Credit Card Risk Assessment Based on Machine Learning

Xuelei Niu, Yufu Zheng

Department of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730000, China
1833499492@qq.com

Abstract. Our credit card business has developed a rapidly rising stage after a slow market incubation period. The rapid development of the credit card business on one hand led to the rapid development of Chinese economy, on the other hand, however, it brought the bank a large dose of credit risk. The number of credit cards issued by commercial banks is increasing. Customer credits risk issues receive more attention. Credit rating has become an important technical means for banks to improve their risk management. This paper proposes a machine learning technology to model and analyze credit card users, and the data is oversampled by SMOTE method. Handling of abnormal values and missing values are performed and the variables are standardized. The range of values is made to fall within the same range. Finally, a logistic regression model is established. The random forest algorithm is used to verify its feasibility and effectiveness.

1. Introduction
Commercial banks act as intermediaries for financial services. Its main benefit comes from the lending of funds. Relative to the borrower, banks face the problem of information asymmetry. This special way of doing business determines that a commercial bank is a business with management and management risks. The level of risk management not only affects the survival and survival of commercial banks and the level of business development. It is also the basis of the stable and orderly operation and development of the entire financial market. Modern society has gradually stepped into the credit society. Credit is the cornerstone of the market economy. With the continuous development of credit card business, the risk issue is also becoming increasingly prominent.

There are multiple risks in credit card business. Credit risk is the main risk of credit cards. Personal credit status is receiving more and more attention and attention. A sound credit system has become a solid foundation for the healthy and sustainable development of the country's economy. A credit card is a way of paying for a non-cash transaction. As a simple credit service, it is one of the fastest growing financial businesses today. It has the following characteristics: credit cards have both payment and credit functions. The credit function is the source of credit card risk. Personal credit scores is the key to the construction of personal credit system.

Personal credit score is the key to building a credit system. Individual credit score is increasing with that rapid development of personal consumption credit and credit card business in recent year. The market is increasingly desperate for personal credit score. To study the theory and method of personal credit score, to build a personal credit score model suitable for China's national conditions, and to strengthen the research on personal credit score are the inevitable requirements of maintaining the sustained. The increased research on the personal credit score is a necessary part of the international financial crisis to keep our economy and financial sustainability, healthy and stable.
credit card risk assessment model discussed in this paper is one of the personal credit assessment models.

2. Logistic regression

2.1. Logistic regression

Logistic regression model is mainly used to analyze the relationship between independent variables and discrete dependent variables. In this study, the main research content is personal credit risk assessment. The dependent variable y is a binary variable with values of 0 and 1. Respectively. y=1 represents a customer who has breach of contract, y=0 represents a customer who has no breach of contract. Assuming a given training data set in the Logistic model \( T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\} \).

After machine learning, the classifier of the Logistic model will derive the weighting coefficients of a set of independent variables \( \{\beta_1, \beta_2, \ldots, \beta_m\} \). The result of linear weighting by the set of weights and sample data: \( x = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_m x_m \). The purpose of establishing a credit scoring model with logistic regression is to try to estimate the probability that a customer is a good or a bad customer by:

\[
p(y = 1 | X) = \frac{\exp(\beta_0 + \beta^T X)}{1 + \exp(\beta_0 + \beta^T X)}
\]

Logistic is essentially a decision model based on conditional probability. Therefore, the sigmoid function are introduced here as a discriminant function. The Sigmoid function is as shown in Fig.1.

\[
f(x) = 1 / (1 + e^{-x})
\]

Calculate the results from Figure 1. Use 0.5 as the demarcation point. If the result X>0.5, a positive class of a category value of 1. If X<0.5, a negative class of a category value of 0. Put the above fitting results: \( x = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_m x_m \). Input into the sigmoid function, in turn, a value of 0 to 1 is obtained.

2.2. Maximum Likelihood Estimation

Since the logistic regression model is a nonlinear model. It cannot be estimated using the usual least squares method. The most common method of estimating logistic regression is the maximum likelihood estimation method. The principle can be briefly described as follows.

Suppose there is N individuals: \( Y_1, Y_2, \ldots, Y_N \). Randomly extract n samples from it. Its observations are recorded as \( y_1, y_2, \ldots, y_N \). Set \( p_i = p(y_i = 1 | x_i) \) the conditional \( x_i \) probability of the result obtained under the given conditions. Then the conditional probability of the result \( y_i = 0 \) is obtained
under the same conditions \( p(y_i = 0 \mid x_i) = 1 - p_i \). So, the probability that we can get an observation is:

\[
p(y_i) = p_i^{y_i} (1 - p_i)^{1 - y_i} \tag{3}
\]

Since the observations are independent of each other. The joint distribution can be expressed as:

\[
L(\beta_0, \beta_1, \ldots, \beta_m) = \prod_{i} p_i^{y_i} (1 - p_i)^{1 - y_i} \tag{4}
\]

As shown in equation (3), equation (4) is also called a likelihood function. Among them

\[
p(y_i) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im})}
\]

The ultimate goal is to find an estimate of the parameter that will maximize the likelihood function. It is often difficult to directly find the maximum value of equation (3). A common method is to take the natural logarithm of equation (3). Because of the monotonic function \( L(\theta) \). Therefore, the parameter that achieves the maximum value \( \ln[L(\theta)] \) must also achieve the maximum value. Express the resulting maximum likelihood estimate \( \hat{\beta}_j (j = 0, 1, 2, \ldots, m) \). The corresponding estimate of the conditional probability is expressed as \( \hat{p}_j (j = 0, 1, 2, \ldots, m) \). This value \( x_i \) is an estimate of the probability of given conditions \( y_i = 1 \).

It represents the fitted or predicted value of a logistic regression model.

It can be proved that the maximum likelihood estimation of logistic regression model has the characteristics of consistency, validity and asymptotic normality in the case of random samples. Equation (1) not only overcomes some of the shortcomings of linear regression methods, but also its practical significance is obvious. It is easy to get by formula (1):

\[
\ln \frac{p(y = 1 \mid X)}{1 - p(y = 1 \mid X)} = \beta_0 + \beta^T X \tag{5}
\]

The left side of the obtained formula (5) are the logarithm of the occurrence ratio of the defaulting customer. The left side of the obtained formula (5) is the logarithm of the occurrence ratio of the customers in default. Obviously, the higher this value is, the more likely the customer is to default.

3. Data Preprocessing

3.1. Data Screening

The sample data set used in this paper is derived from the data submitted by a cardholder in a construction bank of Beijing using a credit card. This data set shows transactions that occurred within two days. Of the 284807 transactions 492 were stolen. The data set shown in Figure 2 is very unbalanced. The active class (stolen) accounted for 0.172% of all transactions. The target variable "Class" has a large difference between the normal and stolen brush categories. It can be a problem with model learning. For example, if there are 100 samples, only one of them was stolen and sampled. The remaining 99 are all normal samples. Then the learner just has to make a simple method: that is, all samples are determined to be normal samples. So before the data generation into the model of training, we must first solve the problem of unbalanced samples. The method proposed in this paper to deal with sample imbalance is oversampling. The specific operation uses SMOTE (Synthetic Minority Oversampling Technique). The basic principles of SMOET are: Sampling the nearest neighbor algorithm. Calculate the K neighbors of each minority class sample. Randomly select N samples from K neighbors for random linear interpolation. Construct a new few samples. At the same time, the new sample is synthesized with the original data to generate a new training set.

3.2. Feature Engineering

According to the Basel II standard. It only contains the digital input variables as a result of the PCA conversion. Unfortunately, due to confidentiality issues. We are unable to provide raw features and
more background information about the data. Features V1, V2,...,V28 is the main component obtained using PCA. The only features that are not converted to PCA are "Time" and "Amoun". The feature "Time" contains the number of seconds between each transaction in the data set and the first transaction. The feature "Amount" is the transaction amount, which can be used for instance-dependent cost-aware learning. The feature class is the response variable, which takes a value of 1 if it is stolen, otherwise 0.

In this paper, seven variables (V18, V9, V18, V14, V3, V17, V2, V12) are selected to complete the principal component analysis through random forest. And extract a comprehensive variable (default or not) to qualitatively reflect the risk. Before building a specific model. We perform person correlation analysis and significance analysis on the default or non-variable variables for variables other than V18, V9, V18, V14, V3, V17, V2 and V12. The remaining variables were found to be significantly correlated at least at the 0.01 level (both sides). The range of values of the Amount variable and the Time variable differs greatly from other variables. So to scale the feature Sklearn.Preprocessing.StandardScaler sorts the importance of the feature. To further reduces the variables. Using the random forest (GBDT) gradient to promote the decision tree for feature importance ranking is shown in Fig.3. So these variables are used to build the model.

4. Card Risk Assessment Model Construction

4.1. Logistic Regression

The sample is divided into two sets of training set and test set in a ratio of 7:3. Training the model with the test set, Logistic regression model is established by Sklearn in Python. Adjust the parameter penalty factor C by k-fold cross-validation. At the same time, using the regularization penalty terms L1 regularity, the model is obtained:

\[
\text{LogisticRegression}(\text{C}=1.0, \text{class_weight}=\text{None}, \text{dual}=\text{False}, \text{fit_intercept}=\text{True}, \text{intercept_scaling}=1, \text{max_iter}=100, \text{multi_class}=\text{warn}, \text{n_jobs}=\text{None}, \text{penalty}=\text{\'l2\'}, \text{random_state}=\text{None}, \text{solver}=\text{\'warn\'}, \text{tol}=0.0001, \text{verbose}=0, \text{warm_start}=\text{False})
\]

Fig.4 below shows the confusion matrix modeled after sampling:
4.2. GridSearchCV regression

The sample is divided into two sets of training set and test set in a ratio of 7:3. The test set is used to train the model. The GridSearchCV regression model is established by Sklearn in Python to obtain the model:

\[
\text{GridSearchCV}(cv='warn', \text{error_score}='raise-deprecating', \text{estimator}=\text{LogisticRegression}(C=1.0, \text{class_weight}=\text{None}, \text{dual}=\text{False}, \text{fit_intercept}=\text{True}, \text{intercept_scaling}=1, \text{max_iter}=100, \text{multi_class}=\text{warn}, \text{n_jobs}=\text{None}, \text{penalty}=\text{l2}, \text{random_state}=\text{None}, \text{solver}=\text{warn}, \text{tol}=0.0001, \text{verbose}=0, \text{warm_start}=\text{False}), \text{fit_params}=\text{None}, \text{iid}=\text{warn}, \text{n_jobs}=\text{None}, \text{param_grid}=[\{'\text{tol}': [0.001, 0.0001, 1e-05], 'C': [1, 0.1, 10, 100]\}], \text{pre_dispatch}=2*n_{\text{jobs}}, \text{refit}=\text{True}, \text{return_train_score}=\text{warn}, \text{scoring}=\text{None}, \text{verbose}=0)
\]

It can be seen from Figure 5 that the Logistic Regression regression of oversampling determines 83366 negative samples. Only 143 positive samples were judged to be positive. At the same time, there are 1909 positive samples and 25 positive samples.

It can be seen from Fig. 6 that the oversampled GridSearchCV regression determines that the negative sample is negative as 83364. The positive sample is positive as only 143. The negative sample is positive as 1911 and the positive sample is determined as there are 25 positive ones. The LogisticRegression model at the moment is not optimal. We can see that the model has determined many negative samples into positive samples. Although it is not a serious problem with the perspective of risk, from the bank's point of view, this will only cause its benefits. Loss, when we re-model by adjusting the threshold. The simulation results are shown in Fig.6.

5. Conclusion

Based on the results and discussions presented above, the conclusions are obtained as below:

(1) This article passes an example of a construction bank of Beijing. SSMTE is used to oversample the data. Outliers and the handling of missing values. And standardized the variables. Make the range of values falls within the same range. Finally, logistic regression and GridSearchCV regression are established. And the simulation compares the recall rate of logistic regression and GridSearchCV regression. The feasibility and effectiveness of the logistic algorithm is verified.

(2) Since this paper studies a bank credit card model assessment. The ultimate meaning is to determine whether the user will overdue or not repay. So identifying a positive sample is crucial. So the bigger the number of positive samples is the better. The oversampled LogisticRegression model is superior to the GridSearchCV model before oversampling.
(3) A credit card risks assessment model based on logistic regression, which has good reliability and validity. User non-default rate can be predicted based on user data. It is fully stated that logistic regression can be applied to the construction of credit card risk assessment models.

Fig.6  Adjusting the confusion matrix of threshold Logistic

References
[1] Wu Yawei. Research on Financial Risk Evaluation and Control of China's Listed Commercial Banks[D]. Tianjin University of Finance and Economics, 2012. (in Chinese)
[2] Xiao Zhi, Li Wenjuan. Personal Credit Evaluation Based on Principal Component Analysis and Support Vector Machine[J]. technical economy,2010,03:69-72. (in Chinese)
[3] Xu Chong, Dai Fuchu, Xu Suning, et al. Risk assessment and inspection of Wenchuan earthquake landslide based on logistic regression model[J].Hydrogeology and Engineering Geology,2013,40(03):98-104. (in Chinese)
[4] Li An. On the Risk Management of Commercial Banks in China[J]. Accounting Research, 2012, 18:78-80. (in Chinese)
[5] Gao Yijie. Research on credit risk control and management of credit card in China [D]. Suzhou University, 2007. (in Chinese)
[6] Yan Nan, Zhang Qiwen. Empirical Research on Credit Card Risk Management Based on Logistic Model. Business Economy, 2012,(7): 24-25. (in Chinese)
[7] Zhao Zhan. On the risk of credit card cash and its prevention. Cooperative economy and technology, 2014, 24 (11): 168-172. (in Chinese)
[8] Zhao Zhihong. Research on Risk Prevention of Credit Card Business in Process Bank. China Credit Card, 2014, 32(9): 53-56. (in Chinese)