

Sharpening the Focus: Using Cash-Flow Data to Underwrite Financially Constrained Businesses

Empirical White Paper

*FinRegLab in collaboration with
Sabrina T. Howell and Siena Matsumoto,
New York University*

About FinRegLab

FinRegLab is a nonprofit, nonpartisan innovation center that tests new technologies and data to increase access to responsible financial services that help drive long-term economic security for people and small businesses. With our research insights, we facilitate discourse across the financial ecosystem to inform market practices and policy solutions.

Acknowledgments

This empirical white paper is part of a broader research project to evaluate ways to improve credit access for small businesses through the use of non-traditional data sources and mission-based lenders. Concurrent with this report, we are updating a qualitative study of mission-based lenders' adoption of cash-flow data and platform technology to increase lending to small businesses. Subsequent work with Emmanuel Yimfor of Columbia University will analyze loans issued by mission-based lenders that are using cash-flow data in their underwriting processes. Other reports in this series are available at:

<https://finreglab.org/research/innovations-for-underwriting-small-businesses>.

Support for this publication and other aspects of FinRegLab's research on innovations for underwriting underserved small businesses was provided by the US Department of Commerce, Minority Business Development Agency, Visa, and Plaid. Detailed information can be found on the inside back cover.

The authors of this report are Sabrina T. Howell (New York University Stern School of Business and the National Bureau of Economic Research), Siena Matsumoto (New York University Stern School of Business), and Kelly Thompson Cochran (FinRegLab).

We would like to acknowledge Paula Ochiel, Karla Renschler, and Zishun Zhao for their contributions to this report and the broader project



The accessible text version of this report provides simply formatted text, tables, charts, and figures for use with an accessibility screen reader. Document interactivity is more limited than in the main version of the report, which can be found at finreglab.org.

CONTENTS

1. Introduction	2
2. Data Sources And Full Sample Summary Statistics.....	6
2.1 Lending data sources	6
2.2 Summary statistics	7
3. Financial Constraint Proxies: Young Firms And Low Personal Credit Scores.....	9
3.1 Young firms	9
3.2 Low personal credit scores.....	12
4. Do Cash-Flow Variables Predict Default? Regression Analysis	15
5. Who Benefits From Cash-Flow Variables? Machine Learning Analysis	20
5.1 Machine learning model for default prediction	20
5.2 Application of Tail Analysis for Comparative Outcomes (TACO).....	24
6. Concluding Thoughts	27
References.....	28
APPENDIX A: Supplementary Tables and Figures	31
APPENDIX B: Geographic Analysis	34

1. INTRODUCTION

In underwriting small business loans, lenders commonly rely on the owner's personal credit score and the firm's time in business. These factors can help to predict loan defaults but can also make it more challenging for smaller and younger firms as well as economically disadvantaged entrepreneurs to access credit. This has important implications for local communities and the broader U.S. economy because small businesses play a vital role in job creation and wealth building. For example, at a time of significant economic uncertainty, small business lending practices could help determine the future of a surge of new businesses formed since the COVID-19 pandemic as they seek to bridge short-term gaps in cash flows and take advantage of expansion opportunities.

New developments in the use of alternative data may help lenders extend credit to small businesses which have historically been considered high risk. In this report, we explore the value of cash-flow variables drawn from recent bank statements in underwriting small businesses as a supplement to the owners' personal credit scores. We draw directly from the analysis developed in Hair et al. (2025), and are motivated by FinRegLab (2019a)'s work on cash-flow underwriting. Although the study relies on data from fintech lenders, the findings are relevant to banks and mission-based lenders as well.

Access to credit is a major determinant of small business survival and growth. It can be critical for bridging invoicing delays, unexpected expenses, and other short-term gaps, as half of small businesses only have cash reserves to cover about a month of operating expenses (Farrell and Wheat, 2016; Bartik et al., 2020). Credit access is also a major determinant in positioning small firms to invest in equipment and cover other expenses needed to support expansion, growth, and job creation (Fracassi et al., 2016; Herkenhoff et al., 2021; Brown and Earle, 2017; Consumer Financial Protection Bureau, 2017; Aktug et al., 2020; Delis et al., 2024).

However, two common lending practices make it difficult for certain small businesses to obtain loans. First, since the 2008 financial crisis, banks have increasingly imposed longer minimum time in business thresholds in order to be eligible for a loan (Federal Deposit Insurance Corporation, 2018, 2024; Mills and McCarthy, 2014; Chen et al., 2017). Second, lenders of all types rely heavily on business owners' personal credit scores, which more directly measures the past personal debt repayment behaviors of the owner than the current solvency of their business and tends to limit loan access to those with substantive and positive personal credit histories. Yet about one third of U.S. adults are estimated to have thin or no traditional credit files, and another 25 percent have subprime scores (Hepinstall et al., 2022; Equifax, 2022). This includes entrepreneurs who have substantially damaged their personal credit in the course of starting new businesses.

These lending practices and market dynamics are likely to have an especially significant impact on the fate of post-pandemic startups, given their short time in business and other characteristics. Research analyzing small business applications for tax identifier numbers suggests that the surge in formation has taken place outside of traditional entrepreneurial ecosystems such as Silicon Valley and the Boston corridor. This includes high numbers of applications in the Southeast, which tends to have larger concentrations of economically distressed communities and populations with lower credit scores than other parts of the country (McSwigan, 2022; Economic Innovation Group, 2023; Van Dam, 2023; Scott et al., 2025). Business ownership also surged among low-income, Black, and Hispanic families after 2019, all of whom are more likely to have lower credit scores as well (Edelberg and Steinmetz-Silber, 2024; Toh, 2024).

The recent application of electronic cash-flow data to underwriting may improve access to credit for younger businesses and entrepreneurs with low credit scores. Cash-flow underwriting incorporates variables from recent bank statements (or, less frequently, other sources of revenue and expenditure information) to assess the financial health and repayment ability of a business. It evaluates real-time financial activity—such as the timing and magnitude of deposits, withdrawals, and financial distress indicators—offering a more direct measure of a business’s ability to pay off debt. This distinction is particularly important for small firms, which often experience fluctuations in liquidity that are not reflected in traditional credit scores. The approach gained traction in the early 2010s as fintech lenders responded to market gaps left by traditional banks. While initially concentrated in riskier market segments, use of electronic cash-flow data has increasingly been adopted in mainstream lending; for example, credit bureaus and other companies rolled out several new cash-flow products in 2024 and 2025.¹ If cash-flow data improves the ability of lenders to accurately predict loan defaults while simultaneously expanding access for entrepreneurs who have promising firms yet struggle to access credit under traditional underwriting approaches, it may represent a win-win for lenders, borrowers, and the broader economy.

This paper examines the implications of incorporating recent cash flows in small business loan underwriting for younger firms, especially those started by entrepreneurs with low personal credit scores. We use data on small business loan applications, originations, and loan performance from two fintech (non-bank) lenders that extend credit to a substantially wider population than other small business lenders may be willing to underwrite. Our dataset includes information used in underwriting, such as credit scores, bank statement-based variables, and industry classification, as well as additional borrower characteristics, including the business zip code. We focus on the borrower population to examine how well traditional variables and cash-flow metrics predict subsequent loan performance. In total, our analysis dataset includes around 38,000 originated loans, spanning from August 2015 to May 2024.² The data are comprehensive from their respective sources, covering a broad range of small business borrowers across different industries and geographic regions.

We consider a set of specific bank statement variables, including revenues, withdrawals, balances, standard deviations of balances and credits, and distress indicators such as low or negative ending balances and insufficient funds transactions. Based on conversations with practitioners, these variables represent a common baseline among fintech lenders, though each lender generally creates its own model and may use more or less features. Importantly, any lender could use these data provided there is sufficient infrastructure to distill the information from traditional bank statements provided by the customer or to facilitate electronic access to the data with customer permission.³

When a firm has a project whose net present value is positive, yet struggles to access sufficient funds to finance it, we call the firm “financially constrained.” We look at two measures of financial constraints in our data, age of the firm and personal credit score of the owner, as well as the interaction between those factors. We define young firms as those with fewer than five years of operating

¹ See e.g., Crosman (2024) and Lawler (2024). Many of the products are focused in the consumer context, where some integrate bank statement attributes into scores, while others just provide features for lenders to incorporate into their own underwriting processes (Experian, 2024, 2025; VantageScore, 2024). Plaid announced a product in May 2025 that categorizes transactions in small business accounts, which could in turn be used to build underwriting variables. (Taylor and Sriram, 2025).

² There are no Paycheck Protection Program loans or otherwise subsidized or mission-driven loans in the sample. To address concerns about pandemic-related bias, we control for the calendar quarter of an application.

³ The United States has developed a robust infrastructure for such electronic data access over several decades though it has not been enshrined in federal regulations. Implementation of federal rules governing consumer- permissioned data transfers is scheduled to begin in 2026, but has been challenged in court and the Consumer Financial Protection Bureau is now seeking to vacate the rules. The regulations would not apply to business checking or transaction accounts, although small business lenders often rely on the same types of intermediaries to access such data. See FinRegLab (2019b, 2020) and Consumer Financial Protection Bureau (2024).

history and low score owners as having less than a 700 FICO score. Appendix B contains additional analyses based on the location of the small business.

We use three methods to explore how traditional credit scores and cash-flow variables differentially predict default for our constrained groups. First, we estimate ordinary least squares regressions predicting loan default as a function of both traditional credit scores and cash-flow variables, controlling for firm and loan characteristics. Across all models, cash-flow variables exhibit strong predictive power in economically intuitive directions. Businesses with stronger financial health—measured by higher bank credits and stable balances—are less likely to default, while indicators of financial distress—such as frequent low balances, higher withdrawals, and reliance on high-cost financing—are associated with higher default risk.

We split the sample by the constrained characteristics. In many cases, the magnitude of cash-flow variables' predictive power is larger for younger businesses and lower score entrepreneurs. For example, one standard deviation increase in balances (about \$64,000) is associated with more than a 2 percentage point (pp) lower likelihood of default for young firms but only a 1 pp lower likelihood for older firms. These differences are magnified when we focus on younger firms that also have low score owners, among whom the credit score becomes less predictive of default and the cash-flow variables become, in some cases, more so. This pattern suggests that lenders relying solely on historical credit scores may underestimate the repayment ability of these firms, while adding cash-flow data provides a more accurate assessment of financial strength.

To further evaluate the informativeness of cash-flow data, we employ random forest machine learning models to predict loan default, comparing a Baseline model using credit score, business age, and other limited features to a Cash-Flow model that additionally incorporates recent bank statement data. We assess performance with two standard classification metrics: ROC AUC (discrimination ability) and the H-measure (overall predictive performance).

We find that incorporating the cash-flow data leads to statistically significant improvements in classification performance across all borrower segments, with a 0.011 improvement in AUC in the overall sample. As we would expect, the performance improvements tend to be larger (0.015 AUC) among borrowers with low credit scores than the overall population, but the magnitude of the gains are further increased among low score owners whose businesses are young (0.023 AUC). These results are also substantially larger than for young firm only, suggesting that cash-flow data can be helpful in evaluating entrepreneurs who are facing both types of financial constraints simultaneously.

In sum, the regression and machine learning prediction exercises suggest that incorporating cash-flow data improves the accuracy of traditional underwriting models that are dependent on owners' personal credit scores. This can give lenders more confidence in underwriting younger firms and entrepreneurs with lower credit scores. Machine learning models incorporating cash-flow variables more effectively identify credit risk, especially among certain borrower segments that struggle to access credit under traditional underwriting methods.

Finally, we examine whether constrained groups have lower predicted default rates under models that incorporate cash-flow data as well as traditional credit scores. We employ a new method that assesses how models reallocate credit risk across groups. Tail Analysis for Comparative Outcomes (TACO), originally developed in Hair et al. (2025), identifies which groups experience the largest changes in predicted default probability when shifting from one model to another, in this case from the Baseline model to the Cash-Flow model. The TACO Ratio quantifies whether a group is disproportionately likely to benefit from cash-flow underwriting. TACO ratios above one mean that the model built with the new data source is more likely to indicate that default risk levels are

lower than predicted by other information, while ratios below one mean that the model is likely to detect risks that are not being picked up by other information sources. Importantly, the TACO results do not imply zero-sum losses and gains; just because one group benefits does not mean that the counterpart group loses out to the same degree.

Similar to the accuracy results, we would expect the TACO ratio for borrowers with low scores to exceed 1. In this case, the Cash-Flow model produces a TACO ratio of 3.626 for the low score population as a whole. We find very strong beneficial impacts for low score borrowers with young firms (7.573 ratio), and still substantial impacts for low score borrowers with mature firms (2.722 ratio). Again, comparisons to the ratio for young firms overall (1.735 ratio) underscore the potential magnitude of impacts that models incorporating cash-flow data could have on entrepreneurs who face constraints both because their businesses are young and because their personal credit scores are low.

Taken together, these results confirm that cash-flow underwriting systematically shifts risk classification in ways that may benefit groups that often struggle to access credit under traditional underwriting approaches. Across multiple empirical approaches—OLS regressions, machine learning models, and the TACO framework—we find that incorporating recent bank statement data improves risk assessment, particularly for younger businesses and low score borrowers. Importantly, these findings suggest that adding cash-flow data to traditional credit scores can give lenders more confidence in lending to businesses that have historically been viewed as higher risk and more economically distressed.

In this paper, we first describe the data we use in Section 2 and then discuss the different types of financial constraints faced by younger businesses and low score entrepreneurs in Section 3. We then provide ordinary least squares results in Section 4 and machine learning informativeness and benefits in Section 5. We then conclude.

2. DATA SOURCES AND FULL SAMPLE SUMMARY STATISTICS

This section first describes the data we use in analysis (Section 2.1), then discusses summary statistics about the loans and firms in our sample (Section 2.2).

2.1 Lending data sources

We employ data from two non-bank online lenders. To maintain confidentiality, these firms have requested anonymity. Each exclusively serves U.S. small businesses and, like many other fintech entrants, uses electronic cash-flow data derived from bank statements in underwriting. They provided us with detailed loan application, approval, origination, and performance data, which are comparable across the two lenders in terms of the borrower characteristics they assess and the types of loans they issue. As neither lender is a traditional bank, they rely on debt vehicles rather than deposits for funding. Additionally, both require a personal guarantee and impose a blanket UCC lien on business assets. While their loans are technically collateralized—secured by non-revenue and non-real estate business assets—collateral verification is not conducted, nor are loans tied to specific assets. This contrasts with traditional real estate lending models, where verified collateral value is a key part of the underwriting process.

We refer to these lenders as “Lender A” and “Lender B.” Lender A offers both term loans and lines of credit.⁴ Its underwriting is highly automated and incorporates a risk scoring model. In contrast, Lender B only issues term loans and follows a more manual underwriting process, where a dedicated underwriter evaluates each case. Lender A’s thresholds are a 600 personal credit score and one year in business. Lender B’s thresholds are a 660 credit score and one to two years in business depending on the product. This means that we cannot study the impact of cash-flow data on applicants below the minimum cutoffs, although analyzing the predictiveness of different data elements against loan performance in the originated sample is still informative, especially since the fintechs extend credit to a substantially broader range of borrowers than many other lenders.

We obtained three types of data from both companies: applications, third party underwriting inputs, and loan characteristics. We describe each in turn. First, we have information about the applicant firm and owner, some of which is used in underwriting and some of which is not. This is collected from the firm directly, such as owner first name, business zip code, owner age, firm founding date, and industry.⁵ The owner’s first name allows us to construct a gender indicator to evaluate the general representativeness of the sample, though it is not a primary focus of analysis. We also observe underwriting inputs collected from third parties, specifically the FICO score from a credit bureau and bank statement information, which comes from the third party company Plaid for Lender A and third party company Oculous for Lender B. In the latter case, we employ the original JSON files containing bank statement variables that are gathered from the Oculous API. Lender A acquires three months of bank statements, while Lender B acquires six months. We understand that three months is a common industry practice and fintech lenders rarely acquire more than six months.

⁴ Loan type is included as a control in our analysis, though it does not materially affect the results.

⁵ Industry is in the form of NAICS codes. These are aggregated to 17 NAICS industries appearing in our data. We group sectors “Management of Companies and Enterprises” and “Administrative and Support and Waste Management and Remediation Services” with “Professional, Scientific, and Technical Services”.

Finally, for both companies, we observe whether a loan was approved, information on the offered loan terms—amount, maturity, and interest rate—and whether the loan was originated (i.e., taken up). There are a total of 104,150 applications and 18,434 loans for Lender A, spanning 11/19/2013 to 6/25/2024. For Lender B, there are a total of 58,668 applications and 25,762 loans, spanning 8/15/2015 to 5/20/2024. Among the 44,196 originated loans, we exclude those that are too new to have a status. This leaves us with 38,021 loans spanning from 2/13/2015 to 1/19/2024 that have been either paid off or considered to be non-performing (charged off, more than 60 days past due, or the borrower has received forbearance and a modified loan).

2.2 Summary statistics

This section presents summary statistics for the sample of originated loans from Lenders A and B in Table 1, focusing on borrower characteristics, loan performance, and financial health.

LOAN PERFORMANCE. The first set of variables focus on loan performance and terms. We construct an indicator variable for whether the loan is “non-performing” or not. This takes a value of one if the loan is charged off, more than 60 days past due, or the borrower has received forbearance and a modified loan. For parsimony, we refer to “non-performing” as “default.” We observe that 17% of loans default. The average interest rate among originated loans from Lenders A and B is 16%.⁶

FICO AND BANK STATEMENTS. Borrowers in our sample generally have strong credit scores reflecting early-stage screening at Lender B, with an average FICO of 728, but significant financial variation emerges from bank statement data. Bank statement variables are averaged across the three months prior to application. Credits (i.e., inflows) have a mean of roughly \$130,000 for borrowers. Withdrawals are higher than overall credits at an average of around \$170,000. The average balance in the account is around \$42,000 for borrowers. Three variables may signal distress. One is the number of overdrafts or insufficient funds (NSF) transactions, which occur when the balance falls below zero. Some banks impose an insufficient funds fee and reject the transaction, while others permit the account to go negative and impose an overdraft fee. The second is the number of low or negative ending balances, which occur when the balance goes below zero or below \$1,000. Consistent with strong selection on these variables, the averages are about five times higher among applicants to Lenders A and B than among borrowers. The third is the number of daily pay loans, which are merchant cash advances (MCAs) that typically have very high interest rates and are reflected in the bank statement by daily withdrawals transferred to the MCA lender. We also observe the standard deviation of credits and balances, which shed light on operational volatility.

BUSINESS CHARACTERISTICS. We observe key dimensions from an underwriting perspective beyond financial information. For the CEO or primary owner, we observe applicant first name and ages. This permits us to construct an indicator for gender based on first names. The sample is predominantly male (77%), aligning with broader trends in small business ownership. The average owner is 49 years old. Looking at all new companies in the U.S., Azoulay et al. (2020) find an average owner age of 42 years at the time of founding. Since many of our companies are not new, it is natural that the average age is somewhat higher. We also observe firm employment, which serves as a proxy for firm size. Borrower firms have 10 employees and are 11 years old on average. The firms operate in diverse industries, with the highest concentration in Retail Trade.

ZIP CODE LEVEL CHARACTERISTICS. Zip code level characteristics in our originated loan sample seem to be generally similar to national averages in Table A.1. However, the median per capita

⁶ There are two types of loans: Lines of credit and term loans. These compose 41% and 59% of the loans, respectively. In our analysis, we combine them and include a control for loan type.

income in originated loan zip codes (\$35,000) is higher than the national median (\$29,000). Similarly, the percentage of residents with a bachelor's degree is 35% in loan applicant zip codes compared to the national average of 21%. At the same time, zip codes originated loan sample have higher median population shares of Black, Hispanic, and Asian households than the national median and home ownership rates are slightly lower, with median home ownership of 68% relative to a national median of 75%. This aligns with studies showing that higher levels of education and income lead to higher rates of self-employment and longer-lasting businesses (Bates, 1990; Block et al., 2013).

TABLE 1: SUMMARY STATISTICS

This table contains summary statistics about originated loans using data from Lender A and Lender B (N = 38,021).

	N	MEAN	MEDIAN	SD
Loan Variables:				
Requested Loan Amount (Th\$)	38,021	109	80	98
APR (%)	36,252	16	15	5.11
Non-Performing Loan (%)	38,021	17		
Originated Loan Amount (Th\$)	38,021	115	78	104
Loan Maturity (Years)	38,021	3.15	3.00	1.41
Credit Score & Cash-Flow (Bank Statement) Variables:				
FICO	38,021	728	726	49
Credits (Th\$)	38,021	131	73	170
Balance (Th\$)	38,021	42	20	64
(#) Insuff. Funds	38,021	0.05	0.00	0.25
(#) Low or Neg. Bal.	38,021	0.42	0.00	1.06
Withdrawals (Th\$)	20,353	168	93	209
11 (Daily Pay Loan)	23,950	0.06		
S.D. Credits (Th\$)	38,021	13	7.63	14
S.D. Balance (Th\$)	38,021	5.86	3.39	6.15
Borrower Characteristics:				
Female	37,820	0.23		
Owner Age	20,190	49	48	11
Business Characteristics:				
Business Age (Years)	38,021	11	7.71	7.64
Number of Employees	38,021	10	6.00	13
Professional Services Industries	38,021	0.33		
Capital Intensive Industries	38,021	0.27		
Retail, Food, Healthcare & Other Service Industries	38,021	0.40		
Zip Code Level Variables:				
Per Capita Income (Th\$)	37,950	39	35	18
Pct Black Pop (%)	38,021	11	5.30	16
Pct Hisp. Pop (%)	38,021	18	10	19
Pct Asian Pop (%)	37,951	6.62	3.40	9
Pct Bachelors (%)	37,951	38	35	18
Pct Home Owner (%)	37,943	65	68	19
Pct Unemployed (%)	37,944	5.01	4.50	2.54
Density (Pop/km2)	37,725	1,050	481	2,992

3. FINANCIAL CONSTRAINT PROXIES: YOUNG FIRMS AND LOW PERSONAL CREDIT SCORES

Small businesses often face financial constraints that limit their ability to access credit, but these challenges are more pronounced for younger firms and entrepreneurs with lower personal credit scores since traditional lending models often rely on both factors in assessing default risk. In this section, we examine why these groups experience greater financing challenges under traditional underwriting models, and why they might benefit from cash-flow underwriting.

3.1 Young firms

Young businesses play a critical role in job creation, innovation, and regional economic growth (Haltiwanger et al., 2016; Decker et al., 2014; Berger and Udell, 1998; Crews et al., 2020; Hyatt, 2022). Compared to more established firms, young firms disproportionately drive net employment gains and contribute to economic dynamism (Decker et al., 2016b). Moreover, small business ownership has long been a key avenue for upward mobility in the U.S. (Audretsch, 2002). However, this dynamism has slowed somewhat in recent decades, with declines in formation rates and the share of young firms in the economy, their pace of job creation, and the share of activity for which they account (Decker and Haltiwanger, 2023; Ewing Marion Kauffman Foundation, 2019). The causes for this slowdown in recent years remain an open question, with financial frictions considered one possible contributor (Decker et al., 2016a).

These historical patterns are one of the reasons that observers are watching the post-pandemic increases in business formation carefully to see whether the surge can be sustained and further expanded. In sharp contrast to the aftermath of the 2008 financial crisis, small businesses rebounded faster and in greater numbers than before the pandemic. After an initial wave of business closures, a combination of factors—including government relief and stimulus funding, substantial shifts in consumer behavior that created new market opportunities, and loss of full-time employment—contributed to record-breaking levels of entrepreneurial activity through 2023 (Edelberg et al., 2023; Sedlacek and Shi, 2024; Decker and Haltiwanger, 2023; Fikri and Newman, 2024). While full data is not yet available for subsequent time periods, business applications for taxpayer identification numbers have continued at high rates, with a total of 21 million applications between 2020 and 2025 (U.S. Census Bureau, 2025). This figure is remarkable considering there were only 30 million small businesses in the U.S. in 2019 (U.S. Small Business Administration, 2019).

However, securing external financing is often a substantial constraint for young firms (Robb and Robinson, 2014; Nanda and Phillips, 2023). Without an established financial track record, they are perceived as riskier by lenders, leading to higher denial rates, stricter loan terms, or outright exclusion from formal credit markets (Cook et al., 2023; Fairlie et al., 2022). Part of lenders' reluctance stems from statistics showing that nearly half of new businesses close within five years (Bureau of Labor Statistics, 2024), although difficulty in obtaining financing may contribute to that pattern. Credit constraints tend to be severe for smaller and younger businesses during economic downturns and can have negative impacts on both firm growth and job creation/destruction (Dinlersoz et al., 2019; Hyatt, 2022).

Bank surveys confirm that years in business is a primary screening criterion for small business loan applications and that many banks impose minimum thresholds that effectively exclude younger firms (Federal Deposit Insurance Corporation, 2018). FDIC survey data shows that 34 percent of banks do not lend to startups, rising to about 45 percent among institutions (including larger banks) that base their small business lending programs primarily on credit scores and other credit bureau information. Even among lenders that do not have a specific minimum threshold—including both bank and non-bank lenders—underwriting models often accord substantial weight to both years in business and entrepreneurs' personal credit scores. (Federal Deposit Insurance Corporation, 2018, 2024). Moreover, young firms that do qualify often receive smaller loan amounts at higher interest rates due to their perceived lack of creditworthiness. These constraints can hinder their ability to bridge short-term liquidity gaps, invest in expansion, and sustain operations during early volatility.

Given these challenges, alternative credit assessment methods, such as cash-flow underwriting, could improve credit access for young firms by offering a real-time measure of financial health. Unlike traditional credit scores—which primarily reflect past borrowing behavior by the entrepreneur or the business, depending on whether a personal or commercial score is used—cash-flow variables provide a more direct measure of the business's liquidity, revenue stability, and payment capacity. This distinction is important for younger businesses, which may demonstrate strong financial performance but lack sufficient credit history to generate a commercial credit score or to qualify based on their owner's personal score under traditional models. Prior research suggests that integrating cash-flow data into underwriting models could help mitigate information gaps and improve lending decisions, especially for borrowers challenged by conventional credit assessment methods (FinRegLab, 2019a; Hair et al., 2025).

To evaluate how these dynamics affect small business lending, we define Young Firms as those with fewer than five years of operating history at the time of loan application. This threshold is motivated by empirical research showing that financial volatility is highest in the first few years after business formation, with liquidity constraints playing a major role in firm survival (Farrell et al., 2018; Robb and Robinson, 2014). Since nearly half of new businesses close within the first five years, this cutoff is a natural way to classify new and emerging small businesses.⁷

Young firms differ significantly from mature businesses in loan size, owner characteristics, and financial constraints. Table 2 compares loan characteristics, financial health, and borrower demographics for young firms (< 5 years old) and mature firms (≥ 5 years old). The data reveals substantial differences in loan terms, credit access, and financial stability between these groups.

⁷ While many banks use two- or three-year thresholds, we were unable to conduct a separate analysis at lower cutoffs due to limited sample size.

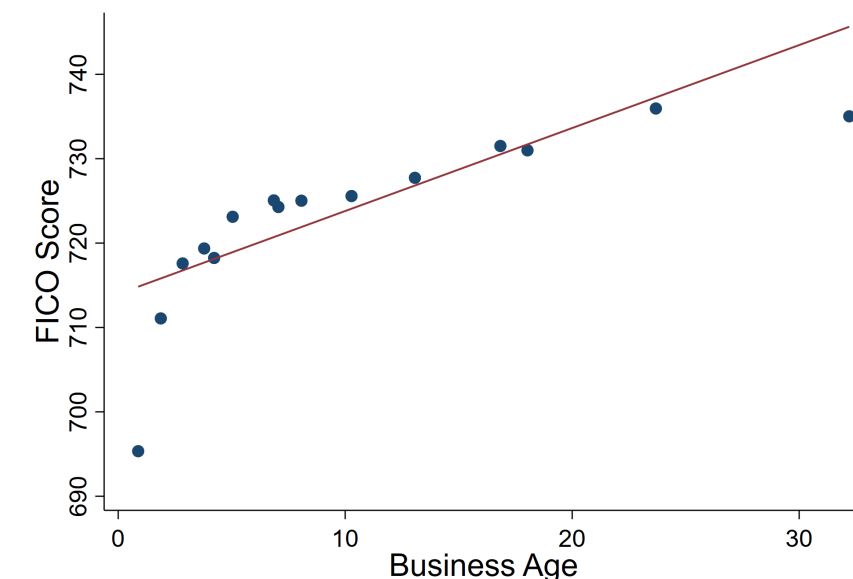
TABLE 2: SUMMARY STATISTICS BY FIRM AGE

This table compares Young Firms (< 5) with Mature Firms (≥ 5) in the sample of originated loans from Lender A and Lender B (N = 38,021).

	YOUNG FIRM (< 5)				MATURE FIRM (≥ 5)			
	N	MEAN	MEDIAN	SD	N	MEAN	MEDIAN	SD
Loan Variables:								
Requested Loan Amount (Th\$)	9,870	91	75	82	28,151	115	90	103
APR (%)	9,216	17	16	5.19	27,036	16	15	5.07
Non-Performing Loan (%)	9,870	18.9			28,151	15.9		
Originated Loan Amount (Th\$)	9,870	93	60	86	28,151	123	88	109
Loan Maturity (Years)	9,870	2.90	2.50	1.35	28,151	3.24	3.00	1.42
Credit Score & Cash-Flow (Bank Statement) Variables:								
FICO	9,870	725	723	50	28,151	729	727	49
Credits (Th\$)	9,870	99	56	134	28,151	142	80	179
Balance (Th\$)	9,870	35	17	53	28,151	44	21	67
(#) Insuff. Funds	9,870	0.05	0.00	0.26	28,151	0.05	0.00	0.25
(#) Low or Neg. Bal.	9,870	0.44	0.00	1.09	28,151	0.41	0.00	1.05
Withdrawals (Th\$)	4,445	123	69	164	15,908	180	101	218
1 (Daily Pay Loan)	6,769	0.05			17,181	0.06		
S.D. Credits (Th\$)	9,870	11	6.42	12	28,151	14	8.19	14
S.D. Balance (Th\$)	9,870	5.14	2.90	5.70	28,151	6.11	3.59	6.28
Borrower Characteristics:								
Female	9,811	0.24			28,009	0.22		
Owner Age	6,120	44	42	10	14,070	51	51	11
Business Characteristics:								
Business Age (Years)	9,870	2.99	3.00	1.15	28,151	13	12	7.08
Number of Employees	9,870	8.11	4.00	11	28,151	11	7.00	13
Professional Services Industries	9,870	0.29			28,151	0.34		
Capital Intensive Industries	9,870	0.29			28,151	0.26		
Retail, Food, Healthcare & Other Service Industries	9,870	0.42			28,151	0.40		
Zip Code Level Characteristics:								
Per Capita Income (Th\$)	9,853	39	35	17	28,097	40	35	18
Pct Black Pop (%)	9,870	12	5.70	16	28,151	11	5.20	16
Pct Hisp. Pop (%)	9,870	18	10	19	28,151	17	10	19
Pct Asian Pop (%)	9,854	6.84	3.40	10	28,097	6.55	3.40	8.87
Pct Bachelors (%)	9,854	38	35	18	28,097	38	36	18
Pct Home Owner (%)	9,854	64	67	19	28,089	65	68	19
Pct Unemployed (%)	9,854	5.01	4.50	2.47	28,090	5.00	4.50	2.56
Density (Pop/km2)	9,823	1,075	500	2,320	27,902	1,041	474	3,196

FIGURE 1: RELATIONSHIP BETWEEN CREDIT SCORE AND AGE OF FIRM

This figure uses applicant data from Lender A and Lender B ($N = 162,818$) to show the relationship between entrepreneurs' personal FICO scores and the age of the business in years. The plot contains a binscatter with 20 equal-sized age bins and a line of best fit.



Young firms tend to have younger owners (44 vs 51), aligning with findings from Hair et al. (2025) that younger entrepreneurs are more financially constrained and benefit more from incorporating cash-flow variables into credit models. Their slightly lower average FICO scores (725 vs. 729) may reflect this age difference rather than inherently riskier financial behavior. Young firms also tend to be smaller (8 employees vs. 11 for mature firms) and are less concentrated in professional services industries, suggesting differences in industry composition.

On average, young firms both request and receive (\$93K vs. \$123K) smaller loans with shorter maturities (2.90 vs. 3.24 years) and higher default rates (18.9% vs. 15.9%), reinforcing their constrained financial position. While they have lower credits and balances, they also show lower withdrawals, consistent with smaller-scale operations and lower loan requests. Despite these differences, indicators of credit distress—daily pay loans, low or negative balances, and insufficient funds—are similar across young and mature firms. Daily pay loans are slightly lower (0.05 vs 0.06), although the number of low or negative balances is slightly higher (0.44 vs 0.40). Notably, young and mature firms do not differ in zip code-level characteristics, suggesting that these financial constraints are distinct from those driven by geographic and demographic factors.

3.2 Low personal credit scores

In part because commercial credit scoring systems tend to focus on larger businesses with more established footprints, lenders have come to rely heavily on owners' personal credit scores in recent decades as an input to small business lending models. This is particularly true where lenders have worked to move away from labor-intensive judgmental or relationship underwriting systems toward the use of standardized predictive models.

For example, bank surveys find that personal credit scores are the most commonly considered types of information for small business underwriting, with more than 80 percent of bank lenders considering such scores for most or all of their loans. Large banks rely heavily on such information for smaller loans, with 97 percent reporting that they evaluate personal credit scores and 59 percent ranking them as the most important factor compared to other types of information such as the firm's financial position or collateral. In contrast, while 87 percent of small banks also consider personal scores for smaller loans, only 11 percent rank them as the most important factor. (Federal Deposit Insurance Corporation, 2024, 2018).

However, while the consumer credit reporting system is more standardized and broad-based than commercial credit reporting options, it is still subject to certain data gaps and limitations. About one third of U.S. consumers are estimated to have no or "thin" credit files, with about 20 percent of adults being unable to be scored under the most widely used models. About 25 percent of consumers have credit scores in the subprime range, which may cause applicants to be rejected or charged substantially higher rates. Information gaps are higher and average scores are lower among lower income households, Black and Hispanic consumers, and consumers under age 25 (Toh, 2024; Hepinstall et al., 2022; Equifax, 2022; Brevoort et al., 2016).

Scarcity of data, higher rates of data errors from collections items, and other "noisy" data patterns also tend to make traditional credit scores less accurate in predicting default risk for low-income consumers, minorities, and consumers with thin credit files. (Blattner and Nelson, 2024; Federal Trade Commission, 2012; Avery et al., 2004). Low credit scores can be a challenge for small business owners who have damaged their personal credit in an earlier unsuccessful launch or during the early years of standing up a new business, when they are most likely not to take a regular salary and to draw on personal savings or credit to try to keep both their households and companies afloat (Federal Reserve Board, 2017; Hwang et al., 2019; Chava et al., 2023).

More broadly, while personal scores can provide a sense of the entrepreneur's past financial history and habits, they are not based directly on information about the current finances of the business. Historically, lenders often collected both substantial financial documentation and "soft" information through their employees to gauge whether small businesses were likely to succeed in paying a new loan, but this can be both time and labor intensive to gather and analyze. While some small banks and mission-based lenders continue to rely more heavily on soft information, large banks and fintechs tend to emphasize "hard" financial data and increasingly to look for sources that can be gathered quickly through electronic means (Federal Deposit Insurance Corporation, 2024; FinRegLab, 2025). For the same reasons described above, cash-flow data from small businesses' transaction accounts that can be accessed electronically is appealing for both efficiency and performance reasons.

We define **Low FICO** as an entrepreneur having a score below 700. Firms owned by such entrepreneurs in the sample differ from firms owned by entrepreneurs with higher scores with regard to loan size and financial constraints as shown in Table 3, but are not substantially differentiated based on borrower, business, and geographic characteristics. The data reveals substantial differences in loan terms, credit access, and financial stability between these groups.

The differential in average FICO scores is 80 points between the low- and high-score group (672 vs. 752). On average, low score firms both request and receive somewhat smaller loans (\$108K vs. \$118K), with higher interest rates (18% vs. 15%) and somewhat longer loan terms (3.30 vs. 3.09 years). The differential in default rates between the two groups is substantial (22.3% vs. 14.4%). Cash-flow metrics also show financial differentials between the two groups, for example in balances, insufficient funds, low or negative balance patterns, and daily pay loans. However, differences are relatively small with regard to owner characteristics, the age of firm and number of employees, and business zip code.

Finally, we consider **Low FICO & Young Firms** as those meeting the thresholds for both low FICO (< 700) and firm age (< 5 years). These firms face multiple constraints, making them strong candidates for underwriting models based on real-time financial performance.

TABLE 3: SUMMARY STATISTICS BY CREDIT SCORE

This table compares Low FICO entrepreneurs (< 700) with High FICO entrepreneurs in the sample of originated loans from Lender A and Lender B (N = 38,021).

	LOW FICO (< 700)				HIGH FICO (≥700)			
	N	MEAN	MEDIAN	SD	N	MEAN	MEDIAN	SD
Loan Variables:								
Requested Loan Amount (Th\$)	11,343	100	75	91	26,678	112	85	101
APR (%)	11,074	18	17	5.20	25,178	15	15	4.89
Non-Performing Loan (%)	11,343	22.3			26,678	14.4		
Originated Loan Amount (Th\$)	11,343	108	71	97	26,678	118	82	107
Loan Maturity (Years)	11,343	3.30	3.00	1.38	26,678	3.09	3.00	1.41
Credit Score & Cash-Flow (Bank Statement) Variables:								
FICO	11,343	672	676	22	26,678	752	747	36
Credits (Th\$)	11,343	114	67	144	26,678	138	76	179
Balance (Th\$)	11,343	32	16	50	26,678	46	22	68
(#) Insuff. Funds	11,343	0.07	0.00	0.28	26,678	0.05	0.00	0.24
(#) Low or Neg. Bal.	11,343	0.56	0.00	1.21	26,678	0.36	0.00	0.99
Withdrawals (Th\$)	6,746	142	83	176	13,607	181	99	222
1 (Daily Pay Loan)	6,039	0.07			17,911	0.05		
S.D. Credits (Th\$)	11,343	12	6.87	12	26,678	14	8.01	14
S.D. Balance (Th\$)	11,343	5.10	2.92	5.58	26,678	6.18	3.63	6.35
Borrower Characteristics:								
Female	11,263	0.24			26,557	0.22		
Owner Age	5,058	48	47	11	15,132	49	49	11
Business Characteristics:								
Business Age (Years)	11,343	10	7.02	7.27	26,678	11	8.06	7.78
Number of Employees	11,343	9	5.00	12	26,678	10	7.00	13
Professional Services Industries	11,343	0.34			26,678	0.32		
Capital Intensive Industries	11,343	0.29			26,678	0.26		
Retail, Food, Healthcare & Other Service Industries	11,343	0.37			26,678	0.42		
Zip Code Level Characteristics:								
Per Capita Income (Th\$)	11,322	39	34	18	26,628	40	36	18
Pct Black Pop (%)	11,343	12	5.60	16	26,678	11	5.10	15
Pct Hisp. Pop (%)	11,343	18	10	20	26,678	17	10	19
Pct Asian Pop (%)	11,322	6.13	3.10	8.51	26,629	6.83	3.50	9
Pct Bachelors (%)	11,322	37	34	18	26,629	38	36	18
Pct Home Owner (%)	11,320	65	67	18	26,623	65	68	19
Pct Unemployed (%)	11,322	5.05	4.50	2.62	26,622	4.99	4.50	2.50
Density (Pop/km2)	11,231	981	471	3,031	26,494	1,078	485	2,976

4. DO CASH-FLOW VARIABLES PREDICT DEFAULT? REGRESSION ANALYSIS

Our first analysis explores the direction and magnitude of cash-flow variables' predictive power over default using ordinary least squares and logistic regression models.⁸ If an allowed variable significantly predicts default, it is expected to be incorporated into underwriting decisions. In addition, the R^2 of the regressions, which quantifies how much of the variation in the outcome is explained by the independent variables, sheds some light on informativeness. We show that cash-flow variables often have more explanatory power for constrained groups than for other populations, and their informativeness does not entirely overlap with the information contained in FICO scores.

We predict default (i.e., a non-performing loan) among originated loans at Lenders A and B using variants of the model in **Equation 1**:

$$1(\text{Non-Performing}_i) = \delta \text{FICO}_i + \beta'_1 \text{Cash-Flow Chars}_i + \gamma'_1 \text{Emp}_i + \gamma'_2 \text{Loan Type}_i + \text{Industry}_i + \text{State}_i + \alpha_t + \varepsilon_i.$$

The coefficients of interest are the FICO and cash-flow variables. To make interpretation straightforward, we standardize to z-scores, so that the coefficient represents the effect of a one standard deviation change in the independent variable on the outcome variable.⁹ We include controls for characteristics that the lenders observe and that are permissible to use in underwriting under fair lending laws. These are firm size (number of employees), as well as fixed effects for the applicant's industry, state, calendar quarter of application (we observe 36 quarters), and loan type. We cluster standard errors by industry and quarter.

Low FICO borrowers represent a financially constrained group that cuts across all business and geographic categories. While much of our analysis focuses on business age, one of our key findings is that entrepreneurs with low personal credit scores stand to benefit substantially from the inclusion of cash-flow variables in underwriting models. This finding is logical, as any additional relevant information that is uncorrelated with a traditional credit score will help to predict default more accurately than the score alone. Our analysis attempts to measure the magnitude of this performance gain, helping to determine what types of information can be relevant and timely for predicting default and providing a baseline for evaluating effects on entrepreneurs who face additional financial constraints because their firms are young.

We first compare firms on the basis of their owners' credit scores in Table 4. Odd-numbered columns show baseline regressions that rely heavily on FICO, while even-numbered columns also incorporate cash-flow variables. Columns 1–2 and 5–6 restrict the sample to low FICO owners (< 700), while the remaining columns consider high FICO owners. Finally, columns 5–8 are different from 1–4 because they include the full set of fixed effects described in Equation 1.

⁸ Logistic regression results are presented in Appendix A. Lenders have frequently relied upon logistic regression in building traditional predictive underwriting models, in part because it has the advantage of being transparent and interpretable while being fairly competitive with regard to predictive power. However, some lender segments are increasingly migrating to supervised machine learning models, which are discussed further below.

⁹ For any missing values of the cash-flow variables that are not fully populated, we impute the mean and include a binary indicator to account for missingness in the model.

TABLE 4: PREDICTING DEFAULT FOR LOW FICO OWNERS (< 700)

This table uses data from Lender A and Lender B (N = 38,021) to show how credit score, cash-flow, and borrower characteristics predict default conditional on origination. Low FICO is defined as an owner having a FICO score below 700. All bank variables and FICO score are standardized to z-scores and can be interpreted as the change in the dependent variable from 1 standard deviation of change. # Insuff. Funds is the number of insufficient funds transactions. # Low or Neg. Bal. is the number of low or negative ending balances across the statements. Missing values are replaced with median values. Standard errors are clustered by industry and quarter. ***, **, * correspond to statistical significance at the 1, 5, and 10 percent levels, respectively.

DEPENDENT VARIABLE	IS NON-PERFORMING (%)							
	YES (1)	YES (2)	NO (3)	NO (4)	YES (5)	YES (6)	NO (7)	NO (8)
Low FICO:								
FICO	-5.44*** (0.91)	-4.97*** (0.90)	-4.33*** (0.28)	-3.90*** (0.29)	-4.91*** (0.92)	-4.62*** (0.91)	-4.01*** (0.27)	-3.74*** (0.28)
Credits (Th\$)		-4.50*** (0.82)		-2.62*** (0.39)		-4.27*** (0.90)		-2.53*** (0.38)
Withdrawals (Th\$)		2.81*** (0.65)		1.62*** (0.35)		2.34*** (0.72)		0.78** (0.36)
Balance (Th\$)		-1.54** (0.69)		-1.16*** (0.36)		-1.44** (0.69)		-1.20*** (0.35)
1 (Daily Pay Loan)		1.03** (0.44)		1.21*** (0.28)		0.79* (0.43)		1.01*** (0.27)
(#) Low or Neg. Bal.		1.45*** (0.35)		1.53*** (0.29)		1.54*** (0.35)		1.54*** (0.28)
(#) Insuff. Funds		-0.21 (0.36)		0.24 (0.29)		-0.11 (0.36)		0.18 (0.28)
S.D. Credits (Th\$)		2.12*** (0.70)		1.78*** (0.36)		2.35*** (0.67)		2.14*** (0.36)
S.D. Balance (Th\$)		1.13* (0.68)		1.12*** (0.38)		1.29* (0.67)		1.34*** (0.36)
Number of Employees					-0.13*** (0.03)	-0.10** (0.04)	-0.07*** (0.01)	-0.05*** (0.02)
Observations	11,343	11,343	26,678	26,678	11,338	11,338	26,678	26,678
Industry FE	No	No	No	No	Yes	Yes	Yes	Yes
State FE	No	No	No	No	Yes	Yes	Yes	Yes
Quarter FE	No	No	No	No	Yes	Yes	Yes	Yes
Loan Type FE	No	No	No	No	Yes	Yes	Yes	Yes
R-squared	0.004	0.009	0.008	0.014	0.028	0.033	0.033	0.039
R-squared Improv.		170.1		74.8		18.6		17.8
Y-mean	22.27	22.27	14.36	14.36	22.28	22.28	14.36	14.36

The first pattern to note is that FICO scores are negatively associated with default, with the effect being stronger among low FICO borrowers, who also have higher baseline default rates (22.3% vs. 14.4%). This effect is more pronounced in the minimal model where FICO score is one of the only inputs, but in the fully specified model with cash-flow variables we see that a one standard deviation increase in FICO (49 points) is associated with a 4.6 percentage point (pp) decrease in the odds

of default for low FICO owners and 3.7 pp for high FICO owners. However, we also see that most of the cash-flow variables, such as credits, withdrawals, and balance, have a greater magnitude of predictive power for low score owners than in the high FICO segment. We also report the improvement in R^2 from adding cash-flow variables at the bottom of the table and see that the improvement for young firms is 18.6% from column 5 to 6 and relatively less for mature firms (17.8% from column 7 to 8). This pattern is consistent with low score owners having thinner or noisier credit files, making their default risk more difficult to estimate.

YOUNG VS. MATURE FIRMS. Next, we study whether the age of a business could lead to differences in the predictive power of FICO scores and cash-flow variables. We compare younger and older firms in Table 5.

Focusing on the tightly specified estimates in columns 6 and 8, we see that a one standard deviation increase in FICO in the fully specified model with cash-flow variables is associated with a 4.8 percentage point (pp) lower likelihood of default for young firms, and a 3.9 pp lower likelihood for older firms. The second thing to notice is that the cash-flow variables also exhibit strong predictive power. For example, higher credits and balances reduce default risk. Notably, a one standard deviation increase in balances (about \$64,000) is associated with more than a 2 pp lower likelihood of default for young firms but only a 1 pp lower likelihood for older firms. Conversely, higher withdrawals, credit and balance volatility, daily pay loans, and frequent low balances predict higher default for both groups, but are generally slightly more powerful for younger firms as well.

At the bottom of Table 5, we report the improvement in R^2 from adding the cash-flow variables. They are relatively more informative for younger firms; for example, the improvement is 17.6% between columns 5 and 6 and only 13.6% between columns 7 and 8. In sum, the table suggests that cash-flow variables help to predict default for both younger and older firms, but that they often have larger magnitudes and are more informative for younger firms. When we run a logistic regression (Table A.2), we see better improvement in pseudo- R^2 for younger firms (16.6% vs. 13.5%) and larger magnitudes of changes in predicted default for most cash-flow variables, although the decrease of FICO's predictiveness in the cash-flow model is the same for young and older firms.

TABLE 5: PREDICTING DEFAULT FOR YOUNG BUSINESSES (< 5)

This table uses data from Lender A and Lender B ($N = 38,021$) to show how credit score, cash-flow, and borrower characteristics predict default conditional on origination. A business is defined as young if it has operated for less than 5 years. All bank variables and FICO score are standardized to z-scores and can be interpreted as the change in the dependent variable from 1 standard deviation of change. # Insuff. Funds is the number of insufficient funds transactions. # Low or Neg. Bal. is the number of low or negative ending balances across the statements. Missing values are replaced with median values. Standard errors are clustered by industry and quarter. ***, **, * correspond to statistical significance at the 1, 5, and 10 percent levels, respectively.

DEPENDENT VARIABLE	IS NON-PERFORMING (%)							
	YES (1)	YES (2)	NO (3)	NO (4)	YES (5)	YES (6)	NO (7)	NO (8)
Young Firm (<5):								
FICO	-5.67*** (0.39)	-5.10*** (0.40)	-4.26*** (0.22)	-3.90*** (0.23)	-5.13*** (0.40)	-4.80*** (0.41)	-4.14*** (0.23)	-3.88*** (0.23)
Credits (Th\$)		-3.08*** (0.75)		-2.91*** (0.38)		-2.77*** (0.77)		-2.83*** (0.38)
Withdrawals (Th\$)		3.73*** (0.74)		1.60*** (0.33)		2.33*** (0.78)		0.93*** (0.34)
Balance (Th\$)		-2.23*** (0.62)		-0.97*** (0.34)		-2.12*** (0.65)		-1.00*** (0.33)
1 (Daily Pay Loan)		1.29*** (0.48)		1.13*** (0.27)		1.34*** (0.47)		0.86*** (0.27)
(#) Low or Neg. Bal.		2.14*** (0.47)		1.32*** (0.27)		2.20*** (0.47)		1.39*** (0.26)
(#) Insuff. Funds		-0.54 (0.40)		0.27 (0.25)		-0.69* (0.39)		0.28 (0.25)
S.D. Credits (Th\$)		2.64*** (0.70)		1.58*** (0.35)		3.08*** (0.71)		1.88*** (0.35)
S.D. Balance (Th\$)		0.89 (0.67)		1.20*** (0.40)		1.17* (0.66)		1.38*** (0.38)
Number of Employees					-0.10*** (0.03)	-0.10*** (0.03)	-0.08*** (0.02)	-0.0*** (0.02)
Observations	9,870	9,870	28,151	28,151	9,869	9,869	28,149	28,149
Industry FE	No	No	No	No	Yes	Yes	Yes	Yes
State FE	No	No	No	No	Yes	Yes	Yes	Yes
Quarter FE	No	No	No	No	Yes	Yes	Yes	Yes
Loan Type FE	No	No	No	No	Yes	Yes	Yes	Yes
R-squared	0.022	0.031	0.013	0.019	0.046	0.054	0.036	0.041
R-squared Improv.		42.8		39.0		17.6		13.6
Y-mean	18.92	18.92	15.95	15.95	18.92	18.92	15.95	15.95

COMPOUND CONSTRAINTS. We next turn to the subset of young businesses which are especially constrained: those with low FICO scores. In Table 6, we repeat the analysis from Table 5 but split the sample according to Low FICO & Young Firms vs. all others.

Here we see a much more dramatic result. Cash-flow variables substantially improve predictive accuracy for this group. Default rates are significantly higher for this group (25.4% vs. 15.9% for other firms). In the fully specified model with cash-flow variables, the coefficient on FICO here is smaller for young firms whose owners are low score than for all young firms in the previous table (column 6). A one standard deviation increase in FICO reduces default by just 15% of the mean, compared to 25% in the previous table. Meanwhile, the predictive power of some of the cash-flow variables increases, notably credits and withdrawals. For instance, a one standard deviation increase in credits lowers default risk by 4.9 pp for young firms with low score owners, relative to 2.8 for other businesses (columns 6 and 8). Similarly, a one standard deviation increase in withdrawals, is associated with a 3.7 pp for young firms with a Low FICO and only 1.0 for others. While there is a large improvement in R^2 for young, low FICO firms in the minimal model (columns 1–4), the percentage increase in R^2 is about even between the two groups under the fully specified model

(columns 5–8). Similarly, in the logistic regression results (Table A.3), we see greater improvement in pseudo- R^2 for Low FICO & Young Firms (18.7% vs. 14.8%), although similar changes in the predictive power of FICO when cash-flow data is added.

TABLE 6: PREDICTING DEFAULT FOR YOUNG BUSINESSES WITH LOW FICO OWNERS

This table uses data from Lender A and Lender B (N = 38,021) to show how credit score, cash-flow, and borrower characteristics predict default conditional on origination. A business is defined as young if it has operated for less than 5 years. Low FICO is defined as an owner having a FICO score below 700. All bank variables and FICO score are standardized to z-scores and can be interpreted as the change in the dependent variable from 1 standard deviation of change. # Insuff. Funds is the number of insufficient funds transactions. # Low or Neg. Bal. is the number of low or negative ending balances across the statements. Missing values are replaced with median values. Standard errors are clustered by industry and quarter. ***, **, * correspond to statistical significance at the 1, 5, and 10 percent levels, respectively.

DEPENDENT VARIABLE	IS NON-PERFORMING (%)							
	YES (1)	YES (2)	NO (3)	NO (4)	YES (5)	YES (6)	NO (7)	NO (8)
Low FICO & Young Firm (<5):								
FICO	-5.87*** (1.69)	-5.12*** (1.67)	-4.24*** (0.21)	-3.84*** (0.22)	-4.23** (1.69)	-3.86** (1.68)	-4.12*** (0.21)	-3.86*** (0.22)
Credits (Th\$)		-5.69*** (1.71)		-2.89*** (0.36)		-4.89*** (1.82)		-2.75*** (0.36)
Withdrawals (Th\$)		4.98*** (1.54)		1.79*** (0.32)		3.74** (1.65)		1.02*** (0.33)
Balance (Th\$)		-1.53 (1.41)		-1.21*** (0.34)		-1.17 (1.53)		-1.21*** (0.32)
1 (Daily Pay Loan)		0.99 (0.92)		1.18*** (0.26)		1.19 (0.93)		0.95*** (0.25)
(#) Low or Neg. Bal.		1.95*** (0.72)		1.49*** (0.24)		2.18*** (0.72)		1.52*** (0.24)
(#) Insuff. Funds		-0.81 (0.81)		0.16 (0.23)		-0.94 (0.77)		0.15 (0.23)
S.D. Credits (Th\$)		3.32** (1.52)		1.77*** (0.34)		3.40** (1.57)		2.10*** (0.33)
S.D. Balance (Th\$)		0.60 (1.41)		1.15*** (0.37)		0.78 (1.52)		1.35*** (0.35)
Number of Employees					-0.18*** (0.07)	-0.16** (0.08)	-0.08*** (0.01)	-0.05*** (0.02)
Observations	3,180	3,180	34,841	34,841	3,178	3,178	34,839	34,839
Industry FE	No	No	No	No	Yes	Yes	Yes	Yes
State FE	No	No	No	No	Yes	Yes	Yes	Yes
Quarter FE	No	No	No	No	Yes	Yes	Yes	Yes
Loan Type FE	No	No	No	No	Yes	Yes	Yes	Yes
R-squared	0.004	0.013	0.013	0.018	0.045	0.052	0.035	0.041
R-squared Improv.		196.4		45.8		15.8		15.3
Y-mean	25.44	25.44	15.92	15.92	25.46	25.46	15.92	15.92

5. WHO BENEFITS FROM CASH-FLOW VARIABLES? MACHINE LEARNING ANALYSIS

Having established that individual cash-flow variables predict default in intuitive ways, and often have more predictive power for constrained groups, we next use machine learning tools to deliver our key results. First, we more definitively and flexibly explore informativeness (Section 5.1). Second, we employ a new method to assess benefits to constrained groups from cash-flow underwriting (Section 5.2).

5.1 Machine learning model for default prediction

We first use a machine learning model, drawn from Hair et al. (2025), to compare the informativeness of cash-flow variables over default across groups of borrowers. Machine learning tools are helpful because the traditional linear models such as those used in Section 4 impose restrictive functional forms, potentially leading to misspecification when relationships are nonlinear or interactive (Sadhvani et al., 2021; Barbaglia et al., 2023). Machine learning models do not suffer from these issues. We use a random forest algorithm (Ho, 1995), a nonparametric method that minimizes prediction error while capturing complex interactions.¹⁰ We take the following steps to train and evaluate the model:

- 1. Hyperparameter Selection:** We tune hyperparameters—parameters that control model complexity—using cross-validation on random subsamples.¹¹
- 2. Data Splitting:** The dataset is stratified by outcome and randomly split into training (80%) and testing (20%) sets. Within the training set, 20% is further set aside for validation to refine the model and mitigate overfitting.
- 3. Bootstrapping:** To obtain robust performance estimates and standard errors, we re-estimate the models over multiple resampled datasets.

To evaluate the added value of cash-flow data, we compare two models: a Baseline model aligned with traditional underwriting practices and a Cash-Flow (CF) model that incorporates additional financial variables. The Baseline model includes FICO; firm size, age, and industry; and limited other features approximating a conventional underwriting model, while the CF model incorporates additional economically relevant transformations of cash-flow variables (see Table 7).

TABLE 7: MACHINE LEARNING MODEL FEATURES

This table lists the features included in the Machine Learning analysis. The Cash-Flow Model consists of all of the Baseline Features plus the additional features listed under Cash-Flow Model. Note that this includes transformations of the cash-flow variables included in the OLS regressions.

¹⁰ Compared to alternative ML models such as XGBoost, random forests offer similar predictive accuracy while being computationally efficient and easier to implement in our small sample setting.

¹¹ Similar to bandwidth selection in regression discontinuity designs or lag selection in VAR models, appropriate hyperparameters ensure stable and interpretable predictions.

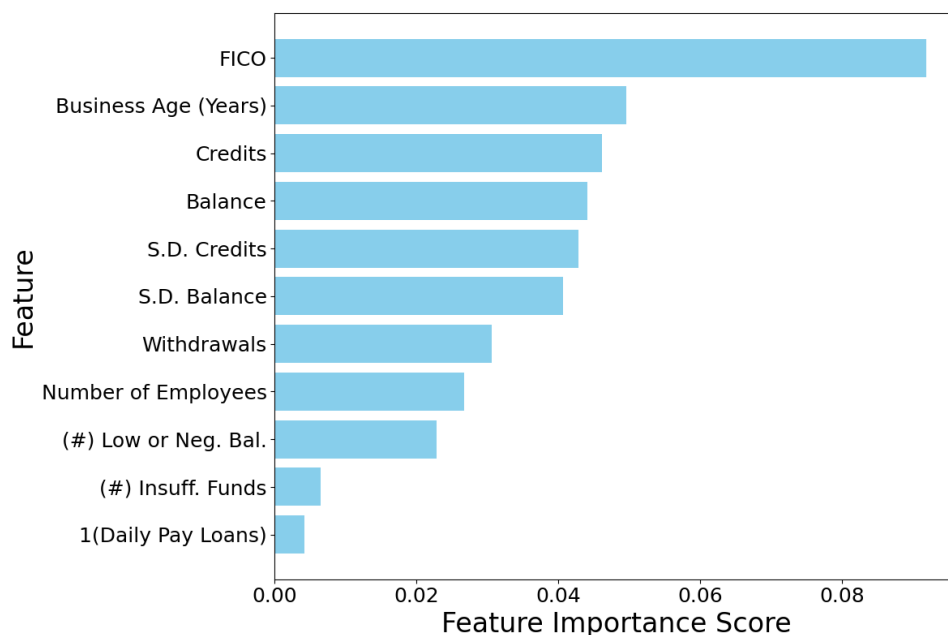
Baseline Model	
FICO Score	Industry
Business Age (Years)	Quarter Number (1-4)
Number of Employees	Late Quarter (After Median)
Lender ID	Region (NE, Midwest, South, West)
Requested Loan Amount (Log)	Loan Type
State	
Cash-Flow Model	
Credits (Log)	Balance (Log)
Withdrawals (Log)	(#) Insuff. Funds
(#) Low or Neg. Balance	1 (Daily Pay Loan)
S.D. Credits	S.D. Balance
Missing Withdrawals	Missing Daily Pay Loans
Credits (less new debt) (Log)	Missing Credits (less new debt)
Debits to Credits Ratio	Balance \times Credits
(#) Insuff. Funds \times (#) Low or Neg. Balance	Low Credit Utilization
Coeff. Variation Balance	Coeff. Variation Credits
1 (Daily Pay Loan) to Balance Ratio	Never Low or Neg. Balance
Never Insuff. Funds	(#) Insuff. Funds > 5
Balance Volatility Ratio	Credits to Balance Ratio

Performance Metrics: We evaluate models using the following metrics, reported in Table 8:

- 1. Receiver Operating Characteristic (ROC) AUC:** Measures a model's ability to distinguish between defaulters and non-defaulters. The ROC curve maps the trade-off between the true positive rate (TPR) and false positive rate (FPR), and the AUC quantifies overall discrimination. A random classifier achieves an AUC of 0.5, while higher values indicate stronger predictive performance.
- 2. H-Measure:** An alternative to ROC AUC, introduced by Hand (2009), which accounts for class imbalance more effectively.

FIGURE 2: FEATURE IMPORTANCE WITH RANDOM FOREST ML MODEL

This figure shows the importance of applicant features in the random forest Cash Flow prediction model, restricting to CF and FICO features using data from Lender A and Lender B on originated loans (N = 38,021) to present the relative importance of CF and FICO in predicting loan defaults. Higher numbers indicate greater importance.



Machine learning models do not provide direct coefficient estimates, but we assess variable importance using the mean decrease in impurity (MDI), which quantifies the contribution of each feature to model performance.¹² Figure 2 shows that while FICO and business age in years remain the two most influential predictors, several cash-flow variables exhibit high importance and collectively make a larger predictive contribution. This suggests that cash-flow data provides substantial additional information in assessing borrower risk.

In the overall sample, the CF model achieves higher performance values over the baseline, increasing both the ROC AUC and H-Measure by .011 as compared to the Baseline model (.652 AUC and .081 H-Measure). These differences are also statistically significant at the 1% level.¹³

TABLE 8: MACHINE LEARNING MODEL PERFORMANCE

This table presents our performance evaluation of the Baseline and Cash-Flow random forest models for predicting loan default for different zip code-level demographic characteristics. This table uses data from Lender A and Lender B (N = 38,021) on originated loans where loan performance is available. Performance metrics are calculated as the mean of 100 bootstrap iterations. Definitions of the performance metrics ROC AUC and H-Measure are provided in Section 5; larger numbers indicate better predictive performance. ***, **, * correspond to statistical significance at the 1, 5, and 10 percent levels, respectively.

¹² Feature importance is computed using scikit-learn's RandomForestClassifier.

¹³ Statistical significance is established via 100 bootstraps using the "Corrected Resampled t-Test" (Nadeau and Bengio, 1999).

	ROC AUC	H-MEASURE
Full Sample		
FICO Model	0.652	0.081
Cash-Flow Model	0.663	0.092
Difference	0.011***	0.011***
Firm < 5		
FICO Model	0.658	0.095
Cash-Flow Model	0.666	0.107
Difference	0.009***	0.011***
Firm ≥ 5		
FICO Model	0.647	0.078
Cash-Flow Model	0.659	0.089
Difference	0.012***	0.011***
Low FICO & Firm < 5		
FICO Model	0.599	0.063
Cash-Flow Model	0.622	0.083
Difference	0.023***	0.020***
High FICO or Firm ≥ 5		
FICO Model	0.648	0.079
Cash-Flow Model	0.660	0.089
Difference	0.011***	0.011***
Low FICO		
FICO Model	0.611	0.056
Cash-Flow Model	0.627	0.069
Difference	0.015***	0.013***
High FICO		
FICO Model	0.647	0.080
Cash-Flow Model	0.659	0.091
Difference	0.012***	0.011***

We next explore whether the CF model is more informative for the groups of borrowers we expect to be more financially constrained. As expected, the gain for low score borrowers is larger than for high score borrowers, increasing ROC AUC by .015 and H-Measure by .013. The performance improvement is not relatively more informative for younger businesses in general than for older businesses. In fact, if anything, the improvement in the CF model is smaller in the younger group. However, a very different pattern emerges when we consider Low FICO & Young Firms. Here, the ROC AUC improvement is 0.023, roughly double the gains observed for the High- FICO or Mature Firms (0.012), with similarly large improvements in H-Measure (0.020). This group also starts with substantially lower baseline model performance (ROC AUC = 0.599), highlighting that cash-flow data is especially valuable where traditional credit scores provide the least predictive power.

In sum, the machine learning results generally align with the earlier regression analysis. They show that cash-flow data is valuable for small business borrowers who struggle to access credit under traditional models, especially business owners who are constrained by both low scores and younger businesses.

5.2 Application of Tail Analysis for Comparative Outcomes (TACO)

Our final main analysis applies the Tail Analysis for Comparative Outcomes (TACO) method, originally developed in Hair et al. (2025), to examine how model selection influences default predictions across borrower subgroups. TACO provides a structured approach for identifying groups most affected by a transition between predictive models, especially when the outcome variable is continuous and differences are most pronounced in the distribution tails. It is important to emphasize that TACO does not yield zero-sum losses and gains; just because one group benefits does not mean that the counterpart group loses out to the same degree. Moreover, to the extent that TACO ratios indicate that new data sources are helping to identify default risks that other information sources miss, such a result can potentially benefit both lenders and borrowers to the extent that it makes it less likely that borrowers are extended credit that they are unlikely to succeed in repaying, leading to negative outcomes such as damaged credit scores and collections activity.

We summarize the TACO procedure as follows:

1. Model Estimation: We estimate two predictive models, f and g , which can differ in methodology or feature composition but target the same outcome. In our case, we examine default probability as an outcome under the *Baseline model* which excludes cash-flow (CF) variables, and the *CF model* which incorporates them.

2. Prediction Differences: For each borrower, we compute the change in predicted default probability when moving from the Baseline to the CF model:

$$h_i = g(X_i) - f(X_i)$$

where h_i captures the shift in predicted outcomes due to the inclusion of cash-flow features.

3. Tail Identification: We identify the most affected borrowers by selecting the top and bottom $p\%$ (in our setting $p = 10$) of observations based on h_i :

$$T^+ = \{X : h(X) > Q_{1-p}(h)\}, \quad T^- = \{X : h(X) < Q_p(h)\}$$

where Q_p represents the p -th percentile of $h(X)$. Borrowers in T^+ experience the largest predicted increases in default risk under the CF model, while those in T^- see the greatest reductions.

4. Comparative Analysis: To assess distributional impacts, we compare the characteristics of the borrowers in T^+ and T^- . Specifically, we calculate the TACO Ratio, which summarizes whether a characteristic x (e.g., an indicator for a business having been started in the past 5 years) is overrepresented in one tail relative to the other:

$$\text{TACO Ratio} = \frac{\frac{1}{|T^-|} \sum_{T^-} x_i}{\frac{1}{|T^+|} \sum_{T^+} x_i}$$

A TACO Ratio of 1 implies that the characteristic is proportionally represented in both groups, suggesting that the model transition does not create imbalances with regard to very large increases or decreases in predicted risk levels. Deviations from 1 indicate that model choice disproportionately impacts predicted risk levels positively or negatively for certain borrower subgroups.¹⁴

¹⁴ Given that models may differ in calibration or scaling, TACO's reliance on percentile-based ranking rather than absolute score differences ensures robustness in comparisons. Further details on the theoretical framework and implementation can be found in Hair et al. (2025).

Table 9 reports the TACO ratios and key inputs for firms based on FICO score, age of business, and both attributes combined. The first two columns describe the share of each group in the overall tails (T^+ and T^-), while the next four columns disaggregate those borrowers whose predicted risk levels increase the most (T^+) and decrease the most (T^-) in the shift from the Baseline to the CF model. The last column reports the TACO ratio, where values greater than one indicate a net benefit from the CF model and values below one indicate that more members of the group are reclassified as substantially higher risk. Note that the tails here are defined as the top and bottom 10%. It is straightforward to use alternative tails, up to splitting the sample in half. Our results are qualitatively similar using alternative tails.

TABLE 9: TACO RESULTS ON BENEFIT OF CASH-FLOW VS BASELINE MODEL

This table shows the results from implementing Tail Analysis for Comparative Outcomes (TACO). We compare two random forest models to predict default: a Baseline model (containing FICO, firm size, firm age, and industry among others) and the Cash-Flow (CF) model, which adds bank statement variables to the Baseline model. The table uses data from Lender A and Lender B ($N = 38,021$) on originated loans. The observation counts represent the sum across 1,000 bootstrap holdout samples. The first two columns ("Tails") show the group's share in the decile tails population of bootstrapped sample observations. The next two columns restrict to the 10% of each bootstrap sample with the highest increase the predicted likelihood of default between the Baseline and the CF model, who are thus adversely affected by switching from the Baseline to the CF model. The mean shows the share of young owners in this group, which can be compared to the "Tails" mean column. The next two columns show the same metric for the bottom 10% (the group that most benefits from switching to the CF model). The last column shows the ratio between the means for the adversely affected and benefited tails, which we call the TACO ratio. A ratio of one implies no implication of switching models, a ratio less than one implies that the group is adversely affected, and a ratio greater than one implies that the group benefits. We calculate standard errors for the TACO ratio using the percentile bootstrap.

	TAILS		TOP 10% DEFAULT INCREASES W/ CF MODEL (HURT)		BOTTOM 10% DEFAULT INCREASES W/ CF MODEL (BENEFIT)		TACO RATIO
	N	MEAN	N	MEAN	N	MEAN	
FICO Split:							
Low FICO (< 700)	1,522,000	0.405	761,000	0.175	761,000	0.635	3.626**
High FICO (≥ 700)	1,522,000	0.595	761,000	0.825	761,000	0.365	0.442**
Full Sample:							
Young Firm	1,522,000	0.302	761,000	0.221	761,000	0.383	1.735**
Mature Firm	1,522,000	0.698	761,000	0.779	761,000	0.617	0.792**
Low FICO (< 700):							
Young Firm	1,522,000	0.140	761,000	0.033	761,000	0.247	7.573**
Mature Firm	1,522,000	0.265	761,000	0.143	761,000	0.388	2.722**
High FICO (≥ 700):							
Young Firm	1,522,000	0.162	761,000	0.188	761,000	0.136	0.722*
Mature Firm	1,522,000	0.433	761,000	0.637	761,000	0.229	0.359**

We can now interpret the results in Table 9. The first two rows offer a FICO benchmark. Mechanically, high FICO borrowers benefit from reliance on credit scores, while low FICO borrowers benefit from CF-based underwriting; this latter group is 3.6 times more likely to be in the group with substantially lower risk predictions than in the group with substantially higher risk predictions under the cash-flow model. In contrast, borrowers with high credit scores have a TACO ratio of .442, indicating that they are more likely to be in the group whose risk predictions increase substantially using the CF model. This indicates that the model is providing risk signals that are not being captured by traditional credit scores.

Next, we see that young firms also benefit; they have a TACO ratio of 1.74, meaning they are 74% more likely to appear in the group where risk levels decrease substantially than in the group whose predicted risk levels rise. In contrast, mature firms have a ratio of .79, indicating that they are slightly more likely to be in the group whose risk predictions increase substantially.

Much as with the performance results, the size of impacts is dramatically amplified for young firms with low FICO scores, where the TACO ratio reaches 7.57. Mature firms with low scores are also more likely to see their risk levels decrease substantially under the CF model, with a TACO ratio of 2.72. For young firms with high FICO scores, the ratio is 0.72. Again, while this is less than one (i.e., more borrowers' predicted risk levels increase than decrease), it is much higher than the benchmark 0.44 for high FICO borrowers in general. This suggests that CF-based underwriting helps substantial numbers of young firms across the credit score distribution.

These findings suggest that incorporating cash-flow data into credit models can improve access to credit for categories of borrowers that lenders often consider high risk. Overall, the CF model leads to a substantial reallocation of predicted risk, which benefits young firms, low FICO borrowers, and entrepreneurs who face both constraints, a pattern consistent with our OLS and model performance results.

6. CONCLUDING THOUGHTS

In this report, we consider the implications of incorporating basic variables drawn from bank statements in small business lending with a particular focus on two important types of firms that face financial constraints for different reasons: Younger firms and firms led by entrepreneurs with low personal credit scores. Traditional underwriting heavily emphasizes entrepreneurs' personal credit scores, which are tied to the age and socioeconomic status of the business owner. Using data and methods drawn from Hair et al. (2025), we show that incorporating cash-flow variables enables lenders to predict default more accurately in general and especially for these constrained groups.

These results have market and policy implications, since use of electronic cash-flow data is growing across the small business lending market but still varies substantially among different types of lenders. Concurrent with this empirical analysis, FinRegLab is releasing an updated qualitative study of mission-based lenders' experiences in adopting electronic cash-flow data and platform technologies to expand their small business lending programs in light of their historical technology and resource constraints (FinRegLab, 2025). The results also underscore that the ability to securely and easily share banking data and other timely information about applicants' ability to repay with third party lenders can have important implications for access to credit.

These considerations are especially important for small businesses, where financial constraints often prevent promising firms from bridging short-term gaps, operating at full capacity, or expanding to new ventures. Improving access to and use of timely and relevant financial data to improve allocation of credit toward financially viable firms could have potential benefits for lenders, borrowers, and the broader national economy.

References

- Aktug, Emrehan, Devrim Ikizler, and Timur Hulagu, 2020, Effects of small loans on bank and small business growth, Technical report, Office of Advocacy, U.S. Small Business Administration.
- Audretsch, David B, 2002, The dynamic role of small firms: Evidence from the US, *Small Business Economics* 18, 13–40.
- Avery, Robert B, Paul S Calem, and Glenn B Canner, 2004, Credit report accuracy and access to credit, Technical report, Federal Reserve Board.
- Azoulay, Pierre, Benjamin F Jones, J Daniel Kim, and Javier Miranda, 2020, Age and high-growth entrepreneurship, *American Economic Review: Insights* 2, 65–82.
- Barbaglia, Luca, Sebastiano Manzan, and Elisa Tosetti, 2023, Forecasting loan default in Europe with machine learning, *Journal of Financial Econometrics* 21, 569–596.
- Barca, Alaina, and Harry Hou, 2024, US bank branch closures and banking deserts, Report, Federal Reserve Bank of Philadelphia.
- Bartik, Alexander W, Marianne Bertrand, Zoe Cullen, Edward L Glaeser, Michael Luca, and Christopher Stanton, 2020, The impact of COVID-19 on small business outcomes and expectations, *PNAS* 117, 17656–17666.
- Bates, Timothy, 1990, Entrepreneur human capital inputs and small business longevity, *The Review of Economics and Statistics* 72, 551–559.
- Berger, Allen N, and Gregory F Udell, 1998, The economics of small business finance: The roles of private equity and debt markets in the financial growth cycle, *Journal of Banking & Finance* 22, 613–673.
- Blattner, Laura, and Scott Nelson, 2024, How costly is noise? Data and disparities in consumer credit, Working Paper No. 3978, Stanford University, Graduate School of Business.
- Block, Joern H, Lennart Hoogerheide, and Roy Thurik, 2013, Education and entrepreneurial choice: An instrumental variables analysis, *International Small Business Journal* 31, 23–33.
- Brevoort, Kenneth P, Philipp Grimm, and Michelle Kambara, 2016, Credit invisibles and the unscored, *Cityscape* 18, 9–34.
- Brown, J David, and John S Earle, 2017, Finance and growth at the firm level: Evidence from SBA loans, *The Journal of Finance* 72, 1039–1080.
- Bureau of Labor Statistics, 2024, Entrepreneurship and the US economy, Accessed: 2025-05-27.
- Chava, Sudheer, Manasa Gopal, Manpreet Singh, and Yafei Zhang, 2023, The dark side of entrepreneurship, Georgia Institute of Technology.
- Chen, Brian S, Samuel G Hanson, and Jeremy C Stein, 2017, The decline of big-bank lending to small business: Dynamic impacts on local credit and labor markets, Technical report, National Bureau of Economic Research.
- Consumer Financial Protection Bureau, 2017, Key dimensions of the small business lending landscape. Consumer Financial Protection Bureau, 2024, Required rulemaking on personal financial data rights, *Federal Register*, Vol. 89, No. 222, pp. 90838–90956, Docket No. CFPB–2023–0052; RIN 3170–AA78.
- Cook, Lisa D, Matt Marx, and Emmanuel Yimfor, 2023, Funding black high-growth startups, Columbia Business School Research Paper No. 4279986.
- Crews, Jonas, Ross DeVol, Richard Florida, and Dave Shideler, 2020, Young firms and regional economic growth, Heartland Forward.
- Crosman, Penny, 2024, How banks can overcome challenges of cash flow underwriting technology, Technical report, American Banker.
- Decker, Ryan, John Haltiwanger, Ron Jarmin, and Javier Miranda, 2014, The role of entrepreneurship in US job creation and economic dynamism, *Journal of Economic Perspectives* 28, 3–24.
- Decker, Ryan A, and John Haltiwanger, 2023, Surging business formation in the pandemic: Causes and consequences?, *Brookings Papers on Economic Activity* 2023, 249–316.
- Decker, Ryan A, John Haltiwanger, Ron S Jarmin, and Javier Miranda, 2016a, Declining business dynamism: What we know and the way forward, *American Economic Review* 106, 203–207.
- Decker, Ryan A, John Haltiwanger, Ron S Jarmin, and Javier Miranda, 2016b, Where has all the skewness gone? The decline in high-growth (young) firms in the US, *European Economic Review* 86, 4–23.
- Delis, Manthos, Fulvia Fringuellotti, and Steven Ongena, 2024, Credit and entrepreneurs' income, Federal Reserve Bank of New York Staff Reports, no. 929.
- Dinlersoz, Emin, Sebnem Kalemli-Ozcan, Henry Hyatt, and Veronika Penciakova, 2019, Leverage over the firm life-cycle, firm growth, and aggregate fluctuations, Working Paper 2019-18, Federal Reserve Bank of Atlanta.
- Economic Innovation Group, 2023, Distressed Communities Index, Accessed: 2025-05-27.
- Edelberg, Wendy, and Noadia Steinmetz-Silber, 2024, The changing demographics of business ownership, The Brookings Institution.
- Edelberg, Wendy, Jon Steinsson, Janice C Eberly, and John Haltiwanger, 2023, Is the post-pandemic surge in business dynamism here to stay?, The Brookings Institution.

- Equifax, 2022, Achieve more inclusive credit: Expanding your credit reach to marginalized populations can help them and you.
- Ewing Marion Kauffman Foundation, 2019, Kauffman Firm Survey, Accessed: 2025-05-27. Experian, 2024, New Experian tool empowers financial inclusion through open banking insights.
- Experian, 2025, Launch of Experian's Cashflow Score signals new era of open banking-powered lending.
- Fairlie, Robert, Alicia Robb, and David T Robinson, 2022, Black and white: Access to capital among minority- owned start-ups, *Management Science* 68, 2377–2400.
- Farrell, Diana, and Chris Wheat, 2016, Cash is king: Flows, balances, and buffer days, JPMorgan Chase & Co Institute.
- Farrell, Diana, Chris Wheat, and Chi Mac, 2018, Growth, vitality, and cash flows: High-frequency evidence from 1 million small businesses, JPMorgan Chase & Co Institute.
- Fazio, Catherine E, Jorge Guzman, Yupeng Liu, and Scott Stern, 2021, How is COVID changing the geography of entrepreneurship? Evidence from the Startup Cartography Project, NBER Working Paper 28787, National Bureau of Economic Research.
- Federal Deposit Insurance Corporation, 2018, Small Business Lending Survey. Federal Deposit Insurance Corporation, 2024, Small Business Lending Survey.
- Federal Reserve Board, 2017, Report to the Congress on the availability of credit to small businesses, Board of Governors of the Federal Reserve System.
- Federal Trade Commission, 2012, Report to Congress under Section 319 of the Fair and Accurate Credit Transactions Act of 2003.
- Fikri, Kenan, and Daniel Newman, 2024, How the pandemic rebooted entrepreneurship in the US, Harvard Business Review.
- FinRegLab, 2019a, The use of cash flow data in underwriting credit: Empirical research findings, Technical report, FinRegLab.
- FinRegLab, 2019b, The use of cash-flow data in underwriting credit: Small business spotlight, Technical report, FinRegLab.
- FinRegLab, 2020, The use of cash-flow data in underwriting credit: Market context & policy analysis, Technical report, FinRegLab.
- FinRegLab, 2025, Transforming small business credit: Technology and data adoption in mission-based lending, Technical report, FinRegLab.
- Fracassi, Cesare, Mark J Garmaise, Shimon Kogan, and Gabriel Natividad, 2016, Business microloans for US subprime borrowers, *Journal of Financial and Quantitative Analysis* 51, 55–83.
- Hair, Christopher, Sabrina T Howell, Mark J Johnson, and Siena Matsumoto, 2025, Modernizing access to credit for younger entrepreneurs: From FICO to cash flow, NBER Working Paper 33367, National Bureau of Economic Research.
- Haltiwanger, John, Ron S Jarmin, Robert Kulick, and Javier Miranda, 2016, High growth young firms: Contribution to job, output, and productivity growth, in *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges*, 11–62 (University of Chicago Press).
- Hand, David J, 2009, Measuring classifier performance: A coherent alternative to the Area Under the ROC Curve, *Machine Learning* 77, 103–123.
- Hand, Mark C, Vivek Shastri, and Varun Rai, 2023, Predicting firm creation in rural Texas: A multi-model machine learning approach to a complex policy problem, *PLoS ONE* 18.
- Hepinstall, Mike, Chaitra Chandrasekhar, Peter Carroll, Nick Dykstra, and Yigit Ulucay, 2022, Financial inclusion and access to credit, Oliver Wyman.
- Herkenhoff, Kyle, Gordon M Phillips, and Ethan Cohen-Cole, 2021, The impact of consumer credit access on self-employment and entrepreneurship, *Journal of Financial Economics* 141, 345–371.
- Ho, Tin Kam, 1995, Random decision forests, in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, 278–282, IEEE.
- Hwang, Victor, Sameeksha Desai, and Ross Baird, 2019, Access to capital for entrepreneurs: Removing barriers, Ewing Marion Kauffman Foundation.
- Hyatt, Henry R, 2022, Firm age and job creation in the US, *IZA World of Labor* 501.
- Lawler, Ryan, 2024, Plaid introduces cash flow data-based credit report, Axios.
- McSwigan, Curran, 2022, Who is behind America's startup surge?, Third Way.
- Mills, Karen Gordon, and Brayden McCarthy, 2014, The state of small business lending: Credit access during the recovery and how technology may change the game, Harvard Business School Working Paper No. 15-004.
- Nadeau, Claude, and Yoshua Bengio, 1999, Inference for the generalization error, *Advances in Neural Information Processing Systems* 12.
- Nanda, Ramana, and Gordon Phillips, 2023, Small firm financing: Sources, frictions, and policy implications, in *Handbook of the Economics of Corporate Finance*, volume 1, 107–135 (Elsevier).
- Next Street Financial, 2024, Access to capital and credit for entrepreneurs and small businesses in Appalachia, Prepared for the Appalachian Regional Commission under contract CO-20993-23.

- Nguyen, Hoai-Luu Q., 2019, Are credit markets still local? Evidence from bank branch closings, *American Economic Journal: Applied Economics* 11, 1–32.
- Robb, Alicia M, and David T Robinson, 2014, The capital structure decisions of new firms, *The Review of Financial Studies* 27, 153–179.
- Sadhwani, Apaar, Kay Giesecke, and Justin Sirignano, 2021, Deep learning for mortgage risk, *Journal of Financial Econometrics* 19, 313–368.
- Scott, Jacob, Ambika Nair, and Claire Kramer Mills, 2025, Credit insecurity in the United States: 2018–2023, Technical report, Federal Reserve Bank of New York.
- Sedlacek, Petr, and Chenchuan Shi, 2024, *Work from Home, Business Dynamism and the Macroeconomy* (Centre for Economic Policy Research).
- Taylor, Niki, and Rohan Sriram, 2025, Introducing transactions for business: Powering real-time business tools, Plaid.
- Toh, Ying Lei, 2024, Addressing traditional credit scores as a barrier to accessing affordable credit, *Economic Review* Third Quarter 2023.
- Toussaint-Comeau, Maude, Yi David Wang, and Robin Newberger, 2019, Impact of bank closings on credit extension to businesses in low-income and minority neighborhoods, *The Review of Black Political Economy* 47.
- U.S. Census Bureau, 2025, Business formation statistics, Accessed: 2025-05-27.
- U.S. Small Business Administration, 2019, Small Business Profiles.
- Van Dam, Andrew, 2023, Why the south has such low credit scores, *The Washington Post*.
- Van Leuven, Andrew J, Dayton Lambert, Tessa Conroy, and Kelsey L Thomas, 2024, Do “banking deserts” even exist? Examining access to brick-and-mortar financial institutions in the continental United States, *Applied Geography* 165.
- VantageScore, 2024, Credit bureaus + bank data = VantageScore 4 Plus, VantageScore Website.
- Wang, Ye, and Shuang Wu, 2024, Impact of mobile banking on small business lending after bank branch closures, *Journal of Corporate Finance* 87.

APPENDIX A:

Supplementary Tables and Figures

TABLE A.1: NATIONAL SUMMARY STATISTICS

This table contains zip code-level summary statistics for the U.S. population in 2019. Each observation represents one zip code. Population statistics come from the U.S. Census Bureau American Community Survey.

	N	MEAN	MEDIAN	SD
Total Population	41,037	12,858	4,617	16,654
Per Capita Income	40,308	32	29	16
Pct Black Pop (%)	40,620	9	1.70	17
Pct Hisp. Pop (%)	40,620	11	4.09	18
Pct Asian Pop (%)	40,620	2.77	0.60	6.27
Pct Bachelors (%)	40,581	26	21	18
Pct Home Owner (%)	40,184	71	75	20
Pct Unemployed (%)	40,391	5.57	4.50	5.80
Density (Pop/km ²)	33,759	2,316	19	54,400

TABLE A.2: PREDICTING DEFAULT (LOGISTIC) FOR YOUNG BUSINESSES (< 5)

This table uses data from Lender A and Lender B (N = 38,021) to show how credit score, cash-flow, and borrower characteristics predict default conditional on origination. A business is defined as young if it has operated for less than 5 years. All bank variables and FICO score are standardized to z-scores and can be interpreted as the change in the dependent variable from 1 standard deviation of change. # Insuff. Funds is the number of insufficient funds transactions. # Low or Neg. Bal. is the number of low or negative ending balances across the statements. Missing values are replaced with median values. Standard errors are clustered by industry and quarter. ***, **, * correspond to statistical significance at the 1, 5, and 10 percent levels, respectively.

DEPENDENT VARIABLE	IS NON-PERFORMING (%)							
	YES (1)	YES (2)	NO (3)	NO (4)	YES (5)	YES (6)	NO (7)	NO (8)
Young Firm (<5):								
FICO	-0.38*** (0.03)	-0.35*** (0.03)	-0.33*** (0.02)	-0.30*** (0.02)	-0.35*** (0.03)	-0.33*** (0.03)	-0.33*** (0.02)	-0.31*** (0.02)
Credits (Th\$)		-0.23*** (0.06)		-0.24*** (0.03)		-0.20*** (0.06)		-0.24*** (0.03)
Withdrawals (Th\$)		0.26*** (0.05)		0.14*** (0.03)		0.16*** (0.05)		0.09*** (0.03)
Balance (Th\$)		-0.18*** (0.05)		-0.09*** (0.03)		-0.17*** (0.05)		-0.09*** (0.03)
1(Daily Pay Loan)		0.08*** (0.03)		0.08*** (0.02)		0.08*** (0.03)		0.06*** (0.02)
(#) Low or Neg. Bal.		0.12*** (0.02)		0.08*** (0.02)		0.13*** (0.03)		0.09*** (0.02)
(#) Insuff. Funds		-0.03 (0.03)		0.02 (0.02)		-0.04 (0.02)		0.02 (0.02)
S.D. Credits (Th\$)		0.18*** (0.05)		0.13*** (0.03)		0.22*** (0.05)		0.15*** (0.03)
S.D. Balance (Th\$)		0.07 (0.05)		0.10*** (0.03)		0.09* (0.05)		0.11*** (0.03)
Number of Employees					-0.01*** (0.00)	-0.01*** (0.00)	-0.01*** (0.00)	-0.00*** (0.00)
Observations	9,870	9,870	28,151	28,151	9,866	9,866	28,126	28,126
Industry FE	No	No	No	No	Yes	Yes	Yes	Yes
State FE	No	No	No	No	Yes	Yes	Yes	Yes
Quarter FE	No	No	No	No	Yes	Yes	Yes	Yes
Loan Type FE	No	No	No	No	Yes	Yes	Yes	Yes
Pseudo R-squared	0.023	0.032	0.015	0.022	0.051	0.060	0.043	0.049
R-squared Improv.		42.9		39.4		16.6		13.5
Y-mean	18.92	18.92	15.95	15.95	18.92	18.92	15.96	15.96

TABLE A.3: PREDICTING DEFAULT (LOGISTIC) FOR YOUNG BUSINESSES WITH LOW FICO OWNERS

This table uses data from Lender A and Lender B (N = 38,021) to show how credit score, cash-flow, and borrower characteristics predict default conditional on origination. A business is defined as young if it has operated for less than 5 years. Low FICO is defined as an owner having a FICO score below 700. All bank variables and FICO score are standardized to z-scores and can be interpreted as the change in the dependent variable from 1 standard deviation of change. # Insuff. Funds is the number of insufficient funds transactions. # Low or Neg. Bal. is the number of low or negative ending balances across the statements. Missing values are replaced with median values. Standard errors are clustered by industry and quarter. ***, **, * correspond to statistical significance at the 1, 5, and 10 percent levels, respectively.

DEPENDENT VARIABLE	IS NON-PERFORMING (%)							
	YES (1)	YES (2)	NO (3)	NO (4)	YES (5)	YES (6)	NO (7)	NO (8)
Low FICO & Young Firm (<5):								
FICO	-0.29*** (0.08)	-0.26*** (0.08)	-0.32*** (0.02)	-0.29*** (0.02)	-0.21** (0.08)	-0.20** (0.08)	-0.33*** (0.02)	-0.31*** (0.02)
Credits (Th\$)		-0.33*** (0.11)		-0.24*** (0.03)		-0.29** (0.11)		-0.23*** (0.03)
Withdrawals (Th\$)		0.28*** (0.09)		0.16*** (0.03)		0.22** (0.10)		0.09*** (0.03)
Balance (Th\$)		-0.09 (0.08)		-0.11*** (0.03)		-0.07 (0.09)		-0.11*** (0.03)
1(Daily Pay Loan)		0.05 (0.04)		0.08*** (0.02)		0.06 (0.05)		0.06*** (0.02)
(#) Low or Neg. Bal.		0.10*** (0.03)		0.09*** (0.01)		0.11*** (0.03)		0.10*** (0.01)
(#) Insuff. Funds		-0.04 (0.04)		0.01 (0.01)		-0.05 (0.04)		0.01 (0.01)
S.D. Credits (Th\$)		0.18** (0.08)		0.14*** (0.03)		0.18** (0.09)		0.17*** (0.03)
S.D. Balance (Th\$)		0.03 (0.08)		0.09*** (0.03)		0.05 (0.08)		0.11*** (0.03)
Number of Employees					-0.01** (0.00)	-0.01** (0.01)	-0.01*** (0.00)	-0.00*** (0.00)
Observations	3,180	3,180	34,841	34,841	3,132	3,132	34,815	34,815
Industry FE	No	No	No	No	Yes	Yes	Yes	Yes
State FE	No	No	No	No	Yes	Yes	Yes	Yes
Quarter FE	No	No	No	No	Yes	Yes	Yes	Yes
Loan Type FE	No	No	No	No	Yes	Yes	Yes	Yes
Pseudo R-squared	0.004	0.011	0.014	0.021	0.035	0.041	0.042	0.049
R-squared Improv.		208.7		46.0		18.7		14.8
Y-mean	25.44	25.44	15.92	15.92	25.73	25.73	15.93	15.93

APPENDIX B:

Geographic Analysis

Access to finance can vary not only based on firm characteristics, but also across geographies. For instance, research suggest that businesses in areas with greater poverty and unemployment, less Internet connectivity, and fewer banks often face particular challenges accessing credit. This can include businesses in both rural and innercity neighborhoods. (Wang and Wu, 2024; Van Leuven et al., 2024; Barca and Hou, 2024; Nguyen, 2019; Toussaint-Comeau et al., 2019; Hand et al., 2023; Next Street Financial, 2024). Particularly in light of research indicating that much of the post-pandemic surge in entrepreneurship has taken place outside of traditional entrepreneurial ecosystems,¹⁵ we conducted a further analysis focusing on the impacts of incorporating cash-flow data on low score entrepreneurs whose businesses are located in zip codes with the lowest median incomes and highest percentages of Black and Hispanic residents in the sample. We were unable to conduct a separate analysis of rural areas due to the limited number of observations.

As described below, we found stronger impacts for some categories, but that many of the results appeared to be driven primarily by the overall benefit to low score owners rather than location-specific effects. Nevertheless, to the degree certain geographies have higher numbers of residents with lower scores, cash flow-based underwriting could be important for increasing overall credit access in economically disadvantaged communities.

Definitions and summary statistics

For purposes of the geographic analysis we define Low-Income zip codes as below the median per capita income and High-Black and High-Hispanic zip codes as within the top quartile of respective minority population share within the originated loan sample. Zip codes that are in both categories in the case of the Black population are designated as High-Black, Low-Income (HBLI), for example. As reflected in Table 1 and briefly discussed in Section 2.2, zip codes in the originated loan sample have higher median Black, and Hispanic population shares than the national averages reflected in Table A.1. While median per capita income and the percentage of residents with bachelor's degrees are somewhat higher in the sample zip codes, home ownership rates are lower and unemployment rates equal national medians.

Summary statistics in Tables B.1, B.2, and B.3 show that the loans originated in LI, HBLI, and HHLI zip codes, respectively, are associated with greater financial constraints and higher default rates than loans to businesses that are not located in such areas.

¹⁵ In addition to high levels of formation in the Southeast, rates were also particularly strong in areas with higher minority populations (McSwigan, 2022; Economic Innovation Group, 2023; Van Dam, 2023; Scott et al., 2025; Fazio et al., 2021).

TABLE B.1: SUMMARY STATISTICS BY LOW-INCOME ZIP CODE

This table compares businesses located in Low-Income zip codes with those who are not in the sample of originated loans from Lender A and Lender B (N = 38,021).

	LOW INCOME				HIGH INCOME			
	N	MEAN	MEDIAN	SD	N	MEAN	MEDIAN	SD
Loan Variables:								
Requested Loan Amount (Th\$)	18,945	103	75	94	19,005	114	90	102
APR (%)	18,029	16	15	5.13	18,153	16	15	5.09
Non-Performing Loan (%)	18,945	16.9			19,005	16.5		
Originated Loan Amount (Th\$)	18,945	109	72	99	19,005	121	86	109
Loan Maturity (Years)	18,945	3.14	3.00	1.41	19,005	3.16	3.00	1.41
Credit Score & Cash-Flow (Bank Statement) Variables:								
FICO	18,945	726	723	49	19,005	730	728	49
Credits (Th\$)	18,945	126	70	166	19,005	135	78	173
Balance (Th\$)	18,945	38	18	59	19,005	45	22	68
(#) Insuff. Funds	18,945	0.05	0.00	0.26	19,005	0.05	0.00	0.25
(#) Low or Neg. Bal.	18,945	0.45	0.00	1.10	19,005	0.39	0.00	1.02
Withdrawals (Th\$)	9,621	162	88	203	10,678	174	98	213
1 (Daily Pay Loan)	12,298	0.06			11,622	0.06		
S.D. Credits (Th\$)	18,945	13	7.23	13	19,005	14	8.06	14
S.D. Balance (Th\$)	18,945	5.51	3.12	5.95	19,005	6.21	3.69	6.33
Borrower Characteristics:								
Female	18,841	0.22			18,911	0.23		
Owner Age	10,533	49	48	11	9,636	49	48	11
Business Characteristics:								
Business Age (Years)	18,945	11	7.37	7.71	19,005	11	7.98	7.57
Number of Employees	18,945	10	6.00	12	19,005	10	7.00	13
Professional Services Industries	18,945	0.29			19,005	0.36		
Capital Intensive Industries	18,945	0.31			19,005	0.23		
Retail, Food, Healthcare & Other Service Industries	18,945	0.40			19,005	0.40		
Zip Code Level Characteristics:								
Per Capita Income (Th\$)	18,945	27	28	5.30	19,005	52	46	18
Pct Black Pop (%)	18,945	16	7.90	19	19,005	6.99	3.90	9
Pct Hisp. Pop (%)	18,945	23	13	24	19,005	12	8.65	11
Pct Asian Pop (%)	18,945	4.44	2.00	7.44	19,005	8.81	5.40	10
Pct Bachelors (%)	18,945	25	24	10	19,005	51	50	14
Pct Home Owner (%)	18,937	60	63	17	19,005	69	73	19
Pct Unemployed (%)	18,941	6.04	5.50	2.96	19,002	3.98	3.80	1.43
Density (Pop/km2)	18,817	943	451	2,582	18,893	1,156	508	3,349

TABLE B.2: SUMMARY STATISTICS BY HIGH-BLACK AND LOW-INCOME ZIP CODE

This table compares businesses located in High-Black and Low-Income zip codes with those who are not in the sample of originated loans from Lender A and Lender B (N = 38,021).

	HIGH-BLACK & LOW-INCOME				NOT (HIGH-BLACK & LOW-INCOME)			
	N	MEAN	MEDIAN	SD	N	MEAN	MEDIAN	SD
Loan Variables:								
Requested Loan Amount (Th\$)	6,882	102	75	94	31,068	110	80	99
APR (%)	6,531	16	15	5.00	29,651	16	15	5.13
Non-Performing Loan (%)	6,882	18.5			31,068	16.3		
Originated Loan Amount (Th\$)	6,882	107	70	98	31,068	117	80	106
Loan Maturity (Years)	6,882	3.15	3.00	1.41	31,068	3.15	3.00	1.41
Credit Score & Cash-Flow (Bank Statement) Variables:								
FICO	6,882	724	722	48	31,068	729	727	49
Credits (Th\$)	6,882	122	68	163	31,068	133	75	171
Balance (Th\$)	6,882	37	17	58	31,068	43	20	65
(#) Insuff. Funds	6,882	0.05	0.00	0.25	31,068	0.05	0.00	0.26
(#) Low or Neg. Bal.	6,882	0.48	0.00	1.16	31,068	0.41	0.00	1.04
Withdrawals (Th\$)	3,338	157	83	202	16,961	170	95	210
1 (Daily Pay Loan)	4,589	0.05			19,331	0.06		
S.D. Credits (Th\$)	6,882	12	7.09	13	31,068	13	7.76	14
S.D. Balance (Th\$)	6,882	5.46	3.05	5.97	31,068	5.95	3.48	6.19
Borrower Characteristics:								
Female	6,825	0.23			30,927	0.22		
Owner Age	3,991	48	48	11	16,178	49	48	11
Business Characteristics:								
Business Age (Years)	6,882	11	7.50	7.78	31,068	11	7.73	7.61
Number of Employees	6,882	9	5.00	12	31,068	10	7.00	13
Professional Services Industries	6,882	0.30			31,068	0.33		
Capital Intensive Industries	6,882	0.29			31,068	0.27		
Retail, Food, Healthcare & Other Service Industries	6,882	0.40			31,068	0.40		
Zip Code Level Characteristics:								
Per Capita Income (Th\$)	6,882	26	26	5.55	31,068	42	39	18
Pct Black Pop (%)	6,882	35	28	20	31,068	6.06	3.80	7.62
Pct Hisp. Pop (%)	6,882	19	12	17	31,068	17	10	20
Pct Asian Pop (%)	6,882	4.06	2.10	5.42	31,068	7.19	3.70	10
Pct Bachelors (%)	6,882	24	24	10	31,068	41	39	18
Pct Home Owner (%)	6,876	55	56	16	31,066	67	71	18
Pct Unemployed (%)	6,879	7.27	6.50	3.44	31,064	4.51	4.20	1.97
Density (Pop/km2)	6,825	1,091	632	3,232	30,885	1,041	446	2,937

FIGURE B.1: HIGH-BLACK AND LOW-INCOME ZIP CODES

This figure uses data from Lender A and Lender B on originated loans (N = 38,021) to show the geographic distribution of High-Black & Low-Income zip codes in orange and other zip codes in the data in blue.

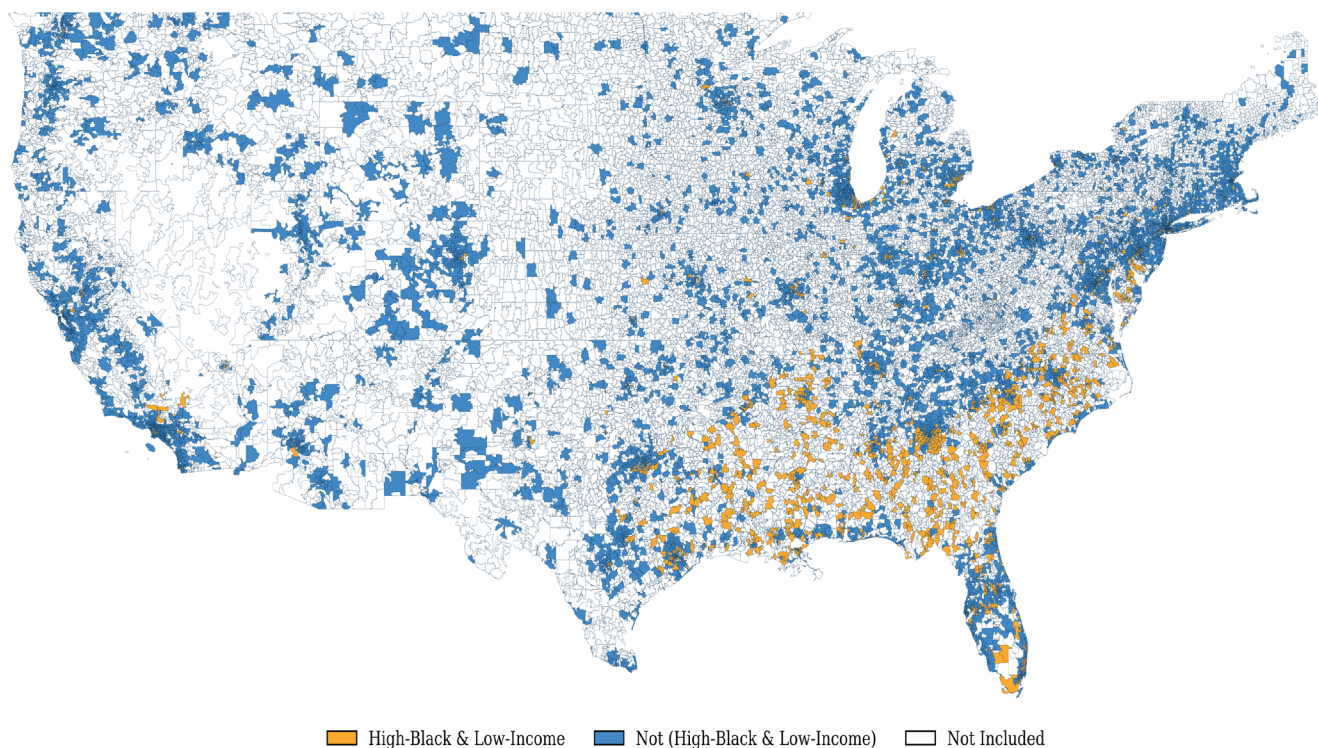


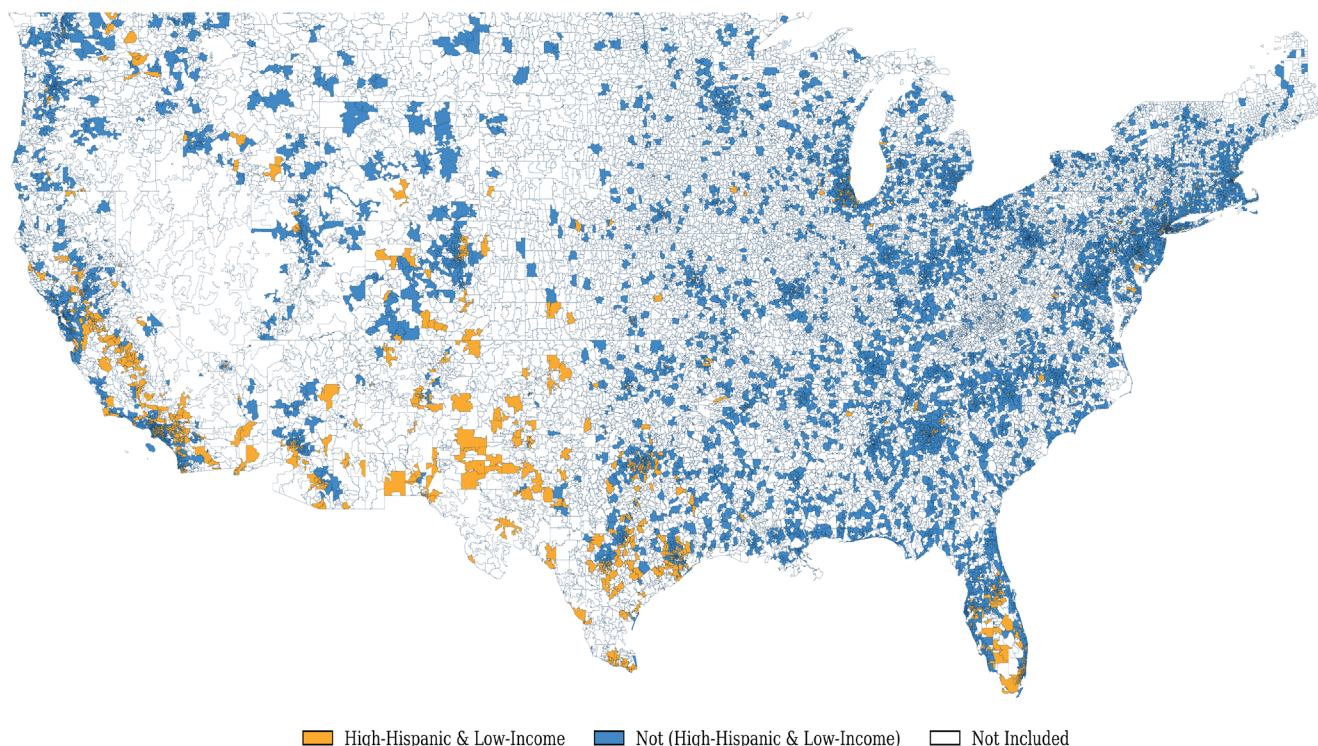
TABLE B.3: SUMMARY STATISTICS BY HIGH-HISPANIC AND LOW-INCOME ZIP CODE

This table compares businesses located in High-Hispanic and Low-Income zip codes with those who are not in the sample of originated loans from Lender A and Lender B (N = 38,021).

	HIGH-HISPANIC & LOW-INCOME				NOT (HIGH-HISPANIC & LOW-INCOME)			
	N	MEAN	MEDIAN	SD	N	MEAN	MEDIAN	SD
Loan Variables:								
Requested Loan Amount (Th\$)	6,935	109	85	97	31,015	108	80	98
APR (%)	6,595	16	15	5.09	29,587	16	15	5.11
Non-Performing Loan (%)	6,935	18.2			31,015	16.4		
Originated Loan Amount (Th\$)	6,935	114	78	103	31,015	115	78	105
Loan Maturity (Years)	6,935	3.11	3.00	1.40	31,015	3.16	3.00	1.41
Credit Score & Cash-Flow (Bank Statement) Variables:								
FICO	6,935	726	723	48	31,015	729	726	49
Credits (Th\$)	6,935	134	74	172	31,015	130	73	169
Balance (Th\$)	6,935	43	20	64	31,015	42	20	63
(#) Insuff. Funds	6,935	0.06	0.00	0.26	31,015	0.05	0.00	0.25
(#) Low or Neg. Bal.	6,935	0.43	0.00	1.08	31,015	0.42	0.00	1.06
Withdrawals (Th\$)	3,448	174	96	210	16,851	167	92	208
1 (Daily Pay Loan)	4,520	0.06			19,400	0.06		
S.D. Credits (Th\$)	6,935	13	7.72	14	31,015	13	7.61	14
S.D. Balance (Th\$)	6,935	5.87	3.35	6.21	31,015	5.86	3.40	6.14
Borrower Characteristics:								
Female	6,888	0.22			30,864	0.23		
Owner Age	3,888	48	48	11	16,281	49	48	11
Business Characteristics:								
Business Age (Years)	6,935	10	7.00	7.42	31,015	11	7.95	7.69
Number of Employees	6,935	10	6.00	12	31,015	10	6.00	13
Professional Services Industries	6,935	0.29			31,015	0.33		
Capital Intensive Industries	6,935	0.31			31,015	0.26		
Retail, Food, Healthcare & Other Service Industries	6,935	0.40			31,015	0.40		
Zip Code Level Characteristics:								
Per Capita Income (Th\$)	6,935	25	26	5.58	31,015	43	39	18
Pct Black Pop (%)	6,935	12	7.50	13	31,015	11	4.90	16
Pct Hisp. Pop (%)	6,935	49	45	20	31,015	10	7.77	10
Pct Asian Pop (%)	6,935	6.45	3.30	8.59	31,015	6.66	3.40	9
Pct Bachelors (%)	6,935	23	23	9	31,015	41	40	18
Pct Home Owner (%)	6,932	53	54	17	31,010	67	71	18
Pct Unemployed (%)	6,935	6.27	5.70	2.70	31,008	4.72	4.20	2.41
Density (Pop/km2)	6,901	1,577	1,052	2,513	30,809	932	397	3,078

FIGURE B.2: HIGH-HISPANIC AND LOW-INCOME ZIP CODES

This figure uses data from Lender A and Lender B on originated loans (N = 38,021) to show the geographic distribution of High-Hispanic & Low-Income zip codes in orange and other zip codes in the data in blue.



Loans originated in zip codes with below median incomes are associated with slightly higher default rates than higher income neighborhoods (16.9% vs 16.5%). They have somewhat lower credit scores (726 vs. 730) and lower cash-flow variables such as average monthly credits (\$126K vs. \$135K), average balances (\$38k vs. \$45k), and average withdrawals (\$162k vs. \$174k). They also have a higher frequency of low or negative balances (0.45 vs. 0.39), although their averages for insufficient funds events and daily pay loans do not differ. Differences in zip code-level characteristics further highlight financial constraints, including rates of bachelor's degree attainment (25% vs. 51%), home ownership (60% vs. 69%), and unemployment (6.0% vs. 4.0%). Average density is lower than higher-income zip codes (943 vs. 1,156 population per square kilometer).

Loans originated in HBLI zip codes show further indicators of financial constraints. They are associated with higher default rates, with 18.5% of loans non-performing, compared to 16.3% in non-HBLI areas. Regarding the cash-flow variables, borrowers in HBLI zip codes have lower average monthly credits (\$122K vs. \$133K), lower average balances (\$37K vs. \$43K), and a higher frequency of low or negative balances (0.48 vs. 0.41). Differences in zip code-level characteristics also further highlight financial constraints. These areas have lower rates of home ownership (55% vs. 71%), lower bachelor's degree attainment (24% vs. 39%) and higher unemployment rates (7.3% vs. 4.5%). The average FICO score in HBLI areas (724) is also somewhat lower than in non-HBLI areas (729).

As shown in Table B.2, HBLI zip codes differ from non-HBLI zip codes. While Low-Income areas are more evenly distributed nationwide, High-Black zip codes are geographically concentrated in the South and historically Black urban centers such as Los Angeles, Oakland, Chicago, and the Northeast. Figure B.1 illustrates that racial and economic distress intersect most acutely in the South and specific urban pockets, where over 70% of High-Black zip codes also qualify as Low-Income.

Loans in HHLI zip codes also exhibit higher default rates, with 18.2% of loans becoming nonperforming, compared to 16.4% in non-HHLI areas. FICO scores are slightly lower (726 vs. 729). Cash-flow measures show modest differences, with businesses in HHLI zip codes having slightly higher average credits (\$134K vs. \$130K) and balances (\$43K vs. \$42K), but similar volatility in balances and credit flows. Borrowers in HHLI zip codes also show similar rates of insufficient funds transactions (0.06 vs. 0.05) and low or negative balances (0.43 vs. 0.42), suggesting only slightly greater financial strain.

Zip code-level characteristics reflect meaningful variation in socioeconomic conditions. By definition, these areas have a higher share of Hispanic residents (49% vs. 10%) and lower per capita income (\$25K vs. \$43K). Educational attainment is also lower, with only 23% of residents holding a bachelor's degree, compared to 41% in non-HHLI areas. Additionally, homeownership rates are lower (53% vs. 67%) and unemployment is higher (6.3% vs. 4.7%). HHLI zip codes appear predominantly in Southern California, Texas, and Florida (see Figure B.2). While financial exclusion affects both Black and Hispanic entrepreneurs, the mechanisms may differ. Hispanic entrepreneurs in our sample are often more concentrated in lower-income immigrant communities with limited formal credit access.

Regression results

Tables B.4, B.5, and B.6 present the OLS results for low score borrowers whose businesses are located in LI, HBLI, and HHLI zip codes. Results are generally directionally similar but with different intensity across the different geographies. The FICO coefficient drops with addition of the cash-flow data, except for the low score borrowers whose businesses are located in HHLI zip codes under the fully specified model (columns 2 and 6).

TABLE B.4: PREDICTING DEFAULT FOR BUSINESSES IN LOW-INCOME AREAS WITH LOW FICO OWNERS

This table uses data from Lender A and Lender B (N = 38,021) to show how credit score, cash-flow, and borrower characteristics predict default conditional on origination. Low-Income is defined as zip codes with total per capita income below the median among originated loans. Low FICO is defined as an owner having a FICO score below 700. All bank variables and FICO score are standardized to z-scores and can be interpreted as the change in the dependent variable from 1 standard deviation of change. # Insuff. Funds is the number of insufficient funds transactions. # Low or Neg. Bal. is the number of low or negative ending balances across the statements. Missing values are replaced with median values. Standard errors are clustered by industry and quarter. ***, **, * correspond to statistical significance at the 1, 5, and 10 percent levels, respectively.

DEPENDENT VARIABLE	IS NON-PERFORMING (%)							
	YES (1)	YES (2)	NO (3)	NO (4)	YES (5)	YES (6)	NO (7)	NO (8)
Low FICO & Low-Income:								
FICO	-5.89*** (1.21)	-5.45*** (1.19)	-4.63*** (0.23)	-4.23*** (0.24)	-5.40*** (1.22)	-5.10*** (1.19)	-4.49*** (0.23)	-4.23*** (0.24)
Credits (Th\$)		-3.99*** (1.06)		-2.92*** (0.38)		-3.91*** (1.18)		-2.79*** (0.37)
Withdrawals (Th\$)		3.15*** (0.89)		1.66*** (0.33)		2.57*** (0.96)		0.90*** (0.34)
Balance (Th\$)		-0.69 (0.98)		-1.28*** (0.34)		-0.74 (0.98)		-1.26*** (0.33)
1(Daily Pay Loan)		0.90 (0.57)		1.21*** (0.26)		0.86 (0.56)		0.98*** (0.26)
(#) Low or Neg. Bal.		1.23*** (0.46)		1.62*** (0.26)		1.40*** (0.48)		1.64*** (0.26)
(#) Insuff. Funds		-0.30 (0.52)		0.15 (0.25)		-0.31 (0.52)		0.11 (0.24)
S.D. Credits (Th\$)		1.91** (0.96)		1.86*** (0.35)		2.03** (0.95)		2.22*** (0.34)
S.D. Balance (Th\$)		0.51 (0.93)		1.20*** (0.36)		1.00 (0.93)		1.38*** (0.34)
Number of Employees					-0.11** (0.05)	-0.09* (0.05)	-0.08*** (0.01)	-0.05*** (0.02)
Observations	5,932	5,932	32,018	32,018	5,928	5,928	32,018	32,018
Industry FE	No	No	No	No	Yes	Yes	Yes	Yes
State FE	No	No	No	No	Yes	Yes	Yes	Yes
Quarter FE	No	No	No	No	Yes	Yes	Yes	Yes
Loan Type FE	No	No	No	No	Yes	Yes	Yes	Yes
Pseudo R-squared	0.004	0.009	0.014	0.020	0.034	0.038	0.037	0.042
R-squared Improv.		116.5		44.8		13.5		16.0
Y-mean	22.05	22.05	15.71	15.71	22.06	22.06	15.71	15.71

At the same time, the magnitude of the coefficients for credits and withdrawals is larger for the low score borrowers in the respective geographies than for other firms (columns 4 and 8). The impacts are particularly strong for low score owners whose businesses are located in HBLI areas; for instance, under the fully specified model the FICO coefficient (column 6) is smaller than in any other table and

no longer significant after the addition of the cash-flow data. We see that a one standard deviation increase in credits under the fully specified model lowers default risk by 6.3 pp relative to 2.7 for other businesses (columns 6 and 8). Improvements in R^2 are also larger for this group relative to others.

TABLE B.5: PREDICTING DEFAULT FOR BUSINESSES IN HIGH-BLACK AND LOW-INCOME AREAS WITH LOW FICO OWNERS

This table uses data from Lender A and Lender B ($N = 38,021$) to show how credit score, cash-flow, and borrower characteristics predict default conditional on origination. High-Black is defined as zip codes in the top 25th percentile for the percentage of Black residents across originated loans. Low-Income is defined as zip codes with total per capita income below the median among originated loans. Low FICO is defined as an owner having a FICO score below 700. All bank variables and FICO score are standardized to z-scores and can be interpreted as the change in the dependent variable from 1 standard deviation of change. # Insuff. Funds is the number of insufficient funds transactions. # Low or Neg. Bal. is the number of low or negative ending balances across the statements. Missing values are replaced with median values. Standard errors are clustered by industry and quarter. ***, **, * correspond to statistical significance at the 1, 5, and 10 percent levels, respectively.

DEPENDENT VARIABLE	IS NON-PERFORMING (%)							
	YES (1)	YES (2)	NO (3)	NO (4)	YES (5)	YES (6)	NO (7)	NO (8)
Low FICO & High-Black and Low-Income:								
FICO	-4.03* (2.07)	-3.32 (2.03)	-4.55*** (0.22)	-4.18*** (0.23)	-3.46 (2.14)	-3.02 (2.11)	-4.42*** (0.22)	-4.17*** (0.23)
Credits (Th\$)		-6.33*** (1.62)		-2.89*** (0.37)		-6.32*** (1.83)		-2.74*** (0.36)
Withdrawals (Th\$)		6.99*** (1.49)		1.64*** (0.33)		6.54*** (1.64)		0.89*** (0.34)
Balance (Th\$)		-0.30 (1.62)		-1.26*** (0.33)		-0.24 (1.71)		-1.26*** (0.32)
1(Daily Pay Loan)		0.17 (0.98)		1.20*** (0.25)		0.30 (0.99)		0.98*** (0.25)
(#) Low or Neg. Bal.		1.43* (0.74)		1.52*** (0.24)		1.60** (0.77)		1.57*** (0.24)
(#) Insuff. Funds		0.59 (0.97)		0.04 (0.23)		0.66 (0.98)		0.01 (0.22)
S.D. Credits (Th\$)		2.01 (1.64)		1.87*** (0.34)		1.64 (1.67)		2.22*** (0.33)
S.D. Balance (Th\$)		0.45 (1.60)		1.14*** (0.35)		0.76 (1.64)		1.33*** (0.33)
Number of Employees					-0.01 (0.08)	-0.03 (0.09)	-0.09*** (0.01)	-0.06*** (0.02)
Observations	2,230	2,230	35,720	35,720	2,225	2,225	35,718	35,718
Industry FE	No	No	No	No	Yes	Yes	Yes	Yes
State FE	No	No	No	No	Yes	Yes	Yes	Yes
Quarter FE	No	No	No	No	Yes	Yes	Yes	Yes
Loan Type FE	No	No	No	No	Yes	Yes	Yes	Yes
Pseudo R-squared	0.002	0.015	0.015	0.020	0.051	0.062	0.036	0.042
R-squared Improv.		766.4		38.9		22.6		14.9
Y-mean	23.68	23.68	16.27	16.27	23.69	23.69	16.27	16.27

TABLE B.6: PREDICTING DEFAULT FOR BUSINESSES IN HIGH-HISPANIC AND LOW-INCOME AREAS WITH LOW FICO OWNERS

This table uses data from Lender A and Lender B (N = 38,021) to show how credit score, cash-flow, and borrower characteristics predict default conditional on origination. High-Hispanic is defined as zip codes in the top 25th percentile for the percentage of Hispanic residents across originated loans. Low-Income is defined as zip codes with total per capita income below the median among originated loans. Low FICO is defined as an owner having a FICO score below 700. All bank variables and FICO score are standardized to z-scores and can be interpreted as the change in the dependent variable from 1 standard deviation of change. # Insuff. Funds is the number of insufficient funds transactions. # Low or Neg. Bal. is the number of low or negative ending balances across the statements. Missing values are replaced with median values. Standard errors are clustered by industry and quarter. ***, **, * correspond to statistical significance at the 1, 5, and 10 percent levels, respectively.

DEPENDENT VARIABLE	IS NON-PERFORMING (%)							
	YES (1)	YES (2)	NO (3)	NO (4)	YES (5)	YES (6)	NO (7)	NO (8)
Low FICO & High-Hispanic and Low-Income:								
FICO	-7.59*** (2.31)	-7.34*** (2.31)	-4.49*** (0.23)	-4.09*** (0.23)	-7.07*** (2.47)	-7.07*** (2.45)	-4.36*** (0.22)	-4.10*** (0.23)
Credits (Th\$)		-3.26** (1.56)		-3.05*** (0.37)		-3.46** (1.71)		-2.85*** (0.36)
Withdrawals (Th\$)		2.61** (1.28)		1.82*** (0.33)		1.63 (1.42)		1.05*** (0.34)
Balance (Th\$)		-2.05 (1.47)		-1.18*** (0.33)		-1.75 (1.57)		-1.17*** (0.32)
1(Daily Pay Loan)		0.82 (0.97)		1.18*** (0.26)		0.62 (0.99)		0.98*** (0.25)
(#) Low or Neg. Bal.		0.15 (0.90)		1.62*** (0.24)		0.23 (0.92)		1.66*** (0.24)
(#) Insuff. Funds		-0.50 (0.86)		0.11 (0.23)		-0.43 (0.88)		0.09 (0.23)
S.D. Credits (Th\$)		2.64 (1.61)		1.81*** (0.35)		3.05 (1.65)		2.10*** (0.34)
S.D. Balance (Th\$)		-1.52 (1.39)		1.27*** (0.36)		-1.56 (1.41)		1.48*** (0.34)
Number of Employees					-0.10 (0.08)	-0.01 (0.09)	-0.08*** (0.01)	-0.06*** (0.02)
Observations	2,172	2,172	35,778	35,778	2,168	2,168	35,777	35,777
Industry FE	No	No	No	No	Yes	Yes	Yes	Yes
State FE	No	No	No	No	Yes	Yes	Yes	Yes
Quarter FE	No	No	No	No	Yes	Yes	Yes	Yes
Loan Type FE	No	No	No	No	Yes	Yes	Yes	Yes
Pseudo R-squared	0.006	0.012	0.014	0.020	0.052	0.057	0.036	0.042
R-squared Improv.		97.3		42.9		9.5		15.9
Y-mean	23.94	23.94	16.26	16.26	23.99	23.99	16.26	16.26

Machine learning results

Table B.7 presents the results for the machine learning models as applied to low score borrowers whose businesses are located in low-income, HBLI, and HHLI zip codes. Compared to the performance gains for low FICO business owners overall as reported in Table 6 (.015 AUC and

.013 H-measure), the performance gains are similar for the low score borrowers whose businesses are located in low-income zip codes (.015 AUC and .012 H-Measure) and substantially larger for those located in zip codes that are HBLI (.019 AUC and .013 H-Measure) and HHLI (.026 AUC and .019 H-Measure). This closes or at least helps to narrow some of the predictive gaps, since the Baseline model is less reliable when applied to low score borrowers in the latter two geographies than to low score borrowers across all low-income zip codes. As expected, the improvements are also larger for the low score borrowers in the respective geographies than across all borrowers in the same zip codes. For example, across all borrowers in HBLI and HHLI zip codes, the AUC improvements are .015 and 0.012, respectively.

Table B.8 presents the TACO results for the same groups. Compared to the gains for low FICO business owners overall as reported in Table 9 (3.626), the ratios for low score owners located in LI, HBLI, and HHLI zip codes are 3.808, 4.044, and 4.325, respectively. These do not suggest the same magnitude of interaction effects as between low score and young firm, where the TACO ratio increased to 7.535 as discussed in main text.

TABLE B.7: SUPPLEMENTARY RESULTS ON MACHINE LEARNING MODEL PERFORMANCE

This table presents our performance evaluation of the Baseline and Cash-Flow random forest models for predicting loan default for additional zip code-level demographic characteristics. This table uses data from Lender A and Lender B (N = 38,021) on originated loans where loan performance is available. Performance metrics are calculated as the mean of 100 bootstrap iterations. Definitions of the performance metrics ROC AUC and H-Measure are provided in Section 5; larger numbers indicate better predictive performance. ***, **, * correspond to statistical significance at the 1, 5, and 10 percent levels, respectively.

	ROC AUC	H-MEASURE
Low FICO & Low-Income Zip		
FICO Model	0.614	0.066
Cash-Flow Model	0.629	0.077
Difference	0.015***	0.012***
Not (Low FICO & Low-Income Zip)		
FICO Model	0.652	0.083
Cash-Flow Model	0.662	0.094
Difference	0.010***	0.011***
Low FICO & High-Black and Low-Income Zip		
FICO Model	0.610	0.079
Cash-Flow Model	0.630	0.092
Difference	0.019***	0.013***
Not (Low FICO & High-Black and Low-Income Zip)		
FICO Model	0.651	0.081
Cash-Flow Model	0.662	0.092
Difference	0.010***	0.011***

	ROC AUC	H-MEASURE
Low FICO & High-Hispanic & Low-Income Zip		
FICO Model	0.584	0.060
Cash-Flow Model	0.610	0.079
Difference	0.026***	0.019***
Not (Low FICO & High-Hispanic & Low-Income Zip)		
FICO Model	0.652	0.082
Cash-Flow Model	0.663	0.093
Difference	0.011***	0.011***
Low-Income Zip		
FICO Model	0.654	0.088
Cash-Flow Model	0.664	0.099
Difference	0.011***	0.011***
High-Income Zip		
FICO Model	0.650	0.083
Cash-Flow Model	0.661	0.094
Difference	0.010***	0.011***
High-Black & Low-Income Zip		
FICO Model	0.651	0.097
Cash-Flow Model	0.666	0.110
Difference	0.015***	0.014***
Not (High-Black & Low-Income Zip)		
FICO Model	0.652	0.082
Cash-Flow Model	0.661	0.092
Difference	0.009***	0.010***
High-Hispanic & Low-Income Zip		
FICO Model	0.641	0.085
Cash-Flow Model	0.653	0.096
Difference	0.012***	0.011***
Not (High-Hispanic & Low-Income Zip)		
FICO Model	0.654	0.084
Cash-Flow Model	0.664	0.095
Difference	0.010***	0.011***

TABLE B.8: TACO RESULTS ON BENEFIT OF CASH-FLOW VS BASELINE MODEL FOR ADDITIONAL CHARACTERISTICS

This table shows the results from implementing Tail Analysis for Comparative Outcomes (TACO). We compare two random forest models to predict default: a Baseline model (containing FICO, firm size, firm age, and industry among others) and the Cash-Flow (CF) model, which adds bank statement variables to the Baseline model. The table uses data from Lender A and Lender B (N = 38,021) on originated loans. The observation counts represent the sum across 1,000 bootstrap holdout samples. The first two columns ("Tails") show the group's share in the decile tails population of bootstrapped sample observations. The next two columns restrict to the 10% of each bootstrap sample with the highest increase in predicted likelihood of default between the Baseline and the CF model, who are thus adversely affected by switching from the Baseline to the CF model. The mean shows the share of young owners in this group, which can be compared to the "Tails" mean column. The next two

columns show the same metric for the bottom 10% (the group that most benefits from switching to the CF model). The last column shows the ratio between the “Hurt” and “Benefit” means, which we call the TACO ratio. A ratio of one implies no implication of switching models, a ratio less than one implies that the group is adversely affected, and a ratio greater than one implies that the group benefits. We calculate standard errors for the TACO ratio using the percentile bootstrap.

	TAILS		TOP 10% DEFAULT INCREASES W/ CF MODEL (HURT)		BOTTOM 10% DEFAULT INCREASES W/ CF MODEL (BENEFIT)		TACO RATIO
	N	MEAN	N	MEAN	N	MEAN	
Full Sample:							
Low-Income	1,519,727	0.506	760,034	0.489	759,693	0.523	1.069
High-Black	1,522,000	0.257	761,000	0.250	761,000	0.264	1.055
High-Hispanic	1,522,000	0.256	761,000	0.237	761,000	0.275	1.162*
High-Black & Low-Income	1,519,727	0.187	760,034	0.176	759,693	0.198	1.125
High-Hispanic & Low-Income	1,519,727	0.188	760,034	0.170	759,693	0.205	1.210*
Low FICO (< 700):							
Low-Income	1,519,727	0.213	760,034	0.089	759,693	0.337	3.808**
High-Black	1,522,000	0.109	761,000	0.044	761,000	0.174	3.925**
High-Hispanic	1,522,000	0.107	761,000	0.041	761,000	0.173	4.224**
High-Black & Low-Income	1,519,727	0.082	760,034	0.032	759,693	0.131	4.044**
High-Hispanic & Low-Income	1,519,727	0.080	760,034	0.030	759,693	0.129	4.325**
High FICO (≥ 700):							
Low-Income	1,519,727	0.293	760,034	0.400	759,693	0.186	0.463**
High-Black	1,522,000	0.148	761,000	0.206	761,000	0.090	0.435**
High-Hispanic	1,522,000	0.149	761,000	0.196	761,000	0.102	0.523**
High-Black & Low-Income	1,519,727	0.106	760,034	0.144	759,693	0.067	0.466**
High-Hispanic & Low-Income	1,519,727	0.108	760,034	0.140	759,693	0.076	0.545**



Additional Acknowledgments


With support from:

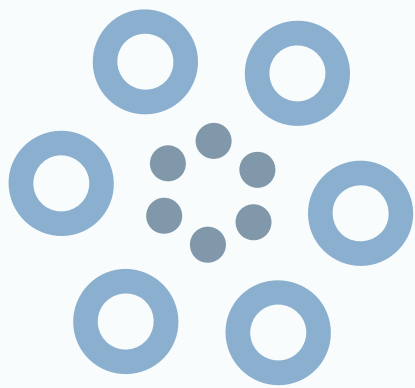
This research initiative was funded by a grant from the US Department of Commerce, Minority Business Development Agency, and support from Visa. In addition, Plaid provided account connectivity services to the pilot programs by mission-based lenders to allow loan applicants to authorize their account data to be accessed for underwriting.

The **Minority Business Development Agency** was created to promote the growth of minority business enterprises through the mobilization and advancement of public and private sector programs, policy, and research.

Visa (NYSE: V) is a world leader in digital payments, facilitating transactions between consumers, merchants, financial institutions and government entities across more than 200 countries and territories. Our mission is to connect the world through the most innovative, convenient, reliable and secure payments network, enabling individuals, businesses and economies to thrive. We believe that economies that include everyone everywhere, uplift everyone everywhere and see access as foundational to the future of money movement. Learn more at visa.com.

Plaid is a global data network that powers the tools millions of people rely on to live a healthier financial life. Our ambition is to facilitate a more inclusive, competitive, and mutually beneficial financial system by simplifying payments, revolutionizing lending, and leading the fight against fraud. Plaid works with thousands of companies including fintechs like Venmo and SoFi, several of the Fortune 500, and many of the largest banks to empower people with more choice and control over how they manage their money. Headquartered in San Francisco, Plaid's network spans over 12,000 institutions across the US, Canada, UK and Europe.





Copyright 2025 © FinRegLab, Inc.

All Rights Reserved. No part of this report may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

Digital version available at finreglab.org

Published by FinRegLab, Inc.

1701 K Street NW, Suite 1150
Washington, DC 20006
United States