AN APPROACH OF ESTIMATING THE PROBABILITY OF BEING GOOD FOR NEW BORROWERS

Vesela Mihova, Velizar Pavlov

University of Ruse, 8 Studentska Street, 7017 Ruse, Bulgaria

Abstract

Statistical models are commonly used in the banking industry in order to assess the credit risk associated with the approval of people applying for certain products (loans, credit cards, etc.). Based on data from the past, these models try to predict what will happen in the future. This work has studied the causal link between the conduct of an applicant upon payment of the loan and the data that he completed at the time of application. A linear regression is used to estimate the probability of being good for new borrowers, and a scorecard is obtained from the linear model to assess new customers in the time of application.

Key words: *credit risk, modelling, scorecards, data analysis*

1. INTRODUCTION

Credit risk is the risk of default on a debt due to a borrower failing to make required payments (on credit line, loan, mortgage, etc.). Duties of the borrower include everything, connected with the loan (credit): principals, coupons, interest, etc.

The existence of credit risk raises a number of consequences. At first the banking and financial industry, which is affected by this type of risk, in most countries is allowed to collect personal data about the borrowers. This data is used to build mathematical models for assessment of the credit risk. Moreover, changes in the behavior of financial companies arise, based on assessments of credit risk. These changes include restructuring of clients' accounts, portfolio diversification, activity specialization, changes in interest rates and individual interest levels to different customers, depending on their solvency. Given the number of corporate relationships that arise due to the presence of credit risk, that risk is essential for the success of the financial institutions.

Statistical models are commonly used in the banking industry in order to assess the credit risk associated with the approval of people applying for certain products (loans, credit cards, etc.). Based on data from the past, these models try to predict what will happen in the future (Montrichard 2008, p. 1). As a result of the built statistical model each candidate receives estimate in the form of points. The points of all candidates draw up a point system or the so-called "scorecard".

1.1. How do the scorecards work?

A relation between the information that each client submits at the time of application for a product and the behavior of that client after he has been approved for the product is made. A regression model is used to assess how the behavior of the approved customers (good or bad customers) depends on the data submitted in the application form for the product. The resulting relation can be used to assess new customers even in the time of application.

The purpose of the scorecard is to model the resulting mechanism. When the model is based on the data submitted in the application form, not on the repayment behavior, then this mechanism could be modeled as the probability the candidate to be good or bad (Mok 2009, p. 19).

The above statement is described mathematically as follows:

$$p = P(y = 1/x) = 1 - P(y = 0/x),$$

where x – a vector that contains the information from the application form; p – the probability a candidate with characteristic vector x to be good; $y=1$ – the event "the candidate is good"; $y=0$ – the event "the candidate is bad".

This work has studied the causal link between the conduct of an applicant upon payment of the loan (the dependent variable, *y*) and the data that he completed at the time of application (independent variables $x = (x_1,..., x_k)$). A linear regression is used to estimate the probability of being good for new borrowers, and a scorecard is obtained from the linear model to assess new customers in the time of application.

2. PROBLEM FORMULATION

A database of 100 borrowers from a commercial bank is used for the purposes of the study. They have applied for loans in the first quarter of 2015. The available data includes information from the time of application and credit history while paying off the loan. Customers are divided into three groups, based on the credit history: "Good", "Bad", "Indeterminate". If in the last 18 months a borrower has no more than one missed payment, he is "Good"; if he has two missed payments, he is "Indeterminate"; and if there are three or more missed payments, the candidate is "Bad". "Indeterminate" candidates most frequently are those for which there is insufficient credit history to determine their behavior (i.e. to be classified as "Good" or "Bad").

A linear regression is applied to the data. A variable named *Good Bad Flag (Y),* which contains the information mentioned above, is modeled as a dependent variable. As a result from the linear regression model, a scorecard is made. The weak points of the model are pointed. New iterations are made and final model is obtained, where all the coefficients are statistically significant and meet economic expectations. The scorecard from the final model could be used to assess new borrowers – each applicant could be passed through the scorecard and will receive certain points (formed on the basis of the available data for the person at the time of application). The bank could set a cut off level of points, bellow which to reject the applications and above it – to accept them.

Independent variables / indicators that are analyzed are presented in Table 1.

Indicators / Variables	Designation	Units of Measurement
Time with Bank	X1	Months
Time in Employment	X2	Months
Gross Annual Income	X3	Euro
Loan Amount	X4	Euro
Age of Applicant	X5	Years
Residential Status	X6	Category
Loan Payment Method	X7	Category
Occupation Code	X8	Category
Account Type	X9	Category
Marital Status	X10	Category

Table 1. List of independent variables / indicators

The indicators X1, X2, X3, X4 and X5 are quantitative, while X6, X7, X8, X9 and X10 are qualitative. The qualitative variables have the following categories:

- Residential Status (X6): H (Homeowner), L (Living with parents), T (Tenant), O (Other);
- Loan Payment Method (X7): B (Bank Payment), Q (Cheque), S (Standing Order), X (Not Given);
- Occupation Code (X8): P (Pensioner), B (Self-employed), M (Employee), O (Other);

- Account Type (X9): FL (Fixed Loan), VL (Variable Loan);
- Marital Status (X10): D (Divorced), M (Married), S (Single), W (Widow), Z (Not Given).

3. MODELLING

The present data analysis has been done by SPSS. Basic steps for statistical analysis in SPSS are presented by Goev (1996), Manov (2001), Pavlov & Mihova (2016), etc.

Dummy variables were created for each of the qualitative variables in order to quantify the qualitative data. Dummies are variables that take the value of 0 (for the absence of an attribute) and 1 (for its presence) and are used in the same way in regression analysis as the quantitative variables. The dummies correspond to the number of classes / categories of the qualitative variable - 1. Thus, the category for which no dummy variable is created, is used as a reference category (Gujarati 2005, p. 302).

Let's take for an example the variable *Residential Status (X6)*. The category "Homeowner" is selected as a reference, since there falls the largest percentage of the observations. No dummy was created for this category. For the remaining categories dummy variables were created as follows:

- *Residential Status Living With Parents (X6.1)* – set to 1 if the applicant lives with his parents and to 0 in all other cases;
- *Residential Status Tenant (X6.2)* – set to 1 if the applicant rents a house and to 0 in all other cases;
- *Residential Status Other (X6.3)* – takes value of 0 if the applicant is a homeowner, lives with parents or rents a house, and 1 in all other cases.

In order to decide which of the independent variables should be used in the model, a correlation analysis has been made.

Due to the presence of qualitative (nominal) variables, Cramer's V (Cramér 1946, p. 282) was used to measure the correlation. The existing correlations between the used indicators and the degree of their significance (** - significant correlation at P = 99%, * - significant correlation at P = 95%) have been established using the "everyone against everyone" method (Table 2).

	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Y	1,00										
X1	0,55**	1,00									
X2	0,43	0,44	1,00								
X3	0,74	0,77	0,82**	1,00							
X4	0,81*	0,73*	0,73	0,79**	1,00						
X5	0,60	0,57	0,64**	0,79*	0,72	1,00					
X6	0,39**	0,56**	0,43	0,89**	0,74	0,78**	1,00				
X7	0,40**	0,39	0,52**	0,88**	0,79*	0,71**	0,21	1,00			
X8	0,27*	0,40	0,34	0,81	0,67	0,70**	0,14	0,28	1,00		
X9	0,11	0,39	0,46	0,76	0,77	0,56	0,22	0,32*	0,11	1,00	
X10	0,31**	0,46	0,45	0,77	0,69	0,78**	0,61**	0,35**	0,31**	0,184	1,00

Table 2. The correlation matrix

Based on the results of the correlation analysis, some of the indicators are excluded from the model. X2, X3, X5 and X9 were excluded from the modelling process due to statistically insignificant correlation with the dependent variable Y. Due to statistically significant correlation between the independent variables that could lead to displacement of scores and hence wrong conclusions from the

model, X4 (strong correlation with X1 and X7) and X10 (moderate correlation with X6 and X8) are excluded from the modelling process.

A linear regression model was built with the rest variables (Y, X1, X6, X7 and X8). It has the following general form:

$$Y = b_0 + b_1X_1 + b_{6.1}X_{6.1} + b_{6.2}X_{6.2} + b_{6.3}X_{6.3} + b_{7.1}X_{7.1} + b_{7.2}X_{7.2} + b_{7.3}X_{7.3} + b_{8.1}X_{8.1} + b_{8.2}X_{8.2} + b_{8.3}X_{8.3}$$

The values of the coefficients in front of each of the indicators are listed in Table 3. The table shows also that more than a half of the coefficients are statistically insignificant at the significance level of 5% and even at 10%.

Indicator	Coefficient (b)	t-value	Significance level
Intercept	1,744	7,367	0,000
X6.1	-0,355	-0,479	0,633
X6.2	0,868	2,897	0,005
X6.3	0,483	2,222	0,029
X8.1	-0,180	-0,758	0,450
X8.2	-0,249	-1,347	0,181
X8.3	0,530	1,325	0,188
X1	-0,003	-1,812	0,073
X7.1	-0,232	-0,846	0,400
X7.2	-0,381	-0,689	0,493
X7.3	0,808	2,704	0,008

Table 3. Coefficients in the linear regression model and their significance

Remark: The indicators are sorted in the order they enter the model.

By replacing in the general form of the model with the relevant coefficients from Table 3 the following equation is obtained:

$$Y = 1{,}744 - 0{,}003X_1 - 0{,}355X_{6.1} + 0{,}868X_{6.2} + 0{,}483X_{6.3} - 0{,}232X_{7.1} - 0.381X_{7.2} + 0{,}808X_{7.3} - 0{,}180X_{8.1} + 0{,}249X_{8.2} + 0{,}530X_{8.3}$$

The correlation coefficient is 0.602, which means that there is a moderate correlation between the dependent variable and the indicators. The coefficient of determination is 0.363, which shows that only 36.3% of the change in the behavior of the applicant (the dependent variable Y) is determined by a change in the modeled independent variables (Gujarati 2005, Pavlov & Mihova 2016). However, the linear model adequately reflects the dependency of the variables, even at a level of significance of 1%, which can be seen from Table 4.

Variations	Sum of squares	Degrees of freedom	Sum of squares / Degrees of freedom	F-value	Significance level
ESS	26,841	10	2,684	5,066	0,000
RSS	47,159	89	0,530		
TSS	74,000	99			

Table 4. ANOVA table

As mentioned earlier, the model conducted above gives the probability the candidate to be a good payer. The Normal P-P Plot of regression standardized residual (Figure 1) shows the observed versus expected cumulative probability of being good. It could be seen that the expected one is the straight line and the points are what is really observed. The graphs of the expected and observed cumulative probabilities of being good are close to each other at their edges and are at a greater distance from one another in the middle.

The scorecard, corresponding to the considered model, can be seen in Table 5 (the coefficients are multiplied by 1000 for clarity).

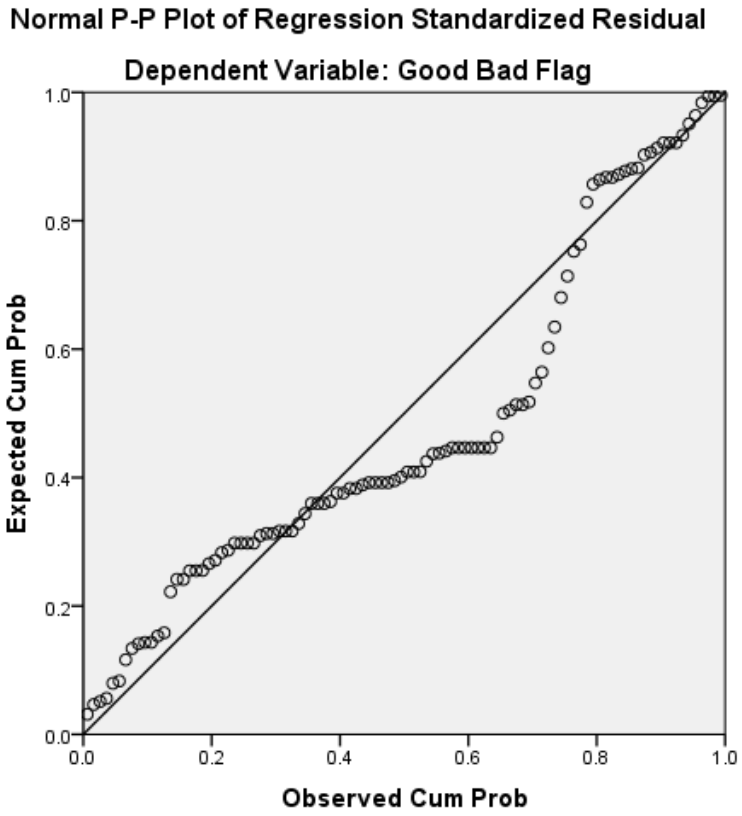

Figure 1. Normal P-P Plot of Regression Standardized Residual

Accordingly, each borrower can be passed through this scorecard system. In case the coefficients are statistically significant, the model could be interpreted as follows. Each candidate starts at 1744 points.

No.	Variable	Points
	Intercept	**1744**
1	**Residential Status (X6)**	
	Homeowner	**0**
	Living with Parents (X6.1)	**-355**
	Tenant (X6.2)	**868**
	Other status (X6.3)	**483**
2	**Occupation Code (X8)**	
	Pensioner (X8.1)	**-180**
	Self-employed (X8.2)	**-249**
	Employee	**0**
	Other occupation (X8.3)	**530**
3	**Time with Bank (X1)**	***(-3)**
4	**Loan Payment Method (X7)**	
	Bank Payment	**0**
	Cheque (X7.1)	**-232**
	Standing Order (X7.2)	**-381**
	Not Given (X7.3)	**808**

Table 5. The scorecard

For example, if the applicant lives with his parents, he loses 355 points. If he rents a house, he will get +868 points (the variable *Residential Status*). If the borrower is client of the bank for 12 months, he will receive -36 points (the variable *Time with Bank*, 12 * (- 3)).

In this case, due to the low coefficient of determination and statistically insignificant coefficients for most of the variables, it is better to make a new model. Moreover, it is noteworthy that some of the coefficients do not have the expected sign. For example, the longer a person is a client of the bank, the more points he should get. But the variable *Time with Bank* enters the scorecard with a coefficient of -3, which means that if an applicant is a client of the bank for one year, his score will be deducted with 12 * 3 = 36 points, and if he is a client for two years, his score will be deducted with 24 * 3 = 72 points. This does not meet the economic expectations and it is better to remove the variable from the model. Also, as it comes to the qualitative variables, it is worthy to consider adding to the reference group the categories "Not given" and "Other", covering candidates with no information given, and some of the rest categories – if the points they receive do not meet the economic expectations.

After removing *Time with Bank* from the model, several iterations have been tested with changing the reference groups of *Residential Status*, *Loan Payment Frequency* and *Occupation Code*, so that categories of these variables could enter the model with logical points. Also, the categories with statistically insignificant coefficients have been removed.

The values of the coefficients in front of each of the indicators in the final linear model are listed in Table 6. It can be seen that the coefficients are statistically significant at the significance level of 10%.

Indicator	Coefficient (b)	t-value	Significance level
Intercept	2,156	8,553	0,000
X6.2	0,762	2,528	0,013
X8.2	-0,310	-1,811	0,073
X7.0	-0,483	-1,733	0,086
X7.1	-0,976	-2,661	0,009

Table 6. Coefficients in the final linear regression model and their significance

By replacing in the general form of the model with the relevant coefficients from Table 6 the following equation is obtained:

$$Y = 2{,}156 + 0{,}762X_{6.2} - 0{,}483X_{7.0} - 0.976X_{7.1} - 0{,}310X_{8.2}$$

The linear model adequately reflects the dependency of the variables, even at a level of significance of 1%, which can be seen from Table 7.

Variations	Sum of squares	Degrees of freedom	Sum of squares / Degrees of freedom	F-value	Significance level
ESS	14,687	4	3,672	5,881	0,000
RSS	59,313	95	0,624		
TSS	74,000	99			

Table 7. ANOVA table for the final model

The correlation coefficient is 0.446, which means that there is a moderate correlation between the dependent variable and the indicators. The coefficient of determination is 0.198, which shows that only 19.8% of the change in the behavior of the applicant (the dependent variable Y) is determined by a change in the modeled independent variables (Gujarati 2005, Pavlov & Mihova 2016). The lower coefficient of determination (compared to the first model) is due to the fact that less variables figure in the final model.

The Normal P-P Plot of the regression standardized residual (Figure 2) shows the observed versus expected cumulative probability of being good for the final model. The trend looks like the one from Figure 1, but the fluctuations look like lines (i.e. constant levels of expected cumulative probability of being good for different values of the observed cumulative probability of being good). This is due to the fact that there is no quantitative data in the final model.

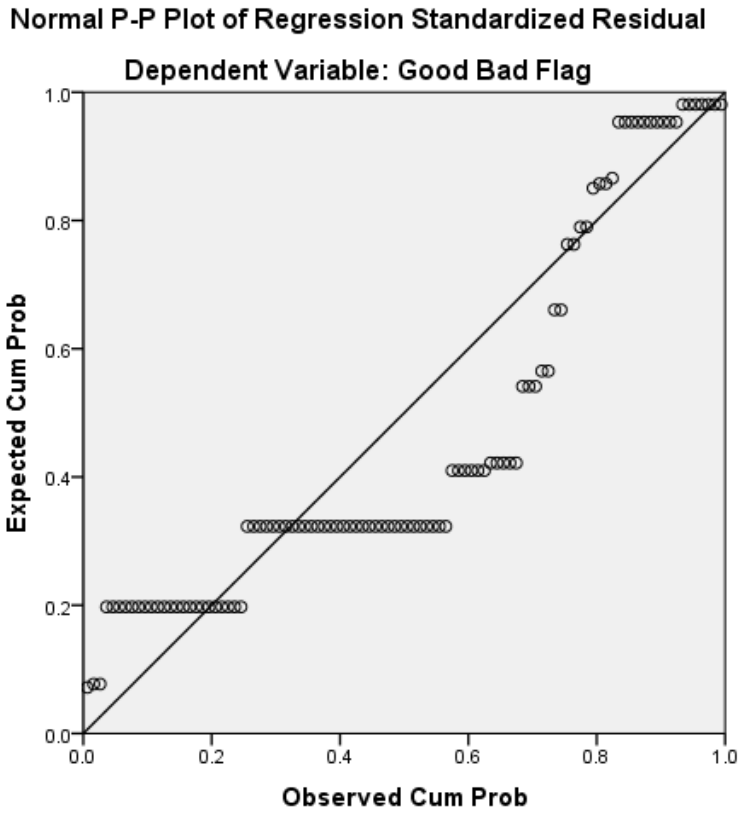

Figure 2. Normal P-P Plot of Final Regression Standardized Residual

The scorecard, corresponding to the final model, can be seen in Table 8 (again the coefficients are multiplied by 1000 for clarity).

The final model could be interpreted as follows. Each candidate starts at 2156 points. If the applicant rents a house, he will get +762 points; if he is a homeowner or lives with parents or something other, he will get no points on the variable *Residential Status*. The logic behind this is that people who live with their parents are not so independent; those who own a home, are not so motivated to have a good job by all means. For the category "other" there is no enough data to decide what kind of points could be expected – positive or negative. Another way to use the variable *Residential Status* is to give the homeowners the most positive points, as they are property owners and are more stable from this point of view. Unfortunately, when adding this category to the model, the homeowners get negative points and that's why in that case it is preferred to put them in the reference group.

If the applicant is self-employed, he will get -310 points; if he is an employee, pensioner or has other occupation, he will get zero points on the variable *Occupation Code*. The logic behind this is that the self-employed people are often very dependent on season, term work on projects, etc. As a matter of fact, the category pensioner has entered the model too - with negative points (which meets the expectations), but the coefficient for this category was insignificant and therefore it was excluded from the model.

If the applicant wants to pay the loan with cheque, his score will be deducted with 976 points. If he wants to pay with bank payment, his score will be deducted with 483 points. If he wants to pay with standing order or another way, he will get no points on the variable *Loan Payment Method*. Standing order is an instruction the payer gives to his bank to pay a set amount at regular intervals from one account to another. It may be taken as the best mentioned option, so the points on the variable *Loan Payment Method* meet the expectation ("Not given" are too small group to take any points).

No.	Variable	Points
	Intercept	**2156**
1	**Residential Status (X6)**	
	Homeowner (X6.0)	**0**
	Living with Parents (X6.1)	**0**
	Tenant (X6.2)	**762**
	Other status (X6.3)	**0**
2	**Occupation Code (X8)**	
	Employee (X8.0)	**0**
	Pensioner (X8.1)	**0**
	Self-employed (X8.2)	**-310**
	Other occupation (X8.3)	**0**
3	**Loan Payment Method (X7)**	
	Bank Payment (X7.0)	**-483**
	Cheque (X7.1)	**-976**
	Standing Order (X7.2)	**0**
	Not Given (X7.3)	**0**

Table 8. The final scorecard

4. CONCLUSION

The relationship between the information that each borrower submitted at the time of application for a product and his behavior, after he has been approved for the product, can be used to estimate the probability of being good for new customers. Such an assessment is made in this work - based on a linear regression model and using the scorecard from this model. The correlations between the variables should be taken into consideration during the modeling, as well as the economic logic of the coefficients of the variables. Taking into account the last mentioned considerations, the weak points of the model are pointed. New iterations are made and final model is obtained, where all the coefficients are statistically significant and meet economic expectations, although the correlation coefficient is very low. The presented algorithm could be used for other data in a similar situation.

REFERENCES

Cramér, H 1946, *Mathematical Methods of Statistics*, Princeton University Press, Princeton.

Goev, V 1996, *Statistical Processing and Data Analysis from Sociological, Marketing and Political Studies with SPSS (Bulgarian)*, Sofia.

Gujarati, DN 2005, *Basic Econometrics*, Fourth edition, McGraw Hill, New York.

Manov, A 2001, *Statistics with SPSS (Bulgarian)*, Trakia-M, Sofia.

Montrichard, D 2008, 'Reject Inference Methodologies in Credit Risk Modeling', *SESUG 2008: The Proceedings of the SouthEast SAS Users Group*, St Pete Beach, FL, <http://analytics.ncsu.edu/sesug/2008/ST-160.pdf>

Mok, JM 2009, 'Reject Inference in Credit Scoring', *Amsterdam: BMI paper*, <https://www.few.vu.nl/nl/Images/werkstuk-mok_tcm243-91398.pdf>

Pavlov, VT & Mihova, VM 2016, *Applied Statistics with SPSS (Bulgarian)*, Avangard Print, Ruse.