Research on Online Loan Default Prediction Model Based on Ensemble Learning

Tao Zhang[1a], Wenhao Sun[*2b]

[a]zhangtony8888@qq.com; [b]sunwenhaochina@163.com

[1]Beijing University of Technology, Beijing, China,
[*2]Beijing University of Technology, Beijing, China,

Abstract: In recent years, with the rise of internet finance and the promotion of early consumption awareness, personal credit services have witnessed rapid development. Traditional banks, due to reasons such as complex borrowing process and long cycle, can no longer meet the growing demand for loans from small and micro enterprises and individuals. Thanks to the combination of the Internet and other traditional industries, online lending has gradually emerged, among which P2P has developed particularly rapidly. However, due to information asymmetry between lending platforms and borrowers, most online lending platforms have closed down due to bad debts and defaults. Therefore, accurately predicting the default probability of borrowers has become an urgent problem to be solved in the online lending industry.

Compared with statistical methods, machine learning has more accurate and efficient prediction capabilities. However, a single model is easy to over-fit, has poor stability, and may not make full use of the information in the data. In order to improve the prediction performance, this study constructs an ensemble model based on logistic regression algorithm and SVM algorithm, aiming to improve the prediction accuracy. Logistic regression is a widely used statistical method for classification problems, while SVM is a machine learning algorithm based on maximum margin classification, which is suitable for processing high-dimensional datasets. This study first analyzed the characteristics of borrowers based on the data published by the P2P platform Lending Club, and then cleaned and processed the data. Then, the data was trained and tested using traditional logistic regression model and SVM model respectively, and the prediction results of the two single models were obtained. Finally, logistic regression and SVM were integrated to train and predict the dataset. By comparing the prediction accuracy of the single model and the ensemble model, it was found that the accuracy of the ensemble model was significantly higher than that of the single model, demonstrating the effectiveness of the ensemble model.

Keywords: P2P, default prediction, Logistic Regression, SVM, ensemble learning

1. INTRODUCTION

With the rapid development of Internet technology, various traditional industries are combined with the Internet to form an 'Internet+' development model. Due to the high threshold and complicated process of traditional bank loans, the financial industry and the Internet have derived online loan services, among which P2P has developed the most rapidly. In 2005, the UK established the world's first P2P platform. In 2007, China's first P2P platform Paipaidai was established in Shanghai. At the end of 2012, there were 132 normal online lending platforms in China.In 2015, there were 3464 normal online lending platforms. By the end of 2018, the

MSEA 2023, May 26-28, Nanjing, People's Republic of China
Copyright © 2023 EAI
DOI 10.4108/eai.26-5-2023.2334378

cumulative transaction of China's online loan platform exceeded the seven trillion mark[1]. With the rapid development of online lending platforms, various problems have been exposed, and most lending platforms have gone bankrupt.

Since the emergence of online loans, many scholars have studied this[2]. As early as the 1980 s, Ohlson[3] used Logistic regression model to identify risk users. On this basis, Wang Qian uses the logistic regression model based on Lasso penalty, Ridge penalty and Elastic-net penalty to empirically analyze the data of Paipai loan borrowers. The results show that these three penalty-based models are better than the logistic regression model. Ralf et al.applied support vector machine models based on different kernel functions to user default prediction, and compared the prediction effects of different kernel support vector machine models[4]. Lakshmi proved that the random forest model is better than the single decision tree model through empirical analysis[5].

Through the above research and the comparative study of single model and integrated model, this thesis uses logistic regression algorithm and SVM algorithm to integrate, constructs a credit default prediction model, and uses the data set published by Lending Club to train and predict the model, so as to meet the demand of online loan platform for user loan prediction under financial risk control.

2. RELATED WORK

2.1 Logistic Regression Model

Logistic regression is a commonly used binary classification algorithm. Its basic idea is to classify data by mapping data to a probability value between 0 and 1, usually with 0.5 as the threshold for classification. As early as the 1960 s, Edward Altman[6] used logistic regression algorithm to classify companies based on their financial data and predict the probability of default.

The logistic regression algorithm combines the sigmoid function and the linear regression model, and uses the output of the linear regression model as the input of the sigmoid function[7]. Linear regression model formula (1). Y_i is the result of whether an event occurs in the ith sample, β_k is the regression coefficient of the kth index, and x_{ik} is the kth index of the ith sample. The output result Y_i of formula (1) is used as the input of the sigmoid function of formula (2), that is, the range is mapped from the range to the [0, 1] interval. P_i is the probability of an event in the i th sample. The function image is S curve, as shown in Figure 1.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \tag{1}$$

$$P_i = \frac{1}{1+e^{-n}} \tag{2}$$

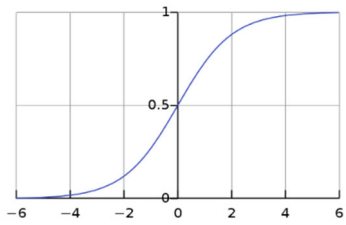

Figure 1. Sigmoid function diagram

The disadvantage of the logistic regression algorithm is that it cannot effectively solve the nonlinear problem, and when judging the probability of an event in the sample, most of the probability values deviate from both ends, and the number of predicted samples near 0.5 is small.

2.2 SVM Model

SVM is a machine learning algorithm developed on the basis of statistical learning theory, proposed by Cortes and Vapnik[8]. SVM is mainly used for classification and regression tasks, with good generalization performance and robustness, especially in small samples and high-dimensional data.

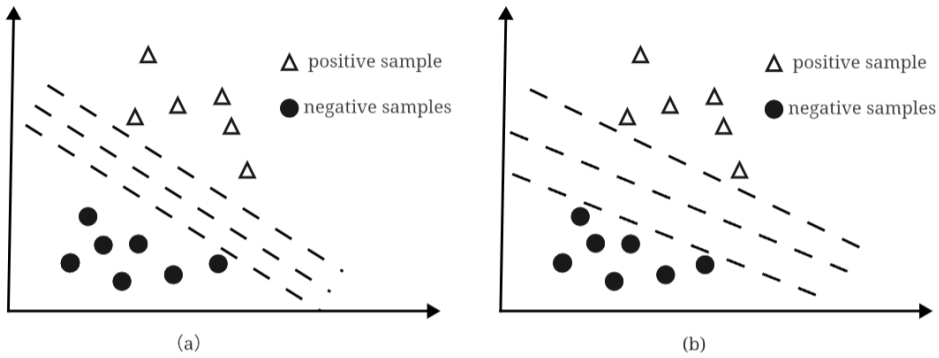

Figure 2. Ordinary(a) and maximum(b) interval classification hyperplane

The basic idea of SVM is to find an optimal hyperplane in the feature space to separate different categories of samples. For linearly separable samples, SVM finds the hyperplane by maximizing the margin, so that it can be better generalized to new samples[9]. For example, given a binary data set $T\{x_i, y_i\}(i = 1,2 \dots n)$, where $x_i \in R_n$ and $y_i \in \{-1, 1\}$, For binary linear separable data sets, in R_n space, The linear classification hyperplane with the expression of $\omega^T + b = 0$ divides the two types of samples in the data set on both sides of the hyperplane. there are many such expressions.

After satisfying the condition that each sample is correctly classified on both sides of the hyperplane, it is expected that the sample points on both sides of the hyperplane can maximize the geometric interval between the hyperplane, as shown in Figure 2. This ensures that the points closer to the hyperplane can be classified correctly.

2.3 Ensemble Learning Model

Ensemble learning is a machine learning method that combines multiple classifiers to build a more powerful classifier[10]. The basic idea is to weighted average or vote the prediction results of multiple models when making decisions, so as to improve the overall prediction accuracy and robustness. The logistic regression model regresses the relationship between the probability of occurrence of the corresponding variable and the explanatory variable by fitting the nonlinear relationship between the corresponding variable and the explanatory variable, but the logistic regression model is not good at dealing with nonlinear problems. The SVM model is an analytical tool based on structural risk minimization and solved by optimization methods. The main advantage is that the prediction accuracy is high, but its stability is poor, and the computational complexity increases sharply when the number of samples is too large. Both models have strong robustness, and both have different decision boundaries, which means that they may perform differently in different situations. Using their differences, classification problems can be better captured[11].

3. Experiments and Results Analysis

3.1 Data Set Introduction

Founded in 2006, Lending Club is the first and the largest P2P platform in the world. The platform uses Internet technology to provide online lending transaction services for investors and borrowers. In 2015 alone, it generated more than $ 8 billion in loans. The platform will publicize the data set after desensitization every year for scholars to analyze and study. In this thesis, the data from 2017 to 20193 published on the Lending Club platform contains 1311,070 data and 144 variables. Among them, the target variable is 'loan_status', which represents the loan status. There are 7 values in total, and the specific meaning of each value is shown in Table 1. The samples with values of 'Fully Paid' and 'Current' are defined as good samples, and the rest are bad samples.

Table 1. Loan status enumeration

Enumeration	Description	Data volume (Percentage)	Classification
Fully Paid	Has been fully repaid	267198(20.38%)	Performing
Current	In the repayment	940844(71.76%)	Performing
In Grace Period	Within the grace period	10402(0.79 %)	Abnormal
Late (16-30 days)	16-30 days overdue	4117(0.31%)	Abnormal
Late (31-120 days)	31-120 days overdue	15988(1.22%)	Abnormal
Default	The loan has been defaulted	661(0.05%)	Abnormal
Charged Off	The loan has been bad debts	71860(5.48%)	Abnormal

Data set variables can be divided into loan information, borrower information, loan status information, repayment plan information and other information. There are 35 character variables and 108 numerical variables.

3.2 Data Cleaning

In order to ensure that the data set can be more standardized, reliable and suitable for analysis. This topic in advance of the data set may exist errors, missing processing, thereby improving the accuracy and reliability of the data, thereby improving the efficiency of data analysis and the quality of analysis results. This topic on the data set cleaning mainly includes deleting irrelevant variables (post-loan variables, meaningless variables or co-lender information), deleting variables with missing rate greater than 0.6, filling variables with missing rate less than 0.1, deleting abnormal values and deleting variables with the same value ratio greater than 0.9. There are 62 variables in the data set after cleaning, and the number of samples is 869910.

3.3 Imbalanced Data Processing

There is an imbalance between the good and bad samples in this topic As shown in Figure 3 (A), represents the proportion of good and bad samples in the original data set. It can be seen that the ratio between the two samples in the data set is nearly 12 : 1, and there is a clear imbalance.

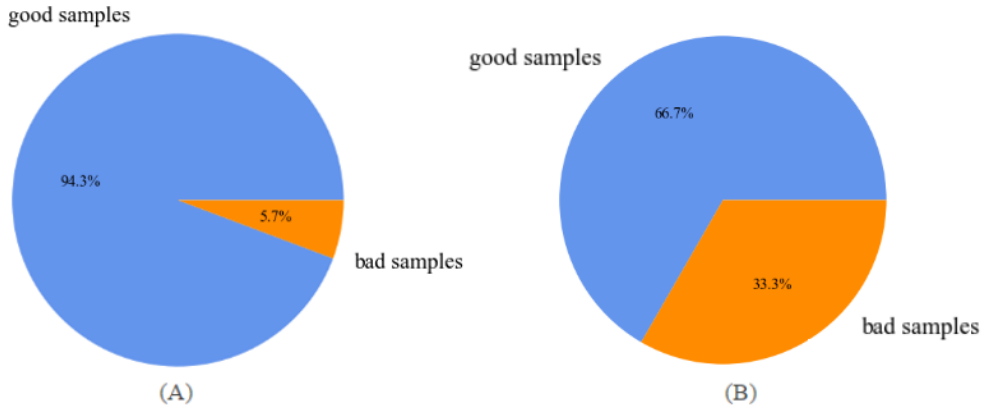

Figure 3. Positive and negative sample proportion chart

In this thesis, the Borderline-SMOTE algorithm is used to oversample the data set, and the parameter of synthesizing a small number of samples is set to 0.5, that is, the minority samples after oversampling account for 1/2 of the majority samples, and the ratio of majority samples and minority samples after oversampling is 3:2, shown in Figure 3 (B) .

3.4 Model Construction

This topic refers to the experiments of many researchers, and finally adopts the ratio of training set and test set 7:3 to segment the data. This topic uses grid optimization for hyperparameter tuning. The grid search method is a hyperparameter tuning method to improve the performance and effect of the model.

The main parameters of the SVM model are 'kernel', 'gamma' and 'C', where 'kernel' represents the kernel function of the model, and 'gamma' represents the kernel function coefficient. The larger the value, the easier the over-fitting, and the smaller the under-fitting. 'C' is the penalty

parameter, and the larger the value is, the greater the penalty slack variable is. Table 2 shows the first three sets of parameters with the highest score of SVM model.

Table 2. SVM model grid search tuning part of the results

kernel	gamma	C	score
rbf	0.1	10	0.98183
rbf	0.1	1	0.97736
rbf	0.001	100	0.83103

The main parameters of logistic regression are 'penalty', 'C', 'solver', 'max_iter' and 'class_weight'. 'Penalty' represents the regularization type, l1 regularization is mainly used to realize the feature selection function, and l2 regularization can avoid the collinear between features. The 'C' is the regularization coefficient, the smaller the stronger the regularization, the 'solver' is the optimization algorithm, the 'max_iter' represents the maximum number of iterations, and the 'class_weight' represents the category weight parameter. The weight of the minority class is raised to deal with the imbalanced data set. Table 3 shows the first three sets of parameters with the highest scores in the logistic regression model.

Table 3. Logical regression model grid search method tuning parameter part of the results

penalty	solver	C	max_iter	score
l2	saga	0.1	50	0.70495
l1	liblinear	1.0	100	0.704431
l1	liblinear	0.1	50	0.704283

About the construction of the integrated model, this topic adopts two methods:

(1) The logistic regression model or SVM model is used to predict the samples, and the predicted probability is used as an explanatory variable and other user information as the input of another model, and then the model is trained and predicted. The flowchart of the method is shown in Figure 4.

Figure 4. Integrated first method flow chart

(2) The logistic regression model or SVM model is used to select the original features of the data set, and the selected feature variables are used as the input of another model to improve the performance of the model. Its flow chart is shown in Figure 5.

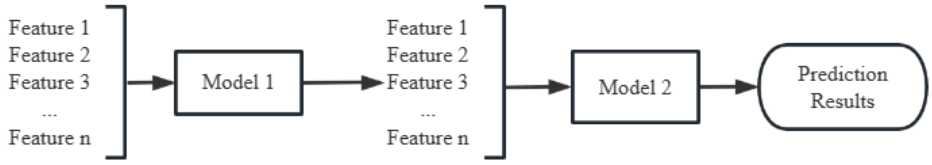

Figure 5. Integrated second method flow chart

For method one. Firstly, the SVM model is used to train and predict the data set, and then the probability value of the prediction stage is output. It should be noted that the parameter 'probability' needs to be set to 'True', which determines whether each possible probability is output according to the probability, and the default is 'False'. The output probability value is taken as a feature to the original data set, and the logistic regression is trained and predicted. This experiment is recorded as Experiment a. Then logistic regression is used to train and predict the data set, and the output probability value is saved as a feature to the original data set. Then the SVM model is used to train and predict it, which is recorded as experiment b.

In the second method, when SVM is used for feature selection, the kernel function cannot select the optimal kernel function 'rbf' selected by this topic, and the linear kernel function needs to be selected. Therefore, this method only uses logistic regression as the feature selection model, and the selected features are input into the SVM model for training and prediction. The experiment is experiment c. Table shows the results of three experiments.

Table 4 is the prediction results of two single model experiments and three integrated model experiments.

Table 4. Comparison of SVM model prediction results under different data sets

Experiments	Acc	Pre	Recall	F1	AUC
SVM Model Experiments	0.9373	0.9293	0.9448	0.937	0.9448
Logistic Regression Model Experiments	0.7037	0.6933	0.7223	0.7075	0.7599
Integration Model Experiment a	0.9358	0.9472	0.9236	0.9352	0.9771
Integration Model Experiment b	0.9405	0.9348	0.9471	0.9409	0.9808
Integration Model Experiment c	0.9377	0.935	0.9413	0.9381	0.9783

3.5 Experimental Comparison and Analysis

It can be seen from Table 4 that the performance of the integrated model is obviously due to the traditional single model. The accuracy of the three integrated experiments is more than 0.92, and the highest accuracy in experiment b exceeds 0.94. From the perspective of F1 value and AUC value, the integrated model is also significantly beyond the single model. In particular, compared with the traditional logistic regression model, the ensemble method of inputting the predicted results of the SVM model as a variable into the logistic regression model improves the accuracy by 24 percentage points. It is inferred that the performance of the SVM model is better, thus weakening the nonlinear relationship between features. From the perspective of

integrated methods, the differences between the three experiments are relatively small, and the differences in each index are within 1 percentage point. In experiment b, the results of logistic regression prediction of the training set are input into the SVM model as a feature of the training set. The overall performance of this integrated method is the best, and its AUC value is 0.9808, which is 0.0037 and 0.0025 higher than that of experiment a and experiment c, respectively. In terms of accuracy, Experiment a and Experiment c are higher than Experiment b, which are 0.0124 and 0.0002 higher respectively. However, in terms of recall rate, Experiment b is 0.0235 and 0.0058 higher than Experiment a and Experiment c respectively. From the perspective of F1 value, it can also be seen that the results of experiment b are better than those of experiment a and experiment c.

4. CONCLUSION

In this thesis, SVM and logistic regression are used to construct a single default prediction model and an integrated default prediction model respectively, and these models are trained by using the lending data of Lending Club for three years. According to the prediction results of the test set, the performance of the integrated model is better than that of the single model. At the same time, it has better stability than single model. Improving the accuracy of credit default prediction is of great significance to financial risk prevention and control.

REFERENCES

[1] Huang, R.H. Online P2P Lending and Regulatory Responses in China: Opportunities and Challenges[J]. Eur Bus Org Law Rev, 2018(19): 63-92. DOI:10.1007/s40804-018-0100-z.

[2] Bo Peng. Research on internal risk control system of P2P network lending[J]. International Journal of Computing Science and Mathematics, 2021, 13(2): 126-135.

[3] Ohlson J A . Financial Ratios and the Probabilistic Prediction of Bankruptcy[J]. Journal of Accounting Research, 1980, 18(1):109-131.

[4] Stecking R, Schebesch K B. Comparing and selecting svm-kernels for credit scoring[M]. Data and Information Analysis to Knowledge Engineering. Springer, Berlin, Heidelberg, 2006: 542-549.

[5] Devasena C L. Comparative analysis of random forest, REP tree and J48 classifiers for credit risk prediction[J]. International Journal of Computer Applications, 2014: 0975-8887.

[6] Altman, E. I. Automated Credit Scoring: A Review. Journal of Finance[J]. 1968, 23(4), 611-616.

[7] Dao Bac, Kermanshachi Sharareh, Shane Jennifer, Anderson Stuart, Damnjanovic Ivan. Developing a logistic regression model to measure project complexity[J]. Architectural Engineering and Design Management, 2022, 18(3): 226-240.

[8] Vapnik, V. N., Cortes, C. Support-vector networks. Machine learning[J]. 1995: 20(3), 273-297.

[9] Debing Wang, Guangyu Xu. Research on the Detection of Network Intrusion Prevention With Svm Based Optimization Algorithm[J]. Informatica, 2020, 44(2).

[10] Fernandez-Delgado M, Cernadas E , Barro S , et al. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems[J]. Journal of Machine Learning Research, 2014, 15: 3133-3181.

[11] Lian Cuiting, Wang Yan, Bao Xinyu, Yang Lin, Liu Guoli, Hao Dongmei, Zhang Song, Yang Yimin, Li Xuwen, Meng Yu, Zhang Xinyu, Li Ziwei. Dynamic prediction model of fetal growth restriction based on support vector machine and logistic regression algorithm[J]. Frontiers in Surgery, 2022, 9.