

# 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<sup>1</sup>**

Central University of Finance and Economics

Peking University

**Zirun Zhang**

Central University of Finance and Economics

*December 10, 2024*

## **Abstract**

We study the impact of artificial intelligence (AI) and big data on the banking sector, using a comprehensive dataset of over 4.5 million loans from a major commercial bank in China spanning 2015-2023. Our analysis reveals that the integration of AI and big data significantly enhances credit rating accuracy and reduces loan default rates, particularly benefiting small and medium-sized enterprises (SMEs). Specifically, the rate of undetermined credit ratings decreases by 2.4 percentage points, while the loan default rate drops by 2.7 percentage points. By examining the timeline of the bank's adoption of different technologies, we find that integrating big data with AI models and recognition technologies has a more profound impact than traditional FinTech models. Our findings highlight the informational advantage of AI and big data, providing empirical evidence of their role in enhancing operational efficiency and risk management capabilities in financial institutions.

**JEL Classification:** G20, G21, G32

**Key words:** artificial intelligence, big data, credit rating, default rate, interest payment

---

<sup>1</sup> Corresponding author, email: junjiexia@nsd.pku.edu.cn

## 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 cutting-edge technologies are revolutionizing FinTech by providing remarkable capabilities in data analysis and decision-making processes. While existing finance studies on AI and big data have highlighted their importance in various areas, such as fund performance, corporate culture, market microstructure, distributional effects, and vendors and small business (e.g., Easley et al., 2021; Li et al., 2021; Fuster et al., 2022; DeMiguel et al., 2023, Hau et al., 2024), there remains a relative scarcity of empirical evidence on how these technologies are reshaping banking operations.

This paper addresses this gap by leveraging a large-scale dataset of over 4.5 million loans spanning 2015-2023 from a major state-owned commercial bank in China. This extensive dataset offers a unique opportunity to explore this topic in depth. An exogenous policy mandate required the bank to implement FinTech in its operation,<sup>2</sup> which allows us to pinpoint the exact timeline when the bank transitioned from conventional human-driven approaches to early FinTech models, and subsequently integrated big data along with AI models and recognition technologies. This temporal information is crucial for investigating the broader issue of comparing the effectiveness of new technologies against traditional models in the banking sector.

Specifically, we focus on credit rating, as the bank underwent two significant stages in adopting new technologies: First, in July 2019, the bank introduced machine learning algorithms like logistic regression models. Then, in October 2020, it integrated big data along with advanced AI models like artificial neural network (ANN) and federated

---

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

learning model (FLM), along with recognition technologies like optical character recognition (OCR) and natural language processing (NLP). These distinct phases enable us to conduct a more nuanced analysis of the differences and impacts of various FinTech technologies, offering insights into their respective roles and effectiveness in transforming banking operations.

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

During our sample period from 2015 to 2023, a striking 88.8% of the "undetermined" credit ratings were attributed to SMEs. This highlights the disproportionate impact of information gaps on smaller businesses, which are already at a disadvantage in accessing credit. The inability to obtain reliable data not only hampers the credit evaluation process but also restricts SMEs' access to much-needed financing.

Addressing these information asymmetries is therefore critical to improving the inclusivity and efficiency of credit markets.

Following the government's policy mandate issued in July 2019, the bank undertook a comprehensive reform of its credit rating system, replacing traditional human decision-making approaches with advanced financial technologies, particularly AI and big data analytics. This transformation marked a significant milestone in optimizing and modernizing the bank's credit assessment processes. Among the 4.5 million loan credit ratings evaluated during this period, the overall "undetermined" rate dropped substantially, from approximately 6.7% to 2.4%. Notably, SMEs contributed the most to this decline, reflecting the transformative impact of these technologies on addressing the information challenges faced by smaller businesses.

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

---

<sup>3</sup> Hau et al (2019) provide similar predictions by constructing a simple model to show that FinTech credit is often used among weak creditors.

Therefore, to further investigate how AI and big data influence credit rating, we 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 adoption of AI and big data as an external shock to the credit rating system and employ a difference-in-difference (DID) approach, which compares the changes in credit ratings before and after the adoption between SMEs and large firms. Our analysis reveals that the rate of undetermined credit ratings among SMEs decreases by 2.4 percentage points compared to large firms. This finding is consistent with our thesis that AI and big data enable the bank to enhance its information accessibility.

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

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

We continue to explore the impact on bank credit accessibility and interest payments. We find that, compared to large firms, SMEs obtain more credit in the post-AI adoption period, indicating that the bank is more willing to provide loans when it has access to more information. Additionally, interest payments were higher among SMEs compared

to large firms in the pre-AI adoption period. However, this gap has been narrowing as the bank becomes more capable of gathering comprehensive information.

We 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 in areas characterized by greater linguistic diversity, such as regions with more dialects. Thirdly, we observe that the results are particularly significant for firms with limited 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 scarce 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. SMEs often face information asymmetries and have limited credit

history, rely on "soft" information that is difficult to quantify, and lack collateral, making it challenging for them to meet bank lending criteria under strict capital requirements (Petersen and Rajan, 1994; Berger and Udell, 1995). Frost et al. (2020) 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.

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

## **2. Background and Data**

### **2.1 Institutional Background**

Our data sample is sourced from one of the largest commercial banks in China.<sup>4</sup> 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 an implementation of advanced financial technologies to align with the government's objectives.

Traditionally, banks relied on human decision-making through conventional credit rating methods, such as shadow ratings and hierarchical analysis, which were heavily dependent on human judgment and the quality of data inputs. However, while these methods have been foundational in the banking sector, they possess inherent limitations that can affect their effectiveness in accurately assessing credit ratings and evaluating loan risk.

First, traditional models often rely on a limited set of financial metrics and historical data, which may not capture the full picture of a borrower's creditworthiness. Human analysts are constrained by the volume of data they can process, leading to potential oversight of critical information that could indicate risk. Second, these models are typically based on fixed criteria and rules that do not easily adapt to changing market conditions or borrower circumstances. Third, due to insufficient information or ambiguous data, traditional models often result in a high number of "undetermined" credit ratings. This uncertainty necessitates further human intervention, which can delay decision-making and lead to either overly cautious or risky lending practices. Lastly, human models often face difficulties in addressing the problem of asymmetric information, where borrowers have more information about their financial situation

---

<sup>4</sup> Due to our data usage agreement, we are unable to disclose any information regarding the bank's identity.



than lenders. This can lead to adverse selection and moral hazard, increasing the likelihood of defaults.

In response to a government policy mandate issued in July 2019, the bank began implementing 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 enhanced its capabilities by incorporating big data analytics alongside sophisticated machine learning models, such as artificial neural networks (ANN) and federated learning models (FLM). Additionally, it employed advanced text recognition technologies like Optical Character Recognition (OCR) and Natural Language Processing (NLP).

These advanced AI models are capable of processing and analyzing large volumes of structured and unstructured data much faster and more accurately than traditional methods. The big data sources include financial statements, transaction histories, social media, and external large-scale databases, such as the National Business Registration System, which contains registration information for all Chinese enterprises, and the National Intellectual Property Administration database. Furthermore, it encompasses unstructured data that requires advanced text recognition technologies, such as firm-to-firm transaction receipts and images. These big data technologies provide real-time monitoring of borrower behavior and market conditions. This continuous surveillance allows banks to detect early warning signs of financial distress, such as sudden changes in transactions and spending patterns, enabling timely intervention to improve the accuracy and efficiency of the credit rating system and mitigate potential defaults. Consequently, since the third quarter of 2020, the bank has fully integrated machine learning and big data analytics into its system.

## **2.2 Data**

Our sample comprises approximately 4.53 million loans for 475,301 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.<sup>5</sup> Table 1 presents the summary statistics of our data sample. Panel A displays the distribution of total numbers of firms and loans across years, while Panel B compares key variables for large firms and SMEs before and after the bank's adoption of AI and big data technologies. 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 undetermined 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 undetermined 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 undetermined 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

---

<sup>5</sup> 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%.

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 undetermined credit rating, an indicator that equals one if a loan application does not have a credit rating (marked as undetermined in the data) and zero otherwise.  $SME_i$  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.  $\varphi_f$ ,  $\gamma_j$ ,  $\theta_t$ , and  $\delta_r$  represent the fixed effects on firm, industry, time and region, respectively.  $\varepsilon_{i,t}$  is the error term.

In this empirical analysis, one potential concern is the absence of firm-level control variables, such as those found in financial statements. There are two primary reasons for this omission. Firstly, firm-level variables from financial statements are integral to the credit rating process. Essentially, if a bank has access to a firm's financial statements,

the issue of an "undetermined" credit rating is likely to be resolved. This is because the availability of detailed financial information typically allows the bank to make a more informed and definitive credit assessment. Secondly, our dataset is primarily focused on loan-level data and contains limited information on firm-level characteristics. Including firm-level controls would significantly reduce the number of observations, thereby diminishing the statistical power of our regression analysis. This reduction in data points could lead to less reliable and less generalizable results.<sup>6</sup>

To address these challenges, we have incorporated a comprehensive set of fixed effects in our regression models. These include firm-level, industry-level, region-level, and quarter-level fixed effects, which help control for unobservable factors that might influence the results. Additionally, we allow for clustering of observations at the firm level to account for potential serial correlation within the data. This methodological approach ensures that our analysis remains robust despite the absence of specific firm-level controls. The primary focus of our analysis is the estimate of  $\beta_3$ , which captures the effect of interest in our study. By employing these strategies, we aim to provide a thorough and reliable examination of the factors influencing the "undetermined" credit rating situation.

### **3.2 Baseline results**

Table 2 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 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,

---

<sup>6</sup> By matching the city statistical yearbook data, we also include city-level control variables such as GDP and fiscal revenue. The estimates in Table A3 indicate that our baseline results is still valid.

controlling for quarter fixed effects as well as other dimensions of fixed effects, ensuring a robust estimation of the treatment effect.

[Table 2 about here]

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

The findings underscore the substantial impact of advanced financial technologies in improving the accuracy and efficiency of credit evaluations. SMEs, which often face greater informational asymmetries and higher barriers to accessing credit, appear to benefit disproportionately from these innovations. By leveraging AI and big data, financial institutions can process a broader range of structured and unstructured data, thereby mitigating the reliance on subjective human judgment and reducing the uncertainty associated with SME creditworthiness.

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 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 between SME and Post is negative and statistically significant at the 1% level across all specifications. This finding indicates a robust relationship between the adoption of AI and big data technologies and a reduction in loan default rates, particularly for SMEs. In terms of economic magnitude, the results

suggest that, compared to large firms, the loan default rate for SMEs decreases by 2.7 percentage points.

[Table 3 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 suffer from higher default risks due to limited access to formalized financial data and greater informational asymmetries. 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 enables a more accurate assessment of borrower creditworthiness, reducing the likelihood of misclassification and improving the overall quality of lending decisions.

We perform a series of robustness checks to show that baseline results are continue to hold. First, we perform a parallel trends test to validate the key identification assumption. Figure 1 illustrates the dynamic responses to the introduction of AI and big data. Panel A and Panel B present the corresponding estimates for undetermined credit ratings and loan default rates, respectively. Each dot in the figure represents the estimated coefficient, along with the associated 95% confidence intervals, derived from the leads and lags regression specified in Equation (1) of the paper. The comparison group is set to time -1, representing the period immediately before 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. 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 external shock is assumed to have occurred in other time frames. Table 4 presents the corresponding results, where we hypothetically set the implementation of AI to one year earlier—specifically, in the first or second quarter of 2018. Our analysis indicates that the coefficient of the core variable is either statistically insignificant or exhibits a very small magnitude for both undetermined credit ratings and loan default rates. These findings reinforce the robustness of our main results and suggest that the observed effects are indeed attributable to the actual timing of the AI implementation.

[Table 4 about here]

Third, we perform a placebo test for the treatment group using the Monte Carlo permutation method. Specifically, we randomly assign individual observations to the treatment group and repeated the regression analysis 500 times, generating 500 sets of regression results (including the estimated coefficients, standard errors, and p-values). We plot the distribution of the 500 estimated coefficients alongside their corresponding p-values to visually illustrate the results of the placebo test. Figure 2 presents the results, with Panel A showing the distribution for undetermined credit ratings and Panel B displaying the distribution for loan default rates. In both panels, the distributions are centered around zero, indicating no systematic bias in the placebo tests. Furthermore, the estimated coefficients from our baseline analysis (-0.024 for undetermined credit ratings and -0.027 for loan default rates) are significantly smaller than the values observed in the placebo distributions, as shown on the horizontal axis. These findings

provide strong evidence supporting the validity of our baseline estimates for undetermined credit ratings and loan default rates.

[Figure 2 about here]

### **3.3 Credit accessibility and interest payment**

We next to investigate the influence of AI and big data adoption on SMEs' access to bank loans and their interest payments. 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 gain improved access to bank credit while benefiting from reduced borrowing costs.

Table 5 presents the corresponding estimation results, which align closely with the findings in Table 2 and Table 3. The coefficient for the interaction term between SME and Post is negative and statistically significant at the 1% level, even after incorporating various dimensions of fixed effects. This robustness underscores the reliability of the results. Specifically, the findings 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 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 5 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, banks can process a wider range of data, including alternative and non-traditional data sources, to



better evaluate the creditworthiness of SMEs. This enhanced assessment reduces the perceived risk associated with lending to SMEs, enabling banks to extend more credit at lower interest rates. For SMEs, this may translate into improved access to financial resources, which can be critical for their growth, innovation, and competitiveness. The reduction in interest rates for SMEs has significant implications for their financial sustainability and long-term viability. Lower borrowing costs reduce the financial strain on SMEs, allowing them to allocate more resources toward productive investments.

### 3.4 Early Machine learning models verse big data analytics

The bank experienced two significant phases in its adoption of AI and big data technologies. 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.

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

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 undetermined credit rating and loan default rate. Undetermined credit rating is an indicator that equals one if a loan application does not have a credit rating (marked as undetermined in the data) and zero

otherwise.  $SME_i$  is an indicator that equals one if a firm is a SME and zero otherwise.  $Post1_t$  is a time indicator that equals one if the time is after the second quarter of 2019 and zero otherwise.  $Post2_t$  is a time indicator that equals one if the time is after the third quarter of 2020 and zero otherwise.  $\varphi_f$ ,  $\gamma_j$ ,  $\theta_t$ , and  $\delta_r$  represent the fixed effects on firm, industry, time and region, respectively.  $\varepsilon_{i,t}$  is the error term.

Table 6 presents the corresponding estimation results, focusing on the key coefficients of interest for the two interaction terms. Column (1) provides the estimates for undetermined 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.

[Table 6 about here]

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 undetermined 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.

#### **4. Heterogeneous analysis**

We further conduct four sets of heterogeneity analyses to illustrate the information advantage provided by AI and big data. Firstly, we investigate the heterogeneity in loan types based on the presence of collateral. We hypothesize that loans without collateral rely more heavily on soft information compared to loans with collateral. To test this, we construct a binary indicator that equals one if a loan is secured (i.e., backed by collateral) and zero otherwise. This dummy variable is then interacted with our core term—the interaction between SME and post—allowing us to perform a triple-difference (DDD) analysis. Column (1) of Table 7 reports the results for undetermined credit ratings, showing that the coefficient for the DDD estimator is negative and statistically significant. This indicates that the impact on credit ratings is more pronounced for unsecured loans, which lack collateral. These findings suggest that the bank increasingly relies on AI and big data to mitigate information asymmetries and address moral hazard issues, rather than depending solely on collateral, as discussed by Aghion and Bolton (1992).

[Table 7 about here]

In addition, we employ two region-level proxies to evaluate information availability: the level of economic development and linguistic diversity. We hypothesize that regions with higher levels of economic development are characterized by more robust financial markets and reduced information asymmetries. To test this, we construct a binary indicator that equals one if a firm is located in a less developed city, and zero otherwise. For linguistic diversity, we use the number of dialects spoken in a city as a proxy. Prior research in urban and development literature (e.g., Falck et al., 2012; Desmet et al., 2017) has shown that cities with a greater number of dialects tend to

exhibit more complex urban structures and interpersonal relationships. This complexity makes it more challenging for lenders to gather reliable information compared to cities with a simpler and more uniform linguistic environment. Based on this, we construct another binary indicator that equals one if a firm is located in a city with more than two dialects, and zero otherwise.

The results, presented in Columns (2) and (3) of Table 7, reveal that the DDD estimator is negative and statistically significant for both proxies. These findings align with our hypothesis, suggesting that AI and big data provide a significant informational advantage in regions where traditional information collection is hindered by lower economic development or greater linguistic diversity. This highlights the potential of AI-driven technologies to mitigate information asymmetries and improve decision-making in complex environments.

Finally, if a borrower is a state-owned enterprise (SOE), it is more likely to have publicly available information, as the Chinese government requires SOEs to disclose key corporate information to the public. 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 (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

Table 8 presents the results for loan default rates, which are consistent with the findings in Table 7. One distinction, however, is that SOEs exhibit a slightly higher default rate compared to non-SOEs. This difference may be attributed to the intrinsic characteristics of firm ownership structures. For instance, SOEs may face less pressure to maximize efficiency or profitability due to government backing, which could influence their financial performance and risk-taking behavior.

[Table 8 about here]

Overall, these findings further support our hypothesis regarding the information advantage of AI and big data. In environments with limited available information, these technologies significantly improve the accuracy and reliability of credit assessments. Our analyses highlight the transformative potential of AI and big data in overcoming information barriers, thereby enhancing decision-making processes in financial institutions.

## **5. Conclusion**

In conclusion, this study provides compelling evidence of the transformative impact of AI and big data on the banking industry, particularly in enhancing credit assessment processes. By analyzing a comprehensive dataset from a major commercial bank in China, we demonstrate that integrating these technologies significantly reduces the incidence of "undetermined" credit ratings, especially among small and medium-sized enterprises (SMEs). This reduction is achieved through improved accuracy and efficiency in credit assessments, facilitated by the advanced data processing capabilities of AI and big data.

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

Furthermore, our heterogeneity analyses underscore the broader applicability of these technologies, particularly in regions with lower levels of economic development and in areas characterized by greater linguistic diversity, and among firms with limited publicly available information. These findings suggest that AI and big data can play a

crucial role in democratizing access to credit and fostering financial inclusion, thereby supporting economic growth and development in underserved areas.

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. Future research could explore the long-term impacts of these technologies on financial stability and customer satisfaction, further enriching our understanding of their role in modern finance.

## 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
- 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
- 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[J]. *American Economic Review* 107(9): 2479-2513.
- Easley, D., M. Lopez de Prado, M. O'Hara, and Z. Zhang. 2021. Microstructure in the machine age. *Review of Financial Studies* 34: 3316–63.
- Erel, I., L. Stern, C. Tan, and M. S. Weisbach. 2021. Selecting directors using machine learning. *Review of Financial Studies* 34: 3226–64.
- 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
- Falck O., S. Heblich, A. Lameli, and J. Sudekum. 2012. Dialects, cultural identity, and economic exchange. *Journal of Urban Economics* 72(2-3): 225-239
- Frost, J., L. Gambacorta, Y. Huang, H. S. Shin, and P. Zbinden. 2020. BigTech and the changing structure of financial intermediation. *Economic Policy* 34(100): 761-799.
- Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther. 2022. Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance* 77(1): 1-808

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

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

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.



**Table 1 – Summary Statistics****Panel A: Loan and firm distribution**

Year	2015	2016	2017	2018	2019	2020	2021	2022	2023	Total
<i>Firms</i>	95291	79448	75266	80416	74611	73353	120429	166237	254611	475301
<i>Loans</i>	417163	333368	321521	352866	305670	281315	523831	776723	1217431	4529888

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

**Panel B: Comparison between large firms and SMEs**

	Before		After	
	Large	SMEs	Large	SMEs
<i>Firms</i>	7395	162991	4366	369701
<i>Loans</i>	176504	1398131	53139	2902114
<i>Undetermined credit rating loans</i>	10978	94243	6887	51982
<i>Rate of undetermined credit rating</i>	0.72%	5.95%	0.31%	2.08%
<i>Rate of loan default</i>	6.31%	9.12%	5.67%	2.14%
<i>Interest rate</i>	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 – Credit rating**

Variables	<i>Dependent Variable: Undetermined Credit Rating</i>		
	(1)	(2)	(3)
<i>SME</i> × <i>Post</i>	-0.117*** (-3.86)	-0.025*** (-6.75)	-0.024*** (-5.62)
<i>Post</i>	0.067** (2.22)	0.015*** (3.72)	
<i>SME</i>	0.005 (0.18)		
Constant	0.062** (2.13)	0.036*** (49.63)	0.045*** (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 <sup>2</sup>	0.018	0.703	0.706

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

**Table 3 – Loan default rate**

Variables	<i>Dependent Variable: Loan Default Rate</i>		
	(1)	(2)	(3)
<i>SME</i> × <i>Post</i>	-0.062*** (-5.18)	-0.027** (-2.01)	-0.027** (-2.12)
<i>Post</i>	-0.015 (-1.30)	0.023* (1.76)	
<i>SME</i>	0.029* (1.76)		
Constant	0.065*** (3.99)	0.044*** (62.72)	0.059*** (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 <sup>2</sup>	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 – Placebo test (non-exist time)**

Variables	<i>Undetermined Credit Rating</i>		<i>Loan Default Rate</i>	
	(1) 2018Q1	(2) 2018Q2	(3) 2018Q1	(4) 2018Q2
<i>SME</i> × <i>Post</i>	-0.001 (-0.70)	-0.001 (-0.48)	-0.001 (-0.11)	-0.000 (-0.03)
Constant	0.069*** (116.19)	0.069*** (122.53)	0.062*** (23.75)	0.059*** (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 <sup>2</sup>	0.932	0.918	0.853	0.854

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

**Table 5 – Bank credit accessibility and interest payment**

Variables	<i>Loan Amount</i>		<i>Interest Payment</i>	
	(1)	(2)	(3)	(4)
<i>SME</i> × <i>Post</i>	0.049*** (2.81)	0.048*** (2.78)	-0.348*** (-6.16)	-0.336*** (-7.09)
<i>Post</i>	-0.053*** (-3.02)		0.388*** (9.78)	
<i>SME</i>				
Constant	14.851*** (4,577.78)	14.818*** (1,369.86)	4.389*** (345.99)	4.633*** (152.69)
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	1,591,811	1,591,811	4,378,065	4,378,065
R <sup>2</sup>	0.780	0.781	0.793	0.807

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

**Table 6 – Early machine learning models V.S. big data analytics**

Variables	<i>Undetermined Credit Rating</i>	<i>Loan Default Rate</i>
	(1)	(2)
<i>SME</i> × <i>Post1</i>	-0.016*** (-5.21)	-0.015 (-1.10)
<i>SME</i> × <i>Post2</i>	-0.020*** (-6.35)	-0.028** (-2.03)
<i>Constant</i>	0.051*** (29.97)	0.067*** (11.72)
Firm F.E.	YES	YES
Industry F.E.	YES	YES
Region F.E.	YES	YES
Quarter F.E.	YES	YES
Observations	4,378,847	4,358,006
R <sup>2</sup>	0.706	0.708

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

**Table 7 – Heterogeneous analysis on undetermined credit rating**

Variables	<i>Unsecured</i>	<i>Less developed</i>	<i>Dialects</i>	<i>Non-SOE</i>
	<i>loans</i>	<i>districts</i>	<i>districts</i>	
	(1)	(2)	(3)	(4)
<i>Dummy</i> × <i>SME</i> × <i>Post</i>	-0.041*** (-8.44)	-0.012** (-2.25)	-0.034*** (-6.44)	-0.021** (-2.48)
<i>SME</i> × <i>Post</i>	-0.051*** (-10.08)	-0.020*** (-4.21)	-0.012*** (-2.64)	0.028*** (8.75)
<i>Dummy</i> × <i>Post</i>	-0.006 (-1.29)	0.009* (1.69)	0.014*** (2.93)	0.006 (0.95)
<i>Dummy</i> × <i>SME</i>	0.042*** (20.18)	0.020** (2.22)	0.024*** (3.65)	
<i>Dummy</i>	0.005*** (3.11)	-0.018** (-2.14)	-0.008 (-1.26)	
<i>Constant</i>	0.071*** (21.51)	0.043*** (15.53)	0.054*** (5.54)	0.047*** (22.99)
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	4,378,847	4,325,748	4,109,026	4,378,847
R <sup>2</sup>	0.708	0.706	0.712	0.706

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

**Table 8 – Heterogeneous analysis on loan default rate**

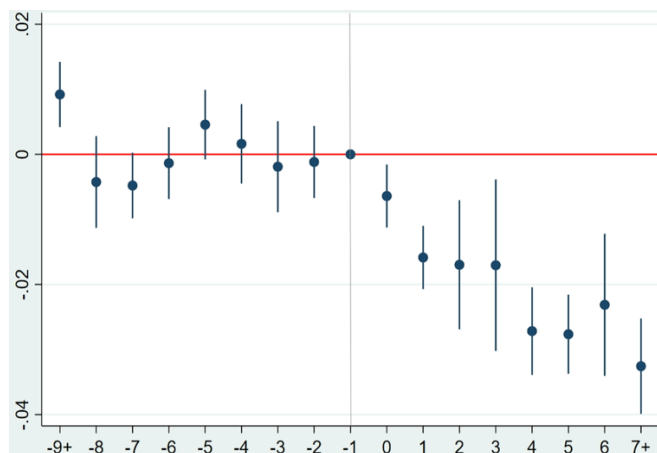
Variables	<i>Unsecured</i>	<i>Less developed</i>	<i>Dialects</i>	<i>Non-SOE</i>
	<i>loans</i>	<i>districts</i>	<i>districts</i>	
	(1)	(2)	(3)	(3)
<i>Dummy</i> × <i>SME</i> × <i>Post</i>	-0.033** (1.97)	-0.035* (-1.90)	-0.026 (-1.39)	0.077*** (-6.23)
<i>SME</i> × <i>Post</i>	0.012 (0.86)	-0.020* (-1.60)	-0.019 (-1.26)	0.069*** (6.07)
<i>Dummy</i> × <i>Post</i>	-0.011 (-0.68)	0.049*** (2.67)	0.030 (1.64)	0.081*** (6.87)
<i>Dummy</i> × <i>SME</i>	-0.017*** (-2.88)	0.001 (0.11)	0.015 (1.42)	
<i>Dummy</i>	-0.008 (-1.40)	-0.009 (-1.08)	-0.022** (-2.14)	
<i>Constant</i>	0.044*** (4.88)	0.053*** (6.82)	0.054*** (5.54)	0.087*** (11.78)
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	4,358,006	4,305,111	4,089,293	4,358,006
R <sup>2</sup>	0.709	0.709	0.712	0.708

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



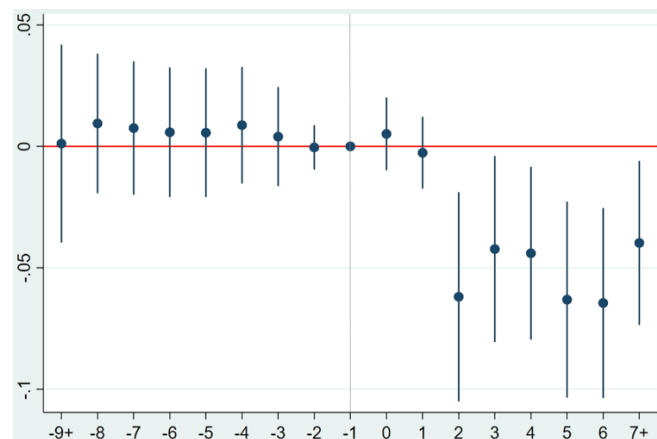
**Figure 1 – Parallel trends test**

**Panel A: Undetermined credit rating**



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

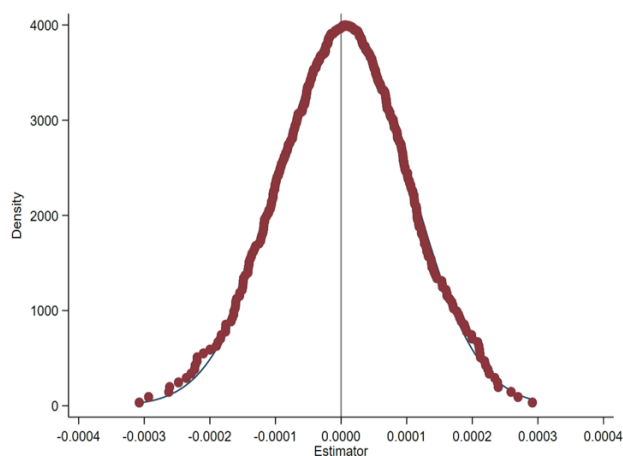
**Panel B: Loan default rate**



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 second quarter of 2019 ( $t = -1$ ).

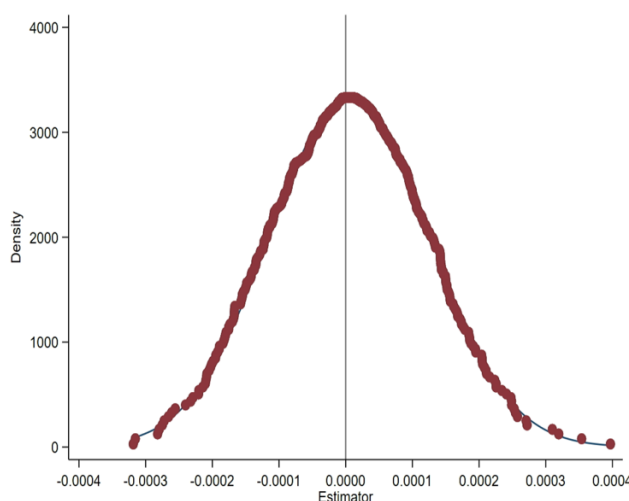
**Figure 2 – Placebo test**

**Panel A: Undetermined credit rating**



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

**Panel B: Loan default rate**



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

## Appendix

**Table A1 Region distribution**

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

**Table A2 Industry distribution**

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

**Table A3 – Baseline regression with including city level control variables**

Variables	<i>Undetermined credit rating</i>	<i>Loan default rate</i>
	(1)	(2)
<i>SME</i> × <i>Post</i>	-0.001 (-0.70)	-0.001 (-0.48)
<i>GDP</i>	0.002 (0.49)	-0.003 (-0.53)
<i>Fiscal Revenue</i>	0.002 (0.49)	-0.003 (-0.53)
Constant	0.069*** (116.19)	0.069*** (122.53)
Firm F.E.	YES	YES
Industry F.E.	YES	YES
Region F.E.	YES	YES
Quarter F.E.	YES	YES
Observations	4,009,378	3,992,325
R <sup>2</sup>	0.689	0.711

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