Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/32/20230863

Data analysis with different variables and credit risk assessment

Ruixin Jin[1,3],Huanyu Zhou[2]

[1]Hangzhou Tianyuan College, Hangzhou, 311121, China
[2]Guangzhou NO.7 Middle School, Guangzhou, 510080, China

[3]alan2873464366@outlook.com

Abstract. Nowadays, credit payment is a very common way to pay, such as credit cards, loans, many people can use their credit as a guarantee to borrow money from the bank, however some people will default. So we have to predict whether the borrower will pay on time, it is known as credit risk assessment. In this paper, we analyze a data set on credit risk to predict whether individuals will be late on their payments, helping financial firms improve their earnings and reduce their losses. We not only made predictions on the data, but also analyzed the relationship between the variables that affect the overdue probability to find some specific associations. Specifically, we performed ANOVA analysis and found that married people borrowed significantly more than other groups, and the delinquency rate of people with higher education was lower, and the delinquency rate of married people was higher than that of unmarried people. In addition, we conducted a binary logistic regression and found that gender had no significant impact on the prediction results, but an individual's amount of bill statement, amount of previous payment, past repayment situation and Amount of the given credit had an impact on the prediction results. Other variables, such as marital status and education, can also impact the predicted results. Our research puts forward more factors affecting credit risk and also different angles that can be used to analyzes individual credit risk. This has a guiding role for financial firms like banks and other companies in the financial industry, providing more ways to help them analyze the credit risk of borrowers.

Keywords: Credit risk assessment, Factor analysis, ANOVA analysis, Binary logistic regression.

1. Introduction

Nowadays, more and more people use credit cards, which are a payment method guaranteed by personal credit. However, banks and financial companies do not know whether individuals will pay their bills on time. Inevitably, banks must take risks in giving loans and credit cards to customers because these are economic drivers [1]. But when debts are not paid on time, the cash flow of financial companies will be affected, and it can also cause some harmful effects on the company. So credit risk assessment is vital for banks; they must ensure that borrowers are able to pay their installments before allocating a loan to them [2]. For example every month almost every adult in the US and the UK is scored several times to enable a lender to decide whether to mail information about new loan products, to evaluate whether a credit card company should increase one's credit limit, and so on [3]. On a daily basis credit/financial analysts have to investigate an enormous volume of financial and non-financial data of firms [4]. Credit

DOI: 10.54254/2755-2721/32/20230863

risk assessment can help financial companies effectively manage risks, make wise investment decisions, and avoid unnecessary losses.

So the predecessors have done many research in this area. Paweł Pławiak et al .applied a novel approach based on deep genetic cascade ensemble of different support vector machine (SVM) classifiers (called Deep Genetic Cascade Ensembles of Classifiers (DGCEC)) to the Statlog Australian data [5]. Charles Guan et al .combined ML classification models trained on limited data with a well established form of "human-in-the-loop" knowledge acquisition based on Ripple-Down Rules (RDR) to construct fair and compliant rules that could also improve overall performance [6]. David West investigates the credit scoring accuracy of five neural network models: multilayer perceptron, mixture-of-experts, radial basis function, learning vector quantization, and fuzzy adaptive resonance. Found that the multilayer perceptron may not be the most accurate neural network model [7]. Stjepan Oreski et al .designed a hybrid system with genetic algorithm and artificial neural networks (GA-NN) for finding an optimum feature subset at retail credit risk assessment that enhances the classification accuracy of neural network classifier [8]. Hussain Ali Bekhet et al .Radial basis function (RBF) and logistic regression model are used to test the validity of a credit scoring model. In terms of overall accuracy, logistic regression model is slightly better than radial basis function model. However, the radial Foundation function is better at identifying those customers who are likely to default [9].

We found a data set on credit risk assessment to study the relationship between default of payment and different factors. The variables in the dataset are Gender, Age (year), Marital status, Education, Amount of the given credit, History of past payment, Amount of bill statement, Amount of previous payment and default of payment. Gender is divided into male and female, marital status is divided into married, unmarried and other, education is divided into high school, university, graduate and other. Among all variables, these three variables are different (History of past payment, Amount of bill statement, Amount of previous payment), each of them is divided into six periods, representing different states of individuals in the data set in six months. The History of past payment represents if the individual has been overdue in the past or not and, if so, for how long.

In this paper, we used factor analysis, ANOVA and binary logistic regression to analyze the data.By using factor analysis, we can reduce the dimension and compress the six periods of data of History of past payment, Amount of bill statement and Amount of previous payment to a smaller number of data and simplify the data. In order to make the data easier to observation and research, and also to let the model has a better generalization ability. To investigate the relationship between the two different data points, we performed ANOVA analyses and used the resulting data to analyze which factors might be more influential on default of payment. In our research the method was also used to investigate the relationship between other factors. We also analyze the data and use binary logistic regression to predict if the person will default his payment.

2. Factor analysis

2.1. Factor analysis on History of payment

Factor analysis is formed on the History of past payment, and compressed the data through the analysis of component eigenvalues and lithotriptic maps. After analyzing the data, this work found that the set of data can be compressed to at least two components. As can be seen from the rotated component matrix in as shown in Table 1 and 2, component 1 can better summarize phases 3, 4, 5 and 6, while component 2 can better summarize phases 1, 2 and 3. Component 2 is better at summarizing recent conditions. (the period of history of past payment smaller it refers more recent). It can be seen from the common factor variance table that the extracted two components have better generalization ability to the original data, so such compression is reasonable.

DOI: 10.54254/2755-2721/32/20230863

Table 1. Rotated component matrix.

	Component 1	Component 2
History of past payment 1	0.210	0.889
History of past payment 2	0.438	0.797
History of past payment 3	0.624	0.625
History of past payment 4	0.809	0.428
History of past payment 5	0.890	0.311
History of past payment 6	0.877	0.248

Table 2. Communality table.

	Initial	Extraction
History of past payment 1	1.000	0.849
History of past payment 2	1.000	0.821
History of past payment 3	1.000	0.777
History of past payment 4	1.000	0.839
History of past payment 5	1.000	0.884
History of past payment 6	1.000	0.824

2.2. Factor analysis on Amount of bill statement

This work also carried out factor analysis on the Amount of bill statement, and compressed this set of data through the analysis of component eigenvalues and gravel map. After analyzing the data, this work found that the set of data can be compressed to at least two components. As can be seen from the rotated component matrix in as shown in Table 3 and 4, component 1 can better summarize the 4th, 5th and 6th phases, and component 2 can better summarize the 1st, 2nd and 3rd phases. Component 2 is better at summarizing recent conditions. (the time order of 6 phase of Amount of bill statement is the same as that of History of past payment). It can be seen from the common factor variance table that the extracted two components also have good generalization ability to the original data, so such compression is reasonable.

Table 3. Rotated component matrix.

	Component 1	Component 2
Amount of bill statement 1	0.452	0.870
Amount of bill statement 2	0.510	0.844
Amount of bill statement 3	0.611	0.747
Amount of bill statement 4	0.755	0.612
Amount of bill statement 5	0.845	0.507
Amount of bill statement 6	0.868	0.453

DOI: 10.54254/2755-2721/32/20230863

Table 4. Communality table.

	Initial	Extraction
Amount of bill statement 1	1.000	0.961
Amount of bill statement 2	1.000	0.972
Amount of bill statement 3	1.000	0.932
Amount of bill statement 4	1.000	0.945
Amount of bill statement 5	1.000	0.971
Amount of bill statement 6	1.000	0.958

2.3. Factor analysis on Amount of previous payment

When analyzing the Amount of previous payment, the results show that when the data was compressed into two components, the interpretation ability was not good. When combined with the accumulation, when the data was compressed into two components, the accumulation was less than 50%.This data needs to be accumulated to the fourth factor before the accumulation approaches 80%, so this work decided to compress this set of data to four components to make the compressed components more interpretable.

Table 5. Communality table.

Component 1	Total	Cummunality%
Amount of previous payment 1	1.958	32.629
Amount of previous payment 2	0.893	47.519
Amount of previous payment 3	0.852	61.719
Amount of previous payment 4	0.837	75.669
Amount of previous payment 5	0.753	88.225
Amount of previous payment 6	0.707	100.000

This is to the data after compression, from Table 7 communality table, we can conclude four ingredients, compressed on the original data has good generalization ability. As shown in table 5 and 6, the rotated component matrix shows that component 1 can better summarize phases 1, 2 and 3, component 2 can better summarize phase 4, component 3 can better summarize phase 5 and component 4 can better summarize phase 6. (Note: the chronological order of factors from 1 to 6 is the opposite of the first two groups.).

DOI: 10.54254/2755-2721/32/20230863

Table 6. Rotated component matrix.

	Component 1	Component 2	Component 3	Component 4
Amount of previous payment 1	0.731	0.086	-0.034	0.174
Amount of previous payment 2	0.763	-0.29	0.187	0.005
Amount of previous payment 3	0.569	0.376	0.049	0.023
Amount of previous payment 4	0.104	0.946	0.075	0.077
Amount of previous payment 5	0.112	0.082	0.977	0.076
Amount of previous payment 6	0.123	0.075	0.077	0.979

Table 7. Communality table.

	Initial	Extraction
Amount of previous payment 1	1.000	0.573
Amount of previous payment 2	1.000	0.618
Amount of previous payment 3	1.000	0.468
Amount of previous payment 4	1.000	0.917
Amount of previous payment 5	1.000	0.979
Amount of previous payment 6	1.000	0.986

3. ANOVA analysis

3.1. Gender and overdue

In this data set, in the item of whether the loan is overdue, 1 means overdue, and 0 means repayment on time, so in this line chart and other chart where the Y-axis is overdue, the higher the data point, the higher the probability of the group of people being overdue. As can be seen from the Figure 1, the probability of men defaulting on loans is slightly higher than that of women, but the difference is not large, so it can be said that gender has no significant impact on whether they will default.

DOI: 10.54254/2755-2721/32/20230863

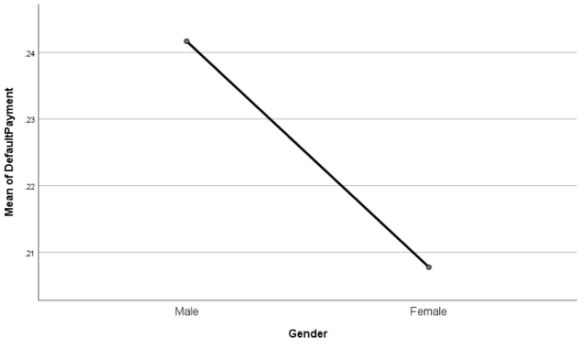

Figure 1. The relationship between Gender and overdue.

3.2. Marital status and overdue

It can be seen from the line chart in Figure 2 that married people are more likely to overdue than unmarried people. We confused why people that are married the possibility of default can be higher, there are two people repay the debts together, so we combined the line chart in Figure 3 (the relationship between marital status and loan amount) for analysis.

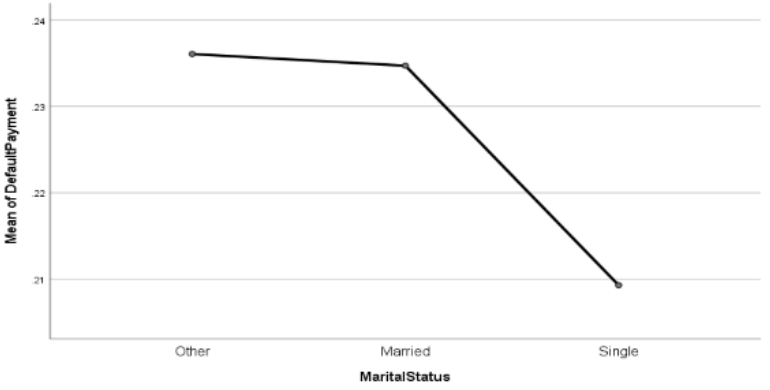

Figure 2. The relationship between marital status and overdue.

From Figure 3, this work shows that married people have the largest amount of loans, and the loan amount of unmarried people is obviously smaller than that of married people. Therefore, married people may be have a bigger possibility to overdue, and it can also be regarded as the group with more loans, the greater the probability of loan overdue. So we can infer that the majority of married people probably spend more than the majority of people in other states, and they have to take on more expenses, maybe a house, a car, or other spending on their children.

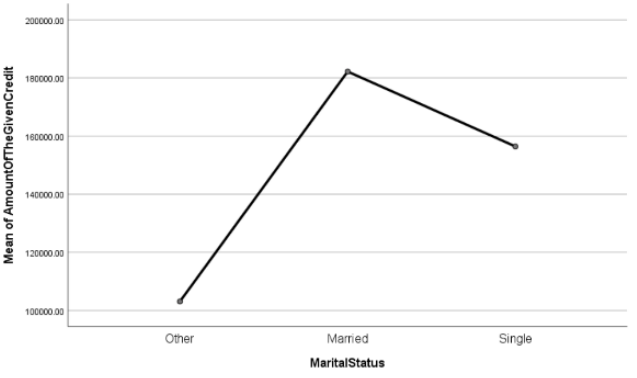

Figure 3. The relationship between marital status and loan amount.

DOI: 10.54254/2755-2721/32/20230863

In addition, this work can relate Figure 4 to the relationship between age and marital status. As Figure 4 shows, marriage rates are higher among those over 33. Combined with the above analysis, married people borrow the most, so it can be guessed that people in this age group consume more and have more expenses.

Figure 4. The relationship between marital status and age.

3.3. Education and overdue

From Figure 5, we can get an intuitive conclusion that people with higher education are less likely to default. But the difference between the default probability of a college degree and a high school degree is not big. The default probability of graduate school students is significantly lower than the previous two, which can be guessed that this group may have higher salary income after graduation, or the number of loans of this group is smaller.

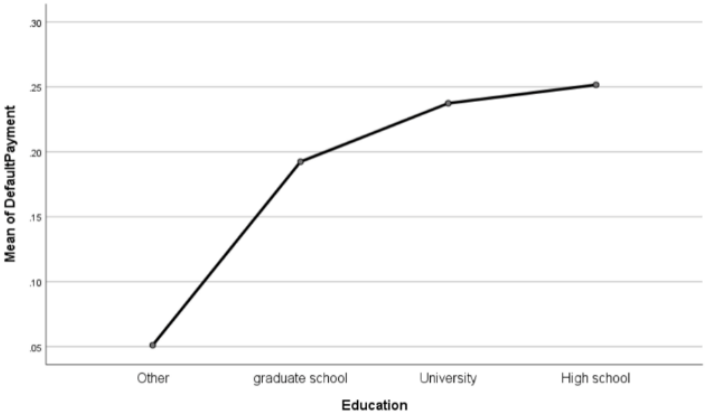

Figure 5. The relationship between education and default.

4. Binary logistic regression

From the table 8, we can see that the B coefficient of gender is -0.105, so gender is an influential factor to predict whether an individual is overdue, and its Exp(B) is 0.9, indicating that the aberration of overdue probability between men and women is not very large.

If the amount of the given credit has a B coefficient of 0, it can be considered not an influencing factor.

The B-coefficient of education is -0.054, so it is also one of the factors influencing the prediction result. Exp(B) shows that the aberration of overdue probability of different education levels is about 0.95 times.

DOI: 10.54254/2755-2721/32/20230863

Marital status is also one of the influential factors of overdue prediction, with a coefficient of -0.152, which has a greater impact on the overall prediction result, and its Exp(B) ranges from 0.806 to 0.915, indicating that the overdue probability of different maretal status differs by about 0.86 times.

The 3456 history of payment had no effect on the forecast, because the sig. is bigger than 0.05.

(In this experiment, the reason why we did not use the compressed amount of previous payment is that only when it is compressed to 4 or 5, can the data have a better generalization, so we used the original data of 6 periods for the experiment.)

Table 8. Binary logistic regression.

				95% C.l.for EXP(B)	
	B	Sig.	Exp(B)	Lower	Upper
Gender	-.105	.001	.900	.848	.956
Amount Of The Given Credit	.000	.000	1.000	1.000	1.000
Education	-.054	.020	.948	.906	.991
Marital Status	-.152	.000	.859	.806	.915
Age	.006	.001	1.006	1.003	1.010
Amount Of Previous Payment1	.000	.000	1.000	1.000	1.000
Amount Of Previous Payment2	.000	.000	1.000	1.000	1.000
Amount Of Previous Payment3	.000	.005	1.000	1.000	1.000
Amount Of Previous Payment4	.000	.000	1.000	1.000	1.000
Amount Of Previous Payment5	.000	.001	1.000	1.000	1.000
Amount Of Previous Payment6	.000	.060	1.000	1.000	1.000
History of payment 3456	.032	.123	1.032	.991	1.075
History of payment 123	-.146	.000	.864	.831	.898
Constant	-.854	.000	.426		

Form the Table 8 and 9, we can see the model has a relatively good predict ability, the overall accuracy is about 70% (In the experiment we set the dividing value as 0.220).

DOI: 10.54254/2755-2721/32/20230863

Table 9. Classification Table.

Observed			Predicted		
			Default Payment		Percentage Correct
			No	Yes	
Step 1	Default Payment	No	16112	6947	69.9
		Yes	2457	4153	62.8
	Overall Percentage				68.3

5. Conclusion

This paper mainly uses SPSS to analyze the influence of different variables on the overdue credit repayment of commercial banks, including factor analysis, binary logistic regression, ANOVA, comparative analysis and others. At the same time, for these different analysis methods, we also made a summary of the results and chart analysis. Through the analysis, we have reached some conclusions, including the higher the degree of the group of people delinquent rate is lower, the probability of married people delinquent rate is higher than unmarried people, the amount of married people loan is higher, the difference between men and women is not significant and so on. At present, with the development of the global economy, the cooperation between enterprises and banks is deepened, and the bank pursues profit as the goal, so it is inevitable that customers will delay their payment in the process. The phenomena told by these data can enable us to predict the lender's loan risk well in advance, thus avoiding many problems and accidents in advance and increasing profits. Therefore, commercial banks should always pay attention to the changes in the global economic situation, collect the information of each customer, understand the development prospects of customer business, establish and improve the credit risk assessment system, so as to reduce risks and promote the development of banks.

References

[1] Moradi, S., Mokhatab Rafiei, F., Financ Innov 5, 15 (2019), A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks, https://doi.org/10.1186/s40854-019-0121-9

[2] J.I. Gusti Ngurah Narindra Mandala, Catharina Badra Nawangpalupi, Fransiscus Rian Praktikto (2012), Assessing Credit Risk: An Application of Data Mining in a Rural Bank, https://doi.org/10.1016/s2212-5671(12)00355-3

[3] Jonathan N. Crook a, David B. Edelman b, Lyn C. (2007), Thomas c Recent developments in consumer credit risk assessment, https://doi.org/10.1016/j.ejor.2006.09.100

[4] M Doumpos a, K Kosmidou a, G Baourakis b, C Zopounidis (2002), a Credit risk assessment using a multicriteria hierarchical discrimination approach: A comparative analysis, https://doi.org/10.1016/S0377-2217(01)00254-5

[5] Paweł Pławiak a, Moloud Abdar b, U. Rajendra Acharya c (2019), Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring, https://doi.org/10.1016/j.asoc.2019.105740

[6] Charles Guan, Hendra Suryanto, Ashesh Mahidadia, (2023), Michael Bain & Paul Compton Responsible Credit Risk Assessment with Machine Learning and Knowledge Acquisition, https://link.springer.com/article/10.1007/s44230-023-00035-1#Sec5

[7] David West (2000), Neural network credit scoring models, https://doi.org/10.1016/S0305-0548(99)00149-5

DOI: 10.54254/2755-2721/32/20230863

[8] Stjepan Oreski a, Dijana Oreski b, Goran Oreski a (2012), Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment, https://doi.org/10.1016/j.eswa.2012.05.023

[9] Hussain Ali Bekhet, Shorouq Fathi Kamel Eletter (2014), Credit risk assessment model for Jordanian commercial banks: Neural scoring approach, https://doi.org/10.1016/j.rdf.2014.03.002