Customer Credit Rating by Machine Learning

Chengyijing Wang[1, *], Haining Jiang[1], Xiaoyan Jin[2], Ziyu Zhou[3]

[1]Department of Social Science, University of California Irvine, Irvine, United Stated

[2]Department of Social Audit, Nanjing Audit University, Nanjing, China

[3]Department of Computer Science and Engineering, University of New South Wales, Kensington, Australia

*Corresponding author: chengw24@uci.edu

These authors contributed equally.

Abstract. Recently, people's consumption attitudes have also changed, being inclined to spend in advance. Banks and other financial institutions use credit rating models as a tool to evaluate the credit score of individuals, determine whether to grant the loan to the applicant. One of the biggest challenges for the banking industry in assessing the customers' credit is that it is unlikely to provide a manual review to classify them because of the huge volume of data on applicants. Therefore, it is necessary to establish a suitable and effective credit rating model to help banks evaluate the quality of applicants. This paper focuses on the problems existing in the development of personal credit rating system and tries to find the best solution in the field of personal credit rating system. By selecting independent variables that are highly correlated with delinquency behavior, using different models for testing, and comparing the results of the models, this paper finally draws the conclusion that different algorithms combined by the group decision method can make better decisions.

Keywords: Credit Card, Credit Rating, Machine Learning, Customer credit

1. Introduction

With the gradual establishment of the market economy system in the world, the credit rating system has covered all market economic activities at three levels: capital market, commercial market, and individual consumers. While the activity of personal consumer credit business has increased significantly, the personal credit risk brought by it has also been decentralized, universal, and non-systematic, which has caused great trouble to financial institutions. Therefore, there is a growing demand for accurate and effective personal credit assessment models from both the state and financial institutions. Credit rating can effectively and quickly identify and classify the quality of customers, and objectively quantify the size of the risk, which is also the standard risk control method in the financial industry. The application credit rating is a model that assesses the applicant's credit standing and ability to repay loans on time. Credit card applicants are rated using machine learning based on the personal information submitted by customers and their credit standing in the data set.

Nowadays, scholars at home and abroad have carried out a lot of research on the problem of credit evaluation, mainly focusing on the study of model methods, hybrid machine learning algorithms [1-4], and a genetic algorithm to optimize the model [5-7]. For example, Estran et al. have only focused on genetic algorithms [8], although the related works have achieved results in optimizing purely expert-based developing credit rating models. And Tsai and Chen's "classification + classification" hybrid model based on logistic regression and neural network provides the highest prediction accuracy and maximizes the revenue model [9]. In this paper, based on them, we adopted the group decision-making algorithm, combined three single algorithms, namely Logistics Regression, Decision Tree, and Neutral Network, and adopted the group voting method to improve the accuracy of the model. The contribution of this paper is to find out which combined methods and techniques of mixing different algorithmic models are most suitable for credit rating to help banks assess the quality of applicants.

This paper is organized as follows. In Section 2, the working principles of the three single models and the advantages of group decision-making algorithm are explained, and briefly introduce the

experimental process based on group decision-making algorithm. Subsequently, our experimental process, that is, the conclusion of comparing group decision-making algorithm with single models and external models, will be elaborated in Section 3. Finally, the effectiveness and superiority of group decision-making algorithm are verified through comparative experiments, which will be presented in the conclusion of Section 4 as an effective method to improve the accuracy of the model.

2. Methodology

2.1 Algorithms

2.1.1 Logistic regression

The logistic regression method is a data analysis method that predicts binary outcomes from previously observed data sets. This model predicts the dependent variable by analyzing the relationship between the independent variables and is a crucial tool in machine learning. It can classify the input data based on historical data. When the amount of relevant data increases, the predictive power of the method will become stronger and stronger [10].

Logistic regression uses a more complex cost function relative to linear regression. The cost function of logistic regression can be called a “sigmoid function” or a “logistic function”, as seen in figure 1.

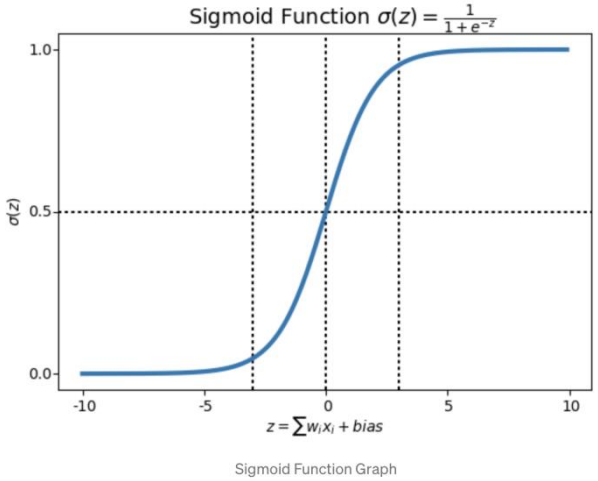

Figure 1. Sigmoid Function Graph

Here is the function for logistic regression:

$$\text{Sig(x)} = \frac{1}{1+e\text{^}(-x)} \tag{1}$$

E is the log base

X is the value to be converted.

Logistic regression assumptions:

The logarithm of the independent variable and the output variable is linear.

Non-collinearity among independent variables. That is, the independent variables are independent of each other.

The output variable is binary.

The cost function of the logistic regression method is defined as:

$$Cost(h_\theta(x), y) = -log\ if\ y = 1 \tag{2}$$

$$Cost(h_\theta(x), y) = -log\big(1 - h_\theta(x)\big)\, if\ y = 0. \tag{3}$$

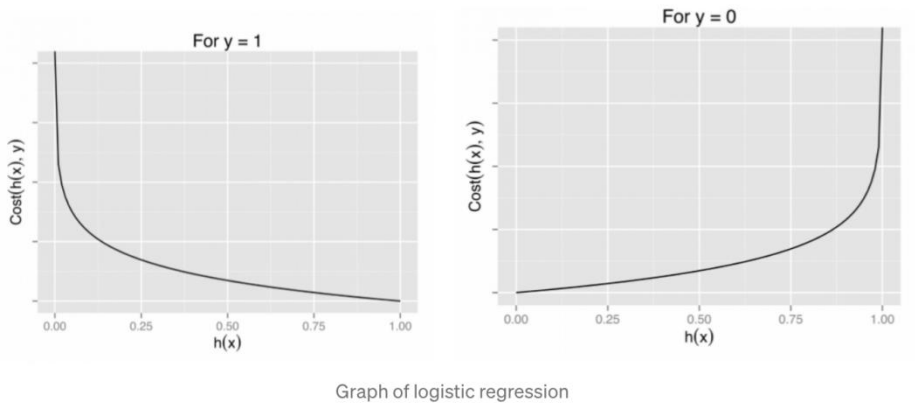

Figure 2. Logistic regression graph

As shown in figure 2, after compressing the two functions, the final function is:

$$J(\theta) = -\frac{1}{m}\sum[y^{(i)} log(h_\theta(x(i))) + (1 - y^{(i)}) \log (1 - h_\theta(x(i)))] \tag{4}$$

2.1.2 Decision tree

The decision tree method is a data mining method commonly used to build a classification system according to multiple variables and a prediction algorithm that mines the target variable. This method divides the population into branching segments and forms an inverted tree with root, inner and leaf nodes. The algorithm can handle extremely complicated and enormous amount of data sets. Research data can be divided into training datasets and testing datasets if the sample size is large enough. And this makes it possible to determine the appropriate tree size for the best and final model, figure 3 shows the structure of decision tree.

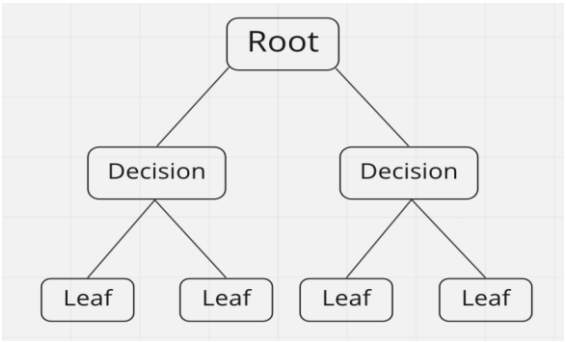

Figure 3. Graph of Decision Tree

2.1.3 Neural Network

The neural network is an algorithm that develops latent relationships in a set of data by simulating the working methods of the human brain. The method recognizes and adapts to changing inputs. Therefore, the neural network method can produce the best results without re-outputting the criteria.

The main goal of this model is to learn by automatically modifying itself to be able to perform complex tasks that traditional rule-based programming cannot. The functional scope of neural networks is very broad, and due to their operation, they are able to approximate any existing function with sufficient training. Figure 4 shows how the simple neural network works:

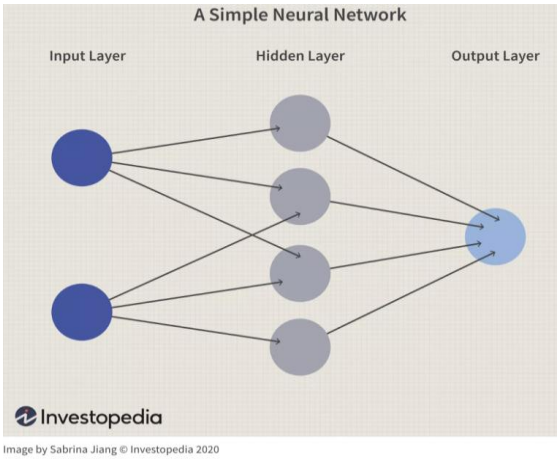

Figure 4. Graph of A Simple Neural Network

2.2 Group decision-making method

Group decision-making describes a situation when individuals collectively make a choice among options before them. The decision then is no longer upon to any single individual who is in the group but will be determined on a group consensus. Thus, some personal biases can be effectively avoided, and a more objective decision will be made. When it comes to decisions involving personal judgements such as whether to approve the credit loans, group decision-making can lead to a more objective outcome.

So, in our experiment, the group decision making method was applied. We combined the above three single algorithms together and made a majority voting to get the final result. Since there already have three predictive outcomes, either is good or bad, the most likely outcome would be the final prediction. By applying such method, the biases and noise brought by each single algorithm can be effectively reduced to a lower level. As each algorithm is of great difference, majority voting helps take advantage of each and consider the overall aspects of the problem.

2.3 Data pre-processing

Figure 5 is a flowchart of how our above methods can be applied to credit card rating.

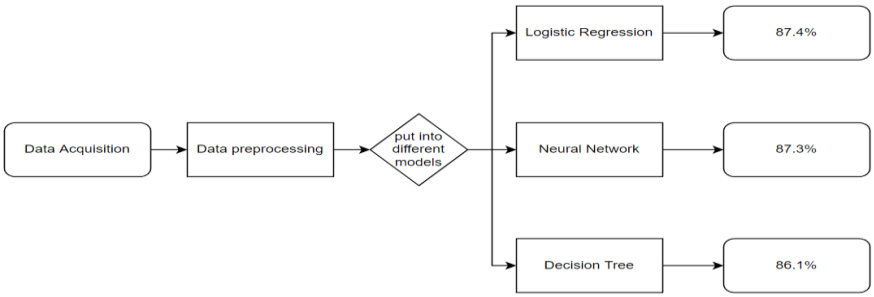

Figure 5. Flowchart of Three Single Algorithms and Results

During the data acquisition, we split the target and features firstly, but we found there are several feature columns are not useful and they will make results over-fitting, so we dropped them and acquired seventeen features. Then, when we split the train dataset and the test dataset, we discovered that the feature, FLAG_MOBILE, has only one value for all customers and some features, FLAG_WORK_PHONE, FLAG_PHONE and FLAG_EMAIL, only have a small effect to determine good or bad, so we dropped these features in our dataset.

For data processing, DAYS_BIRTH is converted to age column and DAYS_EMPLOYED is converted to work years, since the number of days is not easier to read and analyze than the number of years, but the year is more intuitively reflect the age and the number of working years. Also, the

children count feature is dropped, because children count is the subset of family size, so it's not very useful either. To deal with the missing values, we noticed that OCCUPATION_TYPE feature, which is also a categorical data, has missing values in the dataset, so the missing part is filled with the values that are most frequent in the column by using a simple imputer. Finally, to deal with other categorical parts, the categorical features is investigated, and we found OCCUPATIOB_TYPE feature has a high cardinality and the rest features have a low cardinality. To solve this problem, the Target Encoding method is implemented for the OCCUPATION_TYPE, and one hot encoding is used for the rest features to get a better prediction. Target encoding is a process of replacing a categorical value with the mean of the target covariate. Therefore, target encoding is used to transform categorical columns into numeric. Also, different features are in different ranges, so standardization is a very good way to solve this problem. The same range values are achieved after standardization. For example, the AMT_INCME_TOTAL feature has changed from large numbers to decimals. To conduct the low cardinality categorical features, one-hot encoding is used. Each categorical value was transformed into a new categorical column, either 1 or 0. In the end, we acquired 32 column features. Also, the test dataset is dealt with the same thing as train dataset, like dealing with the missing values, categorical values, standardization and so on.

3. Experiment

3.1 Dataset

The dataset is from kaggle, which is a credit card dataset used for machine learning, containing two different tables, named application_record.csv and credit_record.csv. The application record form contains general information about the applicant, such as gender, date of entry, annual income, etc., and has a total of 18 variables: 12 categorical variables, 5 numerical variables, and the last one is the applicant ID, which we use to accommodate two data sets, and these variables can be used as predictive characteristics. The credit record table contains the applicant's monthly transaction history and loan payment records (days past due), which contains the user's credit card behavior. The details can be explained in the following forms of remarks.

3.2 Experimental design

In order to verify the method of group decision-making, we try to compare it from two dimensions, that is within the voting group or out of it. On one hand, we compare the accuracy of group decision-making with the accuracy of three single models to demonstrate the effectiveness of majority voting method. On the other hand, we tried to validate the objectiveness and superiority of the group decision making model and compared the accuracy with external models, such as random Forest and light GBM.

3.2.1 The multiple perspectives approach

(1) Comparison between group decision making and single models:

In the previous model introduction, we have mentioned the advantages and calculation methods of three single models, logistics regression, decision tree and neural network. In this experiment, a single model is used to predict and calculate the customer category.

Table 1. Accuracy and F1 score for group decision making and single models

Model	Accuracy	F1 score
Logistic regression	87.37%	93.25%
Decision tree	86.26%	92.20%
Neural network	87.23%	93.09%
Majority voting	87.69%	93.31%

As can be told from table 1, the accuracy of logistics regression, decision tree and neural network are 87.40%, 87.3% and 86.30%, respectively. Then, we use the groups voting formed by these three models to make group decisions, and the results are shown in the figure above. It can be seen that the Accuracy of group voting is improved by 0.3%, 0.4% and 1.4% respectively compared with the three single models.

(2) Comparison between group decision making and external models:

There are many other models for comparing group decision-making, here we selected the Random Forest model and the lightGBM model as representatives. The advantage of random forests is that they can produce good predictions that are easy to understand for performing regression and classification tasks, and can effectively handle large data sets. LightGBM is a gradient enhancement framework based on the decision tree, which improves the efficiency of the model and reduces memory usage.

Table 2. Accuracy and F1 score for group decision making and external models

Model	Accuracy	F1 score
Light GBM	87.40%	93.27%
Random forest	87.37%	93.26%
Majority voting	87.69%	93.31%

However, the result in table 2, shows that under the same data set and preprocessing progress, the accuracy of the Random Forest model and lightGBM model is 87.4% and 87.4% respectively, while the accuracy of group decision-making using logistics regression, decision tree, and neural network is 87.7%, which is 0.3% higher than the accuracy of the other two models. Therefore, it can be seen intuitively that the group decision model adopted by us has strong advantages and is superior to other single models in the objective level of accuracy.

3.2.2 Performance metrics

To measure the performance of each classification algorithms, both accuracy and F1 score are used to compare the outcomes.

Accuracy is the ratio of true positives and true negatives to all positive and negative observations, which means it can show how often machine learning model used will efficiently predict an outcome out of the total number of prediction times. It can be expressed mathematically:

$$Accuracy\ Score\ =\ (TP\ +\ TN)/\ (TP\ +\ FN\ +\ TN\ +\ FP) \tag{5}$$

However, for accuracy does not take into account how the data is distributed, we use F1-score as supplementary. It is a harmonic mean of precision and recall score and provides accurate results for both balanced and imbalanced data, considering both the precision and recall ability of the model. Mathematically, it is expressed as follows:

$$F1\ Score\ =\ 2\ *\ (Precision\ *\ Recall)\ /\ (Precision\ +\ Recall) \tag{6}$$

$$Precision\ =\ TP\ /\ TP\ +\ FP \tag{7}$$

$$Recall\ =\ TP\ /\ TP\ +\ FN \tag{8}$$

3.2.3 Experiment settings

(1) Parameter setting of the model

After adjusting the parameters to get the optimal performance of each model, we set them as follows.

In the model of decision tree, max depth was set equal to 30, with min_samples_split equal to 5.

In the model of neural network, lbfgs was chosen as solver, learning rate being adaptive, while alpha was 0.001, hidden layer sizes being 20, max iteration 2000.

In random forest, we set max depth to 12， n_estimators to 250, and min samples leaf as 16.

In light GBM, learning rate was equal to 0.02, max depth being 8, and n_estimators was 250.

(2) Data volume division

We use 80% of the total dataset as our training set, the others as testing set.

(3) Mechanism of majority voting

According to the predictions of logistic regression, decision tree, and neural network, we made a vote between them, that was to pick the most likely outcome of each applicant as the final outcome. By this group decision method, the disadvantages of different algorithms can be complemented mutually, and the noise and biases be reduced to a very low level.

In addition to logistic regression, decision tree and neural network in the voting group mentioned above, another two common models out of the voting group, which are often considered as powerful tools to deal with classification problems, are applied as comparisons. One is random forest, an extension of the CART model that combines the output of multiple decision trees to achieve a single result. Often by increasing the bias of the forest and decreasing the variance slightly, it encourages an overall better model. Another is light GBM, which is a gradient boosting method involving two new techniques, Gradient-based One Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).

By comparing models internal and external to the group, the majority voting method is verified under two dimensions.

3.3 Experimental results and analysis

The data after preprocessing will be put into different single model to train, and the results are 87.4%, 87.3% and 86.1% respectively. After the single model learning, we decided to use the group decision-making to improve the accuracy of the results. Figure 6 shows the workflow of this group decision-making is similar to the process of single model described above. The only difference is that the result can consider all the models' results in the group to get the final better result. In the end, the accuracy is improved by 0.3%, reaching 87.7% by implementing majority voting. With regard to F1 score, majority voting method also performs better than the other single model. When dealing with very large data set, the majority voting method would help banks save a huge amount of cost.

Below are the results of the experiment, shown in figure 7 and figure 8.

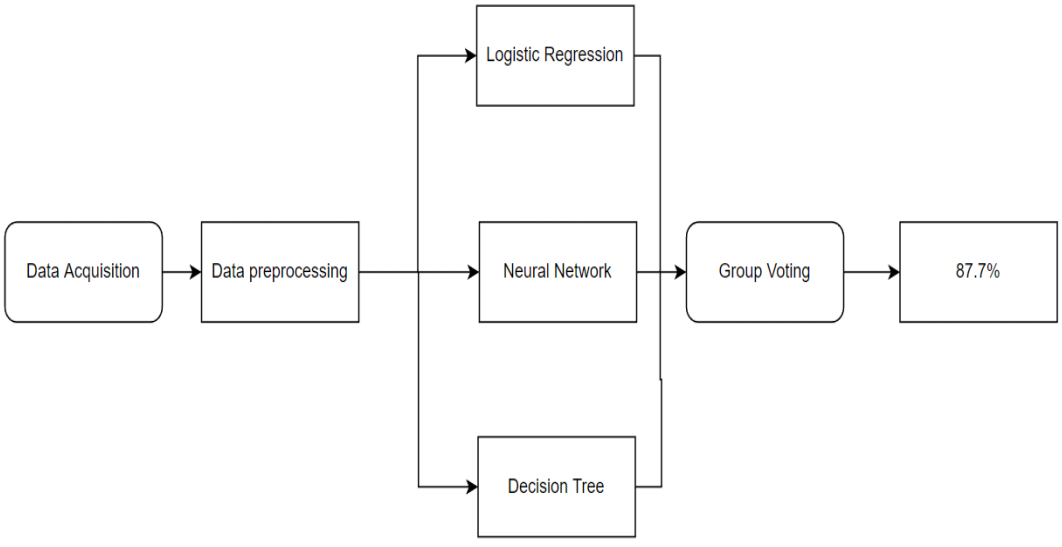

Figure 6. Flowchart of Decision-making Method and Results

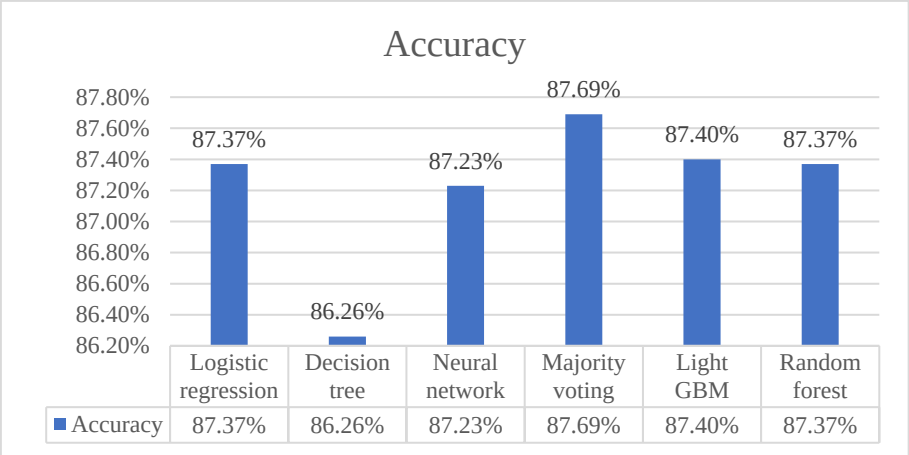

Figure 7. Accuracy for all of single models and external models

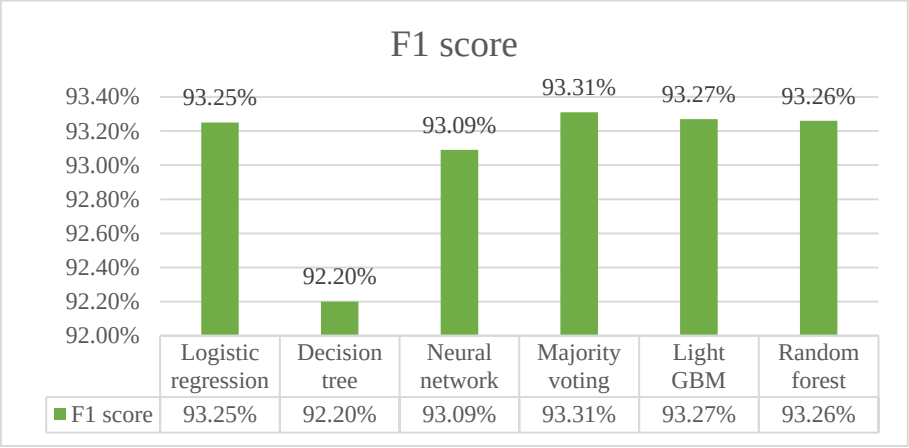

Figure 8. F1 score for all of single models and external models

4. Conclusion

This paper mainly focuses on customer credit rating prediction, since customers spending with credit cards represents a commitment that they will reimburse the credit card provider for the money they owe at some point in the future. And more importantly, the damage of the lack of credit rating will be seen from a concrete level, the bank will suffer huge losses, enterprises or individuals will break their trust, and from a broader level, serious financial risks will occur. Therefore, the importance and significance of credit rating in today's society is indispensable.

In the experiment, we used machine learning algorithms to classify the credit status of credit card applicants by personal information and data submitted by credit card applicants to predict future default and credit card borrowing, and set borrowing limits and repayment rules for users according to their credit scores. In addition, the three models adopted in this paper are logistics regression, decision tree and neural network. To improve the accuracy and verify the effectiveness and superiority of our group decision-making model, we propose a group voting method to get the accuracy of the ensemble algorithm of three single model. It can be seen intuitively that the accuracy of credit card prediction is improved by 0.3-1.4 (for different single models) through group decision-making method.

It is precisely because of the importance of credit rating for commercial banks to judge the degree of loan risk and its ability to serve as the main basis for risk management of credit assets, the results obtained in this study can be used by commercial banks to predict future loan defaults and help banks decide whether to issue a credit card to the applicant. After all, bank credit is also one of the important sources of funds for its production and development, and a superior and effective group decision-making model can effectively predict the quality and behavior norms of the production and operation

activities of individuals and companies, which will directly affect the use and efficiency of bank credit funds.

References

[1] Kui Wang, Meixuan Li, Jingyi Cheng, Xiaomeng Zhou, Gang Li, Research on personal credit risk evaluation based on XGBoost[C], Volume 199, 2022, Pages 1128-1135, ISSN 1877-0509,.

[2] Yunke Cheng, Research on Credit Strategy Based on XGBoost Algorithm and Optimization Problem[D], China: School of Electronic Information of Wuhan University, 2021

[3] Javadpour, A., Saedifar, K., Wang, G. et al. Improving the Efficiency of Customer's Credit Rating with Machine Learning in Big Data Cloud Computing[C]. Wireless Pers Commun 121, 2699–2718 (2021).

[4] Li, X., Sun, Y. Application of RBF neural network optimal segmentation algorithm in credit rating[D]. Neural Comput & Applic 33, 8227–8235 (2021).

[5] Baisong Li, Huiyu Li, Mengmeng Gong, Establishment of a Mathematical Model for Enterprise Credit Risk Recognition and Rating Based on Hybrid Learning Algorithms[C], IOP Conference Series: Materials Science and Engineering, Volume 563, Issue 5

[6] Aurora Y.MU, A Hybrid Machine Learning Model with Cost-function Based Outlier Removal and Its Application on Credit Rating[C], J ournal of Physics: Conference Series, doi:10.1088/1742-6596/1584/1/012001

[7] KIROLOS ATEF, Credit Card Approval Prediction without vintage, https://www.kaggle.com/code/kirolosatef/credit-card-approval-prediction-without-vintage/notebook

[8] Remy Estran, Antoine Souchaud, David Abitbol, Using a genetic algorithm to optimize an expert credit rating model[C], Volume 203, 2022, 117506, ISSN 0957-4174,.

[9] Chih-Fong Tsai, Ming-Lun Chen, Credit rating by hybrid machine learning techniques[C], Volume 10, Issue 2, 2010, Pages 374-380, ISSN 1568-4946.

[10] Liu Z , Tian H , Wang H . Staff's Perception of Credit Bank System Based on Logistics Regressions Regression Model: A Survey of Faculty Members at Model: at 11 Municipal RTVUs in Shanxi Province, Northwest China[J]. Distance Education in China, 2014.