Research Article

Bo Gao* and Vipin Balyan

Construction of a financial default risk prediction model based on the LightGBM algorithm

https://doi.org/10.1515/jisys-2022-0036
received October 18, 2021; accepted January 17, 2022

Abstract: The construction of a financial risk prediction model has become the need of the hour due to long-term and short-term violations in the financial market. To reduce the default risk of peer-to-peer (P2P) companies and promote the healthy and sustainable development of the P2P industry, this article uses a model based on the LightGBM (Light Gradient Boosting Machine) algorithm to analyze a large number of sample data from Renrendai, which is a representative platform of the P2P industry. This article explores the base LightGBM model along with the integration of linear blending to build an optimal default risk identification model. The proposed approach is applicable for a large number of multi-dimensional data samples. The results show that the prediction accuracy rate of the LightGBM algorithm model on the test set reaches 80.25%, which can accurately identify more than 80% of users, and the model has the best prediction performance in terms of different performance evaluation indicators. The integration of LightGBM and the linear blending approach yield a precision value of 91.36%, a recall of 75.90%, and an accuracy of 84.36%. The established LightGBM algorithm can efficiently identify the default of the loan business on the P2P platform compared to the traditional machine learning models, such as logistic regression and support vector machine. For a large number of multi-dimensional data samples, the LightGBM algorithm can effectively judge the default risk of users on P2P platforms.

Keywords: peer-to-peer industry, LightGBM algorithm, default prediction model, P2P network lending, logistic regression, support vector machine

1 Introduction

The development in network technology and Internet-based financial applications has become the backbone of national and international economies in various countries. These small- and medium-sized organizations create the two-thirds of entire employment in the world, creating a strong contribution to economic profitability [1,2]. The advents in the construction of social credit standards improve the credit scoring system by upgrading the rapid advancement in the entire credit system. The relevant progress in science and technology has led to the development of the entire society by improving the objectivity and timeliness of personal credits. However, the advances in the fields of e-commerce and Internet financing have resulted in the evaluation of personal credits meeting the requirements of changing time as per the big data scenario of the current economy [3–6].

* **Corresponding author: Bo Gao,** School of Management Engineering, Henan University of Engineering, Zhengzhou, Henan 451191, China, e-mail: bogao231@outlook.com
Vipin Balyan: Department of Electrical, Electronics and Computer Engineering, Cape Peninsula University of Technology, Cape Town, South Africa, e-mail: balyanv@cput.ac.za

Peer-to-peer (P2P) lending has become an important aspect of the financial channel replacing the tedious traditional loaning system from the financial institutions, making it convenient to the investors and borrowers [7–9]. P2P online lending refers to the formation of a loan contract on the Internet platform by the supply and demand side of funds in an unsecured form. Different from traditional financial loans, P2P online loan is a personal-to-person unsecured small loan [10,11]. The basic process of P2P lending is depicted in Figure 1.

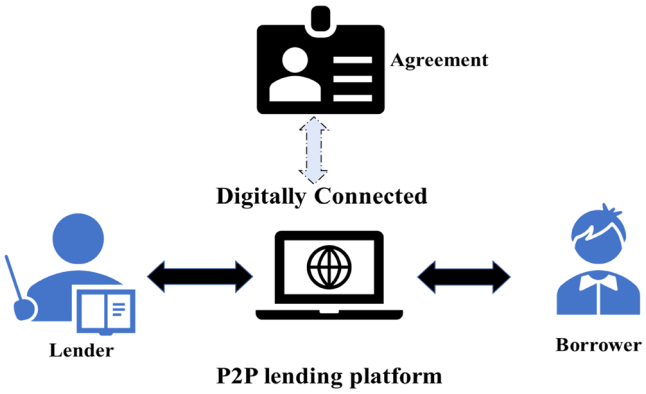

Figure 1: Basic P2P lending mechanism.

Individuals who have spare money on the Internet lend money to individuals who need money based on trust because trust on the Internet is very difficult [12,13]. The things which gave birth to the online loan intermediary platform is two side transaction through online platform enabling dependencies. From the initial barbaric growth, speculators and lawbreakers were attracted to take advantage of the loopholes in control. In recent years, there have been a large number of examples of capital chain breaking and platform managers running away [14]. According to the data of Wangdaizhijia, there are only 458 platforms currently operating normally, which is less than one-tenth of the peak period. The platform problems that run away are mainly concentrated in two aspects. On the one hand, the lack of strict supervision has caused many people who want to make quick money to flood into the market and eventually develop into a "Ponzi scheme"; on the other hand, there are problems with the risk management of the platform itself [15]. The high rate of bad debts leads to the breakdown of funds. Therefore, it is very important to strengthen the risk control of online loans.

There is a contradiction between risk control and expansion in the development of P2P online lending platforms. If the loan review mechanism is too strict, the number of borrowers will be drastically reduced, which limits the scale of the platform's operation, while the loose loan review mechanism will reduce the number of borrowers. But the increase in loan scale will also bring more non-performing loans. Therefore, a more accurate model is needed to help the platform accurately identify the probability of future default by the borrower and determine whether to issue a loan to the borrower based on the probability of the model output and determine the amount and interest rate if the loan is issued to the borrower. Online loan default prediction is a management technique for online loan risk control. The existence of a large number of real data on online loans provides the possibility for the realization of online loan default predictions.

This article contributes to the construction of a financial risk prediction model for default risk reduction in P2P companies, to promote the healthy and sustainable development of the P2P industry. The main contribution can be highlighted as follows:

- This research work presents a LightGBM (Light Gradient Boosting Machine) algorithm to analyze a large number of sample data from a representative (Renrendai) platform of the P2P industry.
- The novelty of this approach lies in the exploration of the base LightGBM model along with the integration of linear blending to build an optimal default risk identification model. The integrated LightGBM

algorithm can efficiently identify the default of the loan business on the P2P platform utilizing a large number of multi-dimensional data samples.
- The proposed LightGBM algorithm can effectually judge the default risk of users on P2P platforms compared to the traditional machine learning models, such as logistic regression and support vector machine (SVM).
- The development of big data and machine learning technology provides feasibility for the improvement of online loan default prediction technology.

Therefore, this article intends to improve the platform's loan default prediction ability and reduce bad debts through the construction of a financial default risk prediction model based on the LightGBM algorithm.

This research article is organized as follows. Literature review is presented in Section 2. The explanation of the research methods like the discussion of P2P loan business analysis, a theoretical model for data mining, and evaluation indicators are given in Section 3. Section 4 presents the result analysis and discussion. The concluding remarks are given in Section 5.

2 Literature review

The birth and growth of credit risks occur in commercial banks. Loans contain risks, and commercial banks' loan approvals are becoming more and more standardized. This naturally leads to small- and medium-sized enterprises and individuals who have difficulty obtaining loans when they need money [16]. Therefore, P2P network loans have also emerged. However, it is not to say that P2P network loans do not contain risks. In fact, P2P network loans have higher risks due to their wide liquidity, high returns, and no collateral. This means that the platform needs more intelligent risk control systems. Countless scholars are also exploring how to build a better risk control system based on P2P network loan data.

The fields of big data and Internet finance have developed tremendously in the twenty-first century. The state's attention to this field has gradually increased. P2P is an innovative lending model that is a powerful supplement to the traditional financial industry. The estimated credit default rate is an absolute prerequisite to ensure the normal operation of relevant financial projects or platforms. Ma et al. have utilized the "multi-observation" and "multi-dimensional" data cleaning methods and applied the modern machine learning algorithm LightGBM to Asia at the end of 2016 and applied XGboost to Lending Club's real P2P transaction data. The strong and innovative loan prediction of default risks is presented by authors [3]. The results obtained from various studies are compared and it is observed through the LightGBM algorithm based on the classification and prediction results of multiple observation data sets is the best. The average performance rate of historical transaction data on the Lending Club platform increased by 1.28% points, which reduced loan defaults by approximately $117 million. Finally, regarding the influencing factors of the default rate, it provides suggestions for the development of Lending Club and other P2P platforms, as well as suggestions for the development of other countries in this field [17].

Li et al. studied a data-driven method to extract the knowledge of default risk from the borrower's demographic information and behavioral characteristics in the loan process, which can be used to reduce the default risk of P2P platforms. The possibility of automation of credit risk ratings can also be studied by estimating the accuracy of forecasts. One huge dataset is analyzed from well-known P2P lending platform in China, and used three default prediction models to conduct data research on discrete input and output pairs, continuous input and output pairs, and continuous input and discrete output pairs. The average hit rate and lift analysis are used to evaluate the accuracy of the prediction. The 2-layer artificial neural network model performed well in continuous input and output data pairs with an average relative error of 0.24. An SVM is highly recognized because of its 89.18% prediction accuracy of discrete input and output data pairs. Decision tree C5.0 (DT) is used to discover some important factors affecting the risk rate and

predict the default risk of borrowers. Based on the results of data mining, some constructive conclusions can be drawn about P2P online loan risk management [18].

Su et al. studied a new LightGBM method combined with a random forest algorithm, which was used to predict the global underground temperature and salinity anomalies within 1,000 m depth based on remote sensing data and Argo floating-point data. These methods combine longitude and use multi-source sea surface parameters (sea surface height anomaly [SSHA], sea surface temperature anomaly [SSTA], sea surface salinity anomaly [SSSA], north and east components of sea surface wind anomaly [USSWA, VSSWA]) Latitude data (LON, LAT) are used as predictors, and Argo grid data are used as training and testing labels for model construction and prediction. This research established a five-parameter model (SSTA, SSHA, SSSA, USSWA, VSSWA), a six-parameter model (LAT, SSTA, SSHA, SSSA, USSWA, VSSWA), a six-parameter model of latitude and longitude (LON, SSTA), SSHA, SSSA, USSWA, VSSWA) and seven-parameter models with longitude and latitude (LON, LAT, SSTA, SSHA, SSSA, USSWA, VSSWA) to analyze and evaluate the role of LON + LAT in the prediction of STA and SSA LightGBM and Random Forest (RF) models [19].

There are various methods reported in the literature for the classification of credit scoring utilizing various statistical methods [20,21], non-parametric techniques [22–24], and neural network techniques [25]. The development in the optimization theory and machine learning approaches [26,27] has led to the improvement of credit assessment-based research. The researchers have also explored the concept of deep learning [28,29], genetic algorithm [30], and ensemble models [31,32] for the improvement of the P2P lending platform. Some of the drawbacks of these methodologies lie in their complex learning for massive data processing. This article uses the LightGBM algorithm for removing the drawback of the current research on this prospect.

The innovation of this article is to use the LightGBM algorithm to establish a model that can efficiently identify the default of the loan business on the P2P platform. LightGBM algorithm is better than traditional machine learning models such as logistic regression and SVM for a large number of multi-dimensional data samples. This article utilizes the LightGBM algorithm as a base model and uses linear blending for model integration to build an optimal default risk identification model. It can effectively judge the default risk of users of P2P platforms and reduce the misestimation of users who have not defaulted.

3 Research methods

3.1 P2P loan business analysis

The procedures and rules of each P2P platform borrowing business are roughly the same. First of all, those in need of funds use the official website of the P2P platform or mobile phone software to apply for an account, and those in need of funds must provide relevant information for review, including ID cards, basic information, work status, and credit records. P2P companies review the data of the fund demand after receiving the data of the fund-demand person and use a manual or model to assess the credit level of the fund-demand person and then determine the loan or not and the loan amount. After that, to effectively withdraw funds, P2P companies will monitor the use of funds by fund demanders, urge them to repay the funds, and adopt corresponding strategies according to the performance of the contract.

Usually, to evaluate the credit level of the fund demander, P2P companies will consider the user's various information. For example, under the premise of getting the user's permission, some mature P2P companies will collect and analyze the lender's electronic payment usage, social media usage records and mobile phone contacts, and other related information. This information and data help P2P companies to investigate and evaluate the loan application of the lender.

After the funds are released, since the previous data review can only reduce the risk of bad debts to a certain extent, the P2P platform needs to focus on monitoring the use of funds by those who need funds. The

information reported by many loan customers is not necessarily true and reliable, and the information of the lender may not remain unchanged. Therefore, the P2P platform must collect the relevant information of the fund demanders as comprehensively as possible, optimize the default risk identification model, and use the model to identify lenders with high default probability, strictly screen this type of group, and conduct strict monitoring. Once a default is found, effective remedial measures should be implemented.

3.2 Theoretical model of data mining

The current common machine learning models include the logistic regression model, SVM, decision tree algorithm, random forest model, XGBoost model, and LightGBM algorithm. This article mainly studies the prediction model based on the LightGBM algorithm. LightGBM is a decision tree-based machine learning algorithm published by Microsoft in 2016, suitable for processing sorting, classification, regression, and other problems. LightGBM is a kind of GBDT, to effectively deal with the latter's shortcomings in the face of large amounts of data, such as low operating efficiency and long time-consuming [33].

The decision tree base model in LightGBM is split according to the way of leaf splitting, so its calculation cost is very small, but the depth of the control tree must be controlled with the minimum sample size on the leaf nodes to prevent the occurrence of overfitting. LightGBM divides the eigenvalue into many small "buckets" and then looks for splits on the small "buckets" to effectively reduce storage and computing costs. LightGBM mainly completes the control and optimization of the model through the following parameters, as shown in Table 1.

Table 1: Significance of the main parameters of LightGBM

Parameter	Parameter meaning
num_leaves	Number of leaves per tree
learning_rate	Rate of learning
n_estimators	Maximum allowed iterations
max_depth	Maximum learning depth
min_date	Minimum data used to assist in control overfitting for leaf nodes
bagging_fraction	Random and non-repeated selection of observations whose value is between 0 and 1 to improve model training speed and control overfitting
feature_fraction	The proportion of the number of selected features to the total number of features, the default is 1

The advantages of the LightGBM algorithm are mainly reflected in two aspects: on the one hand, LightGBM has faster training speed, lower memory consumption, and support for parallel learning when processing massive data, which solve the difficulties of GBDT when facing massive data; on the other hand, while improving efficiency, the LightGBM model has higher model accuracy and the prediction effect is very ideal. Therefore, applying the LightGBM algorithm with an ideal prediction effect to the identification of P2P default risk, its convenience, and effectiveness will greatly promote the long-term development of this field [34,35].

3.3 Balance the data set and evaluation indicators

3.3.1 Balance the data set

In the classic hypothesis of machine learning, it is generally assumed that the number of different types of samples of the target variable in the training set is balanced, and the machine learning model obtained by

training with a data set with a balanced number of samples can usually achieve better prediction results. However, this assumption is often not satisfied in practical applications. When the number of samples in each category of the training set has a large gap, the generated model often cannot effectively classify the samples. The scarce categories in the data set are often easily overlooked during model training, resulting in the model not being able to extract the information of this type of sample, thus affecting the model's prediction results for this type of sample. In the actual P2P online loan business, the number of defaulting users and non-defaulting users is often very different, and the proportion of defaulting users is at a low level. In summary, this article must balance the training samples to make the target variables as balanced as possible.

The method used in this article to deal with unbalanced samples is to divide the positive samples into small sets. The number of positive samples and the number of negative samples in the small sets are roughly the same, and then, the positive samples and negative samples in each small set are used to build models. Finally, the model results based on the small set are calculated by the weighted average method to obtain the prediction results of an overall algorithm model. Figure 2 shows the idea of how this article deals with samples.

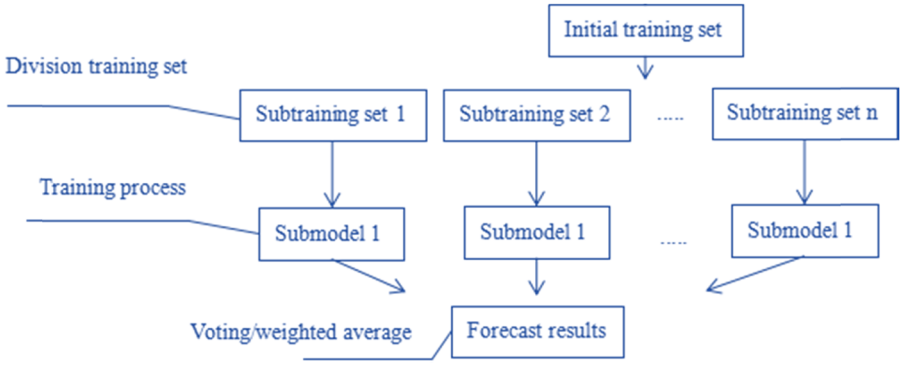

Figure 2: Balanced sample method.

Compared with the up-sampling method, the idea of this method has certain advantages: the method used in this article improves the model effect by optimizing the integration of the model, while the up-sampling method improves the model effect from the data the aspects of the set are optimized. Taking logistic regression as an example, the method used in this article to deal with unbalanced samples has an average prediction accuracy rate of 64.62% and an average time-consuming of 2 min 15 s. The model results of up-sampling balanced samples have a prediction accuracy rate of 64.87% and the average time-consuming of 3 min 16 s. The model prediction accuracy of the two methods is almost the same, but in terms of computational efficiency, the balanced sample method used in this article has increased by 37.78%.

There are 59,166 samples in the original data set of this article, of which only 1,695 are default samples, accounting for only 2.86%. Therefore, this article divides the non-default samples into 44 groups, each with 1,305 samples. At the same time, 1,305 records of the default records are selected as the training set, and the remaining samples are used as the test set for model construction.

3.3.2 Description of evaluation indicators

3.3.2.1 Confusion matrix and accuracy

The training models are generally evaluated using a confusion matrix, which is also called an error matrix. Each column of the matrix represents the predicted category of the sample by the classifier, and each row of

the matrix represents the true category of the sample. The form of the confusion matrix is as follows (Table 2).

Table 2: Confusion matrix

		Forecast category	
		0	**1**
Real category	0	TN	FP
	1	FN	TP

Among them, TP represents the number of true positives, FP represents the number of false positives, TN represents the number of true negatives, and FN represents the number of false negatives. According to the confusion matrix, we can calculate many indicators to evaluate the prediction effect of the classifier. Among them, accuracy is a more common evaluation index, which represents the proportion of samples accurately divided by the classifier. Statisticians can use this index to have a general understanding of the accuracy of the classifier's prediction. The calculation formula is provided in the following equation:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \tag{1}$$

3.3.2.2 Precision, recall, and Fl statistical value

In many cases, statisticians pay more attention to the prediction results of positive samples by the classifier. For the default risk studied in this article, P2P companies pay more attention to the identification of default users. At this time, the two evaluation indicators of recall and precision play a key role. Recall refers to the proportion of positive samples predicted by the classifier to the true positive samples, and precision represents the proportion of real positive samples predicted by the classifier. This article gives the calculation formulas of these two indicators as the following equations:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{3}$$

Recall and accuracy are indicators used to evaluate the predictive effect of classifiers on positive samples. However, there is a certain conflict between the two and it is difficult to improve at the same time. Therefore, we use the recall and precision to calculate the $F1$ statistical value, which takes into account both the recall and precision of the classifier prediction. The calculation formula of the $F1$ statistical value is presented in the following equation:

$$F1 \text{ statistical value} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$

4 Result analysis and discussion

The results are discussed in terms of prediction outcomes for error rate and loss value for the training and testing sets and the later discussion involves the outcomes obtained for various evaluation indicators to assess the performance of the presented LightGBM model.

4.1 Prediction results of the LightGBM model

In the face of massive data scenarios, the LightGBM algorithm has a smaller training cost, which can save a lot of time, and the prediction effect of the model is also better. This article sets the LightGBM algorithm model parameters as follows for the sample data: num_leaves = 25, learning rate = 0.1, n_estimators = 100, feature_fraction = 1, bagging_fraction = 0.5, max_depth = 8, and min_data = 50. The results of the model operation are shown in Figures 3 and 4.

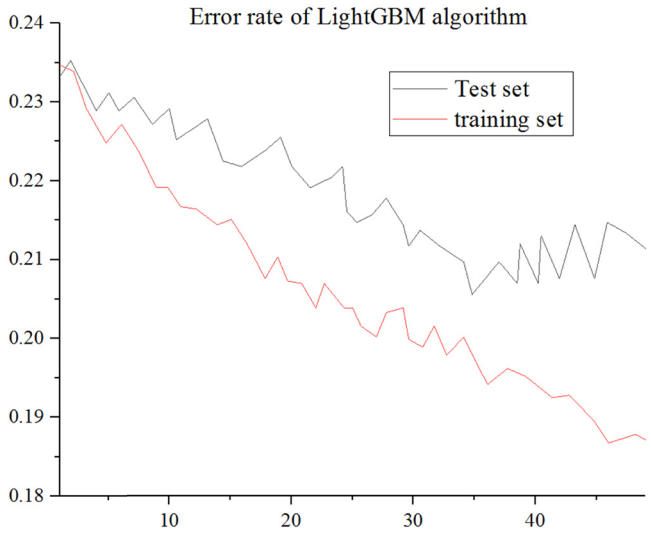

Figure 3: Error rate of LightGBM algorithm.

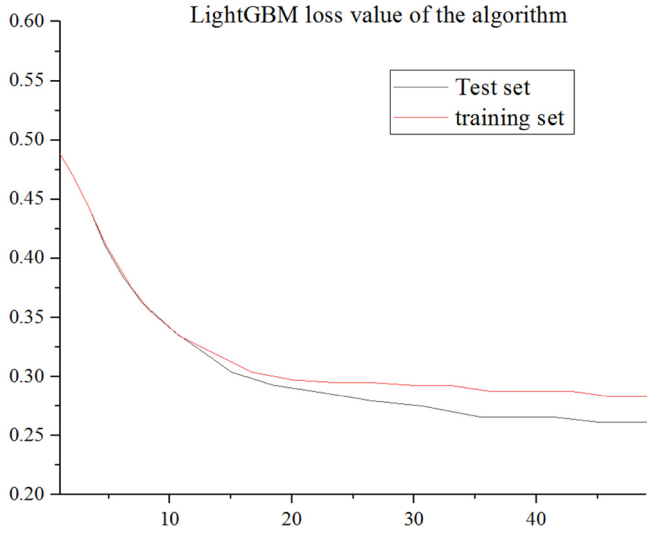

Figure 4: Loss value of LightGBM algorithm.

It can be seen from Figure 3 that when the number of iterations is 36, the error rate of the test set and the training set is synchronously small. Among them, the error rate of the sample test set is about 19.744%, and prediction accuracy rate of the model on the test set reaches 80.256% . This results accurate identification of more than 80% users which the model predicts at best effect. From Figure 4, it is seen that the error and loss

value of the test set and training set have been declining simultaneously, so there is no overfitting, and the prediction results are reliable and effective.

In addition to using statistical indicators, such as accuracy and recall to evaluate models, statisticians usually use receiver operating characteristic (ROC) curves and area under the ROC curve (AUC) values to evaluate models. ROC curve is also called the ROC curve. Each point on it reflects the susceptibility to the same signal stimulus. Two concepts of true class rate (also known as sensitivity) and negative–positive class rate are proposed here. The true class rate refers to the proportion of samples correctly predicted by the classifier in all positive sample instances, which is consistent with the definition of recall. The negative–positive rate refers to the proportion of all negative samples that are predicted by the model to be positive but negative. After calculating the model's predicted true class rate and negative and positive class rate indicators under different threshold settings, the AUC is shown in Figure 5.

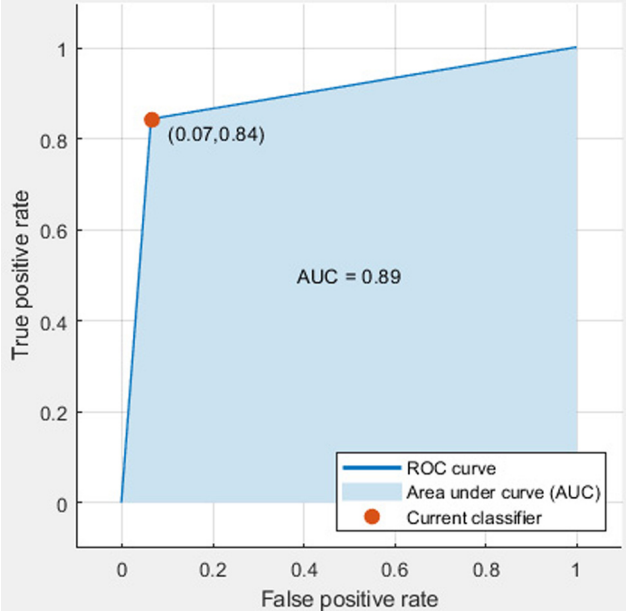

Figure 5: ROC curves for LightGBM algorithm.

This article focuses on the risk of P2P net loan default. On the one hand, we want to increase the real class rate, that is, to filter out all default users to the maximum extent; on the other hand, we want to reduce the negative–positive class rate, that is, to minimize the number of non-default users. So, if the shape of the ROC curve can be closer to the upper left corner of the graph, then the prediction effect of the model is better. Therefore, this article introduces the AUC value here to accurately evaluate the ROC curve. The AUC value refers to the AUC of the model (between 0 and 1). The closer the indicator is to 1, the closer the ROC curve is to the graph and the better is the prediction effect of model refer to upper left corner. Model provides the AUC value of 0.89 which is closer to 1, indicating its competence and feasibility.

4.2 Model integration based on linear blending

The model ensemble is one of the important means to improve the prediction accuracy of the classifier. Compared with a single classifier, the combined prediction effect of the model using ensemble strategy is better, and it is easier to achieve ideal results. However, different algorithms models presents their strengths, and have certain differences, and model integration can bring out the advantages of each model

so that these relatively weak models can be combined in some way to generate a more powerful model achieves the best fitting effect and creates value for major P2P companies.

The integration method used in this article is mainly linear blending. Since this article is studying the two-classification problem, the logistic model is chosen as the secondary learner of linear blending. In the performance of a single model, the LightGBM algorithm model performs the best. Therefore, this article chooses to use the LightGBM method to build a primary learner and then linearly blend it through linear blending. To highlight the differences between the sub-models, this article randomly selects half of the independent variables into the model when building the LightGBM model, so that the input characteristics of each sub-model are different. In this article, a total of ten LightGBM sub-models are established for training, and the parameter settings of the sub-models are kept as the parameter settings of a single model.

After linear blending, the prediction accuracy in the test set is as high as 84.36%, which is higher than the prediction results of all other single models, as shown in Table 3.

Table 3: Predictive effects of several models

Model type	Accuracy (%)	Precision (%)	Recall (%)	*F*1 statistical value
Logistic	64.62	77.14	41.54	0.540
Random Forest	74.36	85.19	58.97	0.697
XGBoost	76.92	88.89	61.54	0.727
LightGBM	80.26	87.82	70.26	0.781
LightGBM + linear blending	84.36	91.36	75.90	0.829

Table 3 depicts the values for different evaluation indicators using various model types ranging from the logistic, random forest, XGBoost, LightGBM, and the integrated linear blending model. The clear depiction of accuracy and *F*1 statistical value are provided in Figure 6, and precision and recall values obtained for different models are shown in Figure 7.

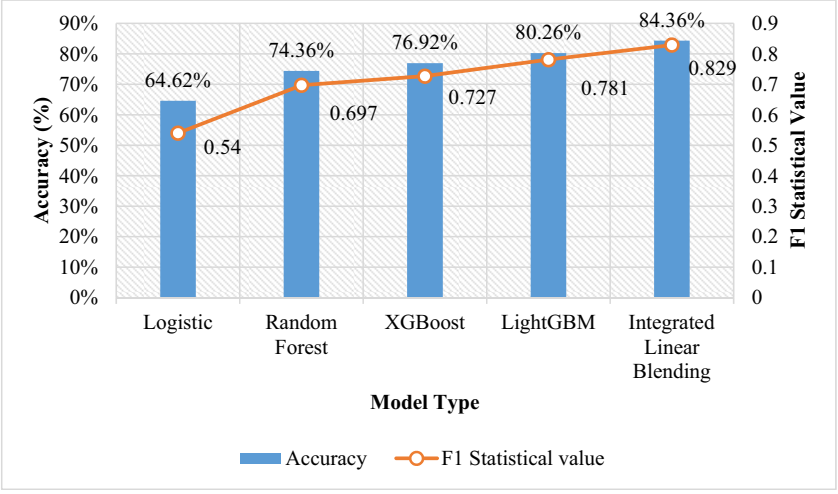

Figure 6: Comparative analysis of accuracy and *F*1 statistical value.

The best outcomes are obtained for the integrated linear blending technique using the base LightGBM model. This approach yields a precision value of 91.36%, a recall of 75.90%, and an accuracy of 84.36%. The integrated model can identify most default users and minimize the omission of default users. The

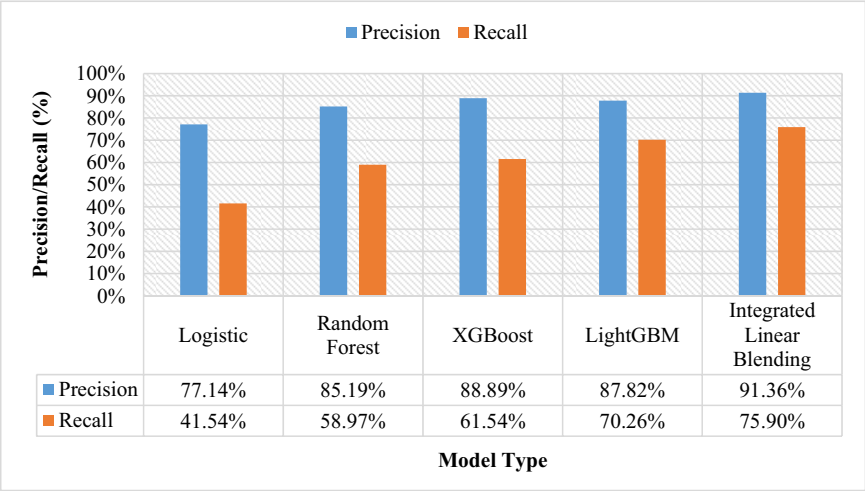

Figure 7: Comparative analysis of precision and recall.

comparison of the model integrated by linear blending with the previous single classifiers reveals that the advantages of the linear blending integrated model are very significant. In addition, the Fl value of the LightGBM model is 0.781, which when combined with the linear blending phenomenon provides a 0.829 *F*1 value. This performance is good, which is the highest value among the built models, and the model has fast running efficiency and short running time, which can save a lot of time and cost.

5 Conclusion and future scope

This article provides a representative platform for the P2P industry by analyzing a large number of sample data of Renrendai utilizing data processing and data modeling methods. It provides a reference method for solving the problem of excessive loan default risk on the current P2P platform and suggests ways for minimizing the risk of default. The following conclusions can be drawn from the article:

(1) Machine learning models, such as LightGBM, have excellent performance in predicting multi-dimensional data. The model studied in this article has a high dimensionality and a large amount of data. Machine learning algorithms provide us with new ideas for solving problems compared to the outdated traditional forecasting methods. This article builds models for more mature machine learning models, such as the SVM model and logistic regression model, as well as current cutting-edge machine learning models, such as the LightGBM model. The results show that the prediction effect of integration of the LightGBM model with linear blending yields better results than that of basic SVM and logistic regression [35]. The precision value of 91.36%, the recall of 75.90%, the accuracy of 84.36%, and the *F*1 statistical value of 0.829 are witnessed using the integrated platform.

(2) The integration method represented by linear blending can effectively improve the predictive ability of the model. From the performance of the integrated model, the predictive ability of the model is significantly improved compared to a single model, which can identify default users to the greatest extent and reduce the misjudgment of non-default users. The integrated model established in this article has certain feasibility and practical significance and has a certain reference effect for the model building of the P2P platform.

Acknowledgments: This research was supported by (1) National Natural Science Foundation of China in 2017, Research on mechanism and simulation and of multi-agent collaborative governance of ecological environment in the water source area of the middle route of South-to-North Water Diversion Project, Project

Number U1704124, Principal Investigator Jingbao Fu; (2) Humanities and Social Sciences project of Ministry of Education of China, Research on coupling coordination mechanism and dynamic regulation of water ecological environment and economy in the water source area of middle route project of South-to-North Water Diversion, Project Number 19YJC630075, Principal Investigator Hongyan Li; and (3) Science and technology innovation team of the Education Department of Henan Province, Public resources and environmental governance, Project Number 19IRTSHN015, Principal Investigator Hongyan Li.

Conflict of interest: Authors declare no conflict of interests for this article.

References

[1] Altman EI, Sabato G, Wilson N. The value of non-financial information in SME risk management. Available at SSRN. 2008;1320612.
[2] Shin GH, Kolari JW. Do some lenders have information advantages? Evidence from Japanese credit market data. J Bank Financ. 2004;28(10):2331–51.
[3] Ma X, Sha J, Wang D, Yu Y, Yang Q, Niu X. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. Electron Commer Res Appl. 2018;31:24–39.
[4] Dhiman G, Kumar VV, Kaur A, Sharma A. DON: Deep Learning and Optimization-Based framework for detection of novel coronavirus disease using X-ray Images. Interdiscip Sci: Comput Life Sci. 2021;13:1–13.
[5] Cornée S. The relevance of soft information for predicting small business credit default: Evidence from a social bank. J Small Bus Manag. 2019;57(3):699–719.
[6] Yuvaraj N, Srihari K, Dhiman G, Somasundaram K, Sharma A, Rajeskannan S, et al. Nature-inspired-based approach for automated cyberbullying classification on multimedia social networking. Math Probl Eng. 2021;2021:2021–12.
[7] Bastani K, Asgari E, Namavari H. Wide and deep learning for peer-to-peer lending. Expert Syst Appl. 2019;134:209–24.
[8] Poongodi M, Hamdi M, Malviya M, Sharma A, Dhiman G, Vimal S. Diagnosis and combating COVID-19 using wearable Oura smart ring with deep learning methods. Personal Ubiquitous Comput. 2021;1–11.
[9] Babaev D, Savchenko M, Tuzhilin A, Umerenkov D. Et-rnn: Applying deep learning to credit loan applications. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2019, July. p. 2183–90.
[10] Kannan S, Dhiman G, Natarajan Y, Sharma A, Mohanty SN, Soni M, et al. ubiquitous vehicular ad-hoc network computing using deep neural network with IOT-based bat agents for traffic management. Electronics. 2021;10(7):785.
[11] Wang C, Han D, Liu Q, Luo S. A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism LSTM. IEEE Access. 2018;7:2161–8.
[12] Niu K, Zhang Z, Liu Y, Li R. Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending. Inf Sci. 2020;536:120–34.
[13] Zhang FP, Huang YP, Luo WX, Deng WY, Liu CQ, Xu LB, et al. Construction of a risk score prognosis model based on hepatocellular carcinoma microenvironment. World J Gastroenterol. 2020;26(2):134–53.
[14] Li Z, Xu H, Xue Y, Pei B. Construction method of flight safety manipulation space based on risk prediction. J Beijing Univ Aerona Astron. 2018;44(9):1839.
[15] Odediran SJ, Windapo AO. Risk-based entry decision into African construction markets: A proposed integrated model. Built Environ Proj Asset Manag. 2018;8:91–111.
[16] Plebankiewicz E, Wieczorek D. Adaptation of a cost overrun risk prediction model to the type of construction facility. Symmetry. 2020;12(10):1739.
[17] Sun X, Liu M, Sima Z. A novel cryptocurrency price trend forecasting model based on LightGBM. Financ Res Lett. 2020;32:101084.
[18] Li XF, Zhang C, Lin XC, Lv TJ, Liu LL. Research on default risk of peer-to-peer online lending based on data mining algorithm. J Computers. 2020;31(2):83–100.
[19] Su H, Lu X, Chen Z, Zhang H, Lu W, Wu W. Estimating coastal chlorophyll-a concentration from Time-Series OLCI data based on machine learning. Remote Sens. 2021;13(4):576.
[20] Altman EI, Sabato G. Modeling credit risk for SMEs: Evidence from the US market. Managing Measuring Risk: Emerg Glob StRegul Fina Crisis. 2013;251–79.
[21] Sohn SY, Kim DH, Yoon JH. Technology credit scoring model with fuzzy logistic regression. Appl Soft Comput. 2016;43:150–8.

[22] Xia Y, Liu C, Li Y, Liu N. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Syst Appl. 2017;78:225–41.
[23] Hsieh NC, Hung LP. A data driven ensemble classifier for credit scoring analysis. Expert Syst Appl. 2010;37(1):534–45.
[24] Zhao Z, Xu S, Kang BH, Kabir MMJ, Liu Y, Wasinger R. Investigation and improvement of multi-layer perceptron neural networks for credit scoring. Expert Syst Appl. 2015;42(7):3508–16.
[25] Ma L, Huo X, Zhao X, Zong GD. Observer-based adaptive neural tracking control for output-constrained switched MIMO nonstrict-feedback nonlinear systems with unknown dead zone. Nonlinear Dyn. 2020;99(2):1019–36.
[26] Deng C, Che WW, Shi P. Cooperative fault-tolerant output regulation for multiagent systems by distributed learning control approach. IEEE Trans Neural Netw Learn Syst. 2019;31(11):4831–41.
[27] Kozeny V. Genetic algorithms for credit scoring: Alternative fitness function performance comparison. Expert Syst Appl. 2015;42(6):2998–3004.
[28] Maldonado S, Pérez J, Bravo C. Cost-based feature selection for support vector machines: An application in credit scoring. Eur J Operational Res. 2017;261(2):656–65.
[29] Finlay S. Multiple classifier architectures and their application to credit risk assessment. Eur J Operational Res. 2011;210(2):368–78.
[30] Wang G, Hao J, Ma J, Jiang H. A comparative assessment of ensemble learning for credit scoring. Expert Syst Appl. 2011;38(1):223–30.
[31] Xia Y, Liu C, Da B, Xie F. A novel heterogeneous ensemble credit scoring model based on bstacking approach. Expert Syst Appl. 2018;93:182–99.
[32] Qiu X, Zuo Y, Liu G. ETCF: An ensemble model for CTR prediction. In 2018 15th International Conference on Service Systems and Service Management (ICSSSM). IEEE; 2018, July. p. 1–5.
[33] Jiang S. Construction of risk prediction model for Alzheimer's disease based on meta-analysis. Open Access Library J. 2019;6(9):1.
[34] Chen X, Metawa N. Enterprise financial management information system based on cloud computing in big data environment. J Intell & Fuzzy Syst (Prepr). 2020;5:1–10.
[35] Guo LW, Li N, Chen HD, Lyu ZY, Feng XS, Wei LP, et al. Progress in construction and verification of colorectal cancer risk prediction models: a systematic review. Zhonghua Yu Fang Yi Xue Za Zhi [Chin J Preven Med]. 2019;53(6):603–10.