ISSN 2616-5902 Vol. 5, Issue 5: 62-74, DOI: 10.25236/AJBM.2023.050509

Research on Personal Credit Risk Assessment Based on Data Fusion

Tian Wei[1,a], Shuai Li[1,b,*]

[1]School of Management Science and Engineering, Central University of Finance and Economics, Beijing, China
[a]cufe_weitian@outlook.com, [b]cufejwc_ls@cufe.edu.cn
[*]Corresponding author

Abstract: *In recent years, the domestic emphasis on systemic risk prevention has continued to increase. Chinese credit department, which is a crucial department to avoid significant systemic risks, has a particular responsibility to manage and control personal credit risks related to the lives of the public. Therefore, strengthening the management of the individual credit department, especially to guard against personal credit risks, build a data-driven quantitative model for scoring applicants' credit, and improve the personal credit identification and evaluation methods on this basis is a vital part of individual credit risk prevention and systemic financial risk. In this paper, we first study the index system of personal credit risk identification, standard personal credit risk identification models, and some data fusion techniques. Then, we build an individual credit risk identification model based on the data fusion method on the above theoretical basis. After that, the selection of indicators was carried out among 20 variables of German credit data. Finally, an empirical study was conducted to compare the effect of the two data fusion methods on the performance of the individual credit risk identification model. The defects of single models in the personal credit risk problem domain are apparent. However, data fusion technology can help classification models to gain better uniformity in accuracy and robustness, but it also needs to be improved in terms of interpretability and promotion.*

Keywords: *Personal Credit Risk; Data Fusion; Classification Model; Bagging*

1. Introduction

1.1 Research Background

In recent years, China has proposed and implemented measures such as financial deleveraging, financial supply-side reform, and anti-corruption actions to deal with possible future economic wars and systemic financial risks so that the overall financial risks in China have gradually entered a stable and controllable zone. However, as a financial sector affecting the general public's lives, the personal credit sector is vulnerable to massive, distressed assets and systemic bank risks if significant management gaps exist. Therefore, it is necessary to strengthen the risk management of the personal credit sector.

The personal credit lending sector has profit incentives and regulatory pressures as motivation. It constructs a data-driven, individual credit scoring or classification model that varies based on customer categories to create a systematic and quantitative private credit assessment system to classify and identify customers. However, personal credit risk identification and assessment is now facing a development bottleneck. First, no new modeling methods have emerged in the field of personal credit risk identification in the last decade or so, and it is difficult to make a significant breakthrough in the critical mathematical and theoretical framework; second, there is less and less room for improvement in the various single risk identification models that can be used at present, and the design tends to be mature, so it is difficult to avoid the inherent defects of each model.

1.2 Literature Review

Risk scoring is the main component of a personal credit risk identification study. It refers to applying various information collected about the customer by the relevant institution to the classification model to score and predict the customer's creditworthiness level. Personal credit scoring is often seen as a customer classification problem based on individual information.

ISSN 2616-5902 Vol. 5, Issue 5: 62-74, DOI: 10.25236/AJBM.2023.050509

A statistical model is a scoring model to identify an individual's credit risk. It is realized in the field based on various classification models developed by statistical theory. It includes multiple linear regression, logistic regression, discriminant analysis, and decision trees. Wiginton (1980) first applied logistic regression to credit scoring, assuming that the dependent variable must be binary and the independent variables must be continuous or pseudo[1]; Makowski (1985) first discovered that the decision tree method, a non-parametric identification method based on statistical theory, could be used for the classification problem of credit scoring and has since applied this method to the field[2]; Quinlan (1986) proposed the ID3 decision tree algorithm using information gain as the node selection criterion. Seven years later, he changed the attribute choice criterion to information gain ratio and proposed the C4.5 decision tree algorithm[3,4]. Speaking of non-statistical methods, which mainly include SVM and artificial neural networks, have gained enough attention after the rise of machine learning. Still, the models built with them need more interpretability and stability.

Data fusion is mainly used in military and sensing fields, which can fuse data from multiple sources and sensors to perform pattern recognition and judgment with better recognition capability than under a single data source[5]. Academics have explained the definition of data fusion from different perspectives. Scholars in this field generally agree with the JDL's (Joint Council of the U.S. Department of Defense Tri-Service Laboratories) definition of data fusion, which is to view data fusion as a technique for processing the correlation and combination of multiple data sources to obtain accurate geographic locations and correctly identify military enemies[6]. Hall and Llinas offer another well-known definition: "Data fusion techniques are methods that can effectively fuse multi-sensor data, which can achieve higher accuracy and more stable determinations than single-sensor situations."[7] Combining the views of previous scholars, in this paper, we can define data fusion as a process or technique to obtain better information by processing a combination of multiple data or information sources.

This characteristic of data fusion technology across multiple disciplines makes the study of its classification often achieved by developing specific criteria. Durrant-Whyte classified data fusion techniques based on the relationship between data sources as applicable to complementary, redundant, and cooperative data, respectively[8]. Dasarathy proposed one of the most famous data fusion classification systems based on the type and nature of input and output data, which consists of the following five categories: data in-data out (DAI-DAO), data in-feature out (DAI-FEO), feature in-feature out (FEI- FEO), feature in-decision out (FEI-DEO), and decision in-decision out (DEI-DEO)[9]. JDL and the U.S. Department of Defense (DoD) proposed dividing the data fusion process into five processing levels, an associated database, and an information bus connecting these five components, constituting the popular JDL conceptual model[6]. In another classification criterion, scholars focus on where to perform the data fusion process when designing data fusion systems and identify four classifications of data fusion techniques: centralized architecture, decentralized architecture, distributed architecture, and hierarchical architecture. In the approach proposed by Luo et al. to classify the types of data fusion techniques based on the layers of data, the method is divided into four layers: signal layer, pixel layer (e.g., Kalman filtering), feature layer (e.g., fuzzy inference), and decision layer (e.g., Dempster- Shafer evidential inference)[10].

It is generally accepted that data fusion methods can be widely used in the case of multiple data sources for parameter estimation[11]. Therefore, data fusion techniques have been commonly used in multi-sensor environments to fuse and aggregate data from various distributed sources with a low error rate and high stability. In addition, data fusion can be applied to other areas, such as text processing.

1.3 Research Aims and Objectives

In this paper, we want to obtain more useful information through data fusion at the feature layer in the case of multidimensional data sources and to make the strengths and weaknesses of a single model complement each other through data fusion at the decision layer, which will significantly improve the discriminative accuracy and robustness of the overall model.

2. Theoretical Models

2.1 Data Fusion on Feature-layer

This technique is mainly applied in the process of index system establishment. In the face of redundant data, many scientific methods can be used to construct optimal indicator subspaces, in addition to the selection from the complex and specific indicators based on qualitative analysis. This technique

ISSN 2616-5902 Vol. 5, Issue 5: 62-74, DOI: 10.25236/AJBM.2023.050509

can eliminate data redundancy and improve risk identification classifiers' learning speed and accuracy. Selecting individual credit score metrics based on importance ranking is a common approach (see Figure 1). It uses a stepwise forward selection, stepwise backward deletion, or a combination of both to optimize the subspace of indicators according to the order of their importance (the statistics chosen to describe it differ in different methods) or guided by the classification accuracy or AUC values.

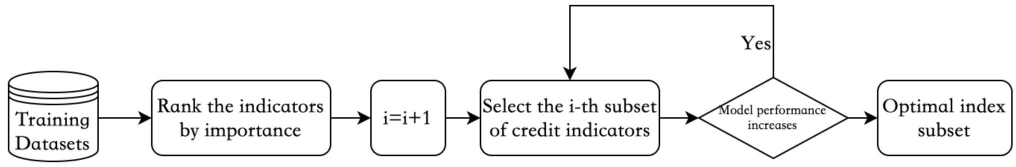

Figure 1: Selection method of personal credit indicators based on importance.

The chi-square statistic, information gain, Relief F method, and BP neural network connection weights are standard methods for selecting metrics based on importance ranking.

2.2 Data Fusion on Decision-layer

2.2.1 Fusion Model for Homomorphic Parallel Structures: Bagging Algorithm

Breiman (1996) constructed a juxtaposition-structured algorithm. It built the Bagging algorithm by randomly sampling the training set with put-backs to obtain multiple training subsets that are different from each other[12]. Afterward, numerous base classifiers were obtained using the same classification algorithm on different training sets, and then they were combined for decision-making using a fusion algorithm.

The detailed steps of the Bagging algorithm are as follows.

①Let i=1,2,3...,n. The training set is randomly sampled with put-back to generate the self-help sample set Ti. Since it is randomly sampled with put-back, some examples are repeatedly drawn while others are not, which results in the variability of the n self-help sample sets.

②Using an unstable classification algorithm, the n base classifiers hi (i=1,2,3...,n) is constructed on the self-help sample set Ti (i=1,2,3...,n).

③Record the predicted categories of n base classifiers for the test sample x. The final type of x is decided using the voting method.

$$H(X) = argmax\left(\sum_{i=1}^{T} h_i(x)\right) = y \quad (1)$$

2.2.2 Fusion Model for Heterogeneous Parallel Structures

The construction of this fusion model is like the previous homomorphic parallel structure. Still, the main difference is that this fusion model is a parallel arrangement of several base classifiers based on different classification models. Although it has a parallel structure, the juxtaposed objects are different classification models. Each of these base classifiers performs the classification process internally, and the classification output is finally used as the input to the fusion machine for the fusion decision. And the final output is the fused classification. The specific structure can be seen in Figure 2.

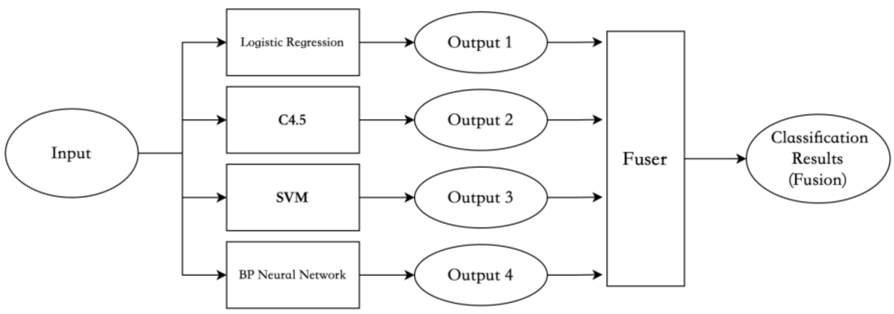

Figure 2: Diagram of fusion model of heterogeneous parallel structure.

ISSN 2616-5902 Vol. 5, Issue 5: 62-74, DOI: 10.25236/AJBM.2023.050509

The fuser chosen to study the default classification problem for personal credit identification in this paper is the weighted voting method. This method multiplies the number of classification votes by the weight of each weak classifier. The weighted votes for each category are then summed, and the category corresponding to the maximum value is the final category. This method can consider the differences in classification patterns among the base classifiers to the maximum extent possible.

3. Methodology

3.1 Research Philosophy

The study adopted the research philosophy of positivism. Specifically, the thesis took the deductive method to discuss. In this paper, the specific levels and characteristics of the main factors affecting the size of individual credit risk are analyzed qualitatively in the study of the indicator system. Indicators are selected using quantitative methods, and hypotheses are made about the advantages and disadvantages of decision fusion. Finally, the analysis is validated using the results of empirical data studies.

3.2 Staged Diagram of Model

The process of scoring individual credit risk usually has a general approach to the stages, which can generally be divided into four stages with logical, sequential relationships. However, this process is not immutable, and sometimes different choices of objectives can change the definition of the problem. Updating data and the number of data sources can also affect the second stage. The specific process relationship is shown in Figure 3 below.

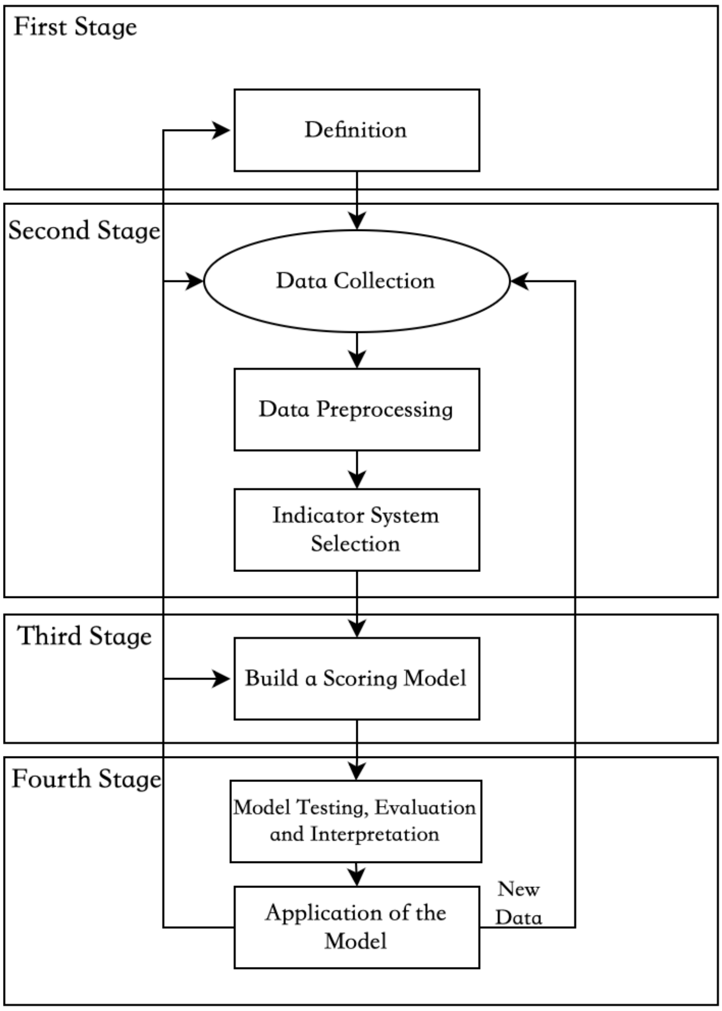

Figure 3: General process of personal credit risk assessment.

ISSN 2616-5902 Vol. 5, Issue 5: 62-74, DOI: 10.25236/AJBM.2023.050509

3.3 Approach to Data Collection and Preprocessing

This paper uses the classic German personal credit data from UCI. It can be obtained from the UCI database by searching for "German," which contains raw and numerical data. We use the raw data before the transformation.

The German personal credit database has no missing data cases and contains 20 dimensions, one customer category attribute, and 1000 data. The dataset will classify people as having good or bad credit risk by attributes. Customers will be classified as good (1) when they have good credit and inadequate (2) when they have bad credit. Some dataset indicators are described in Table 1 below, and detailed indicator descriptions are provided in Appendix 1.

Table 1: Description of some data indicators in the German credit dataset.

Indicator code	Indicator name	Indicator type	Indicator value range
A1	Status of existing checking account	Qualitative	A11:... <0 DM
			A12:0 <= ... <200 DM
			A13:...> = 200DM
			A14:no checking account
A2	Duration in month	Numeric	
A3	Credit history	Qualitative	A30: No credit taken/all credit returned
			A31: All credits with this bank have been repaid
			A32: Existing credit will be repaid on time
			A33: Past repayment delays
			A34: Existing key account/other credit (not at this bank)
A4	Purpose	Qualitative	A40: Car (new)
			A41: Cars (used)
			A42: Furniture/equipment
			A43: Radio/TV
			A44: Household appliances
			A45: Maintenance
			A46: Education
			A47: leave
			A48: Retraining
			A49: Business
			A410: other

The data types in the German personal credit database involve two categories, numeric data, and character-based data. Numeric data are measured using numeric values, such as the seven variables of monthly Period, loan amount, installment rate, current residence, age, loan amount, and the number of obligated supporters in this database. Character-based data is represented using characters such as Chinese and English, ASCII, etc. The other 13 variables in this paper are character-based data. To improve the applicability of the data to more models, this paper will use the LabelEncoder of the Sklearn library to encode the character data.

This paper uses the triple standard deviation detection method for three continuous-type numerical variables: credit period, loan amount, and age. The record where the absolute value of the difference between the value of any numerical variable and the mean value exceeds three times the standard deviation is regarded as an outlier. Finally, the outlier detection situation shown in Table 2 below is obtained. The record where the detected outlier is located is deleted.

Table 2: Outlier detection in German credit dataset.

Variable	Minimum	Maximum	Mean	Standard Deviation	Number of Outliers
Credit Period	4	72	20.90	12.00	14
Loan Amount	250	18,424	3,271.26	2,822.74	24
Age	19	75	35.55	11.38	7

In the process of personal credit scoring, it is often necessary to discretize numerical data for reasons such as protecting sensitive private information, improving the speed and accuracy of classification, and the fact that some classification algorithms, such as ID3 and Parsimonious Bayes, can only handle

ISSN 2616-5902 Vol. 5, Issue 5: 62-74, DOI: 10.25236/AJBM.2023.050509

discrete data. This paper chooses the information entropy-based discretization method for data fusion at the pixel level. We wrote the code for implementing the information entropy-based discretization method using Python to perform data discretization operations on three continuous-type numerical variables: credit term, loan amount, and age. The results are shown in Table 3.

Table 3: Data discretization standards based on information entropy.

Variable	Discretization criteria
A_2(credit period)	A_2≤12; A_2>12
A_5(loan amount)	A_5≤3914; A_5>3914
A_13(age)	A_13≤35; A_13>35

According to the results in Table 3, it can be seen that the credit period is divided into two intervals, A2≤12 and A2>12; the loan amount is divided into two intervals, A5≤3914 and A5>3914; and the age is divided into two intervals A13≤35 and A_13>35.

To address the data imbalance problem, this paper uses Python to write a nearest-neighbor-based SMOTE sampling program and sets the parameter of generated data volume to 200%. As a result, 558 new wrong customers were developed based on the 274 "bad" customer records in the original dataset with outliers removed.

3.4 Pros and Cons of Classification Models

In this paper, four standard personal credit scoring models are used. Here the following conclusions are drawn from a comparative analysis of the characteristics, advantages, and disadvantages of the four models:

a) Logistic regression has more relaxed assumptions compared to other classification methods. It has good explanatory power and stability but lower classification accuracy and is more sensitive to correlation between variables[13].

b) The C4.5 decision tree model has no restrictions on data distribution assumptions, and its tree classification rules can be used for model interpretation. Still, overfitting problems occur when there are many tree levels.

c) The advantage of support vector machines is their higher classification accuracy compared to statistical models. It differs from neural networks because it is suitable for modeling small samples. Its disadvantages are sensitivity to training parameters, lack of intuitive interpretability, and lack of stability.

d) BP neural networks can fit complex nonlinear relationships compared with other classification models. While the assumptions on the data are very relaxed, the classification accuracy is high, and they are suitable for modeling large-volume data. Still, interpretability has been a long-standing problem of BP neural networks.

Because no scoring model can achieve uniformity in classification accuracy and robustness, using data fusion, especially at the decision level, in individual credit scoring, is vital in improving the robustness of decisions and classification accuracy.

3.5 Construction of Model Based on Data Fusion

This paper's credit risk identification models include single classification and decision fusion. The single classification model includes two statistical models and two non-statistical models. The statistical models include Logistic regression and the C4.5 decision tree model, and the non-statistical models consist of an SVM support vector machine and BP neural network. Decision fusion models include four classification models combined with the Bagging algorithm and decision fusion models with heteromorphic parallel structures.

This paper will generate training and test sets using a ten-fold cross-validation approach. Four single classification models, four classification models combined with the Bagging algorithm, and decision fusion models with the heteromorphic parallel structure are trained and validated for model accuracy using the test set. This paper will focus on comparing the performance of the models after using the fusion algorithm with the single model performance to demonstrate that decision fusion can be effective in improving the classification performance for the application of individual credit risk identification.

ISSN 2616-5902 Vol. 5, Issue 5: 62-74, DOI: 10.25236/AJBM.2023.050509

4. Results

4.1 Construction of Indicator System

The importance ranking in Table 4 below was obtained by ranking the 20 indicators in the German personal credit database using the importance-based individual credit score indicator selection method mentioned.

Table 4: Ranking of the importance of German credit data indicators.

Sorting	Chi-square statistic	Info Gain Ratio	Connection Weight	Relief F	Combination
1	A1	A1	A1	A1	A1
2	A3	A2	A4	A14	A3
3	A6	A3	A6	A3	A2
4	A4	A20	A3	A9	A4
5	A2	A5	A2	A2	A5
6	A5	A6	A10	A5	A6
7	A12	A15	A7	A10	A14
8	A7	A14	A20	A4	A7
9	A15	A4	A8	A7	A12
10	A14	A10	A11	A11	A10
11	A9	A12	A12	A13	A9
12	A20	A7	A5	A8	A20
13	A10	A9	A14	A17	A15
14	A8	A8	A15	A15	A8
15	A16	A16	A17	A19	A11
16	A17	A19	A9	A12	A17
17	A19	A17	A16	A18	A16
18	A11	A11	A18	A16	A19
19	A18	A18	A19	A20	A13
20	A13	A13	A13	A6	A18

The results show that the five ranking methods have the same perceived importance for variables A1, A2, A3, and A4. These four variables are almost always in the top 5 positions of all methods' rankings. At the same time, all five scale methods identified four variables, A13, A17, A18, and A19, as unimportant, and nearly every ranking method ranked these four indicators in the last five positions. The variables A6 and A20 are ranked differently in different ways. A6 is ranked last in Relief F but in the top 6 in all other five rankings; A20 is ranked 2nd by the information gain rate method but is further down in different ways.

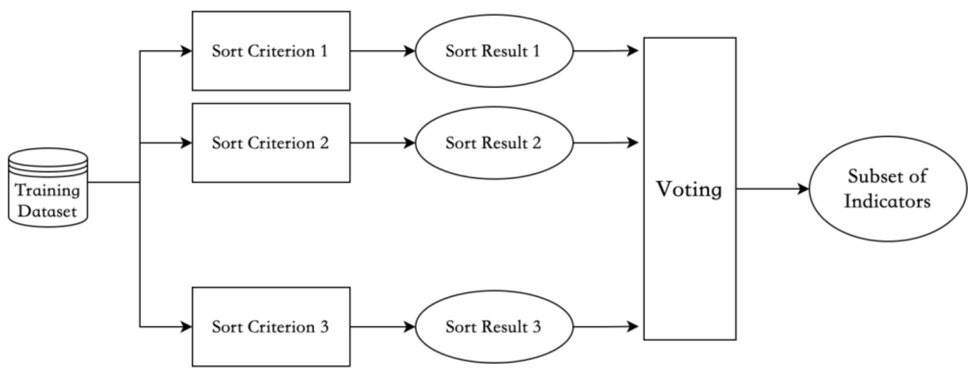

Figure 4: Combination methods using multiple importance rankings.

This paper uses the combination method as in Figure 4 above to obtain more scientific and stable ranking results. It uses weighted voting to rank the importance of all indicators and get the results, as shown in the last column of Table 4.

ISSN 2616-5902 Vol. 5, Issue 5: 62-74, DOI: 10.25236/AJBM.2023.050509

4.2 Results of Single Model

The results of the two statistical models and the two non-statistical models used in this paper were summarized after performing the 10-fold cross-validation to obtain Table 5 and Figure 5.

Table 5: Average classification accuracy of four models in 10-fold cross-validation.

	Category	Total Classification Accuracy	F1 Score	Rate of Change: Total Classification Accuracy
Logistic regression	training set	76.01%	0.7591	0.96%
	test set	75.28%	0.7522	
C4.5 Decision Tree	training set	100.00%	1.0000	20.38%
	test set	79.62%	0.7956	
SVM	training set	86.24%	0.8615	6.26%
	test set	80.84%	0.8070	
BP neural network	training set	94.13%	0.9413	10.55%
	test set	84.19%	0.8408	

From the results of the empirical study, the classification accuracy of non-statistical models such as BP neural network and SVM is significantly higher than the other two models.

Regarding model robustness, the Logistic regression model is the most stable, and the SVM support vector machine is the 2nd. Still, this paper's C4.5 decision tree classification model needs better robustness.

In terms of model interpretability, the results of the empirical study summarize that statistical models perform better than non-statistical models in this regard. Logistic regression has the best interpretability, and the C4.5 decision tree is the 2nd and can directly generate practical personal credit scoring rules. The BP neural network and support vector machine models have the worst interpretability among the four classification models.

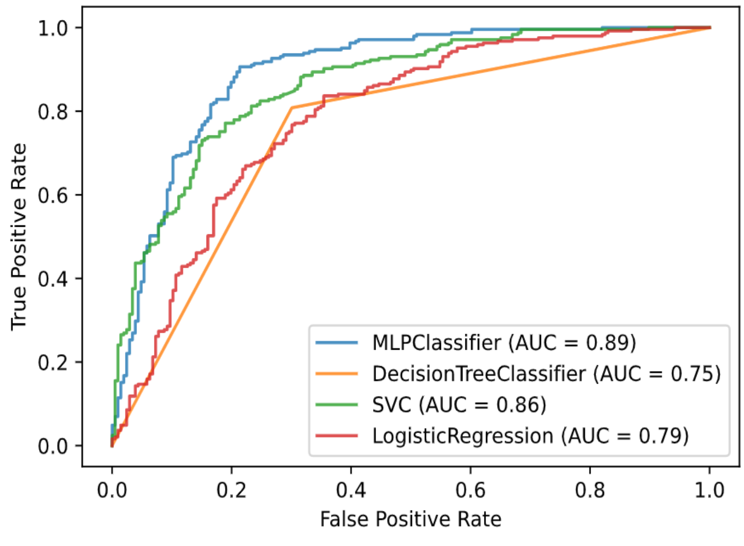

Figure 5: AUC of four classification models.

4.3 Results of Data Fusion Model

4.3.1 Bagging Algorithm Decision Fusion Model with Homomorphic Parallel Structure

In the previous section, four classification models (two statistical and two non-statistical models) were built, and the model performance was analyzed. This section used Bagging Classifier from the scikit-learn library to create an integrated Bagging model based on the above four models. The same sample partitioning method, i.e., ten-fold cross-validation, was used to analyze the model's accuracy. The final classification accuracy of the model was obtained as follows in Tables 6-9, and the AUC is shown in Figures 6-9 below.

ISSN 2616-5902 Vol. 5, Issue 5: 62-74, DOI: 10.25236/AJBM.2023.050509

Table 6: Classification accuracy of the Bagging algorithm model based on Logistic Regression.

	precision	recall	f1-score
1	0.78	0.66	0.72
2	0.74	0.84	0.79
accuracy_train	76.71%		
accuracy_test	75.61%		
Rate of Change: Total Classification Accuracy	1.43%		

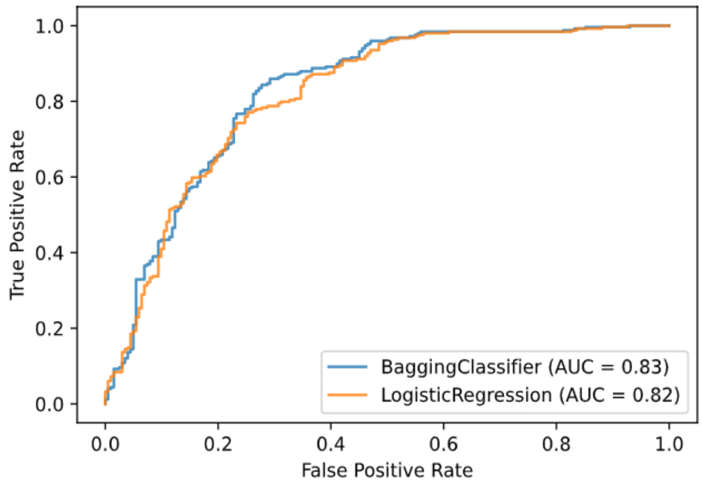

Figure 6: Comparison of AUC of Logistic Regression and its Bagging integrated model.

Comparing Table 5 and Table 6, the logistic model trained by bootstrap sampling using the Bagging algorithm to obtain several different training subsets performs better. Its total classification accuracy on the test set improves by 0.33 percentage points, while the overall accuracy variation rate change is small, and the model is still very robust. Also, when using the ROC curve to measure the model's accuracy, it can be seen in Figure 6 that the improved logistic regression model using the Bagging algorithm has a larger AUC area.

Table 7: Classification accuracy of the Bagging algorithm fusion model based on C4.5 decision tree.

	precision	recall	f1-score
1	0.85	0.76	0.80
2	0.81	0.88	0.84
accuracy_train	100.00%		
accuracy_test	82.48%		
Rate of Change: Total Classification Accuracy	17.52%		

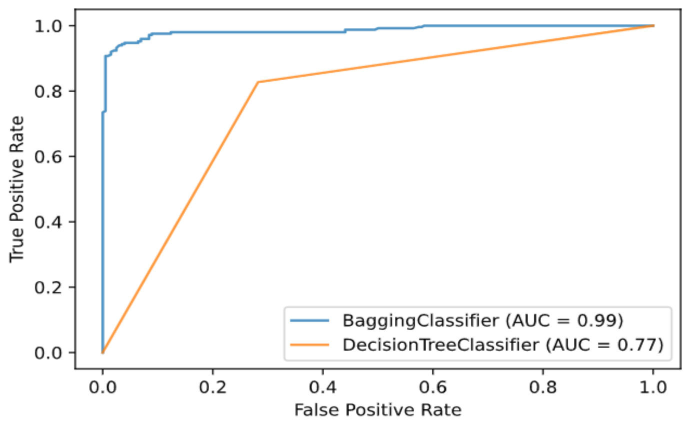

Figure 7: Comparison of AUC of C4.5 Decision Tree and its Bagging integrated model.

Comparing Table 5 and Table 7, we can see that compared to the slight improvement of Logistic

ISSN 2616-5902 Vol. 5, Issue 5: 62-74, DOI: 10.25236/AJBM.2023.050509

regression, the total classification accuracy of the C4.5 decision tree model improved using the Bagging algorithm for the test set has been enhanced by 2.86 percentage points. In the meantime, the overall accuracy variation rate has been reduced by 2.86 percentage points, and the classification accuracy and robustness have been significantly improved. Also, the AUC area in Figure 7 has been increased by 0.22.

Table 8: Classification accuracy of Bagging algorithm fusion model based on SVM support vector machine.

	precision	recall	f1-score
1	0.84	0.72	0.77
2	0.78	0.88	0.83
accuracy_train	87.07%		
accuracy_test	80.49%		
Rate of Change: Total Classification Accuracy	7.66%		

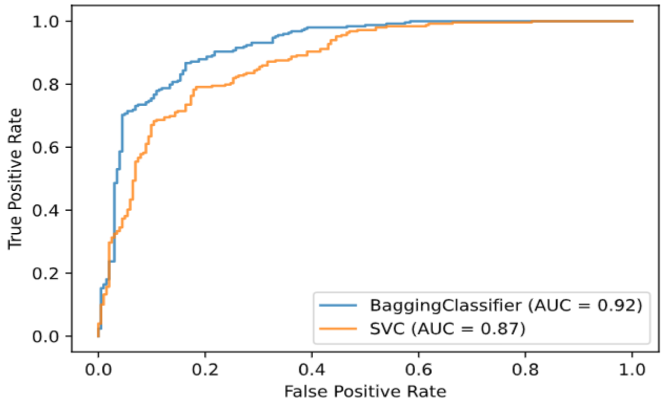

Figure 8: Comparison of AUC of SVM and its Bagging integrated model.

Comparing Table 5 and Table 8, the total classification accuracy of the SVM support vector machine model improved using the Bagging algorithm enhanced by 0.35 percentage points for the test set. Also, the overall accuracy variation rate increased by 1.4 percentage points, and the AUC area in Figure 8 increased by 0.05.

Table 9: Classification accuracy of Bagging algorithm fusion model based on BP neural network.

	precision	recall	f1-score
1	0.9	0.74	0.81
2	0.81	0.93	0.86
accuracy_train	98.67%		
accuracy_test	84.26%		
Rate of Change: Total Classification Accuracy	14.60%		

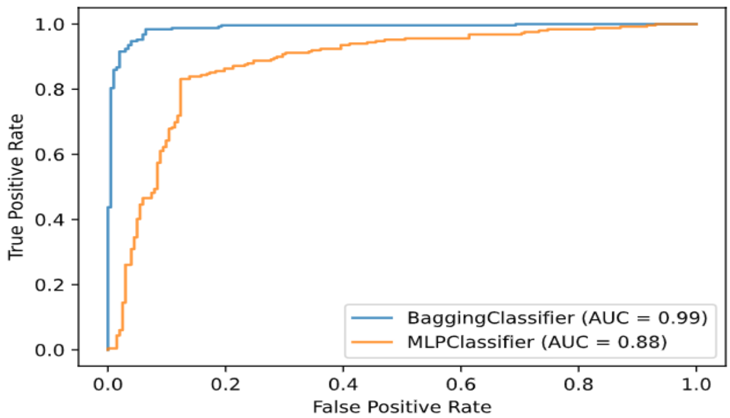

Figure 9: Comparison of AUC of BP neural network and its Bagging integrated model.

ISSN 2616-5902 Vol. 5, Issue 5: 62-74, DOI: 10.25236/AJBM.2023.050509

Comparing Table 5 and Table 9, we can see that the improved BP neural network model using the Bagging algorithm has a slight improvement in the total classification accuracy of the test set, while the overall accuracy variation rate increases by 4.1 percentage points, which shows that the Bagging algorithm used on the BP neural network does not have a significant performance improvement effect.

In general, improving the Bagging algorithm enhances the classification accuracy for all four classification models but at the cost of losing some model robustness for specific models.

4.3.2 Voting Method Decision Fusion Model with Heterogeneous Parallel Structure

The voting method decision fusion model with heteromorphic parallel structure was trained and validated using the same sample data and ten-fold cross-validation method. The results are shown in Table 10 and Figure 10 below.

Table 10: Classification accuracy of the heteromorphic parallel structure fusion model.

	precision	recall	f1-score
1	0.9	0.74	0.81
2	0.81	0.93	0.86
accuracy_train	100%		
accuracy_test	86.25%		
Rate of Change: Total Classification Accuracy	13.75%		

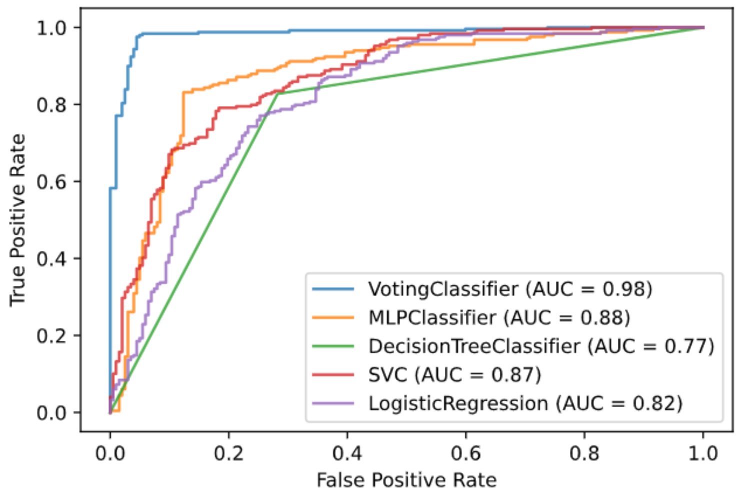

Figure 10: Comparison of the AUC of four single classification models and heterogeneous parallel structure models.

Among all single classification models and decision fusion models, the total classification accuracy of the heteromorphic parallel structured decision fusion model is the highest, reaching 86.25%, while its overall accuracy variation rate is only slightly higher than that of the BP neural network. Regarding robustness assessment, the overall accuracy variation rate of the single-model BP neural network is significantly lower than that of the BP neural network improved with the Bagging algorithm. This dramatically reduces the instability associated with using the C4.5 decision tree model. Thus, the heteromorphic parallel structured voting decision fusion approach can perform decision fusion based on multiple single individual credit scoring models while balancing absolute improvement in classification accuracy (over any single model) and robustness.

Table 11: Comparison of test set classification accuracy between four models and their two fusion models.

	Logistic Regression	C4.5 Decision Tree	SVM	BP neural network
single-model	75.28%	79.62%	80.84%	84.19%
improved by Bagging Algorithm	75.61%	82.48%	80.49%	84.26%
heteromorphic parallel structure	86.25%			

Combining the classification accuracies of the single model and the two fusion models in the test set (see Table 11), several single classification models showed practical improvements in accuracy and stability when either fusion model was used for improvement. The fusion model with the heteromorphic parallel structure presents a good improvement in classification accuracy and model robustness. But the fusion model with the homomorphic parallel structure does not have the same noticeable improvement effect as the other.

5. Discussion

This paper selects the classic German credit database data to study individual credit risk identification. Based on the pre-processing data work, such as outlier detection, data discretization, and imbalanced data processing, this paper first uses a feature-layer data fusion model to select individual credit score indicators based on importance ranking in the section of the indicator system. After that, the application prospects of data fusion technology in personal credit identification are explored by comparing the prediction accuracy and robustness of a single individual credit risk identification model and two parallel structural decision fusion models, homomorphic and heteromorphic.

Specifically, the paper concludes with the followings: (1) The shortcomings of every single model are apparent. Based on the empirical study, the conclusions obtained in this paper are consistent with those of previous scholars: in terms of robustness alone, logistic regression has a unique advantage, but the C4.5 decision tree model, SVM support vector machine model, and BP neural network model are insufficient; the classification accuracy of SVM and BP neural network is high, but they face the problem of poor interpretation. In conclusion, achieving a better unification of classification accuracy and model stability in the commonly used statistical models for individual credit scoring is generally tricky. (2)After using data fusion techniques, the classification models can be better unified in classification accuracy and robustness, but they also need to improve in interpretation and generalization. In this paper, several single classification models are enhanced by using two data fusion methods, and the fusion model with the heteromorphic parallel structure improves classification accuracy and model robustness. However, the weight determination of the heteromorphic parallel structure fusion model depends on the user's experience and needs to be more reproducible.

How to appropriately improve the two decision fusion models constructed in this paper to obtain better interpretability and replicability while ensuring the accuracy and robustness of the model classification is a future research direction that the writer can conduct. In addition, considering more comparative studies on data fusion at the feature layer to evaluate the effect of data fusion at the feature layer is also a direction that can investigate afterward.

References

[1] Wiginton J. C. (1980). A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. Journal of Financial and Quantitative Analysis, 15(3), 757–770.
[2] Makowski M., & Sosnowski J. (1985). A Decision Support System for Planning and Controlling Agricultural Production with a Decentralized Management Structure. In M. Grauer, M. Thompson, & A. P. Wierzbicki (Eds.), Plural Rationality and Interactive Decision Processes (pp. 296–305). Springer.
[3] Quinlan J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106.
[4] Quinlan J. R. (1993). C4. 5: Programming for machine learning. Morgan Kauffmann, 38(48), 49.
[5] Su L., Li Q., Xu X., & Guo X. (2006). Data Fusion Algorithm for Sensor Network Based on D-S Evidence Theory. Journal of Chinese Computer Systems, 1321–1325.
[6] White F. E. (1991). Data fusion lexicon. Joint Directors of Labs Washington DC.
[7] Hall D. L., & Llinas J. (1997). An introduction to multisensor data fusion. Proceedings of the IEEE, 85(1), 6–23.
[8] Durrant-Whyte H. F. (1988). Sensor models and multisensor integration. The International Journal of Robotics Research, 7(6), 97–113.
[9] Dasarathy B. V. (1997). Sensor fusion potential exploitation-innovative architectures and illustrative applications. Proceedings of the IEEE, 85(1), 24–38.
[10] Luo R. C., Yih C.-C., & Su K. L. (2002). Multisensor fusion and integration: Approaches, applications, and future research directions. IEEE Sensors Journal, 2(2), 107–119.
[11] Castanedo F. (2013). A review of data fusion techniques. The Scientific World Journal, 2013.
[12] Breiman L. (1996). Bagging predictors. Machine Learning, 24, 123–140.
[13] REN Xiao, JIANG Minghui, CHE Kai, & WANG Shang. (2016). The research on methods of

ISSN 2616-5902 Vol. 5, Issue 5: 62-74, DOI: 10.25236/AJBM.2023.050509

personal credit scoring combined model selection based on optimized index system. Journal of Harbin Institute of Technology, 67–71.

Appendix 1: Description of German credit dataset

Value	Definition	Type	Range
A1	Status of existing checking account	qualitative	A11 : ...<0 DM A12 : 0<=...<200 DM A13 : ...>=200 DM / salary assignments for at least 1 year A14 : no checking account
A2	Duration in month	numerical	
A3	Credit history	qualitative	A30 : no credits taken/all credits paid back duly A31 : all credits at this bank paid back duly A32 : existing credits paid back duly till now A33 : delay in paying off in the past A34 : critical account/other credits existing (not at this bank)
A4	Purpose	qualitative	A40 : car (new) A41 : car (used) A42 : furniture/equipment A43 : radio/television A44 : domestic appliances A45 : repairs A46 : education A47 : (vacation - does not exist?) A48 : retraining A49 : business A410 : others
A5	Credit amount	numerical	
A6	Savings account/bonds	qualitative	A61 : ... < 100 DM A62 : 100 <= ... < 500 DM A63 : 500 <= ... < 1000 DM A64 : ... >= 1000 DM A65 : unknown/ no savings account
A7	Present employment since	qualitative	A71 : unemployed A72 : ... < 1 year A73 : 1 <= ... < 4 years A74 : 4 <= ... < 7 years A75 : ... >= 7 years
A8	Installment rate in percentage of disposable income	numerical	
A9	Personal status and sex	qualitative	A91 : male:divorced/separated A92 : female:divorced/separated/married A93 : male:single A94 : male:married/widowed A95 : female:single
A10	Other debtors / guarantors	qualitative	A101 : none A102 : co-applicant A103 : guarantor
A11	Present residence since	numerical	
A12	Property	qualitative	A121 : real estate A122 : if not A121 : building society savings agreement/life insurance A123 : if not A121/A122 : car or other, not in attribute 6 A124 : unknown / no property
A13	Age in years	numerical	
A14	Other installment plans	qualitative	A141 : bank A142 : stores A143 : none
A15	Housing	qualitative	A151 : rent A152 : own A153 : for free
A16	Number of existing credits at this bank	numerical	
A17	Job	qualitative	A171 : unemployed/ unskilled - non-resident A172 : unskilled - resident A173 : skilled employee / official A174 : management/ self-employed/highly qualified employee/ officer
A18	Number of people being liable to provide maintenance for	numerical	
A19	Telephone	qualitative	A191 : none A192 : yes, registered under the customers name
A20	foreign worker	qualitative	A201 : yes A202 : no