The Role of Fintech in Loan Origination: Evidence from China

Yao Zhang[1], Fan Zhao[2]*
[1] Institute of International Economy, University of International Business and Economics, Beijing, China
[2] School of Foreign Studies, Zhongnan University of Economics and Law, Wuhan, China
*Corresponding author, e-mail: 181564413@qq.com

***Abstract**: Using loan-level data on a large Fintech firm in China, we investigate whether unstructured data regarding consumers' digital mobile footprint and other soft information such as abundance of loan description, the type of mobile operating system, number of calls, SMSs and contacts proxy for social media connections, numbers of Sales apps, Financial Apps and Travel apps proxy for individual behavior etc., can complement rather than substitute for traditional hard information.*

***Keywords**: Fintech; Loan origination; Digital footprint; Soft information*

Introduction

Information is an essential component in financial markets. Given the importance of information to alleviate moral hazard and adverse selection issues, as the growth of the internet and technology changes the way of information production transmission and collection, it also changes the way to process the loan origination in each step (Fuster et al., 2019; Berg et al., 2019). In addition, there are hundreds of millions of young people who have never obtained a bank loan. They are internet users preferring shop online and even obtaining loan and investing online. Thus, new ways are created to serve these consumers by employing unstructured data and big data techniques to predict their lending behavior (Agarwal et al., 2019). However, there is limited evidence on whether soft information collected by fintech firms can be conducive to loan origination.

This paper uses manually collected data from one of the largest Fintech lending firm in China to examine whether new soft information obtained by Fintech firms can complement traditional hard information in predicting loan outcomes. Firstly, we specify a Probit Model 1 of loan outcomes on all hard information indicators. We then regress the credit scores from the fintech platform on all variables of hard information to obtain consumers' digital mobile footprint and other soft information (as measured by a first-stage residual item). Secondly, we include this residual item as an additionally explanatory variable besides hard information indicators in a Model 2 analyzing loan outcomes. Finally, to discuss implication of our finding for the effect of abundance of loan description, we add this explanatory variable in Model 2 to construct a Model 3. Constructing receiver operating characteristics (ROC) and determining the area under the curve (AUC) is a popular method to judge the predictive power of three models (Iyer et al. ,2016), which in other words, is a formal analysis of the role of consumers' digital mobile footprint and other soft information. Comparing these models, we predict the values of AUC gradually increase when adding extra variables of soft information.

Vol.1 No.11 2020
https://doi.org/10.37420/j.cer.2020.105

Thus, two predictions are expected as follows:

Prediction I.—Digital mobile footprint will be a significantly complementary information for predicting loan outcomes.
Prediction II.—The abundance of loan description also complements for traditional hard information in predicting loan outcomes.

Data and Methodology

We use Python to obtain data on about 283,331 loan applicants between November 2013 and June 2018 from a mobile-only Fintech lending platform named Renrendai operating in China since 2010. The dependent variable is whether the loan is grated or not (State), the failed loan is marked as 0, and the successful loan is marked as 1. The explanatory variables selected in this paper include three categories. First, hard information indicators including loans characteristic variables and borrower characteristic variables are accessible openly from the official website. The loans characteristics indicators are as follows: Interest Rate, which can be adjusted by Renrendai platform within the scope of the upper and lower limits of Interest Rate under the control of the national government; Lending Amount, the loan amount to be raised; Lending Term, borrowing duration of the loan. Borrower characteristics comprise the following variables: (1)Gender, 0 for female and 1 for male; (2) the Age of the borrower; (3)Marriage, unmarried state is labeled as 0, married, divorced or widowed state is labeled as 1; (4) Education, the borrower's education is classified into four categories, high school or below, junior college, undergraduate, graduate and above, assigned to the value of 1--4 successively; (5) Monthly Income, the borrower's monthly income is divided into five categories, i.e., below 2,000 yuan, 2000-5,000 yuan, 5,000-10,000 yuan, 10,000-20,000 yuan, and above 20,000 yuan. (6) Work Experience. The working years of the borrowers are grouped into four categories: less than 1 year, 1-3 years, 3-5 years and more than 5 years. (7) Borrower Type, according to the type of the borrower's job, loans can be divided into salary, online business, private business owner, successively assigned to the value of 1 to 3; (8) Housing, the borrower has the property to take 1, otherwise the value is 0; (9) House Mortgage (House_D), the mortgage has been paid off gets 1, otherwise gets 0; (10) Car, the borrower owns cars takes 1, otherwise the assignment is 0; (11) Car Mortgage (Car_D), the car mortgage has been cleared takes 1, otherwise the value is 0. Year fixed effect (Year) and Region (Provincial) fixed effect are also added.

Second, Berg et al. (2019) document that the digital footprint is a trace of simple, easily accessible information collected to predict consumer payment behavior and defaults even for customers who do not have credit bureau scores and other verifiable hard information. Other than writing text, uploading financial information, or social network data, the mobile action of accessing or registering on an APP leaves behind valuable information, which is practical basis for the fintech firm to grant loans. According to an official announcement on evaluation (auditing) mechanism of the platform’s credit scores, the digital footprint information comprises the type of mobile operating system, number of calls, Short Messaging Service (SMSs) and contacts proxy for social media connections, numbers of Sales apps, Financial Apps and Travel apps proxy for individual behavior etc. They are soft information newly available to proxy for the economic status and characteristic variable of a borrower. However, the digital footprint information is not disclosed publicly, so we construct the variable through a regression (Formula 1). Running the credit scores on all variables

of hard information, the influences of all hard factors on credit score can be removed. Thus, the residual item (*Footprint*) indicates consumers' digital mobile footprint and other soft information that is exclusively leveraged by the fintech firm to predict borrower's credit score.

$$Score_i=Hard\_Info\_colle_i\times a+e_i+\varepsilon_i \quad (1)$$

where subscript *i* identifies a unique customer. *Hard_Info_colle* is all hard information indicators (borrower characteristic variables). *e* is Province fixed effect. ε is the robust error.

Third, loan description is also an unstructured information written by borrowers on official website, very relevant to information disclosure to investors, is highlighted in examining the impact on anticipating lending behaviors (Gao, Lin, and Sias, 2018). The abundance of loan description (measured by String Length of the description) is excluded in forming the Credit Scores in terms of the announcement. Therefore, it is extra soft information variable to further demonstrate the role of soft information newly derived from fintech firm.

To conduct empirical research, we construct three multivariate analysis, and rely on both the economic and statistical significance of individual explanatory variables as well as AUC, a commonly used measure of the predictive power of credit scores. Firstly, we run a Probit regression of loan outcome on hard information variable:

$$Loan\ Outcome_i=\beta_1\times Hard\_Info_i+e_i+\varepsilon_i \quad (2)$$

where the *Loan outcome* is a dummy variable setting as the following: approved loan takes the value one and zero otherwise. *Hard_Info* is a predicted value of *Score* in Formula (1), the reason for constructing the gross hard information variable is to eliminate collinearity between the variables of interest and the control variable (all hard information indicators).

Focusing on the role of digital footprint, we add Footprint (substituted by the residual item in Formula (1) into Formula (2):

$$Loan\ Outcome_i=\beta_1\times Hard\_Info_i+\beta_2\times Footprint_i+e_i+\varepsilon_i \quad (3)$$

At last, *Description* variable (the abundance of loan description) is plugged into formula (3) in order to analyze the impact of extra soft information:

$$Loan\ Outcome_i=\beta_1\times Hard\_Info_i+\beta_2\times Footprint_i+\beta_2\times Description_i+e_i+\varepsilon_i \quad (4)$$

We test the role of soft information newly obtained from fintech firm through the economic and statistical significance of individual explanatory variables. The AUC ranges from 50% (pure random prediction) to 100% (perfect prediction) judging the discriminatory power of three specifications is used to examine whether it complement or substitute for traditional hard information. The AUC refers to the probability of correctly identifying the good case when facing with one random good and one random bad case. It is one of the most important evaluation metrics for checking our model's predictive performance (Berg, Puri, and

Vol.1 No.11 2020
https://doi.org/10.37420/j.cer.2020.105

Rocholl, 2020).

Empirical Results

Table 1 shows the descriptive statistics of the gross hard information variable (also the revised credit score). The mean and median value of revised credit score are 30.8876 and 25.0050 respectively, and the kurtosis is 6.3574, indicating that the sample distributes with characteristics of sharp peaks and thick tails. In addition, the revised credit score (at 99% quantile) is 100.404, and the corresponding credit rating is HR. The economic implication is that borrowers on the lending platforms usually lack sufficient verifiable hard information. The approved rate of non-initial loans is only 3.6812% (10,430/283,331), which suggests that the long-tail group faced a large credit gap.

Table1. Descriptive statistics of gross hard information

Mean	Median	Variance	25% quantile	75% quantile	99% quantile
30.8876	25.0050	396.3488	17.7564	36.5033	41.1728

Table 2 reports descriptive statistics of other variables. The average borrowing interest rate of the total sample is 12.9987%, which is higher than 11.8482% of the successful loans. The W-M test between the mean of failed and successful loans is 118.199, significantly at 1% level, implying that investors on the platform are not completely yield-chasing. The average amount (*Amount*) of the failed order and the successful loan order is 64,704.11 yuan and 47,817.74 yuan respectively, and the standard deviations are 91,367.52 yuan and 42,249.05 yuan respectively. The median w-m test of them is 6.702, which is significant at 1% level, illustrating that the loan amount of the successful loan is significantly lower than that of the failed. The average loan term (*Term*) of Failed lending and successful borrowing are 17.3785 months and 22.4672 months. The standard deviation are 9.5767 months and 12.1434 months respectively, and the median W-M test is -65.859, significantly at the 1% level, showing that the borrowing period of failed loan is significantly lower than the successful, which implies a better liquidity (short term), and the creditor of small scale of loans is less likely to invest.

Table 2. Descriptive statistics of loans characteristics variables and Description

Variable	Total		Failed		Success		W-M
	Mean	S.D	Mean	S.D	Mean	S.D	Test
Amount	63066.82	87958.15	64704.11	91367.52	47817.74	42249.05	6.702***
Interest	12.9987	2.3640	13.1222	2.4145	11.8482	1.3709	118.199***
Term	17.8720	9.9692	17.3786	9.5767	22.4672	12.1434	-65.859***
Description	68.4541	50.0468	44.5541	38.2519	105.0436	43.4271	-124.178***

Notes: *** for $p < 0.01$, ** for $p < 0.05$, * for $p < 0.1$.

Table 3 reports the estimates from our Probit regressions examining the determinants of loan approval. Loan Outcome takes the value one for loan applications that were granted and zero for those that were denied. All of these specifications comprise all control variables in Table 2, region fixed effects and time fixed effects.

We report AUCs in the bottom rows of Table 3 and test for differences in AUCs using the methodology by DeLong, DeLong, and Clarke-Pearson (1988).The specification in Column 1 only includes the gross hard information variable (revised credit score) as the explanatory variable. Column 2 of Table 3 reports results using both revised credit score and as independent variables. Note that the growth of the internet brings novel techniques of information production. Intuitively, it should make it much easier to discriminate between good and bad cases. As expected, the AUC in Column 2 using the revised credit score and Footprint variables is 94.65%, higher than the 83.04% AUC using the revised credit score alone. Column 3 adds Description variable, of which the information is not included in the variable of Footprint but still belongs to soft information obtained by novel techniques. Consistent with the effect of Footprint, the abundance of loan description is significantly relevant to loan outcome. Furthermore, the 98.63% AUC which is higher than the AUC in Column 2 suggests that the abundance of loan description provides supplementary information for predict loan origination.

Table 3. Digital footprint variable, description and loan outcomes

Dependent Variable	Probit	Probit	Probit
	(1)	(2)	(3)
Hard_Info	0.00332***	0.00240***	0.00213***
	(0.00003)	(0.00002)	(0.00010)
Footprint		0.00157***	0.00132***
		(0.00001)	(0.00003)
Description			0.00011***
			(0.00002)
Interest	-0.11657***	0.00801***	-0.02780***
	(0.00176)	(0.00220)	(0.00711)
$Interest^2$	0.00304***	-0.00045***	0.00052***
	(0.00005)	(0.00008)	(0.00028)
Log (amount)	-0.04307***	-0.02865***	-0.02542***
	(0.00050)	(0.00038)	(0.00141)
Term	0.00591***	0.00095***	0.00260***
	(0.00006)	(0.00005)	(0.00021)
Intercept	Control	Control	Control
Time fixed effects	Control	Control	Control
Region effects	Control	Control	Control
Pseudo R2	0.2907	0.6389	0.2275
LR test	43014.07***	36245.84***	16254.15***
Observations	283,331	283,331	24484
AUC	0.8304	0.9465	0.9863
Difference to AUC of 50%/previous model	0.3304***	0.1161***	0.245***

Note: Data in parentheses are robust standard errors for products; *** p < 0.01, ** p<0.05,* p<0.1.

Finally, Figure 1 plots ROC curves of Model1-Model3, illustrating the rising of predictive power.

Vol.1 No.11 2020
https://doi.org/10.37420/j.cer.2020.105

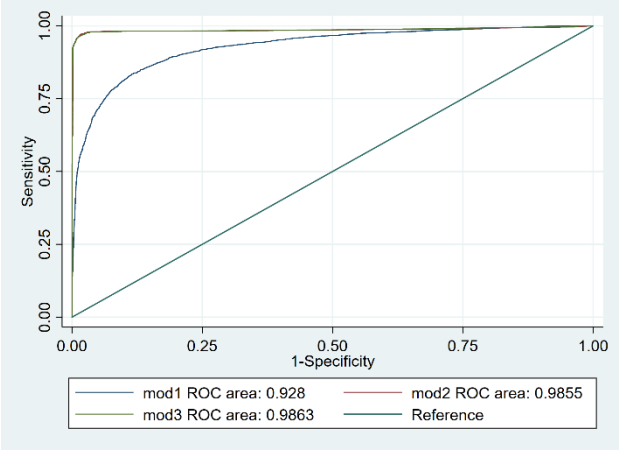

Figure 1. ROC curve of Model 1-3

In that the sample used to examine the effect of the abundance of loan description is limited to 24484 observations, we also test model 1 and model 2 with dataset of the limited sample and obtain similar results with consistent significance and AUC trends.

Robustness

One potential concern is that loan outcome may be also affected by some missing variables, such as the borrower's bank ledger, income stability and debt-to-income ratio etc., which are not disclosed on the platform. We emphasize that the coefficient on Fintech lenders is unbiased across specifications and samples. Following Du et al. (2014), we select the average car ownership of other borrowers at the same income level in the same year and same province as the instrumental variable for gross hard information, and a first-stage residual item as the instrumental variable for Footprint after repeating the construction by Formula (1). Then a two-stage IV-Probit estimation is carried out. These IVs do not directly affect loan outcome, at the same time, borrowers at the same income level generally have similar number of cars. Thus, endogenous explanatory variables and instrumental variables are strongly correlated. Wald test and Stock and Yogo test (Stock and Yog, 2005) prove it statistically. Compare statistical significance of interesting variables as well as AUC in Table 4 to them in Table3, the significance of coefficient remains consistent and signs do not flip. Meanwhile, AUC values from Column 1-3 also show similarly increasing trend. In a nutshell, the result is consistent with the results shown in Table 3, which shows consumers' digital mobile footprint and other soft information such as abundance of loan description play an important role in lending. This highlight the role of Fintech in Loan Origination.

Table 4. Robust Test (IV-Probit estimation)

Dependent Variable	IV-Probit		
	(1)	(2)	(3)
Hard_Info	0.05214***	0.06706***	0.08356***
	(0.00043)	(0.00117)	(0.00821)

Footprint		0.01914**	0.02039***
		(0.00022)	(0.00064)
Description			0.00101**
			(0.00047)
Interest	-0.44835**	0.33597***	-0.35642***
	(0.01722)	(0.02622)	(0.13007)
Interest^2	0.01074**	-0.01304***	0.00494***
	(0.00053)	(0.00089)	(0.00488)
Log (amount)	-0.42395***	-0.59128***	-0.60114***
	(0.03245)	(0.00839)	(0.03669)
Term	0.03245***	0.01354***	0.04677***
	(0.00057)	(0.00073)	(0.00300)
Intercept	Control	Control	Control
Time fixed effects	Control	Control	Control
Region effects	Control	Control	Control
Observations	283,328	283,328	283,328
AUC	0.928	0.9855	0.9863
Difference to AUC=50%	0.428***	0.0575***	0.0008***

Note: Data in parentheses are robust standard errors for products; *** p < 0.01, ** p<0.05,* p<0.1.

Conclusion

As a new type of information intermediary, online lending has brought into novel techniques and ways to collect information. Through evaluation (auditing) mechanism, fintech firms are capable of producing and transmitting hard and soft information at a lower cost hard and soft information of the borrower, through flattening financial organization, reducing transaction costs, and improving the efficiency of financial services. The fintech platform has to act as an information intermediary because the traditional credit intermediary is scarce. In addition, the investors on the platform possess insufficiently professional means to identify the credit quality for the long-tail crowd. Therefore, this lending market faces a serious issue of information asymmetry. For this reason, the platform is crucial to the process of the borrower's hard and soft information production, transmission and supervision, and is vital to relieve the issue of information asymmetry in financial markets.

Funding

This paper was supported by the Postgraduate Research and Innovative Project in University of International Business and Economics [No. 274201921].

References

Agarwal, S., Alok, S., Ghosh, P., & Gupta, S. (2019). Fintech and Credit Scoring for the Millennials: Evidence using Mobile and Social Footprints. Available at SSRN 3507827.

Vol.1 No.11 2020
https://doi.org/10.37420/j.cer.2020.105

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837-845.

Du, Z., Renyu, L. I., He, Q., & Zhang, L. (2014). Decomposing the rich dad effect on income inequality using instrumental variable quantile regression. *China Economic Review*, 31, 379-391.

Fuster, A., Plosser, M., Schnabl, P., & Vickery, J. (2019). The role of technology in mortgage lending. *The Review of Financial Studies*, 32(5), 1854-1899.

Gao, Q., Lin, M., Sias, R. W. (2018). Words matter: The role of texts in online credit markets. Available at SSRN 2446114.

Iyer, R., Khwaja, A. I., Luttmer, E. F., & Shue, K. (2016). Screening peers softly: Inferring the quality of small borrowers. *Management Science*, 62(6), 1554-1577.

Stock, J. H., & Yogo, M. (2002). Testing for weak instruments in linear IV regression (No. t0284). National Bureau of Economic Research.

Berg, T., Puri, M., & Rocholl, J. (2020). Loan Officer Incentives, Internal Rating Models, and Default Rates. *Review of Finance*, 24(3), 529-578.

Berg, T., Burg, V., Gombović, A., Puri, M. (2018). On the rise of fintechs–credit scoring using digital footprints (No. w24551). National Bureau of Economic Research.