Economic Management & Global Business Studies
World Scientific Reports
2023, Volume1, Issue1, 2023, Type: Articles

Research article

An Information Processing and Decision Support model for Credit Default Prediction in Emerging Internet Finance Markets

Latifa Binti Zainal[1*] **Arijit Al-Masri**[1]

1.Universiti Malaysia Perlis, Faculty of Engineering Technology, Malaysia

**latifa.zainal@unimap.edu.my*

P2P (Peer to Peer) online lending is an emerging Internet finance mode that gathers small-amount fund lending to fund demander. This paper draws on the existing credit risk assessment research, combines rationality, science and other principles, according to the characteristics of the famous online loan platform, collects borrower information and combines computer technology to design a borrower credit risk assessment system. In this paper, we make improvements based on the famous LightGBM algorithm (Light Gradient Boosting Machine). Firstly, In the process of data input, the improved Convolutional Neural Network CNN model is adopted to extract features from the data. Specifically, the Global Average Pooling(GAP) layer is adopted to replace the full connection layer to improve the Convolutional Neural Network. This paper first proposes a P2P online loan default prediction model based on GCNN-LightgBM. The model integrates the advantages of the improved Convolutional Neural Network and LightGBM model, and realizes the efficient prediction of network loan default. Then, in order to improve the accuracy of P2P online loan borrower default prediction, this paper proposes a new model based on LightGBM and Bagging (LGB-BAG). LGB-BAG uses LightGBM as the base learner. With the help of LightGBM, which can effectively reduce the deviation of the model, and Bagging, which can reduce the variance of the model, the volatility of the prediction is further reduced (F1 variance), so that the LGB-BAG model has smaller deviation and variance, and the prediction effect is further improved. In our ablation experiment, the proposed model (GCNN-LGB-BAG) obtained an AUC of 0.86 and an accuracy of 0.97, both of which outperformed all benchmark models. This paper uses actual data to identify the loan risks of P2P online lending platforms, aiming to provide investment reference for investors and methodological support for relevant online lending regulators.

Keywords: Peer to Peer online lending; loan default prediction; Light Gradient Boosting Machine; Global Average Pooling;Bagging

Statement: The data in this study can be provided without reservation by the corresponding author, also, the authors have no potential conflicts of interest.

1 Introduction

In 2005, Zopa was born, which is the world's first P2P online loan platform, with more than 130,000 users and £550 million of loans to date. Borrowers enter the amount they want to borrow on Zopa, which offers the highest interest rate. Those with money to borrow can find the amount and interest rate they can accept on the website, which ensures security and fairness. Zopa measures the credit of individuals by spreading out their loans, and if they fail to pay for a long time, it will take mandatory repayment measures to avoid the risk of default. Through cooperation with Equifax credit rating company! Zopa signed relevant agreements to minimize bad debts. Zopa provides repayment protection insurance for loan investors and lenders. Prosper, the world's largest peer-to-peer lending platform, makes money by charging fees to both lenders and borrowers. Prosper strictly reviews the applicant users, who must meet the conditions of having a social security number, personal tax id number, etc. at the same time before they can register on the website. Because Prosper's credit system is so well established that it is fast and efficient for customers to authenticate on the platform, Prosper acts as a simple middleman and does not bear the risk of bad debts. Lending Club, which was launched in 2007 and passed SEC safety standards, will be backed by the U.S. government even in the face of a debt crisis until it goes bankrupt. P2P platforms do not have data to study when rating transactions due to privacy issues. However, after Prosper and Lending Club, two big Lending platforms in the United States, made large amounts of transaction data available to the public, many excellent experts in economics, management, psychology and sociology have devoted themselves to the study of online loan credit evaluation.

2 Research progress of P2P platform default

There are abundant researches on the credit risk of P2P online loan platform. Some scholars have studied the factors that affect the success of individual borrowers in obtaining loans. For example, Pope and Sydnor [1] studied the data of Prosper platform and found that older borrowers are more likely to borrow money than younger borrowers. Moreover, online lending platforms have racial discrimination in loans. Ravina [2] also conducted an empirical study on prosper platform data and found that appearance has a significant impact on loan interest rate, and borrowers with poor appearance will have higher loan interest rate.

Credit risk assessment is very important on financial lending. Lai Hui et al. [3] proposed the PCBS method because of the frequent occurrence of default events of

personal credit customers, and established a dynamic credit evaluation method of personal credit customers on this basis to reduce the risk loss caused by default. Herzenstein looked at data from Prosper, an American P2P lending platform, and found that having words such as "credible" and "successful" in self-rating had no significant effect on default behavior, but was significantly related to whether they paid on time and early. Meanwhile, borrowers who reported financial difficulties were at greater risk of default. At the risk of winning sympathy, platforms need to be more circumspect about such lending [4]. Michels found through research that the descriptive text information voluntarily disclosed by borrowers is significantly correlated with default behavior. People with more content in self-description are more likely to be favored by investors in loan application, and such borrowers are less likely to default, although the authenticity of description content cannot be verified [5]. Stein (2000) found that friendship increases the possibility of successful financing, lowers the interest rate of loans and reduces the post-default rate [6]. Lin (2013) studies show that borrowers with wider social connections can increase the success rate of borrowing, reduce borrowing costs, and be accompanied by a low default rate [7]. Pope and Sydnor (2011) studied the data of Prosper platform and found that the success rate of black people's borrowing was low and accompanied by high interest rate [8]. Emekter and Tu (2015) pointed out that as the credit rating of borrowers decreases, their default risk increases [9]. Puro et al. (2010) found that a larger loan amount is accompanied by a lower success rate of loan, and a higher interest rate can increase the success rate of loan [10]. Some scholars have carried out researches on investors of P2P lending. Burtch et al. (2014) found that lenders tend to invest in borrowers with similar geographical location and educational level [11]. Zhang and Liu(2012), Berkovich (2011), Lee and Lee (2012), and wu jiazhe (2015) found that there is a significant herding effect in P2P online lending mode [12,13,14].

With the development of artificial intelligence (AI), classical single classifiers such as SVM and ANN appear. Therefore, scholars at home and abroad also use these methods to improve the prediction accuracy of personal credit risk assessment. Ensemble learning method has become a hot topic because it can integrate base classifier and improve the classification effect. Nanni and Lumini [15] used credit data from Japan, Australia and Germany and applied Random Subspace, Bagging, Class Switching and Rotation Forest to study the credit problem of banks. Abellan and Castellano [16] used the actual credit data of six countries and five integrated learning methods to build the model. The results showed that compared with a single model, the integrated learning model had stronger warning ability and stability. In addition, florez-Lopez, Ramon-Jeronimo [17], Tsai et al. [18] studied these models. Cao Wei et al. [19] compared these different integrated learning methods. Among these ensemble learning methods, Bag-ging can reduce the variance of the model, but it is difficult to reduce the deviation of the model. Boosting can effectively reduce the deviation of the model. Therefore, a new model based on Light GBM and Bagging, named LGB-BAG, is proposed in this paper, which effectively combines the advantages of Boosting and Bag-ging ensemble learning strategies and improves the classification effect of the model.

At present, deep learning theory has made great progress in the field of Identification and classification [20,21,22,23] , and Convolutional Neural Network (CNN), as one of the important models of deep learning theory, has continuously indicates the potential in financial markets. Although CNN has got well performance in pattern recognition, But softmax layer can not be used to classify and extract features well. Single learning algorithms such as SVM [24], KNN [25] and integrated learning algorithms such as random forest [26] and XGBoost [27] have got well performance on pattern recognition. Nevertheless, in the current environment of big data and high dimension, those identification methods could not both outperformed on efficiency and accuracy. The LightGBM is a Gradient lifting method [28], which is optimized for the identification accuracy in light of Boosting. However, assuming the first sign is straightforwardly input into LightGBM, it will contain numerous repetitive signs without handling, which will consume a lot of memory space during model preparation and effectively objective over-fitting of LightGBM classifier.

To handle this problems, the GlobalAverage Pooling (GAP) layer is adopted to replace the full connection layer to improve the convolutional neural network (hereinafter referred to as GCNN). Finally, a hybrid intelligent model, GCNN-LGB-BAG is build in this paper.

3 Method Proposed

3.1.Gcnn-lightgbm module introduction

3.1.1 Convolution Layer and Pooling Layer

Convolutional layer is the most basic structure in convolutional neural network. Its main function is to extract features from input data. The convolution layer formula is:

$$y^{l(i,j)} = \sum_{j`=0}^{n-1} k_i^{l(j`)} x^{l(j+j`)} m \qquad (1)$$

In the formula, $y^{l(i,j)}$ is the output after convolution; $k_i^{l(j`)}$ is the j`-th weight of the i-th convolution kernel in the l-th layer; $x^{l(j+j`)}$ is the j`-th local region convolved in the l layer; M is the width of the convolution kernel. The operation of convolution is linear, but most of the samples are linearly indivisible. In order to solve the problem that the linear model cannot effectively deal with nonlinear samples, a nonlinear activation function is introduced into the convolution layer. Common activation functions include tanh， sigmoid and ReLU, etc. ReLU function is adopted in this paper, and its expression is

$$\mathrm{a}^{l(i,j)} = max\{0, y^{l(i,j)}\} \qquad (2)$$

In the formula, $\mathrm{a}^{l(i,j)}$ is the value of ReLU function after the convolution output $y^{l(i,j)}$ is activated. Pooling layer, also known as under-sampling layer or under-sampling layer, is mainly used for feature selection and information filtering, maximum pooling is used in this paper, and the maximum value in the region is

selected as the pooled value of the region, and its expression is

$$p^{l(i,j)} = max\{a^{l(i,j)}\}; (j-1)n+1 \leq t \leq jn, \qquad (3)$$

In the formula, : $p^{l(i,j)}$ is the pooled output.

3.2 Global average Pooling layer

Classical convolutional neural networks tend to connect one or more fully connected layers after several times of convolution and pooling, and finally adopt softmax layer for classification. Each neuron in the full connection layer is connected with all neurons in the upper layer to fuse the features extracted from the convolution layer. Due to the characteristics of full connection, the number of parameters of full connection layer is very large, which will not only reduce the training speed of model, but also easily cause overfitting. To make up the defects of the link layer, the literature [29] puts forward the concept of global average pooling layer, the characteristics of convolution output of each figure averaging, make each characteristic figure only one output and does not require training tuning parameters, thus greatly reduce the network parameters, the model is more robust and has a better fitting effect.

Figure 1 indicates the comparison between FC and GAP. Assuming that the last convolution layer outputs a feature graph of 4*2*2 and the output neuron of the full connection layer is 4, a total of 4*2*2*4 = 64 parameters need to be trained, and 4 outputs are also obtained without using any parameters. Therefore, it is easy to see that replacing the full connection layer with the global average pooling layer can greatly reduce the parameter computation of the classical convolutional neural network from this simple comparison structure diagram.

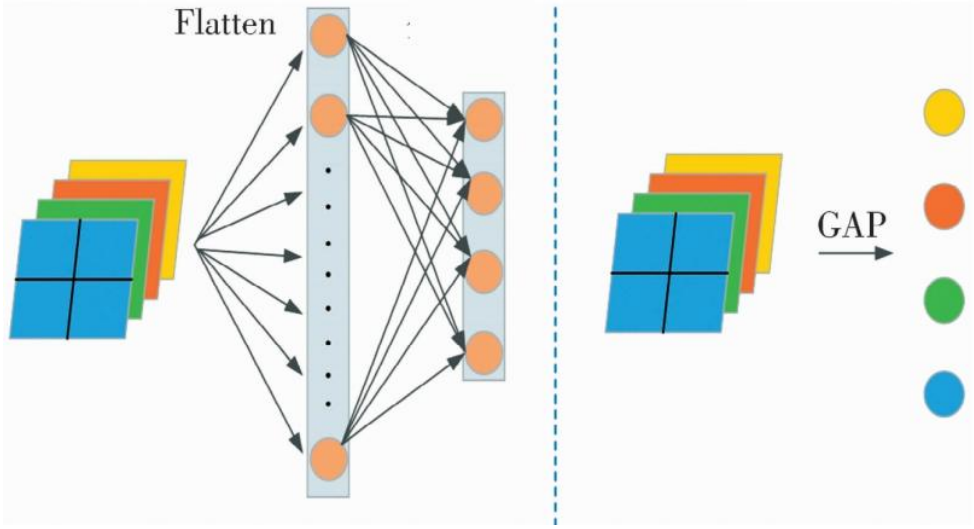

Figure 1 Comparison between FC and GAP

3.3 LightGBM

The key of LightGBM is that gradient-based One-side Sampling （GOSS） and Exclusive Feature Bundling are integrated based on Boosting algorithm (EFB) these two new methods.

3.4 GOSS algorithm

GOSS is an algorithm that reduces the amount of data but maintains accuracy. Each data has a different gradient value, and the smaller the gradient value is, the smaller

the data training error is. If the data with small gradient is completely discarded, the distribution of the data will be changed, thus affecting the accuracy of the training model. GOSS proposed an ingenious sampling method, and the specific algorithm steps are as follows:

1) All the data of the features to be split are arranged in absolute value from largest to smallest;

2) Select the first a % of the largest data;

3) Randomly select B % data from the remaining smaller gradients,

Multiply it by a constant coefficient (1 -a) over b. The above sampling method not only keeps all the large gradient instances, but also ensures that part of the small gradient samples can be trained. By introducing constant coefficient to the small gradient data, it can be as consistent as possible with the total data distribution, so as to ensure the accuracy of training samples and improve the training speed while reducing the number of training samples.

3.5 EFB algorithm

EFB is an algorithm that reduces the number of features but keeps the accuracy. In practice, high-dimensional data generally have sparsity, and EFB uses sparsity to design a clever and non-destructive method to reduce feature dimensions. Usually, the bound of the sparse characteristics are mutually exclusive, such as like one - hot features not to a non-zero value at the same time, such features tied up not lost information, but also exist some imperfect mutually exclusive features, EFB novel algorithm proposed conflict than the index to measure the extent of not mutually exclusive, when conflict is small, fewer feature packages can be obtained by fusing and binding these incompletely mutually exclusive features, which greatly reduces the number of features and improves computing efficiency.

3.6 GCNN-LighTGBM model

3.6.1 Model structure diagram

The GCNN-LightgBM model is mostly made out of the convolutional pooling layer, the worldwide normal pooling layer and the LightGBM classifier. Before the first one-layered vibration signal is contribution to the convolution layer, arbitrary inactivation with likelihood of 0.2 is completed to further develop the speculation capacity of the preparation model and the security of arrangement under factor load conditions. The convolutional pooling layer comprises of two layers. In the main layer, a huge convolutional piece is utilized to acquire more successful data in low recurrence band of the first sign [30]. The component pictures got by two-level convolutional pooling are input into the worldwide normal pooling layer, and the auxiliary element extraction and information aspect decrease are accomplished by averaging each element picture. At long last, the separated low-layered highlights are input into the LightGBM classifier for characterization.

3.7 LGB-BAG model

Bagging and Boosting, an ensemble learning method, and LightGBM as a base learning device are introduced. Finally, a new model, LGB-BAG model, is created based on the fusion of LightGBM and Bagging.

3.7.1 Bagging and Boosting Ensemble Learning Method

Bagging [31] is a method to integrate multiple different base classifiers into one ensemble classifier. Based on bootstrap sampling, Bagging repeatedly samples different data sets and trains base classifiers with high generalization ability and large difference degree on different data sets. When the prediction set obtained from a group of base classifiers predicts a class standard, voting method is adopted to determine the class standard with the most votes as the prediction class standard of this sample. This algorithm is also a parallel integrated learning method to improve the time efficiency of the algorithm.

Boosting [32] is a serial ensemble learning method, whose function model is superposition. The latter base learner will constantly modify or improve the results of the former base learner, and eventually each base learner is superimposed. Among them, Gradient Boosting is an important method in Boosting, which selects the Gradient descent direction during iteration to ensure the best final result. There are many famous algorithms based on this method, such as GBDT, XGBoost and Light-GBM.

3.7.2 LightGBM

LightGBM [33] (Light Gradient Boosting Machine) is an open source algorithm developed by the DMTK team at Microsoft Research Asia. It is an improved model based on decision tree and Gradient Boosting. LightGBM and XGBoost algorithms are known as the "heaven sword" and "dragon sword" in machine learning respectively. LightGBM has many advantages: The algorithm based on histogram has faster training speed and higher efficiency; less memory occupation; supports parallel computing, and has the ability to deal with big data due to the reduction in training time.

Two important innovations in LightGBM are Leaf growth strategies using histogram algorithms and Leaf-wise with depth constraints:

(1) Histogram algorithm. One of LightGBM's innovations is based on the histogram algorithm. During the calculation, the model converts floating-point values into discrete values, generating a histogram. The result of accumulating statistics in the graph with discrete values as indexes is to greatly reduce memory footprint for to find the optimal segmentation point.

(2) Leaf growth strategy of Level-wise with depth limitation. Most decision trees use the Level-wise strategy. However, Level-wise is an inefficient algorithm, which brings a lot of unnecessary overhead for its indiscriminate treatment of leaves at the same layer. Compared with Level-wise strategy, Leaf-wise is more efficient. It has such a cycle: each time from all the current leaves, find the Leaf with the maximum splitting gain, and then split. Therefore, in the case of the same number of splits, Level-wise has lower error and higher accuracy.

The disadvantage of Leaf growth strategy of Leaf-WISE with depth limitation is that when the sample size is small, Leaf-wise may grow relatively deep trees, leading to over-fitting. So LightGBM adds a maximum depth limit to prevent overfitting.

3.7.3 LGB-BAG

On the premise that base classifiers are independent from each other, it can be inferred from Hoeffding inequality [34] that with the increase in the number of learners in the

integration, the error rate of the integration will decrease exponentially and finally approach zero. Bagging re-selects training sets to increase the difference degree of classifier integration, improve generalization ability and reduce the risk of over-fitting.

In this paper, LightGBM (decision tree and Boosting) is used as a base classifier, and Bagging is used as an ensemble learning method to construct THE LGB-BAG model. Lgb-bag is a decision tree based algorithm with Boosting and Bagging. Bagging can reduce the variance of the model, while Boosting can effectively reduce the deviation of the model. Therefore, in theory, LGB-BAG can not only reduce the variance of the model, but also effectively reduce the model deviation.

The LGB-BAG algorithm is as follows: Repeat T times, each time m samples are randomly put back from the training sample set with the size of M, and the base classifier LightGBM is used for training. T base classifiers are obtained by using the same method, and a sequence of classification functions h_1 (x), h_2 (x),... and h_T (x) is obtained. The final classification function H (x) adopts voting method, and the class standard with the most votes is determined as the prediction class standard of this sample. When the unknown sample X is classified, each base classifier gets a classification result, T base classifiers vote, and the class with the most votes is determined as the prediction class of sample X.

The LGB-BAG algorithm is described as figure2 as follows:

Input: Training set $D(x) = \{(x_1,y_1),(x_2,y_2),...,(x_m,y_m)\}$;
Algorithms Bagging and LightGBM;
Number of training wheels T.

Process:

1. For t = 1 , 2 , ⋯ , T ;
2. For training set D, the autonomous sampling method is used, and m samples are randomly selected to create sample set D_t;
3. The base classifier H_T (x) is obtained by using D_t and LightGBM.

Output: $H(x) = \text{argmax} \sum_{t=1}^{T} h_t(x)$
The ensemble classifier H (x) is used to classify the unknown sample X:
1. When the unknown sample X is classified, each classifier H_T (x) gets a classification result. T classifiers vote, and the class with the most votes is determined as the prediction class of sample X.
2. The voting algorithm is used to output the classification results $H(x) = \text{argmax} \sum_{t=1}^{T} h_t(x)$

Figure 2.The introduction of LGB-BAG algorithm

4. Data statistics and pre-treatment

4.1 Introduction to data

As domestic online Lending platforms do not release specific customer desensitization data, this paper uses the customer desensitization data released by

Lending Club in the first half of 2019, which is an early start abroad, for analysis. Data from Lending Club's website. The original data set contains 246,814 customer desensitization data with 150 variable dimensions.

There are 7 loan states in the dataset, including Current, Fully Paid, In Grace Period, Default, Charged Off and Late(16-30 days)、Late(31-120 days). As the object of this study is the default phenomenon which has become an established fact, customers whose loans are not yet due cannot be taken as samples, and customers whose accounts have been cancelled are not related to this study, so customers whose loan status is Current and Charged Off are not included in the statistics. In this paper, customers whose loan status is Fully Paid, In Grace Period and Late(16-30 days) are considered as good credit without default, that is, Y =0, however, customers whose loan status is Default , Charged Off and Late(31-120 days) are considered as poor credit and have Default risk, that is, Y =1. As can be seen, Fully Paid customers account for the largest proportion of 74.13%, and Default customers account for the smallest proportion of 0.037%. Overall, non-default users account for 87.63%, and users with default risk account for 12.37%. For details, see Table 1.

Table 1 Loan status table

Whether there is a default (risk)	State of the loan	count	proportion
NO	Fully Paid	11919	74.13%
	In Grace Period	1463	9.04%
	Late（16-30 days）	717	4.46%
YES	Late（31-120days）	1984	12.34%
	Default	6	0.037%

4.2 Variable distribution statistics

In order to further observe the distribution details of variables, the numerical variables and classified variables are statistically analyzed. The numerical variables count the average value, variance, one quantile, median, third quantile, maximum and minimum value. The distribution of some numerical variables is shown in Table 2.

Table 2 Distribution table of some numerical variables

	mean	std	min	50%	max
loan amnt	15439.85	102339.08	1000	12750	40000
installment	464.08	300.37	30.64	380.66	1664.57
annual	90159.03	112575.84	0	75000	9000000
loan status	0.12	0.33	0	0	1
dti	18.98	18.56	0	17.36	999
open acc	11.87	6.02	1	11	63
pub rec	0.12	0.34	0	0	3
revol bal	16253.68	23787.64	0	10285	652794
total acc	25.20	13.16	2	23	142
tatal bal il	41976.52	49736.36	0	28167.5	703967

max bal bc	5774.17	5695.08	0	4332.5	100146
inq fi	1.44	1.72	0	1	22
total cu tl	1.95	3.21	0	1	54
avg cur bal	16033.93	19072.43	0	9545	338522
pct tl nvr dlq	95.01	8.35	21.4	100	100
total bc limit	28336.81	26598.70	0	20800	313500

It can be seen that the distribution of different numerical variables varies greatly. The variances of annual inc (annual income) and total bal il (current total balance of all installment accounts) are the largest at 112575.84 and 49736.6 respectively. The range of annual inc is the largest, which is $9,000,000, the least annual income is only $0, and the largest is $9,000,000. These outliers are very individual phenomena, which can be removed in the subsequent analysis to exclude the interference of special values.

In this paper, some classification variables are also counted and the word cloud map is drawn by Python. In the word cloud map, the larger the label, the higher the frequency of occurrence. As we can see from Figure 3, debt consolidation (A single loan to pay off all other debts), home improvement and major purchase are very prominent, indicating the large number of people taking out loans for this purpose. At the same time, medical and credit card, etc. also have a high frequency of occurrence.

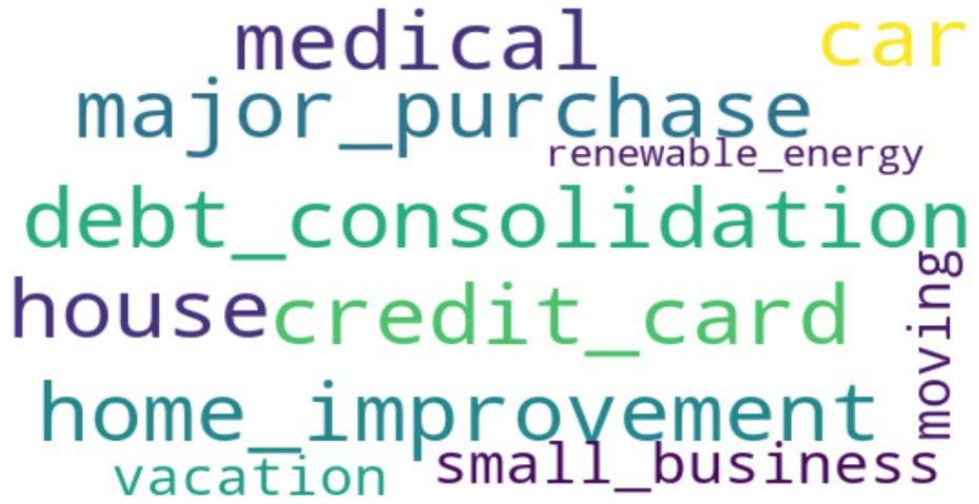

Figure 3 Cloud of loan object words

Table 3 shows the statistical results of the remaining classification variables. In the statistics of working years, the number of people who have worked for 9 years is the least, with 361 people, accounting for 2.49% of the total. The number of people who have worked for 10 years or more is the largest, with 5,008 people, accounting for 34.57% of the total. There are two types of application: Individual and Joint App. There are 12,697 customers who apply independently, accounting for 87.64% of the total number. In the statistics of house ownership, the customers of MORTGAGE account for the majority, followed by RENT and OWN.

Table 3 Statistical table of classification variables in part

Working fixed number of year	classification	<1 year	1 year	2 years	3 years	4 year	5 year
	number of people	1795	1072	1382	1167	922	1049
	proportion	11.91%	7.11%	9.17%	7.74 %	6.12%	6.96 %
	classification	6 year	7 year	8 year	9 year	10+years	
	number of people	657	559	516	361	5008	
	proportion	4.36%	3.71%	3.42%	2.40 %	33.23%	
Application way	classification	Individual	Joint App				
	number of people	12697	1791				
	proportion	84.26%	11.89 %				
The housing situation	classification	MORTGAGE	RENT	OWN	ANY		
	number of people	7729	5011	1592	156		
	proportion	51.29%	33.25 %	10.56 %	1.04 %		

4.3 Analysis of network loan default user portrait

For loan platforms and investors, the most important purpose is to maximize profits. In customers identification, they tend to pay more attention to those customers with default risk. For honest users, the loan platform earns fees and service fees, while for customers who break the contract, the loan platform and investors will cause great losses due to the customer's default behavior. Therefore, after screening the customer data of Loan status =1(that is, there is a default situation), this paper conducts a simple user portrait analysis on the defaulting customers from the perspectives of occupation, working years, annual income, amount of loan, purpose of loan and housing situation.

(I) Occupation and years of work

The word cloud map is made according to the occupational information of defaulting customers, as shown in Figure 4. Since many occupations are not composed of one word, we separate the occupations containing several words to make statistics when making word segmentation, so the words displayed in the word cloud map may be part of the occupations composed of words, but it does not affect us to have a general understanding of the occupations of the defaulting customers. Through

the professional word cloud map of defaulting customers, we can see intuitively that there are a large number of defaulting customers whose professions are Manager, Sales, Driver, etc.

Figure 4 Cloud of professional words of defaulting customers

After the statistics of the working years of the defaulting customers, different colors are given according to different values, as shown in Table 4. The redder the color is, the larger the data is in the column, and the greener the color is, the smaller the data is in the column. From the color changes, it can be seen intuitively that from the perspective of the total number of defaults, the number of default customers who have worked for 10 years or more is the largest, with 503 people, while the number of default customers who have worked for 9 years is the lowest, with 39 people. From the perspective of default rate, the default rate of customers who have worked for less than one year is the highest, accounting for 16.66% of the total number of people who have worked for 10 years or more, while the default rate of customers who have worked for 10 years or more is the lowest, accounting for 10.04%. On the whole, the default rate shows a declining trend with the increase of working years. The reason for this may be that the longer your working years are, the more stable your job is, the better your credit is, and the probability of default is not very high, while the shorter your working years are, the less stable your job is, so the default rate is correspondingly higher. So when identifying customers, you can make a preliminary judgment according to the customer's working years.

Table 4 Statistical table of working years of defaulting customers

emp length	The total	No breach	default	The default rate
<1 year	1795	1496	299	16.66%
1 year	1072	930	142	13.25%
2 years	1382	1215	167	12.08%
3 years	1167	1012	155	13.28%
4 years	922	803	119	12.91%

5 years	1049	917	132	12.58%
6 years	657	591	66	10.05%
7 years	559	494	65	11.63%
8 years	516	462	54	10.47%
9 years	361	322	39	10.80%
10+ years	5008	4505	503	10.04%

(II) Annual income

The annual revenue is divided into 6 segments, the default customers are counted, and the default rate of customers in each segment is calculated. Table 5 lists the specific results. In terms of the total number of defaults, the number of defaults was 567 in the range of annual income between 30,000 and 60,000, while the number of defaults was 96 in the range of annual income less than 30,000. In terms of default rate, customers with annual income less than 30,000 have the highest default rate of 14.41%, while those with annual income more than 150,000 have the lowest default rate of 10.85%. Default rate from the table that a list of the histogram, you can see that with the increase of income, default rate gradually decreases on the whole, so we can think of, the higher the income, the higher the credibility of the customers, the smaller the probability of default, and for earning less loan customers, platform will pay a higher risk of default.

Table 5 Annual income statistics of defaulting customers

Point of division	The total	No breach	default	The default rate
<=30000	666	570	96	14.41%
<=60000	4461	3894	567	12.71%
<=90000	4361	3857	504	11.56%
<=120000	2423	2148	275	11.35%
<=150000	1240	1086	154	12.42%
>150000	1337	1192	145	10.85%
aggregate	14488	12747	1741	12.02%

(III) Borrowing amount

The amount of money borrowed also has an important effect on default rates. According to the dataset, the bar chart of total number of people in different loan amount intervals and the broken line chart of default rate of each interval are drawn. The abscissa is each loan amount interval, and the ordinate is default rate and total number of people respectively. Figure 5 shows the details. From the bar chart, we can see that the number of customers who borrowed $8000-10,000 is the largest, with 1758 people, while the number of customers who borrowed $32000-34,000 is the lowest, with 87 people. From the line chart, it can be seen that the default rate of customers who borrowed $32,000-34,000 is the lowest at 8.05%, while the default rate of those who borrowed $34,000-36,000 is the highest at 18.03%. It can be found that, on the whole, the more the amount of borrowing, the more likely the customer is to default. Therefore, during the daily trading period, special attention should be paid

to customers who borrow large amounts of money. If the amount of money borrowed is inconsistent with their annual income or status, their application should be rejected, or they should be asked to provide more detailed information for assessment.

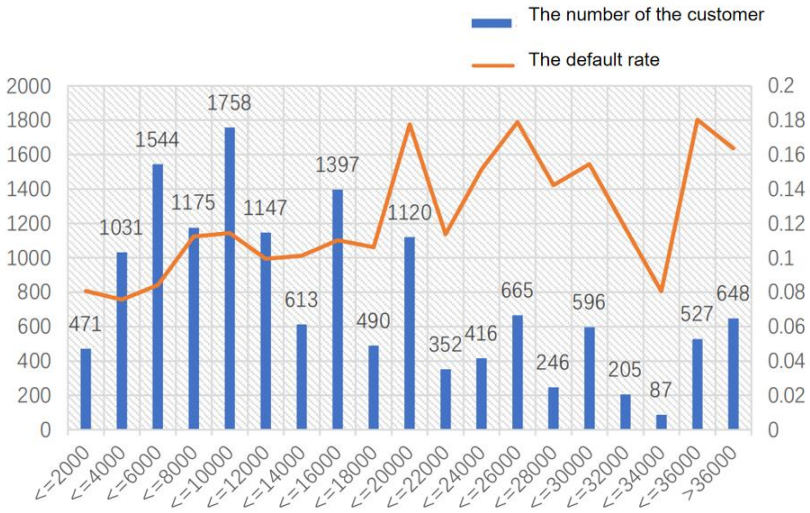

Figure 5 The mixed figure of number of people and default rates in various loan segments

(IV) Purpose of borrowing and housing status

There are mainly 11 borrowing purposes, such as credit card, debt consolidation and home improvement. After statistical sorting, it is found that the default rates of customers borrowing for major purchase and small business are high, 14.47% and 15.28% respectively. The default rates of customers borrowing for car and renewable energy are low, 6.09% and 7.69% respectively. The default rates of customers for other purposes are evenly distributed, about 10%. For details, see Table 6.

Table 6 Default rate statistics for loan purposes

purpose	headcount	No breach	default	The default rate
debt consolidation	7799	6809	990	12.69%
credit card	3096	2741	355	11.47%
home improvement	1122	999	123	10.96%
other	1067	963	104	9.75%
major purchase	387	331	56	14.47%
small business	216	183	33	15.28%
medical	202	180	22	10.89%
moving	198	180	18	9.09%
vacation	146	130	16	10.96%
house	127	111	16	12.60%
car	115	108	7	6.09%
renewable energy	13	12	1	7.69%

From the perspective of housing situation, there are four main housing states of

customers: MORTGAGE, OWN, RENT and ANY. Among them, the total number of MORTGAGE customers is the largest, 7729, and the default rate is also low, 9.85%, which is second only to ANY, 7.69%. The default rate of customers whose houses are leased is the highest, which is 15.29%. For details, see Table 7. This situation is in line with the public perception that the customers of house mortgage loans themselves have good credit, because the bank's loan examination is more stringent, so they can have passed the bank's examination mechanism in the bank loan description. Customers who own houses are not in bad condition, so the default rate of both is relatively low. However, the house is a rental customer, which is unstable and has a high probability of default. The online loan platform can appropriately relax the loan conditions of the customers whose houses are mortgaged, while for the customers whose houses are leased, it should carry out more strict examination, and increase audit variables such as "the number of moving in the last two years" or "the living time in the current house", so as to reduce the default rate.

Table 7 Statistical table of housing default rate

Home ownership	number of people	No breach	default	The default rate
MORTGAGE	7729	6968	761	9.85%
OWN	1592	1390	202	12.69%
RENT	5011	4245	766	15.29%
ANY	156	144	12	7.69%

5.Comparative experiment

Table 8 shows the accuracy rate, recall rate and F1 score of normal and default recognition of various models. SVM classifier and CNN classifier adopt default parameters, and the number of decision tree of RF classifier is set to 200. The specific structure is shown in Table 8.

Table 8 Classification results of different models

Algorithm	Category	Accuracy	Recall	F1-score
RF	Normal	0.925	0.999	0.961
	Default	0.419	0.028	0.053
SVM	Normal	0.925	1.00	0.96
	Default	0.00	0.00	0.00
CNN	Normal	0.94	0.983	0.96
	Default	0.528	0.223	0.311
LighTGBM	Normal	0.945	0.967	0.957
	Default	0.441	0,318	0.359
GCNN-LighTGBM	Normal	0.94	0.95	0.947
	Default	0.315	0.267	0.29
GCNN-LGB-BAG in this paper	Normal	0.978	0.99	0.985
	Default	0.85	0.718	0.777

It can be seen from Table 8 that although the traditional machine learning algorithm

performs well in the recognition of normal repayment on time, it is weak in the recognition of loan default, which is extremely important for the credit default recognition model. In this paper, the ACCURACY rate, recall rate and F1 of THE GCNN-LGB-BAG model for normal and default categories are higher than other models. The comprehensive evaluation of the model classification performance fully proves the superiority of the GCNN-LGB-BAG model proposed in this paper.

Table 9 Comparison of results of different models

Classifier	Accuracy	AUC
RF	0.9235	0.5409
SVM	0.9248	0.5
CNN	0.9269	0.6022
LighTGBM	0.9177	0.6432
GCNN-LighTGBM	0.9046	0.6405
GCNN-LGB-BAG in this paper	0.9705	0.8611

According to the experimental results in Table 9, it can be found that although classification accuracy of SVM, CNN, RF and LighTGBM four traditional credit default recognition methods reaches 91% or more, the AUC value is low, indicating that the classification effect of the model is not ideal. This shows the limitations of traditional machine learning algorithms on the classification performance of high-dimensional unbalanced data sets, that is to say, the learning ability of high-dimensional features is insufficient. Secondly, through further comparison of experimental results, it is found that the accuracy of GCNN-LGB-BAG model is 4.63% higher than that of RF model, and the AUC value is 0.313 higher. Compared with the GCNN-LighTGBM model, its accuracy is improved by 6.52%, and its AUC value is improved by 0.2134. It is proved that the combination of convolutional neural network and stochastic forest algorithm has a good classification effect on high-dimensional unbalanced financial transaction data and the feasibility and advantages of autonomous feature learning using GCNN-LighTGBM. In short, the GCNN-LGB-BAG model proposed in this paper has an incomparable outstanding performance in solving the existing problems of credit default recognition.

6.Conclusion

The method proposed in this paper can be widely applied to the credit risk assessment of borrowers including all kinds of P2P platforms, which can help alleviate the platform crisis caused by the lack of credit investigation system and poor risk control, and promote the steady development of Internet finance. Based on the method proposed by us, P2P lending platforms and investors can identify defaulting customers with the help of their own algorithms. P2P lending platforms are more suitable for using better performance integration algorithms, which can effectively reduce the risk of loss caused by customer default, maximize profits and safeguard the interests of investors.

7.References

[1] POPE D G ， SYDNOR J R .What’s in a picture ？ Evidence of discrimination from prosper [J].Journal of Human Resources ， 2011 ， 46(1) ： 53-92.

[2] RAVINA E .Love & loans ： the effect of beauty and personal characteristics in credit markets [J]. SSRN Electronic Journal ， 2012. DOI ： 10 .2139/ssrn .1107307 .

[3] Lai Hui, Shuai Li, ZHOU Zheng Fang. A new method of credit Evaluation for personal credit customers [J]. Technical and economic ， 2014 ， 33（9） ： 97-103.

[4] Herzenstein, Michal, Sonenshein, Scott, Dholakia, Utpal M. Tell Me a Good Story and I May Lend You Money: The Role of Narratives in Peer-to-Peer Lending Decisions[J]. Journal of Marketing Research, 2011, 48(SPL):S138.

[5] Michels J . Do Unverifiable Disclosures Matter? Evidence from Peer-to-Peer Lending[J]. Accounting Review, 2012, 87(4):1385-1413

[6] Stein J C. Information Production and Capital Allocation: Decentralized versus Hierarchical Firms[J]. Journal of Finance, 2002, 57(5):1891-1921.

[7] 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[M]// Judging Borrowers By The Company They Keep 1 : Social Networks and Adverse Selection in Online Peer-to-Peer Lending. 2013.

[8] Pope D G, Sydnor J R. What's in a Picture? Evidence of Discrimination from Prosper.com[J]. Social Science Electronic Publishing, 2011, 46(1):53-92.

[9] Emekter R, Tu Y, Jirasakuldech B, et al. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending[J]. Applied Economics, 2015, 47(1):54-70.

[10] PURO, Lauri, TEICH, et al. Borrower Decision Aid for people-to-people lending[J]. Decision Support Systems, 2010, 49(1):52-60.

[11] Burtch G, Ghose A, Wattal S. Cultural Differences and Geography as Determinants of Online Pro-Social Lending[J]. Mis Quarterly, 2014, 38(3): 773-794.

[12] Zhang J, Liu P. Rational Herding in Microloan Markets[J]. Management Science, 2012, 58(5):892-912.

[13] Berkovich E. Search and herding effects in peer-to-peer lending: evidence from prosper.com[J]. Annals of Finance, 2011, 7(3):389-405.

[14] Lee E, Lee B. Herding behavior in online P2P lending: An empirical investigation[J]. Electronic Commerce Research & Applications, 2012, 11(5):495-503.

[15] NANNI L,LU MINI A .An experimental comparison of ensemble classifiers for bankruptcy prediction and credit scoring [J] .Expert Systems with Applications, 2009, 36(2): 3028-3033.

[16] ABELL Á N J,CASTELLANO J G . A comparative study on base classifiers in ensemble method for credit scoring [J]. Expert Systems with Applications,2016, 73: 1-10.

[17] FLOREZ-LOPEZ R,RAM ON-JERONIM O J M . Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment: a correlated-adjusted decision forest proposal [J].Expert Systems with Applications, 2015,42(13):5737-5753.
[18] TSAI C F, HSU Y F,YEN D C .A comparative study of classifier ensembles for bankruptcy prediction [J].Applied Soft Computing, 2014,24: 977-984.
[19] Cao Wei, Li Can, He Tingting , et al. Comparative study on credit risk early-warning models of P2P online lending in China based on integrated learning [J]. Data analysis and knowledge discovery, 2018,2(10):65-76.
[20] HOANG D T， KANG H J.A survey on deep learning based bearing fault diagnosis［J］.Neurocomputing，2019， 335：327 －335.
[21] HINTON G E，SALAKHUTDINOV R R.Reducing the dimensionality of data with neural networks ［J］.Science，2006，313（5786）：504 －507.
[22] Wang Lihua, Xie Yangyang, Zhang Yonghong, et al. Fault Diagnosis Method for Asynchronous Motor Based on Deep Learning [J]. Journal of Xi 'an Jiaotong University.2017， 5 1（10）：128 －134.
[23] Zhou Qicai, Shen Hehong, Zhao Jiong, et al. Bearing Fault Diagnosis Based on Improved Stacked Cyclic Neural Network [J]. Journal of Tongji University (Natural Science edition)，2019，47（10）：1500 －1507.
[24] Yang Liusong, HE Guangyu. SVM Fault Diagnosis Method Based on Improved Particle Swarm Optimization [J]. Computer Engineering,2013，39（3）：187 －190，196.
[25] Hou Pingzhi, Zhang Ming, Xu Xiaobin, et al. Fault Diagnosis Method based on K-nearest Neighbor Evidence Fusion [J]. Control and Decision，2017，32（10）：1767 － 1774.
[26] Yan Renwu, Ye Xiaozhou, Zhou Li. Fault Diagnosis technology of Power Electronic Circuit based on Random forest [J]. Journal of Wuhan University (Engineering Science), 2013，46（6）：742 －746.
[27] Jiang Shaofei, Wu Tianji, Peng Xiang, et al. Data Driven Fault Diagnosis Method based on XGBoost Feature Extraction [J]. China Mechanical Engineering,2020， 31（10）：1232 －1239.
[28] KE G L，MENG Q，FINLEY T，et al. LightGBM：a highly efficient gradient boosting decision tree［C］// Advances in Neural Information Processing Systems 30，New York：Curran Associates，2017.
[29] LIN M，CHEN Q，YAN S C.Network in network[J]. Computer Science，2013，arXiv：1312.4400.
[30] ZHANG W，PENG G L，LI C H，et al.A new deep learning model for fault diagnosis with good anti －noise and domain adaptation ability on raw vibration signals[J]. Sensors，2017，17（2）：425.
[31] BREIM AN L .Arcing classifiers [J].The Annals of Statistics ，1998 ，26（3）：801-824.
[32] FREUND Y ， SCHAPIRE R E . A decision-theoretic generalization of on-line learning and an application to boosting [J]. Journal of Computer and System Sciences ， 1997 ，55（1） ：119-139.

[33] KE G ， M ENG Q ， Finley T ， et al .Lightgbm ： a highly efficient gradient boosting decision tree [C]//31st Conference on Advances in Neural Information Processing Systems .Long Beach ， CA ： Neural Information Processing Systems Foundation ， 2017 ： 3146-3154.
[34] Li Hang. Statistical Learning Methods [M]. Beijing: Tsinghua University Press ， 2012 ： 18-20.