MASARYK UNIVERSITY
Czech Republic

Estimating Information Value for Credit Scoring Models

Martin Řezáč

Dept. of Mathematics and Statistics, Faculty of Science,
Masaryk University

Basic notations

- We consider following markings:

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

Number of all clients: N
Number of good clients: n
Number of bad clients: m
Proportions of good/bad clients:

$$p_G=\frac{n}{n+m}\ ,\ p_B=\frac{m}{n+m}$$

- Cumulative distribution functions (CDF):

$$F_0(a)=P(S\le a\mid D=0),$$
$$F_1(a)=P(S\le a\mid D=1),\qquad a\in\mathbb{R}.$$

- Corresponding densities:

$$f_0, f_1$$

- Empirical cumulative distribution functions (CDF):

$$\hat{F}_0(a)=\frac{1}{m}\sum_{i=1}^{N} I(s_i\le a\wedge D=0)$$
$$\hat{F}_1(a)=\frac{1}{n}\sum_{i=1}^{N} I(s_i\le a\wedge D=1),\qquad a\in[L,H],$$

$$I(A)=\begin{cases}1 & A\ is\ true\\ 0 & A\ is\ false\end{cases}$$

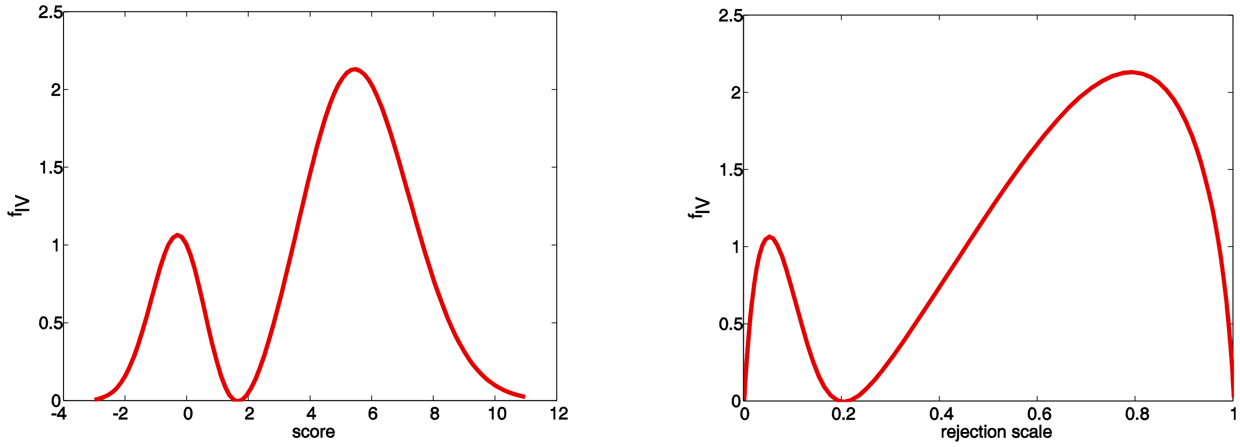
Information value

- The special case of Kullback-Leibler divergence given by:

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

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

The example of $f_{IV}(x)$ for 10% of bad clients with $f_0 \sim N(0,1)$ and 90% of good clients with $f_1 \sim N(4,2)$



I_{val} for normally distributed scores

- Assume that the scores of good and bad clients are normally distributed, i.e. we can write their densities as

$$f_1(x)=\frac{1}{\sigma_g\sqrt{2\pi}}e^{-\frac{(x-\mu_g)^2}{2\sigma_g^2}} \qquad f_0(x)=\frac{1}{\sigma_b\sqrt{2\pi}}e^{-\frac{(x-\mu_b)^2}{2\sigma_b^2}}$$

- Assume that standard deviations are equal to a common value σ :

$$\boxed{I_{val}=D^2}$$

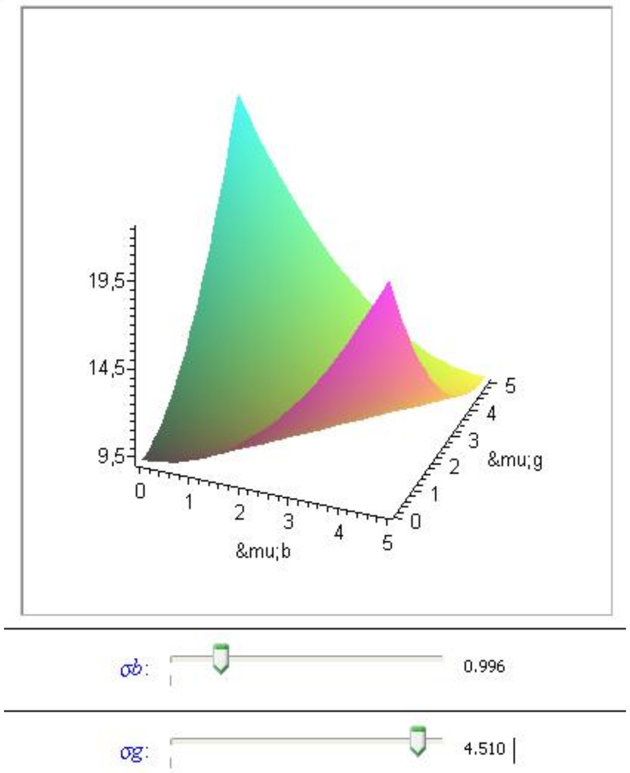
where $D=\dfrac{\mu_g-\mu_b}{\sigma}$

- Generally (i.e. without assumption of equality of standard deviations):

$$\boxed{I_{val}=(A+1)D^{*2}+A-1,\quad A=\frac{1}{2}\left(\frac{\sigma_b^2}{\sigma_g^2}+\frac{\sigma_g^2}{\sigma_b^2}\right)}$$

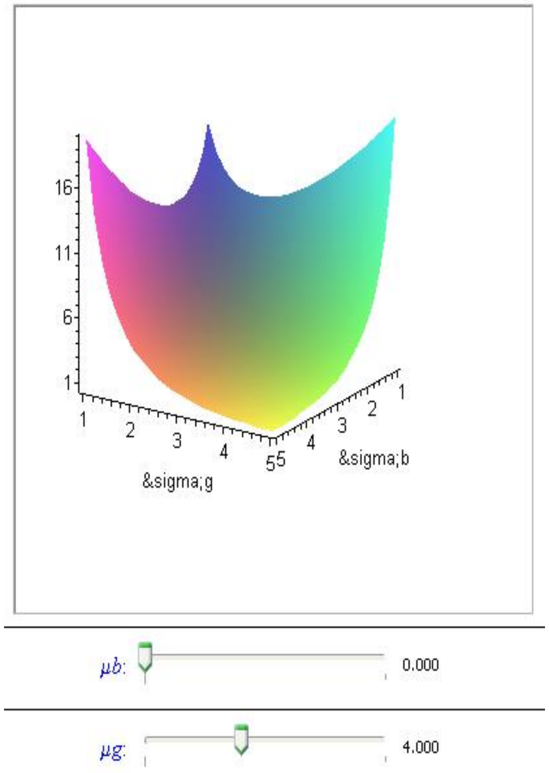
where $D^*=\dfrac{\mu_g-\mu_b}{\sqrt{\sigma_g^2+\sigma_b^2}}$

I_{val} for normally distributed scores



σb: 0.996

σg: 4.510



μb: 0.000

μg: 4.000

- ➢ We can see a quadratic dependence on difference of means.
- ➢ I_{val} takes quite high values when both variances are approximately equal and smaller or equal to 1, and it grows to infinity if ratio of the variances tends to infinity or is nearby zero.

Empirical estimate of I_{val}

The main idea of this approach is to replace unknown densities by their empirical estimates. Let's have m score values s_{0_i}, $i = 1, \dots, m$ for bad clients and n score values s_{1_j}, $j = 1, \dots, n$ for good clients and denote L (resp. H) as the minimum (resp. maximum) of all values. Let's divide the interval $[L, H]$ up to r equal subintervals $[q_0, q_1], (q_1, q_2], \dots, (q_{r-1}, q_r]$, where $q_0 = L, q_r = H$. Set

$$n_{0_j} = \sum_{i=1}^{m} I\left(s_{0_i} \in (q_{j-1}, q_j]\right)$$
$$n_{1_j} = \sum_{i=1}^{n} I\left(s_{1_i} \in (q_{j-1}, q_j]\right), \quad j = 1, \dots, r$$

observed counts of bad or good clients in each interval. Then the empirical information value is calculated by

$$\widehat{I}_{val} = \sum_{j=1}^{r} \left(\frac{n_{1_j}}{n} - \frac{n_{0_j}}{m} \right) \ln \left(\frac{n_{1_j} m}{n_{0_j} n} \right).$$

Empirical estimate of I_{val}

❑ However in practice, there could occur computational problems. The Information value index becomes infinite in cases when some of n_{0j} or n_{1j} 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 numbers of goods or bads by a minimum constant of say 0.0001. Choosing of 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.

Empirical estimate with supervised interval selection

- We want to avoid zero values of n_{0j} or n_{1j} .
- I propose to require to have at least *k*, where *k* is a positive integer, observations of scores of both good and bad client in each interval.
- Set

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

$$q_i = \widehat{\mathrm{F}_0^{-1}}\left(\frac{\mathrm{k}\cdot\mathrm{i}}{\mathrm{m}}\right), i = 1, \dots, \left\lfloor\frac{m}{k}\right\rfloor$$

$$q_{\left\lfloor\frac{m}{k}\right\rfloor+1} = \mathrm{H}$$

where $\widehat{\mathrm{F}_0^{-1}}(\cdot)$ is the empirical quantile function appropriate to the empirical cumulative distribution function of scores of bad clients.

Empirical estimate with supervised interval selection

- 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.
- 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.
- 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^th^ interval, we merge this interval with its neighbor on the right side.
- 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.

Empirical estimate with supervised interval selection

- Very important is the choice of *k*. If we choose too small value, we get overestimted 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

$$k = \left\lceil \sqrt{m} \right\rceil$$

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

where n_{0j} and n_{1j} correspond to observed counts of good and bad clients in intervals created according to the described procedure.

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

Simulation results

❑ Consider 10000 clients, $100p_B$% of bad clients with $f_0 : N(\mu_b, 1)$ and $100(1-p_B)$% 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.

$$MSE = E\left(\left(\hat{\hat{I}}_{val} - I_{val}\right)^2\right) \quad \Rightarrow \quad k_{MSE} = \underset{k}{argmin}\, MSE.$$

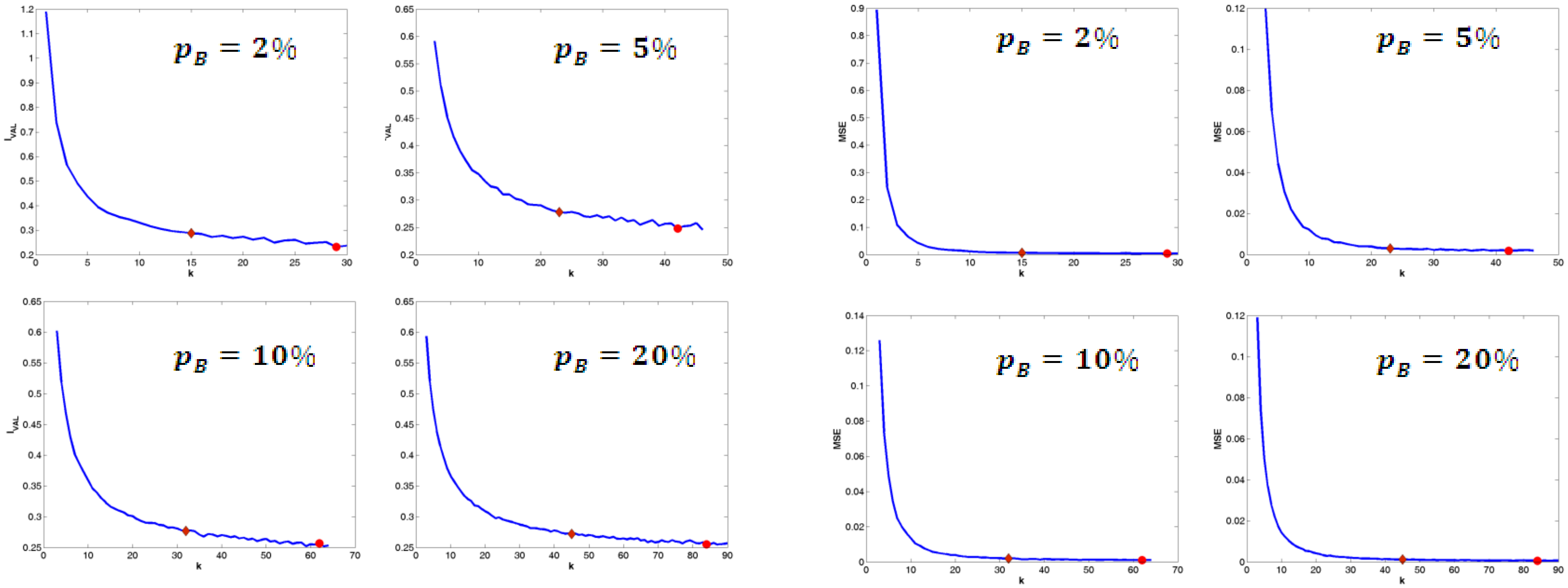
k_{MSE}		p_B			
		0.02	0.05	0.1	0.2
$\mu_g - \mu_b$	0.5	29	42	62	84
	1	12	18	23	32
	1.5	6	9	8	9
$k = \lceil \sqrt{m} \rceil$		15	23	32	45

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 optimal number of bins is quite higher than 10.

Simulation results

- **Dependence of $\hat{\hat{I}}_{val}$ and MSE on *k*,** $\mu_g - \mu_b = 0.5$ **.**

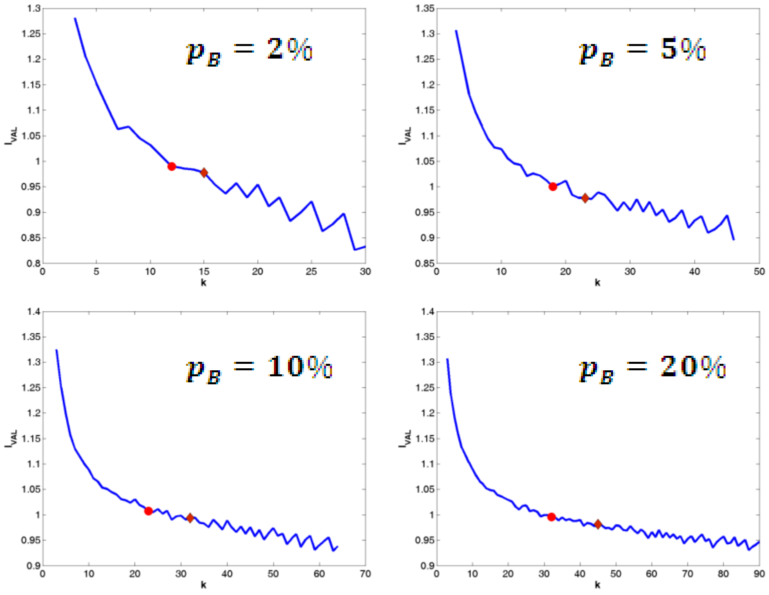


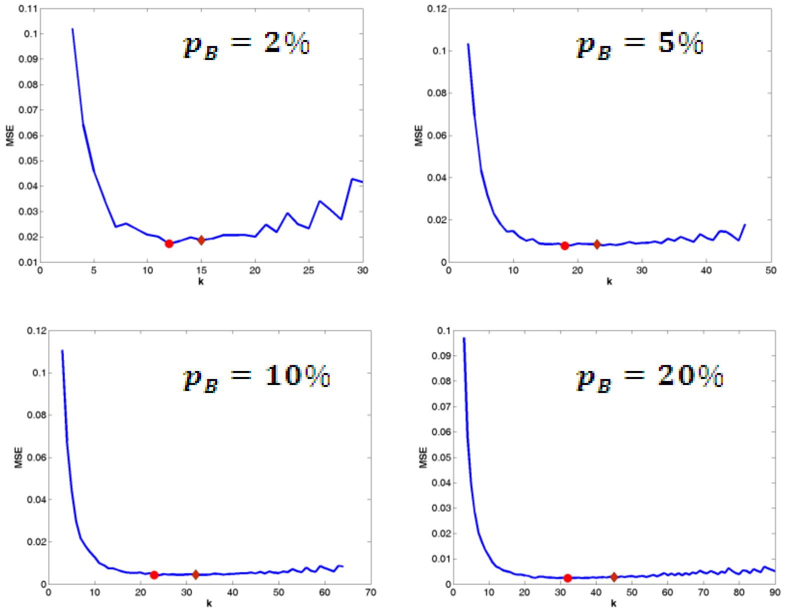
➢ The highlighted circles correspond to values of *k*, where minimal value of the *MSE* is obtained. The diamonds correspond to values of *k* given by $k = \lceil \sqrt{m} \rceil$.

➢ We can see that $\hat{\hat{I}}_{val}$ is decreasing when *k* is increasing. The speed of this decreas is very high for small values of *k*, while it is nearly negligible for values of *k* higher than some critical value. The similar holds for *MSE*.

Simulation results

❑ **Dependence of $\hat{I}_{val}$ and MSE on *k*, $\mu_g - \mu_b = 1$.**

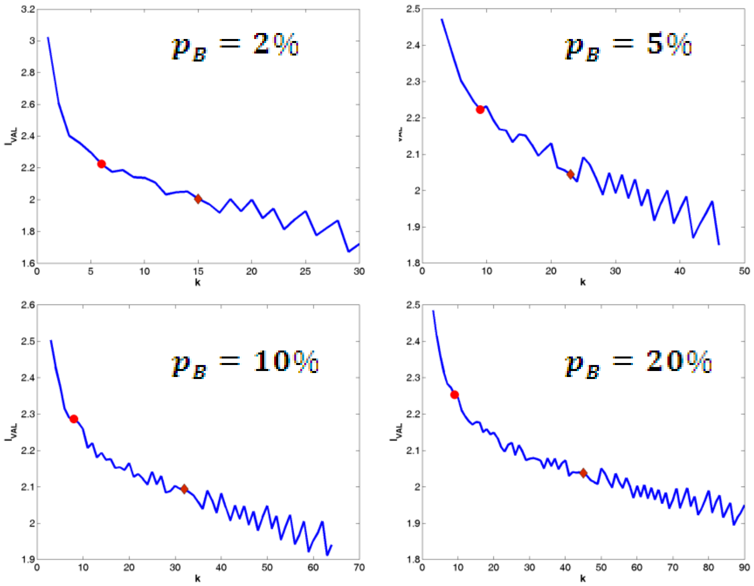


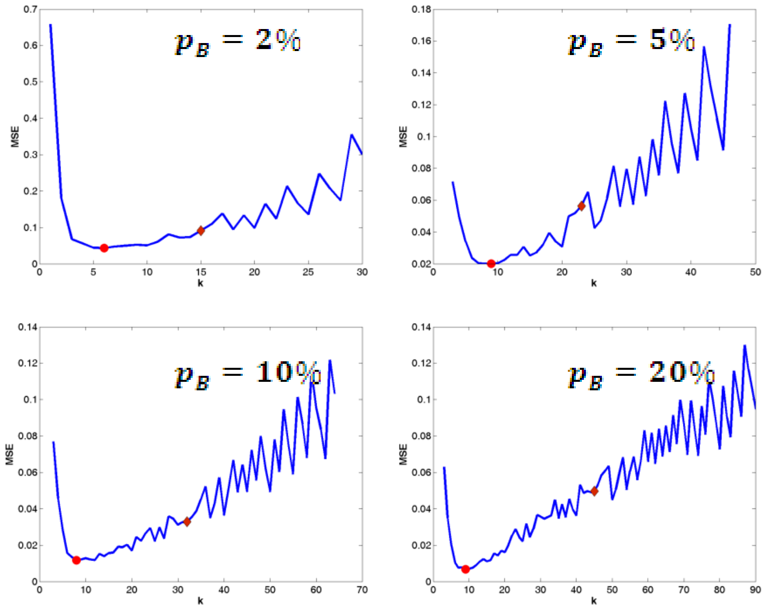


➢ The speed of the decrease is lower compared to the previous case. Furthermore *MSE* is not so flat, especially for p_B =2%. But what is interesting and important here, our choice of *k* is nearly optimal according to *MSE*. Moreover, it is valid for all considered values of p_B.

Simulation results

- **Dependence of $\hat{\hat{I}}_{val}$ and MSE on *k*, $\mu_g - \mu_b = 1.5$.**





➢ The speed of the decrease of $\hat{\hat{I}}_{val}$ is the lowest compared to the previous two cases. The novelty, relative to the previous two cases, is the shape of *MSE*. Especially for the highest considered value of proportion of bad clients, i.e. p_B = 20%, we can see that MSE has really sharp minimum.

Conclusions

□ The most popular method for the Information value estimation is the empirical estimator using deciles of given score.

□ But it can lead to infinite values of I_{val}.

□ The proposed adjustment for the empirical estimate, called the empirical estimate with supervised interval selection, solve this issue.

□ The simulation study showed 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 according to *MSE*.

Thank you for your attentation.