RESEARCH ARTICLE

Evaluating borrowers' default risk with a spatial probit model reflecting the distance in their relational network

Jong Wook Lee, So Young Sohn *

Department of Information and Industrial Engineering, Yonsei University, Seoul, Republic of Korea

* sohns@yonsei.ac.kr

Abstract

Potential relationship among loan applicants can provide valuable information for evaluating default risk. However, most of the existing credit scoring models either ignore this relationship or consider a simple connection information. This study assesses the applicants' relation in terms of their distance estimated based on their characteristics. This information is then utilized in a proposed spatial probit model to reflect the different degree of borrowers' relation on the default prediction of loan applicant. We apply this method to peer-to-peer Lending Club Loan data. Empirical results show that the consideration of information on the spatial autocorrelation among loan applicants can provide high predictive power for defaults.

OPEN ACCESS

Citation: Lee JW, Sohn SY (2021) Evaluating borrowers' default risk with a spatial probit model reflecting the distance in their relational network. PLoS ONE 16(12): e0261737. https://doi.org/10.1371/journal.pone.0261737

Editor: Elisa Ughetto, Politecnico di Torino, ITALY

Received: September 11, 2021

Accepted: December 8, 2021

Published: December 31, 2021

Data Availability Statement: All relevant data are within the paper and its Supporting information files.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2020R1A2C2005026). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors declare no competing interests.

1. Introduction

Credit risk management is very important for service firms in the lending business. To predict the probability of default of loan applicant that is essential for credit risk management, machine learning models use two types of borrower information: standard "hard" information and nonstandard "soft" information [1]. The former directly reflects the loan applicants' financial status or creditworthiness, while the latter includes those that do not have a direct relationship to the credit applicant's financial status or creditworthiness such as age or residence. Existing studies have shown that not only hard information but also soft information, which is less relevant to their financial condition, is helpful in predicting default risk [1–5]. While both hard and soft information has been used in most credit scoring models, what is missing is the potential relation among loan applicants. Relationship among loan applicants that are at high risk of default can also provide valuable information for evaluating default risk [6–8].

In this study, we use a borrower relationship network based on the borrowers' information provided for loan applications. This network is utilized as a spatial weight matrix for a spatial probit model that reflects different degrees of borrowers' relation for the prediction of a loan default. Our proposed approach is applied to peer-to-peer (P2P) lending.

Online P2P lending allows individuals to lend money to other individuals through online platforms without the intervention of a financial institution. These online P2P lending platforms are gaining popularity due to their low operating costs compared with traditional lending programs [9]. However, online P2P lending faces a significant problem, such as information asymmetry between borrowers and lenders, that is, the reliability of a borrower's credit is unknown to the lender [10]. Therefore, the use of relationship information among borrowers beyond those provided on the P2P platform is necessary. As it is difficult to discover realistic relationship information between borrowers in a P2P landing platform, this study defines the data-driven latent relationships between borrowers in terms of the similarity of their hard and soft information. We expect that the data-driven latent relationships information between borrowers can improve default risk prediction.

This paper is organized as follows. Section 2 reviews prior studies on default prediction in online P2P lending. Section 3 explains the methodologies employed, and Section 4 explores the Lending Club Loan (LCL) dataset used for this study. Finally, Section 5 presents the results, and Section 6 discusses the results, limitations, and suggestions for improvement.

2. Literature review

Models for default risk prediction in P2P lending services are divided into three categories: the probability of default (PD), exposure at default (EAD), and loss given default (LGD). Among them, PD models have been explored steadily [11]. The PD model predicts borrower's default using classification models based on the statistical or machine learning approaches. Statistical methods have the advantage of being able to quantitatively show the effect of each factor on the borrowers' default [12]. Emekter et al. [13] used a logistic regression model to predict the default probability of borrowers and found that Fair, Isaac and Company scores are a very important factor. However, statistical methods have the disadvantage of requiring strong assumptions in the observed data [14]. Meanwhile, machine learning methods have strong default prediction performance without requiring any statistical assumptions. These models include neural network [15], support vector machine [16, 17], and random forest [18]. However, these models have a fatal drawback, that is, individual factors do not directly show the effect on borrowers' default.

It is also important to choose the optimal features used to predict default risk. Generally, hard information can reflect borrowers' repayment ability [19], while soft information can reflect borrowers' repayment willingness [20]. Hard information plays an important role in explaining default risk because it directly represents the borrowers' financial status. However, online P2P lending platforms have difficulty collecting sufficient hard information. To overcome these limitations, the importance of soft information that is not related to the borrowers' financial status is increasingly emphasized. Lin et al. [21] discovered that information on gender, age, educational level, and marital status play a significant role in predicting default. Recently, unstructured data, such as text and image information, as well as structured data, have been used as soft information. Dorfleitner et al. [22] used textual soft information containing a description of the loan purpose such as text length, spelling errors, and the presence of positive emotion-evoking keywords. Jiang et al. [23] used a topic model to extract representative features from descriptive text concerning loans.

However, few studies have used information on the relationship among individual borrowers in online P2P lending services. Calabrese et al. [24] defined bank networks by estimating interbank relationships as aggregate claims to predict bank contagion. Agosto et al. [6] defined business networks by estimating inter-company relationships as aggregate trade volumes to predict business default from P2P platforms that specialize in business lending. Unlike for banks and companies, obtaining quantitative indicators of relationships among individuals is difficult. In this study, we propose a network definition among individual borrowers and use this relationship information as independent information.

3. Methodology

3.1 Spatial probit model

Generally, the latent response model is the method used to fit the binary response variable Y as a regression model [25]. The model used in this study is a spatial probit model, which has a spatial autoregressive structure and can be used with a binary response variable. Taking the latent underlying quantity as being represented by a continuous variable $\mathbf{Y}_{\mathbf{i}}^{*}$, we consider the observation mechanism as

$$\mathbf{Y}_{\mathbf{i}} = \begin{cases} \mathbf{1}, & \mathbf{Y}_{\mathbf{i}}^{*} > \mathbf{0} \\ \mathbf{0}, & \mathbf{otherwise} \end{cases} \tag{1}$$

with i = 1, 2, · · ·, n where n is the number of observations. We implement the spatial structure with an autoregressive model specification, such that

$$\mathbf{Y}^{*} = \boldsymbol{\rho}\mathbf{W}\mathbf{Y}^{*} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\mathbf{Y}^{*}$ is a continuous latent vector; $\mathbf{X}$ represents an n × k matrix of explanatory variables with related coefficient vector $\boldsymbol{\beta}$; $\mathbf{W}$ is a spatial lag weights matrix with $\boldsymbol{\rho}$ as the associated coefficient; and $\boldsymbol{\varepsilon}$ is the error term.

This spatial probit model implies heteroskedastic errors $\mathbf{e}$ as follows:

$$\mathbf{Y}^{*} = (\mathbf{I} - \boldsymbol{\rho}\mathbf{W})^{-1}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = (\mathbf{I} - \boldsymbol{\rho}\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{3}$$

where $\mathbf{e} = (\mathbf{I} - \boldsymbol{\rho}\mathbf{W})^{-1}\boldsymbol{\varepsilon}$ with variation: $\mathbf{var}(\mathbf{e}) = \boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{2}[(\mathbf{I} - \boldsymbol{\rho}\mathbf{W})'(\mathbf{I} - \boldsymbol{\rho}\mathbf{W})]^{-1}$.

Calabrese and Elkink [26] reviewed various methods for estimating parameters $\boldsymbol{\rho}$ and $\boldsymbol{\beta}$ in Eq (3). Among them we performed parameter estimation using the generalized method of moments (GMM) proposed by Pinkse and Slade [27], which derive the GMM equations from the likelihood function. This method is extended by Klier and McMillen [28] to the logit model. It is more robust than the maximum likelihood estimation because it does not depend on the assumption that the error term follows a normal distribution [27].

A GMM estimator is defined as follows:

$$\hat{\boldsymbol{\theta}} = \mathbf{arg} \ \min_{\boldsymbol{\theta}} \ \mathbf{u}'\mathbf{Z}\mathbf{M}\mathbf{Z}'\mathbf{u} \tag{4}$$

where $\boldsymbol{\theta} = [\boldsymbol{\rho}, \boldsymbol{\beta}]$, $\mathbf{u_i} = \mathbf{y_i} - \mathbf{p_i}$, $\mathbf{p}_{\mathbf{i}} = \mathbf{Pr}[\mathbf{y}_{\mathbf{i}} = \mathbf{1}] = \frac{\mathbf{exp}((\mathbf{I} - \hat{\boldsymbol{\rho}}\mathbf{W})^{-1}\mathbf{X}^{*}\hat{\boldsymbol{\beta}})}{\mathbf{1} + \mathbf{exp}((\mathbf{I} - \hat{\boldsymbol{\rho}}\mathbf{W})^{-1}\mathbf{X}^{*}\hat{\boldsymbol{\beta}})}$, $\mathbf{X}_{\mathbf{i}}^{*} = \frac{\mathbf{X_i}}{\boldsymbol{\sigma}_{\mathbf{i}}}$; $\boldsymbol{\sigma}_{\mathbf{i}}$ is a diagonal element of covariance matrix $[(\mathbf{I} - \boldsymbol{\rho}\mathbf{W})'(\mathbf{I} - \boldsymbol{\rho}\mathbf{W})]^{-1}$; $\mathbf{Z}$ is a matrix of instruments; and $\mathbf{M}$ is a positive definite matrix that is generally initialized to an identity matrix. We define the instrument matrix $\mathbf{Z} = \{\mathbf{X}, \mathbf{WX}, \mathbf{W}^{2}\mathbf{X}, \mathbf{W}^{3}\mathbf{X}\}$, as proposed by Kelijian and Prucha [29].

To estimate the parameter, $\boldsymbol{\theta}$, we use a two-step estimation procedure:

1. First, fix $\boldsymbol{\rho} = \boldsymbol{\rho}_{\mathbf{0}}$, then estimate the $\boldsymbol{\beta}_{\mathbf{0}}$ with GMM and
2. Find the optimal value of $\hat{\boldsymbol{\theta}} = [\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\beta}}]$ through GMM as the initial value of $\boldsymbol{\theta}_{\mathbf{0}} = [\boldsymbol{\rho}_{\mathbf{0}}, \boldsymbol{\beta}_{\mathbf{0}}]$ found in (1).

The estimated spatial lag $\hat{\boldsymbol{\rho}}$ is used to test the statistical significance of $\boldsymbol{\rho}$ by the Lagrange Multiplier (LM) test proposed by Anselin [30]. The LM statistic for spatial lag $\hat{\boldsymbol{\rho}}$ is defined as:

$$\mathbf{LM}_{\boldsymbol{\rho}} = [\mathbf{u}'\mathbf{W}\mathbf{y}/(\mathbf{u}'\mathbf{u}/\mathbf{n})]^{2}/\mathbf{D} \tag{5}$$

where $\mathbf{D} = [(\mathbf{W}\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{W}\mathbf{X}\boldsymbol{\beta})/\hat{\boldsymbol{\sigma}}^{2}] + \mathbf{tr}(\mathbf{W}^{2} + \mathbf{W}'\mathbf{W})$ with $\hat{\boldsymbol{\sigma}}^{2} = [\mathbf{e}_{0} - \hat{\boldsymbol{\rho}}\mathbf{e}_{\mathrm{L}}]'[\mathbf{e}_{0} - \hat{\boldsymbol{\rho}}\mathbf{e}_{\mathrm{L}}]/\mathbf{n}$, $\mathrm{e}_{\mathbf{0}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, and $\mathbf{e_L} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$.

The spatial lag weights matrix between borrowers on the P2P platform, W, is defined in Section 3.2.

3.2 Borrowers'relation network

In this study, we construct a network with each borrower as a node and the distance between them as an edge to represent the relationship between the borrowers. The distance between them is defined as the degree of similarity in terms of their hard and soft information. Similarity between numeric information is easily defined by Euclidean distance, but defining similarity between categorical information is a challenge. We use a method proposed by Ahmad and Dey [31] to calculate the distance between borrowers with mixed numeric and categorical information.

Let us assume $\mathbf{B_i}$ and $\mathbf{B_j}$ are two borrowers with $\mathbf{m}$ hard and soft information attributes: $\mathbf{X_1}, \ldots, \mathbf{X_m}$. The two borrowers may be represented as $\mathbf{B_i} = \{\mathbf{X_{i1}}, \mathbf{X_{i2}}, \ldots, \mathbf{X_{im}}\}$ and $\mathbf{B_j} = \{\mathbf{X_{j1}}, \mathbf{X_{j2}}, \ldots, \mathbf{X_{jm}}\}$ where the first $\mathbf{m_r}$ attributes are numeric, the next $\mathbf{m_c}$ attributes are categorical, and $\mathbf{m_r} + \mathbf{m_c} = \mathbf{m}$. The distance between $\mathbf{B_i}$ and $\mathbf{B_j}$, denoted by $\mathbf{Dist}(\mathbf{B_i}, \mathbf{B_j})$ is computed as follows:

$$\mathbf{Dist}(\mathbf{B_i},\mathbf{B_j}) = \sum\nolimits_{\mathbf{t=1}}^{\mathbf{m_r}} (\mathbf{s_t}(\mathbf{X_{it}} - \mathbf{X_{jt}}))^2 + \sum\nolimits_{\mathbf{t=m_r+1}}^{\mathbf{m}} (\boldsymbol{\delta}(\mathbf{X_{it}},\mathbf{X_{jt}}))^2. \tag{6}$$

where $\mathbf{s_t}$ is the significance of the t-th numeric attribute, and $\boldsymbol{\delta}(\mathbf{X_{it}}, \mathbf{X_{jt}})$ is a distance function between the t-th categorical attributes in $\mathbf{B_i}$ and $\mathbf{B_j}$. The distance between two distinct values, $\mathbf{c_1}$ and $\mathbf{c_2}$, of any categorical attribute $\mathbf{X}_t$ is given by:

$$\boldsymbol{\delta}(\mathbf{c_1},\mathbf{c_2}) = \left(\frac{\mathbf{1}}{\mathbf{m-1}}\right)\sum\nolimits_{\mathbf{t'=\ 1,\cdots,m,\ t\neq t'}} \boldsymbol{\delta}^{\mathbf{tt'}}(\mathbf{c_1},\mathbf{c_2}) \tag{7}$$

where $\boldsymbol{\delta}^{\mathbf{tt'}}$ $(\mathbf{c_1}, \mathbf{c_2}) = \mathbf{P_t}(\mathbf{c'}\ |\mathbf{c_1}) + \mathbf{P_t}(\sim\mathbf{c'}\ |\mathbf{c_2}) - \mathbf{1}$, $\mathbf{c'}$ denotes a subset $\mathbf{C}$ of values of $\mathbf{X_{t'}}$ that maximizes the quantity $\mathbf{P_t}(\mathbf{c'}\ |\mathbf{c_1}) + \mathbf{P_t}(\sim\mathbf{c'}\ |\mathbf{c_2})$; $\sim\mathbf{c'}$ denotes the complementary set of values occurring for attribute $\mathbf{X_{t'}}$; and $\mathbf{P_t}(\mathbf{c'}\ |\mathbf{c_1})$ denotes the conditional probability that an element having value $\mathbf{c_1}$ for $\mathbf{X_{t'}}$ has a value belonging to $\mathbf{c'}$ for $\mathbf{X_{t'}}$. To compute the significance of normalized numeric attributes, we discretize them to have $\mathbf{L}$ equal intervals: $\mathbf{u}[\mathbf{1}]$, $\mathbf{u}[\mathbf{2}]$, $\cdots$, $\mathbf{u}[\boldsymbol{l}]$. The significance of the t-th numeric attribute, $\mathbf{s_t}$, is computed as:

$$\mathbf{s_t} = \sum\nolimits_{\boldsymbol{l}_1=1}^{\mathbf{L-1}} \sum\nolimits_{\boldsymbol{l}_2>\boldsymbol{l}_1}^{\mathbf{L}} \boldsymbol{\delta}(\mathbf{u_t}[\boldsymbol{l}_1],\mathbf{u_t}[\boldsymbol{l}_2])/(\mathbf{L}(\mathbf{L}-\mathbf{1})/\mathbf{2}). \tag{8}$$

The relationship between two borrowers ($\mathbf{B_i}$ and $\mathbf{B_j}$) is mapped so that the closer the distance is, the stronger the relationship. We use double-power distance weights, and the degree of relationship between $\mathbf{B_i}$ and $\mathbf{B_j}$ is evaluated as follows:

$$\mathbf{W_{ij}} = \begin{cases} [\mathbf{1} - (\mathbf{Dist}(\mathbf{B_i},\mathbf{B_j})/\mathbf{d})^2]^2, & \mathbf{0} \leq \mathbf{Dist}(\mathbf{B_i},\mathbf{B_j}) \leq \mathbf{d} \\ \mathbf{0}, & \mathbf{Dist}(\mathbf{B_i},\mathbf{B_j}) > \mathbf{d} \end{cases} \tag{9}$$

where $\mathbf{d}$ donates the maximum radius of influence (bandwidth). To use $\mathbf{W_{ij}}$ as a spatial weight matrix, row normalization is performed.

3.3 Evaluation metric

To measure the performance of the proposed spatial probit model, we used the following evaluation metrics: accuracy, precision, recall, F1 score, and area under the receiver operator characteristic (ROC) curve. These 4 indicators are the most used indicators for performance evaluation of binary classification tasks such as default prediction. The accuracy is the most intuitive performance indicator of a classification model and is defined as the ratio of correct to total predictions. The precision is the percentage of borrowers that actually defaulted out of those who were predicted to default. The recall is the percentage of borrowers predicted to default out of those actually defaulted. The F1 score is the harmonic mean of the precision and recall. Precision, recall, and f1 score are used as important indicators in a credit scoring task where borrowers with default is much less than borrowers with fully paid [32]. The ROC curve for a binary classification problem represents the true positive proportion as a function of the false positive proportion.

4. Data

We used LCL data from Lending Club, the largest online credit marketplace offering P2P lending worldwide. This data is open to public and provides 2.26 million loan records from June 2007 to December 2018. There are 36-month and 60-month long loans provided by LCL data. Therefore, there exist quite a few borrowers who belong to the "Current" category out of those who received the loan after 2013. Their default record is unknown. Because of these data problems, we only used loans issued in 2012. In the 2012 loan record, Fully Paid, Default, and Charged Off status existed, and in this study, Fully Paid was defined as a good result and the other two were defined as bad results.

In sum, our dataset consists of 51,314 issued loans, including 8,241 defaults. The LCL dataset describes 145 attributes of borrowers but like previous studies, selected only the important attributes with several references [18, 33, 34]. Brief descriptions of the seven numeric and five categorical attributes used in this study are presented in Table 1. Employment length and home ownership are soft information not directly representing borrowers' financial status. We removed the missing values for the 12 variables and obtained 37,012 borrowers with fully paid loans and 7,080 borrowers with defaulted loans.

Table 1. Description of attributes used in this study.

Type	Variable	Definition
Numeric	Annual income	The annual income provided by the borrower during registration
	Debt to income	The borrower's debt-to-income ratio: monthly payments on the total debt obligations, excluding mortgage, divided by self-reported monthly income
	Inquiries in the last six months	The number of inquiries by creditors during the past 6 months
	Loan amount	The listed amount of the loan applied for by the borrower
	Open accounts	The number of open credit lines in the borrower's credit file
	Revolving balance	The total credit revolving balance
	Revolving utilization rate	The amount of credit the borrower is using relative to all available revolving credit
Categorical	Employment length	Employment length in years: integers between 0 and 10, with 0 meaning less than one year and 10 meaning ten or more years
	Grade	Lending Club categorizes borrowers into seven different loan grades from A down to G, A-grade being the safest.
	Home ownership	The home ownership status information provided by the borrower during registration: rent, own, and mortgage
	Loan length	The length of time (years) that workers have been with their current employer: 36 months, 60 months
	Loan purpose	Includes 14 loan purposes: wedding, credit card, car loan, major purchase, home improvement, debt consolidation, house, vacation, medical, moving, renewable energy, educational, small business, and other

https://doi.org/10.1371/journal.pone.0261737.t001

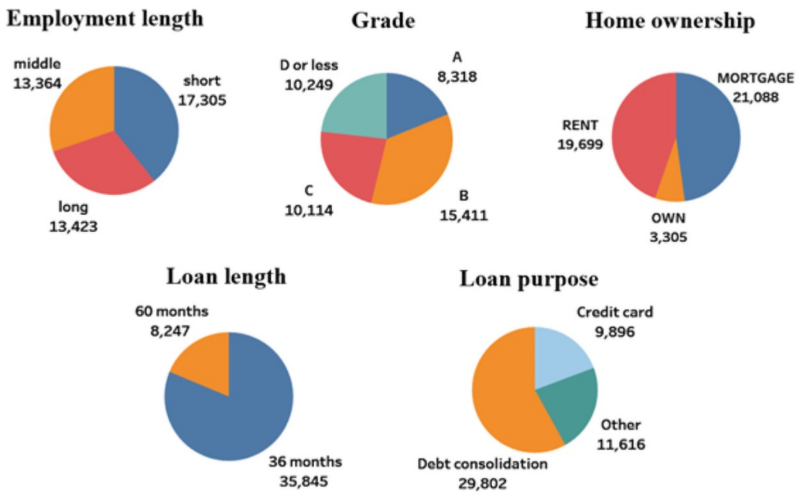

Fig 1. The distribution of categories for each categorical attribute.

https://doi.org/10.1371/journal.pone.0261737.g001

We performed preprocessing, taking into account the dispersion of each attribute. "Annual income," "Loan amount," and "Revolving balance" are log-transformed to reduce variance. Since 77% of all borrowers are classified as A, B, or C in the "Grade" attribute, classifications D to G are combined together as D or less. Since 78% of all borrowers are also concentrated under the categories debt consolidation and credit card in the "Loan purpose" attribute, we combined the remaining categories into the category other. The "Employment length" attribute is newly categorized as short, representing less than five years; middle, five to nine years; and long, 10 years or more. Thus, the categorical variables increased to nine, and their distribution is shown in Fig 1.

We performed the Welch's T test on the difference between borrowers with fully paid loans and borrowers with defaulted loans for numeric attributes, as shown in Table 2. There were no statistically significant differences in the "Revolving balance" attribute under the significance level of 0.05. However, for attributes related to income, borrowers with fully paid loans are observed to be more stable than borrowers with defaulted loans.

We performed a chi-square test to check if being in default in a categorical attribute is independent of its categories. Table 3 shows for each category the number of borrowers with fully paid loans and those with defaulted loans, the ratio of borrowers with defaulted loans to borrowers with fully paid loans, and the chi-square statistic with the corresponding p-value. Depending on the "Grade" and the "Loan length," the default-to-fully-paid ratio was quite

Table 2. Result of the Welch's T test for numeric attributes.

Attributes	Fully paid loans	Defaulted loans	P-value
Annual income (log)	11.0397	10.9587	<0.0001
Debt to income	16.7235	18.2408	<0.0001
Inquiries in the last six months	0.7908	0.9697	<0.0001
Loan amount (log)	9.2938	9.4174	<0.0001
Open accounts	11.1021	10.757	<0.0001
Revolving balance (log)	9.2420	9.2568	0.29
Revolving utilization rate	57.4090	62.2409	<0.0001

https://doi.org/10.1371/journal.pone.0261737.t002

Table 3. Result of the chi-squared test for categorical attributes.

Attribute	Category	Fully Paid Loans	Defaulted Loans	Defaulted / Fully Paid Loans	Chi-squared test
Employment length	Short	14,592	2,713	0.19	4.5902 (0.1)
	Middle	11,148	2,216	0.2	
	Long	11,272	2,151	0.19	
Grade	A	7,757	561	0.07	1589.9 (<0.0001)
	B	13,493	1,918	0.14	
	C	8,251	1,863	0.23	
	D or less	7,511	2,738	0.36	
Home ownership	Mortgage	17,935	3,153	0.18	37.839 (<0.0001)
	Own	2,762	543	0.2	
	Rent	16,315	3,384	0.21	
Loan length	36 months	31,030	4,815	0.16	978.29 (<0.0001)
	60 months	5,982	2,265	0.38	
Loan purpose	Credit card	7,365	1,074	0.15	99.942 (<0.0001)
	Debt consolidation	21,432	4,481	0.21	
	Other	8,215	1,525	0.19	

https://doi.org/10.1371/journal.pone.0261737.t003

different. The "Employment length" did not show a statistically significant difference under the p-value of 0.05.

5. Experiment

In our dataset, borrowers with defaulted loans account for 16% of the total; thus, there is a class imbalance problem. This leads to a problem whereby the classification model is trained to be biased to predict a major class, and significantly reduces the performance of the prediction of a minor class [35]. To alleviate this problem, we utilized the under-sampling method [36]. We sampled 5,000 borrowers with fully paid loans and 5,000 borrowers with defaulted loans. We limited the range of some numeric attributes to control the dispersion of their min-max normalization. Values greater than 3 for "Inquiries in the last 6 months" and 26 for "Open accounts" were excluded from the sampling process. The spatial weight matrix, W, has been built from the sampled dataset, as described in section 3.2. Numeric variables were divided into three sections of equal length (**L**). The bandwidth (**d**) was set to 0.06059, which was the third quantile value of distances between borrowers.

To consider the allowable computation time for parameter estimation, we sampled 2,000 borrowers from the sample dataset, which was divided into 1,500 train datasets and 500 test datasets. Using the train dataset, the parameters: $\hat{\boldsymbol{\theta}} = [\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\beta}}]$ were estimated by GMM. To find the initial $\boldsymbol{\rho}_{\mathbf{0}}$, we observed a change in the "area under the curve" (AUC) for the test dataset by increasing the $\boldsymbol{\rho}_{\mathbf{0}}$ from 0 to 1 at intervals of 0.1. As shown in Fig 2, with an initial $\boldsymbol{\rho}_{\mathbf{0}}$ of 0.5, the test AUC was the highest, at 0.6855. This shows that borrowers are not independent in the borrowers' relation network, and that there is sufficient spatial autocorrelation between borrowers with defaulted loans.

Table 4 compares the baseline model, logistic regression model without spatial component, with the model presented in this study. In the baseline model, ten attributes were statistically significant at the significance level of 0.1. The default probability of the borrower has a strong negative correlation with the "log(Annual income)" and "log(Revolving balance)" attributes. However, it has a positive correlation with the "Debt to income," "Revolving utilization rate," "Grade," "Loan length," and "Loan purpose." In the spatial probit model proposed in this

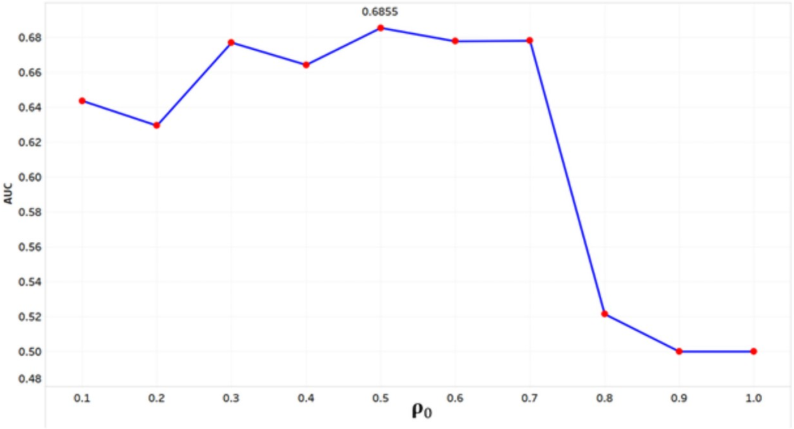

Fig 2. Test AUC variation with initial ρ_0.

https://doi.org/10.1371/journal.pone.0261737.g002

Table 4. Result of the estimation of the baseline and SAR models.

	Baseline model			Spatial probit model		
	Estimate	**Std. Error**	**Pr(>|Z|)**	**Estimate**	**Std. Error**	**Pr(>|Z|)**
Intercept	-0.076	0.563	0.893	-0.481	0.867	0.579
log(Annual income)	-1.714 ***	0.546	0.002	-0.417	0.559	0.455
Debt to income	0.512 *	0.301	0.089	0.963 ***	0.312	0.002
Inquiries in the last 6 months	0.192	0.176	0.276	-0.007	0.180	0.969
log(Loan amount)	0.584	0.384	0.128	-0.905 **	0.404	0.025
Open accounts	0.444	0.361	0.219	-0.557	0.371	0.133
log(Revolving balance)	-2.106 ***	0.786	0.007	0.368	0.813	0.651
Revolving utilization rate	0.591 *	0.327	0.071	-0.766 **	0.334	0.022
Employment length (short)	-0.006	0.131	0.963	-0.039	0.131	0.768
Employment length (long)	0.155	0.142	0.275	0.104	0.143	0.469
Grade (B)	0.457 **	0.194	0.018	0.677 **	0.328	0.038
Grade (C)	0.805 ***	0.213	<0.001	1.085 ***	0.362	0.003
Grade (D or less)	1.081 ***	0.235	<0.001	1.394 ***	0.393	<0.001
Home ownership (Own)	-0.062	0.213	0.771	-0.179	0.213	0.401
Home ownership (Rent)	0.111	0.124	0.391	0.094	0.124	0.446
Loan length (60 months)	0.581 ***	0.163	<0.001	0.488 ***	0.182	0.007
Loan purpose (debt consolidation)	0.269 *	0.155	0.083	0.129	0.154	0.404
Loan purpose (other)	0.395 **	0.189	0.036	0.118	0.187	0.529
Spatial component (**ρ**)				Estimate	LM_ρ	p-value
				0.505 ***	273.282	<0.001
Accuracy	0.624			0.652		
Precision	0.63			0.619		
Recall	0.6			0.792		
F1 score	0.615			0.695		
AUC	0.696			0.713		

*, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

https://doi.org/10.1371/journal.pone.0261737.t004

Table 5. Result of the estimation of the SAR model with 500 repetitions.

Initial Rho	0.2	0.5	0.8
	Mean	Mean	Mean
Accuracy (Baseline model: 0.614)	0.606	0.613	0.592
Precision (Baseline model: 0.612)	0.598	0.590	0.564
Recall (Baseline model: 0.622)	0.647	0.745	0.809
F1 score (Baseline model: 0.617)	0.621	0.658	0.664
AUC (Baseline model: 0.660)	0.650	0.665	0.652

https://doi.org/10.1371/journal.pone.0261737.t005

study, seven attributes were statistically significant at the significance level of 0.1. The "log (Annual income)" and "log(Revolving balance)" attributes were underestimated over the baseline model and were not statistically significant. Instead, "log(Loan amount)" and "Revolving utilization rate" have negative coefficients. In addition, the spatial autocorrelation component between borrowers with defaulted loans was 0.505, which was very significant under the significance level of 0.05. Compared to the baseline model, there was an increase in accuracy and AUC. In particular, the proposed model has remarkably increased recall and F1-score, which can be expected to have significant spatial autocorrelation between borrowers with defaulted loans. The additional consideration of spatial autocorrelation in the borrower relation network significantly improved the performance of logistic regression.

We sampled the training and test dataset 500 times and observed changes in the test performance differences of the baseline and spatial probit models in the entire dataset. To observe the strength of autocorrelation between borrowers with defaulted loans, the initial $\boldsymbol{\rho_0}$ was set to 0.2, 0.5, and 0.8. The results are shown in Table 5. The larger the initial rho, the higher the recall, which means the higher the predictability of the borrowers with defaulted loans. However, too large an initial value creates the risk of reduced accuracy and AUC. In our experiment, when the initial rho is 0.5, the AUC is slightly higher, and the F1-score is significantly higher than the baseline model. Therefore, a consideration of the appropriate level of spatial autocorrelation is expected to contribute significantly to the prediction of the default risk of a borrower.

6. Conclusion

This study proposed a spatial probit model to improve default prediction by reflecting the relationship between borrowers, which is defined by the similarity of their characteristics.

We applied this method to 2012 LCL data. We found an evidence of a high level of spatial autocorrelation between borrowers with defaulted loans. Reflecting the spatial autocorrelation among loan applicants did not result in an overall improvement in the accuracy of the default prediction but instead, a significant improvement in the F1-score. An increase in the F1 score is a very significant contribution, since finding borrowers with high default risk is a more important issue than finding normal borrower. This study showed that the additional information of spatial autocorrelation between borrowers with high default risk can alleviate the class imbalance problem in the loan dataset and provide a high predictive power for high default risk borrowers.

However, this study has some limitations. Since the spatial weighting matrix increases enormously in proportion to the square of the number of observations, there are time and memory difficulties in using all the data. In addition, the calculation of the inverse of $(\mathbf{I} - \boldsymbol{\rho}\mathbf{W})$ in the parameter estimation process using GMM requires a large amount of computation. Because of these constraints on the spatial weighting matrix, we sampled a small number instead of the entire dataset. If the computing power is complemented and the constraints on the spatial weighting matrix are relaxed, then more robust default predictive modeling can be expected.

Supporting information

S1 File.
(ZIP)

Author Contributions

Conceptualization: Jong Wook Lee, So Young Sohn.

Data curation: Jong Wook Lee.

Formal analysis: Jong Wook Lee.

Funding acquisition: So Young Sohn.

Methodology: Jong Wook Lee.

Project administration: So Young Sohn.

Resources: So Young Sohn.

Software: Jong Wook Lee.

Supervision: So Young Sohn.

Visualization: Jong Wook Lee.

Writing – original draft: Jong Wook Lee.

Writing – review & editing: So Young Sohn.

References

1. Angilella S., & Mazzù S. (2015). The financing of innovative SMEs: A multicriteria credit rating model. European Journal of Operational Research, 244(2), 540–554.
2. Kim Y., & Sohn S. Y. (2007). Technology scoring model considering rejected applicants and effect of reject inference. Journal of the Operational Research Society, 58(10), 1341–1347.
3. Jeon H., & Sohn S. Y. (2008). The risk management for technology credit guarantee fund. Journal of the Operational Research Society, 59(12), 1624–1632.
4. Sohn S. Y., Doo M. K., & Ju Y. H. (2012). Pattern recognition for evaluator errors in a credit scoring model for technology-based SMEs. Journal of the Operational Research Society, 63(8), 1051–1064.
5. Ju Y. H., & Sohn S. Y. (2015). Stress test for a technology credit guarantee fund based on survival analysis. Journal of the Operational Research Society, 66(3), 463–475.
6. Agosto A., Giudici P., & Leach T. (2019). Spatial regression models to improve P2P credit risk management. Frontiers in artificial intelligence, 2, 6. https://doi.org/10.3389/frai.2019.00006 PMID: 33733095
7. Wei Y., Yildirim P., Van den Bulte C., & Dellarocas C. (2016). Credit scoring with social network data. Marketing Science, 35(2), 234–258.
8. Óskarsdóttir M., Bravo C., Sarraute C., Vanthienen J., & Baesens B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. Applied Soft Computing, 74, 26–39.
9. Zeng X., Liu L., Leung S., Du J., Wang X., & Li T. (2017). A decision support model for investment on P2P lending platform. PloS one, 12(9), e0184242. https://doi.org/10.1371/journal.pone.0184242 PMID: 28877234
10. Serrano-Cinca C., Gutiérrez-Nieto B., & López-Palacios L. (2015). Determinants of default in P2P lending. PloS one, 10(10), e0139427. https://doi.org/10.1371/journal.pone.0139427 PMID: 26425854
11. 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.

12. 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.

13. 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.

14. Kruppa J., Ziegler A., & König I. R. (2012). Risk estimation and risk prediction using machine-learning methods. Human Genetics, 131(10), 1639–1654. https://doi.org/10.1007/s00439-012-1194-y PMID: 22752090

15. Ma Z., Hou W., & Zhang D. (2021). A credit risk assessment model of borrowers in P2P lending based on BP neural network. PloS one, 16(8), e0255216. https://doi.org/10.1371/journal.pone.0255216 PMID: 34343180

16. Harris T. (2013). Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions. Expert Systems with Applications, 40(11), 4404–4413.

17. Yao X., Crook J., & Andreeva G. (2015). Support vector regression for loss given default modelling. European Journal of Operational Research, 240(2), 528–538.

18. Malekipirbazari M., & Aksakalli V. (2015). Risk assessment in social lending via random forests. Expert Systems with Applications, 42(10), 4621–4631.

19. 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.

20. Abdou H. A., & Pointon J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. Intelligent Systems in Accounting, Finance and Management, 18(2–3), 59–88.

21. Lin X., Li X., & Zheng Z. (2017). Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in China. Applied Economics, 49(35), 3538–3545.

22. Dorfleitner G., Priberny C., Schuster S., Stoiber J., Weber M., de Castro I., 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.

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

24. Calabrese R., Elkink J. A., & Giudici P. S. (2017). Measuring bank contagion in Europe using binary spatial regression models. Journal of the Operational Research Society, 68(12), 1503–1511.

25. Verbeek M. (2008). A guide to modern econometrics. John Wiley & Sons.

26. Calabrese R., & Elkink J. A. (2014). Estimators of binary spatial autoregressive models: A Monte Carlo study. Journal of Regional Science, 54(4), 664–687.

27. Pinkse J., & Slade M. E. (1998). Contracting in space: An application of spatial statistics to discrete-choice models. Journal of Econometrics, 85(1), 125–154.

28. Klier T., & McMillen D. P. (2008). Clustering of auto supplier plants in the United States: generalized method of moments spatial logit for large samples. Journal of Business & Economic Statistics, 26(4), 460–471.

29. Kelejian H. H., & Prucha I. R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. The Journal of Real Estate Finance and Economics, 17(1), 99–121.

30. Anselin L. (1988). Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. Geographical Analysis, 20(1), 1–17.

31. Ahmad A., & Dey L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering, 63(2), 503–527.

32. Li W., Ding S., Chen Y., & Yang S. (2018). Heterogeneous ensemble for default prediction of peer-to-peer lending in China. Ieee Access, 6, 54396–54406.

33. Li Z., Tian Y., Li K., Zhou F., & Yang W. (2017). Reject inference in credit scoring using semi-supervised support vector machines. Expert Systems with Applications, 74, 105–114.

34. Szwabe A., & Misiorek P. (2018, September). Decision Trees as Interpretable Bank Credit Scoring Models. In International Conference: Beyond Databases, Architectures and Structures (pp. 207–219). Springer, Cham.

35. Longadge, R., & Dongre, S. (2013). Class imbalance problem in data mining review. arXiv preprint arXiv:1305.1707.

36. Kotsiantis S. B., & Pintelas P. E. (2003). Mixture of expert agents for handling imbalanced data sets. Annals of Mathematics, Computing & Teleinformatics, 1(1), 46–55.