



Creating a Customer Profile in a Credit Institution

Julian Vasilev*

*Department of Informatics,
Varna University of Economics, Bulgaria*

Abstract— *The purpose of this article is to create a customer profile in a credit institution. The customer profile is built on the basis of historical data. It is used mainly to predict the reliability of a new customer. If a new loan is given, what is the possibility to be returned? A multi-factor model is created. The article is published for the first time. Unlike other articles where questionnaires are used in this paper transactional data for given loans are used. Collected data are analysed by contemporary statistical methods in SPSS. Initially five hypotheses are defined. The assumption is that the possibility of returning a loan depends on the sum of the contract, gender, age, year and month of signing the contract. By applying statistical methods the most important factors are discovered – year, month and age. It is proved that the sum of contract and the gender do not influence the possibility of returning a loan.*

Keywords— *credit profile, credit institution, SPSS, credit risk, customer profile*

I. INTRODUCTION AND LITERATURE REVIEW

Creating a customer profile in a credit institution is a difficult task. A given loan may be returned or may not be returned. Factors are quite lot. Our task is to choose several factors and check their influence on the possibility of returning the loan. In economics many factors affect the result variable. For the sake of the current investigation we have access to loan data. If we choose those factors, which have the strongest influence on the possibility of returning the loan, we will reach the purpose of the article. Credit institutions are widely studied. Sometimes data in credit institutions are collected by questionnaires. Bradutanu (e.g. [1]) analyses the staff structure. She finds out that employees work harder and sometimes overtime to keep their jobs. Credit institutions are sometimes mixed with banks because they have common activity. Thus some articles (e.g. [2]) focus on the lending mechanisms and credit scoring method. They also analyse transactional data in commercial banks. Since all these data are confidential we will make some transformations to hide personal data. The described credit scoring system is quite complex and it usually applies to banks but not to small credit institutions. Nitescu (e.g. [3]) focuses on customer factors influencing bank customer behaviour and their impact on early repayment of loans. Estimating the risk of a loan is a difficult task. Ichim (e.g. [4]) analyses local public loans and refinancing of bonds. Rambaud (e.g. [5]) finds out that banks are giving new type of loan products but they are afraid whether a customer could face corresponding payments. A mathematical expression of the average duration of the loan is calculated.

A lot of researchers try to evaluate credit risk. Hai et al. (e.g. [6]) argue that most Chinese banks have not established a credit risk rating systems when giving loans to farmers. They use correlation analysis and significant discriminant to find indicators which distinguish default customers from non-default ones. They made the same conclusion as it is made in our article – the age is an important factor, but the gender is not. Li et al. (e.g. [7]) make a study of personal credit evaluation method based on PSO-RBF neural network model.

II. DATA COLLECTION

A crediting institution has a lot of data. We have to choose reliable data for our research. We have to transform or exclude personal data in the investigation. Using the ETL method we have extracted the data in a tabular format. Column names are the following: sum of the contract (loan), gender, year, month, finished (“true” if the sum is returned) and age. The year and month are extracted from the date of the initial signing the contract. The age is calculated at the date of signing the contract. Data are collected between 2012, July and 2013 March from a small credit institution in Bulgaria. When a loan is given, there is a period when the credit institution waits to get money back. That is why historical data do not affect loans given during the current 2014 year. Even though data are for 9 months ago, we have actual information about those loans given in 2013, March whether they are returned or not.

Our dataset consists of 1880 records. We have enough detailed data to get tendencies and to mark influences. A small part of the dataset is given in Table 1.

TABLE I: A PART OF THE DATASET USED TO ANALYSE DATA

Sum_contract	Is_Male	Year	Month	Finished	Age
70	FALSE	2012	7	TRUE	43
30	FALSE	2012	7	TRUE	67
40	FALSE	2012	7	TRUE	30

100	TRUE	2012	7	TRUE	33
100	FALSE	2012	7	TRUE	37
225	TRUE	2012	7	FALSE	39

III. DEFINING HYPOTHESIS

Defining hypothesis is a process when we predict the influence of a factor on a result variable. In economics many factors affect the result variable. If we are lucky to choose the best variables, we will get those most influencing factors. But in real life we have data which may be used for formulating hypothesis. We will start with one-factor analysis and we will add factors. We will check for autocorrelations.

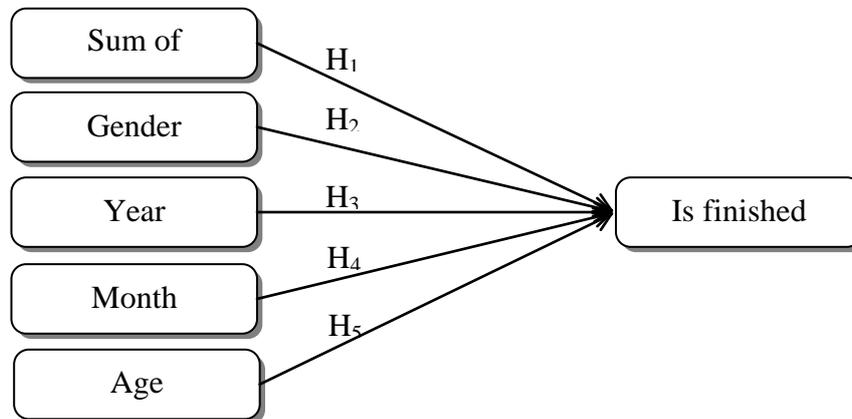


Fig. 1 Initial model for creating customer profile

Our prediction for the loan – to be returned or not starts with defining five hypothesis:

- H1: The initial sum of loan is a factor which results in the reliability of a customer.
- H2: The gender of a customer is a factor which results in the reliability of a customer.
- H3: The year when the loan was given is a factor which results in the reliability of a customer.
- H4: The month when the loan was given is a factor which results in the reliability of a customer.
- H5: The age of the customer is a factor which results in the reliability of a customer.

First we have to check if there is a causal link between each factor and the resulting variable. Second, if there is a casual link we will try to determine the direction by the use of an appropriate correlation coefficient.

IV. ENTERING DATA IN SPSS

SPSS is well-known software for analysing data and checking hypotheses. Before entering data in SPSS we have to use the variable view to define variables: sum_contract, gender, year, month, finished, age. The consequence of variables in SPSS is the same, as it is in the Excel spreadsheet. It is very important to give an appropriate measure for each variable.

TABLE III
DEFINING VARIABLES IN SPSS

Variable	Measure
sum_contract	Scale
Gender	Nominal
Year	Scale
Month	Scale
finished	Nominal
Age	Scale

Since SPSS works with numeric values in Excel we have to do some transformations before Copy-Pasting data from Excel to SPSS. As it is obvious in table 1, we have a column “Is_male” with values “true” and “false”. We have to use a built-in function IF or to use the command “Find and replace” to convert all cells, containing “true” to “1” (which means “male”) and all cells, containing “false” to “0” (which means “female”). We also have to transform all values in the column “finished”. We will convert all cells, containing “true” to “1” (which means “the loan is returned”) and all cells, containing “false” to “0” (which means “the loan is not returned”).

TABLE IIIII
CONVERTED DATA IN EXCEL

Sum_contract	Is_Male	Year	Month	Finished	Age
70	0	2012	7	1	43
30	0	2012	7	1	67

40	0	2012	7	1	30
100	1	2012	7	1	33
100	0	2012	7	1	37
225	1	2012	7	0	39

Now we have prepared data which may be copy-pasted from Excel in SPSS.

V. ANALYZING DATA IN SPSS

A. Testing H1 – does the reliability of customer depend on the sum of the loan

We start data analysis with checking each hypothesis. We use Descriptive Statistics (Analyze/Descriptive Statistics/Crosstab). We mark Chi-square and contingency coefficient because the resulting variable is on a nominal scale. The Chi-square test shows Pearson Chi-square value of 152 which is statistically significant (Asymp. Sig. 2-tailed is 0.000). But the comment is that 63.4% of the cells have expected count less than 5. The minimum expected count is 29. This result shows that for the independent variable “sum_contract” we may transform data from strong to weak scale. The contingency coefficient is 0.274. It is statistically significant (Approx. Sig. is 0.000). Now we can reject H1. Thus the initial sum of the loan does not influence significantly the reliability of the customer – whether he/she will return the loan. We will transform data from column “sum_contract” into new variable “sum_loan” in an ordinal scale – “low” (for sums less than 100), “medium” (for sums between 100 and 200) and “high” (for sums more than 200). Now we may check H1, but assuming the sum of loan is low, medium or high. In this case the Pearson Chi-square value is 5.882. It is not statistically significant (Asymp. Sig. 2-tailed is 0.053 which is more than 0.05). The contingency coefficient 0.056 which is not statistically significant (Approx. Sig. is 0.053 which is more than 0.05). The experiment with the new variable “sum_loan” also leads to rejecting H1.

B. Testing H2 – does the reliability of customer depend on the gender

Again we use descriptive statistics. The Chi-square tests show a comparatively high value of Pearson Chi-square 0.612. But it is not statistically significant (Asymp. Sig. 2-sided is 0.434. It is more than 0.05). Symmetric measures show the value of contingency coefficient 0.018 which is not statistically significant (Approx. Sig. is 0.434 which is more than 0.05). Now we can reject H2. Thus the gender of the customer does not affect significantly the reliability of the customer – whether he/she will return the loan.

Since the independent variable “gender” has dichotomy values (males and females) we may make independent samples t-test to check the variances between males and females, when we estimate their reliability as people who take loans. The grouping variable is “gender”. The tested variable is “finished”. The Leven’s test for equality of variances shows the value of F-test 2.458 with significance of 0.117 (0.117>0.05). It means that equal variances assumed. The t-test for equality of means shows the value of -0.782 with Sig. (2-tailed) 0.434 (0.434>0.05). The t-test shows that there are not significant differences between the means of two groups (males and females). The independent samples t-test confirms our conclusion – we rejected H2. There is not a significant difference in the reliability of customers between males and females.

We may make nonparametric tests “Two-independent-samples tests”. The test type is Mann-Whitney U test. The value of Mann-Whitney U test is 431907. It is not statistically significant (Asymp. Sig. 2-tailed is 0.434). The nonparametric test confirms our conclusion – we have to reject H2.

C. Testing H3 – does the reliability of customer depend on the year

Since all economic events depend on the current social situation we make the assumption that dynamic social and business events influence the reliability of a customer or the reliability of each contract. The Pearson Chi-square value is 156.525. It is statistically significant (Asymp. Sig. 2-sided is 0.000). So there is a link between the year and the possibility of returning the loan. The contingency coefficient is 0.277. It is statistically significant (Approx. Sig. is 0.000). The symmetric measures show that the link between the independent variable “year” and “finished” is weak. There are other factors that influence the resulting variable. We accept H3 – the reliability of the customer depends on the year the loan is given.

The crosstab between variables “year” and “finished” shows the result of some events.

TABLE IV: CROSSTAB BETWEEN VARIABLES “FINISHED” AND “YEAR”

Year	Finished		Total
	0 (no)	1 (yes)	
2012	242	993	1235
2013	304	340	644
Total	546	1333	1879

Crosstab data in table 4 show that 71% of loans are returned (1333/1879). For the year 2012 the percent of returned loans is 80%. For the year 2013 the percent of returned loans is 53%. We may assume that this negative effect (decreasing the part of returned loans) with the increase of unemployment in Bulgaria. We have data for 2012 and 2013. We may make 2-independent samples t-test. The grouping variable is “year”. The test variable is “finished”. The Leven’s

test for equality of variances shows that F value is 368.545 with significance 0.000. It means that equal variances are not assumed. The t-test for equality of means is 12.163 with sig. 2-tailed 0.000. It means that equal means are not assumed. The 2-independent samples t-test shows that the reliability of a customer (when taking a loan) depends on the year.

The value of the nonparametric Mann-Whitney U test is 287784. Asymp. Sig 2-tailed is 0.000. The nonparametric Mann-Whitney U test shows that the year has an influence on returning the loan.

The one-way ANOVA test confirms that the factor “year” influences the resulting variable “finished”.

D. Testing H4 – the possibility of returning a loan depends on the month

Economic phenomena and processes happen throughout the year. The employment in tourism for instance may influence the intention to take a loan and the possibility the loan to be returned later. That is why we assume that the month has an influence on the probability of returning the loan. The Pearson Chi-Square value is 250.041. Asymp Sig. 2-sided is 0.000. Thus the factor “month” influences the reliability of a customer. The symmetric measures show the value of contingency coefficient 0.343. Approx Sig. is 0.000. We accept H4 – the month influences the possibility of returning the loan. The influence is not strong because other factors are included in the model.

In Excel or in SPSS we may calculate the possibility of returning the loan for each month. So we have a table with risky and favourable months for giving loans.

TABLE V
CROSSTAB BETWEEN VARIABLES “FINISHED” AND “MONTH”

month * finished Crosstabulation				
% within month				
		finished		Total
		0 (no)	1 (yes)	
Month	1	31,9%	68,1%	100,0%
	2	40,7%	59,3%	100,0%
	3	69,4%	30,6%	100,0%
	7	7,3%	92,7%	100,0%
	8	23,7%	76,3%	100,0%
	9	23,7%	76,3%	100,0%
	10	19,6%	80,4%	100,0%
	11	13,2%	86,8%	100,0%
	12	19,0%	81,0%	100,0%
Total		29,1%	70,9%	100,0%

Table 5 shows the most risky month – March. 30.6% of loans given in March are returned. Thus 69.4% of loans given in March are not returned. Loans given in January and February are also risky. Best loans are given in July. 92.7% of them are returned. Loans given in the last three months of the year are not returned in less than 20% of the cases.

E. Testing H5 – the possibility of returning a loan depends on the age

The age of a customer may influence the possibility of returning it. The Pearson Chi-Square value is 96.987. Asymp. Sig. 2-tailed is 0.000. So there is an influence of the age on the possibility of returning the loan. The contingency coefficient is 0.222. Its approx. sig. is 0.000. Thus the age is a factor in the model. We accept H5 – the age influences the possibility of returning the loan.

We may transform the age variable into an ordinal scale. Since the younger customer is 20 years old and the older is 75, we make intervals 10 years wide.

TABLE VI
AGE GROUPS AND THE POSSIBILITY OF RETURNING THE LOAN

age_diap * finished Crosstabulation				
% within age_diap				
		finished		Total
		0	1	
age_diap	<30	29,1%	70,9%	100,0%
	30-40	27,0%	73,0%	100,0%
	40-50	31,9%	68,1%	100,0%
	50-60	35,0%	65,0%	100,0%
	60-70	28,5%	71,5%	100,0%
	>70	13,6%	86,4%	100,0%
Total		29,1%	70,9%	100,0%

The most risky groups are people between 50 and 60 years. They take a loan but it is difficult for them to return it. 35% of people between 50 and 60 years do not return their loan. 29.1% of young people (below 30 years) cannot return their loans. The best customers are aged over 70. 86.4% of them return their loan.

After rejecting H1 and H2 and accepting H3, H4 and H5 we may build the final model (fig. 2).

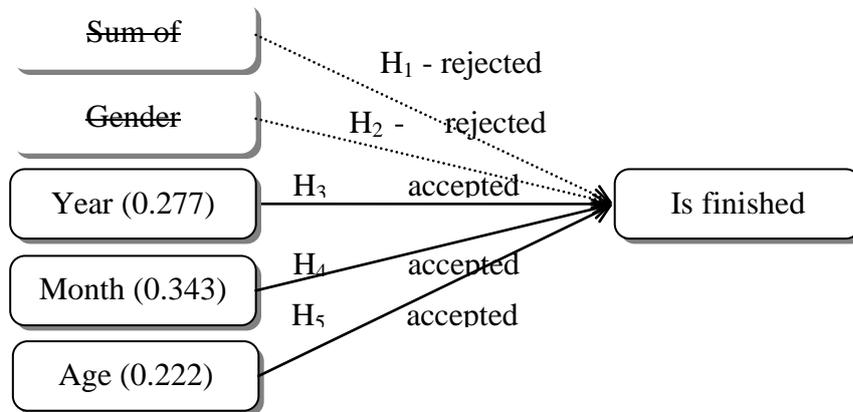


Fig. 2 Final three factor model with influence of each factor

VI. CONCLUSIONS

Creating a customer profile in a credit institution is a difficult task. The possibility of returning the loan depends on many factors. In this study we checked five hypotheses – we assumed that the sum of loan (contract), gender, year, month and age have the best influence on the possibility of returning a loan. The calculations in SPSS showed that the sum of loan and gender are not factors that affect the reliability of a customer. The year, month (of signing the contract for loan) and the age of the customer have an influence. Even though their influence is not strong, some assumptions and recommendations are made – which months and which age groups are risky. Future research on the topic may continue testing other independent variables in the model. The dataset may be imported into a neural network to check the relative importance of different independent variables on the resulting variable. Since the dependent variable has Boolean values a probit regression model may be defined and tested.

REFERENCES

- [1] D. Bradutanu, *Do really employees resist change? Case study at a credit institution*. Challenges of the Knowledge Society, Vol 2, Iss 1, Pp 1263-1268, 2012.
- [2] D. Delia and Alexandru S. *Aspects regarding the mechanism of bank lending for individuals*. Studia Universitatis Vasile Goldis Arad, Seria Stiinte Economice, Vol 22, Iss 1, Pp 47-54, 2012.
- [3] D. Nitescu. *Prepayment risk, impact on credit products*. Theoretical and Applied Economics, Vol 8(573), Iss 8(573), Pp 53-62, 2012.
- [4] C. Icihim, *LOCAL PUBLIC ADMINISTRATION AUTHORITY LOANS*. USV Annals of Economics and Public Administration, Vol 13, Iss 1(17), Pp 245-251, 2013.
- [5] Rambaud, S. *A Financial Analysis of Certain Flexible Loans: Calculation of the Average Duration*. International Journal of Economics and Finance, Vol 5, Iss 4, 2013.
- [6] L. Hai et al. *A Credit Risk Evaluation Index System Establishment of Petty Loans for Farmers Based on Correlation Analysis and Significant Discriminant*. Journal of Software, Vol 8, Iss 9, Pp 2344-2351, 2013.
- [7] S. Li et al. *Study of Personal Credit Evaluation Method Based on PSO-RBF Neural Network Model*. American Journal of Industrial and Business Management, Vol 03, Iss 04, Pp 429-434, 2013.