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 and Peking University

Zirun Zhang
Central University of Finance and Economics

July 28, 2025

Abstract

This paper examines how the integration of artificial intelligence (AI) and big data affects banking operations, emphasizing the crucial role of big data in unlocking the full potential of AI. Leveraging a comprehensive dataset of over 4.5 million loans issued by a leading commercial bank in China and exploiting a policy mandate as an exogenous shock, we document significant improvements in credit rating accuracy and loan performance, particularly for SMEs, which traditionally suffer from greater information asymmetry. The adoption of AI and big data reduces the rate of unclassified credit ratings by 40.1% and decreases loan default rates by 29.6%. Analyzing the bank's phased implementation, we show that while the initial adoption of AI alone yielded modest improvements, the subsequent integration of big data analytics substantially enhanced the effectiveness of the AI models, highlighting big data's pivotal enabling role. Furthermore, we identify significant heterogeneity: improvements are especially pronounced for uncollateralized and short-term loans, borrowers with incomplete financial records, first-time borrowers, long-distance borrowers, and firms located in economically underdeveloped or linguistically diverse regions. These findings underscore the powerful synergy between big data and AI, demonstrating their joint capability to alleviate information frictions and enhance credit allocation efficiency.

JEL Classification: G20, G21, G32
Keywords: artificial intelligence, big data, machine learning, information asymmetry, credit rating, default rate

[*] We are sincerely grateful to Lin William Cong, Ye Guo, Pengfei Han, Yan Ji, Fuwei Jiang, Yan Liu, Zhigang Qiu, Zhao Rong, Sergey Sarkisyan, Xuan Tian and Xiaoyu Zhu, as well as the seminar participants and discussants at Peking University, Xiamen University, Shanghai University of Finance and Economics, Sun Yat-sen University, and various conferences, for their valuable comments and suggestions. Binkai Chen thanks the support from the National Natural Science Foundation of China (72342031). Junjie Xia thanks the support from the National Natural Science Foundation of China (72473172) and the Major Program of the National Social Science Foundation of China (24ZDA057). All remaining errors are ours.

[†] Corresponding Author, email: junjiexia@nsd.pku.edu.cn

1. Introduction

The integration of artificial intelligence (AI) and big data is fundamentally transforming financial institutions by enabling unprecedented advancements in efficiency, accuracy, and financial inclusion. Big data, in particular, is a critical enabler that complements AI models, directly addressing long-standing inefficiencies in banking. Although existing literature has extensively explored AI and big data across various financial applications—such as fund management, 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 limited empirical evidence on precisely how these technologies reshape banking operations and credit decision-making processes.[1]

This paper addresses this gap by leveraging a unique and granular dataset containing over 4.5 million loans originated between 2015 and 2023 from a major Chinese state-owned bank. Our data not only track detailed loan outcomes but also record the bank's internal transition from traditional human-driven methods to machine learning (ML) models, and subsequently to integrated AI and big data technologies. This temporal variation allows us to identify how the synergistic combination of AI and big data alleviates persistent information asymmetries in lending.

Our identification strategy exploits a government-mandated FinTech initiative introduced in 2019 as an exogenous shock.[2] We employ a difference-in-differences (DID) framework, comparing outcomes for small and medium-sized enterprises

[1] Mo and Ouyang (2025) provide a comprehensive review on the interaction between AI and financial economics, noting that despite the proliferation of AI research, micro-level evidence on AI-driven transformations within banks remains scarce.

[2] The bank's decision to implement AI and big data was influenced by the government's Three-year Development Plan (2019-2021) for FinTech, aimed at promoting the development of financial technologies in the banking sector. This marked the first time the Chinese government initiated a development plan specifically for FinTech. For more details, see the related policy announcements at www.pbc.gov.cn. Thus, this serves as a quasi-natural experiment: the timing of the technology adoption is externally imposed rather than chosen by the bank, helping to establish causality.

(SMEs)—which traditionally encounter greater informational opacity—to large firms, which typically have richer financial transparency and collateral. This design allows us to causally isolate the impact of AI and big data adoption on credit ratings and default rates, and uncover the extent to which these technologies mitigate information frictions.

Historically, the bank relied heavily on human-driven credit evaluations, including shadow ratings, hierarchical analyses, and subjective judgment. Such methods perform adequately only when borrower information is abundant and reliable; when data are sparse or incomplete, they produce a high proportion of ‘unclassified’ credit ratings, often resulting in either rejected applications or unfavorable loan terms. SMEs were disproportionately impacted by this issue, representing 89% of unclassified ratings in our sample.

In July 2019, following the government mandate, the bank replaced human-driven credit ratings with ML-based credit evaluation, marking the first stage of digital transformation and substantially reducing subjective judgment. A more holistic integration occurred in October 2020, when the bank incorporated extensive big data inputs—such as textual analysis, VAT invoices, and online transaction data—into advanced AI models. This second phase significantly enhanced the accuracy of credit assessments, leading to a dramatic decline in unclassified credit ratings and default rates, even amid economic disruptions from the COVID-19 pandemic. SMEs particularly benefited from this transition, underscoring big data's unique ability to close informational gaps where traditional methods fall short.

Specifically, our DID analysis reveals that the adoption of integrated AI and big data technologies led to a substantial 2.4 percentage-point (approximately 40.1%) reduction in SMEs' unclassified credit ratings, highlighting the role of these technologies in improving credit rating accuracy. Additionally, loan default rates declined by 2.7 points (a 29.6% reduction), reflecting the bank's improved ability to assess risks and prevent fraud through dynamic monitoring.

We perform extensive robustness checks. Parallel-trend tests confirm that SMEs and large firms had similar pre-trends, and placebo tests show no spurious effects prior to the adoption of AI and big data. Moreover, regional analysis verifies that our results are not driven by concurrent policy interventions targeted at SMEs (e.g., tax incentives or industrial programs).

To enhance the understanding of information advantage channel, we conduct a comprehensive set of heterogenous analysis. We find that the effects of AI and big data adoption are significantly more pronounced for (i) loans with shorter maturities and no collateral, (ii) firms lacking formal financial statement information or public data, and (iii) regions with lower economic development and higher linguistic diversity. These findings illustrate that AI and big data deliver their most significant advantages precisely when traditional information is scarce.

Moreover, we find that these technological advancements improved overall credit accessibility and substantially narrowed the interest rate disparity between SMEs and large firms. In the pre-AI era, SMEs paid significantly higher rates due to perceived risk and informational disadvantages. After the adoption of AI and big data, SMEs' creditworthiness is better revealed, leading to a decline in the risk premium they pay. This change indicates tangible progress toward more equitable access to finance for smaller firms.

Critically, our findings underscore the impactful synergy between AI and big data, wherein big data serves as a pivotal factor that unlocks the full potential of AI models. Our analysis of the bank's two-stage adoption proves the synergy: the initial machine-learning upgrade (without big data) reduced unclassified credit ratings by 1.6 percentage points, but once big data was integrated, the reduction reached 3.6 percentage points. In other words, big data more than doubled the improvement. This shows that AI models alone yield moderate gains, but their full potential in financial applications can only be realized when paired with the informational depth and real-time dynamism provided by big data.

This paper makes several key contributions. First, we contribute to the growing literature on the real effects of machine learning and big data in finance by providing rare micro-level evidence from the banking sector. Prior studies have examined these technologies in various contexts, such as corporate governance and decision-making (Li et al., 2021; Erel et al., 2021), asset management performance and firm growth (DeMiguel et al., 2023; Babina et al., 2024), market microstructure (Easley et al., 2021), and distributional outcomes in lending (Fuster et al., 2022).[3] Yet, there remains limited empirical evidence demonstrating how AI and big data jointly transform credit operations within banks. Leveraging a unique two-stage technology rollout, we find that the initial AI-only adoption yielded notable but relatively modest improvements in credit rating accuracy and loan performance, whereas the subsequent integration of big data led to substantially larger gains.[4] This stark contrast highlights big data's critical role in unlocking the full potential of AI, as the richer, real-time datasets introduced in the second phase substantially amplified the capabilities of the AI models.

Our findings on the complementarity between data and algorithms align with recent literature on the role of data in AI. Mihet et al. (2025) find that advanced AI capabilities disproportionately benefit data-rich firms, whereas enhanced data availability narrows the performance gap between AI leaders and laggards. Similarly, Miao et al. (2023) show that firms with greater data assets derive more value from generative AI, reinforcing the notion of strong data-AI complementarity. Our study offers one of the

[3] 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. Fuster et al. (2019) find that FinTech lenders process loans faster and increase credit supply. Goldstein et al. (2021) also provide an excellent summary for the recent research on big data in finance.

[4] While the second-stage upgrade also introduced a more advanced AI algorithm, the relatively limited gains from the first stage (AI alone) versus the much larger gains after big data's introduction strongly indicate that big data was the pivotal factor. Although we cannot fully disentangle the effects of model refinement from data enrichment, the evidence suggests that it was the infusion of big data that unlocked the AI model's additional performance in the second stage.

first empirical demonstrations that enhancing data resources can significantly boost AI's effectiveness in banking.

Second, our paper adds to the emerging literature on the information advantages offered by FinTech and advanced analytics in credit markets. Theoretical work has long posited that financial innovations can mitigate information problems (e.g., Livshits et al., 2016). Recent empirical studies show increased consumer credit (Balyuk, 2023), improved decision-making with advanced credit tech (Hau et al., 2024). Di Maggio et al. (2022) find that a sophisticated ML underwriting algorithm can approve more loan applicants at lower interest rates. Likewise, Vives and Ye (2025) develop theoretical models showing how IT adoption affects lending competition. Our findings empirically validate and deepen these insights by causally demonstrating that integrating AI and big data substantially expands banks' information-processing capacity, enhancing credit decisions and outcomes.

Third, we enrich the broader literature on the impact of FinTech on SMEs. SMEs often encounter high barriers to credit access due to information opacity (Petersen & Rajan, 1994; Berger & Udell, 1995). Recent studies have begun to explore how FinTech innovations address these challenges. Frost et al. (2020) highlight how FinTech credit supports SME financing in low-competition regions, and Gopal and Schnabl (2022) show that non-bank lenders filled critical gaps after the 2008 crisis. Agarwal et al. (2019, 2022) find that mobile finance platforms expand credit access and stimulate small-business activity. Hau et al. (2024) show that FinTech credit boosts sales and customer satisfaction for riskier entrepreneurs.[5] Our study complements this literature by showcasing how a traditional bank's adoption of AI and big data can also materially improve SME credit access, reinforcing FinTech's broader role in enabling inclusive financial growth.

[5] In the Chinese context, Liu et al. (2022) show that Ant Group's AI-driven lending platform uses alternative data to extend loans to credit-constrained SMEs, providing rapid funding even during shocks like COVID-19.

Finally, our paper contributes to the growing literature on big data in finance. Prior research highlights how data-driven decision-making enhances firm productivity (Brynjolfsson & McElheran, 2016) and improves financial forecasting (Begenau et al., 2018). Other studies show that non-traditional data sources can predict borrower risk as well as conventional credit metrics (e.g., Berg et al., 2020; Liu et al., 2022).[6] Cong et al. (2025) further find that as firms accumulate data, they endogenously shift toward AI-driven innovation strategies. We provide fresh evidence of big data's critical enabling role in traditional banking: big data is not merely supplementary but essential for unlocking AI's full analytical potential, leading to superior credit market outcomes.

In summary, our work bridges multiple strands of research—on AI/ML in finance, FinTech credit, and big data analytics—and offers rare micro-level evidence of how advanced technologies can fundamentally transform banks' lending practices. While our empirical context is China, these insights carry broader implications for the global conversation on AI in finance by illustrating the tangible benefits of combining algorithms with big data in lending, as well as by highlighting areas (such as SME finance) where technology can markedly improve outcomes.

The remainder of the paper proceeds 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 policy implications.

2. Background and Data

2.1 Institutional Background

[6] Liberti and Petersen (2019) also emphasize how hard-information tools (big data) complement or reduce the need for soft information.

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

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

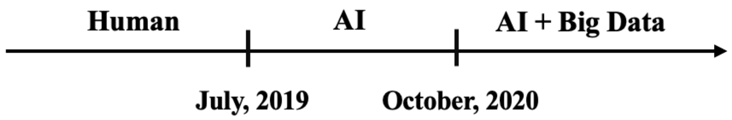
In response to this policy, the bank actively embraced advanced financial technologies, including AI-driven credit assessment models and big data analytics, to align with the government's objectives and maintain its competitive edge in an increasingly technology-driven financial landscape. These efforts 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 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 predominantly 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.

Several critical limitations characterize traditional credit rating models. First, they 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 face significant cognitive limitations regarding the sheer volume and complexity of data they can process effectively, increasing the likelihood of oversight or misinterpretation of critical risk indicators. 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 'unclassified' or '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—a straightforward supervised learning algorithm—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 more sophisticated machine learning models, including Artificial Neural Networks (ANNs) and the use of Federated Learning Models (FLM). ANNs consist of multiple computational layers (far exceeding the single-layer structure of logistic regression) and can learn complex non-linear relationships, were deployed to harness the large, high-dimensional data. FLM, a cutting-edge approach that trains models across decentralized databases. This allowed multiple branches or institutions to collaboratively improve credit scoring algorithms without sharing raw data, thus utilizing a wider information set while preserving privacy. Additionally, to enhance the processing of both structured and unstructured data, the bank adopted cutting-edge text recognition technologies, such as Optical Character Recognition (OCR) mainly designed for text and Natural Language Processing (NLP) mainly designed for unstructured data. The following timeline illustrates the evolution of the bank adopting different approaches in the credit rating system.



The introduction of these advanced AI models and big data tools empowered the bank to process and analyze expansive and complex datasets with unprecedented efficiency, granularity, and predictive accuracy, thereby overcoming traditional methodological limitations. 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, containing 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 textual data from scanned documents, receipts from firm-to-firm transactions, online consumer feedback, and visual records, all of which historically lay beyond traditional analytical reach. 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, integrating external, unstructured, and real-time big data into ANN and FLM methodologies unlocks their full potential and introduces robust real-time monitoring capabilities. For instance, ANN-based credit scoring models augmented by big data allow the bank to dynamically capture borrower behaviors, financial transactions, and market signals, delivering timely and accurate predictive insights. Similarly, FLM, enhanced by external big data, enables the aggregation of rich and diverse insights across multiple institutions or branches without directly compromising data privacy, effectively addressing challenges related to data heterogeneity and limited local feature richness. Such integration can swiftly detect emerging financial distress signals, including abrupt changes in spending behaviors, transactional irregularities, cash flow anomalies, or macroeconomic disturbances, thereby facilitating early warning interventions. These real-time surveillance capabilities significantly reduce the incidence of unclassified credit ratings and improve overall accuracy and responsiveness in lending decisions.

Importantly, these technological advancements have profoundly improved the bank's capacity to support historically underserved segments, notably 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.

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.[7]

Table 1 provides summary statistics of the data. The definition of SMEs used in this study is directly sourced from the bank, which adheres to the official classification established by the Central Bank of China. 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 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]

[7] 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%.

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 unclassified 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 unclassified 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 unclassified credit rate for SMEs dropped sharply to 1.759%. In summary, the rate of unclassified 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 unclassified 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 unclassified 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—characterized by greater information asymmetry—are designated as the experimental group, while large firms serve as the control group. Our identification strategy leverages a government-initiated FinTech adoption as an exogenous shock, 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:

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

where i indexes loan; f indexes firm; j indexes industry; r indexes region and t indexes time. $Y_{i,t}$ refers to the outcome variable, particularly an indicator for unclassified credit rating equaling one if a loan application does not have a credit rating (marked as "unclassified"), indicating insufficient information to assign a rating, and zero otherwise; or an indicator for loan default rate equaling one if the loan is defaulted 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, δ_r and θ_t, represent the fixed effects on firm, industry, region and time, respectively.[8] $\varepsilon_{i,t}$ is the error term. Note that the coefficient

[8] 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.

of interest is β on the interaction *SME* × *Post*, which measures the additional post-policy change in the outcome for SMEs relative to large firms.

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 "unclassified" 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 unclassified 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 unclassified credit ratings in the absence of sufficient information. Thus, this choice keeps the focus on the intended mechanism (information availability) and avoids controlling away part of the treatment effect.

Secondly, our dataset is primarily loan-specific and does not provide comprehensive firm-level characteristics for all firms. 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.[9]

Therefore, we have incorporated a comprehensive set of fixed effects in our regression models. Firm fixed effects absorb static differences between SMEs and large firms (such as baseline riskiness or creditworthiness), focusing identification on within-

[9] To retain the original size of observations, one way to conduct additional robustness check is to incorporate an indicator variable that equals one if a firm has missing financial information and zero otherwise. Furthermore, by merging city-level data obtained from Chinese statistical yearbooks, we also include city-level control variables such as GDP and fiscal revenue. The results presented in Table A3 confirm that our baseline results are still valid.

firm changes relative to the control group's trend. Time fixed effects capture any economy-wide or bank-wide shocks (e.g., macroeconomic changes, overall improvements in the bank's operations, seasonality) that affect all firms in that quarter. By also adding industry and region fixed effects, we control for sector-specific trends or regional economic changes. This comprehensive fixed-effects structure ensures that the DID estimator β is identified purely from the relative change in SMEs vs. large firms, net of any other fixed influences. 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 "unclassified" 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 unclassified credit ratings for SMEs. In terms of economic significance, the coefficient of -0.024 in Column (3) suggest that unclassified credit rating rate among SMEs decreases by 2.4 percentage points relative to large firms. Considering the average unclassified 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%). This improvement in rating accuracy is consistent with evidence that AI-driven credit models capture complex nonlinear risk patterns and outperform traditional linear methods (Sadhwani et al., 2021).

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 information 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 information 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. While SMEs had a higher incidence of unrated loans prior to the policy (reflecting informational gaps), this difference was largely time-invariant; importantly, we verify that SMEs and large firms exhibited parallel pre-policy trends in our outcome variables. This satisfies the key DID assumption that, absent the AI and big data adoption, both groups would have followed a similar trajectory.

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 unclassified 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 unclassified 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 unclassified 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 unclassified 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 unclassified 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 unclassified credit ratings and loan default rates.

[Figure 2 about here]

Fourth, a potential alternative explanation for the observed decline in unclassified 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. During our sample period, other interventions such as industrial support programs, tax incentives, or targeted financial assistance policies might have enhanced SMEs' operational efficiency and financial health independently from technological adoption, thus improving credit ratings and reducing default rates.

To explicitly address this concern, we shift our empirical analysis from the firm level to the regional level, enabling us to account more effectively for these overlapping policy effects. Specifically, we identify regions exhibiting high levels of unclassified credit ratings prior to the adoption of AI and big data technologies and classify these regions as treated groups within our difference-in-differences (DID) framework. We define treated regions as those whose pre-adoption unclassified credit rating rates exceed certain thresholds (e.g., 5% or 10%).[10] This regional-level identification strategy helps mitigate potential confounding effects from concurrent SME-specific policies, thereby isolating the impact of AI and big data technology adoption on credit evaluation outcomes.

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 unclassified credit ratings and loan default rates. These findings align closely with our baseline firm-level analysis and further underscore the pivotal role of AI and big data in effectively mitigating informational asymmetries and improving credit outcomes. For instance, Column (1), which examines the top 5% of regions, shows an estimated coefficient of -0.069 relative to an average pre-adoption unclassified credit rating rate of approximately 11.69%. This implies an economically substantial reduction of about 59% (=0.069/0.1169), comparable to—and indeed exceeding—our baseline firm-level estimate of 40.1%. Thus, this regional-level robustness test alleviates concerns regarding biases arising from contemporaneous SME-supportive policies, reinforcing our conclusion that the

[10] Regions with unclassified credit rating rates higher than 5% fall within the upper 50% of the distribution, while those with rates above 10% belong to the upper 25% group.

integration of AI and big data technologies was critical in achieving significant improvements in credit allocation and loan performance.

[Table 6 about here]

Additionally, within the SME group, we find that smaller firms benefit more than medium-sized firms from the adoption.[11] This suggests that the smallest, most information-opaque businesses see the greatest improvement in credit access and loan performance, which is consistent with our thesis that information frictions drive the technology's impact. Finally, we further include city-by-time and region-by-time fixed effects, allowing us to account for any unobserved, time-varying city-specific and region-specific characteristics, such as local economic cycles, policy interventions, or regional development programs, that could influence both the adoption of AI and big data technologies and the banking outcomes. Importantly, after controlling for these additional fixed effects, the magnitude of our core coefficients increases, and their statistical significance becomes even stronger.[12] This reinforces the robustness and reliability of our baseline findings and suggests that, if anything, our original estimates might understate the true impact of AI and big data adoption.

3.4 Heterogeneous analysis

To further substantiate the informational advantages provided by AI and big data, we perform a comprehensive set of heterogeneous analyses at the loan, firm and region levels. These analyses aim to deepen our understanding of how adopting these advanced technologies mitigates information asymmetries and improves credit evaluation processes under varying conditions.

[11] While large firms are the primary control in our baseline analysis, we also consider alternative comparisons to reinforce our findings. Specifically, we compare small-size firms (treatment group) with medium-size firms (control group). Small-size firms likely have even less structured information and higher default risk, so they might see a larger relative improvement. Table A4 in the appendix shows the coefficient of the interaction term is negative and significant for both unclassified credit rating and loan default rate, supporting our baseline DID setting.

[12] The corresponding results are presented in Table A5 in the appendix.

We begin by examining the effect of AI and big data on firms in four dimensions: whether or not a firm is missing critical financial information; is missing public information; is the first-time borrower; and is a cross-city borrower. First, financial 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 financial information 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) and (2) of Table 7, reveal that the coefficient for the DDD estimator is negative and statistically significant. This finding demonstrates that the reduction in unclassified credit ratings and loan default rate 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 7 about here]

Second, we investigate 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 (3) and (4) of Table 7, 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.

Third, a plausible hypothesis is that AI and big data are most useful for evaluating new clients (with no lending history). In contrast, for repeat borrowers, the bank already has internal data on past repayment behavior. To capture this effect, we include an indicator for loans made to the first-time borrowers and interact it with our *SME* × *Post* treatment term in a triple-difference framework. The results presented in Column (1) and (2) of Table 8 confirm our thesis, indicating that the impact of AI and big data is more profound in the first-time borrowers. Our finding that AI and big data benefit 'thin-file' borrowers most is consistent with recent evidence that data-rich algorithms can identify creditworthy 'invisible primes' overlooked by traditional scoring models (Di Maggio et al., 2022; Ouyang, 2023).

[Table 8 about here]

Forth, we consider geographic distance as a source of information friction in lending. Borrowers applying for credit outside their local area often lack the advantage of proximity, which traditionally facilitates soft-information gathering through personal interactions and local knowledge (Petersen and Rajan, 2002). To examine this heterogeneity, we construct a binary indicator for loans made to borrowers located in a different city than the lending branch. Columns (3) and (4) of Table 8 report the results for these long-distance loans. The coefficient on the triple interaction is negative for the unclassified credit rating outcome and the default rate, indicating that the benefits of AI adoption are indeed stronger for cross-region borrowers.

Taken together, the heterogeneity results in Table 7 and Table 8 underscore that AI and big data technologies deliver the greatest benefits under severe information asymmetries. Whether the information gap stems from incomplete borrower documentation, the absence of collateral, a first-time borrower with no credit history, or geographic distance between the borrower and the bank, the pattern is consistent: the improvements in credit assessments and loan performance are significantly more pronounced in these high information friction scenarios. In essence, the AI-powered credit evaluation system acts as a substitute for traditional informational proxies—financial statements, collateral guarantees, relationship history, or local insight—by extracting predictive signals from alternative data. This capacity allows the bank to mitigate information asymmetry more effectively, yielding sharper reductions in unclassified credit ratings and default rates for the most opaque borrowers. These findings reinforce our central thesis that AI and big data can alleviate informational frictions in lending.

Next, we investigate the heterogeneous effects of AI and big data based on loan types and loan maturity. We categorize each loan as either collateralized or uncollateralized Collateral traditionally serves as a safeguard for lenders, reducing reliance on borrower-specific soft information. Uncollateralized loans, lacking collateral, inherently rely more on soft information (e.g., borrower reputation, behavior). Thus, we hypothesize that AI and big data have a greater impact on uncollateralized loans, since these technologies can process alternative information to substitute for the missing collateral.

Specifically, we construct a binary indicator that equals one if a loan is pledged with collaterals and zero otherwise. Column (1) of Table 9 confirms the hypothesis: the reduction in unclassified ratings is significantly larger for uncollateralized loans than for secured loans. This suggests the bank is now leveraging AI/big data to evaluate borrower quality in cases where it previously would have leaned on collateral. In line with Aghion and Bolton (1992), collateral has limitations in resolving information problems, and our results imply that advanced data analytics can partly substitute for collateral by revealing borrower creditworthiness. Consequently, lending becomes more efficient and inclusive, as the bank can confidently extend uncollateralized credit to worthy borrowers that lack collateral.

[Table 9 about here]

Interestingly, Column (2) of Table 9 shows that the effect of AI and big data on default rates appears positive for uncollateralized loans. This does not contradict our story; rather, it reflects that uncollateralized loans inherently carry higher risk. Secured loans, backed by collateral, tend to involve borrowers with stronger financial positions and lower incentives to default, as the pledged assets act as both a signal of creditworthiness and a mechanism of discipline. In contrast, uncollateralized loans inherently carry higher credit risk, since borrowers are not required to post collateral and thus face fewer financial consequences in the event of default. As such, the elevated default rates in uncollateralized lending reflect structural differences in loan design rather than limitations in the predictive power of AI and big data technologies. While these technologies significantly improve risk identification and monitoring, they cannot fully eliminate the underlying risk differentials that are embedded in loan contracts.

We also examine heterogeneity by loan maturity. Short-term loans (e.g., working capital loans under one year) could benefit more from the adoption of AI and big data, which enable real-time monitoring, whereas long-term loans (often given to more creditworthy borrowers) might see smaller gains. We include a dummy for short-term loans (maturity < 1 year) and interact it with the treatment. Column (3) and (4) of Table 9 show that the coefficient for both unclassified credit rating and loan default rate is negative, suggest that the adoption of AI and big data produces a more pronounced impact on short-term loans. Notably, the effect on default rates is especially strong for short-term loans, consistent with big data's advantage in continuous monitoring. This result makes intuitive sense: long-term borrowers already undergo rigorous screening and tend to be safer, leaving less room for improvement, whereas short-term lending to less-established borrowers gains more from enhanced information.

Finally, 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 Table 10, 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.

[Table 10 about here]

In summary, the empirical findings from the heterogeneous analysis 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.

3.5 Extension – Credit accessibility and borrowing cost

We extend our study 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 11 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 11 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 also 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 information asymmetries or perceived riskiness.

4. The integration of big data and AI models

In this section, we analyze how integrating big data with advanced AI algorithms and sophisticated text recognition technologies can deliver substantially greater impacts on banking operations compared to traditional FinTech models. Specifically, we explore how big data serves as an essential enabler that unlocks the full potential of AI, allowing financial institutions to achieve superior performance in credit evaluation, risk management, and operational efficiency.

The bank experienced two significant phases in its adoption of AI and big data technologies. The first phase commenced in July 2019, during which the bank initially introduced machine learning approaches, particularly logistic regression models, to automate and enhance credit evaluation processes previously reliant on human judgment. This transition significantly improved the consistency and objectivity of credit assessments by reducing manual errors and subjective biases inherent in traditional methodologies.

The second phase followed in October 2020, during which the bank further advanced its technological capabilities by integrating big data analytics and incorporating more advanced AI models (ANN and FLM) and sophisticated text recognition technologies (OCR and NLP). These advancements permitted the comprehensive processing of previously inaccessible or underutilized unstructured and semi-structured datasets—including scanned financial documents, handwritten contracts, and textual transaction records—thereby vastly expanding the informational basis for credit decisions. Recent literature has highlighted that large language models and related AI techniques are particularly effective in extracting predictive signals from such high-dimensional textual 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. By exploiting the temporal variation in the adoption of these distinct innovations, we can rigorously assess their individual and combined contributions, offering nuanced insights into the specific channels through which big data analytics and AI synergistically enhance banking operational efficiency.

To capture these effects in our empirical analysis, 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.

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

where i indexes loan; f indexes firm; j indexes industry; r indexes region; and t indexes time. The dependent variable $Y_{i,t}$ refers to the unclassified credit rating and loan default rate. Unclassified credit rating is an indicator that equals one if a loan application does not have a credit rating (marked as unclassified 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 12 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 12 about here]

Column (1) reports the estimates for unclassified 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 unclassified 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 unclassified 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 unclassified credit ratings after the adoption of these technologies.

This finding underscores the complementary nature of machine learning and big data analytics. Advanced recognition technologies play 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 12 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 AI models and 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.

To further elucidate the powerful role of big data in mitigating information asymmetries, we incorporate firm-level financial information including total assets and total debts into regression equation (2). Our analysis underscores the hypothesis that, despite possessing firm financial data, banks confront greater challenges in SMEs compared to large enterprises due to heightened information asymmetries. The integration of big data and AI models in the later adoption phase offers a noteworthy advancement in resolving these issues. Consequently, we anticipate that the coefficient of the second interaction term—representing the intersection between *SME* and *Post2*—will be negative and exhibit a greater absolute magnitude than the first interaction term, reflecting a more significant impact.

Table 13 provides empirical evidence supporting this hypothesis, illustrating that the use of big data notably deepens its influence, especially for unclassified credit rating and loan default rate. This emphasizes the unique capability of big data to enhance banks' monitoring of dynamic firm activities, showcasing its transformative potential in refining the precision and efficiency of financial evaluations within the banking ecosystem.[13]

[Table 13 about here]

From a practical perspective, these findings underscore the importance of optimizing the value of big data by leveraging advanced recognition technologies and incorporating with advanced AI models. By doing so, the bank can better assess borrowers – especially SMEs lacking formal records – more accurately, thus safely extending credit to a broader client base. In sum, even after accounting for traditional financial metrics, the informational lift from big data is evident, cementing our argument that big data is a crucial tool for reducing SME information frictions.

Overall, we show that big data is not merely supplementary but essential for unlocking AI's full analytical potential, allows financial institutes to overcome informational barriers, enhance risk management, and promote financial inclusion.

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 unclassified credit ratings—a long-standing obstacle to effective risk

[13] As shown in Table 12, incorporating total assets and total debts significantly reduces the number of observations. If we further add firm sales, it results in an even greater decline in the number of observations. Table A6 in the appendix presents the result. Despite the reduction in sample size, the results remain consistent with those presented in Table 12, confirming the robustness of our findings.

evaluation, particularly for small and medium-sized enterprises (SMEs). This improvement reflects enhanced accuracy, granularity, and efficiency in credit assessments, made possible by the synergistic interaction between AI models and big data analytics.

Our findings reveal that the adoption of big data analytics, in conjunction with machine learning algorithms, not only decreases the rate of unclassified 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. Critically, we underscore the foundational role of big data: it is not merely complementary to AI but a necessary enabler that enhances the scope, context, and relevance of AI predictions. By incorporating real-time, high-dimensional data streams—such as VAT invoices, online transactions, and unstructured text—big data empowers AI models to capture dynamic borrower behaviors and latent creditworthiness, thereby unlocking their full potential.

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.

Ultimately, 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 AI provides the computational engine for credit scoring and predictive modeling, it is big data that fuels, contextualizes, and elevates these models to actionable insights. The two must function as an integrated system: AI without data is blind, and data without AI is inert. Their convergence enables financial institutions to overcome entrenched information asymmetries, enhance credit accessibility for SMEs, and strengthen systemic risk detection—outcomes that would not be achievable by either technology 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.

Berg, T., V. Burg, A. Gombović, and M. Puri. 2020. On the Rise of FinTechs: Credit Scoring Using Digital Footprints. *Review of Financial Studies* 33(7): 2845-2897.

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

Cong, L. W., Y. Lu, H. Shi, and W. Zhu. 2025. Automation-induced innovation shift. Available at SSRN 5049949.

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.

Di Maggio, M., D. Ratnadiwakara, and D. Carmichael. 2022. Invisible primes: Fintech lending with alternative data. Harvard Business School Working Paper, No. 22-024.

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., M. Plosser, P. Schnabl, and J. Vickery. 2019. The Role of Technology in Mortgage Lending. *Review of Financial Studies* 32(5): 1854-1899.

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.

Liberti, J.M., and M. A. Petersen. 2019. Information: Hard and Soft. *Review of Corporate Finance Studies* 8(1): 1-41.

Liu, L., G. Lu, and W. Xiong. 2022. The big tech lending model. NBER working paper 30160.

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.

Mihet, R., O. Gomes, and K. Rishabh, 2025. Is AI or data driving market power? *Journal of Monetary Economics*, forthcoming.

Eisfeldt, A., G. Schubert, B. Taska, and M. Zhang, 2025. Generative AI and firm values. Available at SSRN 4436627.

Mo, H., and S. Ouyang. 2025. (Generative) AI in Financial Economics. Working paper.

Ouyang, S. 2023. Cashless Payment and Financial Inclusion. Working paper.

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.

Petersen, M. A., and R. G. Rajan. 2002. Does Distance Still Matter? The Information Revolution in Small Business Lending. *Journal of Finance* 57(6), 2533-2570.

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

Sadhwani, A., K. Giesecke, and J. Sirignano. 2021. Deep learning for mortgage risk. *Journal of Financial Econometrics* 19(2), pp. 313-368.

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
Unclassified credit rating loans	105221	10978	94243	58863	6879	51984
Rate of unclassified 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 unclassified credit rating loans to the number of overall unclassified credit rating loans.

Table 2 – Comparison between SMEs and large firms

Variables	*Unclassified Credit Rating*	*Default Rate*	*Interest Rate*
	(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. *Unclassified credit rating* is an indicator that equals one if a loan application does not have a credit rating (marked as unclassified in the data) and zero otherwise. *Default rate* is an indicator that equals one if a loan is defaulted and zero otherwise. *Interest rate* 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: Unclassified 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 unclassified credit rating, an indicator that equals one if a loan application does not have a credit rating (marked as unclassified 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: 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	*Unclassified Credit Rating*		*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 unclassified credit rating and loan default rate. The data sample is based on the pre and post four quarters of 2018Q1/2018Q2. *Unclassified credit rating* is an indicator that equals one if a loan application does not have a credit rating (marked as unclassified in the data) and zero otherwise. *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	
	Unclassified Credit Rating	Default Rates	Unclassified Credit Rating	Default Rates
	(1)	(2)	(3)	(4)
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 unclassified credit rating and loan default rate. *Unclassified credit rating* is an indicator that equals one if a loan application does not have a credit rating (marked as unclassified in the data) and zero otherwise. *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's pre-adoption unclassified credit rating rates exceeding 5% (or 10%) and zero otherwise. *Post* is a time indicator that equals one if the time is after 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 – Firm-level heterogeneous analysis (I)

Variables	*Missing Information*		*Non-SOE*	
	Unclassified Credit Rating	Default Rates	Unclassified Credit Rating	Default Rates
	(1)	(2)	(3)	(4)
Dummy × SME × Post	-0.028***	-0.056***	-0.021**	-0.077***
	(-6.56)	(-5.52)	(-2.50)	(-6.22)
SME × Post	-0.006**	0.010***	-0.007	0.008
	(-2.31)	(3.56)	(-0.87)	(1.58)
Dummy × Post	0.019***	0.054***	0.006	0.081***
	(4.71)	(5.35)	(0.96)	(6.87)
Dummy × SME	0.010	-0.018**		
	(1.63)	(-2.14)		
Dummy	-0.002	0.031***		
	(-0.32)	(3.75)		
Constant	0.032***	0.026***	0.043***	0.034***
	(18.01)	(12.85)	(11.08)	(15.72)
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	4,378,877	4,358,049
R^2	0.706	0.709	0.706	0.708

Note: This table presents the panel regression results on the influence of the adoption of AI and big data on the unclassified credit rating and loan default rate. *Dummy* is an indicator representing two types of firm-level heterogeneity: first, it equals one if a firm is missing financial information and zero otherwise; second, it equals one if a firm is state-owned 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 8 – Firm-level heterogeneous analysis (II)

Variables	*First-time borrower*		*Cross-region borrower*	
	Unclassified Credit Rating	Default Rates	Unclassified Credit Rating	Default Rates
	(1)	(2)	(3)	(4)
Dummy × SME × Post	-0.001	-0.006***	-0.015*	-0.013
	(-0.31)	(-2.63)	(-1.69)	(-0.73)
SME × Post	-0.024***	-0.027**	-0.023***	-0.026**
	(-5.78)	(-1.98)	(-5.14)	(-1.99)
Dummy × Post	0.003	-0.006***	-0.002	0.005
	(0.85)	(-2.62)	(-0.21)	(0.32)
Dummy × SME			-0.001	0.018*
			(-0.22)	(1.85)
Dummy			0.017***	-0.015
			(-2.16)	(-1.62)
Constant	0.044***	0.062***	0.044***	0.059***
	(16.61)	(7.15)	(15.36)	(6.93)
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	4,378,877	4,358,049
R^2	0.706	0.708	0.706	0.708

Note: This table presents the panel regression results on the influence of the adoption of AI and big data on the unclassified credit rating and loan default rate. *Dummy* is an indicator representing two types of firm-level heterogeneity: first, it equals one if a firm is the first-time borrower and zero otherwise; second, it equals one if a firm is the cross-region borrower 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 9 – Loan-level heterogeneous analysis

Variables	*Uncollateralized loans*		*Short-term loans*	
	Unclassified Credit Rating	Default Rates	Unclassified Credit Rating	Default Rates
	(1)	(2)	(3)	(4)
Dummy × SME × Post	-0.041***	0.033**	-0.011	-0.040***
	(-8.48)	(1.98)	(-1.35)	(-2.73)
SME × Post	-0.011***	-0.045***	-0.012	0.007
	(-9.45)	(-4.07)	(-1.63)	(1.05)
Dummy × Post	-0.006	-0.012	-0.001	0.033**
	(-1.32)	(-0.69)	(-0.10)	(2.29)
Dummy × SME	0.042***	-0.017***	0.024**	-0.003
	(20.22)	(-2.88)	(2.50)	(-0.69)
Dummy	0.005***	-0.008	-0.021**	-0.010**
	(3.10)	(-1.40)	(-2.16)	(-2.30)
Constant	0.030***	0.076***	0.042***	0.054***
	(38.19)	(10.74)	(8.82)	(12.54)
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	4,378,877	4,358,049
R^2	0.708	0.709	0.706	0.708

Note: This table presents the panel regression results on the influence of the adoption of AI and big data on the unclassified credit rating and loan default rate. *Dummy* is an indicator representing two types of loan-level heterogeneity: first, whether a loan is a secured loan (with collateral) and zero otherwise; second, whether a loan is a short-term loan (less than one year) 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 10 – Region-level heterogeneous analysis

Variables	*Less developed districts*		*Dialects districts*	
	Unclassified Credit Rating	Default Rates	Unclassified Credit Rating	Default Rates
	(1)	(2)	(3)	(4)
Dummy × SME × Post	-0.017**	-0.044***	-0.034***	-0.026
	(-3.27)	(-2.91)	(-6.44)	(-1.39)
SME × Post	-0.017***	-0.011	-0.012***	-0.019
	(-3.53)	(-1.06)	(-2.64)	(-1.26)
Dummy × Post	0.011**	0.060***	0.014***	0.030
	(2.18)	(3.98)	(2.93)	(1.64)
Dummy × SME	0.007	0.009	0.024***	0.015
	(0.76)	(0.80)	(3.65)	(1.42)
Dummy			-0.008	-0.022**
			(-1.26)	(-2.14)
Constant	0.040***	0.038***	0.037***	0.054***
	(6.77)	(3.73)	(11.76)	(5.54)
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	4,109,026	4,089,293
R^2	0.706	0.708	0.710	0.712

Note: This table presents the panel regression results on the influence of the adoption of AI and big data on the unclassified credit rating and loan default rate. *Dummy* is an indicator representing two types of region-level heterogeneity: first, whether a borrower is located in a more economically developed region and zero otherwise; second, whether a borrower is located in a district with more than two dialects 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 11 – Bank credit accessibility and interest payment

Variables	*Loan Amount*		*Interest Rate*	
	(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 rate* 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 12 – The synergy between big data and AI models

Variables	*Unclassified Credit Rating*	*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 the unclassified credit rating and loan default rate. *Unclassified credit rating* is an indicator that equals one if a loan application does not have a credit rating (marked as unclassified in the data) and zero otherwise. *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 rate* 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.

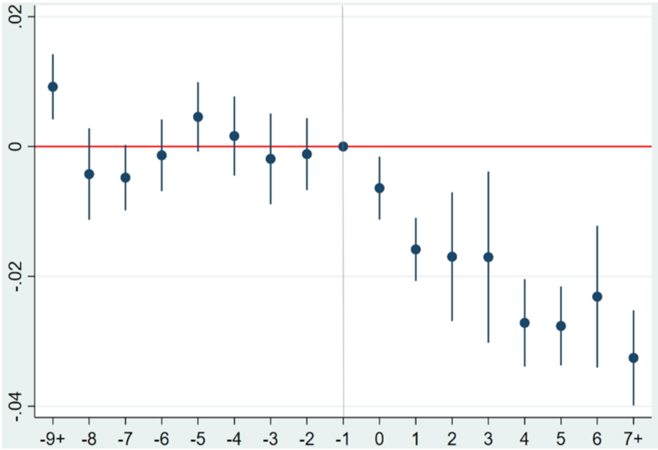
Table 13 – The synergy between big data and AI models (results from a sample with firm-level financial information)

Variables	*Unclassified Credit Rating*	*Default Rate*	*Loan Amount*	*Interest Rate*
	(1)	(2)	(3)	(4)
SME × Post1	-0.005***	-0.018**	0.030	-0.177***
	(-2.88)	(-2.27)	(1.57)	(-13.42)
SME × Post2	-0.012***	-0.020*	-0.013	-0.062**
	(-4.13)	(-1.78)	(-0.51)	(-2.42)
Constant	0.026***	0.081***	15.687***	4.985***
	(43.05)	(32.68)	(2,795.70)	(1,004.11)
Controls	YES	YES	YES	YES
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	1,812,496	1,799,970	670,571	1,811,740
R^2	0.624	0.719	0.769	0.812

Note: This table presents the panel regression results on the influence of the adoption of AI and big data on unclassified credit rating and loan default rate. *Unclassified credit rating* is an indicator that equals one if a loan application does not have a credit rating (marked as unclassified in the data) and zero otherwise. *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 rate* 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. Controls refer to firm-level financial information including total assets and total debts. T-statistics values are shown in parentheses. The superscript ***, **, or * indicates statistical significance at the 1%, 5% or 10% level, respectively.

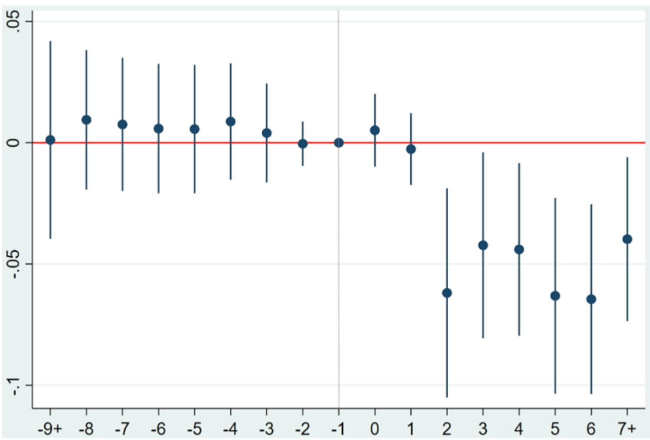
Figure 1 – Parallel trends test

Panel A: Unclassified 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 unclassified credit rating, an indicator that equals one if a loan application does not have a credit rating (marked as unclassified in the data). The estimated coefficients are relative to the one in the first quarter of 2019 (t = -1).

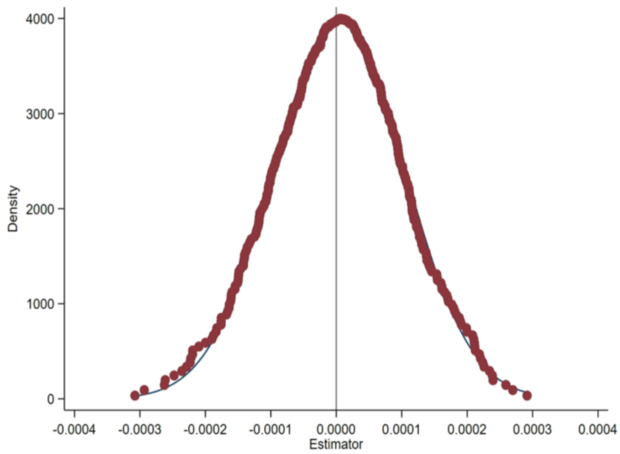
Panel B: Loan default rate



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).

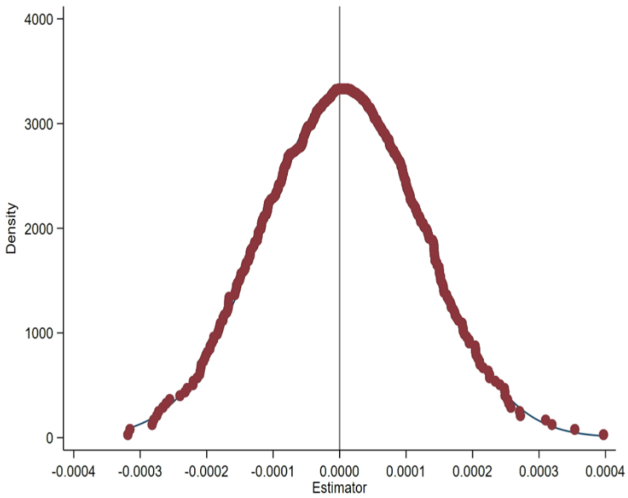
Figure 2 – Placebo test

Panel A: Unclassified 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 unclassified 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	*Unclassified credit rating*	*Default rate*	*Unclassified credit rating*	*Default rate*
	(1)	(2)	(3)	(4)
SME × Post	-0.024***	-0.027**	-0.026***	-0.026**
	(-5.87)	(-2.16)	(-5.33)	(-1.99)
Financial Infor	0.006***	0.015***		
	(4.56)	(7.16)		
GDP			0.002	-0.003
			(0.49)	(-0.53)
Fiscal Revenue			0.000	-0.009**
			(0.03)	(-2.34)
Constant	0.040***	0.046***	0.029*	0.217***
	(14.70)	(4.96)	(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	0.708	0.689	0.711

Note: This table presents the panel regression results on the influence of the adoption of AI and big data on unclassified credit rating and loan default rate. *Unclassified credit rating* is an indicator that equals one if a loan application does not have a credit rating (marked as unclassified in the data) and zero otherwise. *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 second quarter of 2019 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.

Table A4 – Small firms versus medium firms

Variables	*Unclassified Credit Rating*	*Default Rate*
	(1)	(2)
Small × *Post*	-0.023***	-0.034***
	(-10.81)	(-9.42)
Constant	0.044***	0.062***
	(33.61)	(27.52)
Firm F.E.	YES	YES
Industry F.E.	YES	YES
Region F.E.	YES	YES
Quarter F.E.	YES	YES
Observations	4,172,952	4,155,047
R^2	0.703	0.708

Note: This table presents the panel regression results on the influence of the adoption of AI and big data on unclassified credit rating and loan default rate, restricting the data sample of SMEs. *Unclassified credit rating* is an indicator that equals one if a loan application does not have a credit rating (marked as unclassified in the data) and zero otherwise. *Default rate* is an indicator that equals one if a loan is defaulted and zero otherwise. *Small* is an indicator that equals one if a firm is a small-size firm and zero if a firm is medium-size firm. *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 A5 – Adding more fixed-effects

Variables	*Unclassified Credit Rating*			*Default Rate*		
	(1)	(2)	(3)	(5)	(4)	(6)
SME × Post	-0.024***	-0.029***	-0.027***	-0.028**	-0.026***	-0.027***
	(-7.50)	(-8.28)	(-8.23)	(-2.23)	(-3.73)	(-3.74)
Constant	0.045***	0.048**	0.047**	0.060***	0.058***	0.059***
	(21.79)	(21.44)	(22.16)	(7.46)	(13.19)	(12.89)
Firm F.E.	YES	YES	YES	YES	YES	YES
Industry F.E.	YES	YES	YES	YES	YES	YES
Region F.E.	YES	YES	YES	YES	YES	YES
Time F.E.	YES	YES	YES	YES	YES	YES
Industry × Time	YES		YES	YES		YES
Region × Time		YES	YES		YES	YES
Observations	4,378,872	4,378,870	4,378,865	4,358,044	4,358,042	4,358,037
R^2	0.707	0.714	0.715	0.709	0.714	0.714

Note: This table presents the panel regression results on the influence of the adoption of AI and big data on unclassified credit rating and loan default rate. *Unclassified credit rating* is an indicator that equals one if a loan application does not have a credit rating (marked as unclassified in the data) and zero otherwise. *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 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 A6 – The synergy between big data and AI models (restrict sample with firm-level financial information)

Variables	*Unclassified Credit Rating*	*Default Rate*	*Loan Amount*	*Interest Rate*
	(1)	(2)	(3)	(4)
SME × Post1	0.001	-0.006	0.015	-0.005
	(0.12)	(-0.58)	(0.44)	(-0.18)
SME × Post2	-0.013***	-0.029***	-0.008	-0.056**
	(-3.80)	(-2.72)	(-0.28)	(-1.99)
Constant	0.022***	0.070***	15.415***	4.397***
	(3.62)	(5.96)	(474.58)	(167.40)
Controls	YES	YES	YES	YES
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	364,322	362,324	138,712	364,277
R^2	0.576	0.807	0.733	0.836

Note: This table presents the panel regression results on the influence of the adoption of AI and big data on unclassified credit rating and loan default rate. *Unclassified credit rating* is an indicator that equals one if a loan application does not have a credit rating (marked as unclassified in the data) and zero otherwise. *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 rate* 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. Controls refer to firm-level financial information including total assets, total debts and firm sales. T-statistics values are shown in parentheses. The superscript ***, **, or * indicates statistical significance at the 1%, 5% or 10% level, respectively.