The Transformative Role of Artificial Intelligence and Big Data in Banking

Binkai Chen
Central University of Finance and Economics

Dongmei Guo
Central University of Finance and Economics

Junjie Xia[1]
Central University of Finance and Economics
Peking University

Zirun Zhang
Central University of Finance and Economics

March 16, 2025

Abstract

This study examines the impact of artificial intelligence (AI) and big data on the banking sector, leveraging a comprehensive dataset of over 4.5 million loans issued by a major commercial bank in China between 2015 and 2023. Using a policy mandate to adopt AI and big data as an exogenous shock, our analysis reveals that the integration of these technologies significantly enhances credit rating accuracy and reduces loan default rates, particularly benefiting small and medium-sized enterprises. Specifically, the rate of undetermined credit ratings decreases by 2.4 percentage points, while the loan default rate drops by 2.7 percentage points. By analyzing the timeline of the bank's adoption of different technologies, we find that integrating big data with AI models and recognition technologies has a more profound impact than traditional FinTech models. Our findings highlight the informational advantages of AI and big data, providing empirical evidence of their role in enhancing operational efficiency, improving risk management capabilities, and promoting financial inclusion for small businesses in financial institutions.

JEL Classification: G20, G21, G32
Key words: artificial intelligence, big data, credit rating, default rate, interest payment

[1] Corresponding author, email: junjiexia@nsd.pku.edu.cn

1. Introduction

The integration of artificial intelligence (AI) and big data has fundamentally transformed the operations of financial institutions in recent years. These cutting-edge technologies are revolutionizing the FinTech landscape by enabling unprecedented advancements in data analysis and decision-making. While existing finance studies on AI and big data have highlighted their importance in various areas, such as fund performance, corporate culture, market microstructure, distributional effects, and vendors and small business (e.g., Easley et al., 2021; Li et al., 2021; Fuster et al., 2022; DeMiguel et al., 2023, Hau et al., 2024), there remains a relative scarcity of empirical evidence on how these technologies are reshaping banking operations.

This paper aims to address this gap by leveraging a large-scale dataset of over 4.5 million loans spanning 2015-2023 from a major state-owned commercial bank in China. This extensive dataset offers a unique opportunity to explore this topic in depth. A key strength lies in its identification strategy, which leverages an exogenous policy mandate requiring the bank to adopt FinTech in its operations.[2] This natural experiment allows us to precisely isolate the causal effects of FinTech adoption on credit ratings and risk management. Another key advantage is that we observe the exact timeline when the bank first transitioned from conventional human-driven approaches to early FinTech models, and then integrated big data along with AI models and recognition technologies. This temporal information is crucial for investigating the broader issue of comparing the effectiveness of new technologies against traditional models in the banking sector.

Specifically, the bank involves with information the bank underwent two distinct stages in adopting new technologies: In July 2019, the bank introduced machine

[2] The bank's decision to comprehensively implement AI and big data was influenced by the government's Three-year Development Plan (2019-2021) for FinTech, aimed at promoting the development of financial technologies in the banking sector. This marked the first time the Chinese government initiated a development plan specifically for FinTech. For more details, see the related policy announcements at www.pbc.gov.cn.

learning algorithms like logistic regression models. Subsequently, in October 2020, it integrated big data along with advanced AI models, including artificial neural network (ANN) and federated learning model (FLM), as well as recognition technologies like optical character recognition (OCR) and natural language processing (NLP). These distinct phases enable us to conduct a more nuanced analysis of the differences and impacts of various FinTech technologies, offering insights into their respective roles and effectiveness in transforming banking operations.

Historically, the bank has relied on traditional credit rating approaches, such as shadow ratings and hierarchical analysis, to assess borrowers' creditworthiness. While these methods have been effective to some extent, they are heavily dependent on the availability of accurate and comprehensive information. In cases where such information is incomplete or difficult to obtain, these approaches often result in a high proportion of "undetermined" or "unclassified" credit ratings. This lack of clarity forces banks to rely on subjective human judgment and experience, which can introduce biases and inconsistencies into the credit evaluation process. The prevalence of undetermined credit ratings has significant implications for borrowers. For instance, it often leads to the withdrawal of loan applications or the imposition of overly stringent loan terms, as banks seek to mitigate the perceived risks associated with insufficient information. This issue is particularly pronounced for small and medium-sized enterprises (SMEs), which typically lack the financial transparency and detailed operational data that larger corporations can provide. SMEs often operate with limited resources and less formalized reporting structures, making it challenging for banks to accurately assess their creditworthiness.

During our sample period from 2015 to 2023, a striking 89.1% of the undetermined credit ratings were attributed to SMEs. This highlights the disproportionate burden of information gaps on smaller businesses, which are already at a disadvantage in accessing credit. The inability to obtain reliable data not only hampers the credit

evaluation process but also restricts SMEs' access to much-needed financing. Addressing these information asymmetries is therefore critical to improving the inclusivity and efficiency of credit markets.

Following the government's policy mandate issued in July 2019, the bank undertook a comprehensive reform of its credit rating system, replacing traditional human decision-making approaches with advanced financial technologies, particularly AI and big data analytics. This transformation marked a significant milestone in optimizing and modernizing the bank's credit assessment processes. Among the 4.5 million loan credit ratings evaluated during this period, the overall undetermined rate dropped substantially, from approximately 6.68% to 1.99%. Notably, SMEs contributed the most to this decline, reflecting the transformative impact of these technologies on addressing the information challenges faced by smaller businesses.

The remarkable improvement can be attributed to the comparative advantages of AI and big data in gathering, processing, and analyzing vast amounts of structured and unstructured information. Unlike traditional methods, which are constrained by human capacity and the availability of formalized data, these technologies excel at extracting insights from diverse and fragmented data sources. Hau et al. (2024) provides empirical evidence supporting the information advantage of new credit technologies, demonstrating their ability to enhance decision-making in financial markets. Similarly, Livshits et al. (2016) develop a theoretical model illustrating how financial innovation mitigates the challenges of asymmetric information, particularly in the context of diverse default risks and the fixed costs associated with contract distribution.[3] In our case, the introduction of new technologies enhances the accuracy and efficiency of credit assessments by improving the gathering and processing of information and

[3] Hau et al (2019) provide similar predictions by constructing a simple model to show that FinTech credit is often used among weak creditors. Vives and Ye (2025) build a spatial model to study how information technology influences lender competition and lending to SMEs. Balyuk (2023) underscores information spillovers as a new mechanism through which FinTech lending reduces consumers' financial constraints.

unstructured data, thereby reducing the reliance on subjective human judgment and improving access to credit for SMEs.

Therefore, to further investigate how AI and big data influence credit rating, we focus on SMEs as a treatment group, given their disproportionate challenges in accessing credit due to information asymmetries. By comparing SMEs to a control group of large firms, we aim to uncover the extent to which these technologies mitigate information gaps and improve credit accessibility for underserved segments of the economy. We use the bank's initial adoption of AI and big data as an exogenous shock to the credit rating system and employ a difference-in-difference (DID) approach, which compares the changes in credit ratings before and after the adoption between SMEs and large firms. Our analysis reveals that the rate of undetermined credit ratings among SMEs decreases by 2.4 percentage points compared to large firms. This finding is consistent with our thesis that AI and big data enable the bank to enhance its information accessibility.

The information advantage of AI and big data is also evident in risk management, such as transaction monitoring and fraud detection. These technologies enhance the bank's ability to assess credit risk more accurately and efficiently by leveraging vast amounts of structured and unstructured data from diverse sources. Our estimates indicate that the loan default rate decreases by 2.7 percentage points following the introduction of these technologies.

Using the temporal differences in the adoption of technologies between the logistic regression model and the integration of big data with AI models and recognition technologies, we find that the later implementation exerts a larger effect. This is largely because using AI models to analyze big data enables the bank to consolidate a diverse range of internal and external sources. This integration provides a wealth of real-time information, empowering the bank to gain deeper insights into customer behavior, market trends, and potential risks.

We continue to explore the impact on bank credit accessibility and interest payments. We find that, compared to large firms, SMEs obtain more credit in the post-AI adoption period, indicating that the bank is more willing to provide loans when it has access to more information. Additionally, interest payments were higher among SMEs compared to large firms in the pre-AI adoption period. However, this gap has been narrowing as the bank becomes more capable of gathering comprehensive information.

We conduct a series of heterogeneity tests to further highlight how AI and big data address issues of information asymmetry and mitigate the challenges posed by incomplete information. Firstly, we observe that the impact of AI and big data integration is significantly stronger for firms with incomplete financial records. This suggests that these technologies play a critical role in bridging gaps in traditional datasets, enabling more precise credit evaluations even in the absence of robust financial disclosures. Secondly, the analysis indicates that the effect on credit ratings is especially prominent for unsecured loans, where lenders cannot rely on collateral as a safeguard against default. This finding underscores a shift in lending practices, suggesting that banks leveraging AI and big data place reduced emphasis on collateral as a means of mitigating information asymmetry and moral hazard. Thirdly, our analysis reveals that the effect is more pronounced in regions with lower levels of economic development and in areas characterized by greater linguistic diversity, such as regions with more dialects. Lastly, we observe that the results are particularly significant for firms with limited publicly available information. Together, these results emphasize the critical role of AI and big data in enhancing credit assessment processes across varying levels of information availability, particularly in contexts characterized by informational scarcity or asymmetry.

This paper contributes to the literature on the real effects of machine learning and big data technologies in financial markets. By providing empirical evidence on their transformative impact within the banking industry, we extend existing research on

FinTech's role in addressing information asymmetries and improving credit accessibility. While much of the existing research has explored how these technologies influence various domains—such as corporate culture (Li et al., 2021), board decision-making (Erel et al., 2021), fund performance (DeMiguel et al., 2023), firm growth and innovation (Babina et al., 2024), market microstructure (Easley et al., 2021), and disproportionate effects on market participants (Fuster et al., 2022)—our study extends this body of work by providing new empirical evidence on their transformative impact within the banking industry.[4] Specifically, we examine how the adoption of AI and big data technologies affects credit ratings and loan default rates, two critical dimensions of banking operations.

By focusing on these outcomes, our research highlights the potential of machine learning and big data analytics to enhance decision-making processes, reduce information asymmetries, and improve risk management in financial markets. Unlike prior studies that often emphasize the theoretical or macro-level implications of these technologies, our paper provides a granular, micro-level analysis of their operational effects within a financial institution. This contribution broadens the understanding of AI and big data's role in finance, offering insights into their practical applications and measurable benefits.

Our paper also contributes to the broader literature on the impact of FinTech on SMEs, a critical yet often underserved segment of the economy. SMEs frequently face significant barriers to accessing credit due to information asymmetries, limited credit histories, reliance on soft information that is difficult to quantify, and a lack of collateral. These challenges are well-documented in the literature (e.g., Petersen and Rajan, 1994 and Berger and Udell, 1995), and they are further exacerbated by stringent capital requirements imposed on traditional banks. Recent studies have begun to explore how

[4] Philippon (2016) discusses the potential benefits and challenges posed by FinTech in the financial services sector and explain how FinTech can improve efficiency and enhance access to financial services. Goldstein et al. (2021) also provide an excellent summary for the recent research on big data in finance.

FinTech innovations address these challenges. For instance, Frost et al. (2020) examine the factors and outcomes associated with FinTech in global finance, emphasizing that FinTech credit is particularly impactful in regions with less competitive banking sectors. Gopal and Schnabl (2022) document that finance companies and FinTech lenders increased lending to SMEs following the 2008 financial crisis, largely substituting for a reduction in traditional bank lending. Agarwal et al. (2019, 2022) finds that the adoption of mobile payment technology significantly enhances customer acquisition and facilitates business formation among small firms, while also driving an increase in consumer credit card spending. Similarly, Hau et al. (2024) utilize data from individual credit lines on Alibaba's online platform to empirically confirm that access to FinTech credit boosts vendor sales growth, transaction volumes, and customer satisfaction, particularly for vendors facing higher levels of information asymmetry regarding credit risk. These studies collectively underscore the transformative potential of FinTech in addressing the financing challenges faced by SMEs.

Our study aligns with and extends this existing body of literature by demonstrating that the adoption of FinTech, particularly AI and big data technologies, improves SMEs' access to bank credit. By leveraging advanced recognition technologies, banks can incorporate non-traditional data sources into their credit evaluation models, thereby reducing information asymmetries and enabling more accurate assessments of SME creditworthiness. Our findings highlight the role of FinTech as a catalyst for innovation in SME financing, with implications for policymakers and financial institutions seeking to promote economic growth and inclusivity.

Finally, our empirical contributes the literature on the impact of big data in finance and economics. Begenau et al. (2018) examine how big data reshapes financial markets, highlighting its role in enhancing decision-making processes, improving risk assessment, and creating competitive advantages for large firms. Similarly, Brynjolfsson and McElheran (2016) document the growing adoption of data-driven

decision-making practices across organizations. Our study provides fresh empirical evidence on the specific effects of big data adoption in the banking sector. We demonstrate how these technologies address information asymmetries, strengthen risk management, and promote financial inclusion.

This paper is organized as follows. Section 2 describes institutional background and our data sample. Section 3 discusses our empirical methodology. Section 4 provides further heterogenous analysis. Section 5 concludes the paper by emphasizing the transformative potential of AI and big data in reshaping the banking industry, with significant implications for policy and practice.

2. Background and Data

2.1 Institutional Background

Our data is sourced from one of China's largest commercial banks,[5] a pivotal institution in the country's financial sector. In 2019, the Chinese government introduced the Three-Year Development Plan (2019–2021) for FinTech, marking the first nationwide policy aimed at systematically promoting the adoption of advanced financial technologies in the banking industry. This landmark initiative underscores the government's strategic vision to modernize the financial sector, enhance risk management capabilities, and foster innovation to meet the demands of a rapidly digitizing economy.

More specifically, the Three-Year Development Plan was designed to accelerate the integration of cutting-edge technologies into traditional banking operations, enabling financial institutions to improve efficiency, reduce costs, and enhance the accuracy of decision-making processes. It reflects China's broader ambition to establish itself as a global leader in FinTech innovation. By fostering collaboration between financial

[5] Due to our data usage agreement, we are unable to disclose any information regarding the bank's identity.

institutions, technology companies, and regulatory bodies, the policy has created a conducive environment for experimentation and growth. This initiative has not only spurred technological advancements within the banking sector but also set the stage for the development of new financial products and services that cater to the diverse needs of businesses and consumers.

In response to this policy, the bank has actively embraced advanced financial technologies, including AI-driven credit assessment models and big data analytics, to align with the government's objectives and maintain its competitive edge in an increasingly technology-driven financial landscape. These efforts have not only enabled the bank to modernize its operations but also positioned it as a leader in leveraging FinTech to address critical issues such as SME financing and regional economic disparities. By adopting these technologies, the bank has demonstrated its commitment to innovation and its role as a key player in the broader transformation of China's financial ecosystem.

Traditionally, credit rating and loan evaluation processes in Chinese banks relied heavily on human decision-making through conventional methods, such as shadow ratings and hierarchical analysis. These approaches, while foundational to the banking sector, were characterized by their dependence on human judgment and the quality of data inputs. However, they also exhibited several inherent limitations that constrained their effectiveness in accurately assessing creditworthiness and managing loan risk.

First, traditional models often rely on a limited set of financial metrics and historical data, which may not capture the full picture of a borrower's creditworthiness. Human analysts are constrained by the volume of data they can process, leading to potential oversight of critical information that could indicate risk. Second, these models are typically based on fixed criteria and rules that do not easily adapt to changing market conditions or borrower circumstances. Third, due to insufficient information or ambiguous data, traditional models often result in a high number of "undetermined"

credit ratings. This uncertainty necessitates further human intervention, which can delay decision-making and lead to either overly cautious or risky lending practices. Lastly, traditional methods often face difficulties in addressing the problem of asymmetric information, where borrowers have more information about their financial situation than lenders. This can lead to adverse selection and moral hazard, increasing the likelihood of defaults.

Following the government's policy mandate in July 2019, the bank began implementing machine learning techniques, specifically utilizing logistic regression models, to replace human decision-making in its credit rating processes. By October 2020, the bank had further enhanced its capabilities by incorporating big data analytics alongside sophisticated machine learning models, including artificial neural networks (ANN) and federated learning models (FLM). Additionally, it also adopted text recognition technologies, such as Optical Character Recognition (OCR) and Natural Language Processing (NLP), to enhance data processing.

These advanced AI models are capable of processing and analyzing large volumes of structured and unstructured data much faster and more accurately than traditional methods. The big data sources include financial statements, transaction histories, social media, and external large-scale databases, such as the National Business Registration System, which contains registration information for all Chinese enterprises, and the National Intellectual Property Administration database. Furthermore, the bank has incorporated unstructured data sources, such as firm-to-firm transaction receipts, scanned documents, and images, which require advanced text recognition technologies for effective analysis.

The integration of big data analytics has also enabled real-time monitoring of borrower behavior and market conditions. For instance, the bank can now detect early warning signs of financial distress, such as sudden changes in transaction patterns, spending behavior, or cash flow irregularities. This continuous surveillance allows for

timely interventions, improving the accuracy and efficiency of the credit rating system while mitigating potential defaults. By leveraging these technologies, the bank has significantly enhanced its ability to address the limitations of traditional credit rating methods, particularly in reducing the incidence of "undetermined" credit ratings and improving decision-making processes.

Since October, 2020, the bank has fully integrated machine learning and big data analytics into its credit evaluation system. This transformation has not only improved the accuracy and efficiency of credit assessments but has also enabled the bank to better serve SMEs, which often face challenges in accessing credit due to limited financial transparency. By addressing these challenges, the bank has positioned itself as a leader in leveraging AI and big data to drive innovation, reduce risks, and promote financial inclusion in China's banking sector.

2.2 Data

Our sample comprises approximately 4.53 million loans for 475,325 firms, spanning from the beginning of 2015 to the end of 2023. This dataset is rich in loan information, including credit ratings, interest rates, and default rates. The loans cover all provinces and all industries in China.[6] Table 1 presents the summary statistics of our data sample. Panel A displays the distribution of total numbers of firms and loans across years, while Panel B compares key variables for large firms and SMEs before and after the bank's adoption of AI and big data technologies.

[Table 1 about here]

[6] Table A1 and Table A2 present the distributions by region and industry. The regional distribution aligns with the GDP-based distribution. For instance, developed provinces and districts like Guangdong province, Jiangsu province, Zhejiang province, Shandong province, Shanghai district, and Beijing district represent significant loan amounts. In terms of total numbers of loans, manufacturing accounts for about 40.5% and wholesale and retailing accounts for about 31.7%.

As shown in Table 1, the adoption of AI and big data significantly reduces the proportion of undetermined credit ratings, falling from 6.682% during the pre-adoption period to 1.992% in the post-adoption period. Prior to the adoption of these technologies, the share of undetermined credit ratings attributed to large firms was approximately 0.697% of the total, while for SMEs, it was notably higher at 5.985%. After the implementation of AI and big data, both categories experienced declines, with SMEs showing a particularly dramatic improvement. The undetermined credit rate for SMEs dropped sharply to 1.759%.

The loan default rate presents a similar pattern. The average loan default rate for large firms experienced a slight reduction from 6.31% to 5.67%, whereas for SMEs, it significantly decreased from 9.12% to 2.14%. In summary, both the rate of undetermined credit and the rate of loan defaults experience a substantial change after the implementation of AI and big data, especially for SMEs.

Another key observation in the adoption of AI and big data technologies is their role in reducing borrowing costs, particularly the disparity in interest rates between large firms and SMEs. Historically, SMEs have faced significantly higher interest rates compared to larger firms, primarily due to the higher perceived risks and information asymmetries associated with smaller organizations. This disparity has long constrained SMEs' ability to compete on a level playing field for access to credit. In the pre-adoption period, the average interest rate for large firms was approximately 4.64%, while it was 5.35% for SMEs. Following the adoption of AI and big data technologies, these rates declined for both groups, with SMEs experiencing a more pronounced decrease. The average interest rate for large firms decreased to 3.45%, whereas for SMEs, it reduced to 3.94%. This convergence in interest rates suggests that AI and big data have contributed to more equitable lending practices, potentially improving access to more favorable loan terms for SMEs.

To further explore the difference between SMEs and large firms prior to the adoption of AI and big data, we conduct a simple empirical test incorporating a series of fixed effects. The results are presented in Table 2, where Columns (1), (2), and (3) correspond to the estimates for undetermined credit ratings, loan default rates, and interest payments, respectively. The core variable of interest, *SME*, is a binary indicator that equals one if a firm is classified as an SME and zero otherwise. The coefficient for SME is positive and statistically significant at the 1% level across all three columns, indicating that before the implementation of AI and big data, SMEs faced significantly greater challenges compared to large firms. Specifically, SMEs were more likely to receive undetermined credit ratings, experience higher loan default rates, and incur higher interest payments. These results highlight the disadvantages that SMEs encounter in traditional credit evaluation systems, likely due to information asymmetry and limited access to financial resources. This finding aligns with our thesis that information asymmetry disproportionately affects SMEs, making it more challenging for them to secure favorable credit terms.

[Table 2 about here]

Therefore, these results provide an important baseline for understanding the pre-existing disparities between SMEs and large firms, and highlight the role of AI and big data to bridge the informational gap. By addressing these disparities, AI and big data technologies have the potential to transform the credit evaluation process, enabling banks to better assess the risk profiles of SMEs and offer more equitable credit terms. This not only reduces the financial burden on SMEs but also enhances their ability to contribute to economic growth and innovation.

3. Empirical analysis

3.1 Empirical specification

To investigate the impact of AI and big data on credit evaluation outcomes, we adopt a difference-in-differences (DID) methodology. In this framework, SMEs are designated as the experimental group, while large firms serve as the control group. The initial implementation of AI and big data by the bank is treated as an external shock to the credit rating system, providing a natural experiment to evaluate the causal effects of these technologies.

This approach enables us to isolate the influence of AI and big data by leveraging their inherent information advantage. Specifically, the DID framework allows us to compare changes in key outcomes between SMEs and large firms before and after the adoption of AI and big data. We first estimate the impact on credit ratings by utilizing the following regression equation:

$$UnCredit_{i,t} = \beta SME_f \times Post_t + \varphi_f + \gamma_j + \theta_t + \delta_r + \varepsilon_{i,t}, \quad (1)$$

where i indexes loan; f indexes firm; j indexes industry; t indexes time; and r indexes region. $UnCredit_{i,t}$ refers to the undetermined credit rating, an indicator that equals one if a loan application does not have a credit rating (marked as undetermined in the data) and zero otherwise. SME_f is an indicator that equals one if a firm is a SME and zero otherwise. $Post_t$ is a time indicator that equals one if the time is after the second quarter of 2019 and zero otherwise. φ_f, γ_j, θ_t, and δ_r represent the fixed effects on firm, industry, time and region, respectively.[7] $\varepsilon_{i,t}$ is the error term.

Notably, in our empirical analysis, we refrain from including firm-level control variables, such as those derived from financial statements. There are two primary reasons for this omission. Firstly, firm-level variables from financial statements are inherently integral to the credit rating process. Essentially, if a bank has access to a

[7] To account for firms that switch industries during the sample period, we incorporate industry fixed effects into our empirical model. This ensures that our results are not biased by industry-level heterogeneity or structural differences, such as variations in regulatory environments, market dynamics, or risk characteristics across industries.

firm's financial statements, the likelihood of its credit rating being classified as "undetermined" is substantially reduced. The availability of such granular financial information enables the bank to make a more informed and definitive credit assessment, thereby mitigating the uncertainty that leads to an undetermined rating. As a result, including these variables would not only be redundant but could also obscure the very phenomenon we aim to study—namely, the challenges associated with undetermined credit ratings in the absence of sufficient information.

Secondly, our dataset is primarily loan-specific and does not provide comprehensive firm-level characteristics. Introducing additional firm-level controls would substantially reduce the sample size, potentially leading to a loss of statistical power and limiting the robustness of our analysis. This trade-off between statistical validity and additional control variables would undermine the reliability and generalizability of our findings. By focusing on loan-level data and leveraging the DID design, we ensure that our analysis is both methodologically sound and empirically compelling.[8]

Therefore, we have incorporated a comprehensive set of fixed effects in our regression models. These include firm-level, industry-level, region-level, and quarter-level fixed effects, which help control for unobservable factors that might influence the outcomes of interest. By accounting for these fixed effects, we aim to capture the heterogeneity across firms, industries, and regions, as well as temporal variations, thereby reducing the omitted variable bias. Additionally, we allow for clustering of standard errors at the firm level to account for potential serial correlation within the data, ensuring that our statistical inferences remain robust.

The primary focus of our analysis is the estimate of β, which captures the effect of interest in our study. By employing these strategies, we aim to provide a thorough and

[8] By matching the city statistical yearbook data, we also include city-level control variables such as GDP and fiscal revenue. The estimates in Table A3 indicate that our baseline results are still valid.

reliable examination of the factors influencing the "undetermined" credit rating situation.

3.2 Baseline results

Table 3 presents the panel regression results analyzing the impact of AI and big data adoption on credit ratings. The primary coefficient of interest is the interaction term between *SME* and *Post*, which captures the differential effect of the technological adoption on SMEs relative to large firms. Column (1) reports the results without including any fixed effects, while Columns (2) and (3) progressively incorporate fixed effects as specified in Equation (1). Specifically, Column (3) presents our baseline results, controlling for quarter fixed effects as well as other dimensions of fixed effects, ensuring a robust estimation of the treatment effect.

[Table 3 about here]

Across all model specifications, the coefficient for the interaction term is consistently negative and statistically significant at the 1% level. This indicates a strong and reliable relationship between the adoption of AI and big data technologies and the reduction in undetermined credit ratings for SMEs. In terms of economic significance, the results suggest that the rate of undetermined credit ratings among SMEs decreases by 2.4 percentage points relative to large firms. This reduction underscores the transformative role of AI and big data in addressing the unique challenges faced by SMEs in the credit assessment process.

The findings highlight the substantial impact of advanced financial technologies in improving the accuracy, efficiency, and inclusivity of credit evaluations. SMEs, which often face greater informational asymmetries and higher barriers to accessing credit, appear to benefit disproportionately from these innovations. Traditional credit assessment methods often rely heavily on financial statements, credit histories, and

other structured data, which may be incomplete or unavailable for SMEs. By leveraging AI and big data, financial institutions can process a broader range of structured and unstructured data, such as transaction histories, online reviews, and behavioral patterns. This capability mitigates the reliance on subjective human judgment and reduces the uncertainty associated with SME creditworthiness.

To further investigate the implications of AI and big data adoption, we proceed to examine the impact of AI and big data adoption on loan default rates by modifying our regression model to use *Default* as the dependent variable in Equation (1). In this context, *Default* is an indicator that equals one if the loan is defaulted and zero otherwise. Table 4 presents the corresponding estimation results. Across all model specifications, the coefficient for the interaction term between *SME* and *Post* is consistently negative and statistically significant at the 1% level. This finding indicates a strong relationship between the adoption of AI and big data technologies and a reduction in loan default rates, particularly for SMEs. In terms of economic magnitude, the results suggest that, compared to large firms, the loan default rate for SMEs decreases by 2.7 percentage points.

[Table 4 about here]

This empirical finding underscores the effectiveness of AI and big data in improving credit risk assessment and mitigating default risks, and highlights the transformative potential of advanced financial technologies in addressing the unique challenges faced by SMEs in the credit market. SMEs often face higher default risks due to limited access to formalized financial data, greater informational asymmetries, and a lack of collateral or credit history. By leveraging AI and big data, financial institutions can incorporate a wider range of data sources, including non-traditional and unstructured data, into their credit risk models. This expanded scope enables a more nuanced and accurate assessment of borrower creditworthiness, reducing the likelihood of misclassification and improving the overall quality of lending decisions.

3.2 Robustness checks

We conduct a series of robustness checks to confirm the validity and consistency of our baseline results. First, we perform a parallel-trend test to validate the key identification assumption. Figure 1 illustrates the dynamic responses to the introduction of AI and big data. Specifically, Panel A and Panel B present the corresponding estimates for undetermined credit ratings and loan default rates, respectively. Each dot in the figure represents the estimated coefficient, along with the associated 95% confidence intervals, derived from the leads and lags regression specified in Equation (1) of the paper. The comparison group is set to time -1, representing the period immediately prior to the adoption of AI and big data.

[Figure 1 about here]

Both panels reveal no significant pre-trend in the outcomes prior to the adoption of these technologies, indicating that the parallel trends assumption holds. Furthermore, there is a clear and substantial shift in both the magnitude and statistical significance of the coefficients following the adoption of AI and big data. This shift becomes particularly pronounced after the external shock, suggesting that the introduction of these technologies had a meaningful impact on the observed outcomes.

Notably, while the rate of undetermined credit ratings demonstrates an immediate and substantial decline following adoption, the reduction in loan default rates emerges with a visible lag, becoming statistically significant in the second period. This lag is consistent with the nature of default rates, which typically require a longer observation horizon to reflect the effects of upstream improvements in credit assessment processes. Overall, these findings provide strong evidence in support of our identification strategy and reinforce the robustness of our results.

Second, we do not observe similar results from placebo tests conducted using non-existent time periods, where the adoption of AI and big data is assumed to have taken place in alternative, non-existent time periods. Table 5 presents the corresponding results, where we hypothetically set the implementation of AI to one year earlier—specifically, in the first or second quarter of 2018. Our analysis indicates that the coefficient of the core variable is either statistically insignificant or exhibits a very small magnitude for both undetermined credit ratings and loan default rates. These findings reinforce the robustness of our main results and suggest that the observed effects are indeed attributable to the actual timing of the AI implementation.

[Table 5 about here]

Third, to supplement the findings from the prior placebo test, we conduct an additional robustness check using the Monte Carlo permutation method. Specifically, we randomly assign individual observations to the treatment group and repeat the regression analysis 500 times, generating 500 sets of regression results (including the estimated coefficients, standard errors, and p-values). We plot the distribution of the 500 estimated coefficients alongside their corresponding p-values to visually illustrate the results of the placebo test.

Figure 2 presents the results, with Panel A showing the distribution for undetermined credit ratings and Panel B displaying the distribution for loan default rates. In both panels, the distributions are centered around zero, indicating no systematic bias in the placebo tests. Furthermore, the estimated coefficients from our baseline analysis (-0.024 for undetermined credit ratings and -0.027 for loan default rates) are significantly smaller than the values observed in the placebo distributions, as shown on the horizontal axis. These findings provide strong evidence supporting the validity of our baseline estimates for undetermined credit ratings and loan default rates.

[Figure 2 about here]

Fourth, a potential alternative explanation for the decline in undetermined credit ratings and loan default rates is that these improvements might not be driven by the adoption of AI and big data, but rather by other contemporaneous policies aimed at supporting SMEs. For example, during our sample period, there might exist policies, such as industrial support programs or tax incentives, have contributed to improving the operational performance and financial health of SMEs, thereby leading to higher credit ratings and lower default rates independent of the adoption of AI and big data technologies.

To address this concern, we shift the focus of our analysis from the firm level to the regional level to account for such overlapping policy effects. Specifically, we identify regions with the highest levels of undetermined credit ratings prior to the adoption of AI and big data and designate these regions as the treated group in our DID framework. Specifically, we classify the top 5% (or 10%) of regions with the highest pre-adoption undetermined credit rating rates as the treatment group. This regional-level approach mitigates the influence of SME-specific policies while isolating the role of AI and big data in addressing undetermined credit ratings.

The results, as reported in Table 6, show that the interaction term between the regional treatment indicator, *Region* and the post-adoption period indicator, *Post* is negative and statistically significant for both undetermined credit ratings and loan default rates. These findings, consistent with the baseline results, underscore the role of AI and big data in effectively addressing information asymmetry and improving financial outcomes. This analysis mitigates the potential bias arising from contemporaneous SME policies, reinforcing the conclusion that the integration of AI and big data technologies was instrumental in achieving these improvements.

[Table 6 about here]

3.3 Credit accessibility and interest payment

We next to investigate the influence of AI and big data adoption on SMEs' access to bank loans and their borrowing costs. Specifically, we modify our regression model to use *Loan amount* and *Interest* as the dependent variables in Equation (1). Here, *Loan amount* refers to the logarithm of the quarterly total sum of all loans, while *Interest* represents the interest rate of a loan. Given that AI and big data enable banks to gather more comprehensive information and make more accurate assessments of SMEs' creditworthiness, it is anticipated that SMEs will experience improved access to bank credit while benefiting from reduced borrowing costs.

Table 7 presents the corresponding estimation results, which align closely with the findings in Table 3 and Table 4. The coefficient for the interaction term between *SME* and *Post* is negative and statistically significant at the 1% level across all specifications, even after incorporating various dimensions of fixed effects. Specifically, the findings indicate that, compared to large firms, the interest rate for SMEs decreases by 0.323 percentage points following the adoption of AI and big data technologies. This reduction suggests that the gap in borrowing costs between SMEs and large firms has narrowed, highlighting the potential of AI and big data to enhance credit assessment and reduce financial burdens for smaller businesses.

[Table 7 about here]

The results provide compelling evidence of the transformative role of AI and big data in improving financial inclusion for SMEs. By leveraging these technologies, the bank can process a broader range of data, including alternative and non-traditional data sources, to better evaluate the creditworthiness of SMEs. This enhanced assessment reduces the perceived risk associated with lending to SMEs, enabling banks to extend more credit at lower interest rates. For SMEs, this may translate into improved access to financial resources, which can be critical for their growth, innovation, and competitiveness.

The reduction in interest rates for SMEs has significant implications for their financial sustainability and long-term viability. Lower borrowing costs alleviate the financial strain on SMEs, allowing them to allocate more resources toward productive investments, such as technology upgrades, workforce expansion, and market development. This, in turn, enhances their ability to compete with larger firms and contribute to broader economic growth. Furthermore, the narrowing of the borrowing cost gap between SMEs and large firms reflects a more equitable financial system, where smaller businesses are no longer disproportionately disadvantaged due to informational asymmetries or perceived riskiness.

3.4 Early Machine learning models verse big data analytics

The bank experienced two significant phases in its adoption of AI and big data technologies. The first phase commenced in July 2019, when the bank introduced machine learning techniques, specifically employing a logistic regression model to enhance its credit evaluation processes. The second phase followed in October 2020, during which the bank further advanced its technological capabilities by integrating big data analytics and incorporating sophisticated text recognition technologies, such as optical character recognition (OCR) and natural language processing (NLP). These advancements enabled the bank to process unstructured and semi-structured data, such as scanned documents, contracts, and textual information, thereby broadening the scope of its credit evaluation framework. Recent literature has highlighted that large language models are particularly effective at extracting valuable insights from various types of unstructured data (e.g., Bartik et al., 2023; Gabaix et al., 2023; Costello et al., 2024).

We leverage this temporal difference in the adoption of financial technologies to examine how these distinct innovations differentially impact banking operations. To capture these effects, we introduce an additional interaction term into Equation (1). This approach allows us to isolate and analyze the distinct contributions of machine learning and big data analytics to the bank's operational efficiency and decision-making

processes. Accordingly, we estimate the following equation to analyze these impacts in detail.

$$Y_{i,t} = \beta_1 SME_f \times Post1_t + \beta_2 SME_f \times Post2_t + \varphi_f + \gamma_j + \theta_t + \delta_r + \varepsilon_{i,t}, \quad (2)$$

where i indexes loan; f indexes firm; j indexes industry; t indexes time; and r indexes region. The dependent variable $Y_{i,t}$ refers to the undetermined credit rating and loan default rate. Undetermined credit rating is an indicator that equals one if a loan application does not have a credit rating (marked as undetermined in the data) and zero otherwise. SME_i is an indicator that equals one if a firm is a SME and zero otherwise. $Post1_t$ is a time indicator that equals one if the time is after the second quarter of 2019 and zero otherwise. $Post2_t$ is a time indicator that equals one if the time is after the third quarter of 2020 and zero otherwise. φ_f, γ_j, θ_t, and δ_r represent the fixed effects on firm, industry, time and region, respectively. $\varepsilon_{i,t}$ is the error term.

Specifically, the inclusion of the interaction term allows us to estimate Equation (2), which distinguishes between the effects of the machine learning phase (July 2019 onward) and the big data analytics phase (October 2020 onward). By doing so, we can assess whether the incremental adoption of big data analytics and text recognition technologies generates additional benefits beyond those achieved through the initial implementation of machine learning techniques. This distinction is critical for understanding the complementary and potentially synergistic effects of these technologies on the bank's performance.

Table 8 presents the corresponding estimation results, focusing on the key coefficients of interest for the two interaction terms, which capture the distinct effects of the bank's phased adoption of machine learning and big data technologies. Column (1) reports the estimates for undetermined credit ratings, a key indicator of the bank's ability to classify borrowers' creditworthiness. The coefficient for the first interaction term, representing the adoption of machine learning techniques, is -0.016, while the

coefficient for the second interaction term, reflecting the integration of big data analytics and advanced recognition technologies, is -0.02. Both coefficients are statistically significant at the 1% level, indicating that both phases of technological adoption contributed to reducing the incidence of undetermined credit ratings. Notably, the larger magnitude of the second coefficient suggests that the adoption of big data analytics and advanced recognition technologies had a more substantial impact than the initial implementation of machine learning.

[Table 8 about here]

Since the third quarter of 2020, the bank has fully integrated machine learning and big data analytics into its credit rating system. By summing the coefficients of the two interaction terms (0.016 + 0.02 = 0.036), we estimate the combined effects of these technologies. This implies that the incidence of undetermined credit ratings decreases by 3.6 percentage points following the bank's implementation of these advanced technologies. This finding underscores the enhanced accuracy and efficiency in credit rating processes achieved through the complementary use of machine learning and big data analytics. The results highlight the transformative potential of combining structured and unstructured data sources, as well as the ability of advanced recognition technologies, such as OCR and NLP, to process non-traditional data inputs, thereby improving the bank's ability to classify borrowers more effectively.

Column (2) presents the estimates for loan default rates, a critical measure of the bank's risk management performance. The coefficient for the first interaction term, associated with the adoption of machine learning, is -0.015, but it is not statistically significant. In contrast, the coefficient for the second interaction term, representing the integration of big data analytics and advanced recognition technologies, is -0.028 and statistically significant. This suggests that the reduction in loan default rates is primarily driven by the second phase of technological adoption, where the bank incorporated more sophisticated tools to enhance its credit evaluation processes.

In terms of economic impact, the results indicate that the loan default rate decreases by 2.8 percentage points following the adoption of big data analytics and advanced recognition technologies. This finding highlights the effectiveness of integrating more complex technologies in reducing the risk of loan defaults, thereby enhancing the bank's overall risk management capabilities. The ability to process and analyze a broader range of data, including unstructured and alternative data sources, likely contributed to more accurate credit assessments and better-informed lending decisions, reducing the likelihood of defaults.

Overall, the results from Table 8 provide compelling evidence of the incremental benefits of adopting advanced financial technologies in a phased manner. While the initial implementation of machine learning techniques improved the bank's credit evaluation processes, the subsequent integration of big data analytics and advanced recognition technologies delivered more substantial improvements. This suggests that the combination of these technologies is not merely additive but potentially synergistic, as the capabilities of big data analytics build upon and enhance the foundation established by machine learning.

From a practical perspective, these findings underscore the importance of leveraging advanced recognition technologies, such as OCR and NLP, to process non-traditional data sources. By incorporating such data into credit evaluation models, banks can better assess the creditworthiness of borrowers, particularly SMEs and underbanked clients, who may lack formalized financial records. This not only improves the accuracy of credit ratings but also promotes financial inclusion by enabling banks to extend credit to a broader range of clients.

4. Heterogeneous analysis

To further substantiate the informational advantages provided by AI and big data, we perform five sets of heterogeneity analyses. These analyses aim to deepen our

understanding of how adopting these advanced technologies mitigates information asymmetries and improves credit evaluation processes under varying conditions.

We begin by examining the effect of AI and big data on firms with missing critical financial information. Metrics such as firm cash flow, sales, and profits are pivotal in the bank's lending decisions. Lian and Ma (2021) estimate that approximately 80% of corporate debt decisions rely on cash flows generated from firm operations. Therefore, we hypothesize that AI and big data have a disproportionately larger effect on firms lacking financial information since these technologies enable banks to gather additional soft and hard information for credit rating decisions.

To test this hypothesis, we construct a binary indicator that equals one if a firm is missing information on annual sales and profit, and zero otherwise. This dummy variable is then interacted with our core term—the interaction between *SME* and *post*—to perform a triple-difference (DDD) analysis. The results, reported in Column (1) of Table 9, reveal that the coefficient for the DDD estimator is negative and statistically significant. This finding demonstrates that the reduction in undetermined credit ratings is more pronounced for firms with missing financial information, and also suggest that the bank increasingly relies on AI and big data to address information asymmetries.

[Table 9 about here]

Next, we investigate the heterogeneity in loan types by categorizing loans as either secured or unsecured based on the availability of collateral. Collateral traditionally serves as a guarantee for lenders, mitigating risks by reducing reliance on borrower soft information. However, unsecured loans lack such guarantees and are inherently more reliant on borrower-specific soft information, such as behavioral patterns, reputation, and other non-financial indicators. We hypothesize that the transformative impact of AI and big data is more pronounced for unsecured loans, as these technologies are better

equipped to process and analyze soft information, thereby addressing information asymmetries.

Consistent with our hypothesis, the results in Column (2) in Table 9 show that the reduction in undetermined credit ratings is more pronounced for unsecured loans compared to secured loans, and also suggest that the bank increasingly relies on AI and big data to mitigate information asymmetries and address moral hazard issues, rather than depending solely on collateral as a risk mitigation tool. This aligns with Aghion and Bolton (1992), which highlights the limitations of collateral in resolving information asymmetries and moral hazard. By leveraging advanced technologies, the bank can better assess the creditworthiness of borrowers, particularly in cases where collateral is absent, thereby improving the efficiency and inclusivity of its lending practices.

In addition, we employ two region-level proxies to evaluate information availability: the level of economic development and linguistic diversity. We hypothesize that firms located in less developed cities face greater information asymmetries due to weaker financial infrastructure, less transparent markets, and limited access to formal financial records. In contrast, firms in more developed regions benefit from more robust financial markets and greater availability of reliable information. To test this, we construct a binary indicator that equals one if a firm is located in a less developed city, and zero otherwise.

The diversity of spoken dialects within a region reflects cultural and linguistic heterogeneity, which can add further layers of complexity to the information environment. Existing studies (e.g., Falck et al., 2012; Desmet et al., 2017) highlight that greater linguistic diversity complicates interpersonal networks, making it more difficult for lenders to collect and interpret reliable information from borrowers. Using the number of dialects spoken in a city as a proxy for linguistic diversity, we construct

a binary indicator equal to one if more than two dialects are spoken in a given city and zero otherwise.

The corresponding results, presented in Columns (3) and (4) of Table 9, reveal that the DDD estimator is negative and statistically significant for both proxies. These findings align with our hypothesis, suggesting that AI and big data provide a significant informational advantage in regions where traditional information collection is hindered by lower economic development or greater linguistic diversity. This highlights the potential of AI-driven technologies to mitigate information asymmetries and improve decision-making in complex environments.

Finally, we examine the role of firm ownership structure in shaping the availability of information and its implications for credit assessment. SOEs tend to have greater public transparency because they are subject to government mandates requiring the disclosure of corporate information. Thus, if a borrower is a SOE, it is more likely to have publicly available information. To capture this distinction, we construct a binary indicator that equals one if a firm is not state-owned, and zero otherwise. The results, presented in Column (5) of Table 9, indicate that the DDD estimator is negative and statistically significant. This finding further supports our hypothesis that AI and big data provide an informational advantage.

In addition to credit ratings, we analyze loan default rates to further validate our findings. Table 10 presents the results, which are consistent with those in Table 9. However, one notable distinction is that the reduction in default rates for SOEs is slightly lower compared to that of non-SOEs. This difference may be attributed to the intrinsic characteristics of firm ownership structures. For instance, SOEs often benefit from government backing, which may reduce their urgency to maximize efficiency or profitability. This implicit support could lead to moral hazard, influencing their financial performance and risk-taking behavior. In contrast, non-SOEs, which operate

under greater market discipline, may adopt more prudent financial practices to ensure sustainability and competitiveness.

[Table 10 about here]

Overall, these findings provide robust evidence supporting our hypothesis regarding the informational advantage of AI and big data. In environments characterized by limited publicly available information, these technologies significantly improve the accuracy and reliability of credit assessments. Our analyses highlight the transformative potential of AI and big data in overcoming information barriers, thereby enhancing decision-making processes in financial institutions.

5. Conclusion

In conclusion, this study provides compelling evidence of the transformative impact of AI and big data on the banking industry, particularly in enhancing credit assessment processes. By analyzing a comprehensive dataset from a major commercial bank in China, we demonstrate that the integration of these technologies significantly reduces the prevalence of "undetermined" credit ratings, a challenge that has historically hindered effective credit evaluation, especially among SMEs. This reduction is achieved through improved accuracy and efficiency in credit assessments, facilitated by the advanced data processing capabilities of AI and big data.

Our findings reveal that the adoption of big data analytics, in conjunction with machine learning algorithms, not only decreases the rate of undetermined credit ratings but also contributes to a lower loan default rate. Additionally, these technologies help narrow the gaps in credit accessibility and interest payments between SMEs and larger firms. These outcomes underscore the potential of AI and big data to address long-standing issues of information asymmetry, thereby enhancing decision-making processes and fostering a more equitable financial ecosystem.

This paper makes several important contributions to the literature and practice. First, it provides robust empirical evidence on the causal effects of FinTech adoption, leveraging a natural experiment driven by an exogenous policy mandate. By isolating the impact of AI and big data on credit ratings and risk management, our research offers a clear framework for understanding how technological innovations are reshaping the financial sector. Second, our analysis highlights the temporal evolution of FinTech adoption, from early models to advanced AI-driven systems and big data, offering valuable insights into the comparative effectiveness of these technologies over time. Third, our heterogeneity analyses reveal that the benefits of AI and big data are particularly pronounced in regions with lower levels of economic development, areas with greater linguistic diversity, and among firms with limited publicly available information. These findings emphasize the broader applicability of these technologies across diverse contexts and their potential to democratize access to credit.

Additionally, our findings provide important implications for policymakers and financial institutions. First, they highlight the importance of promoting the adoption of AI and big data technologies in the banking sector, particularly in regions and among populations that have historically faced barriers to credit access. Policymakers could consider providing incentives, such as subsidies or tax breaks, to encourage financial institutions to invest in these technologies. Second, the results underscore the need for regulatory frameworks that support the ethical and responsible use of AI and big data in financial decision-making. Ensuring transparency, fairness, and accountability in the deployment of these technologies will be critical to maximizing their benefits while minimizing potential risks, such as algorithmic bias.

While this study provides valuable insights, it also opens up several avenues for future research. First, future studies could explore the long-term effects of AI and big data adoption on SME growth, financial stability, and market competitiveness. Understanding how these technologies influence firm performance and broader

economic outcomes over time would provide a more comprehensive picture of their impact. Second, it would be valuable to investigate whether similar benefits can be observed in other sectors or regions, particularly in developing economies where access to credit remains a significant barrier to growth. Comparative studies across different institutional and regulatory environments could yield important insights into the conditions under which these technologies are most effective. Finally, further research could examine the potential for emerging technologies, such as blockchain and decentralized finance (DeFi), to complement existing FinTech solutions. These innovations could offer additional pathways for improving financial inclusion, reducing transaction costs, and enhancing the efficiency of financial systems.

Overall, this paper contributes to the growing body of literature on the real-world effects of technological integration in finance. By bridging the literature on AI and big data with research on FinTech and SME financing, we provide a comprehensive framework for understanding how these innovations are reshaping the financial sector. Our findings not only advance academic knowledge but also offer actionable insights for policymakers and practitioners seeking to harness the potential of AI and big data to build a more inclusive, efficient, and sustainable financial system.

Reference

Aghion, P., and P. Bolton. 1992. An Incomplete Contracts Approach to Financial Contracting. *Review of Economic Studies* 59(3): 473-494.

Agarwal, S., W. Qian, Y. Ren, H.T. Tsai, and B.Y. Yeung. 2022. The Real Impact of FinTech: Evidence from Mobile Payment Technology. Working Paper.

Agarwal, S., W. Qian, B.Y. Yeung, and X. Zou. 2019. Mobile Wallet and Entrepreneurial Growth. *AEA Papers and Proceedings* 109: 48-53.

Babina, T., A. Fedyk, A. He, and J. Hodson. 2024. Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics* 151, 103745

Balyuk, T. 2023. FinTech lending and bank credit access for consumers. *Management Science* 69(1): 555-575.

Bartik, Alexander, Arpit Gupta, and Daniel Milo. 2023. The costs of housing regulation: Evidence from generative regulatory measurement. Available at SSRN 4627587.

Begenau, J., M. Farboodi and L. Veldkamp. 2018. Big Data in Finance and the Growth of Large Firms. *Journal of Monetary Economics* 97: 71-87.

Berger, A. N., and G. F. Udell. 1995. Relationship Lending and Lines of Credit in Small Firm Finance. *The Journal of Business* 68(3): 351-381.

Brynjolfsson, E., and K. McElheran. 2016. The Rapid Adoption of Data-Driven Decision-Making. *American Economic Review* 106(5): 133–39.

Costello, A., B. Levy and V. Nikolaev. 2024. Uncovering information: Can AI tell us where to look? Working paper.

DeMiguel, V., J. Gil-Bazo, F. J. Nogales, and A. A.P. Santos. 2023. Machine learning and fund characteristics help to select mutual funds with positive alpha. *Journal of Financial Economics* 150(3), 103737

Desmet K, I. Ortuño-Ortín, and R. Wacziar. 2017 Culture, ethnicity, and diversity. *American Economic Review* 107(9): 2479-2513.

Easley, D., M. Lopez de Prado, M. O'Hara, and Z. Zhang. 2021. Microstructure in the machine age. *Review of Financial Studies* 34: 3316–63.

Erel, I., L. Stern, C. Tan, and M. S. Weisbach. 2021. Selecting directors using machine learning. *Review of Financial Studies* 34: 3226–64.

Falck O., S. Heblich, A. Lameli, and J. Sudekum. 2012. Dialects, cultural identity, and economic exchange. *Journal of Urban Economics* 72(2-3): 225-239

Frost, J., L. Gambacorta, Y. Huang, H. S. Shin, and P. Zbinden. 2020. BigTech and the changing structure of financial intermediation. *Economic Policy* 34(100): 761-799.

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther. 2022. Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance* 77(1): 1-808.

Gabaix, X., R. SJ Koijen, R. Richmond, and M. Yogo, 2023. Asset embeddings. Available at SSRN 4507511.

Goldstein, I., C.S. Spatt, and M. Ye. 2021. Big data in finance. *Review of Financial Studies* 34: 3213–325

Gopal, M., and Schnabl, P., 2022. The rise of finance companies and fintech lenders in small business lending. *Review of Financial Studies* 35 (11), 4859–4901.

Hau, H., Y. Huang, H. Shan, and Z. Sheng. 2019. How FinTech Enters China's Credit Market. *AEA Papers and Proceedings* 109: 60-64.

Hau H., Y. Huang, C. Lin, H. Shan, Z. Sheng, L. Wei. 2024. FinTech credit and entrepreneurial growth. *Journal of Finance* 79(5): 3309-3359.

Li, K., F. Mai, R. Shen, and X. Yan. 2021. Measuring corporate culture using machine learning. *Review of Financial Studies* 34: 3265–315.

Lian, C., and Y. Ma. 2021. Anatomy of corporate borrowing constraints. *Quarterly Journal of Economics* 136(1): 229-291.

Livshits, I., J. C. Mac Gee, and M. Tertilt. 2016. The democratization of credit and the rise in consumer bankruptcies. *Review of Economic Studies* 83(4): 1673-1710.

Petersen, M. A., and R. G. Rajan. 1994. The Benefits of Lending Relationships: Evidence from Small Business Data. *Journal of Finance* 49(1):3-37.

Philippon, T. 2016. The FinTech Opportunity. NBER Working Paper 22476.

Vives, X., and Z. Ye. 2025. Information technology and lender competition. *Journal of Financial Economics* 163, 103957.

Table 1 – Summary Statistics

Panel A: Loan and firm distribution

Year	2015	2016	2017	2018	2019	2020	2021	2022	2023	Total
Firms	95291	79448	75266	80416	74611	73353	120429	166237	254644	475325
Loans	417163	333368	321521	352866	305670	281315	523831	776723	1217431	4529888

Note: This table presents the summary statistics for the total number of loans and firms in the data sample from 2015 to 2023.

Panel B: Comparison between large firms and SMEs

	Before			After		
	Overall	Large	SMEs	Overall	Large	SMEs
Number of Firms	170386	7395	162991	374088	4360	369728
Number of Loans	1574635	176504	1398131	2955293	53094	2902199
Undetermined credit rating loans	105221	10978	94243	58863	6879	51984
Rate of undetermined credit rating	6.682%	0.697%	5.985%	1.992%	0.233%	1.759%

Note: This table presents the summary statistics for the data sample from 2015 to 2023. Before refers to the pre-adoption of AI and big data period. After refers to the post-adoption of AI and big data period. Rate of undermined credit rating is the ratio of the number of undetermined credit rating loans to the number of overall undetermined credit rating loans.

Table 2 – Comparison between SMEs and large firms

Variables	*Undetermined Credit Rating*	*Loan Default Rate*	*Interest Payment*
	(1)	(2)	(3)
SME	0.046***	0.025***	0.582***
	(16.40)	(4.50)	(15.36)
Constant	0.019***	0.068***	4.687***
	(7.48)	(12.62)	(119.72)
Firm F.E.	NO	NO	NO
Industry F.E.	YES	YES	YES
Region F.E.	YES	YES	YES
Quarter F.E.	YES	YES	YES
Observations	1,563,285	1,550,496	1,562,563
R^2	0.071	0.071	0.396

Note: This table presents the panel regression results on the difference between SMEs and large firms prior to the adoption of AI and big data. Undetermined credit rating, an indicator that equals one if a loan application does not have a credit rating (marked as undetermined in the data) and zero otherwise. Loan default rate, an indicator that equals one if a loan is defaulted and zero otherwise. Interest payment refers to the interest rate of a loan. *SME* is an indicator that equals one if a firm is a SME and zero otherwise. T-statistics values are shown in parentheses. The superscript ***, **, or * indicates statistical significance at the 1%, 5% or 10% level, respectively.

Table 3 – Credit rating

Variables	*Dependent Variable: Undetermined Credit Rating*		
	(1)	(2)	(3)
SME × *Post*	-0.117***	-0.025***	-0.024***
	(-3.85)	(-6.76)	(-5.63)
Post	0.067**	0.015***	
	(2.22)	(3.71)	
SME	0.005		
	(0.18)		
Constant	0.062**	0.036***	0.045***
	(2.13)	(49.66)	(16.46)
Firm F.E.	NO	YES	YES
Industry F.E.	NO	YES	YES
Region F.E.	NO	YES	YES
Year F.E.	NO	YES	NO
Quarter F.E.	NO	NO	YES
Observations	4,529,928	4,378,877	4,378,877
R^2	0.018	0.703	0.706

Note: This table presents the panel regression results on the influence of the adoption of AI and big data on credit rating. The dependent variable is undetermined credit rating, an indicator that equals one if a loan application does not have a credit rating (marked as undetermined in the data) and zero otherwise. *SME* is an indicator that equals one if a firm is a SME and zero otherwise. *Post* is a time indicator that equals one if the time is after the second quarter of 2019 and zero otherwise. T-statistics values are shown in parentheses. The superscript ***, **, or * indicates statistical significance at the 1%, 5% or 10% level, respectively.

Table 4 – Loan default rate

Variables	*Dependent Variable: Loan Default Rate*		
	(1)	(2)	(3)
SME × *Post*	-0.062***	-0.027**	-0.027**
	(-5.19)	(-2.01)	(-2.12)
Post	-0.015	0.023*	
	(-1.29)	(1.76)	
SME	0.029*		
	(1.76)		
Constant	0.065***	0.044***	0.059***
	(3.99)	(62.70)	(7.19)
Firm F.E.	NO	YES	YES
Industry F.E.	NO	YES	YES
Region F.E.	NO	YES	YES
Year F.E.	NO	YES	NO
Quarter F.E.	NO	NO	YES
Observations	4,507,689	4,358,049	4,358,049
R^2	0.031	0.707	0.708

Note: This table presents the panel regression results on the influence of the adoption of AI and big data on the loan default rate. The dependent variable is loan default rate, an indicator that equals one if a loan is defaulted and zero otherwise. *SME* is an indicator that equals one if a firm is a SME and zero otherwise. *Post* is a time indicator that equals one if the time is after the second quarter of 2019 and zero otherwise. T-statistics values are shown in parentheses. The superscript ***, **, or * indicates statistical significance at the 1%, 5% or 10% level, respectively.

Table 5 – Placebo test (non-exist time)

Variables	*Undetermined Credit Rating*		*Loan Default Rate*	
	(1)	(2)	(3)	(4)
	2018Q1	2018Q2	2018Q1	2018Q2
SME × *Post*	-0.001	-0.001	-0.001	0.000
	(-0.70)	(-0.48)	(-0.11)	(0.03)
Constant	0.069***	0.069***	0.062***	0.059***
	(116.19)	(122.53)	(23.75)	(21.28)
Firm F.E.	YES	YES	YES	YES
Industry F.E.	YES	YES	YES	YES
Region F.E.	YES	YES	YES	YES
Quarter F.E.	YES	YES	YES	YES
Observations	635,898	628,293	629,878	622,978
R^2	0.932	0.918	0.853	0.854

Note: This table presents the panel regression results on the influence of the adoption of AI and big data on undetermined credit rating and loan default rate. The data sample is based on the pre and post four quarters of 2018Q1/2018Q2. Undetermined credit rating, an indicator that equals one if a loan application does not have a credit rating (marked as undetermined in the data) and zero otherwise. Loan default rate, an indicator that equals one if a loan is defaulted and zero otherwise. *SME* is an indicator that equals one if a firm is a SME and zero otherwise. *Post* is a time indicator that equals one if the time is after the first or second quarter of 2018 and zero otherwise. T-statistics values are shown in parentheses. The superscript ***, **, or * indicates statistical significance at the 1%, 5% or 10% level, respectively.

Table 6 – Regional test

Variables	>= top 5% Region		>= top 10% Region	
	Credit Rating	Default Rates	Credit Rating	Default Rates
	(1)	(2)	(3)	(3)
Region × *Post*	-0.069***	-0.005***	-0.035***	-0.007***
	(-56.63)	(-4.74)	(-24.05)	(-5.74)
Constant	0.067***	0.044***	0.046***	0.044***
	(120.15)	(100.09)	(106.73)	(118.48)
Region F.E.	YES	YES	YES	YES
Time F.E.	YES	YES	YES	YES
Observations	4,529,928	4,507,689	4,507,689	4,507,689
R^2	0.044	0.060	0.041	0.060

Note: This table presents the panel regression results on the influence of the adoption of AI and big data on undetermined credit rating and loan default rate. The data sample is based on the pre and post four quarters of 2018Q1/2018Q2. Undetermined credit rating, an indicator that equals one if a loan application does not have a credit rating (marked as undetermined in the data) and zero otherwise. Loan default rate, an indicator that equals one if a loan is defaulted and zero otherwise. *Region* is an indicator that equals one if a region is among the top 5% (or 10%) of regions with the highest pre-adoption undetermined credit rating rates and zero otherwise. *Post* is a time indicator that equals one if the time is after the first or second quarter of 2018 and zero otherwise. T-statistics values are shown in parentheses. The superscript ***, **, or * indicates statistical significance at the 1%, 5% or 10% level, respectively.

Table 7 – Bank credit accessibility and interest payment

Variables	*Loan Amount*		*Interest Payment*	
	(1)	(2)	(3)	(4)
SME × *Post*	0.049***	0.048***	-0.335***	-0.323***
	(2.82)	(2.79)	(-6.17)	(-7.33)
Post	-0.053***		0.367***	
	(-3.04)		(10.11)	
Constant	14.851***	14.818***	4.366***	4.596***
	(4,577.50)	(1,365.97)	(369.58)	(162.94)
Firm F.E.	YES	YES	YES	YES
Industry F.E.	YES	YES	YES	YES
Region F.E.	YES	YES	YES	YES
Year F.E.	YES	NO	YES	NO
Quarter F.E.	NO	YES	NO	YES
Observations	1,591,857	1,591,857	4,378,094	4,378,094
R^2	0.780	0.781	0.867	0.890

Note: This table presents the panel regression results on the influence of the adoption of AI and big data on loan amount and interest payment. Loan amount is the logarithm of the quarterly total sum of all loans. Interest payment refers to the interest rate of a loan. *SME* is an indicator that equals one if a firm is a SME and zero otherwise. *Post* is a time indicator that equals one if the time is after the second quarter of 2019 and zero otherwise. T-statistics values are shown in parentheses. The superscript ***, **, or * indicates statistical significance at the 1%, 5% or 10% level, respectively.

Table 8 – Early machine learning models V.S. big data analytics

Variables	*Undetermined Credit Rating*	*Loan Default Rate*
	(1)	(2)
SME × *Post1*	-0.016***	-0.015
	(-5.22)	(-1.11)
SME × *Post2*	-0.020***	-0.028**
	(-6.30)	(-2.04)
Constant	0.051***	0.067***
	(29.86)	(11.70)
Firm F.E.	YES	YES
Industry F.E.	YES	YES
Region F.E.	YES	YES
Quarter F.E.	YES	YES
Observations	4,378,877	4,358,049
R^2	0.706	0.708

Note: This table presents the panel regression results on the influence of the adoption of AI and big data on credit rating and the loan default. Undetermined credit rating is an indicator that equals one if a loan application does not have a credit rating (marked as undetermined in the data) and zero otherwise. The loan default rate is an indicator that equals one if a loan is defaulted and zero otherwise. *SME* is an indicator that equals one if a firm is a SME and zero otherwise. *Post1* is a time indicator that equals one if the time is after the second quarter of 2019 and zero otherwise. *Post2* is a time indicator that equals one if the time is after the third quarter of 2020 and zero otherwise. T-statistics values are shown in parentheses. The superscript ***, **, or * indicates statistical significance at the 1%, 5% or 10% level, respectively.

Table 9 – Heterogeneous analysis on undetermined credit rating

Variables	*Missing information=1*	*Unsecured loans=1*	*Less developed districts=1*	*Dialects districts=1*	*Non-SOE =1*
	(1)	(2)	(3)	(4)	(5)
Dummy × SME × Post	-0.042***	-0.041***	-0.017**	-0.034***	-0.021**
	(-6.92)	(-8.48)	(-3.27)	(-6.44)	(-2.50)
SME × Post	-0.004	-0.011***	-0.017***	-0.012***	-0.007
	(-1.36)	(-9.45)	(-3.53)	(-2.64)	(-0.87)
Dummy × Post	0.022***	-0.006	0.011**	0.014***	0.006
	(3.61)	(-1.32)	(2.18)	(2.93)	(0.96)
Dummy × SME	0.016***	0.042***	0.007	0.024***	
	(2.82)	(20.22)	(0.76)	(3.65)	
Dummy	-0.008	0.005***		-0.008	
	(-1.50)	(3.10)		(-1.26)	
Constant	0.038***	0.030***	0.040***	0.037***	0.043***
	(18.22)	(38.19)	(6.77)	(11.76)	(11.08)
Firm F.E.	YES	YES	YES	YES	YES
Industry F.E.	YES	YES	YES	YES	YES
Region F.E.	YES	YES	YES	YES	YES
Quarter F.E.	YES	YES	YES	YES	YES
Observations	4,378,877	4,378,877	4,378,877	4,109,026	4,378,877
R^2	0.706	0.708	0.706	0.710	0.706

Note: This table presents the panel regression results on the influence of the adoption of AI and big data on undetermined credit rating. Undetermined credit rating is an indicator that equals one if a loan application does not have a credit rating (marked as undetermined in the data) and zero otherwise. Dummy is an indicator representing five types of heterogeneity: first, whether a firm is missing information on sales and profit; second, whether a loan is secured by collateral; third, whether the lender is located in a more economically developed region; forth, whether the lender is located in a district with more than two dialects; and fifth, whether the borrower is a state-owned enterprise. *SME* is an indicator that equals one if a firm is a SME and zero otherwise. *Post* is a time indicator that equals one if the time is after the second quarter of 2019 and zero otherwise. T-statistics values are shown in parentheses. The superscript ***, **, or * indicates statistical significance at the 1%, 5% or 10% level, respectively.

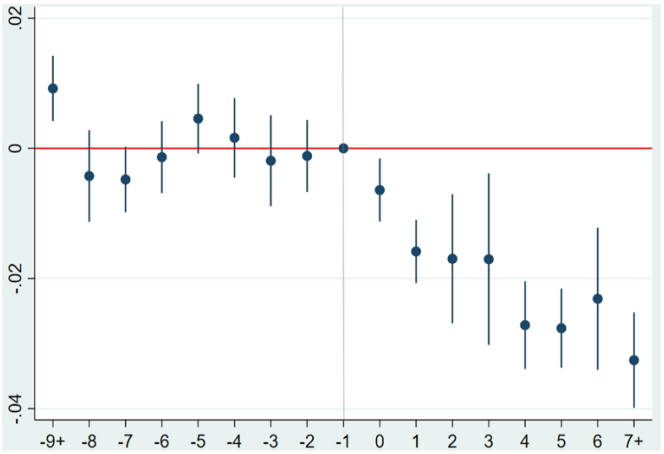
Table 10 – Heterogeneous analysis on loan default rate

Variables	*Missing information=1*	*Unsecured loans=1*	*Less developed districts=1*	*Dialects districts=1*	*Non-SOE =1*
	(1)	(2)	(3)	(4)	(5)
Dummy × SME × Post	-0.046***	0.033**	-0.044***	-0.026	-0.077***
	(-3.86)	(1.98)	(-2.91)	(-1.39)	(-6.22)
SME × Post	-0.011	-0.045***	-0.011	-0.019	0.008
	(-1.19)	(-4.07)	(-1.06)	(-1.26)	(1.58)
Dummy × Post	0.035***	-0.012	0.060***	0.030	0.081***
	(2.91)	(-0.69)	(3.98)	(1.64)	(6.87)
Dummy × SME	-0.000	-0.017***	0.009	0.015	
	(-0.04)	(-2.88)	(0.80)	(1.42)	
Dummy	0.013	-0.008		-0.022**	
	(1.34)	(-1.40)		(-2.14)	
Constant	0.046***	0.076***	0.038***	0.054***	0.034***
	(7.44)	(10.74)	(3.73)	(5.54)	(15.72)
Firm F.E.	YES	YES	YES	YES	YES
Industry F.E.	YES	YES	YES	YES	YES
Region F.E.	YES	YES	YES	YES	YES
Quarter F.E.	YES	YES	YES	YES	YES
Observations	4,358,049	4,358,049	4,358,049	4,089,293	4,358,049
R^2	0.708	0.709	0.708	0.712	0.708

Note: This table presents the panel regression results on the influence of the adoption of AI and big data on the loan default rate. Loan default rate is an indicator that equals one if a loan is defaulted and zero otherwise. Dummy is an indicator representing five types of heterogeneity: first, whether a firm is missing information on sales and profit; second, whether a loan is secured by collateral; third, whether the lender is located in a more economically developed region; forth, whether the lender is located in a district with more than two dialects; and fifth, whether the borrower is a state-owned enterprise. *SME* is an indicator that equals one if a firm is a SME and zero otherwise. *Post* is a time indicator that equals one if the time is after the second quarter of 2019 and zero otherwise. T-statistics values are shown in parentheses. The superscript ***, **, or * indicates statistical significance at the 1%, 5% or 10% level, respectively.

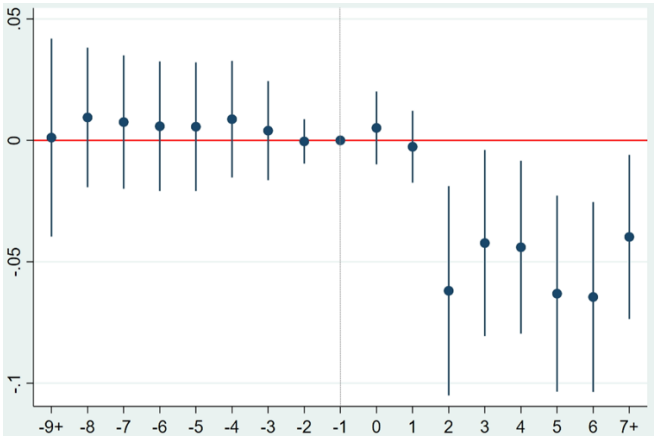
Figure 1 – Parallel trends test

Panel A: Undetermined credit rating



Notes: This figure presents the estimate for parallel trends. Every dot depicts the coefficient, associated 95% confidence intervals, from estimating the leads and lags regression of Equation (1) in the paper. The dependent variable is undetermined credit rating, an indicator that equals one if a loan application does not have a credit rating (marked as undetermined in the data). The estimated coefficients are relative to the one in the first quarter of 2019 (t = -1).

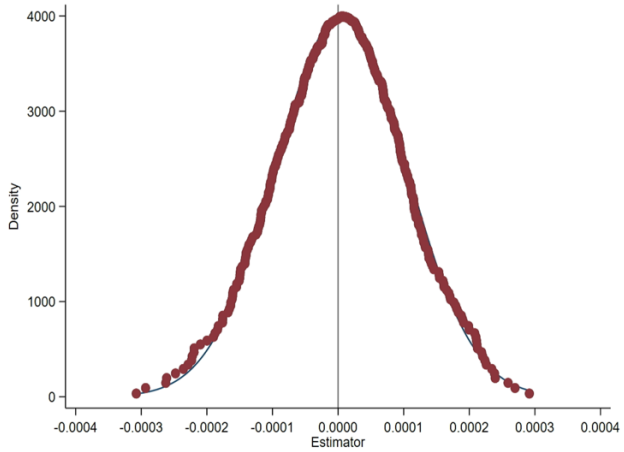
Panel B: Loan default rate



Notes: This figure presents the estimate for parallel trends. Every dot depicts the coefficient, associated 95% confidence intervals, from estimating the leads and lags regression of Equation (1) in the paper. The dependent variable is the loan default rate, an indicator that equals one if a loan is defaulted and zero otherwise. The estimated coefficients are relative to the one in the first quarter of 2019 (t = -1).

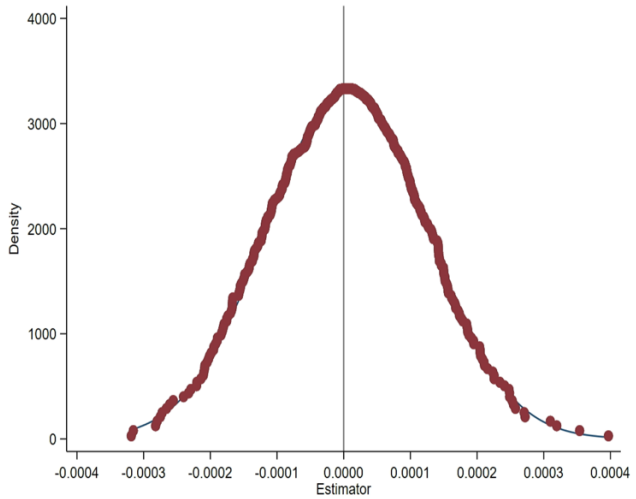
Figure 2 – Placebo test

Panel A: Undetermined credit rating



Notes: This figure illustrates the distribution of the placebo test results for the baseline regression, conducted using the Monte Carlo permutation method. The dependent variable is undetermined credit rating. In this test, individual observations were randomly assigned to the treatment group, and the regression analysis was repeated 500 times. Each dot in the figure represents an estimated coefficient along with its corresponding p-value, providing a visual representation of the placebo test results. The estimated coefficient from the actual baseline regression is -0.024.

Panel B: Loan default rate



Notes: This figure illustrates the distribution of the placebo test results for the baseline regression, conducted using the Monte Carlo permutation method. The dependent variable is loan default rate. In this test, individual observations were randomly assigned to the treatment group, and the regression analysis was repeated 500 times. Each dot in the figure represents an estimated coefficient along with its corresponding p-value, providing a visual representation of the placebo test results. The estimated coefficient from the actual baseline regression is -0.027.

Appendix

Table A1 – Region distribution

District	Loans	Percent	District	Loans	Percent
Beijing	169172	3.73%	*Inner Mongolia*	30480	0.67%
Tianjin	67703	1.49%	*Guangxi*	91487	2.02%
Hebei	197199	4.35%	*Chongqing*	91858	2.03%
Shanghai	198793	4.39%	*Sichuan*	217270	4.80%
Jiangsu	415190	9.17%	*Guizhou*	30409	0.67%
Zhejiang	669140	14.77%	*Yunnan*	46548	1.03%
Fujian	240904	5.32%	*Shaanxi*	114423	2.53%
Shandong	283282	6.25%	*Gansu*	37948	0.84%
Guangdong	635024	14.02%	*Qinghai*	5970	0.13%
Hainan	16370	0.36%	*Ningxia*	17439	0.38%
Shanxi	74172	1.64%	*Xinjiang*	39213	0.87%
Anhui	134114	2.96%	*Liaoning*	90109	1.99%
Jiangxi	81157	1.79%	*Jilin*	67890	1.50%
Henan	144924	3.20%	*Heilongjiang*	32638	0.72%
Hubei	131419	2.90%	*Xizang*	1317	0.03%
Hunan	156366	3.45%			

Note: This table presents the summary statistics for regional distribution of bank loans in the data sample from 2015 to 2023. There are 31 provinces and special districts.

Table A2 – Industry distribution

Industry	Loan	Percent
Agriculture, forestry, animal husbandry, fishery	39322	0.87%
Mining	15665	0.35%
Manufacturing	1832876	40.46%
Electricity, heat, gas and water production and supply	47701	1.05%
Construction Industry	450478	9.94%
Wholesale and retail industry	1435430	31.69%
Transportation, warehousing and postal services	158771	3.50%
Accommodation and Catering Industry	30876	0.68%
Information transmission, software and information technology	100130	2.21%
Real Estate Industry	35904	0.79%
Leasing and business services industry	158431	3.50%
Scientific Research and Technical Services	89510	1.98%
Water, Environment and Utilities Management Industry	44500	0.98%
Resident services, repairs and other services	27007	0.60%
Education	5843	0.13%
Health and social work	9869	0.22%
Culture, sports and entertainment industry	12398	0.27%
Other	35177	0.78%

Note: This table presents the summary statistics for industry distribution of bank loans in the data sample from 2015 to 2023.

Table A3 – Baseline regression with including city level control variables

Variables	*Undetermined credit rating*	*Loan default rate*
	(1)	(2)
SME × *Post*	-0.026***	-0.026***
	(-5.33)	(-1.99)
GDP	0.002	-0.003
	(0.49)	(-0.53)
Fiscal Revenue	0.000	-0.009**
	(0.03)	(-2.34)
Constant	0.029*	0.217***
	(1.65)	(10.83)
Firm F.E.	YES	YES
Industry F.E.	YES	YES
Region F.E.	YES	YES
Quarter F.E.	YES	YES
Observations	4,009,378	3,992,325
R^2	0.689	0.711

Note: This table presents the panel regression results on the influence of the adoption of AI and big data on undetermined credit rating and loan default rate. Undetermined credit rating, an indicator that equals one if a loan application does not have a credit rating (marked as undetermined in the data) and zero otherwise. Loan default rate, an indicator that equals one if a loan is defaulted and zero otherwise. *SME* is an indicator that equals one if a firm is a SME and zero otherwise. *Post* is a time indicator that equals one if the time is after the first or second quarter of 2018 and zero otherwise. GDP is the logarithm of the yearly city level GDP. Fiscal revenue is the logarithm of the yearly city level fiscal income. T-statistics values are shown in parentheses. The superscript ***, **, or * indicates statistical significance at the 1%, 5% or 10% level, respectively.