information

MDPI

Article

Credit Scoring Using Machine Learning by Combing Social Network Information: Evidence from Peer-to-Peer Lending

Beibei Niu, Jinzheng Ren * and Xiaotao Li

College of Economics and Management, China Agricultural University, Beijing 100083, China; oubeigo121@163.com (B.N.); lixiaotao01@126.com (X.L.)

* Correspondence: rjzheng@cau.edu.cn; Tel.: +86-10-6273-8506

Received: 25 November 2019; Accepted: 16 December 2019; Published: 17 December 2019

Abstract: Financial institutions use credit scoring to evaluate potential loan default risks. However, insufficient credit information limits the peer-to-peer (P2P) lending platform's capacity to build effective credit scoring. In recent years, many types of data are used for credit scoring to compensate for the lack of credit history data. Whether social network information can be used to strengthen financial institutions' predictive power has received much attention in the industry and academia. The aim of this study is to test the reliability of social network information in predicting loan default. We extract borrowers' social network information from mobile phones and then use logistic regression to test the relationship between social network information and loan default. Three machine learning algorithms—random forest, AdaBoost, and LightGBM—were constructed to demonstrate the predictive performance of social network information. The logistic regression results show that there is a statistically significant correlation between social network information and loan default. The machine learning algorithm results show that social network information can improve loan default prediction performance significantly. The experiment results suggest that social network information is valuable for credit scoring.

Keywords: credit scoring; peer-to-peer (P2P) lending; social network

1. Introduction

When a consumer attempts to obtain a credit card or auto loan, lenders usually use credit scores to determine whether to approve it. Credit scoring is a statistical analysis performed by financial institutions or credit bureaus to evaluate a borrower's creditworthiness and is based on credit history, demographic data, and credit behavior. In general, borrowers with high credit scores are more likely to obtain the loan and be charged a lower interest rate. Credit scoring is important to financial institutions not only because it can measure default risk but also because any small improvement would produce great profits [1,2]. However, many customers do not have a credit history and many developing countries have an imperfect credit reporting system [3]. Insufficient credit information limits financial institutions in building an effective credit scoring system to distinguish high-risk borrowers from the target population, especially in peer-to-peer (P2P) lending [4]. In addition, customers will be excluded from credit because they lack credit data. With the rapid development of P2P lending, payday loans, and online microlending markets in developing countries, financial institutions should seek more reliable methods to better evaluate a borrower's default risk.

One way to toughen predictive power is to search for an alternative source of credit information. In the past few years, a considerable amount of literature has studied variables related to personality and socioeconomic status that can be used to predict the loan default probability. Jiang et al. [5] use the

Latent Dirichlet Allocation topic model to subdivide the loan statement texts and quantify the text information. The empirical result shows that loan statement text can be used to improve classification accuracy. On some P2P lending platforms, loan applications that include a picture and loan period affect the default probability [6]. Meanwhile, as the role of social networks has received much attention, some researchers have explored the correlation between online social network information and loan default [7,8]. Extending this stream of research, our study is to test the reliability of social network information in predicting loan default.

Mobile phones have become an indispensable part of daily life in modern society. People rely heavily on mobile phones to store a large amount of personal information. Mobile phone data and data collected through mobile phones have been widely used in personal behavior research, poverty identification, and other studies [9,10]. Mobile phones also store personal social network information, so each person's social network information can be obtained from his/her mobile phone for personal credit risk assessment.

In this paper, we extract social network data from mobile phones to test whether social network information can be used for credit scoring. A total of 21,036 P2P loan samples were collected from mobile network operators and a Chinese P2P lending platform. Since almost every person in modern society owns a mobile phone, the social network information extracted from mobile phones ensures the method's versatility. The suggested method is thus expected to reduce the problem of insufficient credit records of borrowers. Additionally, it can enable more people to obtain financial services, thus reducing financial exclusion. At the same time, the results could be beneficial to P2P lending platforms' attempt to design more effective default prediction algorithms. Financial institutions can more accurately evaluate institutional risk. In this way, the P2P lending industry can develop well.

The remainder of the paper is organized as follows. Section 2 presents the literature review. Section 3 provides the hypothesis development. Section 4 is the empirical study. Section 5 discusses the results, offering limitations and future directions.

2. Literature Review

Credit scoring has played a key role in the significant growth of consumer finance over the past 60 years. Many statistical models are used to credit scoring, such as linear or logistic regression, linear discriminant analysis, probit analysis, and naïve Bayes [11,12]. However, these methods often perform poorly when dealing with nonlinear relationships. It is difficult for them to meet the given statistical assumptions in practical application. Therefore, many machine learning and artificial intelligence methods have been applied to credit scoring, and these algorithms have worked better than statistical analysis. These methods include support vector machine (SVM) [13], artificial neural networks (ANN) [14], and random forest [15]. In recent years, ensemble methods such as random forest, AdaBoost (Adaptive Boosting), and GBDT (gradient boosted decision trees) have received much attention in credit scoring [16]. Compared with different algorithms for credit scoring, ensemble methods are typically more advantageous [5], and these algorithms are now regarded as mainstream in credit scoring.

The financial institutions rely on three kinds of data to build credit scoring: demographic information, customer's transactional history data, and credit history data [17]. In recent years, the boom in P2P lending has provided a large number of research materials for credit scoring. With the dramatic changes in credit scoring data sources, a lot of research sought to test the new data sources. Empirical research on Lending Club shows debt information and FICO scores play an important role in loan default prediction [18]. Gao and Lin [19] demonstrated that borrowing reasons text features can impact loan default. For borrowers with strong readability and more objective description, the default probability is lower. Ma et al. [4] extracted information from meta-level phone usage data to build a default prediction method. The results show that phone usage data can improve credit scoring model accuracy.

The social media boom has generated a huge amount of social network data. The data gathered from social media may be considered an important source of information. Hill et al. [20] applied social network data to empirical research. They observed social network data from calling behavior to predict product/service adoption. In order to reduce information asymmetry, many P2P lending platforms encourage borrowers to build online groups. The empirical results of research by Everett [21] reveal that the closer the location of the borrowers in the group, the lower the default probability. In addition, if they have a certain connection in real life, joining the group will significantly reduce the borrower default probability. According to social media exposure by the borrower, borrowers with more friends are more likely to get loans, and the default probability is lower [22]. Based on the assumption that people are more likely to form social relationships with people like themselves, Wei et al. [8] proposed a social network theoretical framework. Their framework showed that social networks can improve the prediction accuracy of borrowers' defaulting. De Cnudde et al. [23] used online social network information extracted from Facebook accounts to build a credit scoring model. Their results show that social network information extracted from Facebook have predictive value. The developing country market has also explored the relationships between individuals' social networks and loan performance. For example, Li et al.'s [24] empirical results reveal that the higher the credit grade of friends, the less likely the borrower is to default. Zhang et al. [25] found that online social network data could predict loan default. Ge et al. [7] tested the predictive effect of self-disclosed online social network information. Guo et al. [26] extracted social network data from Weibo, a Chinese microblogging website. The empirical results show that credit scoring with online social network data outperforms traditional credit scoring methods.

In sum, there has been considerable interest in loan default prediction, and many algorithms and data have been tested. Ensemble algorithms have better performance. In order to improve the performance of the classification model, related research verified "hard information" such as income, age, FICO score, and debt information, and they have also focused on the "soft information" of borrowers. In this aspect, social network information has received much attention. Some scholars have explored the role of social network information, but most only explored the correlation between online social network information and loan default. Although De Cnudde et al. [23] extracted social network data from Facebook to increase the predictive ability, this method cannot be applied in many countries. For example, Facebook is blocked in some countries. In addition, many people are reluctant to provide online social network information or do not have social media accounts, and some people provide invalid data and their social network information cannot be obtained, which leads to the failure of the method and limits the improvement of its risk prediction ability.

Although some people do not have online social network data, nearly every person contacts others by mobile phone, which saves the borrower's social network information. This paper extracts three types of social network information—social network quality, social network stability, and social network exposure—from mobile phones to build credit scoring.

Theoretically, this research enriches the research on social networks and provides a theoretical basis for social network application in credit scoring. In addition, using social network data provided by mobile phones to predict the default probability of the borrower has a huge advantage in practical applications. First, this method improves financial institutions' risk control capabilities. There is a lack of usable credit history for a large number of consumers; therefore, the commonly used variables of loan repayment history and debt ratio in credit risk assessment cannot be used, which greatly reduces the risk control capability of the financial institution. As a tool used by everyone in modern society, social network data extracted from mobile phones ensures the validity and breadth of credit risk assessment. Second, this method has greatly expanded the scope of financial services, enabling more people to access financial services without restrictions.

3. Hypothesis Development

We define the social network data collected from the mobile phone as the following three variables: social network quality, social network stability, and social network exposure. Social network quality counts the number of default borrowers in a borrower's mobile address book. When the borrower applies for a loan, the platform will ask the borrower for permission to access the address book. After the platform reads the address book, the platform will obtain the borrower's real social network information. The platform compares social network information with the platform's database and can get the total number of borrowing defaults within the borrower's social network. The social network exposure information comes from the number of contacts that the borrower fills in on the app or website when applying for a loan, including home phone, work phone, and emergency contact phone number. The borrower must fill in his or her own phone number when registering on the platform. In addition, the borrower can also fill in his home phone number, work phone number, and emergency contact phone number, so that the platform receives the social network exposure information. The social network stability means the length of time the borrower's mobile phone number has been used. The telecom operator knows the customer's mobile phone number usage duration, and the platform can obtain the social network stability information of the borrower through cooperation with the telecom operator. Due to the large volume of social network information, we used MapReduce to carry out the extraction process.

In the process of forming a social network, individuals prefer to build social relationships with people like them [27]. This makes closely connected individuals in a social network not only highly correlated but also highly similar. This similarity is not only reflected in the fact that the group accept the same information, form similar beliefs, and have similar behaviors [28]. Wei et al.'s [8] theoretical framework shows that the credit scoring of individuals in the same social network is relatively close. Compared with the online social network data that the previous research focused on, social relationships in real life can reflect the characteristics of individuals more realistically and accurately [24]. Based on these works of literature, we propose the following hypothesis:

Hypothesis 1 (H1). *Social network quality contains predictability of loan default.*

In China, borrowers know that if they default, the telephone collection will have a great impact on their life. Unless the loan is paid off, the impact will continue. The consequences of defaulting will make the borrower want to pay the loan as on time as possible. If the borrower wants to change their phone number so that the collector cannot contact them, it will cost a lot. The longer the mobile phone number is used, the more people the borrower can contact and the closer relationship there will be in the network. Once he/she leaves the social network, it takes a lot of effort to rebuild. The replacement of the mobile number means that the reconstruction of the social network will cause great inconvenience to the borrower, which is also a shock to the borrower's default. If the borrower's mobile phone number is only used for a short period of time, the borrower's social conversion cost will be lower, and the default will have less impact on the borrower's life.

In addition, the length of time that the mobile phone number is used can reflect his/her economic situation, and this is positively related to the economic conditions of the individual [29]. Economic status is an important indicator of P2P loan default prediction [30]. Therefore, we propose the following hypothesis:

Hypothesis 2 (H2). *Social network stability contains predictability of loan default.*

When borrowers apply for loans, they can choose to disclose many kinds of contact information to the platform. In addition to their mobile phone number, they can also provide home phone number, work phone number, and emergency phone number. The more contact information the borrower provides, the more social information the platform can control. In the event of default, the platform

will have more channels to contact collection, and the information that the borrower cannot pay the loan on time will be communicated to his social network. Whether a person is honest and trustworthy is of great significance to the maintenance of his social relationship [31]. If the loan is defaulted on, the borrower's family and colleagues can perceive the borrower as a dishonest and untrustworthy person, which impacts his social relationships. Compared to the online social network, the borrower's relatives and friends can be directly contacted by telephone, and the social relationship is more effective and binding. Therefore, the borrowers think the number of social relationships they disclose can demonstrate the borrower's willingness to repay his/her debts.

In China's P2P lending market, due to the imperfect credit system, only a small number of people have credit information, and the borrowers lack objective credit data evaluation, which may lead to adverse selection problems in the market. Borrowers with better credit have more ways to raise loans, so they are not willing to over-disclose their social network information. A large number of P2P lending borrowers are high-risk groups identified by other channels and have difficulty obtaining loans, but they also have certain credit needs. Because of the lack of financing channels, borrowers with poor credit are more inclined to adopt a strategy of actively disclosing social network information, revealing more social network verifiable information to make themselves look more credible, and thus attract more lender investment and achieve the purpose of successful borrowing. Herzenstein et al.'s [32] empirical research found that customers with a higher risk of default will disclose more information. Therefore, we propose the following hypothesis:

Hypothesis 3 (H3). *Social network exposure contains predictability of loan default.*

4. Empirical Study

In this section, we show the variables and use a logistic regression model to identify the relationship between borrowers' social network information and loan default. Previous studies have shown that the ensemble methods have better performance for credit scoring [5]. We employed the random forest, AdaBoost, and LightGBM algorithms to build default prediction models. In order to test the discrimination performance of social network data, we applied these algorithms with social network information and without. Then, we showed the feature importance of three algorithms with social network variables.

The experiment was coded using Python 3 and stata 14. The empirical evaluation was carried out using one core of an Intel Core i3-4170 based Windows 10 operating system PC with 8GB RAM.

4.1. Data and Variables

We collected the dataset from a Chinese P2P lending platform. All data were encrypted to protect privacy. A total of 21,036 loan samples were collected, of which 9025 are default loans and 12,011 are good (not default) loans. Every loan contains 21 variables. These variables include the borrower's personal information variables, registration information variables, loan information variables, and social network information variables. All variables are listed in Table 1. The personal information is the borrower's demographic information and financial information, including the borrower's age (Age), gender (Gender), marital status (Marriage), number of children (Children), family member counts (Family), education status (Education), income level (Income), car ownership (Car), income category (Income_type), job title (Jobtitle), house ownership (House), and days of work (Work_days). Registration information contains minutes of registration (Registration_time) and minutes before the borrower changed the document with which he applied for the loan (Document_change). In addition, loan information including the loan amount (Amount), interest rate (Rate), when the borrower applies for a loan during the day (Time) and repayment period (Period). The social network variables are social stability (Social_stability), social exposure (Social_exposure), and social quality (Social_quality).

Table 1. Features used in the analysis.

Variables	Description
Age	21–69 years
Gender	{male (0.63), female (0.37)}
Marriage	{married (0.73), unmarried (0.15), divorced (0.07), widow (0.05)}
Children	{have (0.31), don't have (0.69)}
Family	{Number of family members the borrower has: 1 (0.22), 2 (0.50), 3 (0.18), 4 (0.09), 5 or above (0.01)}
Education	{Junior high school or below (0.76), senior high school (0.03), bachelor/junior college or above (0.21)}
Income	The annual income of the borrower
Car	{yes (0.33), no (0.67)}
Income_type	Income type, six types
House	{own (0.88), parent's house (0.06), rent (0.06),}
Work_days	0–16,061 days
Registration_time	Numbers of minutes before the application the borrower started registration in this platform
Document_change	Numbers of minutes before the application the borrower changed his/her document
Jobtitle	Job title, five types
Amount	Loan amount
Rate	Loan rate
Period	8–45 months
Time	When the borrower applies for a loan during the day
Social_stability	{1 year or below (0.32), 1–3 years (0.31), >3 years (0.37)}
Social_exposure	{1 (0.12), 2 (0.56), 3 (0.21), 4 (0.11)}
Social_quality	{0 (0.87), 1 (0.10), 2 or above (0.03)}

The exploratory analysis of the social network variables is shown in Table 2. The mean, the median, and the standard deviation of defaulting and non-defaulting borrowers are shown in Table 2. From the table, we can see that the average value of the social network quality variable of the defaulting borrower is 0.190, whereas for the non-defaulting borrowers it is 0.141. The results show that the defaulting borrower's social network quality is even worse. In their social network, they have more defaulting friends than non-defaulting borrowers. The defaulting borrower's social network is less stable than the non-defaulting borrowers. The social stable variable's mean value for a defaulting borrower is 824.67, which is less than the non-defaulting borrowers' mean value of 998.47. The defaulting borrower will show more contact information than non-defaulting borrowers. On average, defaulting borrowers expose 2.366 contact information, which is bigger than that of non-defaulting borrowers.

Table 2. Exploratory analysis of variables from the social network.

	Default (N = 9025)			Non-Default (N = 12011)			*p*-Value
	Mean	Median	Std.	Mean	Median	Std.	
Social_quality	0.190	0	0.511	0.141	0	0.440	0.000(<0.01)
Social_stability	824.67	610	770.81	998.47	809	843.52	0.000(<0.01)
Social_exposure	2.366	2	0.803	2.278	2	0.832	0.000(<0.01)

The *t*-test was used to test whether there is a statistically significant difference between the two groups. The last columns display the significant results of the *t*-test. As we can see, defaulting borrowers and non-defaulting borrowers' differences are significant in all the social network information variables, which indicates that social network information may differ between non-defaulting and defaulting borrowers. Based on the exploratory analysis, the social network information hypotheses can be partially supported. Social network information variables contain predictability of loan default.

4.2. Statistical Significance of Social Network Information

We used a logistic regression model to test the influence of the social network information on loan default probability. The default variable is the dependent variable. If the borrower defaulted the default variable is 1 and 0 if the borrower did not default. In accordance with P2P lending industrial inertia, we define a borrower default as when the borrower fails to pay for more than 30 days. The independent variables are all shown in Table 1, which contains social network variables and control variables. Control variables are borrowers' demographic information, registration information,

and loan information. At the same time, in order to test the impact of social network information on the default of the borrower, two models were constructed. Model 1 only contains control variables. Model 2 not only has control variables but also has social network information:

Model 1:

$$\text{Logistic}\ (\text{default}_1) = \alpha_i \text{controlvariables}_i + \varepsilon_i \quad (1)$$

Model 2:

$$\text{Logistic}\ (\text{default}_1) = \alpha_i \text{controlvariables}_i + \beta_i \text{socialnetworkvariables}_i + \varepsilon_i \quad (2)$$

The term α is the coefficient of control variables, β is the coefficient of social network variables, and ε is disturbance.

4.3. Default Prediction Models

Credit scoring is regarded as a classification problem [30]. After the borrower obtains a loan from the P2P lending platform, two final statuses will arise: default or not. Defaulting borrowers are classified as bad and assigned 1, while non-defaulting ones are classified as good and assigned 0. Credit scoring predicts the borrower's final repayment status through the relevant characteristics of the borrower. For a borrower set $D = \{(\boldsymbol{x_1},y_1),(\boldsymbol{x_2},y_2), \ldots (\boldsymbol{x_n},y_n)\}$, $\boldsymbol{x}_i$ is the relevant characteristics of the borrower, and y_i is the final state variable. Ensemble algorithms perform better in credit scoring compared to other algorithms [5]. They are prevalent in credit scoring. Random forest, AdaBoost and GBDT are the most popular ensemble methods. In order to verify the effect of social network information in loan prediction, we selected random forest, AdaBoost, and LightGBM models to examine the changes in predictive ability before and after fusing social network information to credit scoring. The LightGBM algorithm is a framework proposed to implement the GBDT algorithm. In order to show the importance of social network variables in the classification model, we show the feature importance results of random forest, AdaBoost, and LightGBM.

4.3.1. Random Forest

Random forest is an ensemble learning algorithm based on the decision tree. It can be used for both regression and classification tasks and is easy to implement. Random forest uses bootstrap to obtain samples from the original data. Every tree gives a classification, and the forest chooses the classification with the most votes in all the trees. The parameter m is the number of decision trees and determines the degree of randomness. In the random forest, the borrower is assumed to have d attributes. In general, we set $m = \log_2 d$. The CART classification tree uses the Gini index to create split points. If we have n classes, p_i is the probability that an object belongs to class i. The Gini index would be:

$$\text{Gini}(D) = 1 - \sum_{i=1}^{n} p_i^2 \quad (3)$$

The Gini index is used to compute the impurity of data sets. The smaller the Gini index, the higher the purity of the data sets. The Gini index of classification problem:

$$\text{Gini}_{\text{split}}(D) = \frac{|D_1|}{|D|}\text{Gini}(D_1) + \frac{|D_2|}{|D|}\text{Gini}(D_2). \quad (4)$$

4.3.2. AdaBoost

AdaBoost is a powerful classification algorithm that has high accuracy and low generalization error. We can use CART, C4.5, and SVM for AdaBoost as weak learner. AdaBoost aims to convert a set of weak classifiers into a strong one. AdaBoost reduces the error rate of classification by multi-step iteration. After each iteration, the weak classifier weight with small classification error rate is reduced, and the weak classifier weight with a large classification error rate is increased until the predetermined

error rate or iteration round is reached. We choose the decision tree as weak learners. The number of weak learners $h_t(x)$ is T. The weight of every learner is

$$\theta_t = \frac{1}{2}\ln\frac{1-\epsilon_t}{\epsilon_t}, \tag{5}$$

where ϵ_t represents the classification error rate in every weak learner. The weight for each data point is updated as

$$\omega_{t+1}(x_i, y_i) = \frac{\omega_t(x_i, y_i)e^{-\theta_t y_i f_t(x_i)}}{Z_t}, \tag{6}$$

where ω_t is the weight of the data point, and Z_t is a normalization factor that ensures the sum of instance weights is equal to 1.

The weak learner is ensembled to a strong classifier. The linear combination of the basic classifiers is defined as

$$H(x) = \sum_{t=1}^{T} \theta_t h_t(x). \tag{7}$$

The final prediction sums up the weighted prediction of each classifier:

$$G(x) = \text{sign}(\sum_{t=1}^{T} \alpha_t h_t(x)). \tag{8}$$

4.3.3. LightGBM

The LightGBM algorithm is a framework proposed by Microsoft in 2017 to implement the GBDT algorithm. It can be used for classification, regression, and many other machine learning tasks. LightGBM is one of the most popular methods for data scientists and the machine learning online community Kaggle. Compared with the xgboost algorithm, which has shined within data science competition in previous years, LightGBM guarantees accuracy while having faster speed and less memory consumption and supports distributed computing, which can process massive data quickly.

The GBDT algorithm procedure is performed as Algorithm 1. $f(x)$ is the decision tree, $\left\{R_J\right\}_1^J$ are the parameters of the decision tree, $\left\{b_j\right\}_1^J$ is the decision tree's output function value, J is the number of leaf nodes, and F is a collection of decision trees. D is training data, L is target function, and K is the number of iterations.

Algorithm 1. Gradient boosting decision tree algorithm procedure.

Input: N samples: $\{x_i, y_i\}$, K, L, ...
Initialize f_0
for $k = 1$ to K:
$\widetilde{y_i} = -\frac{\partial L(y_i, F_{k-1}(x_i))}{\partial F_{k-1}}$, i = 1, 2, ... , N
$\left\{R_j, b_j\right\}_1^{J^*} = \text{argmin}_{\{R_j, b_j\}_1^J} \sum_{i=1}^{N}\left[\widetilde{y_i} - f_k\left(x_i; \left\{R_j, b_j\right\}_1^J\right)\right]^2$
$\rho^* = \text{argmin}_\rho \sum_{i=1}^{N} L(y_i, F_{k-1}(x_i) + \rho f_k(x_i)) + \Omega(f_k)$
$f_k = \rho^* f_k$, $F_k = F_{k-1} + f_k$
Output: F_k

GBDT continuously fits the residuals and adds them to F. In this process, the residuals become smaller and smaller. GBDT algorithms use a gradient descent method to optimize the loss function min$f(x)$ and min$L(F)$.

Different from many other boosting methods that use pre-sort-based algorithms, LightGBM uses histogram-based algorithms, which bucket continuous variables into discrete bins. This speeds up training and reduces memory usage.

GBDT algorithms are based on decision tree algorithms. Most of the GBDT algorithms split the tree by depth or level rather than by leaf. This can control the complexity of the model and reduce overfitting. However, the information gain of some leaf nodes is lower, and the computation is increased. LightGBM splits the tree with leaf-wise algorithms. The leaf-wise algorithm can reduce loss more than other algorithms. Level-wise splits the leaf nodes with the largest information gain. Level-wise can reduce the error and obtain higher precision than leaf-wise.

4.3.4. Parameter Selection

Hyperparameters are parameters that cannot directly be estimated from training data. Hyperparameters are important because they can control the performance of algorithms. In order to have a good performance for every model, we used the gird search to find the optimal hyperparameters. We use the area under the ROC curve (AUC) to measure how effective a model is. In order to prevent overfitting and identify the validity of the three models, these methods are optimized using a 5-fold cross-validated grid search over a hyperparameter selection. The data set is randomly partitioned into five roughly equal subsets. The algorithm is training on four folds, and the other fold is the test set.

The n_estimator is the number of trees in the random forest algorithms. A parameter grid on [10, 150] is applied to select the parameters. The best n_estimator is 100. A parameter grid on [10, 1000] $\times$ [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1] is applied to select the best parameters of forest size and learning rate in AdaBoost algorithms. The result shows that when forest size is 400, and learning_rate is 0.2, and the performance reaches its best. In LightGBM algorithms, max_depth is the maximum depth of tree, min_data_in_leaf is the minimum number of the records a leaf may have, feature_fraction determines how many parameters LightGBM will select in each iteration for building trees, and bagging_fraction specifies the fraction of data to be used for each iteration. We get the optimal parameter from the grid search: max_depth = 6, min_data_in_leaf = 35, feature_fraction = 0.8, bagging_fraction = 0.5.

4.4. Results

The logistic regression results are shown in Table 3. Model 1 only contains control variables; Model 2 not only has control variables but also has social network information. As can be seen, the variables that are statistically significantly correlated in Model 1 are also significant in Model 2, but some variables differ significantly in their levels.

The coefficients of all the social network information variables are statistically significant at the $p < 0.01$ level. Among the social network information features, social network quality and social network exposure are positively correlated with loan default; the social stable feature is negatively correlated. This indicates that when the borrower has more loan defaults in the social network or exposes more contact information, the more likely the borrower is to default. In contrast, the longer the mobile phone number has been used, the less likely the borrower is to default.

The results show that the social network information variables extracted from the mobile phone have certain relationships with P2P lending loan default. The social network variables can be used to estimate the loan's probability of default.

Table 3. Results of the logistic regression model.

Variable	Model 1		Model 2	
	Coefficient	**Std.**	**Coefficient**	**Std.**
Age	−0.2550 ***	(0.020)	−0.2212 ***	(0.021)
Gender	−0.3797 ***	(0.033)	−0.3802 ***	(0.033)
Marriage	-	-	-	-
Children	0.0180	(0.040)	0.0139	(0.040)
Family	−0.0205	(0.046)	−0.0125	(0.049)
Education	−0.2233 ***	(0.016)	−0.2193 ***	(0.016)
Income	−0.0116	(0.018)	0.0008	(0.017)
Car	−0.1296 ***	(0.016)	−0.1238 ***	(0.016)
Income_type	0.0123	(0.020)	−0.0191	(0.021)
House	−0.0328 **	(0.015)	−0.0367 **	(0.015)
Work_days	−0.1448 ***	(0.017)	−0.1470 ***	(0.017)
Registration_time	−0.1054 ***	(0.016)	−0.1007 ***	(0.016)
Document_change	−0.1466 ***	(0.015)	−0.1369 ***	(0.015)
Jobtitle	−0.0476 ***	(0.017)	−0.0473 ***	(0.017)
Amount	0.0037	(0.023)	0.0165	(0.023)
Rate	0.1986 ***	(0.015)	0.2114 ***	(0.015)
Period	0.0559 ***	(0.021)	−0.0631 ***	(0.021)
Time	−0.0560 ***	(0.015)	−0.0604 ***	(0.015)
Social_stability	-	-	−0.1287 ***	(0.015)
Social_exposure	-	-	0.1162 ***	(0.017)
Social_quality	-	-	0.1020 ***	(0.014)

Notes: ***$p < 0.01$; **$p < 0.05$. Since the dummy variables are large and are not statistically significant, the coefficient results are not listed.

4.4.1. Default Prediction Models and Discussion

In order to verify the improvement of social network information on credit risk assessment, we chose three classification models: random forest, AdaBoost, and LightGBM to see the change of prediction ability of each model before and after adding social network information.

We conducted 50 experiments and every experiment adopted 5-fold cross-validation to ensure the validity of the results. We used AUC, F1 score and the prediction accuracy to evaluate the prediction results. Table 4 summarizes the mean of accuracy, F1 score and AUC results on 50 repeated 5-fold cross-validation experiments for the three models. Every algorithm has two results: contain and not contain social network information.

Table 4. Discrimination performance of three credit scoring models.

	Random Forest			AdaBoost			LightGBM		
	Accuracy	**AUC**	**F1**	**Accuracy**	**AUC**	**F1**	**Accuracy**	**AUC**	**F1**
Not contain	63.57%	0.674	0.635	63.91%	0.681	0.642	65.50%	0.692	0.649
Contain	63.92%	0.689	0.644	64.40%	0.697	0.651	66.22%	0.711	0.659

From Table 4, we can see that the LightGBM algorithm has the best classification effect, regardless of the accuracy of the model or the F1 score and AUC. In the absence of social network information, the AUC value of the original model is 0.692, and the prediction accuracy of the model is 65.50%. After combing social network information, its AUC increased to 0.711, and the prediction accuracy also increased to 66.22%. The F1 score of the benchmark model is 0.649, and the F1 score of the model with social network information is higher at 0.659. The rest of the models are inferior in accuracy, F1 score, and AUC compared to LightGBM, but the final results show that after combing the social network variables, both the prediction accuracy, F1 score, and AUC have a certain degree of improvement. After combing social network information, the AUC of the three algorithms increased by 0.015, 0.016 and 0.019. In the credit scoring industry, it is a significant upgrade [33]. In addition, hyperparameters are searched in the dataset that are not combined with social network information. If we do the grid search on the dataset with social network information, the model's predictive power will be higher.

In order to test whether the AUC of the model with social network variables is not only larger than that of the model not containing social network information but also statistically away from it, we used the nonparametric approach developed by DeLong et al. [34] and implemented the stata routine "roccomp". Table 5 shows the results. The statistical test of random forest and AdaBoost are statistically significant at the $p < 0.05$ level, LightGBM is statistically significant at the $p < 0.01$ level. The results proved that the AUC value of the model with social network information is not only larger than when not combined but also statistically away from it. Regardless of which algorithm, all the AUC values with social network information are better than the results without social network information.

Table 5. The nonparametric test results.

	Do Not Contain	Contain	*p*-Value
Random Forest	0.674 (0.001099)	0.689 (0.001287)	0.018
AdaBoost	0.681 (0.001141)	0.697 (0.001327)	0.014
LightGBM	0.692 (0.001328)	0.711 (0.001535)	0.009

Through the comparison of the model AUC, F1 score and accuracy, we find that combing social network information can significantly improve the accuracy of the prediction and has a good application prospect. Further tests carried out with the nonparametric approach [34] corroborated our initial findings.

4.4.2. Feature Importance

There are 21 features in the classification model. Figures 1–3 show the ordering of the 14 most important variables of the random forest, AdaBoost, and LightGBM. These variables are plotted in decreasing order of the importance value.

As can be seen, although the importance order of the three models is not the same, the period variable is ranked in the top one in all three models. The social network stability ranks 5th, 9th, and 4th, respectively, in the three models; the social network quality variables rank 14th, 13th, and 13th, respectively, in the three models. The social network exposure variable ranked 13th, 12th, and 14th, respectively, in the three models. The results show that social network information plays an important role in the prediction of loan default.

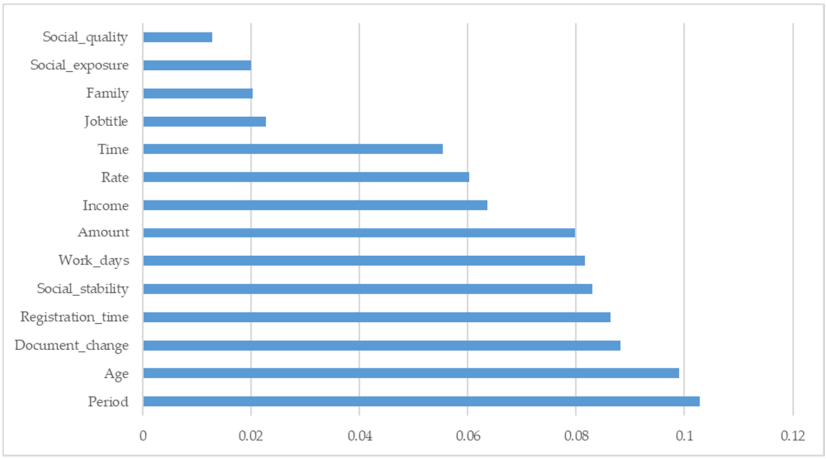

Figure 1. Variable importance of the random forest model.

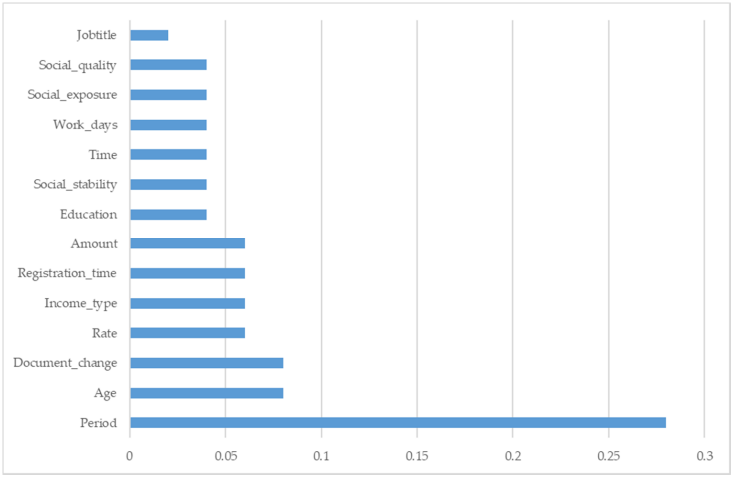

Figure 2. Variable importance of the AdaBoost model.

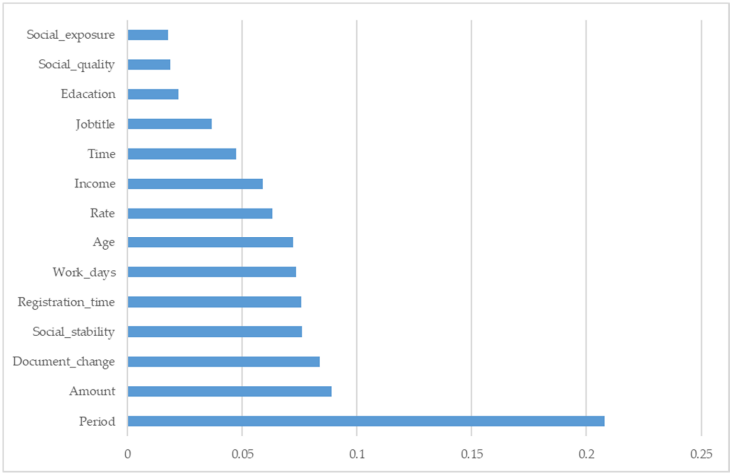

Figure 3. Variable importance of the LightGBM model.

The above results show that social network information can be used in credit scoring, and the predictive ability of the model has been improved. The social network information hypotheses can be supported: the three social network information variables contain predictability of loan default.

Many research and financial projects use different kinds of information and machine learning algorithms to predict loan default [3]. Our research results show the reliability of social network information in predicting credit default risk. This study demonstrates the value of social network information and proves the connection between raw social network data and credit scoring.

5. Conclusions

Loan default prediction is important to financial institutions, which can use credit scoring to distinguish customer default or not. In developing countries, many P2P lending borrowers lack credit history data, making it difficult to predict whether the borrower will default or not. Financial institutions need a more robust method for credit scoring. Modern society is highly interconnected, and social network data can show the credit status of borrowers. Mobile phones are accessible to almost everyone and hold a lot of social network information. This paper extracted social network data from mobile phones to predict borrowers' behavior. The empirical results show the social network variables extracted from mobile phones could be used to improve loan prediction accuracy. The research enhances our understanding of the social network. The platform can also obtain a better judgment of the credit risk of the borrower, improve the industry's ability to resist risks, and facilitate the smooth development of the P2P lending industry.

However, some limitations are worth noting. We did not collect other social network data, such as the frequency of calls, whether they are incoming or outgoing and the strength of social network ties. This limits the research regarding the role of social networks in personal credit risk assessment. In the future, other offline social network data may be collected and used for credit scoring.

Author Contributions: Data curation, B.N.; funding acquisition, J.R.; methodology, B.N.; writing—original draft, B.N.; writing—review and editing, J.R. and X.L.

Funding: This work is supported by the Young Scientists Fund of the National Natural Science Foundation of China (No. 71603259) and the Beijing Social Science Found (No. 19GLA002).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Einav, L.; Jenkins, M.; Levin, J. The impact of credit scoring on consumer lending. *Rand J. Econ.* **2013**, *44*, 249–274. [CrossRef]
2. Blöchlinger, A.; Leippold, M. Economic benefit of powerful credit scoring. *J. Bank Financ.* **2006**, *30*, 851–873. [CrossRef]
3. Aitken, R. 'All data is credit data': Constituting the unbanked. *Compet. Chang.* **2017**, *21*, 274–300. [CrossRef]
4. Ma, L.; Zhao, X.; Zhou, Z.; Liu, Y. A new aspect on P2P online lending default prediction using meta-level phone usage data in China. *Decis. Support. Syst.* **2018**, *111*, 60–71. [CrossRef]
5. Jiang, C.; Wang, Z.; Wang, R.; Ding, Y. Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Ann. Oper. Res.* **2018**, *266*, 511–529. [CrossRef]
6. Dorfleitner, G.; Priberny, C.; Schuster, S.; Stoiber, J.; Weber, M.; de Castro, I.; Kammler, J. Description-text related soft information in peer-to-peer lending—Evidence from two leading European platforms. *J. Bank. Financ.* **2016**, *64*, 169–187. [CrossRef]
7. Ge, R.; Feng, J.; Gu, B.; Zhang, P. Predicting and Deterring Default with Social Media Information in Peer-to-Peer Lending. *J. Manag. Inform. Syst.* **2017**, *34*, 401–424. [CrossRef]
8. Wei, Y.; Yildirim, P.; Van den Bulte, C.; Dellarocas, C. Credit scoring with social network data. *Market. Sci.* **2015**, *35*, 234–258. [CrossRef]
9. Harari, G.M.; Müller, S.R.; Aung, M.S.; Rentfrow, P.J. Smartphone sensing methods for studying behavior in everyday life. *Curr. Opin. Behav. Sci.* **2017**, *18*, 83–90. [CrossRef]
10. Blumenstock, J.; Cadamuro, G.; On, R. Predicting poverty and wealth from mobile phone metadata. *Science* **2015**, *350*, 1073–1076. [CrossRef]
11. Hand, D.J.; Henley, W.E. Statistical classification methods in consumer credit scoring: a review. *J. R. Stat. Soc. A. Stat.* **1997**, *160*, 523–541. [CrossRef]
12. Baesens, B.; van Gestel, T.; Viaene, S.; Stepanova, M.; Suykens, J.; Vanthienen, J. Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring. *J. Oper. Res. Soc.* **2003**, *54*, 627–635. [CrossRef]
13. Harris, T. Credit scoring using the clustered support vector machine. *Expert. Syst. Appl.* **2015**, *42*, 741–750. [CrossRef]
14. Zhao, Z.; Xu, S.; Kang, B.H.; Kabir, M.M.J.; Liu, Y.; Wasinger, R. Investigation and improvement of multi-layer perceptron neural networks for credit scoring. *Expert. Syst. Appl.* **2015**, *42*, 3508–3516. [CrossRef]
15. Malekipirbazari, M.; Aksakalli, V. Risk assessment in social lending via random forests. *Expert. Syst. Appl.* **2015**, *42*, 4621–4631. [CrossRef]
16. Ala'Raj, M.; Abbod, M.F. A new hybrid ensemble credit scoring model based on classifiers consensus system approach. *Expert. Syst. Appl.* **2016**, *64*, 36–55. [CrossRef]
17. Thomas, L.C. Consumer finance: challenges for operational research. *J. Oper. Res. Soc.* **2010**, *61*, 41–52. [CrossRef]
18. Emekter, R.; Tu, Y.; Jirasakuldech, B.; Lu, M. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Appl. Econ.* **2015**, *47*, 54–70. [CrossRef]
19. Gao, Q.; Lin, M. Words Matter: The Role of Texts in Online Credit Markets. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2446114 (accessed on 27 September 2019).

20. Hill, S.; Provost, F.; Volinsky, C. Network-Based Marketing: Identifying Likely Adopters via Consumer Networks. *Stat. Sci.* **2006**, *21*, 256–276. [CrossRef]
21. Everett, C.R. Group Membership, Relationship Banking and Loan Default Risk: The Case of Online Social Lending. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1114428 (accessed on 27 September 2019).
22. Lin, M.; Prabhala, N.R.; Viswanathan, S. Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Manag. Sci.* **2013**, *59*, 17–35. [CrossRef]
23. De Cnudde, S.; Moeyersoms, J.; Stankova, M.; Tobback, E.; Javaly, V.; Martens, D. What does your Facebook profile reveal about your creditworthiness? Using alternative data for microfinance. *J. Oper. Res. Soc.* **2019**, *70*, 353–363. [CrossRef]
24. Li, S.M.; Lin, Z.X.; Qiu, J.X.; Safi, R.; Xiao, Z.Y. How friendship networks work in online P2P lending markets. *Nankai Bus. Rev. Int.* **2015**, *6*, 42–67. [CrossRef]
25. Zhang, Y.; Jia, H.; Diao, Y.; Hai, M.; Li, H. Research on Credit Scoring by Fusing Social Media Information in Online Peer-to-Peer Lending. *Procedia Comput. Sci.* **2016**, *91*, 168–174. [CrossRef]
26. Guo, G.; Zhu, F.; Chen, E.; Liu, Q.; Wu, L.; Guan, C. From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring. *ACM T. Web.* **2016**, *10*, 1–38. [CrossRef]
27. Zeng, Z.; Xie, Y. A preference-opportunity-choice framework with applications to intergroup friendship. *Am. J. Social.* **2008**, *114*, 615–648. [CrossRef] [PubMed]
28. McPherson, M.; Smith-Lovin, L.; Cook, J.M. Birds of a Feather: Homophily in Social Networks. *Annu. Rev. Social.* **2001**, *27*, 415–444. [CrossRef]
29. Pokhriyal, N.; Jacques, D.C. Combining disparate data sources for improved poverty prediction and mapping. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E9783–E9792. [CrossRef] [PubMed]
30. Lessmann, S.; Baesens, B.; Seow, H.; Thomas, L.C. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *Eur. J. Oper. Res.* **2015**, *247*, 124–136. [CrossRef]
31. Baumeister, R.F.; Leary, M.R. The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychol. Bull.* **1995**, *117*, 497–529. [CrossRef]
32. Herzenstein, M.; Sonenshein, S.; Dholakia, U.M. Tell Me a Good Story and I May Lend You Money: The Role of Narratives in Peer-to-Peer Lending Decisions. *J. Market. Res.* **2011**, *48*, S138–S149. [CrossRef]
33. Iyer, R.; Khwaja, A.I.; Luttmer, E.F.P.; Shue, K. Screening Peers Softly: Inferring the Quality of Small Borrowers. *Manag. Sci.* **2015**, *62*, 1554–1577. [CrossRef]
34. Delong, E.R.; Delong, D.M.; Clarke-Pearson, D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **1988**, *44*, 837–845. [CrossRef] [PubMed]