Consumer Default Risk Assessment in a Banking Institution

Eliana Costa e Silva, Isabel Cristina Lopes, Aldina Correia and Susana Faria

Abstract. Credit scoring is an application of financial risk forecasting to consumer lending. In this study, statistical analysis is applied to credit scoring data from a financial institution to evaluate the default risk of consumer loans. The default risk was found to be influenced by the spread, the age of the consumer, the number of credit cards owned by the consumer. A lower spread, a higher number of credit cards and a younger age of the borrower are factors that decrease the risk of default. Clients receiving the salary in the same banking institution of the loan have less chances of default than clients receiving their salary in another institution. We also found that clients in the lowest income tax echelon have more propensity to default.

Keywords: Data analysis, Nonparametric statistical tests, Credit scoring, Financial mathematics, Actuarial sciences
PACS: 02.50.Sk, 02.50.Tt, 29.85.Fj

INTRODUCTION

Credit scoring is the assessment of the risk associated with lending to an organization or an individual [1]. Credit risk modeling is very helpful in consumer credit loan grant decision. A bad customer (*Defaulted*) is commonly taken to be someone who has missed three consecutive months of payments [2]. Models of credit scoring are based on historical information from a data set of existing clients, in order to assess whether the prospective client will have a greater chance of being a good or bad payer. Consumer credit risk assessment involves the use of risk assessment tools to manage a borrower's account, from the time of pre-screening a potential application through, to the management of the account during its life and possible write-off [1].

Credit scoring is used in almost all forms of consumer lending as: credit cards, personal loans, car finance, insurance policies, utility payments. Virtually all major banks use credit scoring with specialized consultancies providing credit scoring services and offering powerful software to score applicants, monitor their performance and manage their accounts [1]. Financial institution systems incorporate models of credit scoring to permit on-line credit evaluation, and thereby getting increasing profits [3].

The ability of a performance measure to capture the true skill of a model is highly dependent on the available data for assessment [4]. Beyond the social-economical characteristics of the individual, the underlying economic conditions also have a major impact on default. These scoring systems raise social issues, for which institutions are accused of discriminating consumers in the access to credit, and although it is illegal to use some characteristics such as race, gender, or religion, some authors defend the use of surrogate variables.

Other studies include statistical techniques such as discriminant analysis, linear regression, logistic regression, classification trees, and Bayesian statistics (e.g. [5, 6, 7, 8]). Also Operational Research based approaches include variants of linear programming, genetic algorithms, nearest neighbor search, and AI modeling approaches such as neural networks and expert systems have been applied to credit risk prediction [1].

To collect data, the lenders usually observe the performance of the repayment of the borrowers, who were made an offer of a loan which was accepted. This information consists on socio-demographic characteristics of the borrower at the time of application (extracted from his/her application form), the repayment performance of each borrower on other loans and of individuals who live in the same neighborhood.

The objective of this paper is to analyze, using statistical inference techniques, such as the Mann-Whitney-Wilcoxon and Pearson Chi-squared independence tests, to infer the factors that influence credit risk from massive payment history data of customers of a banking institution.

DATA ANALYSIS

The data set used in this work contains financial data regarding consumer loans and a brief social characterization of the clients of a banking institution, in the years of 2008 and 2009, in a total of 14 variables:

- 8 quantitative variables — `ContractedCapital` represents the capital negotiated in the loan agreement (in €); `CapitalOutstanding` represents the capital that it is still owed to the bank at the moment (in €); `Spread` the interest rate spread is the interest rate charged by banks on loans to private sector customers minus the interest rate paid by commercial banks for demand, time, or savings deposits (in percent points); `Term` the length of the loan (in years); `MonthlyInstallment` the amount to be paid to the bank each month (in €) `Age` age of the borrower (in years); `Seniority` for how long the borrower has been a client of the banking institution (in years); `NumCreditCards` number of credit cards owned by the client;
- and 6 qualitative variables — `Sex` gender of the borrower; `MaritalStatus` marital status of the client; `Salary` indicates whether the salary of the borrowed is received in an account in the same baking institution in which the loan is made or if it is received in some other banking institution; `OtherCredit` indicates if the client has other credits; `TaxEchelon` the tax echelon of the client (1 is the lowest income echelon, and 6 is the highest income echelon); `Default` indicates if the borrower is in default, i.e., if the borrower has not made a scheduled payment of interest or principal (0 = No, 1 = Yes).

This data set is a simple random sample of all the banking institution records, composed of 3221 individuals, where 319 have defaulted, making an observed default rate of 10%, approximately.

We will compare the two groups according to the variable `Default`.

Table 1 shows the means and standard deviations (between brackets) of each quantitative variable and for each group. At a first sight there seems to be no difference between these two groups. Accordingly to the Kolmogorov-Smirnov test, the data is not normally distributed, therefore the nonparametric Mann-Whitney-Wilcoxon test was used to compare the medians of each variable (see Table 1). The results show that, for a 5% significance level, there are differences between the medians in the groups of defaulters and non-defaulters, for the variables `ContractedCapital`, `CapitalOutstanding`, `Spread`, `Age` and `NumCreditCards`. This shows that these five variables may influence credit risk.

However, for the variables `ContractedCapital` and `CapitalOutstanding`, if a 1% significance level was considered, the null hypothesis would not be rejected, i.e, there are no statistical significant differences between the medians in the groups of defaulters and non-defaulters. Also, when testing the variables `Term`, `Monthly Installment` and `Seniority` there is no statistical evidence, at a 5% significance level, to reject the null hypothesis. Hence these variables may not be relevant to explain the variable `Default`.

TABLE 1. Mean, standard deviation and Mann-Whitney-Wilcoxon test for the two groups defined according to variable `Default` (* p<0.05, ** p<0.01, *** p<0.001).

Variable	All cases Mean (Std.Deviation)	Default 0 (No) Mean (Std.Deviation)	Default 1 (Yes) Mean (Std.Deviation)	Mann-Whitney-Wilcoxon's test p-value
ContractedCapital, €	77101.88 (57214.12)	77663.80 (57194.09)	71990.02 (57232.54)	0.0155 *
CapitalOutstanding, €	71551.64 (53884.45)	71984.82 (53838.46)	67610.97 (54226.98)	0.0386 *
Spread, %	0.925 (0.617)	0.891 (0.586)	1.240 (0.786)	<2.2e-16 ***
Term, years	30.83 (10.04)	30.81 (10.10)	31.05 (9.52)	0.7478
Monthly Installment, €	282.10 (220.66)	282.80 (219.59)	275.75 (230.46)	0.1042
Age, years	42.98 (9.78)	42.73 (9.72)	45.26 (10.07)	1.403e-05 ***
Seniority, years	7.71 (5.40)	7.73 (5.41)	7.46 (5.27)	0.7853
NumCreditCards	0.49 (0.66)	0.54 (0.67)	0.08 (0.33)	<2.2e-16 ***
Number of contracts	3221	2902 (90%)	319 (10%)	

To check if the qualitative variables have some influence on the probability of occurring a default, crosstabulation and the Pearson Chi-squared independence test were used (see Table 2). For the binary variables `Sex`, `Salary` and `OtherCredit`, the Yates' continuity correction was used.

The results show that the credit default risk depends on receiving the salary in the same banking institution. In fact, the crosstabulation of these variables (`Salary` and `Default`) shows that when the salary is received in the

same institution, the percentage of defaulted clients is smaller then when the salary is received in a different banking institution (Table 2).

The results also show that, at a 5% significance level, the tax echelon is not independent of the default risk. In fact, for clients in the lowest tax echelon, the percentage of defaulted clients is larger, comparing to what happens in other echelons. Therefore, both variables, `Salary` and `TaxEchelon`, might be important factors in default risk assessment.

TABLE 2. Crosstabulation of qualitative variables vs. Default and Pearson's Chi-squared Independence test.

Variable	Default 0 (No)	Default 1 (Yes)	Row Total	Pearson's Chi-squared p-value
Sex				0.4331
F (Feminine)	838 (26.05%)	85 (2.64%)	924 (28.69%)	
M (Masculine)	2063 (64.05%)	234 (7.26%)	2297 (71.31%)	
Marital Status				0.1403
0 (Unknown)	188 (5.837%)	27 (0.838%)	215 (6.68%)	
1 (Single)	731 (22.695%)	61 (1.894%)	792 (24.59%)	
2 (Union of fact)	14 (0.435%)	1 (0.031%)	15 (0.47%)	
4 (Married)	1675 (52.002%)	189 (5.868%)	1864 (57.87%)	
5 (Separated)	15 (0.466%)	1 (0.031%)	16 (0.497%)	
6 (Divorced)	254 (7.886%)	35 (1.087%)	289 (8.972%)	
7 (Widowed)	25 (0.776%)	5 (0.155%)	30 (0.931%)	
Salary				6.2e-19 ***
0 (Other Institution)	1162 (36.08%)	211 (6.55%)	1373 (42.63%)	
1 (Same Institution)	1740 (54.02%)	108 (3.35%)	1848 (57.37%)	
Other Credit				0.2923
0 (No)	2070 (64.266%)	218 (6.768%)	2288 (71.033%)	
1 (Yes)	832 (25.830%)	101 (3.136%)	933 (28.966%)	
Tax Echelon				5.4e-63 ***
1	1345 (41.757%)	311 (9.655%)	1656 (51.413%)	
2	160 (4.967%)	1 (0.031%)	161 (4.998%)	
3	379 (11.767%)	2 (0.062%)	381 (11.829%)	
4	334 (10.369%)	4 (0.124%)	338 (10.494%)	
5	129 (4.005%)	0 (0.000%)	129 (4.005%)	
6	555 (17.231%)	1 (0.031%)	556 (17.262%)	
Column Total	2902 (90.10%)	319 (9.90%)	3221	

For the variables `Sex`, `OtherCredit` and `MaritalStatus`, the p-values are greater than 0.05, therefore we do not reject the null hypothesis of independence (see Table 2).

To better understand what is the influence of the tax echelons on the credit default, new variables with less number of categories were created and the corresponding independence tests were performed. In the crosstabulation of `Default` vs. `TaxEchelon`, there were categories with 0 count, hence we decided to group the two highest income echelons and create the variable `TaxEchelon_5Cat`: a qualitative variable with 5 categories, where code 1 is for the lowest income echelon, and code 5 gathers the two highest income echelons (formerly known as codes 5 and 6, in the original variable `TaxEchelon`). The Pearson's Chi-squared Independence Test of `Default` vs. the new variable `TaxEchelon_5Cat` produced no 0 count cells, and resulted in χ^2 = 301.2349, with 4 degrees of freedom, and a p-value of 5.9e-64<0.05, thus for a 5% significance level the null hypothesis is rejected, and therefore the non independence of the variables can be assumed.

However, if we perform the same test using only the records that do not have the lowest income tax echelon, the result will be completely different: χ^2 = 4.839546, with 3 degrees of freedom, and a p-value 0.1839. The null hypothesis is not rejected for a 5% significance level, and the independence of tax echelon and default could be assumed. The reason for this can be identified as the first tax echelon produces a significant influence on the default risk.

For this reason we created a new binary variable named `TaxEchelon_Bin` coded as: 1 if `TaxEchelon`=1 and 0 otherwise. The Pearson's Chi-squared Independence Test with Yates' continuity correction of Default vs. the new binary variable `TaxEchelon_Bin` produced a test statistic of χ^2 = 298.9151, with a p-value of 5.7e-67 < 0.05,

meaning that for a 5% significance level the null hypothesis can be rejected, and therefore being in the first tax echelon or in another echelon has a significant influence on the default risk.

CONCLUSIONS AND FUTURE WORK

Credit risk modeling is very helpful in consumer credit loan grant decision. In this work we have analyzed the factors that influence credit risk. The statistical analysis was performed on a data set containing financial data regarding consumer loans and a brief social characterization of the clients of a banking institution. A random sample of 3221 records of the banking institution was considered and a default rate of approximately 10% (319 defaulters) was observed. A total of 14 variables (8 quantitative and 6 qualitative) were analyzed.

The nonparametric Mann-Whitney-Wilcoxon test was used to compare the medians of each continuous variable, separated in two groups according to the variable `Default`. For a 5% significance level, it was found that there are differences between the medians in the groups of defaulters and non-defaulters, for the variables `ContractedCapital`, `CapitalOutstanding`, `Spread`, `Age` and `NumCreditCards`. While there is no statistical evidence, at a 5% significance level, that the variables `Term`, `Monthly Installment` and `Seniority` influence the risk of default.

The influence of the qualitative variables on occurring a default, was analyzed using crosstabulation and the Pearson Chi-squared independence test. The results show that `Salary` and `Tax Echelon` are important factors in default risk assessment. In fact, the percent of defaulted clients is smaller when the salary is received in the same institution than when it is received in a different banking institution. Also, for clients in the lowest tax echelon, the percent of defaulted clients is larger, comparing to what happens in other echelons. No statistical evidence was found that the variables `Sex`, `OtherCredit` and `MaritalStatus` have influence on credit risk.

In summary, the data suggests that the decision of granting consumer credit should take into account variables such as `CapitalOutstanding`, `Spread`, `Age`, `NumCreditCards`, `Salary` and `Tax Echelon`. However, demographic variables such as `Sex` and `MaritalStatus`, and economic characteristics of the individuals such as `Term`, `Monthly Installment`, `Seniority` and `OtherCredit`, were found not to be relevant to actuate on default.

The questions that a banking institution should include in an application form, in order to identify if a candidate to a loan has high risk of failure, are extremely important to the accuracy of the classification of credit risk. The advantages of using regression models consist of the possibility to perform statistical tests to identify how important is each question and whether two different questions are essentially asking the same thing and getting equivalent responses. This allows to drop unimportant questions, making scorecards more robust, and helps in deciding what questions to ask in new scorecards. In the future we will apply a logistic regression model to credit scoring data to evaluate the default risk of consumer loans.

REFERENCES

1. J. N. Crook, D. B. Edelman, and L. C. Thomas, *European Journal of Operational Research* **183**, 1447–1465 (2007), ISSN 0377-2217.
2. L. C. Thomas, *International Journal of Forecasting* **16**, 149–172 (2000), ISSN 0169-2070.
3. M. A. Gouvêa, and E. B. Gonçalves, "Credit Risk Analysis Applying Logistic Regression, Neural Networks and Genetic Algorithms Models," in *Production and Operations Management Society 18th Annual Conference*, Dallas, Texas, U.S.A., 2007.
4. J. Chen, S. Dhar, D. Duffy, Y. Liu, R. O. Moore, M. Pedneault, A. Pole, Y. Qian, D. Rumschitski, T. Wang, and M. Zyskin, New performance measures for credit risk models, Tech. rep., Standard & Poor's Rating Services following 2014 Workshop on Mathematical Problems in Industry, held at New Jersey Institute of Technology, Newark, N.J., U.S.A. (2014).
5. M. Karan, A. Ulucan, and M. Kaya, *Central European Journal of Operations Research* **21**, 479–494 (2013), ISSN 1435-246X.
6. S. Mestiri, and M. Hamdi, *International Journal of Management Science and Engineering Management* **7**, 200–204 (2012).
7. K. Fang, and H. Huang, *Journal of Computers* **6**, 1868–1874 (2011).
8. C. Yeh, and T. Lee, *International Research Journal of Applied Finance* **IV**, 730–739 (2013).