Credit Scores: Performance and Equity*

Stefania Albanesi, University of Miami, NBER and CEPR
Domonkos F. Vamossy, University of Pittsburgh

March 24, 2026

Abstract

Credit scores are the primary tool for allocating consumer credit in the United States, yet their performance has received little systematic evaluation. We benchmark a widely used credit score against a machine learning model trained on the same credit report information and find that 41% of borrowers are assigned to risk categories inconsistent with their actual default rates, with misclassification concentrated among marginalized groups. Low-income borrowers are 12 percentage points more likely to be misclassified than otherwise comparable higher-income borrowers, controlling for repayment history and credit profile characteristics. Demographic differences arise because credit scores overweight profile characteristics — length of credit history, credit mix, and new credit demand — that are weakly predictive of default but systematically lower for young, low-income, and minority borrowers. We develop a competitive lending model and show that this misweighting generates larger within-score dispersion in true default risk for marginalized groups, producing a scoring system that restricts credit access for borrowers who do not pose the risks attributed to them. A model that correctly weights repayment fundamentals reduces misclassification and substantially improves the standing of non-defaulters in marginalized groups – by over 6 percentiles for low-income borrowers– establishing a complementarity between better screening and equitable credit allocation.

JEL Codes: C45; D14; E27; E44; G21; G24.

Keywords: Consumer default; credit scores; machine learning; equity; fairness.

*We are grateful to Manuel Amador, Marco Bassetto, Sera Linardi, Erzo Lutmer, Jonathan Zinman and many seminar participants for useful comments and suggestions. This research was supported in part by the National Science Foundation under Grant No. SES 1824321. This research was also supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. Correspondence to: stefania.albanesi@gmail.com.

1 Introduction

Credit scores are the primary instrument for allocating consumer credit in the United States. When consumers apply for a loan, lenders check their credit score—a single number, generated by a proprietary model, that ranks borrowers by their likelihood of future default.[1] The amount of credit extended and the interest rate charged depend heavily on this score, even for secured debts such as mortgages and vehicle loans Edelberg (2006), and the emergence of algorithmic underwriting based on credit scores is seen as a key factor in the growth of consumer debt over the past three decades Livshits, Mac Gee, and Tertilt (2016). Credit scores are increasingly used outside the lending sphere—in employment screening, insurance pricing, and rental applications (Corbae and Glover (2018), Chatterjee et al. (2025))—amplifying their consequences for economic opportunity.

Despite this central role, credit score performance has never been subjected to independent evaluation against a well-defined benchmark. Credit scoring models are proprietary, and while score developers release high-level performance summaries, the theoretical literature on consumer default and credit markets typically takes the credit score as an unbiased if imperfect signal of default risk (Chatterjee et al. (2023)).[2] Emerging evidence suggests severe limitations: 11% of consumers lack a credit score, with the unscored concentrated among young, low-income, and minority populations (Brevoort, Grimm, and Kambara (2016), Di Maggio, Ratnadiwakara, and Carmichael (2022)); and mortgage delinquencies rose sharply among high credit score borrowers during the 2007–2009 financial crisis, suggesting the scores did not accurately reflect default risk for these consumers Albanesi, DeGiorgi, and Nosal (2022). Yet absent a benchmark, the nature and consequences of these limitations have remained unclear.

This paper provides the first comprehensive evaluation of credit score performance and its implications for equitable credit allocation. We develop a machine learning model of consumer default that uses the same credit report information available to scoring models, targets the same outcome—90+ days delinquency within eight quarters—and benchmark a widely used credit score

[1] The Fair Isaac Corporation introduced the first credit score in 1958 and launched the widely used FICO score in 1989. The VantageScore, introduced in 2006, is the other major scoring model. The Fair Credit Reporting Act of 1970 and the Consumer Credit Reporting Reform Act of 1996 require disclosure of the main factors affecting a consumer's score but not of the model's statistical performance.

[2] As mandated by Section 215 of the Fair and Accurate Credit Transactions Act of 2003, the Federal Reserve Board conducted a comprehensive study of credit scoring, concluding that credit scores were predictive of loan performance but stopping short of evaluating whether existing scoring models were optimal or comparing them against alternatives (FRB (2007)).

against this model.[3] To interpret our findings, we develop a theoretical framework of competitive lending with noisy screening that formalizes the link between how a score weights its inputs, the resulting misallocation of credit, and its demographic incidence.

The economic mechanism we identify is misweighting: credit scores overweight credit profile characteristics—length of credit history, credit mix, and new credit demand—that are weakly predictive of default but systematically lower for young, low-income, and minority borrowers. This overweighting inflates within-score dispersion in true default risk, pooling borrowers with substantially different default probabilities into the same risk category. A model that correctly weights repayment fundamentals reduces this dispersion, improving performance for all borrowers but delivering the largest gains for those penalized by the current weighting scheme. Better screening and equitable credit allocation are thus complements, not substitutes.

Empirically, the credit score misclassifies 41% of consumers by assigning them to a risk category inconsistent with their realized default behavior. Misclassification is concentrated at the bottom of the distribution: 70% of Near Prime and 47% of Subprime borrowers are misclassified. SHAP-based feature attribution confirms the misweighting mechanism: the credit score places 25% of its weight on credit history length and credit mix, which explain only 13% of variation in our model's rankings. This performance advantage is structural rather than driven by informational recency: a version of our model trained once on 2004Q1 data—predating VantageScore 3.0's training data by seven years—and never updated consistently outperforms the credit score throughout the evaluation period.

The performance gains are sharply larger for marginalized groups, as our theoretical framework predicts. The AUC improvement is 12 percentage points for low-income borrowers (0.86 vs. 0.74), 6 points for young borrowers (0.88 vs. 0.82), and 8 points for minority borrowers (0.88 vs. 0.80), compared to 5 points overall. These gains translate directly into ranking improvements: low-income non-defaulters gain over 6 percentiles (approximately 33 credit score points) when ranked by our model, while low-income defaulters lose 7 percentiles—precisely the pattern predicted by our theoretical model. The mechanism is our model's robustness to low-quality data: unlike the credit score, whose performance degrades sharply for borrowers with delinquency histories, thin files, and

[3]For reviews of machine learning applications in economics, see **?**) and **?**). Our model is a hybrid deep neural network/gradient boosting approach, detailed in Section 3 and Appendix B.

limited credit mix, our model maintains discriminatory power across these attribute categories. Since marginalized borrowers are disproportionately represented in these unfavorable categories, they benefit disproportionately from a model that is robust to data quality. A natural concern with benchmarking against a non-acted-upon score is the endogeneity of lending outcomes; we address this formally in Section 4.1 and Appendix F.1, showing that the acted-upon score problem works against our model in the comparison and cannot explain the demographic pattern of our results.

Our paper contributes to several strands of the literature. First, we contribute to the evaluation of credit score performance. Credit scores are foundational to models of consumer credit markets (Chatterjee et al. (2023), Chatterjee et al. (2025)), yet the assumption that they provide an unbiased signal of default risk has rarely been tested. Bach et al. (2023) find that credit scores contain less information about low-income and minority borrowers; Di Maggio, Ratnadiwakara, and Carmichael (2022) document the consequences of being "credit invisible." We complement these papers by identifying a specific mechanism—misweighting—that degrades performance and showing that correcting it improves both accuracy and equity. Immorlica, Sztutman, and Townsend (2025) show that optimal credit scores may deliberately suppress predictive information to prevent adverse selection unraveling; we document a distinct distortion in which traditional scores misweight information they do use.

Second, we contribute to the debate on algorithmic fairness in screening. Fuster et al. (2022) find that replacing logistic regression with a random forest for mortgage default prediction widens racial disparities in approval rates. Our result is distinct: the improvement comes from reweighting the factors entering the score, not from greater algorithmic flexibility. Blattner and Nelson (2021) argue that sophisticated models cannot overcome compositional differences in marginalized borrowers' attribute distributions. We show that the relevant question is whether the model is robust to these differences—ours maintains performance where the credit score's accuracy breaks down, effectively neutralizing data bias as a source of demographic disparities. More broadly, we extend the growing literature evaluating fairness properties of institutional screening systems—including Obermeyer et al. (2019) on healthcare algorithms, Kline, Rose, and Walters (2024) on employer screening, and Avenancio-León and Howard (2022) on property tax assessment—to credit scoring, where we identify the specific mechanism and show that correcting it improves both accuracy and equity

simultaneously. Methodologically, our approach of benchmarking an institutional screening tool against an algorithm parallels Kleinberg et al. (2018), who compare judicial bail decisions to machine predictions.

Third, credit scores are used to screen job applicants, price insurance, and evaluate tenants Corbae and Glover (2018). Our finding that credit scores misclassify 41% of borrowers—with misclassification correlated with age, income, and race—implies that these non-lending applications may amplify rather than reduce economic disparities, as the informational basis for screening is weaker than commonly assumed.

The remainder of this paper is structured as follows. Section 2 presents the theoretical framework. Section 3 describes the data and our machine learning model. Section 4 compares our model to the credit score. Section 5 examines the relationship between performance and equity. Section 6 concludes.

2 Theoretical Framework

We develop a simple model of competitive lending with credit scoring to illustrate how misweighted credit scores lead to misallocation of credit and how this misallocation falls disproportionately on marginalized populations.

2.1 Economic Environment

Borrowers. There is a unit mass of borrowers, each seeking to borrow an amount B. Borrower i has a true probability of repayment $1-p_i$, where $p_i \in (0,1)$ is the probability of default. Borrowers have limited liability: in the event of default, the lender receives nothing.

True default risk is determined by two classes of observable characteristics:

$$p_i = \beta_x x_i + \beta_z z_i + \varepsilon_i \tag{1}$$

where $x_i \in \mathbb{R}$ captures repayment fundamentals—payment history and amounts owed—and $z_i \in \mathbb{R}$ captures credit profile characteristics—length of credit history, credit mix, and indicators of new credit demand. The noise term ε_i is i.i.d. with mean zero and is independent of (x_i, z_i). The key

assumption is that repayment fundamentals are substantially more predictive of default than credit profile characteristics:

Assumption 1 (Predictive hierarchy)**.** $\beta_x \gg \beta_z > 0$.

This assumption is motivated by the empirical evidence we present in Section 4, where amounts owed and payment history jointly account for 82% of the variation in default risk rankings in our model, compared to 18% for length of credit history, credit mix, and new credit combined.

Demographic groups. There are two demographic groups, $g \in \{M, N\}$, with population shares π_M and π_N, representing marginalized (young, low-income, minority) and non-marginalized borrowers. Group membership is unobservable to lenders, consistent with fair lending regulations that prohibit the use of protected characteristics in credit decisions. The groups differ in the distribution of credit profile characteristics but not in repayment fundamentals conditional on true risk:

Assumption 2 (Demographic differences)**.** *(i)* $E[z_i \mid g = M] < E[z_i \mid g = N]$*;*

(ii) $F(x \mid p, g = M) = F(x \mid p, g = N)$ *for all* x, p*.*

Part (i) states that marginalized borrowers have shorter credit histories, narrower credit mix, and higher credit demand on average. Part (ii) states that, conditional on true default risk, the distribution of repayment fundamentals does not differ across groups. Together, these assumptions imply that group differences in credit scores arise from z, not from inherent differences in creditworthiness.

2.2 Credit Scoring and Lending

Credit scores. A credit scoring model aggregates the observable characteristics (x_i, z_i) into a scalar signal:

$$s_i = w_x x_i + w_z z_i \tag{2}$$

The score is the sole input available to lenders for pricing and allocation decisions. An optimal scoring model sets weights proportional to the true coefficients, $w_x^*/w_z^* = \beta_x/\beta_z$, minimizing the residual variance of default risk conditional on the score. The legacy credit score overweights credit profile characteristics relative to their predictive contribution:

Assumption 3 (Score misweighting)**.** $w_z/w_x > \beta_z/\beta_x$.

This assumption is directly supported by our feature attribution analysis. Table 5 shows that conventional credit scores assign 35% of weight to credit profile characteristics (length of credit history 15%, credit mix 10%, new credit 10%), while our model assigns only 18% (8%, 5%, 5%, respectively).

Competitive lending. A competitive lending sector observes only the credit score s_i and extends to each borrower an amount $q(s_i) \cdot B$, where $q(s_i) \in [0, 1]$ determines the fraction of the requested loan that is funded. The borrower owes repayment of B: with probability $1 - p_i$ she repays B, and with probability p_i she defaults and repays nothing.

Lenders set $q(s)$ to satisfy a zero-expected-profit condition. The lender's expected profit on a borrower with score s is:

$$[1 - \bar{p}(s)]\, B - q(s) \cdot B = 0 \tag{3}$$

where $\bar{p}(s) \equiv E[p_i \mid s_i = s]$ is the average default probability among borrowers with score s. This yields:

$$q(s) = 1 - \bar{p}(s) \tag{4}$$

Higher default risk in a score group lowers $q(s)$, directly reducing the amount of credit extended. A borrower accepts the loan if her expected surplus is non-negative:

$$q(s_i) \cdot B - (1 - p_i)B \geq 0 \tag{5}$$

The first term is the amount received; the second is the expected repayment cost. Substituting (4), participation requires $p_i \geq \bar{p}(s_i)$: borrowers whose true default risk is below the average for their score group—i.e., borrowers who are <u>better</u> than their score implies—decline the loan because the credit extended does not compensate for their expected repayment. These are precisely the borrowers who cross-subsidize higher-risk borrowers within the same score group.

2.3 Misallocation from Misweighted Scores

Within any score group s, there is residual heterogeneity in true default risk. Define the within-score dispersion of default risk as:

$$V(s) \equiv \text{Var}(p_i \mid s_i = s) \tag{6}$$

When the score is optimally weighted, $V(s)$ reflects only the irreducible noise ε_i. When the score overweights z, it groups together borrowers with very different true risks—those with strong repayment fundamentals but poor credit profiles are pooled with those who have weak fundamentals but favorable profiles. This generates two forms of misallocation.

Cross-subsidization. Within a score group, borrowers with $p_i < \bar{p}(s)$ are offered less credit than their true risk warrants—they receive $q(s) \cdot B = [1 - \bar{p}(s)]B$ when they should receive $[1 - p_i]B > [1 - \bar{p}(s)]B$—while borrowers with $p_i > \bar{p}(s)$ receive more credit than warranted. The magnitude of cross-subsidization within score group s is proportional to $V(s)$.

Inefficient exclusion. Borrowers with $p_i < \bar{p}(s_i)$ decline the loan because the credit extended does not compensate for their expected repayment cost. Among these, many are creditworthy borrowers whose true default risk is low, but who are pooled into score groups with high average default rates due to score misweighting. Define the set of inefficiently excluded borrowers as:

$$\mathcal{E} = \{i : p_i < \bar{p}(s_i) \text{ and } p_i < \bar{p}(s_i^*)\} \tag{7}$$

where s_i^* denotes the score under optimal weighting. These are borrowers who would participate under a correctly weighted score—because the optimal score would place them in a group with lower $\bar{p}$, raising q to a level that makes borrowing worthwhile—but are excluded under the legacy score because misweighting inflates the average default rate in their score group.

Proposition 1 (Misallocation increases with misweighting)**.** *The within-score variance $V(s)$ is increasing in the degree of misweighting $w_z/w_x - \beta_z/\beta_x$. Consequently, the measure of inefficiently excluded borrowers $|\mathcal{E}|$ is increasing in the degree of misweighting.*

Proof sketch. When $w_z/w_x > \beta_z/\beta_x$, the score assigns excessive discriminatory power to z, which has low predictive content for p. Borrowers with similar p but different z are assigned to different score groups, while borrowers with different p but offsetting x and z are pooled. Formally, the

projection of p onto s leaves a larger residual when s overweights z, increasing $V(s)$. The result for $|\mathcal{E}|$ follows because higher $V(s)$ raises $\bar{p}(s)$ in score groups that contain a mixture of truly high- and low-risk borrowers, pushing the participation threshold $p_i \geq \bar{p}(s)$ above the true risk of many creditworthy borrowers. □

This proposition provides a structural interpretation of our empirical finding that 41% of borrowers are misclassified by the credit score (Table 3): the within-score dispersion in true default risk is large because the score overweights factors that are weakly correlated with default.

2.4 Demographic Implications

We now show that the misallocation from misweighted scores falls disproportionately on marginalized borrowers.

Proposition 2 (Score compression of marginalized borrowers)**.** *Under Assumptions 1–3, within low score groups, the expected default rate of marginalized borrowers is lower than that of non-marginalized borrowers:*

$$E[p_i \mid s, g = M] < E[p_i \mid s, g = N] \quad \textit{for } s < \textit{median}(s) \tag{8}$$

Consequently, marginalized borrowers receive less credit relative to their true risk, and the rate of inefficient exclusion is higher for group M.

Proof sketch. Consider borrowers in a low score group s. Group N borrowers in this group have relatively high z (by Assumption 2), so they must have poor repayment fundamentals x to generate a low overall score. Group M borrowers, who have low z on average, can arrive at the same low score with better x. Formally, from $s = w_x x_i + w_z z_i$ and $E[z \mid M] < E[z \mid N]$, we have $E[x \mid s, M] > E[x \mid s, N]$ for low s. Since $\beta_x \gg \beta_z$, the group difference in x dominates the group difference in z in determining p, yielding $E[p \mid s, M] < E[p \mid s, N]$. Because the credit extended $q(s) \cdot B$ is set according to $\bar{p}(s)$, group M borrowers—who are better than average within low score groups—are disproportionately excluded by the participation constraint $p_i \geq \bar{p}(s)$. □

This result maps directly to Table 4: among borrowers classified as Subprime by the credit score, those reclassified as Near Prime by our model (who have better repayment fundamentals

despite poor credit profiles) have a realized default rate of 17%, compared to 48% for those who remain Subprime—a gap of over 30 percentage points.

Proposition 3 (Correcting misweighting benefits marginalized groups)**.** *Moving from the legacy score to the optimal score reduces the within-score variance $V(s)$ for all s, but the reduction is larger in score groups where group M is overrepresented. The fraction of borrowers in $\mathcal{E}$ who belong to group M declines, and the average credit extended $q(s_i) \cdot B$ to group M borrowers increases relative to that of group N.*

Proof sketch. Under the legacy score, group M is overrepresented in low score groups due to low z, and the within-score variance in these groups is inflated by the pooling of group M borrowers (with low z but heterogeneous x) and group N borrowers (with high z but poor x). The optimal score, by reducing the weight on z, separates these types: group M borrowers with good x move up in the ranking to score groups with lower $\bar{p}(s)$ and higher $q(s)$, while group N borrowers with poor x move down. Because group M has more mass in the region where the reweighting matters—low score groups where z was the primary reason for a poor rating—the increase in credit extended and the reduction in inefficient exclusion are concentrated in group M. □

This proposition delivers the paper's central finding: better performance and greater equity are complements, not substitutes. The source of the scoring model's poor performance (overweighting z) is the same as the source of its inequitable impact (penalizing groups with low z).

2.5 Welfare

Since lenders earn zero profits, total surplus from lending equals aggregate borrower surplus:

$$W = \int \mathbf{1}[p_i \geq \bar{p}(s_i)] \cdot [q(s_i) - (1 - p_i)]B \, di \tag{9}$$

where the indicator restricts attention to borrowers who accept the loan. The welfare gain from moving to the optimal score decomposes as:

$$\Delta W = \underbrace{\int_{\mathcal{E}} [q^*(s_i^*) - (1 - p_i)]B \, di}_{\text{Inclusion gain}} + \underbrace{\int_{\mathcal{O}} [(1 - p_i) - q(s_i)]B \, di}_{\text{Exclusion gain}} \tag{10}$$

where $q^*(s) = 1 - \bar{p}^*(s)$ denotes the credit fraction under optimal scoring, $\mathcal{E}$ is the set of borrowers inefficiently excluded under the legacy score but included under the optimal score, and $\mathcal{O}$ is the set inefficiently included under the legacy score but excluded under the optimal score. Both terms are positive, and by Proposition 3, the inclusion gain is disproportionately concentrated among group M borrowers.

Discussion Two features of the model merit emphasis. First, the performance–equity complementarity in Proposition 3 is distinct from the finding in Fuster et al. (2022) that more flexible scoring algorithms can widen demographic disparities. In our framework, the improvement comes not from learning group-specific patterns in repayment fundamentals, but from reweighting known factors to better reflect their predictive content—a correction that benefits marginalized groups precisely because they are penalized by the current misweighting. Second, the model addresses the concern in Blattner and Nelson (2021) that sophisticated scoring cannot overcome compositional differences in attribute distributions across groups. Proposition 3 shows that the relevant question is not whether marginalized groups have different distributions of z (they do, by Assumption 2), but whether the scoring model is robust to these differences. A model that correctly weights x relative to z is less sensitive to the distribution of z, and therefore less sensitive to the compositional differences that drive data bias.

3 Prediction Model

We now describe the machine learning model we use to benchmark credit score performance. The model targets the same default outcome as conventional credit scoring models—a 90+ days delinquency on any debt within the subsequent eight quarters—and uses only information from credit reports that is permitted under current legislation. Since credit scores provide only an ordinal ranking of consumers by default risk, we use only the ordinal ranking implied by our model's predicted default probabilities for our comparison.

3.1 Prediction Problem

We adopt a discrete-time formulation for periods $t = 0, 1, \ldots, T$, each corresponding to a quarter. For individual i, let $D_t^i \in \{0, 1\}$ indicate whether the borrower is 90 or more days past due on any trade in quarter t.[4] Our target outcome, following the convention of credit scoring models, is whether the borrower experiences any such delinquency over the subsequent eight quarters:

$$Y_t^i = \begin{cases} 0 & \text{if } D_n^i = 0 \text{ for all } n = t, \ldots, t+7 \\ 1 & \text{otherwise.} \end{cases} \tag{11}$$

We posit that the probability of this outcome depends on a vector of observable credit report variables $X_{t-1}^i \in \mathbb{R}^{d_X}$, which includes the borrower's delinquency status D_{t-1}^i. Fixing a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an information filtration $(\mathcal{I}_t)_{t=0,1,\ldots,T}$, we specify a probability function $f_\theta : \mathbb{R}^{d_X} \to [0, 1]$ satisfying

$$\mathbb{P}[Y_t^i = 1 \mid \mathcal{I}_{t-1}] = f_\theta(X_{t-1}^i) \tag{12}$$

where θ is a vector of parameters to be estimated. The model outputs a continuous value between 0 and 1, interpretable as an estimate of the probability that borrower i will experience a 90+ day delinquency on any existing account within eight quarters, given the credit report information available at time $t - 1$.[5]

3.2 Model Architecture

We adopt a hybrid deep neural network/gradient boosting model. Models in this class display high performance in settings with high-volume data and multidimensional nonlinear interactions among variables.[6] The deep neural network component excels at capturing complex patterns in the data, while the gradient boosting component improves predictive accuracy by iteratively correcting residual errors. Albanesi and Vamossy (2019) extensively compare alternative machine learning architectures for this prediction problem and show that this hybrid approach delivers the best

[4] This is a flow measure of new delinquency in each quarter, not a stock of cumulated past delinquency. A borrower who was delinquent in quarter $t - 1$ but cures by quarter t has $D_t^i = 0$.

[5] Our prediction target encompasses delinquency on all existing accounts, not originations of new credit. This aligns with the account-management function of credit scores.

[6] Sirignano, Sadhwani, and Giesecke (2018) and Fuster et al. (2022) apply deep learning models to mortgage default risk.

performance. Appendix B describes each component model and the rationale for combining them.

3.3 Empirical Implementation

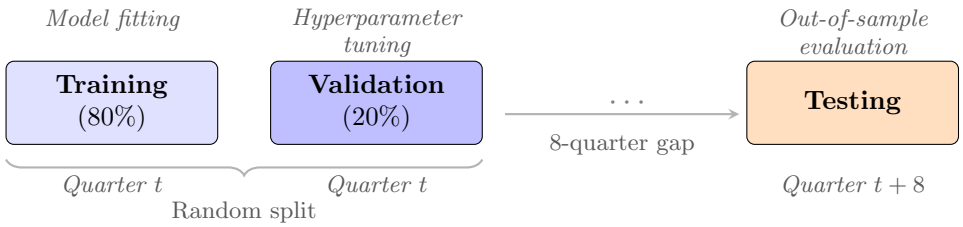
Data. We use anonymized credit report data from the Experian credit bureau, collected quarterly from 2004Q1 to 2018Q2 for a nationally representative panel of approximately one million consumers. The data encompass over 200 variables covering credit and bank cards, other revolving credit, auto and installment loans, business loans, first and second mortgages, home equity lines of credit, student loans, hard inquiries by product type, and collections. For each loan type, the data include the number of trades, outstanding balance and available credit, monthly payments, and delinquency status. The data also include each borrower's credit score for every quarter and an estimate of household labor income based on IRS data. As required for credit report data, demographic characteristics such as gender, marital status, and race are not included, though borrower addresses are available at the zip code level. The data and sampling strategy are described in detail in Appendix A, where Table 11 reports summary statistics for the sample population, including age, household income, credit score, and incidence of default after some technical sample restrictions. The fraction of consumers in default according to our definition is 18.4%.

Features. We select 79 features, listed in Table 12, designed to cover the factors that conventional credit scoring models report as important: payment history, amounts owed, length of credit history, credit mix, and new credit. This broad feature set also maximizes the number of consumers for whom we can generate a prediction.

Training and evaluation. Given the eight-quarter prediction horizon, we ensure fully out-of-sample evaluation by separating training and testing data by eight quarters. Specifically, we train and validate the model on quarter t data (using a random 80/20 split for training and validation) and evaluate it on quarter $t+8$ data. For example, default predictions for 2008Q2 are generated by a model trained exclusively on 2006Q2 data. This temporal separation ensures that no information from the evaluation period contaminates the training process.[7] This procedure yields out-of-sample

[7]This temporal cross-validation approach in the credit scoring context was also adopted by Khandani, Kim, and Lo (2010).

Figure 1: Training and Evaluation Design



Example: Train on 2006Q2 data ⟶ predict defaults in 2008Q2

Notes: The model is trained and validated on data from quarter t, split randomly 80/20. Out-of-sample evaluation uses data from quarter $t + 8$, ensuring full temporal separation equal to the prediction horizon.

Table 1: Y_t^i Frequency and Transitions

Frequency		Transition probability	No default	Default
Default	0.184	No default	0.8729	0.1271
No default	0.816	Default	0.1460	0.8540

Notes: Average default rates and average quarterly transition frequency 2006Q1-2016Q2. Source: Authors' calculations based on Experian Data.

default predictions for every consumer with a non-empty credit record in each quarter from 2006Q1 to 2016Q2. Our strategy is visualized in Figure 1.

Equation 12 defines a theoretical transition process for the default outcome. Table 1 reports its empirical counterpart. The default and non-default states are highly persistent: 85% of consumers in default and 87% of consumers not in default remain in the same state in the subsequent quarter.

The model achieves an average out-of-sample AUC of 0.9076 across all testing quarters (2006Q1–2016Q2), with limited variation over time (standard deviation of 0.008). The average in-sample AUC on training data is 0.9182. The gap of approximately 0.010 between the two reflects the eight-quarter temporal separation between training and evaluation windows: the model is always assessed on an economic environment that differs from the one it was trained on, and performance differences track the degree of regime divergence rather than model instability. Appendix B provides full details.

4 Comparison with Credit Score

We now compare the performance of our model to a widely used conventional credit score. Since the credit score is an ordinal ranking of borrowers by default risk, we compare rankings rather than levels. We segment the credit score, which ranges from 300 to 850, into percentiles. For our model, whose output $\hat{Y}(X)$ is a predicted default probability, we rank consumers by $1 - \hat{Y}(X)$ and segment the resulting distribution into 100 equal-sized bins.

Our comparison is necessarily conceptual rather than fully statistical, as credit scoring models are proprietary: the law requires disclosure only of the broad factors driving score variation, not the specific features, functional form, or training data. Nonetheless, both scores target the same outcome using the same type of information, making the comparison economically informative.

4.1 Validity of the Comparison

A natural concern with comparing two scoring models is the acted-upon score problem: if the credit score being evaluated was used by lenders to make credit decisions, the resulting selection (denial of credit to borrowers the score deems risky) and treatment effects (overlending to borrowers the score overrates) create a mechanical bias against the acted-upon score in any out-of-sample evaluation.[8]

This concern is substantially mitigated in our setting. The credit score in our data is VantageScore 3.0, developed in 2013 using data from 2011–2012. During our evaluation period (2006Q1–2016Q2), the predominant scores used in lending decisions were FICO Score 8 and earlier FICO vintages. VantageScore 3.0 was not widely adopted by lenders until 2018—after our sample

[8]See Appendix F.1 for a formal treatment. The intuition is straightforward: when a score's errors influence who receives credit and on what terms, the correlation between those errors and observed default outcomes biases performance comparisons against the acted-upon score.

ends.[9] Consequently, the selection and treatment biases that arise from acting on a score's errors operated through the FICO scores used in practice, not through the VantageScore 3.0 we evaluate. Both VantageScore 3.0 and our model were effectively not acted upon during our sample period, placing them on a level playing field. Arnold, Dobbie, and Hull (2022) address the closely related problem of constructing nondiscriminatory algorithms when outcomes are observed only for those selected by the existing system; our comparison design, which evaluates two non-acted-upon scores against realized defaults, sidesteps this selection concern directly.

To the extent that VantageScore 3.0 and the FICO scores in use are correlated—as they are, since both draw on similar credit report information—some residual bias may remain. However, this residual bias is attenuated relative to the canonical case where the evaluated score is itself the one used in lending, and it operates symmetrically on VantageScore 3.0 and our model, as both are correlated with FICO through shared inputs without having directly determined lending outcomes.

4.2 Statistical Performance

Figure 2 plots AUC scores for the credit score and our model over time. Our model achieves an average AUC of 0.91, compared to 0.86 for the credit score. The credit score's AUC starts above 0.86 but declines to approximately 0.85 after 2011Q1 before slowly recovering, indicating a deterioration in performance following the Great Recession. Our model's superior discriminatory power is persistent across the full sample and is most pronounced during periods of economic instability. A natural question is whether this advantage reflects our model's weighting of credit report factors or simply its access to more recent training data through quarterly retraining. To

[9]Throughout our sample period, Fannie Mae and Freddie Mac exclusively required Classic FICO scores for mortgage originations, a monopoly not legislatively challenged until the Credit Score Competition Act of 2018, enacted as part of the Economic Growth, Regulatory Relief, and Consumer Protection Act (S. 2155). The FHFA did not issue its first Request for Input on alternative credit score models until 2017—after our sample ends—and VantageScore 4.0 was not validated for GSE use until October 2022. More broadly, FICO accounted for over 90 percent of credit scores sold to firms for lending decisions during this period (Consumer Financial Protection Bureau, "Analysis of Differences between Consumer- and Creditor-Purchased Credit Scores," September 2012; available at `https://files.consumerfinance.gov/f/201209_Analysis_Differences_Consumer_Credit.pdf`). Oliver Wyman market adoption studies, commissioned by VantageScore beginning in 2015, show that VantageScore usage grew approximately 20 percent per year from a low base, reaching 12.3 billion total score pulls only by 2018–2019—a figure that includes non-decisional uses such as marketing, pre-screening, and portfolio monitoring (VantageScore, "Market Adoption: 12.3 Billion and Counting"; available at `https://www.vantagescore.com/resources/knowledge-center/market-adoption-12-3-billion-counting`). Consumer-facing platforms such as Credit Karma adopted VantageScore 3.0 for free credit monitoring, but these scores served an educational rather than decisional role. Also see `https://www.vantagescore.com/lenders/2024-model-assessment/` for adoption timelines. The limited adoption of VantageScore 3.0 for actual credit decisions during our sample period mitigates concerns that our results are affected by acted-upon score bias.

address this, Figure 2 also plots a frozen baseline: a model with the same architecture and features, trained once on 2004Q1 data and never updated — a training vintage that predates VantageScore 3.0's training data by seven years. The frozen model consistently outperforms the credit score throughout the evaluation period, confirming that the performance gap is structural rather than driven by informational recency. The credit score's performance decline after 2011Q1 – coinciding with the end of its training period — has no counterpart in the frozen model, suggesting that the credit score's reliance on weakly predictive factors makes it more vulnerable to distributional drift. Retraining improves our model modestly, but the gap between the frozen baseline and the credit score dominates the gap between the rolling and frozen models. Figures 8 and 9 in Appendix C report consistent results for AUC, as well as Spearman rank and Kendall correlations, for our model at eleven additional freeze dates.

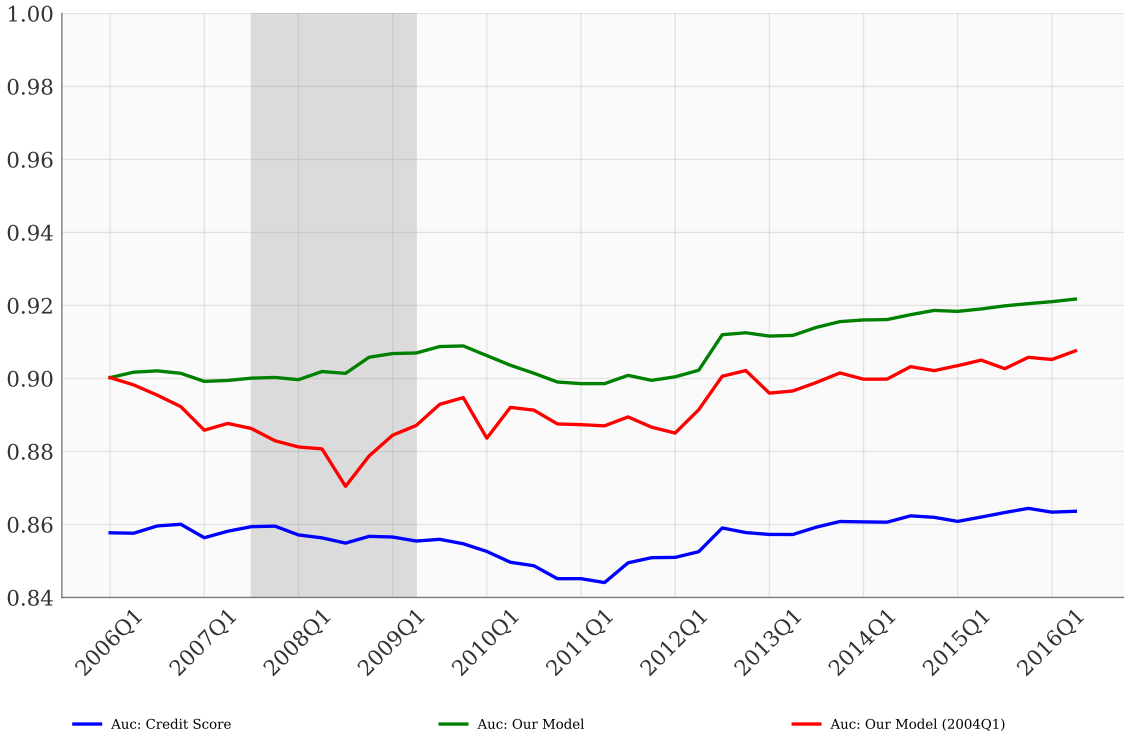


Figure 2: Gini Coefficient, Credit Score and Prediction Model

Notes: AUC scores for the credit score, our model (quarterly retraining), and our model frozen at 2004Q1, by quarter, 2006Q1–2016Q2. Source: Authors' calculations based on Experian data.

A related concern is whether these performance gains would survive the monotonicity constraints that regulators require for real-world deployment (Zoldi (2021))— that is, whether model

predictions must respond in the expected direction to changes in each input (e.g., greater delinquency severity must always increase predicted default risk). Appendix D reports results under two constraint regimes of increasing stringency, the more comprehensive of which constrains 39 of 79 features. Even under the stricter regime, our model's AUC score declines by only 0.005 relative to the unconstrained model and continues to substantially outperform the credit score. Monotonicity constraints thus impose a minimal performance cost and do not affect our conclusions.

4.3 Risk Profile Misclassification

Credit scores place borrowers into five industry-defined risk profiles—Deep Subprime, Subprime, Near Prime, Prime, and Super Prime—that determine credit limits and interest rates. Table 2 reports the credit score thresholds for each profile and the corresponding population shares: 6%, 20.1%, 14.0%, 36.3%, and 23.6% of borrowers, respectively. We construct analogous risk profiles for our model using the same population shares based on predicted default probabilities.

Table 2: Credit Score Risk Profiles

	Deep Subprime	Subprime	Near Prime	Prime	Super Prime
Credit score ranges	300-499	500-600	601-660	661-780	781-850
Credit score percentiles	0-6.0	6.01-27.10	27.11-41.10	41.11-77.40	77.41-100
Percent of borrowers	6.0	20.1	14.0	36.3	23.6

Notes: Industry-defined credit score risk profiles. Definition by credit score range, credit score percentiles. All rates, fractions, and shares in percentage. Total # of observations: 26,240,879. Time period 2006Q1-2016Q2. Source: Authors' calculations based on Experian Data.

Table 3 compares the risk profile assignments. The credit score misclassifies between 26% and 70% of borrowers depending on the risk profile, with misclassification concentrated at the bottom of the distribution. Among Deep Subprime borrowers, 55% would be reclassified into a more favorable category by our model. Among Near Prime borrowers, 70% are misclassified, with 38% placed in Subprime and 28% in Prime by our model. By contrast, misclassification is modest for Super Prime borrowers, with 74% remaining in the same category.

These misclassifications are economically significant. Table 4 shows that borrowers assigned to the same credit score risk profile have sharply different realized default rates depending on our model's classification. For example, Subprime borrowers have an average default rate of 44%, but those whom our model reclassifies as Near Prime default at only 17%, while those our model also

Table 3: Credit Score and Model Risk Profile Comparison

	Model Based Risk Profile					
Credit Score	Deep Subprime	Subprime	Near Prime	Prime	Super Prime	Disagreement
Deep Subprime	45.06	43.57	8.89	2.44	0.03	54.94
Subprime	14.67	52.71	22.46	10.07	0.08	47.29
Near Prime	1.66	38.05	30.35	28.08	1.86	69.65
Prime	0.17	4.43	12.83	66.50	16.07	33.50
Super Prime	0.02	0.05	0.15	25.71	74.08	25.92

Notes: Rows correspond to industry-defined credit score risk profiles, defined in Table 2. Columns are fractions in each corresponding model-based risk profile. Column "Disagreement" reports the fraction of borrowers in each credit score risk profile that would be placed in a different risk profile by our model. All values in percentage. Total # of observations: 26,240,879. Time period 2006Q1-2016Q2. Source: Authors' calculations based on Experian Data.

classifies as Subprime default at 48%. Similarly, among Prime borrowers (average default rate 6%), those our model places in Near Prime default at 13%, while those confirmed as Prime default at 4%. These within-profile differences, often exceeding 10 percentage points, reveal that the credit score fails to distinguish borrowers with substantially different default risk, particularly at the lower end of the distribution.

Table 4 also reports our model's predicted default rates for each credit score/model risk profile combination. Our predicted default rate is always within 5 percentage points of the realized rate, confirming that our model's discriminatory power is high throughout the distribution, as formalized in Proposition 1.

4.4 Feature Attribution

What drives the credit score's misclassification of borrowers? We investigate by comparing how the credit score and our model weight the factors that determine default risk rankings. Credit scoring companies are required to disclose the broad categories of factors driving their scores. For VantageScore 3.0, they cite payment history (35%), amounts owed (30%), length of credit history (15%), credit mix (10%), and new credit (10%) as the main drivers (VantageScore Solutions, LLC (2013)).[10] We group our model's features into these same categories and aggregate absolute SHAP values across each category.[11]

The results, reported in Table 5, reveal a striking pattern consistent with the misweighting

[10]Credit scoring companies do not disclose which specific features enter each factor category.

[11]For details on SHAP values, see Lundberg and Lee (2017). We aggregate over all instances in the testing dataset across all periods.

Table 4: Default Risk Variation by Credit Score Profile

Risk Profile	Default Rate		Risk Profile	Default Rate	
Credit Score	Realized	Predicted	Model Based	Realized	Predicted
Deep Subprime	68.19	66.88	Deep Subprime	94.75	93.86
			Subprime	54.40	52.48
			Near Prime	18.18	17.75
			Prime	6.94	5.99
			Super Prime	5.82	0.94
Subprime	43.68	43.10	Deep Subprime	92.48	92.46
			Subprime	48.25	47.59
			Near Prime	17.09	16.61
			Prime	8.27	7.20
			Super Prime	4.00	1.00
Near Prime	22.38	22.56	Deep Subprime	88.18	90.85
			Subprime	34.89	37.76
			Near Prime	18.31	16.46
			Prime	7.36	5.97
			Super Prime	1.17	1.00
Prime	6.05	6.62	Deep Subprime	90.87	91.53
			Subprime	25.88	32.72
			Near Prime	13.17	15.17
			Prime	4.38	4.40
			Super Prime	0.92	0.91
Super Prime	1.02	1.24	Deep Subprime	92.43	91.48
			Subprime	42.23	44.90
			Near Prime	10.15	14.44
			Prime	2.04	2.39
			Super Prime	0.61	0.76

Notes: Realized and model predicted default rates by credit score and model-based risk profile. Industry-defined credit score risk profiles, defined in Table 2. All values in percentage. Total # of observations: 26,240,879. Time period 2006Q1-2016Q2. Source: Authors' calculations based on Experian Data.

mechanism identified in Section 2. Amounts owed explains 50% of the variation in our model's rankings, compared to only 30% for the credit score. Conversely, our model gives substantially less weight to length of credit history (9% vs. 15%) and new credit (3% vs. 10%). These factors—credit history length and new credit—are weakly predictive of default but strongly correlated with age and, through homeownership patterns and credit market experience, with income and race. The credit score's overweighting of these factors is the mechanism through which it disproportionately misclassifies young, low-income, and minority borrowers, as we document in Section 5.

Table 5: Feature Attribution Differences

Feature Group	# of Features	Our Model	Credit Score
Payment History	25	0.35	0.40
Age and type of credit	10	0.12	0.21
Credit utilization	5	0.12	0.20
Balances	29	0.21	0.11
Recent credit	4	0.05	0.05
Available credit	6	0.15	0.03

Notes: SHAP values for five feature groups. For each prediction window, we compute the SHAP value for 100,000 randomly sampled observations and for each feature. We then calculate the sum of the absolute value for each feature, aggregate it across the feature groups, and report the results for the group. We normalized the results so that for each model the five groups sum up to 1. Feature groups are defined in Table 12 in Appendix B. Credit score variation information from `https://www.vantagescore.com/wp-content/uploads/2022/01/VantageScore3-0_WhitePaper.pdf`. Source: Authors' calculations based on Experian data.

4.5 Credit Score Prevalence

A natural question raised by our findings is why the lending industry has relied on credit scores for decades despite their inferior statistical performance. We identify four non-mutually exclusive explanations.

A natural question raised by our findings is why the lending industry has relied on credit scores for decades despite their inferior statistical performance. We identify four non-mutually exclusive explanations. First, overweighting utilization may serve a moral hazard deterrence function: a consumer who increases borrowing experiences an immediate score drop, raising the cost of future credit and potentially deterring excessive debt accumulation, consistent with Padilla and Pagano (2000).[12] Second, overweighting credit mix and account age generates consumer lock-in, as borrow-

[12]Kiviat (2019) documents an episode in which lenders cut credit limits during the financial crisis, mechanically raising utilization ratios and lowering scores for consumers whose behavior was unchanged—illustrating the emphasis on borrower responsibility embedded in the scoring design.

ers keep unused accounts open to avoid score reductions from shorter credit histories. Third, the proprietary nature of scoring models creates information asymmetry between score developers and lenders: while large banks have developed internal models, smaller lenders lack the analytical capacity to independently evaluate score performance Consumer Financial Protection Bureau (2012). Fourth, FICO held a near-monopoly in credit scoring throughout our sample period, insulated from competitive pressure by GSE requirements. The 2018 Credit Score Competition Act mandated that the FHFA establish a process for validating alternative models, but VantageScore was not approved for GSE lending until July 2025—seven years after the legislative mandate—illustrating the depth of institutional lock-in in this market.

4.6 Costs of Risk Profile Misclassification

We have shown that the credit score's lower statistical performance leads to risk profile misclassification for a large fraction of consumers, concentrated primarily at the low end of the credit score distribution. What are the costs of such misclassification? We provide an initial answer by analyzing how access to credit varies as a function of credit scores for credit cards and mortgages, the most common types of consumer loans.[13]

Table 6 reports age-adjusted estimated coefficients for average credit risk profile effects on credit access. Access to credit varies steeply across risk profiles. The fraction of borrowers with a credit card rises from 25% for Deep Subprime to 96% for Superprime, while mortgage holding rises from 11% to 38%. Among those with credit, limits and balances diverge sharply: average credit card limits range from roughly $6,000 for Deep Subprime borrowers to nearly $40,000 for Superprime borrowers, and conditional mortgage balances rise from $26,000 to $61,000. Success rates at obtaining credit also increase with the credit score: only 6% of Deep Subprime borrowers who inquire about a credit card obtain one, compared to 35% of Superprime borrowers. Figure 10 in Appendix E displays the time variation in these effects, showing that the gap in access to credit between borrowers at the bottom and the top of the credit score distribution widened substantially from 2009 to 2016 in the aftermath of the Great Recession.

Combining these access gradients with the misclassification rates in Table 3 gives a sense of the magnitudes involved. Recall that 22% of Subprime borrowers would be reclassified as Near Prime

[13]Since we use data from credit reports, we do not have access to interest rate information.

Table 6: Access to Credit by Risk Profile

	(1) Fraction with	(2) Credit Limit / Balances	(3) Inquiries	(4) Originations if Inquired
Panel A: Credit Cards				
Superprime	96.13	39.43	5.90	35.15
Prime	79.58	26.30	8.04	33.88
Near Prime	66.88	19.76	12.37	29.87
Subprime	38.15	7.73	11.40	16.70
Deep Subprime	25.00	5.99	10.93	5.81
Panel B: Mortgages				
Superprime	37.86	61.18	4.58	12.24
Prime	32.93	66.42	5.99	9.08
Near Prime	25.92	50.48	6.83	6.10
Subprime	13.64	28.33	4.82	3.06
Deep Subprime	11.14	25.55	4.15	1.77

Notes: This table reports age-adjusted estimated coefficients for the average credit risk profile effects for credit cards (Panel A) and mortgages (Panel B). Deep Subprime, Subprime, Near Prime, Prime and Superprime are indicator variables for the consumer belonging to that category at quarter $t-1$ while outcomes are measured at quarter t. Column (2) is credit limits for credit cards and first mortgage balances for mortgages measured in thousands of USD. All other values in percentage. Column (4) restricts the sample to individuals who inquired in quarters $t-1, t$ and measures originations at quarter t. Full specification and regression results are reported in Appendix E.

by our model, and an additional 8% as Prime. Moving from the Subprime to the Near Prime risk profile is associated with roughly $12,000 in additional credit card limits and a near-doubling of the mortgage origination rate. At the other end, 26% of Superprime borrowers would be reclassified as Prime, a group whose default probabilities are consistent with a less favorable risk rating based on our model. These borrowers enjoyed higher credit limits and more favorable origination terms than their underlying risk warranted, potentially exposing lenders to unexpected losses. Additionally, the fraction with inquiries declines markedly with the credit score for borrowers in the Near Prime-Deep Subprime group. This pattern is consistent with our model's prediction that lower-rated borrowers are systematically overcharged relative to their true default risk: facing credit offers with interest rates and terms calibrated to an overstated risk assessment, these borrowers rationally reduce their demand for credit, as reflected in fewer inquiries. While we do not observe interest rates in our data, the credit access gradients we document almost certainly understate total costs, since industry rate schedules imply that differences across adjacent risk profiles correspond to several percentage points in annual percentage rates on credit card balances and 50 to 100 basis points on mortgage

rates.[14]

These credit market costs represent a lower bound on the total welfare costs of misclassification, because credit scores have diffused well beyond lending. Credit-based insurance scores are used to price automobile and homeowners insurance in most U.S. states (Federal Trade Commission (2007), Kiviat (2019)); Blonz, Hossain, and Weill (2024) show that subprime borrowers pay approximately 30% more for home insurance than super-prime borrowers, and Blonz et al. (2026) find that the effect of credit on insurance costs is comparable in magnitude to its effect on mortgage interest rates. In the labor market, roughly one-quarter of employers use credit reports in hiring decisions (Ballance, Clifford, and Shoag (2020)), despite evidence that adverse credit flags have no predictive power for labor market outcomes conditional on basic observables (Dobbie et al. (2020)). Credit scores also affect access to rental housing and determine whether consumers must post security deposits for utilities. Our finding that credit scores misclassify 26% to 70% of borrowers implies that the informational basis for non-lending screening is weaker than commonly assumed (Corbae and Glover (2018)): a misclassified consumer may simultaneously face restricted credit, higher insurance premiums, diminished employment prospects, and limited housing options. The analysis in this section quantifies the first of these channels; the full cost of misclassification extends considerably further.

5 Performance and Equity

A central concern in the literature on algorithmic scoring is the potential for more sophisticated models to widen demographic disparities. Fuster et al. (2018) find that a random forest model of mortgage default improves predictive accuracy for all racial groups but reduces the relative standing of Black and Hispanic borrowers compared to a logistic regression benchmark. Blattner and Nelson (2021) argue that machine learning models cannot improve performance for marginalized borrowers because the main source of worse performance for these groups is compositional differences in their attribute distributions—what they term "data bias." Our theoretical framework in Section 2 prediction: when the source of poor performance is misweighting of credit profile characteristics rather than model flexibility per se, correcting the weights improves performance most for the

[14]See, for example, Federal Reserve Statistical Release G.19 for average credit card interest rates by account type and the *Freddie Mac Primary Mortgage Market Survey* for rate variation by creditworthiness.

groups penalized by the misweighting (Proposition 3). We now test this prediction.

5.1 Data on Demographic Groups

We consider three dimensions of marginalization: income, age, and minority status. Our credit report data include date of birth and estimated annual income, from which we construct indicators for being under 30 (Young) and for being in the bottom quintile of the income distribution (Income_{p20}). Race is not recorded in credit reports. We address this limitation with two complementary approaches.

First, for the full Experian sample, we construct a minority proxy based on the share of Black and Hispanic residents at the ZIP code level from the 2010 Census, following the methodology developed by the Consumer Financial Protection Bureau.[15] Second, we match our Experian data to the Home Mortgage Disclosure Act (HMDA) database, which records race directly for mortgage borrowers and applicants. The matched sample is restricted to consumers with a first mortgage and is therefore positively selected: these borrowers have lower default rates, higher credit scores, and higher incomes than the full sample (Table 19). Importantly, having a mortgage mechanically boosts credit mix—a factor the credit score overweights—making this a conservative sample for detecting gains from our model. The matching procedure is described in the Online Appendix.

5.2 Variation in Standing

To assess whether our model's superior performance translates into gains for marginalized borrowers, we regress the difference in percentile rankings between our model and the credit score on demographic group indicators, an indicator for future default, and their interaction:

$$Y_{ist} = \alpha + \beta_1 x_{ist} + \beta_2 z_{ist} + \beta_3 z_{ist} \times x_{ist} + \lambda_t + \delta_s + \gamma_{st} + \epsilon_{ist} \tag{13}$$

where Y_{ist} is the ranking difference (positive values indicate our model ranks borrower i more favorably), x_{ist} is a marginalized group indicator, z_{ist} is a future default indicator (90+ days past due within 8 quarters), and we include state (δ_s), quarter (λ_t), and state×quarter (γ_{st}) fixed

[15] See `https://www.consumerfinance.gov/data-research/research-reports/using-publicly-available-information-to-proxy` The indicator equals one if the ZIP code Black and Hispanic population share exceeds 50%.

effects.[16] The interaction β_3 is the key coefficient: it captures whether our model's ranking gains for marginalized borrowers come from better separation of defaulters and non-defaulters within each group. The full regression results are reported in Table 18 in Appendix F.

Figure 3 summarizes the main results. Three findings emerge. First, our model systematically ranks future defaulters lower than the credit score does, by approximately 5.3 percentiles (29 credit score points), confirming its superior discriminatory power.[17] Second, each marginalized group gains in standing on average: low-income borrowers gain approximately 5 percentiles (25 credit score points), young borrowers gain 2 percentiles (11 points), and minority borrowers gain 1 percentile (5 points) in the Experian sample. Third, and most importantly, these gains are concentrated among borrowers who do not default. Low-income non-defaulters gain 6 percentiles (33 points), while low-income defaulters lose 7 percentiles (39 points). The pattern is similar for young and minority borrowers: non-defaulters gain, defaulters lose. This is precisely the prediction of Proposition 3: correcting the misweighting of credit profile characteristics raises the standing of marginalized borrowers with good repayment fundamentals while lowering the standing of those with poor fundamentals.

Results are similar in the HMDA Match sample, with one notable exception. Low-income and young borrowers gain in ranking, with gains again concentrated among non-defaulters. However, minority borrowers in the HMDA sample lose approximately 1 percentile on average, including non-defaulters. This attenuation is consistent with the positive selection inherent in the HMDA sample. Consumers with mortgages benefit disproportionately from the credit score's overweighting of credit mix, and this benefit is likely strongest for minority borrowers given the historical barriers to mortgage access for these populations. In a sample defined by mortgage holding, the credit score's misweighting of credit mix works in minorities' favor, attenuating the gains our model delivers in the broader population.

5.3 Variation in Performance

We now examine why our model delivers larger performance gains for marginalized groups. Table 7 reports AUC scores by demographic group. For the overall sample, our model's AUC exceeds

[16] State-level controls account for variation in regulations affecting the consequences of default (Albanesi and Nosal (2025)).

[17] One percentile in ranking is equivalent to approximately 5.5 credit score points.

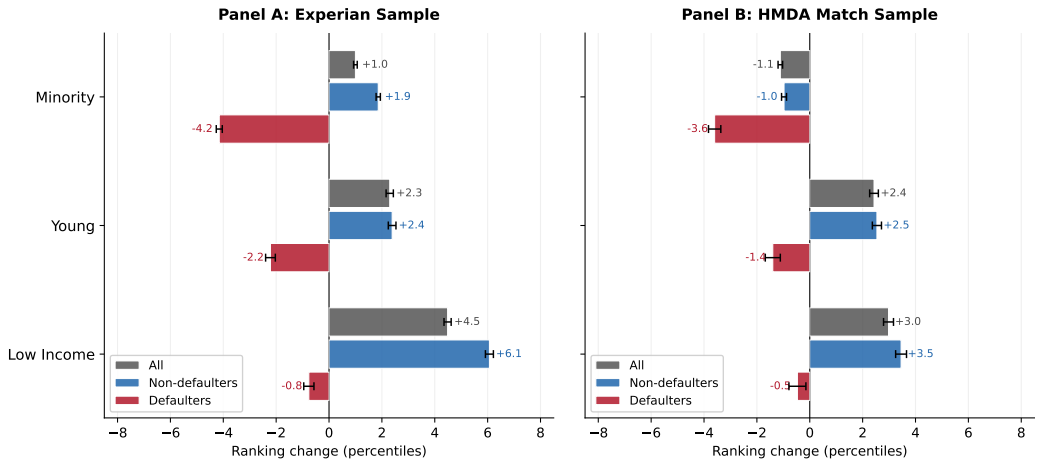


Figure 3: Ranking Gains by Demographic Group and Default Status

Notes: Estimated ranking difference (our model minus credit score, in percentiles) by demographic group and future default status. "All" reports the group coefficient from the specification without interactions (columns 3, 5, 7 of Table 18). "Non-defaulters" and "Defaulters" report the implied ranking change from the interaction specification (columns 4, 6, 8). Panel A: Experian sample with ZIP code-based minority proxy. Panel B: HMDA Match sample with observed race. Error bars show 95% confidence intervals. One percentile ≈ 5.5 credit score points. Source: Authors' calculations based on Experian and HMDA data.

the credit score's by 5 percentage points (0.91 vs. 0.86). The improvement is substantially larger for marginalized groups: 12 points for low-income borrowers (0.86 vs. 0.74), 6 points for young borrowers (0.88 vs. 0.82), and 8 points for minority borrowers (0.88 vs. 0.80). The gap between marginalized and non-marginalized groups in AUC scores is much smaller for our model than for the credit score, indicating that our model's performance is more uniform across the population.

Table 7: Performance by Demographic group

AUC score	All	Income		Age		Minority	
		$Income_{p20}$	$Income_{\geq p20}$	Age < 30	Age ≥ 30	Yes	No
Credit Score	0.856	0.735	0.867	0.824	0.865	0.802	0.864
Our model	0.906	0.855	0.908	0.881	0.913	0.881	0.91

Notes: AUC score for our model ad the credit score for each demographic group, averages for 2006Q1-2016Q2. Source: Authors' calculations based on Experian data.

Table 8 reports SHAP-based feature attribution statistics by demographic group, showing virtually no differences across groups. This implies that the conditional expectation function for default is similar across groups—differences in predictions arise from the distribution of attributes, not from group-specific patterns in the mapping from attributes to default risk, consistent with Blattner and Nelson (2021).

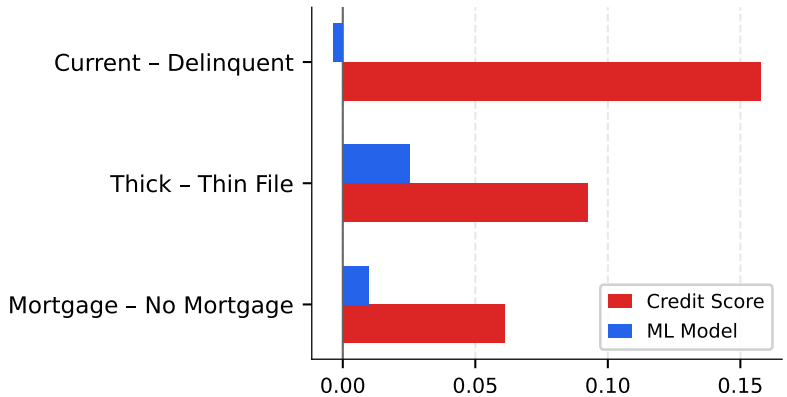
Table 8: Feature Attribution by Demographic group

Feature Group	Income		Age		Minority	
	$Income_{p20}$	$Income_{\geq p20}$	Age < 30	Age ≥ 30	Yes	No
Payment History	0.352	0.349	0.350	0.352	0.351	0.352
Age and type of credit	0.118	0.119	0.119	0.118	0.119	0.119
Credit utilization	0.115	0.116	0.115	0.115	0.115	0.115
Balances	0.212	0.214	0.213	0.212	0.213	0.213
Recent credit	0.053	0.054	0.054	0.053	0.053	0.053
Available credit	0.149	0.149	0.149	0.149	0.149	0.149

Notes: SHAP values for five feature groups. For each prediction window, we compute the SHAP value for 100,000 randomly sampled observations and for each feature. We then calculate the sum of the absolute value for each feature, aggregate it across the feature groups, and report the results for the group. We normalized the results so that the five groups sum up to 1. Feature groups are defined in Table 12 in Appendix B. Time period 2006Q1-2016Q2. Source: Authors' calculations based on Experian data.

What, then, explains the differential performance gains? Following Blattner and Nelson (2021), we decompose performance by attribute composition. We calculate AUC scores separately for borrowers who are current versus delinquent, those with thick versus thin files, and those with versus without a mortgage.[18] Figure 4 displays the performance gap across these attribute categories. For the credit score, performance is dramatically better for borrowers with favorable attributes: the AUC gap between current and delinquent borrowers is approximately 16 percentage points. For our model, these differences are minimal. The credit score's performance degrades sharply with low-quality data; our model's does not.

Figure 4: AUC Score Gaps by Feature Composition



Notes: Difference in AUC scores between favorable and unfavorable attribute categories (Current–Delinquent, Thick file–Thin file, Mortgage–No mortgage) for the credit score and our model. A positive value indicates better performance for the favorable category. Averages 2006Q1–2016Q2. Source: Authors' calculations based on Experian data.

[18]Thin file consumers are those with a credit history of less than 10 years or fewer than 3 types of products on the credit report, among credit cards, bank cards, auto loans, HELOCs, first mortgages, and installment loans.

Table 9 connects this finding to the demographic results. Low-income, young, and minority consumers are substantially overrepresented in the unfavorable attribute categories—delinquent, thin file, and no mortgage—compared to their non-marginalized counterparts. Since the credit score's performance deteriorates precisely in these categories while our model's does not, the demographic composition of unfavorable attribute groups explains why marginalized borrowers benefit disproportionately from our model.

Table 9: Feature Composition by Demographic Group

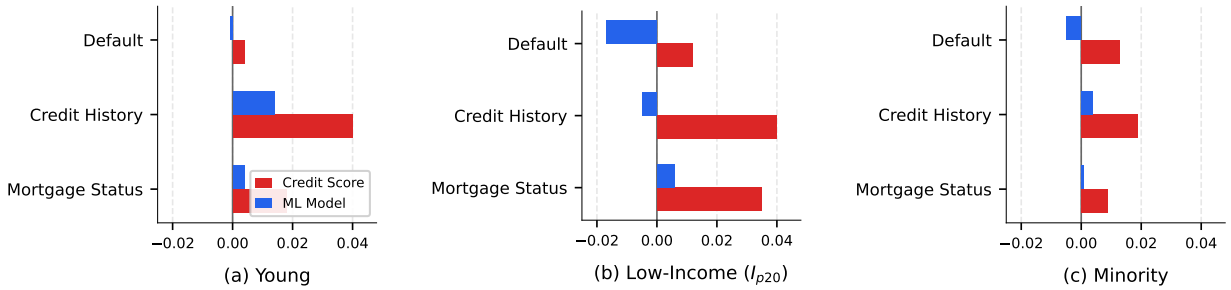
		Demographic Group					
		Income		Age		Minority	
Feature Category	Status	$Income_{p20}$	$Income_{\geq p20}$	Age < 30	Age ≥ 30	Yes	No
Default History	Current	0.64	0.87	0.80	0.83	0.73	0.84
	Delinquent	0.36	0.13	0.20	0.17	0.27	0.16
Credit History	Thin file	0.93	0.39	0.89	0.39	0.64	0.47
	Thick file	0.07	0.61	0.11	0.61	0.36	0.53
Mortgage Status	No Mortgage	0.95	0.61	0.91	0.62	0.78	0.66
	Mortgage	0.05	0.39	0.09	0.38	0.22	0.34

Notes: Share of borrowers in each feature composition category by demographic group. Feature composition categories defined by Default History, Credit History, Mortgage Status and Credit History interacted with Mortgage Status, with shares within each category adding up to 1. Time period 2006Q1-2016Q2. Source: Authors' calculations based on Experian data.

We confirm this mechanism with a counterfactual analysis. For each marginalized group, we compute the AUC score that would obtain if the group had the same attribute composition as the non-marginalized population. Figure 5 reports the results. The gap between actual and counterfactual AUC scores is an order of magnitude smaller for our model than for the credit score, indicating that attribute composition is a major determinant of the credit score's performance shortfall for marginalized groups but plays a negligible role for our model. The variation in AUC scores across attribute composition categories by demographic group, reported in Table 20 in Appendix F, confirms that the performance gap between favorable and unfavorable attribute categories is similar across demographic groups for our model and substantially smaller than for the credit score.

These findings establish that a machine learning model with superior statistical performance can deliver greater performance improvements for groups traditionally marginalized in consumer credit markets. This result is not inconsistent with Blattner and Nelson (2021): attribute composition differences (data bias) do reduce credit score performance for marginalized groups. But whereas they conclude that sophisticated algorithms cannot overcome this problem, our results show that

Figure 5: Contribution of Feature Composition to Model Performance



Gap between counterfactual and actual AUC score for marginalized consumers by group for our model and the credit score by feature component. Counterfactual AUC score applies feature distribution of non-marginalized group to marginalized group. Time period 2006Q1–2016Q2. Source: Authors' calculations based on Experian data.

the relevant question is whether the model is robust to unfavorable attribute compositions. Our model's ability to maintain performance with low-quality data—thin files, delinquency histories, limited credit mix—is what drives the differential gains for marginalized groups. Our findings also complement Fuster et al. (2018): in their setting, increased model flexibility captures group-specific nonlinearities in repayment fundamentals, which can widen disparities. In our setting, the improvement comes from correcting the weighting of known factors, which benefits marginalized groups precisely because they are penalized by the current misweighting. As suggested by Blattner and Nelson (2021), incorporating features more informative for young, low-income, or minority borrowers could potentially improve performance for these groups even further.

6 Conclusion

Credit scores are the primary tool for allocating consumer credit in the United States, yet their performance has never been independently evaluated against a well-defined benchmark. Using a nationally representative panel of credit reports spanning 2004–2016, we document that the credit score misclassifies 41% of consumers, with misclassification concentrated at the bottom of the score distribution. The source is a misweighting of credit profile characteristics—credit history length and credit mix—that are weakly predictive of default but systematically lower for young, low-income, and minority borrowers. Correcting this misweighting improves both performance and equity through a single channel: a model that appropriately weights repayment fundamentals is more accurate for all borrowers, but the gains are largest for groups penalized by the current

weighting scheme. Low-income borrowers see a 12-percentage-point AUC improvement (compared to 5 overall), and low-income non-defaulters gain over 6 percentiles in ranking. These gains are driven by our model's robustness to low-quality data—thin files, delinquency histories, limited credit mix—precisely the attribute categories where marginalized groups are overrepresented and where the credit score's performance breaks down.

Our analysis has limitations that suggest directions for future research. We benchmark against VantageScore 3.0, which was not widely used in lending decisions during our sample period; while this mitigates acted-upon score concerns, our results speak to the structure of conventional scoring models rather than to any single score in use today. Our model uses only credit report information—incorporating alternative data sources could further improve performance for thin-file borrowers (Blattner and Nelson (2021)). Finally, our welfare analysis is partial equilibrium: we do not model general equilibrium effects of a change in the scoring regime on lending standards, interest rates, or borrower behavior.

Despite these caveats, our results carry a clear policy implication. The current structure of widely used credit scores generates predictable patterns of misclassification that fall disproportionately on populations already marginalized in consumer credit markets. Improving the performance of credit scoring models is not merely a technical exercise: it is a direct lever for promoting more equitable access to credit.

References

Albanesi, Stefania, Giacomo DeGiorgi, and Jaromir Nosal. 2022. "Credit growth and the financial crisis: A new narrative." *Journal of Monetary Economics* 132:118–139.

Albanesi, Stefania, and Jaromir Nosal. 2025. "Consumer Default After the 2005 Bankruptcy Reform." *Journal of Law, Economics and Organization*.

Albanesi, Stefania, and Domonkos F Vamossy. 2019, August. "Predicting Consumer Default: A Deep Learning Approach." Working paper 26165, National Bureau of Economic Research.

Arnold, David, Will Dobbie, and Peter Hull. 2022. "Building Nondiscriminatory Algorithms in Selected Data." *American Economic Review: Insights* 4 (2): 231–249.

Avenancio-León, Carlos, and Troup Howard. 2022. "The Assessment Gap: Racial Inequality in Property Taxation." *The Quarterly Journal of Economics* 137 (3): 1209–1270.

Bach, Helena, Pietro Campa, Giacomo De Giorgi, Jaromir Nosal, and Davide Pietrobon. 2023. "Born to be (sub) prime: An exploratory analysis." *AEA Papers and Proceedings*, Volume 113. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 166–171.

Ballance, Joshua, Robert Clifford, and Daniel Shoag. 2020. ""No More Credit Score": Employer Credit Check Bans and Signal Substitution." *Labour Economics* 63:101769.

Blattner, Laura, and Scott Nelson. 2021. "How costly is noise? Data and disparities in consumer credit." *arXiv preprint arXiv:2105.07554*.

Blonz, Joshua, Mallick Hossain, Benjamin J Keys, Philip Mulder, and Joakim A Weill. 2026, February. "Pricing Protection: Credit Scores, Disaster Risk, and Home Insurance Affordability." Working paper 34848, National Bureau of Economic Research.

Blonz, Joshua A., Mallick Hossain, and Joakim Weill. 2024, November. "Pricing Protection: Credit Scores, Disaster Risk, and Home Insurance Affordability." SSRN Working Paper 5026600.

Brevoort, Kenneth P., Philipp Grimm, and Michelle Kambara. 2016. "Credit Invisibles and the Unscored." *Cityscape* 18 (2): 9–34.

Chatterjee, Satyajit, Dean Corbae, Kyle Dempsey, and José-Víctor Ríos-Rull. 2023. "A Quantitative Theory of the Credit Score." *Econometrica* 91 (5): 1803–1840.

Chatterjee, Satyajit, Dean Corbae, Kyle P Dempsey, and José-Víctor Ríos-Rull. 2025, July. "Credit Scores and Inequality across the Life Cycle." Working paper 34027, National Bureau of Economic Research.

Consumer Financial Protection Bureau. 2012, December. "Key Dimensions and Processes in the U.S. Credit Reporting System." Technical Report, Consumer Financial Protection Bureau, Washington, DC.

Corbae, Dean, and Andrew Glover. 2018. "Employer Credit Checks: Poverty Traps versus Matching Efficiency." Working paper 25005, National Bureau of Economic Research.

Di Maggio, Marco, Dimuthu Ratnadiwakara, and Don Carmichael. 2022, March. "Invisible Primes: Fintech Lending with Alternative Data." Working paper 29840, National Bureau of Economic Research.

Dobbie, Will, Paul Goldsmith-Pinkham, Neale Mahoney, and Jae Song. 2020. "Bad Credit, No Problem? Credit and Labor Market Consequences of Bad Credit Reports." *The Journal of Finance* 75 (5): 2377–2419.

Edelberg, Wendy. 2006. "Risk-based pricing of interest rates for consumer loans." *Journal of monetary Economics* 53 (8): 2283–2298.

Federal Trade Commission. 2007. "Credit-Based Insurance Scores: Impacts on Consumers of Automobile Insurance." Report to congress, Federal Trade Commission, Washington, DC.

FRB. 2007. "Report to the congress on credit scoring and its effects on the availability and affordability of credit." *Board of Governors of the Federal Reserve System*.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2018. "Predictably unequal? the effects of machine learning on credit markets." *The Effects of Machine Learning on Credit Markets (November 6, 2018)*.

———. 2022. "Predictably unequal? The effects of machine learning on credit markets." *The Journal of Finance* 77 (1): 5–47.

Gupta, Akhil, Naman Shukla, LLC Deepair, Lavanya Marla, Arinbjörn Kolbeinsson, and Kartik Yellepeddi. 2020. "How to Incorporate Monotonicity in Deep Networks While Preserving Flexibility?"

Immorlica, Nicole, Andre Sztutman, and Robert M. Townsend. 2025, November. "Optimal Credit Scores Under Adverse Selection." Working paper.

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

Kiviat, Barbara. 2019. "The Moral Limits of Predictive Practices: The Case of Credit-Based Insurance Scores." *American Sociological Review* 84 (6): 1134–1158.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133 (1): 237–293.

Kline, Patrick M., Evan K. Rose, and Christopher R. Walters. 2024. "A Discrimination Report Card." *American Economic Review* 114 (8): 2472–2525.

Kvamme, Håvard, Nikolai Sellereite, Kjersti Aas, and Steffen Sjursen. 2018. "Predicting mortgage default using convolutional neural networks." *Expert Systems with Applications* 102:207–217.

Livshits, Igor, James C Mac Gee, and Michele Tertilt. 2016. "The democratization of credit and the rise in consumer bankruptcies." *The Review of Economic Studies* 83 (4): 1673–1710.

Lundberg, Scott M, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4765–4774. Curran Associates, Inc.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–453.

Padilla, A. Jorge, and Marco Pagano. 2000. "Sharing Default Information as a Borrower Discipline Device." *European Economic Review* 44 (10): 1951–1980.

Sirignano, Justin, Apaar Sadhwani, and Kay Giesecke. 2018. "Deep Learning for Mortgage Risk." *Available at SSRN*.

VantageScore Solutions, LLC. 2013, December. "VantageScore 3.0 White Paper." Technical Report, VantageScore Solutions, LLC.

You, Seungil, David Ding, Kevin Canini, Jan Pfeifer, and Maya Gupta. 2017. "Deep lattice networks and partial monotonic functions." *Advances in neural information processing systems.* 2981–2989.

Zoldi, Scott. 2021, February. Trusted AI: The Challenge of Monotonicity and Palatability. FICO Blog. Accessed February 2026.

A Experian Data

The data contain 204 selected Experian credit bureau attributes observed at quarterly frequency. These span several broad categories: dollar-denominated balance and payment variables (credit limits, outstanding balances, amounts past due, and monthly payments), trade counts (number of open, new, delinquent, or charged-off accounts), temporal measures (months since last delinquency, last bankruptcy filing, or oldest account opened), delinquency and derogatory status indicators (worst present or historical status ranked on an ordinal scale, occurrence counts of 30+, 60+, 90+, or 180+ day delinquencies, derogatory, or charged off over various trailing windows), public record fields (counts of Chapter 7 and Chapter 13 bankruptcies filed, discharged, or dismissed), inquiry counts (by trade type and time window), and miscellaneous flags (deceased indicator, government debt presence). The attributes are reported separately by trade type — including bankcard revolving and charge accounts, installment loans, auto loans, first and second mortgages, home equity lines, student loans, business credit, revolving trades, utilities, and collections — as well as at the all-trades aggregate level, with further breakdowns for joint accounts and authorized user trades. Most variables are measured over trailing 6-, 12-, 24-, or 36-month windows.

The data include two sets of attributes that account for jointly held credit obligations. The first set consists of variables which report balances, credit limits, amounts past due, and monthly payments exclusively on joint accounts, disaggregated by trade type. The second set consists of weighted variables , which adjust aggregate balance and payment measures by weighting joint account amounts — typically at 50% — to approximate individual liability rather than reporting the full joint balance. Both sets exclude derogatory trades and authorized user accounts. In our analysis, we use the weighted variants when computing individual-level debt and payment burdens to avoid overstating exposure on shared obligations for mortgage accounts where joint borrowing is prevalent.

The data also include each consumer's VantageScore 3.0 credit score, observed quarterly, as well as their zip code of residence and Experian-provided estimates of individual W-2 income and household income derived from IRS records. The data do not contain information on gender, race, marital status, or other protected demographic categories.

A.1 Sampling Design

The Experian data are constructed in three sequential samples that together form a longitudinal consumer credit panel spanning 2004Q1 through 2018Q2. The samples differ in calendar coverage and acquisition timing but are integrated using a stable individual identifier. Across all draws, the final dataset contains 1,000,000 unique consumers.

Sample 1 (2004Q1–2012Q4). Sample 1 constitutes the baseline panel and covers quarterly archives from 2004Q1 through 2012Q4. Consumers are drawn from Experian's national credit bureau files using a stratified design. The majority of sampled consumers are existing credit users, yielding approximately 800,000 individuals observed in this baseline period. The resulting panel is unbalanced due to entry and exit from the credit bureau over time.

Sample 2 (2013Q1–2015Q4). Sample 2 augments the baseline by adding consumers who newly enter the credit bureau after the end of Sample 1. This sample spans 2013Q1 through 2015Q4 and contributes approximately 200,000 additional consumers. New entrants are identified using archive-specific thresholds for the age of the oldest trade, ensuring that entry into the sample closely mirrors

entry into the underlying credit bureau population. As a result, Samples 1 and 2 together are both cross-sectionally and dynamically representative of the credit-active population at the time of entry.

Sample 3 (2016Q1–2018Q2). Sample 3 completes the dataset by adding quarterly records up to 2018Q2 for consumers already selected in Samples 1 and 2. No new consumers are drawn at this stage. Instead, Sample 3 retroactively completes historical archive coverage when records are available. Consequently, Sample 3 remains cross-sectionally representative of consumers who had entered the credit bureau by the end of 2015, but it is not dynamically representative of post-2015 entry cohorts.

A.2 Sample Restrictions

We discard observations of individuals with missing birth information, and deceased individuals and restrict our analysis to individuals aged between 18 and 85, residing in one of the 50 states or the District of Columbia, with 8 consecutive quarters of non-missing default behavior. This leaves us with 25,386,523 data points. Our itemized sample restrictions are summarized in Table 10 below.

Table 10: Itemized Sample Restrictions

	Observations
Credit Report Data	40,600,000
Remove	
Deceased	- 756,149
Age	- 3,051,161
Residence	- 15,366
Credit Score	- 155,149
2004Q1-2005Q4 & 2016Q3-2018Q2	-9,795,920
Prediction Window	- 585,376
Prediction Sample	26,240,879

Table 11 reports basic demographic information on our sample, including age, household income, credit score, and incidence of default, defined as the fraction of households who report 90 or more days past due delinquency on any trade, excluding collections.

B Model Details

B.1 Model

We use a Hybrid Deep Neural Network-Gradient Boosted Trees model for our analysis. This is an ensemble model, comprised of two components. The first is based on deep learning, in the class used by Sirignano, Sadhwani, and Giesecke (2018). We restrict attention to feed-forward neural networks, composed of an input layer, which corresponds to the data, one or more interacting hidden layers that non-linearly transform the data, and an output layer that aggregates the hidden layers into a prediction. Layers of the networks consist of neurons, with each layer connected by

Table 11: Descriptive Statistics

Feature	Mean	Std. Dev.	Min	25%	50%	75%	Max
Age	45.530	16.751	18.0	31.0	45.0	58.0	84.0
Household Income (Imputed)	78.331	55.092	15.0	43.0	64.0	91.0	316.0
Credit Score	679.317	108.475	300.0	595.0	691.0	776.0	839.0
Credit History (Months)	197.672	129.669	0.0	95.0	175.0	274.0	988.0
$\mathbb{I}_{\{90+ \text{ DPD Debt}\}}$	0.078	0.269	0.0	0.0	0.0	0.0	1.0
$\mathbb{I}_{\{90+ \text{ DPD Debt, next 8 Quarters}\}}$	0.184	0.387	0.0	0.0	0.0	0.0	1.0

Notes: Household income is in USD thousands, winsorized at the 99th percentile. Source: Authors' calculations based on Experian Data.

synapses that transmit signals among neurons of subsequent layers. A neural network is, in essence, a sequence of nonlinear relationships. Each layer in the network takes the output from the previous layer and applies a linear transformation followed by an element-wise nonlinear transformation.

The second component of our model is Extreme Gradient Boosting, which builds on decision tree models. Tree-based models split the data several times based on certain cutoff values in the explanatory variables.[19] Gradient Boosted Trees (GBT) are an ensemble learning method designed to correct the tendency of tree-based models to overfit training data. This is achieved by recursively combining the predictions of multiple, simpler trees. Although each individual shallow tree is a "weak learner" with limited predictive power, the ensemble of these weak learners forms a strong model with improved stability over a single complex tree.

These two components — Deep Neural Network (DNN) and Gradient Boosted Trees (GBT) — are combined to improve predictive performance. Each model is trained independently, and their final predicted probabilities are then averaged. This approach, akin to the method proposed by Kvamme et al. (2018), which combined a convolutional neural network with a random forest by averaging, ensures that our methodology benefits from the strengths of both models. Specifically, we achieve this by following two key steps:

1. For each observation, run DNN and GBT separately and obtain predicted probabilities for each of the models;

2. Compute a weighted average of the predicted probabilities, with the weights determined based on model performance, as explored in Albanesi and Vamossy (2019).

B.2 Model Inputs

Table 12 lists the features from the model we use as inputs from the credit report data. They include information on balances and credit limits for different types of consumer debt, severity and number of delinquencies, credit utilization by type of product, public record items such as bankruptcy filings, collection items, and length of the credit history. In order to be consistent with the restrictions of the Fair Credit Reporting Act of 1970 and the Equal Opportunity in Credit Access Act of 1984, we do not include information on age or zip code, and we do not include any income information, to be consistent with current credit scoring models. Table 12 lists the complete set of features used in our machine learning models.

[19] Splitting means that different subsets of the dataset are created, where each observation belongs to one subset.

Table 12: Model Inputs

Category	Features
Payment History	1. Fraction of 30 days delinquent debt 2. Fraction of 60 days delinquent debt 3. Fraction of 90+ days delinquent debt 4. Months since the most recent 30-180 days delinquency on credit card trades 5. Months since the most recent 30-180 days delinquency on trades 6. Months since the most recent 90+ days delinquency 7. Months since the most recent foreclosure proceeding started on first mortgage trades 8. Number of collections 9. Number of occurrences of 90 days delinquencies in the last 36 months 10. Number of occurrences of 90 days delinquencies in the last 6 months 11. Public record bankruptcies 12. Public record discharged bankruptcies 13. Public record dismissed bankruptcies 14. Public records filed in the last 24 months 15. Trades legally paid in full for less than the full balance 16. Unsatisfied collections 17. Worst ever status on a credit card trade in the last 24 months 18. Worst ever status on a mortgage type trade in the last 24 months 19. Worst ever status on a trade in the last 24 months 20. Worst present status on a credit card trade 21. Worst present status on a mortgage type trade 22. Worst present status on a trade 23. Worst present status on a trade (excluding collections) 24. Worst present status on an installment trade 25. Worst present status on an open trade
Available Credit	1. Credit amount on home equity line of credit trades 2. Credit amount on open credit card trades 3. Credit amount on open deferred student trades 4. Credit amount on open non-deferred student trades 5. Credit amount on open trades 6. Credit amount on revolving trades
Credit Utilization	1. Credit card utilization 2. Heloc utilization 3. Installment utilization 4. Mortgage to total debt 5. Utilization ratio

Continued on next page

Table 12 – continued from previous page

Category	Features
Recent Credit	1. Bankcard revolving and charge inquiries made in the last 3 months 2. Inquiries made in the last 12 months (no deduplication) 3. Mortgage type inquiries made in the last 3 months 4. Ratio of inquiries (no deduplication) to trades opened in the last 6 months
Age and Type of Credit	1. Months since the most recently closed, transferred, or refinanced first mortgage trade 2. Months since the most recently opened credit card trade 3. Months since the most recently opened first mortgage trade 4. Months since the most recently opened home equity line of credit trade 5. Months since the oldest trade was opened 6. Number of auto loan trades 7. Number of credit & bankcards 8. Number of installment trades 9. Number of open mortgage type trades 10. Open home equity line of credit trades

Continued on next page

Table 12 – continued from previous page

Category	Features
Balances	1. Amount past due on bankcard revolving and charge trades presently 30 days delinquent 2. Amount past due on credit card trades presently 90-180 days delinquent 3. Amount past due on installment trades presently 90-180 days delinquent 4. Amount past due on joint mortgage type trades 5. Amount past due on revolving trades presently 30 days delinquent 6. Amount past due on revolving trades presently 90-180 days delinquent 7. Amount past due on trades presently 30 days delinquent 8. Amount past due on trades presently 90-180 days delinquent 9. Balance on bankcard revolving and charge trades presently 90-180 days delinquent 10. Balance on collections 11. Balance on collections placed with the collector in the last 24 months 12. Balance on credit & bankcards 13. Balance on home equity line of credit trades 14. Balance on installment trades 15. Balance on installment trades presently 90-180 days delinquent 16. Balance on open auto loan trades 17. Balance on revolving trades presently 90-180 days delinquent 18. Balance on second mortgage trades 19. Balance on trades presently 30 days delinquent 20. Balance on trades presently 60 days delinquent 21. Balance on trades presently 90+ days delinquent or derogatory 22. Credit amount paid down on open first mortgage trades 23. Monthly payment on credit card trades 24. Monthly payment on debt 25. Monthly payment on joint mortgage type trades 26. Monthly payment on open first mortgage trades 27. Monthly payment on open non-deferred student trades 28. Monthly payment on second mortgage trades 29. Total debt balances

Table 13: Summary Statistics: Selected Features

Feature	Mean	Std. Dev	25%	50%	75%
Balance on collections	700.416	4118.955	0	0	0
Balance on credit & bankcards	7972.449	18413.71	0	1006	7414
Balance on home equity line of credit trades	4192.915	29259.37	0	0	0
Balance on installment trades	8590.208	30966.59	0	0	9659
Balance on open auto loan trades	4272.418	10241.09	0	0	1507
Balance on second mortgage trades	1678.163	14879.45	0	0	0
Balance on trades presently 30 days delinquent	1205.915	17896.94	0	0	0
Balance on trades presently 60 days delinquent	567.826	13012.02	0	0	0
Balance on trades presently 90+ days delinquent or derogatory	2995.097	32610.85	0	0	0
Bankcard revolving and charge inquiries made in the last 3 months	0.098	0.325	0	0	0
Credit card utilization	0.395	3.554	0.029	0.143	0.509
Fraction of 30 days delinquent debt	0.018	0.119	0	0	0
Fraction of 60 days delinquent debt	0.009	0.084	0	0	0
Fraction of 90+ days delinquent debt	0.031	0.162	0	0	0
Heloc utilization	4.964	656.818	0.076	0.594	0.936
Inquiries made in the last 12 months (no deduplication)	1.194	2.17	0	0	2
Installment utilization	0.723	13.832	0.521	0.744	0.904
Monthly payment on debt	828.813	10463.25	0	229	1099
Months since the most recent 90+ days delinquency	28.398	27.18	5	17	52
Months since the most recent foreclosure proceeding started on first mortgage trades	13.855	21.2	2	4	13
Months since the most recently closed, transferred, or refinanced first mortgage trade	58.329	40.703	21	53	97
Months since the most recently opened credit card trade	48.968	66.882	9	25	64
Months since the most recently opened first mortgage trade	99.053	84.897	34	77	143
Months since the most recently opened home equity line of credit trade	115.552	64.765	62	127	155
Months since the oldest trade was opened	197.672	129.669	95	175	274
Mortgage to total debt	0.36	0.438	0	0	0.883
Mortgage type inquiries made in the last 3 months	0.061	0.27	0	0	0
Number of auto loan trades	0.312	0.583	0	0	1
Number of collections	0.372	1.256	0	0	0

Continued on next page

Table 13 – continued from previous page

Feature	Mean	Std. Dev	25%	50%	75%
Number of credit & bankcards	5.064	5.94	0	3	8
Number of installment trades	0.727	1.435	0	0	1
Number of occurrencies of 90 days delinquencies in the last 36 months	0.212	1.122	0	0	0
Number of occurrencies of 90 days delinquencies in the last 6 months	0.079	0.556	0	0	0
Number of open mortgage type trades	0.441	0.784	0	0	1
Open home equity line of credit trades	0.089	0.303	0	0	0
Public record bankruptcies	0.063	0.266	0	0	0
Public record discharged bankruptcies	0.048	0.215	0	0	0
Public record dismissed bankruptcies	0.007	0.102	0	0	0
Public records filed in the last 24 months	0.049	0.295	0	0	0
Ratio of inquiries (no deduplication) to trades opened in the last 6 months	1.576	8.949	0	0	0
Total debt balances	71502.85	167324.9	0	7615	78982
Trades legally paid in full for less than the full balance	0.076	0.392	0	0	0
Unsatisfied collections	0.701	2.113	0	0	0
Utilization ratio	0.649	39.937	0.199	0.63	0.855

Note: For credit history features (such as months since variables, including public record variables), we report conditionals. This means those without any history are not included in the summary statistics.

B.3 Feature Scaling

We normalize all explanatory variables by their means and standard deviations:

$$z_i = \frac{x_i - \mu_x}{\sigma_x} \tag{14}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_k)$, and $\mathbf{z_i}$ is the i^{th} normalized data.

B.4 Train-Test Split

We split the data temporally to account for look-ahead bias, meaning the training set consists of data 8 quarters (8Q) prior to the testing data to ensure that future information does not inadvertently influence the model's predictions. For instance, we calibrate the first out-of-sample model using input data from 2004Q1 with realized delinquencies from 2004Q2 through 2006Q1. The resulting parameter estimates are then applied to the input data in 2006Q1, generating forecasts of delinquencies over the following eight-quarter window from 2006Q2 to 2008Q1. We then roll forward by one quarter, updating both the training and testing windows, following the approach used in Khandani, Kim, and Lo (2010). This rolling-window framework allows for continuous model evaluation over time, adapting to changes in the underlying data patterns while avoiding look-ahead bias. This provides us with 42 calibration and testing periods, spanning from 2006Q1

to 2016Q2.[20]

For each model, we scale the testing data by the mean and standard deviation of the training data. We randomly shuffled the data in each specification to ensure that the mini-batch gradients were unbiased. If gradients are biased, training may not converge, and accuracy may be lost.

B.5 Overfitting Assessment

A natural concern with machine learning models estimated on historical data is overfitting: the model may learn idiosyncratic features of the training sample that do not generalize to new observations. We assess this risk by comparing in-sample and out-of-sample performance across six standard classification metrics.

Table 14 reports pooled results. The in-sample period spans 2004Q1–2014Q2, corresponding to the quarterly rolling windows used to train the model. The out-of-sample period spans 2006Q1–2016Q2, the held-out evaluation sample used throughout the paper. The pooled in-sample AUC of 0.9182 exceeds the pooled out-of-sample AUC of 0.9076 by approximately 0.010. This gap is modest and economically interpretable: the training and evaluation windows are separated by eight quarters, meaning the model is always assessed on a prediction environment that differs from the one it was trained on. When economic conditions shift across that gap — as they do substantially during the financial crisis — some reduction in out-of-sample performance is expected and does not indicate overfitting. Crucially, overfitting would manifest as a degradation of out-of-sample performance that widens over time or that is concentrated in periods of model complexity; neither pattern appears in the data.

Table 14: In-Sample vs. Out-of-Sample Performance

	In-Sample (2004Q1–2014Q2)	Out-of-Sample (2006Q1–2016Q2)
AUC	0.9182	0.9076
Log-Loss	0.2607	0.2723
Accuracy	0.8934	0.8891
F1-Score	0.6648	0.6438
Precision	0.8171	0.7866
Recall	0.5604	0.5449

Notes: Performance metrics are computed on the pooled sample for each period. AUC is the area under the ROC curve. Log-loss is the binary cross-entropy. F1-score is the harmonic mean of precision and recall. The model is retrained quarterly on a rolling window; in-sample metrics are computed on each training window and averaged, while out-of-sample metrics are computed on observations held out from each window.

Figure 6 plots both the in-sample and out-of-sample AUC quarter by quarter, alongside their respective pooled values as reference lines. Three features of the figure support the conclusion that the performance gap reflects economic regime differences rather than overfitting. First, the gap between the two series is largest during 2004–2009, when the training window captures the pre-crisis credit environment and the evaluation window falls eight quarters later into an increasingly distressed one — exactly the period of largest regime divergence. Second, the two series converge

[20]This setup is analogous to Khandani, Kim, and Lo (2010), where they adopt a monthly setup with a 3-month prediction window, yielding 10 splits between January 2008 and October 2008 input dates. Exactly as in our setup, each delinquency period end date matches the input date for the out-of-sample predictions. We follow this structure, as it effectively leverages the available data and provides a feasible framework for quarterly credit scoring.

sharply during 2011Q1–2012Q2: when the training and evaluation environments are both characterized by post-crisis conditions, the gap closes to near zero, confirming that the model generalizes well within a stable economic regime. Third, there is no secular widening of the gap over time — the IS–OOS difference does not grow as the model accumulates complexity through retraining, as would be expected under overfitting. Taken together, the quarterly evidence is consistent with a model that generalizes well, with performance differences driven by the economic distance between training and evaluation environments rather than by memorization of training data.

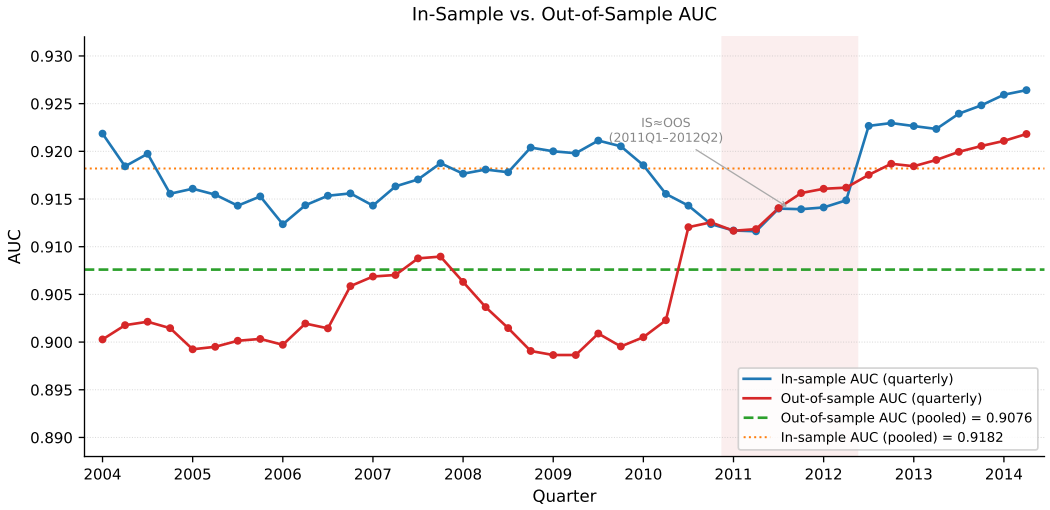
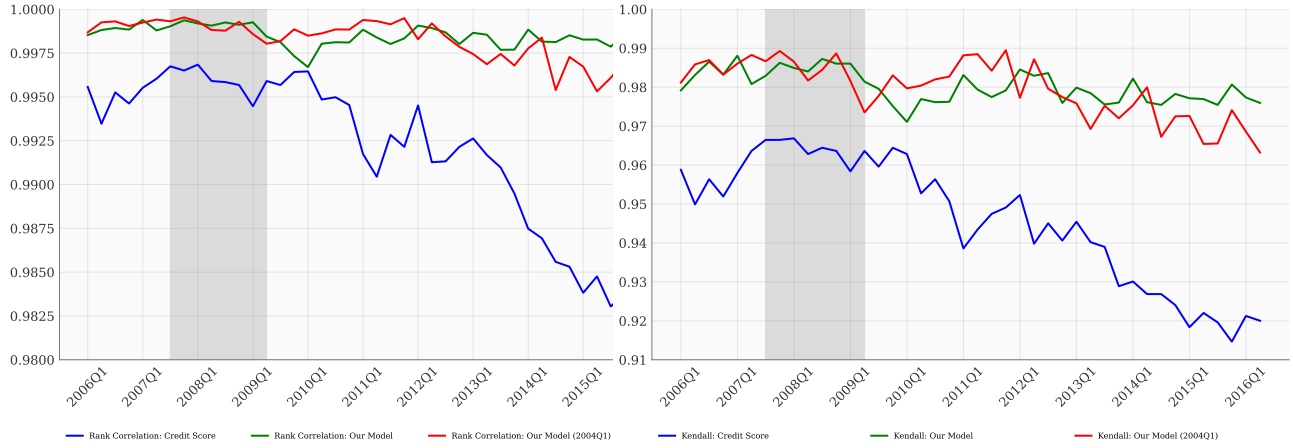


Figure 6: In-Sample vs. Out-of-Sample AUC

Notes: The solid blue line plots the in-sample AUC for each quarterly training window (2004Q1–2014Q2); the solid red line plots the out-of-sample AUC for the corresponding evaluation window (2006Q1–2016Q2). The dashed green line is the pooled out-of-sample AUC (0.9076); the dotted orange line is the pooled in-sample AUC (0.9182). The shaded region marks the 2011Q1–2012Q2 period of IS–OOS convergence. All AUC values correspond to the gradient-boosted model described in Section 3.

C Comparison with Credit Scores

This appendix reports additional performance metrics for the comparison in Section 4.2.



(a) Spearman Rank Correlation (b) Kendall Correlation

Figure 7: Spearman and Kendall Rank Correlation with Realized Default Rate

Notes: Panel a): Spearman's rank correlation for the credit score, our model and a frozen version of our model trained in 2004Q1 by quarter. Panel b): Kendall correlation for the credit score, our model and a frozen version of our model trained in 2004Q1 by quarter. Source: Authors' calculations based on Experian data.

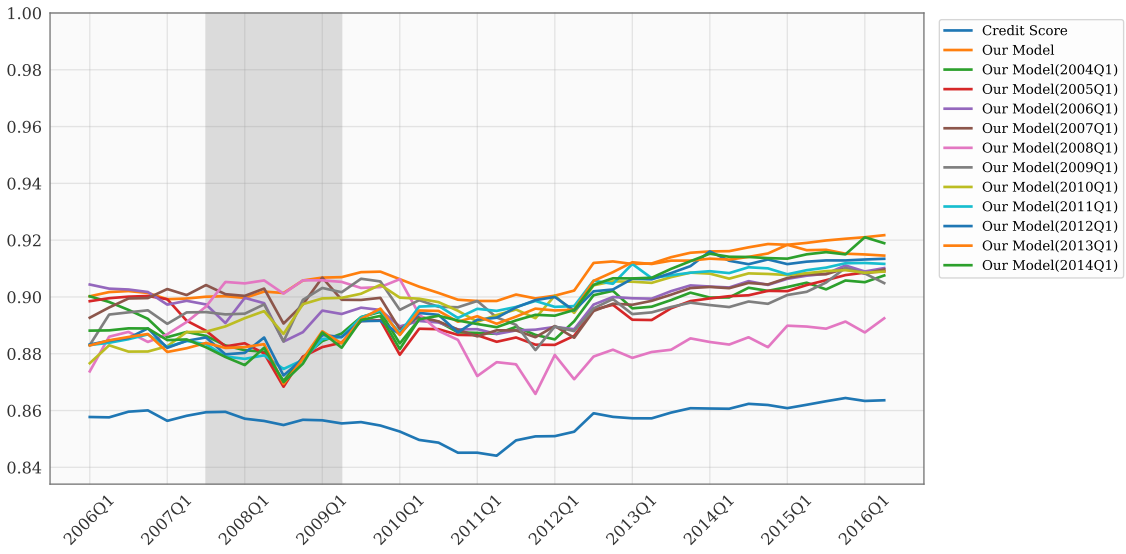


Figure 8: Gini Coefficient, Credit Score and Prediction Model

Notes: AUC scores for the credit score, our model (quarterly retraining), and our model frozen at various training periods, by quarter, 2006Q1–2016Q2. Source: Authors' calculations based on Experian data.



(a) Spearman Rank Correlation

(b) Kendall Correlation

Figure 9: Spearman and Kendall Rank Correlation with Realized Default Rate

Notes: Panel a): Spearman's rank correlation for the credit score, our model and a frozen version of our model trained in different periods by quarter. Panel b): Kendall correlation for the credit score, our model and a frozen version of our model trained in different periods by quarter. Source: Authors' calculations based on Experian data.

D Monotonicity Constraints

We next examine the performance trade-offs from imposing monotonicity constraints on our GBT model. Monotonicity constraints require that the model's predicted default probability respond in a specified direction to changes in each input — for instance, that higher delinquency severity always increases predicted default risk. Such constraints are a practical prerequisite for regulatory acceptance: lenders must be able to demonstrate palatable model behavior to regulators, and unconstrained machine learning models can otherwise produce predictions that move counterintuitively with respect to individual inputs Zoldi (2021).[21] Monotonicity constraints are straightforwardly implementable for GBT by specifying the required direction of each feature's relationship with the output.

We investigate two constraint regimes of increasing stringency. Under Regime I (R I), we constrain all "Worst status" features to be positively related to predicted default probability, requiring that greater delinquency severity always implies higher predicted risk.[22] Under Regime II (R II), we apply a comprehensive set of constraints to 39 of 79 features: credit limit, length of credit history, and temporal proximity to delinquency are negatively constrained, while "Worst status" features, number of delinquencies, bankruptcies, collections, utilization, and the fraction of delinquent balances are positively constrained. Table 15 presents ROC-AUC results alongside other performance metrics across all borrowers (Panel A), current borrowers (Panel B), and delinquent borrowers (Panel C).

The performance cost of monotonicity constraints is small. For all borrowers, Regime I reduces ROC-AUC by less than 0.001 relative to the unconstrained model (0.9065 vs. 0.9073), confirming that the unconstrained GBT already learns near-monotone relationships for delinquency-related features. Regime II entails a somewhat larger but still modest decline (0.9023 vs. 0.9073), reflecting genuine non-linearities among the more extensively constrained variables. The pattern is consistent across current and delinquent borrowers. Crucially, the constrained model under both regimes continues to substantially outperform the credit score, as reported in Table **??** of the main text. Monotonicity constraints thus impose only a minimal performance cost and do not alter the paper's core conclusions.

[21] Imposing monotonicity constraints are not yet possible for standard DNN models. You et al. (2017) proposes deep lattice networks that are monotonic with respect to a user-specified set of inputs as an alternative to standard DNNs, while Gupta et al. (2020) implements a special loss function to achieve partial monotonicity for DNNs.

[22] "Worst status" features range from current → 30 days late → 60 days late → 90 days late → 120–180 days late → derogatory, with current having the lowest value.

Table 15: Constraining Model Behavior: Performance Metrics

Metric	GBT UN	GBT R I	GBT R II
Panel A: All Borrowers			
Default	0.1842 [0.0178]	0.1842 [0.0178]	0.1842 [0.0178]
Model Prediction	0.1853 [0.0156]	0.1855 [0.0155]	0.1851 [0.0153]
Accuracy	0.8889 [0.0114]	0.8886 [0.0114]	0.8858 [0.0113]
Log-Loss	0.2730 [0.0228]	0.2739 [0.0227]	0.2795 [0.0219]
ROC-AUC	0.9073 [0.0079]	0.9065 [0.0079]	0.9023 [0.0074]
Precision	0.7848 [0.0092]	0.7837 [0.0092]	0.7779 [0.0102]
Recall	0.5474 [0.0240]	0.5463 [0.0235]	0.5323 [0.0224]
F1-Score	0.6446 [0.0165]	0.6435 [0.0165]	0.6317 [0.0160]
FPR	0.0341 [0.0053]	0.0342 [0.0049]	0.0345 [0.0056]
Panel B: Current Borrowers			
Default	0.1274 [0.0150]	0.1274 [0.0150]	0.1274 [0.0150]
Model Prediction	0.1305 [0.0110]	0.1307 [0.0108]	0.1308 [0.0107]
Accuracy	0.8903 [0.0125]	0.8900 [0.0124]	0.8877 [0.0123]
Log-Loss	0.2684 [0.0248]	0.2693 [0.0247]	0.2744 [0.0239]
ROC-AUC	0.8669 [0.0110]	0.8656 [0.0111]	0.8595 [0.0107]
Precision	0.6417 [0.0191]	0.6394 [0.0186]	0.6247 [0.0193]
Recall	0.3157 [0.0190]	0.3139 [0.0185]	0.2957 [0.0225]
F1-Score	0.4228 [0.0177]	0.4207 [0.0175]	0.4010 [0.0220]
FPR	0.0258 [0.0044]	0.0259 [0.0041]	0.0261 [0.0048]
Panel C: Delinquent Borrowers			
Default	0.8537 [0.0088]	0.8537 [0.0088]	0.8537 [0.0088]
Model Prediction	0.8321 [0.0132]	0.8325 [0.0130]	0.8247 [0.0145]
Accuracy	0.8712 [0.0059]	0.8706 [0.0061]	0.8627 [0.0072]
Log-Loss	0.3293 [0.0126]	0.3300 [0.0127]	0.3416 [0.0140]
ROC-AUC	0.8086 [0.0111]	0.8074 [0.0114]	0.7966 [0.0115]
Precision	0.9009 [0.0062]	0.9005 [0.0064]	0.8979 [0.0062]
Recall	0.9540 [0.0081]	0.9538 [0.0081]	0.9468 [0.0111]
F1-Score	0.9267 [0.0040]	0.9263 [0.0041]	0.9217 [0.0050]
FPR	0.6140 [0.0506]	0.6167 [0.0490]	0.6306 [0.0535]

Notes: Performance comparison of GBT models of consumer default risk under various monotonicity constraint regimes. UN denotes the unconstrained model, while R I and R II are the models under Regime I and Regime II respectively. The models' predicted probabilities are evaluated against actual outcomes over an eight-quarter period from Q1 2006 to Q2 2016. Data is segmented into three panels: all borrowers (Panel A), current borrowers (Panel B), and delinquent borrowers (those with 90+ DPD) in Panel C. We report quarterly averages and standard deviations (in brackets) of our metrics. Source: Authors' calculations based on Experian Data.

E Costs of Misclassification

Table 16: Descriptive Statistics on Credit Access by Credit Score Risk Profile

	(1) Credit Limit / Balances	(2) % Inquired	(3) Origination	(4) Origination if Inquired	(5) % of Borrowers with	(6) Credit Limit or Balances / Income (%)
Panel A: Credit Cards						
Deep Subprime	7,031.5	11.4	6.2	2.4	1.36	12.0
Subprime	13,170.4	11.6	17.3	6.1	7.36	20.1
Near Prime	27,685.4	12.6	31.0	13.2	9.30	36.0
Prime	32,923.3	8.2	34.5	13.6	28.94	35.8
Superprime	44,710.7	5.0	34.9	14.0	23.01	39.1
Panel B: Mortgages						
Deep Subprime	171,036.42	4.7	0.5	2.3	0.56	268.02
Subprime	172,753.09	5.0	0.8	3.9	2.62	238.81
Near Prime	188,108.00	6.9	1.6	7.6	3.60	224.40
Prime	208,991.91	6.0	1.9	10.6	11.79	195.83
Superprime	170,236.19	4.1	2.1	13.4	9.30	123.35

Notes: The table presents descriptive statistics categorized by credit score risk profiles from 2006Q1 to 2016Q2. The credit score risk profiles are defined in Table 2. The sample consists of 26,147,712 borrowers. The credit risk profile is determined based on the previous quarter. % Originated represents the fraction of borrowers within each credit risk profile category who originated in the current quarter. % Inquired indicates the fraction of borrowers within each credit risk profile category who made an inquiry in the current quarter. % Originated (Borrowers with Inquiries) reflects the fraction of borrowers who had an inquiry in the current or previous quarter and subsequently originated in the current quarter. % of Borrowers with reports the percentage of borrowers in each credit risk category who have a first mortgage or credit limit. For example, 1.36% of borrowers with a non-zero credit limit are classified as deep subprime borrowers, even though they represent 6% of the population. Therefore, only 22.7% of deep subprime borrowers have a non-zero credit limit. The table reports credit limits, first mortgage balances, and their ratios to household income for those with non-zero credit card limits and first mortgage balances. Source: Authors' calculations based on Experian Data.

E.1 Access to Credit

The specification for the access to credit regressions in Section **??** is as follows:

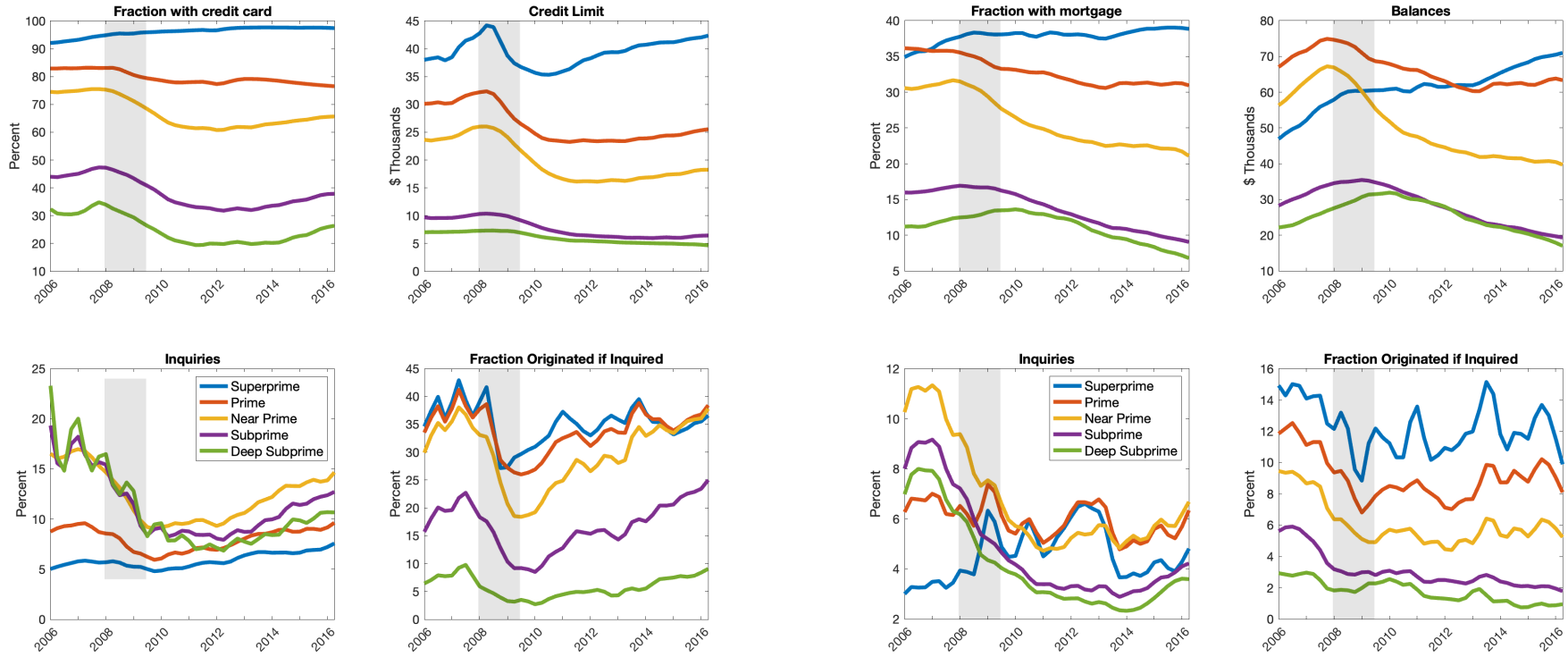
$$y_{it} = \alpha + \beta crp_{i,t-1} + \gamma a_{it} + \lambda_t \times crp_{i,t-1} + \epsilon_{it} \tag{15}$$

where i corresponds to an individual and t corresponds to time in quarters. The main regressor is $crp_{i,t-1}$, an indicator for credit score risk profile at $t-1$, varying from Deep Subprime to Superprime. The credit profile is measured at $t-1$ to avoid joint endogeneity in risk profile and credit outcomes (see Albanesi, DeGiorgi, and Nosal (2022) for a discussion). The variable a_{it} is an indicator for the consumer's age bin. We allow for five age bins, corresponding to the following age ranges in years: 18-29, 30-39, 40-49, 50-59, 60-69 and 70+. The variable λ controls time-varying factors affecting access to credit over the sample period, and is interacted with the credit risk profile, to detect any risk profile specific variations in access to credit over the sample period. Additionally, we include zip code fixed effect to account for geographical variation in home values and consumer prices and other factors that might affect the level of credit card and mortgage balances. We consider four possible outcomes: credit limits for credit cards and first mortgage balances for mortgages; inquiries and originations for credit cards and mortgages; and originations at quarter t among consumers with inquiries at quarter $t-1, t$, to allow for a lag between the posting of an inquiry and the possible realization of an origination. The estimated are reported in Table 17. The age adjusted estimated values for the time effects interacted with the credit risk profile are displayed in Figure 10.

Table 17: Access to Credit by Risk Profile

	Credit Cards				Mortgages			
	(1) Credit Limit / Balances	(2) Inquiries	(3) Origination	(4) Originations if Inquired	(5) Credit Limit / Balances	(6) Inquiries	(7) Origination	(8) Originations if Inquired
Prime	-7917.347	0.037	0.011	-0.011	20126.692	0.033	0.007	-0.031
	(130.652)	(0.001)	(0.001)	(0.005)	(494.724)	(0.001)	(0.001)	(0.005)
Near Prime	-14383.424	0.115	0.015	-0.047	9405.433	0.073	0.010	-0.055
	(173.008)	(0.001)	(0.002)	(0.006)	(646.995)	(0.001)	(0.001)	(0.005)
Subprime	-28332.291	0.143	-0.071	-0.189	-18646.570	0.050	-0.003	-0.093
	(121.642)	(0.001)	(0.001)	(0.005)	(420.904)	(0.001)	(0.001)	(0.005)
Deep Subprime	-31015.469	0.183	-0.109	-0.282	-24759.575	0.040	-0.010	-0.120
	(127.105)	(0.002)	(0.001)	(0.005)	(505.414)	(0.001)	(0.001)	(0.005)
30-39	9685.490	0.008	0.011	-0.001	48881.809	0.037	0.009	-0.002
	(70.860)	(0.000)	(0.000)	(0.001)	(417.223)	(0.000)	(0.000)	(0.001)
40-49	15675.728	0.002	0.013	-0.005	65633.694	0.032	0.008	-0.009
	(108.016)	(0.000)	(0.000)	(0.001)	(608.152)	(0.000)	(0.000)	(0.001)
50-59	18744.131	-0.009	0.009	-0.012	53565.830	0.019	0.005	-0.013
	(123.438)	(0.000)	(0.000)	(0.001)	(601.340)	(0.000)	(0.000)	(0.001)
60-69	17673.841	-0.021	-0.001	-0.013	32022.655	0.006	0.001	-0.018
	(134.698)	(0.000)	(0.000)	(0.001)	(541.962)	(0.000)	(0.000)	(0.001)
70+	7636.702	-0.042	-0.031	-0.021	-30.810	-0.015	-0.006	-0.033
	(123.213)	(0.000)	(0.000)	(0.001)	(404.867)	(0.000)	(0.000)	(0.001)
Constant	-11502.599	0.062	0.087	0.362	11335.228	0.014	0.013	0.159
	(314.503)	(0.001)	(0.001)	(0.003)	(526.828)	(0.001)	(0.000)	(0.004)
Observations	26,147,712	26,147,712	26,147,712	3,761,219	26,147,712	26,147,712	26,147,712	2,235,063
R^2	0.2518	0.0305	0.0367	0.0705	0.1241	0.0208	0.0073	0.0298

Notes: This table reports estimated for equation 15 for credit cards (columns 1 to 4), and mortgages (columns 5 to 8). Column (1) is credit limits for credit cards, while Column (5) is first mortgage balances for mortgages. Column (4) and (8) restrict the sample to individuals who inquired in the current or previous quarter and estimates the coefficients for origination. Credit risk profiles based on previous quarter. Standard errors are clustered by ZIP code.



(a) Credit Cards

(b) Mortgages

Figure 10: Credit Access By Risk Profile Overtime

Age-adjusted quarter effects interacted with credit risk profile from estimation of equation (15) for credit cards and mortgages. Age adjustment obtained by assuming aggregate age average age distribution for each credit risk profile and using age bin effects reported in Table 17. See text for more detail.

F Trade-off Between Performance and Equity

Table 18: Vulnerable Populations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
DV: Our Model Ranking - Credit Score Ranking								
Panel A: Experian Sample								
Default		-5.164***		-6.392***		-5.585***		-5.111***
		(0.032)		(0.038)		(0.036)		(0.028)
Income_{p20}			4.485***	6.067***				
			(0.069)	(0.077)				
Income_{p20} × Default				-0.438***				
				(0.048)				
Young					2.296***	2.387***		
					(0.070)	(0.074)		
Young × Default						0.983***		
						(0.045)		
Minority							0.998***	1.861***
							(0.035)	(0.041)
Minority × Default								-0.902***
								(0.030)
Constant	0.001***	0.951***	-0.857***	0.047***	-0.498***	0.461***	-0.184***	0.643***
	(0.000)	(0.006)	(0.013)	(0.013)	(0.015)	(0.016)	(0.007)	(0.008)
Observations	26,240,879	26,240,879	26,240,879	26,240,879	26,240,879	26,240,879	26,240,879	26,240,879
R^2	0.0019	0.0181	0.0145	0.0391	0.0055	0.0228	0.0024	0.0196
Panel B: HMDA Match Sample								
Default		-3.387***		-3.790***		-3.229***		-3.532***
		(0.052)		(0.057)		(0.059)		(0.061)
Income_{p20}			2.979***	3.454***				
			(0.096)	(0.105)				
Income_{p20} × Default				-0.131				
				(0.112)				
Young					2.430***	2.540***		
					(0.084)	(0.088)		
Young × Default						-0.709***		
						(0.099)		
Minority							-1.113***	-0.973***
							(0.044)	(0.048)
Minority × Default								0.907***
								(0.090)
Constant	-5.375***	-4.986***	-5.746***	-5.366***	-5.979***	-5.614***	-5.183***	-4.831***
	(0.000)	(0.006)	(0.012)	(0.014)	(0.021)	(0.024)	(0.008)	(0.010)
Observations	756,179	756,179	756,179	756,179	756,179	756,179	756,179	756,179
R^2	0.0416	0.0487	0.0475	0.0564	0.0482	0.0556	0.0426	0.0494

Notes: This table reports regressions of the form specified by Equation (13). The dependent variable is the percentile ranking difference between our model and the credit score. The independent variable, "Default", is a binary indicator showing whether the individual defaulted within the next two years. The variable "Young" is a binary indicator for individuals under 30 years old. The variable "Income_{p20}" indicates whether the individual is in the lowest quintile of the estimated income distribution. In Panel A, the Minority indicator is based on ZIP code level Black and Hispanic population being over 50%. In Panel B, the Minority indicator is 1 if the individual is Black or Hispanic. Panel A presents the full Experian sample, while Panel B provides the HMDA Match sample results. We control for State x Quarter fixed effects. Standard errors are clustered by state and quarter. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 19: Descriptive Statistics (Vulnerable Populations)

	Age: < 30	Age ≥ 30	Diff_{Age}	Income_{p20}	$\text{Income}_{>p20}$	Diff_{Income}	Minority	Non-Minority	$\text{Diff}_{Minority}$
Panel A: Experian Sample									
Default	0.23	0.17	-0.06***	0.37	0.14	-0.23***	0.28	0.16	-0.12***
	(0.42)	(0.38)		(0.48)	(0.35)		(0.45)	(0.37)	
Model Percentile	40.46	52.65	12.19***	26.33	55.60	29.27***	39.23	52.46	13.22***
	(24.13)	(29.87)		(18.62)	(28.38)		(27.52)	(28.96)	
Credit Score Percentile	38.68	53.14	14.47***	22.70	56.46	33.76***	38.54	52.61	14.08***
	(24.15)	(29.65)		(18.20)	(27.48)		(27.50)	(28.89)	
Total Balances	22.63	85.08	62.45***	10.15	86.02	75.87***	48.66	76.71	28.05***
	(65.33)	(183.69)		(39.37)	(182.08)		(122.40)	(175.56)	
90 DPD+ Balances	1.25	3.48	2.23***	1.99	3.23	1.24***	4.32	2.69	-1.63***
	(14.03)	(36.10)		(15.06)	(35.51)		(35.51)	(31.91)	
Household Income	46.95	87.05	40.10***	29.68	89.84	60.16***	62.54	81.93	19.40***
	(32.93)	(56.81)		(6.54)	(55.23)		(43.39)	(56.82)	
Age	23.81	51.56	27.76***	32.70	48.57	15.87***	43.72	45.94	2.22***
	(3.31)	(13.71)		(13.17)	(16.06)		(16.15)	(16.86)	
N	5,704,564	20,536,315	26,240,879	5,021,250	21,219,629	26,240,879	4,873,733	21,367,146	26,240,879
Panel B: HMDA Matched Sample									
Default	0.13	0.11	-0.01***	0.22	0.10	-0.12***	0.19	0.10	-0.09***
	(0.33)	(0.31)		(0.42)	(0.30)		(0.39)	(0.30)	
Model Percentile	44.22	52.94	8.73***	30.75	53.62	22.88***	40.22	52.98	12.76***
	(21.45)	(26.17)		(18.30)	(24.94)		(24.71)	(24.94)	
Credit Score Percentile	47.37	59.06	11.68***	33.13	59.43	26.30***	46.68	58.13	11.46***
	(21.45)	(25.08)		(18.72)	(23.74)		(24.62)	(24.31)	
Total Balances	44.52	158.61	114.09***	23.59	145.42	121.83***	104.03	135.73	31.70***
	(83.85)	(187.30)		(59.03)	(180.39)		(139.25)	(180.83)	
90 DPD+ Balances	0.97	2.39	1.42***	1.57	2.11	0.54***	3.38	1.76	-1.62***
	(14.25)	(26.35)		(15.30)	(24.91)		(29.96)	(22.45)	
Household Income	50.82	99.56	48.75***	30.30	95.57	65.27***	71.37	90.81	19.44***
	(30.85)	(57.67)		(6.45)	(55.64)		(44.36)	(58.03)	
Age	24.32	46.87	22.55***	28.04	43.15	15.10***	40.07	41.52	1.45***
	(3.21)	(11.51)		(9.82)	(13.53)		(13.39)	(14.16)	
N	188,005	568,174	756,179	94,176	662,003	756,179	130,869	625,310	756,179

Notes: *** denotes statistical significance at the 1% level. Differences in means between subgroups are tested via a t-test (results in "Diff" columns). Income and debt figures are in thousands. In Panel A, 'Minority status' is based on a ZIP code's Black and Hispanic population being over 50%. In Panel B, it is assigned if an individual is Black or Hispanic.

Table 20: AUC by Feature Composition

AUC Score: Our Model							
Feature Category	Status	Income		Age		Minority	
		$Income_{p20}$	$Income_{\geq p20}$	Age < 30	Age ≥ 30	Yes	No
Default History	Current	0.81	0.86	0.78	0.85	0.81	0.85
	Delinquent	0.84	0.85	0.85	0.85	0.85	0.85
Credit History	Thin file	0.88	0.90	0.85	0.90	0.87	0.90
	Thick file	0.90	0.92	0.84	0.91	0.90	0.92
Mortgage Status	No Mortgage	0.88	0.91	0.85	0.91	0.88	0.91
	Mortgage	0.89	0.91	0.87	0.91	0.89	0.91

AUC Score: Credit Score							
Feature Category	Status	Income		Age		Minority	
		$Income_{p20}$	$Income_{\geq p20}$	Age < 30	Age ≥ 30	Yes	No
Default History	Current	0.76	0.81	0.66	0.80	0.74	0.81
	Delinquent	0.64	0.65	0.61	0.67	0.63	0.65
Credit History	Thin file	0.82	0.81	0.74	0.83	0.77	0.82
	Thick file	0.89	0.91	0.81	0.90	0.88	0.91
Mortgage Status	No Mortgage	0.82	0.85	0.73	0.85	0.79	0.85
	Mortgage	0.88	0.90	0.84	0.90	0.87	0.90

Notes: AUC score for marginalized consumers by group for our model and the credit score by feature component. Source: Authors' calculations based on Experian data.

F.1 The Acted-Upon Score Problem

This appendix formalizes the acted-upon score concern discussed in Section 4.1 and explains why it does not affect our empirical comparison.

Setup. Consider two scoring models: model A (a legacy credit score used in lending) and model B (an alternative not used in lending). A borrower's true default risk is p_i, and the two models generate scores:

$$s_i^A = p_i + \eta_i^A, \qquad s_i^B = p_i + \eta_i^B \tag{16}$$

where η_i^A and η_i^B are scoring errors with mean zero. The lender extends credit if $s_i^A < \bar{s}^A$ (lower scores indicate lower risk). We observe default only for borrowers who receive credit:

$$d_i^{obs} = d_i^* \cdot \mathbf{1}[s_i^A < \bar{s}^A] \tag{17}$$

where d_i^* is the latent default outcome under the terms offered. Acting on model A creates two biases.

Selection bias. When model A underrates a borrower (η_i^A large, so $s_i^A > p_i$), the borrower may be denied credit. Default is never observed for these borrowers, and both models trivially appear to perform well. This censoring is correlated with η_i^A but not with η_i^B.

Treatment bias. When model A overrates a borrower (η_i^A small, so $s_i^A < p_i$), the borrower receives more favorable terms than warranted, which can increase default above what would have occurred under correctly priced credit. Formally:

$$d_i^* = \mathbf{1}[p_i + \gamma \cdot (p_i - s_i^A) > u_i] \tag{18}$$

where $\gamma > 0$ captures the effect of overlending on realized default and u_i is drawn from a standard distribution. The error η_i^A becomes negatively correlated with observed default among borrowers who receive credit, biasing any comparison in favor of model B even if the two models have identical underlying accuracy.

Why this concern does not apply to our setting. The acted-upon score problem requires that the evaluated score was used in lending decisions during the period over which defaults are measured. The credit score in our data is VantageScore 3.0, developed in 2013. VantageScore 3.0 was not widely adopted in lending until 2018, after our evaluation period ends in 2016Q2. During our sample, the predominant scores used in lending decisions were FICO Score 8 and earlier vintages.

Let s_i^F denote the FICO score used in lending, s_i^V the VantageScore 3.0 we evaluate, and s_i^{ML} our model's score. The lending decision was based on s_i^F:

$$\text{Credit extended if } s_i^F < \bar{s}^F. \tag{19}$$

The selection and treatment biases operate through s_i^F, not through s_i^V or s_i^{ML}. The errors η_i^V and η_i^{ML} are both uncorrelated with the lending decision conditional on s_i^F, placing the two scores on equal footing.

To the extent that VantageScore 3.0 and FICO are correlated through shared credit report inputs, some residual bias may remain. However, this bias is attenuated relative to the canonical case where the evaluated score is itself acted upon, and it operates symmetrically on VantageScore 3.0 and our model—neither directly determined lending outcomes during our sample.

Implications for the demographic results. Even if the acted-upon score problem were a concern, the relative performance gains across demographic groups are more robust than the absolute performance comparison. The selection and treatment biases operate through the legacy score's errors, not through group membership, and are therefore symmetric across groups conditional on the legacy score. The cross-group differences in performance improvement—our central finding that the ML model delivers larger AUC gains for young, low-income, and minority borrowers—are driven by the misweighting mechanism formalized in Section 2, not by the endogeneity of lending decisions.

Formally, define the performance gain for group g as $\Delta\text{AUC}_g = \text{AUC}_g^{ML} - \text{AUC}_g^{V}$. Any bias from the correlation of s_i^V and s_i^{ML} with the FICO-based lending decision affects both AUC_g^{ML} and AUC_g^{V} in the same direction for each group g. The difference-in-differences

$$\Delta\text{AUC}_M - \Delta\text{AUC}_N \tag{20}$$

eliminates this common bias under the assumption that the residual correlation with FICO does not differ systematically across groups—a reasonable assumption given that both VantageScore and our model use the same credit report inputs for all borrowers regardless of group membership.