ISSN: 2472-3703 DOI: 10.6911/WSRJ.201911_5(11).0018

Research on Personal Credit Evaluation Based on Multi-model Combination

Qun Ma[1, a, *], Kunlun Li[1] and Jiahui Hu[2]

[1]Department of Electronic Information Engineering, School of Hebei University, Baoding 071000, China

[2]Nortn China Electric Power University, Baoding 071000, China

[a]Corresponding author Email: 1436810029@qq.com

Abstract

In order to improve the accuracy, stability and diversity of the personal credit evaluation model and reduce the instability of the single model, this paper proposes a credit evaluation model for the multi-model combination algorithm. Weighted voting combinations were performed on three stable classifiers, random forest, LightGBM and support vector machine. In the feature extraction, an improved feature extraction method is adopted. Firstly, according to Fisher's ratio, the importance degree of personal credit data features is analyzed, and their features are partitioned according to their different values. In the training model, the improved grid search algorithm is used to optimize the parameters of the model to improve the optimization speed and efficiency. The experimental results show that the performance of LightGBM model is better than that of random forest and support vector machine. The accuracy of the model is higher and the resolution is stronger. The comprehensive performance of the combined model is better than three single models, which can be applied to the field of personal credit evaluation.

Keywords

Combined model; LightGBM; Support Vector Machine; Random Forest; Grid Search; Fisher Criterion.

1. INTRODUCTION

In recent years, with the advent of the era of big data, the Internet financial industry has risen rapidly. P2P online lending platforms, micro-credit, consumer loans and other online lending businesses have sprung up. As a key technical issue in the financial field, personal credit evaluation has received extensive attention in recent years [1]. In order to make the sustainable development of Internet finance, an emerging industry, scientific and rational use of data mining technology to do personal credit assessment is the key. Personal credit evaluation is usually based on the machine learning model, based on the credit information and personal data of the creditor. The selection and use of the model should be studied to solve key technical problems.

Traditional personal credit evaluation models usually use a single classification algorithm for risk prediction. However, due to the diversification of the Internet, with the increasing amount of data and the diversification of data, traditional classification algorithms have been difficult to meet the existing data. Processing and solving practical problems [2]. In recent years, the combined model approach has been considered the most influential development in data mining and machine learning. The personal credit evaluation model based on the combined

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201911_5(11).0018

model algorithm provides a good idea for the risk control of Internet credit institutions. The combined model refers to the combination of multiple models through certain methods, making full use of the classification results of a single model, so as to improve the accuracy of model prediction [3]. In the literature [4], the author proposes a combined model based on neural network and wavelet theory. The combined model was compared with four other single models (TS fuzzy neural network, BP neural network, Elman neural network and RBF neural network). The wavelet neural network combination model can effectively reduce the prediction bias and has higher prediction accuracy. In the literature [5], a multi-model aggregation method based on Lasso, support vector regression, random forest and gradient advancing decision tree is proposed for the multi-point prediction of earth pressure in EPB shield. Use Leave-One-Out to verify predictive performance. The Lasso, Random Forest and Gradient Boosting decision tree models provide feature importance. Experimental results show that the performance of multi-model sets is better than single models. In the literature [6], the authors use Logistic regression, Probit regression, Bayesian classifier and holding vector machine to combine. Firstly, the common models of personal credit scoring models are summarized and summarized, and then the accuracy of different models is illustrated by analysis and comparison. In the "bad sample" distinction and the result judgment using weighting method, a correction algorithm is proposed to determine the index weight in the credit scoring model, to meet the needs of different bank data diversification, and improve the accuracy of the scoring model. In the literature [7], the authors use Logistic regression and BP neural network algorithm. And the two methods are improved respectively. The personal credit evaluation model based on BP-Logistic hybrid strategy is constructed by using Clementine tool.

Using neural network and Bayesian classifier algorithm for model combination, it is mainly used in the field of credit risk assessment to face the following two problems:

(1) The interpretability of the model is relatively poor, and it is difficult for people to find credit decision rules from the model. In the personal credit assessment, not only the accuracy of the credit risk of the loan customer is required to be high, but also the funder needs to be able to understand the credit decision rules. The interpretability of neural networks is poor and does not reflect the importance of different variables. Moreover, compared with other statistical models, the stability of neural networks is relatively poor, and there are serious limitations in the field of credit scoring [8].

(2) The calculation speed of the model is poor and takes a long time. Banks and online lending platforms have huge customers every day, not only need higher accuracy but also need faster computing models to support the business.

(3) The Bayesian classifier needs to know the prior probability, and the prior probability often depends on the hypothesis. The a priori and the data determine the posterior probability to determine the classification, so the classification effect is sometimes not ideal [9]. Moreover, Bayesian classifiers are sensitive to the form of input data.

After the above analysis, it can be seen that a plurality of single models can achieve better results by combining certain methods, but two problems need to be paid attention to when selecting the model: the interpretability of the model and the operation speed of the model. In order to make the credit evaluation model meet the fast calculation speed under the premise of ensuring the accuracy, the relatively good requirements can be explained. In this paper, the three algorithms of random forest, LightGBM and support vector machine are combined to obtain a personal credit evaluation combination model.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201911_5(11).0018

2. MULTI-MODEL COMBINATION METHOD

2.1. Voting Strategy

Voting strategy is the voting method, which is the simplest and most practical method in the model fusion method. Its main idea is to adopt the principle of minority obeying majority. In the multi-model combination scheme, each single model analyzes the training set to generate classification results, and directly votes the classification results of multiple single models according to the voting principle. The most common category of the total number of votes is the final output of the combined model [10].

2.2. Stacking Strategy

The Stacking strategy is a hierarchically integrated model framework in which the base classifiers can be heterogeneous or homogenous. For example, the random forest algorithm is a new algorithm integrated by many decision trees. Stacking usually adopts a two-layer framework structure, as shown in Figure 1. Its execution steps are as follows: first, stacking trains the base classifier from the initial data set, and then "generates" a new data set for training the meta-learner. Finally, the output of the meta-learner is the final output of the combined model [11].

Among them, the output of the base classifier in the new data set is taken as a sample input feature, and the mark of the initial sample is still treated as a sample mark [12]. In the training phase, the training set of the meta-learner is generated by the base classifier. If the training set of the base classifier is used directly to generate the meta-learner training set, the over-fitting risk will be relatively large, so generally by using the crossover A method of verifying or leaving a method to generate a sample of a meta-learner using a sample that is not used in the training-based classifier [13].

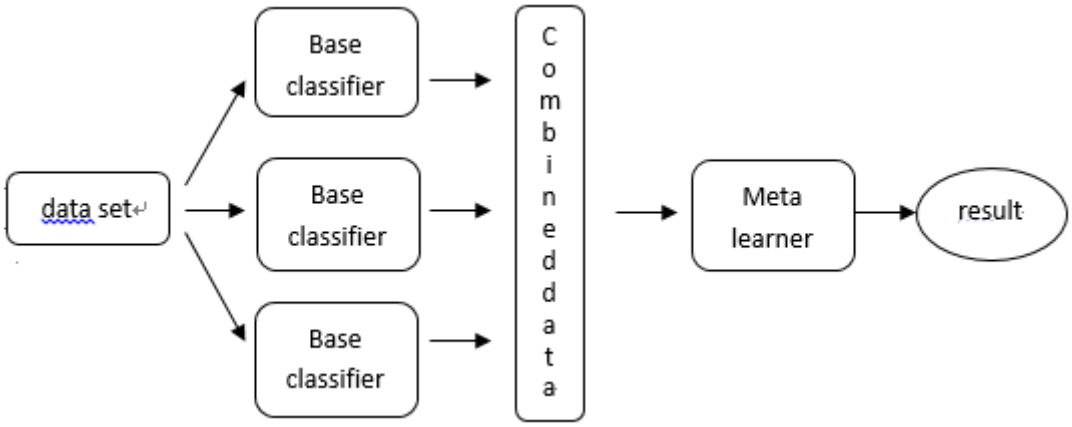

Figure 1. Stacking strategy

2.3. Blending Strategy

Blending is similar to Stacking, except that blending differs in that it does not use cross-validation methods to obtain predictive values to generate features of the second layer model, but instead uses different data sets to train different base learners. Taking a two-layer framework as an example, the specific training steps are as follows: First, the data is divided into two parts: training set train and test set test, and the training set train is divided into S1 and S2. After S1, multiple models are trained with S1 in the first layer, and the prediction results of S2 and test are taken as new features of the second layer. In the second layer, a new model is trained with the new features and tags of S2, and then the new feature of test is entered as the final test set. Finally, the predicted result for test is the value of the final model fusion [14].

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201911_5(11).0018

3. EVALUATION MODEL SELECTION SCHEME

There are two main ideas for the specific selection method of the model:

(1) The single model used belongs to the same type and is homogeneous. For example, the random forest algorithm is constructed by a number of decision trees through a certain method to construct a random forest algorithm.

(2) The single model used is of a different type and is heterogeneous. For example, when solving the classification problem, the same data set is used to combine the different classification algorithms such as XGBoost algorithm, Naive Bayes algorithm and decision tree to obtain the final classification model. There are two basic points to be met when using a heterogeneous model: first, the performance of each single model cannot be too different; second, the similarity between each single model is as small as possible.

In this paper, we use the second method, which uses a heterogeneous model to combine, using three algorithms: random forest, LightGBM and support vector machine. Random forest and LightGBM belong to the integrated decision tree algorithm, but random forest belongs to Baging type algorithm, mainly focusing on reducing the problem of variance. LightGBM belongs to Boosting type algorithm, mainly focusing on reducing the problem of deviation. Support vector machines are not part of the integration algorithm, and mainly focus on the problem of structural risk minimization. It can be seen that the correlation between the three algorithms is small, and the diversity of the combined model can be increased, and the classification performance of the model is also relatively close, which meets the requirements of the combined model.

3.1. Random Forest Model

L. Breiman proposed the Random Forest Algorithm in 2001, which is an improved Bagging algorithm based on the Bootstrap method [15]. The random forest algorithm uses the Bootstrap repetitive sampling technique to randomly extract m subclass samples randomly in the training data set M to form a new training data set W, and then adopt m subclass samples in the new training set W. Generate m decision trees, and finally get the final result of the random forest algorithm by simple voting method. The random forest effectively improves the classification accuracy on the basis of solving the problem that the single decision tree is prone to over-fitting.

3.2. LightGBM Model

LightGBM (Light Gradient Boosting Machine) is an open source, fast and efficient decision tree algorithm based enhancement (GBDT, GBRT, GBM and MART) framework published by Microsoft Research Asia on GitHub in January 2017 [16]. XGBoost has proven to be a very efficient and popular classification algorithm in various application scenarios and major competitions, but LightGBM is a more accurate and efficient new algorithm, which reduces the memory footprint under the premise of ensuring accuracy. About 3 times faster than XGBoost. By introducing this fast, accurate and efficient algorithm in the combined model in order to obtain a better personal credit evaluation model.

LightGBM is a decision tree algorithm based on Histogram, which uses the optimal leaf-learning learning method. However, other lifting algorithm split trees usually do not use this method, but adopt a layer-by-layer learning method. Level-wise Learning).

(1) Level-wise Learning can process the same layer of leaves at the same time, so multi-thread optimization can be realized, and the complexity of the model can be well controlled, and it is not easy to over-fitting. But in fact, because many leaves have low split gain, there is no need to search and split, and Level-wise splits the leaves of the same layer without distinction, which increases the unnecessary computational cost.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201911_5(11).0018

(2) Leaf-wise learning is to find the split with the largest split gain from all the current leaves, and then repeat the steps. Therefore, Leaf-wise can reduce more errors when the number of splits is the same, which is a more efficient strategy. But there is also an obvious disadvantage: the decision tree it generates may be too deep and over-fitting. Therefore, LightGBM adds a maximum depth limit above Leaf-wise, which prevents high-efficiency while preventing over-fitting, making the algorithm faster and more efficient, so it has higher precision, and supports parallel learning. Running memory. Any other existing lifting algorithms are difficult to achieve.

3.3. Support Vector Machine Model

Support Vector Machine is a machine learning algorithm based on statistical learning theory proposed by Vapnik in the mid-1990s [17, 18]. Support vector machine is different from traditional learning method. It is a structural risk. An approximate implementation of the minimization method. Support vector machines can select a portion of a large amount of training data for model building and are generally not sensitive to dimensions. It can better solve problems such as nonlinearity, high dimensionality, and local minimum points.

Suppose the training data set is S={xi, yi}(i=1,2,3,$\cdots$,m), xi$\in$Rn, yi $\in${+1, -1}, where yi is the output. Considering xi as m sample points in n-dimensional space, in order to maximize the isolation boundary between the positive and negative examples, the support vector machine will establish a classification hyperplane as the decision surface. This hyperplane is expressed as:

$$\sum_{i=1}^{m}\omega_i x_i + b = 0 \tag{1}$$

Divide m samples into two categories and maximize the classification interval $\frac{2}{\|\omega\|^2}$. Such a hyperplane is called the optimal classification plane, as shown in Figure 2 below. Among them, the classification hyperplane H1, H2 are respectively expressed as:

$$H_1 = \sum_{i=1}^{m}\omega_i x_i + b = 1 \tag{2}$$

$$H_2 = \sum_{i=1}^{m}\omega_i x_i + b = -1 \tag{3}$$

To maximize the classification interval, the problem can be equivalently converted to minimize $\frac{\|\omega\|^2}{2}$. The problem of seeking the optimal classification hyperplane H is equivalent to solving the following minimization problem:

$$\min \frac{1}{2}\|\omega\|^2 \tag{4}$$

$$\text{s.t.} y_i\left[\omega^T x + b\right] \geq 1\left(i = 1, 2, 3, \cdots, m\right) \tag{5}$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201911_5(11).0018

According to the optimization theory, the classification decision function under linear separable condition is solved. The Lagrange multiplier corresponding to each sample is represented by αi, the classification threshold is represented by b*, and the corresponding sample when αi is not zero is the support vector. Then the decision function is:

$$f(x)=\operatorname{sgn}\left\{\sum_{i=1}^{m}\alpha_i^* y_i\left(x_i^T \mathrm{x}\right)+b^*\right\} \tag{6}$$

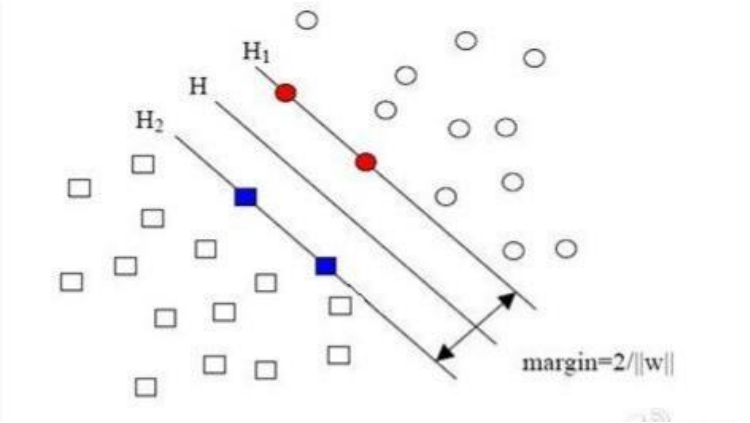

Figure 2. Optimal classification hyperplane

When using support vector machine modeling, for the case of linear indivisibility, the kernel function is generally introduced to solve the problem. The choice of kernel function has a great influence on the final effect of the support vector machine model. By using the appropriate inner product kernel function, the nonlinear separable problem in low-dimensional input space can be transformed into a linear separable problem in high-dimensional feature space [19].

3.4. Combined Model

For machine learning and deep learning, the effect of using a single model is often not as good as the combination of models [20]. The multi-model combination method can improve the accuracy of the model while increasing the diversity of the algorithm and reducing the generalization error. The main combination methods are Voting, Stacking and Blending.

After comparing the three model combination methods, this paper uses the most commonly used method in practice - Voting strategy. Using a fast and efficient weighted voting combination method, the three established models are tested directly on the test set data, and the predicted results are weighted. The better the performance of a single model, the higher the weight, and the formula for weighted voting is:

$$V_{组合}=m_1V_1+m_2V_2+\cdots+m_nV_n \tag{7}$$

$$m_1+m_2+\cdots+m_n=1 \tag{8}$$

Through the above weighted voting combination formula, the three different models selected in this paper are combined, and the personal credit data is analyzed to achieve the purpose of improving the accuracy of the evaluation model.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201911_5(11).0018

4. EXPERIMENTAL ANALYSIS

All the experiments in this paper are run on a computer with 2 CPU Cores i5-5350U, 1.80 GHz, and 8 Gbytes of memory. The code is built using the program toolkit such as Scikitlearn in Python 3.5. This paper uses the open real data of the world's largest P2P platform "Lending Club" to conduct experiments, including more than 6,000 data, 1237 features, through the model to find the bad debt rate of about 13%, and the importance of different characteristics for the model, such as Figure 3 shows.

In order to ensure the practical application of the model, this paper mainly uses the accuracy of the model, the AUC (Area under the Curve of ROC) value and F1 to evaluate the model. Accuracy is the ratio of the number of correctly predicted samples to the total number of samples [21]. The AUC value is the area under the ROC (receiver operating characteristic curve) curve. The ROC curve is usually used to describe the resolving power of the model. It is based on the true ratio of the ordinate and the false positive rate plotted on the abscissa. The AUC value is used because it can better and more intuitively reflect the results expressed by the ROC curve. The higher the prediction accuracy of the model, the larger the AUC value [22]. The F1 score is an index used to measure the accuracy of the two-category model in statistics. The value range [0~1], taking into account the accuracy and recall rate of the classification model, is a kind of model accuracy and recall rate. Weighted average [23].

Bad debt rate: 0.13467336683417086
Select feature total time : 71.123s

	Feature name	Importance
0	loan_amnt	0.151696
1	funded_amnt	0.0663415
2	funded_amnt_inv	0.051483
3	term	0.0453622
4	int_rate	0.00890354
5	installment	0.00406069
6	emp_title	0.00339042
7	emp_length	0.00219396
8	home_ownership	0.00180089
9	annual_inc	0.0010044
10	issue_d	0.000907957
11	purpose	0.000836258
12	zip_code	0.000412031
13	addr_state	0.000194988
14	dti	0.000184878
15	earliest_cr_line	0.000162209
16	mths_since_last_delinq	0.000139732
17	open_acc	0.000137936

Figure 3. Order of importance

4.1. Improvement of Feature Selection Method

In feature extraction, in order to improve the diversity of feature subspace, reduce the correlation between decision trees, improve the over-fitting phenomenon and make the generalization error as small as possible, improve the completely random sampling method, using partition sampling method.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201911_5(11).0018

After the improved random forest algorithm performs feature selection, firstly, the sample is sub-categorized on the basis of the original classification, assuming e is the number of sub-classes; then Fisher is used to analyze and verify the importance of the feature. The method of comparison. Suppose the data set has a total of m samples, belonging to M categories: the set of the ω class, using mω to represent the number of samples, using μc to represent the mean of the c-dimensional features in all samples, and μωc to represent the first in the ω class. The mean of the c-dimensional features, σ2ωc is used to represent the variance of the c-th dimension feature in the ω class. Therefore, the calculation method of the variance between classes is as follows:

$$S_B^c = \frac{1}{m}\sum_{\omega=1}^{M} m_\omega \left(\mu_{\omega c} - \mu_c\right)^2 \tag{9}$$

The intraclass variance can be expressed as:

$$S_W^c = \frac{1}{m}\sum_{\omega=1}^{M} m_\omega \sigma^2{}_{\omega c} \tag{10}$$

Then the Fisher ratio can be expressed as:

$$J_{\text{Fisher}}(c) = \frac{S_B^c}{S_W^c} \tag{11}$$

According to the Fisher comparison, the important feature sets F1, F2, ⋯, Fe of each subclass are calculated, and then the shared feature Fshare, the unimportant feature Fleft, and the private features of each subclass F1, F2, ⋯, Fe are obtained through the set operation. .

Common features can be expressed as:

$$F_{share} = F_1 \cap F_2 \cap \cdots \cap F_{share} \tag{12}$$

Unimportant features can be expressed as:

$$F_{left} = F_{all} - F_1 - F_2 - \cdots - F_{share} \tag{13}$$

The private characteristics of each subclass can be expressed as:

$$\begin{aligned} F_1 &= F_1 - F_{share} \\ &\vdots \\ &\vdots \\ F_e &= F_e - F_{share} \end{aligned} \tag{14}$$

According to the ratio, the features are randomly selected from {Fshare, F1, Fleft}, {Fshare, F2, Fleft}, ⋯, {Fshare, Fe, Fleft} to construct the feature subspace.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201911_5(11).0018

4.2. Grid Search Optimization Parameters

Grid search algorithms are often used to optimize model parameters, but for models with many parameters and a large range of values, using grid search can be time consuming. In order to improve efficiency and ensure optimization, this paper proposes an improved grid search algorithm.

In order to save training time, a method of large-scale optimization and small-scale solution is proposed. First, use a longer spacing to divide a sparse grid over a larger range, and perform a large-scale optimization pre-search in the first step to find the range of areas where the best advantage lies; then use a shorter spacing. Divide the fine mesh in the region where the most advantageous is located, perform the small-scale solution in the second step, and search for the optimal solution again. Repeat the above steps until the target function variation or grid spacing is less than the given value.

Under the premise of ensuring the performance of the random forest algorithm is improved, the accuracy of each decision tree and the diversity of the tree should also be considered [24]. In this paper, when using the improved grid search algorithm to find the optimal parameters for random forests, the out-of-bag score is the generalization ability of the model as the objective function value. The specific steps to improve the grid search method are as follows:

(1) Determine the parameters that need to be optimized, the corresponding parameters are the points on the grid, and set a longer grid spacing to divide the grid;

(2) Traversing each set of parameters on the grid once, using the out-of-bag data scores in the random forest algorithm to evaluate the classification error, and performing preliminary large-scale optimization;

(3) Select a group of data with the highest score outside the bag, that is, the smallest error, and continue to shorten the grid spacing for parameter optimization. If the out-of-bag score or grid spacing meets the requirements, the result is output, otherwise the above steps are repeated.

4.3. Random Forest Modeling

When using random forest algorithm modeling, using improved feature extraction method and partition sampling according to the importance degree of features, the diversity of feature subspace can be improved, and the correlation between base classifiers can be effectively reduced. The voting method of the traditional random forest algorithm is slightly simple. In order to enhance the practicability of the voting method, improve the accuracy of the random forest algorithm, improve the voting method of the random forest algorithm, and introduce a weighted integrated voting method to increase the classification performance. The weight of a good decision tree increases its "discourse power", which makes the voting method more reasonable. Express the weight of each decision tree as follows:

Among them: the probability that classifier Ti divides sample X into class l is represented by Pil, the number of categories is represented by c, and the number of split subsets is represented by l, then the weight is:

$$\omega_i = \frac{\sum_l P_{il}}{\sum_{il} P_{il}} \qquad l = 1,2,\cdots, \quad c \tag{15}$$

Then calculate the confidence of each category:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201911_5(11).0018

$$u_l(x)=\frac{1}{N}\sum_{i=1}^{N}\omega_i P_{il}(x) \quad l=1,2,\cdots, \ \text{c} \tag{16}$$

In the modeling of parameter tuning, the improved grid search algorithm is used to quickly and efficiently coordinate the parameters. Taking the maximum depth md of the decision tree and the minimum number of samples mn required by the internal nodes to be used as an example, the improved grid search algorithm is used to optimize the random forest parameters. First, when the preliminary search for large-pitch is performed, the value range of the maximum depth md of the decision tree is determined as 1<md<30, and the grid spacing is set to 5; the range of values of the minimum number of samples required for internal node subdivision is determined. For 10<mn<201, the grid spacing is set to 50. The search results are shown in Figure 4:

```
([mean: 0.73715, std: 0.02128, params: {'max_depth': 1, 'min_samples_split': 10},
  mean: 0.73715, std: 0.02128, params: {'max_depth': 1, 'min_samples_split': 60},
  mean: 0.73715, std: 0.02128, params: {'max_depth': 1, 'min_samples_split': 110},
  mean: 0.73715, std: 0.02128, params: {'max_depth': 1, 'min_samples_split': 160},
  mean: 0.78927, std: 0.01668, params: {'max_depth': 6, 'min_samples_split': 10},
  mean: 0.78933, std: 0.01669, params: {'max_depth': 6, 'min_samples_split': 60},
  mean: 0.78857, std: 0.01893, params: {'max_depth': 6, 'min_samples_split': 110},
  mean: 0.78698, std: 0.01623, params: {'max_depth': 6, 'min_samples_split': 160},
  mean: 0.79653, std: 0.01515, params: {'max_depth': 11, 'min_samples_split': 10},
  mean: 0.79259, std: 0.01788, params: {'max_depth': 11, 'min_samples_split': 60},
  mean: 0.79221, std: 0.01729, params: {'max_depth': 11, 'min_samples_split': 110},
  mean: 0.79405, std: 0.01550, params: {'max_depth': 11, 'min_samples_split': 160},
  mean: 0.79687, std: 0.01347, params: {'max_depth': 16, 'min_samples_split': 10},
  mean: 0.78993, std: 0.01968, params: {'max_depth': 16, 'min_samples_split': 60},
  mean: 0.79049, std: 0.01623, params: {'max_depth': 16, 'min_samples_split': 110},
  mean: 0.79242, std: 0.01626, params: {'max_depth': 16, 'min_samples_split': 160},
  mean: 0.79638, std: 0.01379, params: {'max_depth': 21, 'min_samples_split': 10},
  mean: 0.79006, std: 0.01984, params: {'max_depth': 21, 'min_samples_split': 60},
  mean: 0.79028, std: 0.01629, params: {'max_depth': 21, 'min_samples_split': 110},
  mean: 0.79242, std: 0.01626, params: {'max_depth': 21, 'min_samples_split': 160},
  mean: 0.79638, std: 0.01379, params: {'max_depth': 26, 'min_samples_split': 10},
  mean: 0.79006, std: 0.01984, params: {'max_depth': 26, 'min_samples_split': 60},
  mean: 0.79028, std: 0.01629, params: {'max_depth': 26, 'min_samples_split': 110},
  mean: 0.79242, std: 0.01626, params: {'max_depth': 26, 'min_samples_split': 160}],
 {'max_depth': 16, 'min_samples_split': 10},
```

Figure 4. Optimization of large spacing parameters

It can be seen from the above figure that the preliminary search result is optimally the maximum depth of 16, and the minimum number of samples is 10. When the maximum depth of the decision tree is md =16, and the internal node divides the minimum number of samples required by mn =10, the extra-bag score at this time is calculated to be 0.788, which increases proportionally, so the mesh is subdivided at a small pitch and the decision is made. The maximum depth md of the tree is determined to be 10<md<20, and the grid spacing is set to 2; the range of the minimum number of samples required for internal node subdivision is determined as 5<mn<20, and the grid spacing is set. Set to 2. The search results are shown in Figure 5:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201911_5(11).0018

```
mean: 0.80674, std: 0.01095, params: {'max_depth': 9, 'min_samples_split': 13},
mean: 0.81117, std: 0.01270, params: {'max_depth': 11, 'min_samples_split': 5},
mean: 0.81117, std: 0.01270, params: {'max_depth': 11, 'min_samples_split': 7},
mean: 0.81117, std: 0.01270, params: {'max_depth': 11, 'min_samples_split': 9},
mean: 0.81117, std: 0.01270, params: {'max_depth': 11, 'min_samples_split': 11},
mean: 0.81117, std: 0.01270, params: {'max_depth': 11, 'min_samples_split': 13},
mean: 0.80901, std: 0.01381, params: {'max_depth': 13, 'min_samples_split': 5},
mean: 0.80901, std: 0.01381, params: {'max_depth': 13, 'min_samples_split': 7},
mean: 0.80901, std: 0.01381, params: {'max_depth': 13, 'min_samples_split': 9},
mean: 0.80901, std: 0.01381, params: {'max_depth': 13, 'min_samples_split': 11},
mean: 0.80901, std: 0.01381, params: {'max_depth': 13, 'min_samples_split': 13},
mean: 0.80809, std: 0.01347, params: {'max_depth': 15, 'min_samples_split': 5},
mean: 0.80809, std: 0.01347, params: {'max_depth': 15, 'min_samples_split': 7},
mean: 0.80809, std: 0.01347, params: {'max_depth': 15, 'min_samples_split': 9},
mean: 0.80809, std: 0.01347, params: {'max_depth': 15, 'min_samples_split': 11},
mean: 0.80809, std: 0.01347, params: {'max_depth': 15, 'min_samples_split': 13},
mean: 0.80911, std: 0.01479, params: {'max_depth': 17, 'min_samples_split': 5},
mean: 0.80911, std: 0.01479, params: {'max_depth': 17, 'min_samples_split': 7},
mean: 0.80911, std: 0.01479, params: {'max_depth': 17, 'min_samples_split': 9},
mean: 0.80911, std: 0.01479, params: {'max_depth': 17, 'min_samples_split': 11},
mean: 0.80911, std: 0.01479, params: {'max_depth': 17, 'min_samples_split': 13},
mean: 0.80925, std: 0.01492, params: {'max_depth': 19, 'min_samples_split': 5},
mean: 0.80925, std: 0.01492, params: {'max_depth': 19, 'min_samples_split': 7},
mean: 0.80925, std: 0.01492, params: {'max_depth': 19, 'min_samples_split': 9},
mean: 0.80925, std: 0.01492, params: {'max_depth': 19, 'min_samples_split': 11},
mean: 0.80925, std: 0.01492, params: {'max_depth': 19, 'min_samples_split': 13},
mean: 0.80900, std: 0.01471, params: {'max_depth': 21, 'min_samples_split': 5},
mean: 0.80900, std: 0.01471, params: {'max_depth': 21, 'min_samples_split': 7},
mean: 0.80900, std: 0.01471, params: {'max_depth': 21, 'min_samples_split': 9},
mean: 0.80900, std: 0.01471, params: {'max_depth': 21, 'min_samples_split': 11},
mean: 0.80900, std: 0.01471, params: {'max_depth': 21, 'min_samples_split': 13},
mean: 0.80900, std: 0.01471, params: {'max_depth': 23, 'min_samples_split': 5},
mean: 0.80900, std: 0.01471, params: {'max_depth': 23, 'min_samples_split': 7},
mean: 0.80900, std: 0.01471, params: {'max_depth': 23, 'min_samples_split': 9},
mean: 0.80900, std: 0.01471, params: {'max_depth': 23, 'min_samples_split': 11},
mean: 0.80900, std: 0.01471, params: {'max_depth': 23, 'min_samples_split': 13}],
{'max_depth': 11, 'min_samples_split': 5},
0.8111674625159752)
```

Figure 5. Small pitch parameter optimization

The search result of the subdivision mesh is that the maximum depth of the decision tree is md =16, and when the internal node is further divided into the minimum number of samples mn = 5, the extra-bag score at this time is calculated to be 0.862, which has become stable. Therefore, the values of md and mn can be preliminarily determined. In addition, other parameters of other models in this paper are also determined by the improved grid search algorithm, and the target points are determined by large-scale search. Finally, the classification results of the random forest model are: the accuracy rate is 93%, the F1 score is 0.79, the AUC value is 0.92, and the ROC curve is shown in Fig. 6.

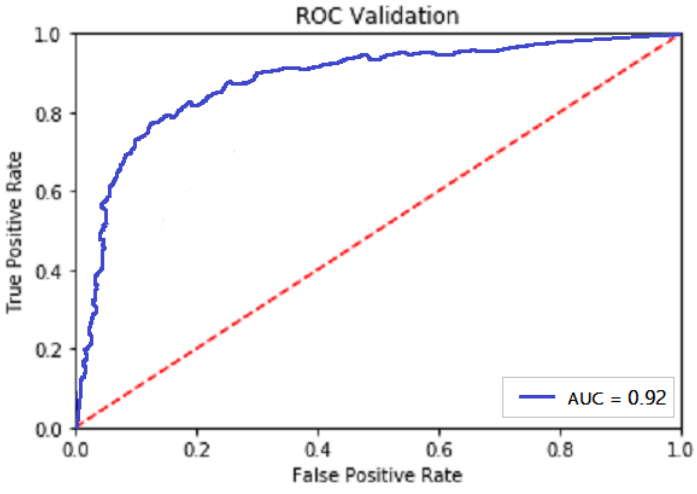

Figure 6. Random forest ROC curve

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201911_5(11).0018

4.4. LightGBM Modeling

When modeling with the LightGBM algorithm, because the algorithm performs very well in all aspects, there is no need to make excessive improvements. When processing data, the improved feature extraction method is used for partition sampling. Because the algorithm has many parameters similar to the random forest algorithm, but the overall robustness is better, so only the improved grid search algorithm is needed, and the parameters of the LightGBM model are optimized by the method of random forest tuning mentioned above.

Finally, the classification results of the LightGBM model are: accuracy rate of 95%, F1 score of 0.81, AUC value of 0.94, ROC curve diagram shown in Figure 7. The model performs better than random forests.

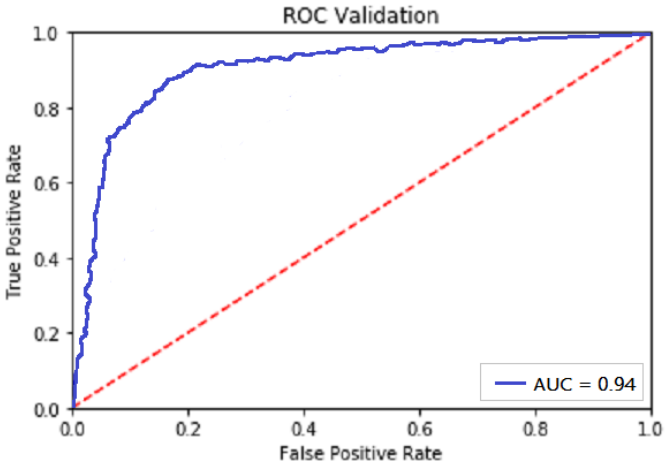

Figure 7. LightGBM ROC curve

4.5. Support Vector Machine Modeling

In the modeling using the support vector machine algorithm, in order to make the model have better classification ability, Gaussian kernel function (also called radial basis and function or RBF kernel) is used as the inner product kernel function of the support vector machine.

$$k\left(x_i, x_j\right) = \exp\left(-\frac{\| x_i - x_j \|^2}{2\sigma^2}\right) \tag{17}$$

Compared with other functions, it has the following advantages: the RBF kernel function can map a sample to a higher-dimensional space, and the linear kernel function is a special case of RBF, that is, if RBF is considered, there is no need to consider linear kernel. Function; compared with polynomial kernel function, RBF needs to determine fewer parameters, the number of kernel function parameters directly affect the complexity of the function; when the order of polynomial is higher, the element value of the kernel matrix will tend to infinity or infinity Using RBF will reduce the computational difficulty of numerical values [25].

In the modeling of parameter tuning, an improved grid search algorithm is used to select the optimal parameters and penalty factor C of the Gaussian kernel function. Finally, the classification results of the support vector machine model are as follows: the accuracy is 91%, the F1 score is 0.76, the AUC value is 0.89, and the ROC curve is shown in Fig. 8. The performance of the model is not as good as the random forest.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201911_5(11).0018

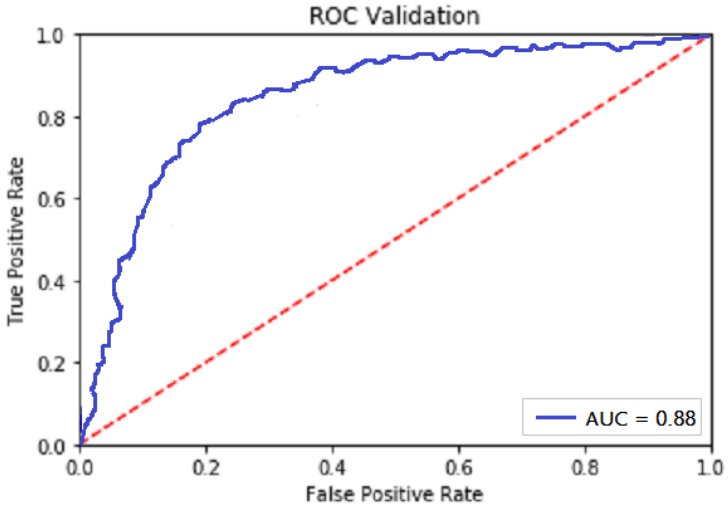

Figure 8. SVM ROC curve

4.6. Model fusion results

The final classification results of three different models of random forest, LightGBM and support vector machine are combined by the weighted voting combination formula above. The experimental results are shown in Table 1. Using grid optimization, the optimal weight ratio of the three models of random forest, LightGBM and support vector machine is 1:1:8. Finally, the accuracy of the combined model is 95%, the F1 score is 0.84, the AUC value is 0.96, and the ROC curve is shown in Figure 9.

Table 1. Experimental results

Mode	Accuracy	F1 score	AUC
RF	93%	0.79	0.92
LightGBM	95%	0.81	0.94
SVM	91%	0.76	0.89
Combined model	95%	0.84	0.96

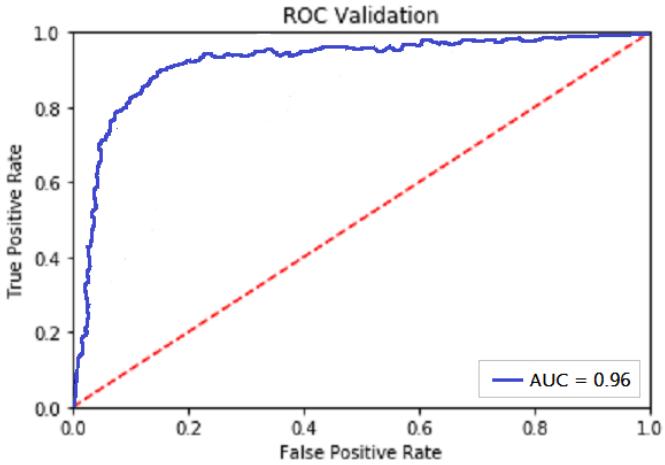

Figure 9. Combined model ROC curve

As shown in Figure 10, the dotted line represents the combined model, the solid line represents LightGBM, the dashed line represents random forest, and the dotted line represents SVM. Comparing the ROC curve and the AUC (area under the curve) values of the combined model and the three single models, the ROC curve of the combined model is significantly higher

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201911_5(11).0018

than the other three single models, and the AUC value is larger, indicating that the classification model has better classification performance.

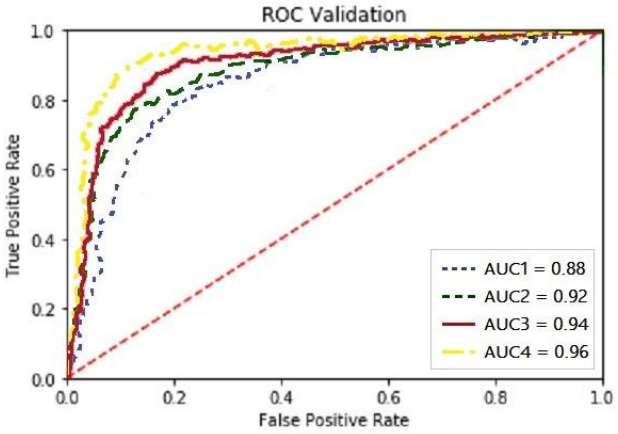

Figure 10. ROC curve comparison chart

In order to increase the contrast, the AdaBoost and XGBoost algorithms were introduced. The experimental results are shown in Table 2, and the ROC curve comparison chart is shown in Fig. 11. The solid line represents the combined model, the dotted line represents XGBoost, and the dotted line represents AdaBoost. By comparing experimental data and graphs, it can be seen that the combined model has better performance in all aspects.

Table 2. Experimental results

Mode	Accuracy	F1 score	AUC
Combined model	95%	0.84	0.96
XGBoost	87%	0.68	0.82
Adaboost	84%	0.62	0.72

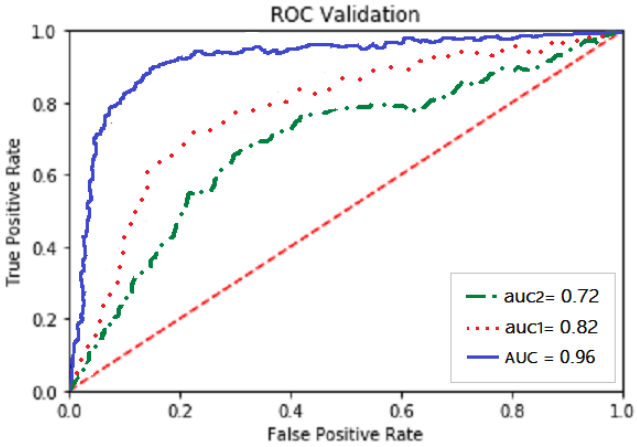

Figure 11. ROC curve comparison chart

5. CONCLUSION

The high efficiency and stability of the multi-model combination method in solving the classification problem make many fields achieve better results after introducing the combined model method. Personal credit assessment is a key technical issue in the financial field. This

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201911_5(11).0018

paper uses the combined model approach to achieve better results in the field of personal credit assessment, mainly due to the following improvements:

(1) By combining three single models, the credit evaluation model of multi-model combination has better stability and diversity, and the comprehensive performance of model performance is better.

(2) In the feature selection, Fisher's comparison is introduced to quantify and analyze the importance of features, which improves the diversity of feature subspace, overcomes the disadvantages of completely random sampling of features in credit evaluation, and reduces the relationship between decision trees. Correlation.

(3) The grid search algorithm is improved, which saves a lot of time required to traverse all grid points, and improves the efficiency of grid search algorithm optimization model parameters.

The multi-model combination credit evaluation method has been applied to the field of personal credit assessment to obtain better results, which provides a feasible reference for credit evaluation.

REFERENCES

[1] Gahlaut A , Tushar, Singh P K . Prediction analysis of risky credit using Data mining classification models[C]International Conference on Computing. IEEE Computer Society, 2017.

[2] Bequé, Artem, Lessmann S . Extreme Learning Machines for Credit Scoring: An Empirical Evaluation[J]. Expert Systems with Applications, 2017:S0957417417303718.

[3] Wang Yang, Jin Junshi, Sun Meifeng, et al. Application of Combination Algorithm Based on Multiple Classifiers in Personal Credit Evaluation[J]. Information Technology, 2016(6).

[4] Tao Li, Xiang Li, Lina Wang, Yongjun Ren, Tingyu Zhang, Meichen Yu. Multi-Model Ensemble Forecast Method of PM2.5 Concentration Based on Wavelet Neural Networks[C] 2018 1st International Cognitive Cities Conference (IC3). IEEE,2018

[5] Zheng Zhang, Zhitao Liu, Hongye Su, Weijie Mao, Longhua Ma. Earth Pressure Multipoint Prediction for EPS Shield Based on Multi-Model Ensemble[C]. 2018 Chinese Automation Congress (CAC). IEEE,2018

[6] Ren Xiao, Jiang Ming-hui, Che Kai, Wang Shang. The research on methods of personal credit scoring combined model selection based on optimized index system[J]. Jouranl of Haerbin Institute of Technology. 2016, 48(5):67-71.

[7] Weidong H , Xiangwei Z , SuQingling. Research on application of personal credit scoring based on BP-logistic hybrid algorithm[C]// International Conference on Computer Application & System Modeling. IEEE, 2010.

[8] Jiang Minghui, Xu Pei, Han Yutong. Research on Personal Credit Score Based on Optimized CBR[J]. China Soft Science, 2014(12): 148-156.

[9] He Ming, Sun Jianjun, Cheng Ying. A Review of Text Classification Based on Naive Bayes[J]. Information Science, 2016, V34(7): 147-154.

[10] Rojarath A , Songpan W , Pong-Inwong C . Improved ensemble learning for classification techniques based on majority voting[C]// IEEE International Conference on Software Engineering & Service Science. IEEE, 2017.

[11] Vishwanath D , Gupta S . Adding CNNs to the Mix: Stacking models for sentiment classification[C]// India Conference. IEEE, 2017.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.201911_5(11).0018

[12] Lin Fei, Zhang Zhan. Prediction of large-scale network course learning results based on downsampling heap model[J]. Journal of Computer Applications and Software, 2018, v.35(07):137-143.)

[13] Gao Yuan, Liu Baiwei. Research on title classification algorithm based on integrated learning[J]. Journal of Computer Applications, 2017, 34(4): 1004-1007.

[14] Lu S , Hwang Y , Khabibrakhmanov I , et al. Machine Learning Based Multi-Physical-Model Blending for Enhancing Renewable Energy Forecast -- Improvement via Situation Dependent Error Correction[C]// Control Conference. IEEE, 2015.

[15] Breiman L. Random Forests [J]. Machine Learning, 2001, 45(1) : 5-32.

[16] Xiaojun M, Jinglan S, Dehua W, et al. Study on A Prediction of P2P Network Loan Default Based on the Machine Learning LightGBM and XGboost Algorithms according to Different High Dimensional Data Cleaning[J]. Electronic Commerce Research and Applications,2018.

[17] Vapnik V. The nature of statistical learning theory [M]. New York:Springer,1995.

[18] Vapnik V. Statistical learning theory[M]. New York:Springer,1999.

[19] Xiao Zhi, Li Wenjuan. Personal Credit Assessment Based on Principal Component Analysis and Support Vector Machine[J].2010(3) : 71-74.

[20] Dudek G . Heterogeneous ensembles for short-term electricity demand forecasting[C]// International Scientific Conference on Electric Power Engineering. IEEE, 2016.

[21] Bernard S, Heutte L, Adam S.Influence of hyperparameters on random forest accuracy[C]Proceedings of the 8th International Workshop on Multiple Classifier System.Berlin, Heidelberg: Springer-Verlag, 2009: 171-180.

[22] Adnan M N, Islam M Z.Optimizing the number of trees in a decision forest to discover a subforest with high ensemble accuracy using a genetic algorithm[J].Knowledge-Based Systems, 2016, 110: 86-97.

[23] Li Z W, Xin X P. Estimating grassland LAI using the Random Forests approach and Landsat imagery in the meadow steppe of Hulunber, China [J]. Journal of Integrative Agriculture, 2017, 16(2) : 286-297.

[24] Malhotra R, Jha M, Poss M. A random forest classifier for detecting rare variants in NGS data from viral populations[J].Computational and Structural Biotechnology Journal, 2017, 35(15) : 388-395.

[25] You Jinming, Wang Junxi, Tang Wei, et al. Research on real-time accident risk of expressway based on support vector machine[J]. Journal of Tongji University(Natural Science), 2017, 45(3): 355-361.