Predictably Unequal?
The Effects of Machine Learning on Credit Markets

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

This draft: March 2020

[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, Francesco D'Acunto, Andrew Ellul, Kris Gerardi, Andra Ghent, Johan Hombert, Ralph Koijen, Andres Liberman, Gonzalo Maturana, Adair Morse, Karthik Muralidharan, Daniel Paravisini, Jonathan Roth, Jann Spiess, Jeremy Stein, Daniel Streitz, Johannes Stroebel, and Stijn Van Nieuwerburgh for useful conversations and discussions, participants at numerous conferences and seminars, and the reviewing team at the Journal of Finance for thoughtful 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.

Electronic copy available at: https://ssrn.com/abstract=3072038

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.

Electronic copy available at: https://ssrn.com/abstract=3072038

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 basic theoretical frameworks to analyze these issues, conducting empirical analysis on a large administrative dataset of loans in the US mortgage market, and undertaking an initial assessment of potential economic magnitudes using a simple equilibrium model.

The essential idea underlying our paper is that a more sophisticated statistical technology (in the sense of reducing predictive mean squared error) produces predictions with greater variance than a more primitive technology. When applied to the context that we study, the insight that this yields is that 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 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.

[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).

Electronic copy available at: https://ssrn.com/abstract=3072038

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 who wins or loses depends on both the functional form of the new technology, and the differences in the distribution of the characteristics across groups. Perhaps the simplest way to understand this point is to consider an economy endowed with a primitive prediction technology which simply uses the mean level of a single characteristic to predict default. In this case, the predicted default rate will just be the same for all borrowers, regardless of their particular value of the characteristic. If a more sophisticated linear technology which identifies that default rates are linearly 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 mapping between permissible variables and default outcomes. Another possibility is that the structural relationship between permissible variables and default is perfectly estimated by the primitive technology, but the

Electronic copy available at: https://ssrn.com/abstract=3072038

more sophisticated technology can *triangulate* the effect of the unobserved restricted variables on the outcome by more effectively and accurately combining the observed permissible variables. In the latter case, particular groups are penalized or rewarded based on realizations of the permissible variables, as the more sophisticated technology "de-anonymizes" the group identities in the data using the permissible variables.[5]

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 joint distributions of characteristics across groups, or 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.[6] 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. In our empirical work, 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.[7]

We estimate a set of increasingly sophisticated statistical models to predict default using these data, beginning with a logistic regression of default outcomes on borrower and loan characteristics, and culminating in a Random Forest machine learning model (Ho, 1998;

[5]While the concept of triangulation has been well-investigated by prior work in the area (see, e.g., Ross and Yinger, 2002; Pope and Sydnor, 2011), we add to this line of research by investigating how the incidence of triangulation is affected by the introduction of more sophisticated prediction technologies.

[6]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.

[7]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.

Electronic copy available at: https://ssrn.com/abstract=3072038

Breiman, 2001).[8]

We confirm that the machine learning technology delivers statistically significantly higher out-of-sample predictive accuracy for default than the simpler logistic models. We also 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) "win", that is, experience lower estimated default propensities under the machine learning technology than the less sophisticated Logit technology, these benefits do not accrue to the same degree to some minority race and ethnic groups (e.g., Black and Hispanic borrowers). We show that these inferences are robust to numerous changes to the set of covariates, the sample used for estimation, and the estimation approach.

We propose simple empirical measures of the extent to which flexibility or triangulation is responsible for these results, by comparing the performance of the naïve and sophisticated statistical models when race and ethnicity are included and withheld from the information set used to predict default. While we find that both flexibility and triangulation are important, in our empirical application, the majority of the predictive accuracy gains from the more sophisticated machine learning model can be attributed to the increased flexibility of the model, with at most 30% attributable to pure triangulation. These findings suggest that simply prohibiting certain variables as predictors of default propensity will likely become increasingly ineffective as technology improves.[9] For one, such regulations will confront the difficulty of prohibiting triangulation in the face of increasingly complicated attempts to model the joint distribution of outcomes, permissible, and restricted characteristics.[10]

[8]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.

[9]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).

[10]We note here that the machine learning models are better than the logistic models at predicting race

Electronic copy available at: https://ssrn.com/abstract=3072038

Another important reason is that such regulations cannot protect minorities against the greater flexibility conferred by the new technology.

How might these changes in predicted default propensities across race and ethnic groups map 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 provide a first evaluation of these questions, we embed the statistical models in a simple equilibrium model of credit provision in a competitive credit market in which rational lenders compete to issue loans. To evaluate magnitudes, we assume that lenders are subject to a constraint arising from the availability of statistical prediction technology.[11] We then compute counterfactual equilibria associated with each statistical technology, and compare the resulting equilibrium outcomes with one another to evaluate comparative statics on outcomes across groups.

In this simple analysis of counterfactuals arising under different technologies, we face a number of obvious challenges to identification. These arise from the fact that the data that we use to estimate the default models were not randomly generated, but rather, a consequence of the interactions between borrowers and lenders who may have had access to additional information whilst making their decisions. We attempt to deal with these challenges in a number of sensible ways by changing the estimation sample and attempting to de-bias our estimates, as we describe later in the paper. We simply caveat here that the results of our elementary computations should not be viewed as a precise prediction, but

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).

[11] 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. In contrast, 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. Our exercise can be viewed as an attempt to map the changes in default probabilities that we find on credit provision along the intensive and extensive margins, which is of interest whether new prediction technology is used by private lenders, or by a centralized entity changing its approach to setting underwriting criteria.

Electronic copy available at: https://ssrn.com/abstract=3072038

instead as a useful first step towards assessing magnitudes.

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 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

[12] Bartlett et al. (2019) study empirically whether "FinTech" mortgage lenders in the US appear to discriminate more across racial groups. Buchak et al. (2018) and Fuster et al. (2019) study other aspects of FinTech lending in the US mortgage market.

Electronic copy available at: https://ssrn.com/abstract=3072038

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 basic theory framework to understand how improvements in statistical technology can affect different groups of households in credit markets, and describes how nonlinear technologies relate to 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 simple equilibrium model of credit provision under different technologies, and discusses how the changes in default predictions affect both the intensive and extensive margins of credit provision. Section 6 concludes. An extensive online appendix contains a few proofs, numerous auxiliary analyses, and robustness checks.

2 A Simple Conceptual Framework

Consider a lender predicting the probability of default, $y \in [0, 1]$, of a loan using a vector x of observable borrower characteristics (e.g., income, credit score) and contract terms (e.g. loan size, interest rate). The lender uses historical data to find a function $\hat{y} = \hat{P}(x)$ which maps x into a predicted y. Each borrower is characterized by x, as well as by her group membership g (e.g., her race). The lender is not permitted to include g in prediction.

[13] These issues have been a major focus of work on mortgages and housing—see, e.g., Berkovec et al. (1994, 1998), Ladd (1998), Ross and Yinger (2002), Ghent et al. (2014), Bayer et al. (2018), or Bhutta and Hizmo (2019). 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).

Electronic copy available at: https://ssrn.com/abstract=3072038

Machine learning techniques such as tree-based models and neural networks can employ a wider range of functional forms $\hat{P}(x)$ in prediction, relative to traditional approaches (e.g., Logit) which are based on linear functions of x. We represent this by assuming that traditional statistical technologies lie in class $\mathcal{M}_1$ of predictive functions (i.e., lenders using these technologies can only choose mappings $\hat{P} \in \mathcal{M}_1$), while machine learning allows consideration of a larger set of functions $\mathcal{M}_2$, where $\mathcal{M}_1 \subset \mathcal{M}_2$.[15] Note that we study the distributional consequences of innovation in statistical technologies given a *fixed* set of observable variables x; we do not consider the effects of expanding this set, say by using borrowers' "digital footprints" (e.g., Berg et al., 2019).

The standard goal of statistical learning is to find predictive functions $\hat{P}(x)$ that converge, given enough data, to the "oracle."[16] The oracle is the optimal predictor in the class of available functions, minimizing a statistical loss function such as the predictor's mean-square error out of sample. Additional machine learning techniques such as regularization allow faster convergence to the oracle because they discipline overfitting in finite samples.

Given this setup, to derive a large-sample approximation of the consequences of the change in technology, we compare the oracle in $\mathcal{M}_1$ to that obtained in the broader class $\mathcal{M}_2$. We find that improvements in statistical technology lead to predictions that are more disperse across borrowers:

Lemma 1. Let $\hat{P}(x|\mathcal{M}_1)$ be the oracle (i.e., the predictor that minimizes mean-square error loss) among functional forms available with traditional statistical technology. Let

[15] In practice, these classes of functional forms are nested only in an approximate sense. For example, tree-based models work by combining simple indicator functions, which can never exactly replicate a smooth functional form such as Logit. However, "simple approximation" results in real analysis state that one can approximate any well-behaved function arbitrarily well with functions that combine sufficiently many indicator functions. Thus, tree-based models (in a manner that is similar to neural networks) are "universal approximators": they can arbitrarily closely represent any functional form if they are allowed enough flexibility (i.e., enough leaves, trees, and splits). We therefore get closer and closer to the "nested models" scenario as the data become larger and the statistician can allow more flexibility.

[16] See, for example, Vapnik (1999) and Friedman et al. (2001) for an exposition of statistical learning theory.

Electronic copy available at: https://ssrn.com/abstract=3072038

$\hat{P}(x|\mathcal{M}_2)$ be the corresponding oracle available with machine learning, with $\mathcal{M}_1 \subset \mathcal{M}_2$. Then, in a population of borrowers, $\hat{P}(x|\mathcal{M}_2)$ is a mean-preserving spread of $\hat{P}(x|\mathcal{M}_1)$.

Proof: See appendix.[17]

The result is intuitive: by definition, improvements in technology yield predictions with a mean-square error at least as small as from pre-existing predictions. These new predictions $\hat{y}$ track true y more closely, and will therefore be more disperse on average.[18]

Lemma 1 shows that 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 (more risky) under the new technology, and therefore experience better (worse) credit market conditions. Put differently, 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. However, this analysis does not yet provide any guidance on the specific groups g of borrowers that will be made better or worse off, a matter to which we later return.

Figure 1 gives an instructive example of possible group-specific effects when borrowers have only one observable characteristic x. There are two groups of borrowers: Blue and Red, with the same mean $x = a$, but different variances of x; the two bell curves show these group-specific distributions of characteristics. Consider traditional statistical technology to consist of linear predictive functions, shown as $\hat{P}_{lin}(x)$. The more sophisticated statistical technology $\hat{P}_{nl}(x)$ is (in this case) convex quadratic in x. In this example, $\hat{P}_{nl}(x) > \hat{P}_{lin}(x)$

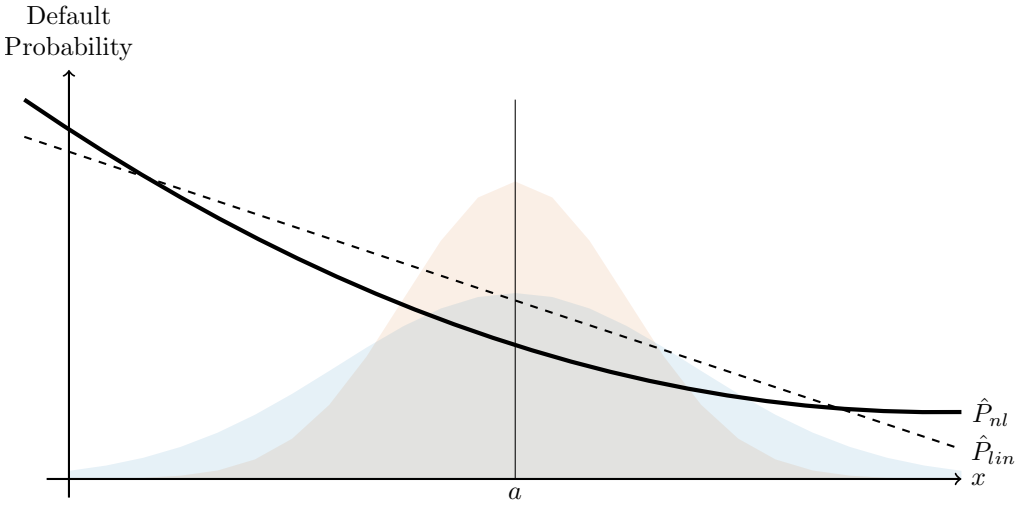
[17] The proof imposes the additional technical condition that both $\mathcal{M}_1$ and $\mathcal{M}_2$ are closed subspaces of the space $\mathcal{L}^2$ of square-integrable functions of x.

[18] The fact that the spread is mean-preserving follows because the oracle is unbiased regardless of technology. This is not necessarily true of predictions achieved in finite samples, where machine learning techniques trade off increases in bias against reductions the variance of the out-of-sample forecast (see, e.g., James et al., 2013). However, these biased predictions still converge to the oracle as the dataset grows large. As a result, the properties discussed here are (once again) approximately informative about the properties of regularized estimators, as long as algorithms are fit on sufficiently large datasets (in our case, $N \approx$10 million).

Electronic copy available at: https://ssrn.com/abstract=3072038

when x is far from its mean a in either direction. It follows that Blue borrowers tend to be adversely affected by new technology, as their characteristics x are more variable and hence more likely to lie in the tails of the distribution, which are penalized by nonlinear technology.

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



This intuition about the factors determining winners and losers generalizes beyond the convex quadratic example, which is used simply for illustrative purposes. More generally, the effect of introducing a more sophisticated technology depends on two factors. These are the higher-order moments of borrower characteristics in each group, and the higher-order derivatives of predictions under sophisticated technology.[19]

What are the underlying sources of unequal effects? We consider two widely discussed possibilities, as mentioned earlier in the introduction. One is that the unequal effects across groups could be driven by the *flexibility* of the new technology. If the true function connecting x and y is nonlinear, while g does not affect y, then the underlying source of the unequal effects is the ability of the new technology to capture this nonlinear structural relationship. Another possibility is that more sophisticated technology can *triangulate* group identity.

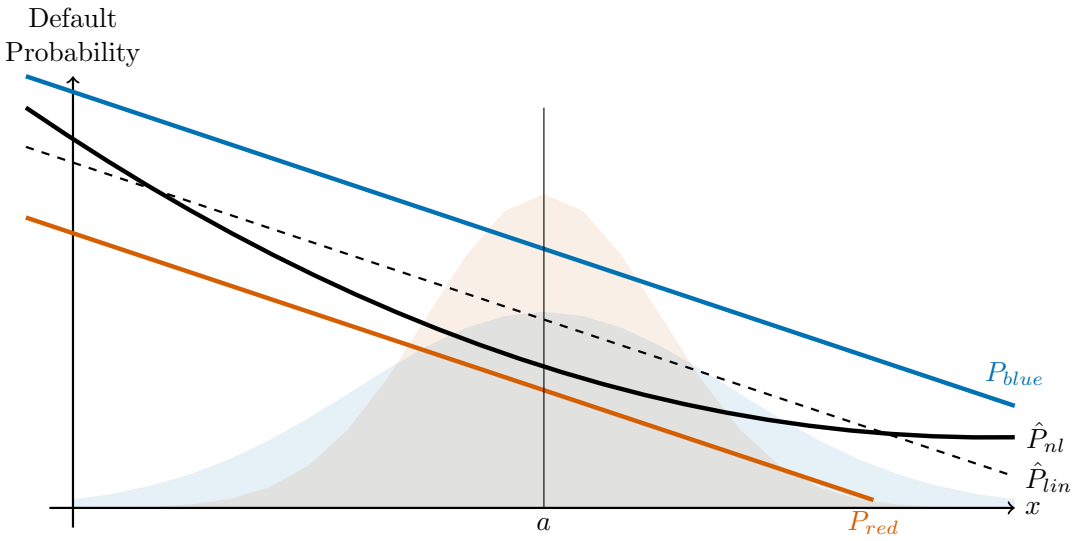
[19]Lemma 2 in the online appendix makes this point formally in the context of predictions that are smooth functions of a single characteristic x.

Electronic copy available at: https://ssrn.com/abstract=3072038

Intuitively, in this case, the more sophisticated technology uses nonlinear functions of x to more effectively proxy for the relationship between the omitted variable g and y, thus resulting in unequal effects under the new technology.[20]

Figure 2 provides an example where unequal effects are generated exclusively by triangulation. Here, true default risk is assumed to be a linear function of x in each group, and higher for the Blue group ($P_{blue} > P_{red}$), while the group-conditional distributions of x are the same as in Figure 1. The linear prediction $\hat{P}_{lin}(x)$ in this case will simply equal the population-weighted average of the true group-specific default probabilities (i.e., the dashed straight line in the figure). In contrast, the nonlinear technology penalizes the Blue group—since extreme realizations of x are more likely to come from Blue borrowers, the technology assigns higher predicted default probabilities to more extreme realizations of x.[21]

Figure 2: **Triangulation**



[20] As discussed in the literature, g could capture the effects of unobservables such as access to informal safety nets (e.g. ability to borrow from family or friends), differential treatment in the labor market, access to other sources of formal credit, or indeed other sources of unobserved income or wealth that affect y. Our addition to this debate is that machine learning would yield differential predictions across groups because it better proxies the predictive power of omitted g for y by using nonlinear combinations of x.

[21] In the online appendix, we provide a mathematical derivation of the pattern shown in this figure.

Electronic copy available at: https://ssrn.com/abstract=3072038

These examples highlight two points that guide our empirical analysis. First, Figure 1 highlights that the group-specific effects are ambiguous *a priori*. Without knowing the precise nature of nonlinearity in machine learning predictions, one cannot anticipate which groups will be better or worse off—for example, a concave quadratic function would deliver precisely the opposite effects to the convex quadratic we posited.[22] This ambiguity implies that we must inspect the data to understand the distributional effects of machine learning, as we do in the next section.

Second, Figure 2 suggests that flexibility and triangulation can deliver unequal effects that are observationally equivalent. The distinction between them is important, however, from a normative perspective. The 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, such as the approaches proposed in Ross and Yinger (2002) or Pope and Sydnor (2011), whereas flexibility might instead push us towards discussing the underlying sources of cross-group differences in the distributions of observable characteristics. In Section 4, we define empirical measures of flexibility and triangulation to attempt to ascertain the extent to which these two sources drive unequal effects observed in the data.

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^{TM} mortgage

[22] That is, new technology could allow a lender to identify good credit risks within a minority group previously assigned uniformly high predicted default rates under the old technology, thus *reducing* inequality across groups. Anecdotally, the credit card company CapitalOne more efficiently used demographic information and expanded lending in such a manner during the decade from 1994 to 2004. See, for example, Wheatley (2001).

Electronic copy available at: https://ssrn.com/abstract=3072038

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.[23] 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 information on the cost of a loan (this is only reported for "high cost" loans).[24]

The McDash^{TM} dataset from Black Knight contains much more information on the contract and borrower characteristics of loans, including mortgage interest rates. Of course, these data are only available for originated loans, which the dataset follows over time. The dataset also contains a monthly indicator of a loan's delinquency status, which has made it one of the primary datasets that researchers have used to study mortgage default (e.g., Elul 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

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

[24] Bhutta and Ringo (2014) and Bayer et al. (2018) 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. Since the 2018 reporting year, additional information has been collected under HMDA; see http://files.consumerfinance.gov/f/201510_cfpb_hmda-summary-of-reportable-data.pdf for details.

Electronic copy available at: https://ssrn.com/abstract=3072038

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.[25] We thus focus on loans originated after the end of the housing boom, which (unlike earlier vintages) did not experience severe declines in house prices. Indeed, most borrowers in our data experienced positive house price growth throughout the sample period. This means that delinquency is likely driven to a large extent by idiosyncratic borrower shocks rather than macro shocks, mapping more closely to our theoretical discussion.

For the origination vintages from 2009-2013, our HMDA-McDash dataset corresponds to 45% of all loans in HMDA. This fraction is driven by the coverage of McDash (corresponding to 73% of HMDA originations over this period) and the share of these McDash loans that can be uniquely matched to the HMDA loans (just over 60%). For our analysis, we impose some additional sample restrictions. We only retain conventional (non-government issued) fixed-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.

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

Electronic copy available at: https://ssrn.com/abstract=3072038

There are 42.2 million originated mortgages on 1-4 family properties (incl. manufactured homes) 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.[26] 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 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

[26]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."(We later check robustness to dropping this category prior to estimation.) 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.

Electronic copy available at: https://ssrn.com/abstract=3072038

of estimated default probabilities) under the new technology relative to the old.

Even though the sample we use for our analysis covers a sizable portion of US mortgage originations during the sample period, one may still be concerned that our sample is not fully representative. In the online appendix we show that at least based on some key characteristics like income and loan amount, there is no evidence that the distributions of these variables across and within groups are not representative of the market (as measured in HMDA). We also verify the robustness of our empirical results to a number of changes to the sample, as we describe later in the paper.

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 another point regarding our data and the US mortgage market more

Electronic copy available at: https://ssrn.com/abstract=3072038

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.[27] 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.[28]

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.

[27]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.

[28]One set of lenders that may have been using more sophisticated models during our sample period are "FinTech" lenders like Quicken Loans, which gained market share over the sample period. In our matched sample, we do not have lender identifiers (due to data restrictions) so we cannot, unfortunately, directly study whether those lenders appear to assess and price risk differently. However, based on the list of FinTech lenders from Buchak et al. (2018) and the full HMDA sample, we note that the market share of these FinTech lenders was very low over the first three years of our sample (2-3% of all originated loans over 2009-2011), before roughly doubling in 2012 and 2013. In our robustness checks, we show that our results are very similar if we restrict the sample to 2009-2011, making it unlikely that the patterns in the data that drive our results were due to FinTech lenders.

Electronic copy available at: https://ssrn.com/abstract=3072038

4 Estimating Probabilities of Default Using Different Statistical Technologies

In this section, we describe the different prediction methods that we employ to estimate default probabilities for originated mortgages in the our dataset. In our description of the estimation techniques, we refer to observable characteristics as x, the loan interest rate as R, and the conditional probability of default as $P(x, R) = Pr(\text{Default}|x, R)$.[29] We subsequently use these estimated default probabilities to understand the impact of different statistical technologies on mortgage lending. In the remainder of this section, given concerns of unobserved private information in issued loan interest rates R, we estimate $\hat{P}(x)$, i.e., we do not include the loan rate at origination in the set of covariates. In practice, as we later demonstrate, our inferences are not much affected by the inclusion or exclusion of R when estimating default probabilities, and we later attempt to more cleanly estimate $\hat{P}(x, R)$ for a simple back-of-the-envelope calculation of potential economic magnitudes on interest rates and loan granting decisions.

We also note here that we restrict our attention in this paper to the prediction of default probabilities. In practice, final outcomes such as interest rates should reflect not just default probabilities, but other aspects like borrowers' prepayment propensities, which may also have a group-specific component.[30] While this is certainly a shortcoming of our approach, we nevertheless abstract from this issue in our analysis, for three main reasons. First, unlike default, prepayment has ambiguous effects on the lender: the lender benefits (suffers) from faster prepayment when the rate on a loan is below (above) the prevailing market rate.

[29] We do not directly estimate lifetime probabilities of default (which are the object of interest in our models in Sections 2 and 5), but rather, three-year probabilities of default. In the online appendix, we discuss the industry-standard assumptions that we use to convert estimated three-year probabilities into lifetime probabilities of default.

[30] Borrowers in the US mortgage market, unlike in almost all other countries, generally have the option to prepay their loan at any time, without compensating the lender for lost income.

Electronic copy available at: https://ssrn.com/abstract=3072038

To the extent that new loans are issued at par, as we later assume in our model (after accounting for credit risk), prepayment propensities do not have first-order effects on loan values.[31] Second, for our purposes in this paper, differences in prepayment behavior must manifest themselves systematically across groups to affect our inferences—and any such differences will have ambiguous effects depending on whether some groups systematically prepay quicker or slower than others conditional on how the market mortgage rate compares to the rate on their current loan.[32] Third, to get a sense of how any estimated differences in prepayment behavior would affect equilibrium interest rates across groups (in the spirit of our calculations in Section 5), one would require rather complicated machinery (e.g., simulations from calibrated interest rate models) which is beyond the scope of this paper.

We now turn to the estimation approaches that we use to contrast traditional and and more sophisticated prediction technologies. 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.[33]

[31] See Gabaix et al. (2007) for a simple model illustrating these points. Boyarchenko et al. (2019) show empirically that prepayment risk premia are close to zero for mortgage-backed securities with prepayment options near-the-money, as is the case for newly issued loans. Consistent with this, in the industry, prepayment modeling is most important for the valuation of older existing loans, which may have mortgage rates well above or below current market rates.

[32] See, for instance, Keys et al. (2016) and Andersen et al. (2019) for recent studies of heterogeneity in mortgage refinancing behavior.

[33] While many different techniques can be classified as belonging (or not) to the class of "machine learning" models, we simply seek to shed light on the effects of access to a flexible nonlinear technology unencumbered by concerns of overfitting. This guides the contrast that we draw between more traditional Logit-based approaches and the Random Forest implemented and tuned with cross-validation and calibration. We also employ the eXtreme Gradient Boosting (XGBoost) model (Chen and Guestrin, 2016), which delivers very similar results to the Random Forest—we describe this alternative model to the online appendix.

Electronic copy available at: https://ssrn.com/abstract=3072038

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:

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

We estimate the model in two ways, varying 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, and 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 the "Logit"in what follows.

In the second type of Logit model, we allow for a more flexible use of the information in the covariates in x, reflecting standard industry practice. In particular, we include the same dummies as in the first model, but instead of all continuous variables entering the model for the log-odds ratio linearly, we bin some of them to allow for the possibility of nonlinear relationships. 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 frequent 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".[34]

[34]We later check robustness to allowing for an even richer set of right-hand-side variables in the Nonlinear Logit, adding interactions between FICO and LTV bins, and further interacting these bins with loan purpose, term, and documentation type. Our inferences are not greatly affected by this change.

Electronic copy available at: https://ssrn.com/abstract=3072038

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	
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 main models. Section 4.5 considers additional specifications. Data source: HMDA-McDash matched dataset of conventional fixed-rate mortgages.

4.2 Tree-Based Models

As an alternative to the traditional models described above, we use machine learning models to estimate $\hat{P}(x)$. 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

Electronic copy available at: https://ssrn.com/abstract=3072038

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.[35]

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

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

Electronic copy available at: https://ssrn.com/abstract=3072038

process can flexibly identify *interactions* between covariates, i.e., bins that identify regions defined by (possibly nonlinear functions of) multiple variables simultaneously, rather than restricting the covariates to enter additively into the link function, or specifying the variable interactions up-front, as with 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 (for textbook treatments, see, e.g., Hastie et al. 2009, and James et al. 2013).

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

Electronic copy available at: https://ssrn.com/abstract=3072038

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 $K = 3$).[36]

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.[37] Our procedure selects a minimum number of observations to split of 200 and requires at least 100 observations in each 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 whether or not the borrower is predicted to default. For many purposes, including credit evaluation, the *probability* of belonging to a class (i.e., the default probability) is also needed, to set interest rates, for example. We therefore need to convert

[36] The choice of the hyperparameter K involves a trade-off between computational speed and variance; with a smaller K, there will be more variance in our estimates of model fit, as we will have fewer 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.

[37] We define our grid from 2 to 500 in increments of 50 (i.e., 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 in a leaf (*min_samples_leaf*).

Electronic copy available at: https://ssrn.com/abstract=3072038

predicted class labels into predicted loan default probabilities to serve as inputs into a model of lending decisions.

In tree-based models such as the Random Forest model, we could estimate these probabilities by counting the fraction of predicted defaults in the training dataset associated with the leaf into which a new borrower is classified. However, such estimates tend to be very noisy, as leaves are optimized for purity, and there are often sparse observations in any given leaf. A frequently used alternative 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").[38] We employ isotonic regression calibration to translate the predicted classifications into probability estimates, providing more details of this procedure in the online appendix.

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.

[38] 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.

Electronic copy available at: https://ssrn.com/abstract=3072038

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.[39]

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).[40]

One drawback of the AUC is that it is less informative in datasets which are sparse in defaulters, since FPRs are naturally low in datasets of this nature (see, for example, Davis and Goadrich, 2006). We therefore also compute the *Precision* of each classifier, calculated as $P(y = 1|\hat{y} = 1)$, and the *Recall*, as $P(\hat{y} = 1|y = 1)$,[41] and draw Precision-Recall curves which plot Precision against Recall for different probability thresholds. To summarize these Precision-Recall curves, we report the average Precision score, which calculates the weighted

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

[40] 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.

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

Electronic copy available at: https://ssrn.com/abstract=3072038

mean of Precision, with weights corresponding to the trade-off in Recall.[42]

Two additional measures we compute are the Brier Score and the R^2. The 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. The Brier Score can be decomposed into three components:

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

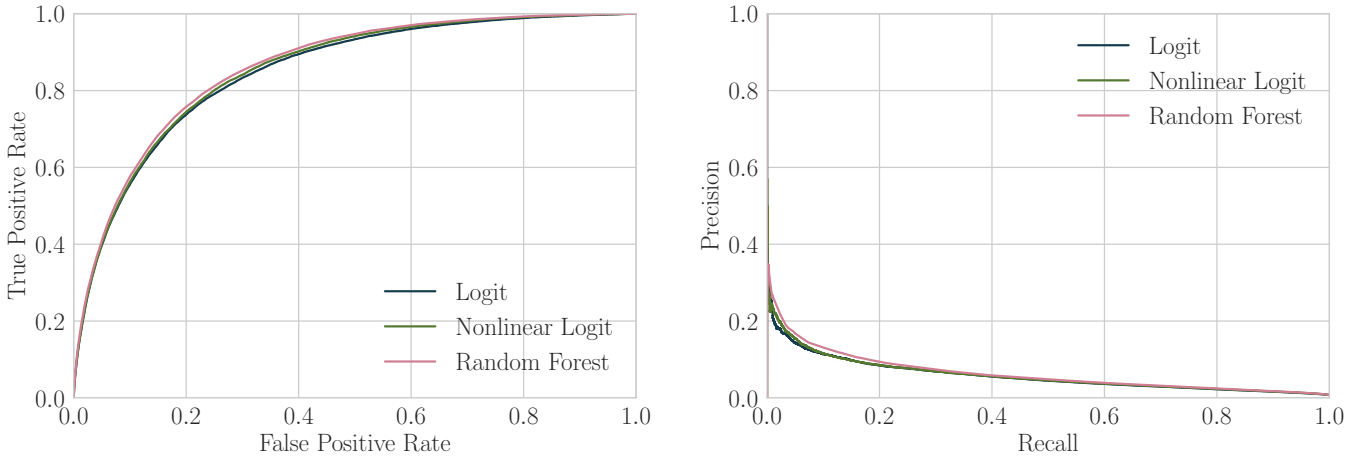
where the predicted values are grouped into K discrete bins, $\hat{y}_k$ is the predicted value within the kth bin, and $\bar{y}_k$ is the true mean predicted value within the kth bin. Uncertainty measures an inherent feature of the outcomes in the prediction problem, Reliability is a measure of the model's calibration, i.e., the distance between the predicted probabilities and the true probabilities, and Resolution is a measure of the spread of the predictions. Larger Resolution is better, while smaller Reliability implies a smaller overall error. In our application, overall Uncertainty is 0.00725, and tends to dominate the overall value of the Brier Score. Finally, the R^2 is calculated as one minus the sum of squared residuals under the model, scaled by the sum of squared residuals from using the simple mean, with the usual 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 show 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 for the traditional models for every

[42]Specifically, average Precision $= \sum_n (R_n - R_{n-1})P_n$, where n denotes each point on the Precision-Recall curves, and R_n and P_n each denote the recall and precision at point n.

Electronic copy available at: https://ssrn.com/abstract=3072038

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



Panel A: ROC

Panel B: Precision-Recall

level of FPR. In Panel B, the Precision-Recall curves, which are better suited for evaluating models on the kind of dataset we consider (sparse in defaulters) show stronger gains for the Random Forest model over the Logit models.

Columns (1), (4), (7), and (10) of Table 3 show that the Random Forest model performs better than the Logit model, suggesting that the machine learning model more efficiently uses information in the training dataset to generate more accurate predictions out of sample. The Random Forest outperforms the Nonlinear Logit model by 5.1% in terms of average precision, 0.8% in terms of AUC, and 14.3% in terms of R^2. The Brier Score, as mentioned, is dominated by the overall uncertainty of the outcome (which is driven by the overall probability of default). After adjusting for this, the change from Nonlinear Logit to Random Forest is more substantial (14.3%), identical to the R^2 improvement. The decomposition revelas that the gains from switching to Random Forest in Reliability are large, with a reduction in Reliability error of 94.3%, but at the cost of a small (3.2%) decline in Resolution.[43]

[43]This result is consistent with Figure A-2 in the online appendix, where we see more spread in the predictions of the Logit model, but slightly worse calibration. When calculating the Brier Score decomposition, we use bins of size 0.0001 and restrict our predicted values in the model to less than 0.15 to avoid the influence of outliers.

Electronic copy available at: https://ssrn.com/abstract=3072038

Electronic copy available at: https://ssrn.com/abstract=3072038

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

	ROC AUC			Precision Score			Brier Score × 100			R^2		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Model	No Race	Race	Race Int.	No Race	Race	Race Int.	No Race	Race	Race Int.	No Race	Race	Race Int.
Logit	0.8486	0.8491	0.8499	0.0579	0.0582	0.0589	0.7181	0.7179	0.7173	0.0233	0.0234	0.0243
Nonlinear Logit	0.8537	0.8541	0.8543	0.0590	0.0593	0.0600	0.7149	0.7148	0.7147	0.0275	0.0277	0.0279
Random Forest	0.8602	0.8610		0.0620	0.0622		0.7120	0.7118		0.0315	0.0318	

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 columns (1), (4), (7), and (10), race indicators are not included in the prediction models; in the other columns, they are included. In columns (2), (5), (8), and (11), the Logit models include the race indicators as simple dummy variables, while in columns (3), (6), (9), and (12), separate Logit models are estimated for each race group (which is equivalent to fully interacting the race indicators with all other variables).

The differences in performance metrics appear modest, and we verify that they are indeed statistically significant using bootstrapping. We hold fixed our estimated models, and randomly resample with replacement 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. Across all metrics, the Random Forest shows better performance than the Nonlinear Logit in 100% of these bootstrap samples.[44] Overall, we conclude with considerable statistical confidence that the machine learning models statistically significantly improve default prediction performance.

4.3.1 Model Performance With and Without Race

Columns (2), (5), (8), and (11) of Table 3 simply add race dummies to the Logit models, and show that this inclusion has positive effects on the performance of the traditional Logit models. In columns (3), (6), (9), and (12), we allow each variable in the Logit models to be *interacted* with the race dummies, which is equivalent to estimating separate Logits for each group. The performance increase from this change is even greater. The table also shows how the more sophisticated machine learning model uses race as an explanatory variable—the model flexibly picks race and the interactions of race with the other included variables, so we simply show the single measure of performance from adding race to the set of features. We find that even this more sophisticated model benefits from the inclusion of race as an explanatory variable.

That having been said, it is worth noting that while all models benefit from the inclusion of race either as a dummy, or when interacted with the other variables in the model, 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

[44] 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.

Electronic copy available at: https://ssrn.com/abstract=3072038

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 augment performance for all models 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 continues to reveal differences between the models—Table 4 confirms that the Random Forest outperforms the other two models, which have very similar scores, by 6.7% in terms of average precision, 0.6% in terms of AUC, 1.1% in terms of Brier score, and 10.8% 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.[45] Whether this ability contributes to triangulation will depend on whether there is considerable variation in true default propensities across race and ethnic groups unrelated (or at least unrelated up to the detection capabilities of the more sophisticated nonlinear model) to observable characteristics. 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

[45]Interestingly, if we redo this exercise including the interest rate R as an additional predictor, we are able to predict race even more precisely using the Random Forest model, with an R^2 improvement of 30% relative to Nonlinear Logit.

Electronic copy available at: https://ssrn.com/abstract=3072038

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

Model	ROC AUC	Precision Score	Brier Score × 10	R^2
Logit	0.7436	0.1837	0.5792	0.0607
Nonlinear Logit	0.7442	0.1857	0.5785	0.0619
Random Forest	0.7485	0.1983	0.5738	0.0694

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, to better understand the differences in distributional consequences between the traditional and sophisticated models.

4.4 Differences in Predicted Default Propensities

We begin our analysis of distributional consequences with Figure 4. This figure, which is based on a subset of the data, illustrates the estimated models and extends our motivating theoretical example in Figure 1 to two dimensions. Each panel of the figure plots level sets (or isoquants) of predicted default propensities as a function of borrower income on the horizontal axis, and FICO score on the vertical axis. We fix other borrower characteristics in the subset of the data, and then obtain predicted default probabilities from our estimated models by combining this fixed constellation with every income-FICO combination on a two-dimensional grid.[46] The level sets of predicted default probabilities for the Nonlinear Logit model are in the top two panels, and for the Random Forest, in the bottom two panels. 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 left panels, and White non-Hispanic, Asian, and other borrowers in the right panels, with darker colors

[46]Specifically, we vary income and FICO for 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.

Electronic copy available at: https://ssrn.com/abstract=3072038

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 suggests (similarly to our theoretical example) that at least for the restricted sample that we consider in these plots, new technology has unequal impacts across racial groups.

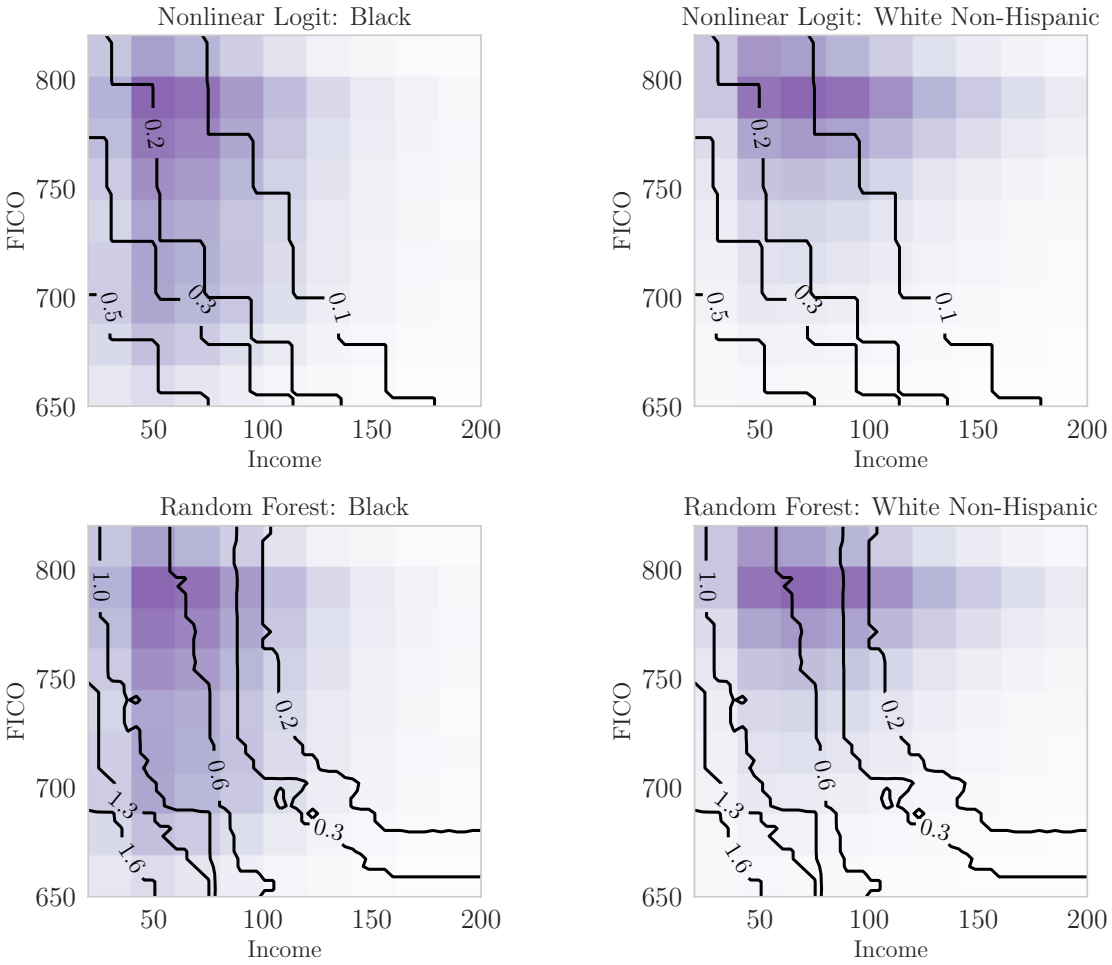
This figure, while intriguing, needs to be backed up with more comprehensive evidence from the full dataset. In particular, many "leaves" of our underlying tree models (e.g., leaves associated with LTV ratios below 70), and many possible interactions (e.g., between FICO and LTV) never come into play in the figure.[47] To more rigorously assess who wins and who loses from the introduction of new technology, we now turn to studying the change in the entire distribution of predicted default propensities in our test sample as estimation technology varies.

In Figure 5, we look at the differences between the entire set of predicted probabilities of default (PDs) from the machine learning model and those from 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

[47]In addition, for the particular constellation of characteristics we have considered in Figure 4, the Random Forest predicts uniformly higher default probabilities than Logit. This estimated difference seen in this subset of the data (used for illustrative purposes) is not representative of the overall population, where the sample average predicted probability of default is 0.74 percent for both Random Forest and Logit predictions, which also coincides with the true average probability of default in our data.

Electronic copy available at: https://ssrn.com/abstract=3072038

Figure 4: **Example of Predicted Default Probabilities Across Models**



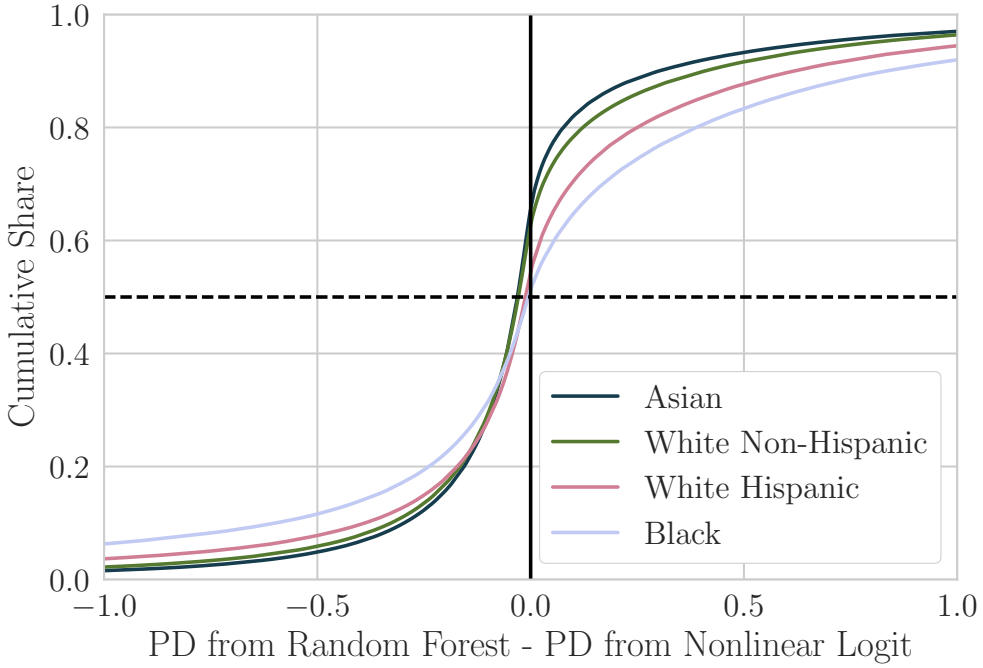
Note: Figure shows level sets of default probabilities (in %) predicted from different statistical models for different values of borrower income and FICO (holding other characteristics fixed as explained in text). Nonlinear Logit predictions are shown in the top row; Random Forest predictions in bottom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

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 the *proportional* benefit for each group.[48]

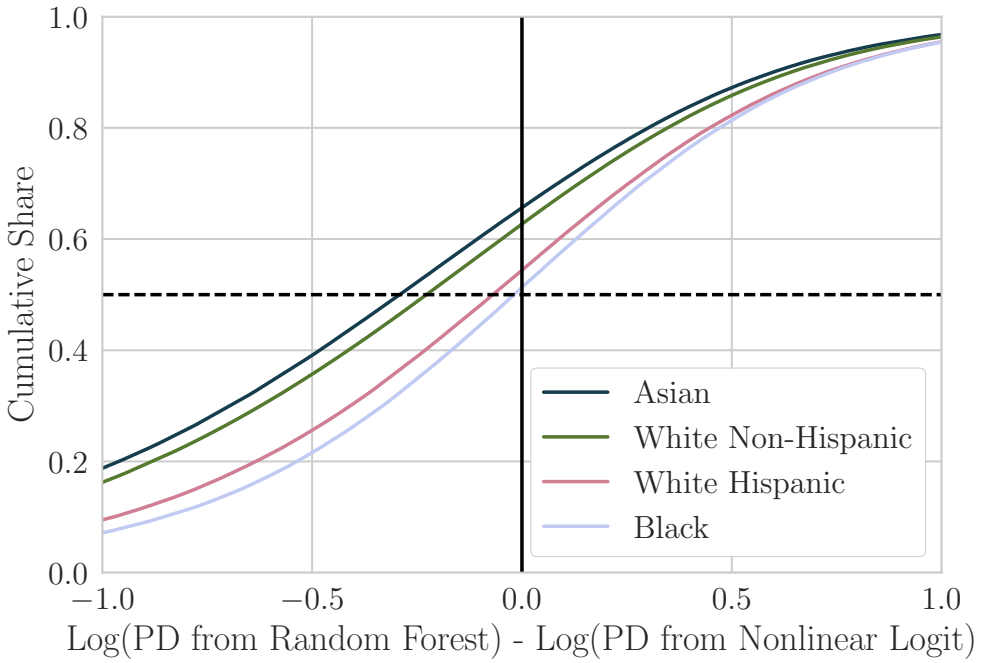
[48]For ease of visual representation, we truncate the x-axes on these plots, as there is a small fraction of cases for which the estimated differences in the default probabilities are substantial.

Electronic copy available at: https://ssrn.com/abstract=3072038

Figure 5: **Comparison of Predicted Default Probabilities Across Models, by Race Groups**



Panel A



Panel B

Electronic copy available at: https://ssrn.com/abstract=3072038

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* groups, the share of borrowers for whom 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 distribution of relative PD differences is 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 changes in predicted default probabilities from the Random Forest model has larger variance than under the Nonlinear Logit model.[49] We return to this finding later, but simply note here that this can be interpreted as introducing an additional element of risk for borrowers in these minority 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. Before proceeding to analyzing this finding further, we first check its robustness.

[49]The distributions are also right-skewed, i.e., the Random Forest model predicts far higher PDs than the Logit model for some borrowers in all groups.

Electronic copy available at: https://ssrn.com/abstract=3072038

4.5 Alternative Specifications and Robustness of the Models

We conduct a number of checks to our results, to verify their robustness to a number of changes to the covariates employed in the models, alternative samples of data, and changes to the specifications employed in our comparisons.

Our first check is conducted in Figure A-3 in the online appendix, which shows the plots in Figure 5, but replacing the predictions from the Random Forest model with those from the XGBoost model (also explained in the online appendix). 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.

Table 5 summarizes the results from various other robustness checks in a consistent format. The rows show the outcome variables that we report for each robustness check—these are across columns (2)-(8), with column (1) showing these outcomes in the baseline version described in the previous section. The outcome variables are, in Panel A, the differences in out-of-sample performance statistics between the Random Forest and Nonlinear Logit model; in Panel B, the percent share of borrowers that appear more creditworthy (i.e., those assigned a lower PD) in the Random Forest versus the Nonlinear Logit model, as a simple way to summarize the differences detected between groups in Figure 5; and in Panel C, the standard deviations of the changes in the PDs between the two models, as a way to summarize the variance in PDs faced by a borrower in each group when moving from the traditional to the more sophisticated technology.

Columns (2) adds interest rates to the set of covariates, and shows that there are insubstantial changes to the baseline outcomes arising from this addition. Column (3) shows what happens when we remove FICO from the baseline covariates—the performance differences here between the machine learning and Logit model are now more substantial, suggesting

Electronic copy available at: https://ssrn.com/abstract=3072038

that the FICO score already contains some of the information gain arising from nonlinearly combining the other covariates. However, while attenuated, it is still the case that the majority White group gains under the machine learning model in terms of lower average PDs, and the standard deviation of PDs is also still higher for the minority groups under the machine learning model.

The next four columns consider changes to the sample and their effects on our inferences. Column (4) removes the observations with "Unknown" race (10.5% of our sample); column (5) re-does our analysis using only the 2009-2011 period in which the market share of Fintech lenders (taken from Buchak et al., 2018) was very low (2-3% of originated loans); column (6) re-does our analysis only for the majority White group; and column (7) restricts the sample to only GSE-backed loans with full documentation (a subsample we will return to in the next section). Columns (4), (5), and (7) show that our inferences remain very similar despite changes to the sample, and column (6) shows that the performance differences are also of a similar magnitude across the machine learning and traditional models when only the majority group is utilized. We return to this finding once again when discussing triangulation and flexibility below.

Finally, in column (8) of Table 5 we use a more flexible version of the Nonlinear Logit: we fully interact the FICO and LTV bin indicators with one another, and further interact the resulting terms with separate dummies for the loan being for a home purchase, being underwritten with full documentation, and having a duration of 15 years or less. If this added flexibility improved the predictive performance of the Logit, we would expect the performance metric differences in panel A to decrease relative to column (1); however, we see that this is only the case for Precision, but not the other three metrics. This indicates that according to the ROC-AUC, R^2, and Brier score metrics, the more complex version of the Nonlinear Logit actually performs worse out of sample than our baseline version. This is perhaps unsurprising given the fact that the more complex Logit model is likely to suffer

Electronic copy available at: https://ssrn.com/abstract=3072038

more from concerns of overfitting. In contrast, the Random Forest model naturally controls variance of the out-of-sample predictions.

Having confirmed that the differences between groups associated with the different models are robust, we seek a better understanding of the sources of these unequal effects in the data. We turn to this in the next subsection.

Electronic copy available at: https://ssrn.com/abstract=3072038

Table 5: **Alternative Models and Subsamples of Data**

		Change Covariates		Alternative Samples				Change Spec.
	Baseline	Int. Rates	No Fico	No Unknowns	2009-2011	White	GSE+Full Doc	More Interactions
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Percentage difference in performance statistics between Random Forest and Nonlinear Logit								
ROC	0.77	0.76	1.64	0.73	0.62	0.60	0.58	0.79
Precision	5.06	4.83	11.57	4.33	1.34	3.36	1.56	2.70
Brier Score	-0.40	-0.45	-0.30	-0.41	-0.40	-0.35	-0.30	-0.66
R^2	14.30	15.64	21.28	14.78	12.06	13.71	11.74	25.66
Panel B: Percent share of borrowers that look more creditworthy in Random Forest vs. Nonlinear Logit								
White Non-Hisp.	12.68	13.07	5.59	12.68	12.69	11.60	9.33	7.10
Asian	15.62	16.14	5.09	15.76	15.44	-	12.00	8.57
Black	1.29	1.04	-0.49	1.48	1.58	-	1.50	-1.44
White Hispanic	4.40	3.77	-0.00	4.54	6.47	-	2.91	-0.30
Panel C: Standard deviation of Δ probability of default between Random Forest vs. Nonlinear Logit								
White Non-Hisp.	0.0071	0.0072	0.0057	0.0071	0.0082	0.0073	0.0061	0.0067
Asian	0.0061	0.0061	0.0052	0.0062	0.0074	-	0.0057	0.0057
Black	0.0136	0.0139	0.0084	0.0137	0.0161	-	0.0120	0.0122
White Hispanic	0.0102	0.0103	0.0069	0.0104	0.0128	-	0.0089	0.0092
# Obs. used for estimation	4,591,292	4,591,292	4,591,292	4,108,819	2,757,163	3,495,156	2,522,670	4,591,292
# Obs. used for testing	2,810,996	2,810,996	2,810,996	2,515,826	1,687,444	2,140,187	1,545,697	2,810,996

Note: Table shows improvements in performance metrics when going from Nonlinear Logit to Random Forest (panel A), the percentage of borrowers (by race group) who get assigned a lower PD by the Random Forest than by the Nonlinear Logit (panel B), and the standard deviation (by race group) of changes in PD when going from Random Forest to Nonlinear Logit (panel C), for different model versions or estimation samples, as described in the text. The last two rows show for each column the number of observation used for the estimation of the model, and the number of observations used for the out-of-sample evaluation (testing).

Electronic copy available at: https://ssrn.com/abstract=3072038

4.6 Flexibility and Triangulation in the Data

In Section 2, we discussed how 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 in an attempt 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. In this case, 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 from our empirical results 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* allowing for the use of 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 15.6% ($\frac{0.0318-0.0275}{0.0275}$). 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

Electronic copy available at: https://ssrn.com/abstract=3072038

of both ways in Table 6. 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 simply adding race dummies to the set of variables available to the Nonlinear Logit, without changing technologies. For example, around 3.4% of the total performance improvement in terms of R^2 (or Brier Score) arise from the inclusion of race dummies as covariates in the Nonlinear Logit model.[50] The second column of the table then allows for the addition of race dummies to the Nonlinear Logit in a more flexible manner, interacting them with all other borrower/loan characteristics, and this improves the performance metrics even further, with an incremental gain of 4.4% in R^2 (and substantially higher for the Precision measure). Finally, the rightmost column of Panel A shows the complement of the sum of the fractions in the previous two columns, which we interpret as the fraction of the total performance improvement attributable to increased flexibility, conditional on the improvement achieved by simply adding race in naïve and more flexible ways. For example, moving from the Nonlinear Logit model with race interactions to the Random Forest model with race delivers roughly 92% of the total 15.6% improvement in R^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 right column shows its complement, attributable to the improvements in performance from adding race conditional on already having a flexible model (i.e., moving from Random Forest without race to Random Forest with race). For two of three performance measures, a larger fraction of the improvement is now attributed to race conditional on new technology than in the Nonlinear Logit model. This is not too surprising, as additional interactions between race and other observables are

[50]Note that the decomposition for R^2 and Brier Score are identical due to their arithmetic relationship. Since $R^2 = 1 - (\text{Brier}/(Pr(Default) \times (1 - Pr(Default))))$, any percentage comparison of R_i^2 and R_j^2 where $Pr(Default)$ is fixed is identical to comparing the Brier Scores.

Electronic copy available at: https://ssrn.com/abstract=3072038

being utilized by the Random Forest. This result suggests that machine learning models capitalize on interactive effects between race and other characteristics.

Table 6: **Decomposition of Performance Improvement**

	Race	Race Int.	Technology
ROC-AUC	6.28	2.04	91.69
Precision	9.05	22.43	68.52
R^2 / Brier Score	3.37	4.39	92.24

Panel A: Race Controls First

	Technology	Race
ROC-AUC	89.77	10.23
Precision	94.14	5.86
R^2 / Brier Score	92.95	7.05

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 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 suggests that when predicting mortgage default, triangulation alone is not at the heart of the performance improvements from machine learning, although triangulation does emerge as a non-trivial factor according to all metrics, and a substantial factor according to one of the three performance metrics, Precision. Indeed, the numbers in the first and second columns of Panel A suggest that knowing race (which is the best that triangulation without additional flexibility could achieve) could yield up to 31.5% of the total performance improvement, though the averages across performance metrics suggest smaller increases.

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

Electronic copy available at: https://ssrn.com/abstract=3072038

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 even larger effects of triangulation.

The fact that unequal effects in this exercise appear to mainly be 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. This flexibility allows new technology to map distributions of characteristics such as income and FICO scores (themselves affected by unobservable, potentially historic determinants correlated with group status) to default probabilities.

5 Illustration of Equilibrium Effects

Our discussion so far has focused on the impact of machine learning technology on predicted probabilities of default across different groups of borrowers. In credit markets, these effects can translate into changes in lenders' decisions both on the extensive margin (i.e., which borrowers to accept for credit, and which to reject) and on the intensive margin (i.e., how to price loans to borrowers who are accepted). To arrive at a tractable "back-of-the-envelope" computation of how these effects on default probabilities might map to these final outcomes, we consider two exercises. Our main approach, described in this section, consists of a simple calibrated one-period equilibrium model.[51] As a reduced-form alternative, we empirically

[51]The online appendix contains a more rigorous analysis of this model, including a basic discussion of how the analysis might change if we were to instead consider the steady state of a dynamic model of credit provision rather than a one-period model.

Electronic copy available at: https://ssrn.com/abstract=3072038

estimate a mapping from default probabilities to interest rates, under the assumption that observed interest rates in our sample were generated by the Logit technology. We then study how the distribution of rates across groups would change if lenders moved to Random Forest technology but kept the mapping constant. The results from this exercise, described in detail in the online appendix, are qualitatively consistent with what follows.

5.1 A Simple Model of Mortgage Market Equilibrium

Our model is not intended for full structural estimation, but rather as an illustration of the potential effects of new statistical technology. To facilitate this exercise, we make three strong assumptions.

First, we assume that the market for mortgages is competitive. Lenders set interest rates R contingent on borrowers' observable characteristics x. Lenders are risk-neutral and discount future repayments at rate ρ. All non-price characteristics of mortgage contracts, such as the loan amount and LTV ratio, are pre-determined for each borrower.[52]

Second, we assume that lenders have access to a statistical technology which provides an unbiased estimate $\hat{P}(x, R)$ of the structural probability of default when a borrower has characteristics x and accepts a mortgage with interest rate R. The dependence of $\hat{P}$ on R captures, in reduced form, the fact that a higher required monthly payment makes it more likely that a borrower finds it beneficial to default on their loan (e.g., after a negative income shock); we discuss potential problems with estimating the relationship between default and R below. Lenders in the model are fully rational, but constrained in their mapping of characteristics and rates to default probabilities by the available statistical technology. As

[52] In reality, these parameters are often dictated, or at least confined to a narrow range, by local property prices and borrower liquidity constraints. In the online appendix, we discuss how endogenous contracting terms would affect our results. 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).

Electronic copy available at: https://ssrn.com/abstract=3072038

technology varies, the mapping that lenders use also changes—this allows us to plug in the estimated functions $\hat{P}(x, R)$ from various technologies into equilibrium conditions to arrive at our back-of-the-envelope estimates.

Third, as discussed earlier, we assume no possibility of prepayment. Each mortgage is either repaid in full at the end of its term, for a repayment of $1 + R$ per dollar of the loan, or ends in default, for a repayment of $1 + R - LGD(x, R)$. Here, $LGD(x, R)$ stands for the loss given default for a borrower with characteristics x, which we assume to be deterministic.

Under these assumptions, each lender estimates net present value (NPV) of offering a mortgage with rate R, per dollar of the loan amount, as:

$$N(x, R) = \frac{1}{1+\rho}\left[1 + R - \hat{P}(x, R) \cdot LGD(x, R)\right] - 1, \tag{2}$$

where $LGD(x, R)$ is the loss given default for a borrower with characteristics x, which we assume to be deterministic.

In a Bertrand equilibrium, borrowers with characteristics x are offered the lowest interest rate $R(x)$ that satisfies the break-even condition $N(x, R(x)) = 0$. Because default probabilities depend on interest rates, it is possible that no break-even rate exists. Intuitively, this has the flavor of market unravelling. If default risk $P(x, R)$ is high and, in addition, if it increases in R, then lenders cannot break even at low interest rates, but also cannot break even by charging higher rates, since doing so would exacerbate default risk. In that case, we say that borrowers with characteristics x are rejected for a loan.

5.2 Empirical Implementation

To explore the equilibrium implications of machine learning in the data, we solve for equilibrium rates $R(x)$ (or rejection, when no break-even R exists) implied by different statistical

Electronic copy available at: https://ssrn.com/abstract=3072038

technologies for borrowers in our dataset. Concretely, we plug predicted default probabilities $\hat{P}(x, R)$ from both traditional technology and machine learning models into equation (2). We further calibrate the lender's discount rate ρ and the loss given default $LGD(x, R)$ directly based on empirical data from the mortgage market. Finally, we solve the equation $N(x, R(x)) = 0$ to find the equilibrium (break-even) interest rate. The online appendix describes the computational approach and our calibration in detail.

This approach requires us to re-run the predictive exercise above, using models that also include the interest rate R as a predictive variable. While our empirical results do not greatly differ between models that do and do not contain observed interest rates (as shown in Table 5), adding R is conceptually questionable because interest rates in our data are not randomly assigned. In order to correctly estimate the sensitivity of default probabilities to interest rates (even in a machine learning model), one needs to assume *no selection on unobservables*: i.e., interest rates need to be assigned independently of default outcomes *conditional on controlling for observable characteristics* x. We state the conditions required for identification more formally in the online appendix, where we also discuss the related issue of adverse or advantageous selection by borrowers.

In reality, a key threat to identification is that interest rates could partly reflect lenders' "soft" information about borrowers—which is unobservable from our perspective. To try to address this issue, we take 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—Table 5 column (7) shows how our results on PDs look in this sample. In this segment, soft information is less likely to be important because lenders mainly focus on whether a borrower fulfils the underwriting criteria set by the GSEs.[53]

[53] 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 informa-

Electronic copy available at: https://ssrn.com/abstract=3072038

Second, we rely on and extend existing work by Fuster and Willen (2017) to adjust our estimates of $\hat{P}(x, R)$ for any residual bias. We do this by estimating the bias as the difference between causal and non-causal estimated effects of interest rate changes on mortgage default.[54]

We finally note that in our data, we cannot observe borrowers not granted mortgages, and therefore restrict our counterfactual statements to populations with distributions of characteristics identical to the one we observe. 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.[55]

5.3 Equilibrium Results

Table 7 summarizes our computation of lending and pricing decisions from the model. 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

tion is likely more important), not for full-documentation loans that we consider.

[54]The adjustment is described in detail in the online appendix, which also describes other relevant details of our equilibrium calculations.

[55]We note that GSE policies have likely led to more loan originations than a purely private market (as assumed in our exercise) would, which helps reduce concerns about extensive margin biases.

Electronic copy available at: https://ssrn.com/abstract=3072038

mortgage increases by about 0.9 percentage points when lenders use Random Forest instead of Logit.[56] 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.

The average interest rate spread when moving to the more sophisticated model falls slightly (by 2 bp) for Asian borrowers and increases (by 4 bp) for Black borrowers. The cross-group standard deviation of spreads increases by almost 50% relative to its baseline value of 2bp under the Logit model. That is, average pricing effects implied by our simple model are also unequal across race groups, consistent with our results on predicted default probabilities in Section 4. While a 6bp differential change in the interest rate only has a modest effect on the resulting interest payment—about $120 over the first year of a typical new mortgage[57]—the total effect across groups is still not negligible, given the size of the US mortgage market.[58]

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, we compute that the dispersion of rates increases by about 21% ($= \frac{0.360-0.298}{0.298}$).

[56] In the online appendix, we further show that the aggregate increase in acceptance rates masks some heterogeneity: About 1.7 percent of the population are newly excluded when moving to the machine learning model, while about 2.6 percent are newly included.

[57] Assumption: 30-year fixed-rate mortgage over $200,000 with baseline interest rate of 4.5%.

[58] For instance, over our sample period 2009-13, 1.8 million mortgages were originated by Black borrowers according to the HMDA data. Taking the flow of originations over these five years as a rough approximation to the stock of outstanding loans, a $100 difference in annual interest payments would correspond to a total cost to Black borrowers of $180 million per year. Using a slightly different calculation, Bartlett et al. (2019) estimate that an increase in the interest rate paid by Latinx and African-American borrowers of 4.6 bp—corresponding to Bartlett et al.'s estimated weighted average discrimination across purchase and refinance loans—costs these borrowers $765 million per year.

Electronic copy available at: https://ssrn.com/abstract=3072038

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 White Hispanic and Black 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.

Table 7: **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		

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. We also note that to take the next step relative to the simple exercise we conduct here, and to arrive at a more realistic set of magnitudes, one would need (1) purely random variation (rather

Electronic copy available at: https://ssrn.com/abstract=3072038

than the bias corrections that we employ) in interest rates to correctly estimate $\hat{P}(x, R)$ and calibrate equilibrium outcomes given a baseline dataset; (2) a dynamic model of borrower behavior in which we can simulate a new cohort of data from first-round outcomes re-calibrate equilibrium, and so forth, ideally until convergence; (3) to observe the behavior of borrowers not granted mortgages to understand the full counterfactual effects under the new technology. We discuss several of these issues in the online appendix in greater detail, but stop here with our simple analysis of likely magnitudes in this paper.

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 of our paper is that a more sophisticated statistical technology 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 machine learning technologies, we find that Black and White

Electronic copy available at: https://ssrn.com/abstract=3072038

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, from a simple equilibrium model that we set up to ascertain magnitudes, in terms of counterfactual interest rates—not only on average, but also in terms of larger within-group dispersion.

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 find that while under some metrics both flexibility and triangulation play important roles, flexibility is the primary source of unequal effects between groups in our empirical application.

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.

Electronic copy available at: https://ssrn.com/abstract=3072038

7 Appendix

7.1 Proof of Lemma 1

We write $\mathcal{L}^2$ for the space of random variables z such that $E[z^2] < \infty$, equipped with the linear product $\langle x, y \rangle = E[xy]$. Let $P(x) = E[y|x]$ and assume that $P(x) \in \mathcal{L}^2$.

Let $\hat{P}_j$ denote the projection of P onto a closed subspace $\mathcal{M}_j \subset \mathcal{L}^2$. It is well known (e.g., chapter 2 of Brockwell and Davis, 2006) that the projection $\hat{P}_j(x)$ minimizes the mean square error $E[(P(x) - \hat{P}(x))^2]$. Hence, the projection coincides with the oracle:

$$\hat{P}_j(x) = \hat{P}(x|\mathcal{M}_1)$$

We need to show that $\hat{P}_2$ is a mean-preserving spread of $\hat{P}_1$, i.e., that $\hat{P}_2 = \hat{P}_1 + u$, where $E[u] = 0$ and $Cov(u, \hat{P}_1) = 0$. By the projection theorem we have

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

for any $m \in \mathcal{M}_j$. Letting $m \equiv 1$, we obtain $E[\hat{P}_j] = E[P]$, which directly gives us $E[u] = 0$. Moreover, note that

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

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 $\hat{P}_1$.

Electronic copy available at: https://ssrn.com/abstract=3072038

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.

Andersen, S., J. Campbell, K. Meisner-Nielsen, and T. Ramadorai (2019): "Sources of Inaction in Household Finance: Evidence from the Danish Mortgage Market." Working paper, CBS, Harvard, and Imperial.

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 (2019): "Consumer-Lending Discrimination in the FinTech Era," Working paper, UC Berkeley.

Bayer, P., F. Ferreira, and S. L. Ross (2018): "What Drives Racial and Ethnic Differences in High-Cost Mortgages? The Role of High-Risk Lenders," *Review of Financial Studies*, 31, 175–205.

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 (2019): "On the Rise of FinTechs: Credit Scoring Using Digital Footprints," *Review of Financial Studies*, forthcoming.

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 A. Hizmo (2019): "Do Minorities Pay More for Mortgages?" Working Paper, Federal Reserve Board.

Electronic copy available at: https://ssrn.com/abstract=3072038

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

Boyarchenko, N., A. Fuster, and D. O. Lucca (2019): "Understanding Mortgage Spreads," *Review of Financial Studies*, 32, 3799–3850.

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 (2018): "Fintech, Regulatory Arbitrage, and the Rise of Shadow Banks," *Journal of Financial Economics*, 130, 453–483.

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.

Electronic copy available at: https://ssrn.com/abstract=3072038

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.

Friedman, J., T. Hastie, and R. Tibshirani (2001): *The Elements of Statistical Learning*, vol. 1, Springer Series in Statistics New York.

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 (2019): "The Role of Technology in Mortgage Lending," *Review of Financial Studies*, 32, 1854–1899.

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

Gabaix, X., A. Krishnamurthy, and O. Vigneron (2007): "Limits of Arbitrage: Theory and Evidence from the Mortgage-Backed Securities Market," *Journal of Finance*, 62, 557–595.

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.

Electronic copy available at: https://ssrn.com/abstract=3072038

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Science and Business Media.

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.

Keys, B. J., D. G. Pope, and J. C. Pope (2016): "Failure to Refinance," *Journal of Financial Economics*, 122, 482–499.

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.

Electronic copy available at: https://ssrn.com/abstract=3072038

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.

Vapnik, V. N. (1999): "An overview of statistical learning theory," *IEEE transactions on neural networks*, 10, 988–999.

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*.

Electronic copy available at: https://ssrn.com/abstract=3072038

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

A.1 Additional Theoretical Analysis of Unequal Effects

This appendix adopts the notation of Section 2 in the paper.

A.1.1 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:

$$E[\hat{P}(x|\mathcal{M}_2) - \hat{P}(x|\mathcal{M}_1)|g] = \sum_{j=2}^{\infty} \frac{1}{j!} \frac{\partial^j \hat{P}(a|\mathcal{M}_2)}{\partial x^j} E\left[(x-a)^j|g\right] - B \tag{3}$$

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

Electronic copy available at: https://ssrn.com/abstract=3072038

$x = a$ 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

$$\hat{P}(x|\mathcal{M}_2) - \hat{P}(x|\mathcal{M}_1) = \sum_{j=2}^{\infty} \frac{1}{j!} \frac{\partial^j \hat{P}(a|\mathcal{M}_2)}{\partial x^j}(x - a)^j - B \tag{4}$$

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

A.1.2 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

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

Electronic copy available at: https://ssrn.com/abstract=3072038

is the regression coefficient of g onto x^2.[1] 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.

A.2 Further Details on Data Representativeness

As discussed in Section 3 of the main text, the HMDA-McDash merged dataset that underlies our analysis covers 45% of all loan originations in HMDA over our sample period 2009-2013. The incomplete coverage is driven by the fact that not all loan servicers are included in McDash, and that among the loans in McDash, not all can be uniquely matched to HMDA loans. Therefore, a possible concern is that the matched loans that we end up with are not fully representative of the market as a whole, and that this may affect our conclusions, especially if our sample does not reflect the across-group and within-group variation of key variables. In this appendix, we show that there is little evidence that our sample is non-representative of the market as a whole.

For the comparison of our sample to the full HMDA data, we impose the same restrictions (to the extent possible—some of the restrictions on our analysis sample are based on information only available in McDash, such as LTV). In particular, we restrict to conventional first-lien loans on 1-4 unit properties, where the applicant income is no higher than $500,000, the loan amount no higher than $1 million, and the occupancy type not marked as "unknown."

First, Table A-1 shows that the shares of the different race and ethnic groups in the two samples are very similar, i.e. there is no evidence that our data over- or underrepresents a particular group when compared to the market as a whole. Second, we turn to summary statistics of two of the key borrower/loan characteristics that we observe in HMDA, namely borrower income and the amount of the mortgage.[2] For each race and ethnic group, we compare three key summary statistics of the distributions of these variables—mean, median, and standard deviation—between our analysis sample and the full HMDA data. Figure A-1 shows that the statistics look very similar across the two samples in terms of the ranking of the groups, and also within each group. There is no evidence that our sample exaggerates the variation across or within groups (which one could worry would lead us to overstate

[1] 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 $Cov(\gamma g, x^2)/Var(x^2) = \gamma\phi$.

[2] Other key underwriting characteristics such as FICO or LTV are not available in HMDA; neither is the interest rate (except for a small set of borrowers with rates well above the market rate). In our analysis in the main text, these characteristics come from the McDash portion of the data set.

Electronic copy available at: https://ssrn.com/abstract=3072038

Table A-1: **Share of Different Groups, Our Sample Vs. Full HMDA Data**

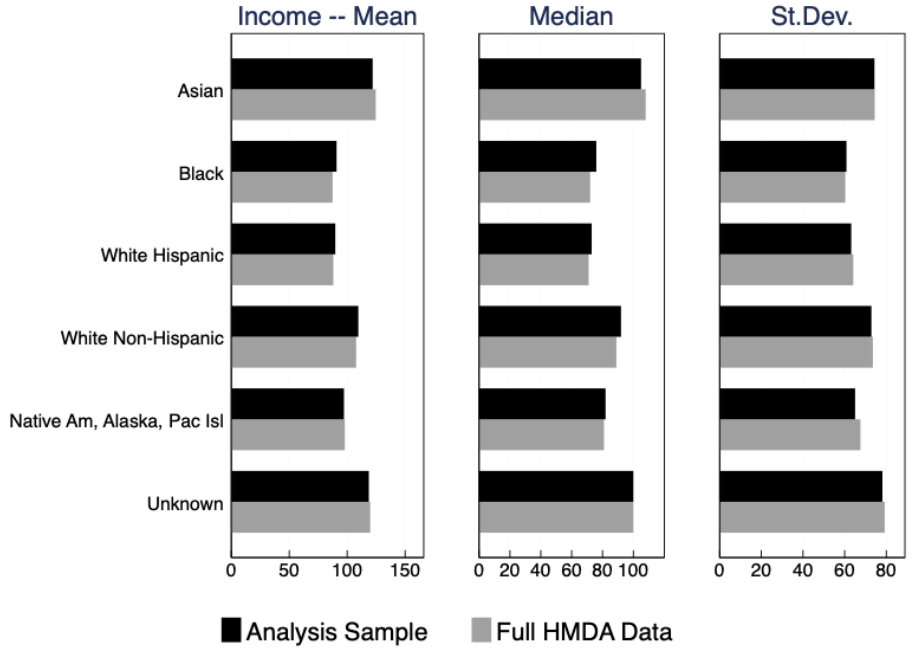
Shares in %	Our Sample	HMDA
Asian	6.13	6.50
Black	2.52	2.55
White Hispanic	4.07	4.11
White Non-Hispanic	76.14	75.48
Native Am, Alaska, Hawaii/Pac Isl	0.63	0.66
Unknown	10.50	10.70
Total	100	100

Note: Shares in our analysis sample correspond to the number of observations shown in Table 1 of the main text. Shares in HMDA data 2009-2013 are calculated after imposing the restrictions discussed in the text.

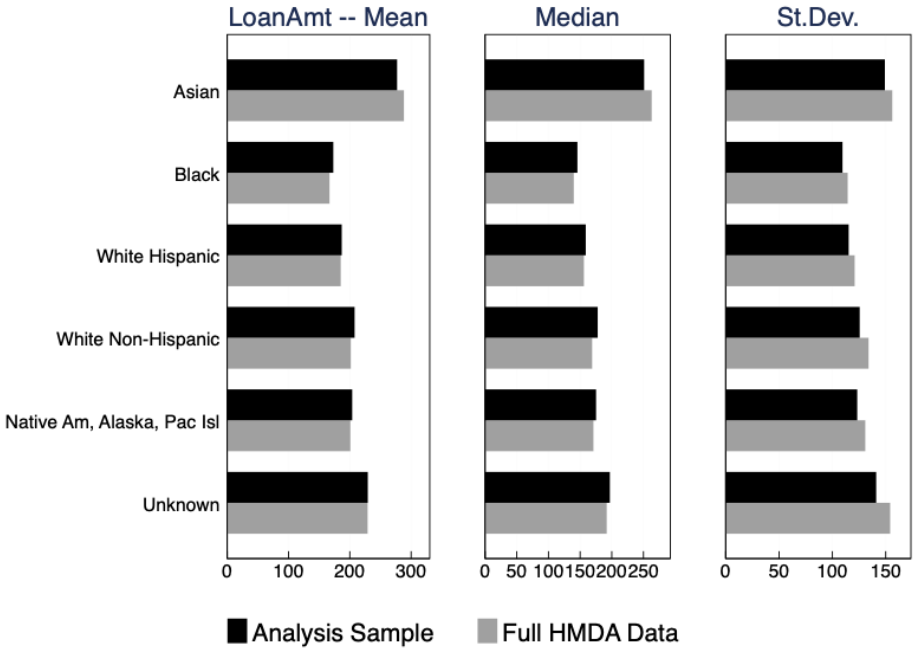
the effects of moving to more flexible statistical models). If anything, there is even more within-group variation in loan amounts in the full data than in our analysis sample.

Electronic copy available at: https://ssrn.com/abstract=3072038

Figure A-1: **Comparison of Summary Statistics on Income and Loan Amount in Our Sample Vs. Full HMDA Data**



Panel A: Borrower Income



Panel B: Loan Amount

Electronic copy available at: https://ssrn.com/abstract=3072038

A.3 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.

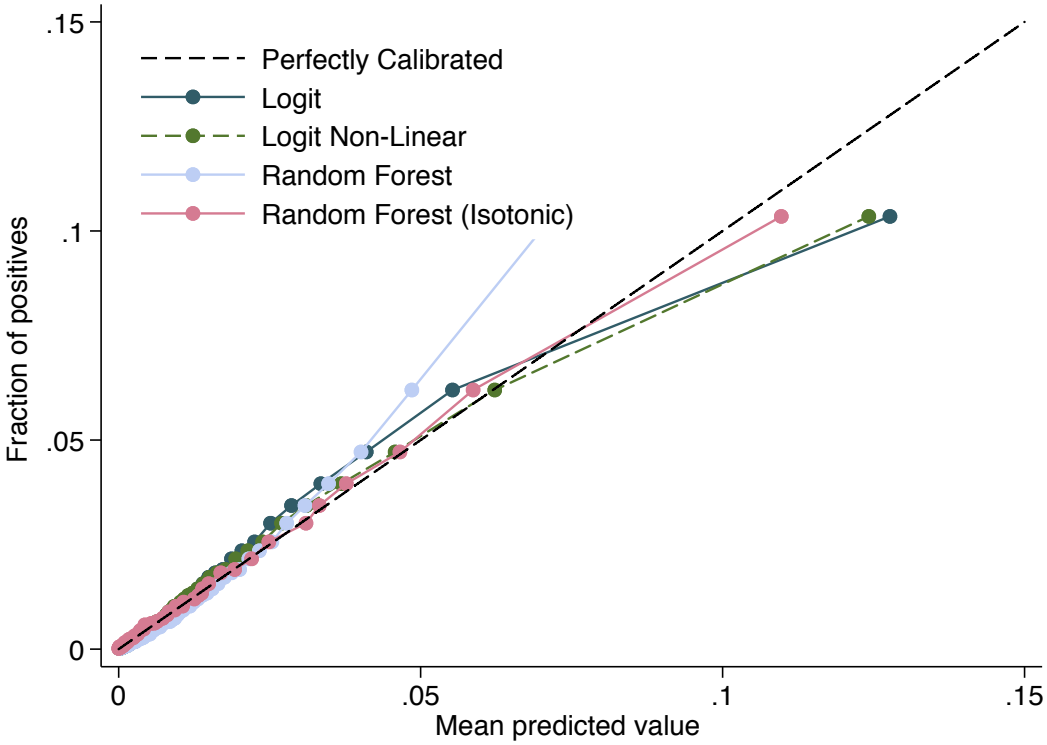
Isotonic regression involves searching across the space of monotonic functions to find the best fit function connecting the noisy estimates with the true values. More concretely, for an individual i, let y_i be the true outcome, and let $\hat{y}_i$ be predicted value from the Random Forest model. Then, 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:

$$\hat{z} = \arg\min_{z} \sum_{i} (y_i - z(\hat{y}_i))^2. \tag{5}$$

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-2. This figure bins the predicted values (either $\hat{y}_i$ or $\hat{z}(\hat{y}_i)$) into 20 bins with equal number of observations in each bin, and takes the average predicted value and the average true default rate for each 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 initially above the predicted values, and then below, 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. However, the calibrated Random Forest - Isotonic model line is almost exactly on the 45° line, suggesting near-perfect calibration.

Electronic copy available at: https://ssrn.com/abstract=3072038

Figure A-2: **Calibration Curve**



Electronic copy available at: https://ssrn.com/abstract=3072038

A.4 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.

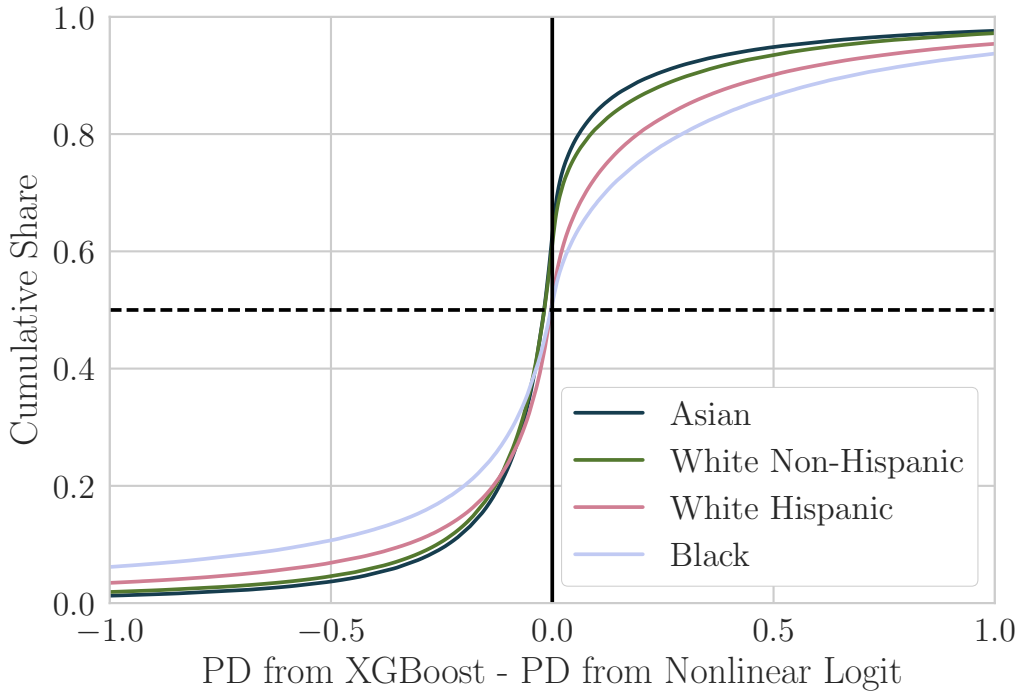
The gradient boosting approach takes a single tree model, similar to those underlying the Random Forest. However, rather than increase the number of trees, the model iterates over the tree by constructing new leafs (branching) and removing leafs (pruning) to continuously improve the tree's predictive 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.[3]

In Figure A-3, we plot a version of Figure 5, 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.

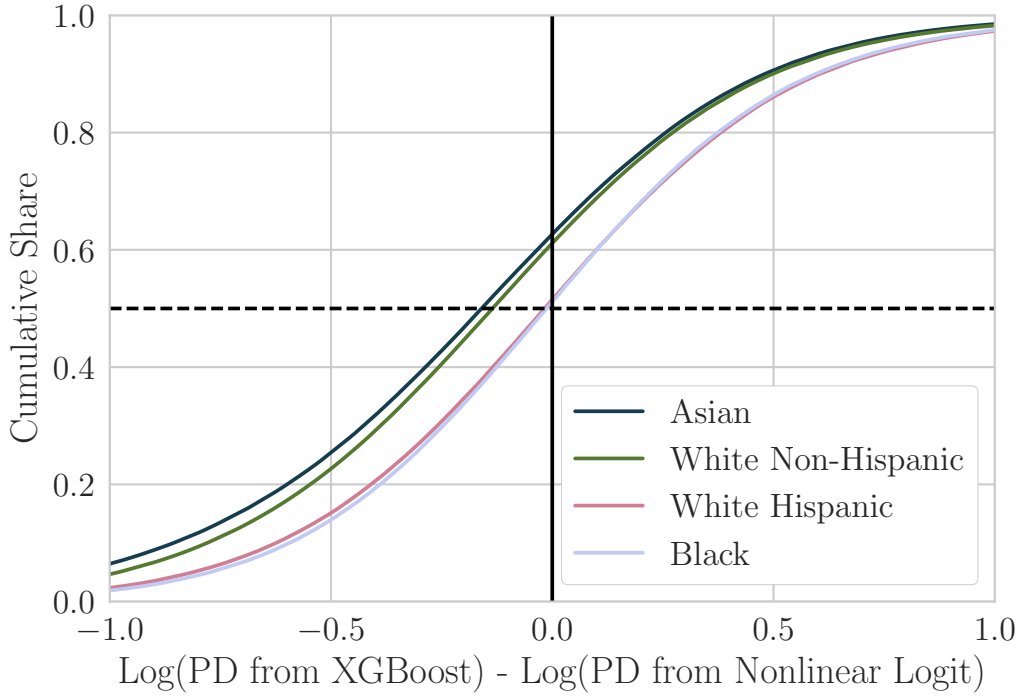
[3]We implement this method in R using the `xgboost` library.

Electronic copy available at: https://ssrn.com/abstract=3072038

Figure A-3: **Comparison of Predicted Default Probabilities — XGBoost vs. Nonlinear Logit**



Panel A



Panel B

Electronic copy available at: https://ssrn.com/abstract=3072038

A.5 Formal Model of Mortgage Market Equilibrium

This section contains a more rigorous description of our equilibrium model in Section 5, including the assumptions on the behavior of actors in the model, assumptions required for the identification of default probabilities, an extension to the realistic case with endogenous contract terms, details on the empirical calibration and the computation of equilibrium, and empirical details on the sample used in this equilibrium exercise.

A.5.1 Model Set-up

In order to discuss the issue of selection by borrowers (into those who accept, and those who reject a loan offer), we now formally model borrower behavior. While this additional formalism is instructive, it does not change the nature of competitive equilibrium, which continues to be determined by the break-even condition for lenders that we develop in the main text.

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.

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,[4] 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

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

[4] 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.

Electronic copy available at: https://ssrn.com/abstract=3072038

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

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

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 loss given default per dollar of the loan is therefore

$$LGD(x,R) = \phi + \max\left\{1 + R - \frac{\gamma}{LTV}, 0\right\}$$

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

Lenders operate under incomplete information. They know 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 $\hat{P}(x,R)$. [5]

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, the expected Net Present Value per dollar of the loan is now

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

where $N(x,R$ is as defined in the text.

Equilibrium Prices. With lenders in Bertrand competition, the equilibrium interest rate offered to borrowers with characteristics x is the lowest interest rate that satisfies $N(x,R) \geq 0$, allowing lenders to break even in expectation given their information. If no such rate exists, then borrowers with characteristics x are not accepted for a loan.

[5] 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.

Electronic copy available at: https://ssrn.com/abstract=3072038

Notice that, although the probability $\alpha(x, R)$ of acceptance matters for profits, it does not matter for competitive equilibrium rates, which are determined by zero profits $N(x, R)$ conditional on acceptance.

Generically, this is a 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.

A.5.2 Empirical Identification of Default Probabilities

Our empirical implementation relies on the identification of 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 $\hat{P}(x_i, R_i)$ from different statistical technologies to compute equilibrium, as summarized above.

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 $R \geq 0$.[6] 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 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

[6] 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.

Electronic copy available at: https://ssrn.com/abstract=3072038

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.

Implicit in the above is a second assumption, which relates to selection by borrowers. Indeed, in our derivation, have continued to use the condition that surplus $S(x,\theta,R) \geq 0$ as the condition under which borrowers are observed to take a mortgage with interest rate R. One caveat to our analysis is that this—unlike in the model—borrowers in the data may have additional considerations when choosing a mortgage. For instance, borrowers may have access to competing contracts, and may compare their surplus across those contracts before making a choice. For identification of the relevant probabilities in the model, we therefore have to assume that the average borrower taking interest rate R in the data has similar default behavior to the average borrower for whom the surplus from such a mortgage is positive.

A.5.3 Extensions

Equilibrium with Endogenous Contracting Terms. In our model, lenders' Net Present Value depends on contracting terms beyond the interest rate. In particular, equation (2) 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

Electronic copy available at: https://ssrn.com/abstract=3072038

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

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

and the average mortgage rate conditional on obtaining credit is

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

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.

Dynamic Model. Here, we briefly discuss how our equilibrium model might change if we explicitly modeled the transition from one statistical technology to the other, and studied how the market would evolve after the initial transition to the new technology. This is a potentially important issue since in the new equilibrium, the set of borrowers that receive credit (and the rate at which they do so) is altered, which in turn changes the set of data points that lenders subsequently use when estimating their default prediction models.

If the baseline data, which lenders use in the first round (when transition occurs), is sufficiently large, and does not suffer from issues of selection on unobservables, lenders' inferences converge to the "oracle" (i.e., the best possible predictor in the permitted class of functions) for structural probabilities. In this case, the inferences that are drawn from the first round of effects will be a reasonable description of the steady-state effects. However, if the baseline data used in the first round is either relatively small or selected on unobservables, then convergence issues may be important.

In such cases, the effects on inequality are potentially nuanced. One might get "once out, always out," i.e., if borrowers with characteristics x did not look good enough in the baseline data, then they are excluded from the additional data (this is analogous to the dystopian case referred to in models in which algorithms inherit the biases they learn from the training

Electronic copy available at: https://ssrn.com/abstract=3072038

data set; for a book-length treatment, see O'Neil 2016). This means that nothing new will be learned about them—and once again, inferences drawn from the first cohort would be a fair characterization of equilibrium.

However, it could also be the case that a good machine learning model estimated on repeated cross-sections of selected borrowers generalizes well beyond the exact constellations that are in the data, and brings some excluded people back into the market, meaning that inferences drawn from a single cohort persist for some period of time, but are then less important as time goes by and greater equality between groups is restored. In this case, we note that minority groups, along the path to convergence, will face higher integrated "losses" along both exclusion and rate dimensions until convergence is finally achieved. This means that the political economy and distributional questions that our analysis raises are still valid, since it is difficult to argue convincingly to populations facing losses over several periods that they need only wait for several cohorts for an uncertain, but hopefully brighter future.

To obtain a sense of likely magnitudes, and to more deeply interrogate this question, one would need to calibrate realistic equilibrium outcomes given a baseline dataset, simulate a new cohort of data from those outcomes (which is a non-trivial task in itself without a full model of borrower behavior), re-calibrate equilibrium, and so forth, ideally until convergence. We leave this important but challenging task for future research.

A.5.4 Calibration and Computational Procedure

We now describe how we calculate equilibrium outcomes in our model. To start, 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 capital plus a fixed spread of 30bp.[7] 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.[8]

Then we use the following procedure to solve for lenders' equilibrium decisions, using both traditional and machine learning methods:

1. Estimate cumulative probability of default up to a time period post-loan issuance of

[7] 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).

[8] 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 smallt downward adjustment to these fractions since we need the loss relative to the original balance.

Electronic copy available at: https://ssrn.com/abstract=3072038

36 months, as a function of borrower observables and spreads at origination (SATO), which maps one-for-one into interest rates R.[9]

2. Transform the estimated 36-month default probability into an imputed cumulative default probability $\hat{P}(x, R)$ over the lifetime of the mortgage. We describe this transformation in detail below.

3. Adjust estimated default probabilities $\hat{P}(x, R)$ for the potential bias driven by the endogeneity of interest rates. We describe this adjustment in detail below.

4. For every borrower i in the sample, solve for the equilibrium interest rate (or rejection, where appropriate), as follows:

   - Use adjusted default probabilities from the previous step to evaluate $N(x_i, R)$ at a grid of 20 values for SATO between -0.4 percent and 1.5 percent
   - 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.

Details of Adjustment of Lifetime Default Probabilities. We use the Standard Default Assumption (SDA) curve[10] in combination with our estimated three year cumulative probabilities of default to estimate the lifetime cumulative probability of default.

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

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

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

We assume that the hazard rates of the mortgages in our sample can be expressed as multiples M of $h(t)$, i.e., as a scaled version of the same basic SDA shape. Using this

[9] This follows because SATO in our data is the difference between the interest rate R and the average interest rate in the respective borrower's cohort

[10] 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.

Electronic copy available at: https://ssrn.com/abstract=3072038

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:

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

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:

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

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:

$$\begin{aligned}P(T) &= 1 - \exp[MH(T)] \\ &= 1 - \exp\left[\frac{1}{h_{max}}\frac{\log(1-\hat{p})}{\hat{t} - \frac{t_1}{2}}H(T)\right]\end{aligned} \tag{6}$$

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

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

These equations allow us to transform the empirical estimate $\hat{p}(x, R)$ into a lifetime cumulative probability of default $\hat{P}(x, R)$ as required.

Electronic copy available at: https://ssrn.com/abstract=3072038

Details of Bias Correction. 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 fixed-rate mortgages 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:

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

Electronic copy available at: https://ssrn.com/abstract=3072038

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.5.5 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.

Table A-3 below shows results from a direct way to check for the prevalence of soft information in this sample. 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-4 shows histograms and kernel density estimates for the SATO residual and the difference between actual and equilibrium rates.

Electronic copy available at: https://ssrn.com/abstract=3072038

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.

The figure shows that the counterfactual equilibrium rates that we predict differ from actual rates, but for the most part, these changes to the predictions lie within the region covered by residual variation, or the "noise" in observed interest rates. It is true that a small fraction of our predictions is driven by extrapolation outside the noise in rates that we observe (the area under the actual rates minus equilibrium rates curve that does not overlap measures this fraction), but the patterns in the plot are broadly reassuring about the fairly limited extent of this extrapolation.[11]

[11] 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.

Electronic copy available at: https://ssrn.com/abstract=3072038

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.

A.5.6 Additional Results

This section contains two further sets of results from our illustration of the equilibrium effects of machine learning.

First, to further explore effects of new technology at the intensive margin, Figure A-5 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 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.

Second, to better understand the changes introduced by machine learning technology at the extensive margin, Table A-4 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).

Electronic copy available at: https://ssrn.com/abstract=3072038

Figure A-4: **Residual interest rate variation**.

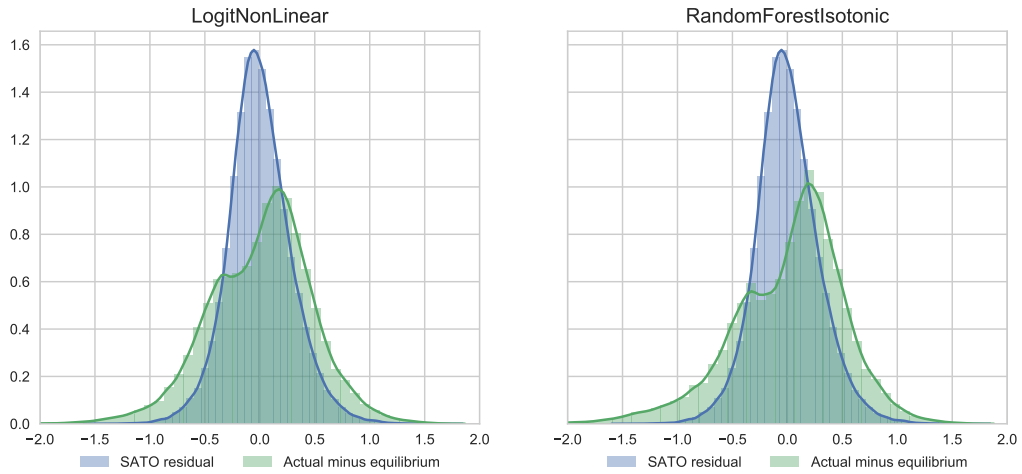
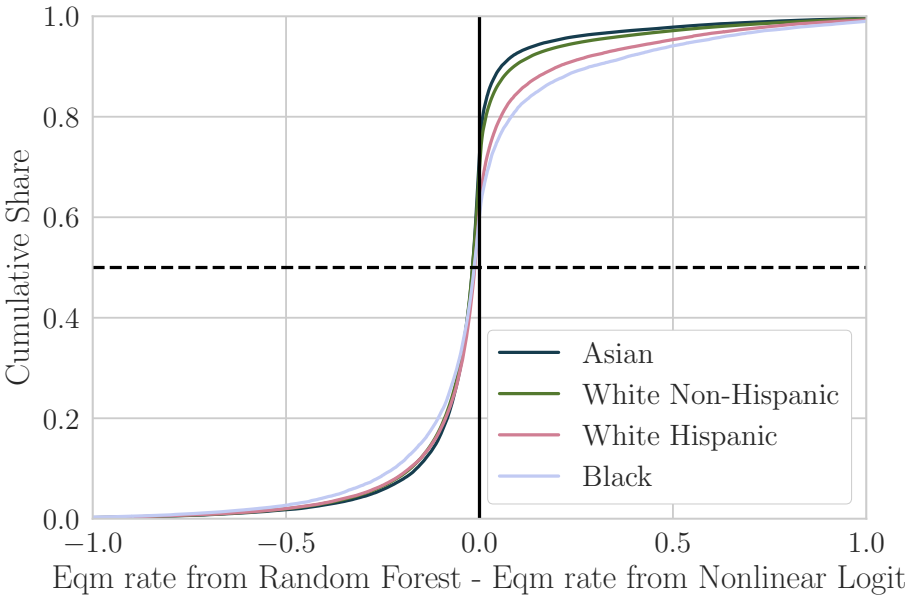


Table A-4: **Decomposition of Equilibrium Effects**

	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

Electronic copy available at: https://ssrn.com/abstract=3072038

Figure A-5: **Comparison of Equilibrium Interest Rates**



The proportions of Inclusions and Exclusions reported in the table reveal that the increase in average acceptance rates (in Table 7) 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.

Electronic copy available at: https://ssrn.com/abstract=3072038

A.6 Alternative Approach to Estimate Effects of Innovation in Technology

As described in detail in Section 5 and the previous section of this appendix, our equilibrium credit market model requires us to make a number of strong assumptions. In this section, we present an alternative, reduced-form approach to gauge the possible effects of a change in estimated default probabilities (as a result of innovation in statistical technology) on credit market outcomes.

Associated with this alternative approach is a different set of assumptions, namely:

1. Lenders in the observed empirical data employ a default probability model that is well-approximated by the Nonlinear Logit model that we estimate in the data (without rates in the estimation of default probabilities).

2. The interest rates that they set on loans (and that we observe in the data) are drawn from the computation of default probabilities in point 1. above, processed through an unknown mapping function.

3. This mapping function can be well-approximated, say by estimating a flexible function linking rates and estimated default probabilities in the data.

4. The mapping function is invariant to the underlying technology used to estimate default probabilities, so can be used to estimate the likely distributional changes on the intensive margin as technology varies.

5. These distributional changes can be estimated by simply estimating the default probabilities from a machine learning model (again, not using rates) and plugging them in to the mapping function estimated in step 3. above.

Relative to our equilibrium model, this alternative approach has a number of limitations. First, it makes it impossible to say anything about the extensive accept/reject margin. Second, it ignores the strong empirical effect of R on default probabilities. Third, it also ignores that for the credit risk premium, it is not just the default probability that matters, but also loss-given-default (LGD), which our equilibrium model captures (albeit in a simple way). For those reasons, we view the equilibrium model as a superior approach, albeit a simple one, to assess possible magnitudes. Nevertheless, we find that what follows provides a useful reduced-form sanity check for the results from the equilibrium model.

We implement the alternative approach as follows: we flexibly estimate the relationship between a borrower's spread at origination (SATO) (i.e., cross-sectionally demeaned rates)

Electronic copy available at: https://ssrn.com/abstract=3072038

from the data and predicted default probabilities from the Nonlinear Logit (excluding R from the regressors used to estimate the default probability). We tried a few different approaches to this mapping, including kernel-weighted local polynomial smoothing, with virtually identical results; in what follows, we report results based on fractional polynomial regressions.[12] We then use this estimated mapping function to generate predicted SATOs from the Random Forest estimated default probabilities (also estimated without using R). Finally, we compare the resulting distributions of rates (i.e., fitted rates from the Nonlinear Logit vs. predicted rates from the Random Forest processed through the estimated mapping function) across groups.

The results from this procedure are summarized in Table A-5. The broad result is that the between-group average rate differentials when moving from the standard (Nonlinear Logit) to the machine learning (Random Forest) approach mirror the patterns found using our equilibrium approach: White Non-Hispanic and Asian borrowers gain more (in terms of interest rates) than White Hispanic and Black borrowers. The increase in the spread between Asian and Black borrowers, which was 6bp in the equilibrium model in the main text, is 5bp in the analysis reported here. Moreover, the standard deviation of rates within all groups rise when moving from the Nonlinear Logit to Random Forest, though the differentials between standard deviations across groups are not as pronounced as under the equilibrium mapping. In sum, under this alternative approach to computing the mapping between default probabilities and rates, we find broadly similar patterns both qualitatively and quantitatively.

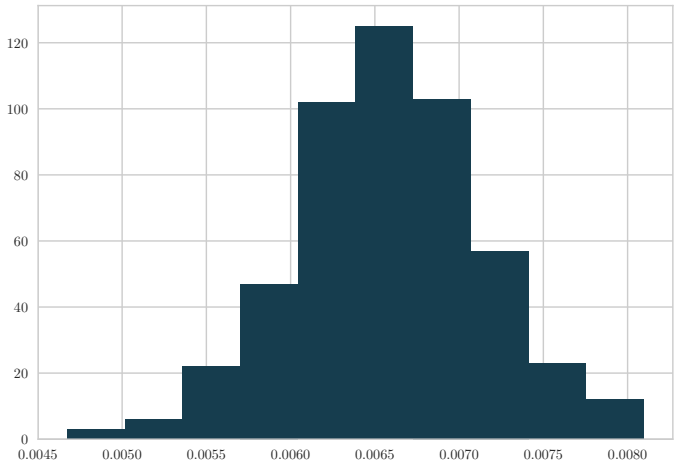
Table A-5: **Equilibrium Effects Under Alternative Approach**

	Mean SATO (%)		SD SATO (%)	
	NL	RF	NL	RF
Asian	-0.064	-0.151	0.188	0.258
White Non-Hispanic	-0.041	-0.118	0.190	0.262
White Hispanic	0.007	-0.036	0.183	0.241
Black	0.055	0.023	0.181	0.232
Other	-0.034	-0.108	0.193	0.264
Population	-0.038	-0.113	0.190	0.262

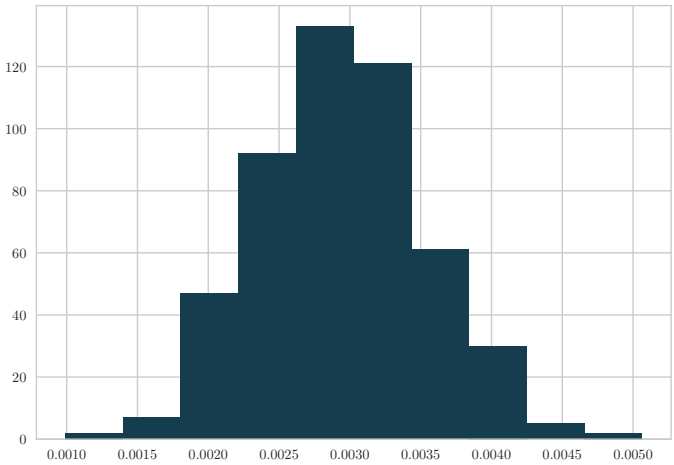
[12] Loosely speaking, fractional polynomial regressions search over different combinations of polynomials of degree z of a continuous explanatory variable x, in our case default probability, where the exponents considered are (-2, -1, -0.5, 0.5, 1, 2, 3) and $\log(x)$, to find the closest in-sample fit for $y = f(x)$. See "fp" in Stata for additional details. In our case, we report results for $z = 2$, though results for $z = 3$ are nearly identical.

Electronic copy available at: https://ssrn.com/abstract=3072038

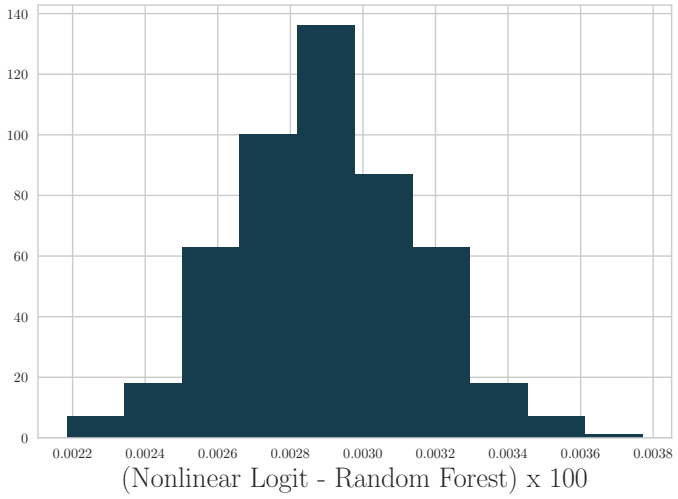
Figure A-6: **Bootstrap Estimates of Differences in Statistics**



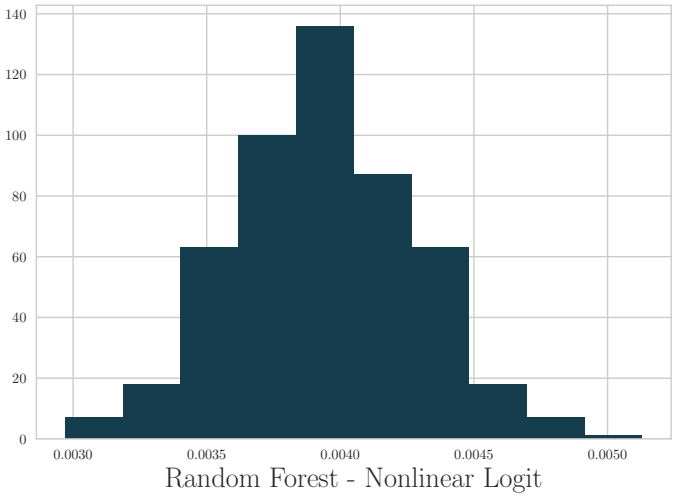
Panel A: Difference in ROC AUC



Panel B: Difference in Average Precision



Panel C: Difference in Brier Score



Panel D: Difference in R^2

Electronic copy available at: https://ssrn.com/abstract=3072038