Understanding of Personalized Customer Credit Risk Based on Selected Attributes

Runqi Jiang

School of Economics and Management, Communication University of China, Beijing, 100020, China

jiangrun7@163.com

Abstract. In today's dynamic business landscape, credit risk assessment has become an essential aspect of financial institutions' success, and a comprehensive understanding of the most important variables that affect creditworthiness is critical. This essay focuses on the analysis of customer credit risk based on some updated techniques, specifically using crosstabulation, factor analysis, and logistic regression to investigate the relationship between credit default and personal characteristics. The study uses a sample of credit card customers and considers several relevant variables, including age, housing status and employment status, to determine their impact on credit default risk. The results show that the aforementioned models improve the accuracy and efficiency of credit risk models by identifying patterns and relationships among variables and predicting credit defaults with better accuracy. Additionally, certain personal characteristics, such as age and employment time, can have a significant impact on credit default risk. The study's findings hold implications for financial institutions, as they can leverage machine learning technology to build more accurate credit scoring models that enable better decision-making. Overall, this analysis provides a valuable perspective on the importance of statistical techniques in improving credit risk assessments, revealing the relationship between customer attributes and credit default risk.

Keywords: credit, risk, management.

1 Introduction

In today's rapidly changing business environment, credit risk assessment plays a critical role in the success of financial institutions. With the increase in the number of customers and transactions, it has become increasingly challenging for financial institutions to determine the creditworthiness of their customers, leading to a surge in the number of credit defaults and resulting in significant losses for these institutions. In recent years, the development of statistical techniques has provided new opportunities for financial institutions to improve their credit risk assessment models, enabling them to make more accurate predictions and decisions. In this essay, we will analyze the

K. Elbagory et al. (eds.), *Proceedings of the 9th International Conference on Financial Innovation and Economic Development (ICFIED 2024)*, Advances in Economics, Business and Management Research 281,
https://doi.org/10.2991/978-94-6463-408-2_28

customer credit risk model based on crosstabulation and factor analysis and evaluate its effectiveness in predicting credit risk.

Credit risk has always been a topic and research direction that attracts much attention. The predecessors have also done a lot of research on credit risk, which provides us with a powerful reference. A Comparative Analysis of Current Credit Risk Models on Journal of Banking & Finance reviewed the proposed industry sponsored Credit Value-at-Risk methodologies before 2000 and systematically analyzed the approach of JP Morgan and McKinsey which focused on the default [1]. Credit Risk Modeling and Introduction to Credit Risk Modeling both focus on the introduction and overview of some credit risk model and make certain extension [2,3]. Compared with the overall research on the credit risk of banks and enterprises, the research on customer credit risk is a more segmented and targeted research direction. There are also many scholars who have made a detailed analysis of customer credit risk. Regarding good performance and the ability of classification, generalization and learning patterns, some studies considered Multi-layer Perceptron Neural Network model trained using various Back-Propagation (BP) algorithms in designing an evaluation model by using SPSS and MATLAB software [4,5]. Some studies considered more in-depth machine learning algorithms to give the best predictive performance using hybid models and demonstrate probability estimation in Random Jungle [6-8]. In this study, I analyzed the customer credit risk model based on some selected attributes and used them to predict credit risk. I chose the information containing various customers' personal information and their credit rating in the Kaggle database, and analyzed the relationship between personal information and credit rating by using the SPSS crosstabs and hypothesis testing. Then, this paper tries to use factor analysis to explore whether there are common factors in multiple categories that affect credit rating. Finally, I conducted a regression analysis using some personal information in the data set, such as age, personal marital status, job status, etc., to predict credit risk.

2 Influencing Factors of Credit Risk

Initially, this study try to discover the relationship with some individual factors such as age, job and whether the person has married or not. Figuring out these specific bond is absolutely helpful for banks to judge borrower’s credit risk and decide whether money should be lent to these people using a simple survey including these following factors. The conclusion helps us to predict in other deeper analysis as well. Some common terms are chosen to analyse the question. In this part, I would find out the relationship between four possible factors of an individual (personal status, employment, housing status and age) and one’s credit risk level respectively using the tools in SPSS Statistics. Chi-square test for the independence of the sample is required before further analysis [9]. Due to diffenrent researching methods of discontinuous factors and continuous factors, using cross table to process discontinuous factor and using simple descriptive analysis to deal with others.

2.1 Personal Status

It is easy to suppose that one's personal status(marital status)would have an influence on the credit risk class because of the diverse economic capability to a large extent. But the hypothesis need to be confirmed. The personal status contains two variables, married and single and the credit risk class contains two variables, good and bad.

Using the Chi-square to confirm whether there is a directly connection between personal status and credit risk and the result is that the figure of Pearson Chi-square is 0.011 and it is lower than 0.05 so it is obvious that there is a significant association between the two variable. Then looking at the cross table to find the specific relationship in the table 1.

According to Table 1, the single people has bigger proportion within class good (57.4%) and look at it in another dimension, people whose credit class is good take up a bigger proportion within single people (73.4%) as well. Consequently,it is notable that people whose personal status is single are more likely to have better credit class and in other word, lower credit risk.

Table 1. The cross table of personal status and credit risk class

			married	single	Total
class	good	Count	298	402	700
		% within class	42.6%	57.4%	100.0%
		% within personal status	65.9%	73.4%	70.0%
	bad	Count	154	146	300
		% within class	51.3%	48.7%	100.0%
		% within personal status	34.1%	26.6%	30.0%
Total		Count	452	548	1000
		% within class	45.2%	54.8%	100.0%
		% within personal status	100.0%	100.0%	100.0%

The prossible reason of that phenomenon seems to be obvious.In our society, married people always have more burden than single people. For instance, most married people have children so they need to raise them and pay for their daily life and education. Also, single people do not have the expenditure of dating and other activties of couples. If borrowers have more expense, it is more possible for them to delay their debt. It is why banks define the two credit class.

2.2 Employment Time

Thinking of the employment years of people, it is easy to understand that maybe people who have worked for long years would have more deposit so they are more resistant to risk. To verify this guess, using the same solution as the previous research is still a suitable choice. The employment years are divided into five variables ($x>7$, $4<x<7$, $1<x<4$, $0<x<1$, $x=0$). The Chi-square is 0.026 which is significantly lower than 0.05 so we could further analyse the cross table.

According to Table 2, 70% and more people whose emlpoyment time longer than 4 years have good credit level. And bad credit class people account for 34.7%, 40.7%,

37.1% within employment time 1<x<4, 0<x<1, x=0 respectively. Clearly, the guess is correct.

Table 2. The cross table of employment time and credit risk class

x>7		4<x<7	1<x<4	0<x<1	x=0	Total
class good Count	189	135	235	102	39	700
% within class	27.0%	19.3%	33.6%	14.6%	5.6%	100.0%
% within employment	74.7%	77.6%	69.3%	59.3%	62.9%	70.0%
bad Count	64	39	104	70	23	300
% within class	21.3%	13.0%	34.7%	23.3%	7.7%	100.0%
% within employment	25.3%	22.4%	30.7%	40.7%	37.1%	30.0%
Total Count	253	174	339	172	62	1000
% within class	25.3%	17.4%	33.9%	17.2%	6.2%	100.0%
% within employment	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

To figure out the possible reason behind this phenomenon, the potential connention between people's employment time and their deposit need to be focused. It is commonly believed that the longer people work, the more money they will save. So usng the same way, the cross table below shows the specific relationship.

According to the Table 3, as we guessed, the people who have longer employment time have higher saving status. So it well explains why people with longer working time have less credit risk——the reason of that is they have more deposit so their financial status would be more stable, they always have capability to repay their loan using their existing money.

Table 3. The cross table of employment time and saving status

	y>1000		500<y<1000	100<y<500	0<y<100	y=0	Total
employment	x>7	Count	14	20	22	133	64	253
	4<x<7	Count	9	9	24	100	32	174
	1<x<4	Count	18	26	33	210	52	339
	0<x<1	Count	7	5	17	120	23	172
	x=0	Count	0	3	7	40	12	62
Total		Count	48	63	103	603	183	1000

2.3 Housing Status

Housing Status is definitely a vital factor which could reflect one's credit risk. It is commonly believed that people who own a house have stronger economic ability to repay their debt. So using the cross table, the result could be clear.There are three different variables about housing: own, rent and free residence (means no house). First, the result of Chi-square is 0.00 and it significantly smaller than 0.05 so the relationship is notable. Looking at the Table 4, the people who own houses have bigger proportion (75.3%) within good class in credit risk and compared with the total percentage 30%,

people who rent house or do not have stable living place are more likely to get bad credit class.

The reason of that is very obvious.It is easily accepted that the people who have a house have better economic capability and they have stronger ability to bear risks when they face some financial crisis. One possible solution is they could easily mortgage their estate to meet the challenge. In conclusion, people who have a steady residence get higher credit standard.

Table 4. The cross table of housing status and credit risk class

own			rent	free	Total
class	good	Count	527	109	64	700
		% within class	75.3%	15.6%	9.1%	100.0%
		% within housing	73.9%	60.9%	59.3%	70.0%
	bad	Count	186	70	44	300
		% within class	62.0%	23.3%	14.7%	100.0%
		% within housing	26.1%	39.1%	40.7%	30.0%
Total		Count	713	179	108	1000
		% within class	71.3%	17.9%	10.8%	100.0%
		% within housing	100.0%	100.0%	100.0%	100.0%

2.4 Age

Figuring out the relationship between age and credit risk seems to be challenging as well. As a continuous variable, analysing age should take a totally different measure. First, do the tests of normality in Table 5 to confirm what test they need to do next. Through the table of the tests of normality, it is easy to discover that the variable of age does not allow the normal distribution so it need to do the nonparametric tests (see Table 6).

Table 5. Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
class		Statistic	df	Sig.	Statistic	df	Sig.
age	good	.112	700	.000	.926	700	.000
	bad	.139	300	.000	.890	300	.000
a.Lilliefors Significance Correction							

Table 6. Hypothesis Test Summary

Null Hypothesis	Test	Sig.	Decision
The distribution of age is the same across categories of class.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
Asymptotic significances are displayed. The significance level is .050.			

Looking at the table of the hypothesis test,the result rejects the null hypothesis which means the distribution of age is diffenrent across categories of class.The meaning of this is the class will be different with the change of the variable.But what is the specific changing direction?

According to Table 7, within the age, the mean figure for good is 36.224 which is obviously bigger than that for bad (33.963). It can be concluded that older people tend to have better credit level. Age is associated with employment time, deposit account etc. All of these could be good explanation.

Table 7. Descriptive of Age

	class			Statistics	Std.Error
age	good	Mean		36.224	.4302
		95% Confidence Interval for Mean	Lower Bound	35.380	
			Upper Bound	37.069	
	bad	Mean		33.963	.6479
		95% Confidence Interval for Mean	Lower Bound	32.688	
			Upper Bound	35.238	

3 Extract Common Factors with Factor Analysis

As we know, there are plenty of factors that would reflect and influence one's credit risk and analysing every single variable could be a complicated work. Factor analysis offers a technique of analysis for studying behavioral phenomena of great complexity and diversity and mode findings into scientific theories [10]. Relying on factor analysis to extract relevant common factors and name them as new factors, we can further simplify our understanding of the relationship between personal information and credit risk and then make predictions.

So this study chooses ten major variables joining in the factor analysis (employment, personal status, property, housing, age, job, existing credit, credit duration, credit amount, credit history) and try to discover whether there are some simple variables behind these exsting variables that could explains most of the data.

After computing the KMO>0.5, this data is confirmed to be quite suitable to do this factor analysis. There are four new terms which could explain 63.35% of the whole data (see Table 8). Then rotate the component matrix and combine some specific variables to new ones (see Table 9). The new variables are renamed as following:

Name the combination of credit amount,duration and job with "credit demand". It is because especially the credit amount and the credit duration describe people's expected money amount and time they hold it. Credit demand is the most important factor which could be interpreted by risk premium. The higher the credit demand is, the higher risk premium is. The demand will definitely influence one's credit decision.

Name the combination of housing and property magnitude with "asset status". The logic behind this is these variables showed the asset especially fixed assets one have.

Name the combination of employment, age and personal status with “demographic description”. The reason of that is all of these three variables well decribe an individual’s personal information.

Name the combination of credit history and exsting credits with “previous credit”. This is because the two variables show one’s credit account in the past to a large extent. It weighs the least important factor since banks seldom measure a person based on their historical performance because there would be a lot of error.

Table 8. Total Variance Explained

Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative%	Total	% of Variance	Cumulative %
2.299	22.994	22.994	2.299	22.994	22.994	1.867	18.673	18.673
1.630	16.302	39.295	1.630	16.302	39.295	1.513	15.134	33.806
1.354	13.542	52.837	1.354	13.542	52.837	1.501	15.010	48.816
1.051	10.512	63.350	1.051	10.512	63.350	1.453	14.533	63.350
.844	8.439	71.789						
.777	7.775	79.563						
.704	7.043	86.607						
.531	5.314	91.921						
.469	4.686	96.607						
.339	3.393	100.000						

Extraction Method: Principal Component Analysis.

Table 9. Rotated Component Matrix[1]

Component				
1		2	3	4
credit_amount	.865			
duration	.862			
job	-.468			
housing		.861		
property		.795		
employment			-.752	
age			.680	
personal status			.659	
credit history				.849
existing credits				.845

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.

4 Credit Risk Prediction

After analysing the relationship and simplifying the factors, finally, this paper focuses on predicting the level of credit risk through some personal information that has appeared above. Binary logistic regression aid in understanding and testing complex relationship among variables and in forming predictive equations [11]. This part we use logistic regression to predict the possible credit risk class result. When the cut value is 0.3, the predicted conclusion would explain almost 67% of the data. Then continue to do the exact prediction in Table 10.

Table 10. The Prediction of The Selected Attributes

B		5.E.	Wald	df	Sig	Exp(B)
Step 1 duration	.039	.008	22.571	1	.000	1.040
credit history			37.955	4	.000	
credit history(1)	.276	.486	.321	1	.571	1.317
credit_history(2)	-.751	.383	3.845	1	.050	.472
credit_history(3)	-1.661	.392	17.944	1	.000	.190
credit_history(4)	-1.040	.426	5.950	1	.015	.353
employment			13.145	4	.011	
employment(1)	-.584	.375	2.427	1	.119	.558
employment(2)	-.988	.401	6.072	1	.014	.372
employment(3)	-,412	.371	1.238	1	.266	.662
employment(4)	-.111	.391	.081	1	.776	.895
personal_status(1)	.387	.166	5.412	1	.020	1.473
age	-.010	.008	1.641	1	.200	.990
housing			4.572	2	.102	
housing(1)	.191	.382	.250	1	.617	1.211
housing(2)	.590	.408	2.095	1	.148	1.805
credit_amount	.000	.000	.002	1	.968	1.000
existing_credits	.313	.170	3.413	1	.065	1.368
job	.119	.144	.691	1	.406	1.127
Constant	-.206	.727	.081	1	.777	.814

a. Variable(s) entered on step 1: duration, credit_history, employment, personal_status, age, housing, credit_amount, existing_credits, job.

According to Table 10, duration, credit history (3)(4), employment (2), personal status (1) can be used to predict unknown credit risk class because their Sig figure (Significance) lower than 0.05. Focus on the B figure and identify whether they have a positive or negative relationship with credit risk. Obviously, if the B figure is negative, the variable have a positive influence on credit risk (we name the class good as 0 and class bad as 1 in the spss). Take duration for an example, due to the 0.039 B figure, the longer credit duration is, the worse one's credit level is. Same as the example, for personal status, because we name married as 0 and single as 1 in spss, the B is 0.387 so married personal status has a negative influence on credit class. From similar arguments, this study also found one's property status could also influence the credit default risk because the figures for Sig are always lower than 0.05. And due to the negative B, the higher the property is, the lower the credit default risk is.

5 Conclusion

In conclusion, this essay has provided an in-depth analysis of customer credit risk based on statistical techniques. Using crosstabulation, factor analysis, and logistic regression models, the study examined the relationship between credit default and personal characteristics, including age, personal status and employment status. The logistic regression revealed that machine learning algorithms can provide improved accuracy and efficiency concerning predicting credit default risk. Additionally, certain personal factors do play a significant role in credit risk assessment. As such, financial institutions can use the insights gained from machine learning models to build more effective credit risk assessment models, helping them make better-informed credit decisions and safeguard against potential losses.

Overall, the results of this analysis underscore the importance of machine learning in assessing credit risk accurately and efficiently, highlighting the critical role it plays in enhancing decision-making in the financial sector. This study's contribution to the literature on credit risk assessment demonstrates that machine learning algorithms can offer an effective strategy for accurately predicting credit defaults, providing substantial value to financial institutions. In conclusion, the findings of this study suggest that financial institutions should incorporate statistical techniques as well as machine learning algorithms into credit risk assessment models and explore the use of other variables to improve their assessments further.

Reference

1. Crouhy, M., Galai, D., Mark, R. (2000) A comparative analysis of current credit risk models. Journal of Banking & Finance, 24: 59-117.
2. Lando, D. (2009) Credit risk modeling. Princeton University Press, Princeton.
3. Bluhm, C., Overbeck, L., Wagner, C. (2016) Introduction to credit risk modeling. Crc Press, Boca Raton.
4. Mohammadi, N., Zangeneh, M. (2016) Customer credit risk assessment using artificial neural networks. IJ Information Technology and Computer Science, 8: 58-66.
5. Nazari, M., Alidadi, M. (2013) Measuring credit risk of bank customers using artificial neural network. Journal of Management Research, 5: 17.
6. Machado, M. R., Karray, S. (2022) Assessing credit risk of commercial customers using hybrid machine learning algorithms. Expert Systems with Applications, 200: 116889.
7. Kruppa, J., Schwarz, A., Arminger, G., Ziegler, A. (2013) Consumer credit risk: Individual probability estimates using machine learning. Expert systems with applications, 40: 5125-5131.
8. Khandani, A. E., Kim, A. J., Lo, A. W. (2010) Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34: 2767-2787.
9. McHugh, M. L. (2013) The chi-square test of independence. Biochemia medica, 23: 143-149.
10. Rummel, R. J. (1988) Applied factor analysis. Northwestern University Press, Evanston.
11. King, J. E. (2008) Binary logistic regression. In: Jason, O.B. (Eds.), Best Practices in Quantitative Methods. Sage Publications, Inc., Thousand Oaks. pp. 358-384.