

Imbalanced credit risk prediction based on data fillers and modified balanced random forest improved by Bayesian optimization

Hongyu Zhang ^{1, a}, Zhenjun Ye ^{1, 2}

¹ Mathematics and Physics School, North China Electric Power University, China;

² Business School, QingHai University of Technology.

^a 294394340@qq.com

Abstract. Based on the distribution characteristics of financial big data, credit risk prediction models often face some problems, such as unbalanced data distribution and difficult data preprocessing process. High-precision models are often accompanied by low model efficiency. Therefore, this paper constructs a complete non-equilibrium credit risk prediction model, namely BO-PBRF, and improves the algorithm to deal with common problems in financial data. In the data preprocessing stage, two missing value fillers are generated according to the original data to facilitate the subsequent new data processing. In the modeling stage, we improve the balanced random forest algorithm, so that the model can not only deal with unbalanced data sets, but also suitable for the background of the explosive development of financial big data, and improve the operation speed of the model. In addition, in the process of establishing the model, we add the Bayesian optimization algorithm to further improve the accuracy of the model, especially in the prediction of default loans. In order to verify the effectiveness of the model proposed in this paper, in the empirical research, we select the credit data from the real world, and compare the model proposed in this paper with the previous models. The experimental results show that the proposed model has the best prediction performance for default data.

Keywords: Credit risk evaluation, Imbalanced classification, Bayesian optimization, Data fillers.

1. Introduction

With the rapid development of economy and the continuous expansion of credit demand, online loan services provided by financial institutions are also growing rapidly. Compared with traditional loan business, consumer loans provide borrowers with more convenient services. As long as its credit value meets the pre-set requirements of the financial institution, it can obtain a more efficient loan method [1]. However, lending activities are often accompanied by potential default risks. Nowadays, data mining and machine learning are effective data processing technologies that are widely used in various fields [2]. Therefore, how to use these technologies to build effective credit scoring models is an important research topic in the field of credit risk.

Recent research shows that ensemble classifiers have better credit assessment performance than single classifiers [3]. However, establishing an ideal credit risk assessment model is still a major challenge due to the high imbalance of data: the number of loan businesses that are repaid on time is usually much larger than the number of loans that are in default. Previous research has solved the class imbalance problem through various methods: methods such as sampling technology, synthetic minority oversampling technology (SMOTE) [4], and cost-sensitive learning [5]. However, these methods have an obvious shortcoming, that is, while improving the identification rate of defaulted loans, the probability of misclassifying non-defaulted loans as defaulted loans also increases. Although the actual business focuses more on high-risk loan applications, consumer lending is a profit-oriented trading activity that should be balanced between avoiding losses caused by high-risk borrowers and accepting the benefits of potentially good loan business. In addition to the accuracy of different dimensions of the model, we also consider the modeling time problem and over-fitting problem, which respectively affect the efficiency of the company's business processing and the accuracy of identifying whether new loan business is in default. Therefore, in this unbalanced actual

business scenario, it is necessary to develop a credit risk assessment model that has both excellent default loan identification performance and good overall classification performance.

In this paper, we propose a new ensemble model for imbalanced credit risk assessment based on the balanced random forest model (BRF) to address the above challenges. In the BRF model, each tree is built based on undersampled balanced data. In the original BRF model, considering that the overall data volume is not large, if only the down-sampling method is used to process the data to achieve balance, most types of data information will be lost. Therefore, [6] recommends not performing the pruning process on the established BRF model. Faced with the rapid growth of today's data volume, useful information for most types of data tends to overflow. Therefore, we add a pruning process in BRF model. In addition, we further optimize the PBRF model through Bayesian optimization algorithm and achieve ideal prediction results.

2. Related Work

2.1 Unbalanced Data Learning Methods

In many real-world classification prediction problems of actual business, such as credit scores, the number of observations belonging to the positive class (minority class) is significantly smaller than the number of negative class (majority class) observed. This class imbalance problem can adversely affect machine learning classification algorithms because most of them minimize the accuracy-based objective function during the training phase. Therefore, classification results on an unbalanced dataset may be biased in favor of the majority class [7].

Data-level technology is the simplest and most basic way to solve the problem of class imbalance, and the number of most class samples and minority samples is basically the same through resampling. The resampling method can be subdivided into downsampling of most samples and oversampling of minority samples. The simplest resampling method: random undersampling (RUS) randomly selects and discards the majority of samples; Random oversampling (ROS) randomly replicates minority samples, but both methods have underfitting and overfitting issues, respectively. In order to solve the problems that arise in random sampling, Chawla et al. propose a new model called Synthetic Minority Oversampling Technique (SMOTE), which uses the nearest neighbor of minority sample data to generate synthetic minority samples to balance the data distribution [8]. For Algorithmic-level approaches, many machine learning classification algorithms assume that the cost of misclassification of minority and majority classes in an unbalanced dataset is unequal [9,10]. In addition, other approaches attempt to modify algorithms to improve their performance in an imbalanced setting. For instance, modified support vector machine (SVM) [11], logistic regression (LR) [12], decision tree (DT) [13], artificial neural network (ANN) [14], and several other studies are very popular artificial intelligence methods for the construction of prediction models. Another approach to the algorithmic level category is ensemble learning methods, which are motivated by the fact that different classifiers have advantages and disadvantages due to inherent differences in classifying the same data. Therefore, they can compensate for each other's shortcomings when making the final prediction [15].

2.2 Credit Risk Prediction Model

With the development of technology and the expansion of data volume, credit assessment models based on machine learning and deep learning are gradually applied in the financial field. The machine learning models initially applied in the field of credit risk assessment were all based on a single classifier, such as using k-nearest neighbors (K-NN) [16] and a DT model combined with rough sets theory [17] to establish a credit assessment model. However, many studies have shown that in credit evaluation, the classification performance of ensemble models is better than that of a single classifier. However, some of the above models do not consider the class imbalance characteristics of credit data. Although these models can achieve good prediction accuracy on the overall data, they perform poorly in identifying minority class samples. With the expansion of data

volume and latitude, and the development of unstructured databases, deep learning is gradually applied to credit risk prediction. However, the deep learning model has poor interpretability [18].

3. Methodology

3.1 Balanced Random Forest

The balanced random forest (BRF) [6] algorithm builds a decision tree based on the balanced data after undersampling, and its theoretical method is described below: Firstly, in the process of each round of bootstrap, random undersampling technique is used to select some samples from a few categories, and the same amount of samples from most categories are also selected to form a balanced data set. Secondly, building the largest decision tree based on the data set without pruning. The decision tree is constructed by CART algorithm and adjusted accordingly: on each node, instead of searching all the attributes to get the best split variable, one attribute is randomly selected as the split variable. Thirdly, Multiple decision trees generated by this method form a balanced random forest. The new data are classified and identified by this balanced random forest classifier, and the final classification results can be obtained by voting. This algorithm can handle unbalanced data better than the Random Forest algorithm.

3.2 Bayesian Optimization

Bayesian Optimization considers previous parameter information to constantly update prior information, and has fast speed. These advantages make it better suited for credit data. Bayesian optimization consists of two parts [19] : a Probability Surrogate model used to establish an objective function and an Acquisition Function used to determine which point to sample.

Here we use the Tree-structured Parzen Estimator Approach (TPE) as the Probability Surrogate model. It divides the observation results according to grades and performs probability distribution modeling in segments. Next is an overview of the theory of Bayesian optimization.

According to

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \quad (1)$$

we can decompose the desired $p(y|x)$ into $p(x|y)$ and $p(y)$. TPE calculates $p(x|y)$ as follows:

$$p(x|y) = \begin{cases} l(x), y < y^* \\ g(x), y > y^* \end{cases} \quad (2)$$

where y^* is a threshold, which is determined by the super parameter γ , that is, $p(y < y^*) = \gamma$.

By (2), we can get:

$$p(x) = \gamma l(x) + (1 - \gamma)g(x) \quad (3)$$

The formula for Expected Improvement (EI) is as follows:

$$EI_{y^*}(x) = \int_{-\infty}^{+\infty} \max(y^* - y, 0) p_M(y|x) dy \quad (4)$$

where EI indicates expectation, and that expectation is a function of x . When x is given, $EI(x)$ is the average increase of y relative to the threshold y^* , and it is used to evaluate the quality of the hyperparameter x . The next step to find the hyperparameter is:

$$x_{new} = \arg \max_x EI_{y^*}(x) \quad (5)$$

According to (1)-(3), (4) can be deduced as follows:

$$EI_{y^*}(x) = \frac{\int_{-\infty}^{y^*} (y^* - y) p(y) dy}{\gamma + (1 - \gamma) \frac{g(x)}{l(x)}} \quad (6)$$

that is: $EL_y(x) \propto \left(\gamma + (1-\gamma) \frac{g(x)}{l(x)} \right)^{-1}$. So, the x we're looking for is the x that maximizes the ratio $\frac{g(x)}{l(x)}$.

4. Empirical Setup

4.1 Data Description

The data set used in this article is selected from the data set of Lending Club company, sourced from the Kaggle competition platform, which provides a variety of information about past loan applicants and their loan status. We mainly selects the loan data accepted by the company in 2018. The dataset has 495242 loan records and 151 columns with a target value named “loan status”. Here we extract 60% of the entire dataset for training and use the remaining 40% for testing. The subsequent data processing is performed on both the training set and the test set respectively.

4.2 Data Preprocessing

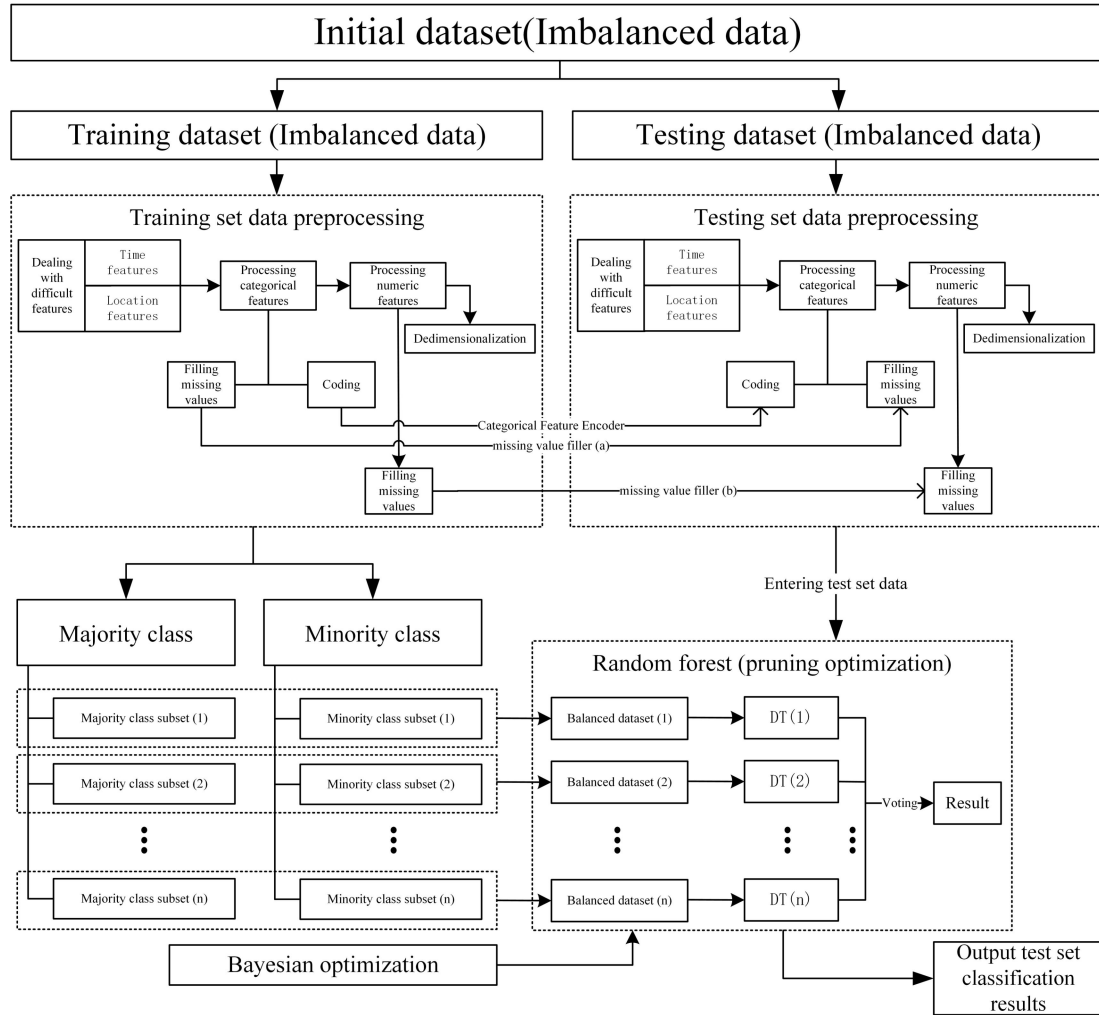


Fig. 1 Framework and calculation process of prediction model

The research method of financial credit risk mainly includes two parts: features preprocessing and modeling analysis. Fig. 1 shows the entire flowchart of these two steps.

In order to standardize the dataset and extract useful information from the time series data, this paper derives the feature features of the training set data and the test set data respectively--extracting the new features of the year, month, and day of the week. For location feature “addr_state”, we divided it into: west, central west, southwest, southeast, and northeast five regions. Next, we deal with classification features and continuous features. First, we propose a

“missing value filler a ” according to the categorical variables of the training set, and then the missing values of the categorical variables of the training set and the test set are filled by this filler. The classification data of the training set is sequentially encoded to generate a “categorical feature encoder”, and finally this categorical feature encoder is used to encode the categorical variables in the test set sequentially. Because in practice, the test set data is constantly updated and added, in this case, the test set must be processed according to the mode and encoder of the training set data. For continuous features, the focus is on missing value processing of the data. Here, we use mean to fill in the missing values, and in the same way as the processing of categorical variables, a “missing value filler b ” is first generated according to the mean of the training set data, and then the missing value of the continuous variable in the test set is filled with this filler. In addition, for category markers, we denote defaulted loans as 1 (minority class) and non-default loans as 0 (majority class).

4.3 Evaluation Measures

Since the number of instances of defaulted loans is much lower than that of non-default loans, the overall accuracy does not fully reflect the predictive performance of the model, especially for defaulted loans. For this reason, we use the area under the ROC curve (AUC), G-means, Maj_Recall, Min_Recall as measures to evaluate the proposed model performance. The confusion matrix of the Table 1 is the basis for calculating various evaluation metrics. For Maj Recall and Min Recall, we let them represent Sensitivity and Specificity respectively.

Table 1. Confusion matrix			
		Prediction class	
		Positive	Negative
Actual class	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

In addition, Bayesian optimization also requires the selection of hyperparameter indicator. Based on the debugging results of the PBRF model, we select the indicators, that is, the sequence x in (6) contains the following values: `n_estimators`, `max_depth`, `max_features`, `min_samples_split`, `min_samples_leaf`. After the Bayesian optimization, the values of these hyperparameters are: 104, 35, 18, 7, 2.

5. Experimental Results and Analysis

In order to verify the effectiveness of our model, we establish two types of classifier models and compare them with the BO-PBRF model designed in this paper. The first classifier models directly on the imbalanced dataset. We select the classifier model which has been proved to be effective in recent years, such as LR, DT, RF, XGBoost, and LightGBM. The comparison results are shown in Table 2. The second category is the classifier modeled after data balancing. We also select several classical unbalanced data processing classifiers to compare with the model designed in this paper, such as RUSBoost, EasyEnsemble, SMOTEBoost, SMOTERF, asymmetric AdaBoost (AsymBoost) and AdaCost. The comparison results are shown in Table 3.

Table 2. Compare to models without a balancer				
Classifier	Evaluation measures			
	Min_Recall	Maj_Recall	AUC	G-mean
BO-PBRF	0.97190	0.96485	0.99476	0.96837
LR	0.63863	0.99783	0.94215	0.79828
DT	0.86714	0.99449	0.93081	0.92863
RandomForest	0.84276	0.99992	0.98955	0.91799

XGBoost	0.86416	0.99929	0.99586	0.92927
LightGBM	0.85318	0.99873	0.99285	0.92309

Table 3. Compare to models including a balancer

Classifier	Evaluation measures			
	Min_Recall	Maj_Recall	AUC	G-mean
BO-PBRF	0.97190	0.96485	0.99476	0.96837
RUSBoost	0.95236	0.96853	0.99318	0.96041
EasyEnsemble	0.91701	0.97587	0.98897	0.94598
SMOTEBoost	0.89096	0.99043	0.99192	0.93938
SMOTERF	0.85114	0.99913	0.99363	0.92217
AsymBoost	0.90435	0.99223	0.99432	0.94728
AdaCost	0.95106	0.85116	0.96322	0.89972

According to the tables obtained here, we have the following findings. First of all, as shown in Table 2 and Table 3, the performance of single classifier and integrated classifier on Maj_Recall is almost better than that of balanced integrated classifier. However, on Min_Recall, they are far inferior to balanced integrated classifiers. This is because for unbalanced data, the standard single classifier and integrated classifier are mainly based on most classes, and assume that the samples between the two classes are evenly distributed. Different from standard single classifier and standard ensemble classifier, balanced ensemble classifier uses resampling method to generate training subset with the same number of majority and minority class samples, or gives different training weights to majority and minority class data. Therefore, the balanced ensemble classifier has a good ability to identify a small number of samples. Then, the classification performance of different balanced integrated classifiers is compared on unbalanced data sets. SMOTERF, SMOTEBoost, EasyEnsemble and AsymBoost are better than other balanced integrated classifiers on Maj_Recall. However, on Min_Recall, they are much lower than other balanced integrated classifiers. For SMOTERF and SMOTEBoost, the reason is that SMOTE method synthesizes new minority samples and does not take into account the data distribution of minority classes, so some newly synthesized samples are hardly helpful to improve the recognition ability of SMOTERF and SMOTEBoost to minority samples. For AsymBoost, because different weights are set for positive and negative samples, these weight parameters need to be debugged many times, so it is very difficult to establish a good model. Through Table 2 and Table 3, compared with all other models, our proposed BO-PBRF has the best performance on Min_Recall, which is the focus of this article. In addition, BO-PBRF is almost optimal on both G-means and AUC. This means that BO-PBRF has a strong ability to identify a small number of samples in unbalanced data sets, and the overall classification performance is the best.

6. Conclusion

In order to improve credit risk prediction models, we propose a new integrated model to predict the credit risk of loan business, namely BO-PBRF. This model first proposes a solution to the missing value problem of financial big data: according to the original data Generates two missing value fillers for handling missing values on new data. During the modeling process, this article improved the balanced random forest algorithm and pruned the random forest tree building process. The model can not only handle unbalanced data sets well, but is also suitable for the background of the explosive development of financial big data today, improving the model calculation speed. Finally, we selected credit data in the real world for empirical research, compared the model designed in this article with 11 existing models (including single models, integrated models, balanced integrated models), and conducted model evaluation from 5 dimensions, verifying the effectiveness of the model in predicting imbalanced credit risk.

References

- [1] Coser, A., Maer-matei, M.M., Albu, C.: Predictive models for loan default risk assessment. *Economic Computation & Economic Cybernetics Studies & Research* 53(2) (2019)
- [2] Korotaev, P., Yanilkin, A.: Steels classification by machine learning and calphad methods. *Calphad* 82, 102587 (2023)
- [3] Xia, Y., Liu, C., Da, B., Xie, F.: A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications* 93, 182–199 (2018)
- [4] Fisher, R.A.: The use of multiple measurements in taxonomic problems. *Annals of eugenics* 7(2), 179–188 (1936)
- [5] Orgler, Y.E.: A credit scoring model for commercial loans. *Journal of money, Credit and Banking* 2(4), 435–445 (1970)
- [6] Chen, C., Liaw, A., Breiman, L., et al.: Using random forest to learn imbalanced data. *University of California, Berkeley* 110(1-12), 24 (2004)
- [7] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G.: Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications* 73, 220–239 (2017)
- [8] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16, 321–357 (2002)
- [9] Nami, S., Shajari, M.: Cost-sensitive payment card fraud detection based on dynamic random forest and k-nearest neighbors. *Expert Systems with Applications* 110, 381–392 (2018)
- [10] Jabeur, S.B., Sadaoui, A., Sghaier, A., Aloui, R.: Machine learning models and cost-sensitive decision trees for bond rating prediction. *Journal of the Operational Research Society* 71(8), 1161–1179 (2020)
- [11] Batuwita, R., Palade, V.: Fsvm-cil: fuzzy support vector machines for class imbalance learning. *IEEE Transactions on Fuzzy Systems* 18(3), 558–571 (2010)
- [12] Maalouf, M., Trafalis, T.B.: Robust weighted kernel logistic regression in imbalanced and rare events data. *Computational Statistics & Data Analysis* 55(1), 168–183 (2011)
- [13] Huang, Y.-P., Yen, M.-F.: A new perspective of performance comparison among machine learning algorithms for financial distress prediction. *Applied Soft Computing* 83, 105663 (2019)
- [14] Zhu, Y., Xie, C., Sun, B., Wang, G.-J., Yan, X.-G.: Predicting china's sme credit risk in supply chain financing by logistic regression, artificial neural network and hybrid models. *Sustainability* 8(5), 433 (2016)
- [15] Kittler, J., Hatef, M., Duin, R.P., Matas, J.: On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence* 20(3), 226–239 (1998)
- [16] Henley, W., Hand, D.J.: A k-nearest-neighbour classifier for assessing consumer credit risk. *Journal of the Royal Statistical Society Series D: The Statistician* 45(1), 77–95 (1996)
- [17] Zhou, X., Zhang, D., Jiang, Y.: A new credit scoring method based on rough sets and decision tree. In: *Advances in Knowledge Discovery and Data Mining: 12th Pacific-Asia Conference, PAKDD 2008 Osaka, Japan, May 20-23, 2008 Proceedings* 12, pp. 1081–1089 (2008). Springer
- [18] Tang, L., Cai, F., Ouyang, Y.: Applying a nonparametric random forest algorithm to assess the credit risk of the energy industry in china. *Technological Forecasting and Social Change* 144, 563–572 (2019)
- [19] Wang, X., Jin, Y., Schmitt, S., Olhofer, M.: Recent advances in bayesian optimization. *ACM Computing Surveys* 55(13s), 1–36 (2023)