ACTA UNIVERSITATIS AGRICULTURAE ET SILVICULTURAE MENDELIANAE BRUNENSIS

Volume LIX 35 Number 2, 2011

ADVANCED EMPIRICAL ESTIMATE OF INFORMATION VALUE FOR CREDIT SCORING MODELS

M. Řezáč

Received: December 17, 2010

Abstract

ŘEZÁČ, M.: *Advanced empirical estimate of information value for credit scoring models.* Acta univ. agric. et silvic. Mendel. Brun., 2011, LIX, No. 2, pp. 267–274

Credit scoring, it is a term for a wide spectrum of predictive models and their underlying techniques that aid financial institutions in granting credits. These methods decide who will get credit, how much credit they should get, and what further strategies will enhance the profitability of the borrowers to the lenders. Many statistical tools are avaiable for measuring quality, within the meaning of the predictive power, of credit scoring models. Because it is impossible to use a scoring model effectively without knowing how good it is, quality indexes like Gini, Kolmogorov-Smirnov statisic and Information value are used to assess quality of given credit scoring model.
The paper deals primarily with the Information value, sometimes called divergency. Commonly it is computed by discretisation of data into bins using deciles. One constraint is required to be met in this case. Number of cases have to be nonzero for all bins. If this constraint is not fulfilled there are some practical procedures for preserving finite results. As an alternative method to the empirical estimates one can use the kernel smoothing theory, which allows to estimate unknown densities and consequently, using some numerical method for integration, to estimate value of the Information value.
The main contribution of this paper is a proposal and description of the empirical estimate with supervised interval selection. This advanced estimate is based on requirement to have at least k, where k is a positive integer, observations of socres of both good and bad client in each considered interval. A simulation study shows that this estimate outperform both the empirical estimate using deciles and the kernel estimate. Furthermore it shows high dependency on choice of the parameter k. If we choose too small value, we get overestimated value of the Information value, and vice versa. Adjusted square root of number of bad clients seems to be a reasonable compromise.

credit scoring, quality indexes, information value, empirical estimate, kernel smoothing

Credit scoring, it is a term for a wide spectrum of predictive models and their underlying techniques that aid financial institutions in granting credits. These methods decide who will get credit, how much credit they should get, and what further strategies will enhance the profitability of the borrowers to the lenders.

Methodology of credit scoring models and some measures of their quality were discussed in papers like Hand and Henley (1997) or Crook *et al.* (2007) and books like Anderson (2007), Siddiqi (2006), Thomas *et al.* (2002) and Thomas (2009). Further remarks connected to credit scoring issues can be found there as well.

Once a scoring model is available, it is natural to ask how good it is. To measure the partial processes of a financial institution, especially their components like scoring models or other predictive models, it is possible to use quantitative indexes such as Gini index, K-S statistic, Lift, Information value and so forth. They can be used for comparison of several developed models at the moment of development. It is possible to use them for monitoring the quality of models after the deployment into real business as well. See Wilkie (2004) or Siddiqi (2006) for more details.

The paper deals primarily with the Information value. Commonly it is computed by discretisation

of data into bins using deciles with requirement on the nonzero number of cases for all bins. As an alternative method to the empirical estimates one can use the kernel smoothing theory, which allows to estimate unknown densities and consequently, using some numerical method for integration, to estimate value of the Information value. The main objective of this paper is a description of the empirical estimate with supervised interval selection. This advanced estimate is based on requirement to have at least k, where k is a positive integer, observations of socres of both good and bad client in each considered interval. A simulation study shows the performance of this estimate compared to "clasical" empirical estimate and kernel estimate. Furthermore, it shows high dependency on choice of the parameter k.

Basic notations

Assume the realization $s \in \mathbf{R}$ of random value S (score) is available for each client. Let D be the indicator of good and bad client

$$D = \begin{cases} 1, & \text{client is good} \\ 0, & \text{client is bad} \end{cases} \tag{1}$$

and let F_0,F_1 denote cumulative distribution functions of score of bad and good clients, i.e.

$$\begin{aligned} F_0(a) &= P(S \le a \mid D = 0), \\ F_1(a) &= P(S \le a \mid D = 1), \text{a} \in \mathbf{R}. \end{aligned} \tag{2}$$

Assume F_0,F_1 and their corresponding densities f_0,f_1 are continuous on $\mathbf{R}$.

In practice, the empirical distribution functions are used

$$\begin{aligned} \hat{F}_0(a) &= \frac{1}{m} \sum_{i=1}^{N} I(s_i \le a \wedge D = 0) \\ \hat{F}_0(a) &= \frac{1}{n} \sum_{i=1}^{N} I(s_i \le a \wedge D = 1), a \in [L, H], \end{aligned} \tag{3}$$

where s_i is the score of i-th client, n, m are the number of good and bad clients, respectively and N=n+m. L is the minimum value of given score, H is the maximum value. Finally, we denote

$$p_B = \frac{m}{N}$$

the proportion of bad clients.

MATERIALS AND METHODS

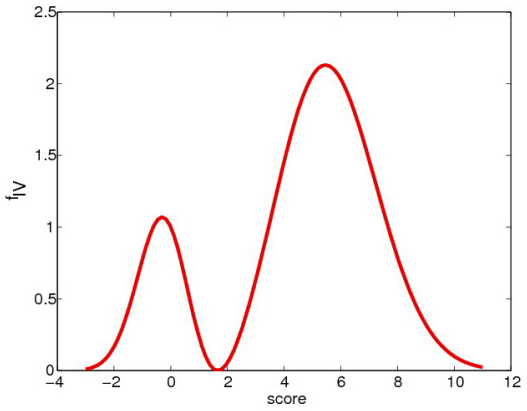
The quality index based on densities is the Information value (statistic) defined as

$$I_{val} = \int_{-\infty}^{\infty} f_{IV}(x)dx, \tag{4}$$

where

$$f_{IV}(x) = (f_1(x) - f_0(x)) \ln\left(\frac{f_1(x)}{f_0(x)}\right). \tag{5}$$

Note that the Information value is also called Divergence. See Wilkie (2004), Hand and Henley (1997) or Thomas (2009) for more details. The example of $f_{IV}(x)$ for 10% of bad clients with f_0:N(0,1) and 90% of good clients with f_1:N(4,2) is illustrated in Fig. 1.



1: *Contribution to Information value*

However, in practice, the procedure of computation of the Information value can be a little bit complicated. Firstly, we don't know the right form of densities f_0, f_1 generally and secondly, mostly we don't know how to compute the integral. I show some approaches to solve these computational problems.

The usual way, how to estimate the information value, is to replace unknown densities by their empirical estimates. Let's have m score values s_{0i}, i=1, …, m for bad clients and n score values s_{1j}, j=1, …,n for good clients and denote L(H) as the minimum (maximum) of all values, respectively. Let's divide the interval [L, H] up to r subintervals $[q_0, q_1],(q_1, q_2], \ldots, (q_{r-1}, q_r]$, where $q_0 = L - 1$, $q_r = H$ and q_i, i =1, …, r – 1 are appropriate quantiles of score of all clients. Set

$$\begin{aligned} n_{0j} &= \sum_{i=1}^{m} I(s_{0i} \in (q_{j-1}, q_j]) \\ n_{0j} &= \sum_{i=1}^{n} I(s_{1i} \in (q_{j-1}, q_j]), j = 1, \ldots, r \end{aligned} \tag{6}$$

observed counts of bad or good clients in each interval. Denote $\hat{f}_{IV}(j)$ the contribution to the information value on jth interval, calculated by

$$\hat{f}_{IV}(j) = \left(\frac{n_{1j}}{n} - \frac{n_{0j}}{m}\right) \ln\left(\frac{n_{1j}m}{n_{0j}n}\right), j = 1, \ldots, r. \tag{7}$$

Then the empirical information value is given by

$$\hat{I}_{val} = \sum_{j=1}^{r} \hat{f}_{IV}(j). \tag{8}$$

However in practice, there could occur computational problems. The Information value index becomes infinite in cases when some of n_{0_j} or n_{1_j} are equal to 0. When this arises there are numerous practical procedures for preserving finite results. For example one can replace the zero entry of num-

bers of goods or bads by a minimum constant of say 0.0001. Choosing the number of bins is also very important. In the literature and also in many applications in credit scoring, the value $r = 10$ is preferred.

I propose the alternative to the previous method, named the empirical estimate with supervised interval selection. This approach builds on ideas in the previous part. Estimation of information value is given again by formulas (6) to (8). The main difference lies in construction of the intervals. Because we want to avoid zero values of n_{0_j} and n_{1_j}, I simply looked for such selection of intervals, which provides such values n_{0_j} and n_{1_j}, which are all positive. This will lead to situation when all fractions and logarithms in (7) are defined and finite.

More generally, I propose to require to have at least k, where k is a positive integer, observations of socres of both good and bad client in each interval, i.e. $n_{0_j} \geq k$ and $n_{1_j} \geq k$ for j=1, …, r. Set

$$q_0 = \mathrm{L} - 1$$

$$q_i = \widehat{\mathrm{F}_0^{-1}}\left(\frac{\mathrm{k} \times \mathrm{i}}{\mathrm{m}}\right), \quad i = 1, \ldots, \left[\frac{m}{k}\right] \tag{9}$$

$$\mathrm{q}_{\left[\frac{m}{k}\right]+1} = \mathrm{H},$$

where $\widehat{\mathrm{F}_0^{-1}}(\cdot)$ is the empirical quantile function appropriate to the empirical cumulative distribution function of scores of bad clients. [x] means lower integer part of number x. Usage of quantile function of scores of bad clients is motivated by the assumption, that number of bad clients is less than number of good clients, which is quite natural assumption. If m is not divisible by k, it is necessary to adjust our intervals, because we obtain number of scores of bad clients in the last interval, which is less than k. In this case, we have to merge the last two intervals. This will lead to situation, when it holds $n_{0_j} \geq k$ for all computed intervals of scores.

Furthermore we need to ensure, that the number of scores of good clients is as required in each interval. To do so, we compute n_{1_j} for all actual intervals. If we obtain $n_{1_j} < k$ for j$^{\text{th}}$ interval, we merge this interval with its neighbor on the right side. This is equivalent with the removal of q_{i+1} from the sequence of borders of the intervals. This can be done for all intervals except the last one. If we have $n_{1_j} < k$ for the last interval, than we have to merge it with its neighbor on the left side, i.e. we merge the last two intervals. However, this situation is not very probable. If we have a reasonable scoring model, we can assume that good clients have higher scores than bad clients. It means that we can expect that the number of scores of good clients is higher than number of scores of bad clients in the last interval. Due to construction of the intervals, number of scores of bad clients in the last interval is greater than k. Thus, it is natural to expect that number of scores of good clients in the last interval is also greater than k. After all, we obtain $n_{0_j} \geq k$ and $n_{1_j} \geq k$ for all created intervals.

Very important is the choice of k. If we choose too small value, we get overestimated value of the Information value, and vice versa. As a reasonable compromise seems to be adjusted square root of number of bad clients given by

$$\mathrm{k} = |\sqrt{\mathrm{m}}|, \tag{10}$$

where [x] means upper integer part of number x.

Denote $\hat{\hat{f}}_{IV}(j)$ the contribution to the information value on j$^{\text{th}}$ interval, calculated by

$$\hat{\hat{f}}_{IV}(j) = \left(\frac{n_{1_j}}{n} - \frac{n_{0_j}}{m}\right) \ln\left(\frac{n_{1_j} m}{n_{0_j} n}\right), j = 1, \ldots, r, \tag{11}$$

where n_{1_j} and n_{0_j} correspond to observed counts of good and bad clients in intervals created according to the procedure described in this chapter. The empirical information value with supervised interval selection is now given by

$$\hat{\hat{I}}_{val} = \sum_{j=1}^{r} \hat{\hat{f}}_{IV}(j). \tag{12}$$

In the previous part, I described some difficulties arisen by computing the Information value. To avoid them one can use another approach, which is proposed in Koláček and Řezáč (2010). As proposed in this paper, it is possible to use the kernel smoothing theory to obtain estimates of unknown densities f_0, f_1. The kernel density estimates are defined by

$$\begin{aligned} \tilde{f}(x, h_0) &= \frac{1}{m} \sum_{i=1}^{m} K_{h_0}(x - \mathrm{s}_{0_i}), \\ \tilde{f}(x, h_1) &= \frac{1}{n} \sum_{i=1}^{n} K_{h_1}(x - \mathrm{s}_{1_i}), \end{aligned} \tag{13}$$

where

$$K_{h_i}(x) = \frac{1}{h_i} \mathrm{K}\left(\frac{x}{h_i}\right), i = 0,1$$

and K is the Epanechnikov kernel

$$\mathrm{K}(x) = \frac{3}{4}(1 - x^2) I_{[-1,1]}. \tag{14}$$

For further details see Wand and Jones (1995). The quality of kernel density estimates is affected mainly by smoothing parameters h_0 and h_1. The estimation of optimal bandwidths h_i can be given by maximal smoothing principal approach, i.e.

$$\begin{aligned} \tilde{h}_{0,m} &= 2{,}5324\, \tilde{\sigma}_0 m^{-\frac{1}{5}} \\ \tilde{h}_{0,m} &= 2{,}5324\, \tilde{\sigma}_0 m^{-\frac{1}{5}}, \end{aligned} \tag{15}$$

where $\tilde{\sigma}_i$, $i = 0,1$ are appropriate estimations of standard deviation for bad and good clients. For more details see Terrell (1990).

The next step is to compute the final integral. To estimate this one can use the composite trapezoidal rule. Set

$$\tilde{f}_{IV}(x) = (\tilde{f}_1(x,h_1) - \tilde{f}_0(x,h_0)) \ln\left(\frac{\tilde{f}_1(x,h_1)}{\tilde{f}_0(x,h_0)}\right). \quad (16)$$

Then, for given $M + 1$ equidistant points $L = x_0$, $x_1, \ldots, x_M = H$ we obtain

$$\tilde{I}_{val} = \frac{H-L}{2M}\left(\tilde{f}_{IV}(L) + 2\sum_{i=1}^{M-1}\tilde{f}_{IV}(x_i) + \tilde{f}_{IV}(H)\right). \quad (17)$$

The value of M is usually set from 100 to 1 000. As one has to trade off between computational speed and accuracy, I propose to use $M = 500$.

RESULTS

Firstly, it is natural to ask which estimate is the best. To compare them and get an answer, a short simulation study was done. Consider n clients, 10% of bad clients with f_0:$N(0,1)$ and 90% of good clients with f_1:$N(1,1)$. Because of normality of scores, we can compute the Information value (see Wilkie (2004)) as

$$I_{val} = \left(\frac{\mu_g - \mu_b}{\sigma}\right)^2,$$

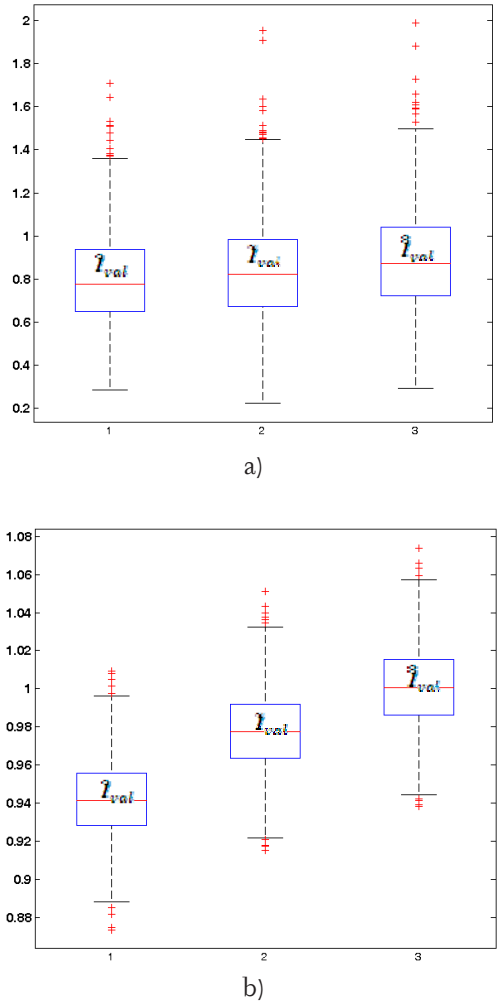
where μ_g and μ_b are means of scores of good and bad clients, σ is their common standard deviation. Given our choice of these parameters we know the value of I_{val}, which is equal to one.

Scores of bad and good clients were generated according to given parameters. Firstly, the number of clients was set to $n = 500$. Estimates $\tilde{I}_{val}$, $\hat{I}_{val}$ and $\hat{\hat{I}}_{val}$ were computed and remembered. These two steps were repeated one thousand times. Finally, averages and interquartile ranges for all three types of estimates were computed. Than the number of clients was set to $n = 100\,000$ and the same computations were made. The results are given in Tab. I.

I: *Average and iqr of $\tilde{I}_{val}$, $\hat{I}_{val}$ and $\hat{\hat{I}}_{val}$ for 500 and 100 000 clients*

n = 500	average	iqr
$\hat{I}_{val}$	0.8008	0.2885
$\tilde{I}_{val}$	0.8410	0.3101
$\hat{\hat{I}}_{val}$	0.8898	0.3154
n = 100 000	**average**	**iqr**
$\hat{I}_{val}$	0.9420	0.0276
$\tilde{I}_{val}$	0.9779	0.0281
$\hat{\hat{I}}_{val}$	1.0010	0.0290

We can see that the best performance was obtained for $\hat{\hat{I}}_{val}$, i.e the empirical estimate with supervised interval selection. The second best estimate was the kernel one. The empirical estimate, which used deciles, was outperformed by both of them. This order was preserved for both $n = 100\,000$ and $n = 500$, i.e. very large and very small range of data. Although the interquartile ranges, as the robust estimate of dispersions, had reversed order, they took very similar values. Over all, $\tilde{I}_{val}$ was significantly better than $\hat{I}_{val}$ and $\hat{\hat{I}}_{val}$ was significantly better than both $\tilde{I}_{val}$ and $\hat{I}_{val}$. Appropriate boxplots are shown in Fig. 2.



2: *Box plots of $\tilde{I}_{val}$, $\hat{I}_{val}$ and $\hat{\hat{I}}_{val}$ - (a) 500 clients, (b) 100 000 clients*

The second part of this chapter is focused on properties of $\hat{\hat{I}}_{val}$ depending on choice of parameter k and depending on proportion of bad clients p_B and difference of means of scores of bad and good clients $\mu_g - \mu_b$. Consider 10 000 clients, $100 \times p_B$% of bad clients with f_0:$N(\mu_b, 1)$ and $100 \times (1 - p_b)$% of good clients with f_1:$N(\mu_g, 1)$. Set $\mu_b = 0$ and consider $\mu_b = 0.5$, 1 and 1.5, $p_b = 0.02$, 0.05, 0.1 and 0.2. The case $\mu_g - \mu_b = 0.5$, i.e. $I_{val} = 0.25$ in our settings, represents weak, $\mu_g - \mu_b = 1$ means high and $\mu_g - \mu_b = 1.5$ very high performance of given scoring model. 2% bad rate ($p_b = 0.02$) represents low risk portfolio, e.g. mortgages (before current crises). 20% bad represents very high risk portfolio, e.g. subprime cash loans.

Appropriate data sets for simulation was randomly generated 1 000 times. Quality of $\hat{I}_{val}$ was assessed using mean squared error given by

$$MSE = E\left((\hat{I}_{val} - I_{val})^2\right). \tag{18}$$

Given this measure, denote

$$k_{MSE} = \underset{k}{\operatorname{argmin}}\ \mathrm{MSE}. \tag{19}$$

Following Tab. II consists of k_{MSE} for all considered values of p_b and $\mu_g - \mu_b$. Values of $k = \lfloor\sqrt{m}\rfloor$ are presented in the last row of the table.

II: *k_{MSE} depending on p_B and difference of μ_g and μ_b*

k_{MSE}		p_B			
		0.02	**0.05**	**0.1**	**0.2**
$\mu_g - \mu_b$	0.5	29	42	62	84
	1	12	18	23	32
	1.5	6	9	8	9
$k = \lfloor\sqrt{m}\rfloor$		15	23	32	45

We can see that k_{MSE} is increasing according to p_B. This is may be somewhat surprising, but it is quite natural. The increasing p_B means increasing number of bad clients, becasuse the number of all clients was fixed to 10 000. If we have enough of bad clients, then too small k leads to too many bins and consequently to overestimated results. But what is surprising, it is the dependence on $\mu_g - \mu_b$. While for weak models it is optimal to take very high number of observation in each bin, the contrary holds for high perfoming models. Overall, $k = \lfloor\sqrt{m}\rfloor$ seems to be a reasonbale comprimise.

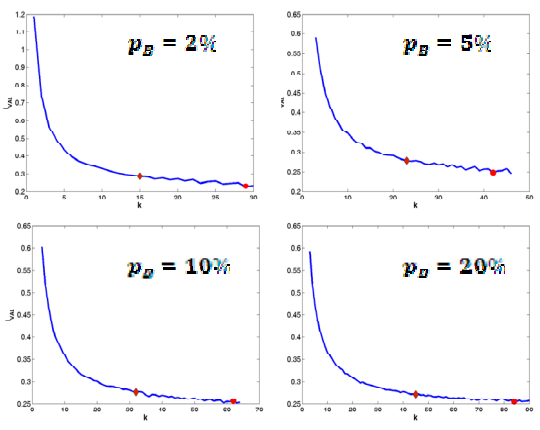
For completeness, Tab. III consists of average numbers of bins for all considered values of p_B and $\mu_g - \mu_b$. We can see that they took values from 8 to 127.

III: *Average number of bins depending on p_B and difference of μ_g and μ_b*

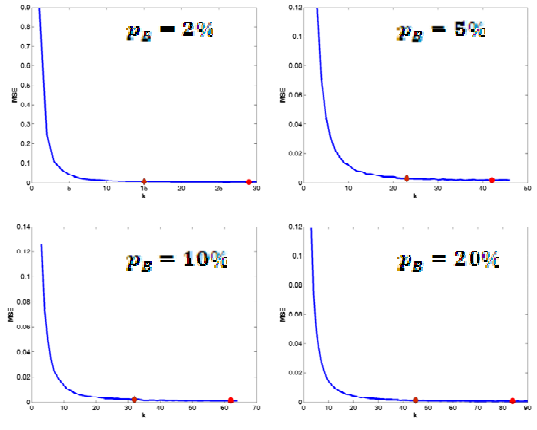
avg. # of bins		p_B			
		0.02	**0.05**	**0.1**	**0.2**
$\mu_g - \mu_b$	0.5	8.00	13.00	18.00	24.90
	1	18.00	28.80	42.76	51.88
	1.5	33.62	50.20	95.96	127.67

The dependece of $\hat{I}_{val}$ on k is illustrated in Fig. 3 to 5. The highlighted circles correspond to values of k, where minimal value of the MSE is obtained. The diamonds correspond to values of k given by (10).

Fig. 3 and Fig. 4 show that curves of MSE are quite flat nearby its minimum. It means that a small deviation of k from k_{MSE} cause a small change in MSE. On the other hand Fig. 5 shows the strong dependence on choice of k.

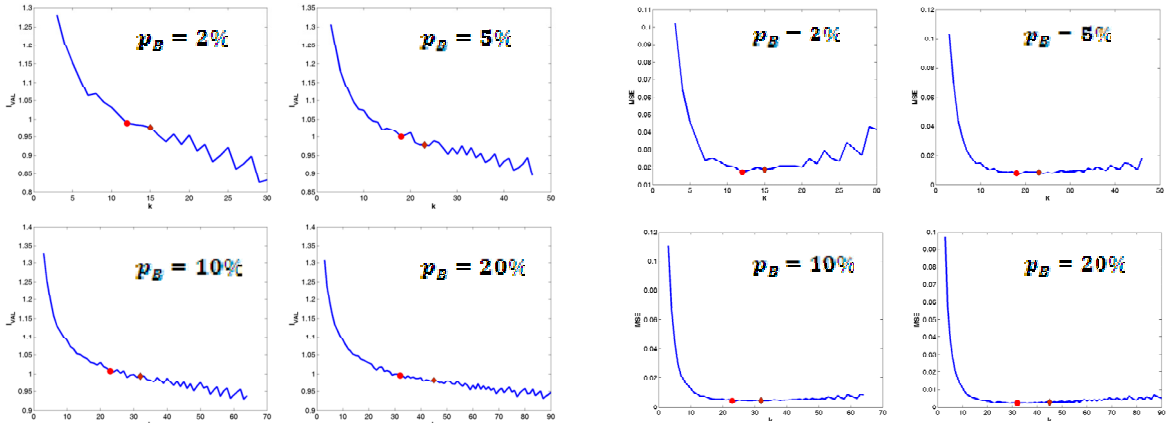


a)



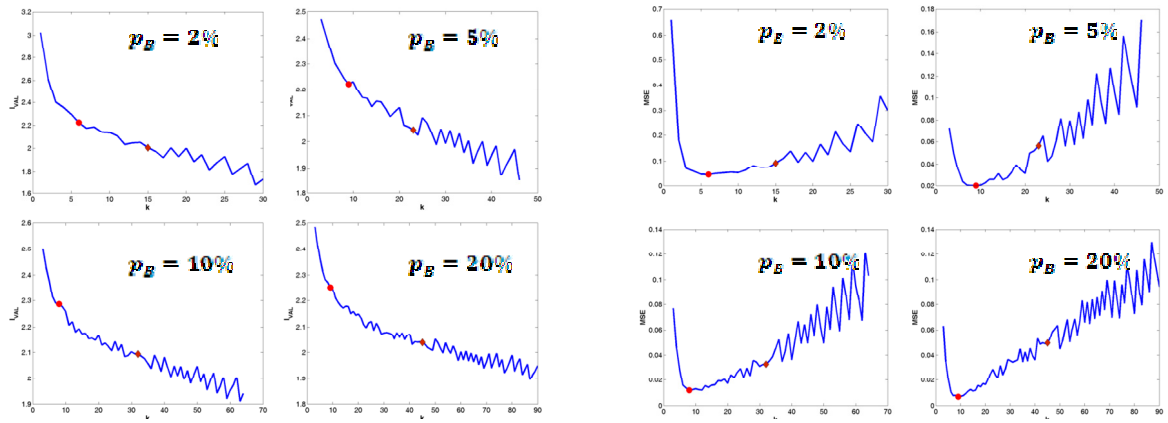
b)

3: *Dependence of $\hat{\hat{I}}_{val}$ and (b) MSE on k, 100 000 clients, $\mu_g - \mu_b = 0.5$*



a) b)

4: *Dependence of (a) $\hat{\hat{I}}_{val}$ and (b) MSE on k, 100 000 clients, $\mu_g - \mu_b = 1$*



a) b)

5: *Dependence of (a) $\hat{\hat{I}}_{val}$ and (b) MSE on k, 100 000 clients, $\mu_g - \mu_b = 1.5$*

SUMMARY

I focused on the Information value and described difficulties of its estimation. The most popular method is the empirical estimator using deciles of given score. But it can lead to infinite values of I_{val} and so a remedy is necessary. To avoid these difficulties the kernel method was proposed. The advantage of this approach is the smoothness of the contribution and easy implementation with a polynomial kernel. Furthermore, I proposed the adjustment for the empirical estimate, called the empirical estimate with supervised interval selection. It is based on the assumption that we have at least some positive number of observed scores in each interval. This directly leads to situation when all fractions and all logarithms are defined and finite. Consequently, I_{val} is defined and finite.

The simulation study showed that for normally distributed scores the empirical estimate with supervised interval selection outperformed both the kernel estimate and the "classical" empirical estimate. This was true for very large and very small range of data files. Furthermore, it is focused on properties of $\hat{\hat{I}}_{val}$ depending on choice of parameter k and depending on proportion of bad clients and difference of means of scores of bad and good clients. It showed some surprising results that were discussed at the end of the paper.

REFERENCES

ANDERSON, R., 2007: The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation. Oxford: Oxford University Press, 731 s. ISBN 978-0-19-922640-5.

CROOK, J. N., EDELMAN, D. B., THOMAS, L. C., 2007: Recent developments in consumer credit risk assessment. European Journal of Operational Research, 183 (3), 1447–1465.

HAND, D. J. and HENLEY, W. E., 1997: Statistical Classification Methods in Consumer Credit Scoring: a review. Journal of the Royal Statistical Society, Series A. 160 (3), 523–541.

KOLÁČEK, J. and ŘEZÁČ, M., 2010: Assessment of Scoring Models Using Information Value. In: Compstat' 2010 proceedings. Paris, 1191–1198. ISBN 978-3-7908-2603-6.

SIDDIQI, N., 2006: Credit Risk Scorecards: developing and implementing intelligent credit scoring. New Jersey: Wiley, 196 p. ISBN 978-0-471-75451-0.

TERRELL, G. R., 1990: The Maximal Smoothing Principle in Density Estimation. Journal of the American Statistical Association, 85, 470–477.

THOMAS, L. C., 2009: Consumer Credit Models: Pricing, Profit, and Portfolio. Oxford: Oxford University Press, 385 p. ISBN 978-0-19-923213-0.

THOMAS, L. C., EDELMAN, D. B., CROOK, J. N., 2002: Credit Scoring and Its Applications. Philadelphia: SIAM Monographs on Mathematical Modeling and Computation, 248 p. ISBN 978-0-898714-83-8.

WAND, M. P. and JONES, M. C., 1995: Kernel smoothing. London: Chapman and Hall, 212 s. ISBN 978-0-41255270-0.

WILKIE, A. D., 2004: Measures for comparing scoring systems. In: THOMAS, L. C., EDELMAN, D. B., CROOK, J. N. (Eds.), Readings in Credit Scoring. Oxford: Oxford University Press, p. 51–62.

Address

Mgr. Martin Řezáč, Ph.D., Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Kotlářská 2, 611 37 Brno, Česká republika, e-mail: mrezac@math.muni.cz