J Inf Process Syst, Vol.17, No.1, pp.75~88, February 2021
https://doi.org/10.3745/JIPS.04.0207

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Default Prediction of Automobile Credit Based on Support Vector Machine

Ying Chen* and Ruirui Zhang*

Abstract

Automobile credit business has developed rapidly in recent years, and corresponding default phenomena occur frequently. Credit default will bring great losses to automobile financial institutions. Therefore, the successful prediction of automobile credit default is of great significance. Firstly, the missing values are deleted, then the random forest is used for feature selection, and then the sample data are randomly grouped. Finally, six prediction models of support vector machine (SVM), random forest and k-nearest neighbor (KNN), logistic, decision tree, and artificial neural network (ANN) are constructed. The results show that these six machine learning models can be used to predict the default of automobile credit. Among these six models, the accuracy of decision tree is 0.79, which is the highest, but the comprehensive performance of SVM is the best. And random grouping can improve the efficiency of model operation to a certain extent, especially SVM.

Keywords

Credit Loan, Default Prediction, Random Forest, Support Vector Machine

1. Introduction

As a pillar industry of the country, automobile industry is a powerful driving force for the development of national economy. In recent years, with the prosperity and development of the automobile industry, automobile credit business has also developed rapidly. It is estimated that in 2018, the penetration rate of automobile credit market has reached 43% [1], and the automobile credit group is mainly the post-90s generation, showing an important feature of younger gradually. However, with the development of automobile credit business, the problem of default in automobile credit has gradually emerged.

Regarding to the study of the causes of automobile credit default, the domestic and foreign scholars mainly focus on two aspects. For example, from the perspective of external economic environment, Wack [2] points out that the more subprime mortgage borrower there are, the more credit defaults there will be. From the perspective of loan-related characteristics, Lim and Yeok [3] singles out that the term of service, the existing relationship with banks, the available guarantors and the interest rate may be related to credit default. From the perspective of automobile sales price, Liu and Xu [4] states that different price ranges have a significant impact on loan default. From the borrower's point of view, Li and Ren [5] states that Logistic model has the highest accuracy in predicting default. Based on large sample data model, Shu

Manuscript received November 28, 2019; first revision January 15, 2020; second revision February 26, 2020; third revision March 23, 2020; accepted April 26, 2020.

Corresponding Author: Ruirui Zhang (zhangruiruisw@163.com)
* School of Business, Sichuan Agricultural University, Chengdu, China (18351925830@163.com, zhangruiruisw@163.com)

and Yang [6] points out that logistic model has the highest accuracy in predicting default. Liu [7] has constructed a practical default discrimination model in accordance with the combination of random forest and logical regression. Walks [8] has built an ordinary least squares (OLS) model to study consumer credit and household income.

According to the data of commercial automobile credit in India, Agrawal et al. [9] has set up a binary logistic regression prediction model. It is found that domestic and foreign scholars have relatively little research on the prediction model of automobile credit default, and that the prediction model is single, mostly using logistic regression. However, there are many models for prediction: binary classifier based on machine and in-depth learning model [10], Bayesian classifier based on improving the accuracy of default probability estimation [11], default prediction model based on support vector machine (SVM) theory [12], topic prediction model based on random forest theory [13], and data block category prediction model based on k-nearest neighbor (KNN) theory [14]. Among many methods of building prediction models, SVM can fit the data well, and random forest can analyze feature importance. So in this paper, we use random forest to choose the features of model, and the data after dimension reduction is brought into the SVM prediction model for analysis. Furthermore, this paper compares the predictive results of SVM with those of random forest, KNN classification, logistic, decision tree, and artificial neural network (ANN).

This paper mainly studies the prediction model of automobile credit default. Firstly, data pretreatment is carried out to delete some features that are not related to label variables. Then, data are processed by feature engineering. The importance of features is analyzed by the method of random forest feature selection. Secondly, the data are grouped randomly by the same number, and SVM, random forest and KNN, logistic, decision tree, and ANN prediction models are used to predict the grouped data and obtain the mean value of performance indicators, respectively. Then the prediction accuracy of these six models are compared. Finally, the classifier model with the best prediction performance indicators is used to help auto finance companies decide whether to borrow or not in the light of customer defaults.

2. Basic Algorithms

2.1 Support Vector Machine

SVM is a supervised learning method based on statistical learning theory, which is proposed by Vapnik [15]. Unlike traditional logistic regression method, SVM can process high-latitude data and solve the problem of dimensional disaster. At the same time, the SVM can avoid falling into the local minimum. SVM has good robustness and has been successfully applied in pattern recognition, signal processing and image analysis. SVM classifies high-dimensional data by kernel method. The basic idea is to map the data samples from the primitive feature space to the higher dimensional feature space through the non-linear function, and the original feature space is shown in Fig. 1.

Fig. 2 shows the distribution of features projected into a high-dimensional space. And then find the optimal classification hyperplane in the high dimensional feature space, so that the nonlinear separable problem can be transformed into a linear separable problem,

SVM can find an optimal classification hyperplane by mapping the original data samples into high-dimensional space through the kernel function $w^T\Phi(x)+b=0$, as shown in Fig. 3.

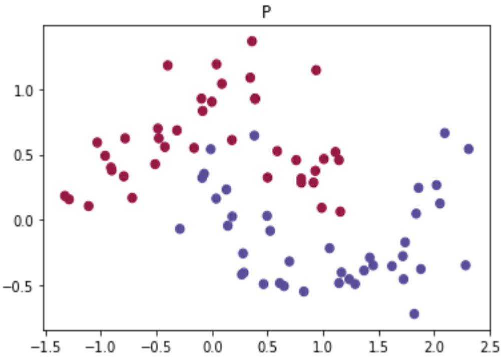


Fig. 1. The primitive feature space P.

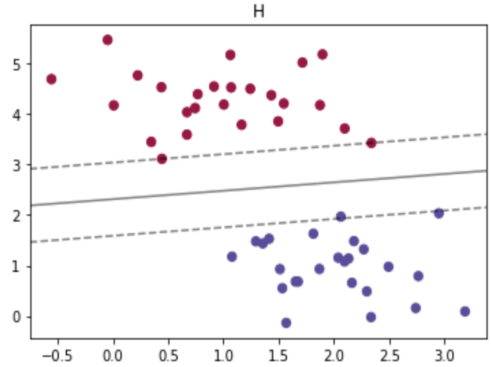


Fig. 2. The higher dimensional feature space H.

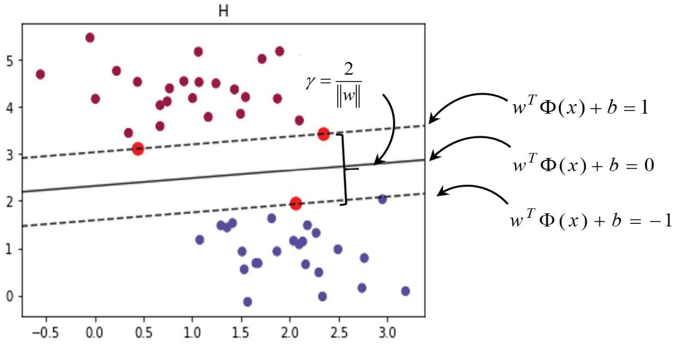


Fig. 3. Classification hyperplane of support vector machine.

The upper boundary of the hyperplane is $w^T\Phi(x)+b=1$, and the sample point is positive in the upper boundary region; the lower boundary is $w^T\Phi(x)+b=-1$, and the sample point is negative in the lower boundary region. The maximum distance between the samples is $\gamma = \frac{2}{\|w\|}$. For the convenience of calculation, the classification problem of SVM can be summarized as the following formula:

$$\begin{aligned} &\max_{a} \ L(a) = \sum_{i=1}^{m} a_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j y_i y_j \Phi(x_i)^T \Phi(x_j) \\ &s.t. \sum_{i=1}^{m} a_i y_i = 0, a_i \ge 0, i = 1,2,..., m \end{aligned} \tag{1}$$

Since the above formula involves the calculation of $\Phi(x_i)^T\Phi(x_j)$, which is the inner product of the mapping function after the original data is mapped to the high-dimensional feature space, and the feature dimension in the feature space is also very high, a kernel function is introduced to replace the inner product operation of the mapping function:

$$\kappa(x_i,x_j) = \Phi(x_i)^T\Phi(x_j) \tag{2}$$

Then the above equation is introduced into formula and written as follows:

$$\begin{aligned} &\max_a \ L(a) = \sum_{i=1}^{m} a_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j y_i y_j \kappa(x_i,x_j) \\ &s.t. \sum_{i=1}^{m} a_i y_i = 0, a_i \ge 0, i = 1,2,..., m \end{aligned} \tag{3}$$

The optimal solution can be obtained by solving the above formula. In practice, the application of different kernels will lead to the SVM classifiers with different accuracy. We must build model many times to improve the performance of the model according to the actual situation. At present, the kernels commonly used in research are:

- Linear kernel:

$$\kappa(x_i,x_j) = x_i^T x_j \tag{4}$$

- Polynomial kernel:

$$\kappa(x_i,x_j) = (x_i^T x_j)^m, m \ge 1 \tag{5}$$

- RBF kernel:

$$\kappa(x_i,x_j) = \exp(-\frac{\left\| x_i - x_j \right\|^2}{2\sigma^2}), \sigma > 0 \tag{6}$$

- Sigmoid kernel:

$$\kappa(x_i,x_j) = \tanh(\lambda x_i^T x_j + \beta), \lambda > 0, \beta < 0 \tag{7}$$

When the above four kinds of kernels are applied in different fields, the prediction accuracy of the SVM classifier is different, which depends on the actual situation to choose which kind of kernels.

2.2 Random Forest

Random forest is a machine learning method that combines multiple decision trees for prediction. Its output category is determined by the mode of the output category of a single decision tree. The random forest algorithm procedure is shown as follows:

(1) In the original training samples, bootstrap method is used to randomly extract *s* samples with playback, forming a new training sample set and constructing *s* classification decision tree. The remaining samples that are not extracted are called out-of-pocket data for prediction and evaluation.

(2) Assuming that the total number of sample features is *n*, *m* features ($m < n$) are randomly selected at the splitting nodes of each decision tree, and the information contained in each feature is calculated. Then the optimal feature is selected among the *m* features for splitting.
(3) Each decision tree is not pruned until it has grown.
(4) All decision trees are combined to form a random forest, and the predicted results are output by the random forest.

Random forests are widely used in pattern recognition [16], feature selection [17,18], and classification prediction [19]. They can process high-dimensional data, run faster, and do not need feature selection in the process of operation. At the end of operation, the importance of each feature will be given directly, and the importance of features can be improved. Random forest is a better feature selection method for row ranking.

2.3 K-Nearest Neighbor

KNN is a basic machine learning method in data mining. It does not need model training and parameter estimation. Compared with other machine learning methods, KNN classification algorithm is simple and easy to implement. It has three basic elements: K selection, distance calculation and classification rules. The core idea is that in the K most similar samples, if most of the samples belong to a certain category, the samples that need to be predicted also belong to that category. In short, "the minority is subordinate to the majority." The steps of KNN classification algorithm are described as follows:
(1) Normalize the data and preprocess other data.
(2) Calculate the distance between training samples and test samples. The calculation methods include Manhattan distance or Euclidean distance. Euclidean distance is a common distance calculation method.
(3) Ranking according to distance, K class labels nearest to the test sample set were selected.
(4) Calculate the frequency of each category in K category labels, and take the category with the highest frequency as the category of test samples.

2.4 Logistic

Logistic model is evolved on the basis of linear regression. It can not only achieve classified prediction, but also carry out regression analysis. In most cases, logistic model is used in the scenario of binary prediction. Sigmoid function, loss function and random gradient algorithm are the core contents of logistic regression model. Sigmoid function can map the variable value between 0 and 1. Loss function is used to confirm the gap between the predicted value of samples and the real value. The smaller the gap, the better the model performance. The random gradient algorithm is used to optimize the loss function. Each of these three parts performs its own duties. When the logistic model deals with the binary classification problem, if the final value is close to 0, the sample prediction result may represent the first category. Otherwise, if the value is close to 1, the sample prediction result may represent the second category.

2.5 Decision Tree

Decision tree is a common machine learning classification method, and its generation algorithms include ID3, C4.5, and CART. Different algorithms follow different principles when choosing split nodes. Generating a complete decision tree includes the determination of root node, the selection of split node and the final leaf node. If there is over fitting, the decision tree needs to be pruned. In addition, the decision tree is the basis of the random forest classifier, which can not only predict the two classifications, but also carry out regression analysis.

2.6 Artificial Neural Network

ANN has always been a research hotspot in the field of artificial intelligence. It is to imitate people's thinking ability to learn data so as to achieve the purpose of analysis. Artificial neural network includes three parts: input layer, middle layer and output layer. In the model, the data sample enters the input layer through the weighted combination of different weights, then goes through the middle layer processing, and finally gets the result from the output layer. With different weights and activation functions, the output of the model may be very different. In recent years, with the development of artificial neural network, it is widely used in different fields, such as image recognition, intelligent machine and risk prediction.

3. Example Application

3.1 Data Selection and Preprocessing

The experimental data are from the kaggle machine learning database on automobile credit default (https://www.kaggle.com/mamtadhaker/lt-vehicle-loan-default-prediction#test.csv). The total number of samples was 225,490, of which 48,967 were positive (default) and 176,523 were negative (no default).

Each sample, including label variables, are a total of 41 features. The feature *Date.of.Birth* is the date of birth of the customer. Since the date of birth of the customer is not clear, the month of birth of the customer is selected while recorded. The feature *Employment.Type* has many missing values, which are deleted when the results are little affected. In feature *Employment.Type*, "salaried" is coded as 0, while "self_employed" is coded as 1. The *DisbursalDate* is a feature of date form. In order to facilitate feature processing, we only record the year. The *perform_cns.score.description* is a multi-class feature, which has 20 values. Before coded, a (very low risk), b (very low risk), c (very low risk), d (very low risk), e (low risk), f (low risk), and g (low risk) are recorded as low risk; h (medium risk) and i (medium risk) are recorded as medium risk; j (high risk), k (high risk), l (very high risk), and m (very high risk) are recorded as high risk; no bureau history available, not scored (no activity seen on the customer (inactive)), not scored (not enough info available on the customer), not scored (only a guarantor), not scored (sufficient history not available), not scored (more than 50 active accounts found), and not scored (no updates available in last 36 months) are recorded as not scored. Then, these four categories are coded. The attribute values of feature *average.acct.age* and feature *credit.history.length* are the string forms of date. The number of years is extracted and converted into months when recording. And the year is calculated into 24 months. For example, "4yrs 8mon" is recorded as 54 months. The feature *loan_default* is the default result, whose value 0 represents no default and whose value 1 represents default. The remaining data are recorded normally. In this experiment, there are 7,661 missing values in the original data samples, which are deleted without much influence on the results.

3.2 Feature Selection

Among the 41 features (including label feature), the feature *uniqueID* is not an input variable because it obviously has nothing with automobile credit default. The attribute values of *branch_id*, *supplier_id*, *manufacturer_id*, *Current_pincode_id*, state_ID, and *Employee_code_ID* are more than ten. Because these six features have little influence on the results, they are also not used as input variables to avoid the problem of “dimension disaster.” In data preprocessing, the attribute value of the *perform_cns.score.description* is re-coded, and this feature with twenty attribute values is replaced by four encoded attribute values when imported into the predicting model. Simply speaking, a feature is processed into four features while the feature is input. So before further feature selecting, there are 36 features in the sample data (excluding label feature). Among many methods of feature selection, random forest can process high-dimensional data. It can not only calculate the importance of a single feature variable (excluding label variables), but also rank it according to its importance. So 36 feature variables are selected by random forest classifier. The results of feature selection for random forest are shown in Fig. 4.

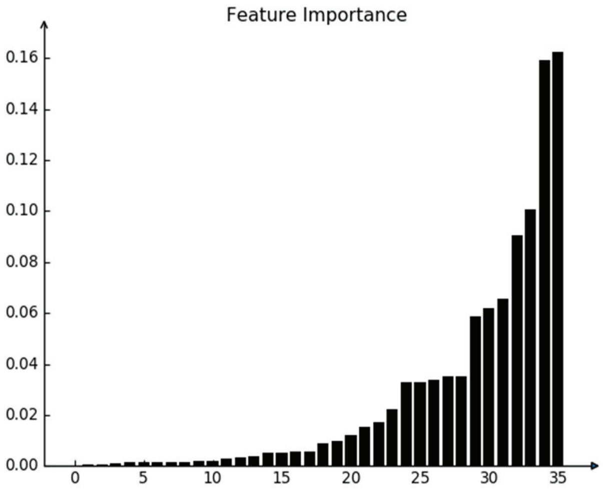


Fig. 4. Random forest feature importance ranking.

Fig. 4 shows that the degree of the most important feature is more than 0.16, and nearly half of features’ importance are less than 0.01. During the experiment, whether the feature importance is greater than 0.01 or not, it has little effect on the performance of the model. This means that those features with lower importance have little effect on the prediction performance of the model. Therefore, in the process of prediction, in order to improve the efficiency of model operation, after comprehensive consideration, 36 features are not chosen as input variables. The experiment chooses the features whose importance is greater than 0.01 as input variables. Finally, there are 17 features chosen as input variables and 1 feature chosen as output variable, whose definitions and importance are described as Table 1.

From the importance of features in Table 1, it can be seen that disbursed amount, asset cost and *DisbursalDate* have a big impact on *loan_default*, and that the importance of these three indicators has exceeded 0.01. This shows that paying attention to these three aspects of borrowers is conducive to reduce credit losses of auto financial institutions.

Table 1. Definition of characteristic variables

No	Feature	Definition	Importance
1	date.of.birth	date of birth of the customer	0.062916
2	Employment.Type	employment type of the customer (salaried/self_employed)	0.014981
3	disbursed_amount	amount of loan disbursed	0.169733
4	asset_cost	cost of the asset	0.175458
5	ltv	loan to value of the asset	0.091125
6	DisbursalDate	date of disbursement	0.102625
7	perform_cns.score	bureau score	0.047888
8	pri.active.accts	count of active loans taken by the customer at the time of disbursement	0.013023
9	pri.no.of.accts	count of total loans taken by the customer at the time of disbursement	0.024398
10	pri.current.balance	total principal outstanding amount of the active loans at the time of disbursement	0.038587
11	pri.sanctioned.amount	total amount that was sanctioned for all the loans at the time of disbursement	0.036265
12	pri.disbursed.amount	total amount that was disbursed for all the loans at the time of disbursement	0.035350
13	primary.instal.amt	EMI amount of the primary loan	0.036934
14	new.accts.in.last.six.months	new loans taken by the customer in last 6 months before the disbursement	0.010882
15	average.acct.age	average loan tenure	0.060287
16	credit.history.length	time since first loan	0.063644
17	no.of_inquiries	time since first loan	0.015904
18	loan_default	-	-

3.3 Standardization and Data Partition

During the study, standardization is used to improve the efficiency and reduce losses of predicting model. Standardization is the conversion of all data to near zero mean with variance of 1, and the formula is:

$$x_j^* = \frac{x_j - x_{mean}}{std}, (j = 1,2,,,n.) \tag{8}$$

In this experiment, the attribute values of *pri.current.balance*, *pri.sanctioned.amount*, *pri.disbursed.amount*, and *primary.instal.amt* are very different, as shown in Fig. 5.

As can be seen from Fig. 5, the range of the attribute values of *pri.disbursed.amount* is about 0 to 1 billion, and the data range of these four features is greatly reduced by standardizing their attribute values, as shown in Fig. 6.

From the comparison of the above two images, we can see that even though the range of attribute values of feature *pri.disbursed.amount* is the largest, the range of attribute values of feature has been greatly reduced by standardization. What’s more, normalization of data is also conducive to the construction of KNN, logistic, ANN, and SVM model. In this experiment, due to the great difference between the characteristic attribute values of sample data, if directly brought into the model for prediction, it will not only cause some errors, but also may affect the prediction accuracy of the model. Therefore, in order to reduce these effects, the sample data are normalized.

The number of sample data that are selected in this study is 225,490. In order to improve the accuracy of model, the sample data are disrupted line by line, and the disrupted sample data are randomly grouped by the same number. Then each group of data is divided into training set and test set. The training set accounts for 60% of the total sample, while the test set accounts for 40% of the total sample. Finally, the predicted results after grouping are recorded, and the average value of the data is calculated to evaluate the classification and prediction performance of each model.

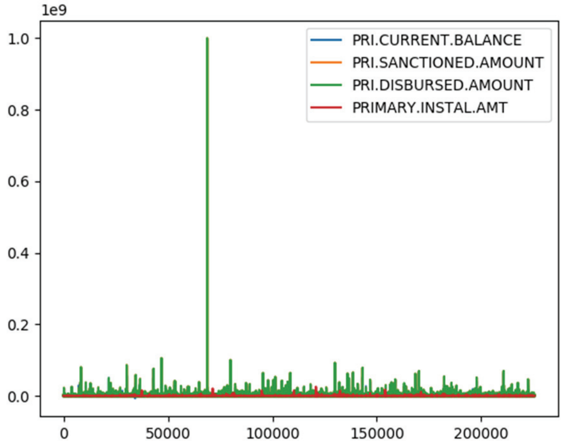


Fig. 5. Data before standardization.

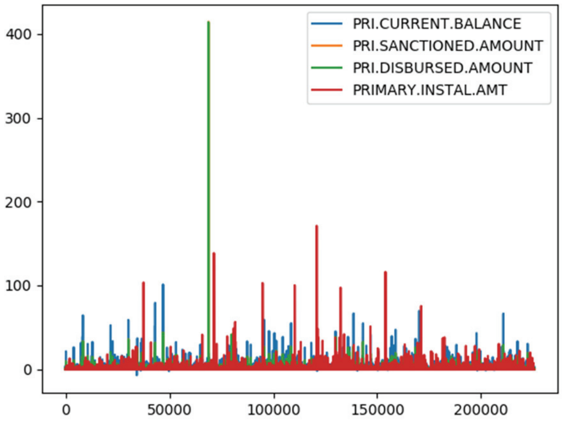


Fig. 6. Data after standardization.

3.4 Results and Analysis

We use PyCharm 2019.1 software to build SVM, random forest, KNN, logistic, decision tree, and ANN model. In order to verify whether these six classification algorithms can successfully predict automobile credit default, the training set and test set samples are imported into these six models to predict, and the corresponding performance indexes are calculated: accuracy (Acc), specificity (Spe), recall, f1_score, and AUC. And their calculation formula is described as below:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

$$Spe = \frac{TN}{TN + FP} \quad (10)$$

$$recall = \frac{TP}{TP + FN} \quad (11)$$

$$f1\_score = \frac{2 * p * r}{p + r} \quad (12)$$

where,
TP: the number of samples that are actually positive examples and also predicted positive examples;
TN: the number of samples that are actually positive examples but predicted negative examples;
FP: the number of samples that are actually negative examples and also predicted negative examples;
FN: the number of samples that are actually negative examples but predicted positive examples;
p: the abbreviation of precision;
r: the abbreviation of recall.

The above indexes are used to evaluate the prediction performance of SVM with different kernel functions, random forest, KNN, decision tree, logistic, and ANN. The predicted results of three kernel function algorithms of SVM are shown in Fig. 7.

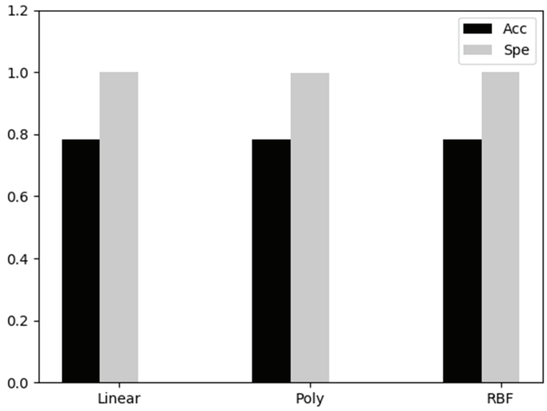


Fig. 7. Prediction results of three kernel function.

As can be seen from Fig. 7, no matter which of the three kernel functions is used by SVM, their accuracy and specificity are almost the same, which shows that these three kernel functions are not important factors affecting the prediction performance of SVM. At the same time, we compare the overall performance of SVM with other prediction algorithms, as shown in Table 2.

The prediction results of these classification algorithms show that the accuracy of decision tree is the highest, reaching 0.79. The recall, f1_score and AUC of SVM, ANN and Logistic are the same. This means that the prediction performance of these three classification algorithms is almost the same under this application situation. However, after comprehensive comparison, it is found that the prediction performance of SVM is slightly better. Due to the problem of data imbalance in this paper, in order to further improve the performance of the model, the method of up sampling is used to balance the samples. The results are shown in Table 3.

Table 2. Prediction results of six classification algorithms (before over sampling)

Classification method	Acc	recall	f1_score	AUC
Overall performance of SVM	0.78	1	0.88	0.78
Random Forest	0.78	0.96	0.87	0.77
KNN	0.74	0.92	0.85	0.74
Decision Tree	0.79	0.77	0.78	0.66
Logistic	0.69	1	0.88	0.78
ANN	0.69	1	0.88	0.78

Table 3. Prediction results of six classification algorithms (after over sampling)

Classification method	Acc	recall	f1_score	AUC
Overall performance of SVM	0.78	0.78	0.69	0.50
Random Forest	0.66	0.66	0.65	0.72
KNN	0.64	0.64	0.63	0.69
Decision Tree	0.65	0.64	0.65	0.71
Logistic	0.57	0.57	0.57	0.61
ANN	0.59	0.59	0.59	0.63

Table 4. Operational results of grouped and ungrouped samples for different models

	Operation time of the model	
	Ungrouped samples	Grouped samples
SVM	8h4m54.90s	7m43.91
Random Forest	14.8s	10.3s
Decision Tree	9.6s	3.5s
KNN	4m11.8s	1m26.0s
ANN	1m58.2s	2m59.1s
Logistic	7.98s	3.00s

It can be seen from Table 3 that the results of Acc, recall and f1_score of SVM are the best after over sampling, but AUC is the lowest of the six prediction algorithms, which indicates that the classification effect of SVM is not ideal. Compared with Tables 2 and 3, it can be found that the results of the data without over sampling are better, and that the performance of SVM is slightly better. In general, when we encounter data imbalance, we will use various methods to solve this problem, or integrate various algorithms to optimize this problem. The results of this paper show that to a certain extent, without data imbalance processing, we can directly predict through the algorithm, and the effect is better.

In the experiment, we find that grouping data can improve the efficiency of the model. As we all know, when dealing with a large number of data, the operational efficiency of SVM will become very low. And there are 233,154 samples in this experiment, so we randomly divide the samples into 10 groups, then the ten groups of data are brought into the six models. The results are shown in Table 4.

In Table 4, we can see that after grouping, except ANN, the operational efficiency of the other five prediction models has been improved. The most obvious improvement of the operational efficiency is SVM. When the samples are not grouped, the model operation time is 8 hours and 4 minutes and 54.9 seconds. After the samples are grouped, the model operation time is 11 minutes and 42.8 seconds. This shows that in dealing with large sample data, the method of random grouping can be used to improve the operation efficiency of the model.

Random grouping only can cut large samples into small samples, which improve the efficiency of the model, but cannot improve the performance of the model. We use Levene test to verify whether the variance of these 10 groups of data is different. The experimental results show that the statistics is 0.900225 and *p*-value is 0.523900. A *p*-value is greater than 0.05, which means that there is no difference in the variance of the ten groups of data. In order to be able to verify that there is no difference in the mean value of the data after grouping, we make *t*-test on these ten groups of data by pairwise comparison analysis. The details are shown in Table 5.

From Table 5, it can be seen that the *p*-value of every two groups of data is greater than 0.05, which shows that there is no difference in the mean value of these ten groups of data after random grouping. It also indirectly proves that random grouping only improves the operational efficiency but has little improvement in other aspects.

Table 5. *t*-test results of every two groups of data

	Stat	*p*-value		Stat	*p*-value
group1 & group2	-1.131921	0.257674	group3 & group10	0.423476	0.671950
group1 & group3	-0.618440	0.536289	group4 & group5	-0.593595	0.552786
group1 & group4	-0.515539	0.606179	group4 & group6	-1.298408	0.194154
group1 & group5	-1.109136	0.267378	group4 & group7	0.068639	0.945278
group1 & group6	-1.813955	0.069691	group4 & group8	-0.991861	0.321271
group1 & group7	-0.446900	0.654949	group4 & group9	0.458171	0.646832
group1 & group8	-1.507405	0.131714	group4 & group10	0.320576	0.748533
group1 & group9	-0.057368	0.954252	group5 & group6	-0.704809	0.480933
group1 & group10	-0.137595	0.845423	group5 & group7	0.662233	0.507825
group2 & group3	0.513479	0.607619	group5 & group8	-0.398265	0.690437
group2 & group4	0.616380	0.537647	group5 & group9	1.051767	0.292912
group2 & group5	0.022785	0.981822	group5 & group10	0.914171	0.360632
group2 & group6	-0.682024	0.495227	group6 & group7	1.367047	0.171617
group2 & group7	0.685019	0.493336	group6 & group8	0.306544	0.759192
group2 & group8	-0.375479	0.707306	group6 & group9	1.756585	0.078995
group2 & group9	1.074553	0.282581	group6 & group10	1.618988	0.105457
group2 & group10	0.936957	0.348786	group7 & group8	-1.060500	0.288923
group3 & group4	0.102900	0.918042	group7 & group9	0.389532	0.696884
group3 & group5	-0.490694	0.623645	group7 & group10	0.251937	0.801091
group3 & group6	-1.195506	0.231896	group8 & group9	1.450036	0.147055
group3 & group7	0.171539	0.863801	group8 & group10	1.312439	0.189379
group3 & group8	-0.888960	0.374029	group9 & group10	-0.137595	0.890561
group3 & group9	0.561072	0.574752	-	-	-

**Significant difference at 0.05 level.

4. Conclusion

Based on SVM theory, this paper constructs three kinds of kernel function prediction models in order to explore which kind of kernel function of SVM makes better prediction results by using the automobile credit data. The results show that when the kernel function is linear or RBF (radial basis function), the prediction results are the same. Then the average value of the predicted results of three kinds of kernels is calculated, and we compare this result with the prediction results of random forest, KNN, logistic, decision tree, and ANN. The test results preliminarily verify that these six algorithms can be applied to predict the default of automobile credit, which is helpful for the automobile financial institutions to evaluate the default risk of loans. However, during the experimental process, we use the over sampling method to solve the problem of data imbalance, and the results show that the performance of the model has not improved, which shows that to some extent, we do not need to focus on solving the problem of data imbalance. At the same time, we find that random grouping can shorten the running time of SVM, logistic, decision tree, random forest, and KNN these five models. Among them, the most obvious improvement of the performing efficiency is SVM, which shows that in the future, if we use SVM to process large sample data, we can use random grouping method. What's more, the research finds that the model of automobile credit default prediction is relatively simple. We need to explore whether other better prediction models can be constructed in this application scenario. At the same time, in the actual situation, the influencing factors of consumer credit default are more complex. Different automobile financial institutions establish different credit mechanisms. We need to use the most representative features to build targeted predicting models in the different scenarios.

Acknowledgement

This study is funded by the Sichuan Provincial Education Department project (No. 2018ZB0460).

References

[1] Y. Li, "2018: What happened to the Chinese car industry?," *Shanghai Enterprise*, vol. 2019, no. 1, pp. 50-52, 2019.

[2] K. Wack, "Large banks assuming more risk in auto lending," 2019 [Online]. Available: https://www.americanbanker.com/news/large-banks-assuming-more-risk-in-auto-lending.

[3] H. E. Lim and S. G. Yeok, "Estimating the determinants of vehicle loan default in Malaysia: an exploratory study," *International Journal of Management Studies*, vol. 24, no. 1, pp. 73-90, 2017.

[4] H. Liu and J. Xu, "An empirical analysis of my country's auto consumer loan default: micro-evidence from commercial banks," *South China Finance*, vol. 1, no. 5, pp. 84-91, 2015.

[5] Y. Li and J. Ren, "Research on the causes of risk of default of personal vehicle loan in automobile finance company," *China Market*, vol. 2011, no. 2, pp. 138-140, 2011.

[6] Y. Shu and Q. Yang, "Research on auto loan default prediction based on large sample data model," *Management Review*, vol. 29, no. 9, pp. 59-71, 2017.

[7] K. Liu, "Application of random forest and logical regression model in default prediction," *China Computer & Communication,* vol. 2016, no. 21, pp. 111-112, 2016.

[8] A. Walks, “Driving the poor into debt? Automobile loans, transport disadvantage, and automobile dependence,” *Transport Policy*, vol. 65, pp. 137-149, 2018.

[9] M. Agrawal, A. Agrawal, and A. Raizada, “Predicting defaults in commercial vehicle loans using logistic regression: case of an Indian NBFC,” *CLEAR International Journal of Research in Commerce & Management*, vol. 5, no. 5, pp. 22-28, 2014.

[10] P. M. Addo, D. Guegan, and B. Hassani, “Credit risk analysis using machine and deep learning models,” *Risks*, vol. 6, no. 2, article no. 38, 2018. https://doi.org/10.3390/risks6020038

[11] R. Kazemi and A. Mosleh, “Improving default risk prediction using Bayesian model uncertainty techniques,” *Risk Analysis: An International Journal*, vol. 32, no. 11, pp. 1888-1900, 2012.

[12] X. Guo and Y Wu, “Analysis of investment decision in P2P lending based on support vector machine,” *China Sciencepaper Online*, vol. 2017, no. 5, pp. 542-547, 2017.

[13] C. Jiang, Z. Wang, R. Wang, and Y. Ding, “Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending,” *Annals of Operations Research*, vol. 266, no. 1, pp. 511-529, 2018.

[14] X. Deng and L. Zhao, “Speeding K-NN classification method based on data block mixed measurement,” *Computer and Modernization*, vol. 2012, no. 12, pp. 47-50, 2016.

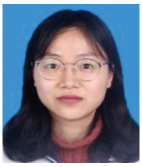
[15] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY: Springer, 2010.

[16] N. Radhika, S. B. Senapathi, R. Subramaniam, R. Subramany, and K. N. Vishnu, “Pattern recognition based surface roughness prediction in turning hybrid metal matrix composite using random forest algorithm,” *Industrial Lubrication and Tribology*, vol. 65, no. 5, pp. 311-319, 2013.

[17] D. Yao, J. Yang, and X. Zhan, “Feature selection algorithm based on random forest,” *Journal of Jilin University (Engineering and Technology Edition)*, vol. 44, no. 1, pp. 137-141, 2014.

[18] B. Gregorutti, B. Michel, and P. Saint-Pierre, “Correlation and variable importance in random forests,” *Statistics and Computing*, vol. 27, no. 3, pp. 659-678, 2017.

[19] D. Denisko and M. M. Hoffman, “Classification and interaction in random forests,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 8, pp. 1690-1692, 2018.



Ying Chen **https://orcid.org/0000-0003-2177-9749**

She received B.S. degree in School of Business from Nanjing University of Posts and Telecommunications in 2017. Since September 2018, she has been studying for M.S. degree in Business Administration at the Business School of Sichuan Agricultural University.

Ruirui Zhang **https://orcid.org/0000-0003-1898-1487**

She received B.S., M.S., and Ph.D. degrees in School of Computer Science from Sichuan University in 2004, 2007, and 2012, respectively. She is a lecturer at the School of Business, Sichuan Agricultural University, China. Her current research interests include network security, wireless sensor networks, intrusion detection and artificial immune systems.