ISSN 2616-5902 Vol. 4, Issue 4: 7-10, DOI: 10.25236/AJBM.2022.040402

A Study on the Measurement and Evaluation of Personal Credit Risk Impact Factors Based on Machine Learning

Chuyu Feng[1], Ling Gu[2]

[1]School of International Business Administration, South China Normal University, Guangzhou, Guangdong, 510000, China
[2]School of Economics and Management, Southwest Jiaotong University, Chengdu, Sichuan, 610000, China

Abstracts: *With the development of large-scale data technology, the Internet finance industry is developing rapidly. Research on personal credit risk assessment models is conducive to strengthening risk control and management and improving the efficiency of accurate services by financial institutions. Based on this, this paper first considers the amount of information, independence and relevance of the data, and uses four measurement to measure the combination of credit impact factors such as basic personal information, basic credit information and credit behaviour information. Four machine learning methods are then used to assess individual credit risk and give a comprehensive comparison of the degree of importance of credit risk impact factors on individual credit risk. The empirical results show that the model is accurate and stable, and can well reflect the degree of influence of individual credit characteristics on credit risk. Finally, providing systematic construction suggestions of risk management in banks and other financial institutions.*

Keywords: *Personal Credit Risk, Impact Factors, Risk Assessment, Machine Learning*

1. Introduction

With the development of the macro economy and the Internet financial industry, an increase of 23.07% over the previous year, with a growth rate significantly higher than the average economic growth rate. To accommodate the increased demand for loans, various lending platforms have emerged. This in turn poses a serious challenge to major banks - how to identify individual credit applications that present a high credit risk. There are still many problems with personal credit in commercial banks in China. On the one hand, people's awareness of credit is relatively weak. On the other hand, the phenomenon of late repayment is worsening. Compared to corporate business, banks face greater uncertainty. From this perspective, it is of great importance to the development of Internet finance in China to effectively construct a measurement and assessment system that can fundamentally improve the credit market environment.

Regarding personal credit risk assessment methods, there are currently more studies by domestic scholars in this area: She Chaobing (2018) used logistic regression algorithm and found that banks could adopt appropriate factor weight according to their business objectives to ensure the stable development of their personal credit business under the experimental background of using weight strategy to solve the sample imbalance problem. Cai, Wenwen, Luo, Yonghao, Zhang, Guanxiang and Zhong, Huiling (2017) found that the fusion model of GBDT and logistic regression had higher credit risk prediction accuracy than other models. In summary, the measurement and evaluation system construction of personal credit risk impact factors based on machine learning methods meets the development needs of the domestic personal credit market. This paper firstly constructs a personal credit risk indicator system based on basic personal information, basic loan information and loan behaviour information, and then uses four measurement methods such as entropy value method, CRITIC method, information quantity weights and independence weight coefficients to measure the combination of personal credit impact factors based on the consideration of data information quantity, volatility and independence. Four models, including Random Forest, Gradient Boosting Tree, XGBoost and LightGBM, are then used for individual credit risk assessment. On the premise that each model is accurate and stable, the most appropriate model to be used in this context is determined by judging the size of the relevant assessment indicators. Finally, using the Ali Tianchi public personal credit dataset as a research sample.

ISSN 2616-5902 Vol. 4, Issue 4: 7-10, DOI: 10.25236/AJBM.2022.040402

2. Personal Credit Risk Indicator System and Data Pre-Processing

2.1. Selection of Indicators and Data Sources

The credit risk indicator system is the most important part of the credit risk control management process. There are also many studies in this area in China. Zhou Xiang, Zhang Wenyu and Jiang Yefeng (2020) proposed a risk indicator system consisting of 75 characteristic values such as current loan status, latest payment information and credit score.In this paper, the 16 indicators are divided into three categories of modules - basic personal information of borrowers, basic credit information and credit behaviour information in accordance with the principles of scientific, comprehensiveness, comparability and practicability.

The data in this paper were obtained from the Ali-Tianchi datasets (https://tianchi.aliyun.com/competition/entrance/531830/information), the data were treated for missing values, outliers, quantitative indicators quantified as numerical data and normalised, and anonymous variables that could not be analysed for variable interpretability and were therefore excluded. The credit records in the sample dataset were obtained according to the established personal credit risk indicator system. While eliminating extreme outliers and dealing with missing values, the sample datasets were normalized and 36915 samples were retained for the study.

Table 1: Personal credit risk indicator system.

Factors	Assessment indicators	Code
Basic information for lenders	Years of employment (years)	X1
	Home ownership status provided by the borrower at the time of registration	X2
	Annual income	X3
	Debt to income ratio	X4
Credit Basic Information	Total current credit limit in the borrower's credit file	X5
	Loan amount	X6
	Loan term (years)	X7
	Loan Rates	X8
	Instalment amount	X9
	Loan Rating	X10
Credit behaviour information	Total credit working balance	X11
	Revolving credit utilization	X12
	Number of defaults over 30 days past due on the borrower's credit file in the past 2 years	X13
	Number of open lines of credit in the borrower's credit file	X14
	Number of derogatory public records	X15
	Number of public records cleared	X16

3. Research Methodology

Firstly, while quantifying the original index X, four sets of weights W1, W2, W3, W4 were obtained after applying each of the four measures of entropy, CRITIC, information weights and independence weight coefficients to it. Secondly, after combining the weights of the four sets of weights W1, W2, W3, W4 corresponding to each original indicator X, the scores F1, F2, F3 of the three modules of basic borrower information, basic loan information and loan behaviour information are calculated.

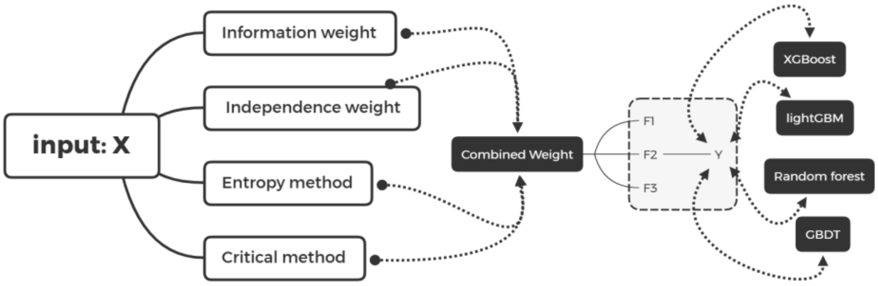

Figure 1: Modelling process.

Next, the outcome Y (whether to default) is predicted using a combination of four model. On the premise that each model has accuracy and stability, the relative degree ranking of the impact of the three major modules F1, F2, F3 on Y is derived based on the output results of feature importance. The specific

ISSN 2616-5902 Vol. 4, Issue 4: 7-10, DOI: 10.25236/AJBM.2022.040402

process is schematically shown in Figure 1. The four weighting methods are programmed with Python's numpy, pandas and math package etc; The machine learning methods are programmed with Python's sklearn package.

4. Research Analysis

4.1. Measure of Individual Credit Risk Impact Factors Based on Portfolio Weights

Firstly, the sample data were normalised, and the four sets of weights W1, W2, W3, W4 were obtained after each original indicator was measured by the entropy value method, CRITIC method, information quantity weight method and independence weight coefficient method in turn. The entropy method measures mainly the amount of information in each indicator, while the CRITIC method, the independence weighting factor and the information weights take into account the volatility of the indicators. Therefore, the combined weights obtained from the combination of these four measures are more comprehensive, stable and secure. Secondly, the combination weights are obtained by combining the four sets of weights for each original indicator. The relative degree of influence of each original indicator on the probability of default was obtained using the combination weight as a measure. The detailed information is shown in Table 2.

Table 2: Combination weights.

Index	X1	X2	X3	X4	X5	X6	X7	X8
weights	0.338417	0.036036	0.000002	0.000065	0.037098	0.007433	0.042156	0.472328
Index	X9	X10	X11	X12	X13	X14	X15	X16
weights	0.01756	0.00873	0.000023	0.039025	0.000015	0.000833	0.000005	0.000274

4.2. Personal Credit Risk Assessment Based on Multiple Machine Learning Methods

Firstly, the combination weights of the original indicators obtained above were applied to the sample data to obtain the quantitative scores of the three modules of basic borrower information, basic loan information and loan behaviour information in each sample. Secondly, the data that received the weighting process were applied to the individual credit risk assessment using Random Forest, Gradient Boosted Tree, XGBoost and LightGBM respectively. On the premise that each model has accuracy and stability, five evaluation metrics - accuracy, F1 score, recall, and prediction accuracy - were used to compare them. The results showed that XGBoost outperformed the other three machine learning methods in a total of five metrics - accuracy, F1 score, recall, prediction accuracy, and AUC score - and each metric was closest to 1. The specific information is shown in Table 3.

Table 3: Scores for each metric of the four machine learning methods.

Index	RF	GBDT	XGBoost	LightGBM
Accuracy	0.814 +/- 0.010	0.808 +/- 0.006	0.859 +/- 0.006	0.828 +/- 0.007
F1 Score	0.745 +/- 0.013	0.728 +/- 0.009	0.803 +/- 0.009	0.762 +/- 0.011
Recall rate	0.691 +/- 0.019	0.641 +/- 0.016	0.719 +/- 0.016	0.689 +/- 0.017
Predictive Accuracy	0.811 +/- 0.016	0.845 +/- 0.018	0.910 +/- 0.012	0.854 +/- 0.012
AUC score	0.882 +/- 0.007	0.869 +/- 0.007	0.911 +/- 0.007	0.893 +/- 0.006

Next, the four machine learning models were applied to determine the importance of features to obtain the AUC score and F1 score respectively. In terms of AUC score, XGBoost was the best with 0.9198, LightGBM was the second best with 0.895846833, Random Forest was the third best with 0.8877 and Gradient Boosting Tree was the last with 0.8762. In terms of F1, XGBoost is still the best performer with a score of 0.8223; LightGBM is second with a score of 0.7728; Random Forest is third with a score of 0.7548; and Gradient Boosting Tree is last with a score of 0.7448.

Table 4: Impact factor scores.

Factor	RF	XGBoost	LightGBM	GBDT
F1	0.326414	0.439214	0.556333	0.498762
F2	0.416011	0.453030	0.211667	0.464556
F3	0.257575	0.107756	0.232000	0.036682

In this paper, we showed that the predictive performance of the integrated learning methods was better. The difference in predictive performance between the four integrated learning methods is not significant, so for the feature importance, it also reflects the consistency of the output feature importance from the

ISSN 2616-5902 Vol. 4, Issue 4: 7-10, DOI: 10.25236/AJBM.2022.040402

perspective of different models. This also represents that this feature has an important role in individual credit risk assessment. the relative importance of the three main modules of basic lending information, basic loan information and loan behaviour information in Random Forest, Gradient Boosting Tree, XGBoost, and LightGBM, respectively, was obtained. From the above, it is clear that XGBoost is the most appropriate to use in this context. Therefore, this conclusion uses the magnitude of F1, F2, F3 scores under XGBoost as the basis for determining the ranking of the relative importance of borrower basic information, loan basic information, and loan behaviour information on the influence of default probability sex. The results show that F2 scores the highest, F1 the second highest and F3 the lowest, i.e. basic loan information has the deepest relative influence on the size of the probability of default, basic borrower information has a relatively weaker influence on the size of the probability of default compared to basic loan information, while loan behaviour information has the weakest ability to influence. The specific information is shown in Table 4.

5. Conclusions and Discussions

This paper evaluates the personal credit risk impact factors through indicator quantification, combination measurement, random forest, gradient boosting tree, XGBoost and LightGBM, and finds that basic loan information has the highest relative impact on personal credit default risk; commercial banks should focus on a total of six indicators including the current total credit limit in the borrower's credit file, loan amount, loan interest rate, installment amount and loan rating. In order to ensure the sound development of credit business, commercial banks should focus on six indicators, including the total number of credit lines in the borrower's credit file, loan amount, loan term (in years), loan interest rate, instalment amount and loan rating; they should also pay attention to the borrower's basic information, which includes four indicators, including the length of employment (in years), home ownership status provided by the borrower at the time of registration, annual income and debt-to-income ratio.

Besides, if commercial banks pursue sound development and wish to avoid risky business as far as possible, their threshold value can be slightly greater than 0.5, while if in addition, XGBoost performs best in all five indicators and is significantly better than the other three methods. Commercial banks can make use of the personal credit risk impact factor measurement and evaluation system constructed in this paper to judge the probability of borrower default in order to facilitate the sound development of commercial banks.

References

[1] Cai Wenwen, Luo Yonghao, Zhang Guanxiang, Zhong Huiling Personal credit risk assessment model and empirical analysis based on the integration of gbdt and logistic regression [J]. Management modernization, 2017, 37 (02): 1-4.
[2] She Chaobing, Application of logistic regression in bank personal credit risk assessment [J]. Technology and innovation, 2018 (19): 113-114 + 118-119.