Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 167 (2020) 2–8

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Intelligence and Data Science (ICCIDS 2019)

Credit Fraud Detection Based on Hybrid Credit Scoring Model

Keqin Chen[a,b]*, Amit Yadav[b], Asif Khan[c,d] and Kun Zhu[b]

[a]*School of Business Administration, Southwestern University of Finance and Economics, Chengdu 611130, China*
[b]*Department of Information and Software Engineering, Chengdu Neusoft University, Dujianyan, Chengdu, Sichuan 611844, China*
[c]*Crescent Institute of Science and Technology, Vandalur, Chennai 600048, India*
[d]*Universty of Electronic Science and Technology of China (UESTC), Chengdu, Sichuan, P.R.China*
**asifkhan@uestc.edu.cn*

Abstract

Credit risk rating can be described by several economic activity indicators. Utilizing these financial movement markers to build a tenable credit scoring model will enormously improve the precision of the model. This method can be used in a series of credit evaluations and specific economic conditions. A reasonable scenario in which the uncertainty is consistent. In this paper, the logistic regression algorithm is joined with weighted evidence to fabricate another credit score model. Through the relationship existing in economic activities, the connection of each economic movement is additionally dissected by utilizing the correlation orthogonal transformation in the weight of proof to improve the exactness of the model. In practice, due to numerous weaknesses in the records, there is significant error in the logistic regression. Hence, building of hybrid scoring model can increase the accurateness of credit score. Thus improved the prediction rate of user credit scores and reducing the occurrence of credit fraud.

Keywords: Logistic Regression; Weight of Evidence; Credit Fraud Detection; Credit Evaluation;

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: asifkhan@uestc.edu.cn

1877-0509
10.1016/j.procs.2020.03.176

1. Introduction

Big data has the characteristics of large quantity, fast speed, strong diversity, low value density and strong authenticity. As the important catalysts for financial development, credit is the essential action of banks everywhere the world, and credit fraud detection is even more important. So far, it is not clear which literature indicates that there is a suitable credit evaluation model [1]. This paper uses the published dataset, applies the logistic regression model in machine learning, and constriction of hybrid model is done with the combination of the weighted evidence to understand the relationship between the current situation of the individual and its fate in the near future, and reduce the phenomenon of credit fraud happened.

In recent years, many financial institutions and scholars have studied many different credit scoring models due to the release of many credit-related data in some countries [2]. Risk management plays significant role in the corporate strategy of many huge financial institutions and corporations, which is usually sustainable [3]. On the off chance that there is just default data, it is hard to choose whether there is any bank fraud or related archive misrepresentation [4]. Customer are additionally uncertain about whether to reimburse their loans. The risk of loan instalment can be adequately diminished by examining client behaviour ahead of time [5]. To assess whether another candidate can get a loan (an applied score), money related organizations commonly utilize a credit scoring model to examine and foresee the probability of a borrower defaulting on a current credit item (behavioural scoring) [6].The two data sets used in this paper are from Kaggle's public data set, which is also a real data set after processing and is very popular in machine learning research and analysis.

In 1989, Goovaerts and Steenackers used the Logistic regression model to construct a system for evaluating personal credit scoring [7]. Although the results were not ideal, the Logistic regression model has been widely used in credit and asset valuation since then. In 1997, Leatham and Ellinger used linear programming and statistical models to study personal credit assessments and established a new credit model with significant results [8]. In 2000, Thomas applied the integrated approach to credit evaluation, incorporating the lender's economic conditions into the credit evaluation system model [9], further improving the accuracy of the credit model, and pointed out that the current credit evaluation model can predict customer property. Further research on risk. In 2018, Jacky used machine learning to detect financial credit fraud [10].At the same year, Huang et al. find connections between entities and use the trading network to detect credit anomalies [11]. Chouiekh and Hassane use the convolutional neural network in deep learning for fraud detection analysis [12].

2. Research methods

In the exercise of modern bank risk management, credit risk assessment method plays significant role [13]. When the relevant data establishes a personal credit rating model, different methods will have different characteristics. Credit fraud detection is mainly used to distinguish the trustworthy and the untrustworthy. Despite the fact that the forecast precision of Logistic regression model isn't tantamount to that of neural network, it has the benefits of good strength, consistent or characterized independent variables, robust illustrative intensity of the model, and the ability to make a linear scorecard. First, the credit scoring model is established by using the weight of evidence, and then different weights are used as variables. One of the variables, after adding other characteristic variables, established a new credit score model based on Logistic regression. The advantage is that due to the high precision of the weight of evidence, its credit rating results contain more information to explain the relationship between variables and dependent variables. The credit assessment result is utilized as one of the logical factors to improve the conjecture precision of the whole scoring model. In conclusion, the logistic regression model can be built to guarantee fortifies and clarification of the model. As a result, this hybrid model is more accurate than the model recognized with a single process.

2.1. Modeling data and feature variables

This paper mixes Kaggle and UCI datasets (https://www.kaggle.com/datasets), uses additional than 100,000 portions of data, and performs missing processing on different data between the two data sets, which is more in line with the actual information asymmetry. Feature selection is very essential in data processing. Chooses credit card account feature of data, this paper discusses the trading behavior of these accounts, and analysis of the distribution of selected attributes (as shown in figure 1), to choose the best amalgamation for better figure results, when the information fit to singular loans consumption for the accompanying essential properties: containing the age of the mortgagor, dissolvability and credit (unpaid days), the situation of the property, loans past due property factors, including time window and the size of family borrower.

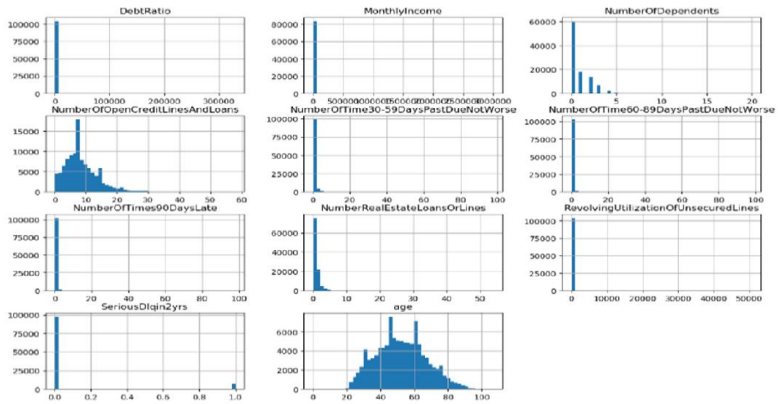

Fig. 1. Borrower Data Distribution Map

2.2. Data Processing

There were 19,371 missing values for the monthly income variable in the data set and 3,924 mislaid values for the quantity of dependents. When developing credit risk rating model, mislaid value needs to be dealt first. Mislaid values are generally treated in the following three ways :(1) the sample data containing missing esteems are erased straightforwardly, (2) the value missing is accomplished by the information mining process (closeness among tests), (3) the value misplaced is accomplished by the relationship among factors. The month to month pay misfortune amount of the variable is more in data collection. Hence, random forest method was used to fill the missing value according to the correlation between variables. The number of mislaid values of a numeric dependent variable is relatively small and is deleted directly. This will not have a big impact on the entire model.

2.3. Variable selection

First, the processing variables (sub boxes) are discretized. The variable box is used to handle the discretization of continuous variables. In this paper, the peak separation of continuous variables is first selected, and the economic significance of indicators is determined by WOE analysis by equating the evasion probability of indicator box and corresponding sub boxes using the variable selection method of credit score model. When the dispersal of continuous variables is not easy to be optimally divided, the interval of continuous variables should be considered. The number of undefined rows, age, debt ratio and monthly income rotation utilization rate in the data set were classified by optimal segmentation. Table 1-2 shows the data elimination results of major variables.

Table 1. Debt Ratio Data Binning

Container	min	max	sum	total	rate	woe	Goodattribute	Badattribute
(-0.001, 0.237]	0.0000	0.2369	22153	23565	0.9401	0.1239	0.3360	0.2968
(0.237, 0.544)	0.2369	0.5438	22060	23565	0.9362	0.0566	0.3346	0.3162

(0.544, 329664)	0.5438	329664	21723	23565	0.9219	-0.1610	0.3295	0.3870

Table 2. Monthly Income Data Bin

Container	min	max	sum	total	rate	woe	Goodattribute	Badattribute
(-0.001, 3400]	0.0000	3400	21851	23728	0.9209	-0.1745	0.3314	0.3946
(3400, 6870)	3401	6870	21670	23403	0.9260	-0.1030	0.3287	0.3643
(6870,3008750)	6871	3008750	22415	23562	0.9513	-0.3435	0.3400	0.2411

2.4. Data analysis

Nowadays risk score assessment model is a commonly used scoring model, which mainly uses value of information to score. In this paper, WOE examination is utilized to partition the list into holders and figure the hub weight estimation of each apparatus to watch the change pattern of the record esteem. Expect a continuous random variable X, the function density is f(x), characterize entropy as:

$$\mathrm{H}(\mathrm{X}) = -\int f(x)\mathrm{In}f(x)\mathrm{d}x = -\mathrm{E}[\mathrm{In}p(\mathrm{X})] \tag{1}$$

Two consecutive random variables X and Y are commonly used to measure the distance between them. The relative entropy, also known as Kullback-Leibler divergence is defined as follows:

$$H(Y \mid X) = \mathrm{E}_Y[\mathrm{In}\frac{f_Y}{f_X}] = \int f_Y(x)\,\mathrm{In}\,\frac{f_Y(x)}{f_X(x)}\mathrm{d}x \tag{2}$$

The imprint scale set by the record is clear utilizing the present articulation, which means the imprint as the logarithm of the proportion. The formula is mentioned below:

$$\mathrm{score}_{\mathrm{total}} = \mathrm{Ascore}_{\mathrm{total}} + \mathrm{B} * \mathrm{In}(\frac{BAD}{GOOD}) \tag{3}$$

The WOE scientific description, WOE- ln (goodattribute/badattribute). Information Value (IV) is info or the worth of info. The value of data be dependent upon Weight of Evidence (WOE), and IV is a noteworthy list to scale the measure of outcome of independent factors on target factors. The equation is beneath:

$$IV = \sum_{\mathrm{i}=1}^{N}(\text{Good Proportion-Bad Proportion})*WOE_i \tag{4}$$

For specific factors, the reliable shirking preliminary thickness intention is communicated as the standard example thickness capacity of that variable. IV is represented to as the total of the association entropy of the avoidance test enduring to the typical sample and that of the standard sample identifying with the default test, i.e:

$$IV = E_d[\mathrm{In}\frac{f_d}{f_{\bar{d}}}] + E_{\bar{d}}[\mathrm{In}\frac{f_{\bar{d}}}{f_d}] = \int (f_{\bar{d}}(x) - f_d(x))\mathrm{In}\frac{f_{\bar{d}(x)}}{f_{d(x)}}\mathrm{d}x \tag{5}$$

The correlation between variables was then observed using the cleaned data, and the selection of variables was tested using model IV (evidence weight). (Figure2 demonstrate IV diagram. Figure3 is the heat map shows the correlation between variables).

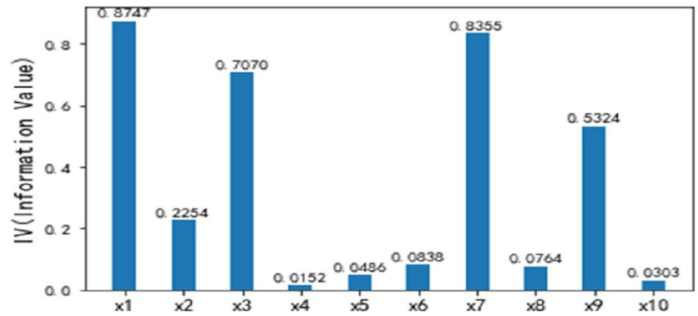

Fig. 2. IV Diagram (X1-X10 corresponds to the borrower's attributes)

The IV values of X1 and X7 are large, and the training effect on the model is significant.
The IV values of X4, X5, X6, X8, and X10 are small, and the training effect on the model is not significant.

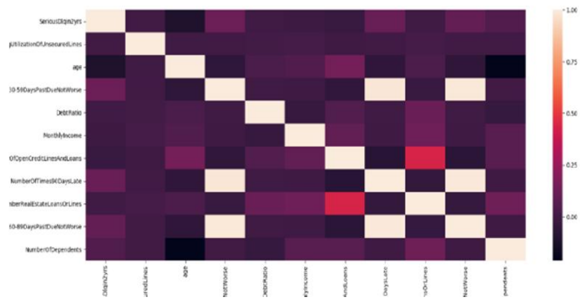

Fig. 3. The Heat Map shows the correlation between variables

The logistic regression model can be transfer by the evidence weight transformation into the typical scorecard setup. The reason WOE transformations are not imported to enhance the superiority of the model is that they do not add value to the model, and there should be no variables in the model. It is usually possible to establish a standard credit score card. The logistic regression model requires to analyse large scale of independent variables, which upsurges the intricacy of the demonstrating program, but the end result is the same as the scorecard. Before illustrating, likewise consider adjusting sieved factors to WOE models for credit ratings.

3. Experiential study outcomes

In Python, you can use metrics in sklearn to easily compare two classifiers and automatically calculate ROC and AUC, as shown in Figure 4 and Figure 5. Compared with the hybrid model, the mixed model prediction results are greatly improved.

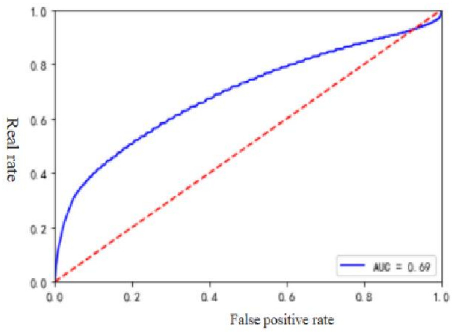

Fig. 4. AUC curve after traditional logistic regression test

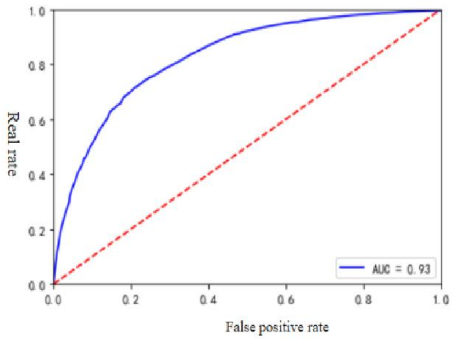

Fig. 5. AUC curve after mixed model test

Taking COEF esteem as the rule, the presentation of Ordinary Least Squares model (OLS), customary logistic model and mixed model was watched, and it was discovered that mixed model was altogether influenced by different models. Table 3 beneath is the constant coefficient (COEF) estimation of each model.

Table 3. Table of Assessment Constant Coefficient results of all model

Model / COEF / Variable	OLS model	Traditional logistic model	Logistic evidence weight mixing model
Constant	0.146317	-1.316169	9.800337

All in all, as one of the descriptive variables of logistic regression algorithm, the scoring model of neural network is increasingly appropriate for the out-dated neural system. This paper checked the quality and likelihood of logistic regression and evidence weighting technique through the model, and furthermore found that this model can retouch the client's personal credit assessment all the more precisely and practically. Credit score organizations can additionally serve clients by entering relating qualities to extra evaluate their potential risks. It very well may be anticipated that with the persistent improvement of computerized reasoning and huge information, the calculation can be continually upgraded. This is likewise helpful for improve the identification of credit misrepresentation and lessen individual credit security risks.

4. Conclusion

Through the study of personal credit fraud detection and the comparison of various credit score models, this paper finds that a progressively powerful forecast model can be accomplished by combination of the evidence weight and Logistic regression model. This mixed scoring model can get progressively exact credit score assessment, yet in addition decrease risk of credit. For economical associations, it is hard to equitably accomplish all the proof of individual clients, which signs to the lacking proof asymmetry. This is one of the main reasons for credit risk and it is believed that this study will help to minimize such risk.

For future examination, this exploration will be stretched out to accentuation on disclosure of treasured information from a lesser amount of information or an enormous amount of missing informational collections. At the same time, the model will be optimized to maintain the stability and robustness of the model and improve the computing speed. Then analysing and predicting of these data will be done to improve the level of credit fraud detection.

References

[1] Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39(3), 3446-3453.

[2] Šušteršič, M., Mramor, D., & Zupan, J. (2007). Consumer credit scoring models with limited data. Expert Systems with Applications, 36(3), 4736-4744.

[3] Bennouna, G., & Tkiouat, M. (2018). Fuzzy logic approach applied to credit scoring for microfinance in morocco. Procedia Computer Science,127, 274–283.

[4] Louzada, F., Moreira, F. F., & Oliveira, M. R. D. (2017). A zero-inflated non default rate regression model for credit scoring data. Communication in Statistics- Theory and Methods(1).

[5] S.Sampath and V. Kalaichelvi. (2014) "Assessment of mortgage application using fuzzy logic". World academy of Science , Engineering and Technology, International Journal of Social, Behavioral, Education, Economic , usiness and Industrial Engineering. Vol 8, No 11.

[6] Djeundje, V. B., & Crook, J. (2018). Incorporating heterogeneity and macroeconomic variables into multi-state delinquency models for credit cards. European Journal of Operational Research.

[7] Steenackers, A., & Goovaerts, M. J. (2006). A credit scoring model for personal loans. Insurance Mathematics & Economics, 8(1), 31-34.

[8] Ziari, H. A., Leatham, D. J., & Ellinger, P. N. (1997). Development of statistical discriminant mathematical programming model via resampling estimation techniques. American Journal of Agricultural Economics,79(4), 1352-1362.

[9] TIMOTHY H. LEE, & MING ZHANG. (2003). Bias correction and statistical test for developing credit scoring model through logistic regression approach. International Journal of Information Technology & Decision Making, 2(02), 299-311.

[10] Chow, J. C. K. (2018). Analysis of financial credit risk using machine learning. Papers.

[11] Huang, D., Mu, D., Yang, L., & Cai, X. (2018). Codetect: financial fraud detection with anomaly feature detection. IEEE Access, 6(99), 19161-19174.

[12] Chouiekh, A., & Haj, E. H. I. E. (2018). Convnets for fraud detection analysis. Procedia Computer Science, 127, 133-138.

[13] Mukid, M. A., Widiharih, T., Rusgiyono, A., & Prahutama, A. (2018). Credit scoring analysis using weighted k nearest neighbor. (Vol.1025, pp.012114)