Construction and application of scoring models

Vladěna Obrová[1]

Abstract. This paper focuses on the definition and practical application of scoring models that are used not only in the bank sector, but also by other institutions to assess whether the client can get a loan or not. If there is an incorrect decision due to wrong definition in this model then the institution is exposed to risk of loss. Therefore when any application for the loan is posted, institution needs to do the best quantification of the risks. Only after that is able to decide, on the basis of available information, about the applicant. There will be used logistic regression method in the process of constructing the models in this paper. The output of this method has to estimate the conditional probability of client repayment of the loan with given characteristics. One of the most studied characters of the scoring model is the ability of diversification represented by the Lorenz curve and quantified by Gini coefficient.

Keywords: scoring model, Gini coefficient, Lorenz curve, credit risk.

JEL Classification: C190, G210
AMS Classification: 62P20

1 History

The history of credit scoring only started some 50 to 70 years ago. In 1936 Ronald Aylmer Fischer for the first time approached the issue of identification of groups within the population. In 1941 David Duran for the first time recognised that Fischer ´s techniques might be used for distinction between good and bad loans. In 1950s Bill Fair and Earl Isaac established the first advisory office. Considerable progress of credit scoring started with the launch of credit cards in 1960s and growth of information technologies. In 1980s a reliable method of logistic regression and linear programming was introduced. Recently expert systems and neuron networks began to be used like in other risk management areas apart from scoring models. There are also other developing methods such as the approximation function, the Bayesian method, the classification trees, the genetic algorithms and others. [2]

2 Scoring model – assumptions

The following article deals with construction of a scoring model. To be able to discuss the issue I need to define the basic terms. The most important term related to scoring models requiring definition is risk. The term risk is connected with 2 more terms: The first is **uncertain result**, which means that there must be at least two variants of the solution for if you know for sure that a loss will be incurred then you cannot speak of a risk. The second is **undesirability** of at least one of the two possible results. Risk is often understood as the danger of occurrence of a certain loss. [9] Theory of finance usually defines as a risk volatility of a financial quantity (portfolio value, earnings etc.) around the expected value as a consequence of changes of numerous parameters.

2.1 Risk classification – credit risk

The notion of scoring model is related not only to risk but also with classification of risks. There are financial and non-financial risks depending on whether the asserted risk factor causes financial loss or not. In the case of the scoring model the risk will always be the financial risk. Financial risk comprises the relation between a subject (individual or organisation) and assets or expected income which may be lost or deteriorated.

Financial risk is usually affected by three factors:

- subject exposed to the risk of loss;
- assets or income whose value reduction, destruction or change of ownership cause the financial loss;
- danger (threat) that may cause the loss. [9]

[1] BUT, Faculty of Business and Management, Kolejní 2906/4, Brno, 612 00, obrova@fbm.vubtr.cz

Financial risks can be divided **to credit risk, operation risk, liquidity risk** and **market risk**. The terms will be defined below, with special focus on credit risks. Operation risk is connected with operation drawbacks or errors causing loss. This risk is defined by the Bank for International Payments with registered offices in Basel as the risk of direct or indirect losses caused by inappropriate or unsuccessful internal processes, employees or systems or external events. [11]

The liquidity risk is related to the ability of the company in question to cater for all its due liabilities at any moment, which in the case of a financial institution means to be able to pay out the due deposits of its clients at any time and in the required form. Market risk follows from changes of market prices and their impact on earnings (equity value) of the company.

The last financial risk is **credit risk,** which ranks among the basic financial risks for loan granting is common and recently represents one of the most frequently discussed themes in all economic areas. The risk is also called loan risk and as one of the current trends is a trend towards multiple loans acquired by clients, the other party (the bank, the company) needs to define certain parameters to assess whether the client will be able to fully refund the loan within the agreed maturity deadline. The following table shows history of loans in the Czech Republic:

SELECTED LOAN TYPES	ABSOLUTE AMOUNT (CZK BILLION) as of 31 March 2012	YEAR-ON-YEAR CHANGE as of 31 March 2012
Total volume of loans	2304.5	5.72%
Loans to non-financial institutions	832.2	5.14%
Short-term (up to 1 year inclusive)	266.2	7.62%
Medium term (1 – 5 years)	150.3	3.60%
Long-term (over 5 years)	415.6	4.16%
Retail loans	1012.5	4.79%
Debit balances of current accounts	12.7	-1.09%
Liabilities from credit cards	24.7	1.81%
Consumer loans	157.4	-2.34%
Housing loans	777.4	6.48%
Other	40.2	6.53%
Loans to traders	37.3	-6.36%

Table 1 History of loans [10]

At present the conditions for loan provision become more and more stringent for the reason of past non-payments by clients. These procedures follow from historic data. Credit risk is one of the most frequent risks and that is why the banks/financial institutions pay a lot of attention to it.

The important indicators checked by the bank before providing a loan to the client include the client ´s creditworthiness, on the basis of which the bank specifies the amount of the loan and the conditions under which the loan may be granted to the client in question. The amount of net income reduced by other loan, overdraft or credit card repayments determines the amount of the mortgage loan to be provided to the applicant. For the financial institution to be able to decide whether to grant a loan to the given client or not the client creditworthiness must be complemented with assessment of other factors as well. Scoring models are used for this purpose of parameter benchmarking and client assessment in practice. The scoring is expressed by a real number, the score. You may also say that the client ´s score means numerical expression of financial reliability of the client. Higher scores are granted by the model to clients with higher creditworthiness while clients with low creditworthiness receive low scores. The higher score - the lower risk of the loan non-repayment. Thanks to imperfections sometimes a good client receives a lower score than a risky client. [3]

2.2 Credit scoring

Scoring model is then a specialised risk rate trying to “squeeze” all risks of an institution/client into a single number. Credit scoring focuses on which clients will not repay the loan and which clients will not pay for provided services. I will continue to deal with just the credit scoring, also used for detection of fraud, on loan application approval/rejection or pre-approval.

Methods of **credit scoring** form a standard part of risk management by financial institutions. On the basis of this model each applicant for a loan receives a score. The score means assessment of the client. The higher the score than better is the client. Sometimes the score is represented by a probability estimate whether the given client is to repay the loan or not. On the basis of the score the institution decides about the conditions under which the loan will be provided. In the case of new clients there is the so called application scoring on the basis of which the bank decides whether to provide a loan to the applying client at all. **Application scoring** models are used by loan institutions to evaluate creditworthiness of potential clients applying for credit product. The aim of scoring models is to classify applicants into two groups: the ones who will not default and the ones who will default. [6] The problem of the application scoring is that the quantity of the assessed clients is different from the quantity of the clients to whom the loan has actually been provided and who become the basis of the application scoring model.

Current clients of the bank are assessed on the basis of the so called behavioural credit scoring allocating a score to all clients, not only to the clients currently applying for a loan or service. As follows from the name of the scoring this model is based on behaviour of the clients of the assessing institution. On the basis of this score selected clients are sent marketing offers of loans or increased overdraft as they have been assessed by the system as clients able to repay higher loans of any kind. In 2007 has Goran Klepac introduced a new methodology of temporal influence measurement (seasonal oscillations, temporal patterns) for behavioural scoring development purposes. His work shows how significant temporal variables can be recognised and then integrated into the behavioural scoring models in order to improve model performance. [5]

Application of the scoring is conditioned by availability of a large volume of homogeneous data, which is one of the reasons why these models have been used by the bigger organizations for unified loans (credit cards, instalments). An important thing is that the scoring model is not to explain the risk, but just to predict it.

Scoring model is usually based on the database of the existing clients to whom a loan has been provided, together with information which clients managed to repay the loan in full. For the sake of simplicity a good client is a client who repaid the loan in full and in time and under the agreed conditions, while a bad client is a client who failed to meet one of its liabilities. Non-repayment is often also called default.

3 Scoring model - construction

The model construction will be based on logistic regression, the difference between logistic regression and the linear regression is, that we did not try to estimate a value of explained variable, but the probability, that the object belong to the one category. [8]

The outputs of the logistic regression method include an estimate of conditioned probability of repayment by the client with the given characteristics. One of the most important analysed features of scoring model is the ability to diversify, i.e. the good/bad client resolution. This feature is expressed in practice by Lorenz curve and quantified by Gini coefficient.

The model construction will be based on logistic regression. The outputs of the logistic regression method include an estimate of conditioned probability of repayment by the client with the given characteristics. One of the most important analysed features of scoring model is the ability to diversify, i.e. the good/bad client resolution. This feature is expressed in practice by Lorenz curve and quantified by Gini coefficient.

Where the scoring model construction is based on the logistic regression model, the simplest way is to start from the binary explained variable Y, which is depends on the explaining variables x_i'. The Y variable, with probability π, acquires value 1, and with probability $(1 - \pi)$, value 0. There is also the explaining variable vector [4]

$$x_i^{'} = [x_{i1}, x_{i2} \dots, x_{ik}\ ], i = 1{,}2 \dots n \tag{1}$$

representing *i-th* combination of the explaining variables $\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{k}$, $X_1, X_2, \dots, X_k$, and then $i - th$ conditioned classification of the Y quantity is alternative [10] with parameter written as $Y \sim Alt(\pi)$ (and median value of the Y quantity, $\mathrm{E}(\mathrm{Y_i}) = \pi_\mathrm{i})$ π_i a pravděpodobnostní funkcí,

$$P(y_i|\pi_i) = {\pi_i}^{y_i}(1 - \pi_i)^{1-y_i} \tag{2}$$

When selecting variables for vector x this probability function is used if for different vectors of x_i values of quantities $X_1, X_2, \dots, X_k$ the conditioned probability ratio of quantity Y (given by parameter π_i) s is the same. Then the Y quantity does not depend on these variables. But if different combinations of values result in different probability ratios π_i, then a certain type of relation between Y and these explaining variables can be assumed and their illustration by the regression model.

3.1 Explaining variables

The explaining variables are expressed with the help of the "score table". The explaining variables may be both quantitative and qualitative. Quantitative variables may be expressed in numbers and are subdivided to discrete (number of client ´s children, number of persons living in the same household as the client) and continuous (monthly income, monthly expenditures). However, most variables are qualitative and cannot be expressed in numbers, such as the highest achieved education, marital status, refunding of previous loans. The score table may include other variables in addition to ownership status, age, loan purpose, and litigation costs. Most variables included in the score table are clearly linked to the risk of non-repayment of the loan. Some variables give an ideal about the client ´s stability. These for example include the length of residence on the current address, or the time with the current employer. Other variables define financial sophistication of the client, such as whether the client has a checking account or a savings account, if h owns any credit cards, and how long the client has been with the current bank. Other variables show the consumer ´s resources. These include ownership status, employment, number of children etc. [3]

These variables are divided into groups by their meaning (such as gender, age, income etc.). Each group consists of several categories. It is further assumed that each client belongs to a single category within each group. To be able to introduce these variables in the model one needs to allocate dummy (binary) variables to them acquiring the value of 0 if the subject does not belong to the category or 1 if the subject falls within the category. Then each binary variable x_j^i of the x set indicates whether the client belongs to the respective $j-th$ category of $i-th$ group ($x_j^i = 1$) či nikoliv ($x_j^i = 0$).

3.2 Logit

There is looked for the relation between π_i and the values of vector x on the basis of the assumption that Y is a binary variable acquiring values 0 or 1 only, and π_i is the probability value from the interval $[0, 1]$. If linear regression is used then the variable can generally acquire any real value. That is why the odds function (or chance) is defined as

$$\text{odds} = \frac{P(Y=1)}{P(Y=0)} = \frac{\pi}{1-\pi}. \tag{3}$$

Thus defined function already acquires a value from the interval $[0,\infty]$. To be able to obtain a function acquiring a value from the whole definition range R it is necessary to calculate the logarithm of the function. Then the function is defined as a logit.

$$\text{logit} = \ln(\text{odds}) = \ln\left(\frac{\pi}{1-\pi}\right) \tag{4}$$

If $\text{logit} = \beta x'$, where $x' = (1, x_1, x_2, .. x_k)\ \beta = (\beta_0, \beta_1, \dots, \beta_k)$, then the logistic regression relation results:

$$\pi = \frac{e^{\beta x'}}{1 + e^{\beta x'}}. \tag{5}$$

The parameters in the linear combination of the explaining variables express the transformed median value of the explained variable – the logit. Logit is the logarithm of the quotient $\pi/1-\pi$ expressing the chances (odds) for the Y quantity to acquire the value of 1. The parameter β_0 expresses the size of the logit for zero values (categories) of all explaining variables. For $\beta_0 = 0$ the chances that $Y = 1$ are one to one, or $\pi = 0.5$. Positive values of the parameter β_0 mean that these odds are higher than one ($\pi > 0.5$) and negative values mean that they are lower than ($\pi < 0.5$). The logit may change in relation to one or more variables. The rate of the change is expressed by parameters $\beta_j, j = 1, \dots k$. In the case of a unit change of a *j-th* explaining variable (on condition that the other variables remain unchanged) the odds are that $Y = 1$, $e^{\beta j}$ times as big. P The maximum plausibility method is applied to logistic regression parameter estimates. [7]

3.3 Odds variables

There is also the $odds$ variable (chances), expressing the ratio of the good clients (G) to the bad clients (B) across the database.

$$odds = \frac{|G|}{|B|}. \tag{6}$$

For the individual features *j* of the individual groups *i* the variables $odds_j^i$, the chances of the feature, are defined as the ratios of the relevant numbers of good and bad clients in the individual categories.

$$odds_j^i = \frac{|G_j^i|}{|B_j^i|} \tag{7}$$

And finally there is the variable $odds\ ratio$, identified as OR_j^i. This variable expresses relative chances of the client within the given category to repay the loan. A value below 1 means that the client ´s chance to repay the loan is under-average, while high values on the other hand show above-average chances.

$$OR_j^i = \frac{odds_j^i}{odds} \tag{8}$$

The main purpose of the credit risk models is to estimate for each potential client with characteristic x o the value of the theoretical characteristic $odds(x)$. In practice it is not recommended to estimate the function $odds(x)$ as the ratio of good to bad clients with characteristic x. on condition of independence of the client in the data set this value will be estimated with the help of the following equation:

$$odds(x) = odds \prod_{(i,j)\in Z} (OR_j^i) x_j^i. \tag{9}$$

That is as the product of the total $odds$ and the respective OR_j^i of the categories where the potential client belongs. Allocations of various weights to the factors in the previous equation can result in different general models.

4 Diversification ability

The ability to diversify, i.e. to separate good clients for bad clients, is one of the most important analysed values of the scoring model. In the ideal case we would like to find a model in which the scoring boundary s_0, would clearly separate all bad clients by allocation of a score lower than s_0 and good clients by allocation of a score higher than s_0. In such a model we would be able on the basis of the calculated score to relatively reliably assess whether the client appears good or not. In practice, however, there is no scoring function faultlessly grasping quality of all clients in the database. There will certainly be clients with a low score who still manage to repay their loans and on the other hand clients who did not repay despite their high scores. The scoring function then only approximately divides the clients to good and bad ones. Quality of a scoring model with regard to its ability to diversify in then assessed according to how well the score is able to separate good clients from bad clients.

4.1 Gini coefficient and Lorenz curve

For graphic representation of the ability to diversify there is for example the Lorenz curve and for numerical quantification the Gini coefficient. Lorenz curve is used not only for graphic representations of scoring models. It also demonstrates the ability of the models to distinguish between good and bad clients. The curve is based on definition of distribution functions of good and bad clients. Gini coefficient is a numerical characteristic of the diversification ability of a scoring model. In the case of credit scoring the Gini coefficient is a benchmark showing how well the score card is able to distinguish between good and bad clients. The final result is the value representing the area under the Lorenz curve, illustrated on the Figure 1.

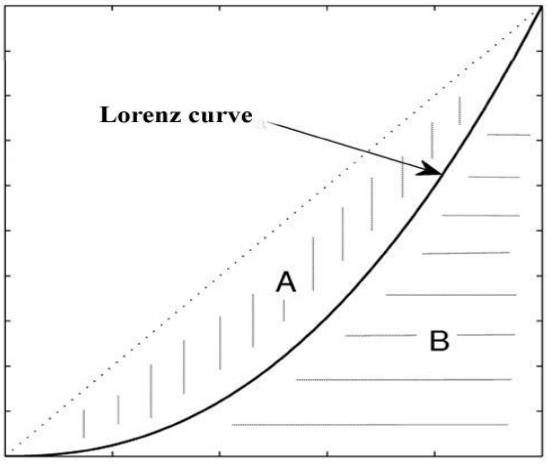

Figure 1 Lorenz curve and Gini coefficient

Let us define for each score value “s” the distribution function of the good client scores FG as the probability that the good client will have the score lower than “s” and the distribution function of the bad clients scores FB at the probability that the bad client will have the score lower than "s". The higher the diversification ability of the model is illustrated by the closer curve to the edges of the given square.

Gini coefficient is defined as the ratio of the oriented area between the Lorenz curve and the diagonal of the unit square (A) to the total area (A+B), i.e.

$$GC = \frac{A}{A + B} \tag{10}$$

Gini coefficient can acquire values from -1 to 1, with values close to 1 representing ideal diversification and values close to 0 representing zero diversification ability and negative values representing opposite classifications of the scoring function (the curve is sagging upwards). Therefore we look for scoring functions with the Gini coefficient as high as possible.

5 Conclusion

The purpose of this article was to describe theoretical assumptions for the construction of scoring models. The essay further draws links between the scoring model and risk, both credit risk and other risk categories. Construction of these models is described in detail including their diversification ability.

Thus defined models can also be used in other sectors apart from the banking sector and can therefore be applied to other areas related to the life of the institution as long as the institution possesses data necessary for the model construction.

Scoring models are irreplaceable for decisions when to lend money to a client and when not, but cannot remain the sole criterion for this decision, as the client can be classified incorrectly, i.e. a client able to repay the loan might be classified in the risky client category and vice versa. And the construction and use of the models should not replace the human factor altogether. At present there are numerous software systems able to serve to other than just banking institutions in using these models in practice.

References

[1] Anděl, J: *The basis of mathematical statistics*. Matfyzpress, Praha, 2007.
[2] Anderson, R: *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation.* Oxford University Press, New York, 2007, p. 731.
[3] Edelman ,D., Thomas, L. and Crook, J.: *Credit scoring and its applications*. Society for Industrial and Applied Mathematics, Philadelphia, 2002, p. 248.
[4] Hosmer, D. W. and Lemeshow, V.: *Applied Logistic Regression*, New York: John Wiley & Sons, 2000.
[5] Klepac, G.: Integrating Seasonal Oscillations into Basel II Behavioral Scoring Models. *Financial Theory and Practice*, vol. 31, no.3, Zagreb, 2007, pp. 281-291.
[6] Majer, I.: Application scoring: logit model approach and the divergence method compared. *Working Papers 17*, 2006.
[7] Pecáková, I.: Logistic regression with multicategorial explaining variable. *Acta Oeconomica Pragensia*, vol. 15, no. 1, Praha, 2007, pp. 86-96.
[8] Řeháková, B.: Do not be afraid for logistic regression. *Sociologic journal*,, Vol. 36 (No., vol. Vol. 36, no. No.4, pp. 475-492, 2000.
[9] Smejkal, V. and Rais, K.: *Risk management in enterprises and other organizations*. Praha: Grada, 2009, p. 360.
[10] The banks and facts - May 2012, "*Czech bank association*," [Online]. Published 11 05 2012, Available: http://www.czech-ba.cz/data/articles/down_43877.pdf. [Accessed 19 05 2012]
[11] "*EUR-Lex Access to law of European union*," [Online]. Available: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52005AE0244:CS:NOT. [Accessed 19 05 2012].