Credit Default Risk Prediction Based On Deep Learning

Xinyu Gao
Southwest University
Yu Xiong
Southwest University
Zehao Xiong
Southwest University
Hailing Xiong (xionghl@swu.edu.cn)
Southwest University

Research Article

Keywords: risk, prediction, deep learning, Logistic regression

Posted Date: August 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-724813/v1

Credit default risk prediction based on deep learning

Xinyu Gao[1], Yu Xiong[1], Zehao Xiong[1], and Hailing Xiong[1,*]

[1]College of Computer and Information Science, Southwest University, Chongqing, 400715, China
[*]xionghl@swu.edu.cn

ABSTRACT

Logistic regression is the industry standard in credit risk modeling. However, when the model is deployed, the lack of negative samples affects the accuracy of the model, and the nonlinear characteristics of the data itself cannot be learned. In this paper, a residual neural network combined with Gan is applied to the lending club public data set to predict credit default. Among them, the number of bad users is very small, which leads to sample imbalance, and then affects the effect of the model. For this problem, we use Gan (general adverse networks) to produce bad user samples, so that the proportion of good user samples and bad user samples reaches 1:1. Finally, the residual neural network is used to predict credit default, and the accuracy is improved by about 5% compared with logistic regression.

Introduction

Losses caused by borrowers defaulting on their credit obligations are part of the banks'normal operating environment. The number and severity of default events will change with the progress of events, and eventually lead to the rising bad debt rate of banks. Credit risk is the main risk faced by financial institutions. The historical financial crisis and the bankruptcy of financial institutions are caused by loans. Therefore, credit risk control is the top priority of banking business. In recent years, with the progress of technology, the explosion of financial data is growing, ushering in the era of big data, a variety of consumer loan tools (such as mortgage loans, auto loans, credit cards or personal loans) emerge in an endless stream. Banks and financial institutions model financial big data to determine whether borrowers are qualified for loans. This shows that the credit scoring model is the basis of bank loan decision. Therefore, the core function of risk control of all financial institutions is to develop good credit scoring models and constantly optimize and update these models. Even if the model is slightly improved, in the era of big data, the improvement of bank business may be huge.

But in this era, the technology widely used in credit default prediction is logistic regression. The simple structure of the shallow layer does get the efficiency dividend, but it is easy to bring the problem of credit fraud, and ignores the rich information of the deeper data. Deep learning has made amazing achievements in image, text and so on. Many years ago, deep learning was not the main solution to deal with such problems. Therefore, in order to further develop intelligent risk control, it is necessary to study the application of deep learning method in the financial field.

In the aspect of credit default prediction, a large number of researches focus on traditional machine learning methods (such as xgboost, multi-layer perceptron, support vector machine), but the exploration of deep learning in this aspect does not start from scratch. The current research on credit risk prediction can be roughly divided into three categories:

- Traditional statistical machine learning methods, the more representative methods are logistic regression and SVM, they are linear classifiers, using 0 and 1 to represent the executor and defaulter respectively:

$$\log(p/1-p) = W^T X + b \tag{1}$$

  Where p is the probability of default, W is the parameter value of logic regression, and X is all the characteristics of the data set.This kind of classification model is widely used in credit default prediction because of its high efficiency,For example, Nie g[1] compared the performance of logistic regression and decision tree in establishing credit risk control score card,The results show that the regression algorithm is slightly better than the decision tree.

- As a machine learning method, ensemble learning combines multiple weak classifiers into a strong classifier in order to better solve the classification problem. According to the current research, this combination has significantly improved the effect of the original single weak classifier. Ensemble learning is mainly divided into boosting[2] and bagging[3] algorithm families,In bagging, the more representative model is random forest[4], which generates multiple different weak classifiers

by random sampling with put back, and finally takes the average result of these weak classifiers, that is, the combination result.Boosting is more like an improvement of bagging, which gives weight to these results, so that after weighted fusion, we can pay more attention to the impact of high weight results. For example, Jin et al[5], Proposed a multi-stage ensemble model, which is based on multiple K-means selective undersampling and used for credit scoring. Compared with the traditional benchmark model, the model results have a good performance.Y Xia et al[6],Proposed a new tree based over fitting cautious heterogeneous integration model (oche). This method can dynamically assign weights to the basic model according to the over fitting measure. Good results have been achieved.

- Deep learning is a branch of machine learning, and it has an amazing development in the field of image and text. The main model used in this paper is convolutional neural network, which is very suitable for mining the potential nonlinear features of data. According to the current research[7], CNN's algorithm has exceeded the accuracy of human raters, but it still has great development potential in the future.In the aspect of credit default prediction, Yu et al[8], Used the algorithm itself to select the appropriate training subset to balance the data set, and then formed a deep belief network (DBN) of SVM, which achieved good classification results, but this choice did not solve the problem that the data itself is incomplete.Ala'raj et al[9], Used bi-directional recurrent neural network to build user's credit score card, and compared with traditional random forest, support vector machine, logistic regression and neural network, which showed that deep learning had obvious advantages in credit default prediction.

next Section 2 describes the credit scoring models examined in the paper. Section 3 describes Lending Club data set. Section 4 presents the empirical results from comparing the models. Section 5 summarizes the paper and makes concluding remarks.

1 Model

In this part, it mainly expounds the feasibility of confrontation generation network and the deep learning model.

1.1 GAN

In the real scene, there are few cases of user default. When learning the model, the effect of the model is greatly reduced because there are few negative samples. Therefore, balancing positive and negative samples is an important step before data import model learning. Gan has better effect than traditional balanced sample method (such as smote).

The main structure of Gan is shown in Figure 1,It mainly consists of two network structures,one is generator, the other is discriminator.First, we know that the distribution of negative samples is $P_{data}(x)$,x is a negative class sample, which can be imagined as a vector. The distribution of this vector set is P_{data}.We need to generate some images that are also in this distribution. If we directly use this distribution, I'm afraid we can't do it.The distribution generated by some current generators can be assumed to be $P_G(x|\theta)$,It's a story by θ The distribution of control,θ Is the parameter of this distribution (if it's a Gaussian mixture model, then θ Is the mean and variance of each Gaussian distribution.Suppose we take some data from the real distribution, $\{x^1,x^2,\cdots,x^m\}$, we want to calculate a likelihood $P_G(x^i|\theta)$.For these data, the likelihood in the generation model is

$$L=\prod_{i=1}^{N}P_G(x|\theta) \tag{2}$$

We want to maximize the likelihood, which is equivalent to maximizing the probability that the generator generates those negative samples. This becomes a problem of maximum likelihood estimation. We need to find one θ^* To maximize the likelihood.

$$\theta^*=\arg\max_{\theta}\prod_{i=1}^{N}P_G(x|\theta) \tag{3}$$

In order to get θ^*, the loss function of Gan is set to

$$\min_G\max_D V(D,G)=E_{x\sim P_{data}(x)}[\log D(x)]+E_{x\sim P_G(x)}[\log 1-D(x)] \tag{4}$$

When $P_G(x) = P_{data}(x)$, G is optimal.

The network is trained by two-layer circulation, the outer layer is fixed with Generator, and the resolution of Discriminator is trained. The inner layer is fixed with Discriminator, and the trained Generator is trained, so that Generator can cheat the classification of Discriminator as much as possible.

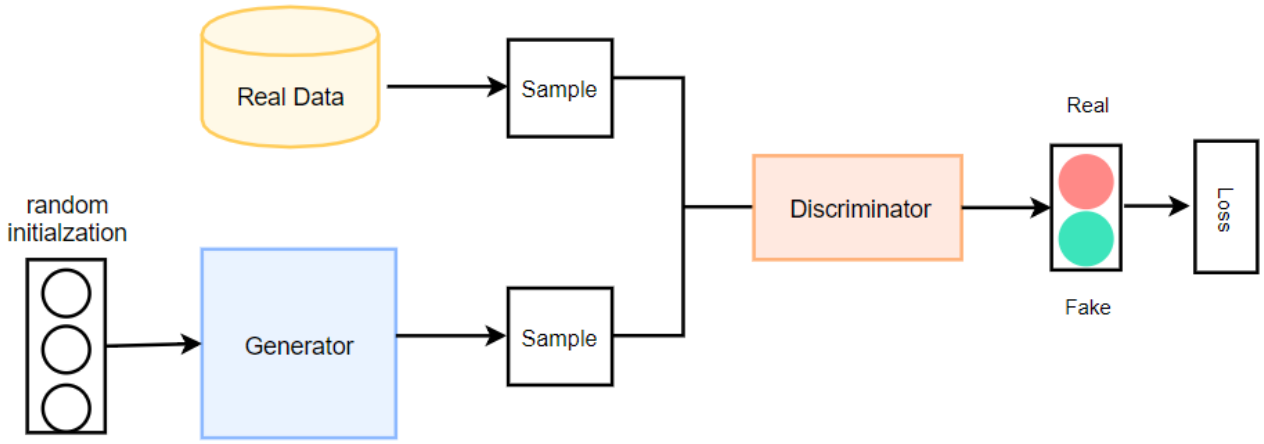

Figure 1. Generative Adversarial Network structure.

1.2 ResNet

Convolution neural network is mostly used in the field of image. Its main idea is to extract nonlinear features in the image by convolution kernel. Figure 2 is the process of feature extraction by convolution neural network. Compared with fully connected neural network, its advantages lie in parameter sharing and sparse connection. But because the data itself is linear data, we need to expand the dimension of the data. In other words, we need to establish the credit attribute portrait of the borrower.

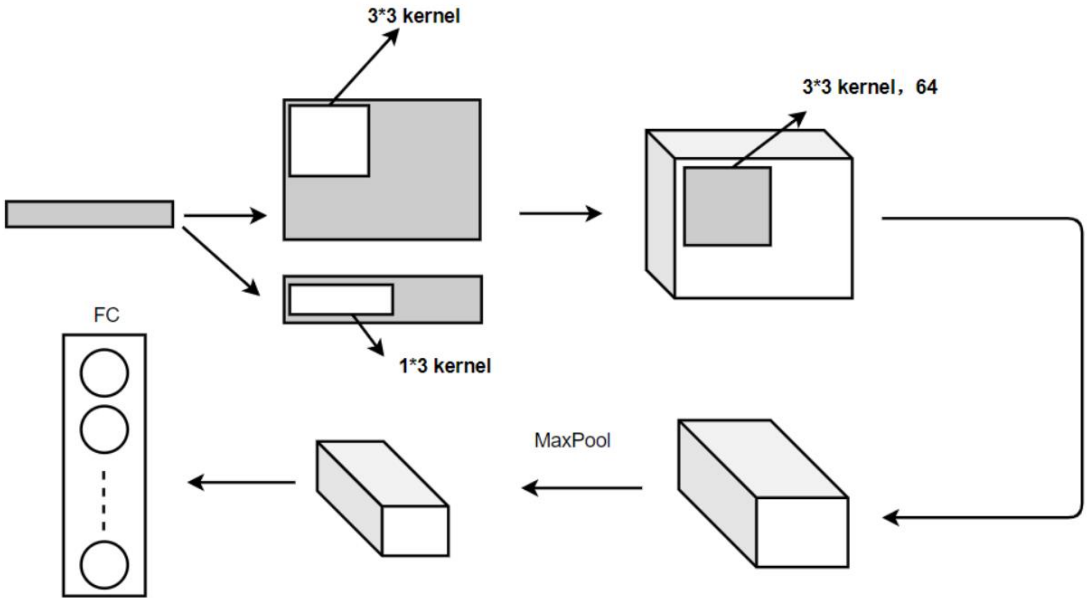

Figure 2. Convolutional neural network structure

Experiments show that the performance of traditional convolutional neural network will be saturated with the deepening of network layers, and the performance of network will begin to degrade with the increase of network layers, but this degradation is not caused by over fitting, because we find that the training accuracy and test accuracy are declining, which indicates that when the network becomes very deep, the depth network becomes difficult to train. The emergence of RESNET is actually to solve the problem of performance degradation after the deepening of the network. The emergence of RESNET[10] is actually to

solve the problem of performance degradation after the deepening of the network.As shown in Figure 3, compared with the traditional convolutional neural network (such as VGG), the same mapping is added between blocks.

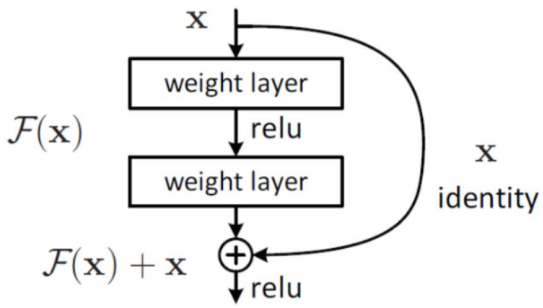

Figure 3. Residual learning: a building block

This mapping transforms the learning of target values into learning of residuals. Intuitively, the learning of residuals requires less learning, because the residuals are generally smaller and the learning difficulty is smaller. To analyze this problem from a mathematical point of view, firstly, the residual element can be expressed as:

$$y_l = h(x_l) + F(x_l, W_l) \tag{5}$$

$$x_{l+1} = f(y_l) \tag{6}$$

Where x_l and x_{l+1} represent the input and output of the L-th residual cell respectively. Note that each residual cell generally contains a multi-layer structure. F is the residual function, which represents the learned residual, while $h(x_l) = x_l$ represents the identity mapping, and F is the relu activation function. Based on the above formula.

2 Data

In this section, we mainly discuss the lending Club dataset used in the experiment, as well as the data cleaning and feature engineering.

2.1 Lending-Club

Lending club is the world's largest P2P Internet lending platform. Its main business is to assess the borrower's default risk and set different loan interest rates according to the borrower's past credit records and other information. The borrower can quickly obtain the loan by submitting an application; By browsing the borrower's past credit records and borrowing purposes, investors decide whether to lend to borrowers with different borrowing rates to earn interest income.

When a loan applicant applies for a loan from the lending Club platform, the lending Club platform allows the customer to fill in the loan application form online or offline to collect the basic information of the customer, including the applicant's age, gender, marital status, educational background, loan amount, applicant's property, etc, Generally speaking, information from third-party platforms such as credit agencies or FICO will also be used.

2.2 Data pre-processing

In order to improve the performance of the model, we need to do some data processing[11].The data set selects the business data of lending Club platform lending from 2007 to the third quarter of 2020, with a total of 2925493 samples and 142 dimensional characteristics. We delete the features with more than 30% missing, use random forest to fill the remaining missing values, and delete the meaningless features and the same valued features "loan_Status" refers to the following types of data set tags, which are encoded respectively: {current : 0, issued : 0, fully paid : 0, in grace period : 1, late (31-120 days) : 1, late (16-30 days) : 1, charged off : 1}.

2.2.1 Balanced sample

The proportion of encoded samples was 1, accounting for 11.56%; 0, accounting for 88.44%, this proportion will affect the effect of model training. Smote method and Gan method are mainly used to balance the data samples. Through the final experimental accuracy comparison, it can be found that the data generated by Gan can obtain better model prediction accuracy. Table 1 shows the comparison results of the two methods after input of logistic regression.

	False Negative	*False Positive*	*Accuracy*
SMOTE	6.05%	7.83%	92.81%
GAN	4.42%	6.28%	94.68%

Table 1. Comparison of accuracy of logical regression classification

3 Results

This paper compares the effects of two balanced sample methods: synthetic minority oversampling technique (smote) and generation countermeasure network (GAN) on the final prediction model. The accuracy comparison results of several traditional machine learning models and deep learning model RESNET are presented.

In order to minimize the influence of data correlation and improve the reliability of estimation, 10 times cross validation is used to create random partition of data set. The training set is divided into 10 equal size subsets. One of them is selected as test set at a time, and the remaining 9 sub sets are used for training, which can greatly improve the reliability of prediction model.The overall default risk prediction model process is shown in Figure 4

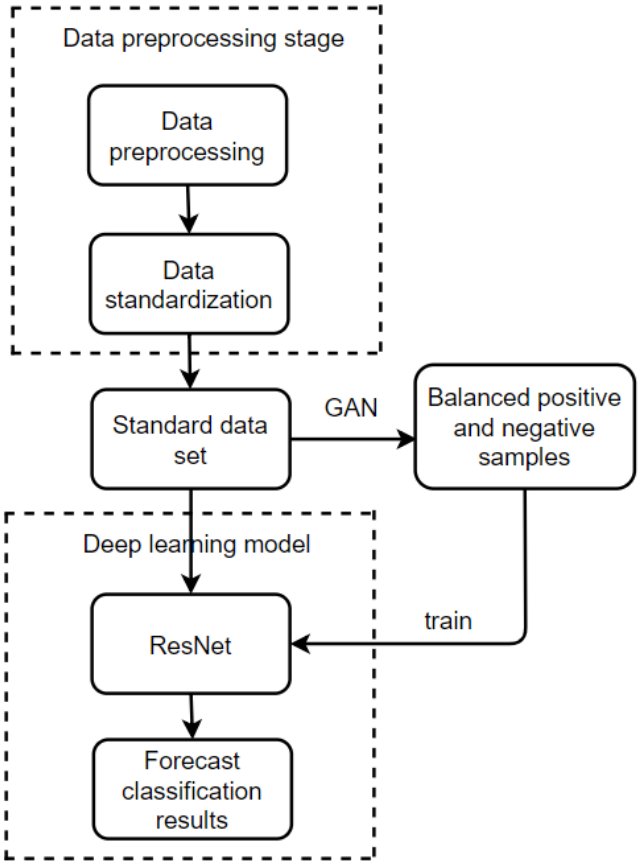

Figure 4. Convolutional neural network structure

3.1 Empirical result

Main comparison models: logistic regression, fully connected neural network and RESNET .Table 2 summarizes our experimental results, showing AUC, FN, FP and accuracy for each model. These results are the average of 10 fold cross validation.

Logistic regression is the most commonly used model in credit business at present.We have restored the standard production process of the industrial rating card. It is worth mentioning that, in addition to the lengthy data preprocessing process, in order

to balance the positive and negative samples of the data set, so that the model will not bias to predict most classes, we use the smote method to simulate the real negative class samples as much as possible by generating the counter network instead of smote method used in the original project, In order to make the model learning data set more universal. Another important feature project is feature box. The essence of feature box is discrete and continuous characteristic variable, so that the business personnel can score customers according to the information filled in by new customers. In addition, the model accuracy is increased by the case splitting, and the discretization features are easier to learn, but the data set information is lost by the case. It is because these data cleaning and feature engineering make logic regression get better precision. But in order to excavate the nonlinear characteristics of data itself, we try to ensure the information of data itself. Therefore, in addition to the logic regression model, we remove the feature box segment of the other comparison models, which also leads to the greatly reduced accuracy of the whole connected neural network, It increases the difficulty of learning model.The fully connected neural network adopts three layers, 128 neurons in the first layer and 256 neurons in the second layer,But with the continuous iterative training, the loss of training set is decreasing, but the accuracy of test set is not improved.The accuracy of the final test set is difficult to exceed 70%.

Accuracy is the standard to measure the classification ability of models. However, it is also a key problem to be solved to find out overdue users in risk control scenarios. Therefore, we introduce FNR and fpr to measure the classification ability of negative and positive samples. We do not want to see that the model tends to classify the new user samples into positive classes, but rather, we should have some generalization ability. Finally, in order to evaluate the comprehensive capability of each model, we compare the AUC index. From table 2, we can see that restnet not only improves the accuracy, but also reduces the classification error between positive and negative samples, which shows that the model is indeed superior to the traditional logic regression.

	AUC	False Negative	False Positive	Accuracy
LR	0.946	4.42%	6.28%	94.6%
ANN	0.727	39.50%	33.82%	71.59%
ResNet	0.971	2.15%	0.89%	97.8%

Table 2. Classification accuracy comparisons.

4 Conclusion

This paper proposes a classification model based on deep learning RESNET, which is used to filter the customers with serious overdue loans. In order to solve the imbalance of training data, the Gan method is adopted to generate the balanced data samples of simulation data, which improves the performance of the model, The classification performance and the final accuracy of negative classes are significantly improved.

The data set used in this paper has been desensitized, and in the era of big data, the collected user data inevitably involves privacy, so there is a certain information loss between the data set and the real data set. The first mock exam is to decide whether to give a user a loan in real scenario. Instead, we will combine a large financial data with multiple models to score a credit rating of a user to determine the loan situation. With the emergence of massive information in the era of big data, data presents the characteristics of high latitude. Traditional machine learning methods can not adapt to the complex financial scene. It is urgent to use deep learning to mine the nonlinear characteristics of data. In recent years, deep learning has made great progress in image, text and other fields, but the overall theoretical system of deep learning is not perfect, so it can not be truly implemented in many financial scenarios. Compared with the traditional machine learning method which is convenient to deploy and does not need massive and multi-dimensional big data, there is a certain gap. But I believe that in the near future, deep learning will be applied to all aspects of the financial field.

References

1. Nie, G., Wei, R., Zhang, L., Tian, Y. & Yong, S. Credit card churn forecasting by logistic regression and decision tree. *Expert. Syst. with Appl.* **38**, 15273–15285 (2011).
2. Elith, J., Leathwick, J. R. & Hastie, T. A working guide to boosted regression trees. *J. Animal Ecol.* **77** (2008).
3. Dietterich, T. G. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems* (2000).
4. Breiman, L. Random forests–random features. *machine learning* (1999).
5. Jin, Y., Liu, Y., Zhang, W., Zhang, S. & Lou, Y. A novel multi-stage ensemble model with multiple k-means-based selective undersampling: An application in credit scoring. *J. Intell. Fuzzy Syst.* **40**, 1–14 (2021).

6. Xia, Y., Zhao, J., He, L., Li, Y. & Niu, M. A novel tree-based dynamic heterogeneous ensemble method for credit scoring. *Expert. Syst. with Appl.* **159**, 113615 (2020).
7. Deep learning for visual understanding: A review. *Neurocomputing* (2016).
8. Yu, L., Zhou, R., Tang, L. & Chen, R. A dbn-based resampling svm ensemble learning paradigm for credit classification with imbalanced data. *Appl. Soft Comput.* **69**, 192–202 (2018).
9. Ala'Raj, M., Abbod, M. F. & Majdalawieh, M. Modelling customers credit card behaviour using bidirectional lstm neural networks. *J. Big Data* **8**, 1–27 (2021).
10. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *2016 IEEE Conf. on Comput. Vis. Pattern Recognit. (CVPR)* (2016).
11. Tsai, C. F., Sue, K. L., Hu, Y. H. & Chiu, A. Combining feature selection, instance selection, and ensemble classification techniques for improved financial distress prediction. *J. Bus. Res.* **130**, 200–209 (2021).