China Economic Review 81 (2023) 102016

Contents lists available at ScienceDirect

China Economic Review

journal homepage: www.elsevier.com/locate/chieco

BigTech credit risk assessment for SMEs

Yiping Huang [a], Zhenhua Li [b], Han Qiu [c,*], Sun Tao [d], Xue Wang [a], Longmei Zhang [d]

[a] Institute of Digital Finance and National School of Development, Peking University, China
[b] Ant Group, China
[c] Bank for International Settlements, Switzerland
[d] International Monetary Fund, United States

ARTICLE INFO

Keywords:
Bigtech
Risk assessment
Asymmetric information

ABSTRACT

Lending by big technology companies (BigTechs) is an important new financial innovation in the digital era. This paper attempts to evaluate robustness and special features of BigTech's credit risk assessment. Using 1.8 million loan transactions for online merchants of a leading Chinese virtue bank, we carry out a horse race analysis between the BigTech approach (i.e., big data and machine learning models) and the bank approach (i.e., traditional financial data and scorecard models) in predicting loan defaults. We show that the BigTech approach better predicts loan defaults, reflecting information and modeling advantages. Though bank approach do well for the firms which have records in credit registry, BigTech's proprietary information can complement or, where necessary, substitute for credit history in predicting defaults, especially for the unbanked borrowers. We further discuss inclusiveness feature of the BigTech approach and the implications for financial inclusion, financial intermediaries' businesses and regulators' policy.

1. Introduction

Promoting financial inclusion for vulnerable households and smaller firms has been a perennial challenge for policy makers globally (Abdulsaleh & Worthington, 2013; Freel, Carter, Tagg, & Mason, 2012; Vos, Yeh, Carter, & Tagg, 2007). An essential element in financial inclusion, access to credit for small- and medium-size enterprises (SMEs) remains quite limited, especially in developing countries (Demirguc-Kunt, Klapper, Singer, Ansar, & Hess, 2018). The main barriers include high cost, physical distance, and lack of proper documentation (Agarwal & Hauswald, 2010; Demirgüç-Kunt & Klapper, 2013), which deter banks from managing the risks in servicing SMEs. Since SMEs are numerous and scattered across locations, commercial banks face a high fixed cost in establishing business connections. Moreover, many SMEs lack high-quality financial data and collateral assets for banks to identify and manage credit risk. As a result, banks typically resort to personal guarantees or relationship lending, relying on local connections and soft information to reduce information asymmetry (Berger & Udell, 2002; Berger & Udell, 2006). Nonetheless, relationship building is also costly, and it is difficult to reach a large business scale, as evidenced by the encouraging but still limited coverage of lending services by

* Corresponding author.
E-mail address: han.qiu@bis.org (H. Qiu).

https://doi.org/10.1016/j.chieco.2023.102016
Received 30 November 2022; Received in revised form 21 June 2023; Accepted 26 June 2023
Available online 3 July 2023
1043-951X/© 2023 Elsevier Inc. All rights reserved.

Muhammad Yunus's Grameen Bank.[1]

In recent years, the surge of BigTech has opened new possibilities for expanding access to credit for SMEs (Liu, Lu, & Xiong, 2022; Hau, Huang, Shan, & Sheng, 2021). Some big technology (BigTech) companies, such as Alibaba and Tencent in China, Mercado Credito in Argentina, Paytm in India, and Amazon in the United States, have extended loans to millions of small borrowers (Agarwal, Alok, Ghosh, & Gupta, 2020; Cornelli et al., 2022; Frost, Gambacorta, Huang, Shin, & Zbinden, 2019). China has been at the forefront of fintech development and is the largest fintech market in the world, with virtual banking being one of its wide-reaching features. Enabled by digital technology and big data, China's big four tech players—Alibaba, Baidu, Tencent, and JD—have made incursions into financial services. During 2014–16, China's banking regulator issued 11 new privately-owned banking licenses. In China, three leading virtual banks—MYbank (affiliate to Alibaba), WeBank (affiliate to Tencent), and XW Bank (affiliate to tech giant Xiaomi)—provides loans to millions of small firms annually, >80% of which have no credit history. Compared with traditional banks, the loans provided by BigTech lenders are much smaller, shorter in duration, and mainly used for operational rather than long-term investment purposes. Hence, BigTech lending has so far played a complementary role to traditional banking by reaching underserved customers.

The wide reach of BigTech lending to SMEs reflects a confluence of factors. First, digital technologies enable BigTechs to use e-commerce, social networks, and mobile payment services to connect to millions of customers at very low marginal costs. This could help overcome barriers to large-scale customer acquisition. In addition, BigTechs can monitor borrowers' activities, most of which occur on the BigTech platforms.

Second, big data, including traditional and proprietary information, could help in credit risk assessment, in the absence of financial history and collateral assets (Boot, Hoffmann, Laeven, & Ratnovski, 2021; Gambacorta, Huang, Li, Qiu, & Chen, 2023). Traditional data include basic information about individuals or firms, such as gender, age, location, profession, and business. Such information is mostly available offline, but it is costly to collect without the support of BigTech. Proprietary information refers to broadly defined digital footprints, such as messages exchanged, payments made, and websites browsed on the BigTech platforms. Such information could be deployed to assess borrowers' financial conditions and behavioral characteristics (Berg, Burg, Gombović, & Puri, 2020; Gambacorta, Huang, Qiu, & Wang, 2019; Jagtiani & Lemieux, 2019). To some extent, these data are similar to soft information in relationship lending (Cornée, 2019).

Third, digital technologies, such as cloud computing and artificial intelligence, allow BigTechs to process massive numbers of loan applications quickly and update risk assessment dynamically, based on real-time data. Furthermore, the big data and machine learning approach enables BigTech lenders to restructure loans quickly at large volumes, which significantly reduces operating costs.[2] The "contact-free feature" of the business process, from customer acquisition to loan underwriting and restructuring, makes the BigTech lending model particularly robust during the COVID-19 pandemic.

Despite the achievements of BigTech lending so far, a central question that remains is whether its credit risk assessment approach, that is, big data plus machine learning method, is more reliable in predicting loan defaults than the bank approach, which features financial data and scorecard models. In addition, given the short history of BigTech lending, it is yet to be tested whether such an approach can withstand business cycles and various economic shocks. Moreover, it is critical to understand the contribution of data versus methodology in credit risk assessment. Can traditional models also fully utilize the information value of big data? Can machine learning algorithms extract more information from traditional data? In particular, for borrowers with no bank credit history and incomplete financial data, is BigTech's proprietary information sufficient to substitute credit registry information in risk assessment?

To shed light on these questions, we employ a unique data set of 1.8 million MYbank SME loans to replicate and compare the BigTech approach with the bank approach in risk assessment. A leading virtual bank in China, MYbank was established in May 2015, with its BigTech lending business inherited from its parent company, Ant Group ("Ant" hereafter), which, in turn, was an affiliated company of the e-commerce giant Alibaba Group ("Alibaba" hereafter). To our knowledge, Alibaba and Ant were the first groups of companies globally to create the BigTech lending business model in 2010. Borrowers in this data set are all online vendors on Alibaba's e-commerce platforms and Ant's Alipay users, mostly small or self-employed merchants. The data set features a large volume of proprietary information, such as business transactions, payments, customer ratings, consumption patterns, and importance in the ecosystem, along with other traditional data. The data set also contains information on borrowers' credit histories with traditional banks, if they have ever borrowed in the past.

Our study shows that the BigTech approach significantly improves the accuracy of loan default prediction, compared with the bank approach. This reflects a combination of information and modeling advantages. BigTech's proprietary information is utilized by machine learning algorithms to reveal the financial condition (i.e., capacity to repay) and behavioral characteristics (i.e., willingness to repay) of the borrower. Applying the traditional scorecard model to big data could also provide reasonably good risk assessment, yet some borrower characteristics would not be fully captured. Similarly, the machine learning approach improves traditional information's predictive power, reflecting its capacity to model more complex interactions among the variables. To be clarified, the paper offer an evidence to shwo the machine learning could improve the predictive power but not necessarily means all machine

[1] Grameen Bank is a microfinance organization and community development bank founded in Bangladesh. It makes small loans to impoverished individuals without requiring collateral. Grameen Bank originated in 1976, through the work of Professor Muhammad Yunus, who launched a research project to study how to design a credit delivery system to provide banking services to the rural poor. As of November 2019, it had 9.6 million members, 97% of whom were women. With 2568 branches, Grameen Bank provides services in 81,678 villages, covering >93% of the total villages in Bangladesh. (http://www.grameen.com/introduction/Grameen)

[2] Loan forbearance and restructuring are resource-intensive and time consuming for traditional banks. Reflecting the small size of SME loans, the unit costs of providing financing services to SMEs are typically high, which deters traditional banks from reaching out proactively to SMEs.

learning methods are better than traditional models.

We then compare the BigTech approach and the bank approach for a subsample of borrowers with bank credit history to examine whether BigTech proprietary information complements or substitutes for the credit history information. We find that bank credit history information is quite effective for predicting defaults. More interestingly, BigTech information alone generates similar results as bank credit history information, if we exclude credit history information in machine learning and scorecard models. This result suggests that, for borrowers without a bank credit history, BigTech proprietary information can effectively substitute credit registry information in risk assessment, hence enhancing financial inclusion.

Finally, we explore the differences in predicted default probabilities between the BigTech approach and the bank approach for SMEs of different sizes and located in different cities. We find that firms that are smaller and in more remote cities benefit more from the BigTech risk assessment model in forms of lower predicted default probabilities. Hence, the BigTech lending model naturally complements traditional banks in broadening firms' access to credit.

The remainder of the paper is structured as follows. Next section reviews the literature and outlines the key analytical steps of the study. The third section briefly introduces the institutional setup, and describes the data set. The fourth section explains the analytical strategy, and discusses the empirical results. The last section concludes.

2. Literature review

The most fundamental challenge for financial transactions, including lending, is information asymmetry, which may cause ex-ante adverse selection and ex-post moral hazard problems. In a way, financial institutions (e.g., banks), market mechanisms (e.g., rating agencies), and regulatory policies (e.g., information disclosure requirements) are all devised to deal with the information asymmetry problem. Commercial banks normally adopt three methods to analyze and mitigate credit risk: financial history, collateral assets, and soft information. Financial history relies on detailed analyses of borrowers' financial data, especially balance sheets, income statements, and cash flow information, to predict probabilities of default. This method is often most useful for large corporations, which can typically provide comprehensive information. Collateral or guarantees are more frequently used in lending to SMEs. Several studies find that pledging collateral may help resolve adverse selection phenomena (Besanko & Thakor, 1987a; Besanko & Thakor, 1987b; Cerqueiro, Ongena, & Roszbach, 2016; Stiglitz & Weiss, 1981) and moral hazard actions (Berger, Frame, & Ioannidou, 2016). However, the fact that many SMEs do not have collateral assets constrains lending to SMEs, especially micro firms and self-employed businesses. Therefore, an alternative model of relationship banking uses soft information about the firm (such as information on its owner and the local community) to predict the probability of default (Agarwal & Hauswald, 2010; Berger & Udell, 2002; Berger & Udell, 2006). This approach can reduce banks' dependence on financial data and collateral assets in credit risk management, but it requires significant investment in physical capital and human resources. Banks need to have broad networks to be close to their potential customers and maintain regular interactions. Therefore, the costs of relationship lending may be high, making it difficult for SMEs to cover these high costs to access financing.

Recent developments in BigTech lending reflect advances in credit risk assessment by employing big data and machine learning models. The BigTech approach for credit risk assessment is still a new but growing field of research. A few papers analyze the role of digital footprints in enhancing credit risk assessment. Using a customer data set from a German online store, Berg et al. (2020) demonstrate that the easily accessible variables from the digital footprints customers leave on the e-commerce platform complement credit bureau information, affect access to credit offered by the online shop, and reduce default rates. The authors show that even simple information on mobile phone operating systems, iOS or Android, reveals a lot about customers' creditworthiness. Berg et al. (2020) speculate that by carefully analyzing digital footprints, financial services might be able to cover the world's 200 million unbanked individuals.

Using unique and proprietary loan-level data from a large BigTech lending firm in India, Agarwal et al. (2020) find that mobile and social footprints have significantly more predictive power for loan approvals and defaults, outperforming the traditional credit scores used by banks. The authors also show that this new approach can expand credit as well as reduce the overall default rate. Similarly, based on a case study in Argentina, Frost et al. (2019) provide evidence that BigTech lenders have an information advantage in credit risk assessment relative to a traditional credit bureau. By analyzing a transaction data set from Lending Club, Jagtiani and Lemieux (2019) prove that the use of alternative information sources allows some borrowers classified as subprime by traditional criteria to be slotted into "better" loan grades and therefore obtain lower-priced credit.

In addition to its use of big data, another salient feature of BigTech risk assessment is the machine learning approach, in contrast to traditional models. Financial institutions have established many traditional methods for decision making, and credit scoring is the most widely used technique (Ahmed & Rajaleximi, 2019; Kithinji, 2010). The retail banking business, such as credit cards, mortgages, and personal loans, makes heavy use of predictive statistical models called scorecards (Thomas, Oliver, & Hand, 2005), which are built using data from past customers and credit histories. The same method can be applied to corporate risk models for SMEs (Miller & Rojas, 2004). The purpose is for banks to create a scorecard with as many borrower characteristics as possible, to enable better decision making. However, the scorecard models have several obvious weaknesses (Hand & Crowder, 2005). Many researchers argue that they require a training period of at least one year. Further, updating the scorecard is time-consuming and resource-intensive (Hopper & Lewis, 2004). In contrast, BigTech lenders often automate their credit assessment process by continuously training machine learning models with updated big data.

The machine learning models adopted by BigTech lenders capture interactions among various explanatory variables. The models turn various digital footprints into data, ranging from social network activities to the physical locations of applicants' activities (Bazarbash, 2019). Khandani, Kim, and Lo (2010) find that machine learning forecasts are considerably more adaptive and capture the

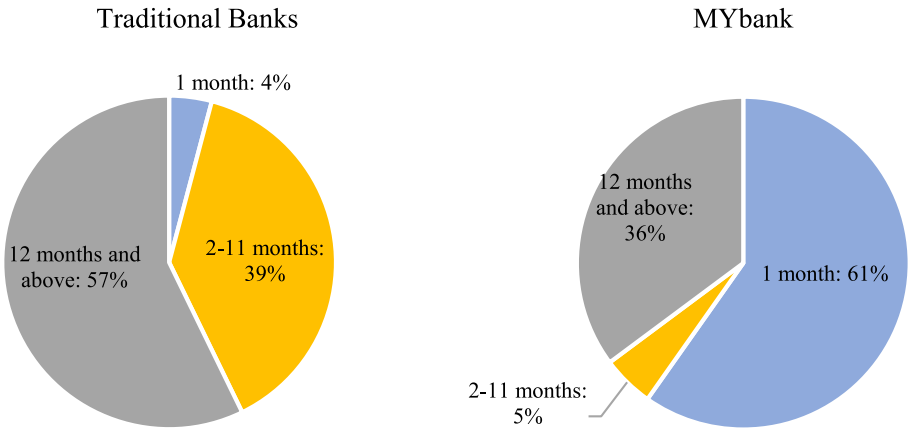


Fig. 1. Distribution of loan duration: MYbank versus traditional banks.
Source: Calculations using data from MYbank.

dynamics of credit cycles as well as absolute levels of default rates. Butaru et al. (2016) provide evidence that decision trees and random forests, the two most popular machine learning methods, outperform logistic regression in out-of-sample and out-of-time forecasts of credit card delinquencies. Analyzing data from a Chinese BigTech firm, Gambacorta et al. (2019) show that a model based on machine learning and nontraditional data is better at predicting losses and defaults, compared with traditional models, by capturing nonlinearities and including additional information. They also find that this conclusion is robust to an exogenous shock to the supply of credit. Fuster, Goldsmith-Pinkham, Ramadorai, and Walther (2020) find that new technologies such as machine learning could help increase predictive accuracy from the improved use of information.

Our study contributes to the literature in the following respects. First, to our knowledge, this is the first study to carry out a horse race analysis between the BigTech approach and the bank approach for credit risk assessment. By controlling the data and models, we quantify the information and model advantages. More importantly, we find that machine learning can better uncover information from non-traditional variables. Second, we find that for businesses with credit history, traditional methods have achieved relatively good results. In other words, big tech risk control methods are more valuable for "credit invisible" individuals without credit records. Third, Specifically, we find that in low tier cities, smaller businesses can benefit more from big tech methods.

3. Institutional setup and data set

3.1. Institutional setup

MYbank, one of the leading virtual banks in China, was founded in 2015 by Alibaba's affiliate firm, Ant Group, through a 30% stake in a joint venture comprising a group of private firms. MYbank uses big data, machine learning, and the associated flexible risk management approach to offer credit to SMEs and manage risks. By harnessing its credit-profiling techniques driven by big data analytics (e.g., e-commerce and cash flow), MYbank mainly manages to offer online loans to micro firms that have previously no access to credit. For example, MYbank has about 20 million SME borrowers, about 80% of whom have never borrowed from banks in the past. The loans are small, short in duration, and used primarily for operational purpose. The loan quota is linked to a firm's cash flows more than its assets, and the historical average nonperforming loan (NPL) ratio has been about 2%.[3]

MYbank mainly serves customers with no access to traditional bank lending and the loans are smaller and shorter in duration compared with traditional banks' loans. In our sample, MYbank's average loan size is RMB 2600 ($380), while the average loan size for SMEs from banks is RMB 1 million ($150,000), >30 times larger than that of MYbank. The duration of MYbank loans is much shorter. For instance, the duration of over 60% of MYbank's loans is less than one month,[4] compared with nearly 57% of traditional bank loans for SMEs being longer than one year (inferred from MYbank borrowers who have also borrowed from traditional banks) (Fig. 1). Reflecting the short duration, these loans are often used as working capital for operational purposes rather than longer-term investment.

The financial inclusion feature of MYbank lending is reflected mainly in credit access, rather than price. MYbank's annualized lending rate is between 10 and 17%, similar to the prevailing private lending rates in China, such as the Wenzhou composite lending rate, but it is higher than the average bank lending rate of 4.35%.[5] The difference reflects several factors. First, SME lending by traditional banks often enjoys preferential regulatory policies, such as cheap central bank funding and explicit government subsidies.

[3] Retrieved from http://stock.xinhua08.com/a/20201025/1960416.shtml?f=arelated; Accessed on January 4, 2020.
[4] This is calculated using data from MYbank's business, not limited to our sample.
[5] Lending rates for MYbank are from the official website: https://render.mybank.cn/p/f/fd-j9fi9ern/index.html. The composite annual lending rate in the Wenzhou informal market was 15.66% in June 2020 (see http://www.wzpfi.gov.cn/).

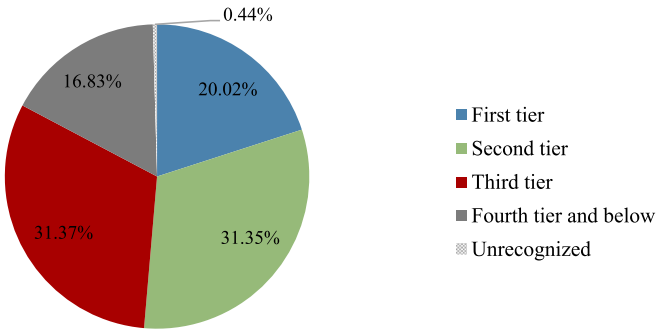


Fig. 2. Geographic distribution of borrowers.
Note: The percentages are the shares of borrowers in each city tier.
Source: Calculations using data from MYbank.

Table 1
Descriptive statistics of the key variables.

Variables	No. obs	Mean	St. dev.	Pct 25	Median	Pct 75	Type of information
House property (0/1)	1,822,423	0.62	0.49	0	1	1	Traditional
Car property (0/1)	1,822,423	0.66	0.47	0	1	1	Traditional
Number of credit cards	1,825,342	2.97	3.84	0	2	4	Traditional
Owner's age (years)	1,825,342	29	6	25	28	31	Traditional
Gender (male = 1; female = 0)	1,825,342	0.65	0.48	0	1	1	Traditional
Firm's age (years)	1,825,342	4.48	2.24	2.70	4.07	5.99	Traditional
City tiers	1,817,380	2.47	1.03	2	2	3	Traditional
Total inflow of funds in Alipay (yuan)	1,825,342	516,708.00	1,149,718.00	138,534.10	264,210.00	534,783.60	Proprietary
Shop rating	1,825,342	4.19	2.96	2	4	6	Proprietary
Transaction volume (last six months, yuan)	1,825,342	9196.80	58,973.99	1964.35	4754.43	9954.34	Proprietary
Log-ins (number, last six months)	1,825,342	68.06	42.54	33	59	99	Proprietary
Security funds in Taobao (yuan)	1,825,342	312.78	827.78	0	0	1000	Proprietary
VIP class	1,825,342	2.94	1.28	2	3	4	Proprietary
Number of good feedbacks from clients	1,825,342	950.94	8002.71	23	118	460	Proprietary
Network effect score	1,825,342	64.48	27.66	48.85	59.95	74.09	Proprietary
Link to credit card (0/1)	1,825,342	0.56	0.50	0	1	1	Proprietary
Daily average Yu'eBao balance (last three months, yuan)	1,825,342	974.84	6656.89	0.00	106.00	532.60	Proprietary
Daily average Alipay wallet balance (last three months, yuan)	1,825,342	871.30	2487.46	95.92	362.18	1117.84	Proprietary
Daily payment activity (index, last year)	1,822,407	1511.85	3682.70	647.00	1007.00	1587.00	Proprietary
Payment activity (index, last six months)	1,821,221	204.72	623.59	81.00	140.00	236.00	Proprietary
Total amount of e-commerce purchases (last six months, yuan)	1,821,221	33,514.44	128,465.10	9132.47	18,080.32	34,788.37	Proprietary
Quarterly consumption of goods (yuan)	1,825,342	838.73	7754.11	0.00	57.70	323.00	Proprietary
Number of fulfilled contracts in daily life	1,822,423	1.07	0.63	1	1	1	Proprietary
Stability of contact information (index, last year)	1,822,423	1.87	1.35	1	2	2	Proprietary
Duration in one location (index)	1815,137	1815.63	762.28	1261.00	1714.00	2318.00	Proprietary

Source: Calculations using data from MYbank.
Notes: 1. We include 32 traditional variables in our analysis. However, only part of the list is shown in this table due to commercial confidentiality issues. 2. *Shop rating* is an evaluation of the borrower's business in Taobao, based on a unique evaluation system. A higher rating indicates a better assessment. 3. *Network effect score* is calculated from variables in several tens of dimensions to capture the relative importance of an individual or firm in the entire social and economic networks. A higher score shows a bigger impact 4. *Daily payment activity last year* and *Payment activity over the last six months* are indexes constructed from payment variables of the borrowers to reflect their transaction activities. 5. *Stability of contact information* and *Duration in one location* are constructed from the duration of using specific contact information and a fixed residential address; a larger index indicates a more stable livelihood.

Lending rates are also set low under government guidance, which does not necessarily reflect the risk premium. In contrast, MYbank's funding cost is higher, given its disadvantage in attracting retail deposits and ineligibility for preferential policies. Establishing a big data infrastructure also requires high fixed costs. Lending rates are set to ensure an adequate profit margin. Second, MYbank borrowers are smaller than typical SMEs with bank relationships and are therefore associated with higher risks and typically excluded by banks.

Table 2
Model definitions.

	Scorecard	Random forest
Traditional information	Model I (Bank approach)	Model III
All information (traditional + proprietary data)	Model II	Model IV (BigTech approach)

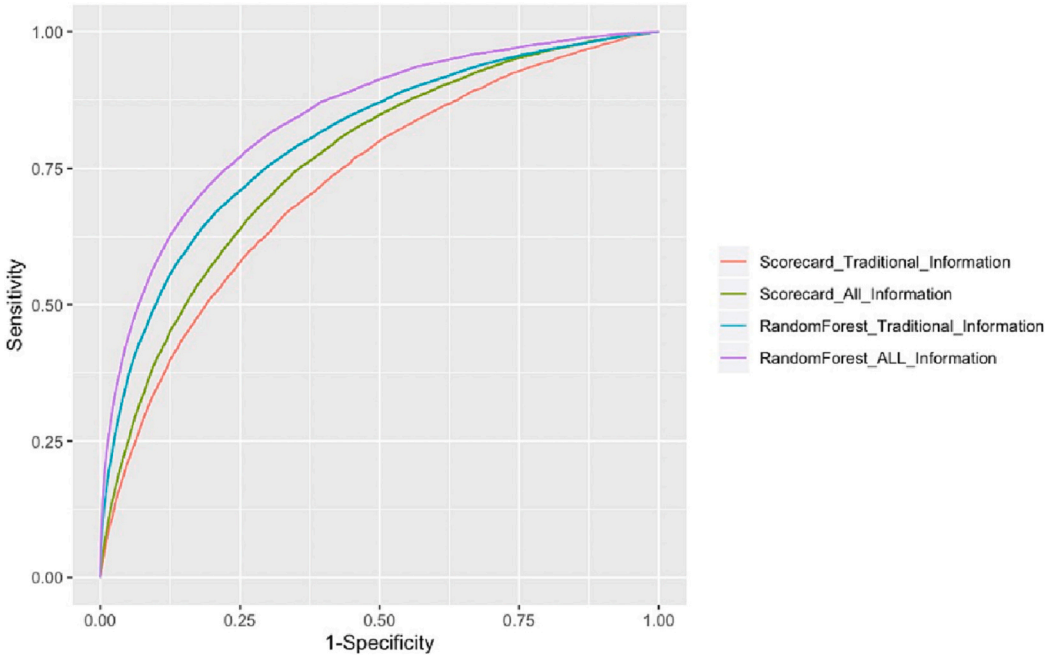


Fig. 3. ROC Curves for Four Models.
Note: The graph compares four model specifications by using ROC curves. ROC = receiver operating characteristics.
Source: Calculations using data from MYbank.

3.2. Data description

This study employs a unique data set of 1.8 million SME loans granted by MYbank between March and August in 2017.[6] All the loans have maturity of one year, and 99.8% of the borrowers are micro firms, that is, firms with annual sales less than RMB 1 million. As shown in Fig. 2, these firms are widely distributed across China, and about half of them are located in less developed cities, that is, Tier-3 and Tier-4 cities. All the borrowers are online vendors on Alibaba's e-commerce platforms and users of Ant's Alipay, which helped generate large volumes of proprietary data on businesses, financing, social networks, and individual behaviors. The data set also includes information on loan repayment through August 2018.

We divide the 76 variables on firm characteristics used in this study into two broad categories: *traditional data* and *proprietary information*, and then further divide them into subcategories. The 32 traditional data variables are classified into four groups: (a) asset-related information, such as housing property; (b) credit history with MYbank; (c) vendor-specific information, such as gender, age, and business; and (d) information on the local (provincial and municipal) economy. The 44 proprietary information variables include transaction volume, network effect score, and other digital footprints. In particular, the network effect score measures a borrower's relative importance in the BigTech network based on their fund flows and social interactions, with 0 representing the lowest impact and 100 the highest impact. Digital footprints also reflect information on borrowers' behavior, such as their online consumption pattern on Alibaba's platform and financial transaction style on Ant's platform. In addition, a firm's real-time customer ratings and reviews are captured. Table 1 provides summary statistics for the key proprietary information variables.

In addition to these variables, we include bank credit history for borrowers who have borrowed from traditional banks in the past. This information allows us to compare the relative importance of traditional data and proprietary information from MYbank and bank credit history for credit risk assessment. Since the majority of the MYbank borrowers are not served by traditional banks, the sample size for this exercise using bank credit history information is much smaller, with 145,109 loan transactions, or about 8% of the entire

[6] Due to business confidentiality, this data set is not available to the public for cross-checking.

Table 3
Comparison of AUCs for the scorecard and random forest models.

Model	(a)(b)	(c)	(a)(b)(c)	(a)(b)(c)(d)	(e)	(a)(b)(c)(d)(e)
Scorecard	0.70	0.58	0.72	0.72	0.69	0.76
Random forest	0.67	0.54	0.76	0.80	0.76	0.84

Source: Calculations using data from MYbank.
Note: 1. The table shows the discriminatory power of different model specifications by providing the AUC. 2. (a) refers to asset and financial information; (b) refers to credit history (only from MYbank); (c) refers to vendor-specific information; (d) refers to local economy information; and (e) refers to BigTech proprietary information. AUC = area under the receiver operating characteristics curve.

Table 4
95% Confidence intervals of the AUCs for four models.

Model	AUC	95% conf. Interval	
Model I (scored card + traditional information)	0.7246	0.7222	0.7269
Model II (scored card + all information)	0.7636	0.7615	0.7658
Model III (random forest + traditional information)	0.8022	0.8001	0.8043
Model IV (random forest + all information)	0.8414	0.8396	0.8433

Source: Calculations using data from MYbank.
Note: AUC = area under the receiver operating characteristics curve.

Table 5
Reduction in NPL ratio from bank approach to BigTech approach.

Threshold (%)	Reduction in NPL ratio (percentage points)
10	1.05
15	0.76
20	0.48
25	0.27
30	0.15

Source: Calculations using data from MYbank.
Note: NPL = nonperforming loans.

sample.

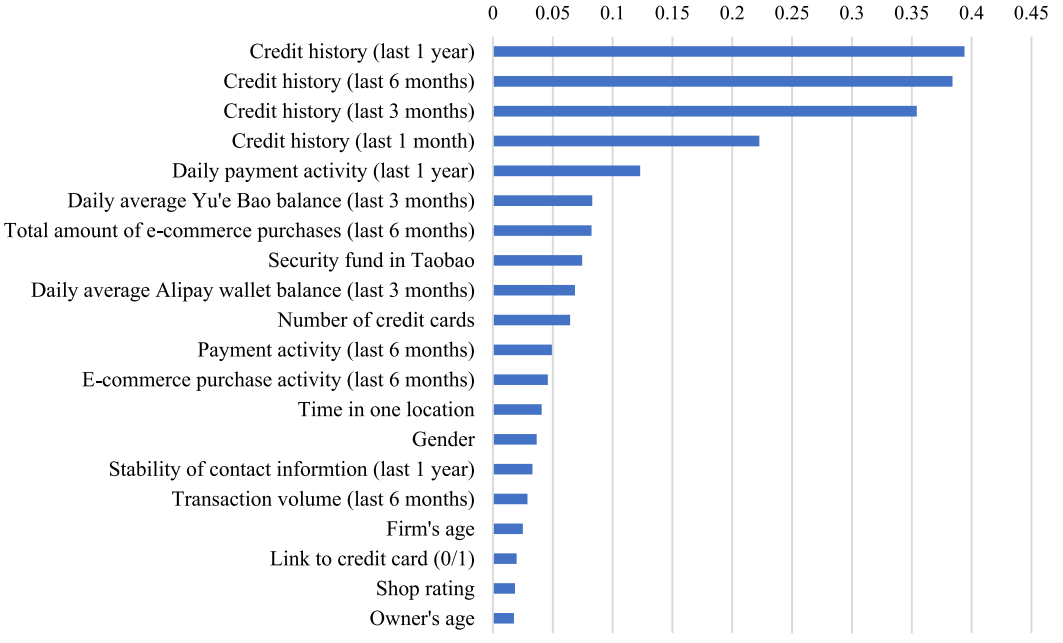
4. Analytical strategy and empirical results

4.1. Analytical strategy

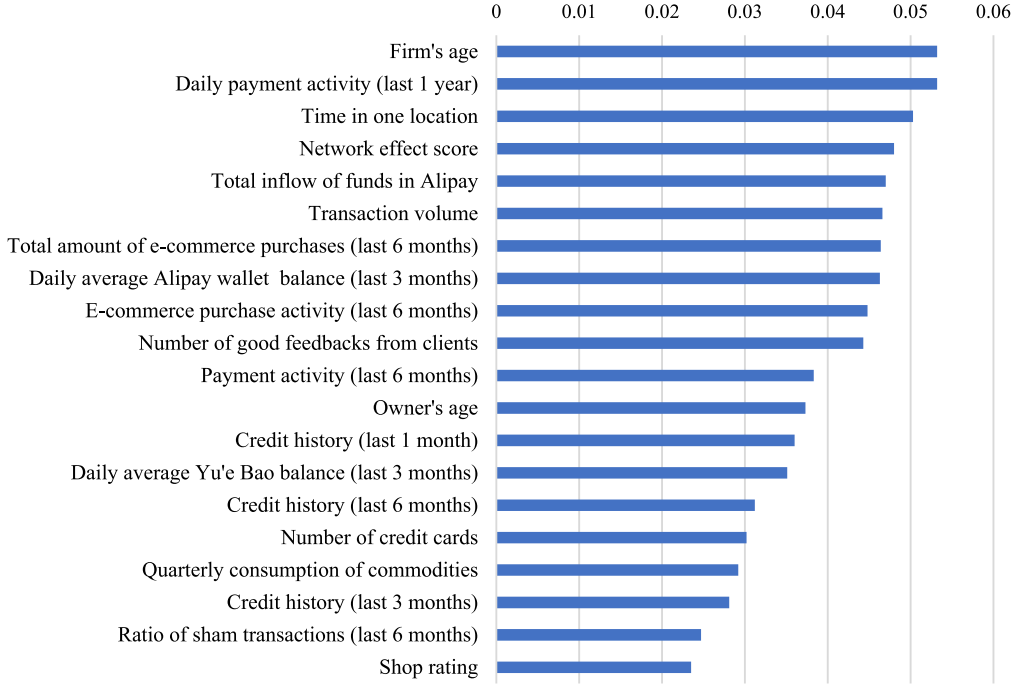
Our empirical analyses include four steps. First, we conduct a horse race between the BigTech approach and the bank approach for credit risk assessment. This should allow us to quantify the information and model advantages by using the same models and data sets. Second, we assess the role of big data in credit assessment, relative to bank credit history information. In particular, we explore whether big data, especially BigTech proprietary information, alone are sufficient for reliable credit risk assessment in the absence of bank credit history information. Third, we test whether the relative outperformance of the BigTech approach can survive an exogenous shock, such as the adoption of a new regulatory policy. Fourth, we explore the inclusive nature of the BigTech approach by comparing its performance with that of the bank approach for subsamples of borrowers of different sizes and in different cities. The data used in our analysis were provided by MYbank on a confidential basis and are not available to the public given protection of the privacy of the customers.

The basic strategy is to conduct four sets of horse races to evaluate the roles of information and modeling methods in assessing credit risk, which is proxied by default. We compare the contributions of traditional data with that of MYbank's proprietary information by controlling for the same models. And we look at the predictive power of traditional scorecard models and machine learning models, by using the same sets of information. The scorecard model is widely used in the financial industry for calculating credit scores. For the machine learning model, we use the random forest model in our baseline, following Butaru et al. (2016). We also use other machine learning models (such as the Gradient Boosting Decision Tree) for robustness checks, and the conclusions remain unchanged. In the following analysis, traditional models refer to standard scorecard models, while machine learning models refer to random forest models. The four models (see Table 2) are model I, or the bank approach (scorecard models + traditional information); model II (scorecard models + all information); model III (machine learning models + traditional information); and moadel IV, or the BigTech approach (machine learning models + all information). The difference in the performance of models I and IV captures the relative advantage of BigTech versus the bank approach, while other pairwise comparisons show the marginal contribution of big data (model IV versus model

a. Ranking of Information Value: Scorecard Model



b. Ranking of Gini Impurity: Random Forest Model



(caption on next page)

Fig. 4. Top 20 information variables in the scorecard and random forest models.
a. Ranking of Information Value: Scorecard Model.
b. Ranking of Gini Impurity: Random Forest Model.
Notes: In panel a, the x-axis shows the information value of each variable in the scorecard model. In panel b, the x-axis shows the Gini impurity of each variable in the random forest model. Both indicators measure the predictive power of the variables, but the values are not directly comparable.
Source: Calculations using data from MYbank.

III and model II versus model I) and the machine learning model (model IV versus model II and model III versus model I).

For the out-of-sample tests, we split the sample into two subperiods (March to May 2017 and July to August 2017).[7] The first subsample is used to estimate the models (training set), and the second subsample is used to test the models (testing set). There are 771,596 loan transactions in the training set and 1,053,748, in the testing set.

4.2. Baseline results

Our empirical analyses include four steps. First, we conduct a horse race between the BigTech approach and the bank approach for credit risk assessment. This should allow us to quantify the information and model advantages by using the same models and data sets.

The baseline results of the four models are reported in Fig. 3 and Table 3. Fig. 3 shows the receiver operating characteristics (ROC) curves. The true positive rate (on the vertical axis) is also known as sensitivity. The false positive rate (on the horizontal axis), known as specificity, is basically the fallout or probability of a false alarm and can be calculated as (1 − specificity). At a given level of specificity (false positive), a superior outcome would have a higher level of sensitivity (true positive). For a given model, it is better if the true positive rate is higher and the false positive rate is lower, which means that the ROC is closer to the upper left-hand corner of the diagram.

The results show that model IV (purple line, random forest + all information) performs the best. Model III (blue line, random forest + traditional information) is second in line, followed by model II (green line, scored card + all information). Model I (red line, scored card + traditional information) is the least efficient among all four. Since we control the size of the sample and the two lists of variables, we can conclude as follows. First, the BigTech approach is more reliable than the bank approach in predicting defaults. Second, if we replace the traditional information in the bank approach with all information (or adding proprietary information to traditional variables), the performance of the model (model II) improves significantly. This may be characterized as an *information advantage*. Third, if we replace the scorecard model with the machine learning model, the performance of the model (model III) improves compared with model I. This may be characterized as a *model advantage*. In summary, comparison of the ROCs of the four models indicates that the BigTech approach exhibits information and model advantages over the bank approach. In this particular case, the model advantage appears to be greater than the information advantage. However, this finding could be sample dependent.

Table 3 reports the area under the ROC curve (AUC), which is widely used to measure the discriminatory power of credit scores (Berg et al., 2020). The AUC ranges from 50% (purely random prediction) to 100% (perfect prediction). The general rule is that the model is reasonably reliable if the AUC is above 60%, and it performs strongly if the AUC is above 70% (Berg et al., 2020). Table 3 reveals some interesting results. First, with relatively few variables, the machine learning models (random forest) show no advantage (illustrated by smaller AUCs for machine learning models than the scorecard models in the first two columns in Table 3). This finding implies that machine learning models are more powerful for large data sets. Second, while relying on BigTech proprietary information ((e) in Table 3) should be sufficient for conducting credit risk assessment (with AUC of 0.76), its advantage over model I (traditional variables + scorecard model, with AUC of 0.72) is limited. Third, other things being equal, adding more variables improves the effectiveness of credit risk assessment. Fourth, the combination of a large data set and a machine learning model greatly improves credit risk assessment. Adding BigTech proprietary information to the bank approach increases the AUC by 5.6%, while applying machine learning techniques adds an additional 11.1% to the AUC. Table 4 shows the 95% confidence intervals of the AUC for the four models (models I to IV) in Fig. 3. The differences in AUCs among the four models are significant.[8]

Table 5 reports the reduction in the NPL ratio when switching from the bank approach to the BigTech approach, using alternative thresholds of the expected default rate. Loans with expected default rates below the threshold will be accepted. We then calculate the default rates (NPL ratios) of loans accepted by the BigTech and bank approaches. The right column shows the reduction in NPL ratios from using the BigTech approach, compared with the bank approach, at given threshold levels. For example, if the threshold is an expected default probability of 10%, the BigTech approach reduces the NPL ratio by 1.05 percentage points. The improvement becomes smaller as we relax the threshold values from 10 to 30%.

When comparing the performance of different modeling methods, we control for the same information set. However, the same variables may contribute to the results differently using different models. To illustrate this point, we report the top 20 contributing variables for the scorecard and random forest models. Since the two models use different criteria in evaluating the importance of the

[7] The results are robust to alternative subsample periods.

[8] To alleviate the concern that the machine learning prediction models may be sensitive to shocks, we tested the robustness of above results by comparing the discriminatory power of the four models in two time periods, that is, before and after an exogenous regulatory shock, respectively. We use the issuance of draft guidelines that aimed to tighten regulations on the asset management activities of financial institutions by the People's Bank of China on November 17, 2017 as the shock, and find that our main conclusion that BigTech approach better predicts loan defaults than the bank approach is still robust.

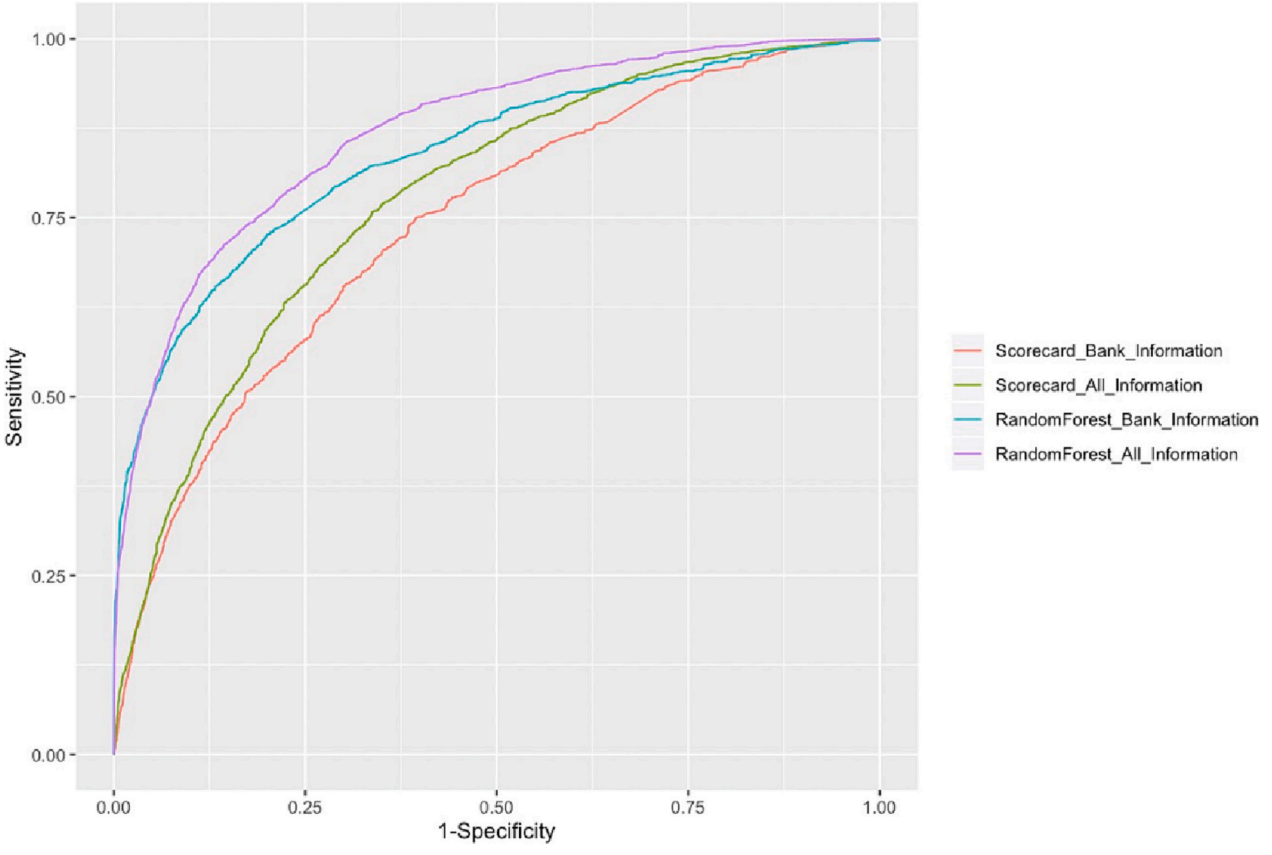


Fig. 5. ROC Curves for different models for traditional bank borrowers.
Notes: The graph shows the discriminatory power of four model specifications by providing the ROC curves. The sample only includes borrowers who have a credit history in traditional banking. ROC = receiver operating characteristics.
Source: Calculations using data from MYbank.

variables, the two sets of parameters are not comparable numerically. The scorecard model results are primarily driven by five variables, four on credit history and one on daily payment activity (Fig. 4, panel a). Among the top 20 variables, most are on transactions and payments. In contrast, in the machine learning model results, information values are more evenly distributed across a wide range of variables (Fig. 4, panel b), and transaction and payment data play a more important role than in the scorecard model. Critically, the key difference in the two models is the role of BigTech proprietary information, such as customers' ratings of vendors and network effect scores in the machine learning models. In a way, these indicators represent the borrowers' "digital collateral" pledged in the Alibaba ecosystem, as their reputation and business operations would be immediately affected in the case of default. These variables do not appear in the list of the top 20 variables in the scorecard model, probably because the relationships between these proprietary information variables and predicted defaults are not linear. Rather, these variables affect the results through interactions with other variables. For example, a high network effect score alone cannot guarantee loan repayment; it is useful only when combined with healthy cash flows.

4.3. Role of bank credit history

Next, we repeat the same exercises for a subsample of borrowers with bank credit history, which accounts for 7.5% of the total. We compare the importance of BigTech information ((a)(b)(c)(d)(e) in Table 3) relative to bank credit history. Again, a comparison of the ROC curves shows that the approach with machine learning models and all information is the most efficient, while the approach with scorecard models and bank credit history information is the least efficient (Fig. 5). Adding 'BigTech's information set to the bank approach or replacing the scorecard models with machine learning models significantly improves credit risk assessment.

Table 6 compares the AUCs of different combinations of models and information sets. First, bank credit history information is quite effective for predicting defaults, with an AUC of 0.74 in the bank approach using the scorecard models. Second, again, applying the machine learning models significantly increases the accuracy of predicting defaults, raising the AUC to 0.84. Third, BigTech information alone generates similar results as bank credit history information, using the scorecard and machine learning models. This finding is especially significant because many SMEs are not registered in the credit registration system. This result suggests that BigTech lenders could cover the massive number of small borrowers that have never been serviced by banks. And finally, the combination of the complete set of information and machine learning models delivers the best outcome for predicting defaults. Table 7 shows the 95% confidence intervals of the AUCs for the four models in Fig. 5. The differences in the AUCs among the four models are significant.

4.4. Inclusiveness of BigTech lending

The stylized facts show that the firms served by MYbank are much smaller than China's average SMEs. Does the advantage of BigTech credit assessment differ for different SMEs in this sample? The answer to this question has important implications for market structure: if MYbank's BigTech lending favors larger firms in the sample, this could indicate that there is a greater chance of competition with bank lending. However, if the BigTech approach favors smaller firms, then there should be greater complementarity between BigTech and bank lending. Even the "larger" firms in our sample are still SMEs.

We apply the cumulative distribution function of increase in estimated default probability in the BigTech and scorecard models in Fig. 6. Each line in the figure represents a subgroup. Borrowers for whom this difference is negative (to the left of 0 on the horizontal axis) are "winners" in the BigTech approach (in the sense of having a lower estimated default probability), and those with a positive difference (to the right of 0 on the horizontal axis) are "losers" (Fuster et al., 2020).

Table 6
Comparison of AUCs for different models for traditional bank borrowers.

Model	Bank credit history information	BigTech information	Bank credit history information + BigTech information
Scorecard model	0.74	0.72	0.78
Random forest	0.84	0.83	0.87

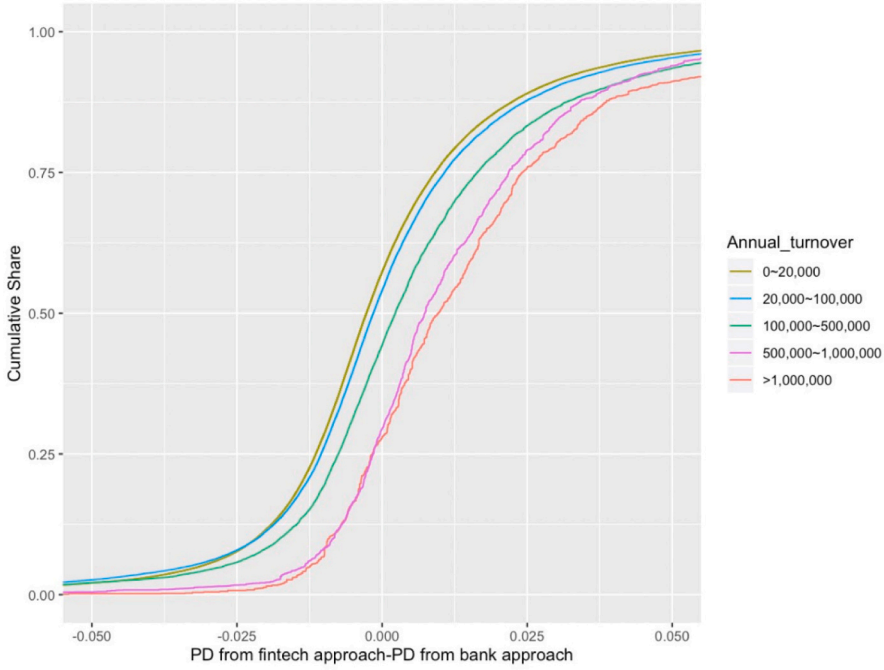
Source: Calculations using data from MYbank.
Notes: The table shows the discriminatory power of different model specifications by providing the AUC. BigTech information includes all the information (traditional information and BigTech proprietary information) we used in the baseline model. The sample only includes borrowers who have a credit history in traditional banking. AUC = area under the receiver operating characteristics curve.

Table 7
95% Confidence intervals of the AUCs for four models.

Model	AUC	95% conf. Interval	
Random forest + bank credit history information + BigTech information	0.8681	0.8636	0.8725
Random forest + bank credit history information	0.8372	0.8318	0.8427
Scorecard model + bank credit history information + BigTech information	0.7767	0.7713	0.7820
Scorecard model + bank credit history information	0.7397	0.7337	0.7456

Source: Calculations using data from MYbank.
Note: AUC = area under the receiver operating characteristics curve.

a. Comparing SMEs of Different Sizes



b. Comparing SMEs in Different City Tiers

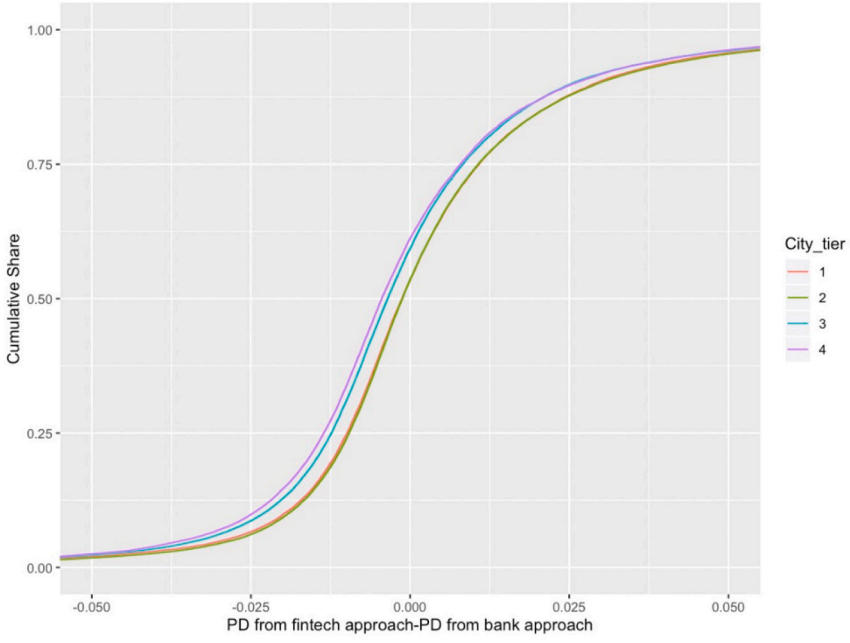


Fig. 6. Comparison of predicted default probabilities across models.
a. Comparing SMEs of Different Sizes.
b. Comparing SMEs in Different City Tiers.
Note: The graphs plot the cumulative distribution function of differences in the predicted probabilities of default between the BigTech model and the traditional model by differentiating SME sizes (panel a) and city tiers (panel b). PD = predicted default; SMEs = small and medium-size enterprises.
Source: Calculations using data from MYbank.

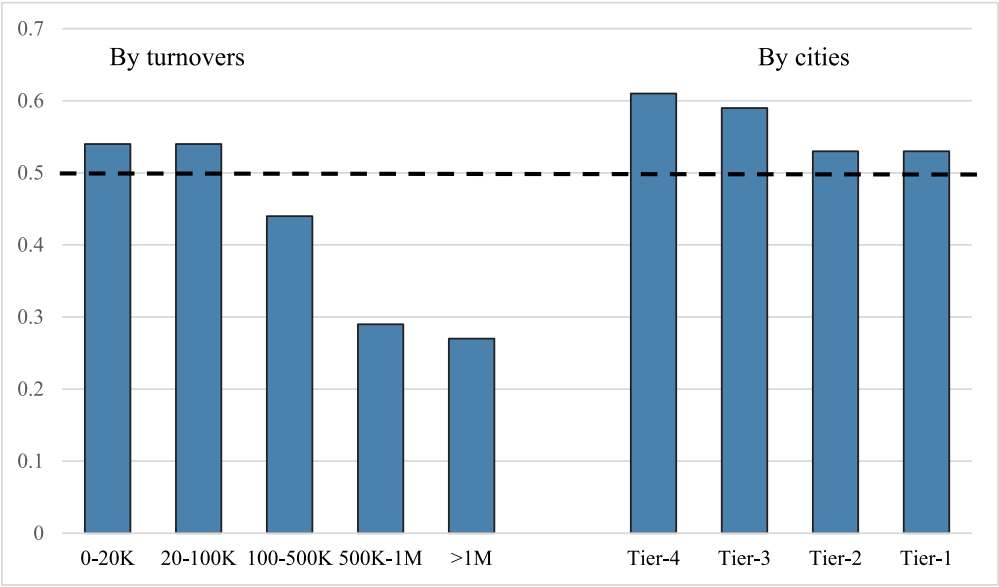


Fig. 7. Share of SMEs with lower predicted default rates using the BigTech Approach, by firm size and city location.
Source: Calculations using data from MYbank.

Table 8
AUCs for different borrower groups under different models.

a. Firm Size (RMB, thousands)					
Borrowers, by annual turnover	< 20	20–100	100–500	500–1000	> 1000
Bank approach	0.72	0.72	0.71	0.8	0.76
BigTech approach	0.84	0.84	0.82	0.89	0.85
BigTech – traditional	0.12	0.12	0.11	0.09	0.09

B. Firm Location				
Borrowers, by city tier	Tier-4	Tier-3	Tier-2	Tier-1
Bank approach	0.71	0.73	0.73	0.73
BigTech approach	0.84	0.85	0.84	0.84
BigTech – traditional	0.13	0.12	0.11	0.11

Source: Calculations using data from MYbank.
Notes: The table shows the discriminatory power of the traditional and BigTech model specifications across groups by providing the AUCs. The BigTech approach refers to a combination of random forest models and all information. The bank approach refers to a combination of scorecard models and traditional information. AUC = area under the receiver operating characteristics curve.

Fig. 6, panel a, compares the differences in predicted default probabilities between the BigTech approach and the bank approach for different subgroups of SMEs (measured by annual sales turnover). Among the smallest borrowers, the share for whom the estimated default probabilities drop using the BigTech model is around 57%, while it is around 27% for the largest borrowers. These findings mean that smaller businesses have lower estimated default probabilities and, therefore, are more likely to obtain loans when the BigTech approach is applied. Thus, compared with the bank approach, the BigTech approach benefits smaller firms proportionately more than larger firms.

Fig. 6, panel b, compares the predicted default probabilities between the BigTech approach and the bank approach for SMEs in different cities. For instance, for borrowers in Tier-4 cities,[9] about 62% have lower predicted default probabilities using the BigTech model, compared with 52% for borrowers in Tier-1 and Tier-2 cities. This finding implies that SMEs in lower-tier cities benefit more from the BigTech approach than those in higher-tier cities, likely because SMEs in smaller cities lack traditional data and rely more on

[9] There is no official government guideline on city classification. The tiering system adopted here is widely used in the media, reflecting a confluence of factors, such as economic development, population size, and administrative hierarchy. Tier-1 cities represent the most densely populated and developed urban areas in China, while Tier-4 cities are small and less developed.

proprietary information to secure small BigTech loans.

To illustrate the results for subgroups of SMEs, Fig. 7 plots the share of SMEs that obtain a lower predicted default rate using the BigTech approach, by size and city location. The BigTech approach favors smaller SMEs, especially those with annual turnovers of less than RMB 100,000. There is no clear advantage of applying the BigTech approach to larger SMEs, as banks may already have sufficient information on these companies. When grouped by city tier, all the SMEs benefit from the BigTech approach, although those in lower-tier cities benefit even more.

These findings confirm that, compared with the bank approach, the BigTech approach benefits smaller SMEs and SMEs in smaller cities. A follow-up question is whether such lower predicted default probabilities are as reliable as the other predictions. Here again, we use the AUCs to measure the reliability of the models for different borrower groups (Table 8). The first row in Table 8 shows the AUCs of the bank approach for borrowers of different sizes. Apparently, the model works better for relatively larger firms, that is, firms with annual turnover of at least RMB 0.5 million. The AUCs shown in the second row for the BigTech approach offer a similar pattern. However, by comparing the first and second rows, we find that the differences are greater for smaller firms than for larger firms (the third row).

5. Conclusion and policy implications

This study conducted analyses of the BigTech approach for credit risk assessment, using big data and machine learning models, relative to the bank approach, featuring standard financial information and scorecard models. Using a unique data set from China's MYbank, the study finds that the BigTech approach has a significant advantage in strengthening credit risk management and promoting financial inclusion for smaller SMEs.

These findings have important implications for policy makers to strengthen credit risk assessment and promote financial inclusion. Given the significant advantage of BigTech lenders in reaching unbanked SMEs, with unique data access and enhanced risk modeling, the government could actively encourage BigTech lending to promote financial inclusion. Meanwhile, the supporting policies should be complemented with strict licensing and regulatory checkups.

Also, as the digital payments are rapidly developing in many countries (such as UPI in India, Pix in Brazil, and PromptPay in Thailand), China's experience and success in BigTech lending model could also provide a useful lesson for other countries where SME lending remains a challenge.

Data availability

The data that has been used is confidential.

References

Abdulsaleh, A. M., & Worthington, A. C. (2013). Small and medium-sized enterprises financing: A review of literature. *International Journal of Business and Management, 8*, 36.

Agarwal, S., Alok, S., Ghosh, P., & Gupta, S. (2020). *Financial inclusion and alternate credit scoring for the millennials: Role of big data and machine learning in BigTech*. Working Paper,. National University of Singapore.

Agarwal, S., & Hauswald, R. (2010). Distance and private information in lending. *The Review of Financial Studies, 23*, 2757–2788.

Ahmed, M. I., & Rajaleximi, P. R. (2019). An empirical study on credit scoring and credit scorecard for financial institutions. *International Journal of Advanced Research in Computer Engineering & Technology, 8*, 275–279.

Bazarbash, M. (2019). *BigTech in financial inclusion: Machine learning applications in assessing credit risk*. Working Paper. Washington, DC: International Monetary Fund.

Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of BigTechs: Credit scoring using digital footprints. *The Review of Financial Studies, 33*, 2845–2897.

Berger, A. N., Frame, W. S., & Ioannidou, V. (2016). Reexamining the empirical relation between loan risk and collateral: The roles of collateral liquidity and types. *Journal of Financial Intermediation, 26*, 28–46.

Berger, A. N., & Udell, G. F. (2002). Small business credit availability and relationship lending: The importance of bank organisational structure. *The Economic Journal, 112*, F32–F53.

Berger, A. N., & Udell, G. F. (2006). A more complete conceptual framework for SME finance. *Journal of Banking & Finance, 30*, 2945–2966.

Besanko, D., & Thakor, A. V. (1987a). Collateral and rationing: Sorting equilibria in monopolistic and competitive credit markets. *International Economic Review, 671–89*.

Besanko, D., & Thakor, A. V. (1987b). Competitive equilibrium in the credit market under asymmetric information. *Journal of Economic Theory, 42*, 167–182.

Boot, A., Hoffmann, P., Laeven, L., & Ratnovski, L. (2021). BigTech: What's old, What's new? *Journal of Financial Stability, 53*.

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking & Finance, 72*, 218–239.

Cerqueiro, G., Ongena, S., & Roszbach, K. (2016). Collateralization, bank loan rates, and monitoring. *The Journal of Finance, 71*, 1295–1322.

Cornée, S. (2019). The relevance of soft information for predicting small business credit default: Evidence from a social Bank. *Journal of Small Business Management, 57*, 699–719.

Cornelli, G., Frost, J., Gambacorta, L., Rau, P. R., Wardrop, R., & Ziegler, T. (2022). Fintech and big tech credit: Drivers of the growth of digital lending. *Journal of Banking & Finance, 148*, Article 106742.

Demirgüç-Kunt, A., & Klapper, L. (2013). Measuring financial inclusion: Explaining variation in use of financial services across and within countries. *Brookings Papers on Economic Activity, 2013*, 279–340.

Demirguc-Kunt, A., Klapper, L., Singer, D., Ansar, S., & Hess, J. (2018). *The global findex database 2017: Measuring financial inclusion and the BigTech revolution*. The World Bank.

Freel, M., Carter, S., Tagg, S., & Mason, C. (2012). The latent demand for bank debt: Characterizing "Discouraged Borrowers". *Small Business Economics, 38*, 399–418.

Frost, J., Gambacorta, L., Huang, Y., Shin, H. S., & Zbinden, P. (2019). BigTech and the changing structure of financial intermediation. *Economic Policy, 34*, 761–799.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2020). Predictably unequal?. In *The effects of machine learning on credit markets*. Rochester, NY: Social Science Research Network. SSRN Scholarly Paper.

Gambacorta, L., Huang, Y., Li, Z., Qiu, H., & Chen, S. (2023). Data versus collateral. *Review of Finance, 27*(2), 369–398.

Gambacorta, L., Huang, Y., Qiu, H., & Wang, J. (2019). *How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese BigTech firm*. BIS Working Papers. Basel, Switzerland: Bank for International Settlements.
Hand, D. J., & Crowder, M. J. (2005). Measuring customer quality in retail banking. *Statistical Modelling, 5*, 145–158.
Hau, H., Huang, Y., Shan, H., & Sheng, Z. (2021). *Fintech credit, financial inclusion and entrepreneurial growth*. Swiss Finance Institute Research Paper No. 21–47, Geneva.
Hopper, M., & Lewis, E. (2004). Behaviour scoring and adaptive control systems. In *Readings in credit scoring: Foundations, developments, and aims*. Oxford University Press.
Jagtiani, J., & Lemieux, C. (2019). The roles of alternative data and machine learning in BigTech lending: Evidence from the LendingClub consumer platform. *Financial Management, 48*, 1009–1029.
Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance, 34*, 2767–2787.
Kithinji, A. M. (2010). *Credit risk management and profitability of commercial banks in Kenya*. Working Paper. Kenya: School of Business, University of Nairobi
Liu, L., Lu, G., & Xiong, W. (2022). *The big tech lending model*. National Bureau of Economic Research. No. w30160.
Miller, M., & Rojas, D. (2004). *Improving access to credit for SMEs: An empirical analysis of the viability of pooled data SME credit scoring models in Brazil, Colombia & Mexico*. Working Paper. Washington, DC: World Bank
Stiglitz, J. E., & Weiss, A. (1981). Credit rationing in markets with imperfect information. *American Economic Review, 71*, 393–410.
Thomas, L. C., Oliver, R. W., & Hand, D. J. (2005). A survey of the issues in consumer credit modelling research. *Journal of the Operational Research Society, 56*, 1006–1015.
Vos, E., Yeh, A. J.-Y., Carter, S., & Tagg, S. (2007). The happy story of small business financing. *Journal of Banking & Finance, 31*, 2648–2672.