The International Journal of Social Sciences and Humanities Invention 5(06): 4820-4823, 2018
DOI: 10.18535/ijsshi/v5i6.09 **ICV 2015: 45.28**
ISSN: 2349-2031

Research Article

Research on bank credit default prediction based on data mining algorithm

Li Ying

School of Business Administration, China University of Petroleum-Beijing, Beijing, 102249, China

ABSTRACT: It is of great importance to identify the potential risks to the bank's loan customers. Based on data mining technology, it is an effective method to classify loan customers by classification algorithm. In this paper, we use Random Forest method, Logistic Regression method, SVM method and other suitable classification algorithms by python to study and analyze the bank credit data set, and compared these models on five model effect evaluation statistics of Accuracy, Recall, precision, F1-score and ROC area. This paper use the data mining classification algorithm to identify the risk customers from a large number of customers to provide an effective basis for the bank's loan approval.

Key words: Bank credit, Risk prediction, Data mining, Classification algorithm, python

INTRODUCTION

In today's information and digit age, bank credit default is still frequent, how to establish an effective model for the prediction whether bank customers will default on the loan for recognition of the risk in bank from a mass of loan applicants is of great significance. At present, many scholars at home and abroad have studied the probability prediction model of bank credit default and put forward some forecasting methods which almost have different restrictions and defects. Beaver proposes to use a single-factor method of financial ratios to analyze the technical credit default prediction of enterprises[1]. Pomp at al constructed a default prediction model using multivariate discriminant analysis (MDA)[2]. Yang at al used Logistic Regression method establishing a probability prediction model of listed companies' credit default, and identified the most influential corporate financial indicators[3]. Zhang at al proposed a SVM model, which constructed the technical credit default prediction model of small and medium-sized enterprises by constructing the evaluation index system of different input variables[4]. A large number of studies have shown that the data mining classification algorithm can find hidden rules from mass data and use this rule to classify and predict new unknown data[5]. Through data mining classification algorithm, banks can set up classification models by using the relevant personal information and consumption data of the loan applicants in the past, and find out the characteristics of risk customers. Then, use the classification model making classification prediction for new loan applicants, from which identify the risk customers, so as to reduce the risk of default repayment.

Therefore, in this paper we use the Random Forest method, Logistic Regression method, SVM method and other suitable classification algorithms to study and analyze the bank credit data set, and compared these models on five model effect evaluation statistics of Accuracy, Recall, precision, F1-score and ROC area to identify the risk customers from a large number of customers and provide effective approaches for the bank's loan approval.

Data sample collection and preprocessing

This paper uses the bank credit data set loan_model_sample in the kaggle as the target data set for the study[6]. There are 11017 samples and 199 attribute features. After the data collection is completed, the data is viewed and pre-processed. The overall framework of data preprocessing is shown in Figure 1.

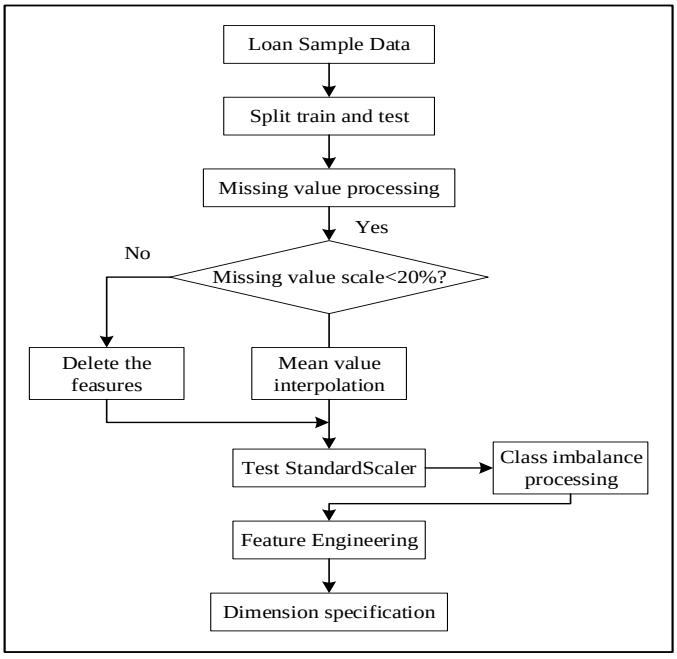

Figure 1 Data pre-processing framework

Firstly，In order to prevent the occurrence of data leakage problem, the data set is divided into two parts: training and testing. Training is used to train the model, and testing is used to test the model classification accuracy. Secondly, view the missing values of the data set and process the missing values. The visualization of the missing values of the original data is shown in Figure 2, in which white lines represent missing data. The threshold is set to 20%, and the features whose missing values are greater than the threshold will be directly deleted, the feature attributes whose missing values are

smaller than the threshold will be averaged using the preprocessing module of sklearn in scikit-learn. The processed data visualization is shown in Figure 3, it can be seen that the entire data set has been filled completely.

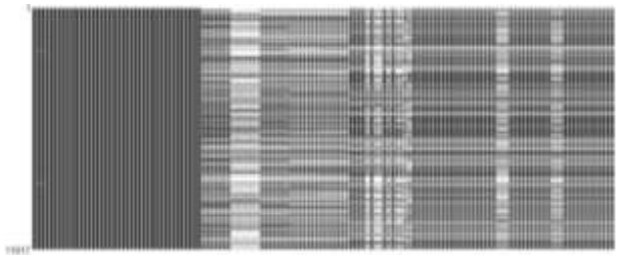

Figure 2 Missing data visualization

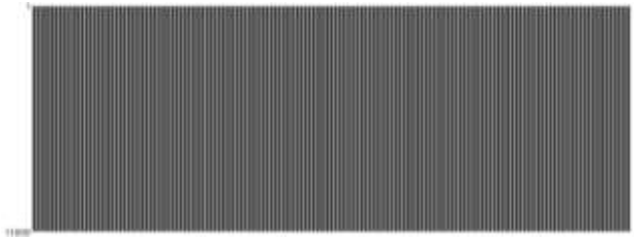

Figure 3 Missing data visualization after processing

Then, use the StandardScaler in scikit-learn to standardize test set data and observe whether the class exists the imbalanced issue. Perform feature engineering on datasets and select feature attributes that are valid for the classification model. Finally, in order to avoid overfitting problems and reduce the complexity of the model, using the PCA method, Pearson correlation coefficient or other automatic screening feature methods to perform dimension reduction on the training set.

Establishment and Evaluation of classification models

Based on the pre-processed data sets, use Random Forest[7], Logistic Regression[8] and SVM[9] method through python programming language to establish classification models respectively and adjust the hyper parameters with GridSearchCV method. Thus, the parameters are obtained when the model classification effect is best. Using five model effect evaluation statistics: Overall Accuracy, Recall, precision, F1-score and ROC area to compared the classification prediction effect of these classification models.

The most straightforward way to evaluate the classification model performance is based on the confusion matrix analysis. Confusion matrix is a concept from machine learning that contains information about actual classifications and predicted classifications done by a classification system. A confusion matrix has two-dimensions, one is indexed by the actual class of an object, the other is indexed by the class that the classifier predicts[10]. In the bank's credit data in this paper, the number 0 represents the credit default customer category, and 1 represents the normal customer category. The confusion matrix is shown in Table Ⅲ.

Table Ⅲ CONFUSION MATRIX

		Predicted Class	
		Class=1	Class=0
Actual Class	Class=1	TP	FN
	Class=0	FP	TN

A series of measures for measuring the performance of learning systems such as Overall Accuracy, Recall, precision and F1-score, can be defined based on the confusion matrix. The definition is as follows.

Accuracy is the correct proportion of the overall number of predictions, it is defined by the formula[11]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and the number of false negatives[12,13].

Precision is to measure the proportion of truly positive samples in samples that are predicted as positive samples by the classification model[14]:

$$Precision = \frac{TP}{TP + FP}$$

Recall is a measure of the ability of prediction model to select instances of a certain class from a data set[7], which is also represent TPR(true positive rate)[15]:

$$Recall = \frac{TP}{TP + FN}$$

F1-score is the harmonic average of precision and recall, it is defined by the formula[11]:

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

FPR(false positive rate) is to describe the proportion of the model negative class in the samples predicted to be positive[14]:

$$FPR = \frac{FP}{FP + TN}$$

ROC(receiver operating characteristic curve) is a technique for visualizing, organizing and selecting classifiers based on their performance[15]. It is a comprehensive index reflecting the continuous variables of sensitivity and specificity. The ROC curve is a two-dimensional curve with FPR as the X axis and TPR as the Y axis, which ranges from (0,0) to (1,1). A common method to compare classifiers is to calculate the area under the ROC curve, abbreviated AUC[16,17]. The larger the AUC, the better the model classification is.

Since the purpose of the bank credit default model is to identify credit default risk customers from a large number of loan application customers, the risk of predicting a risk customer (Class = 0) as a normal customer (Class = 1) is much larger than predicting a normal customer (Class = 1) to customers (Class=0). Based on the above analysis, this paper pays more attention to whether the model can correctly classify Class=0 when assessing model classification results, so the classification algorithms results in this paper are all from Class=0.

Findings and discussions

This paper uses RandomForest method, LogisticRegression method and SVM method to establish classification models. Perform research and analysis on pre-processed bank credit default data sets, and use the GridSearchCV method to search

for the best parameters. The above three classification model construction methods are respectively correspond to RandomForestClassifier, LogisticRegression, and SVC algorithms in scikit-learn.

In the RandomForestClassifier algorithm，what need to perform GridSearchCV to find the best parameter value in order to enhance the prediction effect of classification model is n_estimators，which represents the number of trees in the forest. In addition，with the number of trees in the forest increasing, the classifier is becoming more and more complicated, which probably brings overfitting problem to the model. To decrease the model complexity, we also adjust values of some other parameters in the RandomForestClassifier algorithm, such as max_depth min_samples_split、min_samples_leaf and max_features.

In the LogicsticRegression algorithm, we use GridSearchCV to search the best Parameter combination, one is the penalty(Penalty item) between l1 and l2, the other is C which represents the reciprocal of the regularization coefficient λ. The objective function can be summarized as follows[18]:

$$\omega^* = argmin \sum_i L(y_i, f(x_i, \omega)) + \lambda\Omega(\omega)$$

Where the first item L is the training error, and the second item is the penalty item. The first item is to minimize the training error and get the best fitting data; the second is to simplify the model, prevent overfitting, and get better generalization ability.

SVC algorithm is the classifier model of SVM. In the SVC algorithm, we use GridSearchCV to find the best Parameter combination of kernel {'linear','poly', 'rbf'}、

Table Ⅳ COMPARISON OF CLASSIFICATION ALGORITHMS RESUITS

Algorithm	Accuracy	precision	recall	F1-score
RandomForestClassifier	0.8863	0.18	0.67	0.28
LogisticRegression	0.7130	0.09	0.64	0.16
SVC	0.8580	0.15	0.55	0.24

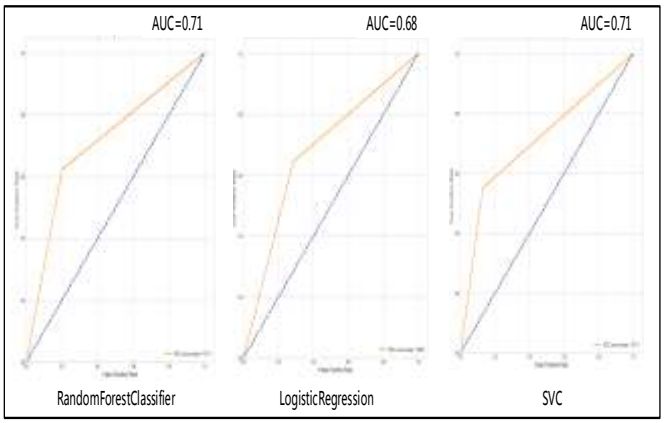

Figure 4 Comparison of algorithm ROC curves

penalty parameter C of the error term and gamma which represents kernel coefficient for 'rbf', 'poly' and 'sigmoid'. C is the penalty factor. The larger its value is, the heavier the penalty for misclassification of training samples and the higher the requirement for correct classification are. Gamma is a parameter of the insensitive loss function. The smaller the value is, the more support vectors are; and the larger the distance between two large edges is, the easier it is to find the maximum hyper-plane Maximum Marginal Hyperplane[5].

After parameter optimization, we obtain the comparison results of RandomForestClassifier, LogisticRegression, and SVC algorithms on the loan_model_sample, as shown in Table Ⅳ and algorithm ROC curves are shown in Figure 4.

The above comparison results show that RandomForestClassifier algorithm has a good classification effect for large sample as well as high dimensional attribute feature datasets，because the value of its Recall, F1-score, Accuracy, Precision and ROC area are all larger than the other Classifiers. And the LogisticRegression algorithm has relatively high recall, but with the lowest value of precision. The SVC algorithm has the relatively high accuracy, precision and F1-score, but with a lowest recall, which to a certain degree shows SVC algorithm may be not suitable for the bank credit default prediction models, because it may well omit risk loan applicants and cause huge losses to banks. Thus, we consider RandomForest algorithm as the most suitable algorithm for the bank credit default prediction problem, especially when the dataset is very large or has high dimensions.

Conclusions

This paper establish bank credit default prediction models using RandomForest, LogisticRegression and SVM classification algorithms under the python language environment. And compare the classification effect of the classifiers through five model effect evaluation statistics: Accuracy, Recall, precision, F1-score and ROC area. Comparative analysis Experimental results show that compared to LogisticRegression and SVM classification algorithms, RandomForest algorithm is more suitable for the bank credit default precision model because its high classification effect for Class=0, especially when the dataset is very large or has high dimensions. This paper provides an effective experimental basis for bank credit approval to identify risk customers from a large number of loan applicants using data mining classification algorithms.

Since the number of appropriate public data sets for bank credit is small, the number of samples in this paper is only 11,017, therefore, the experiment may not be comprehensive. In the future, it is necessary to collect more datasets of large number and features for further improvement.

References

[1] Beaver, W. Financial Ratios as Predictors of Failure.Empirical Research in Accounting: Selected Studied[J]. Journal of Accounting Research, 1966，(4).

[2] Pompe, P.P.M., Bilderbe, J. The Prediction of Bankruptcy of Smalland Medium-sized Industrial Firms[J]. Journal of Business Venturing, 2005,20.

[3] Yang Pengbo, Zhang Chenghu, Zhang Xiang. Prediction model of credit default probability of listed companies based on Logistic regression analysis [J]. economic latitude and longitude,2009(02):144-148.

[4] Zhang Jie, Zhao Feng. [J]. statistics and decision making of SME credit default prediction based on support vector machine,2013(20):66-69.

[5] Mei Mei. Application of data mining classification algorithm in credit card risk management [J]. modern computer,2013(19):13-16.

[6] https://www.kaggle.com/datasets

[7] Liaw A, Wiener M. Classification and regression by randomForest[J]. R news, 2002, 2(3): 18-22.

[8] Hosmer Jr D W, Lemeshow S, Sturdivant R X. Applied logistic regression[M]. John Wiley & Sons, 2013.

[9] Joachims T. Making large-scale SVM learning practical[R]. Technical report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund, 1998.

[10] Deng X, Liu Q, Deng Y, et al. An improved method to construct basic probability assignment based on the confusion matrix for classification problem[J]. Information Sciences, 2016, 340: 250-261.

[11] Ohsaki M, Wang P, Matsuda K, et al. Confusion-matrix-based Kernel logistic regression for imbalanced data classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(9): 1806-1819.

[12] Branco P, Torgo L, Ribeiro R P. A survey of predictive modeling on imbalanced domains[J]. ACM Computing Surveys (CSUR), 2016, 49(2): 31.

[13] Fanshawe T R, Power M, Graziadio S, et al. Interactive visualisation for interpreting diagnostic test accuracy study results[J]. BMJ Evidence-Based Medicine, 2018, 23(1): 13-16.

[14] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves[C]//Proceedings of the 23rd international conference on Machine learning. ACM, 2006: 233-240.

[15] Fawcett T. An introduction to ROC analysis[J]. Pattern recognition letters, 2006, 27(8): 861-874.

[16] Bradley A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms[J]. Pattern recognition, 1997, 30(7): 1145-1159.

[17] Breheny P. Classification and regression trees[J]. 1984..

[18] Hilbe J M. Logistic regression models[M]. CRC press, 2009.