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
Central University of Finance and Economics
Peking University

Zirun Zhang
Central University of Finance and Economics

March 26, 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 (a 40.1% reduction), and loan default rates decline by 2.7 percentage points (nearly 30%). 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 and 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

* Junjie Xia is the corresponding author, email: junjiexia@nsd.pku.edu.cn.
Binkai Chen thanks the support from the National Natural Science Foundation of China (72342031). Dongmei Guo thanks the support from the Major Program of the National Social Science Foundation of China (21&ZD085). Junjie Xia thanks the support from the National Natural Science Foundation of China (72473172). All remaining errors are ours.

1. Introduction

The integration of artificial intelligence (AI) and big data is fundamentally transforming the operations of financial institutions, enabling advancements in efficiency, accuracy, and inclusivity that were previously unattainable. Big data, in particular, is a critical enabler, complementing AI models and addressing long-standing inefficiencies in the banking sector. While the existing finance literature has explored the impact of AI and big data across a range of applications, such as fund performance, corporate culture, market microstructure, distributional effects, and small business financing (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 transforming banking operations and decision-making processes at scale.

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. The dataset not only documents loan outcomes but also tracks the chronological adoption of different technologies within the bank's credit rating systems, encompassing the transition from human-driven approaches to machine learning and the subsequent integration of AI models with big data. This temporal distinction enables a nuanced examination of how the synergy between AI and big data effectively addresses persistent information asymmetries. A key strength of our identification strategy lies in the use of a government policy mandate as an exogenous shock, requiring the bank to initiate a systematic transition toward FinTech solutions in its operations.[1] This natural experiment allows us to isolate the causal effects of AI and big data adoption on credit ratings and loan default rates.

[1] 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

Historically, the bank relied on human-driven methodologies—shadow ratings, hierarchical analyses, and subjective judgment—to assess borrowers' creditworthiness. While these methods provided a foundational approach to credit evaluation, they were highly dependent on access to structured and complete information. In circumstances where borrower data was fragmented, incomplete, or unavailable, these methods often led to a disproportionately high proportion of "undetermined" or "unclassified" credit ratings. This ambiguity pushed the bank to rely on subjective human judgment, introducing potential biases and inconsistencies. Moreover, undetermined credit ratings carried significant implications for borrowers—ranging from the withdrawal of loan applications to the imposition of overly stringent loan terms—resulting in reduced access to financing.

Small and medium-sized enterprises (SMEs), in particular, bore the brunt of these inadequacies, as their limited financial transparency and lack of formalized reporting structures made them especially vulnerable to information asymmetries. Within our sample, SMEs accounted for 89.1% of all undetermined credit ratings, illustrating the disproportionate challenges they face in securing financing. These unresolved information gaps further limited SMEs' access to affordable loans and created inequities in loan terms.

In July 2019, following the government mandate, the bank replaced human-driven credit ratings with machine learning algorithms, marking the first stage of digital transformation. While this transition substantially reduced the reliance on subjective judgment, it further made a more holistic integration of big data with advanced AI models and recognition technologies—implemented in October 2020—that yielded transformative results. This second phase facilitated the incorporation of diverse and real-time data streams, including textual data, transactional VAT invoices, and online

a development plan specifically for FinTech. For more details, see the related policy announcements at www.pbc.gov.cn.

platform activity, into the bank's analytic models. The synergy of AI and big data led to a dramatic reduction in undetermined credit ratings—from 6.68% to 1.99%. Notably, SMEs were the primary beneficiaries, with their undetermined rate declining from 5.99% to 1.76%. These improvements reflect the unique role of big data in resolving information asymmetries and enhancing credit access for underserved borrowers.

Therefore, to rigorously evaluate the causal impact of these technologies, 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 to compare changes in credit ratings before and after the adoption between SMEs and large firms. Our analysis reveals that the adoption of big data and AI reduces the undetermined credit rating rate for SMEs by 2.4 percentage points—an approximate 40.1% reduction—highlighting the role of these technologies in improving credit rating accuracy for smaller businesses. Furthermore, the adoption of these tools reduced loan default rates by 2.7 percentage points (a 30% decline), reflecting the bank's improved ability to assess risks and prevent fraudulent lending through dynamic monitoring.

We perform a comprehensive series of robustness checks to validate the consistency and reliability of our findings, including parallel-trend tests and placebo tests. Our regional-level analysis further confirms that the observed improvements are causally driven by the adoption of AI and big data rather than contemporaneous policy interventions, such as tax incentives and industrial policies aimed at supporting SMEs.

We also demonstrate how the adoption of AI and big data enhances credit accessibility and narrows the disparity in interest payments between SMEs and larger firms. In the pre-AI era, SMEs consistently faced higher interest burdens due to

perceived higher risks and inequities in information availability. However, as these technologies shed light on SMEs' creditworthiness with greater accuracy, lending terms have become more equitable. Our heterogeneity analysis reveals that the benefits of big data and AI integration are particularly pronounced for firms with incomplete financial records, unsecured loans, and limited public information. Additionally, these advantages are magnified in regions with lower economic development and higher linguistic diversity, highlighting the diverse applicability of these tools in overcoming localized challenges.

Finally, our findings underscore the powerful synergy between AI and big data, with big data serving as the critical enabler that unlocks the full potential of AI. While machine learning models alone initially improved accuracy, their effectiveness was significantly enhanced by the integration of diverse and real-time data streams facilitated by big data. By consolidating a wide range of internal and external data sources into richer, multidimensional analytics, big data transforms AI systems from static tools into dynamic, adaptive instruments of analysis. The temporal differences in technology adoption highlight this progression: the adoption of machine learning systems alone reduced the rate of undetermined credit ratings by 1.6 percentage points. However, with the subsequent integration of big data and AI models, this reduction was magnified to 3.6 percentage points. These results emphasize that the full potential of AI in financial applications can only be realized when paired with the informational depth and real-time dynamism provided by big data.

This paper contributes to the growing body of research examining the real effects of machine learning and big data technologies in financial markets. By providing empirical evidence of their transformative role within the banking industry, we extend prior studies on how FinTech addresses information asymmetries and enhances credit accessibility. While much of the existing literature explores macro-level implications of these technologies or focuses on 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),[2] our study offers a distinct contribution by providing micro-level insights into the operational transformation of banking operations. Specifically, we focus on two critical dimensions—credit ratings and loan default rates—and examine how the integration of AI and big data fundamentally alters banking operations.

By analyzing 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. Unlike prior studies that often concentrate on high-level theoretical models or broad industry trends, our paper provides a granular, micro-level analysis of their operational effects within a financial institution, and underscores the importance of big data in enabling the transformative potential of AI in banks by providing richer, more dynamic, and multidimensional informational tools.

Additionally, our findings build on the emerging literature that emphasizes the information advantages offered by FinTech. Hau et al. (2024) empirically demonstrate the enhanced decision-making capabilities enabled by advanced credit technologies, while Livshits et al. (2016) propose a theoretical framework illustrating how financial innovation mitigates asymmetric information challenges, particularly in environments with diverse credit risks and fixed contractual costs.[3] Vives and Ye (2025) build a spatial model to study how information technology influences lender competition and lending to SMEs. Our paper provides fresh empirical evidence consistent with these

[2] 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.
[3] Hau et al (2019) provide similar predictions by constructing a simple model to show that FinTech credit is often used among weak creditors. Balyuk (2023) underscores information spillovers as a new mechanism through which FinTech lending reduces consumers' financial constraints.

findings, showing that the integration of AI and big data in banking operations meaningfully expands the informational capacity of FinTech, thereby improving credit assessments and loan outcomes.

Our paper also adds to the broader literature on the impact of FinTech on SMEs, a critical yet often underserved segment of the economy. SMEs often encounter high barriers to credit access, primarily due to information asymmetries, limited credit histories, reliance on soft information, and insufficient collateral. These challenges are well-documented in the literature (e.g., Petersen and Rajan, 1994 and Berger and Udell, 1995), and are further exacerbated by stringent regulatory capital requirements imposed on traditional banks. Recent studies have begun to explore how FinTech innovations address these challenges. Frost et al. (2020) highlight the role of FinTech credit in facilitating SME financing in regions with less competitive banking sectors, while Gopal and Schnabl (2022) find that non-bank finance companies, including FinTech providers, stepped in to fill the SME credit gap following the 2008 financial crisis. Similarly, Agarwal et al. (2019, 2022) show that mobile payment technologies not only enhance SME business formation and consumer credit spending but also improve access to financial services for underserved borrowers. Hau et al. (2024) demonstrate that access to FinTech credit services on Alibaba's online platform fosters growth in sales, transactions, and customer satisfaction for vendors facing higher credit risk.

Our study complements and extends this stream of literature by examining the specific role of AI and big data in enabling banks to better serve SMEs. We show that by incorporating advanced recognition technologies and utilizing big data, financial institutions can mitigate information asymmetries, reduce reliance on human judgment, and make credit evaluations more comprehensive and accurate. Our findings reveal the potential of FinTech, as catalysts for innovation in SME financing, with important implications for fostering economic growth and financial inclusion.

Finally, our paper contributes to the growing literature on the role 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. Relatedly, Brynjolfsson and McElheran (2016) document the diffusion of data-driven decision-making practices across industries, highlighting its broader implications on organizational productivity and efficiency. Our paper provides fresh empirical evidence on the specific effects of big data adoption in the banking sector. We demonstrate how the adoption of big data, when paired with AI, helps financial institutions overcome persistent challenges from information asymmetries, enhance risk management practices, and expand financial inclusion.

This paper is structured as follows. Section 2 provides an overview of the institutional background and details the data sample. Section 3 outlines the empirical methodology and presents the primary findings. Section 4 delves into additional analyses, exploring the impact of integrating big data and AI models on various aspects of banking operations. Finally, Section 5 concludes the paper and highlights key implications.

2. Background and Data

2.1 Institutional Background

Our data is sourced from one of China's largest commercial banks, a pivotal institution in the country's financial sector.[4] 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

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

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 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.

In response to the government’s policy mandate in July 2019, the bank initiated a transformative shift in its credit evaluation processes by integrating machine learning techniques. Initially, logistic regression models were deployed to replace human decision-making in credit rating, representing a significant step toward automation and data-driven evaluation. By October 2020, the bank had further upgraded its capabilities by incorporating advanced big data analytics and sophisticated machine learning models, including Artificial Neural Networks (ANN) and Federated Learning Models (FLM). Additionally, it adopted cutting-edge text recognition technologies such as Optical Character Recognition (OCR) and Natural Language Processing (NLP) to enhance the processing of both structured and unstructured data.

These advanced AI and machine learning models have enabled the bank to process and analyze vast and complex datasets with unparalleled speed and accuracy,

surpassing the limitations of traditional methods. The sources of big data utilized by the bank encompass a wide array of structured information, such as financial contracts, transaction histories, and external large-scale databases. Prominent external sources include the National Business Registration System, which public information on enterprise registration and ownership structures, and the National Intellectual Property Administration database, offering comprehensive details on patent applications and grants. Beyond structured data, the bank has successfully harnessed unstructured data sources, such as scanned documents, firm-to-firm transaction receipts, and various image-based records. Leveraging tools like OCR and NLP, the bank has been able to extract valuable insights from these previously inaccessible data formats, enabling more informed assessments of creditworthiness and operational performance.

In addition, the integration of big data analytics has introduced real-time monitoring capabilities, providing the bank with dynamic insights into borrower behavior and market conditions. For instance, the system can now detect early warning indicators of financial distress, such as abrupt shifts in transaction patterns, anomalies in spending behavior, or irregularities in cash flows. This real-time surveillance empowers the bank to intervene proactively, mitigating potential risks and improving the accuracy and responsiveness of its credit rating system. Such capabilities have played a key role in addressing traditional challenges, including reducing the proportion of “undetermined” credit ratings and enhancing the overall efficiency and precision of decision-making processes.

Importantly, these technological advancements have had a profound impact on the bank's ability to serve historically underserved sectors, such as small and medium-sized enterprises (SMEs). SMEs often struggle to secure credit due to limited transparency and the high costs of assessing their creditworthiness through conventional approaches. By leveraging AI and big data, the bank has mitigated these barriers, offering SMEs

better access to fair and equitable credit terms. This transformation positions the bank as a pioneer in utilizing AI and big data technologies to drive innovation.

Overall, this comprehensive adoption of advanced technologies has not only revolutionized the bank's credit assessment framework but has also demonstrated broader potential to reduce risks, improve credit accessibility, and promote financial inclusion across China's banking sector. These achievements highlight the critical role of AI and big data in modernizing financial services, fostering economic growth, and strengthening the resilience of the banking system.

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 comprehensive dataset contains detailed loan information, including credit ratings, interest rates, and default rates, covering all provinces and industries in China. Such breadth and depth make the dataset highly representative of the Chinese banking sector and an ideal foundation for examining the impacts of AI and big data technologies on loan issuance and credit evaluation. Furthermore, the dataset's diversity across regions and industries allows us to investigate how the integration of AI and big data influences credit accessibility and risk management across various economic contexts.[5]

Table 1 provides summary statistics of the data. Panel A presents the distribution of firms and loans, highlighting trends over time. Notably, there is a substantial increase in both the number of firms and loans in 2021. which coincides with the bank's full-scale adoption of AI and big data technologies by the bank in October 2020. This surge reflects the technologies' potential to expand credit issuance, especially to previously

[5] 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%.

underserved segments such as SMEs, by improving credit evaluation and operational efficiency. For instance, the number of SMEs increased from 72,009 in 2020 to 119,227 in 2021.

[Table 1 about here]

Panel B provides further comparison for large firms and SMEs before and after the adoption of AI and big data. A notable finding is the sharp reduction in the proportion of loans with undetermined credit ratings—falling from 6.682% in the pre-adoption period to 1.992% in the post-adoption period. Moreover, 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, both categories experienced declines, with SMEs showing a particularly dramatic improvement. The undetermined credit rate for SMEs dropped sharply to 1.759%. In summary, the rate of undetermined credit experiences a substantial change after the implementation of AI and big data, especially for SMEs.

In addition to credit rating accuracy, other loan-level variables, including default rates, loan amounts, and interest rates, exhibit similar positive trends following the adoption of AI and big data technologies. 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%. This notable decrease indicates the potential effectiveness of AI and big data in mitigating risks for smaller firms, which historically faced higher loan default rates due to limited financial transparency and greater operational uncertainties.

Another important observation is the significant reduction in 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.[6] $\varepsilon_{i,t}$ 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 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.[7]

Therefore, we have incorporated a comprehensive set of fixed effects in our regression models. These include firm-level, industry-level, region-level, and quarter-

[6] 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.

[7] For additional robustness checks, we include an indicator, equaling one if a firm is missing financial information, otherwise zero. Furthermore, 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.

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 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 coefficient of -0.024 in Column (3) suggest that undetermined credit rating rate among

SMEs decreases by 2.4 percentage points relative to large firms. Considering the average undetermined credit rating for SMEs before the adoption of AI and big data (5.985%, per Panel B of Table 1), the reduction constitutes an approximately 40.1% decline (=2.4%/5.985%). These results underscore the transformative role of these technologies in addressing the unique challenges faced by SMEs in the credit assessment process.

Importantly, the findings highlight the substantial benefits 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 reduces dependence on subjective human judgment, mitigates uncertainty in assessing SME creditworthiness, and facilitates more equitable access to financial resources.

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. Considering the average loan default rate of 9.12% for SMEs before the adoption of AI and big data, the reduction is an approximately 29.6% (=2.7%/9.12%).

[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.

3.3 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]

3.4 Extension – 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.

3.5 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 8, 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 8 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 8 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 8, 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 8, 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 9 presents the results, which are consistent with those in Table 8. 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 9 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.

4. The integration of big data and AI models

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).

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. We therefore 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.

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

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 10 presents the corresponding estimation results, focusing on the key coefficients of interest for the two interaction terms. These terms differentiate the impact of two distinct technological phases: (1) the initial adoption of general machine learning techniques in July 2019 and (2) the integration of big data analytics and advanced recognition technologies in October 2020. The results provide valuable insights into how each phase influenced the bank's operations, including credit ratings, loan default rate, bank credit accessibility, and borrowing cost.

[Table 10 about here]

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 and statistically significant at the 1% level. This suggests that the initial phase of technological adoption contributed to a 1.6 percentage point decrease in undetermined credit ratings. The second interaction term, associated with the integration of big data analytics and advanced recognition technologies, has a coefficient of -0.02, also statistically significant. This implies a 2.0 percentage point reduction in undetermined ratings in the second phase.

The larger magnitude of the second coefficient highlights the outsized impact of big data analytics and advanced recognition technologies. These tools enhanced the bank's ability to incorporate complex datasets, including unstructured data from scanned documents, firm-to-firm receipts, and images, into the credit evaluation process. By summing the coefficients of both interaction terms, the combined effect is estimated at -0.036, indicating a total 3.6 percentage point reduction in undetermined credit ratings after the adoption of these technologies.

This finding underscores the complementary nature of machine learning and big data analytics. Advanced recognition technologies, such as Optical Character Recognition

(OCR) and Natural Language Processing (NLP), played a pivotal role by enabling the bank to extract meaningful insights from non-traditional data sources, thereby enhancing its capacity to classify borrowers more effectively. The phase-wise adoption illustrates not only a progressive improvement in evaluation accuracy but also the synergistic potential of combining structured and unstructured data in credit modeling, 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, pointing to a 2.8 percentage point decrease in default rates. This suggests that the reduction in loan default rates is primarily driven by the second phase of technological adoption, where the bank incorporated big data with more sophisticated tools to enhance its credit evaluation processes.

The sharp reduction in loan defaults during the second phase can be attributed to the bank’s enhanced ability to process and analyze a broader range of data sources. The integration of big data analytics allowed for more comprehensive borrower profiling and dynamic risk assessment, addressing key challenges such as information asymmetry and adverse selection. Additionally, real-time risk monitoring, enabled by these technologies, helped the bank detect early warning indicators of financial distress, leading to timely interventions and more informed lending decisions. Economically, this reduction translates into enhanced financial stability for the bank while reducing its exposure to risky loans. More accurate credit evaluations not only mitigate the likelihood of defaults but also foster trust between lenders and borrowers, promoting a more secure and sustainable credit ecosystem.

Column (3) explores the impact of FinTech adoption on loan accessibility, demonstrating that the first phase of machine learning adoption did not significantly improve loan accessibility, as the corresponding coefficient is statistically insignificant. However, the second phase, involving big data analytics and advanced tools, shows a statistically significant positive impact on loan accessibility. This finding highlights the transformative potential of using big data analytics to uncover new insights from alternative datasets. By incorporating non-traditional data sources, such as transaction histories and scanned documents, the bank could more accurately assess the creditworthiness of SMEs and underbanked clients who may lack detailed financial records. This enabled the bank to extend credit to a broader range of borrowers, addressing persistent challenges in financial inclusion and SME financing.

The results in Column (4) examine changes in borrowing costs, particularly the gap of interest payment between SMEs and large firms. Unlike loan accessibility, the reduction in borrowing costs is primarily linked to the first phase of adoption, as the coefficient for the first interaction term is statistically significant, while the second interaction term, corresponding to the subsequent integration of big data, is not statistically significant. This suggests that the reduction in borrowing costs for SMEs occurred primarily during the early phase of the bank's FinTech transformation. This narrowing gap could be attributed to the improvements in operational efficiency brought about by the adoption of machine learning algorithms.

Overall, the results from Table 10 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.

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, and highlight that big data not only expands the breadth of information available for decision-making but also enhances the context and relevance of AI predictions, enabling banks to achieve greater accuracy and

efficiency in credit evaluations. This synergy amplifies the benefits of each technology, creating a transformative impact on key financial metrics.

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 highlights the need for financial institutions to move beyond standalone technological solutions by adopting an integrated approach to AI and big data. While machine learning tools provide the analytical framework for identifying patterns and making predictions, their real value lies in the richness, scope, and timeliness of the data they analyze. The synergy between these technologies enables banks to achieve breakthroughs in creditworthiness assessment, risk management, and financial inclusion that would not be possible with either approach in isolation.

As financial systems increasingly adopt AI and big data-driven solutions, ensuring that these tools are deployed responsibly and equitably is of paramount importance. By bridging cutting-edge research on AI and machine learning with dynamic big data applications, this study provides a roadmap for policymakers, practitioners, and researchers seeking to create a more inclusive, efficient, and sustainable financial ecosystem. The future of banking will undoubtedly be shaped by these innovations, and the strategies outlined in this paper offer a pathway for maximizing their transformative potential.

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 provides summary statistics for the total number of loans and firms in the data sample spanning from 2015 to 2023. Each value represents the corresponding count of firms and loans for a specific year.

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 summary statistics for the data sample spanning 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 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. Loan default rate is 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 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. 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. *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
Quarter F.E.	YES	YES	YES	YES
Observations	4,529,928	4,507,689	4,529,928	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. 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. Loan default rate is 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 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 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 – 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 9 – 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.

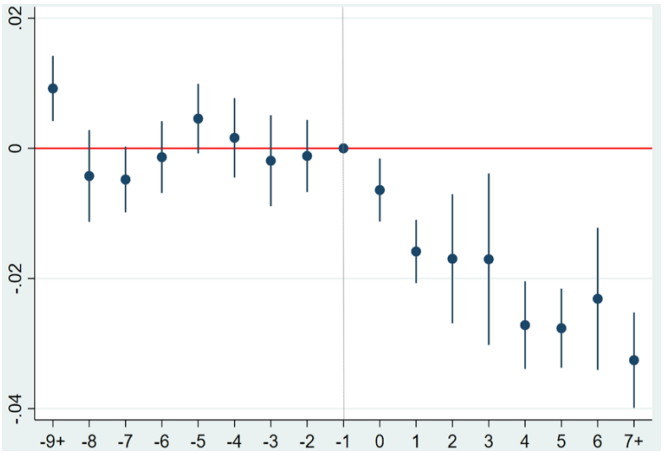
Table 10 – Early machine learning models V.S. big data analytics

Variables	*Undetermined Credit Rating*	*Loan Default Rate*	*Loan Amount*	*Interest Rate*
	(1)	(2)	(3)	(4)
SME × *Post1*	-0.016***	-0.015	0.014	-0.336***
	(-5.22)	(-1.11)	(0.72)	(-5.35)
SME × *Post2*	-0.020***	-0.028**	0.058**	-0.031
	(-6.30)	(-2.04)	(2.43)	(0.55)
Constant	0.051***	0.067***	14.809***	4.587***
	(29.86)	(11.70)	(1,186.42)	(305.11)
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	4,378,877	4,358,049	1,591,857	4,378,094
R^2	0.706	0.708	0.781	0.890

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. 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. *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.

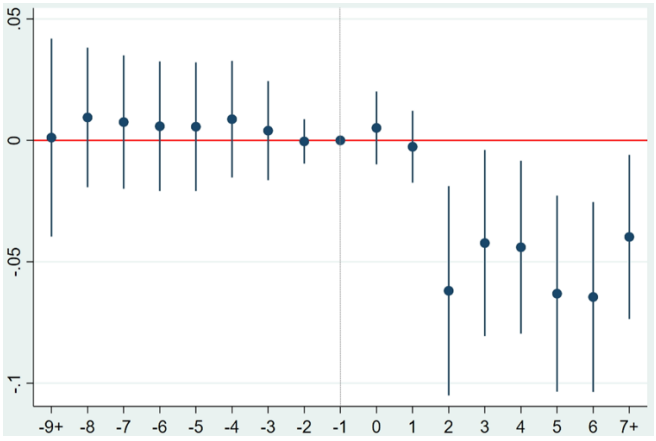
Figure 1 – Parallel trends test

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).

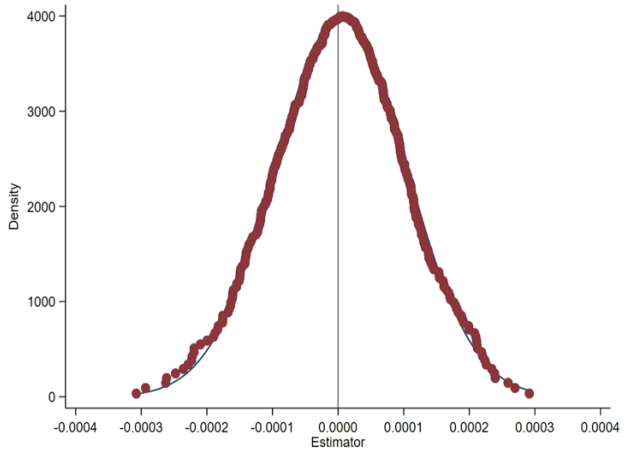
Panel B: Loan default rate



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).

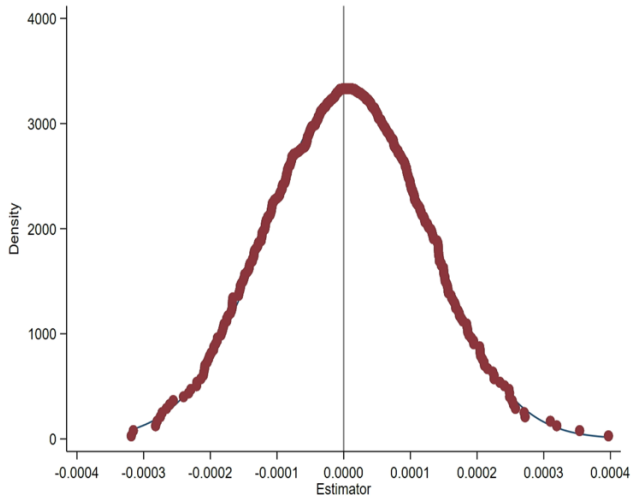
Figure 2 – Placebo test

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 firm-level financial indicator and city level control variables

Variables	*Undetermined credit rating*	*Loan default rate*	*Undetermined credit rating*	*Loan default rate*
	(1)	(2)	(1)	(2)
SME × Post	-0.024***	-0.029**	-0.026***	-0.026**
	(-5.76)	(-2.22)	(-5.33)	(-1.99)
Financial Infor	0.008***	0.020***		
	(9.71)	(17.62)		
GDP			0.002	-0.003
			(0.49)	(-0.53)
Fiscal Revenue			0.000	-0.009**
			(0.03)	(-2.34)
Constant	0.040***	0.048***	0.029*	0.217***
	(14.69)	(5.79)	(1.65)	(10.83)
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	4,378,877	4,358,049	4,009,378	3,992,325
R^2	0.706	0.709	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 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. 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. *Post* is a time indicator that equals one if the time is after the first or second quarter of 2018 and zero otherwise. Financial Infor is an indicator, equaling one if a firm is missing financial information, 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.