

German Credit Risk Data Analysis Report

by Hongliang Yu

Submission date: 08-Mar-2023 01:30AM (UTC+0000)

Submission ID: 198744430

File name: 40826_Hongliang_Yu_German_Credit_Risk_Data_Analysis_Report_482892_2136272160.pdf (1.27M)

Word count: 2827

Character count: 14918

German Credit Risk Data Analysis Report

Abstract

With the rapid development of Internet and other technologies, the competition in the market environment of commercial banks is obviously intensifying year by year. In order to maintain their core competitiveness[1], the importance of asset management and risk management of banks is also gradually increasing. In this paper, we will establish the logistic regression model, decision tree model, random forest and naive Bayesian model of credit card risk factors, and establish the confusion matrix of each model to compare the interpretability and analysis of the prediction performance by using TPR and FPR of each model. In addition, considering that the model involves many independent variables, this paper plans to apply the AIC information criterion and the BIC information criterion to screen the variables, and integrate the interpretability and prediction accuracy of the model to select the model that can most accurately predict the customer's credit risk. Meanwhile, in order to better test the generalization prediction ability of the model, the data set is divided into training set and test set, and ensure the ratio of 0-1 in the dependent variable in the training set and the ratio of 0-1 in the dependent variable in the sample data to improve the training accuracy. Therefore, 70% of the samples from the credit risk and no credit risk data set are randomly selected as the training set, and the rest samples as the test set, train the model with the training set and evaluate the prediction effect on the test set. This paper intends to dig out the significant factors affecting the credit card risk through the model establishment, and accurately predict the customer credit risk, and further optimize the personal credit risk assessment system on the basis of the current bank risk assessment. This is obviously of certain practical significance for optimizing the bank credit rating system, reducing the economic loss of banks and controlling the high credit risk.

Introduction

In recent years, with the continuous development and expansion of banks, various kinds of exposed problems have emerged in an endless stream, among which we have to mention the bank losses caused by personal credit risk. At present, the risk management of Chinese banks mainly focuses on post-loan service, overdue loan collection and other aspects[2], and the response often lags behind the occurrence of problems. However, as the main personal business object of a bank, the default of credit loan customers will undoubtedly bring large economic losses to the bank. Therefore, assessing personal credit risk in advance plays a key role in avoiding losses, which also explains the importance of optimizing the credit risk assessment management system[3].

With the maturity of the credit card market, the credit card business gradually develops in accordance with the requirements of the market, forming an industry link. For the banks, the credit card business is an important part of the profits of the commercial banks[4], which also means that it brings high risks to the banks. In recent years, the risk of credit card business has been rising, which has attracted the attention of many banks. Banks have also issued risk management measures for credit card business to deal with the loss caused by credit risk. Therefore, it is very necessary for us to assess the credit card risk[5].

This paper proposed to establish logical regression, decision tree dig into the significant factors affecting credit card risk, and accurately predict customer credit risk, on the basis of the bank risk assessment further optimize the personal credit risk assessment system[6], which obviously to optimize the bank credit rating system, reduce bank economic losses and control the high credit risk has certain realistic significance.

Data source and description

The dataset selected for this case is the German Credit Data (German credit dataset), which contains data from past credit applicants. German credit data has 1000 data, each with 20 features. These features included 13 qualitative variables and seven quantitative variables. The dependent variable was credibility, and the applicants were classified as good and bad[7]. The model of this data can be used to determine whether a new applicant presents good or bad credit risk. Description of the specific variables are shown in Table Table 2-1.

Table 2-1: Description of data variables

symbol	variable
status.of.existing.checking.account	Status of the current deposit account
duration.in.month	life of loan
credit.history	Credit history
purpose	Lending purpose
credit.amount	credit amount
savings.account.and.bonds	Savings and bonds
present.employment.since	Time of employment at the current employer
installment.rate.percentage	Credit installments as a percentage of disposable income
personal.status.and.sex	Comprehensive information on gender and marital status
other.debtors.or.guarantors	Whether there are other debtors or credit guarantors
present.residence.since	Time of residence in the place of residence
property	The most valuable property ever available

age.in.years	age
other.installment.plans	The installment plan for the supplier outside the bank
housing	Housing type
number.of.existing.credits	The amount of credit owned in the bank
job	The debtor work
number.of.people.provide.maintenance	Number of persons providing for support
telephone	Whether there is a registered phone number
foreign.worker	Whether it is a foreign worker

Descriptive analysis

The variables used in this case are as follows:



Figure 1

The rose chart of some variables is as follows. The relevant variables are the status of the debtor's current deposit account, credit history, loan purpose, and the most valuable property respectively.

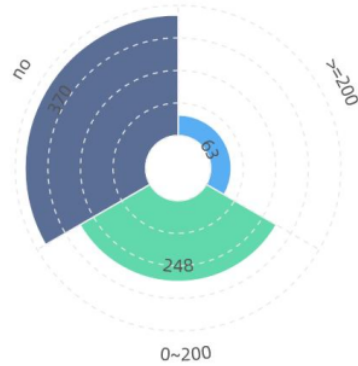


Figure 2

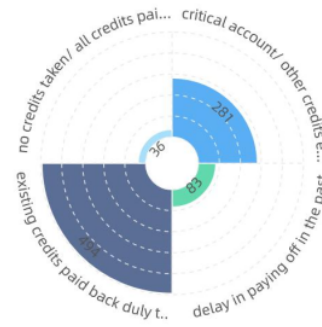


Figure 3

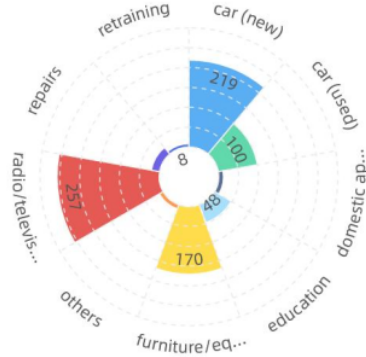


Figure 4

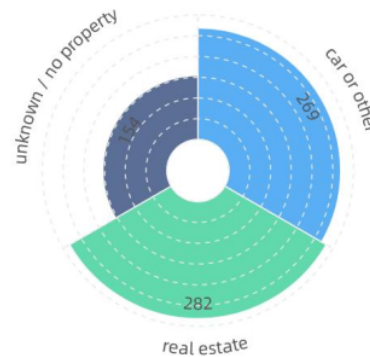


Figure 5

It can be seen intuitively from the figure above that the majority of debtors are in the "no" status of the current deposit account, and there are only 63 debtors with $DW \geq 200$. In addition, according to Figure 2, the majority of people still return their credit cards, while only 36 people have no history of borrowing[8]. Figure 3 shows that the majority of people who borrow from buying new cars, TVs, radios and furniture, while the number of people who borrow from other destinations is very small. As can be seen from Figure 4, most of the most valuable property of debtors is property and vehicles, but about 150 people still have no assets or unknown assets.

Credit card risk assessment model

In order to dig deeper into the significant factors affecting credit card risk and achieve the goal of accurate predict customer credit risk, this case will establish a logical regression model of credit card risk factors, decision tree model, random forest and naive Bayes model, and establish the confusion of each model of matrix, the interpretation of each model and prediction performance comparative analysis. Considering that the model involves many independent variables, this case tries to establish the AIC criteria and BIC criteria for model selection[9], and integrates the complexity and prediction accuracy of the model to select the model that best accurately predicts the customer's credit risk. Also, at the same time, in order to better test the generalization prediction ability of the model, this case divides the data set into training set and test set, and ensure the same 0-1 ratio in the training set and the test variable in the sample data, improve the training accuracy[10], so we randomly selected 70% of the samples from the credit risk data set and no credit risk data set as the training set, the rest samples as the test set.

Data preprocessing

Since the acquired German credit card risk dataset contains 7 continuous variables and 13 categorical variables, we need to set the factor level of each categorical variable to the corresponding data label to facilitate subsequent modeling analysis.

Model building

Establish the logistic regression model

First, this case established a 0-1 logistic model for all variables, and conducted the likelihood ratio test for the overall model, and the significance level of the hypothesis test was taken as 5%. The results are shown in Table 5-1:

Table 5-1: Results of the logistic regression, Fig

variable	regression coefficient	conspicuousness	Interpretation of variables
The intercept term	-1.674		
Status of the current deposit account-1	-4.783e ⁻¹		Benchmark group: <0 DM

Status of the current deposit account-2	-1.372	notable	1: 0<=...<DM 2: ...>=DM
Status of the current deposit account-3	-1.869	notable	3: No checking account
life of loan	3.269e ⁻²	notable	
Credit History-1	-5.842e ⁻¹		
Credit History-2	-1.216		
Credit History-3	-1.155	notable	Base group: all loans repaid
Credit history-4	-1.959	notable	
Lending purpose-1	-1.589	notable	
Lending purpose-2	-8.625e ⁻¹	notable	Benchmark group: for the new car
Lending purpose-3	-1.210	notable	1: Used cars
Lending purpose-4	-7.096e ⁻¹		3: TV
Lending purpose-5	-2.118e ⁻¹		4: Business
Lending purpose-6	-9.394e ⁻¹		5: Education
Lending purpose-7	-6.204e ⁻¹		6: Repair
Lending purpose-8	-2.069		7: Home appliances
Lending purpose-9	-3.970	notable	8: Re-training
Savings and Bonds-1	-1.391e ⁻¹		9: Other
Savings and Bond-2	-4.591e ⁻¹		Base group: less than 100DM
Savings and Bonds-3	-1.173		1: 100~500 DM
Savings and Bond-4	-1.040	notable	2: 500~1000 DM
			3: greater than 1000 DM
credit amount	1.116e ⁻⁴		
Time of employment at the current Employer-1	-6.940e ⁻¹		
Time of employment at the current Employer-2	-6.326e ⁻¹		Benchmark group: Not employed
Time of employment at the current Employer-3	-1.258	notable	1: Less than 1 year
Time of employment at the current Employer-4	-8.704e ⁻¹		2: 1~4 Years
			3: 4~7 Years
By installments as a percentage of disposable income	3.880e ⁻¹	notable	

Comprehensive information on gender and marital status-1	-4.008e ⁻¹	notable	Benchmark group: divorced or separated men
Comprehensive information on gender and marital status-2	-1.008		1: Divorce or separated women 2: Single men 3: Married men
Comprehensive information on gender and marital status-3	-4.629e ⁻¹		
Other debtors or guarantor-1	1.415e ⁻¹		
Other debtors or guarantor-2	-1.006e ⁻¹		Benchmark group: no
Time of the local residency	-3.927e ⁻²		
Most valuable property of the-1	2.165e ⁻¹		
Most valuable property-2	2.534e ⁻¹		Benchmark group: Real estate
age	-3.281e ⁻³		
Other installment payment plans-1	3.856e ⁻¹		
Other installment payment plans-2	-5.407e ⁻¹	notable	Benchmark group: Bank
Type of housing for dwelling in-1	-5.597e ⁻¹	notable	Benchmark group: rent a house 1: Have a house 2: Free house
Type of housing for dwelling in-2	1.446e ⁻¹		
Amount of credit at this bank	2.820e ⁻¹		
Work-1	5.803e ⁻¹		Benchmark group: non-local residents not employed or less competent

Work-2	5.411e ⁻¹	
Work-3	3.544e ⁻¹	
Number of persons providing for support	1.030e ⁻¹	
Whether there is a registered phone number	-3.501e ⁻¹	
Is it a foreigner-1	-1.358	
Model-likelihood ratio test		notable

It can be seen from the above table that the overall model has passed the significance test, and 15 factors significantly affect the credit risk under the significance level of 0.05, such as the term of the loan and the amount of credit in the bank. However, the full model also contains many non-significant factors, so consider using the AIC criteria and the BIC criteria.

Using AIC and BIC selection model drew different conclusions: for logistic regression, AIC model retained 12 variables, respectively: whether there is a registered phone, other debtors and guarantor, credit guarantee amount, whether other installment plan, is a foreign worker, the housing type, loan term, the debtor needs the purpose of credit, installment as a percentage of disposable income, credit history, status of current deposit account. However, BIC only selected two variables, namely: loan term and the status of current deposit account.

The full model, AIC model and BIC models all differ in coefficient estimates. Therefore, further model selection is needed.

Establish a decision tree model

In order to improve the interpretability and robustness of the prediction models, we consider building a decision tree model to predict the customer credit risk. The results are shown in Figure Figure 5-1:

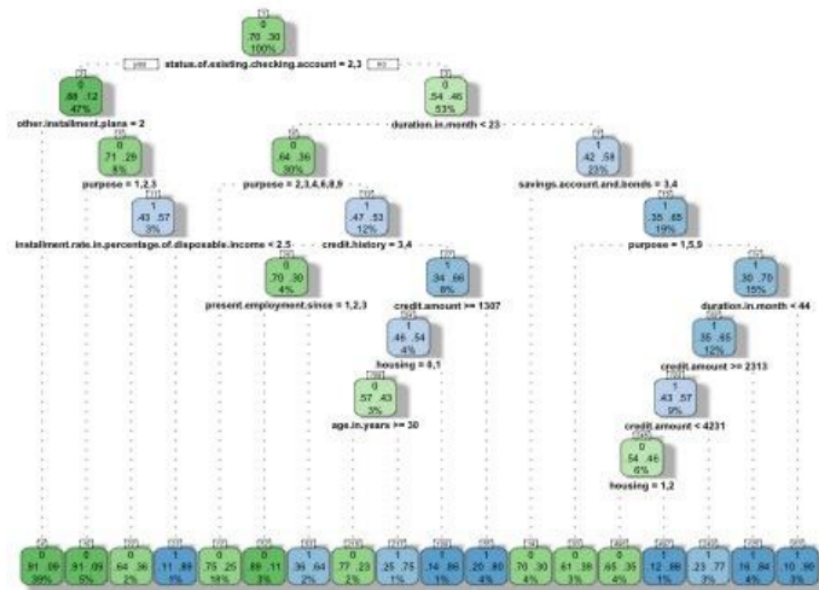


Figure 5-1: Full decision tree model

Can be seen from the above each decision tree node corresponding to different decision conditions, choose after produce different results and again face a new natural state, continue to produce a series of decision-making process, so the decision tree behavior and human thinking behavior, is a kind of effective tool can help decision makers sequence decision analysis, interpretation is strong. However, due to the overabundance of the whole decision tree, we consider pruning the decision tree. First, a 10-fold cross-validation of the full decision tree selects the size of the tree corresponding to the minimum deviation, i. e., the number of leaf nodes. Finally, we built a decision tree model with size 4, and the results are shown in Figure 5-2:

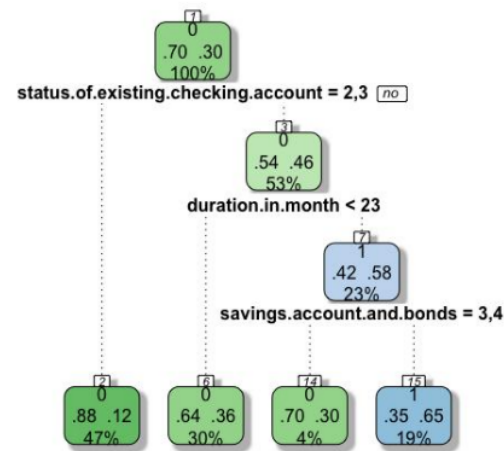


Figure 5-2: Decision tree model after pruning

It can be seen that the decision tree becomes more concise, but because the decision conditions are different, the prediction effect will be different, so we will make further model selection in the future.

Conclusion

In this paper, German credit risk data is selected to build a model to accurately predict customer credit risk, multiple models are selected, and their performance in the test set is used for model selection and evaluation.

First of all, a logistic full regression model was established in this case. Since most insignificant variables were observed in all independent variables, variable selection was needed, significant variables were selected by AIC or BIC, and a stepwise logistic regression model was established. According to the prediction results, the logical AIC model predicted the best, and the logical BIC predicted the worst. This can also be reasonably explained, because there are only two variables after screening by using the BIC criteria, the final prediction effect will be relatively poor.

In order to improve the decipher ability and robustness of the model, this case chose to build the decision tree model. Similarly, due to too many branches, cross-validation is needed to prune the decision tree. According to the prediction results, both the full decision tree and the pruned decision tree have poor predictions on the test set.

For the results of logistic regression, we only observe significant variables and their coefficients. The relatively significant variables include loan maturity and the status of demand deposit accounts. Since the decision tree in this paper is based on the Gini coefficient, the smaller the Gini coefficient is, the greater the contribution of the variable to the classification results. According to the results of the decision tree, we can get more important variables including demand deposit account, borrowing time, savings and bonds. Finally, the conclusion is drawn:

- For the status of demand deposit accounts, people without checking accounts are more prone to credit risk.
- The longer the loan term, the more likely the credit risk tendency.
- People who have not paid off their loans and are in debt for a longer period of time are more likely to be prone to credit risk.

Reflection

In this paper, significant factors affecting credit card risk are explored deeply by establishing mathematical models such as logistic regression and decision tree, and customer credit risk is accurately predicted. However, the accuracy of the model is not particularly high, so it needs to be further improved. In addition, the model established is relatively basic, which has certain practical significance for optimizing bank credit rating system, reducing bank economic losses and controlling high credit risk, but it is of little significance.

Video address

Youtube address: <https://youtu.be/hmSa9RQqZB8>

References:

- [1] Liao Xinting, Xie Lei. *Empirical study on credit card overdue risk assessment based on the comparison between Probit and Logistics model* [J]. Market Forum, 2020 (06): 73-77.
- [2] Sun Ke. *Research on the application of Credit Evaluation in Credit Card Risk Management of Commercial Banks* [D]. Shandong University, 2018.
- [3] Yang Zhong. *Research on credit card credit risk assessment of Commercial Banks in China* [D]. Hunan Agricultural University, 2016.
- [4] Xu Xinyi. *The Application of Data Mining in the Credit Card Fraud Identification Problem* [D]. Shandong University of Science and Technology, 2019.DOI:10.27275/d.cnki.gsdku.2019.000670.
- [5] Li Yangzhi. *Research on credit card risk management based on the decision tree* [J]. Research on Financial Management, 2018 (02): 1-21.
- [6] Wang Congmin. *Research on risk management and consumption behavior based on data mining* [D]. Capital University of Economics and Business, 2013.
- [7] Yan Siyu. *Study of credit card overdue behavior based on random forest classification* [D]. Yunnan University of Finance and Economics, 2018.
- [8] Zhu Zhentao, Sun Min, Shen Jianhong. *Impact factors and behavior prediction of credit card overdue behavior* [J]. Journal of Nanjing Institute of Technology (Social Science Edition), 2019,19(03):46-53.DOI:10.13960/j.issn.2096-238X.2019.03.009.
- [9] Jiang Shiqi. *Research on the credit card anti-fraud model* [D]. Chongqing University, 2019.DOI:10.27670/d.cnki.gcqdu.2019.000094.
- [10] Zhu Jian. *Comparative study on credit card default identification of four data mining algorithms* [D]. Dalian University of Technology, 2017.

German Credit Risk Data Analysis Report

GRADEMARK REPORT

FINAL GRADE

GENERAL COMMENTS

Instructor

53/100

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

GRADING FORM: DATA MINING/ VISUALISATION

HONGLIANG YU

53

KNOWLEDGE AND UNDERSTANDING

Dataset Suitability (What features of the dataset lend themselves to data mining? Has it been mined previously? Why does the student feel that they have the contextual background to work with this data?)



German credit dataset used for analysis. Twenty distinct variables are proposed for future mining. No rationale for choice of dataset and or student prior background is given in assignment.

21

COGNITIVE AND INTELLECTUAL

Objectives (What does the student plan to investigate/expect to find? Are these SMART?)



Plans for future investigation is not given in submission but rather actual analyses are offered. In future please ensure that the assessment requirements are adhered to.

4

PRACTICAL APPLICATION SKILLS

Dataset Description (Where did the data come from? What does the data describe? Is the sample large enough? What data formats and types are involved?)



Dataset location and description provided are sufficient. Relevance of sample size for future analysis and data types are not addressed fully.

23

TRANSFERABLE SKILLS FOR LIFE

Presentation/Slides (Clarity, readability, flow, accuracy, completeness, performance)



Although format and content appears to meet the standard report writing conventions, this submission is inaccurate owing to its ability to adhere to the stipulated assessment requirements in full.

5