Customer Loan Credit Risk Prediction Based on Improved Sparrow Search Algorithm Optimised Random Forest Algorithm

Yuqing Li[1,a,*]

[1]*University of Connecticut, 2101 Mission St, San Francisco, CA, 94110, USA*
a. yuqing.2.li@uconn.edu
**corresponding author*

Abstract: This paper aims to improve the accuracy of credit risk assessment by optimising the traditional random forest model through the sparrow search algorithm. The study used Pearson correlation analysis to explore the extent of association between customer indicators and loan delinquency. The results show that income, homeownership status, and years of employment in the organisation are positively associated with non-delinquency, while age, loan purpose, amount, interest rate, percentage of income, length of credit history, and approval status are negatively associated with loan delinquency. The model achieved 99.87% accuracy in the training set after 15 iterations. The confusion matrix showed 20,021 customers were successfully predicted with only 26 prediction errors. The test set accuracy was 82.55% with 7092 customers successfully predicted and 1499 prediction errors. In summary, the study achieved significant results in credit risk assessment and provided financial institutions with a more accurate basis for decision-making.

Keywords: Credit Risk, Sparrow Search Algorithm, Random Forest

1. Introduction

In the financial environment, assessing credit risk is extremely important for both lenders and financial institutions. For lenders, credit risk assessment has a direct impact on whether they can obtain loans and the level of loan interest rates. For financial institutions, effective credit risk management is one of the keys to maintaining the stability and healthy development of the financial system.

Firstly, for lenders, credit risk assessment is one of the important bases for banks or financial institutions to decide whether to provide loans to them. By evaluating the borrower's credit history, income, asset status, etc., banks can better judge whether the borrower has the ability and willingness to repay the loan [1,2]. If borrowers have good credit, then they are more likely to obtain more favourable loan interest rates and terms; conversely, if borrowers have poor credit, they may face the risk of higher interest rates or loan rejection [3].

Secondly, for financial institutions, effective credit risk management is the basis for ensuring the sustainable development of their business and risk control. By scientifically and systematically assessing and monitoring the credit risk level of customers, financial institutions can identify potential default risks in a timely manner and take appropriate measures to control them. This helps prevent an increase in non-performing assets, reduce losses, and improve overall profitability and market competitiveness [4,5].

When performing credit risk assessment, machine learning algorithms play an increasingly important role in predicting whether a loan applicant will default on a loan. Traditional credit scoring models are often built based on rules and statistical methods, which have limitations when dealing with large-scale data [6]. Machine learning algorithms, on the other hand, can more accurately predict the probability of default of an individual or a group by analysing massive data and identifying hidden patterns and features.

Machine learning algorithms commonly applied in credit risk assessment include logistic regression, random forest, support vector machine and neural network, etc. Various traditional machine learning algorithms have achieved good results in credit assessment risk, and among the traditional machine learning algorithms, the random forest model performs with better results, and in this paper, based on the sparrow search algorithm, we optimise the traditional random forest model, and hope that we can further improve the accuracy of credit risk assessment.

Overall, combining traditional machine learning methods with the latest algorithms can improve the accuracy and efficiency of credit risk assessment. This paper explores this to provide a certain foundation for subsequent research.

2. Source of data sets

The data selected for this paper comes from the Kaggle open source dataset, which is available at (https://www.kaggle.com/datasets/nanditapore/credit-risk-analysis). The dataset contains 32,582 pieces of data, and each piece of data records various indicators of a loan customer with whether or not he/she eventually defaults on the loan, and the indicators include age, income, home ownership status, years of employment in the organisation, the purpose of the loan, the amount of the loan, the interest rate of the loan, the loan as a percentage of the income, the length of the credit history, the status of the loan approval, and whether or not he/she eventually defaults on the loan. The indicators and their meanings are shown in Table 1.

Table 1: The indicators and their meanings.

Parameters	Meanings
Age	The age of the loan applicant.
Incomes	Income of the loan applicant.
House	Home ownership status (owned, mortgaged, rented).
Emp length	Years of employment in years.
Schematic diagram	Purpose of the loan (e.g., education, home improvement).
Sum of money	Amount of loan applied for.
Interest rates	Loan rates.
State of affairs	Loan approval status (paid in full, charged off, current).
Percent income	Loan amount as a percentage of income.
Defaults	Whether the applicant has previously defaulted on a loan (yes, no).
Cred length	The length of the applicant's credit history.

3. Data statistics

The data were statistically summarised by calculating the maximum, minimum, mean, standard deviation and median for age, income, home ownership status, years of employment in the unit, purpose of the loan, loan amount, interest rate of the loan, loan as a percentage of income and length of the credit history for all the data and the results are shown in Table 2.

Table 2: Data statistics.

Variable name	Maximum values	Minimum value	Average value	Standard deviation	Median
Age	144	20	27.727	6.31	26
Income	6000000	4000	66649.372	62356.447	55956
Home	4	1	1.911	0.962	1
Emp length	123	0	4.789	4.155	4
Intent	6	1	3.336	1.68	3
Amount	35000	500	9656.493	6329.683	8000
Rate	23.22	5.42	11.04	3.229	10.99
Status	1	0	0.217	0.412	0
Percent income	0.83	0	0.169	0.106	0.15
Cred length	30	2	5.794	4.038	4

4. Relevance analysis

Pearson correlation analysis is a statistical method used to measure the degree of linear correlation between two variables, with results between -1 and 1. When the correlation coefficient is 1, it means that the two variables are completely positively correlated; when the correlation coefficient is -1, it means that the two variables are completely negatively correlated; when the correlation coefficient is close to 0, it means that there is no linear relationship between the two variables [7]. In this paper, Pearson correlation analysis is used to explore the degree of correlation between customers' indicators and whether they eventually default on their loans. Through this analysis method, we can better understand the degree of influence of different factors on whether a customer defaults on a loan, so as to formulate a targeted risk control strategy or adjust the approval criteria. The correlation heat map is shown in Figure 1.

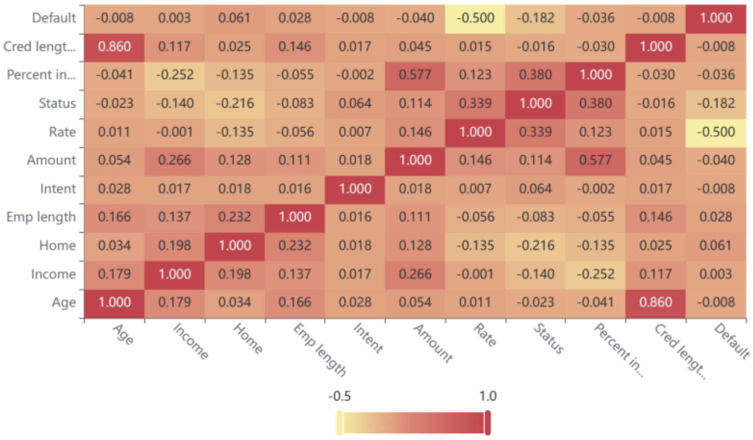

Figure 1: The correlation heat map.
(Photo credit: Original)

From the correlation heat map, it can be seen that the client's income, homeownership status, and length of employment in the organisation have the same and positive (no loan default) direction of influence on whether or not the loan is ultimately defaulted, whereas the client's age, purpose of the loan, amount of the loan, interest rate of the loan, loan as a percentage of income, length of credit history, and status of loan approval have the same and negative (delinquent loan) effect.

5. Method

5.1. Improved Sparrow Search Algorithm

The improved sparrow search algorithm is a heuristic optimization algorithm based on the behavioural characteristics of birds in nature, which improves the search efficiency and convergence by introducing improvement points such as chaotic mapping, dynamic adaptive weighting, inverse learning and Cauchy's variation [8]. Firstly, the use of chaotic mapping for population initialisation makes the initial population more dispersed and diverse; secondly, the dynamic adaptive weighting mechanism is able to adjust the weights according to the performance in the search process to improve the global search ability of the algorithm; finally, the selection and updating strategy of the optimal sparrow is improved by reverse learning and Cauchy's Variation, which increases the local search ability of the algorithm and avoids the risk of falling into the local optimal solution [9].

5.2. Random Forest

Random Forest Classification Model is a machine learning algorithm based on decision tree integration that performs well in dealing with classification problems.

However, the traditional random forest model has some difficulties in parameter tuning, so the introduction of an improved sparrow search algorithm to optimise the random forest classification model can improve its performance [10]. By applying the improved sparrow search algorithm to the random forest model, key parameters such as the number of decision trees, tree depth, feature subset size, etc. can be adjusted more efficiently, so as to enhance the model's accuracy and generalisation ability on the classification task.

In practice, the improved sparrow search algorithm is first combined with the random forest classification model and the relevant parameters are set. Then, the sparrow search algorithm is used to continuously adjust the parameters of the random forest model during the training process, and the performance of the model on the validation set is evaluated after each iteration. Finally, the model with the best performance on the validation set is selected as the final result, and the prediction of the test set is verified.

6. Experiments and Results

6.1. Experimental setup

In terms of data division, firstly, the dataset is divided into training set and test set, and the dataset is divided according to the ratio of 7:3, which is used for training and evaluation of the model respectively; secondly, the data is normalised, and the input data is normalised to map the data into the range of 0-1, which is conducive to the convergence speed and stability of the model. Then set the parameters of the optimisation algorithm, including the number of populations, the maximum number of iterations, the dimension of hyperparameters, etc., and call the SSA (Salp Swarm Algorithm) algorithm to carry out hyperparameter optimisation to find the best combination of hyperparameters. Afterwards, the optimal parameters are extracted, the best hyperparameter values obtained after optimisation are obtained for subsequent model construction, the classification model is constructed using the Random Forest algorithm, and the feature importance information is obtained, and the prediction results are obtained for the training set and test set. Finally, the comparative relationship between the real and predicted values of the training and test sets is shown, and the accuracy information is displayed to show the confusion matrix of the model on the training and test sets, which helps to further evaluate the model performance.

The training process of the model is shown in Fig. 2, and the model went through a total of 15 iterations.

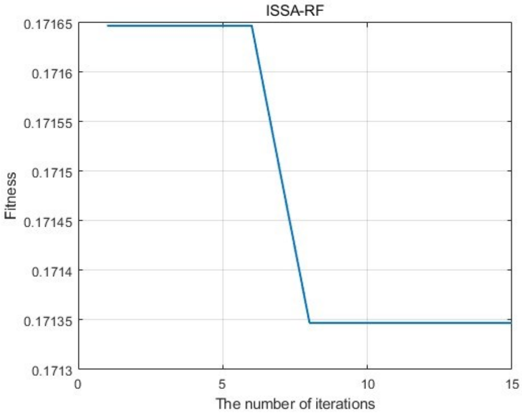

Figure 2: Random Forest Architecture.
(Photo credit: Original)

The confusion matrices for the training and test sets, which hold the results of model training and test predictions, are outputted as shown in Figures 3 and 4, respectively.

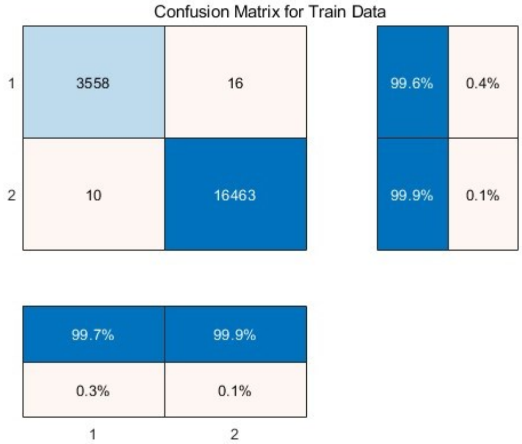

Figure 3: Confusion matrix.
(Photo credit: Original)

From the confusion matrix of the training set, it is clear that the algorithm used in this paper successfully predicted 20,021 customers' delinquencies and only 26 customers failed to be predicted, of which 10 should have been predicted as delinquent but were predicted as not delinquent and 16 should have been predicted as not delinquent but were predicted to be delinquent, and the model's prediction accuracy in the training set was 99.87 per cent.

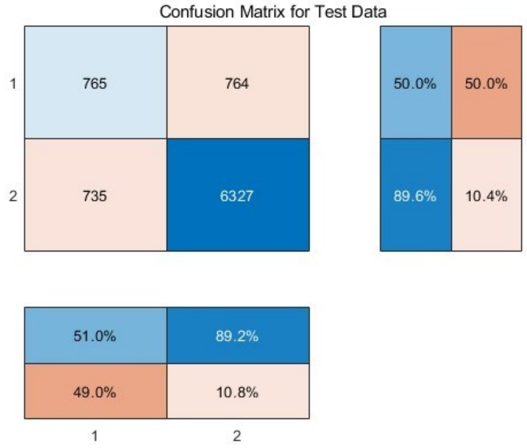

Figure 4: Confusion matrix.
(Photo credit: Original)

From the confusion matrix of the test set, it can be seen that the algorithm used in this paper successfully predicted delinquency for 7092 customers and failed to predict 1499 customers, of which 735 should have been predicted as delinquent but were predicted as non-delinquent, and 764 should have been predicted as non-delinquent but were predicted as delinquent, and the model's prediction accuracy in the test set was 82.55%.

7. Conclusion

In this paper, the random forest model is optimised by sparrow search algorithm to further improve the accuracy of credit risk assessment. Pearson correlation analysis reveals that customers' income, home ownership status, and years of employment in the unit are positively associated with whether they eventually default on their loans, while customers' age, loan purpose, and loan amount are negatively associated with delinquency. After 15 iterations, the model on the training set successfully predicted 20,021 customer delinquencies with only 26 prediction errors and 99.87% accuracy. On the test set 7092 customers were predicted with an accuracy of 82.55%, of which 1499 were predicted incorrectly.

In summary, this paper has achieved significant results in credit risk assessment by optimising the algorithm and random forest model, improving the prediction accuracy and effectively identifying potential delinquency risk customers. These results provide an important reference for financial institutions, which helps to formulate risk control strategies and safeguard financial security more accurately.

References

[1] *Al-Qudah, Anas Ali, et al. "The impact of green lending on credit risk: Evidence from UAE's banks." Environmental Science and Pollution Research 30.22 (2023): 61381-61393.*

[2] *Gilchrist, Simon, et al. "The Fed takes on corporate credit risk: An analysis of the efficacy of the SMCCF." Journal of Monetary Economics (2024): 103573.*

[3] *Wang, Liukai, et al. "Forecasting SMEs' credit risk in supply chain finance with a sampling strategy based on machine learning techniques." Annals of Operations Research 331.1 (2023): 1-33.*

[4] *Duho, King Carl Tornam, Divine Mensah Duho, and Joseph Ato Forson. "Impact of income diversification strategy on credit risk and market risk among microfinance institutions." Journal of Economic and Administrative Sciences 39.2 (2023): 523-546.*

[5] *Ali, Mohsin, Mudeer Ahmed Khattak, and Nafis Alam. "Credit risk in dual banking systems: does competition matter? Empirical evidence." International Journal of Emerging Markets 18.4 (2023): 822-844.*

[6] *Mahbobi, Mohammad, Salman Kimiagari, and Marriappan Vasudevan. "Credit risk classification: an integrated predictive accuracy algorithm using artificial and deep neural networks." Annals of Operations Research 330.1 (2023): 609-637.*

[7] *Rudin, Cynthia, and Yaron Shaposhnik. "Globally-consistent rule-based summary-explanations for machine learning models: application to credit-risk evaluation." Journal of Machine Learning Research 24.16 (2023): 1-44.*

[8] *Shi, Si, et al. "Machine learning-driven credit risk: a systemic review." Neural Computing and Applications 34.17 (2022): 14327-14339.*

[9] *Bussmann, Niklas, et al. "Explainable machine learning in credit risk management." Computational Economics 57.1 (2021): 203-216.*

[10] *Gilchrist, Simon, et al. "The Fed takes on corporate credit risk: An analysis of the efficacy of the SMCCF." Journal of Monetary Economics (2024): 103573.*