Advances in Economics, Business and Management Research, volume 203
Proceedings of the 2021 3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021)

Business Analysis in Modeling of Financial Risk

Dirun Zhang[1,*,a,†], Xiangyi Shan[2,*,b,†],Siqi Li [3,*,c,†]

[1]*Capital University of Economics and Business*
[2]*Dongbei University of Finance and Economics*
[3]*Beijing University of Posts and Telecommunications*
[*]*Corresponding author. Email: [a]32018020116@cueb.edu.cn, [b]2018211092@163.com, [c]2019213667@bupt.edu.cn*
[†] *These authors contributed equally.*

ABSTRACT
The bank's main source of profit is loans, the money lent out by charging interest to make a profit, but with great risk of not being able to recover. In economic globalization, especially in the context of financial internationalization, loan risk control is always an important research topic for banks. In this paper, for the data set of bank loan risk control, we use four statistical analysis models to predict loan defaults: Logistic Regression, Random Forest, Neural Network, and XGB, respectively. Draw ROC curves of the four models and cluster the users. Meanwhile, the AUC values of the four models were calculated. Through simple comparison and analysis, we select the optimal method and probe into the effect and importance of the coefficients.

Keywords: *clustering, RMF model, multicollinearity.*

1. INTRODUCTION

We can use Microsoft SQL Server to hierarchy users with data, a relational database management system. Database Management System (DBMS) organizes data according to a certain data model and manages the Database. The Database application system operates the Database through the interface provided by the DBMS, and the Database Administrator (DBA) manages and operates the Database through the interface provided by the DBMS. The Database Model refers to the data storage structure in the database management system. The database management system stores and manages the data according to the data model. The more common user hierarchy model is the RMF model. RFM is an acronym combination of three indicators, namely, Rencency (the last consumption), Frequency (consumption Frequency), and Monetary (consumption amount). It is an important tool to measure the current user value and stratify users. Rencency means the last consumption refers to the time interval between the last consumption on the platform and the current one. Theoretically, the smaller R is, the more valuable the customer is. Consumption Frequency refers to the number of purchases made by customers within a fixed period. Monetary means the amount of money consumed over some time.

The above is the method of user stratification using data analysis. However, the so-called user stratification is based on the user value (e.g., active users, high-value users) for the center of the segmentation. In the same stratification model, a user will only be in one level. Another term is user clustering based on user attributes (a certain kind of label on the user, for example, the user who likes to read books on the subway). A user may have multiple attributes at the same time. Here are some technical terms to explain:

1.1. Structured data

Structured data, also known as raw data, is logically expressed and realized by a two-dimensional table structure. It is composed of clearly defined data types and strictly follows the data format and length specifications. It is mainly stored and managed through relational databases.

1.2. Big data

The data set that cannot be captured, managed, and processed by conventional software tools within a certain time frame is a massive, high-growth, and diversified information asset that requires new processing modes to have stronger decision-making power, insight discovery power, and process optimization ability.

1.3. Internet finance

a new financial business model in which traditional financial institutions and Internet enterprises utilize Internet technology and information to realize capital financing, payment, investment, and information intermediary services. Internet financial risk information sharing system is a 7*24 hours continuous operation business system that adopts star network structure to connect with access institutions.

1.4. Machine learning

The study of how a computer simulates or implements human learning behavior to acquire new knowledge or skills, allowing a computer to reorganize existing knowledge structures to improve its performance.

1.5Block chain

A new application model of distributed data storage, peer-to-peer transmission, consensus mechanism, encryption algorithms, and other computer technologies.

1.6. Cloud computing

The growth, usage, and delivery of Internet-based related services usually involve providing dynamically scalable virtualized resources over the Internet.

BA (Business Analytics) is based on Business knowledge, mathematical programming as a means, starting from data analysis, creating value by decision optimization, and achieving the Business application of Big Data. Its significance and value created are as follows:

1.7. The significance of business analysis

1.7.1. Quantification

Describe various indicators with clear and explicit data can make the case intuitive and easy to understand. In addition to recording directly, secondary processing based on data can generate more valuable information; multiple data processing can constitute a commercial index system.

1.7.2. Judgment

With enough data, it is easy to make reasonable standards and give an objective reference by running all kinds of statistical methods. Then we can decide whether the research content is reasonable.

1.7.3. Evaluation

By calculating, integrating, comparing, and analyzing, we can improve the confidence and accuracy of the data. After that, we may explore the main causes and countermeasures of the problem by study.

1.7.4Prediction

We are basing history data to record failures and summarize success while predicting future trends and implications. It can also help enterprises adjust plans and make decisions in time. Through analysis, we are likely to put forward goal setting and other helpful advice.

1.8. Value created by business analysis

Segment customers and customize special services for each group. Simulate the real environment to discover new demands and improve the return on investments. Strengthen departmental cooperation and improve the efficiency of the entire management chain and industrial chain. Reduce service costs; discover hidden clues to innovate products and services [1].

2. MATH AND EQUATIONS

Clustering is grouping collections of physical or abstract objects into classes made up of similar objects. Clustering is a process of classifying data into different classes or clusters, so objects in the same cluster have great similarities, and objects between different clusters have great similarities. The goal of clustering analysis is to collect data based on the similarity between data, often used for user segmentation, characterization of user groups, fraud detection, etc. Here are four steps for clustering:

2.1. Select the appropriate clustering algorithm

The key to choosing a clustering algorithm is to look at the amount of data. There are six commonly used clustering algorithms: K-Means clustering, Mean drift clustering, DBSCAN, Max Expectation (EM) clustering with Gauss hybrid model (GMM), Condensed hierarchical clustering, and Graph Detection Community [2]. When the amount of data is very large, you can prioritize K mean clustering and get preliminary results. If the effect is not good, then build more small samples by random sampling method, manual fusion model to improve clustering results, and optimize the model.

2.2. The selection of variables in clustering analysis

Before deciding on variables, we must first clarify the purpose of our analysis. For example, we now have a lot of customer purchase record information and customer personal information. We want to layer the user's purchase situation, and the most analytical value

of the data should be purchase information rather than personal information. The clustering results only for purchase-related information are entirely purchase-driven and should not allow irrelevant information to affect the final clustering results.

2.3. The importance of analyzing variables

Because variable selection is highly subjective, it is difficult to judge the importance of variables. Here are only two ways to think: Consider the correlation between the intrinsic variation of variables and variables. The importance of variables is sorted directly using out-of-the-box algorithms

2.4. Prove that the results of clustering make sense and determine the number of clusters

Since clustering is unsupported learning, there are no specific criteria to determine whether clustering results are correct, generally in three ways:

2.4.1. Human verification of clustering results

Using business logic to explain clustering results, if the results are broadly in line with the views of industry experts, it means that the results are meaningful and can be returned to the real business logic.

2.4.2. Pre-set the criteria for judging

Use some pre-defined functions to make a judgment.

2.4.3. Visualization to prove differences between clusters

There should be some difference between different clusters after visualization, not a cluttered interweaving.

Similarly, determining the number of clusters can be used in these three ways [3].

The main work of exploratory data analysis is to clean the data, describe the data (statistics, chart), check the data distribution, compare the relationship between the data, cultivate the intuition of the data, summarize the data, etc. Exploratory data analysis usually involves the following steps: examining the data, describing the data using descriptive statistics and graphs, and examining relationships between variables. Data types can be divided into numerical type, category type, text type, time series, etc. It mainly refers to the numerical type and category type, in which numerical type can be divided into the continuous type and discrete type. When data is described, continuous variables, disordered discrete variables, and ordered discrete variables are described. You look at relationships between variables. You look at continuous variables versus continuous variables, discrete variables versus discrete variables, and continuous variables [4].

When the values of some variables in the index data set are missing, the missing values are also called NA (not available) values. Pandas use the floating-point value NaN (Not a Number) to represent missing values in floating-point and non-floating-point numbers, and NAT represents missing values in time series. In addition, Python's built-in value of None will be treated as a missing value. Note that some missing values can also be represented in other forms, such as NULL, 0, or infinity (INF).

2.5. Causes of missing values:

An error occurred during data collection. Problems in the data extraction process. The missing value processing methods for service classification are list-wise deletion of variables with many missing values, single imputation, interpolation, model-based imputation [5].

2.6. Logistic Regression

Logistic Regression shows the probability of an event occurring.

2.6.1. Advantage

Simple to complete, widely used in problems for the industry. The algorithm has low computational complexity, fast speed, and small storage resources in classification. It can easily observe the sample probability score. For logistic regression, the problem of multicollinearity can be solved by combining L2 regularization. The computation cost is low, easy to understand, and complete.

2.6.2. Disadvantage

When the feature space is large, the performance of logistic regression is relatively bad. It is easy to underfit and lack of accuracy under normal circumstance. It cannot handle a large number of multi-feature parameters or variables very well. Only dichotomous problems with linearly separable can be solved. For the characteristic of nonlinearity, it needs to be transformed [6].

2.7. Decision Tree

The decision tree is a machine learning algorithm with simple logic. It is composed of the root node (includes all samples), internal node (corresponding characteristic attribute test), and leaf node (represents the result of a decision). During the prediction, a certain attribute value is used to make the judgment at the internal node of the tree. According to the judgment result, the branch node is decided to enter until it

reaches the leaf node and the classification result is obtained.

2.7.1. Three steps of decision tree learning

2.7.1.1. Selection

It determines which features are used to make the judgment. The function of feature selection is to select out the feature of high correlation with the classification results. The criterion commonly used in feature selection is information gain.

2.7.1.2. Generate

After selecting the features, triggering from the root node to calculate the information gain of all features on the node, the feature with the largest information gain is selected as the node feature. It establishes child nodes according to the different values of the feature. Each child node can generate new child nodes in the same way until the information gain is small or there are no features to choose from.

2.7.1.3. Trim.

The main purpose of pruning is to reduce the risk of overfitting by actively removing some branches.

2.7.2. Advantage

Easy to understand, explain and extract rules, can be analyzed visually. It can process nominal and numerical data at once. It is more suitable for processing samples with missing attributes. Able to handle unrelated features. When testing the data sets, the running speed is relatively fast. It can make feasible and effective results for large data sources in a relatively short period.

2.7.3. Disadvantage

Easy to occur overfitting. Easy to ignore the correlation of attributes in the data sets. Different decision criteria will bring different attribute selection tendencies for the data with different sample sizes when attribute division is carried out in the decision tree. The information gain criterion prefers attributes that have a large number of desirable, while the gain rate criterion prefers attributes with a small number of desirable attributes. When the ID3 algorithm calculates information gain, the result is biased to the feature with more values [7].

2.8. Random Forest

Random forest is composed of many decision trees, and there is no correlation between each other [8]. When the classifications task, we input new samples and ask each decision tree in the forest to make judgment and classification, respectively. Each decision tree will get its own classification result. The decision tree which has the most classifications will be regarded as the final result by the random forest [9].

2.8.1. Four steps to constructing a random forest

Select a sample which size is N, sample N times with replacement, draw one at a time and form n samples finally. These specimens are used for one decision tree as samples at the root code. When each decision tree node needs to be split, select m attributes from M (m□M). Then adopt a strategy to select one attribute from m as the split attribute of the node [10]. Repeat step 2 until it can no longer be split. Note that no pruning was done during the entire formation of the decision tree. Follow steps 1-3 to build many decision trees, and then a random forest is formed.

2.8.2. Advantage

It can produce relatively high dimensional data without dimensionality reduction and feature selection. It can evaluate the importance of different features. It can judge whether the features are interacting with each other [11]. It is not easy to overfit [12]. The training speed is relatively fast, and it is easy to make the parallel method. It is simple to implement. For unbalanced data sets, it can balance the errors. If a significant portion of the feature is missing, it can still maintain accuracy.

2.8.3. Disadvantage

Random forests have been shown to overfit some noisy classification or regression problems. For data with attributes of different values, attributes with more value division will have a greater impact on the random forest, so the attribute weights produced by those random forests on such data are not credible [18].

3. FIGURES AND TABLES

Table1.AUC of 4 methods

Model	Logic	Random forest	nn	Xgb figure
Auc	0.5418326741386	0.7401253923971	0.6128691083455	0.7501342587016

The AUC values of the four methods are shown in the figure [14]. It is not difficult to see that the AUC value of the XGB [15] method is the largest. It is similar to the random forest and larger than Logic and NN, which means that the XGB classification method is more likely to rank the positive sample value before the negative sample value, that is, it can be better classified [16].

Table2. Importance of XGB

Feature	Loan Amnt	term	Interest Rate	Annual Income	Post Code
Importance	0.0595021096480206	0.0493022822702696	0.582705844633004	0.0819652400537261	0.0264056598066911

The table above shows the importance score for each variable in the XGB model.[17] Feature importance indicates how useful or valuable each feature is in enhancing the construction of the decision tree in the model. The higher the score, the greater importance of this variable in the model is. It is not difficult to see that interest Rate has the largest importance value, which is far greater than other variables, which means that the change of interest rate has the greatest impact on this model. And the coefficient of interest Rate is positive. That is, the higher the interest rate, the greater the possibility of default [18]. Although the term is smaller than Interest Rate, it is also much larger than other variables, indicating that term change also has a greater impact on the model. Other variables, loan amnt, post Code, and annual Income, have small importance score values, indicating that these variables have little influence on the model.

Table3. Importance of random forest

Feature	Loan Amnt	term	Interest Rate	Annual Income	postcode
Importance	0.034841717125084	0.2090010185857738	0.4847725266428 1	0.059788811983 4469	0.016406688461443

The above table is the importance score for each variable in the random forest's model. The interest rate's importance value is still much larger than other variables, which means that the change of interest Rate has the greatest impact on the random forest's model. Although the importance value of term is smaller than interest Rate, it is far larger than other variables, indicating that term variable also has a greater impact on the changes of the model. The importance of other variables is relatively small, and their changes have no significant impact on the model [19]. The importance index of this model is the first two largest, so for random forest, interest Rate and term variables greatly influence the model [20].

4. CONCLUSION

Doing a good job of loan risk control will improve the bank's profits and bring stability to the whole financial system, which is of great help to prevent large-scale financial risks. Through the brief analysis of the four models of Logistic Regression, Random Forest, Neural Network and XGB respectively, we can know that if the coefficient of a variable is positive, it is more likely to cause people to break the contract. If the importance score of a variable is higher, the greater the impact it has on default forecasts. According to the relevant conclusions, we can select the variables that are crucial to the loan risk control, build the prediction model, and flexibly use it in the actual business according to the characteristics of customers and the financial environment, thereby reducing the loan default rate effectively.

REFERENCES

[1] Informatiom from https://www.sciencedirect.com/science/article/pii/S0007681314000871?casatoken=Ad4u6xnSaCsAAAAA:o1417kjNd7HZhjobQBqri9GglCzBhnm2b4n6TgAwTXCmKvKRRu-urAoqDmiDROwEtMsfkBEPQ, Business analytics:Why now and what next?

[2] Informatiom from https://blog.csdn.net/weixin_42056745/article/details/101287231?ops_request_misc=%257B%2522request%255Fid%2522%253A%25222161857002216780262538468%2522%252C%2522scm%2522%253A%252220140713.130102334

..%2522%257D&request_id=161857002216780262538468&biz_id=0&utm_medium=distribute.pc_search_result.none-task-blog-2~all~top_positive~default-1-101287231.first_rank_v2_pc_rank_v29&utm_term=%E8%81%9A%E7%B1%BB%E7%AE%97%E6%B3%95, Six common clustering algorithms.

[3] Informatiom from https://www.zhihu.com/question/19982667, How do clustering analysis on users?

[4] Informatiom from https://www.cnblogs.com/HuZihu/p/9641248.html,Missing value processing.

[5] Informatiom from https://easyai.tech/ai-definition/logistic-regression/,One article to understand logistic regression.

[6] Information from https://easyai.tech/ai-definition/decision-tree/,One article to understand decision tree.

[7] Yamada, Y., Suzuki, E., Yokoi, H., & Takabayashi, K. (2003). Decision-tree induction from time-series data based on a standard-example split test. In Proceedings of the 20th international conference on machine learning (ICML-03) (pp. 840-847).

[8] Mingers, J. (1989). An empirical comparison of selection measures for decision-tree induction. Machine learning, 3(4), 319-342.

[9] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., & Cho, H. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4).

[10] Zhou, J., Qiu, Y., Armaghani, D. J., Zhang, W., Li, C., Zhu, S., & Tarinejad, R. (2021). Predicting TBM penetration rate in hard rock condition: a comparative study among six XGB-based metaheuristic techniques. Geoscience Frontiers, 12(3), 101091.

[11] Pathy, A., Meher, S., & Balasubramanian, P. (2020). Predicting algal biochar yield using eXtreme Gradient Boosting (XGB) algorithm of machine learning methods. Algal Research, 50, 102006.

[12] Massaoudi, M., Refaat, S. S., Chihi, I., Trabelsi, M., Oueslati, F. S., & Abu-Rub, H. (2021). A novel stacked generalization ensemble-based hybrid LGBM-XGB-MLP model for Short-Term Load Forecasting. Energy, 214, 118874.

[13] Yan, S., Wu, L., Fan, J., Zhang, F., Zou, Y., & Wu, Y. (2021). A novel hybrid WOA-XGB model for estimating daily reference evapotranspiration using local and external meteorological data: Applications in arid and humid regions of China. Agricultural Water Management, 244, 106594.

[14] Cortes, C., & Mohri, M. (2003). AUC optimization vs. error rate minimization. Advances in neural information processing systems, 16, 313-320.

[15] Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography, 17(2), 145-151.

[16] Ling, C. X., Huang, J., & Zhang, H. (2003, August). AUC: a statistically consistent and more discriminating measure than accuracy. In Ijcai (Vol. 3, pp. 519-524).

[17] Informatiom from https://easyai.tech/ai-definition/random-forest/,One article to understand ramdom forest.

[18] Jiménez-Valverde, A. (2012). Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Global Ecology and Biogeography, 21(4), 498-507.

[19] Harbarth, S., Holeckova, K., Froidevaux, C., Pittet, D., Ricou, B., Grau, G. E., ... & Geneva Sepsis Network. (2001). Diagnostic value of procalcitonin, interleukin-6, and interleukin-8 in critically ill patients admitted with suspected sepsis. American journal of respiratory and critical care medicine, 164(3), 396-402.

[20] Schentag, J. J., Nix, D. E., & Adelman, M. H. (1991). Mathematical examination of dual individualization principles (I): relationships between AUC above MIC and area under the inhibitory curve for cefmenoxime, ciprofloxacin, and tobramycin.