabilities that we find on credit provision along the intensive and extensive margins. This is of interest whether new statistical techniques are used by private lenders, or by a centralized entity changing its approach to setting underwriting criteria.

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 which are backed by Fannie Mae and Freddie Mac and originated with full documentation, 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.[11] 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, 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 is predicted to provide a slightly larger number of borrowers access to credit, and to marginally reduce 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—the cross-group disparity of equilibrium rates increases under the machine learning model relative to the less sophisticated logistic regression models. This is accompanied by a substantial increase in within-group dispersion in equilibrium interest rates as technology improves. This rise is virtually double the magnitude for Black and White Hispanic borrowers under the machine learning model than for the White non-Hispanic borrowers, i.e., Black and Hispanic borrowers get very different rates from one another under the machine learning technology. For a risk-averse borrower behind the veil of ignorance, this introduces a significant penalty associated with being a minority.

Overall, the picture is mixed. On the one hand, the machine learning model is a more

[11] In influential work, Keys et al. (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.

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 to slightly reduce cross-group dispersion in acceptance rates. However, the main effects of the improved technology are the rise in the dispersion of rates across race groups, as well as the significant rise in the dispersion of rates within race groups, especially for 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.[12] That said, our exercise is similar in spirit to this work, in the sense that we also seek a clearer understanding of the sources of inequality in household financial markets.[13] Our work is also connected more broadly to theories of statistical discrimination,[14] though we do not model lenders as explicitly having access to racial and ethnic information when estimating borrowers' default propensities.

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, and describes the two sources (flexibility and triangulation) of unequal effects. 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, describes how predicted default probabilities vary across groups, and computes measures of flexibility and triangulation in the data. Section 5 sets up our equilibrium model of credit provision under different technologies, and discusses how the changes in default predictions affect both

[12] Bartlett et al. (2017) study empirically whether "FinTech" mortgage lenders in the US appear to discriminate more across racial groups. Buchak et al. (2017) and Fuster et al. (2018) study other aspects of FinTech lending in the US mortgage market.

[13] These issues have been a major focus on work in household financial markets. In mortgages and housing, see, e.g., Berkovec et al. (1994, 1998), Ladd (1998), Ross and Yinger (2002), Ghent et al. (2014), and Bayer et al. (2017). In insurance markets, see, e.g., Einav and Finkelstein (2011), Chetty and Finkelstein (2013), Bundorf et al. (2012), and Geruso (2016). Also related, Pope and Sydnor (2011) consider profiling in unemployment benefits use.

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

the intensive and extensive margins of credit provision. Section 6 concludes. An appendix included with the paper contains a few proofs, and a more extensive online appendix contains numerous auxiliary analyses and robustness checks.

2 A Simple Theory Framework

Consider a lender who wishes to predict the probability of default, $y \in [0, 1]$, of a loan with a vector of observable characteristics x, which includes both borrower characteristics (e.g., income, credit score) and contract terms (e.g. loan size, interest rate). We start by assuming that the lender takes the contract terms as given when drawing inferences, and study how these inferences are affected by changes in the statistical technology that the lender is 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 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.[15] 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., $\mathcal{M}_1 \subset \mathcal{M}_2$.[16]

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:

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

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

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

[16]Throughout, we focus on improvements in prediction technology given a *fixed* information set; we do not consider the use by lenders of newly available information sources, such as borrowers' "digital footprint" (Berg et al., 2018).

We 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 $\hat{P}(x|\mathcal{M}_1)$:

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

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.[17]

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 the simple case where lenders predict default based on a single variable x.

[17] In practice, machine learning algorithms trade off increases in bias against reductions in the variance of the out-of-sample forecast (see, e.g., James et al., 2013). While we do not discuss how this affects our simple theory in this section, our empirical estimates reflect this feature of the algorithms.

2.1 Unequal Effects of Better Technology

We write g for a vector of dummy variables indicating group membership (e.g., borrowers' race). We continue to assume that the lender can use only observable characteristics x for prediction, and is prohibited from using g directly as an explanatory variable. For simplicity, we suppose that the primitive technology $\mathcal{M}_1$ is the class of linear functions of x.

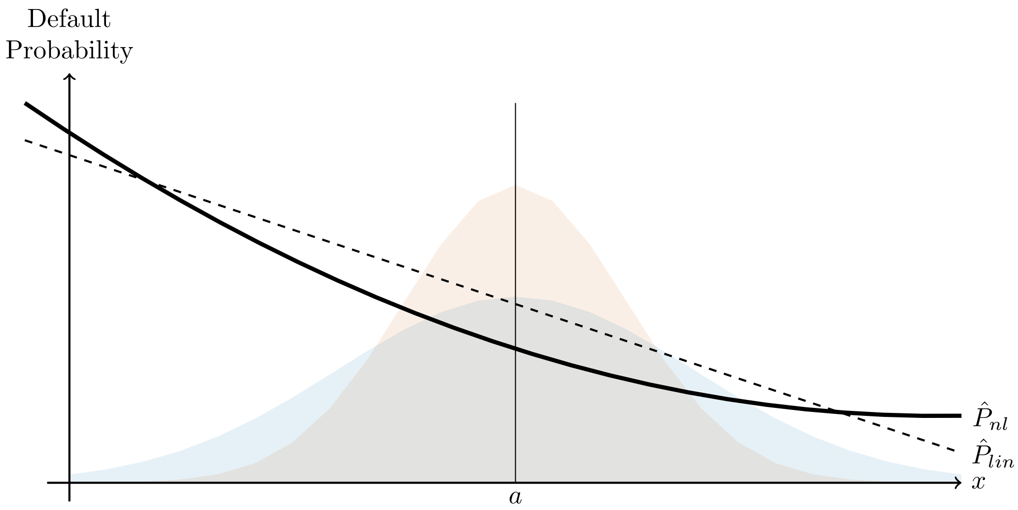
Although group membership g is excluded from prediction, better statistical technology can nevertheless have unequal impacts across groups. Figure 1 gives an example. There are two groups of borrowers: Blue and Red. The two bell curves show the distribution of characteristics x for each group. Specifically, x has the same mean a in both groups, but higher variance in the Blue group.

Suppose that the linear predictor of default $\hat{P}_{lin}(x)$ is a decreasing function of x (e.g., if x were the income or credit score of the borrower). Also suppose that the nonlinear predictor $\hat{P}_{nl}(x)$—associated with more sophisticated statistical technology—is a convex quadratic function of x. The figure shows that in this example, better technology leads to higher predicted default rates ($\hat{P}_{nl}(x) > \hat{P}_{lin}(x)$) when x is far from its mean a in either direction. It follows that Blue borrowers tend to be adversely affected by new technology. This is because their characteristics x are more variable and hence more likely to lie in the tails of the distribution, which are penalized by nonlinear technology.

This intuition about the factors determining winners and losers generalizes beyond the convex quadratic case—this is one possible example, used for illustrative purposes. In the appendix, we formalize this insight by showing that the effect of introducing a more sophisticated technology depends on two factors, namely, the higher-order moments of borrower characteristics in each group, and the higher-order derivatives of predictions under sophisticated technology.[18]

[18]For example, if the distribution of $x|g$ is right-skewed, and the third derivative of $\hat{P}_{nl}(x)$ is positive, then the introduction of $\hat{P}_{nl}(x)$ 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 $\hat{P}_{nl}(x)$ is negative, then the new predictions reward both tails of the conditional distribution,

Figure 1: **Unequal Effects of Better Technology**



2.2 Sources of Unequal Effects

To better understand the sources of unequal effects, it is instructive to consider two special cases. First, suppose that the true data-generating process for y is

$$y = P(x) + \varepsilon, \tag{2}$$

where $P(x)$ is a possibly nonlinear, deterministic function of x, and ε is independent of both x and g.

In this case, group membership g has no direct impact on default risk. Nevertheless, this situation can give rise to unequal effects when a new technology is introduced, as depicted in Figure 1.

To see this, make the additional assumption that in equation (2), the true data-generating function $P(x)$ is quadratic. The new technology, which permits quadratic functions, obviously better approximates the true function, and the estimate $\hat{P}_{nl}$ shown in the figure creates unequal effects across groups g. In this case, the *flexibility* of new technology permits it to better capture the structural relationship between x and y, and is the underlying source of

and members of g will be relatively better off, and so forth.

the unequal effects.

Second, consider another case in which the true data-generating process is linear:

$$y = \beta \cdot x + \gamma \cdot g + \varepsilon, \tag{3}$$

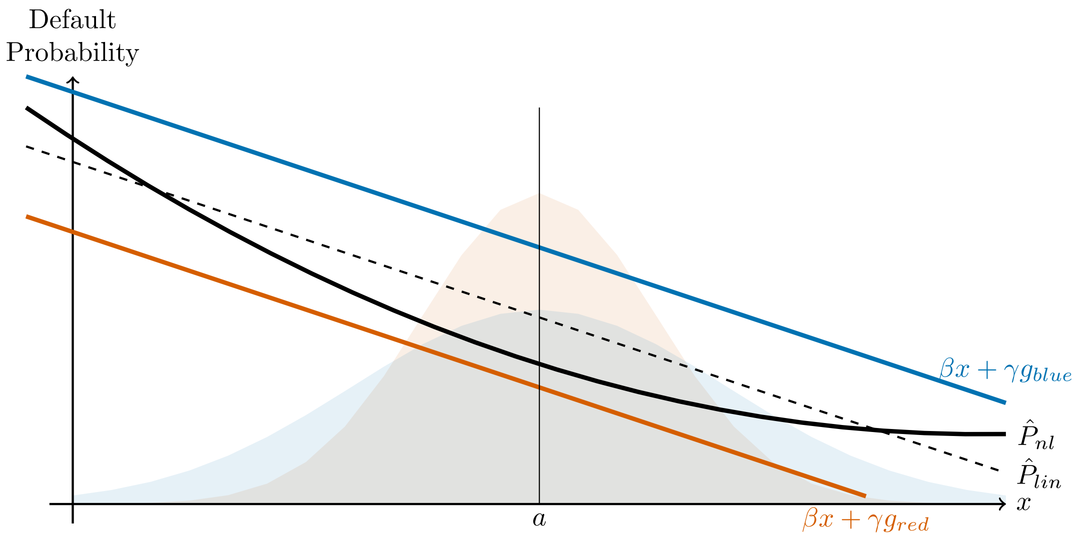
where ε is again independent of x and g. In this case, a linear model achieves the best possible prediction of default if g is available to include as an explanatory variable. The introduction of more sophisticated technology cannot by construction have an impact purely due to its flexibility.

In this case, better technology can still have an unequal impact if g is *restricted* as an explanatory variable. Intuitively, better technology can use nonlinear functions of x to more effectively capture the influence of the omitted variable g. The unequal effects of more sophisticated technology would then arise through more efficient *triangulation* of group membership.

Triangulation can give rise to unequal effects that are similar to those induced by flexibility. Figure 2 provides an example where true default risk is higher for the Blue group. In the figure, we assume that the group-conditional distributions are exactly the same as in Figure 1. Since there is no linear correlation between x and g, the linear prediction $\hat{P}_{lin}(x)$ cannot use x to capture the influence of g and, as a result, it is equal to the population-weighted average of the true Blue and Red default probabilities (the dashed straight line in the figure).

In contrast, the quadratic prediction penalizes the Blue group: since extreme realizations of x are more likely to come from Blue borrowers, the more sophisticated technology assigns higher predicted default probabilities to these extreme realizations of x than to more moderate realizations of x. We provide a formal example of this mechanism in the appendix.

Figure 2: **Triangulation**



2.3 Discussion

The main insights from this 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, we note that better technology can have unequal effects across borrower groups even if lenders are not allowed to include group membership g in predictive modeling. One way this might occur is through the additional flexibility of the new technology to uncover nonlinear, structural relationships between observable characteristics and default rates. Another is that better technology can triangulate unobservable group membership using nonlinear functions of x. Finally, the precise impacts will be jointly determined by the shape of the underlying distribution of $x|g$, and the differences in shape between the new and old functional forms of the predictive equations.

From a positive perspective, Figures 1 and 2 highlight that flexibility and triangulation can both result in similar observed unequal effects. The distinction between them is still important, however, from a normative perspective. For example, if default is truly independent of g, unequal impacts of the introduction of more flexible technology could still arise despite the presence of regulations prohibiting the use of g in prediction. These would simply be a product of an improved ability to uncover structural relationships between permitted observables and y. In contrast, if default is in reality associated with g, unequal impacts could arise from a better ability of the technology to triangulate the omitted variable. These two scenarios would result in a very different set of conversations—triangulation might lead us to consider alternative regulations that are fit for purpose when lenders use highly nonlinear functions, whereas flexibility might push us in a different direction, towards discussing the underlying sources of cross-group differences in the distributions of observable characteristics.

We have thus far considered two special cases: in the pure flexibility case, group membership g had no predictive power for default conditional on the ability to estimate sufficiently flexible functions of x, while in the pure triangulation case, nonlinear functions of x were entirely unable to add predictive power conditional on g. In reality, we should consider a more general data generating process which permits both flexibility and triangulation to generate unequal effects on groups. In Section 4, we define empirical measures of flexibility and triangulation to attempt to bound the extent to which these two sources drive unequal effects observed in the data.

A 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 at this stage predict how changing probabilities of default translate into changes in interest rates, or exclusion of some borrowers from the credit market. We return to this issue in detail following the next section.

Finally, it is worth re-emphasizing that the intuition we have developed using specific functional forms (e.g., convex quadratic) could well be misleading in terms of the true patterns that exist in the data. For example, it could be that the new technology allows a lender a greater ability to identify good credit risks within a minority group previously assigned uniformly high predicted default rates under the old technology. If so, the new technology would *benefit* the minority group on average, though dispersion of outcomes within the group would rise in this case.[19]

[19] Anecdotally, the credit card company CapitalOne more efficiently used demographic information and expanded lending in just this way during the decade from 1994 to 2004. See, for example, Wheatley (2001).

Ultimately, 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 given the significant non-linearities that machine learning allows us to uncover, and their likely complex interactions with the underlying moments of the group-specific multivariate distributions of characteristics. 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.

3 US Mortgage Data

To study how these issues may 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™ 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.[20] 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 et al., 2012).

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

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

information on the cost of a loan (this is only reported for "high cost" loans).[21]

The McDash™ 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 et al., 2010; Foote et al., 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.[22] 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

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

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

some additional sample restrictions. We only retain conventional (non-government issued) fixed-rate first-lien mortgages on single-family and condo units, with original loan term of 10, 15, 20, or 30 years. We furthermore only keep loans with original LTV between 20 and 100 percent, 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 more 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 in the category of 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.[23] 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 (as well as interest rates and the

[23]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.

spreads at origination over average interest rates, known as "SATO") for the Black and White Hispanic borrowers. They also have substantially higher variance in FICO scores than the White Non-Hispanic group. Intuitively, such differences in characteristics make these minority populations look 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	764	122	277	4.24	-0.07	0.42
Asian	Median	775	105	251	4.25	-0.05	0.00
(N=574,812)	SD	40	74	149	0.71	0.45	6.49
	Mean	735	91	173	4.42	0.11	1.88
Black	Median	744	76	146	4.50	0.12	0.00
(N=235,673)	SD	58	61	109	0.71	0.48	13.57
	Mean	746	90	187	4.36	0.07	0.99
White Hispanic	Median	757	73	159	4.38	0.07	0.00
(N= 381,702)	SD	52	63	115	0.71	0.47	9.91
	Mean	761	110	208	4.33	-0.00	0.71
White Non-Hispanic	Median	774	92	178	4.38	0.02	0.00
(N=7,134,038)	SD	45	73	126	0.69	0.44	8.37
Native Am, Alaska,	Mean	749	97	204	4.39	0.04	1.12
Hawaii/Pac Isl	Median	761	82	175	4.45	0.04	0.00
(N=59,450)	SD	51	65	123	0.70	0.46	10.52
	Mean	760	119	229	4.38	0.00	0.79
Unknown	Median	773	100	197	4.50	0.02	0.00
(N=984,310)	SD	46	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.[24] As a result, the lenders retain originated loans in portfolio (i.e., on balance sheet) and thus directly bear the risk of default for less than 10% of the loans in our sample.

As we discuss later in the paper, when we study 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, which is important for identification. Nevertheless, in this section of the paper, we estimate default probabilities using both GSE-securitized and portfolio loans, in the interests of learning 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-based groups in the population of mortgage borrowers.

4 Estimating Probabilities of Default Using Different Statistical Technologies

In this section, we describe the different prediction methods that we employ to estimate $\hat{P}(.)$, the three-year probability of default for originated mortgages in the US mortgage dataset. We subsequently use these estimated default probabilities to understand the impact of different statistical technologies on mortgage lending.[25]

[24] For instance, in addition to their flat "guarantee fee" (i.e., insurance premium), the GSEs charge so-called "loan-level price adjustments" that depend on borrower FICO score, LTV ratio, and some other loan characteristics.

[25] 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

First, we implement two Logit models to approximate the "standard" prediction technology typically used by both researchers and industry practitioners (e.g. Demyanyk and Van Hemert, 2011; Elul et al., 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, the main machine learning model that we consider is a Random Forest model (Breiman, 2001); we use cross-validation and calibration to augment the performance of this model.[26]

Relative to the simple theoretical analysis considered earlier, we make an important change. We include the interest rate at loan origination (as SATO) in the set of covariates used to predict default. That is, we estimate $\hat{P}(x, R)$. At this point, we essentially treat this variable (and indeed, other contract characteristics) the same as all of the other right-hand side variables, and conduct reduced-form estimation assuming there is some noise in initial contract assignment that generates predictive power over and above observable loan and borrower characteristics (we confirm there is indeed incremental predictive power conferred by the use of this variable, as we describe in the online appendix).[27]

4.1 Logit Models

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

$$\log\left(\frac{g(x)}{1-g(x)}\right) = x'\beta. \tag{4}$$

online 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.

[26] We also employ the eXtreme Gradient Boosting (XGBoost) model (Chen and Guestrin, 2016), which delivers very similar results to the Random Forest. We therefore relegate our description of this model to the online appendix.

[27] Later in the paper, we consider the case of endogenously assigned contract characteristics and embed the estimated $\hat{P}(x, R)$ functions in a simple NPV model to structurally evaluate effects on interest rates and acceptance decisions. We also describe later how we correct our estimates of the default sensitivity to interest rates. We do so to bring these reduced form estimates more in line with causal estimates of this sensitivity to facilitate the evaluation of effects within our structural model.

We estimate the model in two ways, which vary how the covariates in x enter the right-hand-side. In the first model, all of the variables in x (listed in Table 2) enter linearly. Additionally, we include dummies for origination year, documentation type, occupancy type, product type, investor type, loan purpose, coapplicant status, and a flag for whether the mortgage is a "jumbo" (meaning the loan amount is too large for Fannie Mae or Freddie Mac to securitize the loan). In addition, we include the term of the mortgage, and state fixed effects. We refer to this model simply as "Logit".

Table 2: **Variable List**

Logit	*Nonlinear 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 600 to 850;) separate dummy for FICO<600)
(with dummy variables for missing values)	
Common Covariates	
Spread at Origination "SATO" (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.

In the second type of Logit 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 include 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 that the relationship is nonlinear. In particular, we assign LTV to bins of 5% width ranging 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 600 to 850 (the maximum). We assign all FICO values between 300 (the minimum) and 600 into a single bin, since there are only few observations with such low credit scores. Finally, we bin income into US $25,000 width bins from 0 to US $500,000. We refer to the resulting model as the "Nonlinear Logit".

4.2 Tree-Based Models

As an alternative to the traditional models, we use 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. For one, the models tend to be nonparametric. 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 nonparametric and nonlinear 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.[28]

[28] For a more in-depth discussion of tree-based models applied to a default forecasting problem see, for example, Khandani et al. (2010).

The Random Forest approach can best be understood in two parts. First, a simple decision tree is estimated by recursively splitting covariates (taken one at a time) 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, for subsequent evaluation on an omitted subset of out-of-sample *test* data. An even more important differentiating factor is that the process can flexibly 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., sampled with replacement) from 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 (i.e., unseen/out-of-sample) 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 (tuning) parameters must be chosen in the estimation of the Random Forest model. Common parameters of this nature include, for example, the maximum number of leaves that the model is allowed to have in total, and the minimum number of data points needed in a leaf in order to proceed with another split. In order to ensure the best possible fit, a common approach is to cross-validate the choice of parameters using K-fold cross-validation. This involves randomly splitting the training sample into K folds or sub-samples (in our case, we use $K = 3$).[29]

For each of the data folds, we estimate the model using a given set of tuning parameters on the remaining folds of the data (i.e., the remaining two-thirds of the training data in our setting with $K = 3$). We then check the fit of the resulting model on the omitted K-th data fold. The procedure is then re-done K times, and the performance of the selected set of tuning parameters is averaged across the folds. The entire exercise is then repeated for each point in a grid of potential tuning parameter values. Finally, the set of parameters that maximize the out-of-sample fit in the cross-validation exercise are chosen. In our application, we cross-validate over the minimum number of data points needed to split a leaf, and the minimum number of data points required on a leaf.[30] Our procedure selects a minimum number of observations to split of 200 and requires at least 100 observations on each leaf.

[29]The choice of the hyperparameter K involves a trade-off between computational speed and variance; with a smaller K, there is more variance in our estimates of model fit, as we will have less observations to average over, while with larger K, there will be a tighter estimate at the cost of more models to fit. As our Random Forest model is computationally costly to estimate with 500 trees, to balance these considerations, we choose $K = 3$ to select tuning parameters.

[30]We define our grid from 2 to 500 in increments of 50 (2, 50, 100, etc.) for the minimum number of data points needed to split (*min_samples_split*), and a grid from 1 to 250 in increments of 50 for the minimum number of data points on a leaf (*min_samples_leaf*).

4.2.1 Translating Classifications into Probabilities

An important difference between the Random Forest model and the Logit models is that the latter naturally produce estimates 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. For many purposes, including credit evaluation, the *probability* of belonging to a class (i.e., 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 predicted class labels into predicted loan default probabilities.

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 use an approach called "calibration," in which noisy estimated probabilities are refined/smoothed by fitting a monotonic function to 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").[31] In our empirical work, 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.

[31] 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.

4.2.2 Estimation

As mentioned earlier, we first estimate both sets of models (the two Logit versions and the Random Forest) 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.

We also further split the training sample into two subcomponents. 70% of the training sample is used as a *model* sample on which we estimate the Logit and Nonlinear Logit models, and train the Random Forest model. We dub the remaining 30% of the training data the *calibration* sample, and use it to estimate the isotonic regression to construct probabilities from the predicted Random Forest class labels as described above. This ensures that both sets of models have the same amount of data used to estimate default probabilities.[32]

4.3 Model Performance

We evaluate the performance of the different models on the test set in several ways. We plot Receiver Operating Characteristic (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; e.g., 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).[33]

One drawback of the AUC is that it is less informative in datasets which are sparse in

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

[33] The TPR is the fraction of true defaulters in the test set that are also (correctly) predicted to be defaulters, and the FPR is the fraction of true non-defaulters in the test set (incorrectly) predicted to be 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.

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 $P(\hat{y} = 1|y = 1)$,[34] and draw Precision-Recall curves which plot Precision against Recall for different probability thresholds.

Two additional measures we compute are the Brier Score and the R^2. Brier Score is calculated as the average squared prediction error. Since this measure captures total error in the model, a smaller number is better, unlike the other metrics. One useful feature of the Brier Score is that it can be decomposed into three components:

$$n^{-1}\sum_{n}(\hat{P}(x_i) - y_i)^2 = \underbrace{n^{-1}\sum_{k=1}^{K} n_k(\hat{y}_k - \bar{y}_k)^2}_{\text{Reliability}} - \underbrace{n^{-1}\sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2}_{\text{Resolution}} + \underbrace{\bar{y}(1-\bar{y})}_{\text{Uncertainty}},$$

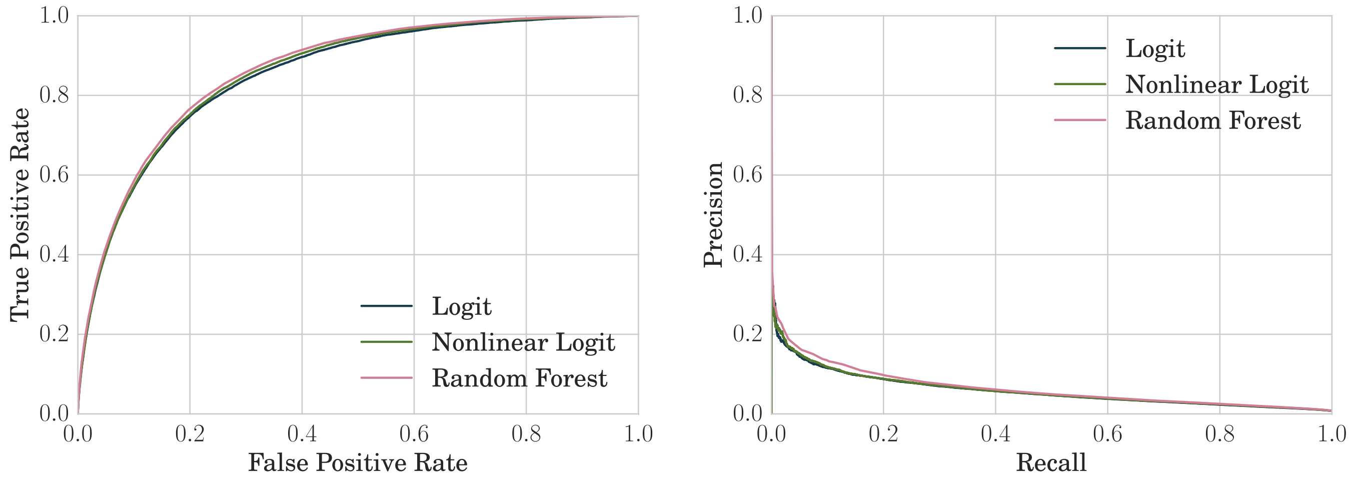
where the predicted values are grouped into K discrete bins, $\hat{y}_k$ is the predicted value within the kth bin, and $\bar{y}_k$ is the true mean predicted value within the kth bin. Uncertainty is an underlying feature of the statistical problem, Reliability is a measure of the model's calibration, and Resolution is a measure of the spread of the predictions. A larger resolution number is better, while a smaller reliability number implies a smaller overall error. It is worth noting that in our application, the overall uncertainty is 0.00725, and tends to dominate the overall value of the Brier Score.

Finally, the R^2 is a well-known metric, calculated as one minus the sum of squared residuals under the model, scaled by the sum of squared residuals from using the simple mean. This gives a simple interpretation as the percentage share of overall variance of the left-hand-side variable explained by a model.

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. Both figures 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.

[34]Note that the *Recall* is equal to the TPR.

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



Panel A: ROC   Panel B: Precision-Recall

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, third, fifth, and seventh columns of Table 3 confirm that the metrics are indeed greater for the Random Forest model than for either Logit model, 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. Indeed, the Random Forest outperforms the Nonlinear Logit model by 5.4% in terms of average precision, 0.8% in terms of AUC, and 15.3% in terms of R^2. The Brier Score, as discussed, is dominated by the overall uncertainty of the outcome, 0.00725. Once that is subtracted, the change from Nonlinear Logit to Random Forest (-0.000104 to -0.000136) is substantial. When we decompose this change into reliability and resolution, we find that the gains from switching to Random Forest in reliability are large, with a 3000% increase, but at the cost of a decrease in resolution, with a roughly 30% decline.[35]

In order to verify that these differences are indeed statistically significant, we use bootstrapping. We hold fixed our estimated models, and randomly resample with replacement

[35]This result is consistent with Figure A-1 in the online appendix, where we see significantly more spread in the predictions of the Logit model, but far worse calibration.

Table 3: **Performance of Different Statistical Technologies Predicting Default**

	ROC AUC		Precision Score		Brier Score × 100		R^2	
Model	(1) No Race	(2) Race	(3) No Race	(4) Race	(5) No Race	(6) Race	(7) No Race	(8) Race
Logit	0.8522	0.8526	0.0589	0.0592	0.7172	0.7171	0.0245	0.0246
Nonlinear Logit	0.8569	0.8573	0.0598	0.0601	0.7146	0.7145	0.0280	0.0281
Random Forest	0.8634	0.8641	0.0630	0.0641	0.7114	0.7110	0.0323	0.0329

Note: Performance metrics of different models. For ROC AUC, Precision score, and R^2, higher numbers indicate higher predictive accuracy; for Brier score, lower numbers indicate higher accuracy. In odd-numbered columns, race indicators are not included in the prediction models; in even-numbered columns, they are included.

from the original test dataset to create 500 bootstrapped sample test datasets. We then re-estimate the performance metrics for all of the models on each such bootstrapped sample. The Random Forest AUC is greater than that of the Nonlinear Logit in 100% of these bootstrap samples, with an average improvement of 0.7 percent; the corresponding Precision score increases in 97.6% of the bootstrap samples, with an average improvement of 5.2 percent; the Brier score improves in 100% of the samples, with an average improvement of 0.4 percent; and the R^2 improves in 100% of the samples, with an average improvement of 15 percent.[36] Overall, we can conclude with considerable statistical confidence that the machine learning models significantly improve default prediction performance.

4.3.1 Model Performance With and Without Race

The second, fourth, sixth, and eighth columns of Table 3 show that the inclusion of race has positive effects on the performance of all three models. This suggests that even the more sophisticated machine learning model benefits from the inclusion of race as an explanatory variable. Moreover, the gain from adding race to the Random Forest model is greater than that obtained from adding it to the Logit models, which is not surprising given the ability

[36]The histograms across bootstrapped datasets of the difference in these scores between the Random Forest and the Nonlinear Logit models are shown in Figure A-6 in the online appendix.

of the Random Forest to compute interactions between any included variable and all of the other variables in the model.

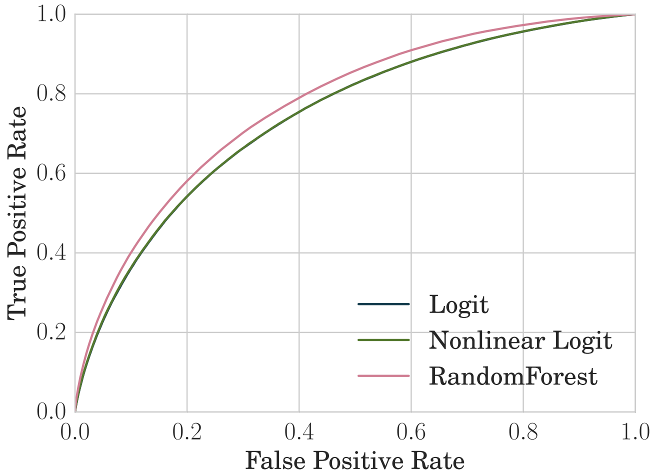
That having been said, it is worth noting that while all models benefit from the inclusion of race, the improvement is quite small relative to that conferred by the increased sophistication of the model that is used. For example, when going from the simple Logit to the Random Forest model, there is an increase in R^2 that dwarfs any improvement obtained from adding race as a variable to any of the models.

Evaluating changes in the 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 such as race, assessments of borrower risk should be colorblind. The fact that race appears to marginally augment model performance suggests that there is still some sense in which this restriction might be helpful. Importantly, even though the performance improvement magnitudes are small overall, this does not mean that the race indicators do not have significant effects on some groups—for instance, average default probabilities in the Nonlinear Logit increase from 0.016 to 0.019 for Black borrowers when race indicators are added, while they decrease for Asian borrowers from 0.006 to 0.004.

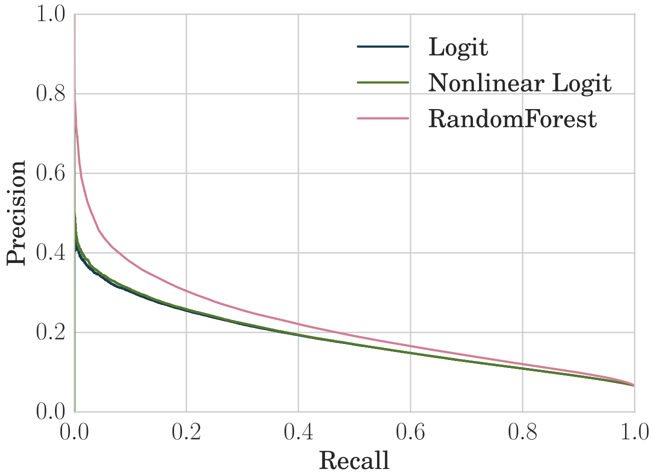
To explore this issue further, we employ the models to predict whether or not a borrower is Hispanic or Black using the same set of variables used to predict default. This exercise reveals striking differences between the models, especially in Panel B of Figure 4. Table 4 confirms that the Random Forest outperforms the other two models, which have very similar scores, by 6.9% in terms of average precision, 0.6% in terms of AUC, 2% in terms of Brier score, and 30.7% in terms of R^2. Put differently, the machine learning model is better able to predict the racial and ethnic identities of borrowers using observable characteristics. Whether this ability contributes to triangulation will depend on whether there is considerable variation in true default propensities across race and ethnic groups that is not non-linearly related to observable characteristics, as in our simple example (equation (3)). We explore this issue more comprehensively when we compute estimates of triangulation and flexibility.

Next, we document how estimated probabilities of default from these models vary across

Figure 4: **ROC and Precision-Recall Curves of Predicting Race**



Panel A



Panel B

Table 4: **Performance of Different Statistical Technologies Predicting Race**

Model	ROC AUC	Precision Score	Brier Score × 10	R^2
Logit	0.7478	0.1948	0.5791	0.0609
Nonlinear Logit	0.7485	0.1974	0.5783	0.0622
Random Forest	0.7527	0.2110	0.5665	0.0813

Note: Performance metrics of different models. For ROC AUC, Precision score, and R^2, higher numbers indicate higher predictive accuracy; for Brier score, lower numbers indicate higher accuracy.

race groups in US mortgage data.

4.4 Differences in Predicted Default Propensities

Figure 5 illustrates the potential impact of new technology on different borrower groups in our sample. Each panel plots predicted default propensities as a function of borrower income on the horizontal axis, and FICO score on the vertical axis. The figure shows the level sets of predicted default probabilities for the Nonlinear Logit model in the top two panels, and for the Random Forest in the bottom two panels. We hold constant other borrower characteristics.[37] These level sets are overlaid with a heatmap illustrating the empirical density of income and FICO levels among minority (Black and White Hispanic) borrowers in the right panels, and White non-Hispanic, Asian, and other borrowers in the left panels, with darker colors representing more common characteristics in the respective group.

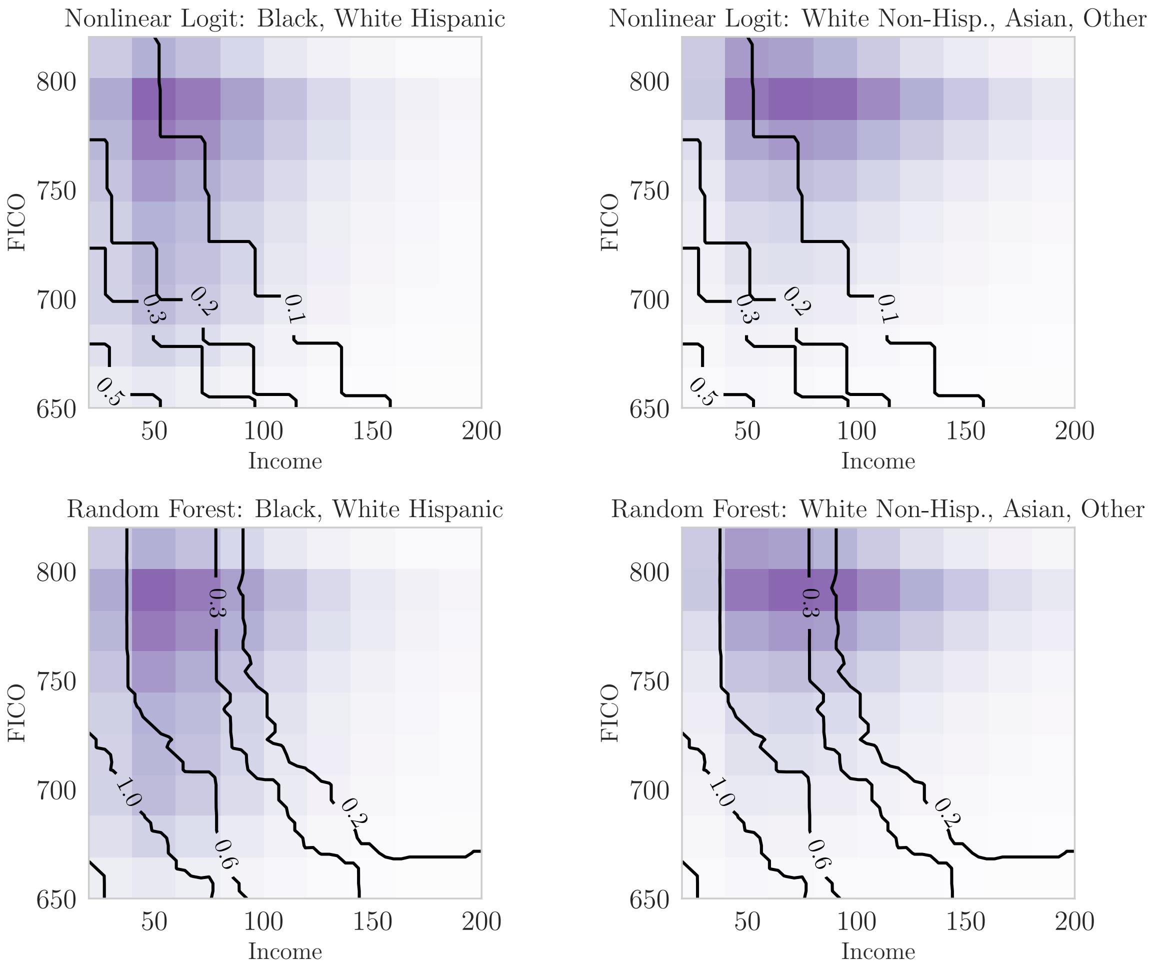
The figure shows that the level sets of the Random Forest predicted default probabilities are highly nonlinear and markedly different from those of the Nonlinear Logit. Moreover, they increase more sharply as we move into the low-income, low-FICO region in the bottom left corner of each chart. The heatmaps show that these regions are also relatively more densely inhabited by minority borrowers.

This graphical example is essentially an extension of our theoretical example in Figure 1 to two dimensions. It suggests, for the restricted sample that we consider in these plots, that new technology has unequal impacts across racial groups. The plot is, however, quite restricted, as it simply shows the effect of varying income and FICO scores, holding all other characteristics constant. To assess more rigorously who wins and who loses from the introduction of new technology, we analyze the change in the entire distribution of predicted default propensities in our test sample as estimation technology varies.

In Figure 6, we look at the differences between the predicted probabilities of default (PDs) from the machine learning model and the traditional Logit model. Panel A of the figure shows the cumulative distribution function (cdf) of the difference in the estimated default probability (in percentage points) between the Random Forest and Nonlinear Logit. Borrowers for whom this difference is negative (i.e., to the left of the vertical line) are "winners"from the new technology, in the sense of having a lower estimated default probability, and those with a positive difference (those to the right of the vertical line) are "losers". For each level of difference in the PDs across the two models listed on the x-axis, the y-axis shows the cumulative share of borrowers at or below that level; each line in the plot shows this cdf for a different race/ethnic group. Panel B plots the log difference in PDs to highlight

[37]Specifically, we vary income and FICO for 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 securitized through Fannie Mae.

Figure 5: **Example of Predicted Default Probabilities Across Models**



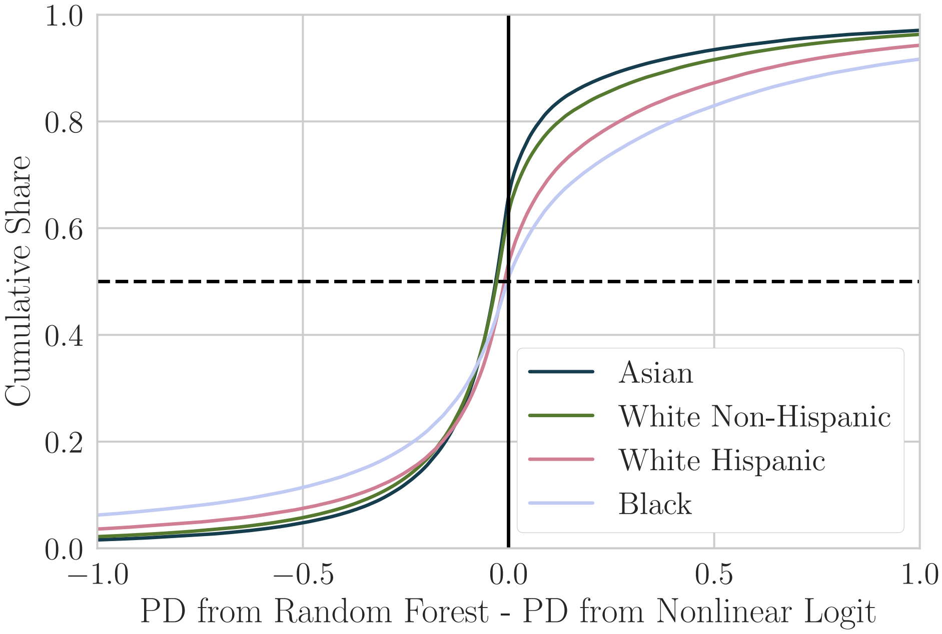
Note: Figure shows level sets of default probabilities (in %) predicted from different statistical models for different values of borrower income and FICO (holding other characteristics fixed as explained in text). Nonlinear Logit predictions are shown in top row; Random Forest predictions in bottom row. Underlying heatmaps show distribution of borrowers within certain race/ethnicity groups: Black and White Hispanic in left column; White Non-Hispanic and Asian in right column.

the *proportional* benefit for each group.[38]

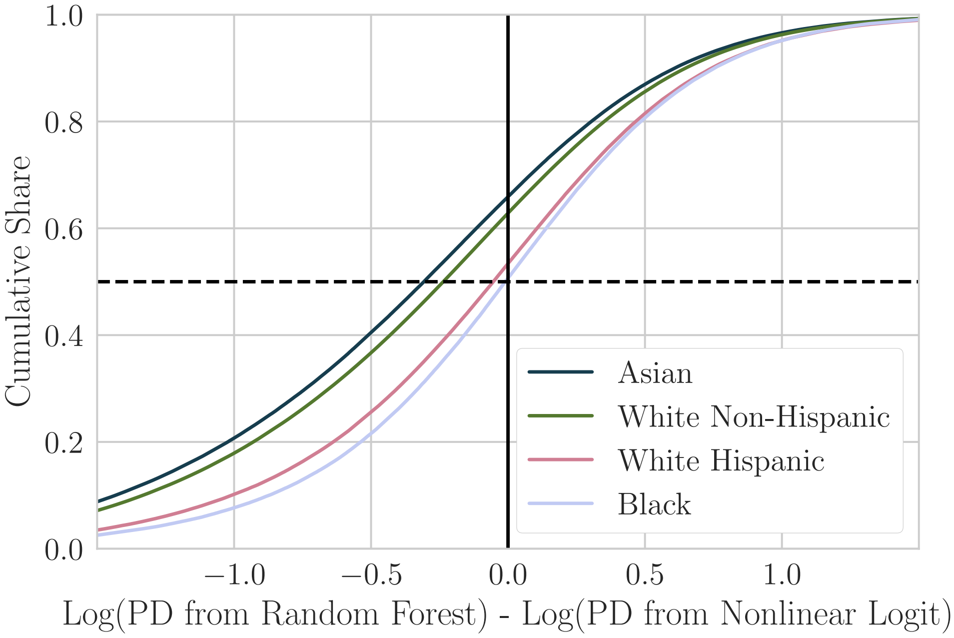
Both panels show that there is a reduction in default risk under the Random Forest model for the median borrower (indicated by the dashed horizontal lines) in the population as a whole. In fact, the plot shows that for all of the groups, the share of borrowers for whom

[38] For ease of visual representation, we have truncated the x-axes on these plots, as there is a small fraction of cases for which the estimated differences in the default probabilities are substantial.

Figure 6: **Comparison of Predicted Default Probabilities Across Models, by Race Groups**



Panel A



Panel B

the estimated probability of default drops under the new technology is above 50%.

However, the main fact evident from this graph are the important differences between race groups in the outcomes arising from the new technology. Panel B makes evident 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 roughly 10 percentage points higher than for the Black and White Hispanic populations, within which there are roughly equal fractions of winners and losers. Furthermore, the entire distributions of relative PD differences are shifted to the north-west for the White non-Hispanic and Asian borrowers relative to minority (Black and White Hispanic) borrowers. This means that there are fewer minority borrowers that see large proportional reductions in predicted default probabilities when moving to the Random Forest model, and more minority borrowers that see large proportional increases.

A further important feature evident especially in Panel A is that for these minority groups, the distribution of predicted default probabilities from the Random Forest model has larger variance than under the Nonlinear Logit model.[39] We return to this finding later.

Figure A-2 in the online appendix shows the same plots replacing the predictions from the Random Forest model with those from the XGBoost machine learning model. The qualitative conclusions are robust to the change of machine learning technique: when moving to the XGBoost model from the Nonlinear Logit model, there are more winners among the White non-Hispanic and Asian borrowers than among the Black and White Hispanic groups.

These figures provide useful insights into the questions that motivate our analysis, and suggest that the improvements in predictive accuracy engendered by the new prediction technology are accompanied by an unequal distribution of the winners and losers across race groups. However, to make further progress, we need to better understand the sources of these unequal effects in the data.

[39]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.

4.5 Flexibility and Triangulation in the Data

In Section 2, we argued that better statistical technology can generate differential effects on different groups in the population arising from increased *flexibility* to learn nonlinear combinations of characteristics that directly predict default, and/or from an enhanced ability to use these nonlinear combinations to *triangulate* hidden variables such as race. In this section, we propose and compute simple empirical measures to gauge the relative importance of these effects in the data.

This is a non-trivial task because in some cases, flexibility and triangulation are observationally equivalent. For example, suppose that the true default probability is $y = f(x) + \varepsilon$ for a nonlinear function $f(x)$ of observable characteristics, but that $f(x)$ is in turn perfectly correlated with group indicators g. Then there is no meaningful distinction between a technology that flexibly estimates $f(x)$ and one that triangulates g to predict default.

Given this identification problem, we do not pursue a unique decomposition. Instead, we aim to provide bounds on the importance of flexibility and triangulation. Consider the performance of the Nonlinear Logit and the Random Forest models reported in the second and third row of Table 3. If we take the Nonlinear Logit without race as a baseline scenario, the greatest increase in predictive power relative to this model will clearly be achieved by simultaneously employing a better technology *and* adding race, i.e., the Random Forest with race. For example, for the R^2 reported in the final two columns of that table, this performance improvement is about 17.5% ($\frac{0.0329-0.0280}{0.0280}$). Our goal is to bound the extent of this total improvement arising from the inclusion of race variables, relative to that arising from the possibility of including nonlinear functions of x (i.e., better technology).

As in any partial regression exercise, there are two possible ways to decompose the explanatory power, depending on the order in which variables enter. We report the results of both ways in Table 5. In Panel A, we add race controls first, fixing the Nonlinear Logit as the statistical technology. The left column reports the percentage of the overall increase in performance that can be achieved by adding race controls without changing technologies. For example, around 2% of the total performance improvement of 17.5% (i.e., roughly 35 basis points) in terms of R^2 arise from the inclusion of race dummies as covariates in the Nonlinear Logit model.[40] The right column of Panel A shows the complement of this fraction, which we interpret as the fraction of the total performance improvement attributable to increased flexibility, conditional on the improvement achieved by simply adding race. For example, moving from the Nonlinear Logit model with race to the Random Forest model with race delivers roughly 98% of the 17.5% improvement in R^2 (i.e., 17.2%).

In Panel B, we add new technology *first*, fixing x (without including race) as the vector of explanatory variables. The left column shows the fraction of the overall improvement that is achieved by changing technology (moving from Nonlinear Logit to the Random Forest, without including race in either model), while the left column shows its complement which is attributable to race conditional on having a flexible model (moving from Random Forest without race to Random Forest with race). A larger fraction of the improvement is attributed to race conditional on new technology than in the Logit model. This is not too surprising, as additional interactions between race and other observables are being utilized by the Random Forest. This result is consistent with the results in Table 3, and suggests that machine learning models capitalize on interactive effects between race and other characteristics.

Table 5: **Decomposition of Performance Improvement**

	Race	Technology
ROC-AUC	5.88	94.12
Precision	7.90	92.10
Brier	3.25	96.75
R^2	2.04	97.96

Panel A: Race Controls First

	Technology	Race
ROC-AUC	91.16	8.84
Precision	77.21	22.79
Brier	90.63	9.37
R^2	87.75	12.25

Panel B: New Technology First

The table yields several interesting observations. To begin with, if $f(x)$ was perfectly correlated with g—the "unidentified" case referred to earlier—the left columns in Panel A

[40]One could consider adding race dummies to the Nonlinear Logit in a more flexible manner, for instance by interacting them with other borrower/loan characteristics. We found that doing so "naïvely" tends to reduce out-of-sample predictive accuracy, due to overfitting. Using regularization methods such as LASSO might partly alleviate this problem, but would blur the distinction between the simple traditional technology and machine learning.

and Panel B would both show 100%, meaning that it would be impossible to tell whether the predictive improvements that we find arise from flexibility or triangulation. This is clearly not the case in our empirical estimates, which consistently imply that a larger share of the increase in accuracy stems from the more flexible technology than from the inclusion of race dummies. This strongly suggests that when predicting mortgage default, triangulation alone is not at the heart of the performance improvements from machine learning.

Indeed, the numbers in the first column of Panel A suggest that knowing race (which is the best that triangulation without additional flexibility could achieve) would yield at most 8% of the total performance improvement. Note that Panel B is not as informative about the share of the overall improvement that is attributable to flexibility, since the improvements generated by the move from the Logit to the Random Forest model could stem from either flexibility or triangulation. The high share of the performance improvement arising from this move simply provides an upper bound of what is achievable by having more flexibility in the model. When predicting mortgage default in our sample, we find that this upper bound is large. That said, we note that in other applications, or indeed in other samples, it may be tighter and suggest larger effects of triangulation.

The fact that unequal effects appear mainly driven by flexibility does not make them less unequal. As discussed earlier, our results potentially hold normative implications, suggesting that simply prohibiting the use of race in the estimation of default propensity may become increasingly ineffective as technology improves. While in some measure this is due to the ability of nonlinear methods to triangulate racial identity, the main effects in our empirical setting seem to arise from the fact that such regulations cannot protect minorities against the additional flexibility conferred by the new technology.

5 Equilibrium Effects of Statistical Technology

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

5.1 A Simple Model of Equilibrium

The model has two dates $t = 0, 1$. There are at least two competing lenders and a population of borrowers. The timing is as follows: At date 0, lenders simultaneously offer mortgage interest rates to borrowers based on their observable characteristics. Each borrower then chooses whether to accept this offer, potentially based on further private information. If borrowers accept, a mortgage is originated. At date 1, if a mortgage has been originated, borrowers either repay the loan (with interest) or default.

The model assumes that non-price characteristics such as the loan amount and LTV ratio are pre-determined. In reality, these parameters are often dictated, or at least confined to a narrow range, by local property prices and liquidity constraints faced by the borrower.[41]

Borrowers. Each borrower has a vector x of observable characteristics and a vector θ of privately known characteristics, drawn from a joint distribution with density $f(x, \theta)$. We describe borrowers in terms of two sufficient statistics. Let $S(x, \theta, R)$ be the surplus that borrower type (x, θ) derives from obtaining a mortgage with interest rate R at date 0,[42] and let $y(x, \theta, R) \in (0, 1)$ be the probability that this borrower defaults on the mortgage at date 1. We assume that $\partial S/\partial R < 0$ and $\partial y/\partial R > 0$. If R is the lowest interest rate offered to a borrower with observables x, then the likelihood of acceptance is

$$\alpha(x, R) \equiv Pr[S(x, \theta, R) \geq 0|x] = \int_\theta 1 \{S(x, \theta, R) \geq 0\} f(\theta|x) d\theta$$

[41] Our focus on a single price for a given contract is motivated mainly by tractability, in common with a large applied literature on insurance contracts (e.g., Einav et al., 2010). The online appendix discusses how this assumption affects our empirical implementation.

[42] A borrower's surplus is defined as difference between the borrower's maximized utility after obtaining a mortgage, and the maximized autarkic utility without a mortgage.

Moreover, if the borrower accepts the loan, then the conditional probability of default is

$$P(x,R) \equiv E[y(x,\theta,R)|S(x,\theta,R) \geq 0, x] = \frac{1}{\alpha(x,R)} \int_{\theta} 1\left\{S(x,\theta,R) \geq 0\right\} y(x,\theta,R) f(\theta|x) d\theta$$

Intuitively, $P(x,R)$ denotes the probability of default in the *subsample* of borrowers who choose to accept the lender's offer. This object is key for the lender's profitability.

Lenders. Lenders are identical, risk-neutral, and discount future cash flows at their cost of capital ρ. Lenders also have the ability to recover part of the loan in the event of default at date 1. Specifically, we assume that lenders recover a fraction γ of the home value at origination, and also further incur a deadweight cost of foreclosure equal to a fraction ϕ of the loan. The recovery value per dollar of the loan is therefore

$$\min\left\{\frac{\gamma}{LTV}, (1+R)\right\} - \phi \equiv \ell(x,R)$$

where LTV is the loan-to-value ratio at origination. Notice that, since LTV is pre-determined for each borrower in our model, we have subsumed it into the borrower's observable characteristics x.

If a lender offers a rate R to a borrower with characteristics x, and if the borrower does not have a better offer from another lender, then the Net Present Value per dollar of the loan is

$$\alpha(x,R) \cdot N(x,R),$$

where

$$N(x,R) = \frac{(1 - P(x,R))(1+R) + P(x,R)\ell(x,R)}{1+\rho} - 1 \tag{5}$$

To compute the total NPV of the loan, the lender multiplies the probability $\alpha(x,R)$ of acceptance with the NPV *conditional* on acceptance, $N(x,R)$. To obtain the conditional NPV, the lender evaluates the promised repayment $(1+R)$, as well as the default recovery value $\ell(x,R)$, weighted by the appropriate conditional probability of default $P(x,R)$.

Statistical Technology. Lenders observe their own cost of capital ρ, as well as the recovery parameters γ and ϕ. By contrast, lenders do not know borrowers' preferences $S(.)$, true default propensities $y(.)$, or the true distribution $f(.)$ of characteristics. In the absence of this information, they cannot directly calculate the Bayesian conditional probability $P(x,R)$ that determines the NPV of a loan. Instead, lenders use their statistical technology to obtain an estimate of $P(x,R)$, which we denote as $\hat{P}(x,R)$. We then define $\hat{N}(x,R)$ as the estimated NPV of a loan conditional on acceptance by the borrower; this is simply the right-hand side of Equation (5) with $\hat{P}(x,R)$ substituted for $P(x,R)$.[43]

Equilibrium Prices. With lenders in Bertrand competition, the equilibrium interest rate offered to borrowers with characteristics x is the lowest interest rate that allows lenders to break even in expectation, given their statistical estimates. More rigorously, if there exist values $R \geq 0$ such that $\alpha(x,R) > 0$ and $\hat{N}(x,R) \geq 0$, then x-borrowers are accepted in equilibrium and offered the lowest rate satisfying these conditions.[44] Note that there can be combinations of characteristics x for which no break-even interest rate exists. Indeed, if $P(x,R)$ is increasing in R, the increased promised repayment as R increases may be more than offset by increased default risk. Borrowers for whom this is the case are not accepted for a loan.

[43] An alternative approach is to estimate a full structural model of borrower characteristics and behavior (e.g., Campbell and Cocco, 2015), and then map its parameters into predicted default rates $\hat{P}(x,R)$. Practitioners usually rely on reduced form models for prediction (see, e.g., Richard and Roll, 1989; Fabozzi, 2016, for mortgage market applications), which tends to achieve better predictive outcomes than structural modeling (e.g., Bharath and Shumway, 2008; Campbell et al., 2008). We therefore posit that lenders take this approach.

[44] Generically, this is the unique Bertrand-Nash equilibrium. If a lender offered anything other than the lowest break-even rate, other lenders would undercut him and make positive profits. The only pathological case is where $N(x,R)$ is tangent to 0 at the smallest break-even rate. In this case, there can be multiple equilibria. We ignore this case because it only applies for knife-edge parameter values.

5.2 Identification and Parameter Choices

5.2.1 Identification

To take the model to the data, we need to identifiy structural default probabilities. Concretely, we use our sample from the HMDA-McDash data, which contain a vector x_i of observable characteristics, a realized interest rate R_i, and a default outcome y_i for mortgages indexed by $i = 1, ..., N$. We then take the predicted default probabilities $\hat{P}(x, R)$ from different statistical technologies, which we have summarized above.[45] Finally, we simulate our equilibrium model assuming that lenders rely on these predictions. Clearly, this is only reasonable if these predictions are well-identified estimates of the underlying default probabilities $P(x, R)$ that affect the lenders' NPV in the model.

In order to derive conditions for identification, we assume that the data are generated according to the following "potential outcomes" model. First, each borrower has a vector θ_i of unobservable characteristics, which determines her structural propensity $y(x_i, \theta_i, R)$ of default and her perceived surplus $S(x_i, \theta_i, R)$ from accepting a mortgage. These are the borrower's *potential outcomes* and are defined for every possible interest rate $R \geq 0$.[46] For every observation i, the variables (x_i, θ_i) determining borrower behavior and the interest rate R_i offered by the lender are an independent draw from a joint distribution $F(x, \theta, R)$. Second, if the draw for observation i implies that $R_i = \emptyset$ (the borrower is not offered a mortgage), or that $S(x_i, \theta_i, R_i) < 0$ (the borrower does not accept her offer), then the mortgage is not originated and discarded from the sample. In short, the econometrician is therefore left with a select sample of borrowers who were offered a mortgage and accepted it.

The standard assumption permitting identification is conditional independence, i.e., given

[45] The 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. However, equation (5) takes as an input lifetime default rates. We therefore convert our empirical estimates into estimates of the lifetime default rate based on the Standard Default Assumptions (SDA) used in the mortgage investor community, as described in the online appendix.

[46] For concreteness, one can think of these objects as arising in an optimizing model of borrower behavior (for example, Campbell and Cocco, 2015), but this formulation can also encompass other behavioral decision rules for households.

observable borrower characteristics x_i, the treatment (interest rate offer) R_i is allocated independently of unobserved borrower types θ_i. In terms of conditional distributions, this assumptions implies that

$$F(\theta|x, R) = F(\theta|x)$$

To see why this is sufficient for identification, let $\hat{P}(x, R)$ be the sample average default rate of borrowers with observables $x_i = x$ and realized interest rate $R_i = R$ in the data. Given enough regularity, the sample average converges in probability to

$$\begin{aligned} E_\theta[y(x, \theta, R)|R_i = R, x_i = x, S(x_i, \theta, R_i) \geq 0] &= \int_\theta 1\{S(x, \theta, R) \geq 0\} y(x, \theta, R) dF(\theta|R, x) \\ &= \int_\theta 1\{S(x, \theta, R) \geq 0\} y(x, \theta, R) dF(\theta|x) \\ &\equiv P(x, R) \end{aligned}$$

Conditional independence thus ensures that the econometrician's estimate $\hat{P}(x, R)$ corresponds exactly to the probability $P(x, R)$ that enters lenders' NPV calculations in our model.

This argument clarifies the roles of selection by lenders and selection by borrowers. On one hand, *selection on observables* by lenders is a natural sufficient condition for identification of $P(x, R)$. If lenders do not base their interest rates R_i on any information that correlates with determinants of borrower behavior other than x_i, then default predictions are identifiable. On the other hand, if this assumption holds, selection on unobservables by borrowers is not a threat to identification. Borrowers' unobserved characteristics θ_i may be systematically related to the interest rates they are offered if there is either adverse or advantageous selection. But this issue does not prevent identification of $P(x, R)$, which is the average probability of default in the *select sample* of borrowers who would accept. Finally, the above derivation clarifies that identification is possible only for combinations (x, R) that actually occur in the data. Thus, we cannot extrapolate our inferences to borrowers with observables that have always lead to rejection in the past, nor to interest rates that are outside the support of the data.

A key concern is that selection on observables is a strong assumption in the context of mortgage markets. Specifically, our empirical estimates $\hat{P}(x, R)$ are not obtained using random variation in interest rates. Hence, we may either over- or understate the interest rate sensitivity of lenders' profits if interest rates were allocated by lenders responding to "soft" information about borrowers—which is unobservable from our perspective.

To try to address this issue, we adopt a two-pronged approach. First, when estimating default probabilities that feed into equilibrium computations, we include only GSE-insured mortgages (i.e., those securitized through Fannie Mae or Freddie Mac) which are marked as having been originated with full documentation of borrower income and assets.[47] In this segment, soft information is less likely to be important because lenders focus on whether a borrower fulfills the underwriting criteria set by the GSEs.[48]

Second, 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 (non-agency hybrid ARMs) and the same hazard model as Fuster and Willen (2017), we estimate a (non-causal) cross-sectional sensitivity of default probabilities to a 50 basis point change in the interest rate spread at origination (SATO) over the first three years of loan life (i.e., before any interest rate resets occur). When we compare the resulting non-causal estimate to their causal estimates, we find that it is roughly 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.

[47] As mentioned earlier, Keys et al. (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), and not for full-documentation loans such as the ones we consider.

[48] In the online appendix, we confirm that the sample of GSE-backed full documentation loans has similar descriptive statistics to the full sample. Moreover, the residual variation in observed interest rates is indeed lower in this subsample, consistent with the idea that soft information is less prominent in this segment. We also confirm that our equilibrium calculations do not place an undue burden of extrapolation on the estimated predictions, in the sense that we mostly consider combinations of x and R that are commonly observed in the data.

The fact that credit risk in the sample used for this exercise is borne by the GSEs rather than the originating lenders is at variance with our model assumptions. Our preferred way to think about our exercise is as an evaluation of how statistical technology might affect lending decisions in a household credit market that is primarily privately funded. While this is currently not the case in the US mortgage market, it is the case for other household credit markets in the US, and also for mortgage markets in almost all other countries. An alternative (more restrictive) interpretation of our work could be that it provides insight into how centralized (GSE) criteria might change as statistical technology improves, and how this development might affect outcomes for different borrower groups.

Since in the data, we cannot observe borrowers not granted mortgages (or at least not many of their key characteristics), we restrict our counterfactual statements to populations with distributions of characteristics identical to the one we observe. 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. However, an advantage of focusing on GSE loans, in addition to soft information playing very little role in underwriting and pricing, is that GSE policies have likely led to more loan originations than in a purely private market. This helps to reduce concerns about extensive margin biases in our exercise.[49]

5.2.2 Calibration of Parameters

We directly calibrate the basic parameters of the lender's objective function. To calibrate ρ, we assume that each quarter, the average interest rate charged by lenders is their cost of

[49] An alternative approach to address issues of potential selection along the extensive margin would be to rely on quasi-exogenous variation in the accept/reject decision, similar in spirit to Kleinberg et al. (2017)'s use of judge leniency in bail decisions. However, no such quasi-exogenous variation is available to us, partly because key borrower characteristics such as FICO are not recorded in HMDA for rejected applications.

capital plus a fixed spread of 30bp.[50] We calibrate recovery values by assuming that lenders can recover $\gamma = 0.75\%$ of the home value at origination, and further incur a deadweight cost of foreclosure equal to $\phi = 10\%$ of the loan, roughly in line with the loss severities that An and Cordell (2017) document for Freddie Mac insured loans originated post 2008.[51] In unreported calculations, we find that none of our qualitative conclusions are sensitive to the calibration of these three parameters. The online appendix describes the computational procedure we use to solve for equilibrium outcomes.

5.3 Equilibrium Results

Table 6 summarizes equilibrium lending and pricing decisions. The first two columns show the average acceptance rate for the Nonlinear Logit (NL) model and the Random Forest (RF) model. The third and fourth columns show the average spread (SATO) charged to borrowers conditional on acceptance, and the final two columns show the dispersion of spreads conditional on acceptance. 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 final row shows the standard deviation of average acceptance rates and spreads across racial groups (this is the cross-sectional spread of the group means relative to the population mean, where each group is weighted by its share in the sample).

We find that the proportion of borrowers in the population that are accepted for a mortgage increases by about 0.9 percentage points when lenders use Random Forest instead of Logit. This increase benefits all racial groups and is particularly pronounced for Black borrowers. Perhaps intuitively, the superior technology is better at screening, and is therefore more inclusive on average, and in a manner that cuts across race groups. Consistent with this intuition, the cross-group standard deviation of acceptance rates decreases for this model.

[50] This corresponds roughly to the average "primary-secondary spread" between mortgage rates and MBS yields over this period, after subtracting the GSE guarantee fee (e.g., Fuster et al., 2013).

[51] An and Cordell show an average total loss severity of roughly 0.4-0.45 of the remaining balance at the time of default, of which about a third are liquidation expenses and carrying costs. We make a small downward adjustment to these fractions since we need the loss relative to the original balance.

The average interest rate spread in the second columns falls slightly (by 2bp) for Asian borrowers and increases (by 4bp) for Black borrowers. The cross-group standard deviation of spreads increases by about 50% relative to its baseline value of 2bp under the logit model. Thus, average pricing effects are unequal across race groups, consistent with our results on predicted default probabilities in Section 4. While these effects are quantitatively modest, they are not negligible—for instance, a 6 basis point difference in the interest rate cumulates to differential interest payments over 10 years of roughly 1% of the original loan amount.

We see larger effects of the more sophisticated technology on the dispersion of interest rates in the population overall, as well as within each group (columns 5 and 6). These facts are reminiscent of our Lemma 1, in which the new technology generates predictions which are a mean-preserving spread of the older technology. Overall, the dispersion of rates increases by about 21% ($= \frac{0.360-0.298}{0.298}$)

The cross-group variation in the within-group *dispersion* of rates is also very different across the models. The breakdown by racial groups reveals that the increase in dispersion is much more pronounced for minority borrowers: the standard deviation of interest rates increases by 5bp and 6bp for Asian and White Non-Hispanic borrowers respectively, but by 8bp and 10bp for Black and White Hispanic borrowers. The *proportional* increases in within-group dispersion are also substantially larger for these minority groups. These patterns suggest that the Random Forest model screens within minority groups more extensively than the Logit model, leading to changes in intensive margin lending decisions associated with the new technology. It also suggests an important form of risk confronting White Hispanic and Black borrowers, namely that their rates are drawn from a distribution with higher variance under the new technology. This introduces an additional penalty for risk-averse borrowers.

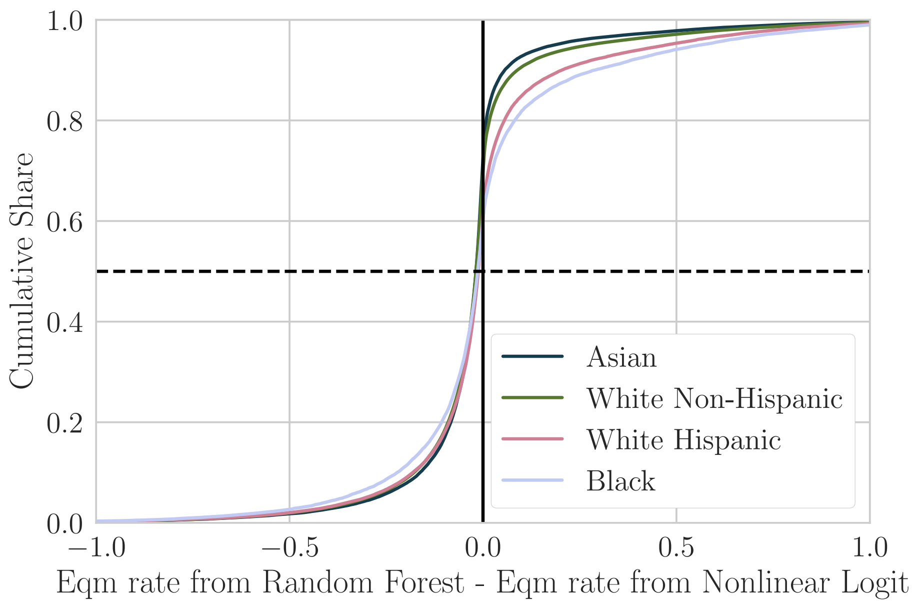
To further explore effects of new technology at the intensive margin, Figure 7 plots the difference of offered rates in equilibrium under the Random Forest model and those under the Nonlinear Logit model, for the borrowers accepted for a loan under both technologies.

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

Table 6: **Equilibrium Outcomes**

	Accept (%)		Mean SATO (%)		SD SATO (%)	
	(1) NL	(2) RF	(3) NL	(4) RF	(5) NL	(6) RF
Asian	92.4	93.3	-0.108	-0.123	0.274	0.322
White Non-Hispanic	90.3	91.1	-0.083	-0.090	0.296	0.356
White Hispanic	85.6	86.4	-0.031	-0.008	0.333	0.414
Black	77.7	79.3	0.022	0.060	0.365	0.461
Other	88.9	89.5	-0.083	-0.088	0.296	0.360
Population	89.8	90.7	-0.081	-0.086	0.298	0.360
Cross-group SD	2.165	2.098	0.020	0.029		

Figure 7: **Comparison of Equilibrium Interest Rates**



equilibrium rate) from the introduction of the new machine learning technology, and vice versa. Once again, the machine learning model appears to generate unequal impacts on different race groups. A larger fraction of White and especially Asian borrowers appear to benefit from the introduction of the technology, being offered lower rates under the new technology, while the reverse is true for Black and Hispanic borrowers.

To better understand the changes at the extensive margin, Table 7 distinguishes across three sets of borrowers: "Inclusions" are rejected for credit under the old technology (Nonlinear Logit) in equilibrium but are accepted under the new (Random Forest). "Exclusions" are accepted under the old technology but rejected under the new technology. The third category are borrowers who are accepted under both technologies. We study the shares of these three categories, as well as their interest rates, for the borrower population as a whole (Panel A) as well as for Asian and White non-Hispanic borrowers (Panel B) and Black and White Hispanic borrowers (Panel C).

Table 7: **Decomposition of Equilibrium Effects**

	Nonlinear Logit		Random Forest	
	Mean SATO	SD SATO	Mean SATO	SD SATO
A. All Groups				
Inclusions (2.61%)			0.96	0.41
Exclusions (1.77%)	0.71	0.39		
Always accepted (88.05%)	-0.10	0.27	-0.12	0.31
B. White Non-Hispanic and Asian Borrowers				
Inclusions (2.52%)			0.95	0.42
Exclusions (1.70%)	0.71	0.39		
Always accepted (88.59%)	-0.10	0.27	-0.12	0.30
C. Black and White Hispanic Borrowers				
Inclusions (3.90%)			1.00	0.38
Exclusions (2.85%)	0.72	0.39		
Always accepted (79.89%)	-0.04	0.31	-0.03	0.37

The proportions of Inclusions and Exclusions reported in the table reveal that the increase in average acceptance rates (in Table 6) masks both winners and losers along the extensive margin. Indeed, 1.8% of the population are losers (who are excluded when moving to the machine learning model) while 2.6% are winners who newly get included. The first row of Panel A shows that the Inclusions are high-risk borrowers, who are charged an average spread that is 96bp larger under Random Forest. These borrowers also have above-average dispersion of equilibrium rates. The second row shows that Exclusions are also high-risk borrowers, but less so than winners. The third row shows that the patterns among borrowers who are always accepted are similar to the population averages.

For the Asian and White non-Hispanic borrowers in Panel B, the shares of Inclusions and Exclusions as well as their rates look similar to the population overall. In Panel C, we see that for Black and White Hispanic borrowers, the shares of both Inclusions and Exclusions are higher, echoing our earlier results on increased dispersion for this group.

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

6 Conclusion

In this paper, we analyze the distributional consequences of changes in statistical technology used to evaluate creditworthiness. Our analysis is motivated by the rapid adoption of machine learning technology in this and other financial market settings. We find that these changes in technology can increase disparity in credit market outcomes across different groups—based, for example, on race—of borrowers in the economy. We present simple theoretical frameworks to provide insights about the underlying forces that can generate such changes in outcomes. We then provide evidence to suggest that this issue manifests itself in US mortgage data, focusing on the distribution of mortgages and rates across race-based groups.

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 (the probability of mortgage default, in our setting). It immediately follows that such dispersion will generate both "winners" and "losers" relative to their position in equilibrium under the pre-existing technology.

Using a large dataset from the US mortgage market, and evaluating a change from a traditional Logit technology to a machine learning technologies, we find that Black and White Hispanic borrowers are predicted to lose, relative to White and Asian borrowers. This is true both in terms of the distribution of predicted default propensities, and, in our counterfactual evaluation, in terms of equilibrium rates.

We outline two possible mechanisms through which such distributional changes could come about. One potential source arises from the increased flexibility of the new technology to better capture the structural relationship between observable characteristics and default outcomes. Another is that the new technology could more effectively triangulate the (hidden) identity of borrowers. With this better ability to triangulate, the technology might then penalize certain groups of borrowers over and above the structural relationship between other observables and default outcomes, if these groups have higher default probabilities even controlling for observables. We suggest a simple way to bound the relative importance of these two sources, and in our empirical analysis find that flexibility is the main source of unequal effects between groups, with triangulation playing a relatively smaller role.

Clearly, increases in predictive accuracy can (and in our setting, do) arise from the improved use of 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, or gender. Our work makes a start on studying these impacts in the domain of credit markets, and we believe there is much more to be done to understand the impacts of the use of these technologies on the distribution of outcomes in a wide variety of financial and goods markets.

7 Appendix

7.1 Proof of Lemma 1

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

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

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

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

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

7.2 General Characterization of Unequal Effects

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:

$$E[\hat{P}(x|\mathcal{M}_2) - \hat{P}(x|\mathcal{M}_1)|g] = \sum_{j=2}^{\infty} \frac{1}{j!} \frac{\partial^j \hat{P}(a|\mathcal{M}_2)}{\partial x^j} E\left[(x-a)^j|g\right] - B \tag{6}$$

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

Proof:

The linear prediction can be written as $\hat{P}(x|\mathcal{M}_1) = \alpha + \beta x$. For the nonlinear technology, let $\underline{\beta} = \min_{x \in [\underline{x}, \bar{x}]} \frac{\partial \hat{P}(x|\mathcal{M})}{\partial x}$ and $\bar{\beta} = \max_{x \in [\underline{x}, \bar{x}]} \frac{\partial \hat{P}(x|\mathcal{M})}{\partial x}$. 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 $\beta = \frac{\partial \hat{P}(a|\mathcal{M}_2)}{\partial x}$. Then, we can write the linear prediction as a shifted first-order Taylor approximation of the nonlinear prediction around a:

$$\hat{P}(x|\mathcal{M}_1) = \hat{P}(a|\mathcal{M}_2) + \frac{\partial \hat{P}(a|\mathcal{M}_2)}{\partial x}(x-a) + B$$

where $B = \hat{P}(a|\mathcal{M}_1) - \hat{P}(a|\mathcal{M}_2)$. Now using a Taylor series expansion around a, we have

$$\hat{P}(x|\mathcal{M}_2) - \hat{P}(x|\mathcal{M}_1) = \sum_{j=2}^{\infty} \frac{1}{j!} \frac{\partial^j \hat{P}(a|\mathcal{M}_2)}{\partial x^j} (x-a)^j - B \tag{7}$$

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

7.3 Example of Triangulation

We prove our claims in the discussion of Figure 2. Suppose that

$$y = \beta \cdot x + \gamma \cdot g + \varepsilon$$

where x is a one-dimensional characteristic, $g \in \{0, 1\}$ is an indicator for the Blue group, and ε is independent of x and g. Suppose that $x|g \sim N(a, v(g))$ and normalize $a = 0$. Let $v(1) > v(0)$ and $\gamma > 0$. There is no linear correlation between x and g, since

$$\begin{aligned} Cov(x, g) &= E[x \cdot g] = E[E[x \cdot g|g]] \\ &= E[E[x|g] \cdot g] = 0 \end{aligned}$$

Hence the projection of y onto linear functions of x is

$$\hat{P}_{\text{lin}}(x) = \alpha_{\text{lin}} + \beta \cdot x$$

where the intercept $\alpha_{\text{lin}} = E[y]$. The projection of y onto quadratic functions of x is

$$\hat{P}_{\text{quad}}(x) = \alpha_{\text{quad}} + \beta \cdot x + \gamma \cdot (\phi \cdot x^2),$$

where

$$\phi = \frac{Cov(x^2, g)}{Var(x^2)}$$

is the regression coefficient of g onto x^2.[52] Note that $E[x^2|g] = v(g)$ is increasing in g, and hence $Cov(x^2, g) > 0$. It follows that the fitted value is a convex quadratic function, as illustrated in the figure.

[52] This follows, for example, from a standard partial regressions argument: Regressing y on $\{1, x\}$ gives residual $\varepsilon_y = \gamma(g - E[g])$. Regressing $z \equiv x^2$ on $\{1, x\}$ gives fitted value $\hat{z} = E[z]$, because $Cov(x, x^2) = 0$ for a mean-zero normal variable, and residual $\varepsilon_z = z - E[z]$. By the Frisch-Waugh-Lovell theorem, the coefficient on x^2 in $\hat{P}_{\text{quad}}(x)$ is the same as in the regression of ε_y on ε_z, namely $Cov(\gamma g, x^2)/Var(x^2) = \gamma\phi$.

References

Agrawal, A., J. Gans, and A. Goldfarb (2018): *Prediction Machines: The Simple Economics of Artificial Intelligence*, Harvard Business Review Press.

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, 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, 29–50.

Berg, T., V. Burg, A. Gombovic, and M. Puri (2018): "On the Rise of FinTechs – Credit Scoring Using Digital Footprints," Tech. rep., Frankfurt School of Finance & Management.

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, 263–294.

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

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

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

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

Breiman, L. (2001): "Random forests," *Machine learning*, 45, 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, 3214–48.

CAMPBELL, J. AND J. COCCO (2015): "A Model of Mortgage Default," *Journal of Finance*, 70, 1495–1554.

CAMPBELL, J. Y., J. HILSCHER, AND J. SZILAGYI (2008): "In Search of Distress Risk," *Journal of Finance*, 63, 2899–2939.

CHEN, T. AND C. GUESTRIN (2016): "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, ACM, 785–794.

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, 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, Elsevier, vol. 5 of *Handbook of Public Economics*, chap. 3, 111 – 193.

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

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.

DEMYANYK, Y. AND O. VAN HEMERT (2011): "Understanding the Subprime Mortgage Crisis," *Review of Financial Studies*, 24, 1848–1880.

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

EINAV, L., A. FINKELSTEIN, AND J. LEVIN (2010): "Beyond testing: Empirical models of insurance markets," *Annual Review of Economics*, 2, 311–336.

ELUL, R., N. S. SOULELES, S. CHOMSISENGPHET, D. GLENNON, AND R. HUNT (2010): "What 'Triggers' Mortgage Default?" *American Economic Review*, 100, 490–494.

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

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., L. Goodman, D. Lucca, L. Madar, L. Molloy, and P. Willen (2013): "The Rising Gap between Primary and Secondary Mortgage Rates," *Federal Reserve Bank of New York Economic Policy Review*, 19, 17–39.

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, 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, 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, 832–844.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013): *An Introduction to Statistical Learning*, Springer.

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, 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, 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., J. Ludwig, S. Mullainathan, and A. Rambachan (2018): "Algorithmic Fairness," *AEA Papers and Proceedings*, 108, 22–27.

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, 41–62.

Mullainathan, S. and J. Spiess (2017): "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31, 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*, IEEE Computer Society, 111–125.

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*, ACM, 625–632.

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, 659–661.

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

Richard, S. F. and R. Roll (1989): "Prepayments on fixed-rate mortgage-backed securities," *Journal of Portfolio Management*, 15, 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," Tech. rep., Stanford University.

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

Wheatley, M. (2001): "Capital One Builds Entire Business on Savvy Use of IT," *CIO Magazine*.

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

A.1 Isotonic Regression and Calibration

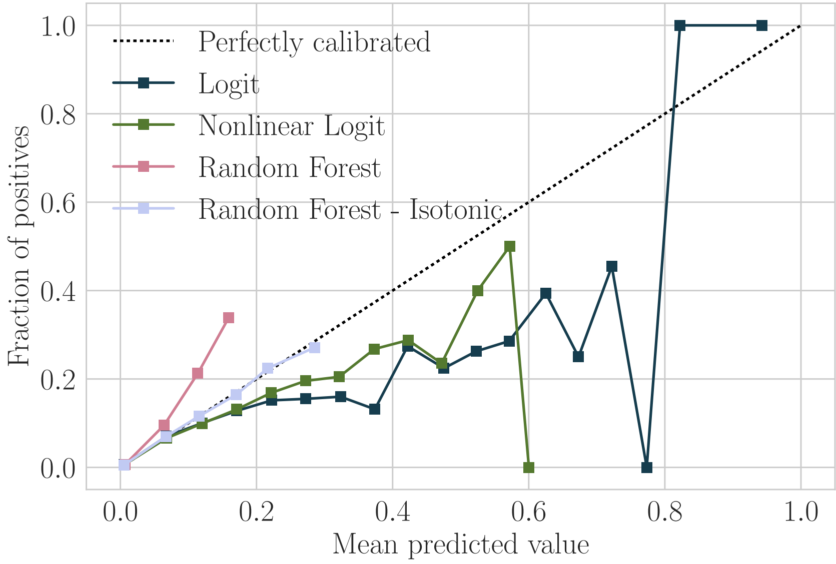
As discussed in Section 4.2.1, the direct estimates of probability that come from tree-based models like the Random Forest model tend to be very noisy. A frequently used approach in machine learning is to "calibrate" these estimated probabilities by fitting a monotonic function to smooth/transform them (see, for example, Niculescu-Mizil and Caruana, 2005). In our empirical work, we employ isotonic regression calibration to translate the predicted classifications into probability estimates.

Isotonic regression involves searching across the space of monotonic functions to find the best fit function connecting the noisy estimates with the true values. More concretely, for an individual i, let y_i be the true outcome, and let $\hat{y}_i$ be predicted value from the Random Forest model. Then, the isotonic regression approach is to find $\hat{z}$ in the space of monotonic functions to minimize the mean squared error over the calibration data set:

$$\hat{z} = \arg\min_{z} \sum_{i} (y_i - z(\hat{y}_i))^2. \tag{8}$$

We estimate this fit over an additional "left-out" dataset, which we call the calibration dataset. In our results, we calculate predicted probabilities as $\hat{z}(\hat{y}_i)$. We examine the improvement from calibration in Figure A-1. This figure bins the predicted values (either $\hat{y}_i$ or $\hat{z}(\hat{y}_i)$) into 20 equally spaced bins, and takes the average true outcome value for each predicted bin. If the model is perfectly calibrated, the two values are equal (denoted by the 45° line). We see that for both Logit and Nonlinear Logit, the true values tend to be below the predicted values, suggesting that for higher predicted values, the Logit models over-predict default. In contrast, the Random Forest line is above the perfectly-calibrated line, suggesting that it is underpredicting default. In contrast, the Random Forest - Isotonic line is almost exactly on the 45° line, suggesting near-perfect calibration.

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



A.2 Comparing Performance With and Without SATO

As discussed in Section 4, we include the interest rate (as SATO) in the set of covariates used to predict default on the right-hand side. In Table A-1, we compare the predictive accuracy of the models with and without the interest rate as a contract characteristic, and see that in most models and cases, adding SATO noticeably improves the accuracy. For both Logit and Nonlinear Logit, in all statistics the model improves with the addition of SATO on the leave-out sample, while in the Random Forest model, adding SATO improves the model performance for all statistics except Precision Score. This implies that there is additional variation in SATO that predicts default that is not already captured by the other borrower and loan observables.

Table A-1: **Performance of Different Statistical Technologies Predicting Default, with and without SATO**

	ROC AUC		Precision Score		Brier Score × 100		R^2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Model	SATO	No SATO	SATO	No SATO	SATO	No SATO	SATO	No SATO
Logit	0.8522	0.8486	0.0589	0.0578	0.7172	0.7181	0.0245	0.0232
Nonlinear Logit	0.8569	0.8537	0.0598	0.0589	0.7146	0.7149	0.0280	0.0275
Random Forest	0.8634	0.8602	0.0630	0.0639	0.7114	0.7120	0.0323	0.0315

Note: Performance metrics of different models. For ROC AUC, Precision score, and R^2, higher numbers indicate higher predictive accuracy; for Brier score, lower numbers indicate higher accuracy. In odd-numbered columns, SATO is included as a covariate in the prediction models; in even-numbered columns, it is not included.

A.3 Additional Machine Learning Estimator: XGBoost

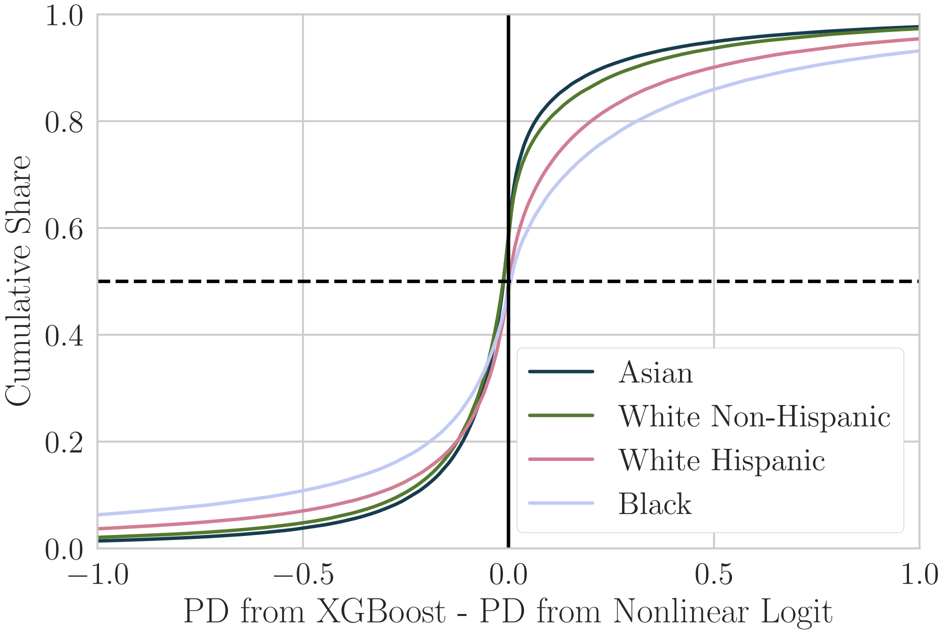
As an additional alternative machine learning method, we also estimate a model known as Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016). In essence, the XGBoost is another nonparametric and nonlinear estimator that uses trees, similar to the Random Forest method. However, unlike the Random Forest method, which aggregates across randomly bootstrapped trees, XGBoost improves its training methods by *boosting* the improvement of a single tree.

The gradient boosting approach takes a single tree model, similar to those underlying the Random Forest. However, rather than increase the number of trees, the model iterates over the tree by constructing new leafs (branching) and removing leafs (pruning) to continuously improve the tree's predictive power. In particular, the formulation of the problem allows the tree to focus on improving where the tree can gain the most by strengthening the "weakness" in the prediction process. Statistically, this method can be viewed as optimizing two components: the training loss (i.e. the mean squared error of the prediction) and the complexity of the model (i.e. avoiding overfitting through a penalization function). The XGboost method allows for a rapid and efficient optimization over these two criteria.[1]

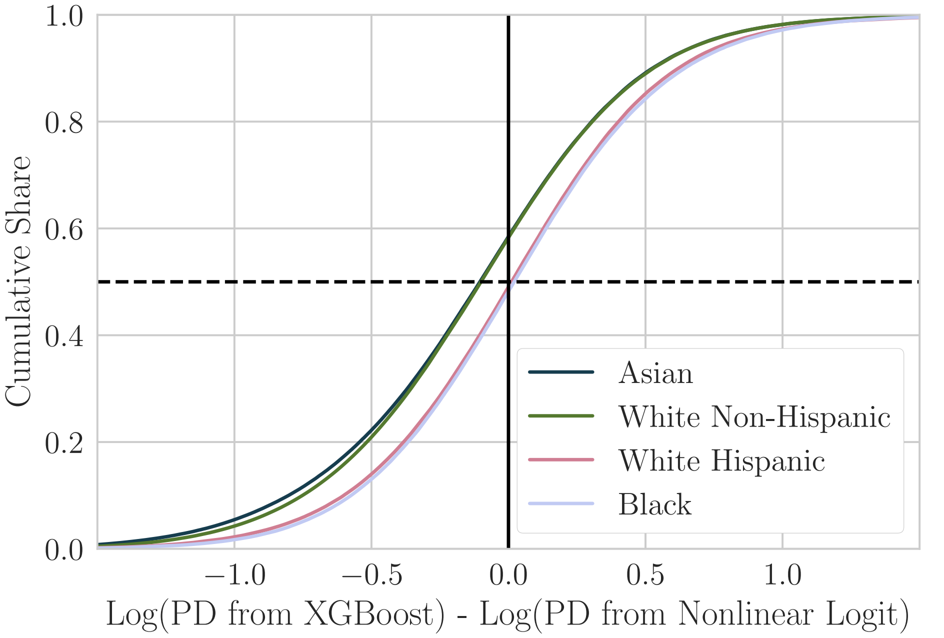
In Figure A-2, we plot a version of Figure 6, replacing the predictions from the Random Forest model with those from XGBoost. The qualitative conclusions are the same: moving to the more complex model, there are more "winners" among the White non-Hispanic and Asian borrowers than among the Black and White Hispanic groups.

[1]We implement this method in R using the `xgboost` library.

Figure A-2: **Comparison of Predicted Default Probabilities — XGBoost vs. Nonlinear Logit**



Panel A



Panel B

A.4 Computational Prodedure

To compute equilibrium, for every borrower i, we evaluate $N(x_i, R)$ at a grid of 20 values for SATO between -0.4 percent and 1.5 percent, using the predicted default intensities $P(x_i, R)$ from every statistical technology. We then use linear interpolation to solve for the equilibrium interest rate. If no such solution exists within the grid of interest rates considered, we conclude that borrower i is rejected.

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 modeling across cohorts of issued loans, with a view to using data up until the present.

This generates the need for further modeling, 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[2] 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:

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

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 $\hat{P}$, 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:

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

[2] 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.

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:

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

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:

$$\begin{aligned}P(T) &= 1 - \exp[MH(T)] \\ &= 1 - \exp\left[\frac{1}{h_{max}} \frac{\log(1-\hat{p})}{\hat{t} - \frac{t_1}{2}} H(T)\right] \\ &\equiv P_T(\hat{P}) \qquad (9)\end{aligned}$$

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 (9) into equation (5) and proceed.

A.5 Endogenous Contracting Terms

In our model, lenders' Net Present Value depends on contracting terms beyond the interest rate. In particular, equation (5) 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.6 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 $\hat{P}^{(j)}$, and

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

where the small number ϵ is introduced to ease computation when taking logarithms. Note that this transformation is invertible with $\hat{P} = 1 - e^{-e^{\epsilon - \Lambda}}$.

We know that our estimates imply a sensitivity $\hat{\phi}$ which is biased, i.e., we can assume that the true sensitivity is $b\hat{\phi}$, where b measures the bias as discussed above.

To adjust our estimates, we transform estimated PDs $\hat{P}^{(j)}$ into $\Lambda^{(j)}$. We assume that the estimates are unbiased for the average interest rate (corresponding to SATO $= 0$ in our dataset), with associated grid point $j = j^{\star}$. Then we obtain the bias-adjusted figure

$$\Lambda^{(j)}_{adj} = b \cdot \Lambda^{(j^{\star})} + (1 - b) \cdot \Lambda^{(j)}$$

and finally invert the transformation to get the bias-adjusted PD

$$\hat{P}^{(j)}_{adj} = 1 - e^{-e^{\epsilon - \Lambda^{(j)}_{adj}}}.$$

A.7 Descriptive Statistics, Equilibrium Sample

We show descriptive statistics for the equilibrium sample (GSE, full documentation) in Table A-2. The table simply confirms that the patterns that are evident in the broader set of summary statistics are also evident for this subsample.

Figure A-3 shows the cumulative distribution functions of the differences between the default probabilities 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.8 Residual Interest Rate Variation

Figure A-4 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 A-3 below shows results from a more direct way to check for the prevalence of

Table A-2: **Descriptive Statistics, GSE, Full Documentation Originations.**

Group		FICO	Income	LoanAmt	Rate (%)	SATO (%)	Default (%)
	Mean	765	121	278	4.16	-0.10	0.35
Asian	Median	775	105	259	4.25	-0.06	0.00
(N=335,892)	SD	39	72	138	0.71	0.45	5.89
	Mean	740	92	181	4.36	0.08	1.57
Black	Median	748	77	155	4.38	0.08	0.00
(N=114,152)	SD	53	60	109	0.71	0.49	12.44
	Mean	748	89	192	4.32	0.06	0.83
White Hispanic	Median	758	74	166	4.38	0.06	0.00
(N=200,543)	SD	47	62	112	0.71	0.48	9.06
	Mean	763	109	212	4.24	-0.04	0.56
White Non-Hispanic	Median	774	92	186	4.25	-0.02	0.00
(N=3,947,597)	SD	42	71	117	0.69	0.43	7.49
Native Am, Alaska,	Mean	751	97	210	4.34	0.01	0.97
Hawaii/Pac Isl	Median	762	82	185	4.38	0.02	0.00
(N=31,275)	SD	47	64	119	0.69	0.46	9.81
	Mean	761	118	233	4.31	-0.03	0.69
Unknown	Median	773	100	206	4.38	-0.02	0.00
(N=520,459)	SD	44	76	128	0.69	0.44	8.29

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 with full documentation securitized by Fannie Mae or Freddie Mac, originated over 2009-2013.

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 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 A-5 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

Table A-3: **Residual Variation in SATO, comparing Full and Equilibrium samples.**

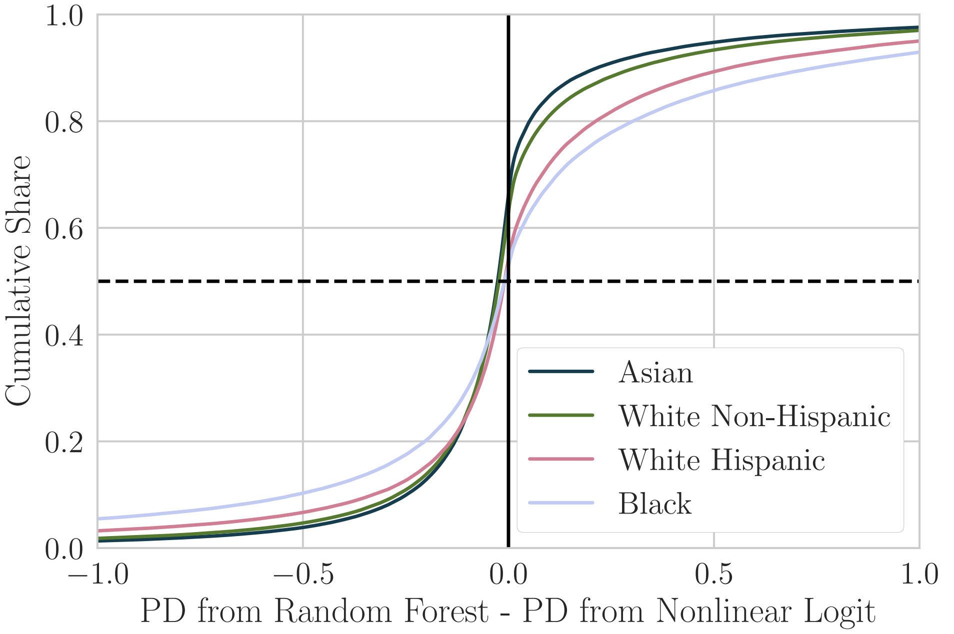
	sato_res	sato
Equilibrium Sample	0.292	0.441
Other	0.312	0.438

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). 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.

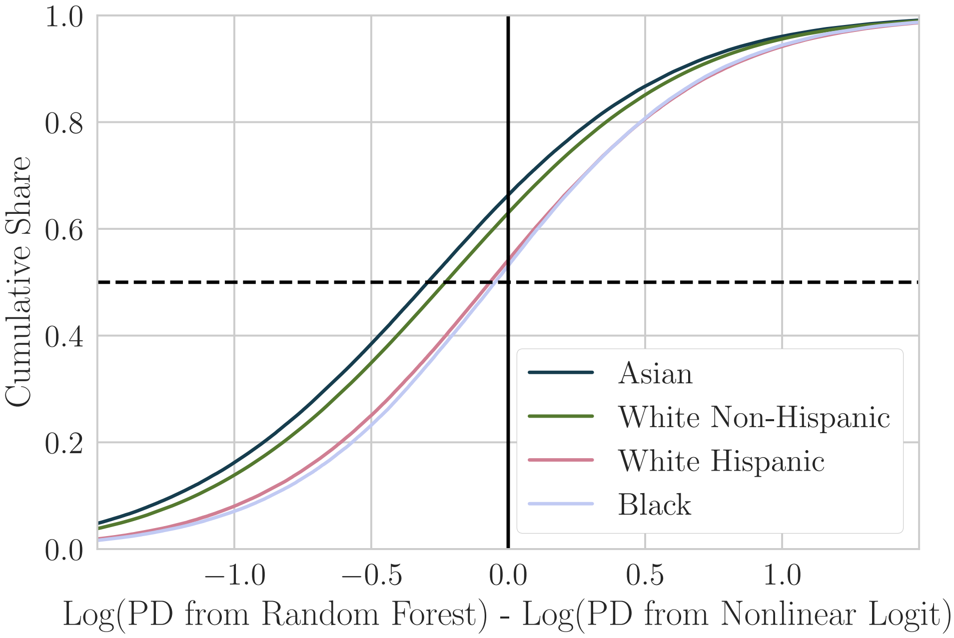
measures this fraction), but the patterns in the plot are broadly reassuring about the fairly limited extent of this extrapolation.[3]

[3] 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 A-3: **Comparison of Predicted Default Probabilities, Equilibrium Sample**



Panel A



Panel B

Figure A-4: **Predicted PD, comparing Full and Equilibrium samples**.

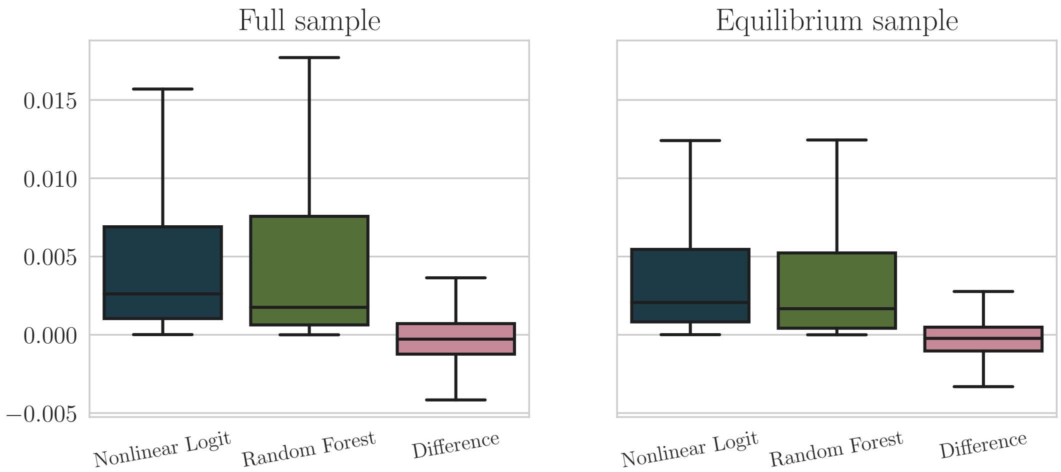


Figure A-5: **Residual interest rate variation**.

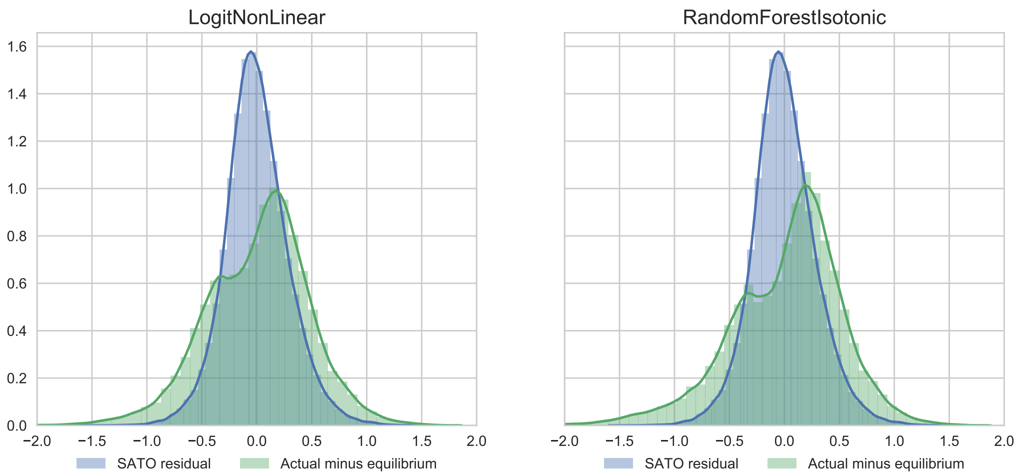
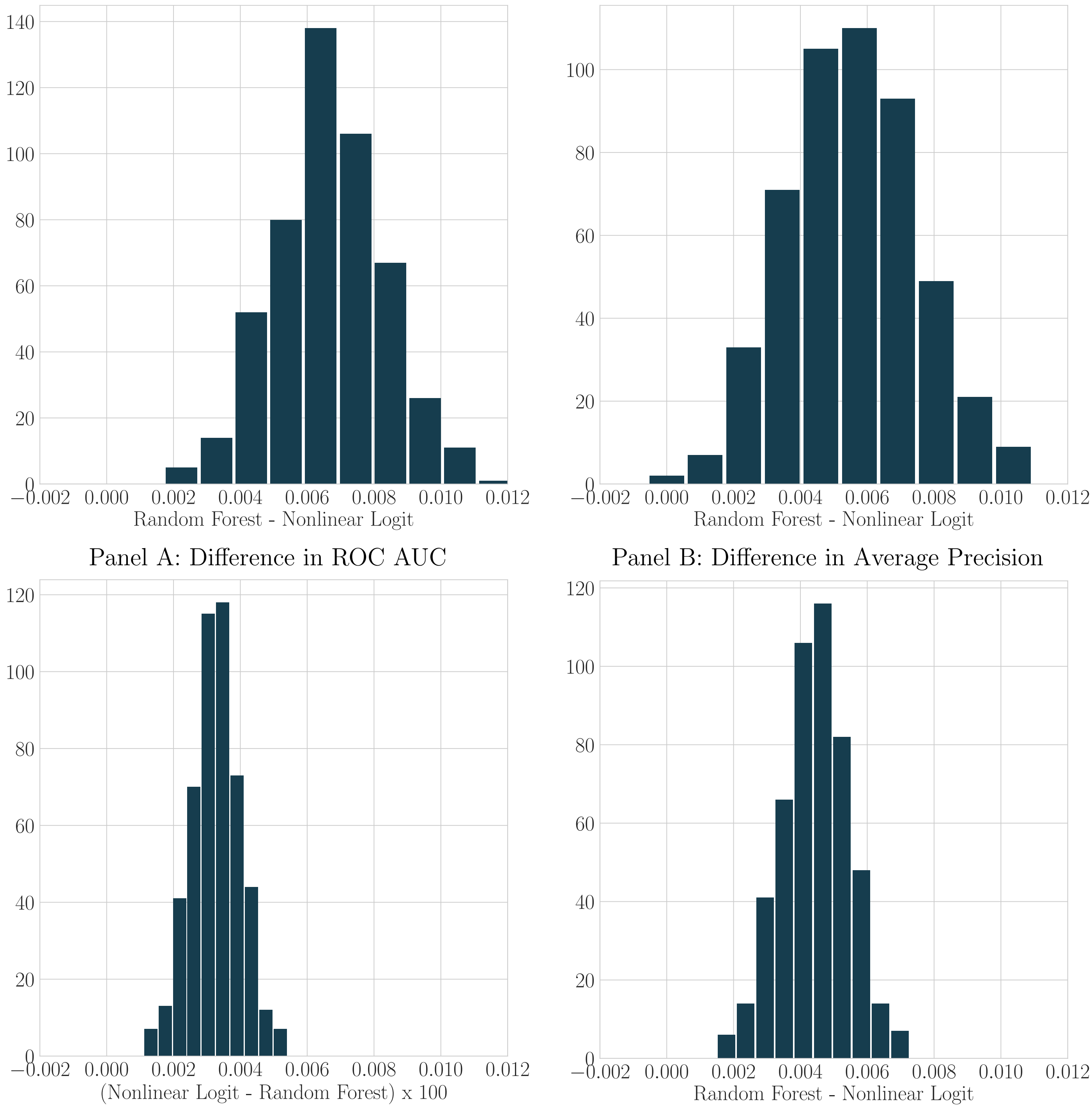


Figure A-6: **Bootstrap Estimates of Differences in AUC and Average Precision**



Panel A: Difference in ROC AUC

Panel B: Difference in Average Precision

Panel C: Difference in Brier Score

Panel D: Difference in R^2