Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 199 (2022) 757–764

Procedia Computer Science

www.elsevier.com/locate/procedia

The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021)

Prediction of loan default based on multi-model fusion

Xingyun Li[a],Daji Ergu[b]*,Di Zhang[b],Dafeng Qiu[b],Ying Cai[b],Bo Ma[b]

[a]College of Business School, Southwest Minzu university, Chengdu 610041,China
[b]key Laboratory of Electronic and Information Engineering, State Ethnic Affairs Commission, Southwest Minzu University,Chengdu 610041,China. * Corresponding author. ergudaji@163.com

Abstract

With the development of Internet technology, online loans continue to enter the public eye, individuals and small businesses must access to more loan opportunities, and it is important for online loan platforms to effectively reduce the credit crisis associated with customer loan defaults. This paper uses the loan default dataset from lending club. The ADASYN (Adaptive synthetic sampling approach) method is adopted to cope with the class imbalance problem of the dataset. In order to improve the prediction accuracy, this paper utilizes the Blending method to fuse three models: Logistic Regression, Random Forest, and CatBoost. After experimental comparison, it is found that the performance of the fusion model proposed in this paper is better than the three models of Logistic Regression, Random Forest, and CatBoost, which can effectively predict the probability of customer loan default through the training of the dataset and reduce the external risk brought by the online loan platform facing customer loan default.

Keywords: Credit crisis, Loan default, Multi-model fusion;

1. Introduction

Along with the prosperous development of Internet technology, the Internet finance industry with Internet technology as the core has emerged. P2P network lending (peer-to-peer lending) is a new type of network lending model developed in this context, using the Internet as a medium to carry out online transactions, acting as a transaction intermediary and information intermediary, to achieve a direct match between investors and lenders. Moreover, the P2P platform can better meet the lending needs of small and medium-sized enterprises and low-income groups that cannot be covered by traditional financial institutions [1], which makes the model develop widely and rapidly.

Investors on the P2P platform invest in loans based on the borrower's personal credit at a certain interest rate, and once they breach the contract, the loan principal will incur a large loss [2] and the credit crisis of the borrower is the main external risk faced by the P2Pplatform, once the borrower defaults, the P2P platform will face a huge credit crisis, for investors, P2P loans are completely completed online, which is a kind of direct financing.

10.1016/j.procs.2022.01.094

Although P2P platforms have a scientific and reasonable credit rating mechanism that can reduce the degree of information asymmetry between borrowers and investors [3], investors can only independently predict the future default probability and expected return of loans [4] to prevent investment risk. Therefore, we must realize that the core of P2P network lending development lies in risk management and reducing default rates is the main aspect of risk management.

Therefore, how to use data mining information timely and effectively, use intelligent methods to identify credit risk, improve the review efficiency of loan users and default prediction accuracy rate, is of great significance to reduce the credit risk of P2P platform and the healthy and stable development of the platform.

2. Literature Review

At present, domestic, and international research on loan default prediction on P2P network lending platforms focuses on using different models to process loan default data. So as to continuously improve the effectiveness of data processing and loan default prediction.

With the advancement of machine learning and artificial intelligence techniques, classification and regression models are also used to predict credit crisis [5]. Galindo compared the performance of decision trees, neural networks, and k-nearest neighbor algorithms in credit default prediction and the results showed that the decision tree algorithm had the best classification results [6]. Malekipirbazari et al. conducted an empirical study in the Lending Club dataset and showed that the random forest model outperformed FICO scores and Lending Club's own credit rating methodology in predicting borrower defaults [7] Sigrist, Fabio et al. obtained the Grabit model by applying the gradient boosting decision tree algorithm to the Tobit model and experimentally found that the model predicts loan default better than other advanced methods [8]

Continuous technological advances have given rise to more research ideas. Some scholars have improved loan default prediction by fusing different models [9, 10, 11], including GBDT, decision tree, XGBoost, and other models. Ma, Xiao Jun et al. conducted a comparative experiment using LightGBM and Boost algorithms on a real transaction dataset of Lending Club and found that the LightGBM algorithm gave the best classification prediction results [12]. In addition, some scholars introduced deep learning models in loan default prediction tasks to improve the results [13] Liang, Long Yue et al. used LSTM neural network for loan default prediction and the experimental results showed that this method has the highest loan default prediction accuracy [14]

Yum pointed out that information asymmetry is the most fundamental problem faced by online lending, as there is no contact between lenders and borrowers, investors don’t have access to sufficient information about the borrower to assess the credit crisis of the borrower [15] Kim and Cho consider an ensemble semi-supervised learning method taking into account both labeled data and unlabeled data [16]. Wei, Xinyuan et al. weighted L1/2 regularization to a logistic regression model to reduce the impact of information asymmetry, which can help investors reduce the impact of information asymmetry in lending decisions [17]

Through literature review, we found that the fusion model based on Logistic Regression, Random Forest, and CatBoost proposed in this paper can provide new methods and ideas for loan default prediction research.

3. Related models and methods

3.1. Related Models Introduction

3.1.1. Logistic Regression

Logistic Regression is a generalized linear regression model.

It achieves the classification task by associating the true marker y of the classification task with the predicted value of the linear regression model (Eq. 3.2) through the sigmoid function (Eq. 3.1)

$$y = \frac{1}{1 + e^{-(w^T x + b)}} \qquad (3.1)$$

$$\ln \frac{y}{1 - y} = w^T x + b \qquad (3.2)$$

where the weight vector w = (w_1; w_2...w_d) and the bias b are parameters that can be learned. If we consider y in Eq. 3.1 as posterior probabilistic to estimate p(y=1 | x), we get:

$$w^T x + b = \ln \frac{P(y = 1 \mid x)}{1 - P(y = 1 \mid x)} \qquad (3.3)$$

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(w^T x + b)}} \qquad (3.4)$$

That is, the log-likelihood of the output y=1 is a model represented by a linear function of the input x, this is the logistic regression model. The logistic regression model estimates the parameters of the model by the great likelihood method, and then maximizes the log-likelihood, which gives the objective function of the model as follows:

$$J(W, b) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log h\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h\left(x^{(i)}\right)\right) \right]$$

3.1.2. Random Forest

Random forest is a classifier that uses multiple decision trees to train and predict samples.

In the process of classification, for each node of the base decision tree, a subset containing k attributes is first randomly selected from the set of attributes of that node, and then an optimal attribute is selected from this subset for division, where the parameter k controls the degree of randomness, generally $k = log_2 d$ [18].

In a random forest, each decision tree is disjoint, and the final classification result is determined by the plurality of the results obtained from each decision tree.

Where the attributes of each decision tree split node are determined by the Gini coefficient, which is formulated as follow:

$$G(X_i) := \sum_{j=1}^{J} \Pr\left(X_i = L_j\right)\left(1 - \Pr\left(X_i = L_j\right)\right) = 1 - \sum_{j=1}^{J} \Pr\left(X_i = L_j\right)^2$$

For a general decision tree, if a total of samples are to be divided into J categories, where Pr $\left(X_i = L_j\right)$denotes the probability that sample X_i belongs to the L_j category. The smaller the Gini coefficient is, the more thorough the data partitioning is.

3.1.3. CatBoost model

CatBoost [19] is one of the open source boosting algorithms in 2017. CatBoost is a GBDT framework with few parameters, support for categorical variables, and high accuracy based on oblivious trees algorithm.

In order to use all samples for training and to better handle the category features CatBoost relies on the ranking principle and uses a more efficient strategy. First, all samples are randomly sorted, and then for a certain value taken in the category-based feature, each sample with that feature converted to a numerical value is averaged based on the category label ranked before that sample, with the addition of a weighting factor for priority and precedence. The formula example is as follows:

$$\frac{\sum_{j=1}^{p-1}\left[x_{\sigma_j,k}=x_{\sigma_p,k}\right]Y_{\sigma_j}+a\cdot P}{\sum_{j=1}^{p-1}\left[x_{\sigma_j,k}=x_{\sigma_p,k}\right]+a^k}$$

3.2. Multi-model fusion

3.2.1. Blending Method

Blending is a model fusion method. Blending fuses models by training a new learner. The implementation principle is as follows：

The original data set is divided into the original training set and the original test set, and the training set needs to be divided again into a new training set and a new test set.

Use the new training set to train the model in the first layer, and then make predictions on the new test set, keeping the predictions as the training set features for the second layer model.

The first layer model is used to predict the original test set, and the prediction results are used as features of the second layer model test set, and the results predicted by the second layer model are the final results. The Blending method is illustrated in the following diagram.

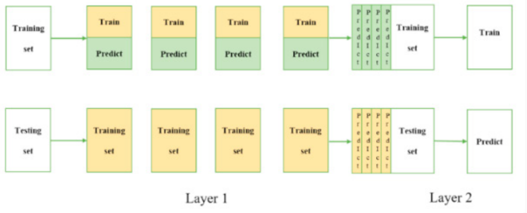

Fig. 3.1. Blending schematic

3.2.2. Multi-Model Fusion Framework

In this paper, logistic regression, random forest, and CatBoost, are used as the models in the first layer of the Blending method. The model fusion process is as follows:

Firstly, the results of the two predictions of the first layer model are output as the training set features and test set features of the second layer model, respectively.

Then GBDT is chosen as the model for the second layer, which is trained and predicted again.

Finally, the second layer model outputs the final prediction results to complete the model fusion. The multi-model fusion framework diagram in this paper is shown in the following figure:

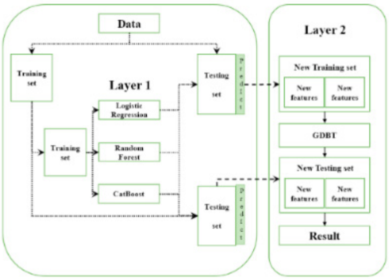

Fig. 3.2. Multi-model fusion framework

4. Experiment

4.1. Dates Processing

The dataset used in this paper is derived from real loan data from Lending Club for Q4 2019 publicly available on the Kaggle website. The original dataset contains 128,262 loan data with 150 columns of attributes. The data processing in this paper includes the following steps:

- Data cleaning and pre-processing. We removed the attributes with a proportion of missing values greater than 40% and which do not contribute much to the prediction effect of the model. After that, the remaining missing values were filled in by means of mode interpolation.
- Feature engineering. In this paper, we construct a new feature 'installment_feat' based on two features of the original dataset, installment (the amount of monthly installment of the loan) and annual_inc (annual income). A larger value of 'installment_feat' means the greater the pressure on the lender to repay the debt and the greater the probability of default.

$$installme_{feat} = \frac{12 \times installment}{annual_{inc}} \qquad (4.1)$$

- Feature Selecting. The first dimensionality reduction is achieved by filtering 30 features with the strongest correlation to the target variable by Recursive Feature Elimination (RFE) method and gradually eliminating the features. Next, the redundant features are identified by bootstrapping the Pearson correlation coefficient. 19 features were finally filtered out, as shown in Table 1.
- Finally, the random forest algorithm is adopted to determine the feature importance to mine which variables are more important. From Figure 4.1, we can see that the importance of the new features (installation_feat) ranks fifth, which proves that the new features constructed in this paper have contributed to improving the prediction effect of the model.

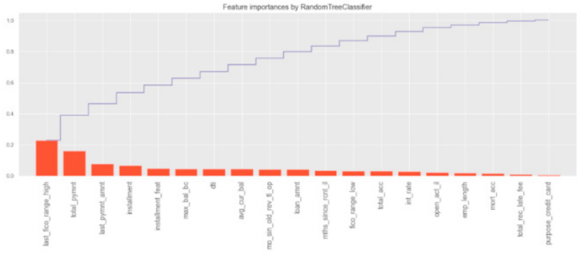

Fig. 4.1. The ranking of features importance

Table 1. Features Description

Features	Description
last_fico_range_high	The highest fico score of the borrower in the last repayment period
total_pymnt	All the money paid by the borrower.
Last_pymnt_amnt	Total payment amount received last time
installment	The monthly amount owed by the borrower。
installment_feat	Feature of structure
max_bal_bc	Max balance of bankcard accounts
dti	Debt ratio
avg_cur_bal	Current average balance of all accounts
mo_sin_old_rev_tl_op	The recent month to open a credit card account
loan_amnt	Amount of loan applied
mths_since_rcnt_il	Number of months to open an installment account
fico_range_low	Fico lowest rating
total_acc	Total amount of borrower's credit line
int_rate	Loan interest rate (high interest rate and high risk)
open_act_il	The number of open credit lines in active installment
emp_length	Years of work
mort_acc	Number of collateral accounts
total_rec_late_fee	Late fees received to date
purpose_credit_card	A category provided by the borrower for the loan request.

4.2. ADASYN Algorithm

An important feature of the dataset in this paper is that there is a serious class imbalance in the dataset. Samples of the dataset with normal markers is 99.34%, but samples with default markers is only 2.66%.

In this paper, the ADASYN (Adaptive Synthetic Sampling) algorithm [20] is adopted, which aims to solve the problem of performance degradation due to unbalanced data. It can assign different weights to different minority classes of samples, thus generating different sample sizes. The algorithm principle is as follows:

For a minority class sample, first, find its K nearest neighbors and calculate the percentage of the number of samples to be synthesized for this minority class sample relative to the total amount of synthesis.

$$r_i = \Delta_i / K$$

Normalize r_i as weights and calculate the number of new samples that need to be generated for each minority class samples g.

$$g = \hat{r}_i \times G$$

G is the total number of samples to be synthesized.

When each minority class sample starts to generate a new sample, it needs to loop from 1 to g. Generate the synthetic data example:

$$s_i = x_i + (x_{zi} - x_i) \times \lambda$$

s_i is the synthesized sample, x_i is the i th sample in the minority class sample, and x_{zi} is a minority class sample randomly selected among the K nearest neighbors of x_i. $\lambda \in [0,1]$ random number

The most important feature of ADASYN is that it can automatically decide how many synthetic samples need to be generated for each minority class sample. Therefore, the ADASYN algorithm can effectively avoid the phenomenon of sample mixing in the data, which leads to poor classification results.

4.3. Results and discussion

In order to verify the validity of the fusion model proposed in this paper, accuracy, recall, F1-score, and roc curve are adopted as the validation metrics of the model performance. Logistic Regression, Random Forest and CatBoost, are used as benchmark models to compare with the fusion model proposed in this paper.

From the ROC curves of the four models shown in Figures 4.2, it can be shown that the ROC curve of the fusion model is closest to the upper left corner and has the largest area under the curve, indicating the best performance.

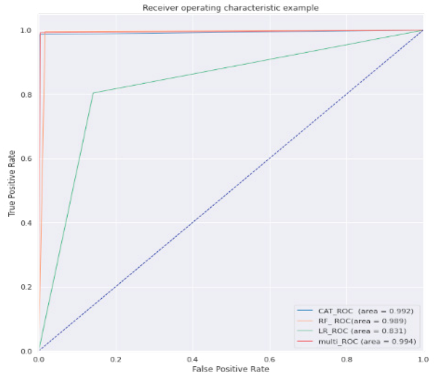

Fig. 4.2. ROC performance comparison of the four models

The comparison results of each metric of the four models can be seen in Table 2. CatBoost has the best performance in the comparison of the three single model metrics. And the fusion model has a higher Accuracy compared to CatBoost model. In general, we can conclude that the performance of the fusion model proposed in this paper is better than the other three models.

Table 2. Evaluation metrics comparison of the four models

Rank	model	Accuracy (%)	Recall	F1-score
1	Fusion model	99.4	0.99	0.99
2	CatBoost	99.2	0.99	0.99
3	Random Forest	98.9	0.98	0.99
4	Logistic Regression	83.1	0.83	0.83

5. Conclusion

In this paper, we propose a fusion model based on Logistic Regression, Random Forest and CatBoost for loan default prediction. Theoretically, it provides new model fusion ideas and technical routes for loan default prediction research. Focusing on and dealing with the problem of data set class imbalance, this article makes the model have better risk screening and pre-warning, which provides new ideas for risk pre-warning in the field of P2P network lending in China and provides a reference for post-loan credit crisis monitoring in China. In terms of practical application, the prediction results of the fusion model are compared with other three single models. It is found that the classification accuracy and performance of the fusion model proposed in this paper are better than the other three models, and it can effectively achieve loan default prediction and reduce the external credit crisis caused by customer default faced by online lending enterprises.

Acknowledgements

This research has been partially supported by grants from the National Natural Science Foundation of China (Nos. 71774134, U1811462). And supported by the Fundamental Research Funds for the Central Universities, Southwest Minzu University (Grant Number 2020NGD04,2020PTJS27004, and 2018NZD02), and the Innovation Scientific Research Program for Graduate student in Southwest Minzu University (No. CX2020SZ69).

References

[1] Liao L, Zhang W Q. An empirical study on peer-to-peer lending: A literature review. Journal of Tsinghua University (Philosophy and Social Sciences Edition), 2017, 32(2): 186–196. (in Chinese)

[2] Wang, H., Chen, K., Zhu, W., Song, Z., 2015. A process model on P2P lending. Financial Innovation 1.. doi:10.1186/s40854-015-0002-9.

[3] Klaff,M. Peer to eer Lending: Auctioning Micro credits over the Internet[R]. Proceedings of the 2008 International Conference on Information System, Technology and Management (ICISTM 08), 2008.

[4] M. Leo, S. Sharma, and K. Maddulety, "Machine learning in banking risk management: A literature review," Risks, vol. 7, p. 29, 2019.

[5] Wang P, Zheng H , Chen D , et al. Exploring the critical factors influencing online lending intentions[J]. Financial Innovation, 2015, 1(1):1-11.

[6] Galindo J , Tamayo P . Credit Risk Assessment Using Statistical and Machine Learning[J]. Computational Economics, 2000.

[7] Malekipirbazari M , Aksakalli V . Risk assessment in social lending via random forests[J]. Expert Systems with Applications, 2015, 42(10):4621-4631.

[8] Sigrist, F., Hirnschall, C., 2019. Grabit: Gradient tree-boosted Tobit models for default prediction. Journal of Banking & Finance 102, 177–192. doi: 10.1016/j.jbankfin.2019.03.004.

[9] Li Y, Chen W. Entropy method of constructing a combined model for improving loan default prediction: A case study in China[J]. Journal of the Operational Research Society, 2019(4):1-11.

[10] Zhou, J., Li, W., Wang, J., Ding, S., Xia, C., 2019. Default prediction in P2P lending from high-dimensional data based on machine learning. Physica A Statistical Mechanics and its Applications 534, 122370. doi: 10.1016/j.physa.2019.122370.

[11] Zurada J. Data mining techniques in predicting default rates on customer loans[M]// Databases and Information Systems II. Springer Netherlands, 2002.

[12] Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., Niu, X., 2018. Study on a prediction of P2P network loan default based on the machine learning LightGBMand XGboost algorithms according to different high dimensional data cleaning. Electronic Commerce Research and Applications 31, 24–39.. doi:10.1016/j.elerap.2018.08.002.

[13] Li, W., Ding, S., Wang, H., Chen, Y., Yang, S., 2020. Heterogeneous ensemble learning with feature engineering for default prediction in peer-to-peer lending in China. World Wide Web 23, 23–45.. doi:10.1007/s11280-019-00676-y.

[14] Liang, L., Cai, X., 2020. Forecasting peer-to-peer platform default rate with LSTM neural network. Electronic Commerce Research and Applications 43, 100997.. doi:10.1016/j.elerap.2020.100997.

[15] Yum H , Lee B , Chae M . From the wisdom of crowds to my own judgment in microfinance through online peer-to-peer lending platform[J]. Electronic Commerce Research and Applications, 2012.

[16] Kim, A.; Cho, S.B. An ensemble semi-supervised learning method for predicting defaults in social lending. Eng. Appl. Artif. Intell. 2019, 81, 193–199.

[17] Wei, X., Yu, B., Liu, Y., 2020. Accessing Information Asymmetry in Peer-to-Peer Lending by Default Prediction from Investors' Perspective. Symmetry 12, 935.. doi:10.3390/sym12060935.

[18] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[19] Prokhorenkova L , Gusev G , Vorobev A , et al. CatBoost: unbiased boosting with categorical features[J]. 2017.

[20] He H , Bai Y , Garcia E A , et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]// Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on. IEEE, 2008.