Journal of Empirical Finance 62 (2021) 28–45

Contents lists available at ScienceDirect

Journal of Empirical Finance

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

Deciphering big data in consumer credit evaluation

Jinglin Jiang [b], Li Liao [b], Xi Lu [c], Zhengwei Wang [b], Hongyu Xiang [a,b,<]

[a] Business School, Beijing Normal University, PR China
[b] PBC School of Finance, Tsinghua University, PR China
[c] BaiRong Financial Information Service Co., Ltd, PR China

ARTICLE INFO

JEL classification:
G10
G21
G23

Keywords:
Big data
FinTech
Personal credit
Large-scale alternative data
Income exaggeration

ABSTRACT

This paper examines the impact of large-scale alternative data on predicting consumer delinquency. Using a proprietary double-blinded test from a traditional lender, we find that the big data credit score predicts an individual's likelihood of defaulting on a loan with 18.4% greater accuracy than the lender's internal score. Moreover, the impact of the big data credit score is more significant when evaluating borrowers without public credit records. We also provide evidence that big data have the potential to correct financial misreporting.

1. Introduction

Large volumes of alternative data, or big data, have recently become available, leading to a profound transformation of both economic research and practice (Einav and Levin, 2014).[1] In the context of consumer lending, financial institutions have begun using large-scale external data to evaluate the creditworthiness of potential borrowers. The increased prevalence of online loan application during the pandemic highlight the importance of this trend. Compared to simple digit footprints, using big data for lending decisions makes it more costly for people to change their behaviors. These data include behavioral loan tracking, location-based information, mobile app data, and much more, and they differ from traditional sources of information (e.g., financial information) in that they are granular, real data that are not self-reported. Despite the increased use of big data in practice, there is limited research on the performance of big data in a real business context. It remains an open question that whether the new wave of big data provide new information value in financial services and improve traditional business practices.

In this paper, we examine whether the availability of large-scale alternative data improve personal credit assessment for a traditional financial institution. Our study is based on a double-blinded test of an anonymous traditional lender, by comparing its own internal rating with the big data score constructed by BaiRong, a big data service company in China. The internal score is based on credit records from the public credit reporting system, account-level data, and self-reported demographic and income, while the big data credit score incorporates multiple dimensions that are unavailable to the lender. We are, therefore, able to assess the predictive power of the big data credit score both separately, compared to the traditional internal score, and jointly with the internal score.

< Corresponding author.
E-mail addresses: jiangjl.14@pbcsf.tsinghua.edu.cn (J. Jiang), liaol@pbcsf.tsinghua.edu.cn (L. Liao), xi.l654@hotmail.com (X. Lu), wangzhw@pbcsf.tsinghua.edu.cn (Z. Wang), xianghy.11@pbcsf.tsinghua.edu.cn (H. Xiang).

[1] A formal definition of big data is ``the information asset characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value'' (De Mauro et al., 2016).

https://doi.org/10.1016/j.jempfin.2021.01.009
Received 7 August 2020; Received in revised form 3 December 2020; Accepted 20 January 2021
Available online 10 February 2021

Financial institutions may choose to use alternative big data to complement China's underdeveloped credit reporting system. The financial credit reporting system, run by the Credit Reference Center of the People's Bank of China (PBCCRC), only covers one-quarter of the Chinese population, meaning that around one billion Chinese individuals lack a financial credit profile. The PBCCRC system provides credit records for licensed financial institutions. For financial institutions, no common credit scores, like the FICO score, is available. In practice, financial institutions develop their own internal score system based on public credit records and other in-house available information. Severe information asymmetry is observed between financial institutions and individuals without public credit records. External information from big data firms might help refine their credit evaluation.

In theory, however, it is unclear whether the big data credit score has the potential to outperform traditional lenders' internal score. The latter is based on core financial information, such as historical credit records at licensed financial institutions and financial account activities, which are essential for predicting default (e.g., Mester et al., 2007; Norden and Weber, 2010). Despite its low dimensionality, this financial information directly captures a borrower's ability to repay in the future. Thus, the traditional internal score could have greater information value in predicting defaults. However, the big data credit score could also outperform the internal rating because it has greater coverage and uses thousands of variables that cannot be easily manipulated. Traditional rating relies on official credit records that are not available for most individuals in China, which limits the coverage of traditional rating. Self-reported financial information, which is widely used in traditional credit risk assessment, is likely to be manipulated and subject to falsification.

We empirically examine the efficiency of the two scores using a real business sample, and we compare their respective predictive power. We find that the big data credit score substantially outperforms the lender's internal score. The AUC obtained by only using the lender's internal score is 0.761, while the AUC obtained using the big data credit score is 0.809. The comparison indicates that the big data credit score predicts a borrower's likelihood of delinquency with 18.4% greater accuracy than the internal score. Combining both scores predicts a borrower's delinquency likelihood with 22.6% greater accuracy than the model using only the lender's internal score. In terms of economic value, the big data credit score significantly raises the expected profit per applicant by CNY 1500–2500 (or USD 220–360). The magnitude of the expected profit is of great economic significance, as the querying fee for the big data credit score is usually less than CNY 50 (or USD 7) per applicant.

Further, we investigate who is rescued by big data credit scores. Among borrowers with below-median traditional ratings, those who submit more consistent identity information in multiple loan applications, fewer loan applications in online non-bank lenders, and allow a longer time interval between two consecutive loan applications are more likely to be granted higher big data credit scores.

One possible explanation for these findings is that the big data credit score incorporates an individual's very frequent and real-time behavioral information, such as online cash loan applications, online shopping, and internet surfing. First, these elements generate information that can better reflect an individual's general profile. While public credit records are the main source of an individual's financial credit profile, they only aggregate loan records at formal financial institutions. Besides, public credit records have limited coverage, which often prevents those without public credit records from accessing money from formal financial institutions. The results in this study are consistent with this interpretation. We find that, for borrowers *without* public credit records, using the big data credit score alone can achieve 99.6% of the predictive power of the combined model, which highlights big data's potential for those who lack a credit history.

Second, big data may provide an opportunity to correct misreporting. Potential borrowers find it more costly to manipulate big data credit scores than simple variables since the former incorporate a large set of variables. In contrast, self-reported financial information, such as income (Jiang et al., 2014), are more vulnerable to manipulation. We focus on borrower's income misreporting and argue that big-data-based income information may be a better proxy for real income than self-reported income. We provide supportive evidence for this argument: big-data-based income is significantly and positively associated with delinquencies, while self-reported income is not. Also, borrowers with self-reported income above the estimated income based on big data are more likely to become delinquent.

Our research's main contribution is to shed light on the value of large-scale alternative data, or ``big data'', in the context of personal credit assessment. To the best of our knowledge, this paper is the first to investigate how large-scale data improve credit evaluating in a real-world business scenario. Thus, this paper contributes to the literature by using such a setting to understand the information value of big data and its potential impact on participants in a real-world market.

The closest research to this study is Berg et al., 2020, who demonstrate that digital footprints perform better in predicting consumer loan delinquency than credit bureau scores. Our paper differs from and complements their findings along several dimensions. First, our paper investigates a machine-learning-based big data score that aggregates various types of information, including credit history in both non-bank and bank lenders, online shopping, and web browsing records. The score is constructed as many FinTech lenders and credit service providers typically do in real business. Instead of only using digital footprints, resorting to big data score deepens the current understanding of the information value of big data and its impact on consumer credit evaluation. Besides, the proposed big data score is constructed from 3312 variables. This approach overcomes the scope and frequency limitations of existing data sources of single digital footprint variables (which can be easily manipulated) and makes this study's conclusions less subject to Lucas critique (Lucas, 1976). Second, given that banks may use other information (e.g., income, bank account, and credit usage information) besides credit bureau scores for evaluating default risk, credit bureau scores likely underestimate a bank's credit assessment ability. Our paper relies on the lender's real internal rating, which summarizes the information that the lender uses in risk assessment. Internal rating thus serves as an appropriate benchmark for evaluating how lenders' screening power is improved by big data. Third, this study's sample individuals are minimally screened before being granted loans. Thus, the sample is more representative of the consumer loan applicant population.

This study is also related to several other strands of the literature. First, our research is related to the literature on how big data in alleviating information asymmetries in the consumer loan market, which exist both in developed and emerging economies (Adams et al., 2009; Stiglitz and Weiss, 1981). Oliver Wyman (Carroll and Rehmani, 2017) estimates that around 50 million people in the U.S. lack an informative credit score, which may lead to denials when applying for mainstream credit. The situation might be even worse in emerging markets, where the credit scoring system is often underdeveloped. This study contributes to this stream of the literature by providing evidence that large-volume information from alternative sources has the potential to fulfill this demand for credit information, thus expanding the access to credit for those without credit records.

Second, our research relates to the literature on financial intermediaries in the consumer lending market. The existing research has attached great importance to the ability of intermediaries' internal data, such as credit history and account data, to assess individual borrowers' risk (Mester et al., 2007; Norden and Weber, 2010). Marshall et al. (2010) includes customer loan approval process information in predicting loan performance. Khandani et al. (2010) construct a consumer credit score model via machine learning, but their model only incorporates bank account transactions and credit bureau information. Our model differs from theirs by including a large-scale alternative information. Our results show that, while internal information performs well in predicting delinquency, the use of external alternative data significantly improves credit evaluation. Our findings suggest that alternative data may provide meaningful insight for intermediaries making business decisions in the consumer lending market.

Third, this study contributes to the growing body of research on the role of big data and data analytics in economics (see Einav and Levin (2014) for an overview). The financial industry is heavily dependent on data. The advent of big data and analytics represents a major advance, with tremendous potential for real-world business. Recent studies have documented the impact of big data's application in capital markets, including serving as a corporate governance mechanism (Zhu, 2019), and the measurement of the FinTech innovation value based on capital market reaction (Chen et al., 2019). This study contributes to this stream of research by deciphering big data's value in consumer lending market.

The remainder of the paper is organized as follows. Section 2 provides an overview of the institutional background. Section 3 details the data and information underlying the two scores. Section 4 reports the main results and discusses their economic implications. Section 5 describes the possible mechanisms. Section 6 compares our paper to Berg et al., 2020, and Section 7 concludes.

2. Institutional background

2.1. China's financial credit information system

China's public credit bureau is run by an arm of the central bank, the Credit Reference Center at the People's Bank of China (PBCCRC), which maintains a credit reporting system on Chinese individuals. As stipulated by Regulation on Credit Reporting Industry enacted in March 15, 2013, this system acts as the Financial Credit Information Basic Database established by the State. According to the PBCCRC's release, the credit reporting system's coverage of individual borrowers was 361 million as of April 2015, around one-quarter of the population.

Licensed financial institutions, such as commercial banks, have an obligation to report credit information (i.e., loan borrowers' payments and defaults) to the PBCCRC system regularly. In turn, the PBCCRC system gathers information from licensed financial institutions and provides them with credit reports of individuals, which only contain credit records. The PBCCRC does not analyze the credit information or produces an aggregate credit score. Financial institutions need to analyze and produce credit scores on their own. Typically, they combine the information in credit reports with other information they proprietarily own to produce an internal credit score. We will discuss this issue in greater detail in Section 2.3. The credit reports are only available to financial institutions on a complementary basis, which means that BaiRong has no access to the PBCCRC credit reporting system.

2.2. BaiRong setting

BaiRong is one of the leading big data service providers in China. Founded in 2014, BaiRong provides big data services to multiple lenders to improve the efficiency of lending. Its customers include over 3500 financial institutions in China, around 300 of which are banks, and the others are FinTech lenders. BaiRong has invested significant resources in building a comprehensive database, covering a wide range of information regarding a borrower's creditworthiness. It comprises approximately 700 million individuals, over half of the nation's population.

Multiple data sources contribute to this database: (1) historical inquiry records accumulated through BaiRong daily businesses, (2) data gleaned from third-party data partners, such as personal identity information and blacklists maintained by an organization operated under the Ministry of Public Security, and (3) data aggregated from the internet using BaiRong's proprietary data collection technologies, with due authorization from the prospective borrowers. BaiRong has developed several proprietary automated programs that can search, aggregate, and process large amounts of data from different sources in a short period, producing and updating a ``big data credit score'' for individuals daily. We provide a more detailed description of the big data credit score in Section 3.2.

BaiRong provides credit scores for financial institutions to screen loan applicants. During formal cooperation, a lender first uploads a loan applicant's identifying information, that is, name, national ID number, and phone number, to BaiRong's credit evaluation system. The system queries the database, matches the application with data based on an encrypted and anonymized identifier, aggregates the data, and finally uses its proprietary credit evaluation model to generate a credit score for the applicant. The system then provides the lender with big data credit scores for the loan applicants. During the whole inquiry process, the identifier of borrowers is encrypted and anonymized.

2.3. How financial institutions do in-house credit screening and how they cooperate with BaiRong

Traditionally, financial institutions only rely on in-house information: borrowers' past credit records extracted from the PBCCRC system (only if the borrower has a borrowing history with licensed financial institutions), financial account balances and cash flows, and information on the application form (i.e., socio-demographic information and self-reported income information). Based on such information, financial institutions often employ a logistical model to produce a score that helps them make a loan decision. Such modeling strategies guarantee adequate performance when banks extend loans to relatively big customers who have abundant historical credit information in licensed financial institutions.

Commercial banks are now increasingly trying to attract small customers, such as individuals and small businesses, based on conversations with industry professionals. Small customers typically lack credit information in licensed financial institutions, which limits the ability of the traditional in-house screening methods to identify good clients. This is a critical issue as about three-quarters of the Chinese population have no credit history with licensed financial institutions. Thus, one available solution for a commercial bank is to seek external data sources with greater coverage of the population, for example, by cooperating with big data firms such as BaiRong. Complementing their in-house score with BaiRong's big data credit score is one typical way for commercial banks to make their loan decisions on small customers.

Before formal cooperation, a commercial bank would typically conduct a double-blinded test to compare the screening performances of its in-house score versus BaiRong's big data credit score. A typical test proceeds as follows: The bank first decides a pool of loan applicants, to which the bank only applies minimal checks. Thus, the test result is least affected by the bank's techniques and, to a larger extent, can be generalized to other applicants. The bank assigns these applicants an internal credit rating based on in-house information. Meanwhile, BaiRong provides its big data credit scores for the same applicants. BaiRong simply inputs applicants' identity information (i.e., name, national ID number, and phone number) into testing system based on an encrypted and anonymized identifier. The system reports big data credit scores for these individuals. Importantly, neither the bank nor BaiRong knows the other's score when they produce their own assessment.

After both scores are produced, the bank grants loans to these applicants and observes the loan performance. The bank then compares the big data credit score's ability to predict delinquency with that of its own internal score and determines whether to integrate BaiRong's big data credit score into its credit underwriting.

2.4. Advantages and challenges of using big data

One advantage of the big data credit score is that the underlying information covers real behaviors of borrowers such as location-based information and loan applications. Manipulating all these behaviors is both very difficult and costly. In addition, in the digital context, this information is available for a considerably large fraction of the entire population. In contrast, information used by traditional models, such as self-reported information, is usually uncertified and likely to be misreported (Garmaise, 2015). In addition, public credit reporting from the central bank does not cover three-quarters of the Chinese population. For those who have never accessed formal credit before, financial institutions can only rely on background information (e.g., age or gender) to evaluate creditworthiness.

Another advantage of big data derives from advanced data analytics. Using big data in credit evaluation usually implies applying advanced machine learning algorithms that have the desired predictive power. In contrast, traditional lending institutions focus more on a model's interpretability due to their hierarchical organizational structure, and thus, they prefer a traditional logistic model to advanced data analytics, which are usually harder to interpret.

Using big data in credit evaluation, however, also generates significant challenges. The first and foremost challenge is how to efficiently collect, store, and manage large-scale data. Computing credit scores from big data using traditional econometric models may also be challenging. Processing high-dimensional data directly is computationally expensive and prone to overfitting (Stanimirova et al., 2007; Yu et al., 2009; Bingham and Mannila, 2001).[2]

The second challenge is that, although based on a larger pool of data, the big data credit score lacks core financial data for borrowers. For example, unlike traditional financial institutions, big data companies have no access to borrowers' financial information, such as income and cash reserves. As account information helps predict loan performance (Norden and Weber, 2010), some might argue that the big data credit score performs worse than a traditional credit score.

3. Sample and two scores

3.1. Sample

Our sample is an anonymous traditional lender's testing sample with BaiRong. The lender is a typical lender in China's consumer loan market. As we discussed in Section 2.3, before formal cooperation, the lender should decide whether to introduce BaiRong's big

[2] Therefore, in processing high-dimensional data, we should first adopt efficient dimensionality reduction techniques to compress the data. However, zeros or missing values are often observed in high-dimensional data, which makes traditional dimensionality reduction techniques, such as principal component analysis (PCA), inefficient in processing such data. For instance, PCA is usually solved by the eigenvalue decomposition of a covariance matrix of variables. However, estimating the covariance matrix in the presence of many zeros or missing values is challenging. In traditional methods, variables or observations with many missing values may be simply deleted. In the big data context, however, the phenomenon of missing values is quite universal. The deleting process may, thus, lead to a significant loss of relevant information. As for model selection, readers of interest can refer to Huang et al. (2014). Huang et al. (2014) discuss how to select models for high-dimensional problems in great details.

Table 1
Information underlying the big data credit score.

Information type	Description	Examples
Inquiry records accumulated through BaiRong's daily business	Unique information on borrower's detailed loan application records across multiple lenders, such as peer-to-peer (P2P) lending platforms, cash loan firms, and finance companies.	Number of new loan applications in P2P platforms within seven days
Derogatory information across different types of lenders	A blacklist obtained by detecting any fraudulent behaviors, including borrowers with liens, judgments, settlements, historical payment and rejection records across multiple lenders, and extreme cases such as cheating on loans.	Whether the borrower has been rejected by cash loan firms
Information searched and aggregated from the internet with due authorization	Borrower's online shopping and payments from their accounts on certain popular Chinese e-commerce websites, locations and use of multiple devices, and website browsing data.	Whether the borrower was using his/her most-frequently-used cell phone when applying for credit

data credit score to enhance its credit evaluation model in future business by exploiting double-blinded tests. Specifically, the lender randomly selected 7838 auto loan applicants from its applicants' pool to be the testing sample. Only minimal authenticity checks were applied to the sample applicants. Then, the lender acquired these applicants' big data credit scores from BaiRong by submitting only their name, national ID number, and phone number into BaiRong's system. The lender also produced its own internal scores for these applicants. Finally, the lender granted loans to these applicants to observe their loan performance.[3]

One key advantage of this procedure is that, for the lender, the test results can be more reliably generalized to the lender's other similar loan applicants, as per conversations with industry professionals. If the test result is positive, the lender introduces BaiRong's big data credit score at an early stage of credit underwriting.

The full sample in our study comprises information from the lender's tests and covers 7838 borrowers. The sample has two characteristics that provide a rare chance to examine the performance of big data credit scores and traditional ratings in a consumer credit context. First, the sample's representativeness is least affected by the lender's techniques, given that no hard work was put into screening these borrowers other than minimal checks. According to the 2018 China Consumer Finance Industry report by Tongdun, 69% of consumer credit borrowers are male, and 72% of borrowers are between 22 and 40 years old. In our sample, 71.7% of borrowers are male, and 78.6% are between 22 and 40 years old (Table 2 Panel C), confirming the sample representativeness.

Second, since only identity information is submitted, BaiRong cannot incorporate the unique information owned by the lender into its big data credit scores. Therefore, the sample provides an ideal context to compare the performance of the two credit evaluation procedures.

Our dataset includes two credit scores both ranging from 0 to 10 with higher scores implying higher credit quality and lower loan delinquency rate: one is the lender's internal score. While the modeling technique of the lender's internal score is confidential, it is known to be constructed on the lender's information about the borrowers, including credit records from the public credit bureau, as well as the lender's proprietary information (i.e., account data and socio-demographic data).

The other is BaiRong's big data credit score, computed using the algorithm described in Appendix A. BaiRong have already had a training sample from the accumulated data of the past auto loan businesses before cooperating with the traditional lender in our paper. The big data credit model is trained using that training sample. For the lender, before deciding whether to introduce in the BaiRong's big data credit scoring model, the lender needs to test the performance of the model. To do so, the lender provides 7838 applicants' identity information to BaiRong to obtain the big data credit scores of these borrowers. The repayment information of the 7838 borrowers (i.e., the *y* variable) is *not* included when training the model. Therefore, it is an ``out-of-sample'' test of the performance of the BaiRong's big data credit scoring model. This out-of-sample test helps the lender compare the performance of two credit scores and decide whether to cooperate with BaiRong.

3.2. Information underlying the two scores

The big data credit score's underlying information comprises three broad dimensions, as shown in Table 1. The first dimension is collected from every piece of inquiry record accumulated through BaiRong's daily business. This dimension comprises unique information on a borrower's detailed loan application records from multiple lenders, such as peer-to-peer (P2P) lending platforms and cash loan platforms. The records include each inquiry since the business' inception, with identifiers for the subscriber (lender), time stamps, borrowers' names, national ID number, and phone number (encrypted). The data include, for example, the number of new loan applications on P2P lending platforms within a certain period, whether numerous applications are registered within a certain period, and the number of phone numbers related to a given national ID number, based on these original records.

The second dimension is collected from third-party data partners (i.e., multiple lenders and government agencies), including derogatory information across different types of lenders. BaiRong maintains a blacklist of any fraudulent behaviors, including: (1) historical payment and rejection records across multiple lenders, especially online lenders, such as P2P lending platforms and cash loan platforms (historical payment records include normal and past delinquency), (2) extreme cases such as cheating on loans, that

[3] The lender still grants the loans to individuals with low scores since this current test needs to know whether the individuals with lower scores will actually have a higher default rate in the future. Besides, only in this case can the result of this test be reliably applied to any future applicants, whether they are with high or low scores.

Table 2

Panel A. Big dat

Fig. 1. Big Data Credit Score, Traditional Rating, and Loan Performance. This figure plots the relationship between credit scores and loan performance. The data cover 7838 borrowers randomly selected among the consumer loan applicants of an anonymous traditional lender. *Traditional rating* is the internal credit score assigned by the lender. *Big data credit score* is based on variables provided by BaiRong. Both credit scores range from 0 to 10, with higher scores implying higher credit quality and a lower loan delinquency rate. Panel A plots the average overdue rate (line) and the number of borrowers (bars) in each big data credit score interval. Panel B plots the average overdue rate (line) and the number of borrowers (bars) in each traditional rating interval.

6.880. Panel B shows that the Pearson (Spearman) correlation coefficient between the two scores is 0.545 (0.466) in the full sample, implying considerable differences in the information content between the two credit scores. The correlation coefficient is as high as in the full sample for borrowers *with* a public credit record, but is much lower for borrowers *without* a public credit record. Panel C presents the gender and age distributions of the borrowers. 72% (28%) of borrowers are male (female). Nearly half (46%) of the borrowers are between 22 and 29 years old, one third are between 30 and 39 years old, and 16.1% of the borrowers are between 40 and 54 years old.

Fig. 1 shows the number of borrowers and loan performance for borrowers in different credit score intervals. The average overdue rate decreases monotonically with both the big data credit score and the traditional rating, indicating that both scores have considerable predictive power for loan delinquency.

In terms of the distribution of scores, the scores of these borrowers have a relatively wide range. This fact is consistent with the lender's practice in this test: granting loans to applicants with low scores, observing all individuals' performance, and comparing the two scores' predictive power of future default.

Table 3
Big data credit score, traditional rating, and loan performance. This table presents the comparison of the predictive powers of traditional rating, big data credit score for loan performance. The dataset covers 7838 borrowers randomly selected among the consumer loan applicants of an anonymous traditional lender. *Traditional rating* is the internal credit score assigned by the lender. *Big data credit score* is based on variables provided by BaiRong and constructed using the algorithm introduced in the Appendix A. Both credit scores range from 0 to 10, with higher scores implying higher credit quality and a lower loan delinquency rate. Panel A present the in-sample results of the logistics regressions that examine the relationship between credit scores and loan performance. The dependent variable is *Overdue*, a dummy variable, which takes the value of one when a loan is overdue. Panel B presents out-of-sample test results. We randomly divide the sample equally into training and test subsamples. We run each of the three logistic regressions, with the traditional rating, big data credit score, and combined score as the independent variable(s), on the training sample. The estimated parameters are then applied to the test sample to calculate the out-of-sample pseudo R^2. We repeat this process 500 times and report the mean of the 500 pseudo R^2s. T-statistics are reported in parentheses. Marginal effects are estimated at the average value of the explanatory variables.

Panel A. In-sample tests

Dependent variable:	(1)		(2)		(3)	
Pr(Overdue = 1)	Coefficients	Marg. Effect	Coefficients	Marg. Effect	Coefficients	Marg. Effect
Traditional rating	*0.424***	*2.395%			*0.137***	*0.658%
	(*23.438)				(*6.019)	
Big data credit score			*0.733***	*3.635%	*0.623***	*2.993%
			(*29.537)		(*20.653)	
Constant	*0.457***		2.147***		2.093***	
	(*5.878)		(14.721)		(14.433)	
N	7838		7838		7838	
Pseudo R^2	0.132		0.232		0.240	

Panel B. Out-of-sample tests

	(1) Traditional rating	(2) Big data credit score	(3) Both scores	Difference (2)-(1)	Difference (3)-(1)
Pseudo R^2	0.131	0.231	0.238	0.100***	0.108***
				(190.9)	(239.9)

*** indicates the difference is different from zero at the 1% level, ** at the 5% level, and * at the 10% level.

4.2. The predictive power of the big data credit score

4.2.1. Goodness-of-fit measures

Table 3 Panel A starts with in-sample tests. We run a logistic regression of the traditional internal score on whether the loan is ex-post overdue. Column (1) shows that the overdue rate decreases with the traditional rating. We measure the quality of the internal score as a prediction tool with a simple goodness-of-fit (pseudo R square) from the logistic regression. The pseudo R square of the traditional rating model is 0.132. Column (2) shows that the big data credit score also has a significantly negative correlation with overdue rate. The pseudo R square of the big data credit score model is 0.232. Thus, the predictive power of the big data credit score is nearly twice that of the traditional internal score model, suggesting that the big data credit score alone performs better in predicting delinquency than the traditional internal rating. The pseudo R square of the combined model in Column (3), is 0.240. These results suggest that traditional lenders in our sample can improve their screening process by incorporating the big data credit score (i.e., the information provided by big data complements the traditional score).

To alleviate the overfitting concerns on in-sample test, we conduct an out-of-sample test. The sample is randomly divided into two equal subsamples: the training sample and the test sample. We run each of the three logistic regressions, similar to the specifications reported in Table 3 Panel A, on the training sample. The estimated parameters are then applied to the test sample to calculate the out-of-sample pseudo R square. We repeat this process 500 times. Table 3 Panel B shows that the out-of-sample tests yield very similar results to the in-sample tests. The predictive powers of the model using big data credit score alone and the combined model are both significantly higher than that of the traditional rating. Fig. 2 Panel A plots the distribution of the differences between the two pseudo R squares estimated from the 500 simulations of the out-of-sample test (big data credit score vs. traditional rating) and Panel B between the two pseudo R squares of combined model and traditional rating model. The distributions indicate that the differences are significantly larger than zero. These results indicate that the main finding is robust to out-of-sample tests.

4.2.2. AUC metrics

We also use ROC (Receiver Operating Characteristics) curve and AUC (Area under curve) to evaluate the predictive power of the traditional rating, big data credit score, and the combination of the two scores. A higher AUC implies greater predictive power. A ``perfect'' predicting tool, one that always makes a correct prediction, has an AUC of 1, while an essentially random predicting tool has an AUC of 0.5. As suggested by Iyer et al. (2016), a 0.01 improvement in AUC is considered a noteworthy gain, an AUC of 0.7 is generally considered desirable in information-rich environments, and AUCs of 0.6 or greater are the goal in information-scarce environments.

Fig. 3 presents the ROC curves and AUC measures of the big data credit score, the traditional rating, and the combination of the both scores. The AUC of the traditional rating is 0.761, significantly different from a random predicting tool (AUC of 0.5). This result is higher than the 0.683 AUC of the credit bureau score alone, as documented in a consumer loan sample from a German E-commerce company (Berg et al., 2020), and higher than the 0.625 AUC obtained using the Experian credit score alone in a Prosper sample

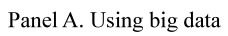

Fig. 2. Distribution of the Prediction Power Difference: Out-of-Sample Test. This figure plots the distribution of prediction power differences (measured by differences of pseudo R^2) using the full sample. The data cover 7838 borrowers randomly selected from among the consumer loan applicants of an anonymous traditional lender. *Traditional rating* is the internal credit score assigned by the lender. *Big data credit score* is based on variables provided by BaiRong and constructed using the algorithm introduced in Appendix A. Panel A plots the distribution of prediction power differences between the models using only the big data credit score and using only the traditional rating. Panel B plots the distribution of prediction power differences between the models using both credit scores and using only traditional rating. Each distribution is estimated from an out-of-sample test. We randomly divide the sample equally into training and test samples. Then we run three logistic regressions on the training sample, respectively, with traditional rating, big data credit score, and both scores as the independent variable(s). The estimated parameters are then applied to the test sample to calculate pseudo R^2 and compute the prediction power differences. We repeat this process 500 times and plot the distribution.

(Iyer et al., 2016). The larger AUC measure may reflect that the traditional rating is based on the lender's proprietary information on borrowers (i.e., account data, socio-demographic data, and self-reported income) in addition to the credit records from the public credit bureau. These results suggest that the traditional lender's internal rating has significant predictive power, and we use 0.761 as a benchmark for the big data credit score in the following comparison.

Notably, the AUC of the big data credit score alone is 0.809, implying that the big data credit score predicts a borrower's likelihood of delinquency with 18.4% greater accuracy than the internal lending score.[5] When using both scores, the AUC is 0.820, higher than the AUC of each of the stand-alone models. The combined model predicts a borrower's delinquency likelihood with 22.6% (= (0.820-0.5)/ (0.761-0.5)-1) higher accuracy than the lender's internal rating. These results suggest that the big data credit score improves upon the internal rating's prediction ability for a traditional lender (i.e., the big data credit score provides additional information and complements the internal score).

4.3. Economic analysis

We conduct a simulation study to assess the economic value that the big data credit score can provide to the traditional lender. The simulation estimates the economic value added (EVA) by big data credit score, as computed by the expected NPV the lender earns when combining the big data score with internal rating, minus the expected NPV when using internal rating alone.

[5] Following Iyer et al. (2016), we calculate the percentage improvement as (0.809-0.5)/(0.761-0.5) = 1.184.

Fig. 3. ROC Curve and AUC: Full Sample. This figure plots the ROC curves of the traditional rating, big data credit score, and both scores, respectively, in the full sample. The x-axis is the false positive rate (FPR). The y-axis is the true positive rate (TPR). To compute the ROC curve of both the traditional rating and big data credit score, we first estimate a logistic model, with whether the loan is overdue as the y variable and both scores as the x variables. AUC is the area under the corresponding ROC curve.

We rely on the following assumptions to simplify the simulation: (1) all the loans offered by the lender have the same size, maturity, and interest rate, and (2) all the loans have a lump sum payment schedule. Under these assumptions, the realized NPV of a repaid loan and that of an overdue loan are, respectively, given by:

$$NPV_{Repaid} = Amount \bullet \left[\frac{.1 + Interest/^{Maturity}}{.1 + Discount/^{Maturity}} * 1 \right]; \quad (1)$$

$$NPV_{Overdue} = Amount \bullet [\frac{1 * Loss}{.1 + Discount/^{Maturity}} * 1]; \quad (2)$$

where Amount **denotes the loan size,** Interest **is the annual compounded interest rate of the loan,** Discount **is the discount rate that the lender uses to evaluate the loan, and** Loss **is the percentage of loss against loan size, conditional on overdue payment.**

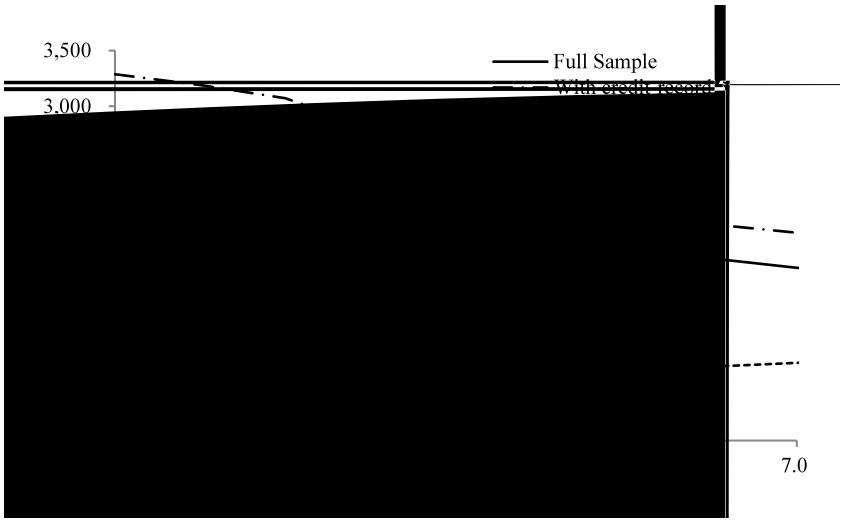

Fig. 4. Economic Value Added. This figure plots the economic value added by using the big data credit score. The economic value added is defined as the expected NPV when using both the traditional rating and big data credit score as the screening tools minus that when using only the traditional rating. The full sample covers 7838 borrowers randomly selected from among the consumer loan applicants of an anonymous traditional lender. The ``with-official-credit-record subsample'' contains 5,468 borrowers with a public credit record in the full sample. The ``without-official-credit-record subsample'' contains 2370 borrowers without a public credit record in the full sample. We assume that each loan has a lump sum payment schedule, with a size of CNY 150,000, a maturity of three years, and an interest rate of 15%.

100%. Considering the difficulty in calibrating the lender's discount rate, we choose a series of discount rates (from 3% to 7%) and calculate the EVA separately.

The results of the simulation in Fig. 4 show the EVA obtained by incorporating the big data credit score, with a discount rate between 3% and 7%. The solid line indicates that, in the full sample, the EVA achieved by incorporating the big data credit score is CNY 1500–2500 (or USD 220–360) per applicant. This amount is of great economic significance, given that the per-applicant querying fee for a big data credit score is usually less than CNY 50 (or USD 7).

4.4. Who is rescued by the big data credit score?

One potential benefit of using big data credit scores is to expand credit access to credit-constrained individuals, especially those whose loan applications are likely to be rejected by using traditional credit rating but approved by using big data credit scores. This raises the question of which borrower characteristics are related to the likelihood of being rescued by big data scores.

To address this question, we focus on borrowers who are granted below-median traditional ratings, assuming that these borrowers are more likely to be rejected by the traditional credit rating. We split these borrowers into two groups: (1) the rescued group, consisting of borrowers with above-median big data credit score, who are more likely to be rescued based on the big data credit score, and (2) the unrescued group, consisting of borrowers with below-median big data credit scores, whose loan applications are less likely to be approved based on the big data credit score.

Then, we compare the differences between these two groups in the 3312 variables. Given the high dimensional data, we only present the variables with the largest positive and smallest negative standardized difference between the two groups. For each variable, the standardized difference measure is computed as the mean difference between the rescued group and the unrescued group scaled by the sample standard deviation of the variable. A higher absolute value of the standardized difference suggests a more prominent difference in the corresponding variables.

Table 4 presents the variables with the 15 most positive and the 15 most negative standardized differences. These findings highlight three features of rescued borrowers. First, borrower in the rescued group have used more different names or phone numbers in past applications (using ID number as the identifier). In other words, borrowers submitting more consistent identity information in previous loan applications are more likely to be rescued by the big data credit score. In practice, using inconsistent names or phone numbers in past loan applications, usually at different lenders, is a sign of using faked identity and high-risk borrowers. This information plays a significant role in big data credit assessment but is barely used in traditional rating.

Second, rescued borrowers have applied for fewer loans from non-bank lenders (e.g., cash loan lenders and P2P lenders). More loan applications show urgent money demand and possibly less ability to repay. Interestingly, applying for more loans from banks suggests higher credit quality for individuals with below-median traditional ratings. One possible explanation is that high-quality borrowers choose to apply from banks, instead of non-bank lenders, seeking for lower interest rate.

Third, the time interval between consecutive loan applications is larger for rescued borrowers than unrescued borrowers. More frequent loan applications imply more urgent demand for credit, which is usually related to higher risk.

Table 4
Variables with the largest difference between rescued and unrescued applicants. This table reports 30 variables with the largest standardized difference between the rescued and unrescued applicants by the big data credit score. For each variable, the standardized difference is the mean difference between rescued and unrescued applicants, divided by the sample standard deviation. Rescued applicants are those with below-median traditional ratings and above-median big data credit scores. Unrescued applicants are those with below-median traditional ratings and below-median big data credit scores. Panel A reports the variables with the smallest 15 negative standardized differences, and Panel B reports the largest 15 positive standardized differences.

Panel A. Variables with the smallest 15 standardized difference	
Standardized difference	**Variable definition**
*1.414	Whether an individual is granted middle risk rating by BaiRong based on his/her bank loans (including credit card) records (using ID number as the identifier for the individual).
*1.229	Whether any previous loan application information of an individual is recorded in BaiRong's database.
*1.205	Whether any the identity information of an individual is recorded in BaiRong's database.
*1.138	Whether any of previous loan information (e.g., application, number of loans, and outstanding loans) of an individual is recorded in BaiRong's database.
*0.975	Whether the information on the number of online cash loan institutions that an individual has applied to is recorded in BaiRong's database (using the phone number as the identifier for the individual).
*0.964	The number of names that have been used by an individual in loan applications within the last 12 months (using ID number as the identifier for the individual).
*0.892	The number of names that have been used by an individual in loan applications within the last six months (using ID number as the identifier for the individual).
*0.835	The number of names that have been used by an individual in loan applications within the last three months (using ID number as the identifier for the individual).
*0.767	The number of names that have been used by an individual in past loan applications (using ID number as the identifier for the individual).
*0.756	Whether an individual's cash loan application information is record in BaiRong's database (using ID number as the identifier for the individual).
*0.726	The number of phone numbers that have been used by an individual in loan applications within the last 12 months (using ID number as the identifier for the individual).
*0.704	The number of months in which an individual has applied for loans in non-bank institutions within the last three months (using ID number as the identifier for the individual).
*0.698	The number of phone numbers that have been used by an individual in loan applications within the last six months (using ID number as the identifier for the individual)
*0.678	The number of months in which an individual has applied for loans in non-bank institutions within the last six months. (using ID number as the identifier for the individual)
*0.678	The number of phone numbers that have been used by an individual in loan applications within the last three months (using ID number as the identifier for the individual)
Panel B. Variables with the largest 15 standardized difference	
Standardized difference	**Variable definition**
0.929	The number of non-bank institutions from which an individual applied for loans at night within the last month (using ID number as the identifier for the individual).
0.935	The number of non-bank institutions from which an individual applied for loans at weekends within the last 15 days (using ID number as the identifier for the individual).
0.935	The number of non-bank institutions from which an individual applied for loans at weekends within the last 15 days (using the phone number as the identifier for the individual).
0.943	Days since the latest change of an individual's phone number (using ID number as the identifier for the individual).
0.954	The maximum time interval between two consecutive applications within the last three months (using ID number as the identifier of individuals).
0.980	The minimum time interval between two consecutive applications to banks within the last three months (using ID number as the identifier of individuals).
1.077	The number of an individual's loan applications to banks (excluding online bank). in the 10th month before (using ID number as the identifier for the individual).
1.077	The number of an individual's loan applications to banks in the 10th month before (using ID number as the identifier for the individual).
1.117	Days since the latest query with another phone number.
1.125	Days since the latest loan application with the same ID number and phone number.
1.168	The number of an individual's loan applications to banks (excluding online bank) in the 8th month before (using ID number as the identifier for the individual).
1.168	The number of an individual's loan applications to banks in the 8th month before (using ID number as the identifier for the individual).
1.201	The number of an individual's loan applications to banks in the 4th month before (using ID number as the identifier for the individual).
1.201	The number of an individual's loan applications to banks (excluding online bank) in the 4th month before (using ID number as the identifier for the individual).
1.226	The minimum time interval between two consecutive applications within the last three months (using the phone number as the identifier of individuals).

Table 5
Comparisons of the two scores' predictability: with vs. without-public-credit-record subsample. This table examines the relationship between the two credit scores and loan performance in both the ``with-public-credit-record subsample'' and ``without-public-credit-record subsample,'' respectively. The two subsamples are obtained from the full sample comprising 7,838 borrowers, which are randomly selected from the consumer loan applicants of an anonymous traditional lender. The ``with-public-credit-record subsample'' comprises 5,468 borrowers The ``without-public-credit-record subsample'' comprises the 2370 borrowers. Panel A shows result for in-sample test and Panel B out-of-sample test. The dependent variable is *Overdue,* a dummy variable with a value equal to one when the loan is overdue (ex-post), and zero otherwise. *Traditional rating* is the internal credit score assigned by the lender. *Big data credit score* is based on variables provided by BaiRong and constructed using the algorithm introduced in the Appendix A. Both credit scores range from 0 to 10, with higher scores implying higher credit quality and a lower loan delinquency rate. T-statistics are reported in parentheses.

Panel A. In-sample tests

	With-public-credit-record subsample			*Without-public-credit-record subsample*		
	(1) Pr(Overdue=1)	(2) Pr(Overdue=1)	(3) Pr(Overdue=1)	(4) Pr(Overdue=1)	(5) Pr(Overdue=1)	(6) Pr(Overdue=1)
Traditional rating	*0.383***		*0.128***	*0.441***		*0.203**
	(*20.967)		(*5.568)	(*5.781)		(*2.375)
Big data credit score		*0.687***	*0.579***		*0.849***	*0.759***
		(*26.357)	(*18.302)		(*9.548)	(*7.961)
Constant	*0.420***	2.013***	1.934***	*1.083***	2.403***	2.832***
	(*5.378)	(13.440)	(12.966)	(*2.896)	(4.187)	(4.767)
N	5468	5468	5468	2370	2370	2370
Pseudo R^2	0.136	0.233	0.241	0.043	0.126	0.134

Panel B. Out-of-sample tests

	(1) Traditional rating	(2) Big data credit score	(3) Both scores	Difference (2)-(1)	Difference (3)-(1)
With-public-credit-record subsample					
Pseudo R^2	0.135	0.232	0.240	0.097***	0.107***
				(162.6)	(211.0)
Without-public-credit-record subsample					
Pseudo R^2	0.038	0.123	0.125	0.085***	0.088***
				(73.3)	(86.2)

*** indicates the difference is different from zero at the 1% level, ** 5% level, and * 10% level.

5. How does the big data credit score improve prediction?

In this section, we try to figure out the way in which big data enhances model predictability. We provide suggestive evidence for the following two possible channels. First, incorporating alternative data may mitigate the information asymmetry between the lender and borrowers, especially those lacking information. Second, high-frequency online behavior data may help better assess those variables that are typically subject to misreporting.

5.1. Providing information for those without credit records

Based on a large scale of alternative data, the big data score covers more individuals than public credit system. Therefore, one explanation for the big data score's better performance is that it reveals credit quality information on borrowers without public records. We test this explanation by dividing our sample into two subsamples. One subsample, labeled ``with-public-credit-record subsample'', includes 5,468 borrowers. The other, labeled ``without-public-credit-record subsample'', includes the remaining 2370 borrowers.

Table 5 Panel A presents the logistic regression results of the in-sample test. For borrowers *with* a public credit record, the big data credit score model or the combined model has nearly twice the power of the traditional rating model in predicting overdue loans. For borrowers *without* a public credit record. The predictive power of the big data credit score model or the combined model is around three times that of the traditional rating model. These results provide evidence that adding a big data credit score to the traditional rating model substantially increases the model's predictive power for borrowers *without* a public credit record. To mitigate the overfitting problem, Panel B shows the results for the out-of-sample test whose procedure is similar to that in Section 4. The in-sample finding is robust to out-of-sample test. Notably, the difference between the average pseudo R^2 of the combined model and that of the big data credit score model is only 0.002, which suggests that for borrowers *without* a public credit record (i.e., the traditional rating is based only on the lender's proprietary information), the availability of a traditional internal rating only makes a very marginal contribution to the prediction when the big data score is available.[6]

We also plot the ROC curve and conduct an AUC analysis of the two subsamples. Fig. 5 Panel A presents the AUC measures for borrowers *with* a public credit record. The AUC of the traditional rating is 0.767, while the AUC of the big data credit model is 0.811. This indicates that the big data credit score has a 16.48% (= (0.811-0.5)/ (0.767-0.5)-1) higher predictive power than the traditional rating in the with-public-credit-record subsample. In Panel B, we present the AUC for the without-public-credit-record subsample.

[6] Appendix B shows the distributions of the differences between pseudo R^2s in subsamples, confirming the results.

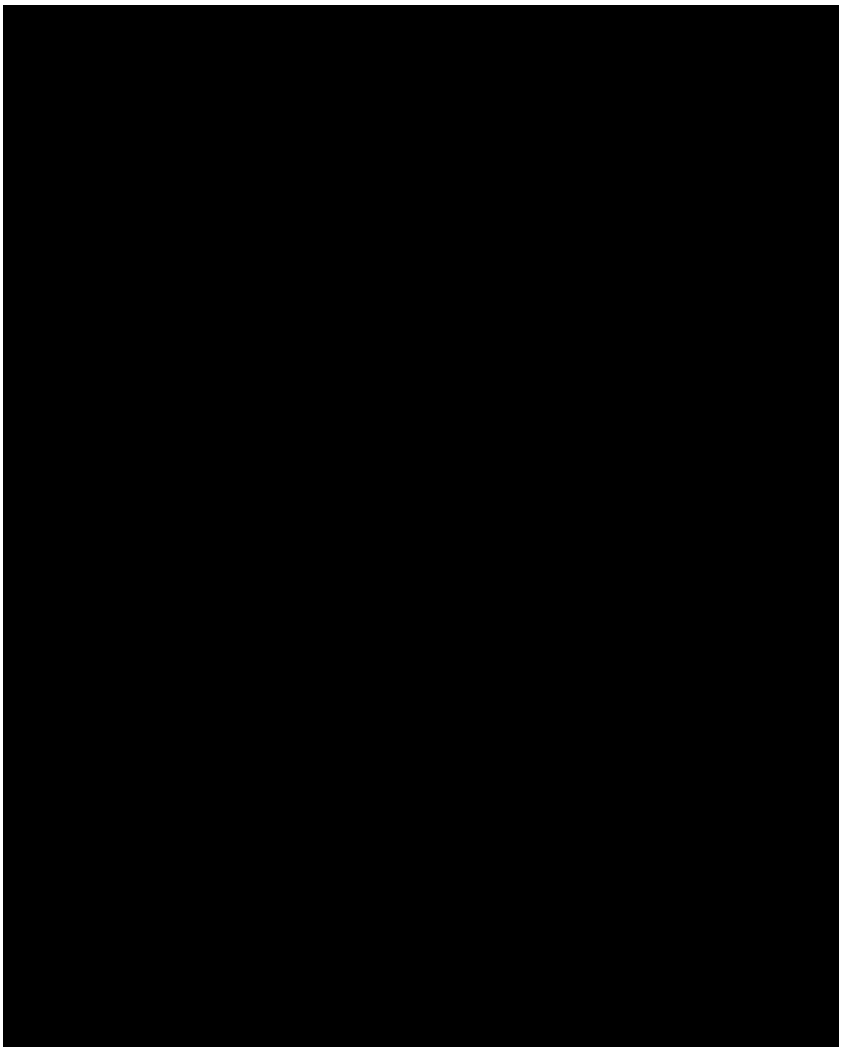

Fig. 5. ROC Curve and AUC: Subsamples with/without credit records. This figure plots the ROC curves of the traditional rating, the big data credit score, and the combination of both the scores, respectively. The subsamples are obtained from the full sample comprising 7838 borrowers that are randomly selected from the consumer loan applicants of an anonymous traditional lender. *Traditional rating* is the internal credit score assigned by the lender. Panel A plots the ROC curves for the ``with-public-credit-record subsample'' comprising 5,468 borrowers with a public credit record. Panel B plots ROC curves for the ``without-public-credit-record subsample'' comprising 2370 borrowers without a public credit record.

The AUC of the traditional rating alone is 0.667, while that of the model using the big data credit score alone is 0.752, which corresponds to an increase of 50.9% (=(0.752-0.5)/(0.667-0.5)-1). This result suggests that the big data credit score outperforms the internal rating by a larger magnitude for individuals *without* public credit records.

Using a similar approach to that in Section 4, we also evaluate the economic value added by employing the big data credit score for two subsamples, the ``with-public-credit-record subsample'' and ``without-public-credit-record subsample''. The upper dashed line in Fig. 4 represents the subsample of borrowers with a public credit record, while the lower dashed line indicates borrowers without a public credit record. The results show that the big data credit score adds more value for borrowers with public credit records than those without.

5.2. Correcting financial misreporting: Income exaggeration

In this section, we provide evidence that a significant portion of the credit rating improvement induced by the big data credit score may result from the detection of financial misreporting. Despite ample anecdotal evidence,[7] there are no formal empirical analyses of how employing big data may help identify borrower information falsification. Our paper fills this gap.

[7] See, for example, EY reporting ``How big data and analytics are transforming the audit'', available at https://www.ey.com/en_gl/assurance/how-big-data-and-analytics-are-transforming-the-audit, accessed on September 15, 2019, or CFA Institute ``Data and Technology: Transforming the Financial Information landscape'',

Table 6
Self-reported income, big-data-based income, and loan delinquencies. This table reports the results of Logistic regressions that examine the relation between delinquency and self-reported income vs. big-data-based income. *INC_reported* is individual borrowers' self-reported income and takes the value of 0 if self-reported income is missing. *INC_reported_miss* is a dummy variable for missing self-reported income. *INC_bigdata* is the income measure estimated from BaiRong's big data and takes the value of 0 if the estimated income is missing. *INC_bigdata_miss* is a dummy variable for a missing big-data-based income measure. *Female* is a dummy taking the value of one for female borrowers and zero for male borrowers. *Age* is the age of the borrower on the loan inception date. *Marginal effects* are the marginal effect of 1 standard deviation increase in the corresponding variable and are estimated at the sample mean.

	(1)	(2)	(3)	(4)
Ln (INC_reported)	0.140	0.143		
	(1.542)	(1.472)		
INC_reported_miss	1.175	1.260		
	(1.304)	(1.331)		
Ln (INC_bigdata)			*0.438***	*0.438***
			(*7.587)	(*7.461)
INC_bigdata_miss			*2.883***	*2.855***
			(*6.555)	(*6.393)
Female		*0.309***		*0.397***
		(*3.166)		(*4.071)
Age		*0.017***		*0.011**
		(*3.106)		(*2.011)
Province dummies		Yes		YES
N	7838	7838	7838	7838
Pseudo R^2	0.001	0.022	0.020	0.042
Marginal effects				
Ln (INC_reported)	4.89%	4.90%		
Ln (INC_bigdata)			*2.67%	*2.47%

Income reflects a person's financial wherewithal and is, thus, widely used to inform loan decisions, affecting loan qualification and pricing. However, in practice, self-reported income is often subject to falsification. Jiang et al. (2014) examine borrower income falsification and its impact on mortgage delinquency in the 2004–2008 mortgage crisis and find that income exaggeration exhibits a significant positive relation with delinquency.

While self-reported income is subject to falsification due to the low cost of manipulation (such as a misreported proof of income), online behavioral information might be more difficult or costly to manipulate in a comprehensive way because online behaviors occur more frequently. BaiRong estimates borrowers' monthly income by applying machine learning algorithms to individuals' shopping behavior on E-commerce platforms, social security payments, and wage information collected from job posting websites.

When collecting the data for this study, we also utilized big-data-based income (hereafter, ``big data income'') for the sample borrowers from BaiRong.[8] We argue that this measure can be a better proxy for real income than self-reported income, as it is based on an individual's real daily transactions. We present two pieces of indirect supporting evidence for this argument. First, we examine how self-reported income and big data income are associated with loan delinquencies. As shown in Columns (1) and (2) of Table 6, self-reported income is not significantly related to loan delinquencies. By contrast, as presented in Columns (3) and (4), big data income is negatively related to overdue rates at a 1% significance level. Given that true income should help predict credit quality (e.g., Van Order and Zorn, 2000), these results imply that big data income is more likely closely related to individuals' true income than self-reported income.

Second, we construct a dummy variable, *INC_Exaggeration*, which takes the value of one if self-reported income is higher than big data income, and zero otherwise. If big data income reflects individuals' real income, this dummy variable will be a good measure of income exaggeration and should be positively related to loan delinquencies (Jiang et al., 2014). This is consistent with the results in Table 7. We further split the borrowers with both self-reported income and estimated income available into two subsamples: the income exaggeration subsample, consisting of the borrowers whose self-reported incomes are higher than their estimated incomes, and the non-exaggeration subsample, comprising the borrowers whose reported incomes are no higher than their estimated ones. We examine the goodness-of-fit of the big data credit score and traditional rating models, respectively. Table 8 shows that the big data credit score model outperforms the traditional rating model in both subsamples. The difference in the AUC is 2.8 percentage points in the income-exaggeration subsample, compared to the 0.9 percentage points in the non-exaggeration subsample. The results suggest that the big data credit score performs relatively better for individuals whose self-reported income is higher than their estimated true income, also suggesting that big data may help detect income exaggeration.[9]

available at https://www.cfainstitute.org/-/media/documents/article/position-paper/data-technology-transforming-financial-information-landscape.ashx, accessed on September 15, 2019.

[8] Big data income is an ordinal variable taking the value of an integer between 1 and 100. The variable takes the value of m (m $\leq$ 100) when the estimated monthly income ranges from CNY 1000m—1000 to 1000m. In what follows, we use the middle point of the range as the estimated income. The results in this subsection are robust when we use the minimum of the income range as the estimated monthly income.

[9] One possible concern is that some unique information in traditional ratings, such as bank account information, can also be informative in estimating borrowers' income. We find that, after controlling for traditional ratings, big data income can still have significant predictive power in terms of loan delinquencies, thus indicating that big data income provides additional information that is not reflected in traditional ratings.

Table 7
Income exaggeration and loan delinquency. This table reports the results of the Logistic regressions that examine how the income exaggeration measure is associated with loan delinquencies. The full sample comprises 7838 borrowers. *INC_exaggeration* is a dummy that takes a value of one if the self-reported income is higher than the big data income. *INC_bigdata* is the income measure estimated from BaiRong's big data and takes the value of 0 if the estimated income is missing. *INC_bigdata_miss* is a dummy variable for a missing big-data-based income measure. *Female* is a dummy taking the value of one for female borrowers and zero for male borrowers. *Age* is the age of the borrower on the loan inception date. *Marginal effects* are the marginal effect of 1 standard deviation increase in the corresponding variable and are estimated at the sample mean.

	(1)	(2)
INC_exaggeration	0.695***	0.683***
	(2.983)	(2.907)
Ln (INC_bigdata)	*0.333***	*0.334***
	(*4.958)	(*4.919)
INC_bigdata_miss	*1.480**	*1.477**
	(*2.350)	(*2.317)
Female		*0.392***
		(4.009)
Age		*0.011**
		(*1.971)
Province dummies		Yes
N	7838	7838
Pseudo R^2	0.023	0.044

Table 8
Goodness-of-fit of models: Income exaggeration. This table presents the AUC of the traditional rating and big data credit score for borrowers with and without income exaggeration. The income exaggeration subsample comprises the 3252 borrowers whose self-reported incomes are higher than their estimated real incomes. The non-exaggeration subsample comprises 990 borrowers whose self-reported incomes are not higher than their estimated real incomes.

	N	(1) AUC: big data credit score	(2) AUC: Traditional rating	Difference (1)-(2)
Income exaggeration subsample	3252	79.7%	76.9%	2.8%
Non-exaggeration subsample	990	71.0%	70.1%	0.9%

6. Discussion: Comparison to Berg et al., 2020

In this section, we briefly discuss three differences between our paper and Berg et al., 2020, who compare the performance in predicting loan delinquency of digital footprints and a credit bureau score. First, in contrast with Berg et al., 2020, who focus on eight digital footprint variables, this study investigates a machine-learning-based big data score that aggregates various types of information, including credit history in both non-bank and bank lenders, online shopping, and web browsing records. These are typical practices in many FinTech lenders and credit service providers. Thus, the present research deepens the current understanding of whether big data improve individual credit evaluation.

Second, in evaluating the big data score's performance, we use the lender's real internal score as the benchmark, whereas Berg et al., 2020 use a score from a private credit bureau. As financial institution lenders are likely to use both the credit bureau score and the in-house information (e.g., income, bank account information, and credit usage) that is not included in credit bureau scores, lenders' internal scores are possibly a more appropriate benchmark for evaluating how lenders' screening power is improved by big data.

Third, this study uses a financial institution's double-blinded test sample of borrowers, who are subject to minimal screening before receiving loans, whereas Berg et al., 2020 use a sample of pay-by-invoice furniture buyers in a German E-commerce company, who have been screened through digital footprints and information from two private credit bureaus. Thus, this study reduces the concern of selection bias by comparing two competing scores' predictive power in a less pre-screened sample.

7. Conclusion

We use a proprietary double-blinded sample from a traditional financial institution lender to evaluate the potential impact of big data on the consumer credit assessment. The dataset provides an ideal context for comparing the big data credit score and the lender's internal rating due to a minimally screened sample and two independently constructed scores. In line with previous studies, we find that alternative data have information content in predicting consumer default. In particular, the big data credit score significantly outperforms the lender's internal score.

Moreover, we are in a unique position to decipher the big data credit score based on large-scale data and observe borrowers' income. We analyze the underlying variables and find that among borrowers with lower scores in the lender's internal score, those who submit more consistent identity information in multiple loan applications, fewer loan applications in online non-bank lenders, and allow longer time interval between two consecutive loan applications are more likely to score higher in the big data credit score. Hence, they are more likely to be rescued. Finally, we identify two possible ways in which big data might play an improving

role. First, incorporating big data can improve the prediction power for borrowers without a public credit record. Second, big data can be used to estimate true income, which contributes to the detection of income misreporting.

This study's research question is relevant for emerging economies, and the findings have important policy implications, even though the present analysis is solely based on Chinese data. Credit constraints are perceived as one of the main drivers of inequality (Jappelli, 1990; [illegible]). According to the World Bank's estimation, approximately one third of adults worldwide lack access to formal credit, while the proportion is [illegible] in low-income economies. One reason is the lack of credit bureau records ([illegible] 2012). [illegible] highlight the opportunity for developing economies to make more people access formal credit, especially when 40% of adults in developing [illegible]

pevansciumstis3J0 0 0 rg 0 0 0 RG0 0 0 rg 0 0 0 RG 0 -10.439 Td [(inp)aionc, ecimg397(e)leingeinu e incomlus [illegible]

Norden, L., Weber, M., 2010. Credit line usage, checking account activity, and default risk of bank borrowers. Rev. Financ. Stud. 23, 3665–3699. http://dx.doi.org/10.1093/rfs/hhq061.

Stanimirova, I., Daszykowski, M., Walczak, B., 2007. Dealing with missing values and outliers in principal component analysis. Talanta 72, 172–178. http://dx.doi.org/10.1016/j.talanta.2006.10.011.

Stiglitz, J.E., Weiss, A., 1981. Credit rationing in markets with imperfect information. Amer. Econ. Rev. 71, 393–410.

Van Order, R., Zorn, P., 2000. Income, location and default: Some implications for community lending. Real Estate Econ. 28 (3), 385–404.

Yu, K., Zhang, T., Gong, Y., 2009. Nonlinear learning using local coordinate coding. Adv. Neural Inf. Process. Syst. 222, 3–2231.

Zhu, C., 2019. Big data as a governance mechanism. Rev. Financ. Stud. 32 (5), 2021–2061.