Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2021, 9, 151-161
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Research on Personal Credit Evaluation Based on Mobile Telecommunications Data

Shaoyong Hong[1], Yan Zhang[2,3], Chun Yang[1]*

[1]Guangzhou Huashang College, Guangzhou, China
[2]Guangdong Teachers College of Foreign Language and Arts, Guangzhou, China
[3]Guangdong Foreign Trade Vocational Technology School, Guangzhou, China
Email: shy2002021@163.com, *michael_0227@126.com

How to cite this paper: Hong, S.Y., Zhang, Y. and Yang, C. (2021) Research on Personal Credit Evaluation Based on Mobile Telecommunications Data. *Journal of Data Analysis and Information Processing*, **9**, 151-161.
https://doi.org/10.4236/jdaip.2021.93010

Received: July 1, 2021
Accepted: July 30, 2021
Published: August 2, 2021

Abstract

With the rapid development of big data technology, the personal credit evaluation industry has entered a new stage. Among them, the evaluation of personal credit based on mobile telecommunications data is one of the hotspots of current research. However, due to the complexity and diversity of personal credit evaluation variables, in order to reduce the complexity of the model and improve the prediction accuracy of the model, we need to reduce the dimension of the input variables. According to the data provided by a mobile telecommunications operator, this paper divides the data into a training sets and verification sets. We perform correlation analysis on each indicator of the data in the training set, and calculate the corresponding IV value based on the WOE value of the selected index, then binning data with SPSS Modeler. The selected variables were modeled using a logistic regression algorithm. In order to make the regression results more practical, we extract the scoring rules according to the results of logistic regression, convert them into the form of score cards, and finally verify the validity of the model.

Keywords

Credit System, Weight of Evidence, Information Value, K-S Test, Logistic Regression

1. Introduction

Credit investigation refers to the collection, sorting, preservation, and processing of credit information of natural persons, legal persons and other organizations in accordance with the law, and the provision of services such as credit reports, credit evaluations, and credit information consultations [1]. Credit investigation can be divided into personal credit investigation and enterprise credit investigation. Based on residents' family income and assets, previous loans and repayments, credit overdrafts, penalties and litigation in the event of bad credit, personal credit investigation is to evaluate, record and archive personal credit ratings at any time, so as to facilitate the supplier of personal credit deciding whether to provide credit or how much to provide [2]. Personal credit evaluation is to identify the behavior of individual customers, screen out the evaluation variables that have a strong relationship with the behavior of individual customers, and use the few selected variables to establish the necessary credit evaluation model to make a prejudgment of individual credibility [3], rank customers and then distinguish between "good" and "bad" customers, which aims to offer a scientific and reasonable technical reference and decision-making basis for enterprises [4]. Credit risk assessment has become an urgent problem to be solved [5].

With the rapid development of information technology, people's ability of statistical analysis and summary of data is increasing. A credit score card model based on historical data and using statistical methods to assess customer risk begins to emerge [6]. At present, the credit evaluation models in foreign markets mainly include FICO credit score [7], Zest Finance credit score [8], and NCTUE credit evaluation. Domestic mature personal credit rating products in China include Sesame Credit [9], Jingdong Baitiao [10], and Credit Score (China Mobile) [11], etc. Personal credit score is usually regarded as a classification problem in pattern recognition. The general way to study such problems is to divide customers into good customers and bad customers [12]. Good customers include customers without downtime and customers who pay after downtime, their characteristics are: network access time is relatively long; the user value of the contact circle is high; more active days; more traffic usage; the number of overdue fees is less. Bad customers include those who do not pay in time after the overdue shutdown, their characteristics are: The duration of network access time is short; the number of active days and traffic usage is small; the value of users in the contact circle is not high; the number of overdue fees is more. In recent years, SVM (support vector machine) has been rapidly developed and widely used in the field of personal credit evaluation. Tony and Harris made an empirical analysis of loans information from customers of a financial institution by utilizing the SVM method. It turned out that the SVM model is beneficial to the research of small sample data [13] [14]. In China, there are also many researches on personal credit evaluation. Combined with principal component analysis，Li Meng constructed a Logistic model for commercial bank credit risk assessment, which proved that the Logistic model has high recognition and predictive capabilities and suffices to function effectively in commercial bank credit risk assessment [15].

The Internet of Everything has accelerated the rapid growth of the mobile telecom industry, as well as brought unprecedented transformation challenges to traditional telecom operators [16]. It is advisable that Telecom operators need to change the traditional operation methods to provide customers with faster and more personalized ones. Given that the complexity and diversity of credit evaluation variables, and the accuracy of the logistic regression model, first, the weight of evidence-information value (WOE-IV) method will be employed to select the variables [17]. Second, the dimensionality reduction variables and the logistic regression method will be utilized to record data on the customer behavior of a communication operator in China, aiming to establish a statistical analysis model for personal credit evaluation to differentiate between "good" customers and "bad" customers. Then, in accordance with the judgmental results, it's easier to provide customers with personalized marketing plans. For customers with poor credit records, increasing control can effectively reduce the risk of arrears and bad debts; for valued customers with good credit, some preferential packages and other services should be launched to attract more users to come back again, thereby enhancing the competitiveness of the enterprise [18].

2. Data and Processing Methods

In simple terms, mobile telecommunication data is the data generated by the mobile phone users of the operators. According to the data source, it can be roughly divided into identity data, terminal data, location data, billing data, call list data, communication data and Internet data. The characteristics of mobile telecom data: wide coverage; high authenticity; large amount of data; strong timeliness; multiple data dimensions.

A total of 10,185 samples were obtained from the business data records of a Chinese communications operator for the six months from January to June. Based on the basic customer information provided by mobile telecommunications operator and referring to existing scoring models at home and abroad, we roughly divide the data into six dimensions: identity characteristics, behavior preference, performance ability, credit history, relationship, external data, including customer age, basic information of overdue payment, communication behavior, online behavior, circle of friends and other data. These data are mainly structured data· For the reason that the data recording range and measurement scale of different numerical variables are not the same, so it is necessary to normalize the numerical variables. The transformation formula is as follows:

$$x'_{ij} = \frac{x_{ij} - \min x_{ij}}{\max x_{ij} - \min x_{ij}} \tag{1}$$

where x_{ij} represents the original variable value, x'_{ij} represents the value obtained after normalization, $\min x_{ij}$ represents the minimum value of all sample data in the i variable, $\max x_{ij}$ represents the maximum value of all sample data in the first variable.

The data period is divided into observation period and performance period· The evaluation index constructed by using the basic situation and behavior characteristics of customers during the observation period is called the independent variable, and the performance of whether customers owe fees in the performance period is called dependent variable. In this paper, we select the basic customer information data provided for operators from January to September, and takes June as the observation point. The sample observation period is six months, from January to June, and the sample performance period is July to September, then the window length of the performance period is three months.

The 10185 samples contained 5810 "good" customers (marked as 0) and 4375 "bad" customers (marked as 1). For the needs of credit evaluation modeling, the 10,185 sample data is randomly divided into 8148 training set samples and 2037 test set samples at a ratio of about 4:1. We use one part of the data (80%) as the training set for the establishment of the model, and the other part of the data (20%) as the verification set for the verification of the model. For the training set, the last column of indicators (whether they are owed) is the dependent variable y, and the other indicators are independent variables. The autocorrelation analysis of these indicators shows that the correlation between the two indicators is relatively high and only one of them needs to be selected. The selected indicators are divided into bins, calculate the woe value of each file, and then calculate the corresponding IV value according to the calculated woe value.

The WOE value is the weight of evidence. The higher the value of woe means the higher the probability of arrear. For the category *i* of a variable, the WOE value is calculated as follows:

$$\text{WOE}_i = \ln\left(\frac{G_i/G}{B_i/B}\right) \tag{2}$$

where G represents good customers, B represents bad customers, G_i/G represents the proportion of good customers in the category of variable *i*, and B_i/B represents the proportion of bad customers in the category of variable *i*. Use the above formula (2), we redefine the WOE expression of the variable X as:

$$\text{WOE}(X) = \beta_1\text{WOE}_1 + \beta_2\text{WOE}_2 + \cdots + \beta_r\text{WOE}_r \tag{3}$$

where $\beta_1, \beta_2, \cdots, \beta_r$ are binary dummy variables. That is, for all categories of variables $i = 1, 2, \cdots, r$, if the value of X belongs to the *i*-th class, then $\beta_i = 1$, and when X does not belong to the *i*-th class, $\beta_i = 0$.

When we calculate the woe value of a variable, we need to grade the indicators according to the following points:

First, the number of groups should be moderate, not too much or too little;

Second, in order to ensure that there are enough good and bad customer samples in each group, the number of records in each subfile should be reasonable, not too much or too little.

Third, combining with dependent variables, the segmentation should be able to show obvious trend characteristics.

Fourth, the distribution difference of dependent variables between adjacent sub grades should be as large as possible.

IV is information value. According to the credit evaluation system model, it is generally assumed that when IV < 0.1, the indicator has no effect. When 0.1 < IV < 0.3, the index has a certain effect. When IV > 0.3, the index has a significant effect. The IV value of the variable is calculated by:

$$\mathrm{IV} = \sum_{i=1}^{r} \mathrm{WOE}_i \left(\frac{G_i}{G} - \frac{B_i}{B} \right) \tag{4}$$

According to the magnitude of the IV value, the variables that have no effect are deleted, and the variables that have a certain effect are retained, so that the variables can be filtered out. We can merge groups with too few sample points or unreasonable hops with neighboring groups. Finally, SPSS Modeler was used to complete the classification. The final results are shown in Table 1.

According to the data preprocessing, the six independent variables were selected: network access time, active days, number of overdue fees, contact circles, number of traffic used, age.

3. Analysis and Discussion

The flowchart of the construction of the credit score card model is shown in Figure 1.

1) Logistic regression

The event of arrears is represented by the variable y, when $y = 1$, it is bad customer, and when $y = 0$, it is good customer. Our purpose is to use the existing sample data to build a model to predict the probability p of the rate of arrears. No matter whether we predict a new customer to be a good customer or a bad customer, the result of using logistic regression analysis is not simply to give yes or no, but to give a probability of this event.

2) Conversion of scorecard

In order to make the results of logistic regression more practical, we need to convert the results into the form of scores. So we use SPSS modeler to transform the result of logistic regression into the form of score card, as shown in Table 2.

The score should meet the following requirements:

First, control the score within a certain range, and draw up a range according to your own business needs, such as 0 to 1000 points.

Second, at a certain score, good customers and bad customers have a certain proportional relationship. There is a special statistic in statistics-odds to represent this proportional relationship. For example, when we expect a score of 500, the ratio of good and bad customers is 50:1.

Third, the increase in score value should reflect the change in the ratio between good and bad customers. For example, it is hoped that for every 50 points increase in score value, the odds will also double.

The value relationship of credit score is:

$$\mathrm{score} = \ln(\mathrm{odds}) * \mathrm{factor} + \mathrm{offset} \tag{5}$$

Based on the company's own business, we independently set the value of the ratio of good to bad customers, that is, the odds ratio, and the increase of the

Table 1. WOE value and IV value of the selected variables.

Index name	Index grading	WOE	IV
network access time	[6, 7]	1.06	0.66
	(7, 11]	0.44	
	(11, 24]	−0.24	
	(24, 36]	−0.66	
	(36, 48]	−0.81	
	(48, 72]	−1.03	
	(72, 150]	−1.32	
	>150	−1.81	
active days	(0, 2]	1.44	0.56
	(2, 10]	0.55	
	(10, 20]	−0.02	
	(20, 25]	−0.39	
	(25, 29]	−0.73	
	>29	−0.83	
number of overdue fees	0	−0.56	0.76
	1	0.93	
	2	1.58	
	3	1.69	
	4	2.02	
	5	1.95	
	6	1.90	
contact circle	(0, 1]	0.86	0.34
	(1, 4]	−0.12	
	(4, 8]	−0.38	
	(8, 15]	−0.53	
	(15, 30]	−0.69	
	≥31	−0.73	
number of traffic used	(0, 30]	0.79	0.26
	(30, 200]	0.03	
	(200, 500]	−0.29	
	(500, 1000]	−0.45	
	(1000, 2000]	−0.56	
	>2000	−0.27	
age	≤18	0.80	0.12
	(18, 25]	0.14	
	(25, 50]	−0.13	
	>50	0.29	

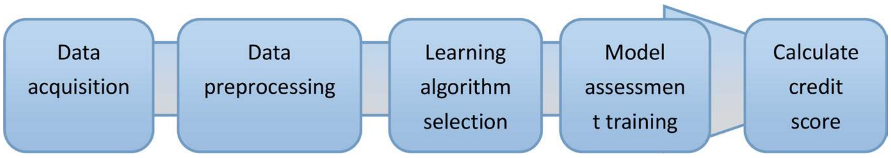

Figure 1. Flowchart of credit card scoring.

Table 2. score card.

Identity characteristics	Age						
	(0, 18]	(18, 25]	(25, 50]	>50			
	0	20	50	30			
Behavioral preferences	Active days (days)						
	≤2	(2, 10]	(10, 20]	(20, 25]	(25, 29]	>29	
	0	23	46	66	100	155	
	Number of traffic used (M)						
	[0, 30]	[30,200]	[200,500]	(500,100 0]	(1,000,200 0]	>2000	
	0	20	40	80	100	60	
Performance capability	Number of overdue fees in recent 6 months (times)						
	0	1	[2, 3]	[4, 6]			
	320	100	50	0			
Credit history	Network access time (month)						
	[6, 7]	(7, 11]	(11, 24]	[24] 13	[48,100]	[100,180]	>180
	0	20	50	80	120	150	245
Connections	Contact circles (ones)						
	≤1	(1, 4]	(4, 8]	(8, 15]	(15, 30]	≥31	
	0	30	40	50	80	128	

score value when the odds doubles. In this paper, the proposed value is debugged several times in combination with the operator's own business. Finally, it is determined that when the value of good customers is 30:1 compared with bad customers, the corresponding score is 500 points, and when the score value is increased by 50 points, the odds are doubled. Therefore, according to the scoring formula, we can get:

$$500 = \ln(30) \cdot \text{factor} + \text{offset} \tag{6}$$

$$550 = \ln(60) \cdot \text{factor} + \text{offset} \tag{7}$$

Using the above formula, we can get the value of factor and offset. The formula for calculating the score value of each file is:

$$\text{score} = \left(\text{WOE} * \beta + \frac{\alpha}{n}\right) * \text{factor} + \frac{\text{offset}}{n} \tag{8}$$

where α and β respectively represent the intercept value and coefficient value of the logistic regression results, n is the number of input variables. WOE,

α , β change with different grades of the calculated variables

3) Model verification

K-S (kolmogorov-smironov) test index is a common test index of the current industry scoring model. It mainly verifies the ability of the model to distinguish good customers from bad customers by calculating the maximum difference of the cumulative percentage of the two types of customers, whose detailed calculation process is shown in Figure 2.

Obviously, the value of KS is between [0,1]. In theory, we calculate the level of KS, which represents the effectiveness of the model. In practical application, the KS value of the model up to 0.2 is acceptable, while the value up to 0.4 indicates that the model has good distinguishing ability, while the value above 0.5 indicates that the model has strong distinguishing ability. The K-S value of this model is shown in Table 3, and the corresponding diagram is shown in Figure 3.

It can be obtained that the most obvious difference between good and bad customers is in the [300 - 400] range. The KS value of the model is 60.27%, which shows that the model works well.

According to the probability value (P value) predicted by the model, the "good" customer and the "bad" customer are estimated. When P > 0.5, they are classified as "bad" customers (Y = 1). When P ≤ 0.5, they are classified as "good" customers (Y = 0). The confusion matrix between the actual value of the original sample data and the predicted value of the model is shown in Table 4.

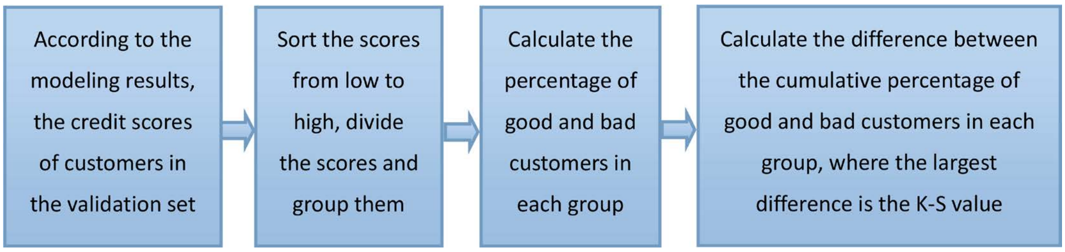
According to the modeling results, the credit scores of customers in the validation set
Sort the scores from low to high, divide the scores and group them
Calculate the percentage of good and bad customers in each group
Calculate the difference between the cumulative percentage of good and bad customers in each group, where the largest difference is the K-S value

Figure 2. Calculation process of K-S value.

Table 3. the results of K-S value.

Score grading	Ks-good	Ks-bad	Ks value
0 - 100	0.29%	16.52%	16.23%
100 - 200	2.21%	45.86%	43.65%
200 - 300	5.31%	56.97%	51.66%
300 - 400	13.78%	74.04%	60.27%
400 - 500	34.71%	86.11%	51.41%
500 - 600	64.10%	93.19%	29.08%
600 - 700	86.74%	97.38%	10.64%
700 - 800	97.73%	99.50%	1.77%
800 - 900	99.93%	99.99%	0.05%
900 - 1000	100.00%	100.00%	0.00%

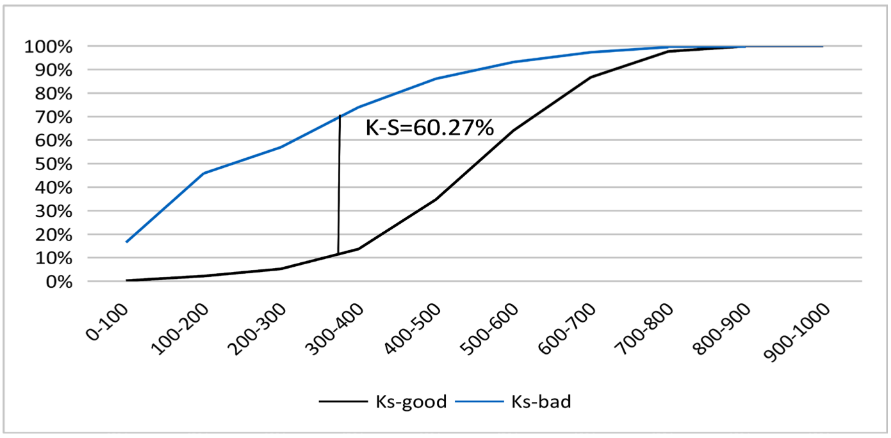

Figure 3. K-S index diagram.

Table 4. Classification results of logistic regression model based on WOE-IV selecting characteristic variables

Training sets		Prediction		Classification accuracy (%)	Testing sets		Prediction		Classification accuracy (%)
		0	1				0	1	
Real date	0	4290	311	93.24	Real date	0	1031	161	86.49
	1	794	2753	77.61		1	146	699	82.72
Total classification accuracy (%)				86.44	Total classification accuracy (%)				84.93

4. Conclusions

We perform correlation analysis on each indicator of the data in the training set, and calculate the corresponding IV value based on the WOE value of the selected index, then binning data with SPSS Modeler. The selected variables were modeled using logistic regression algorithm. From the results of model analysis, logistic regression models have the following advantages: 1) Better stability and stronger robustness. 2) The model is intuitive. The meaning of coefficient is easy to explain and understand. 3) When the effect of the model we built has declined, the logical model can better diagnose the cause of disease.

Through the evaluation of personal credit, the user group can be differentiated according to user credit level to adopt the corresponding marketing operation plan for different groups to achieve precise marketing. By identifying and strengthening the control of poorly valued customers, the risk of arrears and bad debts can be effectively reduced. For high-quality customers with good credit, we can push some preferential packages and other services, so as to improve the stickiness of these users. There are many methods to establish credit evaluation model, each method has its own advantages and disadvantages. In this paper, the linear method is used to establish the evaluation model, which has good robustness and model interpretation ability, but the linear method cannot extract the nonlinear relationship in the data, which is not conducive to the processing of large-scale sample data. How to organically combine machine learning methods with traditional logistic regression methods will be the focus of the later research in this article.

Acknowledgements

This work was supported by Guangdong University Youth Innovation Talent Project under Grant NO. 2019KQNCX213, the Scientific Project of Guangzhou Huashang College under Grant NO. 2019HSDS25.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Liu, X.G. and Wang, T.Y. (2016) International Development of Credit Information and Its Application in China's Insurance Industry. *Financial Computerization*, No. 10, 48-50.

[2] Wu, K. (2010) Construction of China's Personal Credit System. Southwestern University of Finance and Economics, Chengdu.

[3] Zhang, C. and Wan, X. (2019) Construction of Personal Credit Evaluation System and Evaluation Model under the Background of Big Data. *Credit Information*, No. 10, 66-71.

[4] Galindo, J. and Tamayo, P. (2000) Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. *Computational Economics*, **15**, 107-143. https://doi.org/10.1023/A:1008699112516

[5] Zhou, Y.S., Cui, J.L., Zhou, L.Y., *et al.* (2020) Research on Personal Credit Risk Assessment Based on Improved Stochastic Forest Model. *Credit Reference*, No. 1, 28-32.

[6] Li, Y.H. (2010) The Establishment of Credit Score Card Model. *Science and Technology Information*, **37**, 48-49.

[7] Berger, A.N., Frame, W.S. and Miller, N.H. (2005) Credit Scoring and Availability, Price, and Risk of Small Business Credit. *Journal of Money Credit & Banking*, **37**, 191-222. https://doi.org/10.1353/mcb.2005.0019

[8] Liu, X.H. and Ding, W. (2015) Big Data Credit Investigation Practice of American Zest Finance Company. *Credit Investigation*, No. 8, 27-32.

[9] Sun, J.Y., Zhang, M.M. and Wan, S.Y. (2017) SWOT Analysis on the Development of China's Credit Economy—Taking the Alipay Platform Ant Flower Bai as an Example. *China Business Review*, No. 8, 148-149.

[10] Kit (2015) Big Data Credit Reference Is the Cornerstone of JD White Slip. Science and Technology Daily, 2015-07-08, 011.

[11] Wang, B., Chen, B., Wei, Y.H., *et al.* (2016) Research on Construction Method and Business Model of Credit Evaluation System Based on Telecom Big Data. *Mobile Communications*, **40**, 75-79.

[12] Chen, J. and Yang, T.N. (2005) UML Modeling Method for Personal Credit Evaluation System for College Students. *Journal of Chongqing University*, No. 11, 62-64.

[13] Tony, D. and Gestel, V. (2003) A Support Vector Machine Approach to Credit Scoring. *Bank Financiewezen*, **12**, 73-82.

[14] Harris, T. (2015) Credit Scoring Using the Clustered Support Vector Machine. *Ex-*

pert Systems with Applications, **42**, 741-750. https://doi.org/10.1016/j.eswa.2014.08.029

[15] Li, M. (2005) Application of Logit Model in Credit Risk Assessment of Commercial Banks. *Management Science*, No. 2, 33-38.

[16] Wu, H.Q. (2020) New Features of Network Society in 5g Era and Challenges Facing Industry. *Journal of Chongqing University of Posts and Telecommunications* (*Natural Science Edition*), **32**, 171-176.

[17] Chen, Z.Y. (2020) Perfect Combination: Research on the Credit Scoring Card Model of Online Lending Based on Machine Learning. *Wuhan Finance*, No. 3, 42-50.

[18] Chen, Q.H., Yang, H.R. and Cui, H.J. (2020) Personal Credit Scoring Model and Statistical Learning after Variable Screening. *Mathematical Statistics and Management*, **39**, 368-380.