The Application Study of Consumer Credit risk model in Auto Financial Institution Based on Logistic Regression

Cuizhu Meng, Bisong Liu and Li Zhou
China national institute of standardization, Beijing, China

Keywords: Consumer credit risk, Loan, Logistic regression, Random forest, Auto financial.

Abstract. Credit scoring technology is a kind of statistical model, which is widely used for Risk Assessment scoring for loan applicants, which can predict credit risk of applicants, based on information provided by customers, historical data of customers and data from third-party platforms (sesame score, Wechat score, etc.). Based on the data provided by an auto finance institution, this paper completes data processing, feature variable selection, variable WOE coding discretization, logistic regression model development and evaluation, credit scoring card establishment, which provides a reference for the risk control of this auto finance institution.

Introduction

With the rapid development of China's auto financial market, consumers' awareness and trust in auto finance are gradually enhanced with the professional, efficient and convenient service advantages of auto financial participants. More gratifying is that, the laws and regulations of the state have made many adjustments to promote the development of auto finance. At the end of March 2016, the People's Bank of China and the CBRC jointly issued the Guiding Opinions of the CBRC of the People's Bank of China on Enhancing Financial Support in the New Consumption Area, Allow auto finance companies to provide loans (or financial leases) to consumers while providing financing for additional products (including purchase tax, insurance, and even decoration) attached to the purchased vehicles according to consumers' wishes.

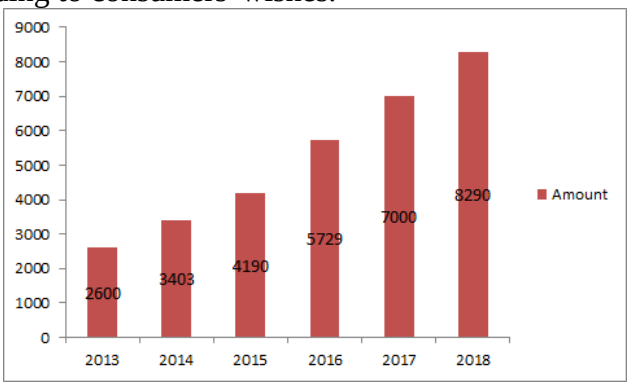

Figure 1. Demostic Auto Market Trend

Introduction to Credit Risk in Auto Finance

For auto finance, it also faces the risk of credit evaluation. Most auto loan companies have their own credit management system and credit evaluation technology, but due to their different situations, together with the lack of credit management dimension and unreasonable, the core credit management dimensions of vehicle value, credit status, work and operation status, family stability, debt status, and bad habits are not in place, which leads to post-loan risk.

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 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 the 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.

Data Processing

Data Description

The data used in this practice analysis are randomly extracted from the actual application data of a domestic mainstream auto finance company in the past two years. The data amount is 30w, including 14 variables. The general situation is shown in the table below:

Table 1. Data Description

S/N	Variable Name	Description	Type
1	SeriDin2yrs	Target variable	Y/N
2	TotalDebet	Total balance of credit card and personal credit balance, minus real estate and debt without disagreement payments	percentage
3	Age	age	integer
4	NumOfTime30-59	The number of times a borrower is 30-59 days overdue, but there has been no worse credit history in the last 2 years	integer
5	DebtRatio	debt ratio	percentage
6	MonthlyIncome	monthly income	real
7	NumOfCreditAndLoan	Number and credit limit of open loans	integer
8	NumOfTime90	Number of times a borrower is 90 days or more overdue	integer
9	NumOfRealEstate	Number of mortgages and real estate loans	integer
10	NumOfTime60-89	The number of times a borrower is 60-89 days overdue, but there has been no worse credit history in the last 2 years	integer
11	NumOfHoney	Number of family members (spouses, children, etc.)	integer
12	ThirdScore1	Third-Party credit score	integer
13	ThirdScore2	Third-Party credit score	integer
14	FraudScore	Fraud score of applicant	integer

Data Acquisition And Viewing

From the data view, the results show that the number of missing characteristic quantities MonthlyIncome is 59462, while NumOfHoney is less, the number is 7848.

Data Preprocessing

Missing Value Processing. The methods of dealing with missing values include the following: a) Direct deletion of samples with missing values, b) Fill in missing values according to similarities between samples, 3) Fill in missing values according to the correlation between variables.

The missing rate of variable MonthlyIncome is relatively high, so we fill the missing value according to the correlation between variables and adopt random forest method to fill it. NumOfHoney 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 statistics, the values below Q1-1.5IQR and higher than Q3+1.5IQR are often regarded as abnormal values. In this dataset, abnormal values can be clearly seen by drawing box diagrams, such as:

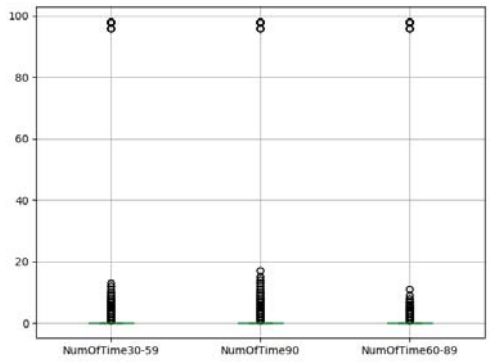

Figure 2. Data Box Diagram

It is obvious that two of the three features deviate from the distribution of other samples and can be removed. In addition, we found that the sample with age 0 is obviously not in line with common sense and should be discarded as abnormal value.

Univariate Exploratory Analysis. Before building models, exploratory data analysis (EDA) is usually performed on existing data. In this paper, we analyzed the characteristic quantities of age and MonthlyIncome as follows:

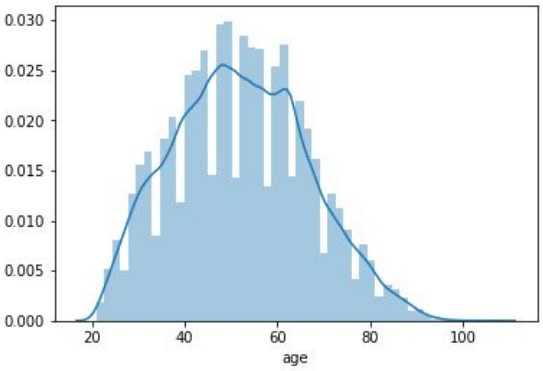

Figure 3. Characteristic Quantities of Age

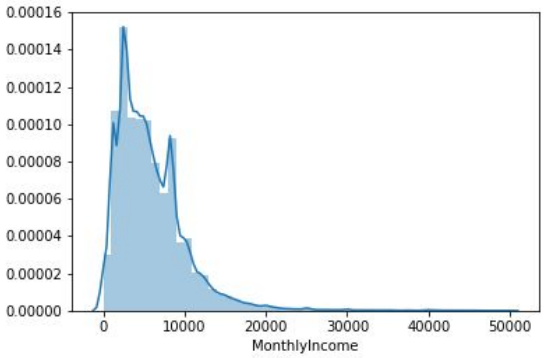

Figure 4. Characteristic Quantities of MonthlyIncome

It can be seen that the distribution of age is approximately normal distribution, which is in line with the statistical analysis hypothesis. Similarly, the distribution of MonthlyIncome is approximately normal, which conforms to the statistical analysis hypothesis.

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. For centralized variables such as TotalDebet、age、DebtRatio and MonthlyIncome, we use optimal binning to divided them.

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.

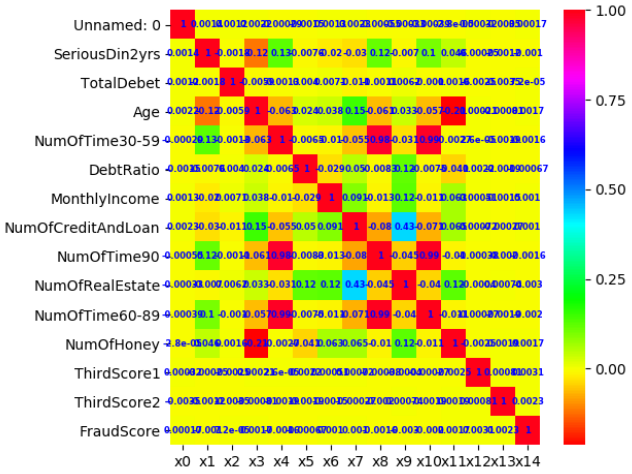

Figure 5. Variable Correlation Analysis

We can see from the figure above:

1) The correlation between the respective variables is very small. In fact, Logistic regression also needs to test the multiple collinearity problems, but here due to the small correlation between the variables, it can be preliminarily judged that there are no multiple collinearity problems.

2) Also, we can see three features have a strong correlation with the value desired we want to predict: NumOfTime30-59，NumOfTime90 and NumOfTime60-89.

IV Prediction

IV value is a quantitative indicator that measures the predictive ability of an independent variable, and the IV values of each variable are predicted below.

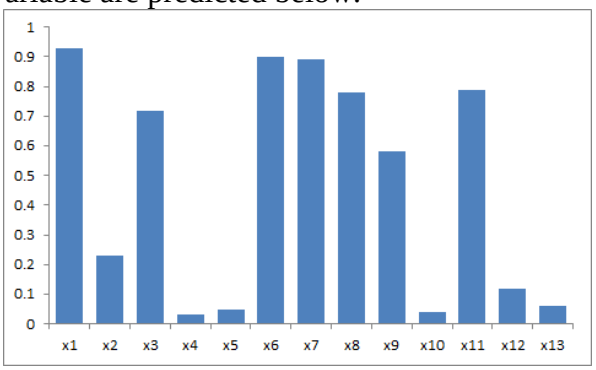

Figure 6. IV Prediction

Obviously, the IV value of DebtRatio, MonthlyIncome, NumOfCreditAndLoan, NumOfRealEstate and NumOfHoney 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

WOE Conversion

WOE conversion can transform the logistic regression model into a standard scoring card format. In this case, the logistic regression model needs to handle a larger number of independent variables.

Although this increases the complexity of the modeling program, the resulting scoring cards are the same.

Establishing Logistic Regression Model

Model Evaluation

After the model is established, the ROC curve is drawn to determine the accuracy of the model by importing the data of the test set.

WOE Conversion of Test Set Data

Fitting the Model, Drawing the ROC Curve to Obtain the AUC Value

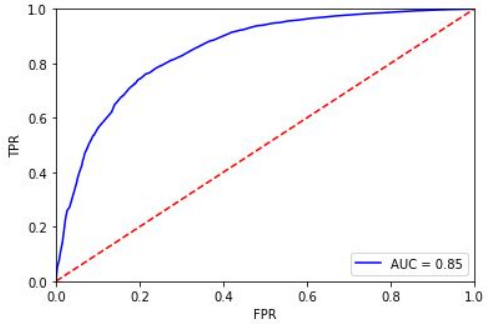

Figure 7. ROC curve

From the figure above, the AUC value is 0.85, which indicates that the prediction effect of the model is good, and the correct rate is high.

Credit Score Card Creation

The credit score card created is as follows:

Table 2. Credit Score Card

TotalDebet	Score	NumOfTime30-59	Score	Age	Score
<=0.0319]	25	<=0]	9	<=21]	-8
(0.0319,0.1613]	22	(0,2]	-14	(21,32]	38
(0.1613,0.5589]	6	(2,4]	-29	(32,37]	42
>0.5589	-19	(4,6]	-38	(37,42]	36
		>6	-44	(42,48]	28
				(48,53]	19
				(53,60]	8
				>60	0
NumOfTime90	Score	NumOfTime60-89	Score	ThirdScore1	Score
<=0]	8	<=0]	3	<=400]	0
(0,1]	-28	(0,1]	-17	(400,469]	28
(1,3]	-32	(1,2]	-32	(469,537]	38
(3,5]	-42	>2	-52	(537,642]	47
>5	-62			>642	55

Conclusion

In this paper, we have presented a general credit risk model, which can be used by local auto finance institution. Above practice is based on machine learning algorithm, through the replacement of data and the improvement of the algorithm to realize the model self-building, so that the institution's credit scoring system more and more powerful. Now this model has already deployed on their online system, and provides a significant suggestion for approval.

References

[1] J. H. Albert and S. Chib, Bayesian analysis of binary and polychotomous response data, Journal of the American Statsitcal Association 88 (1993), 669–679.

[2] R.B. Avery, P.S. Calem, and G.B. Canner, Consumer credit scoring: Do situational circumstances matter?, Journal of Banking & Finance 28 (2004), 835–856.

[3] S. P. Brooks and A. Gelman, Alternative methods for monitoring convergence of iterative simulations, Journal of Computational and Graphical Statistics 7 (1998), 434–455.

[4] P. Burns and C. Ody, Validation of consumer credit risk models, Conference Summary, Federal Reserve Bank of Philadelphia & Wharton School's Financial Institutions Center, 2004.

[5] M. J. Daniels and J. W. Hogan, Missing data in longitudinal studies: strategies for bayesian modeling and sensitivity analysis, Chapman and Hall, 2008.