The Practice Study of Consumer Credit Risk Based on Random Forest

Cui-zhu MENG, Bi-song LIU and Li ZHOU
China national institute of standardization, Beijing, China

Keywords: Consumer credit risk, Loan, Random forest, Auto finance, Data mining.

Abstract. How to evaluate and identify the potential default risk of the borrower before issuing the loan is the basis and important link of the credit risk management of modern financial institutions. Based on the data provided by an auto finance institution, This paper mainly studies how to analyze the historical loan data of auto financial institutions with the help of the idea of unbalanced data classification, and predicts the possibility of loan default based on Random forest classification model, which provides a reference for the risk control of this institution.

Introduction

According to the data of China auto industry association, the sales volume of China's auto market in 2015 was 24.597.76 million units, an increase of 4.7% year on year, is the lowest growth rate since 2012. On the contrary, the growth rate of auto finance business has maintained a high level. Relevant data show that in 2014, the size of auto financial market exceeded 700 billion, and the penetration rate of auto finance exceeded 20%. In 2015, the overall size of China's auto financial market was about 800 to 900 billion, and the overall penetration rate was about 35%.

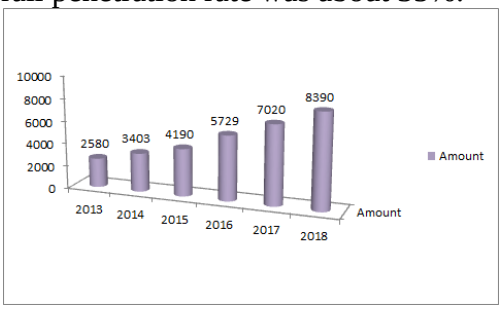


Figure 1. Demostic auto market trend

Introduction to Credit Risk in Auto Finance

At present, in the credit risk management of auto finance companies, subjective judgment is the main way to identify and evaluate the risk, which means based on experience and full of randomness. The basic data used in the model mostly come from the qualitative judgment of credit personnel, which cannot achieve the ideal effect of risk management. In future business operation, in order to improve the technical level of credit risk management, most auto financial institutions willing focus on quantitative indicators, establish a risk control mechanism using loan risk degree model and behavioral scoring model as tools, and use mathematical statistics model to measure and analyze risks, so as to achieve a reasonable offset of risks. Under this context, this paper provides a reference for credit granting of auto financial institutions by data modeling of an auto financial company.

Processing the Data

Viewing Data

The data used in this paper are randomly extracted from actual application data of a domestic mainstream auto finance company in the past three years. The data amount is 45w, including 14 variables.

Data Acquisition

From the data view, the results show that the number of missing characteristic quantities MonIncome is 89,194, while NumOfFamily is less, the number is 11,772.

Data Preprocessing

Missing Value Processing. The missing rate of variable MonIncome is relatively high, so we fill the missing value according to the correlation between variables and adopt random forest method to fill it. NumOfFamily are missing less and deleted directly.

Abnormal Value Processing. Except the missing values are processed, abnormal values also need to be processed. Abnormal values generally refer to values that deviate from the data. In this dataset, abnormal values can be clearly seen by drawing box diagrams, such as:

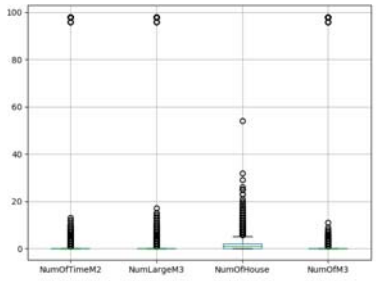


Figure 2. Data box diagram

It is obvious that three of the four features deviate from the distribution of other samples and can be removed.

Univariate Exploratory Analysis. Before building models, exploratory data analysis (EDA) is usually performed on existing data; this practice mainly analyzes the distribution of default rate on each independent variable, and generates the frequency distribution table shown in following tables.

Table 1. Frequency distribution of variable Age

	TotalNumber	Ratio	OverdueNumber	RatioInCategory
<25	6056	2.12%	676	11.26%
26-35	36916	12.33%	4106	11.18%
36-45	59638	19.96%	5256	8.91%
46-55	73380	25.50%	5572	7.71%
56-65	66812	22.48%	3062	4.72%
>65	57198	19.22%	1380	2.49%

As can be seen from table 1, people younger than 25 years of age and 26-35 years of age have a default rate of more than 10%. As age increases, default rates fall.

Table 2. Frequency distribution of variable Num of House

	TotalNumber	Ratio	OverdueNumber	RatioInCategory
<5	447165	99.37%	29826	6.67%
6-10	2205	0.49%	404	18.30%
11-15	540	0.12%	123	22.85%
16-20	40	0.01%	9	22.40%
>20	41	0.01%	9	21%

As can be learned from table 2, the number of real estate and mortgages of 99.37% borrowers is less than 5, but the default rate for more than 5 borrowers has increased significantly, with more than 10 of borrowers having a default rate of more than 20%.

Table 3. Frequency distribution of variable Num of TimeM2

	TotalNumber	Ratio	OverdueNumber	RatioInCategory
0	378495	84.11%	18925	5%
1	48240	10.72%	6754	14%
2	13905	3.09%	3657	26.30%
3	5490	1.22%	1927	35.10%
4	2160	0.48%	877	40.60%
5	990	0.22%	416	42%
6	405	0.09%	206	50.90%
>6	315	0.07%	151	48.07%

As can be seen from table 3, the default rate for borrowers who did not have 30-59 days overdue was only about 4%, but the default rate increased significantly as the number of overdue increases. The other two variables are the same trend as in table 4 for the frequency distribution of the number of overdue 60-89-day overdue and borrower occurrences of 90 and above. It can therefore be concluded that the greater the number of overdue borrowers, the higher the default rate.

Data Partitioning. In order to test the model better, we divide the data into training set and test set. The test set takes 30% of the original data.

Variable Selection

Feature variable selection is very important for data analysis and machine learning practitioners. In this paper, we compare the default probability of index sub-boxes and corresponding sub-boxes to determine whether the variable meets the statistics significance.

Variable Binning (Variable Discretization)

Variable binning is a term for continuous variable discretization. We first choose the optimal binning of continuous variables, and then consider the algorithm of better grouping of continuous variables by equal-length intervals when the distribution of continuous variables does not meet the requirements of optimal binning.

After grouping, for group i, the calculation formula for WOE is as follows:

$$\mathrm{woe}_i = \ln\left(\frac{\mathrm{p_{y1}}}{\mathrm{p_{y0}}}\right) = \ln\left(\frac{\#B_i/\#B_T}{\#G_i/\#G_T}\right) \tag{1}$$

Among them, p_{y1} is the response customer in this group (in the risk model, which refers to the value of the predicted variable in the model as "yes" or 1 of the individuals) accounted for the proportion of all responding customers in all samples, p_{y0} is the proportion of unresponsive customers in this group who account for all unresponsive customers in the sample. $\#B_j$ is the number of responding customers in this group, $\#G_j$ is the number of unresponsive customers in this group, $\#B_T$ is the number of all responding customers in the sample, $\#G_T$ is the number of all unresponsive customers in the sample.

IV The calculation formula is as follows:

$$IV_i = \left(\frac{\#B_i}{\#B_T} - \frac{\#G_i}{\#G_T}\right) * \ln\left(\frac{\#B_i/\#B_T}{\#G_i/\#G_T}\right)$$
$$\mathrm{IV} = \sum_{k=0}^{n} \mathrm{IV}_i \tag{2}$$

Variable Correlation Analysis

Before modeling, it is necessary to examine the dependencies between variables, and if there is a strong correlation between the independent variables, the accuracy of the model is affected, and if there is a strong correlation between the independent variable and the dependent variable, you should pay more attention.

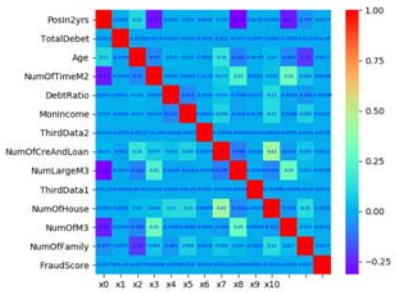


Figure 3. Variable Correlation Analysis

We can see from the figure above:

1) The correlation between the respective variables is very small.

2) Also, we can see three features have a strong correlation with the value desired we want to predict: NumOfTimeM2，NumLargeM3 and NumOfTimeM3.

IV Prediction

IV value is a quantitative indicator that measures the predictive ability of an independent variable, and from the IV values of each variable, obviously, the IV value of DebtRatio, MonIncome, NumOfCreAndLoan, NumOfHouse and NumOfFamily is very low, so we directly delete them. Thus we choose x1, x2, x3, x7 and x9 as useful variables used for following model construction.

Model Development and Evaluation

Introduction to Random Forest

Random forest algorithm is a random method to establish a forest, which is a combinatorial learning algorithm based on decision tree. The basic idea of random forests is that in the process of constructing a single tree, some variables or features are randomly selected to participate in tree node division, repeat multiple times and ensure the independence between these trees. After the random forest is obtained, when a new input sample enters, each decision tree in the forest will judge the sample, get the result of which class the sample belongs to, and finally see which category of the whole forest belongs to the highest, predict which category the sample belongs.

Principle and Characteristics of Random Forest Algorithm

Random forest algorithm works in four steps:

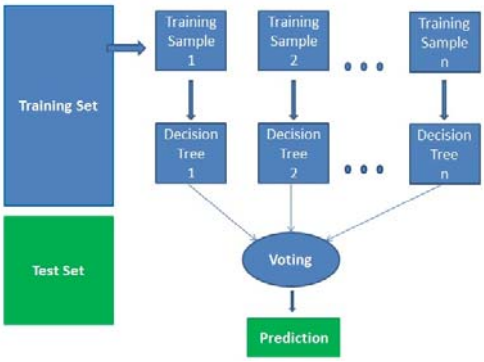
a. Select random samples from a given dataset.

b. Construct a decision tree for each sample and get a prediction result from each decision tree.

c. Perform a vote for each predicted result.

d. Select the prediction result with the most votes as the final prediction.

Table 4. Random forest steps



Random Forests vs Decision Trees

a. Random forests is a set of multiple decision trees

b. Deep decision trees may suffer from overfitting, but random forests prevents overfitting by creating trees on random subsets.

c. Decision trees are computationally faster

d. Random forests is difficult to interpret, while a decision tree is easily interpretable and can be converted to rules

Establishing Model

In this paper, we established the random forest model using sklearn. ensemble. Random Forest Classifier in Python. Some parameters are set to:

n_estimators: The number of decision trees is set to 100

min_samples_split: When dividing nodes according to attributes, the minimum number of samples per partition is set to 3

n_jobs: Parallel number, set to -1

bootstrap: Whether to use bootstrap sample sampling, set to True

Model Evaluation

After the model is established, we use AUC (region under the ROC curve) value for model evaluation. AUC is defined as the area under the ROC (Receiver Operating Characteristic) curve, and it is clear that the value of this area will not be greater than 1. The horizontal axis of the ROC curve is False Positive Rate, the longitudinal axis is True Positive Rate, and because the ROC curve is generally above the y=x line, the AUC values range from 0.5 and 1. The AUC value is used as the evaluation criterion because many times the ROC curve does not clearly indicate which classifier works better, and as a value, the larger classifier corresponding to AUC works well.

In this practice, the AUC value is 0.85, which indicates that the prediction effect of the model is good, with high correct rate.

Credit Scorecard Creation

Based on above model, we create the credit scorecard as follows:

Table 5. Credit score card

TotalDebet	Score	NumOfTimeM2	Score	Age	Score
<=0.0321]	24	<=0]	9	<=21]	-8
(0.0321,0.1623]	21	(0,2]	-16	(21,31]	37
(0.1623,0.5689]	7	(2,4]	-28	(31,37]	42
>0.5689	-19	(4,6]	-37	(37,41]	35
		>6	-45	(41,47]	27
				(47,54]	19
				(54,61]	8
				>61	0
NumLargeM3	Score	NumOfM3	Score	ThirdScore1	Score
<=0]	8	<=0]	3	<=400]	0
(0,1]	-27	(0,1]	-18	(400,479]	29
(1,4]	-33	(1,2]	-33	(479,538]	39
(4,6]	-43	>2	-53	(538,652]	48
>6	-62			>652	56

Conclusion

This paper mainly studies the common problem of loan default in auto financial institution, and uses random forest method of unbalanced data classification to establish the model of predictive loan default, and the basic idea of random forest is to randomly select some variables or features to participate in tree node division in the process of constructing a single tree. Repeat multiple times and ensure the independence between these trees, for unbalanced data, through parameter adjustment so that the stochastic forest method can automatically adjust the weight according to Y value, so as to effectively solve the classification of unbalanced data. Now this model has already deployed on their online system, and provides a significant suggestion for approval.

Acknowledgement

This research was financially supported by the 2017YFF0207602.

References

[1] D. Amaratunga, J. Cabrera, and Y.S. Lee. Enriched random forests. Bioinformatics, 24:2010–2014, 2008.

[2] An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring[J]. Loris Nanni, Alessandra Lumini. Expert Systems With Applications. 2008 (2)

[3] G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. Journal of Machine Learning Research, 9:2015–2033, 2008.

[4] G. Blanchard. Different paradigms for choosing sequential reweighting algorithms. Neural Computation, 16:811–836, 2004.

[5] A.M. Bruckstein, D.L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Review, 51:34–81, 2009.