Internet Financial Credit Evaluation Based on the Fusion of GBDT and LR

Tao Zhang[1, a] Shuo Meng[1, b]

[1] Beijing University of Technology, Beijing 100124, China.

[a]zhangtony8888@qq.com, [b]2247277148@qq.com.

Abstract. With the rapid development of Internet financial credit business, credit evaluation has become a hot spot in the development of the industry. Aiming at the problem of complex data type and large amount of data in original data set, an evaluation model based on Gradient Boosting Decision Tree (GBDT) and Logistic Regression (LR) fusion is proposed. LR model is a very practical model in credit evaluation. The model is simple and fast, but it has high requirement for feature processing. GBDT has natural advantages in processing multi data type data, and it can extract new features from raw data. The fusion of the two can not only fully excavate the information of the data set, but also improve the training efficiency. Compared with other models on Internet financial credit data, it has higher accuracy, recall and Score.

Keywords: Internet financial; Credit evaluation; Feature combination; the fusion of GBDT and LR.

1. Introduction

Now the Internet financial industry has developed rapidly. The credit platform requires a credit assessment of the user to determine whether a loan is issued. It is obviously inefficient to rely solely on the subjective judgment of the staff, and there is a possibility of cheating by the wind control staff. It is very important to construct the credit evaluation model. The accurate credit evaluation of the loan applicant can help the merchant to avoid the risk effectively [1]. In recent years, LR, KNN, Decision Tree, SVM and so on have been applied to credit evaluation, and have achieved some results [2], however, these methods do not fully exploit data set characteristics. In this paper, the model of GBDT and LR is used in the Internet financial data. New features are constructed by the natural nature of GBDT's processing of multiple types of data sets. Then the new features and the original features are trained with LR, and better classification results are achieved compared with other models.

2. Model Theory

2.1 Using GBDT to Construct New Features

GBDT is a common nonlinear model based on the idea of Boosting. Each iteration of the GBDT is to establish a new decision tree in the direction of the gradient of the residuals [3], as shown in the Fig. 1. If the previous round of strong learner iteration obtained is ft-1(x), the loss function is L(y, ft-1(x)), the goal of this iteration is to find a weak learner hot (x) of a decision tree model to minimize the loss of L (fat(x)) =L(y, ft-1(x) +hot(x)). GBDT uses the Shrinkage strategy to avoid over fitting by setting the step length. Shrinkage thought that the effect of a small step approaching the result is more likely to avoid over fitting than the effect of a large step approaching the result. This idea is not entirely dependent on a decision tree. It is considered that every decision tree is a small part of truth. When accumulating, only a small part is accumulated, and the number of corresponding decision trees will be more. The residuals of each tree are gradual, not steep. In essence, Shrinkage is equivalent to giving each decision tree a weight, multiplied by this weight.

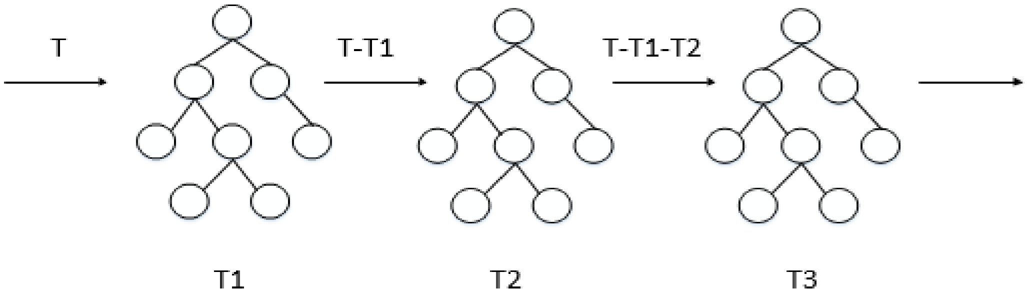

Fig 1. GBDT algorithm thought

GBDT's new features refer to each leaf node of each decision tree as one dimension of the new feature. The new feature’s dimension is the same as the number of GBDT leaf nodes. The feature of leaf node that the sample falls to is 1, and the rest are 0. Fig. 2 is two decision trees in the GBDT model. The input sample X falls to the second leaf node of Tree1 and the first leaf node of Tree2. The two leaf nodes are 1 and the rest are 0, and the new features are [0, 1, 0, 1, and 0].

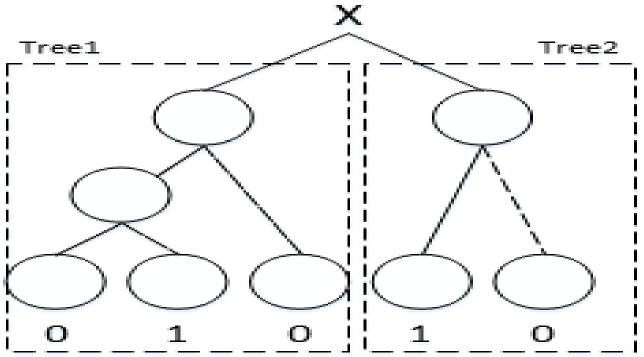

Fig 2. GBDT structures combination features

2.2 The Fusion of GBDT and LR

Logistic regression is a logarithmic linear model. The odds of an event is the ratio of the probability of the event to the probability that the event does not occur. If the probability of the event is p, the probability of the event is (p/ (1-p)) [4].If you want the log odds of an event to be the linear model, then the event distribution function is the sigmoid function, y=log (p/ (1-p)). The logical regression is to apply the sigmoid function on the basis of linear regression, and maps the value of the function to the 0~1 interval, and the function value after the mapping is the probability of classification [5]. Logistic regression is a basic binary classification model [6].In this experiment, the label Y is {0, 1}, Y=1 means good credit can be lent, Y=0 means credit is poor and no loan is granted. X = {x1, x2, up} is the characteristic of p dimension, obtained from training data.

The learning ability of linear model is limited, and a large number of feature engineering is required to analyze the effective features and combination features in advance, so as to indirectly enhance LR's non-linear learning ability. Relying on manual experience alone will not necessarily lead to improved results. The idea of GBDT gives it a natural advantage to find a variety of distinguishing features and feature combinations. The path of decision tree can be used directly as the input feature of LR, eliminating the steps of manual search feature and feature combination. As shown in Fig. 3.

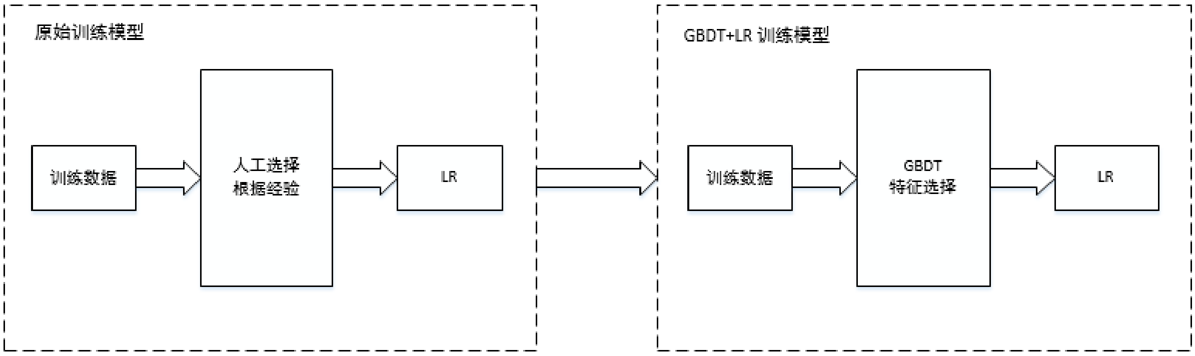

Fig 3. The frame diagram of GBDT and LR fusion

3. Empirical Analysis

3.1 Experimental Data

The experimental data used in this paper is the credit data of "Give me some credit" competition on the Cagle website [7]. The data dictionary is shown in Table 1. The data is divided into two parts: the training data and the test data. The training data contains 150,000 samples, and the test data contains 101503 samples. Each sample point contains 11 variables, including 10 features and 1 label. X1 (SeriousDlqin2yrs) is a label variable, one is 0, representing a non-default customer, and the other is 1, representing a default customer. As the label variable X1 of the test data is unknown, this experimental data only uses the training data. In order to facilitate calculation, the value of X1 is changed in this paper, and the default customer value is changed from "1" to "-1", and the non-defaulting customer is changed from "0" to "1".In training data, there are 139,974 non-defaulting sample points, accounting for 93.32%. The total number of default samples was 1, 0026, accounting for 6.68%.

Table 1. The data dictionary for the kaggle credit database.

Dimension	Variable Name	Description	Type
X1	SeriousDlqin2yrs	Person experienced 90 days past due delinquency or worse.	Y/N
X2	RevolvingUtilization OfUnsecuredLines	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits.	percentage
X3	Age	Age of borrower in years.	integer
X4	NumberOfTime30-59 DaysPastDueNotWorse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years.	integer
X5	DebtRatio	Monthly debt payments, alimony, living costs divided by monthy gross income.	percentage
X6	MonthlyIncome	Monthly income.	real
X7	NumberOfOpenCredit LinesAndLoans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards).	integer
X8	NumberOfTimes90 DaysLate	Number of times borrower has been 90 days or more past due.	integer
X9	NumberRealEstate LoansOrLines	Number of mortgage and real estate loans including home equity lines of credit.	integer
X10	NumberOfTime60-89 DaysPastDueNotWorse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.	integer
X11	NumberOfDependents	Number of dependents in family excluding themselves (spouse, children etc.).	integer

3.2 Data Processing

The transformation of the original data into a suitable form can represent the actual problems of model processing and improve the generalization ability of the algorithm model. Discretization of data, data reduction, creation of new variables and data normalization are commonly used methods for data processing. According to the credit data of Cagle, by analyzing the age variables, It is can be found that there is only one sample for the age of less than 18, and the minors have no default. Therefore, it is possible to deduce that the distribution of default is significantly different between minors and adults, so a new variable X12 (Low age) can be added.

$$\mathrm{X12} = \begin{cases} 0, \mathrm{age} < 18 \\ 1, \mathrm{age} \geq 18 \end{cases} \tag{1}$$

The range of age variables is [0,109], which is relatively large. Numerical data is easy to be influenced by dimensionality. When a variable is larger, it will have a great influence on credit evaluation. The maximum and minimum normalization, piecewise linear transformation and logarithmic change are the common conversion methods. The logarithmic change method was used to deal with this variable, X3 = log(X3-17). And there's a sample that's less than 17 that can't calculate with this formula, and replace it with 0.

The range of X6(Monthly Income) is the largest, far greater than the other variables, and is processed with minimum and maximum normalization. The calculation formula is as follows:

$$\mathrm{Xnew}=\frac{X-Xmin}{Xmax-Xmin} \tag{2}$$

In formula (2), Xin refers to the minimum value of the variable, and IMAX refers to the maximum value in the variable. The variable X6 is transformed into a linear transformation, which is mapped to [0, 1].

3.3 Parameter Tuning

In this paper, the parameters of GBDT are selected by means of experiment. The two main parameters are the max number of leaves in each decision tree and the number of decision trees [8], α and β are used to represent the two parameters. The single variable factor method is used in this paper. First of all, three groups of experiments were carried out. The value of the βare 10, 30, and 50, respectively. The raw data is trained by GBDT to produce new combination features, training in LR with the original features. The relationship between the predicted accuracy and the maximum number of leaves of each decision tree is shown in Fig. 4. It can be found that the accuracy of the prediction significantly improved when α<9. The prediction accuracy tends to be stable when α>=9, therefore, the max number of leaves in each decision tree is 9. Then fixed the value of α is 9, the relationship between the prediction accuracy and the number of decision trees is shown as Fig. 5.When β<22, the prediction accuracy increase, when β>=22, the prediction accuracy tends to be stable, so the value of β is 22.

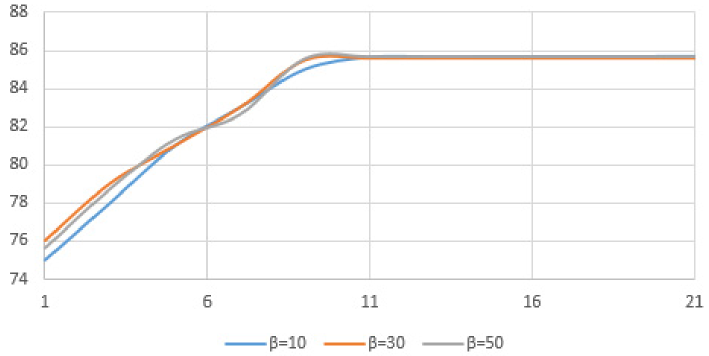

Fig 4. The relationship between the accuracy of prediction and the max number of leaves

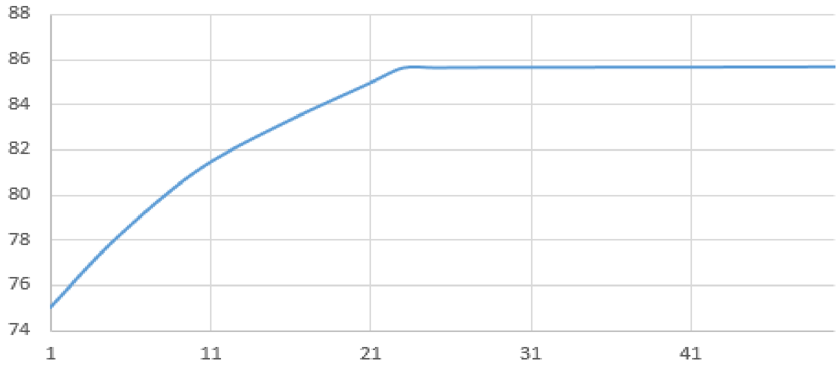

Fig 5. The relationship between the prediction accuracy and the number of decision trees

3.4 Result Contrast

In this experiment, LR, GBDT and the fusion of LR and GBDT were used to compare the performance of credit classification. This experiment uses the 10 fold cross-validation method, experimental results take the average of 10 times.

The evaluation indexes were precision (TP/ (TP+FP)), recall (TP/ (TP+FN)) and Score (2*P*R/ (P+R)) [9], where TP is true positive, FP is false positive, FN is false negative. The results of the test are shown in Table 2.

Table 2. Evaluation effect of LR, GBDT and LR+GBDT on credit data

	LR			GBDT			GBDT+LR	
Precision	Recall	F_Score	Precision	Recall	F_Score	Precision	Recall	F_Score
81.2%	80.3%	80.5%	83.5%	83.1%	83.5%	85.5%	86.2%	85.8%

As can be seen from the Table 2, the evaluation model of the fusion of LR and GBDT compared with the separate LR and GBDT in precision , recall and Score has significantly improved, The feature combination of GBDT is better for mining information in Internet financial data.

4. Summary

With the rapid development of Internet finance industry, all kinds of Internet finance enterprises need to carry out credit evaluation to effectively avoid risks [10]. All kinds of credit evaluation models have a certain effect [11]. The general data amount of the Internet financial credit data is large and the data type is complex, so the LR, which has a faster processing speed, is adopted. LR is a linear model with limited learning ability and high requirements for features. The characteristics of GBDT algorithm can be used to explore the distinguishing feature and feature combination, reducing labor costs in feature engineering, the two can be properly combined. This paper introduces the fusion of GBDT and LR into the credit evaluation of Internet finance, and establishes a model based on combinatorial classifier [12]. According to the empirical research, the fusion of GBDT and LR has higher classification performance than other models. In order to solve the problem of large and complex data sets, the GBDT+LR fusion model can be used to try.

References

[1]. Burton D. Credit scoring, risk, and consumer lending capes in emerging markets [J]. Environment and Planning A, 2012, 44 (1): 111- 124.

[2]. Deng Y, Xiaomin X U. P2P personal credit evaluation model based on Internet behavior information[J]. Journal of Beijing Information Science & Technology University, 2017.

[3]. Zhou Z H. Ensemble Methods: Foundations and Algorithms [M]. Taylor & Francis, 2012.

[4]. Li H . Statistical learning method(in Chinese) [M] . Tsinghua university press, 2012.

[5]. He X, Pan J, Jin O, et al. Practical lessons from predicting clicks on ads at Facebook [C]. Proceedings of 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, 2014: 1-9.

[6]. Abid L, Masmoudi A, Zouari-Ghorbel S. The Consumer Loan's Payment Default Predictive Model: an Application of the Logistic Regression and the Discriminant Analysis in a Tunisian Commercial Bank[J]. Journal of the Knowledge Economy, 2016(1):1-15.

[7]. Information on: http://www.kaggle.com/c/GiveMeSomeCredit/data.

[8]. Ke G L. Study on the parallel learning algorithm of Gradient Boosting Decision Tree (in Chinese) [D]. Xiamen University, 2016.

[9]. Fawcett T. An introduction to ROC analysis[J]. Pattern Recognition Letters, 2006, 27(8):861-874.

[10]. Zhang X, Dai D. Research on Credit Evaluation of Internet Financial Small and Micro-enterprise Customers[J]. Journal of Xiangnan University, 2016.

[11]. Xiao J, Chen L Y. Improved research on P2P credit customer credit evaluation model[J]. Information Technology, 2016.

[12]. Chen J, Wang S, Zhao Z, et al. The Prediction of CTR Based on Model Fusion Theory[M]// Geo-Spatial Knowledge and Intelligence. 2017.