ESGI99 report: Credit scorecard for Corporate Clients based on industries (Problem posed by the company: KOMERCIJALNA BANKA)

Stefana Janicijevic* Zorana Luzanin* Sergiy Pereverzyev†
Irena Stojkovska‡ Andreja Tepavcevic*

February 24, 2014

1 Introduction

Komercijalna Banka is the largest domestic bank in Serbia and has a huge number of clients. Among corporate clients, there are companies of each size: small, medium and large enterprizes, and from different industry sectors: agriculture, mining, manufacturing, trade, transportation etc.

When a client apply for a loan in the bank, the bank is supposed to predict its creditability, to estimate the level of risk associated with this new applicant for a loan. A credit risk scoring is also conducted for existing clients. The bank uses credit scorecards to estimate the probability that a client will display a defined behavior (e.g. loan default) with respect to its current position, [3, 6, 9]. The client's credit is defined as default if he is more than 90 days in delay of repaying its credit i.e. we define the default as

$$\text{default } (Y) = \begin{cases} 0, & \text{if the number of days in delay} \leq 90 \\ 1, & \text{if the number of days in delay} > 90 \end{cases}. \tag{1}$$

Current scorecard, that the Bank uses, is based on both qualitative and quantitative client's characteristics obtained by different kind of available data sources (the bank core system, balance sheets, questionnaires, accounts managers opinion). The Bank also noticed the difference in clients capability in repaying their loan, that might depend on the industry. The main questions that were asked by the Bank are:

*University of Novi Sad, Serbia, e-mail: {stefana@turing.mi.sanu.ac.rs, zorana@dmi.uns.ac.rs, andreja@dmi.uns.ac.rs}
†University of Innsbruck, Austria, e-mail: {Sergiy.Pereverzyev@uibk.ac.at}
‡Ss. Cyril and Methodius University, Macedonia, e-mail: {irenatra@pmf.ukim.mk}

- How to "put" information about the industry in the credit scorecard? Is it possible to group industries and give a separate credit scorecards? What if there is a small number of clients in some industry?

- How to deal with missing data for the client, when calculating the score according to the credit scorecard?

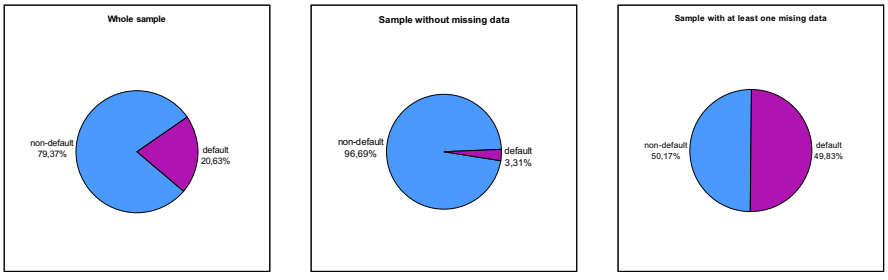
The current credit scorecard needs improvement in a way that it will not reject "good" clients, and it will not approve a loan to a "bad" one. The adequate mathematical model is needed.

2 Preprocessing data

The Bank has data for each client obtained from different sources, the income statement and the balance sheet for the last two years, data from questionnaires and other data. Original sample S had size of 2351 clients, in which 1866 (79.37%) clients are not in default and 485 (20.63%) clients are in default (Figure 1 a)). After removing the missing data, the sample S_1 without missing data has size of 1481 clients, in which 1432 (96.69%) clients are not in default and 49 (3.31%) clients are in default (Figure 1 b)). We had also analyzed the remaining sample S_2 with at least one missing data, the sample size is 870 clients, in which 434 (49.89%) clients are not in default and 436 (50.11%) clients are in default (Figure 1 c)). Then the conditional probability that the client has some missing data, if he is in default is

$$P\{\text{missing data} \mid \text{default}\} = 89.90\%, \tag{2}$$

which shows that the missing data in clients' portfolios, is very big issue for the Bank in the moment. We propose the following way in dealing with the missing data: if there is a small number of missing information about the client, then replace the missing data with the average value from the clients that have similar values for other characteristics.



a) b) c)

Figure 1: Pie charts for samples S, S_1, S_2

Since the inclusion of qualitative variables improves the prediction power of models, [8], we consider both: categoric and numeric input variables (trend of income, maturity adjustment, cash flow, net profit ratio, business profit ratio, business prior period, number of owners, seasonality, etc.). We labeled the

variables by $x_1, x_2, ..., x_{32}$. According to descriptive statistics done with STATISTICA, none of the input numerics variables are normally distributed, so nonparametric tests are supposed to be done.

STATISTICA Median Test and Krushkal-Wallis Test showed that there is no significant difference in default among industry sectors ($p = 0.2504 > 0.05$ and $p = 0.2511 > 0.05$ respectively), and neither there is a significant difference in default among industry branches ($p = 0.9919 > 0.05$ and $p = 0.9921 > 0.05$ respectively). So, the influence of industry seems to be neglectable, taking into account the available data.

Descriptive statistics detected a lot of outliers for each input variable. After removing all outliers from the sample S_1, the sample S_3 without missing data and without outliers has size of 427, in which 426 (99.77%) clients are not in default and only 1 client (0.23%) is in default. So, we decided to proceed with the sample S_1 on which we performed a sampling procedure to obtain a training sample, and a test sample.

3 Mathematical modeling

There are different techniques used for modeling the credit scorecards, both statistical and operational research based, [3, 4, 9]. Scoring methods indicates how sensitivity a likelihood function $L(\theta, x)$ depends on its parameter θ. The score for θ is gradient of the log likelihood with respect to θ, i.e.

$$S = S(\theta, x) = \frac{\partial \log L(\theta, x)}{\partial \theta} = \frac{1}{L} \cdot \frac{\partial L(\theta, x)}{\partial \theta}. \tag{3}$$

Log likelihood ratio test compare the fit of two models. The objective is to assign credit applicants to one of two groups: a "well credit group" that is likely to repay the financial obligation or a "default credit group" that should be denied credit because of a high likelihood of risk propensity. We discuss several credit score approaches and statistical methods - Bayesian decision model, neural networks, PCA, discriminant analysis, fuzzy logic, etc.

We decided to use two approaches where we expected the best results: logistic regression (LR) and support vector machines (SVM).

3.1 Logistic regression (LR)

Other mentioned statistical methods could be also used for scoring and decision making in this process, but we decided to use logistic regression because of the classical binary representation of the target dependent variable. The main reason is that the Bank required that the target variables should be in relation with the time repaying with the cut off value of 90 days, see formula (1).

Logistic regression [5] (logit regression) is a probabilistic statistical classification method that is used to predict a binary response used for predicting the outcome of a categorical dependent variable that is based on one or more predictor variables. The possible outcomes are modeled by probabilities, as a function of the predictor variables, using a logistic function.

Logistic regression gives a relationship between a categorical dependent variable and independent variables by using probability scores as the predicted values of the dependent variable. Binomial or binary logistic regression deals with

situations in which the observed outcome for a dependent variable has two possible types. In multinomial logistic regression, the outcome can have three or more possible types. Here we use binomial logistic regression, since we have two possible outcomes (default and not default).

Logistic regression use one or more predictor variables that may be either continuous or categorical data and it is used for predicting binary outcomes of the dependent variable.

Unlike others, the logistic model does not require multivariate normality or the equality of covariance matrices of two populations. In this scenario there are two types of customers - those who will repay the loan and those who will not repay. By that, we classify customers at the term of two way classification. Using ideal scheme loan officer place customer into right category. Using binary logistic regression of the credit approval process we develop criteria from the bank perspective.

Let us denote the probability for client's default based on input data x by $p(x) = P\{Y = 1 \mid x\}$. Then the logistic function can be expressed as follows:

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \beta^T x, \tag{4}$$

i.e.,

$$p(x) = \frac{1}{1+e^{-\beta^T x}} = \frac{e^{\beta^T x}}{1+e^{\beta^T x}}. \tag{5}$$

In practical implementation of the logistic regression technique, we first choose a cut off value $p_c \in (0,1)$, called *threshold probability*, then we estimate the probability for client's default $p(x)$, and if it is less than the threshold probability, we classify the client as a "good" client, the one that will be able to repay the loan.

Decision procedure:

1. Set the cut off value p_c.
2. If $p(x) < p_c$, then give a loan.
3. If $p(x) \geq p_c$, then do not give a loan.

3.1.1 Practical implementation

Logistic regression model (4) was fitted with the IBM SPSS v20 and STATISTICA 12 software on PC machines.

The LR method computes the maximum likelihood estimators of the $n+1$ parameters by an iterative least squares algorithm. The stepwise procedure is performed in order to select the most significant variables, as well as forward and backward procedures. Stepwise logistic regression finds the most parsimonious set of predictors that are most effective in predicting. Variables are added to the logistic regression equation one at a time using statistical criterion of reducing the included variables. After accepting the variables for the model, another testing is performed. Forward LR is used for a large number of explanatory variables. Variables are entered one at a time, at each step adding the predictor with the largest score, whose significance value is less than 0.05. Variables are removed based on the likelihood ratio test.

LR can be fully embedded in a formal decision framework, but in order to perform a comparison with other models, a threshold probability must be specified. The value 0.5 was chosen for the threshold probability.

In order to select most appropriate variables for the logistic functions, we processed 5 random samples (of size 100 each).

We have selected variables for the logistic functions using STATISTICA Spearman rank R, Gamma and Kendall tau correlation tests and SPSS Forward conditional method.

Finally, as the most appropriate, we selected the following independent variables: $x_1 = B, x_{12} = D, x_{16} = E$.

After testing proposed model on the five random samples, several combinations of coefficients were obtained, with the accuracies given in the table below.

If $S > 0$ after inserting the related values in formulas below, then we estimate that the client with these parameters would be in default, and if $S \leq 0$, then the client would not be in default.

1. $S = -0.037 * B + 0.049 * D + 0.382 * E - 0.146$
2. $S = -0.039 * B + 0.041 * D + 0.131 * E + 0.805$
3. $S = -0.042 * B + 0.035 * D + 0.064 * E + 1.407$
4. $S = -0.035 * B + 0.042 * D + 1.331 * E + 0.158$
5. $S = -0.038 * B + 0.046 * D + 0.339 * E + 0.099$

Finally, we accepted an approximation of the formula, that has approximately the same accuracy as the others five. We decided to use the final formula, because it can be transformed in a way to be of practical usage:

$$S = -0.04 * B + 0.04 * D + 0.4 * E + 1 \tag{6}$$

We provide an equivalent relation for "default":

$$B - D - 10 * E < 25 \tag{7}$$

and for not "default":

$$B - D - 10 * E \geq 25 \tag{8}$$

In the Table 1, the accuracies of five models and the finial model are given.

We proposed the final formula as the most convenient one, to be used for the credit scorecard.

Therefore, an estimation using the obtained data and the proposed model is that there are 0.7% chances that the loan would be granted and not repaid.

The estimate for the probability for client's default, based on the logistic regression approach using the last model, is:

$$p(x) = \frac{1}{1 + e^{-(-0.04*B+0.04*D+0.4*E+1)}}. \tag{9}$$

Table 1: Accuracies of five samples and the final model.

MODEL	correctly predicted (Observed=Predicted)		Observed=0 (not default) Predicted=1 (default)		Observed= 1 (default) Predicted= 0 (not default)	
sample 1	1180	79.7%	291	19.6%	10	0.7%
sample 2	1152	77.8%	319	21.5%	9	0.6%
sample 3	1204	81.3%	264	17.8%	12	0.8%
sample 4	1113	75.2%	360	24.3%	8	0.5%
sample 5	1198	80.9%	273	18.4%	10	0.7%
FINAL MODEL	1112	75.1%	358	24.2%	10	0.7%

3.2 Support vector machines (SVM)

The problem of assigning the credit score for a bank client can be viewed as a multiclass classification problem in the Statistical Machine Learning Theory (SMLT) [2], where clients with the same score are considered as a class.

The number of days in delay of repaying the credit can be used in defining the credit score and the corresponding class of clients. Indeed, let i be the identification number of a client, and N_i be its number of days in delay. Assume that one decided to use 5 possible scores for clients: 1,2,...,5, where 1 should be assigned to the best clients, and 5 — to the worst, for example to those who could be in default. Then, by setting the delay levels N^1, N^2, N^3, N^4, we could define the class of clients with the score 1 as $C^1 := \{\, i \mid N_i < N^1 \,\}$; with score 2 as $C^2 := \{\, i \mid N^1 \le N_i < N^2 \,\}$; and so on. Now, the problem of the credit score assignment is to predict the class of a client using the values of its characteristic variables mentioned in Section 2.

The multiclass classification problem can be considered as a sequence of binary classification problems [2]. Therefore, we consider here the problem of assigning a client to one of the classes

$$C^1 := \{\, i \mid N_i < N^1 \,\} \quad \text{or} \quad C^2 := \{\, i \mid N_i \ge N^1 \,\}. \tag{10}$$

In SMLT, SVM is a well-known method for the binary classification. Let us outline its application in this context.

Consider the sample S_1 from Section 2 with 1400 clients (random 81 clients were taken out). Each client $i \in S_1$ has a vector $\bar{x}_i = (x_{1,i}, x_{2,i}, \ldots)$ with the values of its characteristic variables. Let us take a training subsample $S_{1,\text{train}} \subset S_1$ with 100 randomly chosen clients. For each client in the training subsample $S_{1,\text{train}}$, we know its score $y_i \in \{\, 1, -1 \,\}$, where $y_i = 1$ if $i \in C^1$, and $y_i = -1$ if $i \in C^2$. Then, in SVM, using the training data $\{\, (\bar{x}_i, y_i) \mid i \in S_{1,\text{train}} \,\}$, one constructs a decision function $f(\bar{x})$ such that the training data is well represented, that is

$$f(\bar{x}_i) = y_i \tag{11}$$

for as many $i \in S_{1,\text{train}}$ as possible. The hope is that this decision function $f(\bar{x})$ would also predict well the score of other clients $i \in S_{1,\text{test}} := S_1 \setminus S_{1,\text{train}}$.

There are several choices of the form of the decision function. It can be a function of the linear form

$$f(\bar{x}) = \text{sign}(\, \bar{w} \cdot \bar{x} + b \,), \tag{12}$$

where $\bar{w}$ is a weights vector. This choice could give a good prediction if the data is expected to be linearly separable. This can not be expected to be the case in general; therefore, the nonlinear functions of the form

$$f(\bar{x}) = \text{sign}\left(\sum_{i=1}^{m} c_i K(\bar{x}_i, \bar{x}) \ + b \right) \tag{13}$$

are used. Here, m is the size of the training set, i.e. $m = \#(S_{1,\text{train}})$, and $K(\bar{x}_i, \bar{x})$ is a kernel function [10]. The frequent choice of K is the so-called Gaussian radial basis function (RBF) $K(\bar{x}_i, \bar{x}) = \exp(\ -\|\bar{x}_i - \bar{x}\|\ )$.

It should be noted that even if the form of the decision function is specified, there can be several decision functions that satisfy the condition (11). In the SVM method, the decision function is selected such that the so-called margin band is maximized [2].

For the numerical results below, we used the MATLAB realization of the SVM method in the Statistics Toolbox. We observed that the decision functions (13) with RBF give better results than the linear decision functions (12). We choose $N^1 = 10$ in (10). With such a choice the classes C^1 and C^2 have approximately the same number of clients.

We considered the following two measures for describing the accuracy of the decision function f:

- description accuracy (DA):

$$100\% \frac{1}{\#(\text{Train})} \sum_{i \in \text{Train}} (\ 1 - |y_i - f(x_i)|\ ),$$

  where $'\text{Train}'$ is the training set;

- prediction accuracy (PA):

$$100\% \frac{1}{\#(\text{Test})} \sum_{i \in \text{Test}} (\ 1 - |y_i - f(x_i)|\ ),$$

  where $'\text{Test}'$ is the test set.

Unless it is stated otherwise, in the examples below, the description accuracy is 100%.

With the training set $S_{1,\text{train}}$, for the test set $S_{1,\text{test}}$, we obtained the prediction accuracy 52%. It should be noted that this accuracy on some subsets of $S_{1,\text{test}}$ is considerably higher. This indicates that the sample S_1 contains clients with considerably different characteristics and behavior, which requires different decision functions for different subsets of clients.

Indeed, consider subsets $S_1^{(1)}, S_1^{(2)}, S_1^{(3)}, S_1^{(4)}$ of the sample S_1 that contains clients in 4 biggest industry sectors. For each subset $S_1^{(i)}$, we construct the decision function using SVM based on the corresponding training subsample $S_{1,\text{train}}^{(i)} \subset S_1^{(i)}$. Then, as one sees in Table 2, the prediction accuracy increases.

Also, as one sees in Table 3, the size of the training set has a considerable influence on the prediction accuracy.

Finally, we would like to note the influence of the set of the client's characteristic variables on the accuracy. So far, we considered the full set of the characteristic variables mentioned in Section 2 (v.set1). We tested also two subsets of this set:

Table 2: Prediction accuracy on the sample S_1 and on its subsets $S_1^{(i)}$ corresponding to the biggest industry sectors.

sample	size	#(Train)	PA
S_1	1400	100	52
$S_1^{(1)}$	500	100	54
$S_1^{(2)}$	300	100	54
$S_1^{(3)}$	120	60	58
$S_1^{(4)}$	90	45	58

Table 3: Influence of the size of the training set on the prediction accuracy.

sample	size	#(Train)	PA
S_1	1400	100	52
S_1	1400	50	54
$S_1^{(2)}$	300	100	54
$S_1^{(2)}$	300	50	57

- set of only numeric (not categoric) variables (v.set2);
- set of 3 variables taken in the LR method (v.set3).

Table 4: Influence of the set of characteristic variables on the accuracy.

v.set	DA	PA
1	100	52
2	83	54
3	64	56

We considered the full sample S_1, and the size of the training set was taken to be 100. As one can see in Table 4, although the description accuracy decreases as the set of variables becomes smaller, the prediction accuracy increases. These results suggest that one should not try to describe the data as good as possible. In fact, a very high description accuracy arises in the so-called over-fitting phenomenon. In SMLT, it is known that this phenomenon should be often avoided.

However, the standard realization of the SVM method designs the decision function that should describe the training data as close as possible. So, it is interesting to consider the modifications of the SVM method that allow flexibility in describing the training data.

4 Future work

Ideas for future research with this kind of data might be non linear and non parametric regression as well as metaheuristic methods for global optimization. Objective function can be binary 0-1 scoring with linear constraint which represents the cut off delay.

Other types of methods that can be used are based on neural networks [1] and fuzzy logic approach [7] (fuzzy classifiers).

As further methods, we can establish evaluation criteria: we can consider an area under the receiver operating curve (AUC) like performance measure of each model. The AUC can be computed with the aid of the ROCR method. The prior probabilities and the misclassification costs should also be considered. Cost associated with a Type I error (a client with good credit is misclassified as a client with bad credit) and a Type II error (a client with bad credit is misclassified as a client with good credit) are usually very different. the expected misclassification cost (EMC) is defined as follows:

$$EMC = c_{21}P_{21}\pi_1 + c_{12}P_{12}\pi_2, \tag{14}$$

where π_1 and π_2 are the prior probabilities of good and bad credit populations, P_{21} and P_{12} are measures, the probability of making Type I errors and Type II errors. These parameters can be estimated using the proportion of clients with good and bad credits.

The results obtained by the SVM method suggest that good decision functions can be designed for small subsets of clients. An appropriate selection of these subsets is an interesting issue for the future research. Also, the selection of the client's characteristic variables should be studied in detail.

Finally, as it was already mentioned, it is worthwhile to consider modifications of the SVM method that allow flexibility in the design of the decision function.

5 Conclusion

We proposed two new solutions for credit scorecard, based on the logistic regression statistical analysis and support vector machines tool. With statistical testing we find no significant difference in default among industry sectors, neither among industry branches. We showed that the missing data in clients' portfolios, is very big issue for the Bank in the moment, so we propose the following way in dealing with the missing data: if there is a small number of missing information about the client, then replace the missing data with the average value from the clients that has similar values for other characteristics.

References

[1] A. Blanco, R. Pino-Mejias, J. Lara, S. Rayo, *Credit scoring models for the microfinance industry using neural networks: Evidence from Peru,* Expert Systems with Applications 40 (2013) 356-364.

[2] C. Campbell, *An Introduction to Kernel Methods.* Radial Basis Function Networks: Design and Applications. Eds.: R. J. Howlett and L. C. Jain, pp. 155 – 192, Springer, 2000.

[3] M-D. Cubiles-De-La-Vega, A. Blanko-Oliver, R. Pino-Mejías, J. Lara-Rubio, *Improving the management of microfinance institutions by using scoring models based on Statistical Learning techniques,* Expert Systems with Applications 40 (2013) 6910–6917.

[4] Yu-Chiang Hu, J. Ansell, *Measuring retail company performance using credit scoring techniques,* European Journal of Operational Research 183 (2007) 1595-1606.

[5] D. G. Kleinbaum, M. Klein, *Logistic Regression, A Self-Learning Text*, 3rd Edition, Springer, 2010.

[6] K. D. Majeske, T. W. Lauer, *The bank loan approval decision from multiple perspectives,* Expert Systems with Applications, 40 (2013), 1591–1598.

[7] P. Pardalos, G. Baourakis, *Fuzzy Sets in Management, Economics and Marketing.* Singapore; World Scientific Publishing Co. 2001.

[8] M. Schreiner, *Scoring arreas at a microlender in Bolivia,* Journal of Microfinance, 6 (2004) 65–88.

[9] L. C. Thomas, *A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers,* International Journal of Forecasting 16 (2000) 147–172.

[10] Wikipedia, *Positive-definite kernel,* The Free Encyclopedia, 2014.