The Transformative Role of Artificial Intelligence and Big Data in Banking

Junjie Xia
Central University of Finance and Economics
Peking University
(junjiexia@nsd.pku.edu.cn)

Zirun Zhang
Central University of Finance and Economics

September 12, 2024

Abstract

Utilizing a comprehensive dataset of over 4.5 million loans from a major commercial bank in China, we find that the integration of artificial intelligence (AI) and big data significantly improves credit rating accuracy and reduces loan default rates. Specifically, the rate of undefined credit ratings decreases by 3.6 percentage points, and the loan default rate decreases by 2.8 percentage points. By leveraging the temporal differences in the bank's adoption of various technologies, we identify that the implementation of big data with advanced recognition technologies has more profound impact than machine learning. Our findings highlight the informational advantage of AI and big data and provide evidence of the transformative role that AI and big data play in enhancing the operational efficiency and risk management capabilities of financial institutions.

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

1. Introduction

The integration of artificial intelligence (AI) and big data into the banking sector has marked a significant turning point in how financial institutions operate in recent years. These technologies offer unprecedented capabilities in data analysis and decision-making. While existing finance studies on AI and big data have highlighted their importance in various areas, such as fund performance, corporate culture, market microstructure, and distributional effects (e.g., Easley et al., 2021; Li et al., 2021; Fuster et al., 2022; DeMiguel et al., 2023), there remains a relative scarcity of research on how AI and big data influence the banking sector.

This paper addresses this gap by leveraging a large-scale dataset of over 4.5 million loans from a major commercial bank in China. This dataset provides a valuable opportunity to explore this topic in depth. One of the primary advantages is that it allows us to pinpoint the exact timeline when the bank introduced AI and big data technologies, especially in the realm of credit rating. This temporal information is crucial for investigating the broader issue of comparing the effectiveness of new technologies against traditional models in the banking sector.[1] Furthermore, the bank underwent two significant phases in its adoption of AI and big data: the introduction of machine learning and algorithms in July 2019, followed by the integration of big data analytics in October 2020. These distinct phases enable us to conduct a more nuanced analysis of the differences and impacts of various AI technologies, offering insights into their respective roles and effectiveness in transforming banking operations.

Historically, the bank has relied on traditional models such as shadow ratings and hierarchical analysis to assess creditworthiness. These methods, while effective to some

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

extent, often resulted in a high incidence of "undefined/undetermined" credit ratings due to insufficient information, ultimately necessitating reliance on human judgment and experience. Such ratings frequently led to either the withdrawal of loan applications or the imposition of particularly stringent loan terms. This issue is especially prevalent among small and medium-sized enterprises (SMEs), as their financial and operational data are less comprehensive and more difficult to obtain compared to larger corporations. During our sample period from 2015 to 2023, SMEs accounted for 89.1% of the "undefined/undetermined" credit ratings.

The advent of AI and big data has revolutionized this landscape. By integrating these technologies in June 2019, the bank completely reformed its credit rating system, replacing human decision-making with AI algorithms. This transformation has significantly optimized and upgraded its credit assessment models. Prior to the introduction of AI and big data, "undefined/undetermined" credit ratings accounted for approximately 6% of over 4.5 million loan credit ratings. This figure decreased to about 2% following the implementation of these advanced technologies. This transformation underscores the potential of AI and big data to enhance the accuracy and efficiency of credit assessments by improving the gathering and processing of information and unstructured data, thereby reducing the reliance on subjective human judgment and improving access to credit for SMEs. The information advantage of big data is consistent with the theoretical prediction by Livshits et al. (2016) and Hau et al. (2019).

We therefore consider SMEs as a treatment group that benefits more from AI and big data, compared to the control group of large firms. We use the bank's initial introduction of AI and big data as an external shock to the credit rating system and employ a difference-in-difference (DID) approach to examine their impact on the bank. Our analysis reveals that the rate of undefined credit ratings among SMEs decreases by 2.4 percentage points compared to large firms. Using the temporal differences in the adoption of technologies between machine learning algorithm (logistic regression

model) and big data analytics, we find that the implementation of big data with advanced recognition technologies exerts a larger effect than machining learning. This is largely because big data, combined with advanced recognition technologies such as optical character recognition (OCR) and natural language processing (NLP), enables the bank to consolidate a diverse range of internal and external data sources, including unstructured data. This integration provides a wealth of real time information, empowering the bank to gain deeper insights into customer behavior, market trends, and potential risks. Furthermore, we are able to estimate the combining effects of integrating machine learning and big data analytics. Specifically, the incidence of undefined credit ratings decreases by 3.6 percentage points.

In addition to examining credit ratings, we explore the impact on the loan default rate and interest payment. Our estimates indicate that the loan default rate decreases by 2.8 percentage points following the introduction of these technologies. This reduction is primarily driven by the adoption of big data analytics with advanced recognition technologies. Additionally, interest payments were higher among SMEs compared to large firms in the pre-AI adoption period. However, this gap has been narrowing as the bank becomes more capable of gathering comprehensive information.

We further conduct a series of heterogeneity analyses to demonstrate the information advantage provided by AI and big data. Firstly, we find that the impact on credit ratings is more significant for unsecured loans, which lack collateral. This suggests that lenders are placing less reliance on collateral to mitigate information asymmetries and moral hazard issues, as discussed by Aghion and Bolton (1992). Secondly, our analysis reveals that the effect is more pronounced in regions with lower levels of economic development and among firms with less publicly available information. This underscores the potential of AI and big data to enhance credit assessment processes, particularly in environments where traditional information sources are limited or less reliable.

Our paper contributes to several strands of literature. First, it is closely aligned with research on the real effects of adopting machine learning and big data. While most existing studies have focused on how these technologies influence various aspects such as corporate culture (Li et al., 2021), board decision-making (Erel et al., 2021), fund performance (DeMiguel et al., 2023), firm growth and innovation (Babina et al., 2024), market microstructure (Easley et al., 2021), and disproportionate effects (Fuster et al., 2022), our study offers new empirical evidence on how AI and big data are transforming the banking industry. By examining the impact on credit ratings and loan default rates, our research highlights the potential of these technologies to enhance decision-making processes and reduce information asymmetries in financial markets. This contribution not only broadens the understanding of AI and big data's role in finance but also underscores their significance in shaping the future of banking operations and strategies.

Our paper also contributes to the broader literature on how FinTech access impacts small businesses. Frost et al. (2019) explore the factors and outcomes associated with FinTech in global finance, highlighting that FinTech credit is particularly significant in regions with less competitive banking sectors. Their study provides initial insights into the informational benefits of FinTech and its influence on product choices. Building on this, Agarwal et al. (2019, 2022) utilize data from individual credit lines on Alibaba's online platform to demonstrate that mobile payment technology significantly affects small businesses by enhancing customer acquisition and formation, while also encouraging increased credit card usage among consumers. Furthermore, Hau et al. (2024) empirically confirm that access to FinTech credit boosts vendor sales growth, transaction growth, and customer satisfaction, particularly for vendors facing greater information asymmetry regarding credit risk. Our study aligns with this existing body of literature by showing that SMEs benefit from the adoption of FinTech, particularly in terms of improved access to bank credit. This underscores the transformative potential of FinTech in leveling the playing field for SMEs, enabling them to compete more effectively in the financial landscape.

In sum, the primary contribution of our paper lies in its comprehensive examination of the effects of technological integration within a real-world banking environment, offering compelling empirical evidence of the transformative role that artificial intelligence (AI) and big data play in the banking industry. By leveraging a large-scale dataset from a major commercial bank, our study not only highlights the significant improvements in credit assessment accuracy and efficiency brought about by these technologies but also underscores their potential to revolutionize traditional banking operations.

This paper is organized as follows. Section 2 describes institutional background and our data sample. Section 3 discusses our empirical methodology. Section 4 discusses the results and provides further analysis and robustness checks. Section 5 concludes the paper.

2. Background and Data

2.1 Institutional Background

Our data sample is sourced from one of the largest commercial banks in China.[2] In an effort to advance the development of financial technology and align with international standards, the Chinese government announced a Three-year Development Plan (2019-2021) for FinTech. This plan represents the first nationwide policy aimed at encouraging the adoption of AI and big data within the banking sector. In response to this policy, the bank has undertaken a strategic implementation to align with the government's objectives.

The bank began its efforts by focusing on the credit rating system. Traditionally, the bank relied on conventional models such as shadow ratings and hierarchical analysis, which heavily depended on human judgment and the quality of data inputs. Therefore,

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

these methods often faced limitations, particularly when there was insufficient input data, resulting in outputs that were frequently labeled as "undefined" or "undetermined."

In July 2019, the bank initially implemented machine learning techniques, specifically utilizing a logistic regression model, to replace human decision-making in its credit rating processes. By October 2020, the bank had further advanced its capabilities by incorporating big data analytics, in conjunction with sophisticated text recognition technologies such as Optical Character Recognition (OCR) and Natural Language Processing (NLP). The integration of big data involved accessing external large-scale databases, including the National Business Registration System, which contains registration information for all Chinese enterprises, and the National Intellectual Property Administration database. Additionally, it encompassed unstructured data that necessitated advanced text recognition technologies, such as firm-to-firm transaction receipts and images. This comprehensive approach enabled the bank to enhance its data processing capabilities and improve the accuracy and efficiency of its credit rating system. Therefore, since the third quarter of 2020, the bank has integrated machine learning and big data analytics into its credit rating system.

2.2 Data

Our sample comprises approximately 4.53 million loans for 544,453 firms, spanning from the beginning of 2015 to the end of 2023. This dataset is rich in loan information, including credit ratings, interest rates, and default rates. The loans cover all provinces and all 2-digit industries in China.[3] Table 1 presents the summary statistics of our data sample. We compare key variables for large firms and SMEs before and after the bank's

[3] 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%.

adoption of AI and big data technologies. As shown in the table, SMEs account for a larger number of loans compared to large firms.

[Table 1 about here]

Prior to the adoption of AI and big data, the average rate of undefined credit ratings among large firms was approximately 0.72%, while for SMEs, it was about 5.95%. Following the implementation of AI and big data technologies, these rates declined for both categories, with a particularly notable decrease for SMEs. The average undefined credit rate for SMEs dropped to 2.08%. A similar pattern was observed in the loan default rates. The average loan default rate for large firms experienced a slight reduction from 6.31% to 5.67%, whereas for SMEs, it significantly decreased from 9.12% to 2.14%. In summary, both the rate of undefined credit and the rate of loan defaults experience a substantial change after the implementation of AI and big data, especially for SMEs.

In addition, the difference in interest payments between large firms and SMEs has been narrowing. 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.

3. Empirical analysis

3.1 Empirical specification

To investigate the influence of AI and big data, we employ a difference-in-difference (DID) methodology, where SMEs are considered the experimental group and larger

enterprises serve as the control group. We leverage the bank's initial implementation of AI and big data as an external shock to the credit rating system. This approach allows us to isolate the effects of AI and big data based on their information advantage. We first estimate the impact on credit ratings by utilizing the following regression equation:

$$UnCredit_{i,t} = \beta_1 SME_i + \beta_2 Post_t + \beta_3 SME_i \times Post_t + \varphi_f + \gamma_j + \theta_t + \delta_r + \varepsilon_{i,t} \quad (1)$$

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

We incorporate a comprehensive set of fixed effects, encompassing firm-level, industry-level, region-level, and quarter-level effects, to control for unobservable factors in our regression models. We also allow for clustering of observations at the firm level to account for possible serial correlation in the data. The estimate of β_3 is our focal point of interest.

3.2 Baseline results

Table 2 presents the panel regression results analyzing the impact of AI and big data adoption on credit ratings. The key coefficient of interest is the interaction term between *SME* and *Post*. Column (1) does not include any fixed effects, whereas Columns (2) and (3) incorporate all fixed effects as specified in Equation (1). Column (3) presents our baseline results, controlling for quarter fixed effects and other dimensions of fixed effects. Across all specifications, the results indicate that the coefficient for the interaction term is negative and statistically significant at 1% level. In terms of

economic magnitude, it indicates that the rate of undefined credit ratings among SMEs decreases by 2.4 percentage points compared to that of large firms. This finding underscores the substantial impact of AI and big data technologies in improving credit assessment accuracy, particularly benefiting SMEs by reducing their undefined credit rates.

[Table 2 about here]

We proceed to examine the impact of AI and big data on the loan default rate by modifying our regression model to use *Default* as the dependent variable in Equation (1). Here, *Default* is a binary indicator that equals one if the loan is defaulted and zero otherwise. Table 3 presents the corresponding estimation results. The coefficient for the interaction term is negative and statistically significant at the 1% level. In terms of economic magnitude, this finding suggests that, compared to large firms, the loan default rate for SMEs declines by 2.7 percentage points. This result highlights the effectiveness of AI and big data technologies in enhancing credit risk assessment, thereby reducing the likelihood of loan defaults, particularly for SMEs.

[Table 3 about here]

We next to investigate the influence of AI and big data on interest payments by modifying our regression model to use *Interest* as the dependent variable in Equation (1). Here, *Interest* refers to the interest rate of a loan. Table 4 presents the corresponding estimation results, which are consistent with the findings in Table 2 and Table 3. Specifically, the coefficient for the interaction term is negative and statistically significant at the 1% level, even after incorporating various dimensions of fixed effects. The results indicate that, compared to large firms, the interest rate for SMEs decreases by 0.336 percentage points following the adoption of AI and big data technologies. This suggests that the gap in interest payments between the two groups has narrowed,

highlighting the potential of AI and big data to enhance credit assessment and reduce borrowing costs for SMEs.

[Table 4 about here]

3.3 Machine learning verse big data

The bank experienced two significant phases in its adoption of AI and big data technologies. Initially, in July 2019, the bank implemented machine learning techniques, specifically utilizing a logistic regression model. By October 2020, the bank had further enhanced its capabilities by incorporating big data analytics alongside advanced text recognition technologies, including OCR and NLP.

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

$$Y_{i,t} = \beta_1 SME_i + \beta_2 Post1_t + \beta_3 Post1_t + \beta_4 SME_i \times Post1_t + \beta_5 SME_i \times Post2_t + \varphi_f + \gamma_j + \theta_t + \delta_r + \varepsilon_{i,t} \quad (2)$$

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

Table 5 presents the corresponding estimation results, focusing on the key coefficients of interest for the two interaction terms. Column (1) provides the estimates for undefined credit ratings. The coefficient for the first interaction term is -0.016, while the coefficient for the second interaction term is -0.02. Both coefficients are statistically significant at the 1% level. This indicates that the adoption of big data with advanced recognition technologies has a more substantial impact than the initial implementation of machine learning technology.

Since the third quarter of 2020, the bank has integrated machine learning and big data into its credit rating system. Our results allow us to estimate the combined effects of these technologies, calculated as 0.016 + 0.02 = 0.036. This suggests that the incidence of undefined credit ratings decreases by 3.6 percentage points following the bank's implementation of machine learning and big data analytics. This finding underscores the enhanced accuracy and efficiency in credit rating processes achieved through the combination of these advanced technologies.

Column (2) presents the estimates for loan default. The coefficient for the first interaction term is -0.015, but it is not statistically significant. In contrast, the coefficient for the second interaction term is -0.028 and is statistically significant. This suggests that the primary effect is largely driven by integration of big data analytics with advanced recognition technologies. In terms of economic impact, the loan default rate decreases by 2.8 percentage points following the adoption of AI and big data. This highlights the effectiveness of integrating more complex technologies in reducing the risk of loan defaults, thereby enhancing the bank's risk management capabilities.

[Table 5 about here]

4. Heterogeneity and Robustness

4.1 Heterogeneous analysis

We further conduct three sets of heterogeneity analyses to illustrate the information advantage provided by AI and big data. First, we examine the heterogeneity in loan types based on the provision of collateral. Our findings indicate that the impact on credit ratings is more pronounced for unsecured loans, which lack collateral. This suggests that lenders are increasingly relying on AI and big data to address information asymmetries and moral hazard issues, rather than depending solely on collateral, as discussed by Aghion and Bolton (1992).

Second, we utilize two additional proxies to assess information availability: the level of economic development and the borrower's ownership structure. We hypothesize that regions with higher levels of economic development possess more robust financial markets and experience fewer information asymmetries. Additionally, if a borrower is a state-owned enterprise (SOE), it is likely to have more publicly available information, as the Chinese government mandates SOEs to disclose key corporate information to the public. This analysis underscores the potential of AI and big data to enhance credit assessment processes, particularly in environments where traditional information sources are either limited or more readily available.

Table 6 presents the corresponding results. Our analysis reveals that the effect is more pronounced in regions with lower levels of economic development and among firms with less publicly available information. This demonstrates the potential of AI and big data to enhance credit assessment processes, particularly in environments where traditional information sources are limited or less reliable.

[Table 6 about here]

4.2 Robustness

We conduct robustness checks on our baseline results, specifically by performing a parallel trends test. Figures 1 and 2 present the corresponding estimates for undefined credit ratings and loan default rates. Both figures indicate that there is no significant trend in these outcomes prior to the adoption of AI and big data. Furthermore, there is a noticeable shift in both the magnitude and significance of the coefficients following the adoption of these technologies. This evidence supports our identification assumption, reinforcing the validity of our findings.

[Figure 1 and Figure 2 about here]

5. Conclusion

In conclusion, this study provides compelling evidence of the transformative role of AI and big data in 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 incidence of "undefined/undetermined" credit ratings, especially among small and medium-sized enterprises (SMEs). This reduction is achieved through the improved accuracy and efficiency of credit assessments, which are facilitated by the advanced data processing capabilities of AI and big data.

Our findings reveal that the adoption of big data analytics, combined with machine learning algorithms, not only decreases the rate of undefined credit ratings but also lowers the loan default rate and narrows the interest payment gap between SMEs and larger firms. These outcomes highlight the potential of AI and big data to mitigate information asymmetries and enhance decision-making processes in financial institutions.

Furthermore, our heterogeneity analyses underscore the broader applicability of these technologies, particularly in regions with lower economic development and among firms with limited publicly available information. This suggests that AI and big

data can play a crucial role in democratizing access to credit and fostering financial inclusion.

Overall, our research contributes to the growing body of literature on the real-world effects of technological integration in finance, offering new insights into how AI and big data are reshaping the banking industry. As financial institutions continue to embrace these innovations, our study underscores the importance of leveraging AI and big data to drive efficiency, reduce risks, and enhance the overall competitiveness of the banking sector.

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

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

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.

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

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

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

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

Table 1 – Summary Statistics

	Before		After	
	Large firms	SMEs	Large firms	SMEs
Firms	7395	162991	4366	369701
Loans	176504	1398131	53139	2902114
Undefined credit rating loans	10978	94243	6887	51982
Rate of undefined credit rating	0.72%	5.95%	0.31%	2.08%
Rate of loan default	6.31%	9.12%	5.67%	2.14%
Interest rate	4.64%	5.35%	3.45%	3.94%

Note: This table presents the summary statistics for the data sample from 2015 to 2023. Before refers to the pre-adoption of AI and big data period. After refers to the post-adoption of AI and big data period.

Table 2 – Undefined credit rating

Variables	*Dependent Variable: Undefined Credit Rating*		
	(1)	(2)	(3)
SME × *Post*	-0.117***	-0.025***	-0.024***
	(-3.86)	(-6.75)	(-5.62)
Post	0.067**	0.015***	
	(2.22)	(3.72)	
SME	0.005		
	(0.18)		
Constant	0.062**	0.036***	0.045***
	(2.13)	(49.63)	(16.40)
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	4,529,888	4,378,847	4,378,847
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 undefined credit rating, an indicator that equals one if a loan application does not have a credit rating (marked as undefined 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 3 – Loan default rate

Variables	*Dependent Variable: Loan Default*		
	(1)	(2)	(3)
SME × *Post*	-0.062***	-0.027**	-0.027**
	(-5.18)	(-2.01)	(-2.12)
Post	-0.015	0.023*	
	(-1.30)	(1.76)	
SME	0.029*		
	(1.76)		
Constant	0.065***	0.044***	0.059***
	(3.99)	(62.72)	(7.20)
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	4,507,637	4,358,006	4,358,006
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 4 – Interest rate

Variables	*Dependent Variable: Interest Rate*		
	(1)	(2)	(3)
SME × *Post*	-0.230	-0.348***	-0.336***
	(-0.96)	(-6.16)	(-7.09)
Post	-1.185***	0.388***	
	(-4.95)	(9.78)	
SME	0.713***		
	(4.83)		
Constant	4.637***	4.389***	4.633***
	(31.45)	(345.99)	(152.69)
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	4,529,100	4,378,065	4,378,065
R^2	0.302	0.793	0.807

Note: This table presents the panel regression results on the influence of the adoption of AI and big data on the interest rate. The dependent variable is interest rate, referring 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 5 – Comparison between machine learning and big data

Variables	*Undefined credit rating*	*Loan default*
	(1)	(2)
SME × *Post1*	-0.016***	-0.015
	(-5.21)	(-1.10)
SME × *Post2*	-0.020***	-0.028**
	(-6.35)	(-2.03)
Constant	0.051***	0.067***
	(29.97)	(11.72)
Firm F.E.	YES	YES
Industry F.E.	YES	YES
City F.E.	YES	YES
Quarter F.E.	YES	YES
Observations	4,378,847	4,358,006
R^2	0.706	0.708

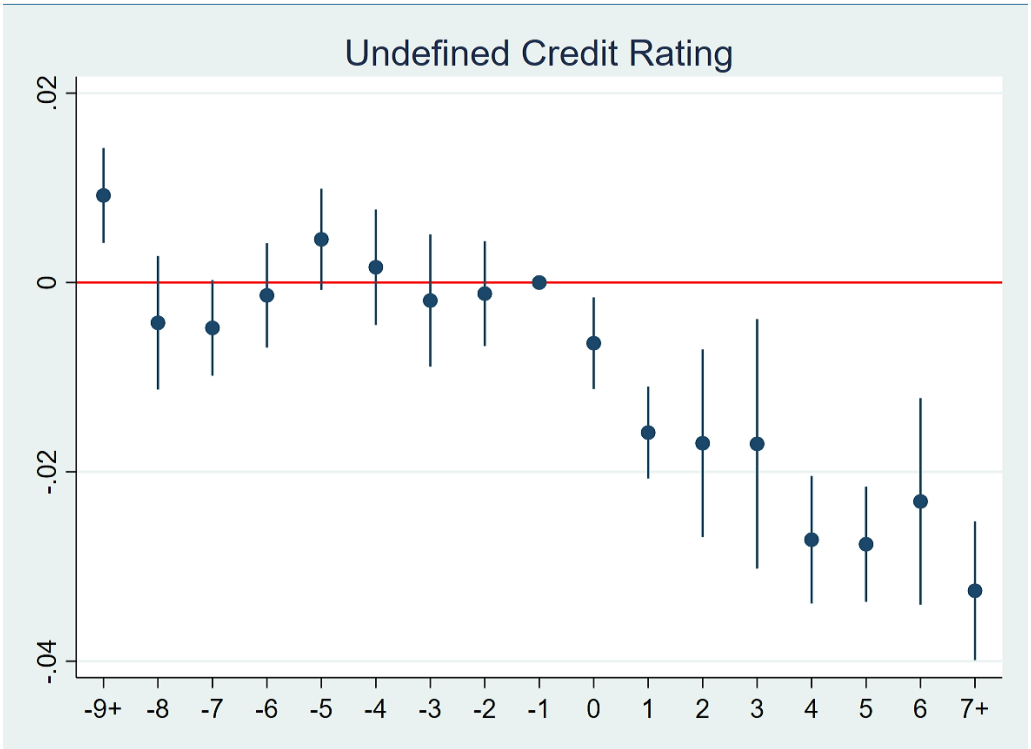
Note: This table presents the panel regression results on the influence of the adoption of AI and big data on credit rating and the loan default. Undefined credit rating is an indicator that equals one if a loan application does not have a credit rating (marked as undefined in the data) and zero otherwise. The loan default rate is an indicator that equals one if a loan is defaulted and zero otherwise. *SME* is an indicator that equals one if a firm is a SME and zero otherwise. *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 6 – Heterogeneous analysis

Variables	Secured loans		SOE		Developed District	
	Undefined credit	Loan default	Undefined credit	Loan default	Undefined credit	Loan default
	(1)	(2)	(3)	(4)	(5)	(6)
Dummy × *SME* × *Post*	0.041***	-0.033**	0.021**	0.077***	0.017***	0.044***
	(8.44)	(-1.97)	(2.48)	(6.23)	(3.22)	(2.92)
SME × *Post*	-0.051***	-0.012	-0.028***	-0.069***	-0.035***	-0.055***
	(-10.08)	(-0.86)	(-8.75)	(-6.07)	(-14.52)	(-4.63)
Dummy × *Post*	0.006	0.011	-0.006	-0.081***	-0.011**	-0.060***
	(1.29)	(0.68)	(-0.95)	(-6.87)	(-2.14)	(-3.99)
Dummy × *Post*	-0.042***	0.017***	-	-	-0.007	-0.009
	(-20.18)	(2.88)			(-0.83)	(-0.80)
Dummy	-0.005***	0.008	-	-	-	-
	(-3.11)	(1.40)				
Constant	0.071***	0.044***	0.047***	0.087***	0.053***	0.085***
	(21.51)	(4.88)	(22.99)	(11.78)	(12.20)	(7.88)
Firm F.E.	YES	YES	YES	YES	YES	YES
Industry F.E.	YES	YES	YES	YES	YES	YES
City F.E.	YES	YES	YES	YES	YES	YES
Quarter F.E.	YES	YES	YES	YES	YES	YES
Observations	4,378,847	4,358,006	4,378,847	4,358,006	4,378,847	4,358,006
R^2	0.708	0.709	0.706	0.708	0.706	0.708

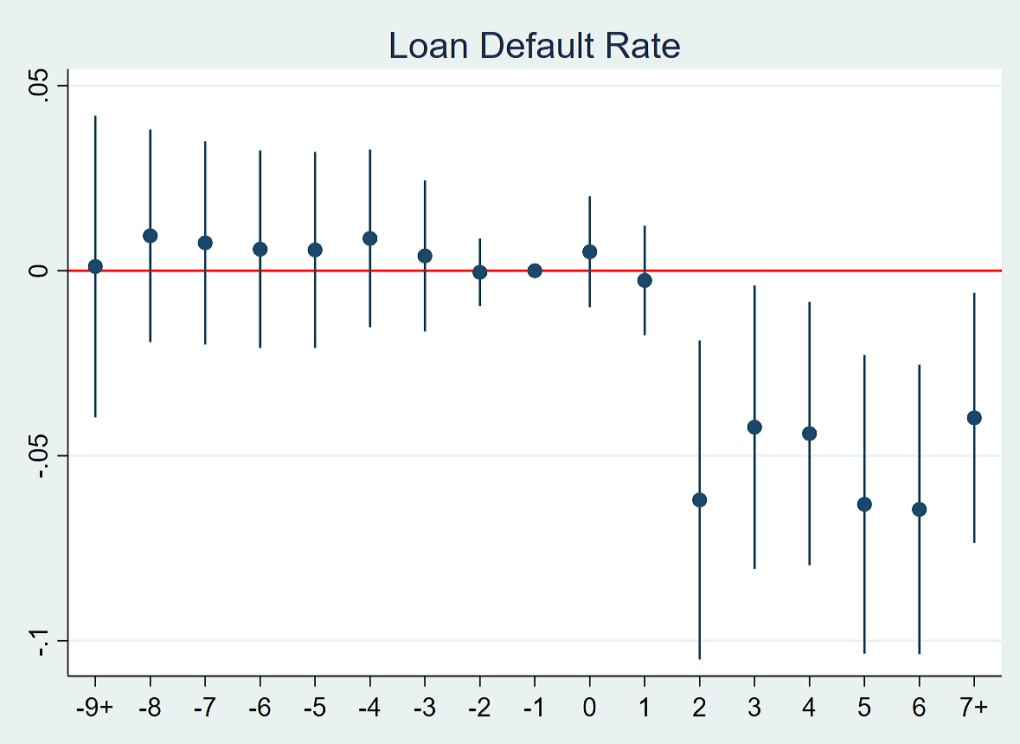
Note: This table presents the panel regression results on the influence of the adoption of AI and big data on credit rating and the loan default. Undefined credit rating is an indicator that equals one if a loan application does not have a credit rating (marked as undefined in the data) and zero otherwise.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

Figure 1 – Parallel trends test for undefined 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 undefined credit rating, an indicator that equals one if a loan application does not have a credit rating (marked as undefined in the data). The estimated coefficients are relative to the one in the second quarter of 2019 (t = -1).

Figure 2 - Parallel trends test for 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 second quarter of 2019 (t = -1).

Appendix

Table A1 Region distribution

District	Loans	Percent	District	Loans	Percent
Beijing	169201	3.74%	*Inner Mongolia*	30475	0.67%
Tianjin	67703	1.49%	*Guangxi*	91487	2.02%
Hebei	197199	4.35%	*Chongqing*	91852	2.03%
Shanghai	198749	4.39%	*Sichuan*	217266	4.80%
Jiangsu	415196	9.17%	*Guizhou*	30410	0.67%
Zhejiang	669188	14.77%	*Yunnan*	46548	1.03%
Fujian	240896	5.32%	*Shaanxi*	114408	2.53%
Shandong	283272	6.25%	*Gansu*	37948	0.84%
Guangdong	635016	14.02%	*Qinghai*	5970	0.13%
Hainan	16370	0.36%	*Ningxia*	17433	0.38%
Shanxi	74172	1.64%	*Xinjiang*	39207	0.87%
Anhui	134111	2.96%	*Liaoning*	90109	1.99%
Jiangxi	81155	1.79%	*Jilin*	67890	1.50%
Henan	144924	3.20%	*Heilongjiang*	32638	0.72%
Hubei	131418	2.90%	*Xizang*	1310	0.03%
Hunan	156367	3.45%			

Table A2 Industry distribution

Industry	Loan	Percent
Agriculture, Forestry, Animal Husbandry and Fishery	39316	0.87%
Mining	15663	0.35%
Manufacturing	1832946	40.46%
Production and Supply of Electricity, Heat Power, Gas and Water	47673	1.05%
Construction	450487	9.94%
Wholesale and Retail Trades	1435411	31.69%
Hotels and Catering Services	30877	0.68%
Information Transmission, Software and Information Technology Services	100111	2.21%
Finance	158758	3.50%
Real Estate	35899	0.79%
Leasing and Business Services	158425	3.50%
Scientific Research and Technical Services	89512	1.98%
Water Conservancy, Environment and Public Facilities Management	44481	0.98%
Services to Households, Repair and Other Services	27004	0.60%
Education	5843	0.13%
Health and Social Service	9868	0.22%
Culture, Sports and Entertainment	12397	0.27%
Other	35217	0.78%