Journal of computer science and software applications
https://www.mfacademia.org/index.php/jcssa
ISSN:2377-0430
Vol. 4, No. 7, 2024

Enhancing Credit Risk Prediction in Financial Services Using Logistic Regression and Artificial Intelligence Techniques

Cheng Wang[1], Qing Xu[2]
[1]Columbia University, New York, USA
[2]Columbia University, New York, USA
Correspondence should be addressed to Cheng Wang; chengzizi.w09l@gmail.com

Abstract:

The advent of artificial intelligence (AI) and big data has transformed the financial industry, particularly in credit risk analysis and prediction. Traditional manual methods of assessing credit risk are no longer feasible for managing the increasing complexity of data and growing customer base in modern banking. This study examines the application of logistic regression and other AI-driven models, such as K-nearest neighbors (KNN), in predicting credit defaults. Findings demonstrate that logistic regression surpasses KNN in accuracy by 15.49%, showing greater efficiency in processing large datasets for credit assessment. Additionally, the performance metrics—ROC (receiver operating characteristic) curve and AUC (Area Under the ROC Curve)—confirm logistic regression's robustness, making it a practical solution for real-world applications in financial risk management. While logistic regression is advantageous for binary classification tasks within the financial sector, limitations include its reliance on data quality and applicability within varied banking environments.

Keywords:

Artificial intelligence, Financial credit, Risk prediction

1. Introduction

A few decades ago, computers were scarce, and calculating long-term benefits manually using comatrices required a bunch of graduate students to spend a lot of time. This is not only a waste of human resources, but also leads to missed opportunities for investment. With the use of intelligent statistical computing in the financial industry, its operational efficiency has been greatly improved, freeing up talent for development and other research. In recent years, artificial intelligence such as SVM, neural network and deep learning have been used to make stock prediction. The optimal portfolio is configured by multi-factor risk control model, signal monitoring and quantitative means. Big data is used to collect and identify users' personal preferences for better customized financial planning. The acquisition of information through big data and the Internet is more accurate and comprehensive than the traditional communication between financial advisers and customers to infer risk preference through the information disclosed by customers, and it can keep up with the changes of the latest data. From the rapid expansion of artificial intelligence in the financial industry, we can see that the rapid development of Internet finance, data become complex and diverse.

In the financial field, the share of bank credit market is growing rapidly. The traditional risk control model driven (internal experience) or artificially driven cannot meet the current demand of bank credit default risk prediction, leading to the occurrence of many default events. At the same time, information asymmetry makes banks unable to clarify the credit status of customers, so the credit risk of banks is also increasing, which has caused losses to many banks. Therefore, it is necessary to introduce AI algorithm (machine learning algorithm) to improve the credit analysis system, to evaluate the credit of customers, to reduce the credit.

2. Method

Logistic regression is a statistical method which is able to classify new data from AI learning of previous data collected. Logistic regression uses basically the same arithmetic as linear regression, a regression with only first power hypotheses, characteristics, and results. The expression for linear regression is

$$h_\theta(x) = \theta_0 + \theta_1 x_1 \cdots \theta_n x_n$$

In this equation, θ_0 and θ_1 represent the relationship between x_1 and $h_\theta(x)$. This expression can also be written as

$$h_\theta(x) = \theta^T x$$

where ∂ illustrates the step length. Furthermore, linear regression is a linear function model with one or multiple independent variables and dependent variables, and these variables have linear relationships. Thus, the linear regression has the capability to predict new data with AI learning of the linear predicting model. During this process, training set and learning algorithm is necessary. However, the model for logistic regression is not linear, because it also includes sigmoid function (not linear), also called logistic regression function. The equation is

$$h_\theta(x) = g(\theta^T x)$$

which has an enormous number of local minimum or local maximum. In a word, both logistic and linear regression are able to predict date from learning of the training set, however the logistic regression are more significant than the other one in the situation that we are talking about.

3. Model building

3.1 Determine prediction function

Logistic regression algorithm is a typical supervised binary classification algorithm model in machine learning algorithm. The cost function is established, and then the best model parameters are obtained by optimization iteration method. Then test to verify the advantages and disadvantages of the model. After its data processing, only 0 and 1 forms are analyzed. So as to predict whether the bank can lend or not. In the model, we use normalization to make the data get the results between 0 and 1.

3.2 Fitting analysis

Because the curve obtained by logistic regression will deviate greatly from the actual situation, we need to carry out fitting optimization after obtaining the prediction function model. After using the prediction model to output the results, you can observe multiple groups of results, find out whether fitting optimization is needed, and put forward the best scheme.

The fitting optimization is divided into two steps: the first step is to construct the cost function, and the second step is to calculate the parameters.

In the logistic regression model, we use the likelihood logarithm function, obtain the distribution density according to the logistic regression model, and deduce the parameters with samples. According to the prediction model, the probability is:

$$P(y=1|x，n)=\text{accuracy}$$

$$P(y=0|x，P)=1-\text{accuracy}$$

The two formulas can be combined to obtain the probability formula and the cost function got by other methods:

$$J(\theta)=1/p\sum p[-y(n)\log(h\,\theta\,(x(n)))-(1-y(n))\log(1-h(\theta)(x(n)))]$$

3.3 Overall algorithm application of the model

1).KNN algorithm forms a contrast, only gives the result data without graph

2).LR algorithm obtains standard parameters, data image and probability.

3).CSV is used to read data and process useless data

4).Data cleaning, missing value processing

3.4 Data needed to process the model

A data set is a collection of data. We have a total of 55,967 sets of data provided by a bank, which are divided into the training set (30,000) and the test set (25867). Each set of data contains 359 specific variables, including the age of users and other datatized information.

1).Read the data in the CSV file and count the number of positive and negative samples。

2).For the missing part of the data, take the average instead.

3).Normalize the data we gained. The maximum and minimum normalization methods are adopted to eliminate the influence of the differences in dimensions and value ranges among indicators. By scaling the data in proportion, it falls into a range between [0,1], which is convenient for comprehensive analysis. The normalized equation is as follows:

$$x^* = \frac{x - min}{max - min}$$

4).One of the columns we need to determine is column 321, which can only be 1 or -1. 1 means that the user has good credit and can lend money; - 1 instead. The rest of the data should be preprocessed and relevant features should be selected. Otherwise, the result generation speed will be slow due to the large amount of data and high computational complexity. Therefore, the random forest algorithm is used to evaluate the importance of each feature in the classification problem and find out the least important 150 features, which can be deleted when building the model to reduce the calculation burden. Of the remaining data, there are 50 columns of data that have nothing to do with model building, so ignore them. The rest of the features are what we need to build the model, read the data and record it.

3.5 The concrete steps of model(interpose the pseudocode)

Defines the important feature about x,y and remove the less important of the 150 data. Define the data in the file .calculate and choose data by the function of F. Doing this process repeated.[trainx,trainy;testx,testy]Reading the data of .csv. Ignoring the data of …column.

Reading the data of other column.

Table 1: Pseudocode for the modeling process

```
Input: the data we gain
Output: The precision, recall and accuracy of the selected algorithm and the graph
of LR

1. Define unimportant features of x and y and remove 150 of them that are less
important
2. Read the data and select the training set and test set # [train_x, train_y; test_x,
test_y]
```

```
3.  Ignore 50 columns of data that have nothing to do with model building
4.  Read the data for the remaining columns
5.  Return data
6.  Count the number of the positive and negative samples in the training set
7.  Data preprocessing for missing value
8.  Assigning SimpleImputer(missing_values=np.nan, strategy='mean') to imp
9.  Assigning data_preprocessing(sample_x) to sample_x
10. Assigning np.array(sample_x) to sample_x
11. Assigning np.array(sample_y) to sample_y
12. Assigning sample_x to train_x
13. Assigning sample_y to train_y
14.   Data preprocessing for normalization
15. Assigning preprocessing.MinMaxScaler( ) to min_max_scaler
16. Assigning min_max_scaler.fit_transform(train_x) to train_x
17. Feature selection
18. Assigning get_low_important_feature(train_x, train_y) to low_150_rf
19. Assigning np.delete(train_x,low_150_rf,1) to train_x
20. Assigning'../ data /test_data.csv' to data_csv_file_test
21. Assigning read_csv_data(data_csv_file_test) s csv_data_test
22. Assigning [ ],    [ ] to ample_x_test, sample_y_test
23. Count the number of the positive and negative samples in the training set
24. Assigning data_preprocessing(sample_x_test) to sample_x_test
25. Assigning np.array(sample_x_test) sample_x_test
26. Assigning np.array(sample_y_test) sample_y_test
27. Assigning sample_x_test test_x
28. Assigning sample_y_test to test_y
29. Dimensionality Reduction
30. Assigning np.delete(test_x, low_150_rf, 1) to test_x
31. Assigning preprocessing.MinMaxScaler() to min_max_scaler
  Assigning min_max_scaler.fit_transform(test_x) to test_x
32. Feeding back the data
```

(a) Number of positive and negative sample in the Statistical training value

We set the train_p=0,train_n=0 when the sample number and row number=0

We will do the following manipulate to the item in the sample:

Make item any value, then use Numpy, Nan to average and append to generate a multidimensional array so that the multidimensional array of the sample is equal to the number of rows item is the multidimensional array of the number of rows_x floating value if the label data is 1, the multidimensional array of the sample is also 1, training set P +1 simultaneous sample

(b)missing value handling

First, pre-process the special data in the training sample, and then convert it into the form of Numpy array. In the same way, change the corresponding label data into the form of Numpy array, and then assign the values to the training sets of x and Y respectively. Finally print(sample x) print(sample y)

(c) Data Normalization

Table 2: The result of code run

Method	AUC	Precision	Recall	Accuracy
LR	0.96	94.23%	97.18%	96.30%
KNN	0.79	80.31%	70.84%	80.81%

Using the normalization method:X*=(X-min)/(max-min)

(d)The choice of importance

The value lower than 150RF is determined as low importance eigenvalue training set X remove the eigenvalue of training set 1 or less than 150RF

(e) print(sample X)

Extract samples from the file and print the samples

(f) The number of positive and negative samples in the test value is calculated

To make test set initial value of 0 n test set p initial value is zero Rows about samples for the initial value of zero entries in processing to remove the entry of a value to average processing of X line, and remove items index of X line, end the process if X X line 1 index to the samples under test If the tag is equal to 1, 1 index to the samples Y test set Test set P = original test set P +1. And the test set that (-1) indexes to sample Y makes the test set N equal to the original test set N +1. Preprocess the data for the test set of sample X, convert the test set of sample X into numpy array, and do the same for the test set of sample Y. Finally, assign the test set of sample X to test set X, and assign the test set of sample Y to test set Y again and again

(g) Data Reduction

4. Conclusion

It can be concluded from the Table 2 that the method logistic regression has better performance than the method named KNN with 15.49 percent more accurate. The method KNN has much lower efficiency, due to the calculation of the training data and test data in each step of classification or regression which means this method is not suitable for the large amount of data. In contrast, the logistic regression method is simpler, more efficient and requires less computation which is more suitable for analyzing the big data of the credit analysis in financial industry. After running the code, we can get the Figure 1 below including the ROC curve (receiver operating characteristic curve) and AUC (Area Under the ROC Curve). We plot the true case rate (TPR) as the vertical axis and the false positive case rate (FPR) as the horizontal axis to obtain the ROC curve, and the AUC is the area under the ROC curve. The AUC index is often used as the most important evaluation index to measure the stability of**.**

The model in the model evaluation stage in dichotomies. The ROC curve increases monotonically, and each point is above the 45-degree slash (curve area is greater than 0.5), which means: TP > FP. Besides, the further the ROC curve is from the 45-degree oblique line, the better the classification effect is. Above all, the logistic regression method is more suitable for financial credit industry.

The purpose of this research is analyzing users' credit and determine whether they will be able to repay in the future through artificial intelligence. This can not only reduce the bank's lending risk, but also improve service efficiency of the bank.

On the one hand, logistic regression models can be applied to many fields, such as the research driving forces of forestland use change, in which the independent variable (drivers) described by a series of natural and social factors and the dependent variable (whether forestland changes) described by the binary classification variables data representation (1 means other land to forest land or other land to forest land，0 means not change). The process of building this model is very much the same as ours. On the other hand, the logistic regression model ca also used to solve the problem of soft binary classification. In that thesis, the probability of secondary heart attack of hospital patients is taken as a research case. Firstly, the principle of logistic regression model is analyzed, the corresponding logistic regression model is obtained by using the data onto secondary incidence, and its cross validation is carried out. Finally, the validation results are compared with the results of other models such as linear regression.

But there are still some drawbacks, including the small data sets, which means only one test could be performed. Moreover, this method may not be suitable for all situations and banks. Also, banks can only judge whether to lend, not how much.

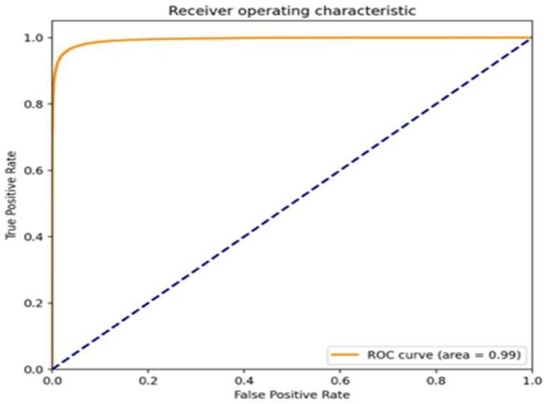

Figure 1: The ROC curve

References

[1] Jin Haoran & Ma Pingping & Liu Shenghe , Application of Logistic Regression Model in the Research on Driving Forces of Forestland Use Change [J] , World Forestry Research , 2016 , 29 (3):12-16.

[2] Xu, K., Wu, Y., Xia, H., Sang, N., & Wang, B. (2022). Graph Neural Networks in Financial Markets: Modeling Volatility and Assessing Value-at-Risk. Journal of Computer Technology and Software, 1(2).

[3] Liu, B., Li, I., Yao, J., Chen, Y., Huang, G., & Wang, J. (2024). Unveiling the Potential of Graph Neural Networks in SME Credit Risk Assessment. arXiv preprint arXiv:2409.17909.

[4] Pear to pear networks. WIKIPEDIA. [online] Available at: https://zh.m.wikipedia.org/zh-cn/%E5%B0%8D%E7%AD%89%E7%B6%B2%E8%B7%AF.

[5] Zhou Wei. What is intelligent interest. ZHIHU. [Online] 13.07.2017.

[6] Tang Xuan. Application of Logistic regression model in RISK assessment of P2P platform. IXUESHU. [Online] Null 2017.

[7] Jiang, M., Lin, J., Ouyang, H., Pan, J., Han, S., & Liu, B. (2024). Wasserstein Distance-Weighted Adversarial Network for Cross-Domain Credit Risk Assessment. arXiv preprint arXiv:2409.18544.

[8] Liu Xinhai. Credit score 60 years: change in the history. [R]School of Finance, Southwestern University of Finance and Economics, China. 18.11.2016. Page 1.

[9] Zhang Qi. A Case Study of Credit Risk Analysis and Modeling for SMEs In an Internet Finance Setting[D]. Arizona state: Arizona State University, may 2016: 1-114.

[10] Yun She Qu. Bank risk control case: Logistics model predicts bank loan default. [Online] 14.03.2018.

[11] Jun Xie, logistic regression models for machine learning and research applications,2018-5-18,p1-2.

[12] Xu, Z., Pan, J., Han, S., Ouyang, H., Chen, Y., & Jiang, M. (2024). Predicting Liquidity Coverage Ratio with Gated Recurrent Units: A Deep Learning Model for Risk Management. arXiv preprint arXiv:2410.19211.