Borrowers in the Shadows:
The Promise and Pitfalls of Alternative Credit Data

November 2024

Preliminary and Incomplete, Not for Circulation

ABSTRACT

More than 45 million U.S. adults lack traditional credit histories, creating a gap that alternative financial service data, such as payday lending records, could potentially fill. Using the staggered adoption of the largest alternative credit database, we examine the data's impact on automotive lenders in the subprime auto loan market. Alternative credit scores predict loan performance, leading lenders to offer better loan terms to higher-scoring borrowers. However, a history of using alternative financial services, even with relatively high alternative credit scores, comes with significant downsides: borrowers with payday loans histories experience higher delinquency rates, face higher interest rates, and reduced loan origination rates after the adoption of the alternative credit data. A flexible machine learning model indicates that only 6% of alternative financial service users possess sufficiently strong credit histories to offset the stigma of using these services. Consequently, alternative credit data limits credit availability and raises traditional loan costs for most users of alternative financial services. Alternative financial services are more commonly used in lower-income areas and communities with higher shares of Black residents, raising concerns that the adoption of alternative credit data may have disproportionate negative impacts on these populations. Our results contribute to the policy debate on credit data, consumer privacy, and financial inclusion.

JEL Classification: D12, D14, G23, G51.

Does alternative credit data enable financially underserved households to establish credit with traditional lenders? The United States has one of the most sophisticated and developed financial services industries in the world, but a significant fraction of Americans still lack access to traditional credit, such as bank loans or credit cards. For these financially underserved consumers, alternative financial services (AFS), such as payday lending, have become a significant source of financing. One of the greatest barriers to financial inclusion is the lack of available credit information about non-prime borrowers. In particular, more than 45 million Americans are unscorable because their financial history lacks sufficient detail or is missing altogether (Consumer Financial Protection Bureau, 2017). Rapid growth in the AFS industry has spurred parallel growth in alternative credit data. This paper studies how the adoption of alternative credit data affects automotive lending decisions and consumer access to credit.

Historically, credit transactions of AFS products were not tracked by the three major consumer credit bureaus (i.e., Experian, Equifax, and TransUnion). The main reason for this lack of coverage is that AFS lenders do not report to them. AFS lenders do not rely on traditional credit scores (e.g., FICO and VantageScore) because they have little predictive power in explaining defaults in the AFS space. Moreover, since AFS borrowers often access short-term credit from multiple AFS sources, the typical three-week reporting time for a new loan or charges to appear on a credit report is insufficient. This contrasts with the real-time credit updates used to assess AFS borrowers.

With the rise in demand for AFS products came the need for alternative data to evaluate loan requests by non-prime applicants. Usually, these data are constructed from three sources: (1) credit data from AFS lenders, (2) information regarding the timeliness of household payments such as rent, cell phone bills, utility bills, and insurance, and (3) public records and social media. Our focus in this paper is on the first of these data sources, credit data from AFS lenders. These data function similarly to traditional credit data and are used by AFS lenders to assess credit risk. More recently, traditional lenders have started acquiring this data to assist in assessing credit risk for traditional loan products such as mortgages and automobile loans.[1]

AFS data has been welcomed by many policymakers as a way for people with limited traditional credit history to build a credit history that can be used to support access to traditional credit

[1] For example, Fannie Mae incorporated rental payments (a form of alternative data) into its automated underwriting system in September 2021 (U.S. Government Accountability Office, 2021).

from lenders (U.S. Government Accountability Office, 2021). Moreover, many advocates for AFS data argue that these data are especially beneficial for low-income and minority borrowers, since these are the groups that have traditionally been excluded from the credit mainstream (Turner et al., 2006). While AFS data offer potential benefits, using alternative credit sources such as payday loans can also create a stigma that could persist even with a positive repayment history. By revealing the usage of alternative credit products, AFS data may hurt more AFS users than it helps. As alternative data are incorporated into more lending decisions, understanding the implications of using such data becomes critically important.

We use proprietary data from one of the largest alternative credit data providers. These data include more than 250 million consumer loan transactions and are used for predictive modeling and risk management by peer-to-peer lenders, AFS lenders, and automotive lenders from which the data are primarily drawn. Most loan transactions found in the data are generated by individuals with non-prime credit scores. This setting serves as an ideal laboratory to answer our three main research questions: (1) To what extent do alternative credit data predict credit risk for automobile loans, and how is the data used by lenders once they become available? (2) How are financially underserved consumers affected when lenders use alternative data in their credit decisions? (3) Are some consumer segments affected differently by the adoption of alternative credit scoring? Comprehensive and conclusive answers to these questions are important for the financial services industry, academics, and the general public.

Despite the rapid adoption of alternative credit data in lending, little is empirically known about the direction and magnitude of the impact of alternative credit data on lenders and consumers. The theory that motivates the use of credit data is decidedly mixed. For example, classic information theory informs us that more data leads to more informative decisions and less adverse selection, achieving more efficient outcomes for both lenders and consumers (Athreya et al., 2012; Livshits et al., 2016). In contrast, moral hazard is reduced by partial sharing of information rather than a policy of full sharing of information (Vercammen, 1995; Padillaa and Pagano, 2000). Furthermore, segmentation between the three major credit bureaus and alternative credit data providers may hinder non-prime consumers from transitioning to traditional, lower-cost financial services, limiting their ability to leverage positive credit outcomes in the alternative credit market as a pathway to mainstream credit access.

We find that there is significant predictive power in alternative credit scores for auto loan performance, demonstrating that higher alternative credit scores are associated with significantly lower delinquency probabilities, even after controlling for traditional credit scores, loan characteristics, and borrower demographics. Moreover, this relationship holds when we separate our sample of borrowers into terciles of traditional credit scores, indicating that alternative credit data provide incremental predictive value beyond traditional credit models. These findings underscore the utility of alternative credit scores in enhancing risk assessment for lenders, particularly for borrowers with low or moderate traditional credit scores.

Next, we investigate the effect of alternative credit data on consumers. We show that lenders' use of alternative credit data significantly impacts loan terms. Using a difference-in-differences framework, we examine changes in auto loan pricing when lenders use alternative credit data. The treatment group comprises lenders who adopted alternative credit data, whereas the control group comprises lenders who have not used the data. A critical feature of the data is that when lenders sign up for the service with the data provider, they are required to provide a sample of previous loan data. This allows us to measure portfolio changes around the adoption of the new credit data. After gaining access to alternative credit data, lenders decrease interest rates for borrowers with high alternative credit scores. In particular, a one-standard-deviation increase in alternative credit score reduces interest rates by 17.6 basis points (bp), even when accounting for traditional credit scores. Focusing on the post-adoption period, in which we observe both loans and loan applications, a higher alternative credit scores are also associated with higher loan origination probability, suggesting that alternative credit data impacts both loan approval and loan pricing. These effects are particularly pronounced for borrowers in the highest VantageScore tercile, suggesting that alternative credit scores complement traditional measures to aid in credit decisions.

Our results also reveal a downside to being included in alternative credit data. On average, borrowers with a history of using alternative financial products, as indicated by a "hit" in the alternative data, have higher delinquency rates. Because using alternative credit products is a negative signal of credit quality, these borrowers face higher interest rates and lower loan origination probabilities after lenders adopt the alternative credit data compared to individuals without a hit in the data. While the alternative score partially offsets this negative signal by improving loan approval probabilities and reducing interest rates for higher scores, the overall effect of being in the data remains significant, and having a hit in the data is generally associated with higher interest rates and lower probability of loan origination even for alternative credit users with the highest possible alternative credit scores.

Individuals who lack traditional credit scores (also known as "ghosts"), represent a significant segment of the population with limited access to affordable credit, making them highly reliant on alternative financial services. Providing more data on ghosts is crucial for improving lenders ability to assess their creditworthiness, expand financial inclusion, and reduce borrowing costs for this underserved group. For ghosts, inclusion in alternative credit data has similar effects. Ghosts with a hit in alternative data face higher delinquency rates and borrowing costs than those without a hit. However, alternative scores improve loan approval probabilities for ghosts, with the highest-scoring ghosts achieving slightly better access to credit compared to those without a hit. Despite these benefits, the effect on loan approval probabilities is modest, and there is no evidence that higher alternative credit scores reduce interest rates for ghosts.

These findings highlight the dual role of alternative credit data: While they enhance lenders ability to predict credit risk and improve credit access for high-scoring borrowers, they also signal increased risk, leading to adverse outcomes for borrowers included in the data. For ghosts, alternative credit data offer some benefit in terms of loan access, but the lack of impact on borrowing costs suggests limited advantages for this group.

Does alternative credit data allow at least some borrowers to build a positive credit history that can be helpful for obtaining traditional credit? To assess this question, we estimate flexible machine learning models using gradient boosting to assess the impact of alternative credit data on interest rates and loan origination probabilities. Consistent with our previous results, most loan applicants receive higher interest rates and experience lower loan origination probabilities compared to counterfactual estimates if they had not used alternate credit. While some AFS users have high enough alternative credit scores to overcome this stigma, this is only true for 6.13% of borrowers for interest rates and 3.86% of applicants for loan origination probability. In general, usage of alternative credit products is a negative signal that is difficult to overcome. On average, compared to the predicted interest rates and loan origination probabilities if AFS users had no alternative credit history, a hit in the AFS data increases interest rates by 0.76 percentage points (ppt) and decreases loan origination probability by 0.97 ppt.

Finally, we investigate whether lenders' use of a more data-driven underwriting process inadvertently has a disparate impact on certain consumer groups. Additional credit data can lead to more impartial and less discriminatory decision-making based hard information, which leaves less discretion to loan officers. However, adoption of alternative credit data adds a negative signal to AFS users. Because AFS usage is not randomly distributed in the population, this will have a bigger impact on some consumer groups than on others. In particular, we find that zip codes with lower incomes and higher minority population shares tend to have higher hit rates and lower credit scores in the alternative credit data. This result is particularly pronounced in zip codes with high Black population shares. As a result, adoption of alternative credit data likely contracts credit supply and increases interest rates for low-income Black borrowers. To the extent that policymakers are concerned about the distributional impacts of changing lending technology, this could be a concern.

A potential challenge to our identification strategy is that prior to the adoption of alternative credit data, we observed only approved loans, not loan applications. However, since auto dealerships submitting applications are unaware of the data adoption, it is reasonable to assume that the applicant pool remains largely consistent within a six-month window around the data adoption.

Our paper provides policymakers, such as the Consumer Financial Protection Bureau (CFPB), with additional guidance when setting regulations on the use and reporting of credit information. This is particularly important because technological advances in the last decade have dramatically expanded the amount of information collected about consumers and the ability of financial institutions to process it. For example, this has had important implications in terms of dispersion in access to credit (e.g., Fuster et al., 2022; Bartlett et al., 2022).

Our paper contributes to a growing body of literature studying the implications of using alternative credit data in lending decisions. Di Maggio et al. (2022) show that fintech models that integrate alternative data approve borrowers with low traditional credit scores but a low propensity to default, improving economic outcomes for these borrowers and returns for fintech lenders. Agarwal et al. (2023) examine data from a large fintech lender in India and emphasize the role of digital footprint data in improving credit access for underserved borrowers. Similarly, Berg et al. (2020) analyze purchases from a German E-commerce company and find that the digital footprint that individuals leave online after registering on a website predicts consumer default and complements credit bureau information. We add to this research by showing that alternative credit scores complement traditional credit scores in aiding lending decisions in the U.S. auto market. Importantly, we show that alternative credit data do not necessarily increase access to credit for all borrowers, as indicated by the lower credit access of borrowers with a history of using alternative financial products.

Recent research reveals a nuanced interaction between alternative and mainstream credit markets in the high-cost credit space. Fonseca (2023) find that restricting debt collection reduces credit access and increases payday loan borrowing, suggesting that regulatory restrictions may inadvertently push consumers towards higher-cost products that may subsequently affect their creditworthiness.[2] Liberman et al. (2021) discuss how high-cost debt, such as payday loans, can create a self-reinforcing stigma that negatively impacts future access to credit, even if they do not default on their high-cost loans. Consistent with payday loans creating a negative stigma, we show that the use of payday loans acts as a negative signal to lenders, who restrict credit to these borrowers regardless of whether they repaid their payday loans. Our results are also consistent with the broader literature that highlights the importance of asymmetric information in credit markets (e.g., Stiglitz and Weiss, 1981).

Finally, our paper contributes to the recent and growing literature on auto lending. A major area of focus in this literature has been the understanding of the lending decisions of auto lenders. For example, Melzer and Schroeder (2017) find that auto lenders change loan terms to stay below the usury limit and Jansen et al. (2022) find that dealership sales personnel face incentives to shift weak borrowers from used to new cars, increasing loan defaults.[3] Jansen et al. (2024) examine finance and vehicle margins and find that deep subprime borrowers with the lowest credit scores generally fare better than subprime borrowers.[4] Closest to our paper, Einav et al. (2013) study how automated credit scoring in automotive finance increases profits for lenders. They show that lender profits improve by screening high-risk borrowers more effectively and marketing to low-risk

[2] Allcott et al. (2022) show, however, that barring payday lending would negatively impact welfare compared to the existing regulation and Miller and Soo (2020) shows that improving access to traditional credit does not reduce payday loan use among subprime borrowers.

[3] Additional evidence that sales representatives have incentives that are not aligned with the interests of consumers can be found in Pierce (2012).

[4] Grunewald et al. (2023) show that auto borrowers are more sensitive to vehicle charges than finance charges. In addition, subprime auto borrowers are sensitive to down payments (Einav et al., 2012) and loan maturity (Argyle et al., 2020)

borrowers.

1 Data and Sample Selection

The total market for auto lending has increased to a record \$1.626 trillion in Q1 of 2024 (Federal Reserve Bank of New York, August 06, 2024). Significantly, the market for non-prime borrowers, which is the setting for our study, constitutes approximately 40% of this total. As the market for used cars has grown, so has the number of auto lenders competing to finance non-prime customers. There are approximately 65,000 financial institutions competing to finance vehicle purchases to non-prime consumers (Baines and Courchane, 2014). The auto finance industry is significantly less concentrated (The cumulative market share of originations by the top 10 was 37.7% in 2013) than the mortgage industry in which 52% of originations were accounted for by 10 originators. The Herfindahl-Hirschman Index (HHI) was approximately 200 for auto finance companies that issued at least 1000 yearly originations over the last 10 years.

Our analysis uses alternative credit data that were adopted by subprime automobile lenders on a staggered basis. The alternative credit data come from a company that collects and maintains individual credit information, providing reports and scores to lenders and other institutions to assess creditworthiness and financial behavior. In total, the database contains more than 200 million loan records from more than 20 million US adults. The data are sourced from a wide range of AFS providers including payday lenders, title lenders, and pawn shops. The database includes information on loan inquiries, employment data, wages, bank account information, and loan payment data. Non-prime consumers found in our data use on average 28 different types of financial services to spend, save, borrow, and plan during their financial lives. In aggregate, the fee and interest revenue for these alternative financial products was \$415 billion in 2023 (Gdalman et al., 2024). Significantly, over one quarter of all fees and interest from financial services were generated by subprime auto loans (\$108.8 billion) which is the nexus of our study.[5] Our analysis primarily focuses on two variables from the alternative credit data, an indicator variable for whether or not a person is in the data and the individual's alternative credit score, which the data provider calculates to predict loan

[5]This excludes Buy Here, Pay Here (BHPH) auto loans, which are in-house financed at high interest rates for subprime or no-credit borrowers at \$9.6 billion.

default probabilities using all of the underlying data.

Our auto loan data come from individual loans and loan applications from six auto lenders serving mostly subprime borrowers that started using alternative credit data for loan underwriting on a staggered basis between 2015 and 2018. Subprime auto lending is a good setting for examining how alternative credit data is used because it overlaps with alternative credit products such as payday lending, and the use of alternative credit data has grown significantly in this market over the past ten years. The data provide a snapshot of consumer finances at the time of application between 2014 and 2018 and again in June 2022. For each individual, we observe auto loan application information, such as prospective borrowers' monthly income, length of employment, age, and zip code. For loans that are originated, we observe interest rates and other loan terms such as loan size and maturity, as well as loan performance data such as subsequent delinquencies. The data also include traditional credit scores (VantageScore 4.0, hereafter Vantage score) and alternative credit scores, both as of the date of the loan application and as of June 2022.

The lenders we study began using alternative credit data between 2014 and 2016. Before using alternative credit data to assess the prospective borrowers, auto lenders used information about the borrower's traditional credit score, income, family and work status, scheduled debt payments, and ability to make a down payment. Loan origination decisions were made based on this reported information, in addition to any 'soft information' they could gather at the time of the loan request. This approach to lending was typical before the advent of alternative data to assess prospective borrowers in subprime automotive finance. We observe all loans and loan applications for each lender after it adopts the alternative credit data until the end of the study in 2017. Before adoption, we observe a random sample of approximately 700,000 loans for five of the six lenders.

1 illustrates the geographic coverage of the auto loan application data in our sample. Panel A provides a visual representation of the distribution of loan applications in different geographic regions. The density of applications is depicted, highlighting areas with varying levels of loan activity. Panel B scales the loan application number by population in each county. This highlights the penetration of the loan application process in relation to the size of the population in different regions. Overall, the figure highlights that the sample includes borrowers and loan applicants from a wide range of geographic areas.

I presents summary statistics for the sample, covering key metrics related to credit scores, loan characteristics, and demographic factors. The alternative credit score (Alt Score) has a mean of 723.61 with a standard deviation of 29.13 and ranges from 500 to 838. The Vantage score, measured for approximately 8 million observations, shows a mean of 552.97 and a standard deviation of 61.59, with values ranging from 300 to 850. The sample also includes 265,000 individuals that do not have a Vantage score. Approximately 45% of individuals in the sample are flagged with a recorded hit (Hit = 1), and 6% of the sample received an auto loan (*GotLoan* = 1) from the lender providing the application data.[6]

The average loan interest rate (*Rate*) is 21.48% with considerable variation (standard deviation of 5.69). The average loan amount is $17,220, with a standard deviation of $6,580. Loan terms have a mean of 68.27 months. The mean delinquency rate (*delinq*) is 0.39, indicating that 39% of borrowers were at some point delinquent on their loans.

Based on matching loan application zip codes to census data, the median household income at the zip-code level (*ZipIncome*) has a mean of $64,240 with a standard deviation of $21,950, and median zip-code income levels in the data range from $5,700 to $441,280. The table also includes several other zip-code-level demographic variables: the proportion of Black individuals in zip codes in the sample is 0.23 on average (*Black*), and in sample zip codes an average of 9% of individuals live below the poverty line (*Poverty100*) and 17% receive food stamps (*Foodstamp*). Payday loan shops (*PDshops*) are prevalent in the samples regions, with an average of 1.59 locations per zip code. Regarding education, zip codes in the sample have an average high school completion rate of 17% (*HighSchool*).

Finally, the data include indicators for different lenders. The majority of loans (85.6%) come from Lender 1, while smaller proportions are associated with Lenders 2 (10.1%), 3 (3.4%), 4 (0.5%), and 5 (0.5%). These summary statistics provide a comprehensive overview of the dataset, illustrating the distribution of credit scores, loan characteristics, and borrower demographics.

Figure 2 depicts the distribution of the alternative credit score for auto loan applicants. The majority of applicants have scores clustered around 700. The distribution is relatively narrow, with a concentration of scores in the upper 600s to lower 700s and a sharp peak at a score of 708, which is frequently associated with AFS users who submitted one application but did not receive

[6] Loan applications that do not result in a loan origination could be due to a denial from the lender or a decision from the borrower to forego the loan or to obtain a loan from a different lender.

an AFS loan. A vertical red dashed line is plotted at 744, indicating the alternative credit score for borrowers who were not previously present in the alternative database, indicating that they had no prior interactions with any AFS lenders. This figure shows that borrowers who enter the database by transacting with an AFS lender typically have lower alternative scores than borrowers without an AFS credit history.

Figure 3 examines the relationship between alternative credit data and traditional Vantage scores. The binscatter shows how alternative scores and hits vary with Vantage score. In addition to mean alternative scores, the plot also includes box plots summarizing the distribution of alternative scores within each Vantage score bin. The yellow squares on the plot indicate the probability of having a hit in the alternative credit data within each Vantage score bin. Across the full sample, the average hit probability is 45%. Loan applicants without a Vantage score have a hit probability of 24%. For applicants with a Vantage score, hit probability decreases with Vantage score, consistent with the usage of alternative credit products being the highest for deep subprime borrowers. In particular, borrowers with Vantage scores below 525, have a hit probability of 58.36%, compared to a hit probability of 46.07% for borrowers with Vantage scores in the 525 to 574 range and a hit probability of 31.78% for borrowers in the top decile of the Vantage score spectrum with scores 575 and above. The dots and box plots in Figure 3 plot the distribution of alternative credit scores within each Vantage score bin, conditional on having a hit in the AFS data. We find that the median alternative score is relatively stable across the Vantage score bins, with some variation in the interquartile range and outliers excluded from the boxes.

2 Main Results

In this section, we assess how well alternative credit scores predict auto loan defaults and to what extent lenders use alternative credit scores to set interest rates and make loan origination decisions. We start by analyzing alternative credit scores conditional on being in the alternative credit data and then add borrowers and applicants who are not in the alternative credit data to assess the potential negative signal of being in the data.

2.1 Does Alternative Credit Data Predict Auto Loan Performance?

For alternative credit data to be useful to lenders, they must demonstrate predictive value to assess loan default probabilities, even after controlling for other observables such as income and traditional credit scores. We start by assessing whether or not this is the case using panel data on auto loan performance. The loan data include loans that were originated both before and after the adoption of the alternative credit data. For this set of results, we limit the sample to borrowers with a hit in the alternative credit data.

Table II estimates regressions of the form:

$$\begin{aligned}\text{Delinquency}_{i,j,t} = {} & \beta_1 \text{AltScore}_{i,j,t} + \beta_2 \text{Vantage4}_{i,j,t} + \beta_3 \text{LoanAmount}_{i,j,t} + \beta_4 \text{LoanTerm}_{i,j,t} \\ & + \beta_5 \text{ZipIncome}_{j,t} + \alpha_j + \delta_t + \varepsilon_{i,j,t},\end{aligned}$$

where i indexes the loan, j indexes the county, and t indexes the time period. The dependent variable is an indicator for whether a loan becomes delinquent. All regressions in Table II include lender-by-county fixed effects to control for differences in delinquency across lenders and over time. Column (1) regresses delinquency on alternative credit score with no other control variables. The estimated coefficient of -0.0351 is highly significant and indicates that a one-standard-deviation increase in alternative credit score is associated with a 3.5 ppt decrease in delinquency probability relative to the mean delinquency rate of 46.1%. Is this new information incremental to other data already available to lenders? In particular, do alternative credit scores have additional predictive power for delinquency after controlling for traditional credit scores? To answer this question, column (2) includes both alternative and traditional credit scores.

In column (2), both the alternative score and the Vantage score are included as independent variables to assess whether the alternative score adds predictive value beyond traditional credit scores. The coefficient for alternative score remains negative and significant with a magnitude of -0.0300, which is only slightly lower compared to column (1). This suggests that while traditional credit scores (measured by the Vantage score) explain part of the variation in delinquency, the alternative score still provides incremental predictive power. The coefficient on the Vantage score is -0.0658, indicating that a one-standard-deviation increase in the Vantage score reduces delinquency probability by 6.6 ppt, which is highly significant.

Column (3) adds additional controls for the loan amount, loan term, and zip code income. The coefficient on the alternative score remains stable, showing that the alternative score continues to significantly predict delinquency even after accounting for these loan-specific factors. The traditional credit score (Vantage score) retains its strong negative association with delinquency (-0.0634), while the loan amount has a small but significant negative effect on delinquency, and longer loan terms are associated with a higher probability of delinquency.

Columns (4) through (6) restrict the sample to different terciles of the Vantage score distribution to assess whether the predictive power of the alternative score varies across traditional credit score groups. In all three Vantage score terciles, alternative score has a negative and significant relation to delinquency with a slightly stronger effect for the middle Vantage tercile. The coefficient in high Vantage score borrowers in column (6) is similar to the other terciles. This indicates that even among borrowers with high traditional credit scores, the alternative score provides additional information to predict delinquency. Across all columns, the inclusion of lender by month and county fixed effects controls for time-invariant differences across lenders and counties.

Overall, the results in Table II demonstrate that the Alternative Score provides significant predictive power for delinquency, both on its own and in conjunction with traditional credit scores. These findings highlight the value of alternative credit data for lenders seeking to assess credit risk, even for borrowers with relatively high credit scores according to traditional scoring models.

Figure 4 graphically illustrates the relation between delinquency and alternative credit score with binscatter plots of delinquency rates by alternative score broken out by Vantage score tercile. Panel A presents the relationship between delinquency and the alternative credit score controlling for loan amount, loan terms, average zip code income, lender-month fixed effects, and county fixed effects. The downward-sloping lines in the plot show a clear negative correlation between alternative credit scores and delinquency rates, meaning that as credit scores increase, the likelihood of delinquency decreases. This relation is monotonic appears to be close to linear. The three-color groups represent the different Vantage score terciles (low, medium, and high). The consistent negative slopes across all three Vantage score groups indicate that alternative credit scores consistently predict delinquencies across a wide range of applicants with different traditional credit profiles. Panel B displays a similar relationship while additionally controlling for Vantage score, which is why there is little vertical separation between the three lines. The pattern remains consistent, with a negative correlation between delinquency and alternative credit score, reinforcing the conclusion that higher alternative credit scores are associated with lower delinquency rates. The different Vantage score terciles remain close together, showing that the overall relationship between credit score and delinquency is robust across regions and Vantage score terciles.

2.2 How do Lenders Use the Alternative Credit Data?

In this section, we evaluate how lenders use alternative credit data, exploiting the fact that lenders gained access to the data on a staggered basis. As in the previous section, we restrict the sample to applicants with activity in the alternative credit score data to assess how lenders use the credit scores.

The availability of detailed loan origination and performance data from the time around the adoption of alternative credit data by lenders allows us to test the impact of such data on both lenders and borrowers. We use a stacked-cohort diff-in-diff empirical approach to identify the average treatment effects of alternative data adoption on lenders, focusing on a window of [-18 months; +18 months] around the adoption of the new credit data. Specifically, we estimate the following OLS model:

$$y_{i,c,t} = \beta(d_{i,c} \times p_{t,c}) + \alpha_{i,c} + \delta_{t,c} + u_{i,c,t} \tag{1}$$

where each pair of treated and matched control lenders i forms a unique cohort c at time t. $d_{i,c}$ identifies the treated lender, and $p_{t,c}$ the period after the dataset adoption. We add unit-cohort fixed effect dummies $\alpha_{i,c}$ to compare the outcome within the same unit around the time of the data adoption. We also add the time-cohort fixed effect dummies $\delta_{t,c}$ so that the treatment unit is compared only with the matched control at each point in time. The coefficient β represents the diff-in-diff effect of using alternative credit data on the outcome variable $y_{i,c,t}$. As outcome variables, we use measures of financial inclusion and expansion of access to credit, such as the likelihood that the borrower received a loan and the interest rate of the loan. We also use the loan performance measure, delinquency.

Table III reports the results. First, we assess how loan delinquency relates to alternative credit scores before and after lenders adopt the data. We expect that alternative scores should have a similar predictive value whether or not lenders observe the data, and this is what we observe. Column (1) of Table III reports regressions of delinquency on alternative credit score separately before and after the adoption of the data. In both cases, delinquency rates decrease with alternative score, and the magnitudes are similar to the overall magnitude estimated in Table II.

We next consider how alternative credit scores affect the interest rates by regressing the loan interest rate on alternative credit score before and after the adoption of the data. Column (2) of Table III shows no relationship between alternative credit score and loan interest rates before the adoption of the data set. In contrast, once lenders have access to the alternative credit scores, we find that a one-standard-deviation increase in alternative credit score is associated with a decrease in interest rate of 17.6 bp. This decrease represents 32.1% of the decrease in interest rates associated to a one-standard-deviation increase in Vantage score of 54.8 bp.

If alternative credit scores are informative of creditworthiness and are used by lenders, then a higher alternative credit score should translate into a higher probability of loan applications being granted. Column (3) of Table III shows that this is indeed the case. A one-standard-deviation increase in alternative credit score is associated with a 0.24 ppt increase in the likelihood that a loan application is approved. This analysis is restricted to being after the adoption of alternative credit data because we only observe loan applications after adoption.[7] For economic context, the standardized alternative score coefficient is about 57% of the coefficient of 4.2 ppt associated with standardized Vantage score.

In Appendix Table A.1, we repeat our regressions of loan interest rate origination probability in subsamples of loans split by Vantage score, after adoption of the alternative credit data. Alternative score is associated with lower interest rates and higher loan origination probabilities in all samples. The effect on interest rates is especially pronounced in the subsample of high-Vantage score loans, and the effect on loan probability is highest in the subsample of low-Vantage score loans.

[7]We observe all loans and loan applications after the adoption of the alternative credit data. Lenders provided a random sample of loans originated before adoption, but we do not have application data for non-originated loan applications before adoption.

2.3 Are Alternative Credit Users Penalized?

Thus far, we have examined the impact of alternative credit scores conditional on their presence in the data. We now add borrowers and applicants who are not in the alternative credit data to assess the potential negative signal of using alternative financial services.

Table IV regresses delinquency, interest rates, and loan probability on an indicator variable, Hit, which takes a value of 1 if someone is in the alternative credit score data. Column (1) reports the results of regressing delinquency on Hit, controlling for Vantage score and other loan characteristics. Borrowers with a history of using alternative financial products have a delinquency rate that is 5.9 percentage points higher than those without such a history, indicating that inclusion in alternative credit data serves as a signal of lower creditworthiness. This effect is large relative to the mean delinquency rate of 39%.

Column (2) adds alternative credit score in addition to the Hit indicator. To separate the effect of Hit from the effect of alternative score, we assign a standardized alternative score of zero to borrowers and loan applicants without a hit in the alternative credit score data.[8] The Hit coefficient of 0.0607 indicates that a borrower with the mean alternative credit score has a delinquency rate that is 0.06 ppt higher compared to an equivalent borrower with no history of using alternative credit. Compared to the mean alternative credit user, a one-standard-deviation higher alternative score is associated with a 0.31 ppt lower delinquency rate.

Column (3) shifts the focus to loan interest rates and shows that borrowers included in the alternative credit data pay interest rates that are 82.3 bp higher than those not in the data. This large and significant effect suggests that being included in alternative credit data is strongly associated with higher borrowing costs, potentially reflecting the perceived higher risk of these borrowers. Importantly, this coefficient is substantial compared to the mean interest rate of 21.48%. Column (4) includes the standardized alternative score alongside the Hit indicator to account for the effect of the alternative score on interest rates. A one-standard-deviation increase in the alternative score reduces interest rates by 11.4 bp, indicating that the alternative score provides lenders with additional information that mitigates some of the risk reflected in the Hit indicator. These findings further support the role of alternative scores in providing predictive value beyond the signal of

[8]Note that this differs from the data vendor's practice of assigning a score of 744 for individuals without a hit in the data.

inclusion in the data. However, to offset the 82.3 bp effect of Hit, a borrower would need to have an alternative score that is more than seven standard deviations above the mean.

Columns (5) and (6) repeat the same regressions for the loan probabilities. Column (5) examines the relationship between the likelihood of loan approval (GotLoan) and the Hit indicator. Borrowers included in the alternative credit data are 0.99 ppt less likely to have their loan applications approved compared to borrowers not included in the data. This finding suggests that inclusion in alternative credit data, independent of the alternative score itself, acts as a negative signal to lenders when making loan decisions. The coefficient is significant in relation to the mean loan approval rate of 6%. Column (6) includes the standardized alternative credit score along with the Hit indicator to disentangle the effects of being included in the data from the predictive power of the alternative score. A one-standard-deviation increase in the alternative score is associated with a 0.20 ppt higher likelihood of loan approval. These results highlight the predictive value of the alternative score in offsetting some of the negative signaling effects of the Hit indicator. However, the magnitude of the alternative score coefficient indicates that an applicant's alternative score would need to be five standard deviations above the mean to offset the negative effect of having a hit in the data.

Figure 5 illustrates the relationship between loan origination terms and the alternative score. Panel A shows a positive relationship between the alternative score and the probability of receiving an auto loan. The figure plots different terciles of Vantage score. Across all terciles, there is an upward trend, indicating that as the alternative credit score increases, the likelihood of receiving a loan increases. Borrowers without an alternative score (represented by the horizontal dashed lines for each tercile) exhibit higher loan approval probabilities compared to those included in the alternative data. This suggests that inclusion in the alternative data, while predictive of creditworthiness, may also act as a negative signal to lenders, reducing the likelihood of loan approval.

Panel B displays a negative relationship between the alternative credit score and the interest rate on auto loans. We observe that higher alternative credit scores for each Vantage score tercile are associated with lower interest rates. However, the borrowers not in the alternative data appear to benefit from systematically lower interest rates, suggesting that lenders may perceive them as less risky despite the absence of an alternative score. Together, the panels highlight the dual role of alternative credit data in providing predictive value while also signaling increased risk, leading to divergent loan outcomes for borrowers included in and excluded from the dataset.

Figure 6 shows the effect of being included in alternative credit data ("Hit") on loan interest rates over time, with month 0 marking the adoption of the data by lenders. Before adoption, the coefficients hover around zero, indicating that "Hit" was not priced by lenders. This absence of a pre-trend suggests that the observed post-adoption pricing effects are not driven by endogenous market factors. After adoption, the coefficients increase steadily, suggesting that lenders progressively incorporate the information from alternative credit data, leading to higher interest rates for borrowers included in the data.

2.4 Ghosts

The analysis thus far indicates that information about past alternative credit usage hurts most users of alternative credit. We next turn to a population that may have the most to gain from alternative credit data: individuals who lack traditional credit scores, frequently referred to as "ghosts" in the lending industry. One might expect that alternative data are more valuable to lenders with a greater degree of asymmetric information with regard to the borrower's type. Similarly, consumers with fewer records from traditional credit bureaus could benefit the most from alternative data. In this subsection, we examine whether the effects of using alternative credit data are particularly important for these types of consumers.

Table V examines the relationship between delinquency, loan approval probability, interest rates, and the Hit indicator for a sample restricted to borrowers and applicants without a Vantage score, referred to as "ghosts." This table replicates the analysis from Table IV assessing whether the inclusion in alternative credit data disproportionately affects individuals lacking traditional credit scores.

Column (1) explores the relationship between delinquency and the Hit indicator. Borrowers included in the alternative credit data are 6.3 ppt more likely to experience delinquency compared to those excluded from the data. This highly significant result highlights that even among ghosts, inclusion in alternative credit data signals higher credit risk. The effect is substantial relative to the mean delinquency rate of 36.3% for this sample. Column (2) adds the standardized Alternative credit score, which is insignificant in the ghost subsample, likely due to lack of power.

Columns (3) and (4) focus on interest rates as the dependent variable. Borrowers included in alternative credit data pay significantly higher interest rates, 75.0 bp in column (3) compared to those not included. Standardized Alternative credit score is insignificant when it is added in column (4), again likely due to a lack of power. These findings indicate that, for ghosts, inclusion in alternative credit data increases borrowing costs significantly, even when controlling for additional loan characteristics. The alternative score does not affect loan interest rates in a meaningful way, and its coefficient remains insignificant in rate regressions.

Columns (5) and (6) examine the likelihood of loan approval (GotLoan). In both specifications, the coefficient on Hit is negative and significant, with borrowers in the alternative credit data being 0.16 ppt less likely to have their loans approved. At the same time, the information contained in the alternative score significantly increases the likelihood that a ghost receives a loan.

Figure 7 shows a graphical representation of how interest rates and loan probabilities vary with the alternative credit score for borrowers who lack a traditional credit score, controlling for Vantage score, zip-code income, and month-lender and county fixed effects. In Panel A, there is no clear relation between interest rates and alternative credit scores for ghosts. Loan origination probabilities are plotted in Panel B. The average probability of loan origination for ghost applicants without a hit in the alternative credit data is 2.1%. Ghosts with a hit in the alternative data have a lower probability of loan origination. However, there is an upward-sloping relation between loan origination and alternative score, and ghosts with alternative scores above around 710 have slightly higher loan origination probabilities than ghosts without a hit in the alternative data.

Overall, the results in Table V and Figure 7 indicate that alternative credit data have an effect on ghosts that is similar to their effect on borrowers more generally. Ghosts who have used alternative credit are generally riskier than borrowers without a hit in the alternative credit data. Ghost loan applicants with the highest alternative credit scores may experience slightly better access to auto credit compared to those without alternative credit history. However, the effect on loan probability is small, and there is no evidence that high alternative credit scores lead to lower interest rates for ghosts.

3 Building Credit with Alternative Financial Services

One of the promises of alternative data is that it may also provide a path for non-traditional borrowers to establish a positive credit history. In this section, we assess how often this is the case. In general, using alternative credit is a negative signal. How often do loan applicants have strong enough alternative credit scores to overcome the negative signal of having a hit in the data?

The results in the previous section suggest that a hit in the alternative financial services data is typically a negative signal even when the alternative credit score is relatively high. However, a limitation of these linear models is that they might miss out on important interactions and nonlinearities that could affect actual lending decisions. To estimate more flexible models, we implement machine learning models for all three outcome variables.

Focusing first on predicting loan interest rates, we evaluate over 200 model configurations in four classes of machine learning models: gradient boosting (XGBoost), random forests, neural networks, and elastic net regression. Each model is trained using 5-fold cross-validation to prevent overfitting, with the best model selected based on out-of-sample RMSE. The models incorporate borrower characteristics, including traditional and alternative credit scores, loan characteristics (amount and term length), borrower zip code demographics (income and race), and fixed effects for lender and month. XGBoost emerges as the best performing model, with optimal parameters including 400 trees, maximum depth of 4, learning rate of 0.1, minimum child weight of 3, and L2 regularization parameter of 1.0. This configuration achieves an RMSE of 4.17 percentage points, capturing important nonlinear patterns while maintaining strong out-of-sample performance. Following the same procedure, we also estimate XGBoost models along the same lines to predict delinquency and loan origination.

To understand how the models use alternative credit information, we employ SHAP (SHapley Additive exPlanations) values. SHAP values provide a unified measure of the importance of the characteristics that shows how each variable contributes to individual predictions. For each prediction, SHAP values decompose the difference between the model's prediction and the average prediction into contributions from each feature. This allows us to understand not just which features are important, but how they affect predictions across their range of values. In our context, SHAP values help quantify how much having a hit in the alternative data and different alternative credit scores contribute to predicted delinquency rates, interest rates, and loan probabilities relative to baseline predictions. By analyzing the distribution of these effects among borrowers, we can identify when alternative credit information helps versus hurts different types of borrowers.

Figure 8, Panel A shows how the combined effect of having a hit in alternative credit data and one's alternative credit score impacts lending decisions. The y-axis plots the SHAP values representing the total impact of alternative credit information (both having a hit and the alternative score), while the x-axis shows the alternative credit score. The figure shows a strong non-linear relationship. For borrowers with alternative scores below 700, the combined effect is positive but diminishing, suggesting that higher alternative scores partially offset the negative impact of being in the data. There is a sharp decline in the combined effect for scores between 700 and 750, after which the impact stabilizes near zero. The rightmost data points show the impact of not having a hit in the alternative data (labeled "No Hit"), which serves as a reference level. Notably, even borrowers with very high alternative scores (above 800) rarely achieve SHAP values comparable to having no hit in the data, indicating that the negative signal of using alternative financial services is difficult to overcome even with extremely high alternative scores. The vertical differences in SHAP values for borrowers with the same alternative score come from interactions with different covariates, which the model flexibly estimates.

Panel B of Figure 8 plots SHAP values for other covariates. The plots indicate that interest rates generally decrease with Vantage score, median income (at the zip-code level), and loan amount, and interest rates tend to increase with the percentage of the population in a zip code that is Black. In addition to estimating the direction and magnitude of these relations, the models account for nonlinearities and interactions, represented by a range of SHAP values for same covariate value.

We next estimate the overall distribution of how the use of alternative credit data affects predicted outcome variables using both OLS and gradient-boosting machine learning models. For OLS estimates, we regress delinquency, interest rate, and loan origination probability on Hit, alternative score, Vantage score, and zip code income with fixed effects for lender-month and applicant county. Estimates for rate and delinquency also control for loan amount and loan term. For each borrower with a hit in the alternative credit data, we then use the model to estimate the combined effect of Hit and the borrower's alternative score relative to a counterfactual in which the borrower has no hit in the alternative credit data and has an alternative score that is the same as other borrowers with no hit in the data.

Column (a) of Figure 9 plots histograms of the OLS results. For predicting delinquency, the usage of alternative credit is associated with higher delinquency rates 94.18% of the time. To have a negative effect on delinquency rates, borrowers need an extremely high alternative score, and this only occurs 5.82% of the time, as plotted in green in the figure. The results are even sharper for interest rates and loan probabilities. Based on the linear model, the usage of alternative credit appears to increase interest rates and decrease loan probabilities 100% of the time. On average, the usage of alternative credit is associated with interest rates that are 0.82 ppt higher and loan probabilities that are 1.02 ppt lower.

Column (b) of Figure 9 plots the results for the XGBoost models. With more flexible models, alternative credit usage has a greater ability to improve credit outcomes. For delinquency, alternative credit usage is associated with a lower delinquency probability 6.05% of the time compared to 5.82% of the time using the linear model.

For interest rates and loan probabilities, the XGBoost models indicate that alternative credit usage enhances credit outcomes for some borrowers, but this is rare. Only 6.13% of borrowers have alternative scores that are high enough to decrease their interest rate, and only 3.86% of applicants have alternative scores that are high enough to increase their loan probability. On average, the usage of alternative credit is associated with interest rates that are 0.76 ppt higher and loan probabilities that are 0.97 ppt lower, which is similar to estimates from the linear models.

4 Differences Across Demographic Groups

The previous section indicates that the vast majority of borrowers with alternative credit usage are hurt by the revelation of this information. To the extent that some demographic groups are more likely to use alternative credit products, this could have differential effects on different groups.

To evaluate this possibility, we regress Hit and the alternative score (conditional on having a hit) on zip-code-level demographic variables. The results are reported in Table VI. In column (1), we regress Hit on the average income of zip codes with a standardized coefficient of -0.0141, indicating that zip codes with one-standard-deviation higher income are associated with 1.4 ppt lower probability of having a hit in the alternative credit data. Column (2) adds race to the regression with a positive significant result for minority population share. The estimate indicates that a one-standard-deviation increase in minority population share is associated with a 3.0 ppt higher probability of having a hit. In column (3), we add more detailed racial characteristics with a large positive relation between Hit and Black population share.

In columns (4) to (6), we repeat the same regressions for alternative score, conditional on having a hit in the alternative data. Across all three specifications, income has a positive relation to the alternative score. Minority population share has a negative relation with alternative score, particularly with respect to Black population share.

The results in Table VI show a clear pattern of more frequent alternative credit usage and lower alternative scores in zip codes with lower incomes and higher Black population shares. As a result, these areas are likely to have the largest negative effects when lenders adopt alternative credit data.

5 Conclusion

This study highlights the transformative potential of alternative credit data for improving credit risk assessment and loan pricing, with important implications for financial inclusion. By providing incremental predictive power for delinquency beyond traditional credit metrics, alternative credit scores enable lenders to make more informed decisions, particularly for underserved borrowers. However, the dual role of these data–as both a predictor of risk and a potential negative signal–underscores the complexity of their impact on borrower outcomes.

The findings in this paper reveal a dual dynamic in the use of alternative credit data, where its positive impact on borrowers is offset by the negative signal associated with inclusion in the data. On one hand, alternative credit data enhances lenders ability to make informed decisions by reducing interest rates and increasing loan approval probabilities for borrowers with higher alternative scores, particularly those with lower traditional credit scores. This highlights the potential of alternative data to improve financial inclusion by expanding credit access and offering better loan terms to underserved populations. On the other hand, inclusion in the alternative credit data itself acts as a negative signal and is associated with higher delinquency rates. In response to this credit risk, lenders set higher interest rates for borrowers with a hit in the alternative credit data, and loan applicants with a hit also have a lower probability of loan origination. This negative signal is partially offset for borrowers and loan applicants with high alternative credit scores, but the net effect is typically negative even for borrowers with relatively high alternative credit scores. As a result, very few borrowers build a positive credit history with alternative credit data, and the increased adoption of these data is likely a hindrance to financial inclusion for users of alternative financial services, even for individuals who lack any traditional credit history.

The adoption of alternative credit data also has significant distributional implications. Because the usage of alternative financial services is higher in areas with low incomes and high shares of the population that are Black, the availability of alternative credit data is likely to increase average interest rates and decrease the availability of credit for these populations.

References

Agarwal, Sumit, Shashwat Alok, Pulak Ghosh, and Sudip Gupta, 2023, Financial inclusion and alternate credit scoring: Role of big data and machine learning in fintech, *Working Paper* .

Allcott, Hunt, Joshua Kim, Dmitry Taubinsky, and Jonathan Zinman, 2022, Are high-interest loans predatory? theory and evidence from payday lending, *The Review of Economic Studies* 89, 1041–1084.

Argyle, Bronson S, Taylor D Nadauld, and Christopher J Palmer, 2020, Monthly payment targeting and the demand for maturity, *Review of Financial Studies* 33, 5416–5462.

Athreya, Kartik, Xuan S Tam, and Eric R Young, 2012, A quantitative theory of information and unsecured credit, *American Economic Journal: Macroeconomics* 4, 153–183.

Baines, Arthur P, and Marsha J Courchane, 2014, Fair lending: Implications for the indirect auto finance market, *Study Prepared for the American Financial Services Association* .

Bartlett, Robert, Adair Morse, Richard Stanton, and Nancy Wallace, 2022, Consumer-lending discrimination in the fintech era, *Journal of Financial Economics* 143, 30–56.

Berg, Tobias, Valentin Burg, Ana Gombović, and Manju Puri, 2020, On the rise of fintechs: Credit scoring using digital footprints, *The Review of Financial Studies* 33, 2845–2897.

Consumer Financial Protection Bureau, 2017, Who are the Credit Invisible?, [Accessed November 6, 2024].

Di Maggio, Marco, Dimuthu Ratnadiwakara, and Don Carmichael, 2022, Invisible primes: Fintech lending with alternative data, Technical report, National Bureau of Economic Research.

Einav, Liran, Mark Jenkins, and Jonathan Levin, 2012, Contract pricing in consumer credit markets, *Econometrica* 80, 1387–1432.

Einav, Liran, Mark Jenkins, and Jonathan Levin, 2013, The impact of credit scoring on consumer lending, *The RAND Journal of Economics* 44, 249–274.

Federal Reserve Bank of New York, August 06, 2024, Household debt increased moderately in q2 2024; auto and credit card delinquency rates remain elevated, *News Release* www.newyorkfed.org/newsevents/news/research/2024/20240806.

Fonseca, Julia, 2023, Less mainstream credit, more payday borrowing? evidence from debt collection restrictions, *Journal of Finance* 78, 63–103.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther, 2022, Predictably unequal? The effects of machine learning on credit markets, *Journal of Finance* 77, 5–47.

Gdalman, Hannah, MK Falgout, Necati Celik, and Meghan Greene, 2024, Finhealth spend report 2024: Record spending on credit services puts pressure on vulnerable households, *FinHealth Spend* finhealthnetwork.org/wp-content/uploads/2024/08/FinHealth-Spend-Report-2024-FHN.pdf.

Grunewald, Andreas, Jonathan Lanning, David Low, and Tobias Salz, 2023, Auto dealer loan intermediation: Consumer behavior and competitive effects, Working paper.

Jansen, Mark, Samuel Kruger, and Gonzalo Maturana, 2024, Dealer financing in the subprime auto market: Markups and implicit subsidies, *Management Science, Forthcoming* .

Jansen, Mark, Hieu Nguyen, Lamar Pierce, and Jason Snyder, 2022, Product sales incentive spillovers to the lending market, *Management Science, Forthcoming* .

Liberman, Andres, Daniel Paravisini, and Vikram Pathania, 2021, High-cost debt and perceived creditworthiness: Evidence from the uk, *Journal of Financial Economics* 142, 719–736.

Livshits, Igor, James C Macgee, and Michele Tertilt, 2016, The democratization of credit and the rise in consumer bankruptcies, *Review of Economic Studies* rdw011.

Melzer, Brian, and Aaron Schroeder, 2017, Loan contracting in the presence of usury limits: Evidence from automobile lending, *Working Paper* .

Miller, Sarah, and Cindy K Soo, 2020, Does increasing access to formal credit reduce payday borrowing?, Technical report, National Bureau of Economic Research.

Padillaa, A. Jorge, and Marco Pagano, 2000, Sharing default information as a borrower discipline device, *European Economic Review* 44, 1951–1980.

Pierce, Lamar, 2012, Organizational structure and the limits of knowledge sharing: Incentive conflict and agency in car leasing, *Management Science* 58, 1106–1121.

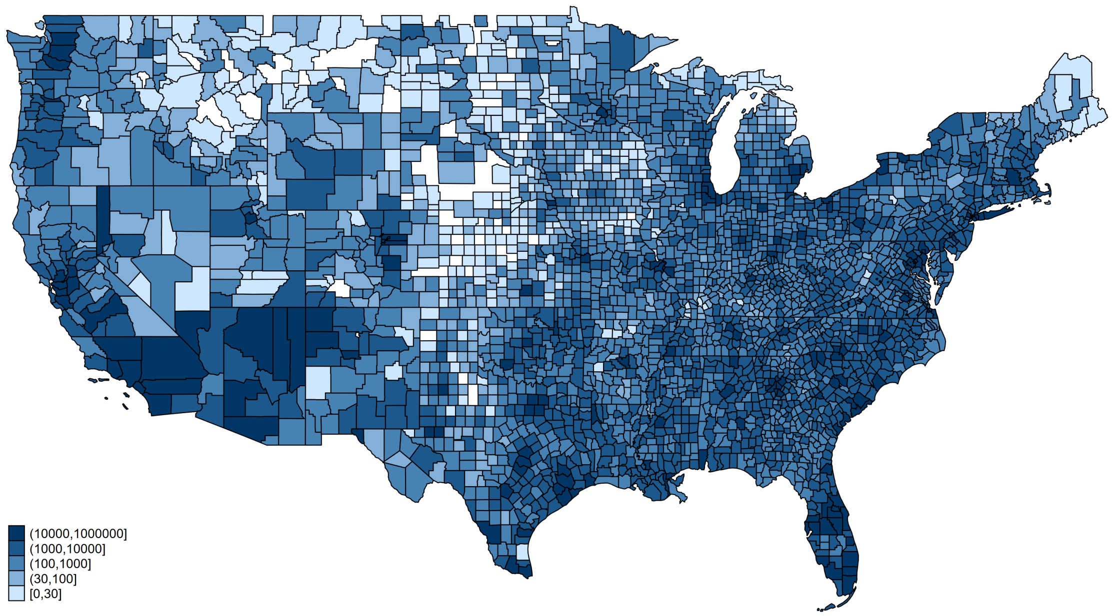
Stiglitz, Joseph E, and Andrew Weiss, 1981, Credit rationing in markets with imperfect information, *American Economic Review* 71, 393–410.

Turner, Michael, Alyssa Stewart Lee, Ann Schnare, Robin Varghese, and Patrick D Walker, 2006, Give credit where credit is due: Increasing access to affordable mainstream credit using alternative data, *The Brookings Institution, Washington, DC* .

U.S. Government Accountability Office, 2021, Mortgage Lending: Use of Alternative Data Is Limited but Has Potential Benefits.

Vercammen, James A., 1995, Credit bureau policy and sustainable reputation effects in credit markets, *Economica* 62, 461–478.

(a) A. Number of Applications



(b) B. Percentage of County Population

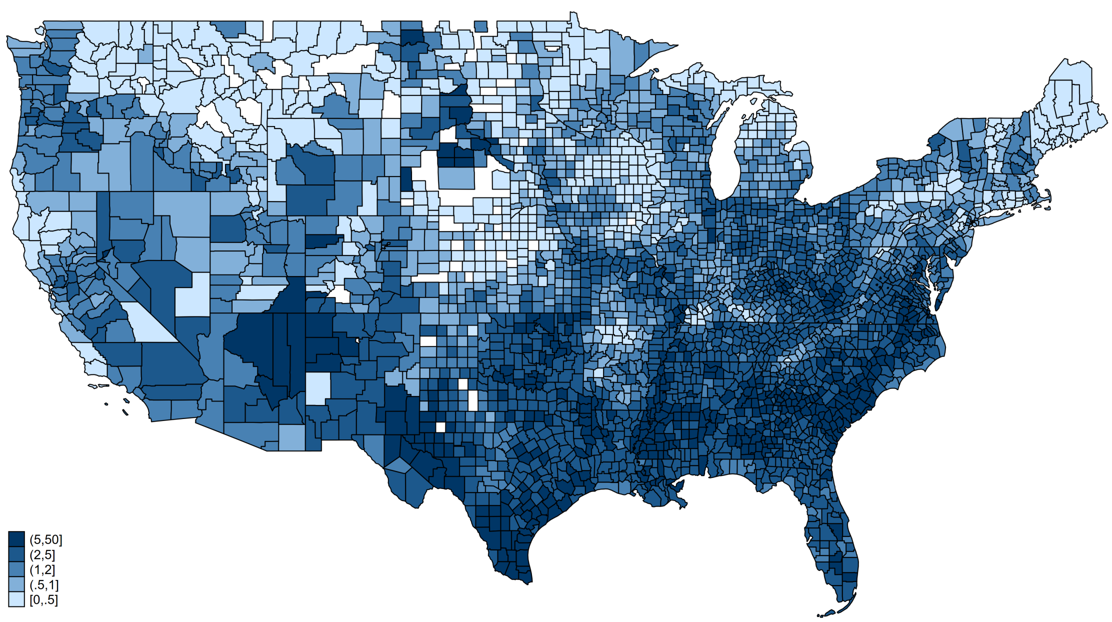


Figure 1. Geographic Coverage. This figure shows the geographic coverage of the loan applications in our data. Panel A shows the raw number of loan applications for each county, and Panel B shows the number of applications as a percentage of each county's population. The county population data is from 2010 Census data.

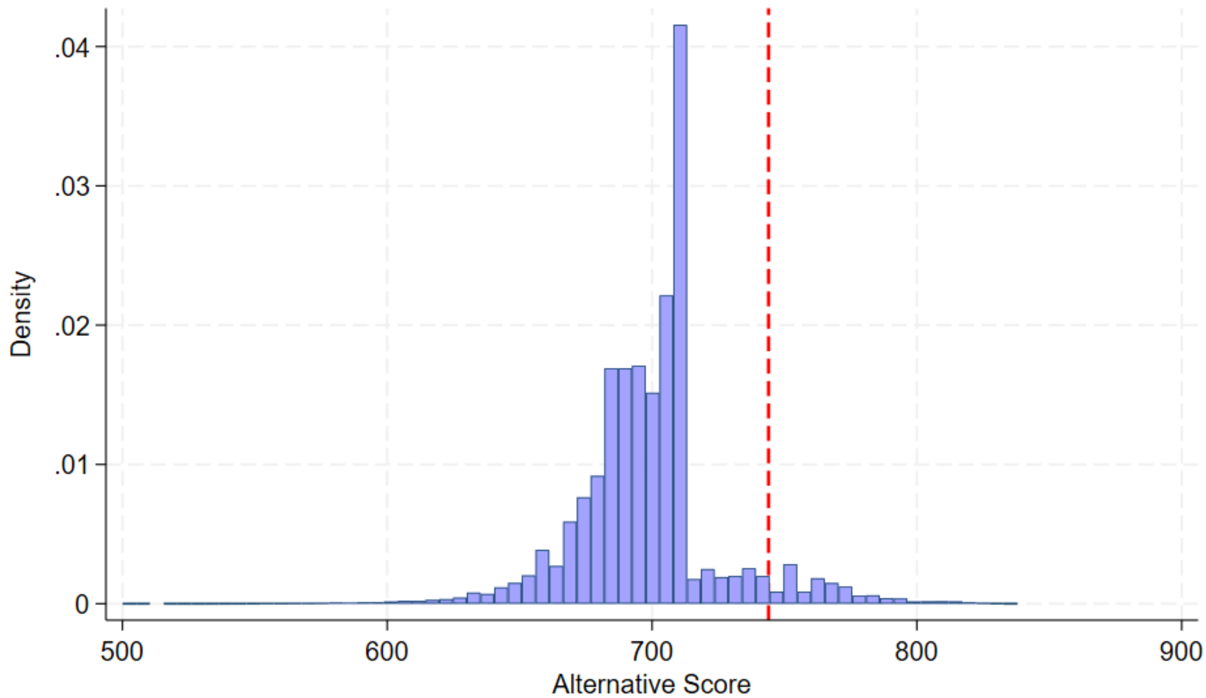


Figure 2. Distribution of the Alternative Score for Rated Loan Applicants. This figure shows the distribution of the alternative credit score for applicants who have a hit in the alternative data, and as a result, receive an alternative credit score.

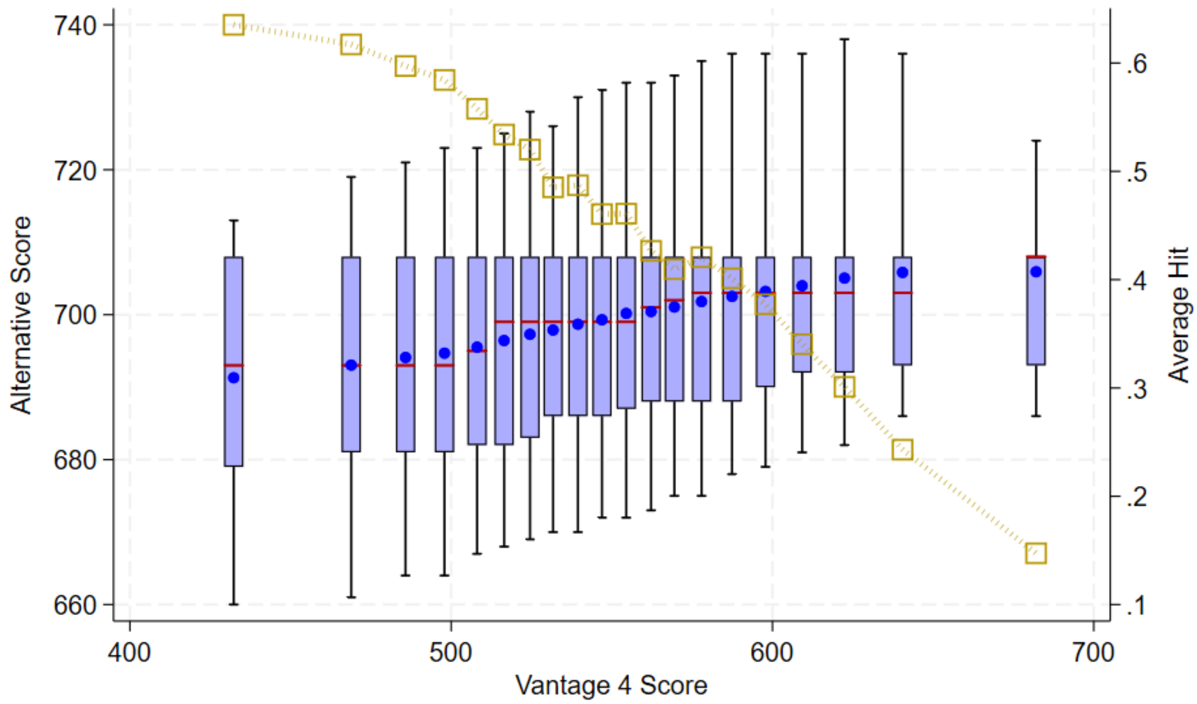
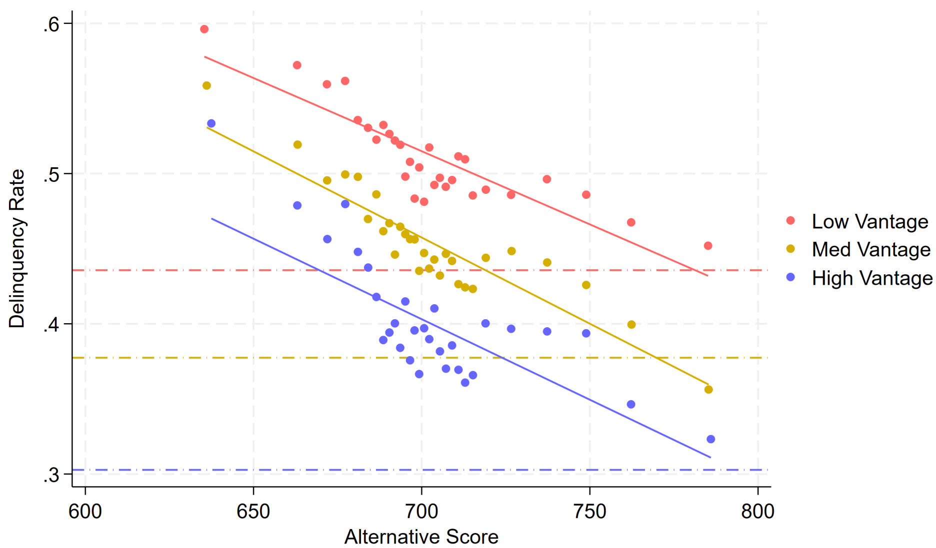


Figure 3. The Relationship between the Alternative Score and Vantage4 Score. Panel A shows the 10- to 90-percentile range of borrowers' alternative scores for 20 quantiles of Vantage 4 score. Boxes represent the interquartile range of the alternative score. The horizontal red line in each box represents the median value, and the blue circles show the mean value. The gold hollow squares show the average ratio of the hit population for each quantile of the Vantage 4 score.

(a) No vantage score control



(b) Controlling for vantage score

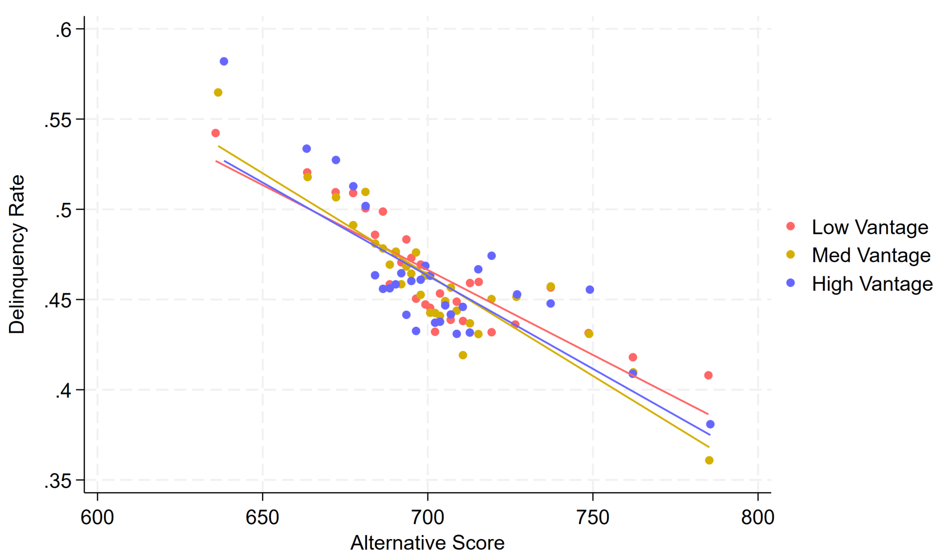
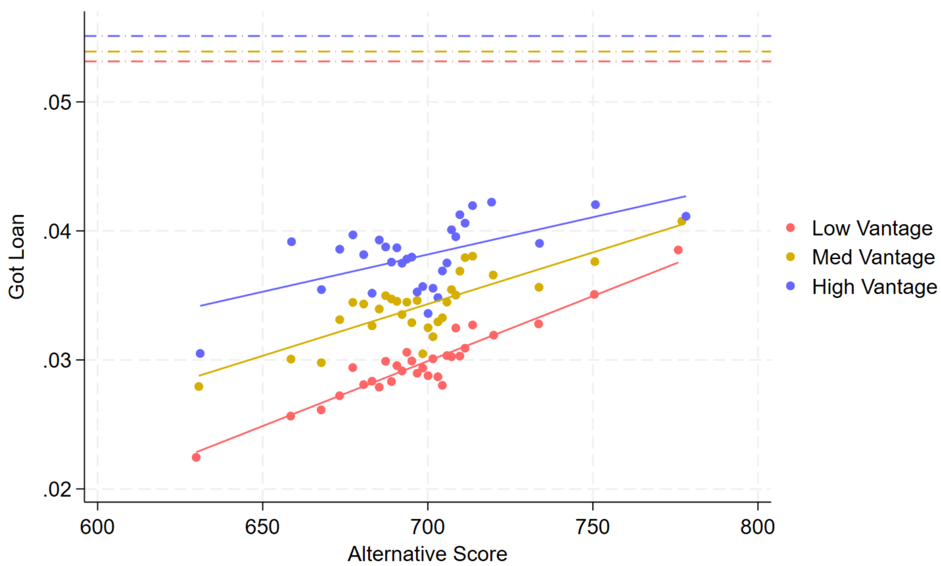


Figure 4. Relationship between Loan Delinquency and the Alternative Score for Different Terciles of Vantage Score. This figure shows the binscatter of loan delinquency over the alternative score. The red dots and lines correspond to the lowest tercile of the Vantage score, gold to the middle tercile, and blue to the highest tercile. The results in panel A control for the loan amount, loan terms, average zip code income, and month-lender and county fixed effects. Panel B additionally controls for the vantage 4 score. The horizontal dashed line shows the average value for applicants in each Vantage group who do not hit the alternative credit data.

(a) Loan approval rate



(b) Interest rate

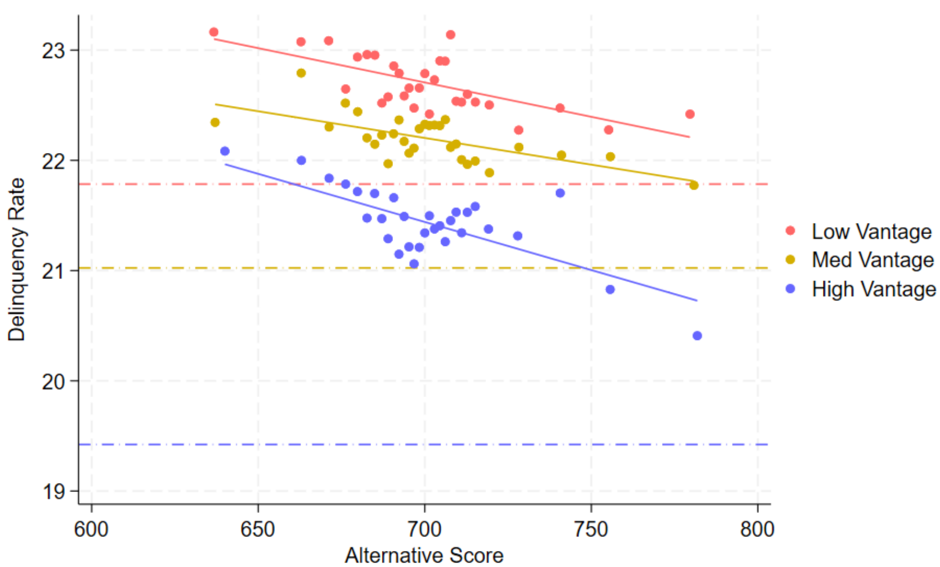


Figure 5. Relationship between Loan Origination and Loan Rate and the Alternative Score for Different Terciles of Vantage Score. Panel A shows the binscatter of loan approval rate over the alternative score. Panel B shows a similar graph for loan interest rate and the alternative score. The red dots and lines correspond to the lowest tercile of the Vantage score, gold to the middle tercile, and blue to the highest tercile. Both panels control for the vantage score, average zip code income, and month-lender and county fixed effects. Panel B additionally controls for the loan amount and loan terms. The horizontal dashed line shows the average value for applicants in each Vantage group who do not hit the alternative credit data.

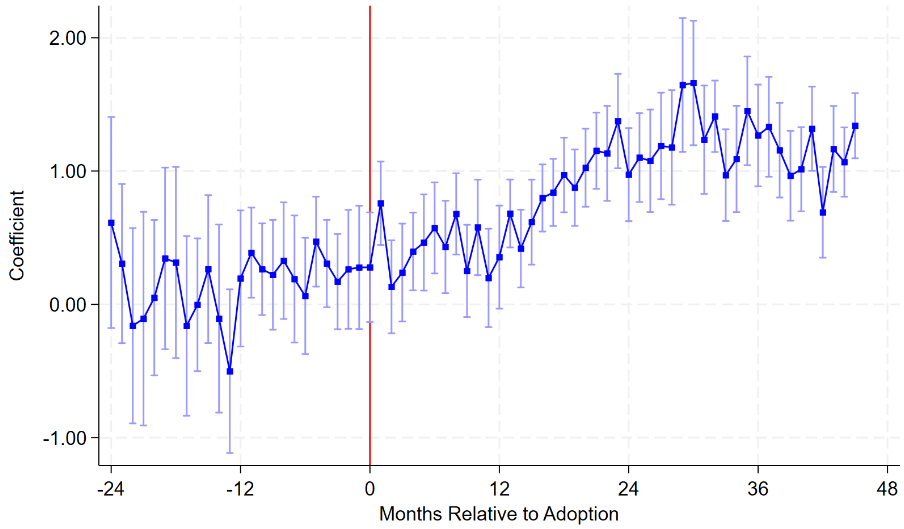
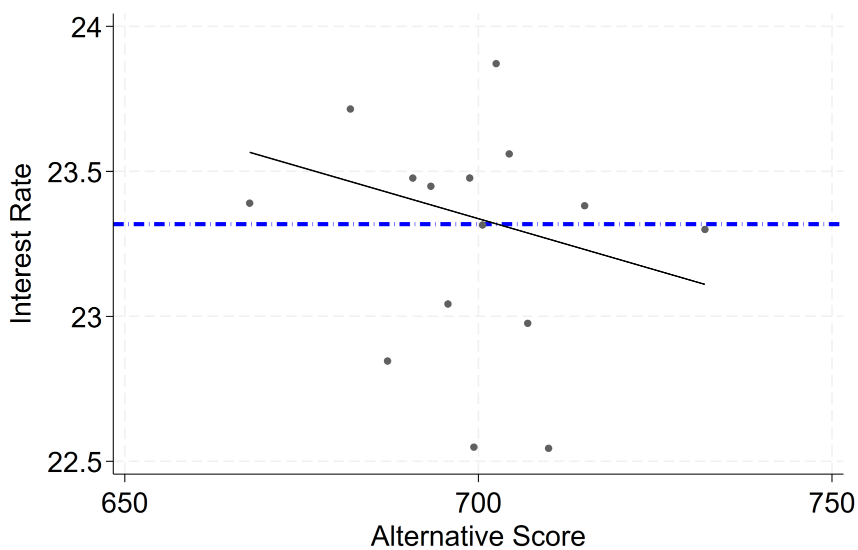


Figure 6. The Effect of Hit on Loan Rates Over Time. This figure plots the coefficients of a regression of loan rates on hit, control variables including vantage score, loan amount, loan terms, zip code income, and lender-month and county fixed effects. The regression is run for each month relative to the time of adoption separately and then graphed over time. The blue vertical lines show the 95% confidence interval for the coefficients.

(**a**) Interest rate



(**b**) Got Loan

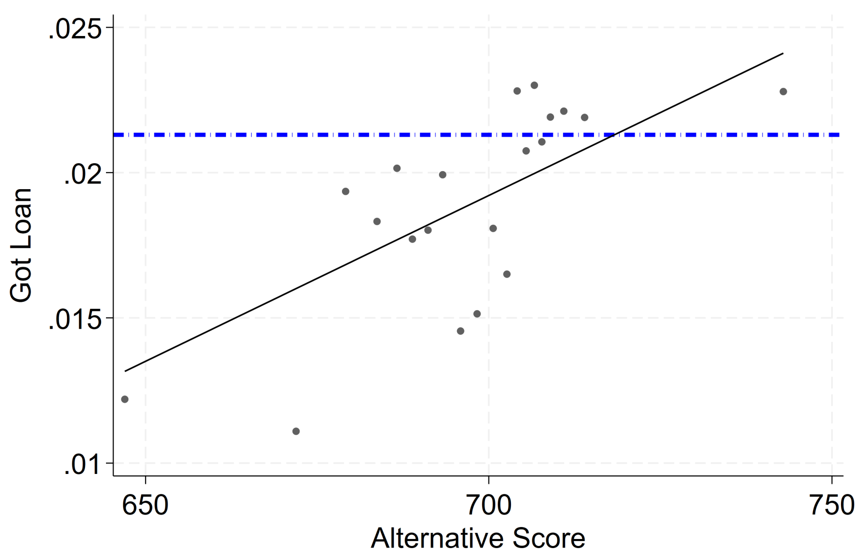
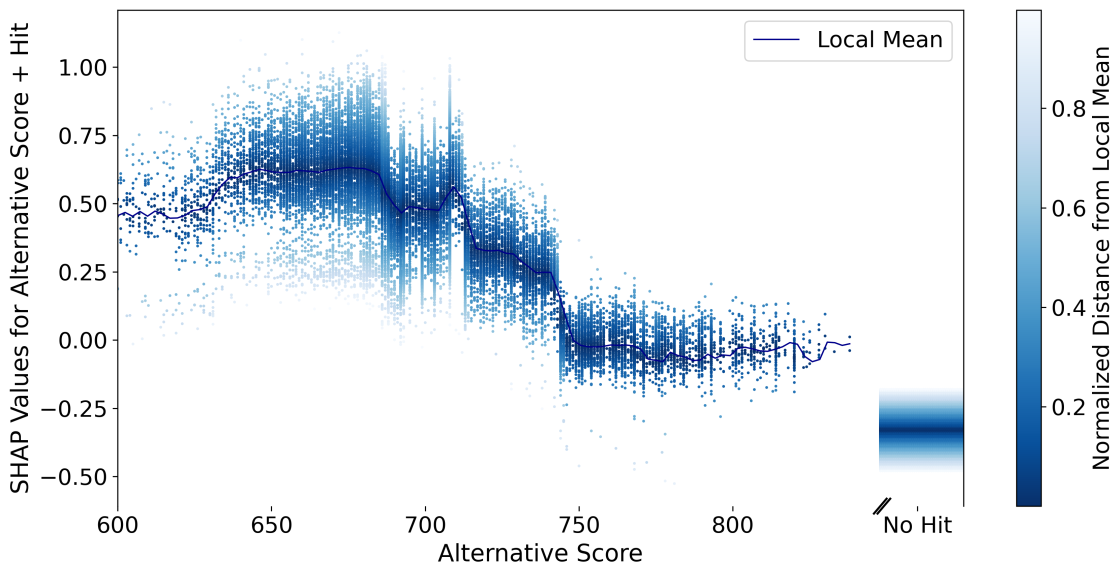


Figure 7. Relationship between Loan Approval, and Loan Rate and the Alternative Score for Ghosts. Panel A shows the binscatter of loan interest rate over the alternative score for the ghost sample, and Panel B shows a similar graph for the likelihood of getting a loan. Both panels control for the vantage score, average zip code income, and month-lender and county fixed effects. Panel A additionally controls for the loan amount and loan terms. The horizontal dashed line shows the average value for the population that does not hit the alternative data.

(a) SHAP values for Alt Score + Hit



(b) SHAP values for other variables

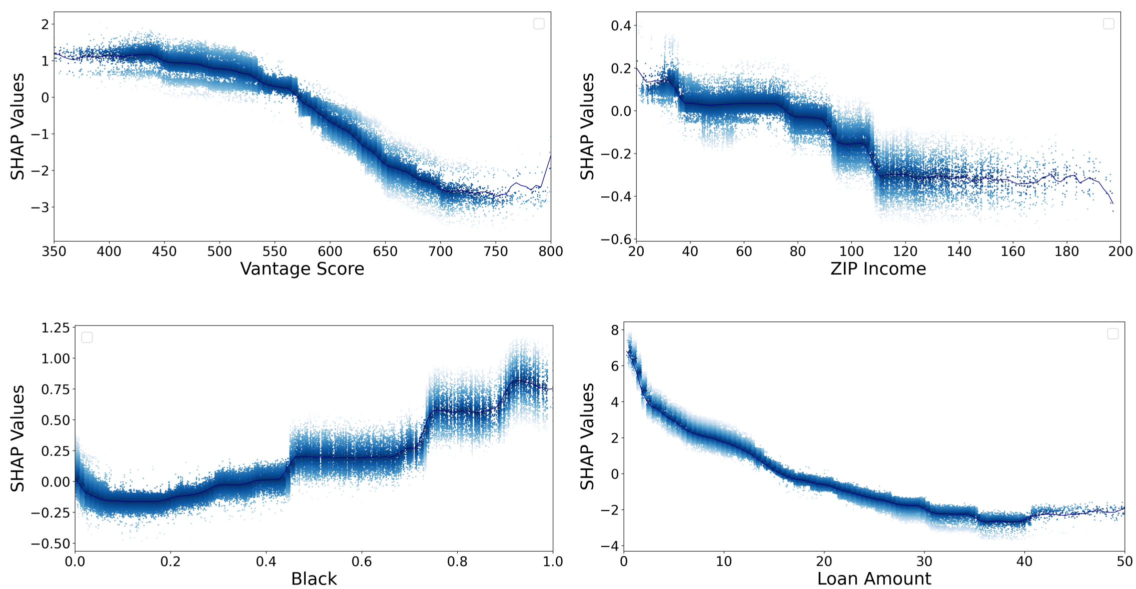


Figure 8. The Relationship between Alternative Score and Loan Rate Predictions. This figure shows how the alternative score, hitting the alternative data, and other borrower characteristics influence loan rate predictions from the selected XGBoost model. In panel A, the Y-axis shows the sum of SHAP values for hit and alternative score. The color intensity indicates the distance from the local mean SHAP value, with darker blues representing closer proximity to the mean. The dark blue line shows the local mean SHAP values. The figure displays values for borrowers without alternative data hits after the axis break. Panel B plots the SHAP values for Vantage score, zip code income, percentage of Black population in zip codes, and loan amount against their respective values.

(a) Linear Model **(b)** XGBoost Model

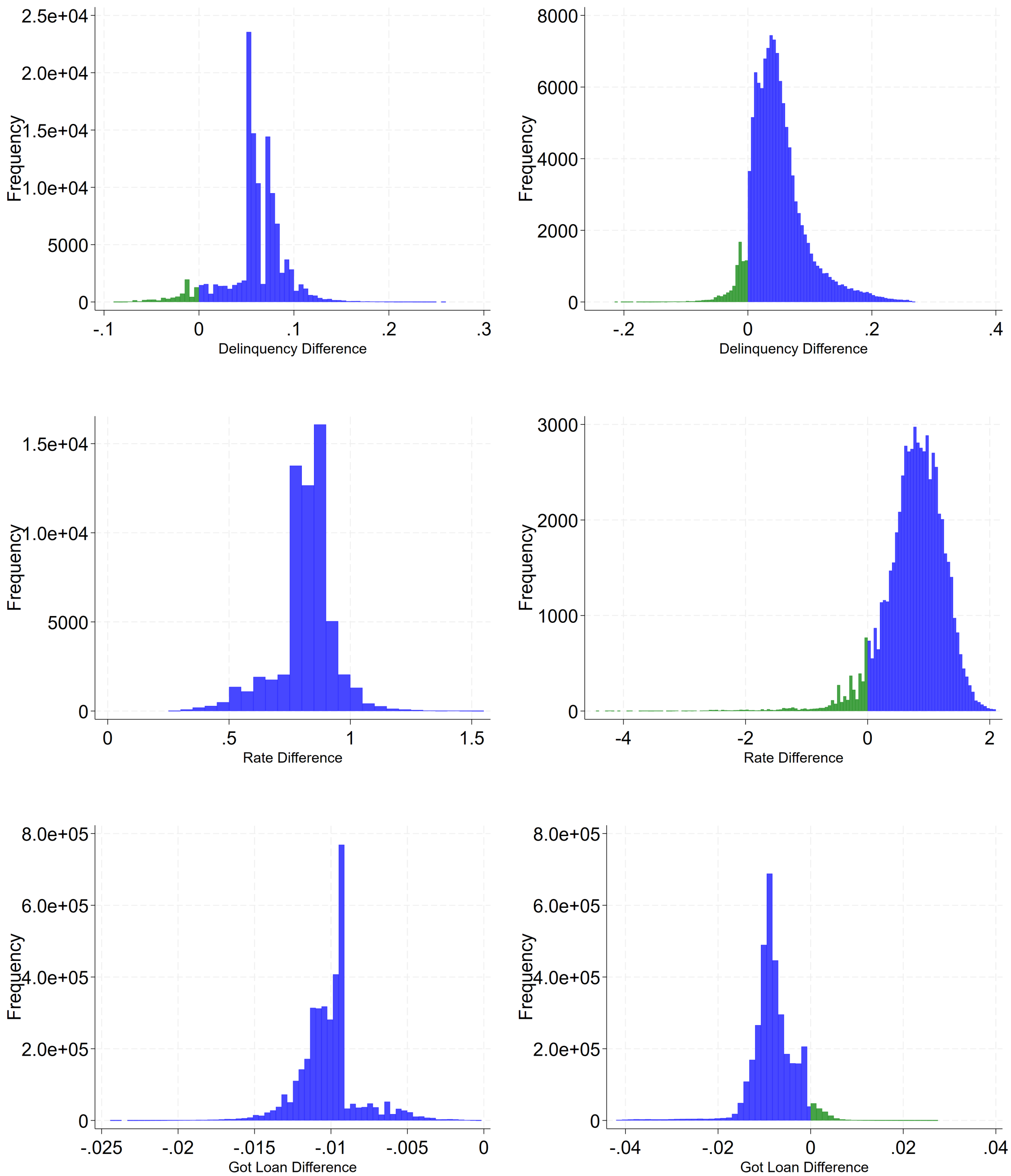


Figure 9. The Estimated Impact of Hit on Delinquency, Interest Rate, and Got Loan. This figure plots the difference between the predicted delinquency, loan rates, and loan origination probability for applicants who hit the data and their corresponding predicted values if they had not been hit, everything else constant. Column (a) shows the distribution of the rate difference for the OLS model and Column (b) for the XGBoost model.

Table I. Summary Statistics.

This table reports summary statistics for all the loan applications and loans in our sample starting from April 2009 to April 2017. The number of observations, mean, standard deviation, and the minimum and maximum value for each variable are reported for the applicants.

	N Obs	Mean	SD	Min	Max
Application Data					
Alt Score	8213863	723.61	29.13	500.00	838.00
Vantage 4	7978250	552.97	61.59	300.00	850.00
Hit	8213863	0.45	0.50	0.00	1.00
GotLoan	8213863	0.06	0.24	0.00	1.00
Lender 1	8213863	0.86	0.35	0.00	1.00
Lender 2	8213863	0.10	0.30	0.00	1.00
Lender 3	8213863	0.03	0.18	0.00	1.00
Lender 4	8213863	0.00	0.07	0.00	1.00
Lender 5	8213863	0.01	0.07	0.00	1.00
Loan Data					
Rate	255012	21.48	5.69	0.01	42.00
Loan Amount (K)	501579	17.22	6.58	0.00	91.86
Loan Term	501571	68.27	11.13	1.00	135.00
Delinquency	501579	0.39	0.49	0.00	1.00
Zip Code Demographic Data					
ZipIncome (K)	8213863	64.24	21.95	5.70	441.28
Black	8213863	0.23	0.25	0.00	1.00
Poverty100	8213863	0.09	0.05	0.00	1.00
Foodstamp	8213863	0.17	0.10	0.00	1.00
PDshops	8213863	1.59	2.13	0.00	17.00
HighSchool	8213863	0.17	0.05	0.00	1.00

Table II. Alternative Score and Delinquency.
This table estimates OLS regressions where the dependent variable is loan delinquency. Columns 4-6 restrict the sample to three different terciles of Vantage 4 score, Column 4 representing the lowest and Column 6 representing the highest Vantage score. All regressions include lender-month and county-fixed effects. Standard errors, reported in parentheses, are double-clustered by state and loan application year-month. * indicates 10% significance, ** indicates 5% significance, and *** indicates 1% significance.

		All		Low Vantage	Med Vantage	High Vantage
Dep Var:	Delinquency	Delinquency	Delinquency	Delinquency	Delinquency	Delinquency
AltScore STD	-0.0351***	-0.0300***	-0.0300***	-0.0274***	-0.0325***	-0.0305***
	(0.0018)	(0.0017)	(0.0016)	(0.0021)	(0.0024)	(0.0027)
Vantage 4 STD		-0.0658***	-0.0634***	-0.0468***	-0.0711***	-0.0577***
		(0.0017)	(0.0018)	(0.0047)	(0.0108)	(0.0039)
Loan Amount			-0.0013***	-0.0000	-0.0014***	-0.0019***
			(0.0003)	(0.0005)	(0.0004)	(0.0004)
Loan Term			0.0053***	0.0038***	0.0060***	0.0062***
			(0.0003)	(0.0004)	(0.0003)	(0.0004)
ZipIncome			-0.0008***	-0.0008***	-0.0008***	-0.0009***
			(0.0001)	(0.0001)	(0.0001)	(0.0001)
Dep Var Mean	.461	.461	.461	.533	.454	.373
Lender-Month FE	YES	YES	YES	YES	YES	YES
County FE	YES	YES	YES	YES	YES	YES
Observations	157057	156062	156055	56910	56152	41802
Adjusted R^2	.0732	.0852	.0919	.0858	.0736	.0745

Table III. Alternative Score and Loan Terms Before and After Adoption.
This table estimates OLS regressions where the dependent variables are the loan delinquency in Column (1), loan interest rate in Column (2), and loan origination probability in Column (3). All regressions include lender-month and county fixed effects. Standard errors, reported in parentheses, are double-clustered by state and loan application year-month. * indicates 10% significance, ** indicates 5% significance, and *** indicates 1% significance.

	(1)	(2)	(3)
Dep Var:	Delinquency	Rate	GotLoan
AltScore-Pre	-0.0230***	0.0363	
	(0.0024)	(0.0537)	
AltScore-Post	-0.0331***	-0.1760***	0.0024***
	(0.0019)	(0.0150)	(0.0002)
Post	0.0125	-0.1338	
	(0.0121)	(0.2193)	
Vantage 4 STD	-0.0633***	-0.5475***	0.0042***
	(0.0018)	(0.0562)	(0.0004)
Loan Amount	-0.0013***	-0.1834***	
	(0.0003)	(0.0086)	
Loan Term	0.0053***	-0.0054	
	(0.0003)	(0.0098)	
ZipIncome	-0.0008***	0.0003	0.0001***
	(0.0001)	(0.0009)	(0.0000)
Dep Var Mean	.461	22.3	.0332
Lender-Month FE	YES	YES	YES
County FE	YES	YES	YES
Observations	156055	85534	3576677
Adjusted R^2	.092	.382	.0855

Table IV. Alternative Data, Delinquency, and Loan Terms.

This table estimates OLS regressions where the dependent variables are the loan delinquency in Columns 1 and 2, loan interest rate in Columns 3 and 4, and loan origination probability in Columns 5 and 6. All regressions include lender-month and county fixed effects. Standard errors, reported in parentheses, are double-clustered by state and loan application year-month. * indicates 10% significance, ** indicates 5% significance, and *** indicates 1% significance.

	(1)	(2)	(3)	(4)	(5)	(6)
Dep Var:	Delinquency	Delinquency	Rate	Rate	GotLoan	GotLoan
Hit	0.0592***	0.0607***	0.8230***	0.8295***	-0.0099***	-0.0100***
	(0.0034)	(0.0034)	(0.0521)	(0.0524)	(0.0008)	(0.0008)
AltScore STD		-0.0312***		-0.1135***		0.0020***
		(0.0020)		(0.0123)		(0.0001)
Vantage 4 STD	-0.0698***	-0.0688***	-0.9281***	-0.9234***	0.0030***	0.0029***
	(0.0013)	(0.0013)	(0.0439)	(0.0441)	(0.0005)	(0.0005)
Loan Amount	-0.0004	-0.0004	-0.1597***	-0.1596***		
	(0.0004)	(0.0004)	(0.0054)	(0.0054)		
Loan Term	0.0063***	0.0063***	0.0179**	0.0179**		
	(0.0003)	(0.0003)	(0.0087)	(0.0087)		
ZipIncome	-0.0010***	-0.0010***	-0.0061***	-0.0061***	0.0000***	0.0000***
	(0.0001)	(0.0001)	(0.0008)	(0.0009)	(0.0000)	(0.0000)
Dep Var Mean	.405	.405	21.2	21.2	.0446	.0446
Lender-Month FE	YES	YES	YES	YES	YES	YES
County FE	YES	YES	YES	YES	YES	YES
Observations	350827	350827	160710	160710	7864056	7864056
Adjusted R^2	.0697	.071	.398	.398	.122	.122

Table V. Alternative Data, Delinquency, and Loan Terms for the Ghost Sample.
This table estimates OLS regressions where the dependent variable are the loan delinquency in Columns 1 and 2, loan interest rate in Columns 3 and 4, and loan origination probability in Columns 5 and 6. All regressions include lender-month and county fixed effects. Standard errors, reported in parentheses, are double-clustered by state and loan application year-month. * indicates 10% significance, ** indicates 5% significance, and *** indicates 1% significance.

	(1)	(2)	(3)	(4)	(5)	(6)
Dep Var:	Delinquency	Delinquency	Rate	Rate	GotLoan	GotLoan
Hit	0.0631***	0.0641***	0.7502***	0.7502***	-0.0016**	-0.0016**
	(0.0134)	(0.0132)	(0.1905)	(0.1898)	(0.0008)	(0.0008)
AltScore STD		-0.0247		-0.0005		0.0021***
		(0.0200)		(0.1522)		(0.0006)
Loan Amount	0.0059**	0.0059**	-0.1569***	-0.1569***		
	(0.0027)	(0.0026)	(0.0232)	(0.0232)		
Loan Term	0.0056***	0.0056***	0.0032	0.0032		
	(0.0014)	(0.0014)	(0.0246)	(0.0246)		
ZipIncome	-0.0012***	-0.0012***	-0.0069*	-0.0069*	0.0000	0.0000
	(0.0003)	(0.0003)	(0.0039)	(0.0040)	(0.0000)	(0.0000)
Dep Var Mean	.363	.363	23.4	23.4	.0207	.0207
Lender-Month FE	YES	YES	YES	YES	YES	YES
County FE	YES	YES	YES	YES	YES	YES
Observations	4388	4388	1872	1872	235399	235399
Adjusted R^2	.0436	.0439	.301	.301	.0171	.0171

Table VI. Alternative Score and Demographic Characteristics.

This table estimates OLS regressions where the dependent variables are hitting the alternative data in Columns 1-3, and the alternative score for the hit population in Columns 4-6. All independent variables are standardized. Standard errors are reported in parentheses. * indicates 10% significance, ** indicates 5% significance, and *** indicates 1% significance.

	(1)	(2)	(3)	(4)	(5)	(6)
Dep Var:	Hit	Hit	Hit	AltScore	AltScore	AltScore
ZipIncome	-0.0141***	-0.0104***	-0.0059***	1.1940***	1.1576***	0.8861***
	(0.0009)	(0.0009)	(0.0010)	(0.0734)	(0.0743)	(0.0840)
Minority		0.0296***			-0.2944***	
		(0.0008)			(0.0503)	
Black			0.0337***			-0.8025***
			(0.0008)			(0.0448)
American Indian			0.0059***			-0.0275
			(0.0007)			(0.0512)
Asian			-0.0006			0.4163***
			(0.0008)			(0.0595)
Hispanic			0.0002			0.1622***
			(0.0007)			(0.0576)
Other			0.0027***			0.4387***
			(0.0008)			(0.0534)
Dep Var Mean	.373	.373	.373	700	700	700
Observations	24477	24477	24477	24477	24477	24477
Adjusted R^2	.0117	.0639	.0787	.0164	.0173	.0283

Appendix for:
"Borrowers in the Shadows: The Promise and Pitfalls of Alternative Credit Data"

Table A.1. Relationship between Alternative Score and Loan Terms for Terciles of Vantage Score. This table estimates OLS regressions in the post-adoption period, where the dependent variables are the likelihood of loan approval in Columns 1-3 and loan interest rate in Columns 4-6 for three different terciles of Vantage 4 score, Columns 1 and 4 representing the lowest and Columns 3 and 6 representing the highest Vantage score. All regressions include lender-month and county fixed effects. Standard errors are double-clustered by state and loan application year-month. *t*-Statistics are reported in parentheses. * indicates 10% significance, ** indicates 5% significance, and *** indicates 1% significance.

	(1)	(2)	(3)	(4)	(5)	(6)
	GotLoan	GotLoan	GotLoan	Rate	Rate	Rate
AltScore STD	0.0027***	0.0023***	0.0018***	-0.1649***	-0.1095***	-0.2901***
	(0.0002)	(0.0003)	(0.0003)	(0.0203)	(0.0238)	(0.0335)
Vantage 4 STD	0.0050***	0.0030***	0.0020***	-0.4398***	-0.6533***	-0.8387***
	(0.0003)	(0.0009)	(0.0006)	(0.0760)	(0.1164)	(0.1010)
ZipIncome	0.0001***	0.0001***	0.0000**	0.0017	-0.0010	-0.0034*
	(0.0000)	(0.0000)	(0.0000)	(0.0016)	(0.0017)	(0.0019)
Loan Amount				-0.1561***	-0.1628***	-0.1868***
				(0.0089)	(0.0097)	(0.0106)
Loan Term				0.0050	0.0200	0.0408**
				(0.0131)	(0.0145)	(0.0177)
Dep Var Mean	.0279	.0355	.0395	23.3	22.2	20.7
Lender-Month FE	YES	YES	YES	YES	YES	YES
County FE	YES	YES	YES	YES	YES	YES
Observations	1504119	1199820	862093	22965	21521	16385
Adjusted R^2	.0756	.0863	.0971	.31	.345	.382