Does Numerical Information Matter for Predicting Default Rates? Evidence from a P2P Lending Platform

Long Ma[*] Yuchao Peng[†] Xu Wei[‡]

This Version: July 2021

Abstract

Using 2012 to 2014 data from the Lending Club, this paper investigates the impact of numerical information in descriptive texts on the ex post default rate in online P2P lending. We find that borrowers providing numerical information with higher density are less likely to default. However, this predictive power is weakened if more round numbers are included in numerical information. In addition, numerical information can enhance the predictive power of traditional indicators like FICO, annual income, and DTI on default rate. Finally, using big data algorithms, we verify that such numerical information can provide additional value in predicting defaults.

JEL classification: G17, G32, G41
Keywords: P2P lending, default rate, numerical information, round number

[*]School of Management and Economics, Chinese University of Hong Kong (Shenzhen), Shenzhen, China.
[†]School of Finance, Central University of Finance and Economics, Beijing, China.
[‡]School of Finance, Central University of Finance and Economics, Beijing, China.

1 Introduction

Online P2P lending has in some cases become a substitute for bank lending, especially in serving infra-marginal bank borrowers (Tang, 2019). It also complements the work of banks in the area of small loans. The key advantage P2P lending holds over traditional bank lending is that soft information can be employed by lenders to predict borrowers' default probability, which reduces information asymmetry between borrowers and lenders. One important example of soft information is descriptive text, which is one or several sentences describing economic status, loan purpose, and other personal information of borrowers. This information is provided by borrowers voluntarily on P2P platforms and cannot be verified. However, a growing body of evidence suggests that the linguistic characteristics of descriptive texts (such as content, emotion, and length) can both affect lenders' investment behavior and reflect the credit status of borrowers, even after traditional credit indicators are considered (Larrimore et al., 2011; Herzenstein, Sonenshein, and Dholakia, 2011).

In this paper, we investigate an element of descriptive text that is largely neglected in the literature: numerical information, i.e., information displayed via mathematical numbers. Alongside the verbal information in natural language found in descriptive text, numerical information provides concise details about borrowers and delivers to lenders who will comprehend these details directly, easily, and differently than the verbal content. That is, numerical information can improve communication efficiency. The use of numerical information reflects the inclination of borrowers to create unobstructed channels with lenders, which alleviates the information asymmetry between these parties. Therefore, borrowers who provide (more) numerical information are supposed to have better credit and be less likely to default.

To test this hypothesis, we employ data from the Lending Club, a leading online P2P lending platform in the US. Since descriptive text was gradually removed from personal information for compliance reasons beginning in 2014, we only use samples from 2012 to 2014. Our empirical analysis begins by examining the relationship between the use of numerical information and borrowers' default probability. The use of numerical information is measured by the density of numbers (proportion of numbers to total text length). Consistent with the above hypothesis, we find that borrowers who include more numerical information in descriptive texts are less likely to default. This suggests that the use of numerical information can predict borrowers' credit status.

We then analyze how the accuracy of numerical information affects its predictive power on borrowers' default. Our paper measures the inaccuracy of numerical information by the ratio of round numbers in total numerical information, on the intuitive basis that round

numbers are less accurate than specific numbers. Increased reliance on round numbers may suggest that borrowers do not have a deep understanding of the real situation behind the data, that they lack confidence in the information due to uncertainty, or that they are trying to hide vital facts. It follows that such borrowers may present less clear and thus less credible information about their credit status than borrowers who use more specific numbers. Consistent with this hypothesis, we find that the predictive power of numerical information decreases when the proportion of round numbers increases.

In addition, we study the relationship between the numerical information from descriptive texts and traditional hard information, such as FICO score, annual income and Debt-to-Income ratio (DTI). It is found that the predictive power of hard information on default probability is strengthened by considering numerical information. That is, numerical information and hard information prove complementary. One reason for this result is that hard information is not perfectly accurate (Hilscher and Wilson, 2017), and numerical information, which reflects the borrower's attitude, can indirectly verify the authenticity of hard information.

Finally, we investigate the predictive power of numerical information on borrowers' default probability. After considering the numerical information indicators, we find that our prediction models perform better than models lacking these indicators. This again proves that numerical information holds value for default rate prediction and default risk management.

It is worth mentioning that our results don't mean that the numerical information is particularly important in and of itself, since it can only prove a borrower's habits in expression and communication. Ultimately, our results suggest that these habits deserve our attention because they offer crucial insights into the overall attitudes of the borrower. Different habits in the use of numbers in descriptive texts reflect quite different attitudes towards the loan. We find that more detailed and accurate numbers can signify reliability and hence better predict borrowers' default propensity.

Our findings contribute to the existing literature in two respects. First, we employ numerical information to extend the range of predictive variables for borrowers' defaults in P2P lending, which supplements the extant literature on hard information or limited soft information (Duarte, Siegel, and Young, 2012; Lin, Prabhala, and Viswanathan, 2013; Carmichael, 2014; Croux et al., 2020; Huang et al., 2021). Second, numerical information provides a new perspective on borrowers' behavior, and supplements research on the behavior of micro entities, especially borrowers and lenders in P2P market (Freedman and Jin, 2011; Zhang and Liu, 2012; Lin and Viswanathan, 2016; Iyer et al., 2016; Mollick and Nanda, 2016; Vallee and Zeng, 2019; Li et al., 2020).

The remainder of the paper is organized as follows: in Section 2, we review the relevant literature around this topic area. In Section 3, we develop three hypotheses on the effect of numerical information on borrowing. In Section 4, we show our data set and introduce variable constructions and empirical tactics. Section 5 presents our descriptive statistics, empirical results, and the robustness checks. We then discuss the predictive ability of our main indicators in Section 6, before concluding the paper in Section 7.

2 Literature Review

When trying to predict the default rates for online P2P lending, previous researchers focused mainly on borrowers' standardized information, such as credit grade (FICO score or credit grade issued by the platform), debt-to-income ratio, annual income, term, and loan amount. These indicators do have a significant impact on loan performances (Emekter et al., 2015). In addition, after controlling for these "traditional" credit indicators, some studies found that "non-traditional" standardized information (such as loan purpose, borrower's title, etc.) also significantly impact the default rate (Croux et al., 2020). Standardized information is therefore, at least to some extent, an objective, authentic, and easy to quantify predictor of default rates; however, used in isolation, it leaves a very limited space for platforms and investors to assess credit and predict defaults because it doesn't capture all the relevant information.

To acquire more information and gain a fuller picture of the borrower, some platforms give borrowers the right to voluntarily provide publicly observable commentaries whose form and content are fairly unrestricted. Borrowers can convey loan purpose, economic conditions, personal morality, family status or other personal "soft" information that might help them attract investors by building emotional ties or other means (Morse, 2015). Even though the authenticity of these commentaries cannot be verified, narrative information, which is much more unique and diverse than standard information, may provide important signals of future loan performance.

For regulatory reasons, some P2P platforms (like Lending Club) gradually eliminated the chance to provide descriptive texts; however, the texts themselves were helpful in judging credit status and attracted a lot of research. One strand of literature focused on the specific content of the texts and established a category system of different meanings: Herzenstein, Sonenshein, and Dholakia (2011) defined six identity claims in loan descriptions-trustworthy, economic hardship, hardworking, successful, moral, and religious-and found that "trustworthy" and "successful", compared with other identity claims, can increase the probability of obtaining loans but have no impact on loan performance. Larrimore et al. (2011) counted

the amount of words and number of different word properties in descriptive texts and found that using more articles, prepositions, certain specific adjectives ("many" or "lots of") and words related to money enhanced the success rate of funding, while words conveying personal information had an adverse effect. Jiang et al. (2018) introduced a topic model to obtain valuable information from the text and found that default prediction performance improved when the descriptive text information was considered in the prediction model. Finally, spelling errors in descriptive text can reflect the education level of the borrower and thus affect their credit status (Dorfleitner et al., 2016).

Another strand of research on descriptive text centered on the length of the text with the understanding that it might signify the traits of borrowers. Firstly, in terms of the impact of text length on funding success rate, Prystav (2016) found that investors were less likely to invest in loans accompanied by longer the descriptive texts. And yet, in terms of loan performance, Dorfleitner et al. (2016) point out that the length of descriptive text has no significant impact on the default rate. Li et al. (2019) do, however, confirm the existence of a quadratic relationship between the length of descriptive text and default possibility.

Emotion analysis is also a popular area of descriptive text research. For example, adopting psychology text mining techniques, Gao, Lin, and Sias (2018) locate linguistic tip-offs in narratives that may predict the loan performance; they find that the ease of reading narratives and narrative complexity are both highly correlated with default rates. Moreover, by regarding the emotion as a bipolar variable, they also find default rates are lower among borrowers that use more positive words.

3 Background and Hypothesis Development

3.1 What is numerical information?

Numerical information is a supplement to traditional hard information in descriptive text for P2P lending; it symbolizes the borrower's attempt to establish a positive relationship with lenders. Here, we present several examples of numerical information as used in descriptive text. In Table1, the numbers in bold depict many aspects of a borrower's information: income, interest rate, family members, number of credit accounts, credit history, occupation history, and even cars purchase date and durable years (which can be seen as detailed information of their assets).

3.2 Why does numerical information matter?

Table 1 above shows five examples of proper numbers that convey information about loan applicants. If such information is not expressed in numbers, but instead adopts more complicated linguistic elaborations, the recipients of the information (the lenders) will face an extremely complex interpretive task. Complex linguistic information may lead lenders to abandon part of the samples and narrow the selection range before making their final decision (Payne, 1976). Under such circumstance, the final choice may not be the optimal one for either party. Numerical information, by contrast, improves text readability and thus slashes communication costs, with a resulting negative impact on the default rate (Gao, Lin, and Sias, 2018). Other studies have shown that in business communication, when compared with verbal information, scale-value and natural-value numerical information is more likely to result in attribute-based processing that helps consumers make better decisions (Stone and Schkade, 1991). Childers and Viswanathan (2000) have also verified that numerical information is more accurately represented in memory, proving that numerical information promotes transaction efficiency.

By using numbers in the descriptive text for P2P lending, borrowers can provide more details about themselves and deliver more concise information to lenders who will comprehend their materials directly and easily. Therefore, the use of numerical information rather than verbal information (or even the use of more numerical information) reflects the willingness of borrowers to open up unobstructed channels with lenders and thus allay the problem of information asymmetry. Borrowers with such characteristics-those who tend to provide numerical information-should, we hypothesize, have better credit and be less likely to default. Based on the above, we present the following hypothesis:

Hypothesis 1. *The use of numerical information is negatively correlated with the borrower's default rate; that is, using numerical information will reduce the borrower's default possibility.*

3.3 Does the accuracy of numerical information affect its predictive power?

The specific meaning behind the numerical information is important to some extent, and yet the borrower's habit of using numerical information deserves our attention because the habit itself can reflect the borrower's characteristics. For instance, in Table 1 we see that when the number is large, the fourth borrower is accustomed to using approximate figures which (also called "round numbers") to describe the information. However, the fifth borrower prefers to provide more concrete figures, even accurate to two decimal places.

To capture differences in the accuracy of numerical information, this paper adopts the concept of round numbers. A round number conveys information vaguely and crudely as opposed to a more specific number. In particular, it refers to a number that is divisible by some special integer (e.g., 50, 100, 500, 1000, etc.). In general, more specific information suggests a greater deal of knowledge about that information on the part of the holder. Therefore, the use of less-specific round numbers may denote a kind of information loss or cognitive loss. D'acunto et al. (2019) point out that people with low intelligence quotients are more inclined to use round numbers to predict the rate of inflation. Kuo, Lin, and Zhao (2015) find that traders who rely too much on round numbers perform poorly in their investments. And finally, Lin and Pursiainen (2021) find that in reward crowdfunding, using round numbers as target amounts can be negatively correlated with the probability of successfully funding.

In our case, borrowers who use round numbers may not have a deep understanding of the real financial situation represented by the data. Otherwise, round numbers may indicate a lack of confidence in the borrower about the information, which may have led the borrower to hide vital facts. Along the same lines, such borrowers may intentionally present less credible or less clear information about their credit status than borrowers who use more specific numbers. Based on the above, we present the following hypothesis:

Hypothesis 2. *Using round numbers in descriptive text will weaken the original effect of numerical information. In other words, compared with more specific numbers, round numbers have less predictive power on default rate.*

3.4 How do numerical information and hard information interact?

Hard information such as credit grade, FICO score, annual income, and etc. is always regarded as the most powerful cue for predicting the default rate in P2P lending markets since it directly represents the financial situation and credit history of the loan applicant. Regarding traditional financial markets, Hilscher and Wilson (2017) point out that while credit ratings are sometimes inaccurate measures of raw default probability, they do nevertheless contain relevant information because they are related to a measure of exposure which is also related to CDS risk premium. However, in P2P lending markets, credit ratings like FICO scores or the sub-grades of the Lending Club are often irrelevant to financial risk management because they are based only on the publicly available information and relied on the big data algorithm. Thus, credit ratings in the P2P field are sometimes inaccurate as they tend to reach a plateau in their predictive power.

We suggest that as a manifestation of the borrower's attitude, numerical information in the descriptive text can indirectly verify the authenticity of hard information. This information can, in other words, enhance the credibility of other hard information indicators. For some borrowers, including numerical information in their descriptive text may strengthen the effect of their traditional credit indicators. From this perspective, this paper uses debt-to-income ratios (DTI), historical credit scores (FICO), and annual income levels to represent borrowers' hard credit information and to examine the interference effect of the numerical information. The third hypothesis that informs this methodology is as below:

Hypothesis 3. *For borrowers who use numerical information in their descriptive texts, the predictive power of their hard information indicators will be enhanced.*

4 Data and Methodology

4.1 Data

The data in this paper was obtained from Lending Club, an American P2P lending platform (www.lendingclub.com). Lending Club went public on December 3, 2014, and then gradually phased out the opportunity for borrowers to submit descriptive text. At that time, joint-liability lending also emerged on the platform, although this paper does not focus on the latter. Individual loan samples with descriptive text are our main concern, and to this end, we retained only samples from 2012 to 2014 and discarded the rest[1].

The loan status is the key dependent variable in this paper. According to the classification system used by Lending Club, loan status can be divided into 6 categories: *Charged Off*, *Fully Paid*, *Current*, *In Grace Period*, *Late (16-30 Days)* and *Late (31-120 Days)*. With the exception of *Charged Off* and *Fully Paid*, all other loan categories remain active in various states of their life cycles. For this reason, to ensure precision and consistency, only samples from these two categories (*Charged Off* and *Fully Paid*) will be selected. Furthermore, taking into consideration the negative influence of extreme values, we delete the top 5% of samples with the longest descriptive text, and the bottom 5% of samples with the shortest descriptive text, respectively. The observations with missing values are also eliminated. In this paper, we ultimately retain 85,975 observations. The distribution of samples in each year is shown in Figure 1.

[1]As mentioned in the introduction, there is no data available on the successful funding rate for Lending Club. As a result, we use only the default rate as our main dependent variable.

4.2 Variable construction

4.2.1 Indicators of numerical information

This paper measures the amount of numerical information in descriptive text with three variables ranging from vague to accurate. The definitions are listed below:

(1) Dummy indicator (*dum_num*): 1 denotes that the descriptive text of each observation contains numerical information, while 0 denotes that there is no numerical information in the text. This kind of binary measurement is less accurate than the following ones.

(2) Density indicator 1 (*den_num1*): the ratio of the number of digits to the total number of letters. This variable, which is more meaningful than the dummy indicator, represents the proportion of numerical information in the descriptive text. Take the fifth row of Table 1 as an example: the length of this text is 65, and it includes 12 digits. Thus, the density indicator 1 for this text will be 12/65.

(3) Density indicator 2 (*den_num2*): the ratio of the number of whole numerical information pieces to the total number of words. Splitting a whole number into fragments may destroy the fuller meaning of the information. Therefore, we construct a more accurate proportion indicator by text mining and word cutting. Again, consider the example above. There are 15 words in the descriptive text and only two pieces of complete numerical information. The density indicator 2 will therefore be 2/15.

4.2.2 Indicators of round numbers

Round number refers to a number that approximates numerical information. Referencing the samples, this paper sets four levels of round numbers: numbers divisible by 50, 100, 500 or 1000. Using this standard, we design four density indicators (*round50*, *round100*, *round500* and *round1000*) to measure the proportion of round numbers in each descriptive text. At the 1000 level, for instance, the indicator round1000 for the first observation in Table 1 is 1/34, while the fifth is 0/15. In the fourth observation, the four indicators are 3/69, 3/69, 1/69 and 1/69.

4.2.3 Control variables

In this paper, we select the control variables with reference to previous studies. As discussed, many factors can affect borrower's default rate, including two broad groupings: the borrower's hard information, and additional soft information.

First, to control the impact of text length on default rates and test their "U" shaped relationship, this paper puts text length and its square into the models. At the same time, the emotions of text may also influence the default rate. Thus, using two emotion dictionaries (Afinn and Bing[2]) we design two variables to depict the emotion of the text and include them in our models.

Second, we only consider the borrower's sub-grade, as provided by Lending Club, as an indicator of credit status because this variable is completely based on the borrower's FICO score and other important credit information. In addition, in order to exclude the impact of macroeconomic factors, we use years and state names of the US as dummy variables[3] in our models.

Finally, to check the robustness of the results, we use the following items to replace the sub-grade: borrower's FICO score, debt-to-income ratio, annual income, the number of delinquencies in the past two years, the number of credit inquiries in the past six months, employment length, home ownership, credit history and other related variables[4]. All the variable definitions and classifications are presented in Table 2.

4.3 Empirical model

4.3.1 The impact of numerical information

The binary response models provide us with a proper method to predict the propensity to default and also to verify the impact of numerical information on the default rate. Thus, we adopt the following logistic model (1) and probit model (2) to test our hypotheses:

$$\ln\left[\frac{\Pr(y=1)}{\Pr(y=0)}\right] = \alpha + \beta X + \boldsymbol{\gamma}^{'}\boldsymbol{C} \tag{1}$$

$$\Phi^{-1}[\Pr(y=1)] = \alpha + \beta X + \boldsymbol{\gamma}^{'}\boldsymbol{C} \tag{2}$$

where y is the dependent variable *loan_status*. y equals 1 if the loan has defaulted and equals 0 otherwise. X denotes the core independent variables-the numerical information indicators-

[2]The vocabulary in Afinn dictionary is graded from - 5 to 5. The greater the grade, the more positive the emotion. We aggregate the weighted (based on the grade) amount of emotion words to obtain the first net emotion indicator. Bing used a dichotomy method to simply divide the words into positive group and negative group. We defined the second net emotion indicator as the difference between the number of positive words and the number of negative words. The proportions of these two indicators in descriptive text are two density variables of text emotion.

[3]Table 13 in appendix shows per capita GDP in 2012-2014 and the default rate on Lending Club of each state. According to our test, GDP is high correlated with default rate. Therefore, controlling the region and the time roughly equals to controlling the effects of macroeconomy factors.

[4]The interest rate in Lending Club is determined by the credit grade. To avoid perfect collinearity, we discard the interest rate.

with which we are concerned, while $\boldsymbol{C}$ represents a vector of control variables. Following Hypothesis 1, we expect the coefficient β in the above two models to be significantly negative, indicating that the use of numerical information is negatively related to the default rate.

To consolidate our findings, we also adopt the Cox regression model, which is a common method for survival analysis. The model is as follows:

$$h(t) = h_0(t) \cdot \exp\left(\beta X + \boldsymbol{\gamma}^{'}\boldsymbol{C}\right) \tag{3}$$

where t denotes the duration of the loans, $h(t)$ is the risk function and $h_0(t)$ presents the basic risk. Also, β is supposed to be significantly negative.

4.3.2 The effect of round numbers

Our strategy for identifying the effect of round numbers on default rates is to introduce the interaction term between numerical information and round numbers into the model. The specification is below ((4) for logistic model and (5) for Cox model):

$$\ln\left[\frac{\Pr(y=1)}{\Pr(y=0)}\right] = \alpha + \beta_1 X + \beta_2 Z + \beta_3 X \cdot Z + \boldsymbol{\gamma}^{'}\boldsymbol{C} \tag{4}$$

$$h(t) = h_0(t) \cdot \exp\left(\beta_1 X + \beta_2 Z + \beta_3 X \cdot Z + \boldsymbol{\gamma}^{'}\boldsymbol{C}\right) \tag{5}$$

where Z is a round number indicator (including *round50* to *round1000*). For robustness reasons, we also control the Z, although the coefficient β_2 is not as important as the coefficient β_3 which, if the reasoning behind hypothesis 2 holds true, ought to be significantly positive.

4.3.3 Improved predictive power of traditional indicators

To see how numerical information affects the predictive power of traditional indicators, M, including FICO, annual income, and DTI, we also add interaction terms. According to Hypothesis 3, the coefficients of these interaction items are supposed to be significantly negative (for FICO and annual income) or positive (for DTI). The logistic specification is similar to (4) where Z is replaced by M.

5 Empirical Result

5.1 Descriptive statistics

Table 3 shows the summary statistics of independent variables and the t-test results for default samples (*Charged Off*) and non-default samples (*Fully Paid*). This table indicates that the mean values of the numerical information indexes (*dum_num*, *den_num1*, *den_num2*) of the non-default sample are significantly higher than those of the nondefault samples. The proportions of round numbers divisible by 50, 500 and 1000, respectively, of the non-default sample are also significantly higher.

For other quantitative control variables, we can draw the following conclusions: (1) on average, the non-default sample has a longer descriptive text; (2) the default sample is significantly higher than the non-default sample in terms of loan amount and loan term; (3) the non-default sample has a more positive tone in the descriptive text; (4) the statistical characteristics of the credit information between these two groups are quite distinct, especially in sub-grade, FICO score, annual income, credit history, debt-to-income ratio (DTI), and so on.

In addition, the frequency statistics of the type variables (*purpose* and *home_ownership*) are listed in Table 4. As seen below, (1) mortgage type accounts for a larger proportion in the nondefault sample than in the default sample, while in the default sample, the percentage of renting type is slightly higher; (2) the number of non-default observations with consumption purpose is higher than that of default observations.

The pairwise correlations are also provided. In Table 5, most of the core independent variables and control variables are significantly correlated with the dependent variable *loan_status*. Specifically, the numerical information indicators are all negatively correlated with the possibility of default. Further, the correlation coefficient between *sub-grade* and the default rate is the greatest and, at the same time, the other control variables are highly correlated with *sub-grade*. Thus, as mentioned, the variable *sub-grade* is fully capable of covering most of the information we need.

Although there are correlations observed between some control variables and the numerical information indicators, or between round number indicators and numerical information indicators, the coefficients are small (all below 0.7), implying a limited multicollinearity.

5.2 Regression results

5.2.1 Testing Hypothesis 1

Table 6 reports the results of the binary regression and Cox regression for testing the pure effect of numerical information on default rate. As seen in specifications 1 to 3, the dummy variable of numerical information is negatively related to the default rate (the coefficients are -0.144, -0.0768 and -0.104, respectively, which are significantly negative on the 0.01 level), and therefore indicates whether providing numerical information or not will affect the borrower's default rate. In other words, borrowers with numerical information in their descriptive text are less likely to default.

Specifications 4 to 6 adopt the density variable *den_num1* and find evidence that more numerical information in units of descriptive text reflects a lower default rate (the coefficients are all significantly negative at 0.01 or 0.05 level). Specifications 7 to 9 adopt the density variable *den_num2* and present the same results. That is, the significant negative effect of numerical information persists. As for hazard ratio in the Cox models, if the proportion of numerical information in descriptive text increases by a marginal unit, the default risk will diminish by $57.4\%(1-0.426) \sim 76.7\%(1-0.233)$. Therefore, more numerical information in descriptive texts suggests lower default risk. Thus, our results are consistent with Hypothesis 1.

For control variables, the following findings are worth noting: (1) the emotion indicators do not have significant effects; (2) there is a U-shaped relationship between text length and default rate; and (3) *sub-grade*, as the strongest indicator of borrower's credit status, has a significant negative impact on default rate. These findings are all robust and largely consistent with the conclusions drawn from previous research.

5.2.2 Testing Hypothesis 2

The effect of round numbers is displayed in Table 7. It is shown that after adding the interaction term between numerical information and round number, the original negative effect of numerical information effect remains significant. However, all the coefficients of interaction items are significantly negative, which implies that the use of round numbers tends to reduce the negative effect of numerical information. This result is consistent with Hypothesis 2.

5.2.3 Testing Hypothesis 3

As seen above, the numerical information is significantly and negatively correlated with the probability of default, which suggests that it reflects the attitude of borrowers. Here, we want to know whether the role of numerical information can be "icing on the cake", i.e., whether it can improve the predictive power of traditional indicators. The regression results are shown in Table 8, which shows that the coefficients of the interaction term between DTI and numerical information are significantly positive (specifications 1 to 3), and that the coeffects of the interaction terms between FICO and numerical information (specifications 4 to 6) and between annual income and numerical information (specifications 7 to 9) are significantly negative. These results suggest that the hard information of borrowers who also provide more numerical information in their descriptive texts has a greater influence on ex post default rate. That is: numerical information enhances the predictive power of hard information. These results are consistent with Hypothesis 3.

5.3 Robustness check

In this section, we discuss the robustness of our results by replacing our control variables and key independent variables, and by addressing the problem of sample selection bias.

5.3.1 Missing information

In our main regression, we only use one variable grade to control the credit status of borrowers. Thus, there may be missing information or unobserved variables. Hence, in this section, we add a group of variables (loan amount, loan term, FICO, DTI, annual income, delinquency status, inquiry status, credit history and length of employment) to replace the former variable *grade* and thus fully control the credit status[5]. The text length, emotion indicators and macro information are still controlled. Table 9 presents the regression results; our results do not change significantly when the control variables are altered.

5.3.2 Measurement errors

The main regression of our paper only considers Arabic numerals when accounting for the amount of numerical information. However, there are also numbers in the form of words in descriptive texts, such as "one", "hundred", and "thousand". After careful cleaning, we establish two new indicators (*den_new1* and *den_new2*) of numerical information that

[5]Actually, the variable grade is a composite variable that considers the whole group of variables we add here. Hence, we don't include this variable in the regression of our robustness check.

include such numbers in words to replace *den_num1* and *den_num2* in the main regression. We report part of these results with a new index variable in Table 10[6]. The main conclusions are all consistent with the benchmark regression model.

5.3.3 Sample selection bias

In our main regression, we only employ the sample of borrowers who provide descriptive texts. However, since borrowers decide whether or not to provide the texts, their decision may correlate with the borrower's default rate. To rectify this bias, we first select a new random sample (60% of the sample with descriptive text) and run our main regressions. The results are in Table 11.

Next, we employ the Heckman two-step regression mode, which sets up a selection equation and a response equation. The results are presented in Table 12. As seen in Tables 11 and 12, our main results are not affected by the sample change or by the regression model rectification. Even though the inverse Mills ratios of the specifications in Table 12 are significant, the sign and the significance of the coefficients on numerical information indicators remain similar to our main regression under the Heckman settings. So, our results are robust and unaffected by the change in external conditions.

6 Default Prediction

The ultimate goal of this paper is to more accurately predict the default rate. To this end, we further explore how numerical information can improve the prediction model. In particular, we discuss the prediction abilities of different models with and without numerical information indicators (NII).

First, we adopt a traditional logit model with 4-fold cross validation (dividing the samples into 4 parts, taking each part as a prediction set and the rest as a training set, and then repeating the process four times) to predict the default rate. We evaluate two aspects of the prediction model (Figure 2): first, the prediction accuracy, which is the proportion of the correctly predicted sample in relation to the total sample; second, lift value which refers to the ratio of the (1-predicted default rate) to the (1-original default rate), and which measures the degree to which the prediction model mitigates the default propensity.

Xgboost, a supervised learning algorithm based on decision trees, is chosen as another prediction model. For this model, we select 70% of the sample as the training set and 30%

[6]To make the paper more concise, we only report representative results in Table 10. Please contact the author if the rest of the results are needed.

as the prediction set. Accuracy and lift values under different threshold levels are provided in Figure 3. As seen in these charts, the accuracy values are approximately 84% and the lift values are greater than 1. It's worth noting that the bars on the right of both logit models and xgboost models are higher than those on the left representing the result without NII, which suggests that adding numerical information indicators to our prediction models can improve accuracy and lift.

In addition, to ensure robustness, we verify our conclusions by using SVM. Figure 4 display the accuracy results under two different kernel functions (sigmoid and rbf). The bars on the right in these three charts are all higher than those on the left, showing once again that we can obtain a more accurate prediction result by considering numerical information. That is: the numerical information in descriptive text can contribute additional value to default rate predictions.

7 Conclusion

Descriptive texts on P2P platforms can provide valuable additional information that is absent from standardized "hard information". This paper explored an overlooked type of information in descriptive text: numerical information. Employing Logit/Probit regressions and a Cox model, we find that numerical information predicts a lower default propensity in borrowers. However, round numbers decrease the credibility of the numerical information. Another important finding is that numerical information and hard information are complementary in the sense that numerical information enhances the predictive power of hard information.

The complexity of information on P2P lending platforms often makes it difficult for lenders to make decisions. This paper provides a new perspective on the available materials that can help them extract more valuable information from the descriptive text; that is, that can help them judge the credit status of borrowers according to the density and the accuracy of numerical information. It is also worth noting that for borrowers with good credit status as measured by standardized indicators, numerical information is an additional benefit. And finally, for the manager of the P2P platform, credit risk management is always important, not only for security reasons but also for the long-term development of the platform. Our findings may thus complement their risk management toolkit and help managers more accurately predict credit risk.

References

Carmichael, D. (2014). Modeling default for peer-to-peer loans. *Available at SSRN 2529240*.

Childers, T. L. and Viswanathan, M. (2000). Representation of numerical and verbal product information in consumer memory. *Journal of Business Research*, 47(2):109–120.

Croux, C., Jagtiani, J., Korivi, T., and Vulanovic, M. (2020). Important factors determining fintech loan default: Evidence from a lendingclub consumer platform. *Journal of Economic Behavior & Organization*, 173:270–296.

D'acunto, F., Hoang, D., Paloviita, M., and Weber, M. (2019). Iq, expectations, and choice. Technical report, National Bureau of Economic Research.

Dorfleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., de Castro, I., and Kammler, J. (2016). Description-text related soft information in peer-to-peer lending–evidence from two leading european platforms. *Journal of Banking & Finance*, 64:169–187.

Duarte, J., Siegel, S., and Young, L. (2012). Trust and credit: The role of appearance in peer-to-peer lending. *The Review of Financial Studies*, 25(8):2455–2484.

Emekter, R., Tu, Y., Jirasakuldech, B., and Lu, M. (2015). Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending. *Applied Economics*, 47(1):54–70.

Freedman, S. M. and Jin, G. Z. (2011). Learning by doing with asymmetric information: Evidence from prosper. com. Technical report, National Bureau of Economic Research.

Gao, Q., Lin, M., and Sias, R. W. (2018). Words matter: The role of texts in online credit markets. *Journal of Financial and Quantitative Analysis, forthcoming*.

Herzenstein, M., Sonenshein, S., and Dholakia, U. M. (2011). Tell me a good story and i may lend you money: The role of narratives in peer-to-peer lending decisions. *Journal of Marketing Research*, 48(SPL):S138–S149.

Hilscher, J. and Wilson, M. (2017). Credit ratings and credit risk: Is one measure enough? *Management science*, 63(10):3414–3437.

Huang, J., Sena, V., Li, J., and Ozdemir, S. (2021). Message framing in p2p lending relationships. *Journal of Business Research*, 122:761–773.

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

Jiang, C., Wang, Z., Wang, R., and Ding, Y. (2018). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals of Operations Research*, 266(1-2):511–529.

Kuo, W.-Y., Lin, T.-C., and Zhao, J. (2015). Cognitive limitation and investment performance: Evidence from limit order clustering. *The Review of Financial Studies*, 28(3):838–875.

Larrimore, L., Jiang, L., Larrimore, J., Markowitz, D., and Gorski, S. (2011). Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research*, 39(1):19–37.

Li, E., Liao, L., Wang, Z., and Xiang, H. (2020). Venture capital certification and customer response: Evidence from p2p lending platforms. *Journal of Corporate Finance*, 60:101533.

Li, Z., Zhang, H., Yu, M., and Wang, H. (2019). Too long to be true in the description? evidence from a peer-to-peer platform in china. *Research in International Business and Finance*, 50:246–251.

Lin, M., Prabhala, N. R., and Viswanathan, S. (2013). Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management science*, 59(1):17–35.

Lin, M. and Viswanathan, S. (2016). Home bias in online investments: An empirical study of an online crowdfunding market. *Management science*, 62(5):1393–1414.

Lin, T.-C. and Pursiainen, V. (2021). The round number heuristic and entrepreneur crowdfunding performance. *Journal of Corporate Finance*, 68:101894.

Mollick, E. and Nanda, R. (2016). Wisdom or madness? comparing crowds with expert evaluation in funding the arts. *Management science*, 62(6):1533–1553.

Morse, A. (2015). Peer-to-peer crowdfunding: Information and the potential for disruption in consumer lending. *Annual Review of Financial Economics*, 7:463–482.

Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational behavior and human performance*, 16(2):366–387.

Prystav, F. (2016). Personal information in peer-to-peer loan applications: Is less more? *Journal of Behavioral and Experimental Finance*, 9:6–19.

Stone, D. N. and Schkade, D. A. (1991). Numeric and linguistic information representation in multiattribute choice. *Organizational Behavior and Human Decision Processes*, 49(1):42–59.

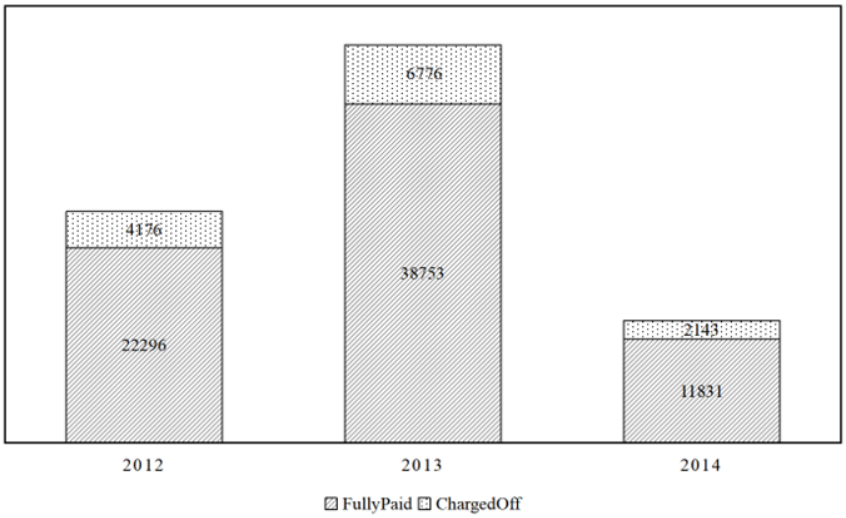
Tang, H. (2019). Peer-to-peer lenders versus banks: substitutes or complements? *The Review of Financial Studies*, 32(5):1900–1938.

Vallee, B. and Zeng, Y. (2019). Marketplace lending: A new banking paradigm? *The Review of Financial Studies*, 32(5):1939–1982.

Zhang, J. and Liu, P. (2012). Rational herding in microloan markets. *Management science*, 58(5):892–912.

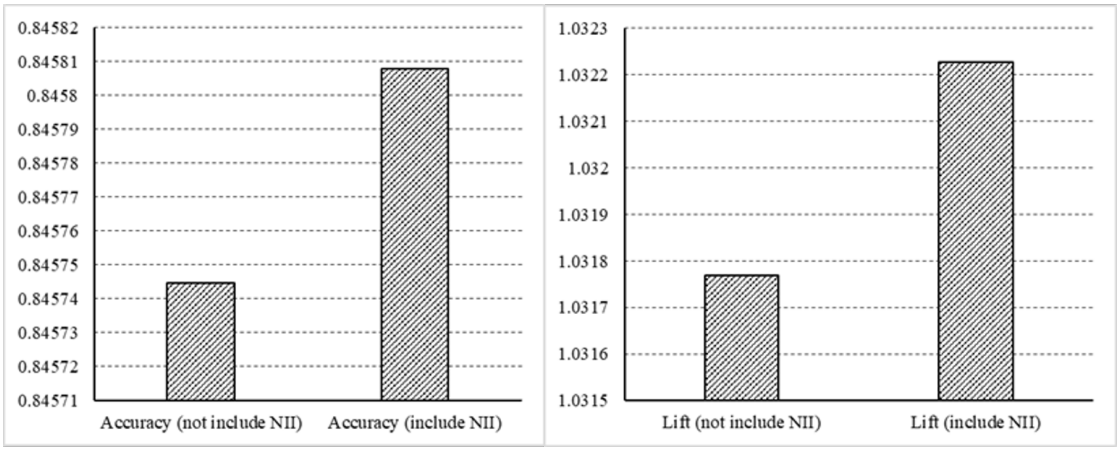
Figure Appendix

Figure 1: Sample distribution in each year



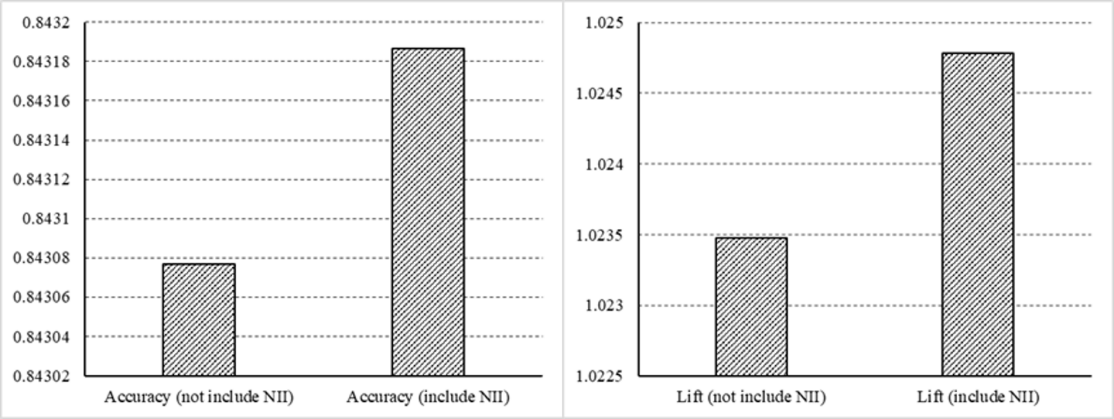
Note: Figure 1 shows the sample volume from 2012 to 2014, including the proportion of Fully-Paid samples and that of Charged-Off samples. Although the volume varies over time, the default rate experiences only slight changes, which is 15.78% in 2012, 14.88% in 2013 and 15.34% in 2014.

Figure 2: Accuracy and Lift of logit prediction model with 4-fold cross validation



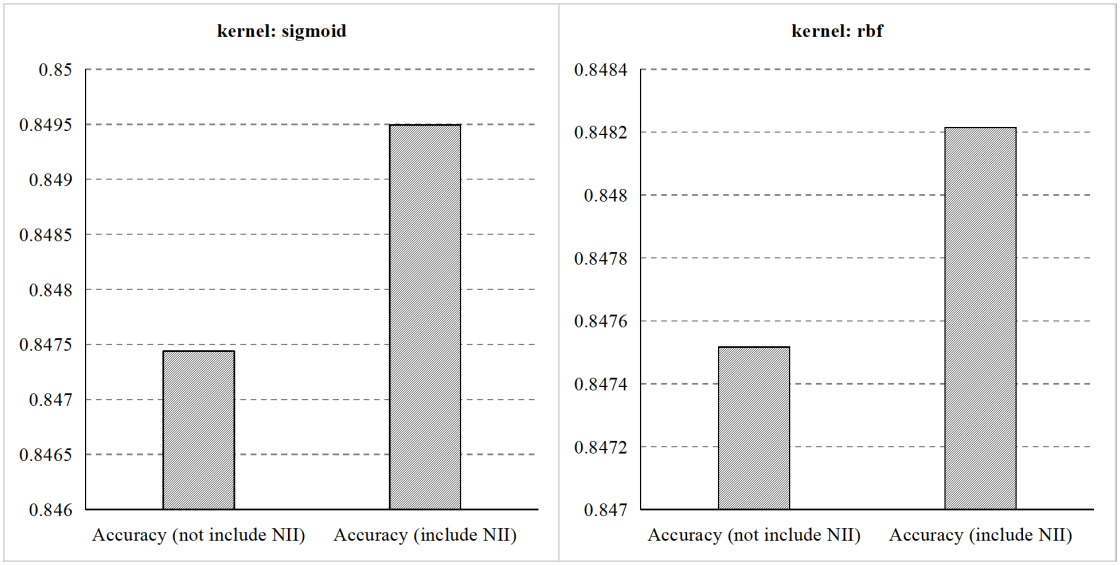
Note: Figure 2 reports the average accuracy and lift of logit prediction model with different threshold levels (0.41 to 0.49). The left bar represents the results of models including NII, and the right bar represents the results of models without NII.

Figure 3: Accuracy and Lift of xgboost prediction model



Note: Figure 3 reports the average accuracy and lift of xgboost model with different threshold levels (0.40 to 0.50). The training set is randomly selected for 70% of the total sample and the rest of the sample is the prediction set. In these two bar charts, the left bar represents the results of models including NII, and the right bar represents the results of models without NII.

Figure 4: Accuracy of SVM prediction models with different kernel functions



Note: Figure 4 reports the accuracy of SVM models under 2 different kernel functions (sigmoid and rbf). The training set is randomly selected for 70% of the total sample and the rest of the sample is the prediction set. In these two bar charts, the left bar represents the results of models including NII, and the right bar represents the results of models without NII.

Table Appendix

Table 1: Examples of Descriptive Text

	Descriptive Text
1	My **2001** car finally passed away with **160000** miles. I would like to purchase a used car. Right now I have to take a bus to the railroad in order to get to work.
2	Great job for **15**+ years. Homeowner for **10**+ years. Paying off small debt and Finishing Kitchen renovation.
3	I have **4** credit cards with high interest rates that I want to refinance. I found this loan helpful to save some money. Thank you for your support.
4	Well I am a **44** year old with a family of 5 children $\cdots$ and good stable job earning about **85000** dollar per year. I have textbftwo large accounts with a major lender and a good payment history yet their interest rate is **18%** and **13%** respectively on my accounts and i am currently paying a little over **400** a month in interest ALONE. Credit score at or near **700**.
5	I have **two** loans I am consolidating. One for **5197.62** and the other for **8014.28**.

Note: Table 1 displays five descriptive text examples from Lending Club. No changes are made to the content, so grammar and punctuation mistakes probably exist. But some unrelated words are deleted for the sake of simplicity. All the numerical information is shown in bold.

Table 2: Description of variables

Variable	Definition
loan_status	1 for *Charged Off*; 0 for *Fully Paid*
dum_num	1 denotes that there is numerical information; 0 denotes that there is no numerical information in descriptive text
den_num1	Ratio of the number of digit to the total number of letters
den_num2	Ratio of the number of whole numerical information pieces to the total number of words
round50	Proportion of round numbers (divisible by 50) in descriptive text
round100	Proportion of round numbers (divisible by 100) in descriptive text
round500	Proportion of round numbers (divisible by 500) in descriptive text
round1000	Proportion of round numbers (divisible by 1000) in descriptive text
n_words	Number of words in descriptive text (text length)
den_emo1	Proportion of net emotion (Affinn dictionary) in descriptive text
den_emo2	Proportion of net emotion (Bing dictionary) in descriptive text
grade	Lending Club Sub-grade (G5 to A1, 35 levels)
date	Issue date (year) of the loans
state	State of the US where the borrower comes from
amount	Amount of the loan
term	Term of the loan (0 for 36 months; 1 for 60 months)
fico	FICO score of the borrower
dti	Debt-to-income(monthly) ratio
delinquency	Number of delinquencies in the past two years
inquiry	Number of credit inquiries in the past 6 months
income	Annual income of the borrower
cred_his	Credit history (the number of days between the opening date of first credit account and issue date of the loan on Lending Club)
bank_card	Bank card utilization
employment	Employment length (0 for less than 1 year; n for n years; 10 for more than 10 years)
purpose	Loan purpose (0 for other purposes; 1 for production purpose; 2 for consumption purpose; 3 for debt consolidation)[7]
home_ownership	Home ownership status of the borrower (0 for other situations; 1 for renting; 2 for mortgage; 3 for owning)

Note: Table 2 displays all the variables and their definitions. Discrete variables include *dum_num*, *date*, *state*, *term*, *purpose*, and *home_ownership*. The rests are continuous variables.

[7]Loan purpose can be classified into: production, consumption, debt consolidation and other purpose.

Table 3: Summary statistics I

Variables	All listings				*Fully Paid*	*Charged Off*	t	p
	Mean	Std.	Min	Max	Mean	Mean		
dum_num	0.237	0.425	0	1	0.242	0.208	8.68	0.000
den_num1	0.005	0.017	0	0.509	0.006	0.005	5.95	0.000
den_num2	0.013	0.032	0	0.643	0.013	0.011	6.72	0.000
round50	0.002	0.013	0	0.5	0.002	0.002	2.1	0.035
round100	0.002	0.012	0	0.5	0.002	0.002	1.51	0.132
round500	0.001	0.009	0	0.5	0.001	0.001	1.77	0.077
round1000	0.001	0.008	0	0.5	0.001	0.001	2.04	0.042
n_words	25.3	20.2	2	109	25.4	24.7	4.07	0.000
den_emo1	-0.021	0.117	-1	1	-0.02	-0.024	3.05	0.002
den_emo2	-0.049	0.24	-1.5	2	-0.047	-0.057	4.05	0.001
amount	14855.4	8024.4	1000	35000	14644.7	16028.3	-17.67	0.000
term	0.23	0.421	0	1	0.2	0.393	-42.6	0.000
grade	24.81	6.32	1	35	25.42	21.43	63.82	0.000
fico	699.3	30.5	662	847.5	700.8	690.7	41.31	0.000
dti	17.21	7.53	0	34.99	16.95	18.62	-23.54	0.000
delinquency	0.248	0.716	0	22	0.245	0.262	-2.34	0.019
inquiry	0.816	1.043	0	8	0.789	0.965	-16.86	0.000
income	74660.2	52710.3	7000	7141778	75759.5	68542.3	17.68	0.000
cred_his	5731.9	2516.1	1096	22827	5767.7	5532.8	10.05	0.000
bank_card	66.1	26.3	0	339.6	65.3	70.6	-22.23	0.000
employment	6.1	3.6	0	10	6.1	6.2	-1.53	0.126

Note: Table 3 displays the result of descriptive statistics of all continuous variables. T-statistics and p values are the results of Bi-variate means test between two different samples.

Table 4: Summary statistics II

Variables	Type	All listings		*Fully Paid*		*Charged Off*	
		Amnt.	Freq. (%)	Amnt.	Freq. (%)	Amnt.	Freq. (%)
home_ownership	Other	45	0.05	35	0.05	10	0.08
	Renting	33467	38.93	27823	38.18	5644	43.1
	Mortgage	45736	53.2	39335	53.97	6401	48.88
	Owning	6727	7.82	5687	7.8	1040	7.94
purpose	Other	3281	3.82	2671	3.66	610	4.66
	Production	2560	2.98	2114	2.9	446	3.41
	Consumption	7704	8.96	6683	9.17	1021	7.8
	Debt Cons.	72430	84.25	61412	84.26	11018	84.14

Note: Table 4 shows the result of descriptive statistics of discrete variables and presents the amount and the frequency of different type of home ownership status and loan purpose.

Table 5: Pairwise correlations

	y	X_1	X_2	X_3	Z_1	Z_2	Z_3	Z_4	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}
y	1																				
X_1	-0.03	1																			
X_2	-0.02	0.58	1																		
X_3	-0.02	0.71	0.88	1																	
Z_1	-0.01	0.32	0.68	0.53	1																
Z_2	-0.01	0.31	0.65	0.50	0.96	1															
Z_3	-0.01	0.24	0.54	0.40	0.78	0.82	1														
Z_4	-0.01	0.21	0.48	0.35	0.70	0.74	0.92	1													
C_1	-0.01	0.41	0.11	0.12	0.08	0.07	0.05	0.04	1												
C_2	-0.01	0.14	0.07	0.08	0.04	0.03	0.02	0.02	0.30	1											
C_3	-0.01	0.13	0.06	0.08	0.04	0.03	0.02	0.02	0.28	0.92	1										
C_4	0.06	0.02	-0.01	0	-0.01	0	0	0.01	0.02	-0.06	-0.06	1									
C_5	0.16	-0.02	-0.02	-0.02	0	0	0	0	-0.03	-0.05	-0.05	0.42	1								
C_6	-0.23	0.05	0.04	0.04	0.02	0.02	0.02	0.01	0.04	0.06	0.04	-0.23	-0.51	1							
C_7	-0.12	0.04	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.05	0.03	0.11	-0.01	0.53	1						
C_8	0.08	-0.01	-0.02	-0.01	-0.02	-0.02	-0.02	-0.02	0.01	-0.03	-0.02	0.05	0.09	-0.15	-0.11	1					
C_9	0.01	-0.02	-0.01	-0.01	0	-0.01	0	0	-0.03	-0.02	-0.02	0.01	0.01	-0.09	-0.17	-0.01	1				
C_{10}	0.06	0	0	0	0	0	0	0	-0.02	0	0	0.02	0.05	-0.24	-0.04	0.01	0.03	1			
C_{11}	-0.05	0.02	0.02	0.02	0.01	0.01	0.01	0.02	-0.03	-0.01	-0.02	0.36	0.07	0.02	0.11	-0.20	0.06	0.08	1		
C_{12}	0.07	-0.01	-0.03	-0.02	-0.02	-0.02	-0.02	-0.02	0.01	-0.05	-0.04	0.04	0.06	-0.37	-0.54	0.22	-0.01	-0.08	-0.03	1	
C_{13}	0.01	0	0	0	-0.01	0	0	0	-0.09	-0.04	-0.03	0.11	0.09	-0.04	0	0.04	0.03	-0.01	0.06	0.04	1

Note: Table 5 shows the result of pairwise correlation between all continuous variables. In this table, y is *loan_status*, X_1 to X_3 are *den_num*, *den_num1* and *den_num2*, Z_1 to Z_4 are *round50*, *round100*, *round500* and *round1000*, C_1 to C_{13} are *n_words*, *den_emo1*, *den_emo2*, *amount*, *term*, *grade*, *fico*, *dti*, *delinquency*, *inquiry*, *income*, *bank_card*, and *employment*. Due to the limit of space, p-values are intentionally omitted in this table.

Table 6: Effect of numerical information on default rate in binary response models and Cox model

Variables	Spec. 1	Spec. 2	Spec. 3		Spec. 4	Spec. 5	Spec. 6		Spec. 7	Spec. 8	Spec. 9	
	Logit	Probit	Cox		Logit	Probit	Cox		Logit	Probit	Cox	
	Coef.	Coef.	Coef.	HR	Coef.	Coef.	Coef.	HR	Coef.	Coef.	Coef.	HR
dum_num	-0.144***	-0.0768***	-0.104***	0.901								
	(0.0262)	(0.0142)	(0.0237)									
den_num1					-2.256***	-1.125***	-1.459**	0.233				
					(0.716)	(0.37)	(0.6)					
den_num2									-1.379***	-0.710***	-0.854***	0.426
									(0.352)	(0.186)	(0.304)	
den_emo1	0.0276	0.0169	0.0321	1.033	0.0325	0.0195	0.0357	1.036	0.0347	0.0207	0.0369	1.038
	(0.0433)	(0.0239)	(0.0384)		(0.0432)	(0.0239)	(0.0384)		(0.0432)	(0.0239)	(0.0384)	
n_words	-0.0039***	-0.0023***	-0.0029**	0.997	-0.0053***	-0.0031***	-0.0040***	0.996	-0.0052***	-0.0030***	-0.0039***	0.996
	(0.0015)	(0.0008)	(0.0013)		(0.0014)	(0.00078)	(0.0013)		(0.0014)	(0.0008)	(0.0013)	
n_words^2	0.00005***	0.00003***	0.00004**	1	0.00006***	0.00003***	0.00004***	1	0.00006***	0.00003***	0.00004***	1
	(0.00002)	(0.00001)	(0.00002)		(0.00002)	(0.00001)	(0.00002)		(0.00002)	(0.00001)	(0.00002)	
grade	-0.0909***	-0.0521***	-0.0681***	0.934	-0.0909***	-0.0522***	-0.0681***	0.934	-0.0909***	-0.0522***	-0.0681***	0.934
	(0.0014)	(0.0008)	(0.0012)		(0.0014)	(0.0008)	(0.0012)		(0.0014)	(0.0008)	(0.0012)	
Constant	0.0879	0.0386			0.0945	0.0426			0.0985	0.0446		
	(0.209)	(0.11)			(0.209)	(0.11)			(0.209)	(0.11)		
date	Yes	Yes	Yes		Yes	Yes	Yes		Yes	Yes	Yes	
state	Yes	Yes	Yes		Yes	Yes	Yes		Yes	Yes	Yes	
Log L.	-34513.784	-34463.923	-139903.09		-34522.961	-34473.145	-139909.72		-34520.293	-34470.451	-139908.72	
Pseudo R^2	0.0592	0.0605	/		0.0589	0.0603	/		0.059	0.0604	/	
Observations	85,974[8]	85,974	85,975		85,974	85,974	85,975		85,974	85,974	85,975	

Note: Table 6 shows the result of regression in which the numerical information indicators (*dum_num*, *den_num1*, *den_num2*) are core independent variables. Spec.1, Spec.4 and Spec.7 are logit models. Spec.2, Spec.5 and Spec.8 are probit models. Spec.3, Spec.6 and Spec.9 are Cox models. Issuing date and state information are controlled in all models. Especially for Cox models, coefficient in left part of the column and hazard ratio in right. Robust standard errors in parentheses. $*p < 0.1$, $**p < 0.05$, $***p < 0.01$.

[8]The observations here in logit and probit models are 85974 and one less than the original sample because the control variable State of one observation is unique and there are no other observations with the same value. Thus, it needs to be removed in regression.

Table 7: Effect of round numbers in Logit model and Cox model

Variables	Spec. 1	Spec. 2		Spec. 3	Spec. 4		Spec. 5	Spec. 6		Spec. 7	Spec. 8	
	Logit	Cox		Logit	Cox		Logit	Cox		Logit	Cox	
	Coef.	Coef.	HR	Coef.	Coef.	HR	Coef.	Coef.	HR	Coef.	Coef.	HR
den_num2	-1.687***	-1.120***	0.326	-1.763***	-1.204***	0.3	-1.524***	-0.986***	0.373	-1.436***	-0.913***	0.401
	(0.404)	(0.358)		(0.396)	(0.35)		(0.369)	(0.327)		(0.363)	(0.321)	
round50	-1.073	-0.668	0.513									
	(1.411)	(1.217)										
den_num2×round50	11.99**	9.019**	8259.641									
	(5.402)	(4.394)										
round100				-0.765	-0.309	0.734						
				(1.454)	(1.256)							
den_num2×round100				12.97**	9.972**	21408.58						
				(5.477)	(4.303)							
round500							-3.382*	-2.437	0.087			
							(1.829)	(1.626)				
den_num2×round500							20.72***	15.70***	6605705			
							(6.638)	(4.586)				
round1000										-3.096	-2.121	0.12
										(2.28)	(1.89)	
den_num2×round1000										17.79*	13.16**	519076.5
										(9.389)	(5.886)	
Constant	0.0894			0.09			0.0897			0.0888		
	(0.209)			(0.209)			(0.209)			(0.209)		
Control Variables	Yes	Yes		Yes	Yes		Yes	Yes		Yes	Yes	
Log L.	-34517.448	-139906.73		-34516.385	-139905.38		-34515.422	-139904.52		-34518.365	-139907.21	
Pseudo R^2	0.0591	/		0.0591	/		0.0591	/		0.0591	/	
Observations	85,974	85,975		85,974	85,975		85,974	85,975		85,974	85,975	

Note: Table 7 shows the result of regression including the interaction items of numerical information indicators and round number indicators. Spec.1, Spec.3, Spec.5 and Spec.7 are logit models. Spec.2, Spec.4, Spec.6 and Spec.8 are Cox models. Round number has been divided into four categories, indicating four levels of information accuracy. The length of descriptive text and its quadratic item, credit grade, issuing date and state information are controlled in all models. Especially for Cox models, coefficient in left part of the column and hazard ratio in right. Robust standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 8: Numerical information and hard information

Variables	Spec. 1	Spec. 2	Spec. 3		Spec. 4	Spec. 5	Spec. 6		Spec. 7	Spec. 8	Spec. 9	
	Logit	Probit	Cox		Logit	Probit	Cox		Logit	Probit	Cox	
	Coef.	Coef.	HR	Coef.	Coef.	HR	Coef.	Coef.	HR	Coef.	Coef.	HR
den_num2	-2.839***	-1.457***	-2.333***	0.097	-1.474***	-0.732***	-0.945***	0.389	-1.194***	-0.613***	-0.682**	0.506
	(0.887)	(0.469)	(0.816)		(0.365)	(0.189)	(0.319)		(0.352)	(0.185)	(0.304)	
dti	0.0193***	0.0105***	0.0112***	1.011								
	(0.0014)	(0.0008)	(0.0013)									
den_num2×*dti*	0.0826*	0.0426*	0.0813**	1.085								
	(0.0445)	(0.0241)	(0.0401)									
fico					-0.00251***	-0.0012***	-0.0041***	0.996				
					(0.0005)	(0.0002)	(0.0004)					
den_num2×*fico*					-0.0210*	-0.0097	-0.0189*	0.981				
					(0.0118)	(0.0060)	(0.0114)					
income									-0.0004***	-0.0002***	-0.0004***	0.9996
									(0.00003)	(0.00002)	(0.00003)	
den_num2×*income*									-0.0019*	-0.0009*	-0.0016*	0.998
									(0.0011)	(0.0005)	(0.0008)	
Constant	-0.306	-0.177			-0.0256	-0.0188			0.11	0.0445		
	(0.211)	(0.111)			(0.21)	(0.11)			(0.21)	(0.11)		
Control Variables	Yes	Yes	Yes		Yes	Yes	Yes		Yes	Yes	Yes	
Log L.	-34401.655	-34354.997	-139854.9		-34496.949	-34452.623	-139845.1		-34353.047	-34316.952	-139734.47	
Pseudo R^2	0.0622	0.0635	/		0.0596	0.0609	/		0.0636	0.0646	/	
Observations	85,974	85,974	85,975		85,974	85,974	85,975		85,974	85,974	85,975	

Note: Table 8 shows the result of regression including the interaction items of numerical information indicators and round number indicators. Spec.1, Spec.4 and Spec.7 are logit models. Spec.2, Spec.5 and Spec.8 are probit models. Spec.3, Spec.6 and Spec.9 are Cox models. We adopt *dti*, *fico* and *income* as hard information. In addition, *income* and *fico* score are decentralized so that we could draw consistent conclusions. The length of descriptive text and its quadratic item, credit grade, issuing date and state information are controlled in all models. Especially for Cox models, coefficient in left part of the column and hazard ratio in right. Robust standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 9: Robustness: Alternative control variables

Variables	Spec. 1	Spec. 2	Spec. 3	Spec. 4	Spec. 5	Spec. 6	Spec. 7	Spec. 8	Spec. 9	Spec. 10
dum_num	-0.124***									
	(0.0262)									
den_num1		-1.684**								
		(0.685)								
den_num2			-1.136***	-1.443***	-1.508***	-1.297***	-1.203***	-2.631***	-1.283***	-1.190***
			(0.344)	(0.401)	(0.393)	(0.365)	(0.358)	(0.891)	(0.363)	(0.349)
den_num2× round50				8.105						
				(5.332)						
den_num2×round100					8.964*					
					(5.338)					
den_num2×round500						15.57***				
						(6.029)				
den_num2×round1000							12.68			
							(8.306)			
den_num2×fico								-0.0226*		
								(0.0129)		
den_num2×dti									0.0819*	
									(0.045)	
den_num2×income										-0.00141
										(0.00114)
Constant	6.451***	6.467***	6.465***	6.447***	6.448***	6.449***	6.448***	-2.316***	6.476***	6.029***
	(0.399)	(0.399)	(0.399)	(0.399)	(0.399)	(0.399)	(0.399)	(0.214)	(0.399)	(0.402)
Original Variable	/	/	/	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Replaced Control Variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Log L.	-34313.532	-34321.324	-34318.802	-34319.535	-34318.642	-34318.273	-34320.328	-34317.21	-34317.155	-34317.732
Pseudo R^2	0.0646	0.0644	0.0645	0.0645	0.0645	0.0645	0.0645	0.0645	0.0646	0.0645
Observations	85,974	85,974	85,974	85,974	85,974	85,974	85,974	85,974	85,974	85,974

Note: Table 9 shows the result of logit models which include a bundle of new control variables. Spec.1 to 3 are for examining the pure effect of numerical information. Spec.4 to 7 are for testing round number effect. Spec.8 to 10 are for verifying the impact of previous credit status. *dti*, *fico* and *income* are also considered in Spec.1 to 7 as control variables, but their coefficients are not displayed in the table for reasons of brevity. Original variable is yes if the original level of variables *round50* $\sim$ *round1000*, *dti*, *fico* or *income* is controlled. Robust standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 10: Robustness: Alternative measurement of independent variables

Variables	Spec.1	Spec.2	Spec.3	Spec.4
den_new1	-0.834***			
	(0.259)			
den_new2		-1.983***	-3.066***	-2.363***
		(0.673)	(0.828)	(0.738)
round100			-0.783	
			(1.496)	
round100×*den_new2*			22.59***	
			(8.064)	
round500				-3.675*
				(1.881)
round500×*den_new2*				34.21***
				(9.986)
Constant	Yes	Yes	Yes	Yes
Control Variables	Yes	Yes	Yes	Yes
Log L.	-34523.58	-34523.866	-34519.021	-34518.409
Pseudo R^2	0.0589	0.0589	0.059	0.0591
Observations	85,974	85,974	85,974	85,974

Note: Table 10 shows a part of the results of logit models which adopt refined numerical information indicators. Spec.1 and Spec.2 are for examining the pure effect of numerical information. Spec.3 and Spec.4 are for testing round number effect. The length of descriptive text and its quadratic item, credit grade, issuing date and state information are controlled in all models. Robust standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 11: Robustness: Regression results with 60% of the original sample

Variables	Spec.1	Spec.2	Spec.3	Spec.4
den_num1	-2.398***			
	(0.917)			
den_num2		-1.383***	-1.439***	-1.231***
		(0.447)	(0.507)	(0.474)
round100			-1.891	
			(1.871)	
round100×*den_num2*			11.27*	
			(6.517)	
round500				-4.939**
				(2.499)
round500×*den_num2*				17.61**
				-8.454
Constant	Yes	Yes	Yes	Yes
Log L.	-20779.034	-20777.838	-20777.007	-20776.048
Pseudo R^2	0.0577	0.0577	0.0577	0.0578
Observations	51,583	51,583	51,583	51,583

Note: Table 11 shows a part of the results of logit models with 60% of the original sample. Spec.1 and Spec.2 are for examining the pure effect of numerical information. Spec.3 and Spec.4 are for testing round number effect. The length of descriptive text and its quadratic item, credit grade, issuing date and state information are controlled in all models. Robust standard errors in parentheses. $*p < 0.1$, $**p < 0.05$, $***p < 0.01$.

Table 12: Robustness: Heckman two-step regression model

Variables	Spec.1	Spec.2	Spec.3
	Response equation (*loan_status*)		
dum_num	-0.0165***		
	(0.0028)		
den_num1		-0.2454***	
		(0.0718)	
den_num2			-0.1545***
			(0.0373)
Constant	Yes	Yes	Yes
Control Variables	Yes	Yes	Yes
	Selection equation (*dum_desc*)		
fico, *inq_last_6mths*, *bc_util*, *emp_length*	Yes	Yes	Yes
Inverse Mills ratio	0.0326**	0.0337**	0.0334**
total sample	380,516	380,516	380,516
selected sample	87,274	87,274	87,274

Note: Table 12 shows the result of Heckman model. The dependence variable of the selection equation is a dummy variable indicating whether the borrower provides a descriptive text. Control variables in benchmark model are all adopted in response equation and four other control variables-*fico*, *inq_last_6mths*, *bc_util* and *emp_length* are considered in selection equation. Robustness standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 13: Default rate and per capita GDP in different states

State	Default rate	GDP 2012	GDP 2013	GDP 2014	State	Default rate	GDP 2012	GDP 2013	GDP 2014
NE	50.00%	50,733	51,694	52,317	WI	14.87%	45,639	45,811	46,676
AR	19.33%	35,753	36,571	37,252	WY	14.71%	60,362	59,961	60,189
AL	18.45%	36,695	36,865	37,083	MO	14.64%	41,756	42,231	42,582
KY	18.27%	38,310	38,564	38,860	CT	14.32%	63,178	62,847	63,650
TN	18.02%	41,469	41,642	41,955	GA	14.18%	41,740	42,082	42,694
OK	17.85%	41,769	42,475	43,588	UT	14.17%	42,282	42,796	43,154
NY	17.35%	62,777	62,610	63,217	WA	14.15%	53,622	53,760	54,662
IN	17.24%	43,056	43,774	44,539	IL	14.08%	52,264	51,812	52,359
HI	17.19%	49,216	49,035	49,073	CA	14.04%	52,660	53,487	54,606
MD	17.14%	53,824	53,342	53,879	TX	13.90%	50,356	51,695	52,742
NJ	17.05%	55,571	55,562	55,841	KS	13.82%	45,701	45,173	45,570
FL	16.93%	37,702	38,090	38,497	MA	13.70%	62,205	61,352	62,145
NC	16.40%	42,746	42,852	43,332	OR	13.36%	49,403	48,323	48,264
MI	16.27%	40,537	41,117	41,695	SC	12.74%	35,512	35,519	35,962
PA	16.26%	46,648	47,159	48,203	MT	11.81%	38,318	38,395	38,820
RI	15.90%	46,681	46,909	47,852	CO	11.71%	50,547	50,807	52,060
MN	15.88%	51,598	52,479	53,499	VT	11.36%	42,978	42,771	43,435
NV	15.78%	43,209	42,505	43,216	SD	11.17%	47,171	47,172	46,917
NM	15.59%	40,226	39,808	40,866	NH	11.11%	47,990	47,975	48,982
OH	15.48%	44,394	33,736	46,137	WV	10.83%	35,479	36,038	36,337
LA	15.36%	45,813	44,408	45,067	AK	10.69%	74,289	69,746	67,366
AZ	15.22%	38,732	38,303	38,438	DC	8.33%	74,125	72,499	72,092
VA	15.17%	51,942	51,347	51,101	ME	0.00%	37,748	37,447	37,807
DE	15.17%	62,327	61,004	62904					

Note: Table 13 shows the default rate and per capita GDP of each state in the United States. The correlation between the default rate and the per capita GDP is significantly negative, meaning that controlling the state information represents the partly covering of macroeconomy impacts.