Who Benefits from Big Data in Lending Markets?*

Menaka Hampole Adam Jørring

March 27, 2025

Abstract

We study the distributional effect of big data in a model of competitive lending markets, where lenders use statistical learning techniques to assess borrower risk and determine loan terms. As lenders gain access to more data variables, they offer more granular pricing, reducing borrower pooling. Crucially, average consumer welfare can improve even as average interest rates rise if lenders reduce credit rationing. To test our model's predictions, we analyze the introduction of FICO scores in U.S. mortgage markets. Leveraging early adopters of the GSEs' automated underwriting systems for identification, we find that FICO scores increased pricing granularity and reduced credit rationing. Distributionally, while low-income and minority borrowers faced higher interest rates, they also gained greater access to credit, expanding financial inclusion. Our findings challenge the notion that interest rate differentials alone are a sufficient statistic for changes in consumer welfare.

Keywords: Big Data, Statistical Learning, Credit Rationing

JEL Classification Codes: G21, D82, O33

*We would like to thank Simcha Barkai, Maryam Farboodi, Ben Matthies and seminar participants at Boston College, Notre Dame, the Financial Research Association conference, and the Junior Household Finance Brownbag for their comments and suggestions. We also thank Trevor Haynes and Song Zhang for outstanding research assistance.
Hampole is at Yale School of Management. Email: menaka.hampole@yale.edu.
Jørring is at University of Massachusetts Amherst. Email: ajorring@umass.edu.

The rapid growth of big data and use of machine learning in credit markets has sparked an intense policy debate about its implications for consumer welfare, financial inclusion, and market fairness. Advanced statistical learning techniques allow lenders to better assess borrower risk, potentially leading to more efficient credit allocation and reduced default rates. However, increased data availability may also lead to greater price dispersion and potentially exacerbate existing inequalities, particularly among historically underserved borrower groups (Waddell, 2016; Vigdor, 2019; Bartlett et al., 2022). Despite these growing concerns, there remains limited systematic analysis quantifying how enhanced data availability affects market outcomes and borrower welfare.

This paper addresses this gap by theoretically and empirically examining how increased data availability affects mortgage lending decisions, pricing, and credit availability. We develop a model that integrates statistical learning techniques into a lending framework and use it to derive novel empirical predictions. Empirically, we test these predictions by leveraging the historical introduction of FICO credit scores into U.S. mortgage underwriting through automated underwriting systems (AUS) by Fannie Mae and Freddie Mac. Exploiting variation across lenders based on the timing of AUS adoption, we causally estimate the impact of increased data availability on lending outcomes.

We document three key results. First, increased data availability leads lenders to more precisely estimate borrower risk, resulting in significant interest rate differentiation. While average interest rates rise due to previously excluded, higher-risk borrowers gaining access to credit, safer borrowers experience lower rates. Second, enhanced borrower segmentation substantially reduces credit rationing, disproportionately benefiting historically underserved populations. Finally, despite higher average rates, overall consumer welfare improves significantly, driven by increased access to credit among borrowers previously excluded from the market.

We begin our analysis by developing a theoretical model of a competitive lending market. The model features heterogeneous borrowers differentiated by credit risk, lenders

that use statistical learning to assess borrower risk, and equilibrium outcomes determined through Nash bargaining (Collard-Wexler et al., 2019). In equilibrium, lenders partition borrowers into discrete risk groups and set uniform interest rates within each group, introducing endogenous credit rationing for some borrowers due to adverse selection—a common result in the literature (Jaffee and Modigliani, 1969; Stiglitz and Weiss, 1981).

Next, we incorporate enhanced data availability into the model, examining how improved statistical information refines risk assessment. Increased granularity allows lenders to subdivide borrower risk partitions, leading to finer price discrimination and reduced credit rationing. The model generates specific predictions: increased data availability raises average interest rates as credit expands to previously rationed higher-risk borrowers, reduces loan rejection rates, and differentially impacts borrower groups based on credit risk.

Our model focuses directly on the effect of increasing data availability rather than other types of changes in screening technology studied in the literature.[1] A noteworthy distinction of our framework is that it simultaneously models both the intensive and extensive margins of lending decisions. Unlike Jiang et al. (2024) and Gao et al. (2024), whose models primarily emphasize how improved screening technology impacts either approval rates or interest rate dispersion separately, our approach explicitly incorporates both margins within a unified theoretical framework. Specifically, Jiang et al. (2024) model lenders who use a cutoff strategy for loan approvals, focusing primarily on how better screening leads to higher approval rates for marginal borrowers and increased dispersion in interest rates. Similarly, Gao et al. (2024) analyze the transition from human to algorithmic underwriting in mortgage markets, highlighting improvements in credit access for marginal borrowers without significantly changing overall default risks. By

[1] Hauswald and Marquez (2003) examine how technological advancements influence competition among financial institutions, showing how improved screening technology may lower interest rates through increased competition or raise them through enhanced market power. Vanasco (2017) investigates how enhanced asset screening capability can paradoxically reduce market liquidity due to heightened informational asymmetries.

modeling these margins jointly, our framework provides a more integrated understanding of how finer-grained borrower information influences market outcomes and welfare effects across borrower segments.

Empirically, we test the model's predictions using detailed loan-level data from Home Mortgage Disclosure Act (HMDA) merged with loan-level origination and performance data from a mortgage servicing aggregator. Exploiting a difference-in-differences strategy based on lender adoption of AUS, we find robust evidence supporting our theoretical predictions. Lenders adopting AUS exhibit increased pricing granularity, with average interest rates rising significantly post-adoption. However, previously rationed borrowers gain access to credit markets, markedly decreasing overall loan rejection rates. Furthermore, our results highlight distributional implications: interest rates increase, particularly affecting higher-risk borrowers, yet access to credit also expands disproportionately among minority and low-income groups.

These findings contribute to multiple strands of literature. First, we build upon seminal theoretical work examining credit rationing and the role of asymmetric information in credit markets (Jaffee and Modigliani, 1969; Stiglitz and Weiss, 1981; Leland and Pyle, 1977; Diamond, 1984). Our model extends this tradition by explicitly modeling how increased data availability, enabled through statistical learning, affects market equilibria and credit allocation. Relatedly, our theoretical analysis complements recent theoretical work on credit rationing that considers the interaction between borrower information and lender decisions under varying institutional constraints (Thakor, 1996; Jiménez et al., 2006).

Second, our paper contributes to the emerging empirical literature on the effects of fintech, big data, machine learning, and open banking (Bartlett et al., 2022; Fuster et al., 2022; He et al., 2023). While previous studies have primarily emphasized the distributional implications of advanced statistical technologies in terms of prices, especially concerning race, ethnicity, or gender disparities (Waddell, 2016; Buolamwini and Gebru,

2018; Vigdor, 2019; NYDFS, 2021), our study uniquely highlights how granular pricing due to increased data availability affects overall credit supply, credit rationing, and consumer welfare beyond interest rate changes. Our findings provide empirical evidence supporting the hypothesis that the welfare implications of enhanced statistical precision are multifaceted, and consumer surplus may improve even when average interest rates rise.

Third, we inform the literature studying financial technology's historical impacts on mortgage markets, notably the transition from manual underwriting to automated underwriting systems (AUS) in the 1990s (Maselli, 1994; Nixon, 1995; Harney, 1996; Straka, 2000; Johnson and Tzur-Ilan, 2025). Unlike existing literature, which largely focuses on how automation altered loan approval and processing efficiency (Fuster et al., 2019; Foote and Willen, 2018), our analysis demonstrates how the introduction of borrower credit scores fundamentally altered lenders' risk segmentation capabilities and thus impacted market outcomes and financial inclusion.[2]

The remainder of the paper proceeds as follows. Section 1 introduces our baseline model and defines the full-information equilibrium. Section 2 integrates statistical learning into the model, analyzing how increased data availability affects equilibrium outcomes and consumer welfare. Section 3 derives empirical predictions from the theoretical framework. Section 4 details our empirical strategy, including data sources and identification methodology. Section 5 presents empirical results validating theoretical predictions and quantifying distributional impacts. Section 6 concludes and discusses policy implications.

[2]Moreover, our results contribute directly to contemporary policy debates concerning algorithmic fairness and regulatory frameworks governing data usage in financial services (Gupta, 2019; Warren and Jones, 2019). By quantifying the trade-offs between granular pricing, financial inclusion, and consumer welfare, we provide policymakers with empirical insights necessary to balance efficiency gains from advanced statistical methods against potential increases in disparities across borrower groups. Importantly, our empirical results underscore that interest rate differentials alone are insufficient for assessing consumer welfare implications, advocating for broader metrics that also incorporate credit access and rationing.

1 Model

This section introduces our baseline model of a credit market with lenders and borrowers bargaining over the interest rate. We begin by introducing the borrowers and lenders, then we describe the bargaining process, before outlining the equilibrium in a full information baseline. We summarize the timing of our model with the diagram provided in Figure 1 below. For convenience, Table B1 in the Appendix lists the notation used in this paper.

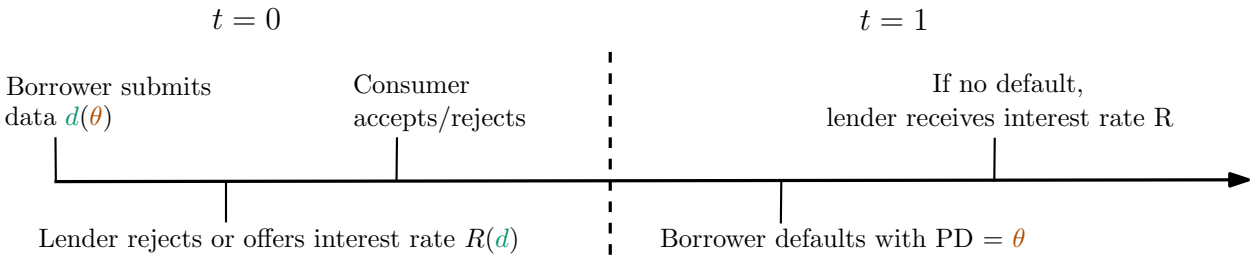
1.1 Borrowers

The model consists of a continuum of risk-neutral borrowers of measure one, each looking for a loan of \$1. Borrowers vary by their type θ, which is private information, and each borrower knows her own type. Borrowers are distributed uniformly across types, $\theta \sim U\left(\underline{\theta}, \bar{\theta}\right)$, where $\underline{\theta}$ is larger than or equal to 0 and $\bar{\theta}$ is less than or equal to 1. The type θ governs both their probability of default and their loss given default. Their probability of default (PD) is equal to their type, and their loss given default (LGD) is a linear transformation of their type: $\text{PD}(\theta) = \theta$ and $\text{LGD}(\theta) = \alpha + \eta\theta$, where α and η are positive constants.[3]

At the beginning of the first period, each borrower submits a mortgage application, applying to borrow \$1, and their application contains data $d = d(\theta)$, which is a function of the borrower's type. If the lender accepts the borrower's loan application, the lender will offer a contract with an interest rate, $R(d)$, that depends on the data in the application. If the borrower accepts the interest rate, she receives the \$1 loan, and she buys a house. In the second period, the borrower is hit by a shock. With probability $1 - \theta$ she does not

[3] As in He et al. (2023), who also study credit market competition, θ is the probability of default. In their model, however, a share θ of the borrowers are high types who always repay, while the remaining 1-θ are low types who never repay. In contrast, in our model, even though the borrowers know their type, there is still uncertainty about whether or not they will default. This difference leads the borrowers in our model to have a continuous demand function over interest rates.

FIGURE 1: Model timing



Note: This figure provides the model timing.

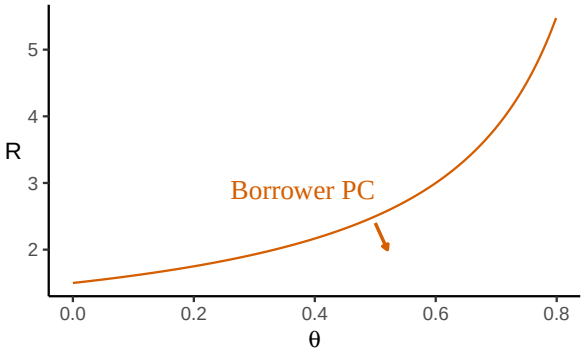
default, and she pays back the lender R. If the borrower accepts the loan contract, she will pay the interest rate of R at $t = 1$, as long as she doesn't default.

The benefit for the borrower from accepting the loan is that she receives utility of 1 (the value of the loan) at $t = 0$ and, if she does not default, a continuation value from owning the house, b, at $t = 1$. Since the borrower knows her own type, she knows that she will default with probability θ, and hence she accepts the loan offer if and only if

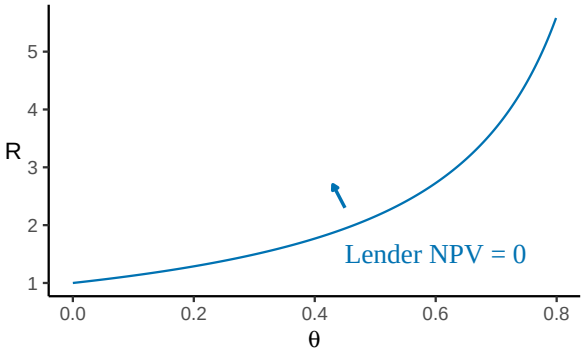
$$\underbrace{(1-\theta)R}_{\text{Expected interest payment}} \leq \underbrace{1}_{\text{Upfront benefit}} + \underbrace{(1-\theta)b.}_{\text{Expected continuation value}} \tag{PC}$$

For convenience, we refer to this condition as the borrower's "participation condition" (PC). To understand this condition, consider how borrowers across the type distribution respond to contracts with varying interest rates. In Panel A in Figure 2, we plot the borrower (PC) in the (θ, R) space (where we set $b = 50\%$). The borrower (PC) indicates the highest interest rate that a borrower of type θ is willing to accept. Intuitively, higher risk borrowers are willing to accept higher interest rates, as they know they will be more likely to default.

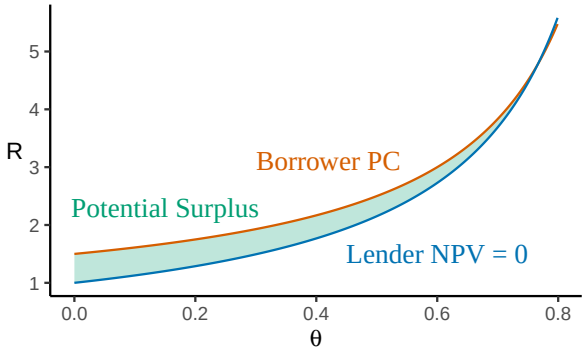
FIGURE 2: BORROWER AND LENDER CONSTRAINTS



Panel A: Borrower Participation Constraint



Panel B: Lender NPV Constraint



Panel C: Total Potential Surplus

Note: This figure plots the borrower participation constraint (PC) in Panel A, the lenders zero Net Present Value (ZNPV) constraint in Panel B. In Panel C, we plot both constraints. The shaded green area between the two constraints represents the total potential surplus in the market.

1.2 Lenders

A continuum of risk-neutral lenders compete. Each lender randomly receives a loan application from a borrower with data d. Based on the available data in the application and the lender's statistical model, the lender decides either to reject the loan application or to offer an interest rate R. The lender's objective function is described as follows:

$$\max_{R} \underbrace{E\Big[[1-PD(d)]R - PD(d)[1-LGD(d)]\Big] - 1}_{\text{Expected Net Present Value}} \quad \text{s.t.} \quad \underbrace{R\,[d(\theta)]}_{\text{Statistical technology}}.$$

In the section below, we will describe in detail how the mapping between data d and interest rate R depends on the statistical technology that the lender uses. The lender will only offer an interest rate R where it has at least a zero net present value. For convenience, we refer to this constraint as the (ZNPV) constraint. It describes the set of interest rates that the lender is willing to offer as a function of the borrower's type (see Appendix B for the proofs):

$$R \geq \frac{1+\theta(1-\alpha)+\eta\theta^2}{1-\theta}. \tag{ZNPV}$$

We plot the lender's zero NPV constraint in Figure 2 Panel B, where the arrow indicates that any combination of R and θ above the constraint will yield a positive expected NPV. In Panel C, we combine the (PC) and (ZNPV) constraints. The shaded green area between the two constraints represents the total potential surplus in the market, and the surplus is divided between the borrower and lender depending on their bargaining power.

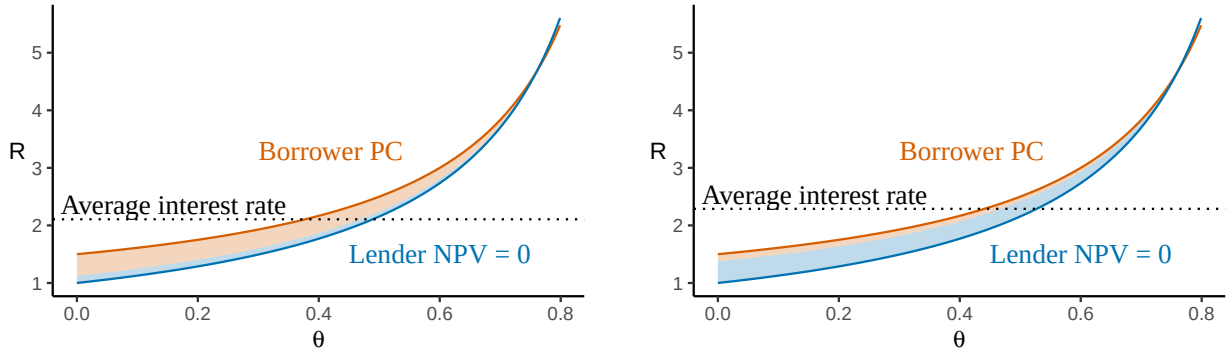
1.3 Bargaining

The total surplus available for bargaining, which we denote by S, is given by the difference between the maximum interest rate the borrower is willing to accept and the minimum rate the lender is willing to offer:

$$S = R^{(PC)} - R^{(ZNPV)} = \left(\frac{1+(1-\theta)b}{1-\theta}\right) - \left(\frac{1+\theta(1-\alpha)+\eta\theta^2}{1-\theta}\right).$$

To make the model tractable, we follow the literature and assume that borrowers and lenders are matched one-to-one and negotiate bilaterally, and that the lender and the borrower engage in Nash bargaining, where $\lambda \in [0,1]$ represents the borrower's bargaining power, while $(1-\lambda)$ represents the lender's bargaining power. These assumptions and the resulting Nash bargaining outcome are standard in the literature and can be micro-

FIGURE 3: EQUILIBRIUM INTEREST RATES UNDER FULL INFORMATION



Panel A: $\gamma = 25\%$ Panel B: $\gamma = 75\%$

Note: This figure plots the division of surplus in the full information baseline case in a concentrated market (Panel A) where $\gamma = 25\%$ and a competitive market (Panel B) where $\gamma = 75\%$. In both panels, the shaded orange region represents the borrower surplus and the shaded light blue region represents the lender surplus. The average interest rate across the entire market is denoted by the dotted line.

founded in several ways, including through frictional search, stochastic re-matching, informational frictions, commitment limitations, or through repeated interactions and reputation effects.[4] In this Nash bargaining equilibrium, the interest rate R^* is determined as $R^* = R^{(PC)} - \lambda S$.

[4]For example, one approach is to introduce a frictional search environment, where borrowers and lenders match randomly, and re-matching involves a costly delay, discouraging renegotiation, resulting in the Nash bargaining solution as shown by Cho and Matsui (2013). Another possibility is a dynamic market with stochastic re-matching, as in Bayati et al. (2011), who develop a bargaining model in exchange networks where agents negotiate iteratively based on their estimated best alternatives, leading to outcomes that generalize Nash bargaining to networked settings. Alternatively, lenders may face informational frictions that slow the borrower screening process, limiting their ability to serve many borrowers simultaneously. Stevens (2007) shows how the efficiency of matching depends on search frictions and the time spent evaluating potential matches. In another variation, commitment limitations could allow borrowers to make take-it-or-leave-it offers while lenders decide whether to accept, leading to an implicit bargaining process. Collard-Wexler et al. (2019) formalize this idea in the "Nash-in-Nash" framework, which extends the Rubinstein bargaining model to multi-agent settings, demonstrating conditions under which pairwise Nash bargaining emerges as a stable outcome. Lastly, repeated interactions and reputation effects can discipline lenders into offering fair interest rates to sustain long-run profits, as analyzed by Damdinsuren and Zaharieva (2023), who model labor markets with on-the-job search, Nash bargaining, and learning dynamics, showing how reputation influences wage-setting and job transitions.

1.4 Full Information Baseline

To illustrate the effect of bargaining on the equilibrium interest rate, we begin by considering a full information baseline version of the model. In this baseline version, the lender perfectly observes the borrower's type. When the lender perfectly observes the borrower's type, it can offer each borrower an individual interest rate, and the equilibrium interest rate R^* is a direct function of the type θ:

$$\begin{aligned} R^* &= R^{(PC)} - \lambda S \\ &= \frac{1 + (1-\theta)b - \lambda[(1-\theta)b - \theta(1-\alpha) - \eta\theta^2]}{1-\theta}. \end{aligned}$$

Figure 3 shows a graphical representation of the effect of competition. In Panel A, we consider a relatively competitive market, where $\gamma = 25\%$. In this market, the borrowers capture the majority of the surplus, and for each borrower type, the interest rate is close to $R^{(ZNPV)}$ (the lowest possible rate). In comparison, we consider a concentrated market in Panel B, where $\lambda = 75\%$. Here, the lender captures most of the surplus, and the interest rate is close to $R^{(PC)}$ (the highest possible rate). In both panels, the shaded orange region represents the borrower surplus, the shaded light blue region represents the lender surplus, and R^* is where the two shaded regions meet. We plot the average interest rate across the entire market and denote this by the dotted black line. As we would expect, the average interest rate is higher in the concentrated market.

2 Learning From Big Data

The main innovation of our model is to introduce a model of statistical learning into the lenders' decision problem.

In contrast with the full information baseline discussed above, we now assume that the lender only knows the information supplied in the loan application. Specifically, we

assume that when a borrower submits a loan application, the application contains a set of N data variables $D = \{d_1, d_2, \ldots, d_N\}$, and the lender is then tasked with deciding whether to accept or reject the application based on D, and if it accepts, which interest rate to offer.

2.1 Training Dataset and Credit Risk Assessment

Lenders possess a historical dataset of past borrowers, containing observations on their types θ and the corresponding values of their observable data variables D. From this dataset, the lender estimates the statistical relationship between each data variable d_i and the borrower's default probability θ. Given this information, the lender constructs a statistical model to infer a borrower's type based on D. The objective is to classify borrowers into risk categories, which then determine the interest rate offered.

To achieve this classification, the lender employs a decision tree algorithm, a non-parametric statistical method that recursively partitions the data space into homogeneous subsets. Specifically, the lender utilizes a Classification and Regression Tree (CART) algorithm (Breiman et al., 1984) to segment borrowers into risk categories. The number of risk categories is determined dynamically using the Freedman-Diaconis rule (Freedman and Diaconis, 1981):

$$B = \max\left(2, \frac{\max(D) - \min(D)}{\frac{\kappa \cdot IQR(D)}{|D|^{1/3}}}\right),$$

where $IQR(D)$ is the interquartile range of the data, and κ is a tuning parameter that adjusts the granularity of partitioning. The lender then assigns risk labels P by discretizing true credit risk: $P = \text{digitize}(\theta, \text{bins} = \text{linspace}(0, 1, B + 1))$.

Given a dataset with features D and labels P, the algorithm selects a splitting rule s that maximizes the Information Gain (IG):

$$IG(S, s) = H(S) - \sum_j \frac{|S_j|}{|S|} H(S_j),$$

where $H(S)$ is the entropy of the dataset before the split and $H(S_j)$ is the entropy of each resulting subset after applying split s.[5] The entropy function is given by $H(S) = -\sum_k q_k \log_2 q_k$, where q_k is the probability of a data point belonging to category k.

The tree is trained on a feature matrix:

$$X = \begin{bmatrix} d_{1,1} & d_{1,2} & \dots & d_{1,N} \\ d_{2,1} & d_{2,2} & \dots & d_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{M,1} & d_{M,2} & \dots & d_{M,N} \end{bmatrix},$$

and labels $P = [P_1, P_2, \dots, P_M]$.

The lender does not observe the borrower's type θ directly but instead uses the historical relationship between observed features and credit risk to classify borrowers into risk categories. By processing the feature matrix through a decision tree, the lender determines which risk category each new borrower belongs to.[6]

2.2 Loan Pricing

After categorizing each borrower into a risk category, the lender first decides whether to accept or reject the loan application. This decision depends on whether the expected return from lending to a borrower in a given risk category meets the lender's required threshold for profitability.

If the lender rejects the application, the borrower does not receive a loan. If the lender accepts the application, it proceeds to assign an interest rate R to the borrower based on the assigned risk category. This modifies the equilibrium described in Section 1. Specif-

[5] See Quinlan (1986) for an introduction to the use of decision trees in machine learning.

[6] The feature matrix X organizes the data available to the lender in a structured format, where each row represents a different borrower (i.e., a separate loan application), and each column represents a different feature (i.e., a specific data variable reported in the loan application). For example, if the loan application includes information on income and credit score, then one row of X corresponds to a single borrower's values for these features, such as Income $= 50,000$ and Credit Score $= 700$.

ically, rather than offering a continuous schedule of interest rates as a function of θ, the lender now groups borrowers into discrete risk categories and assigns a uniform interest rate within each category. For each category P, the lowest possible interest rate that a lender is willing to offer is the interest under which the lender's expected Net Present Value (NPV) is exactly zero:

$$R^{ZNPV}(P) = \frac{1 + (1-\alpha)\mathbb{E}[\theta|P] - \eta\mathbb{E}[\theta^2|P]}{1 - \mathbb{E}[\theta|P]}.$$

If the lender further assumes that borrower risk types within each risk category P are uniformly distributed between two specific borrower types θ_1 and θ_2, the above general expression simplifies explicitly to:

$$R_P^{ZNPV} = \frac{2\eta\frac{\theta_1^2+\theta_1\theta_2+\theta_2^2}{3} + (\alpha-1)(\theta_1+\theta_2) - 2}{\theta_1+\theta_2-2}.$$

This discretization introduces potential distortions relative to the full-information case and affects both borrower participation and lender profits. The next section explores how this statistical learning affects the distribution of surplus in the market.

2.3 Endogenous Credit Rationing

In this subsection, we illustrate how the discretization of borrower risk into coarse categories endogenously leads to credit rationing. To build intuition, we first consider a simplified scenario in which lenders have no borrower-specific information, corresponding to the case where there is only a single risk label P_0 that applies to all borrowers.

Initially, the lender calculates the interest rate under the assumption that all borrowers participate. Specifically, the lender computes a uniform interest rate using the zero NPV condition:

$$R^{ZNPV}(P_0) = \frac{1 + (1-\alpha)\mathbb{E}[\theta|P_0] - \eta\mathbb{E}[\theta^2|P_0]}{1 - \mathbb{E}[\theta|P_0]}.$$

This rate assumes full participation, meaning all borrowers—including those with relatively low default probabilities—accept the offered interest rate.

However, at this initial interest rate, the participation constraint (PC) of the safest borrowers (those with the lowest risk types θ) is not satisfied. Formally, the participation constraint of a borrower with type θ is given by:

$$U(\theta) = 1 - \alpha - \eta\theta - R(1-\theta) \geq 0.$$

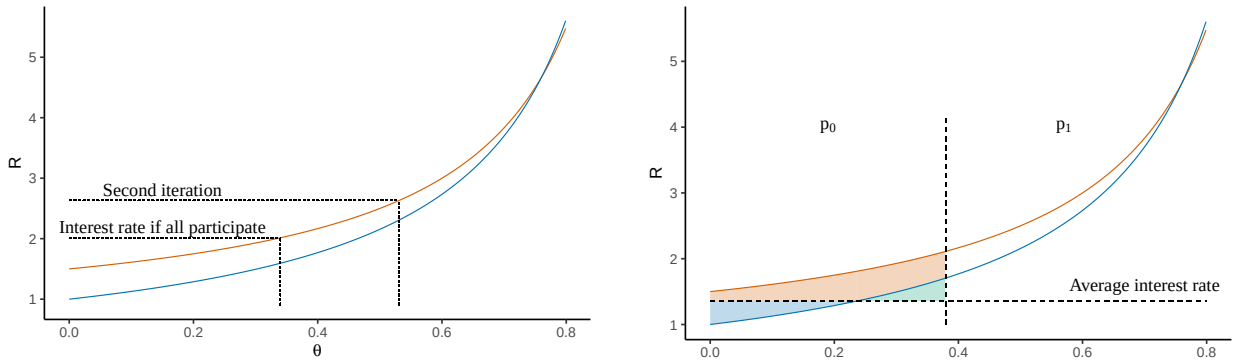
Given the uniform rate $R^{ZNPV}(P_0)$ calculated above, the safest borrowers with the lowest default probabilities (lowest θ) have negative utility, as their cost of borrowing outweighs the benefit of receiving a loan. Consequently, these borrowers rationally choose not to participate. Figure 4, Panel A illustrates this effect: the equilibrium interest rate set to satisfy the zero NPV condition for the entire borrower population is too high to retain the lowest-risk borrowers.

Recognizing this issue, the lender recalculates the interest rate based on the now smaller and riskier sample of borrowers remaining in the market. The new calculation, again using the zero NPV condition but now conditional on the reduced borrower pool, results in a higher interest rate. This second iteration exacerbates the problem, as the safest among the remaining borrowers now also fail to satisfy their participation constraint at the higher interest rate. This process continues iteratively, causing progressively safer borrowers to exit the market at each step, leading the market toward unraveling. Panel A of Figure 4 shows the average interest rate in both the initial and second iterations, illustrating the endogenous credit rationing.

In this way, coarse risk categorization, by imposing uniform interest rates on heterogeneous borrower pools, inherently leads to the exclusion of low-risk borrowers and an endogenous tightening of credit availability due to the adverse selection problem.[7]

[7]The concept of credit rationing has been extensively explored in economic theory, particularly in the context of information asymmetries between borrowers and lenders. Seminal contributions by Jaffee and Modigliani (1969) and Stiglitz and Weiss (1981) laid the foundation for understanding how such asymme-

FIGURE 4: EQUILIBRIUM INTEREST RATES UNDER FULL INFORMATION



Panel A: No information

Panel B: Partinioning into P_0 and P_1

Note: Panel A illustrates the case with no information, where all borrowers are pooled together, leading to an equilibrium interest rate that causes the safest borrowers to exit, triggering market unraveling. The dashed line represents the average interest rate across all borrowers. Panel B shows the equilibrium when borrowers are segmented into risk groups P_0 and P_1. Within P_0, all borrowers receive the same interest rate, as determined by Nash bargaining, ensuring full participation. In contrast, in P_1, the market unravels similarly to Panel A, with credit rationing persisting among the riskiest borrowers. The shaded areas in Panel B correspond to borrower and lender surpluses, with endogenous pooling occurring within P_0, where safer borrowers implicitly subsidize riskier borrowers.

2.4 Lending and Welfare under Discrete Pricing

In this subsection, we examine how the lender's statistical learning algorithm, described in Section 2, shapes equilibrium lending decisions and welfare under discrete risk-based pricing. Specifically, we analyze lending quantities, interest rate structures, and welfare implications resulting from segmenting borrowers into distinct risk groups, P_0 and P_1.

A key result in this framework is that borrowers classified into the safer partition P_0 face no credit rationing. Within this group, all borrowers receive identical loan offers, reflecting a uniform interest rate determined through Nash bargaining as described previously. To be precise, the lender sets the interest rate to extract a fraction γ of the surplus from the safest borrower in partition P_0. By construction, this interest rate is below the maximum acceptable rate implied by the participation constraint (PC) of the safest bor-

tries can lead to credit rationing. Recent work includes theoretical advances on the role of capital regulation and collateral constraints in shaping credit access (Thakor, 1996; Jiménez et al., 2006), as well as empirical studies documenting credit rationing in small business lending and mortgage (Canales and Nanda, 2012; Ambrose et al., 2016).

rower, thus ensuring their participation and preventing market unraveling.

This uniform pricing within P_0 contrasts with the riskier partition P_1, where the lending market unravels similarly to the no-information scenario analyzed earlier. Within P_1, the uniform interest rate exceeds the maximum acceptable rate of safer borrowers in this partition, resulting in endogenous credit rationing. Consequently, when comparing Panels A and B in Figure 4, we observe that introducing partial borrower segmentation reduces the extent of credit rationing but does not eliminate it entirely.

Panel B of Figure 4 displays these equilibrium outcomes explicitly, highlighting the average interest rate and uniform pricing within each risk group. Crucially, all borrowers within P_0 are offered the same interest rate, while in P_1 only riskier borrowers remain in equilibrium.

We now turn to welfare implications. The total surplus generated from lending transactions can be decomposed into borrower (consumer) and lender surplus. For borrowers, the surplus in each partition is the integral of the difference between their participation constraint (PC) interest rate and the equilibrium interest rate offered, aggregated over all borrowers that accept the offer:

$$\text{Borrower Surplus} = \int_{\theta \in P_i} \left(R^{PC}(\theta) - R_{P_i} \right) d\theta,$$

where R_{P_i} denotes the equilibrium interest rate offered to partition P_i, and $R^{PC}(\theta)$ is the borrower's participation constraint.

Analogously, the lender's surplus is given by the integral of the difference between the equilibrium interest rate and the lender's zero net present value (ZNPV) constraint, aggregated across accepted borrowers:

$$\text{Lender Surplus} = \int_{\theta \in P_i} \left(R_{P_i} - R^{ZNPV}(\theta) \right) d\theta.$$

In terms of Figure 4 Panel B, the areas representing these surpluses are as follows:

the sum of the red and green shaded areas illustrates the consumer surplus, reflecting borrowers' gains from trade, while the lender's surplus is represented by the difference between the blue and green shaded areas. Specifically, the green shaded area depicts a transfer from lender surplus to borrower surplus within partition P_0.

This graphical representation highlights a key feature of endogenous pooling: safer borrowers within P_0 implicitly subsidize riskier borrowers. Those at the lowest end of the risk distribution within P_0 pay higher interest rates than under full-information conditions, effectively subsidizing the riskier borrowers who pay interest rates below their individual ZNPV constraints. Indeed, borrowers at the high-risk end of P_0 may receive loan terms that imply a negative expected return to the lender, indicating an endogenous cross-subsidization within the partition. Thus, discrete pricing under statistical learning leads to implicit borrower insurance through endogenous pooling, with implications for pricing, quantity, and welfare.

In the next Section, we explore the effects of increasing the amounts of data variables available to the lender.

3 The Effect of Increased Data Availability

This section presents the impact of increased data availability in lending markets. We first describe how increased data availability impacts risk classification before exploring the effect on equilibrium prices, quantities, and welfare. Finally, we examine distributional implications across borrower groups.

3.1 Modeling Data Growth

We model data growth by representing borrower data as partitions. Specifically, borrowers of type $\theta \in \Theta$ submit data $d(\theta) \in D$. Given the available dataset D, lenders partition the set of borrower types Θ into a family of subsets P that satisfy three properties:

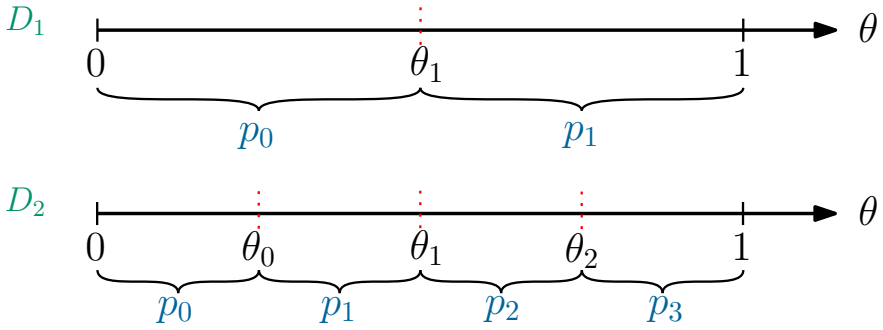
1. $\emptyset \notin P$

2. $\bigcup_{x \in P} x = \Theta$

3. For all $x, y \in P$, $x \neq y \Rightarrow x \cap y = \emptyset$

For each partition subset $p \in P$, the lender sets the optimal interest rate $R(P)$ based on the inferred borrower risk within that subset, following the uniform interest rate assignment and minimum interest rate equations established in Section 2.2 above.

FIGURE 5: Data Growth

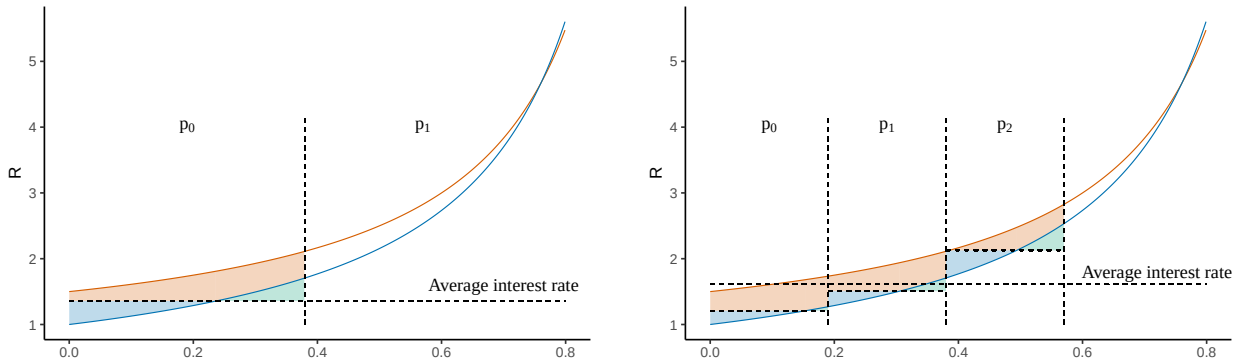


Note: This figure provides an example of data growth and the resulting partitioning of borrower types.

As the lender acquires additional data variables, the resulting dataset grows from an initial dataset D_1 to a larger dataset D_2, with $D_1 \subseteq D_2$. The larger dataset allows for finer partitioning of borrower types, making the sets in P more granular. Consequently, each subset $p \in P$ shrinks in size, and lenders assign more precise and tailored interest rates within each subset, as implied by the equations presented previously. Formally, data growth can be described as a filtration $\mathcal{F} = (p_i)_{i \in I}$ indexed by i, with the property that for indices $i \leq j$, it follows that $p_i \subseteq p_j$. Thus, previous information is always retained and refined as new data becomes available. We show an example of increasingly granular partitions in Figure 5.

Intuitively, as datasets become richer, lenders gain more precise information about borrower risk, enabling them to make more finely tuned lending decisions, thereby increasing pricing granularity as demonstrated in Section 2.2.

FIGURE 6: THE EFFECT OF INCREASED DATA



Panel A: Partinioning into P_0 and P_1 Panel B: Partinioning into P_0, P_1, P_2, and P_3

Note: This figure illustrates the theoretical effects of increased data availability on equilibrium interest rates and credit rationing. Panel A shows equilibrium outcomes under limited data (fewer borrower partitions), while Panel B depicts outcomes with enhanced data availability (additional partitions). Borrowers are segmented into risk partitions (P_i), each assigned a uniform equilibrium interest rate. Red and green shaded areas represent consumer surplus; blue shaded areas minus green areas represent lender surplus.

3.2 The Effect on Prices and Quantities

Figure 6 Panels A and B illustrate how an increase in data availability affects equilibrium outcomes.

Effect on Prices. The equilibrium interest rate within each partition P_i is denoted by R_{P_i}. The introduction of a new data variable mechanically increases the average equilibrium interest rate across all borrowers, as previously rationed higher-risk borrowers are now offered loans. Formally, we have:

$$\Delta\overline{R} = \frac{\sum_i |P_i| R_{P_i}}{\sum_i |P_i|} - \frac{\sum_j |P_j| R_{P_j}}{\sum_j |P_j|} > 0,$$

where the second sum is taken over partitions before data growth and the first sum after data growth, illustrating Prediction 1:

Prediction 1. *The average interest rate for all accepted loans increases as data availability increases.*

Next, we examine Prediction 2: the change in interest rates with respect to data availability is increasing in borrower credit risk θ. To derive this mathematically, recall that the equilibrium interest rate is set via Nash bargaining at the participation constraint (PC) of the safest borrower in partition P_i, thus:

$$R_{P_i}(\gamma) = \gamma \cdot PC(\underline{\theta}_{P_i}) + (1-\gamma) \cdot ZNPV(\underline{\theta}_{P_i}),$$

where $\underline{\theta}_{P_i}$ is the lowest risk borrower in P_i, and $\gamma \in [0,1]$ determines surplus sharing.

Substituting the explicit forms from earlier sections, we have:

$$PC(\underline{\theta}_{P_i}) = \frac{1+(1-\alpha)\underline{\theta}_{P_i}}{1-\underline{\theta}_{P_i}},$$
$$ZNPV(\underline{\theta}_{P_i}) = \frac{1+(1-\alpha)\underline{\theta}_{P_i} - \eta\underline{\theta}_{P_i}^2}{1-\underline{\theta}_{P_i}}.$$

Thus, the interest rate in partition P_i explicitly is:

$$R_{P_i}(\gamma) = \frac{1+(1-\alpha)\underline{\theta}_{P_i} - \eta(1-\gamma)\underline{\theta}_{P_i}^2}{1-\underline{\theta}_{P_i}}.$$

Differentiating with respect to $\underline{\theta}_{P_i}$, we get:

$$\frac{\partial R_{P_i}}{\partial \underline{\theta}_{P_i}} = \frac{(1-\alpha)(1-\underline{\theta}_{P_i}) + \left[1+(1-\alpha)\underline{\theta}_{P_i} - \eta(1-\gamma)\underline{\theta}_{P_i}^2\right] - 2\eta(1-\gamma)\underline{\theta}_{P_i}(1-\underline{\theta}_{P_i})}{(1-\underline{\theta}_{P_i})^2},$$

which simplifies to a clearly positive expression for realistic parameter values, confirming interest rates rise with borrower risk.

Next, the second cross-derivative regarding additional data and borrower risk is:

$$\frac{\partial^2 R_{P_i}}{\partial D \partial \underline{\theta}_{P_i}} > 0,$$

because finer partitions imply higher-risk borrowers (previously rationed) enter the mar-

ket at higher rates, while safer borrowers experience a smaller or even negative rate change. Thus, we formally confirm Prediction 2:

Prediction 2. *The change in interest rates with respect to data availability is positively correlated with borrower credit risk.*

Effect on Quantities. Credit supply increases due to previously rejected borrowers (partition P_2) gaining credit access. This expansion is clearly visible by comparing Panels A and B of Figure 6.

In welfare terms, consumer surplus is represented by the sum of the red and green areas, while lender surplus equals the blue area minus the green area. The welfare change from introducing the new data variable is:

$$\Delta W = \int_{\theta \in P_2} [PC(\theta) - ZNPV(\theta)] d\theta > 0,$$

representing the positive surplus from newly included borrowers.

This yields the third empirical prediction:

Prediction 3. *Introducing additional data reduces loan rejection rates, particularly among higher-risk borrowers.*

3.3 Group Comparisons

Now consider borrower groups labeled W (white) and B (black), with the assumption that on average:

$$\text{For each } i, j, \quad \mathbb{E}[\theta|P_i] \geq \mathbb{E}[\theta|P_j] \Rightarrow \mathbb{E}[g = B|P_i] \geq \mathbb{E}[g = B|P_j].$$

Under this condition, our model predicts: When lenders obtain more data, the relative quantity of credit for group B borrowers weakly increases, and the interest rate gap

between groups W and B also weakly increases:

$$\Delta\left(\frac{Q_B}{Q_W}\right) \geq 0, \quad \Delta(R_B - R_W) \geq 0.$$

Detailed proofs of these predictions are provided in Appendix B.

This concludes the theoretical framework, summarizing that increased data availability leads to higher interest rates, expanded credit supply, reduced but persistent rationing, and nuanced distributional effects. We now empirically test these predictions.

4 Empirical Design

This section outlines institutional design, the datasets, and the empirical methodology used to test the theoretical predictions from our model regarding the distributional effects of increased data availability in mortgage lending.

4.1 Background on the U.S. Mortgage Market

In the early 1990s, the process of mortgage origination relied heavily on manual underwriting procedures. Loan applications were individually reviewed by human underwriters, who applied standardized guidelines provided by government-sponsored enterprises (GSEs), such as Fannie Mae and Freddie Mac, or other secondary market participants. These manual reviews were labor-intensive, resulting in personnel constraints and slow processing times, especially during periods of high application volumes, such as refinancing booms (Straka, 2000; Johnson and Tzur-Ilan, 2025). Additionally, because underwriting standards were primarily based on direct historical experience and simple heuristic rules, the scope for fine-grained risk assessment was limited.

The introduction of automated underwriting systems (AUS) in the mid-1990s transformed the U.S. mortgage market substantially. Freddie Mac and Fannie Mae introduced

their respective systems—Loan Prospector (LP) and Desktop Underwriter (DU)—in pilot phases in 1994, with wider public releases following in 1995 (Maselli, 1994; Straka, 2000). Although both systems aimed to streamline loan processing, their initial implementations differed significantly in terms of underwriting methodology.

Freddie Mac's Loan Prospector represented a major departure from traditional manual underwriting by employing statistically-informed rules derived from detailed analysis of historical loan performance data. LP introduced complexity in credit assessment, allowing interactions among various risk factors to determine borrower eligibility (Maselli, 1994; Harney, 1996). Notably, LP reduced emphasis on mortgage payment-to-income ratios, historically a cornerstone of underwriting standards but shown empirically to have only a weak relationship with mortgage default risk (Foote and Willen, 2018). As a result, lenders using LP were able to extend credit at significantly higher debt-to-income ratios, expanding credit access, especially for borrowers with otherwise strong credit profiles (Johnson and Tzur-Ilan, 2025).

In contrast, Fannie Mae's Desktop Underwriter initially automated the existing manual underwriting rules, improving application processing speed but without altering underwriting standards in a meaningful way (Nixon, 1995; Straka, 2000). DU's statistical and complex underwriting rules were adopted later, toward the end of the 1990s, aligning its underwriting standards more closely with those of LP.

This distinction in the timing and nature of AUS adoption between LP and DU provides a valuable source of variation exploited in empirical analyses. Specifically, Johnson and Tzur-Ilan (2025) document that counties exposed to early adopters of LP experienced an increase in lending at high loan-to-income ratios by approximately 18 percent, compared to counties exposed to early adopters of DU. This expansion in lending, driven by LP's adoption, translated directly into higher house prices, significantly contributing to the housing boom observed between 1993 and 2002.

Thus, the transition from manual underwriting practices to statistically-informed au-

tomated underwriting represented a significant institutional shift within the U.S. mortgage market during the 1990s, profoundly affecting both credit availability and housing market dynamics.

4.2 Data

We utilize two primary datasets in our empirical analysis: (1) the Home Mortgage Disclosure Act (HMDA) data and (2) a proprietary loan performance dataset from a mortgage servicer aggregator. The HMDA data provide near-universal coverage of mortgage applications, originations, and borrower demographics across the U.S.. The loan performance dataset that we use is analogous to the widely used Black Knight (McDash) dataset, capturing detailed information on mortgage characteristics, borrower credit scores, loan performance, and servicing data. Similar datasets have been employed extensively in prior literature (see, e.g., Fuster et al., 2019; Adelino et al., 2016).

To create a comprehensive empirical framework, we merge the HMDA data with the loan performance dataset using methods similar to those described in Buchak and Jørring (2025), which merges HMDA with GSE records. The resulting dataset allows us to match detailed mortgage performance outcomes to borrower demographic and origination characteristics.

4.3 Empirical Design

Our empirical strategy exploits the staggered adoption of automated underwriting systems—specifically, Desktop Underwriter (DU) and Loan Prospector (LP)—as a natural experiment to test the predictions of our theoretical model. This adoption provides a clear empirical counterpart to the theoretical concept of lenders gaining access to an additional data variable, namely the FICO score, thus enabling more granular risk assessment.

We follow Johnson and Tzur-Ilan (2025) by identifying early adopters of DU and LP

using public announcements and leveraging this variation to construct a difference-in-differences (DID) framework. Specifically, we use two complementary approaches. First, we consider the general adoption of either DU or LP as a broad measure of the shift toward automated underwriting. Second, we exploit the differential impact of adopting LP versus DU to isolate the specific effects attributable to adopting statistically-informed underwriting rules, which were initially unique to LP. This dual approach provides robust identification because early adopters of LP effectively constitute a treatment group exposed to both automation and statistically-informed underwriting rules, while early DU adopters serve as a control group exposed only to automation.

Baseline Regression Model: Collapsed Difference-in-Differences (DID). Our baseline empirical specification is a collapsed difference-in-differences approach that compares changes in characteristics in originated loans and loan outcomes before and after the introduction of the new data availability, contrasting early adopters with non-adopters. Specifically, we estimate:

$$\begin{aligned} Outcome_{i,b,c,t} = \beta_1 Post_t + \beta_2 EarlyAdopter_b + \beta_3 Post_t \times EarlyAdopter_b \\ + \boldsymbol{\eta}' \boldsymbol{X}_i + \boldsymbol{\zeta}' \boldsymbol{X}_b + \boldsymbol{\mu}' \boldsymbol{X}_{c,t} + \gamma_c + \epsilon_{i,b,c,t} \end{aligned} \tag{1}$$

Here, the dependent variable $Outcome_{i,b,c,t}$ captures loan decisions (accept or reject) or characteristics (e.g., Debt-to-Income ratio (DTI), Loan-to-Value ratio (LTV), FICO score, or Default). The indices denote individual loans (i), banks (b), counties (c), and years (t). The variable $Post_t$ is an indicator for observations occurring after the treatment year, and $EarlyAdopter_b$ identifies banks that adopted the new data technology early. The coefficient of interest, β_3, captures the differential effect of the new data technology. We include controls at the loan level ($\boldsymbol{X}_i$), bank level ($\boldsymbol{X}_b$), and county-year level ($\boldsymbol{X}_{c,t}$), along with county fixed effects (γ_c).

Saturated Difference-in-Differences (DID) and Event-Study. We further explore robustness through a saturated difference-in-differences specification, which incorporates bank, county, and year fixed effects:

$$\begin{aligned} Outcome_{i,b,c,t} &= \beta_3 Post_t \times EarlyAdopter_b \\ &+ \boldsymbol{\eta}' \boldsymbol{X}_i + \boldsymbol{\zeta}' \boldsymbol{X}_b + \boldsymbol{\mu}' \boldsymbol{X}_{c,t} + \gamma_b + \gamma_c + \gamma_t + \epsilon_{i,b,c,t} \end{aligned} \tag{2}$$

In addition, we adopt an event-study approach to dynamically capture the adoption effects over time:

$$\begin{aligned} Outcome_{i,b,c,t} &= \sum_{t \neq 1993} \beta_t \times EarlyAdopter_b \\ &+ \boldsymbol{\eta}' \boldsymbol{X}_i + \boldsymbol{\zeta}' \boldsymbol{X}_b + \boldsymbol{\mu}' \boldsymbol{X}_{c,t} + \gamma_b + \gamma_c + \gamma_t + \epsilon_{i,b,c,t} \end{aligned} \tag{3}$$

Here, γ_b, γ_c, and γ_t represent bank, county, and year fixed effects, respectively, allowing us to account for unobserved heterogeneity and trends that may confound our estimated treatment effects.

Effect by Subgroup. Finally, we explore heterogeneous effects across borrower subgroups, particularly by race. Specifically, we estimate the following interaction model:

$$\begin{aligned} Rejection_{i,b,c,t} = \beta EarlyAdopter_{b,t} + \eta NonWhite_i + \mu EarlyAdopter_{b,t} \times NonWhite_i \\ + \boldsymbol{\eta}' \boldsymbol{X}_i + \boldsymbol{\zeta}' \boldsymbol{X}_b + \boldsymbol{\mu}' \boldsymbol{X}_{c,t} + \gamma_b + \gamma_c + \gamma_t + \epsilon_{i,b,c,t} \end{aligned} \tag{4}$$

In this specification, the coefficient μ captures the differential effect of early adoption on rejection rates for nonwhite borrowers relative to white borrowers. As before, we control comprehensively for loan-specific ($\boldsymbol{X}_i$), bank-specific ($\boldsymbol{X}_b$), and county-year-specific ($\boldsymbol{X}_{c,t}$) characteristics, as well as fixed effects for banks (γ_b), counties (γ_c), and years (γ_t).

This empirical setup directly aligns with our theoretical model, specifically examining the impact of increased lender access to borrower credit information (FICO scores) on the

market outcomes predicted in the model, including pricing granularity and credit access.

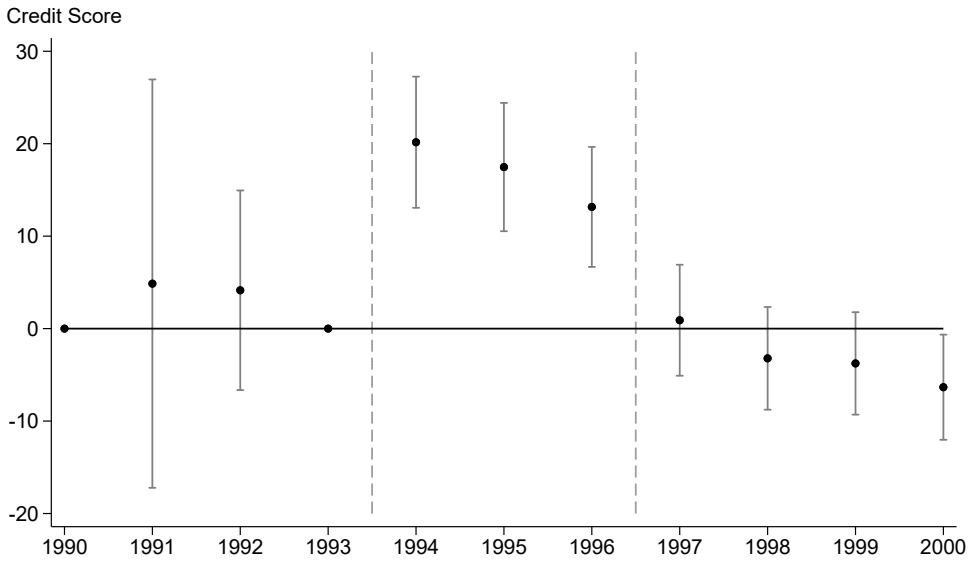
5 Results

In this section, we present our empirical findings. We first present the results on the effect of FICO scores by directly comparing LP with DU. Then, we present the findings that assess the impact of adopting automated underwriting systems (AUS) broadly on mortgage market outcomes. These results are organized into three main areas: the effect of AUS adoption on loan pricing, its effect on credit availability, and the distributional consequences across borrower subgroups.

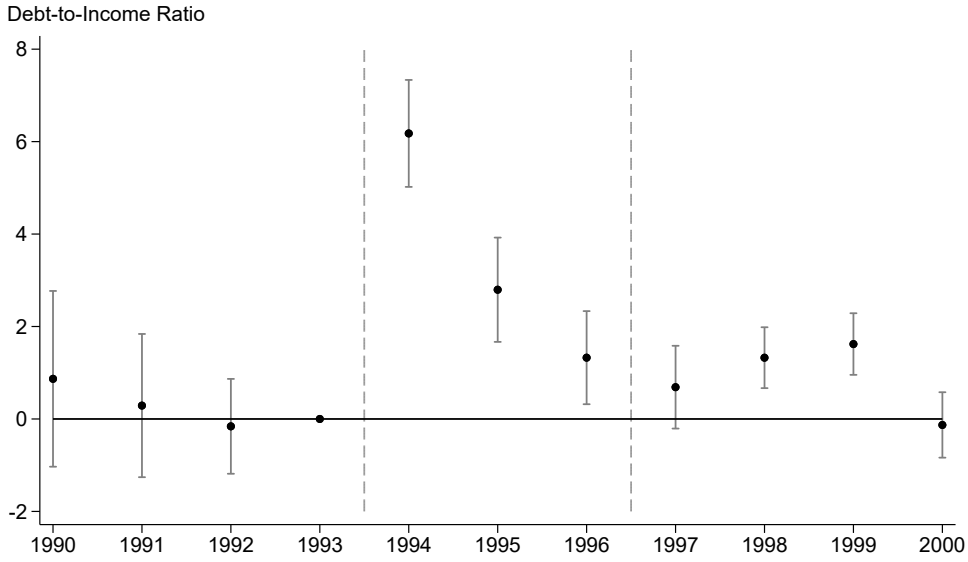
Effect on loan characteristics when comparing LP with DU. Figure 7 provides evidence by illustrating the impact of AUS adoption on credit scores and debt-to-income (DTI) ratios. The figure shows the event-study DID coefficients when comparing LP with DU. Panel A shows a notable improvement in borrower credit scores, and Panel B reveals an increase in DTI ratios following AUS adoption, suggesting a relaxation of previously binding underwriting standards.

Did AUS adoption lead to increased interest rates? Next, we examine whether AUS adoption increased interest rates on average. As illustrated in Table I, interest rates increased significantly after AUS adoption, by approximately 69 to 75 basis points. Notably, when comparing the effects of adopting Loan Prospector (LP) versus Desktop Underwriter (DU), we find that the magnitude of the LP effect is roughly half that of the general early adopter effect. This indicates that about half of the total increase in interest rates was attributable to automation alone, with the other half arising from the introduction of additional data variables.

FIGURE 7: EFFECT ON LOAN CHARACTERISTICS



Panel A: Credit Scores



Panel B: Debt-to-Income Ratios

Note: This figure presents the empirical impact of automated underwriting systems adoption on key borrower characteristics, including FICO scores (Panel A) and Debt-to-Income ratios (DTI, Panel B). The vertical dashed line marks the adoption year. The figure demonstrates significant shifts in loan characteristics consistent with increased lender precision in borrower risk assessment following adoption.

TABLE I: The Effect of AUS on Interest Rates

	Loan Prospector			Desktop Underwriter			LP vs. DU
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Post	0.031	0.052		0.031	0.058		
	(2.57)	(2.82)		(2.57)	(3.43)		
Early adopter		-0.073			-0.065		
		(-1.67)			(-2.85)		
Post X Early adopter		0.75	0.69		0.31	0.32	0.34
		(3.24)	(3.13)		(2.13)	(1.99)	(2.32)
Controls	✓	✓	✓	✓	✓	✓	✓
County FE	✓	✓	✓	✓	✓	✓	✓
Lender FE			✓			✓	✓
Year FE			✓			✓	✓
N	8,624,17	8,624,17	8,624,17	8,624,17	8,624,17	8,624,17	548,92
R^2	0.02	0.10	0.27	0.03	0.10	0.27	0.08

Standard errors are clustered by lender and year, and t statistics are reported in parentheses.

Did AUS adoption lead to increased credit supply? Turning to credit availability, we find strong evidence of increased credit supply following AUS adoption. Figure A3 shows the regression coefficients when the outcome variable is rejection rates. The results demonstrates that AUS adoption significantly reduced mortgage application rejection rates, particularly among historically marginalized borrower groups.

5.1 Soubgroup comparisons

We investigate distributional effects across borrower subgroups, focusing on race, gender, and income.

Figure A2 summarizes the interest rate effects for different borrower groups. AUS adoption disproportionately benefited nonwhite, female, and low-income borrowers by significantly reducing their rejection rates. Specifically, Tables B2, B3, and B2 show that the gap in rejection rates between white and nonwhite borrowers, between male and female, and between high-income and low-income borrowers were substantially reduced post-AUS adoption.

6 Conclusion

This paper develops a theoretical framework and provides empirical evidence on how the rapid increase in data availability, combined with sophisticated statistical learning techniques, reshapes credit market outcomes. Using a novel model of mortgage lending, we illustrate that increased data availability allows lenders to more precisely assess borrower risk, leading to significant reductions in credit rationing and substantial expansions of credit supply.

Our empirical analysis leverages the historical introduction of automated underwriting systems and FICO scores in the U.S. mortgage market. Consistent with the model's predictions, we find robust evidence that greater data availability increased pricing granularity, raised average interest rates, and substantially decreased rejection rates, especially for historically underserved borrowers. Critically, while increased data precision resulted in higher average interest rates due to expanded lending to riskier borrowers, consumer welfare nevertheless improved significantly, driven by increased access to credit among borrowers previously excluded from the market.

These results highlight the complex distributional impacts of increased data availability and underscore the inadequacy of interest rate disparities alone as a measure of consumer welfare. Policymakers and regulators should therefore consider broader welfare metrics that incorporate credit access alongside pricing effects when evaluating the impacts of big data and machine learning in lending markets.

Future research is needed to investigate how additional data sources, beyond credit scoring, influence lending outcomes and distributional fairness across other financial products and borrower segments.

References

Adelino, Manuel, Antoinette Schoar, and Felipe Severino, "Loan Originations and Defaults in the Mortgage Crisis: The Role of the Middle Class," *Review of Financial Studies*, 2016, *29* (7), 1635–1670.

Ambrose, Brent W., James Conklin, and Jiro Yoshida, "Credit Rationing, Income Exaggeration, and Adverse Selection in the Mortgage Market," *Journal of Finance*, 2016, *71* (6), 2637–2670.

Bartlett, Robert, Adair Morse, Richard Stanton, and Nancy Wallace, "Consumer-lending discrimination in the FinTech Era," *Journal of Financial Economics*, 2022, *143* (1), 30–56.

Bayati, Mohsen, Christian Borgs, Jennifer Chayes, Yashodhan Kanoria, and Andrea Montanari, "Bargaining dynamics in exchange networks," 2011.

Breiman, L., J. Friedman, C.J. Stone, and R.A. Olshen, *Classification and Regression Trees*, Taylor & Francis, 1984.

Buchak, Greg and Adam Jørring, "Competition with Multi-Dimensional Pricing: Evidence from U.S. Mortgages," 2025. Working Paper.

Buolamwini, Joy and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in Sorelle A. Friedler and Christo Wilson, eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Vol. 81 of Proceedings of Machine Learning Research PMLR New York, NY, USA 23–24 Feb 2018, pp. 77–91.

Canales, Rodrigo and Ramana Nanda, "A Darker Side to Decentralized Banks: Market Power and Credit Rationing in SME Lending," *Journal of Financial Economics*, 2012, *105* (2), 353–366.

Cho, In-Koo and Akihiko Matsui, "Search theory, competitive equilibrium, and the Nash bargaining solution," *Journal of Economic Theory*, 2013, *148* (4), 1659–1688.

Collard-Wexler, Allan, Gautam Gowrisankaran, and Robin S. Lee, ""Nash-in-Nash" Bargaining: A Microfoundation for Applied Work," *Journal of Political Economy*, 2019, *127* (1), 163–195.

Damdinsuren, Erdenebulgan and Anna Zaharieva, "Expectation formation and learning in the labour market with on-the-job search and Nash bargaining," *Labour Economics*, 2023, *81*, 102311.

Diamond, Douglas W., "Financial Intermediation and Delegated Monitoring," *The Review of Economic Studies*, 07 1984, *51* (3), 393–414.

Foote, Christopher L. and Paul S. Willen, "Mortgage-Default Research and the Recent Foreclosure Crisis," *Annual Review of Financial Economics*, 2018, *10*, 59–100.

Freedman, David and Persi Diaconis, "On the histogram as a density estimator: L 2 theory," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 1981, *57* (4), 453–476.

Fuster, Andreas, Matthew Plosser, Philipp Schnabl, and James Vickery, "The Role of Technology in Mortgage Lending," *Review of Financial Studies*, 2019, *32* (5), 1854–1899.

_ , Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther, "Predictably Unequal? The Effects of Machine Learning on Credit Markets," *The Journal of Finance*, 2022, *77* (1), 5–47.

Gao, Janet, Hanyi Livia Yi, and David Zhang, "Algorithmic Underwriting in High Risk Mortgage Markets," *Working Paper*, 2024.

Gupta, Alisha Haridasani, "Are Algorithms Sexist?," *The New York Times*, November 15, 2019. Available at: https://www.nytimes.com/2019/11/15/us/apple-card-goldman-sachs.html (Accessed: February 19th, 2025).

Harney, Kenneth R., "Automated Underwriting Systems Revolutionize Lending," *Mortgage Banking*, 1996, *56* (10), 39–47.

Hauswald, Robert and Robert Marquez, "Information Technology and Financial Services Competition," *Review of Financial Studies*, 2003, *16* (3), 921–948.

He, Zhiguo, Jing Huang, and Jidong Zhou, "Open banking: Credit market competition when borrowers own the data," *Journal of Financial Economics*, 2023, *147* (2), 449–474.

Jaffee, Dwight M. and Franco Modigliani, "A Theory and Test of Credit Rationing," *The American Economic Review*, 1969, *59* (5), 850–872.

Jiang, Sheila, Adam Jørring, and Douglas Xu, "Bank Technology Adoption and Loan Production in the U.S. Mortgage Market," *Working Paper*, 2024.

Jiménez, Gabriel, Vicente Salas, and Jesús Saurina, "Determinants of Collateral," *Journal of Financial Economics*, 2006, *81* (2), 255–282.

Johnson, Stephanie and Nitzan Tzur-Ilan, "Financial Technology and the 1990s Housing Boom," *Working Paper*, 2025.

Leland, Hayne E. and David H. Pyle, "Informational Asymmetries, Financial Structure, and Financial Intermediation," *The Journal of Finance*, 1977, *32* (2), 371–387.

Maselli, Mark, "Automated Underwriting: Freddie Mac's Loan Prospector," *Mortgage Banking*, 1994, *54* (8), 18–24.

Nixon, Clair, "Fannie Mae's Desktop Underwriter: Expediting Mortgage Processing," *Mortgage Banking*, 1995, *55* (7), 28–32.

NYDFS, "Report on Apple Card Investigation," *New York State Department of Financial Services*, March, 2021.

Quinlan, J. Ross, "Induction of decision trees," *Machine learning*, 1986, *1*, 81–106.

Stevens, Margaret, "NEW MICROFOUNDATIONS FOR THE AGGREGATE MATCHING FUNCTION," *International Economic Review*, 2007, *48* (3), 847–868.

Stiglitz, Joseph E. and Andrew Weiss, "Credit Rationing in Markets with Imperfect Information," *The American Economic Review*, 1981, *71* (3), 393–410.

Straka, John W., "A Shift in the Mortgage Landscape: The 1990s Move to Automated Credit Evaluations," *Journal of Housing Research*, 2000, *11* (2), 207–232.

Thakor, Anjan, "Capital Requirements, Monetary Policy, and Aggregate Bank Lending: Theory and Empirical Evidence," *Journal of Finance*, 1996, *51* (1), 279–324.

Vanasco, Victoria, "The Downside of Asset Screening for Market Liquidity," *Journal of Finance*, 2017, *72* (5), 1937–1982.

Vigdor, Neil, "Apple Card Investigated After Gender Discrimination Complaints," *The New York Times*, November 10, 2019. Available at: https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html (Accessed: February 19th, 2025).

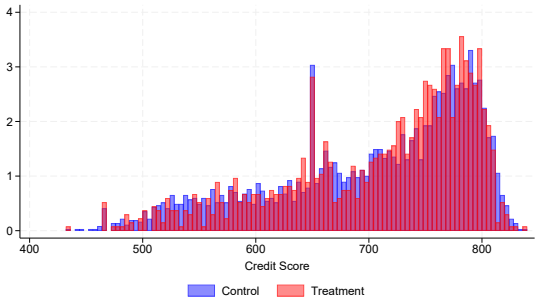
Waddell, Kaveh, "How Algorithms Can Bring Down Minorities' Credit Scores," *The Atlantic*, December 2, 2016. Available at: https://www.theatlantic.com/technology/archive/2016/12/how-algorithms-can-bring-down-minorities-credit-scores/509333/ (Accessed: February 19th, 2025).

Warren, Elizabeth and Doug Jones, "Letter to Regulators on Fintech," *United States Senate*, June 10, 2019.

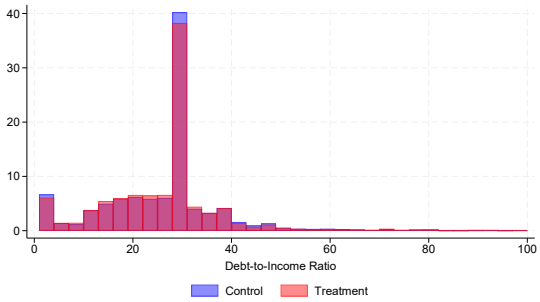
Supplemental Appendix

A Additional Figures and Tables

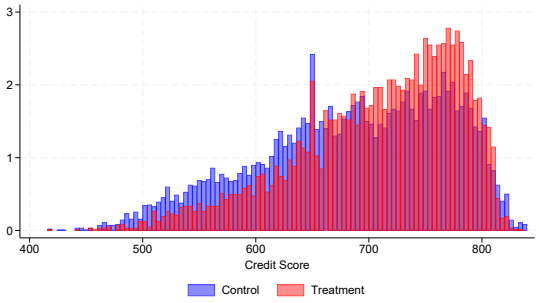
FIGURE A1: DISTRIBUTIONS OF BORROWER CHARACTERISTICS



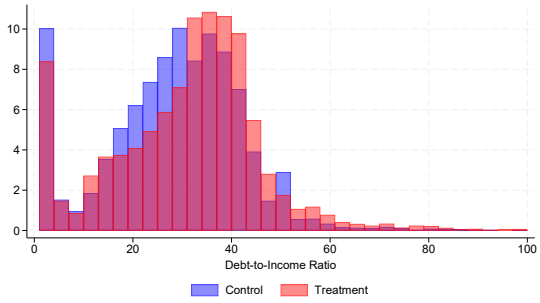
Panel A: Credit Score (1990 - 1993)



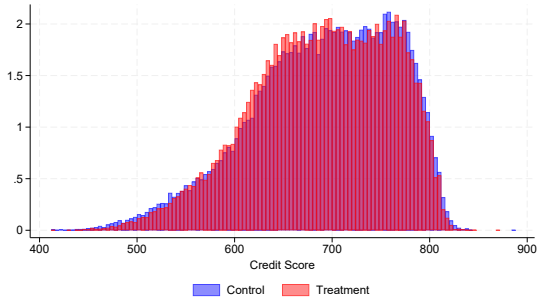
Panel B: DTI (1990 - 1993)



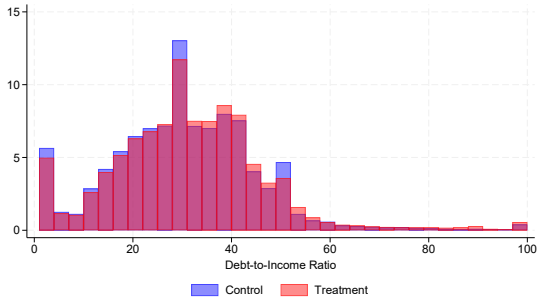
Panel C: Credit Score (1994 - 1996)



Panel D: DTI (1994 - 1996)



Panel E: Credit Score (1997 - 2000)



Panel F: DTI (1997 - 2000)

FIGURE A2: Effect of AUS on Interest Rates

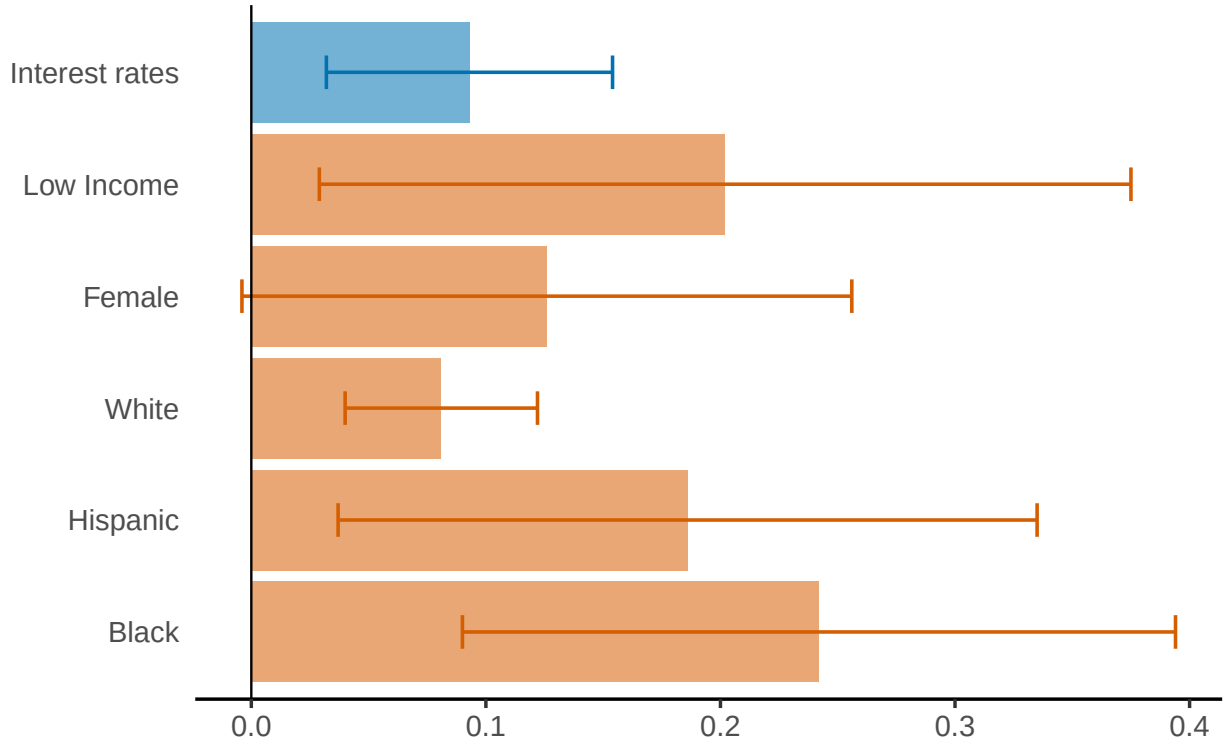


FIGURE A3: Effect of AUS on Rejection Rates

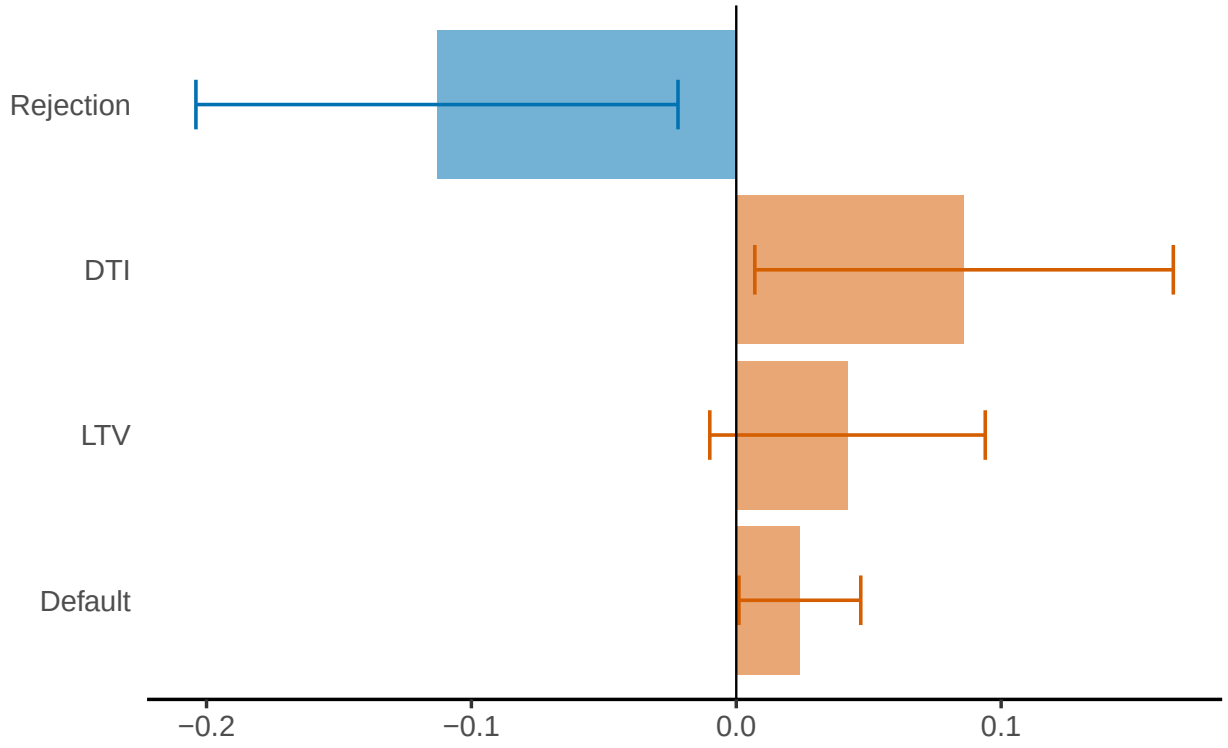


FIGURE A4: (A) Race

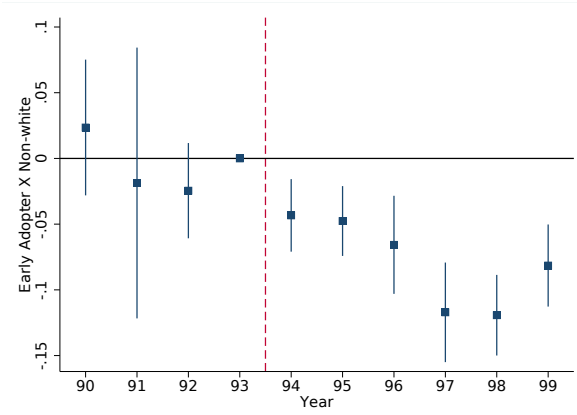
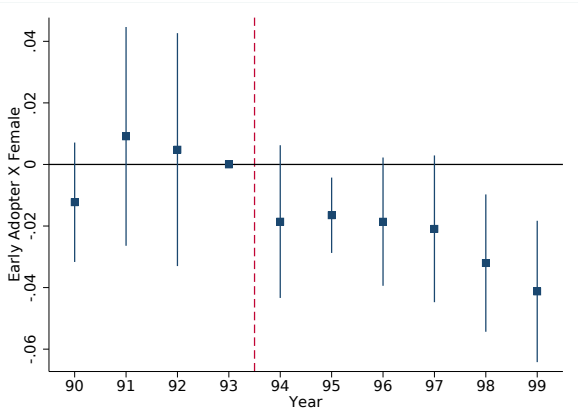


FIGURE A5: (B) Gender



Note: This figure plots the regression coefficients and 95% confidence interval for regressions of the probability of a mortgage application being rejected. Panel A reports the regression coefficients on the triple interaction between "Non-White", "Early Adopter", and each year. Panel B reports the regression coefficients on the triple interaction between "Female", "Early Adopter", and each year. The data source is HMDA.

TABLE B1: Model Notation

Notation	Description
θ	Borrower type (default probability)
$\underline{\theta}, \bar{\theta}$	Minimum and maximum borrower types
$PD(\theta)$	Probability of default for borrower type θ
$LGD(\theta)$	Loss given default for borrower type θ
α, η	Parameters defining loss given default (LGD)
b	Continuation value for borrower if loan does not default
R	Interest rate offered by the lender
$R^{PC}(\theta)$	Maximum interest rate borrower type θ accepts (participation constraint)
$R^{ZNPV}(\theta)$	Minimum interest rate lender offers borrower type θ (zero-NPV constraint)
S	Total surplus from lending
λ	Borrower bargaining power in Nash bargaining
P_i	Borrower partition i determined by lender using statistical learning
R_{P_i}	Uniform equilibrium interest rate in partition P_i
$E[\theta\|P]$	Expected borrower type conditional on partition P
D	Set of data variables observed by lender
N	Number of data variables available to the lender
γ	Fraction of surplus extracted by lender

TABLE B2: Effect of AUS on Credit Supply by Race

	Loan Prospector		Desktop Underwriter		LP vs. DU
	(1)	(2)	(3)	(4)	(5)
Early adopter	-0.10	0.0026	-0.080	-0.015	0.0017
	(-3.53)	(0.17)	(-2.90)	(-0.68)	(0.17)
Non-white	0.14	0.089	0.13	0.087	0.070
	(19.14)	(16.92)	(12.04)	(15.25)	(11.88)
Non-white X Early adopter	-0.14	-0.059	-0.014	-0.0093	-0.038
	(-4.56)	(-2.39)	(-0.75)	(-1.05)	(-2.46)
Controls	✓	✓	✓	✓	✓
County FE	✓	✓	✓	✓	✓
Lender FE		✓		✓	✓
Year FE		✓		✓	✓
N	33,678,822	33,678,434	33,678,822	33,678,434	2,271,536
R^2	0.11	0.27	0.11	0.27	0.08

Standard errors are clustered by lender and year, and *t* statistics are reported in parentheses.

TABLE B3: Effec of AUS on Credit Supply by Gender

	Loan Prospector		Desktop Underwriter		LP vs. DU
	(1)	(2)	(3)	(4)	(5)
Early adopter	-0.15	-0.0040	-0.12	-0.025	0.0061
	(-4.48)	(-0.17)	(-3.68)	(-1.03)	(0.43)
Male	-0.066	-0.023	-0.066	-0.023	-0.0092
	(-9.84)	(-11.23)	(-9.79)	(-11.29)	(-4.95)
Male X Early adopter	0.046	0.011	0.042	0.011	-0.0022
	(6.19)	(2.13)	(7.22)	(3.76)	(-0.49)
Controls	✓	✓	✓	✓	✓
County FE	✓	✓	✓	✓	✓
Lender FE		✓		✓	✓
Year FE		✓		✓	✓
N	32,059,794	32,059,400	32,059,794	32,059,400	1,993,062
R^2	0.10	0.26	0.10	0.26	0.08

Standard errors are clustered by lender and year, and *t* statistics are reported in parentheses.

TABLE B4: Did AUS increase credit supply for low-income borrowers?

	Loan Prospector		Desktop Underwriter		LP vs. DU
	(1)	(2)	(3)	(4)	(5)
Q2 X Early adopter	0.10	0.038	0.096	0.036	-0.0013
	(5.56)	(3.63)	(10.17)	(3.81)	(-0.09)
Q3 X Early adopter	0.17	0.055	0.16	0.051	-0.0057
	(6.20)	(4.11)	(7.58)	(4.17)	(-0.34)
Q4 X Early adopter	0.21	0.067	0.20	0.059	-0.0068
	(5.49)	(3.92)	(6.11)	(3.91)	(-0.34)
Controls	✓	✓	✓	✓	✓
County FE	✓	✓	✓	✓	✓
Lender FE		✓		✓	✓
Year FE		✓		✓	✓
N	3,297,665	3,296,459	3,297,665	3,296,459	222,765
R^2	0.15	0.29	0.15	0.29	0.09

Standard errors are clustered by lender and year, and *t* statistics are reported in parentheses.

B Appendix for the Theoretical Model

In this section, we provide proofs for the theoretical model.

B.1 Baseline NPV constraint (ZNPV)

In this section, we derive the break-even interest rate $R^{ZNPV}(\theta)$ that satisfies the Net Present Value (NPV) constraint for the lender. The NPV is defined as:

$$\text{NPV} = (1 - PD(\theta))R - PD(\theta)(1 - LGD(\theta)) - 1.$$

Setting NPV $= 0$, we obtain the constraint:

$$(1 - PD(\theta))R - PD(\theta)(1 - LGD(\theta)) - 1 = 0.$$

We assume that the probability of default (PD) and loss given default (LGD) are given by:

$$PD(\theta) = \theta, \quad LGD(\theta) = \alpha + \eta\theta.$$

Substituting these into the NPV equation:

$$(1 - \theta)R - \theta(1 - (\alpha + \eta\theta)) - 1 = 0.$$

Expanding the second term:

$$(1 - \theta)R - \theta(1 - \alpha - \eta\theta) - 1 = 0.$$

$$(1 - \theta)R - \theta + \alpha\theta + \eta\theta^2 - 1 = 0.$$

Rearranging for R,

$$(1 - \theta)R = 1 + \theta(1 - \alpha) - \eta\theta^2 \implies$$

$$R^{ZNPV}(\theta) = \frac{1 + \theta(1-\alpha) - \eta\theta^2}{1-\theta}.$$

This represents the equilibrium interest rate that satisfies the zero NPV constraint for a given realization of θ. □

B.2 Deriving $R^{ZNPV}(P)$ for a Risk Group P

We now generalize the zero-NPV condition when the lender does not know the borrower's exact type θ, but only knows that the borrower belongs to a risk category P. Taking expectations over borrower types conditional on the category P, the lender sets an interest rate $R^{ZNPV}(P)$ such that the expected NPV across all borrowers in P is zero:

$$\mathbb{E}[(1-\theta)R - \theta[1-(\alpha+\eta\theta)] - 1 \mid P] = 0.$$

Expanding the expectations and simplifying, we have:

$$(1 - \mathbb{E}[\theta|P])R - \mathbb{E}[\theta|P] + \alpha\mathbb{E}[\theta|P] + \eta\mathbb{E}[\theta^2|P] - 1 = 0.$$

Rearranging terms, we obtain the general zero-NPV interest rate formula for a risk group P:

$$R^{ZNPV}(P) = \frac{1 + (1-\alpha)\mathbb{E}[\theta|P] - \eta\mathbb{E}[\theta^2|P]}{1 - \mathbb{E}[\theta|P]}.$$

This represents the zero-NPV condition for a given risk group P. □

B.3 Deriving $R^{ZNPV}(P)$ for Uniformly Distributed θ

Finally, suppose that the lender further assumes borrowers within each risk category P have types θ uniformly distributed between two endpoints θ_1 and θ_2. Thus, $\theta \sim U[\theta_1, \theta_2]$. The conditional expectations become:

$$\mathbb{E}[\theta|P] = \frac{\theta_1 + \theta_2}{2}, \quad \mathbb{E}[\theta^2|P] = \frac{\theta_1^2 + \theta_1\theta_2 + \theta_2^2}{3}.$$

Substituting these explicitly into the general formula, we have:

$$R_P^{ZNPV} = \frac{1 + (1-\alpha)\frac{\theta_1+\theta_2}{2} - \eta\frac{\theta_1^2+\theta_1\theta_2+\theta_2^2}{3}}{1 - \frac{\theta_1+\theta_2}{2}}.$$

Simplifying the denominator, we obtain the final simplified form of the zero-NPV interest rate under the assumption of a uniform distribution:

$$R_P^{ZNPV} = \frac{2 + (1-\alpha)(\theta_1 + \theta_2) - 2\eta\frac{\theta_1^2+\theta_1\theta_2+\theta_2^2}{3}}{2 - (\theta_1 + \theta_2)}.$$

This formula gives the interest rate that ensures that the lender's expected net present value from lending is zero under the assumption that θ is uniformly distributed between θ_1 and θ_2. □

B.4 Deriving the Effect of Increased Data Availability

B.4.1 Prediction 1: Interest Rates

Given additional data, the equilibrium interest rates satisfy $R_{P_2} > R_{P_1} > R_{P_0}$, directly from monotonicity in borrower risk:

$$\Delta R = R_{P_1} - R_{P_0} > 0,$$
$$R_{P_2} > R_{P_1} > R_{P_0}. \qquad \square$$

B.4.2 Prediction 2: Quantities

The increased quantity from new data equals the integral over previously rejected borrowers:

$$\Delta Q = \int_{\theta \in P_2} d\theta > 0,$$
$$\Delta W = \int_{\theta \in P_2} [PC(\theta) - ZNPV(\theta)] d\theta > 0,$$

which is positive given $PC(\theta) > ZNPV(\theta)$. □

B.4.3 Prediction 3: Group Comparisons

For each i, j:

$$\mathbb{E}[\theta|P_i] \geq \mathbb{E}[\theta|P_j] \Rightarrow \mathbb{E}[g = B|P_i] \geq \mathbb{E}[g = B|P_j].$$

Thus:

$$\Delta \left(\frac{Q_B}{Q_W} \right) \geq 0, \quad \Delta(R_B - R_W) \geq 0. \qquad \square$$