

Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending

Cuiqing Jiang¹ · Zhao Wang¹ · Ruiya Wang¹  · Yong Ding¹

Published online: 4 October 2017
© Springer Science+Business Media, LLC 2017

Abstract Predicting whether a borrower will default on a loan is of significant concern to platforms and investors in online peer-to-peer (P2P) lending. Because the data types online platforms use are complex and involve unstructured information such as text, which is difficult to quantify and analyze, loan default prediction faces new challenges in P2P. To this end, we propose a default prediction method for P2P lending combined with soft information related to textual description. We introduce a topic model to extract valuable features from the descriptive text concerning loans and construct four default prediction models to demonstrate the performance of these features for default prediction. Moreover, a two-stage method is designed to select an effective feature set containing both soft and hard information. An empirical analysis using real-word data from a major P2P lending platform in China shows that the proposed method can improve loan default prediction performance compared with existing methods based only on hard information.

Keywords P2P lending · Default prediction · Soft information · Topic model

1 Introduction

P2P (Peer-to-Peer) lending, the practice of lending money to individuals or businesses through online services that match lenders with borrowers, is emerging as a novel mode of Internet finance. Because the peer-to-peer lending companies offering these services generally operate online, they operate with lower overhead and can provide less expensive services than traditional financial institutions such as banks. Given its feasibility and convenience, P2P lending has experienced rapid development in recent years, in terms of both the number of P2P platforms and the volume of transactions. However, P2P lending also faces challenges, such as information asymmetry. Considering the characteristics of the online platform, the

✉ Ruiya Wang
wrylr@163.com

¹ School of Management, Hefei University of Technology, Hefei, Anhui, China

P2P model has greater risk when extending credit than do traditional lending modes (Pope and Sydnor 2011). It is particularly important to use an effective method to predict the default probability of each loan and evaluate the borrower's credit risk more carefully.

For borrowers, P2P lending offers lower access thresholds than traditional bank lending. In P2P lending, the borrowers are mostly small and medium business owners, entrepreneurs, and some low-income earners: some are borrowers who have been rejected by banks because they have no collateral or guarantee. The characteristics of these customer groups limit for the traditional personal credit evaluation methods. Malekipirbazari and Aksakalli (2015) suggested that even borrowers with high credit ratings according to their financial information may also have a high probability of default. Some important features, such as guarantees, may lose their ability to determine bad borrowers in P2P lending. To compensate for these limitations, P2P platforms must attempt to discover other factors that influence loan defaults in rich user data such as soft information (Angilella and Mazzù 2015).

In economic activities, if sufficient objective quantitative information (i.e., hard information) concerning the parties to the transaction is not available or the information is unreliable, subjective and qualitative information (e.g., text information) can be used to support decision making (Michels 2012). Specifically, in the P2P market, soft information associated with borrowers' repayment ability and repayment willingness may contain valuable factors related to loans that could mitigate the hazards from a lack of hard information on default prediction (Cornée 2017). Soft information is often communicated in text (e.g., descriptive loan text) and could be derived from it (Liberti and Petersen 2017). The descriptive text of a loan application, which is generated during the transaction process and uploaded by the users themselves, contains a detailed description about the loan made during the loan application and may include the borrower's background, loan purpose, loan type and so on. Consequently, such text can be a good information source for credit risk evaluation (Wang et al. 2016).

In this paper, we propose a default prediction method that considers soft information extracted from descriptive P2P lending text. Employing the descriptive text of a loan application as an example, we use a text analysis method and latent Dirichlet allocation (LDA) model to extract soft features (i.e., credit-related topics) from the description text to complement the applicant's hard information. Considering the complexity and diversity of the data (i.e., hard and soft information) collected by P2P platforms, a two-stage feature selection method, using a variety of metrics, is designed to select the appropriate feature set before modeling. We also evaluate our proposed method using a large real-world dataset from a major P2P institution in China. We build four default prediction models, including logistic regression, Naïve Bayes, support vector machines and random forest, and use a full factorial design (i.e., hard information only, soft information only and a combination of hard and soft information) to evaluate the models' discrimination performances.

The remainder of this paper is organized as follows. In Sect. 2, we present a review of the literature on credit evaluation in P2P lending. Section 3 provides a description of the data used for analysis. The procedure to extract soft features is introduced in Sect. 4. Section 5 is divided into two parts, the first part describes the feature selection method for both soft and hard information, and the second part describes the prediction modeling method. An empirical analysis is presented in Sect. 6. Finally, we conclude the paper and discuss future research directions in Sect. 7.

2 Literature review

Credit in P2P lending is typical individual consumer credit. The scoring model generally involves the probability of default (PD), the exposure at default (EAD), and the loss given default (LGD). PD models are especially well researched and continue to attract considerable interest (Lessmann et al. 2015). At present, most of the research on P2P credit evaluation regards it as a classification problem: a loan is classified as either credit worthy or credit worthless according to the PD. Evaluation methods are divided into two categories: statistical analysis methods and machine learning methods. Statistical analysis methods such as discriminant analysis and logistic regression have become the standard method for default forecasting (Crook et al. 2007). Emekter et al. (2015) used the logistic model in combination with FICO points and credit levels available to the platform to predict the default probability for P2P borrowers. However, the statistical analysis method imposes strict statistical assumptions on the evaluation data (Kruppa et al. 2012). In practical applications, most credit features meet these conditions only with difficulty. Therefore, machine learning and artificial intelligence models have gradually been applied, obtaining a good prediction effect. These models include neural network (Hájek 2011), random forest (Malekipirbazari and Aksakalli 2015), support vector machine (Harris 2013; Yao et al. 2015) and so on.

In recent years, ensemble models have become a new hotspot in credit evaluation models (Kruppa et al. 2013). An ensemble model consists of several base classifiers arranged according to some ensemble strategy. Finlay (2011) designed ensemble models with three different structures to evaluate the consumer's credit risk. The results showed that not all the ensemble models are superior to a single classifier model, but the bagging and boosting model achieved better classification results. Similarly, Wang et al. (2012) proposed two ensemble credit scoring models using decision trees based on the ensemble strategies of bagging and random subspace and achieved a relatively higher classification accuracy than other standard models. These methods can reduce the effects of noise and redundant features in the data. In general, the non-statistical analysis models based on machine learning are more robust than the statistical models, and ensemble classification models are typically more advantageous for credit evaluation problems.

Data for the individual credit model stem mainly from application forms, customer demographics, and transactional data from the customer's history (Thomas 2010). It is critical to choose appropriate features for credit modeling. In P2P lending, credit evaluation applies to an individual; therefore, its characteristics are related to individual credit aspects such as repayment ability and repayment willingness (Shao et al. 2012). Financial and asset status can effectively reflect repayment ability (Paul 2014), while features such as gender, age, marital status, and educational level can effectively reflect a user's repayment willingness (Abdou and Pointon 2011). Several existing studies have used features based on financial factors such as income levels, assets and liabilities to predict P2P borrowers' credit (Emekter et al. 2015; Guo et al. 2015); but P2P lending platforms have difficulty obtaining sufficient borrow financial data. Many other features can also reflect credit status directly or indirectly. The credit bureau Zestfinance collected a large amount of data from the Internet and from third parties and extracted abundant credit features that can reflect users' credit in Internet finance. Michels (2012) reported that subjective and qualitative information may offer additional judgment criteria when the available objective information is insufficient. Soft information is such a form of subjective information. It is generally non-standardized or personalized and difficult to quantify, test and summarize (Liberti and Petersen 2017). In contrast, the relative information is called hard information. In the financial scene, hard information and

soft information are not clearly delineated. Some studies define financial information as hard information and non-financial information as soft information (Angilella and Mazzù 2015).

There have been some studies on the relationship between soft information and borrower behavior (Puro et al. 2010). Lin et al. (2013) analyzed the role of social relations in assessing credit risk and found that strong social network relationships were important factors in determining the success of borrowing and in reducing the risk of default. Gao and Lin (2016) found that the readability of the loan application text and its inclusion of positive emotions and fraud clues were related to the probability of default in P2P borrowing. Iyer et al. 2015 further analyzed the default prediction performance of descriptive text characteristics, such as whether the user voluntarily stated the loan purpose, to predict the probability of default and demonstrated that soft information has some influence on P2P lending behavior. Dorfleitner et al. (2016) examined the relationships between soft features (such as text length, spelling errors, and the presence of positive emotion-evoking keywords) and the funding and default rate of P2P lending. Although the results of this study show that textual soft features had little predictive value for default probabilities, the mention of keywords related to business and education were significantly related to the probability of default. This indicates that—although the impact of linguistic text features on default is not significant—the content of text may be able to provide more valuable information for default prediction.

3 Data analysis

The data used in our analysis were collected from a major P2P lending platform (Eloan) of China. Farmers, individual businesses and small micro-owners are the main services targets, and the platform itself is not a participant but an information provider. After a borrower uploads their personal identification information and loan application information, the platform will carry out an examination including information validation and fraud detection (i.e., check whether the information is valid and the borrower is a fraudster). After the examination (i.e., information validation and fraud detection) is complete, the platform will provide a rating of the borrower for investors and an interest rate suggested by platform for the loan application, then the loan information is published. Subsequently, lenders choose whether to invest in accordance with the corresponding information. When a loan is funded successfully and the borrower has confirmed receipt of the money, the loan enters the repayment period.

The data were collected for the period between January 2014 and January 2016 and contain approximately 132 K borrower records. To obtain the repayment status for each sample, we chose records for which the loan was funded successfully and the repayment period had ended. The number of such records was approximately 40 K. A loan default was defined as one for which the repayment period had expired but overdue funds remained; in that case, the borrower has defaulted. According to the overdue records in the dataset, we divided the loans whose repayment period had ended into 2 categories. If the overdue amount was 0, the loan was considered non-default; otherwise, it was considered as a default. Accordingly, there were 39,538 loan records in the sample for the evaluation, of which 32,738 were good (non-default) loans and 6800 were default loans.

The lending platform (Eloan) contains a total of 27 features, of which 9 are personal information features, 11 are financial status features, and 6 are loan information features. The personal information is the borrower's demographic information, including the borrower's age (Age), gender (Gender), occupation (Occupation), job title (JobTitle), years of work (WorkYears), education status (Education), marital status (Marriage), province (Province)

and type of residence (*ResidenceType*). The financial status responds to the borrower's ability to repay and includes features such as the borrower's income level (*Income*) and house ownership (*House*). The guarantees situation reflects whether the borrower is using a house as guarantee (*GuaranteeHouse*), a car (*GuaranteeCar*), or has a guarantor (*Guarantor*). The insurance purchase situation includes business insurance (*BInsurance*) and Social insurance (*SInsurance*). It also contains the borrower's historical loan information, such as the remaining repayment period of existing loans (*LoanRecord*) and the number of failed bids (*FailBids*). The borrower's credit card limit (*CreditLimit*) and whether there a credit report from a bank exists (*CreditReport*) are also important indicators of a borrower's financial status. In addition, some loan information features may exist, including the loan amount (*Amount*), interest rate (*IR*), and repayment period (*RepaymentPeriod*). Depending on the borrowing purpose, loans can be divided into different classes (*LoanType*), such as agricultural loans, enterprise loans and capital turnover. The *PaymentMode* denotes the borrower's repayment method, including average capital, average capital plus interest and repay capital with interest. In addition, there is a short descriptive text for each loan application (*LoanDescription*) regarding the loan's purpose and/or the applicant's personal situation.

The loan application's descriptive text is a subjective description of the loan when the borrower applied for it. The text content may include information related to the borrower's personal background, borrowing purposes, asset status and so on. This descriptive content can both enhance the transparency of the borrower's information and reflect the borrower's attitude to the loan application. The more detailed the description of the loan description is, the more useful the information involved is. Thus, such information indirectly reflects a borrower's willingness to pay back the loan. Because text is a type of unstructured information, an effective quantification and extraction method is needed to analyze its relationship with loan defaults and user credit.

The summary statistics of the features used in our analysis are shown in Tables 1 and 2. Because a loan description is a type of unstructured text, we describe the extraction of several quantifiable features (i.e., soft information) from it in the next section. These features can then be used to predict the borrower's default probability in P2P lending.

4 LDA-based soft features extraction

4.1 Latent Dirichlet allocation model

The LDA model (Blei et al. 2003) is a kind of unsupervised theme model, whose main idea is to regard each document as a mix of the probability distributions of all topics and each topic as a probability distribution of words. It consists of three layers: document sets, documents, and words. A graphical representation of the LDA model is shown in Fig. 1.

Here, the Dirichlet prior parameter α reflects the relative strength between latent topics, while β represents the probability distribution of all latent topics. The parameter θ is the probability distribution of the target text in the latent topic, and φ is the probability distribution of the topic in words. T is the number of topics, and z indicates the number of latent topics of each document assigned to each word. W denotes the word vectors for the document, N represents the number of words in a document, and M is the number of documents in the document set.

Table 1 Category features used in analysis

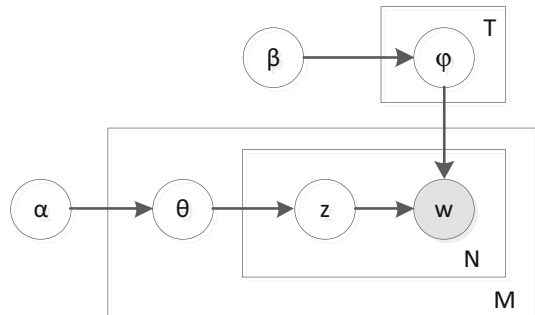
No.	Features	Description
1	Gender	{male (0.83), female (0.17)}
2	Occupation	{company staff (0.72), civil servant (0.24), other (0.04)}
3	JobTitle	{student/uncertain title (0.57), junior executive (0.17), laborial staff (0.09), middle executive (0.09), senior executive (0.08)}
4	Work Years	{< 3 years (0.57), 3–5 years (0.17), > 5 years (0.22), unemployed (0.04)}
5	Education	{junior high school or below (0.52), senior high school/technical secondary school (0.36), bachelor/junior college or above (0.12)}
6	Marriage	{married (0.82), unmarried (0.11), divorced (0.07)}
7	Province	{Shandong (0.22), Hebei (0.16), Inner Mongolia (0.18), Shanxi (0.07), Henan (0.12), Liaoning (0.09), other (0.16)}
8	ResidenceType	{rural residential (0.42), city community (0.21), urban community (0.34), other (0.03)}
9	Income	{< 6w (0.08), 6.1–12w (0.27), 12.1–24w (0.47), 24.1–40w (0.15), > 40w (0.03)}
10	House	{own (0.77), mortgage (0.12), parent's house (0.07), rent/no house (0.04)}
11	GuaranteeHouse	{no (0.96), yes (0.04)}
12	GuaranteeCar	{no (0.75), yes (0.25)}
13	Guarantor	{no (0.97), yes (0.03)}
14	BInsurance	{no (0.68), yes (0.32)}
15	SInsurance	{no (0.16), yes (0.84)}
16	LoanRecord	{no record (0.54), 1–3 years (0.29), 3–5 years (0.07), > 5 years (0.10)}
17	CreditLimit	{none (0.58), < 2w (0.30), > 2.1w (0.12)}
18	CreditReport	{no (0.68), yes (0.32)}
19	LoanType	{capital turnover (0.67), personal consumption (0.15), other (0.18)}
20	PaymentMode	{repay capital with interest (0.95), average capital plus interest (0.04), constant amortization (0.01)}

The unit of money is RMB (¥) 1w = ¥10,000

Table 2 Continuous features used in analysis

No.	Features	Min	Max	Mean	SD
1	Age	20	64	38.174	8.655
2	FailBids	0	8	0.057	0.27
3	Amount	500	20,000,00	55,212.77	28,102.55
4	IR (%)	0.01	25	19.067	1.261
5	RepaymentPeriod	2	24	10.297	2.407

Fig. 1 A graphical representation of the LDA model



If j is a latent topic, and w_i is the i th word in document d , the probability that w_i belongs to j is computed by:

$$P(w_i) = \sum_{j=1}^T P(w_i|z_i = j) P(z_i = j),$$

where $P(w_i)$ is the probability that w_i appears in a given document d , which is an observable known variable for each document. $P(w_i|z_i = j)$ is the probability that w_i belongs to j , that is, φ . $P(z_i = j)$ indicates the probability that j is a topic in d . Both of these variables follow the Dirichlet distribution of parameters α and β . Similarly, the probability that the document d contains word w can be computed as follows:

$$P(w|d) = \sum_{j=1}^T \varphi_w^j \cdot \theta_j^d.$$

The LDA model is used to estimate the parameters of θ and φ . Common estimation methods are the EM algorithm, Gibbs sampling and so on. In this paper, Gibbs sampling (Wei and Croft 2006) is used to estimate the parameters, and the samples close to the limit distribution are sampled by constructing a Markov chain, which converges to a certain limit distribution.

4.2 Soft features abstracted from descriptive loan text

In this paper, we used the LDA model to extract topics from the descriptive text that accompanies loan applications and generate topic-features. The documents include all the descriptive text paragraphs related to borrowing. We calculate the probability distribution of the text in each topic. The topic-features are the quantitative results (i.e., soft information) present in the descriptive loan text, and the probability is the value of each corresponding feature. In

Table 3 Topic features abstracted from descriptive loan text

Topic	Words with high occurrence probability
T1 Asset	房产 (House), 车 (Car), 楼房 (Building), 货车 (Truck), 资金 (Funds), 自有 (Self-owned), 承包 (Contract)
T2 Income	收入 (Income), 稳定 (Stable), 借款 (Loan), 还款 (Repayment), 逾期 (Overdue), 元 (Yuan), 万 (Million)
T3 Work	公司 (Company), 工作 (Job), 担任 (Served), 上班 (Working), 经理 (Manager), 职务 (Duty), 业务 (Business), 行业 (Industry)
T4 Family	家庭 (Family), 和睦 (Harmonious), 已婚 (Married), 子女 (Children), 幸福 (Happiness), 关系 (Relationship), 家人 (Family member)
T5 Business	经营 (Manage), 资金 (Funds), 设备 (Equipment), 销售 (Sales), 进货 (Stock), 批发 (Wholesale), 客户 (Client), 零售 (Retail)
T6 Agriculture	种植 (Planting), 规模 (Scale), 养殖 (Breeding), 田 (Farmland), 亩 (Acre), 化肥 (Fertilizer), 土地 (Land), 蔬菜 (Vegetables)

other words, this approach measures the degree of relevance of the descriptive text of the loan application to each topic.

We used the LDA model implemented by the open source Gibbs-LDA project to train the topic model using each loan description, in which the important parameter is the number of topics. To determine whether a parameter was appropriate, we tested extracting different numbers of topics and evaluated whether the extracted subject contained well-defined and related words. Finally, we determined that the model generated the best results when the number of topics was 15. There are six topics whose meanings are obvious: assets, income and expenses, work, family, business and agriculture. The topics extracted from the descriptive loan text and the words that have a high occurrence probability for each topic are shown in Table 3. The second column (i.e., words with high occurrence probability) list the keywords that have a high probability of appearing in a certain topic. (e.g., the word “Yuan” is highly likely to appear in the topic “Income”). Note that the high occurrence probability words in a topic may also appear in other topics, but their occurrence probability is relatively lower (e.g., the word “Yuan” may also appear in the topic “Asset”).

The number of characters in the text (Length) is another important feature of the loan description and is a linguistic feature of the text. Combined with the above six topic-features, we converted the descriptive loan application text to a total of 7 soft features.

5 Default prediction model

In this paper, we constructed default prediction models combined with the soft features extracted from the descriptive loan text in the P2P platform based on 4 different classifica-

tion methods. One critical step in any classification effort is to select features that have good predictive power for the response variable, especially considering that large numbers of unrelated or redundant features will increase training complexity and decrease model accuracy (Hajek and Michalak 2013). Therefore, we designed a feature selection method for hard and soft information to ensure the effectiveness of the prediction model.

5.1 Two-stage feature selection method

When selecting methods for default prediction, the Pearson coefficient, statistical tests (one-way analysis of variance), and stepwise procedures have been used previously in various combinations (Hajek and Michalak 2013). However, these methods simply filter features individually, ignoring any feature redundancy or various types of feature combinations. Because each selection method has its own metrics, there are no obvious advantages or disadvantages between them. Therefore, we combined a variety of criteria to obtain a comprehensive measure of features; this combination also reduces the rate at which feature subsets change. By combining criteria, it is possible to obtain a better result and improve the stability of the credit evaluation results than when using a single method.

We designed a two-stage selection method that involves sequence procedure first and wrapper procedure second. First, we use three metrics to sequence all features according to their importance and obtain 3 sorted results. Then, we integrate them by voting to obtain a comprehensive ranking. Second, to ensure the utility of the combined features, a wrapper performs feature subset selection based on the AUC of the classification model. The detailed procedure is shown in Fig. 2.

(1) Comprehensive ranking

An information metric is a nonparametric and nonlinear method that can quantify the uncertainty associated with features used for classification while simultaneously providing both numeric and nominal features. In previous studies, statistical values such as the chi-square test and Pearson correlation coefficient have been used to measure feature importance and determine the statistical correlations among them. However, the data distribution of soft information is uncertain, and large nonlinear relations exist; therefore, it is difficult to measure the relations of soft features to loan default accurately based solely on statistical correlations. Here, we combine statistical correlation metrics with information metrics and select three types of criteria, the chi-square test, information gain and information gain ratio, and use these to sequence features according to their importance. We then conduct a vote on the three ranking results to obtain the final comprehensive ranking results.

(2) Wrapper selection

To eliminate the redundant features in subsets and ensure the utility of combining hard and soft information, a wrapper is added to the feature set after sequencing. First, all features in the set serve as input features to the model. The overall utility of each set is evaluated based on the model accuracy. Then, by combining the ranking results from (1), we use a sequential backward selection (SBS) method to successively eliminate the features that have the lowest rank, which generates a new subset that is input into the model to obtain a new accuracy. The preceding steps are repeated and the classification accuracy of all the feature subsets is compared. Finally, we select the feature set with the highest accuracy to serve as the final feature set for the default prediction model.

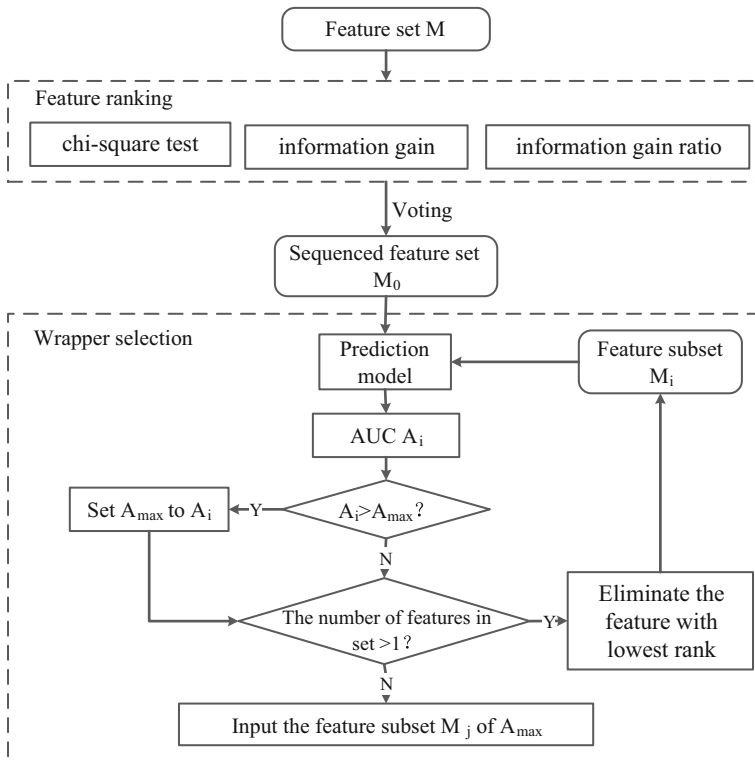


Fig. 2 Flow of two-stage selection method

5.2 Default prediction models

Default prediction is generally regarded as a classification problem that determines whether a borrower will default. The existing modeling methods can be divided into statistical analysis methods and machine learning methods. The traditional statistical analysis model generally has strict statistical assumptions or data distribution requirements; in the actual modeling process, most of the features meet these conditions only with difficulty. However, machine learning methods can ameliorate this deficiency. In this paper, we used Logistic Regression, Naïve Bayes, Support Vector Machine and Random Forest models to build default prediction models. Then, we compared their classification performances through empirical analysis.

(1) Logistic regression

The logistic regression (LR) model (Landwehr et al. 2005) is a generalized linear model that can solve classification problems. The model uses the maximum likelihood estimation method to establish a linear regression classification model to classify binary or multiple valued variables. Assuming that the dependent variable, y , is a binary variable (0, 1), x is the independent variable, and y follows the following probability distribution, then the formula is:

$$P(y_i = 1|x_i) = P[\varepsilon_i \leq (\alpha + \beta x_i)] = \frac{1}{1 + e^{\alpha + \beta x_i}}$$

where β is a vector of $n \times 1$, corresponding to different explanatory independent variables of the coefficient vector.

In the default prediction model, borrower default is a categorical variable, and the borrower’s personal information, financial information and other features are explanatory variables. The model conducts a regression analysis for each explanatory variable and determines the borrower’s default status according to the resulting probability. The default forecast model is shown below:

$$y_i = \ln \left(\frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n,$$

$$p_i (y_i = 1|X) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)]}.$$

(2) Naïve Bayes

The Naïve Bayes (NB) classification model (Friedman et al. 1997) is based on Bayesian theory and assumes that the features are conditionally independent. Based on training samples, the prior probabilities of each class and the conditional probability of each feature are obtained. Then, the Bayesian formula is used to calculate the probability of the sample belonging to each class, and finally, the sample category is determined according to the probabilities. If sample X has a total of v categories, C_v , composed of an attribute vector $A = \{a_1, a_2, \dots, a_m\}$, the probability that sample x_i belongs to class c_i is:

$$P(C_i|X) > P(C_j|X) \quad 1 \leq j \leq v, i \neq j,$$

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)},$$

$$P(X|C_i) = \prod_{k=1}^n P(X_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \dots \times P(x_n|C_i).$$

(3) Support vector machine

The support vector machine (SVM) (Cortes and Vapnik 1995) is a machine learning method based on statistical learning theory that has unique advantages for solving small sample, nonlinear and high-dimensional spatial pattern recognition problems. The basic idea is to map the input space to a high-dimensional feature controlled through a non-linear mapping, ϕ , using the principles of structural risk minimization and classification interval maximization to determine the optimal hyperplane for classification, $f(x) = w^T \phi(x) + b$, in this high-dimensional space, where w and b are the weight and thresholds of the hyperplane, respectively. The classification problem can be expressed as the following constrained quadratic programming problem:

$$\max \left(\sum_{l=1}^n \alpha_l + \frac{1}{2} \sum_{l,r=1}^n y_l y_r \alpha_l \alpha_r K(x_l, x_r) \right),$$

$$s.t. \quad 0 \leq \alpha_l \leq C, l = 1 - n, \sum_{l=1}^n y_l \alpha_l = 0,$$

where C governs the relative cost of each objective in the optimization process and α is a Lagrange multiplier associated with instance l .

(4) Random forest

The random forest (RF) model (Breiman 2001) is a combination of the CART (classification and regression tree) and bagging methods. It uses bootstrap sampling to take samples from the original sample collection and builds a decision tree for each sample. Finally, it combines all the trees and takes their voting results as the final classification. The RF method

has high accuracy, tolerates outliers and noise well, and is not easy to overfit. Some model parameters are important during training, including the number of split attributes, m , and the forest size, k . When M variables exist in the sample, usually $m \approx \sqrt{M}$ or $m \approx \log_2 M$. Considering the imbalance of default data, splitting is done with respect to the Gini Index, which is described below.

For a tuple D with N classes, the Gini index can be calculated as follows:

$$Gini(D) = 1 - \sum_{i=1}^N p_i^2,$$

where p_i is the probability that the tuple in D belongs to class i . If $Gini(D) = 0$, all the tuples in D belong to the same class. When D is divided into D_1 and D_2 , the Gini index of the split is

$$Gini_{\text{split}}(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2),$$

where $|D|$ indicates the number of tuples in D .

6 Empirical analysis

6.1 Experiment design

We evaluated the performance of soft features extracted from descriptive text on the complete dataset, including the extracted soft features described in Sect. 4.2 and the hard features from the dataset (i.e., Tables 1, 2).

Before modeling, we first used the proposed feature selection method described in Sect. 5.1 to find the optimal feature subset for each classifier; then, we evaluated the discrimination performance of each soft feature using fivefold cross-validation (CV). It has been shown that k -fold cross-validation can generally provide a good trade-off between model overfitting and underfitting (Huang et al. 2007). During fivefold cross validation, the original dataset is randomly partitioned into five subsets of equal size. Of the 5 subsets, a single subset is retained as test data, while the remaining 4 subsamples are used as training data. This cross-validation process is then repeated 5 times. To assess the performance of the classifiers, we took the average performance over the 5 CV folds of the following two metrics: (1) overall classification accuracy rate (ACC) on the test slice in the fold, and (2) the area under the receiver operating characteristic curve (AUC). In addition, we validated the statistical significance of all the soft features using the logistic regression model to determine the impact of the soft features extracted from the descriptive loan text.

6.2 Feature selection results

Our proposed two-stage method was performed to select features before modeling. We use three types of criteria to sequence features and obtain the final comprehensive ranking results by vote as shown in Table 4. The sorting results show that there are both some differences and some commonalities between the three sorted results. The province, family topic-feature, repayment period, and repayment method are most important in the three metrics, while the importance of guarantor, gender, age, job title and marital status is lower when using these three metrics. An analysis of the statistical information of these features indicates that this

Table 4 Sequence and wrapper results of features

Rank	Chi-square	IG	GR	Comprehensive rank
1	Province	Province	PaymentMode	Province
2	T4-family	T4-family	GuaranteeCar	T4-family
3	RepaymentPeriod	RepaymentPeriod	GuaranteeHouse	RepaymentPeriod
4	T3-Work	PaymentMode	Province	PaymentMode
5	T1-Asset	T3-Work	CreditReport	T3-Work
6	T2-Income	T1-Asset	RepaymentPeriod	T5-Business
7	T6-Agriculture	T2-Income	T4-family	T1-Asset
8	T5-Business	T5-Business	SInsurance	T2-Income
9	PaymentMode	T6-Agriculture	T5-Business	T6-Agriculture
10	Text-Length	Text-Length	BInsurance	SInsurance
11	Amount	Amount	FailBids	Text-Length
12	CreditLimit	SInsurance	IR	Amount
13	SInsurance	CreditLimit	T3-Work	IR
14	ResidenceType	ResidenceType	T6-Agriculture	CreditLimit
15	WorkYears	IR	T1-Asset	ResidenceType
16	IR	WorkYears	T2-Income	BInsurance
17	Occupation	Occupation	Amount	WorkYears
18	Income	Income	Text-Length	Occupation
19	House	House	CreditLimit	FailBids
20	LoanRecord	LoanType	Occupation	GuaranteeCar
21	BInsurance	LoanRecord	ResidenceType	House
22	Education	Education	WorkYears	CreditReport
23	LoanType	BInsurance	House	Income
24	FailBids	FailBids	LoanType	GuaranteeHouse
25	JobTitle	JobTitle	Education	LoanType
26	Marriage	Marriage	Income	LoanRecord
27	Age	Age	LoanRecord	Education
28	Gender	Gender	Marriage	Marriage
29	CreditReport	CreditReport	Guarantor	JobTitle
30	GuaranteeCar	GuaranteeCar	JobTitle	Age
31	GuaranteeHouse	GuaranteeHouse	Gender	Gender
32	Guarantor	Guarantor	Age	Guarantor

result may occur because P2P loans are a network-based platform whose user groups are mainly young and middle-aged people: for example, there are no obvious age differences between borrowers. In addition, P2P borrowers mainly have low incomes, most of whom cannot provide guarantees; consequently, these features do not currently aid in identifying P2P loan defaults.

It is apparent that the soft features extracted from the loan description text play a relatively important role among all features. The topic-features of family, work and business gain high rankings among all features, much better than most of the hard features. Based on the sorting results, we used the previously described wrapper to select the most effective

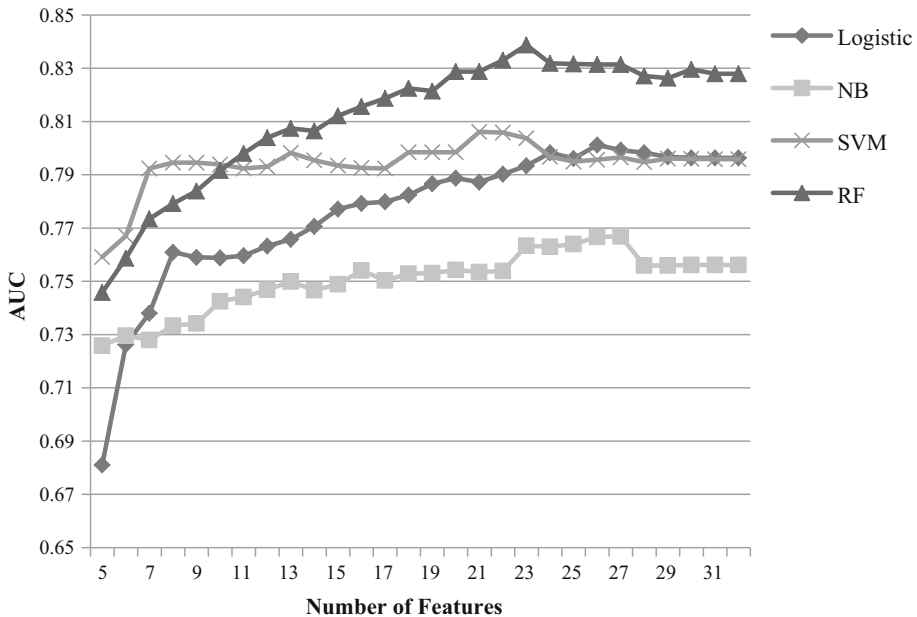


Fig. 3 The influence of the number of features on classification performance

feature subsets for different classifiers. Because the number of features is a hyper-parameter, we constructed a validation set to search for the optimal number of features. In the hyper-parameter optimization process, a new training set and validation set were first generated from the original training set in the fivefold cross validation using a bootstrapping method. After sampling with replacement, a bootstrap sample with the same sample size was obtained to train different classifiers, while the unselected (out-of-bag) samples eventually formed a validation set. The classification performance in terms of AUC resulting from different numbers of features are shown in Fig. 3.

The optimal number of features corresponding to each model is different. For the logistic regression model, the optimal number of features is 26; for NB it is 27; for SVM it is 21; and for RF it is 23. The figure shows that the performance of model does not necessarily improve as the number of features increases. When an invalid feature is included, model classification performance can even decline.

6.3 Discrimination performance of soft features

According to the feature selecting result, we selected 23 features, including 16 hard features and 7 soft features, to evaluate the effectiveness (i.e., discrimination performance) of soft information extracted from the descriptive text. We compared 3 types of feature sets, including hard feature, soft features, and their combination using the logistic regression, Naïve Bayes, SVM and RF models. The four models were trained and evaluated via fivefold cross validation, and the splitting of folds was kept identical across models. The results of discrimination performance in terms of accuracy and AUC are shown in Table 5.

The four models built with both hard and soft features exhibited the optimal classification performance, and the soft features contributed to improvement in discrimination performance

Table 5 Discrimination performance of four prediction models

Feature set	Logistic		Naïve Bayes		SVM		RF	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Hard features	0.834	0.780	0.727	0.754	0.842	0.754	0.857	0.817
Soft features	0.837	0.685	0.809	0.672	0.840	0.669	0.844	0.737
Hard + Soft features	0.852	0.797	0.776	0.759	0.854	0.800	0.860	0.840

Table 6 Discrimination performance of each soft feature

Set	Features	Logistic		Naïve Bayes		SVM		RF	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
1	Hard	0.834	0.780	0.727	0.754	0.842	0.754	0.857	0.817
2	Set 1 + T1	0.836	0.782	0.728	0.754	0.845	0.762	0.858	0.819
3	Set 2 + T2	0.835	0.782	0.736	0.749	0.846	0.765	0.857	0.820
4	Set 3 + T3	0.837	0.784	0.757	0.750	0.848	0.765	0.857	0.822
5	Set 4 + T4	0.848	0.792	0.782	0.757	0.851	0.782	0.858	0.827
6	Set 5 + T5	0.850	0.792	0.779	0.758	0.852	0.787	0.859	0.833
7	Set 6 + T6	0.850	0.794	0.775	0.756	0.852	0.791	0.859	0.837
8	Set 7 + LEN	0.852	0.797	0.776	0.759	0.854	0.800	0.860	0.840

for all four models. Specifically, the RF method achieved the best performance; the accuracy and AUC of the RF model based entirely on hard features were 0.857 and 0.817, respectively, but after adding the soft features extracted from the descriptive text, its prediction accuracy and AUC increased to 0.860 and 0.840, respectively. The SVM model obtained the second highest accuracy and AUC (i.e., 0.854 and 0.800 respectively) when soft features were added. Moreover, modeling using soft features alone achieved approximately a 0.691 (i.e., mean of the AUC of four models) level of AUC. The results show that soft features have a definite classification relationship with loan default.

We further evaluated the discrimination performance of each soft feature by adding the soft features one by one to the hard feature subset and also adopted accuracy and AUC as the assessment metrics. The results of discrimination performance are shown in Table 6. After adding the soft features, the accuracy and AUC values of all four classifiers showed an upward trend. The increase in the AUC value demonstrates that soft features have a certain level of discriminating power for default or non-default loan classification.

6.4 Statistical significance of soft features

To further determine the effect of the extracted soft features on loan default classification, we constructed a regression model to verify their statistical significance using complete dataset without splitting (N = 39,538). The default status is the dependent variable, where 1 denotes default and 0 denotes non-default. The independent variables include both text feature variables and control variables. The text feature variables are the topics extracted from the loan description text, including Asset, Income, Work, Family, Agriculture, and Length. The control variables are the remaining features (i.e., 16 hard features) in the optimal subset obtained from Sect. 6.2. We remained the control variables unchanged and added a single soft

Table 7 Results of soft features in logistic regression analysis

	M1	M2	M3	M4	M5	M6	M7	M8
T1 Asset	-3.562* (-2.531)							-0.595 (-0.410)
T2 Income		0.892* (0.906)						3.739** (3.361)
T3 Work			8.935*** (7.710)					10.050*** (8.452)
T4 Family				16.820*** (15.554)				16.930*** (15.176)
T5 Business					-2.671*** (-2.886)			-2.343*** (-2.348)
T6 Agriculture						-6.394*** (-4.661)		-4.253*** (-3.024)
Length							-0.01*** (-8.74)	-0.01*** (-7.946)
Constant	-19.130	-19.310	-20.000	-20.750	-19.180	-19.070	-18.170	-20.530

The symbols *, **, and *** express significance at the 10, 5 and 1% level. Z-statistics are shown in parentheses

feature or all the soft features together, resulting in 8 logit models (i.e. M1–M8). Confirming the inexistence of endogeneity concerns, a binary logistic regression model is constructed as follows:

$$\begin{aligned} \text{Logit}(\text{Default}_1) = & \alpha_0 + \alpha_1 \text{Asset} + \alpha_2 \text{Income} + \alpha_3 \text{Work} + \alpha_4 \text{Family} \\ & + \alpha_5 \text{Business} + \alpha_6 \text{Agriculture} + \alpha_7 \text{Length} \\ & + \beta_i \text{ControlVariables} + \varepsilon \end{aligned}$$

The logistic regression results of soft features are shown in Table 7, reflecting the effects of the topic-features on P2P loan defaults. The Z-statistics, indicating the significance of coefficients, are shown in parentheses. The coefficients of all the soft features except T1-asset are significant at the 1% level ($p < 0.01$ and $|z| > 1.96$) in M8. And the T1-asset is significant at the 5% level in M1. Among all the soft features, the T1-asset, T5-business, T6-agriculture and text-length features are negatively correlated with loan default; the other features are positively correlated. This indicates that when more asset, business and agriculture description information is included in the loan description or when the description text is lengthy, the borrower is less likely to default. In contrast, the more extensive the borrower's description of family, work, and income-related issues in the loan text, the more likely it is that the borrower will default. The results also show that the soft features extracted from descriptive loan text have certain relationships with P2P loan default status and can serve as additional input data for default prediction.

7 Conclusions

Loan default has become an important factor restricting the development of P2P lending. Because of the unique characteristics of the customer group in P2P lending, the features that are effective in traditional lending, such as those used for bank loans, tend to be less useful for determining the borrowers most likely to default in P2P lending. To solve this problem, this paper presented a default prediction method that combines traditional features with soft features extracted from descriptive loan text to improve prediction accuracy with respect to loan default prediction. First, we employed the LDA model to extract the credit-related topics from the descriptive loan text as the soft features to construct prediction models. Second, to accommodate the differences in soft and hard information, we designed a two-stage feature selection method to generate effective feature subsets for modeling. Finally, we built loan default prediction models using four classification methods and conducted an empirical analysis with real data from a P2P lending-broker platform in China. The results demonstrate that soft information extracted from the descriptive loan text in the P2P platform can help identify borrowers who may default on loans. These features can be integrated into a default prediction model to improve discrimination performance and decrease the risk of defaults to benefit investors and P2P platforms. The soft features have a significant ability in discriminating loan defaults.

Considering the diversity of soft information in an online environment, this paper still has some deficiencies: we analyzed the topic features and statistical linguistic feature (i.e., text length) extracted from descriptive loan text for loan default prediction; however, a vast variety of unstructured data (e.g., descriptive text) may also contain other valuable information for default prediction, such as other linguistic features (e.g., Part-of-Speech features and sentiment features) and social relationship information. How to extract additional valuable

factors that influence loan defaults from the abundant unstructured data requires further research.

Acknowledgements The authors gratefully acknowledge the assistance provided by the constructive comments of the anonymous referees, which considerably improved the paper in terms of quality and clarity. This work was funded primarily by the National Natural Science Foundation of China (Grant Nos. 71571059, 71331002 and 71731005), and the Humanities and Social Sciences Fund Projects of the Ministry of Education (Grant Nos. 13YJA630037, 15YJA630010).

References

- Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting Finance & Management*, 18(2–3), 59–88.
- Angilella, S., & Mazzù, S. (2015). The financing of innovative SMEs: A multicriteria credit rating model. *European Journal of Operational Research*, 244(2), 540–554.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *JMLR.org*.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Cornée, S. (2017). The relevance of soft information for predicting small business credit default: Evidence from a social bank. *Journal of Small Business Management*. doi:10.1111/jsbm.12318.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3), 1447–1465.
- Dorffleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., Castro, I. D., et al. (2016). Description-text related soft information in peer-to-peer lending—Evidence from two leading european platforms. *Journal of Banking & Finance*, 64, 169–187.
- Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending. *Applied Economics*, 47(1), 54–70.
- Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, 210(2), 368–378.
- Friedman, N., Dan, G., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, 29(2–3), 131–163.
- Gao, Q., & Lin, M. (July 15, 2016). Economic value of texts: Evidence from online debt crowdfunding. Available at SSRN: doi:10.2139/ssrn.2446114.
- Guo, Y., Zhou, W., Luo, C., Liu, C., & Xiong, H. (2015). Instance-based credit risk assessment for investment decisions in p2p lending. *European Journal of Operational Research*, 249(2), 417–426.
- Hajek, P., & Michalak, K. (2013). Feature selection in corporate credit rating prediction. *Knowledge-Based Systems*, 51(1), 72–84.
- Harris, T. (2013). Quantitative credit risk assessment using support vector machines: Broad versus narrow default definitions. *Expert Systems with Applications*, 40(11), 4404–4413.
- Huang, C. L., Chen, M. C., & Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, 33(4), 847–856.
- Iyer, R., Khwaja, A. I., Luttmer, E. F., & Shue, K. (2015). Screening peers softly: Inferring the quality of small borrowers. *Management Science*, 62(6), 1554–1577.
- Hájek, P. (2011). Municipal credit rating modelling by neural networks. *Decision Support Systems*, 51(1), 108–118.
- Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, 40(13), 5125–5131.
- Kruppa, J., Ziegler, A., & König, I. R. (2012). Risk estimation and risk prediction using machine-learning methods. *Human Genetics*, 131(10), 1639–1654.
- Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine Learning*, 59(1–2), 161–205.
- Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136.
- Liberti, J. M., & Petersen, M. A. (2017). Information: Hard and Soft. Working Paper.
- Lin, M., Prabhala, N. R., & 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.
- Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*, 42(10), 4621–4631.

- Michels, J. (2012). Do unverifiable disclosures matter? Evidence from peer-to-peer lending. *The Accounting Review*, 87(4), 1385–1413.
- Paul, S. (2014). Creditworthiness of a borrower and the selection process in micro-finance: A case study from the urban slums of India. *Margin: The Journal of Applied Economic Research*, 8(1), 59–75.
- Pope, D. G., & Sydnor, J. R. (2011). What's in a picture? Evidence of discrimination from prosper.com. *Journal of Human Resources*, 46(1), 53–92.
- Puro, L., Teich, J. E., Wallenius, H., & Wallenius, J. (2010). Borrower decision aid for people-to-people lending. *Decision Support Systems*, 49(1), 52–60.
- Shao, H., Ju, X., Wu, C., Xu, J., & Liu, M. (2012). Research on commercial bank credit risk evaluation model based on the integration of the probability distribution theory and the bp neural network technology. *International Journal of Advancements in Computing Technology*, 4(22), 115–128.
- Thomas, L. C. (2010). Consumer finance: Challenges for operational research. *Journal of the Operational Research Society*, 61(1), 41–52.
- Wang, G., Ma, J., Huang, L., & Xu, K. (2012). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems*, 26, 61–68.
- Wang, S., Qi, Y., Fu, B., & Liu, H. (2016). Credit risk evaluation based on text analysis. *International Journal of Cognitive Informatics & Natural Intelligence*, 10(1), 1–11.
- Wei, X., & Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. In *International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 178–185). ACM.
- Yao, X., Crook, J., & Andreeva, G. (2015). Support vector regression for loss given default modelling. *European Journal of Operational Research*, 240(2), 528–538.