DISCUSSION PAPER SERIES

DP14259

HOW DO MACHINE LEARNING AND NON-TRADITIONAL DATA AFFECT CREDIT SCORING? NEW EVIDENCE FROM A CHINESE FINTECH FIRM

Leonardo Gambacorta, Yiping Huang, Han Qiu and Jingyi Wang

FINANCIAL ECONOMICS

ISSN 0265-8003

HOW DO MACHINE LEARNING AND NON-TRADITIONAL DATA AFFECT CREDIT SCORING? NEW EVIDENCE FROM A CHINESE FINTECH FIRM

Leonardo Gambacorta, Yiping Huang, Han Qiu and Jingyi Wang

Discussion Paper DP14259
Published 28 December 2019
Submitted 22 December 2019

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

HOW DO MACHINE LEARNING AND NON-TRADITIONAL DATA AFFECT CREDIT SCORING? NEW EVIDENCE FROM A CHINESE FINTECH FIRM

Abstract

This paper compares the predictive power of credit scoring models based on machine learning techniques with that of traditional loss and default models. Using proprietary transaction-level data from a leading fintech company in China for the period between May and September 2017, we test the performance of different models to predict losses and defaults both in normal times and when the economy is subject to a shock. In particular, we analyse the case of an (exogenous) change in regulation policy on shadow banking in China that caused lending to decline and credit conditions to deteriorate. We find that the model based on machine learning and non-traditional data is better able to predict losses and defaults than traditional models in the presence of a negative shock to the aggregate credit supply. One possible reason for this is that machine learning can better mine the non-linear relationship between variables in a period of stress. Finally, the comparative advantage of the model that uses the fintech credit scoring technique based on machine learning and big data tends to decline for borrowers with a longer credit history.

JEL Classification: G17, G18, G23, G32

Keywords: Fintech, credit scoring, non-traditional information, Machine Learning, credit risk

Leonardo Gambacorta - leonardo.gambacorta@bis.org
Bank for International Settlements and CEPR

Yiping Huang - yhuang@nsd.pku.edu.cn
Institute of Digital Finance and National School of Development, Peking University

Han Qiu - hqiu93@163.com
Institute of Digital Finance and National School of Development, Peking University

Jingyi Wang - w-asx@163.com
School of Finance, Central University of Economics and Finance; and Institute of Digital Finance, Peking University

Acknowledgements
We would like to thank Sebastian Doerr, John V. Duca, Jon Frost, Xiang Li, Julapa Jagtiani and, in particular, one anonymous referee for comments and suggestions. We would also like to thank seminar participants at the University of Basel, the Bank for International Settlements, Bocconi University and the Irving Fisher Committee – Central Bank of Malaysia for useful comments. We thank Giulio Cornelli for excellent research assistance. Yiping Huang and Han Qiu's work is supported by the Chinese National Social Science Foundation (Project 18ZDA091). The views in this paper are those of the authors only and do not necessarily reflect

those of the Bank for International Settlements. The authors wish to highlight that the data and analysis reported in this paper may contain errors and are not suited for the purpose of company valuation or to deduce conclusions about the business success and/or commercial strategy of the anonymous Chinese fintech firm. All statements made reflect the private opinions of the authors and do not express any official position of the anonymous fintech firm or its management. The analysis was undertaken in strict compliance with Chinese privacy law. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper. The anonymous fintech firm did not exercise any influence on the content of this paper, but has ensured the confidentiality of the (raw) data.

1. Introduction

Financial technology (fintech) is taking on an ever more important role in lending decisions, while lending by fintech companies is gaining a significant share of certain market segments. In the United States, for instance, online lenders now account for about 8–12% of new mortgage loan originations, with Quicken Loans being recognised as the country's largest mortgage lender in terms of flow at the end of 2017 (Buchak et al (2017); Fuster et al (2018)). China is a country where new fintech credit is relatively well developed, representing around 3% of total outstanding credit to the non-bank sector at the end of 2017 (BIS (2019)).

New credit scoring models used by fintech lenders differ from traditional models in two key ways. The first is that technology allows financial intermediaries to collect and use a larger quantity of information. Fintech credit platforms may use alternative data sources, including insights gained from social media activity (U.S. Department of the Treasury (2016); Jagtiani and Lemieux (2018a)) and users' digital footprints (Berg et al (2018)). In the case of large technology companies (big techs) with existing platforms, data collection extends to orders, transactions and customer reviews (Frost et al (2019)).

The second difference is the adoption of machine learning techniques. In contrast to traditional linear models such as the logit model, machine learning can mine the non-linear information from variables. For example, Khandani et al (2010) construct a non-linear, non-parametric forecasting model for consumer credit that is based on machine learning techniques and find that this new model can outperform other models in a range from 6% to 25% of total losses. However, the prediction capability of machine learning models has mainly been demonstrated in applications with a stationary external environment. Their performance also needs to be verified in the case of a structural shock that changes the main relationships between the variables.

This paper contributes to the literature by addressing the following four questions:

i) Are machine learning-based fintech credit scoring models better able to predict borrowers' losses and defaults than traditional empirical models?

ii) What is the information content of non-traditional sources such as digital applications on mobile phones and e-commerce platform data?

iii) How do the different models perform in the event of an (exogenous) shock?

iv) How do the different models perform for customers with a different credit history?

The first two questions have also been analysed by other papers, with mixed results. Our contribution is mostly to highlight and explain differences in the results using a more comprehensive set of control variables. The third and fourth questions are completely new and represent the main contribution of the paper.

To answer these four questions, we use a unique data set from a leading Chinese fintech company at loan-transaction level for the period between May and September 2017. The fintech firm has requested to remain anonymous but has given us access to a very comprehensive data set. Compared to previous studies, this data set allows us

to disentangle the effects of traditional bank-type information (credit card information) and non-traditional information (obtained from the use of digital applications on mobile phones and e-commerce platforms). Moreover, we can assess the performance of the credit scores calculated by the fintech company using machine learning methods and such large volumes of data. Papers based on data from Renrendai, a Beijing-based company providing P2P financial services (see, for example, Braggion et al, 2019) cannot use credit card transaction information because Renrendai's borrowers typically do not have a current account with a bank.

Furthermore, unlike other fintech companies, in which borrower information is self-reported by the users themselves (see for example Berg et al (2018)), our fintech company is able to read both credit card transactions and digital app information directly from the system (with the user's permission). The information is therefore collected more comprehensively to include both credit card information and additional non-traditional information.

We analyse personal loans, most of which are repayable in up to one year. We also observe the borrowers' repayment record until October 2018 in order to track the status (viable or defaulted) of each loan after origination. This enables us to evaluate the performance of each loan ex post in terms of losses and defaults.

In order to answer the third question, we analyse the effects of a largely unexpected regulatory change that occurred in China in the period under review. On 17 November 2017, the People's Bank of China (PBoC) – the Chinese central bank – issued specific draft guidelines to tighten regulations on shadow banking. This regulatory change has led many financial intermediaries to increase their lending requirements, causing credit conditions for borrowers to deteriorate. In particular, the aggregated data indicate a significant increase in the default rate and a drop in lending after the shock. A similar pattern can be observed at our fintech company, which enables us to study how the different models performed during this stress period.

The main conclusions of our paper can be summarised as follows:

i) The fintech's machine learning-based credit scoring models outperform traditional empirical models (using both traditional and non-traditional information) in predicting borrowers' losses and defaults.

ii) Non-traditional information improves the predictive power of the model.

iii) While the models perform similarly well in normal times, the model based on machine learning is better able to predict losses and defaults following a negative shock to the aggregate credit supply. One possible reason for this is that machine learning can better mine the non-linear relationship between variables in the event of a shock.

iv) The predictive power of all the models improves when the length of the relationship between bank and customer increases. However, the comparative advantage of the model that uses the fintech credit scoring technique based on machine learning tends to decline when the length of the relationship increases.

2. Literature review

A few studies have started to analyse how credit supplied by fintech firms and their scoring models perform compared with traditional bank lending. Jagtiani and Lemieux (2018a) compare loans made by a large fintech lender and similar loans that were originated through traditional banking channels. Specifically, they use account-level data from LendingClub and the Y-14M data reported by bank holding companies with total assets of $50 billion or more. They find a high correlation between interest rate spreads, LendingClub rating grades and loan performance. Interestingly, the correlations between the rating grades and FICO scores have declined from about 80% (for loans that were originated in 2007) to only about 35% for recent vintages (originated in 2014–2015), indicating that LendingClub has increasingly used non-traditional alternative data.

Using market-wide, loan-level data on US mortgage applications and originations, Fuster et al (2018) show that fintech lenders process mortgage applications about 20% faster than other lenders, even when controlling for detailed loan, borrower and geographic observables. It is interesting to note that faster processing does not come at the cost of higher defaults. Furthermore, fintech lenders adjust their supply more elastically than other lenders in response to exogenous mortgage demand shocks, thereby alleviating the capacity constraints associated with traditional mortgage lending. Buchak et al (2018) compare the pricing of online (fintech) lenders in the US mortgage market with the pricing of banks and (non-fintech) shadow banks; they find that fintech lenders charge a premium of 14–16 basis points over bank mortgages. Jagtiani et al (2019) find that fintech lenders in the United States tend to supply more mortgages to consumers with weaker credit scores than do banks; they also have greater market shares in areas with lower credit scores and higher mortgage denial rates.

While banks usually incentivise borrowers to pay their loans back by requiring them to pledge tangible assets (eg real estate) as collateral, fintech credit is typically uncollateralised. This makes the use of big data particularly relevant when considering a loan application. Preliminary evidence based on credit data for China suggests that big data can act as a substitute for collateral: the volume of corporate loans supplied by big techs does not correlate with asset prices, whereas bank loans do. The left-hand panel of Figure 1 shows that the elasticity of bank credit to firms with respect to asset prices is close to one for collateralised credit and 0.5 for bank credit to SMEs, whereas credit to small firms it is not statistically different from zero in the case of big tech.

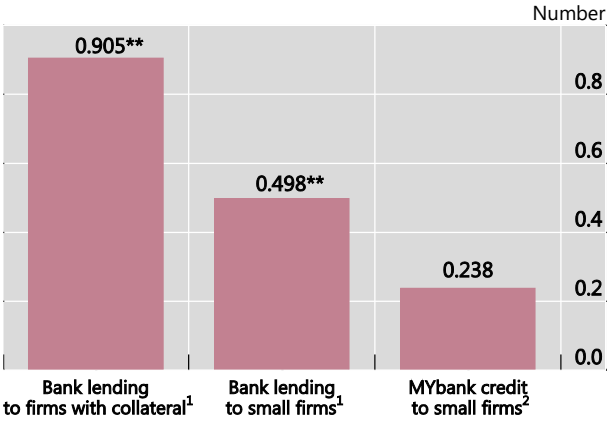
Frost et al (2019) use data for Mercado Credito, which provides credit lines to small firms in Argentina on the e-commerce platform Mercado Libre. They find that, when it comes to predicting loss rates, credit scoring techniques based on big data and machine learning have so far outperformed credit bureau ratings. A key question here is whether this outperformance will persist through a full business and financial cycle. Indeed, fintech credit could give rise to new forms of non-prudent risk-taking that needs to be tested in the event of an adverse shock. For example, De Roure et al (2016) find that online lenders in Germany substitute bank loans for high-risk consumer loans. For US consumer credit markets, Tang (2019) finds that online lending substitutes for bank lending by serving marginal borrowers, but complements bank lending in terms of loans. Interestingly, the performance of online lenders seems to

depend on the quantity and quality of information to which online lenders have access.

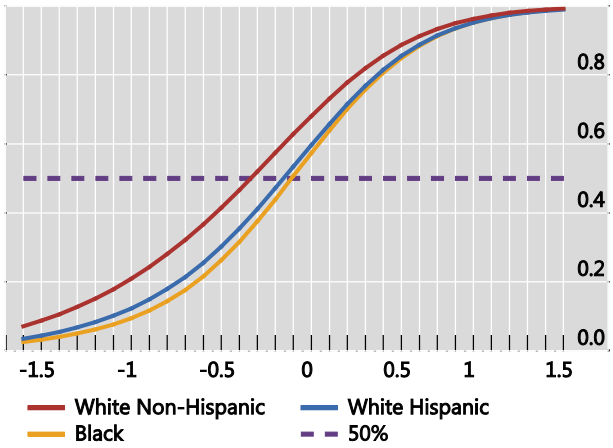
Elasticity of fintech credit to asset prices, and winners and losers of machine learning credit allocation

Figure 1

Elasticity of credit with respect to house price in China



Machine algorithm bias[3]



** indicates significance at 5% level.

[1] Period of estimation 2005–13. [2] Period of estimation 2011–17. [3] The change in the log predicted default probability as lenders move from traditional predictive technology (a logit classifier) to machine learning technology (a random forest classifier) is reported on the horizontal axis. The cumulative share of borrowers from each racial group who experience a given level of change is reported on the vertical axis.

Sources: Fuster et al (2019); authors' calculations (left-hand panel).

Some of the literature looks at the informational content of digital soft information and credit performance. Dorfleitner et al (2016) study the relationship between soft factors in P2P loan applications and financing and default outcomes. Using data on the two leading European P2P lending platforms, Smava and Auxmoney, they find that soft factors influence the funding probability but not the default probability. Jagtiani and Lemieux (2018a) find that the rating grades assigned on the basis of alternative data perform well in predicting loan performance over the two years after origination. The use of alternative data has allowed some borrowers who would have been classified as subprime by traditional criteria to be slotted into "better" loan grades, enabling them to benefit from lower priced credit. In addition, for the same risk of default, consumers pay smaller spreads on loans from LendingClub than from credit card borrowing. Berg et al (2018) show that digital footprints are a good predictor of the default rate. Analysis of simple, easily accessible variables from digital footprints is equal to or better than the information from credit bureau scores.

Another stream of the literature analyses *unfair* price discrimination. In particular, sophisticated machine learning algorithms may not be as neutral as their mathematical nature suggests at first glance. Even though artificial intelligence and machine learning algorithms are neither trained nor fed with protected characteristics such as race, religion, gender or disability, they are able to triangulate such information. Using data on US mortgages, Fuster et al (2019) find that Black and Hispanic borrowers are disproportionately less likely to gain from the introduction of machine learning in credit scoring models, suggesting that the algorithm may develop differential effects across groups and increase inequality (see the right-hand panel of

Figure 1, taken from their study). Borrowers to the left of the solid vertical line represent "winners", who are classed as less risky by the more sophisticated algorithm than by the traditional model. Reading off the cumulative share around this line, we see that about 65% of White Non-Hispanic and Asian borrowers win, compared with about 50% of Black and Hispanic borrowers.

3. Data description

We access data on proprietary loan transactions from a leading fintech company in China for the period between May 2017 and September 2017.[1] To obtain credit through the platform, customers need to provide the fintech company with bank credit card information and additional non-traditional information (via platform services). For each customer, the fintech company calculates a credit score that assesses risk on the basis of machine learning technology and information provided by customers (via their credit card transactions, the digital apps on their phones and e-commerce platform data).[2]

In our analysis, we will try to disentangle the information content of the credit score, the credit card information (which is typically what a traditional bank observes) and non-traditional information (accessible via social media and platforms use).

The fintech company decides whether to grant a loan or not on the basis of the fintech score. However, the threshold value is not fixed and is adjusted in line with general economic and funding conditions. For example, the fintech company increased the threshold value for the credit scoring when credit conditions deteriorated in the wake of the regulatory change in November 2017. We will follow up on this in more detail later in the paper.

The fintech company provides personal loans with a maturity of up to 24 months, although the vast majority (more than 80%) mature after one year. In order to analyse performance, we also access the loan repayment records up to October 2018 to let us evaluate loan defaults and the losses incurred by the fintech firm. Table 1 provides descriptive statistics of the data set.

Some of the variables seem quite skewed, with rather extreme outliers. For instance, the average frequency of credit card usage over the past year is 6.65, but the max is at 2,637. Similarly, there appear to be some people with extremely high numbers of defaults on their credit card (eg 31 over the past 3 months) and the max repayment is RMB 1.3 million, while the average is 15. Such skewness in the variables could lead the simple "linear" models to do worse than if the independent variables had more "normal" distribution. Therefore, we will use winsorised variables, at the 1% level in the regressions.

[1] We did not use the volume of credit extended by the fintech company in October 2017, before the regulatory shock, for two reasons: first, to avoid criticism that the fintech company could have anticipated the regulatory change in November; and second, because the fintech company changed the rules for its credit score in October 2017.

[2] The fintech company used a decision tree approach applied to a database of 300 variables. The company ended up using 20 variables to calculate fintech credit scores. One of the reasons why they did not include all 300 variables was to avoid an overfitting problem.

Descriptive statistics Table 1

Variables	Obs	Mean	Std dev	Min	Max
Default dummy (0/1)	343,976	0.2199194	0.4141924	0	1
Loss rate	343,976	0.1302757	0.2711493	0	1
Fintech credit score	343,976	623.7213	29.97468	576	815
Interest rate	343,976	3.966122	1.071682	2.489046	8.47319
Number of bank accounts (last 3 months)	343,976	0.6274188	0.6599821	0	13
Number of bank accounts (last 12 months)	343,976	4.643048	2.690386	0	18
Frequency of credit card usage (last 12 months)	343,976	6.658348	22.0309	0	2637
Frequency of credit card usage (last 3 months)	343,976	1.5239	6.665101	0	494
Large payment counts	343,976	35.93432	53.57181	0	3155
Credit line (RMB)	343,976	41062.79	39863.85	0	3,500,000
Credit card defaults (last 12 months)[1]	343,976	0.3961759	1.150032	0	59
Credit card defaults (last 3 months)[1]	343,976	0.0529601	0.3297239	0	31
Repayments (RMB)	343,976	15.34175	2970.462	0	1,281,800
Credit history (months)	343,976	25.89493	17.83306	0	126
Salary deposited in current account	343,976	902.7588	8,272.06	0	1,500,000
Gender (0=male)	343,976	0.244	0.4295	0	1
Max call duration (last 12 months)	343,976	2739.147	1873.6	0	10157
Call times to family (last 12 months)	343,976	287.2105	446.2912	0	2340
Frequency of calls (3 months)	343,976	1090.113	873.6526	0	4329
Frequency of calls (daily)	343,976	6.155257	4.488499	0	22.21557
Taobao payments (daily) (RMB)	343,976	518.6343	1206.691	0	6914.921
Defaults (Taobao) [2]	343,976	0.0003082	0.0214299	0	5

[1] Number of credit card defaults. [2] Number of times a borrower has not paid/delivered goods on the Taobao e-commerce platform.

4. A horse race between different credit scoring models

4.1 Empirical strategy

Our first goal is to assess whether fintech credit scoring models (based on machine learning plus big data) are better able to predict borrowers' losses and defaults than linear models (based on traditional and non-traditional data).

We start by estimating different models to predict total losses:

$$L_{i,t} = \alpha CS_{i,t} + \mu_P + \mu_T + \varepsilon_{i,t} \quad (1)$$

$$L_{i,t} = \beta X_{i,t} + \mu_P + \mu_T + \varepsilon_{i,t} \quad (2)$$

$$L_{i,t} = \beta X_{i,t} + \delta Y_{i,t} + \mu_P + \mu_T + \varepsilon_{i,t} \quad (3)$$

where $L_{i,t}$ indicates the loss rate (as a percentage of the origination volume) on a loan. The first information set includes the fintech credit score for borrower *i* at time *t* ($CS_{i,t}$). The second information set includes a vector of variables obtained through the credit card ($X_{i,t}$). This set of traditional information is typically also available to a bank. The third set of information also includes a vector of non-traditional variables ($Y_{i,t}$) obtained by the fintech company through customers' mobile phone apps and their activity on the e-commerce platform. All models include province (μ_P) and time (μ_T) fixed effects and $\varepsilon_{i,t}$ is an error term. Equations (1) to (3) are estimated using a tobit model given the censored nature of data (either 0 or positive).

The second set of equations are

$$p(D_{i,t}) = \Phi(\alpha CS_{i,t} + \mu_P + \mu_T + \varepsilon_{i,t}) \quad (1')$$

$$p(D_{i,t}) = \Phi(\beta X_{i,t} + \mu_P + \mu_T + \varepsilon_{i,t}) \quad (2')$$

$$p(D_{i,t}) = \Phi(\beta X_{i,t} + \delta Y_{i,t} + \mu_P + \mu_T + \varepsilon_{i,t}) \quad (3')$$

where $p(D_{i,t})$ indicates the probability for the borrower of a loan to default (and to generate a loss). Equations (1') to (3') are estimated using a logit model, which is more appropriate than probit models for large sample sizes.

To sum up, we consider three different models with different information sets. Model I only uses the fintech score as the independent variable, while Model II only uses the traditional bank-type information set as independent variables. Model III includes both traditional and non-traditional information as independent variables. We need to stress that for Models II and III we use the same explanatory variables as are used in the machine learning model (13 traditional and 7 non-traditional variables). These explanatory variables were selected from more than 300 series, using a data selection process based on their highest predictive power.[3]

It is worth emphasizing that in the "horse race" between the three models, the comparison is not completely one-for-one. In a sense, the fintech credit score (Model I) is tested "out of sample", while Models II and III are estimated "in sample". So this would in principle produced a bias against Model I. On the other hand, Model I uses more data for training than the data used in Models II and III, so that may be one reason for its better performance. In other words, in Models II and III we use the same set of data selected to be used for the machine learning model, under the assumption that they would be also the best ones for the linear models. We will address some of these points in the robustness check section.

4.2 Results

Table 2 presents the results of equations (1) to (3) that consider the three different information sets. The model in the first column uses only the fintech score as the

[3] For instance, the popular lasso belongs to this class of estimators that produce sparse representations of predictive models (see Belloni et al (2011) for a recent survey and examples of big data applications of these methodologies in economics). By contrast, Giannone et al (2018) point to the need to use dense-modelling techniques that recognise that all possible explanatory variables might be important for prediction, although their individual impact might be small.

independent variable (Model I), while the model in the second column provides the result using the traditional credit card information as independent variables (Model II), and the third column provides the result using all variables (Model III). All models are estimated using a Tobit regression model. The fintech score is a highly significant predictor of the loss rate. The credit card/bank and non-traditional information are also useful. However, the pseudo R^2 of Model I (0.0367) is almost double that of Model III (0.0217).

Loss rate regressions — Table 2

	Loss rate		
	I	II	III
Variables	Fintech credit score	Traditional information only	All information
Fintech credit score	–0.00845***		
	(8.30e–05)		
Traditional information			
Number of bank accounts (12 months)		–0.00198**	–0.00195*
		(0.00100)	(0.00100)
Frequency of usage (12 months)		–2.14e–05	–0.000119
		(0.000150)	(0.000150)
Frequency of usage (3 months)		–0.000756	–0.000671
		(0.000476)	(0.000474)
Large payment count		–0.00126***	–0.00100***
		(5.87e–05)	(5.80e–05)
Credit line		–2.97e–07***	–1.52e–07**
		(6.78e–08)	(6.69e–08)
Defaults (12 months)		0.0121***	0.0159***
		(0.00197)	(0.00196)
Defaults (3 month)		0.00978	0.0117*
		(0.00674)	(0.00669)
Repayments (12 months)		–3.60e–07	–4.32e–07
		(7.71e–07)	(7.96e–07)
Number of bank accounts (3 months)		0.0140***	0.0155***
		(0.00347)	(0.00351)
Credit history		–0.00624***	–0.00600***
		(0.000149)	(0.000150)
Salary in debit card		–4.55e–06***	–4.64e–06***
		(5.13e–07)	(5.21e–07)
Non-traditional information[1]			
Max call duration			–3.55e–10***
			(1.16e–10)
Call times to family			2.97e–06
			(4.70e–06)
Frequency of calls (daily)			–0.0191***
			(0.000978)
Taobao payments (daily)			–1.69e–05***
			(1.64e–06)
Observations	310,919	310,919	310,919
Pseudo R^2	0.0367	0.0169	0.0217

[1] The model with non-traditional information also includes the number of defaults on the Taobao platform, the number of calls in the last three months and gender. All models include monthly and province fixed effects.

Table 3 has a similar structure to Table 2, but presents the estimates of equations (1') to (3'), ie the probability that a customer will default. All the models in Table 3 are estimated using a logistic regression model. Consistent with Table 2, Model I – which only uses the fintech credit score – has the highest pseudo R^2, 0.0399. Model II has a pseudo R^2 of 0.0180, while model III has a pseudo R^2 of 0.0231.

Default rate regressions

Table 3

	Loss rate		
	I	II	III
Variables	Fintech credit score	Traditional information only	All information
Fintech credit score	–0.0178***		
	(0.000179)		
Traditional information			
Number of bank accounts (12 months)		0.00111	0.000834
		(0.00209)	(0.00211)
Frequency of usage (12 months)		0.000121	–6.65e–05
		(0.000310)	(0.000313)
Frequency of usage (3 months)		–0.00206**	–0.00195*
		(0.00101)	(0.00102)
Large payment count		–0.00283***	–0.00225***
		(0.000133)	(0.000131)
Credit line		–7.97e–07***	–5.09e–07***
		(1.47e–07)	(1.46e–07)
Defaults (12 months)		0.0236***	0.0309***
		(0.00406)	(0.00405)
Defaults (3 month)		0.0209	0.0252*
		(0.0138)	(0.0138)
Repayments (12 months)		–6.20e–07	–7.20e–07
		(1.62e–06)	(1.69e–06)
Number of bank accounts (3 months)		0.0371***	0.0409***
		(0.00722)	(0.00734)
Credit history		–0.0123***	–0.0118***
		(0.000312)	(0.000316)
Salary in debit card		–1.06e–05***	–1.12e–05***
		(1.25e–06)	(1.29e–06)
Non-traditional information[1]			
Max call duration			–7.76e–10***
			(2.88e–10)
Call times to family			1.07e–05
			(1.01e–05)
Frequency of calls (daily)			–0.0397***
			(0.00212)
Taobao payments (daily)			–3.67e–05***
			(3.70e–06)
Observations	310,910	310,910	310,910
Pseudo R^2	0.0399	0.0180	0.0231

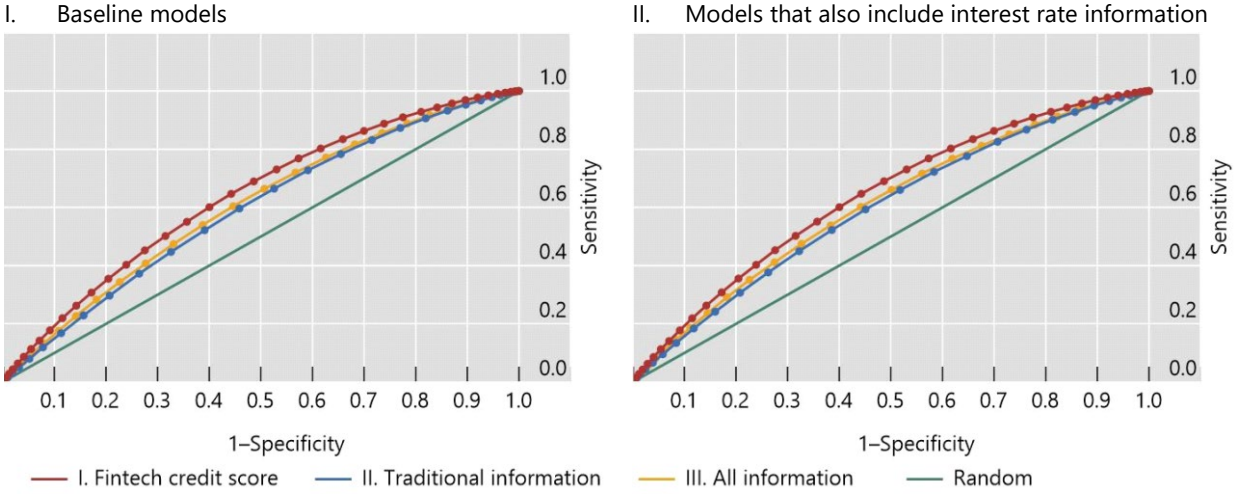
[1] The model with non-traditional information also includes the number of defaults on the Taobao platform, the number of calls in the last three months and gender. All models include monthly and province fixed effects.

Figure 2 and Table 4 present a comparison between the three models with different information sets. Figure 2 shows the receiver operating characteristics (ROC) curve of each model. The ROC curve is created by plotting the true positive rate (TPR)

against the false positive rate (FPR) at various threshold settings. The TPR is also known as sensitivity. The FPR is also known as the fall-out or probability of false alarm and can be calculated as (1 − specificity). Table 4 reports the area under the ROC curve (AUROC) for every model. The AUROC is a widely used metric for judging the discriminatory power of credit scores. The AUROC ranges from 50% (purely random prediction) to 100% (perfect prediction).

ROC curves for different models

Figure 2



Source: Authors' calculations.

A horse race between the three different models

Table 4

Panel I. Baseline Models

	AUROC	Std err	95% conf. interval	
I. Fintech credit score	0.6391	0.0012	0.63686	0.64143
II. Traditional information	0.5939	0.0012	0.59149	0.59621
III. All information	0.607	0.0012	0.60462	0.60932

Panel II. Models that also include interest rate information

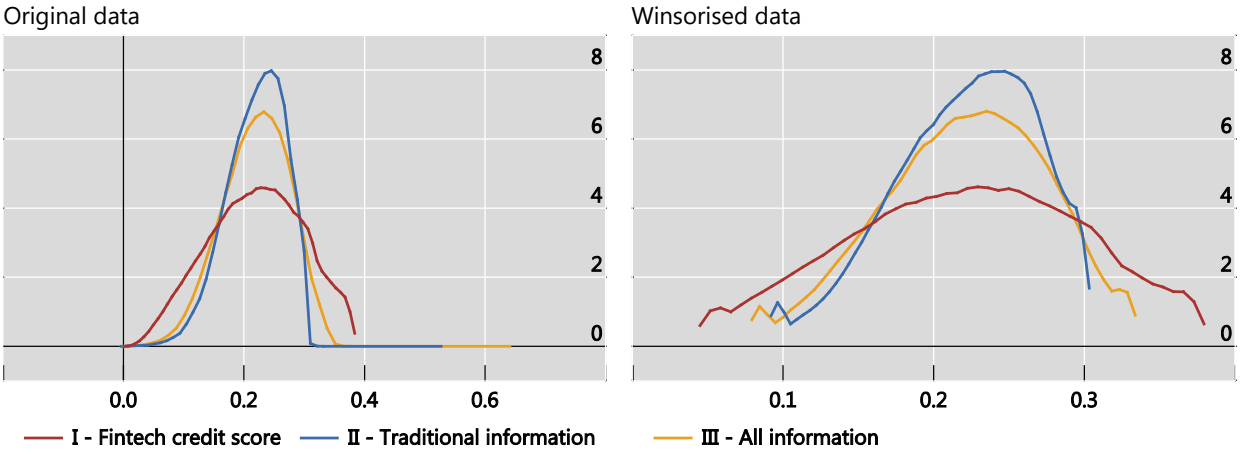
	AUROC	Std err	95% conf. interval	
I. Fintech credit score	0.6391	0.0012	0.63686	0.64144
II. Traditional information	0.5971	0.0012	0.59477	0.59951
III. All information	0.6095	0.0012	0.60712	0.61183

The left-hand panel of Figure 2 reports the results for the three different models. The results show that Model I performs better than the other two models. Model III is the second-best. Model II performs worst. This means that the model based on the fintech credit score (Model I) is better than the traditional model that use bank-type information (Model II) in predicting default rates for this sample of borrowers. But Model I, which uses machine learning techniques, is also superior to logit regressions that use also non-traditional information (Model III). The better performance of the fintech company in predicting defaults could depend on both: (i) specific selection of the variables that better fit Model I than Model II or III; and (ii) the use of machine

learning techniques that are able to capture relevant non-linearities among the variables. The three models are statistically different at the 5% level, as verified by the test at the bottom of Table 4. In terms of contribution of non-traditional data and machine learning to predictive power, non-traditional data contribute an additional 2.2% of the AUROC (=(0.607-0.5939)/0.5939), while applying machine learning techniques provides an additional 5.3% of the AUROC (=(0.6391-0.607)/0.607).

We conducted two additional tests with a view to shedding further light on this result. First, we considered the distribution for the expected default rate for the three different models over both the whole original sample and the winsorised sample. The results reported in Figure 3 indicate that Model I has a greater discriminatory power than Model II and III (ie the expected default rates encompass a larger set of plausible data).

Distribution of the expected default rate for the three different models — Figure 3



Note: The graphs show that the expected default rate for the whole sample is more dispersed for the fintech credit score model. This implies that Model I better captures the heterogeneity among borrowers.

Source: Authors' calculations.

Second, we performed similar tests using the information content of the interest rates. In particular, as the interest rate is highly correlated with the fintech score, we have included the residual of a regression of the interest rate on the credit score in Models II and III. The test aims to control for the fact that the interest rate could take into account additional information not included in the list of explanatory variables but that the fintech company can use in its assessment. As can be seen from the second panel of Figure 2 and the second panel of Table 4, the results are qualitatively very similar.

5. Performance of the models in the event of an (exogenous) change in regulation

In this section, we want to test model performance in the event of an exogenous change in regulation. The current debate highlights one possible problem for machine

learning models.[4] In particular, some of the literature stresses that the machine learning technology could only be useful in situations where the relationship between inputs and outputs remains the same – but this is often not the case in financial applications.

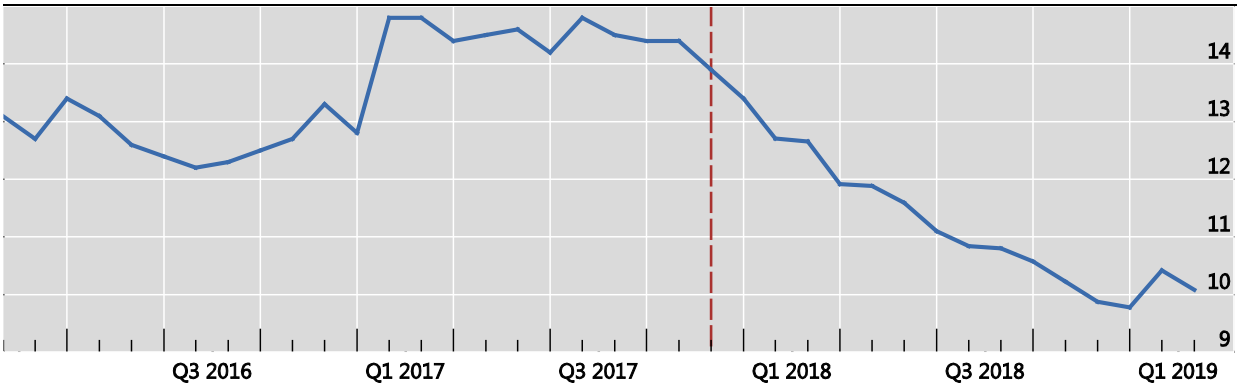
So far, machine learning, especially supervised learning, has been applied successfully in applications where there are stationary patterns. For example, when a CT scan is performed, we know that well trained doctors will make the same diagnosis every time they see a certain pattern in the scan. Application to financial data does not respond to this situation of stable correspondence. In credit scoring, for example, the relationship between the characteristics of the borrowers and whether they defaulted or are delinquent might not be stable at all. So the performance of machine learning models in stress situations remains to be fully explored.

In this section, we analyse the impact of a regulatory change in China on the performance of credit scoring models. On 17 November 2017, the PBoC issued draft guidelines to tighten regulations on financial institutions' asset management activities, a key component of the country's growing shadow banking sector. The main aim of the new rules, which affected $15 trillion of asset management products, was to unify regulatory practices across the financial industry. These changes were largely unexpected and caused a significant impact on fintech firms' business models. In particular, starting 17 November 2017, financial institutions have not been allowed to use asset management products to invest in commercial banks' credit assets or provide "funding services" for other institutions (such as fintech companies) to bypass regulations. The new rule had a huge impact on fintech companies' funding sources. The PBoC also set a limit on the interest rates charged by P2P lending companies. All annualised interest rates, which include the upfront fees charged for loans, were capped at 36%. The effects of these new rules were reinforced on 1 December 2017 when China's *Internet Financial Risk Special Rectification Work Leadership Team Office* rolled out strict measures concerning online micro-lending.

Total credit to the Chinese economy (yearly credit growth)

In per cent

Figure 4



The vertical dashed line indicates 17 November 2017, when the PBoC issued specific draft guidelines to tighten regulations on shadow banking.

Source: The People's Bank of China.

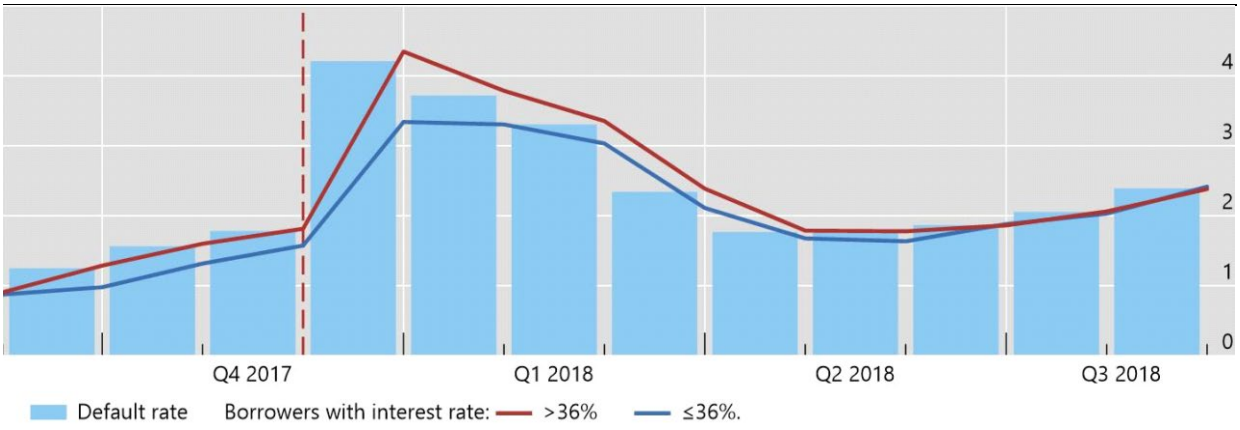
4 https://www.risk.net/risk-management/4120236/academics-warn-against-overuse-of-machine-learning#cxrecs_s

The impact of the regulatory changes was to reduce loan supply, especially to riskier borrowers. Figure 4 shows that the pace of growth in total credit in the Chinese economy fell by 4 percentage points in less than one year after these regulatory changes. Moreover, the sudden freeze on rolling over credit lines to risky borrowers caused many sole proprietorships to default. The histograms in Figure 5 show that the default rate of the loans supplied by the fintech company analysed in this study increased by 3 percentage points at the end of December 2017, and then decreased smoothly to pre-shock levels after one year. Some of the borrowers might also have defaulted strategically, especially those with interest rates in excess of 36% whose credit could not be rolled over at the same conditions. However, the lines in Figure 5 indicate that the default rate of borrowers with an interest rate in excess of 36% – despite being structurally higher because of the higher associated risk – followed a similar pattern to the default rate of the other borrowers.

Default rate for the fintech company

In per cent

Figure 5



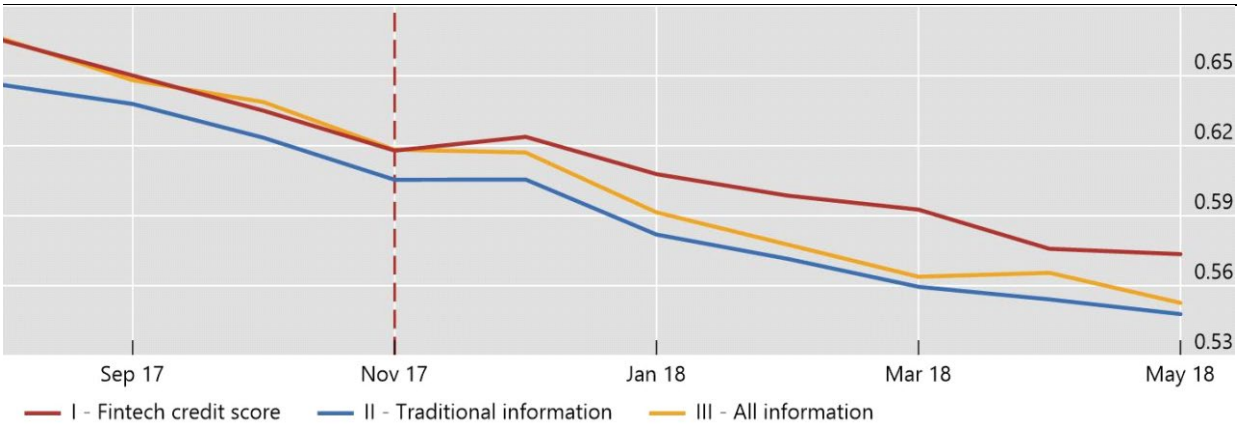
The vertical dashed line indicates 17 November 2017, when the PBoC issued specific draft guidelines to tighten regulations on shadow banking. Among these new rules, the PBoC set also a limit on the interest rates charged by P2P lending companies. All annualised interest rates, which include the upfront fees charged for loans, were capped at 36%. This figure wants to analyse if those borrowers with credit contracts with an interest rate greater than 36% reacted strategically and defaulted by more with respect to the others.

Source: Authors' calculations based on an anonymous Chinese fintech company.

We use the loan repayment records to calculate the discriminatory powers of the three models (Model I: fintech score; Model II: traditional information; Model III: all information) in every month. The discriminatory power is measured by the area under the ROC curve (AUROC). Figure 6 shows the results. The vertical line represents the date of the regulatory shock. We find that Model I and Model III perform better than Model II before the regulatory change and the difference between Model I and Model III is not statistically significant. We find that, after the regulatory shock, the discriminatory powers of all three models decline. However, Model I performs better than Model II and Model III in relative terms.[5]

[5] We needed to consider how to include the time-fixed effects in the analysis of the explanatory power of the three different models in response to the exogenous regulatory shock. The results reported from Figure 6 onwards include the average effects for the time dummies. Another possibility would have been to estimate the models up to month t and then make a prediction for month t+1 under the assumption that the month fixed effect in month t +1 is identical to the one in month t. The results obtained using this second assumption are very similar and not reported for the sake of brevity.

Discriminatory powers of the models before and after the regulatory shock Figure 6



The vertical axis reports the area under the ROC curve (AUROC) for every model. The AUROC is a widely used metric for judging the discriminatory power of credit scores. The AUROC ranges from 50% (purely random prediction) to 100% (perfect prediction). The vertical dashed line indicates the date of the shock. In particular, it refers to a largely unexpected regulatory change that occurred in China on 17 November 2017, when the PBoC issued specific draft guidelines to tighten regulations on shadow banking. This regulatory policy has led many financial intermediaries to increase their lending requirements, resulting in deteriorating credit conditions for borrowers.

Source: Authors' calculations.

Figure 7 shows the gap between the discriminatory powers of Model I (based on the credit scoring obtained using machine learning with big data) and Model II (traditional bank model). We decompose the gap into two parts. The first part (light blue) is the value added provided by non-traditional information (the gap between Model II and Model III). The second part (dark blue) is the gain obtained from machine learning technology (the gap between Model I and Model III).[6] Based on this graph, non-traditional information represents the main reason why Model I performs better than Model II prior to the shock. The contribution of machine learning technology is particularly relevant after the shock. This result could be due to the fact that machine learning technology can mine richer information from the variables during a period of stress. This may be due to the non-linearity of the model, which better captures dynamic relationships after the regulatory shock.

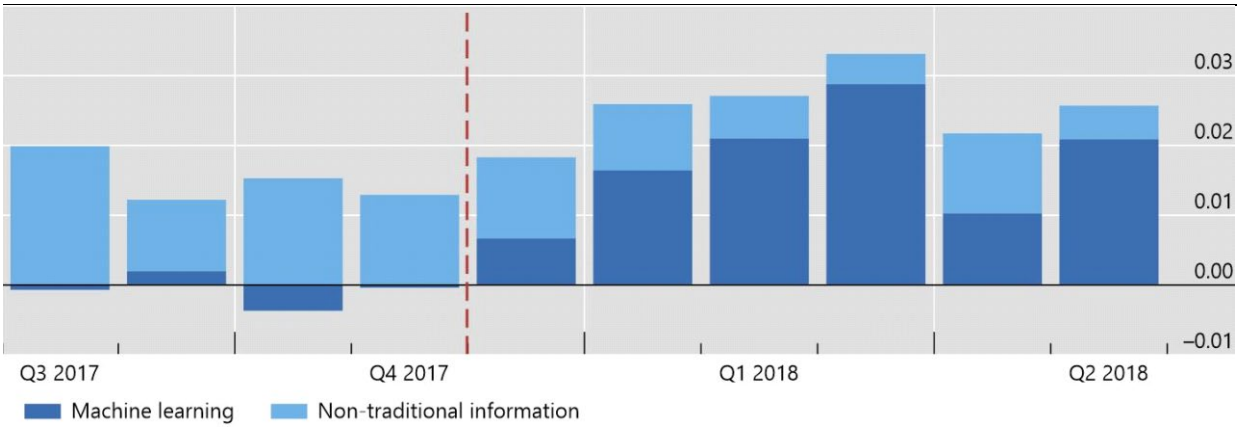
One concern could be that the above results only hold for the period of estimation (May 2017 – September 2017) and that different results could be obtained by estimating the logit models for a different period. As a robustness test, we therefore estimate the coefficient of the logit models for the period from January 2015 to December 2016 using a random sample of 10,000 customers of the fintech company. We then apply these coefficients to the explanatory variables of the borrowers in the period from May 2017 to September 2017 to verify any possible changes. The results reported in Figure 8 indicate that, in relative terms, Model I performs better than Model III even in "normal" times, but the difference between the two models widens significantly after the regulatory shock.

[6] The selection of the parameters to be used in the machine learning algorithm requires not only a knowledge of technical aspects but also experience in the selection of the appropriate weights and variables. In doing so, the technology officers may use also their own experience (soft information) in the evaluation. This means that this gap captures not only technological aspects but also experience (which is not easily replicable).

The contribution of machine learning and non-traditional information

In per cent

Figure 7

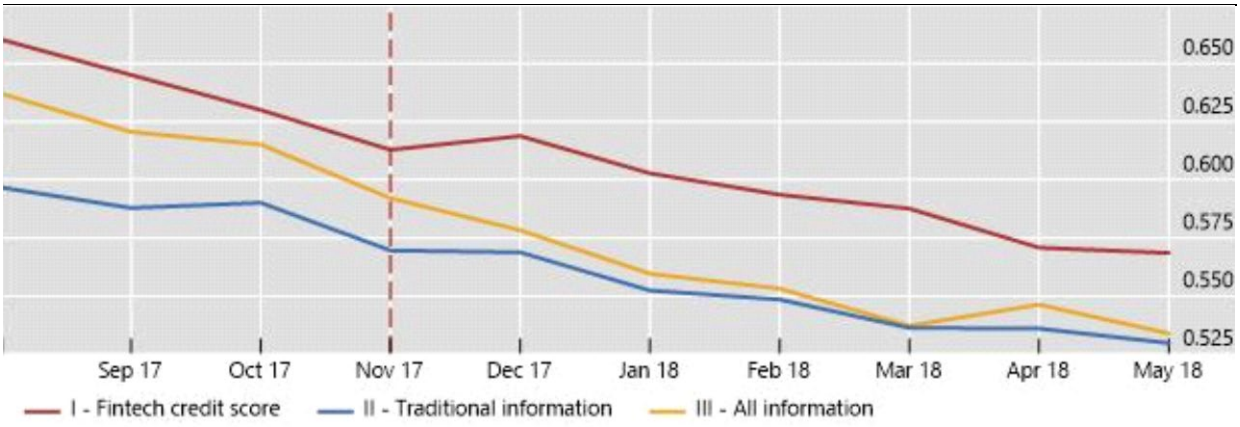


The vertical dashed line indicates 17 November 2017, when the PBoC issued specific draft guidelines to tighten regulations on shadow banking.

Source: Authors' calculations.

Robustness check using a different estimation period for the logit models

Figure 8



The vertical axis reports the area under the ROC curve (AUROC) for every model. The AUROC is a widely used metric for judging the discriminatory power of credit scores. The AUROC ranges from 50% (purely random prediction) to 100% (perfect prediction). The vertical dashed line indicates the date of the shock. In particular, it refers to a largely unexpected regulatory change that occurred in China on 17 November 2017, when the PBoC issued specific draft guidelines to tighten regulations on shadow banking. This regulatory policy has led many financial intermediaries to increase their lending requirements, resulting in deteriorating credit conditions for borrowers.

Source: Authors' calculations.

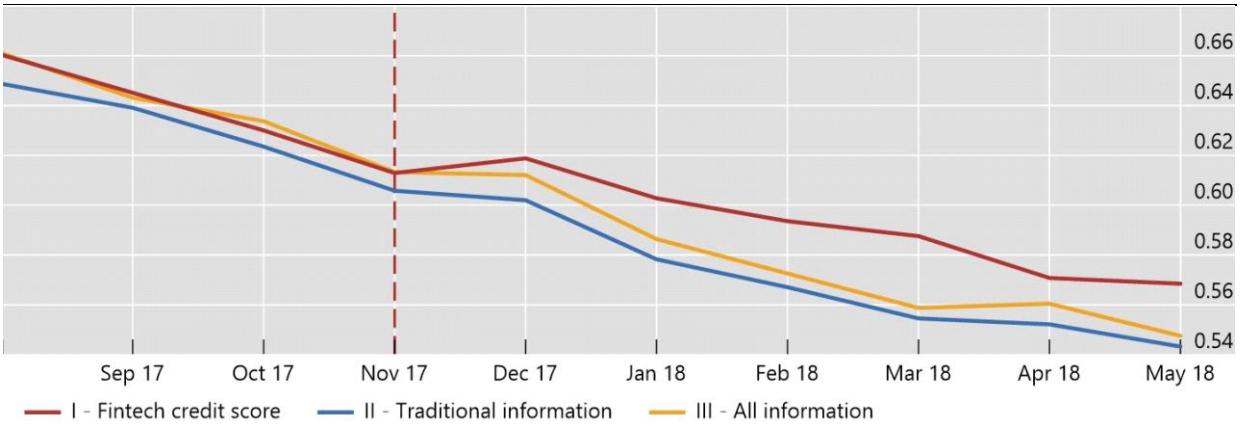
Another possible concern regards the role of the gender variable in evaluating the differences between Model II and Model III. In the results presented so far, we have considered the gender variable in the non-traditional information. Banks may not include this variable in the set of traditional information out of concern for discrimination issues.[7] However, a borrower's gender is easily detectable and could be highly informative. We have therefore re-run the model by including the gender variable in traditional information. The results reported in Figure 9 indicate that, also in this case, the model based on machine learning is better able to predict losses and

[7] Similarly the US Fair Housing Act (FHA) and Equal Credit Opportunity Act (ECOA) prohibit discrimination based on race, national origin, sex or religion.

defaults after the regulatory shock. However, the difference between Model II and Model III becomes less evident.

Robustness check including gender among traditional information

Figure 9

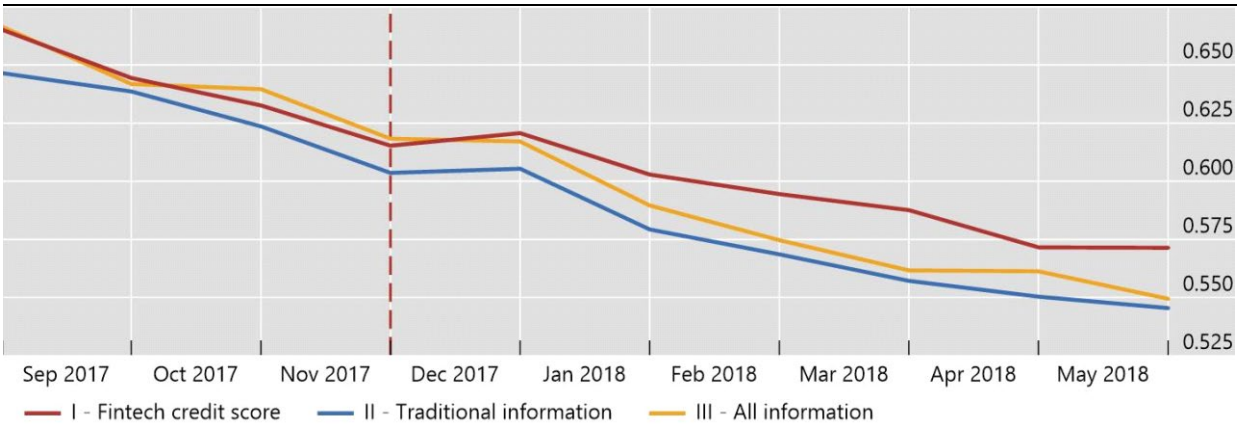


The vertical axis reports the area under the ROC curve (AUROC) for every model. The AUROC is a widely used metric for judging the discriminatory power of credit scores. The AUROC ranges from 50% (purely random prediction) to 100% (perfect prediction). The vertical dashed line indicates the date of the shock. In particular, it refers to a largely unexpected regulatory change that occurred in China on 17 November 2017, when the PBoC issued specific draft guidelines to tighten regulations on shadow banking. This regulatory policy has led many financial intermediaries to increase their lending requirements, resulting in deteriorating credit conditions for borrowers.

Source: Authors' calculations based on an anonymous fintech company data.

Robustness check: loans up to one year

Figure 10



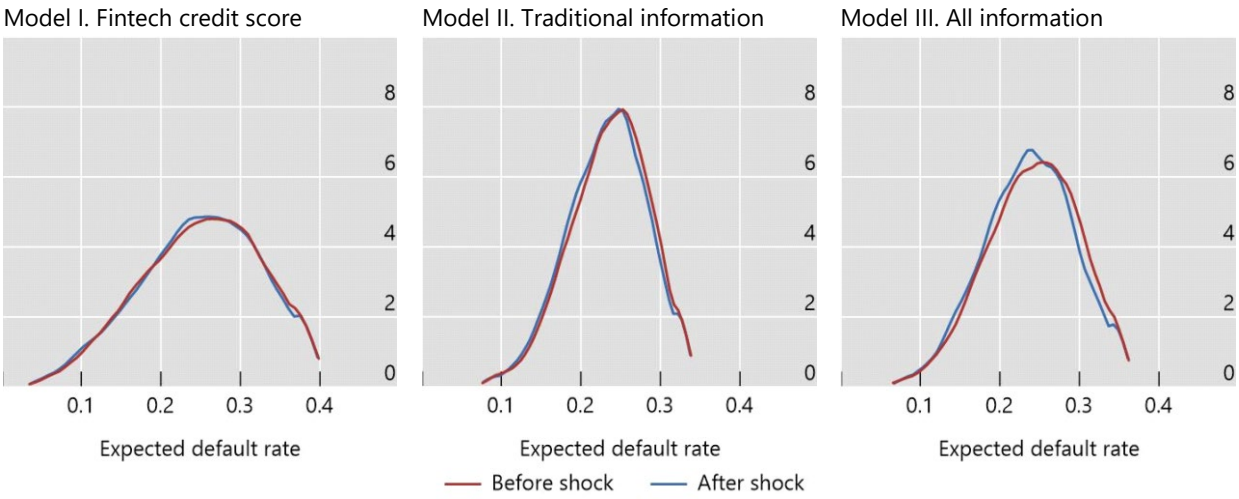
See note in Figure 9.

Source: Authors' calculations based on an anonymous fintech company data.

We also carried out some checks for loans with differing maturities. In particular, 80.3% of the loans have a maturity up to 12 months, 5.4% have a maturity between 12 and 18 months, and 14.3% have a maturity between 18 and 24 months. As we evaluate loans originated in the period between May 2017 and September 2017, until October 2018 we are only able to observe the whole life cycle of credit maturing in up to one year. For this reason, we conducted the same structural break analysis only for those loans with a maturity of up to one year. These loans are observed over their whole life. (The last credit line extended in September 2017 expires in October 2018.) The results of this test are reported in Figure 10 and are qualitatively very similar. This is also in line with the statistical observation that most of the defaults take place in the first months of a loan contract.

The distribution of expected default rate before and after shock

Figure 11



Note: These graphs compare two-month windows of defaults before and after the shock for the three different models. The comparison makes it possible to see whether the distribution of expected default rates changes after the shock. If the distribution moves to the left, this means that the power of the model has reduced. Models tend to underestimate the probability of default after the shock.

Source: Authors' calculations.

Based on the above checks, the result that the model based on machine learning is better able to predict losses and defaults after the regulatory shock seems quite robust. We look at this aspect by considering the loans that defaulted in the two-month window before and after the shock for the three different models. In particular, Figure 11 plots the distribution of the three models' expected default rate before and after the regulatory shock. The sample prior to the shock includes the months of October and November 2017, while that after the shock covers December 2017 and January 2018. If the shock does not affect the model's predictive power, the distributions of the two samples should not be significantly different. Figure 11 shows that the distributions based on the fintech credit model are qualitatively very similar, while those based on Model II and Model III shift to the left after the shock. This means that, prior to the shock, these models were too optimistic regarding customers' capacity to repay their loans. A more precise evaluation is reported in Table 5 using the results of a quantile regression. The dependent variable is the expected default rate over the four months from October 2017 to January 2018. The right-hand side includes a dummy variable that takes the value of one in the post-shock period. For Models II and III, the dummy has a negative value for all quantiles, indicating that people who defaulted after the shock have a lower expected default rate based on

Quantile regression before and after shock

Table 5

Variables	Fintech credit score			Traditional information			All information		
	q25	q50	q75	q25	q50	q75	q25	q50	q75
After shock	7.50e–05	–0.00172	–0.00145	–0.004***	–0.004***	–0.004***	–0.005***	–0.006***	–0.007***
	(0.00160)	(0.00114)	(0.00118)	(0.00086)	(0.00076)	(0.00072)	(0.00134)	(0.00086)	(0.00098)
Constant	0.197***	0.255***	0.307***	0.203***	0.240***	0.272***	0.202***	0.245***	0.285***
	(0.00117)	(0.00083)	(0.00098)	(0.00074)	(0.00075)	(0.00076)	(0.00134)	(0.00086)	(0.00102)
Observations	30,216	30,216	30,216	30,216	30,216	30,216	30,216	30,216	30,216

traditional models. In other words, those people who have a higher evaluation based on the traditional model defaulted. This effect is not significant in the fintech credit model, indicating more stability.

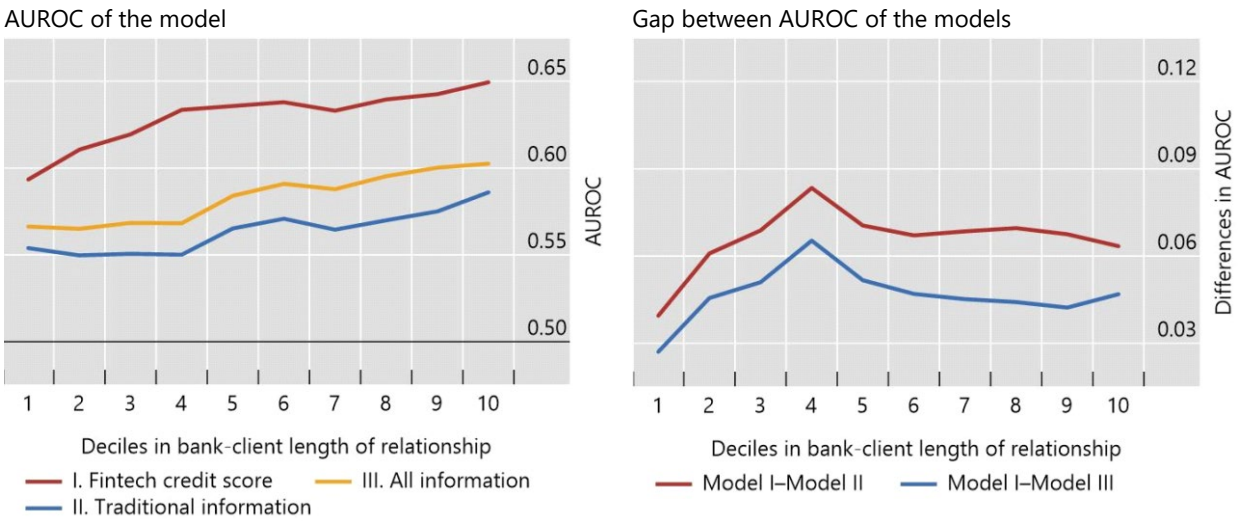
6. Credit scoring and relationship lending

The provision of financial services (lending, insurance, wealth management etc) traditionally relies on trust and human interaction – they are relationship-based. By contrast, fintech lending is transaction-based and does not involve human intervention or a long-term relationship with the customer.

The loans offered by fintech lenders are strictly transactional, typically short-term credit lines that can be automatically cut if a customer's condition deteriorates. It is therefore interesting to study how the model's performance evolves for customers with different credit histories.

Figure 12 highlights how the comparative advantage of the model that uses the credit scoring technique based on machine learning and big data could be modified by the length of the relationship between bank and customer. Please note that we use the length of the relationship between borrower and bank to calculate the borrower's credit history. This is because borrowers typically enter into a credit relationship with a bank first. In particular, we use the number of months from the opening of the bank account as a proxy of the bank-customer relationship. We have divided the sample into ten deciles according to the length of the relationship and calculate the predictive powers (level and gap) of the three models for the ten different buckets. We find that the performance of the three models – measured by the AUROC – improves with the length of the relationship (see Figure 12, left-hand panel). On the right-hand side of Figure 12, we compare the predictive power of the model based on the fintech score (Model I) with the model that considers only traditional information (Model II) and the model that includes all information (Model III). Interestingly, the comparative advantage of Model I over Models II and III tends to increase for low levels of the

Predictive power of the models and length of the bank-customer relationship — Figure 12



Source: Authors' calculations.

bank-customer relationship. However, when the relationship becomes stronger, the differences between Model I and the other two models decrease. This tallies with the idea that a longer relationship between bank and customer tends to attenuate asymmetric information problems. This is also reflected in the relationship between borrower and fintech company.

7. Conclusion

The main goal of this paper is to compare the predictive power of credit scoring models based on machine learning techniques and big data with that of traditional loss and default models. Using a unique data set at the transaction level from a leading fintech company in China, we test the performance of different models to predict losses and defaults both in normal times and when the economy is hit by a shock. In particular, we analyse the case of an (exogenous) change in regulatory policy on shadow banking in China that caused lending to contract and credit conditions to deteriorate.

We find that the model based on machine learning and big data is better able to predict losses and defaults than traditional models in the event of a negative shock to the aggregate credit supply. One possible reason for this is that machine learning can better exploit the non-linear relationship between variables in a period of stress. By analysing different types of data, we find that non-traditional information, obtained from mobile phone applications and e-commerce platforms, has high predictive value. Finally, the comparative advantage of the model that uses the fintech credit scoring technique based on machine learning and big data tends to decline for those borrowers with a longer credit history.

References

Bank for International Settlements (2019): "Big tech in finance: opportunities and risks", *BIS Annual Economic Report*, June, 55–79.

Belloni, A, V Chernozhukov and C Hansen (2011): "Inference for high-dimensional sparse econometric models," in *Advances in Economics and Econometrics*, 10th World Congress of the Econometric Society.

Berg, T, V Burg, A Gombović and M Puri (2018): "On the rise of fintechs – credit scoring using digital footprints", *NBER Working Papers*, 24551, April.

Braggion, F, Manconi A, Z Haikun (2019): "Can technology undermine macroprudential regulation? Evidence from online marketplace credit in China", mimeo.

Buchak, G, G Matvos, T Piskorski and A Seru (2018): "Fintech, regulatory arbitrage, and the rise of shadow banks", Journal of Financial Economics, 130(3), 453-483.

De Roure, C, L Pelizzon and P Tasca (2016): "How does P2P lending fit into the consumer credit market?", *Deutsche Bundesbank Discussion Papers*, 30.

Dorfleitner, G, C Priberny, S Schuster, J Stoiber, M Weber, I de Castro and J Kammler (2016): "Description-text related soft information in peer-to-peer lending – Evidence from two leading European platforms", *Journal of Banking and Finance*, 64, March, 169–-87.

Frost, J, L Gambacorta, Y Huang, H S Shin and P Zbinden (2019): "BigTech and the changing structure of financial intermediation", *Economic Policy*, forthcoming (also published as BIS Working Papers, 779).

Fuster, A, M Plosser, P Schnabel and J Vickery (2018): "The role of technology in mortgage lending", *Federal Reserve Bank of New York Staff Reports*, 836, February.

Fuster, A, P Goldsmith-Pinkham, T Ramadorai and A Walther (2019): "The effect of machine learning on credit markets", VoxEU, 11 January 2019.

Giannone D, M Lenza and G Primiceri (2018): "Economic predictions with big data: The illusion of sparsity", *Federal Reserve Bank of New York Staff Reports*, 847, May.

Jagtiani, J and C Lemieux (2017): "Fintech lending: Financial inclusion, risk pricing, and alternative information", mimeo.

Jagtiani, J and C Lemieux (2018a): "The roles of alternative data and machine learning in fintech lending: evidence from the LendingClub consumer platform", *Federal Reserve Bank of Philadelphia Working Papers*, 18–15, April.

Jagtiani, J and C Lemieux (2018b): "Do fintech lenders penetrate areas that are underserved by traditional banks", *Federal Reserve Bank of Philadelphia Working Papers*, 18–13, March.

Jagtiani J, L Lambie-Hanson and T Lambie-Hanson (2019): "Fintech lending and mortgage credit access", *Federal Reserve Bank of Philadelphia Working Papers*, 19–47, November.

Khandani, A, A Kim and A Lo (2010): "Consumer credit-risk models via machine-learning algorithms", *Journal of Banking & Finance*, 4(11), 2767–87.

Tang, H (2019): "Peer-to-peer lenders versus banks: substitutes or complements?", *Review of Financial Studies*, forthcoming.

US Department of the Treasury (2016): "Opportunities and challenges in online marketplace lending", May.