A Loan risk assessment model with consumption features for online finance

Li Li¹, Ding Huang²

¹,² School of Software Engineering, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China

*Corresponding author’s e-mail: s181231024@stu.cqupt.edu.cn

Abstract. While online finance is rapidly growing, it is a big challenge to evaluate the loan risk of users based on the data on the Internet. In this study, we used basic user information and added the user's consumption features to construct a loan risk assessment model that integrates features. We extract consumption features from two aspects: firstly, construct consumption portrait features through statistical analysis and clustering; secondly, combine convolutional neural networks (CNN) and bidirectional long short-term memory (BiLSTM) networks to extract sequence features in consumption, and add attention mechanisms to improve the evaluation effect. Finally, the features are combined and fed into the fully connected layer, and the probability of default loan is calculated through the activation function. Our data set comes from an online finance company. The experimental results show that the loan assessment model combined with consumption features does improve the accuracy of default loans under the AUC and KS indicators.

1. Introduction

In recent years, the development of the Internet industry has also led to the rapid development of online finance. Online loan is also one of the products of online finance. It is different from the credit of traditional banking finance, it can’t communicate face to face, so it will cause asymmetric information when lending, which is also one of the main sources of risk of online loan. With the development of big data, to reduce the asymmetry of information between loan platforms and borrowers. These online finance company begins to integrate external data to evaluate the loan risk of users on the premise of protecting users' privacy, which is also one of the main trends of this study. For example, adding ‘soft Factors’ lenders’ loan descriptions to the loan risk assessment, and extracting text features using natural language processing to improve loan assessment results [1],[2] Added the user's credit card consumption information, and evaluated the user's loan risk based on the user's credit card consumption. [3] Using the behavior sequence of borrowers in the loan platform website for risk assessment, the experiment shows that adding the user's behavior sequence can improve the final evaluation effect.

Loan risk assessment is a kind of financial risk. In recent years, there have been more and more researches on loan risk assessment. The commonly used loan assessment methods include expert scoring, statistical learning method, machine learning method and deep learning method. [4] Analytic Hierarchy Process (AHP) is used in mortgage loan evaluation to build evaluation criteria and provide decision-makers with a more transparent mortgage risk evaluation system. In addition, there are also improvements in AHP's improved methods used in loan risk assessment[5]. Among the methods that use machine learning, methods commonly used for risk assessment are logistic regression(LR), support vector machines (SVM), decision tree( DT), XGBoost, LightGBM, etc. The XGBoost method is applied...
to imbalanced credit risk assessment, and combine the Kmeans method to deal with the imbalance of data to improve the model effect in [6]. The effects of XGBoost, LightGBM, and CatBoost models are compared on the Home Credit Dataset in [7]. As for the deep learning method, due to the introduction of external data, CNN and LSTM models are gradually being used in loan risk assessment [8-11]. However, there are few studies on the influence of the consumption information on the user's loan risk. The consumption information reflects the user's economic level and thus reflects the user's loan risk.

In this study, we discussed the role of user consumption data in loan risk assessment. More specifically, we studied the impact of users' historical order information on consumer platforms on loan risk. Previously, there have been studies on the impact of user bank card consumption information on loan assessment [2]. We added the user’s historical consumption orders on the consumption platform to the loan risk assessment to reduce information asymmetry and make full use of user’s historical consumption data. User's historical consumption information often reveals a lot of valuable information. The stability of historical consumption and the interval of consumption time can sometimes reflect the user's economic situation, so as to analyze according to the user's economic situation and provide a reference for loan risk assessment. Therefore, this study constructed a loan risk assessment model that integrates consumption data.

The main contributions of our study are as follows:
(1) This study adds the user's historical order information on the consumer platform into the loan risk assessment dimension
(2) We construct consumption features from two aspects, consumption portrait features and consumption sequence characteristics.
(3) We experimented on real data sets, and the proposed model integrating consumption data improves the evaluation effect of loan risk.

The structure of this paper is as follows. The second section introduces the techniques used in the proposed method, include CNN, LSTM and attention mechanism. The third section introduces the proposed method. The forth section presents the experimental setting and results from analysis. The last section summarizes our study.

2. Theory and Model

2.1 CNN
CNN is inspired by biological natural visual cognition mechanism [12], and is frequently used in image recognition, classification and other fields. Its advantage is that it can accurately extract the local correlation of data features, thereby improving the accuracy of feature extraction. It is an excellent deep learning classification algorithm.

As a deep learning model, CNN can achieve supervised learning through multi-layer backpropagation. Its structure usually includes the Input layer, Convolutional Layer, Activation layer, Pooling layer, Fully Connected Network, and Output layer. Among them, the input layer can be one-dimensional to multi-dimensional data, and correspondingly, one-dimensional to multi-dimensional convolution can be used. One-dimensional and two-dimensional convolutions are currently the most widely used. One-dimensional convolution is usually used to process time series data and text data. Two-dimensional convolution is often used in the field of graphics and image computer vision.

In recent years, research has gradually introduced the CNN method into the field of credit loan risk assessment, and by using CNN to extract features, improve the effectiveness of risk identification [13].

2.2 LSTM
LSTM is one of the variants of Recurrent Neural Network (RNN) model [14], RNN is a commonly used deep learning model. It introduces state variables to store past information, and it determines the current output together with the current input. It is often used to process sequence data, such as a piece of text or sound, or even a row or column of pixels in an image. Therefore, cyclic neural networks have
extremely wide practical applications, such as text classification, machine translation, speech recognition, etc.

RNN can make better use of the information that the traditional neural network structure cannot model, but at the same time, it also brings greater technical challenges—long-term dependence. LSTM has a special network structure with three “gate” structures[15]. LSTM relies on the "gate" structure to allow information to selectively affect the state of the recurrent neural network at every moment. There are three gate structures in the LSTM structure: input gate, forget gate and output gate. In order to make the recurrent neural network more effective in preserving long and short-term memory, "forget gate" and "input gate" are essential, they are the core of the LSTM structure. The function of the "forgotten gate" is to make the recurrent neural network "forget" previously useless information.

2.3 Attention Mechanism
The attention mechanism is essential to mimic the way humans observe objects[16]. By scanning the global image quickly, human vision can obtain the target area that needs to be focused on, which is commonly known as the focus of attention, and then invest more attention resources in this area to obtain more detailed information of the target that needs to be focused on, while suppressing other useless information. The attention mechanism in deep learning is essentially similar to the human selective visual attention mechanism. The core goal is to select the information that is more critical to the current task goal from a large number of information[17].

The attention mechanism machine consists of two parts: the attention mechanism needs to determine which part of the entire input needs more attention; extract features from the key parts to obtain important information. In [18] the application of attention in various scenarios is introduced. However, in the current risk assessment, the attention mechanism is rarely applied to the loan risk assessment.

3. Proposed model
This study proposes to extract consumption features from consumption data and build a loan risk assessment model that integrates consumption features. The historical consumption data of users on the consumption platform can reflect the economic level of users over a period of time. For example, users continue to purchase high-priced goods on the consumer platform, which may reflect that the user’s consumption level is high and the economic situation is good, so the user’s loan risk may be lower; on the contrary, if the user spends less time on the consumer platform or the consumption level is unstable, the user's loan risk may be higher.

Therefore, we propose to add the dimension of user consumption data in the loan risk assessment. On the basis of using basic user information (age, income, occupation, etc.), add user consumption data and the feature of consumption are extracted from two aspects, one is to construct consumption portrait features through statistical analysis of consumption information, and the other is to extract sequence features in consumption data through deep learning methods. The structure of the proposed model is shown in Figure 1 below. We take loan risk assessment as a binary classification in this study.

![Figure 1. Model Structure](image-url)
3.1. Basic user information features
The basic information of users includes demographic information such as age, marital status, occupation, etc. The numerical information is normalized and the category variables is encoded to a low dimensional vector through the corresponding embedding layer. Then the features are extracted through the fully connected layer, and finally input into the final activation function through feature connection.

3.2. Features of consumption portrait
In user consumption data, we construct user FCP(Features of Consumption Portrait) by analyzing user consumption information. The characteristics of user consumption portrait are constructed from four aspects, which are user consumption level: by constructing the user's cumulative consumption amount, the consumption amount in the past month and three months; consumption activity: the cumulative consumption times, the past month and three months number of consumption; consumption stability: average consumption amount of users, average monthly consumption amount, maximum consumption, etc.; potential problem users: analyze the number of cancellations, the number of clip coupons, and the analysis of the user's monthly consumption and monthly income relationship.

Then clustering is performed according to the features extracted in each dimension, for example, the user consumption level is clustered into high, medium, and low feature labels according to the features in the user consumption level, and the remaining dimensions are clustered. Then the consumer portrait is fully connected to extract the feature, and finally, the feature is stitched into the final activation function.

3.3. Consumption sequence feature
However, if we only use consumption information to construct the consumption portrait features, analyze the largest, smallest, average, and other similar characteristics, lacking the dynamic characteristics related to the time series between consumption, and analyzing the sequence relationship between consumption may give a better understanding user consumption, better analyze the user’s economic situation, and then reflect the user's loan risk.

First, we construct a consumption vector. Each consumption includes fields such as consumption time, consumption amount, and payment method and so on. Table 1 shows the consumption order data. We use the information of N consumption orders before the loan, normalize the numerical variables in consumption, and encode the category variables with one hot. Among them, we use the consumption timestamp as the category variable (hour, week, month), and each transaction is formed by concatenating the normalized numerical variable and the encoded category variable.

| Amount   | 258 | 999 | 86 |
|----------|-----|-----|----|
| Pay Type | online payment | installment payment | online payment |
| Order Status | succeed | canceled | succeed |
| Time     | 16:40:23 | 22:01:45 | 23:33:17 |
| Date     | 2016-06-18 | 2016-06-18 | 2016-07-02 |
| Address  | Guang Xi | Guang Xi | Guang Zhou |

Then we use 1D-CNN to extract the important information in the consumption vector. When input to the network, the one-hot encoded vector is sparse, and CNN is used to process the vector after each transaction. Then input the features extracted by CNN into the BiLSTM network, because sometimes the prediction may need to be jointly determined by the previous inputs and the following inputs, which will be more accurate. The BiLSTM network expands the second hidden layer on the basis of the standard LSTM network. The two hidden layers process input data from opposite directions.

Finally, we input the features extracted by bidirectional LSTM into the attention mechanism. Here, we apply the attention mechanism to the loan risk assessment and classification using historical consumption order information, which is used to evaluate the importance of the last n different consumptions of loan users to the final loan result, and calculate the weight vector of each consumption.
When outputting the result, we use the Sigmoid activation function to splice the user's basic information feature vector, the user's consumption portrait feature, and the weighted consumption sequence feature, and input it into the activation function to obtain the default probability of the loan user.

4. Experiments

4.1. Datasets

The data used in this study is from a financial technology company. The data content includes the user's registration information on the loan platform and the user's consumption records on the consumption platform. In order to protect the privacy of users, the key personal information in the data has been desensitized. The data contains a total of 32,000 users, the platform registration information contains the user's demographic information (age, occupation, income, etc.), and the consumption record contains the consumption order information from the user registration to the consumption platform to the time before the loan.

4.2. Baselines

First of all, in order to prove the effectiveness of adding consumption data, that is, adding consumption number dimension can improve the accuracy of loan risk assessment, we use four machine learning methods to prove. Here we only use the basic information of users and the FCP generated by aggregation. We compare the experiment of adding FCP with the experiment of using only user basic information features. Our experimental results are obtained based on 5 times of 10-fold cross-validation, the experimental results are shown in Table 2.

In order to evaluate the ability of the model to distinguish between non-performing loans and good loans. We use two evaluation criteria commonly used in loan risk assessment: AUC(Area under curve) and KS(Kolmogorov-Smirnov)[19].From the experimental results, we know that after we add FCP, AUC, and KS values are improved in LR, SVM, XGBoost and lightgbm, which proves that consumer number plays a positive role in loan risk assessment.

| Features                  | Base | FCP | Base+FCP |
|---------------------------|------|-----|----------|
| LR AUC                    | 0.6950 | 0.591 | 0.7350 |
| KS                        | 0.3639 | 0.223 | 0.4059 |
| SVM AUC                   | 0.6227 | 0.5634 | 0.6354 |
| KS                        | 0.3047 | 0.165 | 0.3327 |
| LightGBM AUC              | 0.7859 | 0.602 | 0.8045 |
| KS                        | 0.4454 | 0.2804 | 0.4676 |
| XGBoost AUC               | 0.8036 | 0.6324 | 0.8166 |
| KS                        | 0.4606 | 0.2856 | 0.4953 |

Table 2. Performance of classifiers with basic information and consumption portrait features

| Mean Rank | p-value of Pairwise Comparison |
|-----------|-------------------------------|
| Base      | -                             |
| FCP       | <.001                         |
| Base+FCP  | <.001                         |
| Friedman X² | 638.185                       |

Table 3. Results of full pairwise comparison
Also, on the basis of the cross-validation experiment we conducted, we tested the statistical significance of adding consumption features to distinguishing loan evaluation performance. In Table 3 we show the comparison results of three feature sets. Here we use Friedman test[20], which puts the results of classification models (i.e., LR, SVM, Lightgbm and XGBoost) and performance metrics (i.e., AUC, KS) together, and performs rank-sum test on the experimental results of 5 cross-validation (sample size is 50 * 4 * 2 = 400). In terms of the results, the difference between the three feature sets is statistically significant, and the addition of consumption portrait features improves the accuracy of loan risk assessment.

4.3. sequences features
When extracting sequence features from consumption data, first of all, we compare the effect of the feature extraction using different network architectures when we use the last 30 consumption times of users. The feature information extracted by BiLSTM is more comprehensive, and then the attention mechanism is added to distinguish the weight of different consumption for classification. Considering that before inputting into BiLSTM, the consumption vector is composed of numerical normalization and one-hot categorical variables, so a sparse vector is generated, so we add 1D-CNN processing before inputting into the BiLSTM, used to extract more important information. The experimental results of several network architectures are shown in Table 4 below. The experimental results show that using CNN_BiLSTM_attention to extract features of consumption sequences has the best effect.

| Models              | AUC    | KS     |
|---------------------|--------|--------|
| LSTM                | 0.6042 | 0.1724 |
| BiLSTM              | 0.6103 | 0.1855 |
| BiLSTM-Attention    | 0.6227 | 0.2005 |
| CNN-BiLSTM-Attention| 0.639  | 0.224  |

Secondly, we compared the impact of choosing different consumption times on loan risk assessment. If the selected consumption times are small (short period), it will not be a good reflection of the user's consumption level and consumption stability within a period of time; if the selected consumption time is longer (long period), then the current economy The situation has little effect. Therefore, we choose to compare 20-70 consumptions. The experimental results are shown in Figure 2, shows that choosing the most recent 40 consumptions for loan risk assessment has the best result.

![Figure 2. AUC & KS of different consumptions times](image)

4.4. Feature fusion
This article compares the use of a single model, an integrated model, and the deep model that we propose to integrate consumption data.

Single model: The single model uses XGBoost. Because it cannot learn the sequence features in the consumption data, the single model only uses the user's basic information features and clustered consumption features.
Ensemble model[21]: Here we use a stacking-based Ensemble model, which integrates the output results of the base classifier of XGBoost and CNN_BiLSTM-attention, and then input it into the meta-classifier, where we choose LR as the meta-classifier.

Our model: According to the feature fusion model introduced in Section 3, using user basic information features, and consumption portrait features and sequence features extracted from consumption data, the classification results are finally obtained through feature fusion.

The experimental results are shown in Table 5. The experimental results show that the model we have built can improve the final risk assessment task of loan users by integrating the user’s historical consumption features and the sequence features of the extracted consumption data while using basic user information.

Table 5. Results of integrated consumption features model

| Model     | AUC   | KS     |
|-----------|-------|--------|
| XGBoost   | 0.8166| 0.4953 |
| Ensemble  | 0.8233| 0.51   |
| Our model | 0.831 | 0.54   |

5. Conclusion
At present, the main study directions in the field of loan risk assessment are divided into two categories, one is the study on loan assessment methods, the other is the study on loan assessment dimensions. Therefore, this study from the perspective of dimensions and methods, by adding the user's consumption data in the consumption platform to the loan evaluation dimension, constructs a loan evaluation model integrating consumption data. Through experiments on real data sets, the evaluation effect of our proposed fusion model is better than other models. In view of the user's historical consumption data, we extract consumption features from two aspects, analyze consumption information, construct consumption portrait features, and consider the sequence information in consumption. By constructing CNN_BiLSTM, we can extract the sequence features in consumption data well. Considering that each consumption has different impact on the user's economic level and loan risk, We introduce an attention mechanism to further enhance the classification effect.

In this study, when we fused the user's basic information features and consumption features, we only considered the sequential splicing of feature vectors. It has not considered the use of weighted fusion or fusion based on the attention mechanism. In the future, the weighted fusion method will be considered, which can well integrate the features of other dimensions into the loan risk assessment and improve the accuracy of loan risk assessment.

References
[1] Wang Z, Jiang C, Zhao H, et al. Mining semantic soft factors for credit risk evaluation in Peer-to-Peer lending[J]. Journal of Management Information Systems, 2020, 37(1): 282-308.
[2] Babaev D, Savchenko M, Tuzhilin A, et al. ET-RNN: Applying Deep Learning to Credit Loan Applications[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 2183-2190.
[3] Wang C, Han D, Liu Q, et al. A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism LSTM[J]. IEEE Access, 2018, 7: 2161-2168.
[4] Ferreira F A F, Santos S P, Dias V M C. An AHP-based approach to credit risk evaluation of mortgage loans[J]. International Journal of Strategic Property Management, 2014, 18(1): 38-55.
[5] Song P, Li L, Huang D, et al. Loan risk assessment based on Pythagorean fuzzy analytic hierarchy process[C]//Journal of Physics: Conference Series. IOP Publishing, 2020, 1437(1): 012101.
[6] Qiu W. Credit Risk Prediction in an Imbalanced Social Lending Environment Based on XGBoost[C]//2019 5th International Conference on Big Data and Information Analytics (BigDIA). IEEE, 2019: 150-156.
[7] Al Daoud E. Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit
Dataset[J]. International Journal of Computer and Information Engineering, 2019, 13(1): 6-10.
[8] Giesecke K, Sirignano J, Sadhwani A. Deep learning for mortgage risk[R]. Working paper, Stanford University, 2016.
[9] Zhang Y, Wang D, Chen Y, et al. Credit risk assessment based on long short-term memory model[C]/International conference on intelligent computing. Springer, Cham, 2017: 700-712.
[10] Neagoe V E, Ciocre A D, Cucu G S. Deep convolutional neural networks versus multilayer perceptron for financial prediction[C]/2018 International Conference on Communications (COMM). IEEE, 2018: 201-206.
[11] Jiang C, Wang Z, Wang R, et al. Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending[J]. Annals of Operations Research, 2018, 266(1-2): 511-529.
[12] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.
[13] Deng S, Li R, Jin Y, et al. CNN-based feature cross and classifier for loan default prediction[C]/2020 International Conference on Image, Video Processing and Artificial Intelligence. International Society for Optics and Photonics, 2020, 11584: 115841K.
[14] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
[15] Gers F A, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM[J]. 1999.
[16] Mnih V, Heess N, Graves A. Recurrent models of visual attention[J]. Advances in neural information processing systems, 2014, 27: 2204-2212.
[17] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
[18] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]/Advances in neural information processing systems. 2017: 5998-6008.
[19] Chatzis S P, Siakoulis V, Petropoulos A, et al. Forecasting stock market crisis events using deep and statistical machine learning techniques[J]. Expert systems with applications, 2018, 112: 353-371.
[20] Wang Z, Jiang C, Zhao H, et al. Mining semantic soft factors for credit risk evaluation in Peer-to-Peer lending[J]. Journal of Management Information Systems, 2020, 37(1): 282-308.
[21] Polikar R. Ensemble learning[M]/Ensemble machine learning. Springer, Boston, MA, 2012: 1-34.