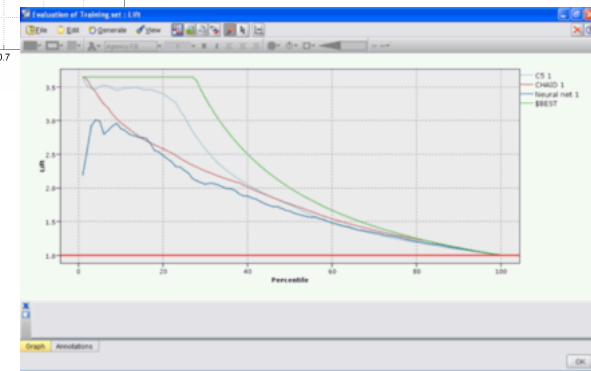
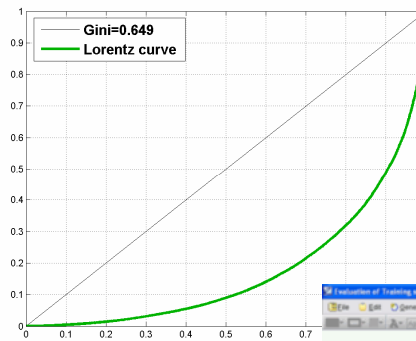
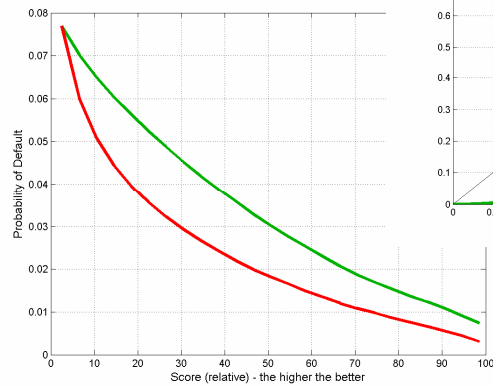
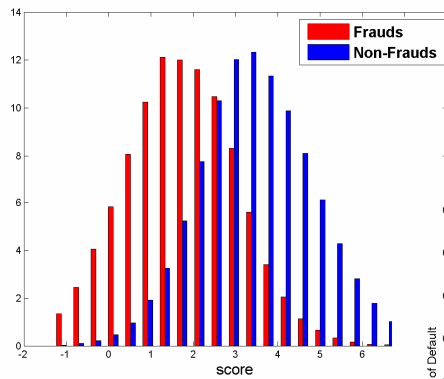


Credit scoring – evaluate modelu



Introduction



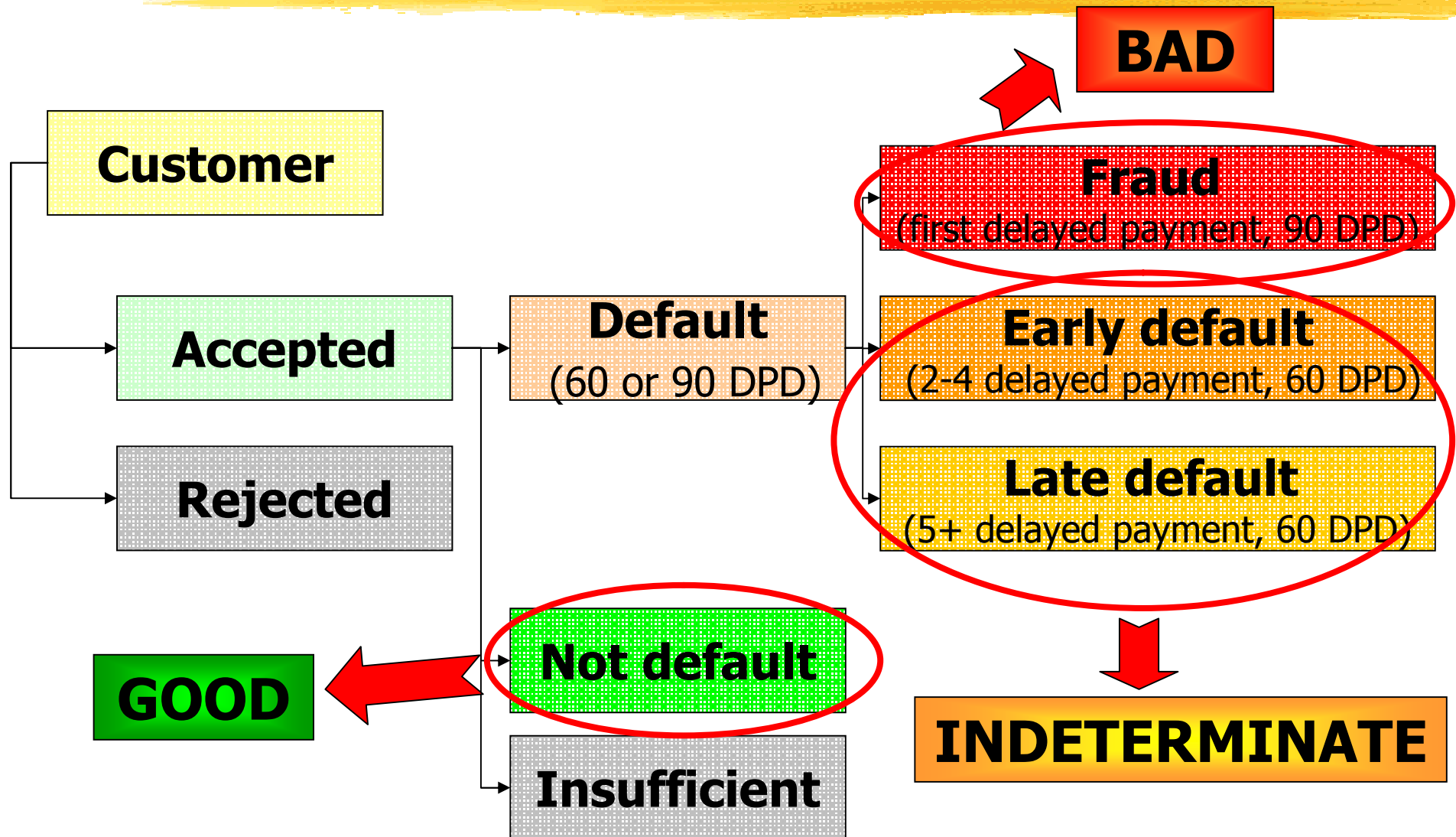
- ❑ It is impossible to use scoring model effectively without knowing how good it is.
- ❑ Usually one has several scoring models and needs to select just one. The best one.
- ❑ Before measuring the quality of models one should know (among other things):
 - good/bad definition
 - expected reject rate

Good/bad client definition



- ❑ Good definition is the basic condition of effective scoring model.
- ❑ The definition usually depends on:
 - days past due (DPD)
 - amount past due
 - time horizon
- ❑ Generally we consider following types of client:
 - Good
 - Bad
 - Indeterminate
 - Insufficient
 - Excluded
 - Rejected.

Good/bad client definition



Measuring the quality

□ Once the definition of good / bad client and client's score is available, it is possible to evaluate the quality of this score. If the score is an output of a predictive model (scoring function), then we evaluate the quality of this model. We can consider two basic types of quality indexes. First, indexes based on cumulative distribution function like

- Kolmogorov-Smirnov statistics (KS)
- Gini index
- C-statistics
- Lift.

The second, indexes based on likelihood density function like

- Mean difference (Mahalanobis distance)
- Informational statistics/value (I_{val}).

Indexes based on distribution function

$$D_K = \begin{cases} 1, & \text{client is good} \\ 0, & \text{otherwise.} \end{cases}$$

Number of good clients: n

Number of bad clients: m

Proportions of good/bad clients: $p_G = \frac{n}{n+m}$, $p_B = \frac{m}{n+m}$

➤ Empirical distribution functions:

$$F_{n,GOOD}(a) = \frac{1}{n} \sum_{i=1}^n I(s_i \leq a \wedge D_K = 1)$$

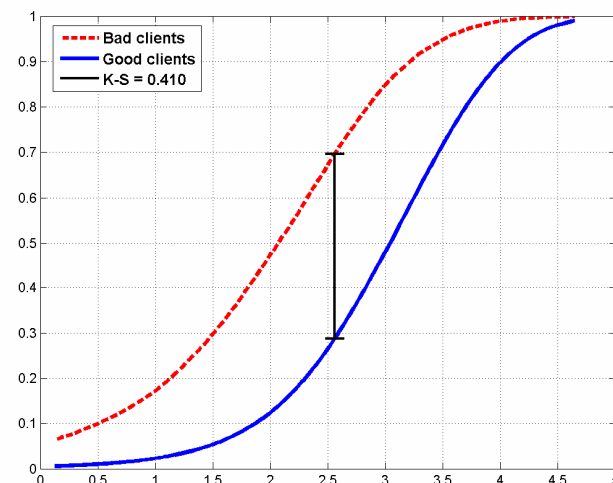
$$F_{m,BAD}(a) = \frac{1}{m} \sum_{i=1}^m I(s_i \leq a \wedge D_K = 0)$$

$$F_{N,ALL}(a) = \frac{1}{N} \sum_{i=1}^N I(s_i \leq a) \quad a \in [L, H]$$

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

➤ Kolmogorov-Smirnov statistics (KS)

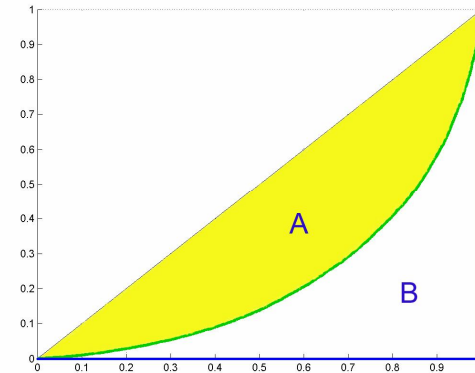
$$KS = \max_{a \in [L, H]} |F_{m,BAD}(a) - F_{n,GOOD}(a)|$$



Indexes based on distribution function

➤ Lorenz curve (LC)

$$x = F_{m.BAD}(a)$$
$$y = F_{n.GOOD}(a), a \in [L, H].$$



➤ Gini index

$$Gini = \frac{A}{A+B} = 2A$$

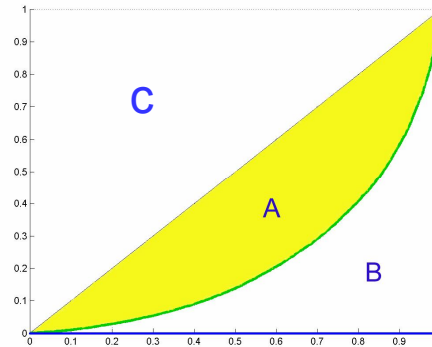
$$Gini = 1 - \sum_{k=2}^{n+m} (F_{m.BAD_k} - F_{m.BAD_{k-1}}) \cdot (F_{n.GOOD_k} + F_{n.GOOD_{k-1}})$$

where $F_{m.BAD_k}$ ($F_{n.GOOD_k}$) is k-th vector value of empirical distribution function of bad (good) clients

Indexes based on distribution function

➤ C-statistics:

$$c-stat = A + C = \frac{1 + Gini}{2}$$



It represents the likelihood that randomly selected good client has higher score than randomly selected bad client, i.e.

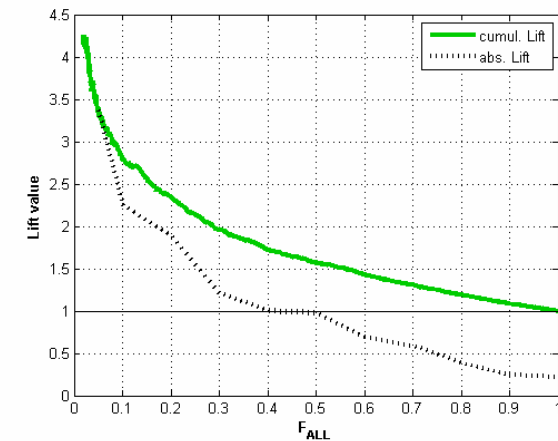
$$c-stat = P(s_1 \geq s_2 \mid D_{K_1} = 1 \wedge D_{K_2} = 0)$$

Indexes based on distribution function

□ Another possible indicator of the quality of scoring model can be *cumulative Lift*, which says, how many times, at a given level of rejection, is the scoring model better than random selection (random model). More precisely, the ratio indicates the proportion of bad clients with less than a score a , $a \in [L, H]$, to the proportion of bad clients in the general population. Formally, it can be expressed by:

$$Lift(a) = \frac{CumBadRate(a)}{BadRate} = \frac{\frac{\sum_{i=1}^{n+m} I(s_i \leq a \wedge Y = 0)}{\sum_{i=1}^{n+m} I(s_i \leq a)}}{\frac{\sum_{i=1}^{n+m} I(Y = 0)}{\sum_{i=1}^{n+m} I(Y = 0 \vee Y = 1)}} = \frac{\sum_{i=1}^{n+m} I(s_i \leq a \wedge Y = 0)}{\sum_{i=1}^{n+m} I(s_i \leq a)} = \frac{n}{N}$$

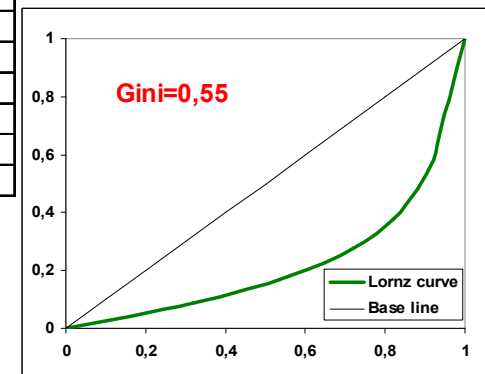
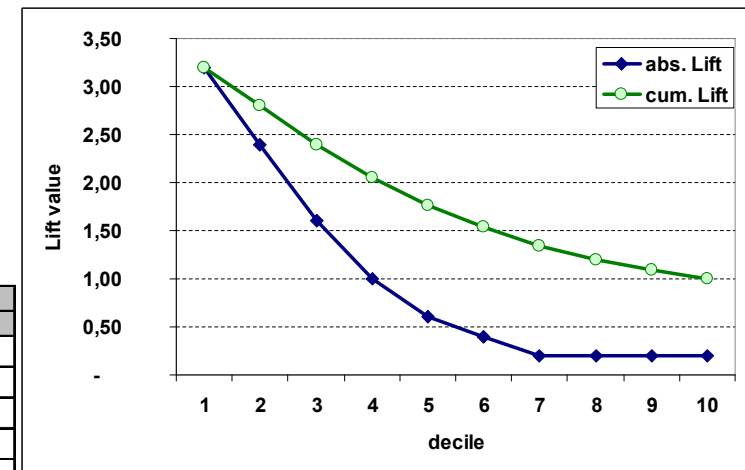
$$absLift(a) = \frac{BadRate(a)}{BadRate}$$



Indexes based on distribution function

□ Usually it is computed using table with numbers of all and bad clients in some bands (deciles).

decile	# cleints	absolutely			cumulatively		
		# bad clients	Bad rate	abs. Lift	# bad clients	Bad rate	cum. Lift
1	100	16	16,0%	3,20	16	16,0%	3,20
2	100	12	12,0%	2,40	28	14,0%	2,80
3	100	8	8,0%	1,60	36	12,0%	2,40
4	100	5	5,0%	1,00	41	10,3%	2,05
5	100	3	3,0%	0,60	44	8,8%	1,76
6	100	2	2,0%	0,40	46	7,7%	1,53
7	100	1	1,0%	0,20	47	6,7%	1,34
8	100	1	1,0%	0,20	48	6,0%	1,20
9	100	1	1,0%	0,20	49	5,4%	1,09
10	100	1	1,0%	0,20	50	5,0%	1,00
All	1000	50	5,0%				

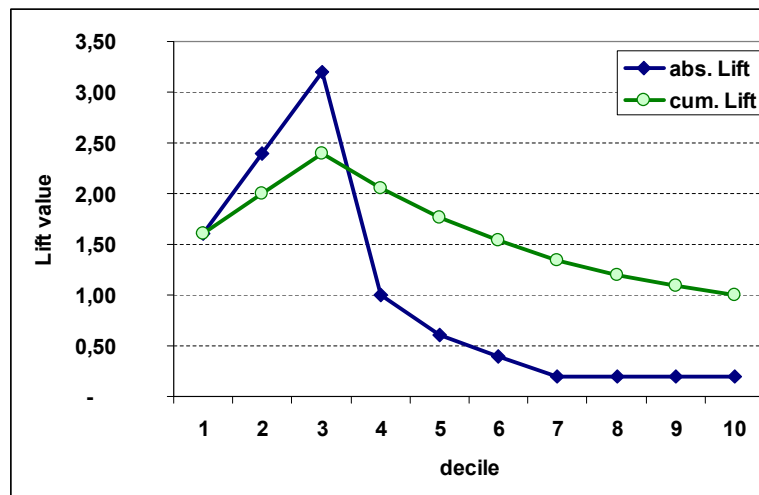
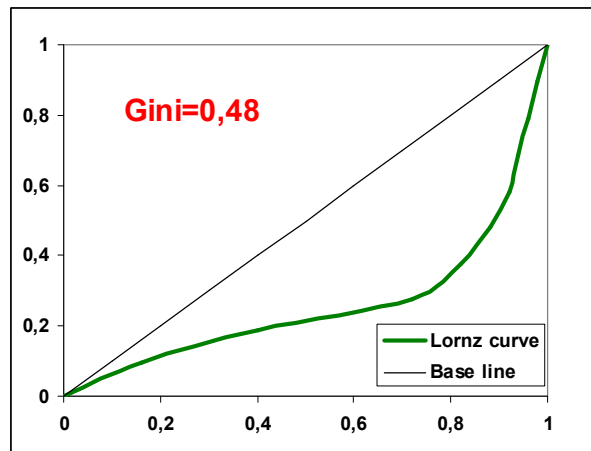


Indexes based on distribution function

□ When bad rates are not monotone:

- LC looks fine
- Gini is slightly lowered
- Lift looks strange

decile	# cleints	absolutely			cumulatively		
		# bad clients	Bad rate	abs. Lift	# bad clients	Bad rate	cum. Lift
1	100	8	8,0%	1,60	8	8,0%	1,60
2	100	12	12,0%	2,40	20	10,0%	2,00
3	100	16	16,0%	3,20	36	12,0%	2,40
4	100	5	5,0%	1,00	41	10,3%	2,05
5	100	3	3,0%	0,60	44	8,8%	1,76
6	100	2	2,0%	0,40	46	7,7%	1,53
7	100	1	1,0%	0,20	47	6,7%	1,34
8	100	1	1,0%	0,20	48	6,0%	1,20
9	100	1	1,0%	0,20	49	5,4%	1,09
10	100	1	1,0%	0,20	50	5,0%	1,00
All	1000	50	5,0%				



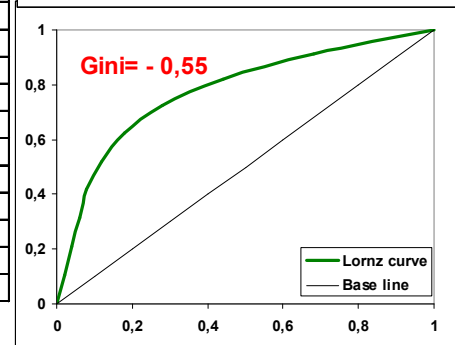
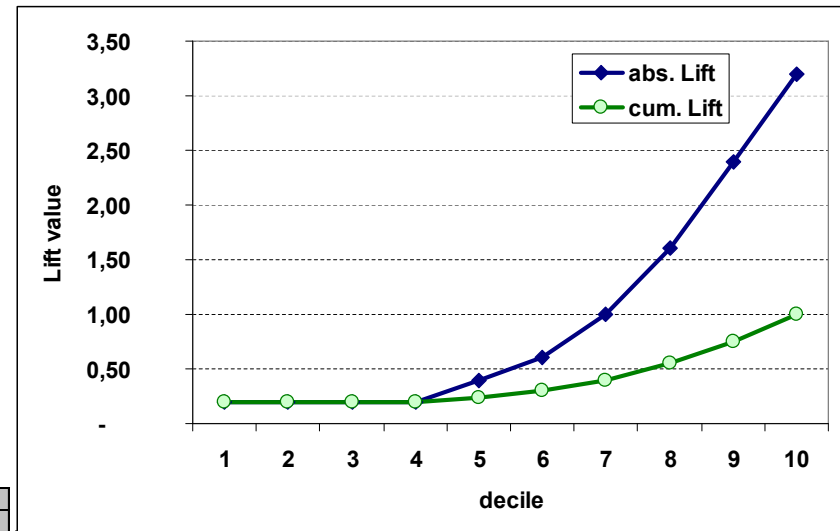
Indexes based on distribution function

□ When score is reversed, we obtain reversed figures.

decile	# cleints	absolutely			cumulatively		
		# bad clients	Bad rate	abs. Lift	# bad clients	Bad rate	cum. Lift
1	100	16	16,0%	3,20	16	16,0%	3,20
2	100	12	12,0%	2,40	28	14,0%	2,80
3	100	8	8,0%	1,60	36	12,0%	2,40
4	100	5	5,0%	1,00	41	10,3%	2,05
5	100	3	3,0%	0,60	44	8,8%	1,76
6	100	2	2,0%	0,40	46	7,7%	1,53
7	100	1	1,0%	0,20	47	6,7%	1,34
8	100	1	1,0%	0,20	48	6,0%	1,20
9	100	1	1,0%	0,20	49	5,4%	1,09
10	100	1	1,0%	0,20	50	5,0%	1,00
All	1000	50	5,0%				



decile	# cleints	absolutely			cumulatively		
		# bad clients	Bad rate	abs. Lift	# bad clients	Bad rate	cum. Lift
1	100	1	1,0%	0,20	1	1,0%	0,20
2	100	1	1,0%	0,20	2	1,0%	0,20
3	100	1	1,0%	0,20	3	1,0%	0,20
4	100	1	1,0%	0,20	4	1,0%	0,20
5	100	2	2,0%	0,40	6	1,2%	0,24
6	100	3	3,0%	0,60	9	1,5%	0,30
7	100	5	5,0%	1,00	14	2,0%	0,40
8	100	8	8,0%	1,60	22	2,8%	0,55
9	100	12	12,0%	2,40	34	3,8%	0,76
10	100	16	16,0%	3,20	50	5,0%	1,00
All	1000	50	5,0%				

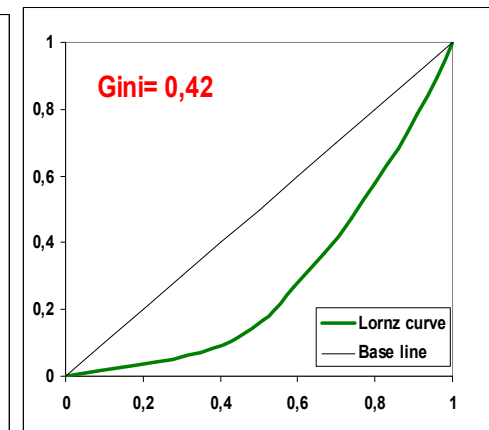
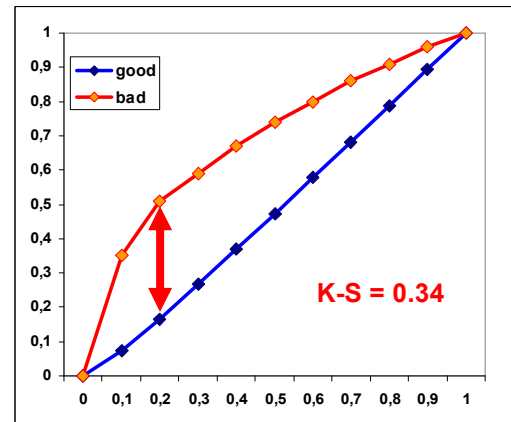


Indexes based on distribution function

❑ The Gini is not enough!!!

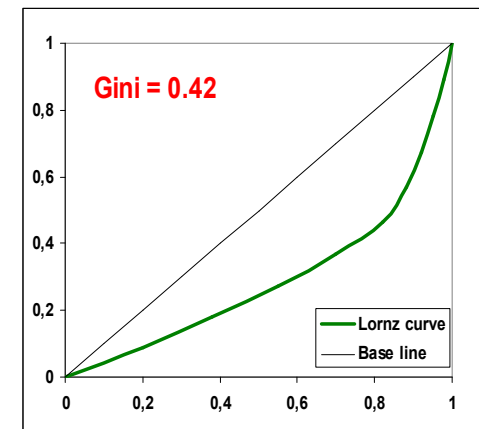
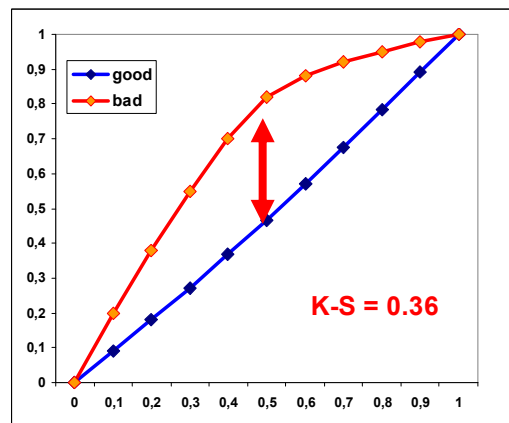
➤ SC 1:

decile	# cleints	# bad clients	Bad rate
1	100	35	35,0%
2	100	16	16,0%
3	100	8	8,0%
4	100	8	8,0%
5	100	7	7,0%
6	100	6	6,0%
7	100	6	6,0%
8	100	5	5,0%
9	100	5	5,0%
10	100	4	4,0%
All	1000	100	10,0%



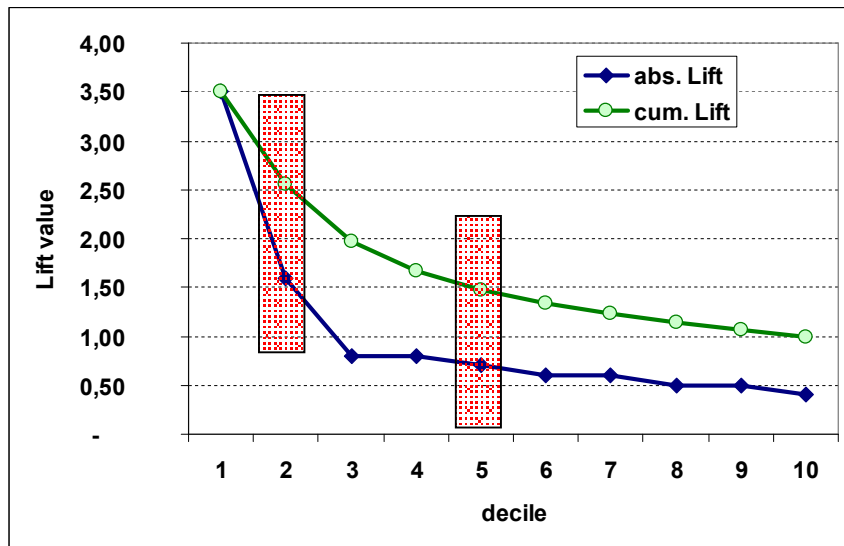
➤ SC 2:

decile	# cleints	# bad clients	Bad rate
1	100	20	20,0%
2	100	18	18,0%
3	100	17	17,0%
4	100	15	15,0%
5	100	12	12,0%
6	100	6	6,0%
7	100	4	4,0%
8	100	3	3,0%
9	100	3	3,0%
10	100	2	2,0%
All	1000	100	10,0%



Indexes based on distribution function

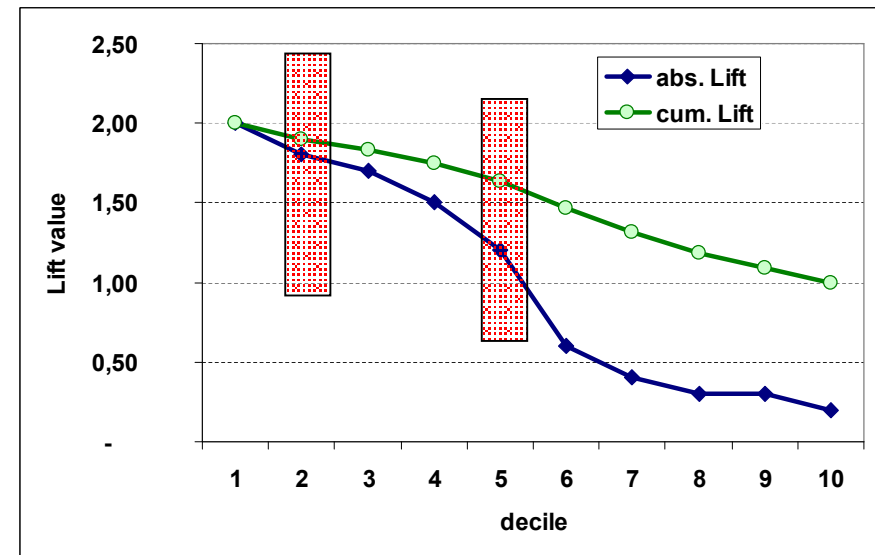
➤ SC 1:



$$\text{Lift}_{20\%} = 2.55$$

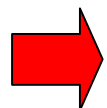
$$\text{Lift}_{50\%} = 1.48$$

➤ SC 2:



$$> \text{Lift}_{20\%} = 1.90$$

$$< \text{Lift}_{50\%} = 1.64$$



SC 2 is better if reject rate is expected around 50%.
SC 1 is much more better if reject rate is expected by 20%.

Indexes based on distribution function

- Lift can be expressed and computed by formulae:

$$Lift(a) = \frac{F_{n.BAD}(a)}{F_{N.ALL}(a)} \quad a \in [L, H]$$

$$Lift_q = \frac{F_{n.BAD}(F_{N.ALL}^{-1}(q))}{F_{N.ALL}(F_{N.ALL}^{-1}(q))} = \frac{1}{q} F_{n.BAD}(F_{N.ALL}^{-1}(q))$$

$$F_{N.ALL}^{-1}(q) = \min\{a \in [L, H], F_{N.ALL}(a) \geq q\}$$

$$Lift_{10\%} = 10 \cdot F_{n.BAD}(F_{N.ALL}^{-1}(0.1)).$$

Indexes based on density function

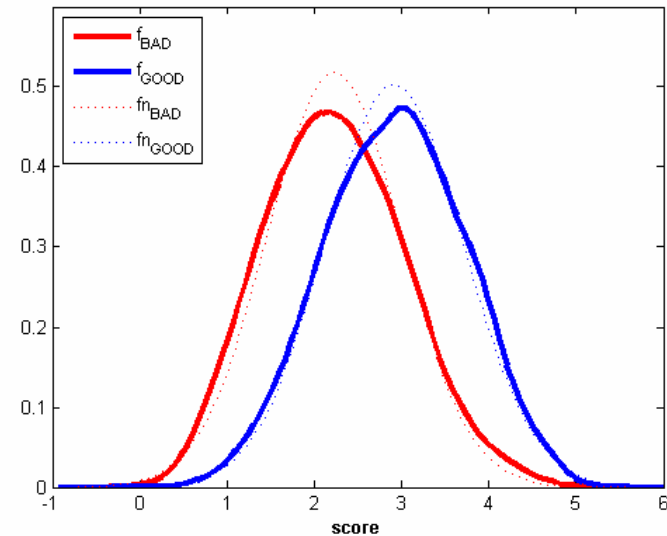
➤ Likelihood density functions: $f_{GOOD}(x)$ $f_{BAD}(x)$

➤ Kernel estimates: $\tilde{f}_{GOOD}(x, h) = \sum_{\substack{i=1, \\ D_K=1}}^n \frac{1}{n} K_h(x - s_i)$ $\tilde{f}_{BAD}(x, h) = \sum_{\substack{i=1, \\ D_k=0}}^m \frac{1}{m} K_h(x - s_i)$

➤ Optimal bandwidth (maximal smoothing):

$$h_{OS,k} = \left[\frac{(2k+1)! k (2k+5)^{k+3/2}}{(2k+3)!} \right]^{\frac{1}{2k+1}} \cdot \tilde{\sigma} \cdot n^{\frac{1}{2k+1}}$$

where: k is the order of kernel function
(e.g. 2 for Epanechnikov kernel)
 n is number of actual cases
 $\tilde{\sigma}$ is an estimate of standard deviation



Indexes based on density function

➤ Mean difference
(Mahalanobis distance):

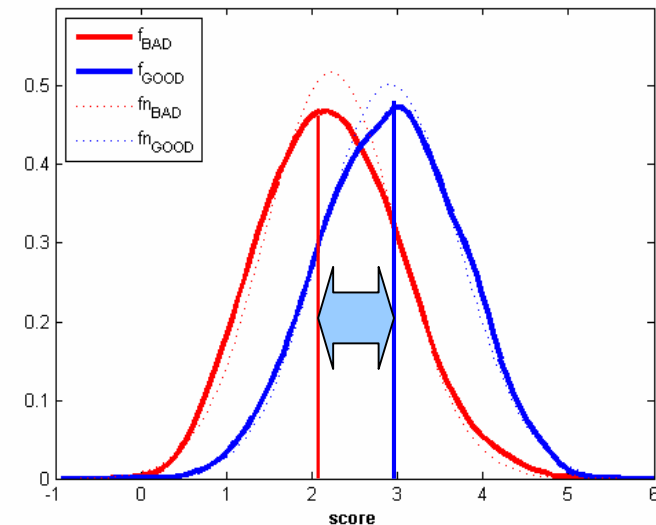
$$D = \frac{M_g - M_b}{S}$$

where S is pooled standard deviation:

$$S = \left(\frac{nS_g^2 + mS_b^2}{n + m} \right)^{\frac{1}{2}}$$

M_g, M_b are means of good (bad) clients

S_g, S_b are standard deviations of good (bad) clients



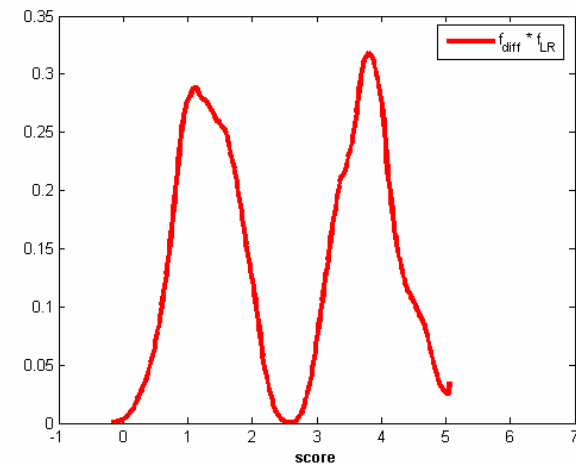
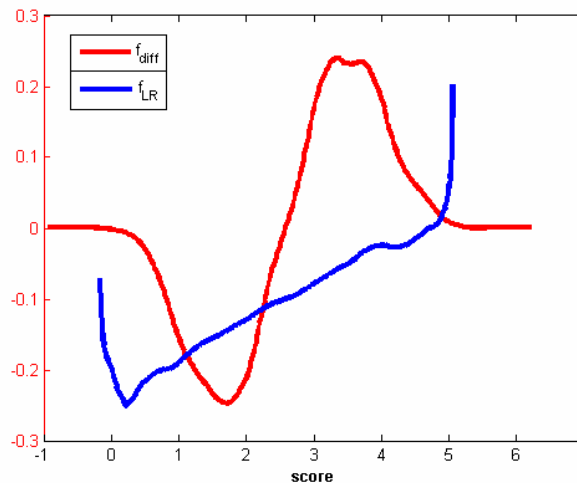
Indexes based on density function

- Information value (I_{val}) – continuous case (Divergence):

$$I_{val} = \int_{-\infty}^{\infty} (f_{GOOD}(x) - f_{BAD}(x)) \ln \left(\frac{f_{GOOD}(x)}{f_{BAD}(x)} \right) dx$$

$$f_{diff}(x) = f_{GOOD}(x) - f_{BAD}(x)$$

$$f_{LR}(x) = \ln \left(\frac{f_{GOOD}(x)}{f_{BAD}(x)} \right)$$

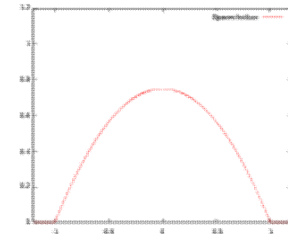


Indexes based on density function

➤ Information value (I_{val}) – discretized continuous case:

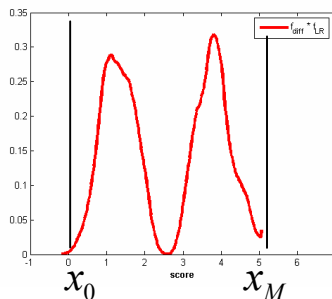
- Replace density functions by their kernel estimates and compute integral numerically (e.g. by composite trapezoidal rule).

- Using Epanechnikov kernel, given by $K(x) = \frac{3}{4}(1-x^2) \cdot I(x \in [-1, 1])$ and optimal bandwidth $h_{OS,k}$ we have



$$\tilde{f}_{IV}(x) = (\tilde{f}_{GOOD}(x, h_{OS,2}) - \tilde{f}_{BAD}(x, h_{OS,2})) \ln \left(\frac{\tilde{f}_{GOOD}(x, h_{OS,2})}{\tilde{f}_{BAD}(x, h_{OS,2})} \right)$$

- For given $M+1$ points x_0, \dots, x_M we obtain



$$I_{val} = \frac{x_M - x_0}{2M} \left(\tilde{f}_{IV}(x_0) + 2 \sum_{i=1}^{M-1} \tilde{f}_{IV}(x_i) + \tilde{f}_{IV}(x_M) \right)$$

Indexes based on density function

➤ Information statistics/value (I_{val}) – discrete case:

- Create intervals of score – typically deciles. Number of goods (bads) in i-th interval is marked by g_i (b_i).

- It must hold $g_i > 0, b_i > 0 \quad \forall i$

- Then we have

$$I_{val} = \sum_i \left(\frac{g_i}{n} - \frac{b_i}{m} \right) \ln \left(\frac{g_i m}{b_i n} \right)$$

score int.	# bad clients	#good clients	% bad [1]	% good [2]	[3] = [2] - [1]	[4] = [2] / [1]	[5] = ln[4]	[6] = [3] * [5]
1	1	10	2,0%	1,1%	-0,01	0,53	-0,64	0,01
2	2	15	4,0%	1,6%	-0,02	0,39	-0,93	0,02
3	8	52	16,0%	5,5%	-0,11	0,34	-1,07	0,11
4	14	93	28,0%	9,8%	-0,18	0,35	-1,05	0,19
5	10	146	20,0%	15,4%	-0,05	0,77	-0,26	0,01
6	6	247	12,0%	26,0%	0,14	2,17	0,77	0,11
7	4	137	8,0%	14,4%	0,06	1,80	0,59	0,04
8	3	105	6,0%	11,1%	0,05	1,84	0,61	0,03
9	1	97	2,0%	10,2%	0,08	5,11	1,63	0,13
10	1	48	2,0%	5,1%	0,03	2,53	0,93	0,03
All	50	950					Info. Value	0,68

Indexes based on density function

□ Information value for our example of two scorecards:

➤ SC 1:

decile	# cleints	# bad clients	#good	% bad [1]	% good [2]	[3] = [2] - [1]	[4] = [2] / [1]	[5] = ln[4]	[6] = [3] * [5]	cum. [6]
1	100	35	65	35,0%	7,2%	-0,28	0,21	-1,58	0,44	0,44
2	100	16	84	16,0%	9,3%	-0,07	0,58	-0,54	0,04	0,47
3	100	8	92	8,0%	10,2%	0,02	1,28	0,25	0,01	0,48
4	100	8	92	8,0%	10,2%	0,02	1,28	0,25	0,01	0,49
5	100	7	93	7,0%	10,3%	0,03	1,48	0,39	0,01	0,50
6	100	6	94	6,0%	10,4%	0,04	1,74	0,55	0,02	0,52
7	100	6	94	6,0%	10,4%	0,04	1,74	0,55	0,02	0,55
8	100	5	95	5,0%	10,6%	0,06	2,11	0,75	0,04	0,59
9	100	5	95	5,0%	10,6%	0,06	2,11	0,75	0,04	0,63
10	100	4	96	4,0%	10,7%	0,07	2,67	0,98	0,07	0,70
All	1000	100	900						Info. Value	0,70

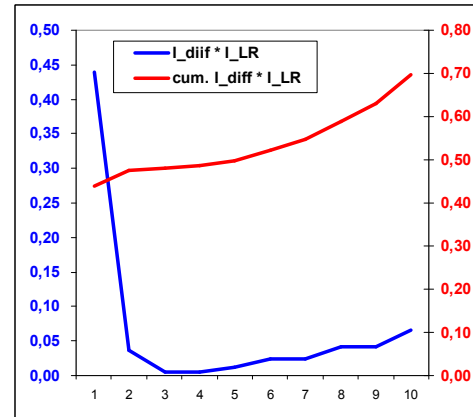
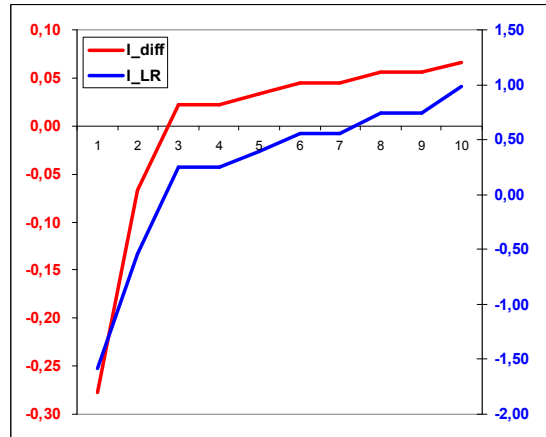
➤ SC 2:

decile	# cleints	# bad clients	#good	% bad [1]	% good [2]	[3] = [2] - [1]	[4] = [2] / [1]	[5] = ln[4]	[6] = [3] * [5]	cum. [6]
1	100	20	80	20,0%	8,9%	-0,11	0,44	-0,81	0,09	0,09
2	100	18	82	18,0%	9,1%	-0,09	0,51	-0,68	0,06	0,15
3	100	17	83	17,0%	9,2%	-0,08	0,54	-0,61	0,05	0,20
4	100	15	85	15,0%	9,4%	-0,06	0,63	-0,46	0,03	0,22
5	100	12	88	12,0%	9,8%	-0,02	0,81	-0,20	0,00	0,23
6	100	6	94	6,0%	10,4%	0,04	1,74	0,55	0,02	0,25
7	100	4	96	4,0%	10,7%	0,07	2,67	0,98	0,07	0,32
8	100	3	97	3,0%	10,8%	0,08	3,59	1,28	0,10	0,42
9	100	3	97	3,0%	10,8%	0,08	3,59	1,28	0,10	0,52
10	100	2	98	2,0%	10,9%	0,09	5,44	1,69	0,15	0,67
All	1000	100	900						Info. Value	0,67

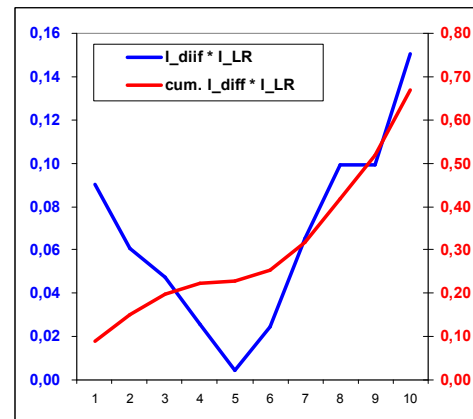
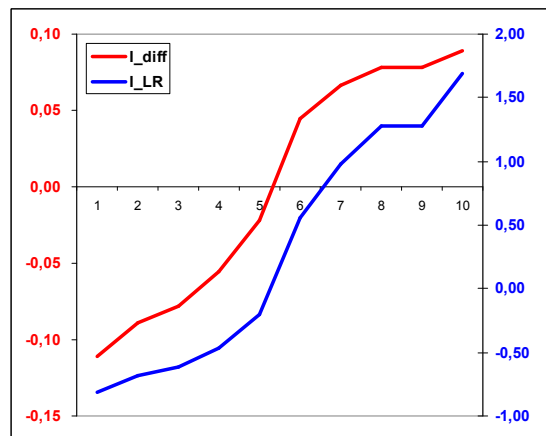
Indexes based on density function

Using markings $I_{diff_i} = \left(\frac{g_i}{n} - \frac{b_i}{m} \right)$ $I_{LR_i} = \ln \left(\frac{g_i m}{b_i n} \right)$ we have:

➤ SC 1:



➤ SC 2:



K-S	= 0.34
Gini	= 0.42
Lift _{20%}	= 2.55
Lift _{50%}	= 1.48
I _{val}	= 0.70
I _{val20%}	= 0.47
I _{val50%}	= 0.50
K-S	= 0.36
Gini	= 0.42
Lift _{20%}	= 1.90
Lift _{50%}	= 1.64
I _{val}	= 0.67
I _{val20%}	= 0.15
I _{val50%}	= 0.23

Some results for normally distributed scores

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

$$f_{GOOD}(x) = \frac{1}{\sigma_g \sqrt{2\pi}} e^{-\frac{(x-\mu_g)^2}{2\sigma_g^2}} \quad f_{BAD}(x) = \frac{1}{\sigma_b \sqrt{2\pi}} e^{-\frac{(x-\mu_b)^2}{2\sigma_b^2}}$$

- Estimates of parameters μ_g, μ_b, σ_g and σ_b :

M_g, M_b are means of good (bad) clients

S_g, S_b are standard deviations of good (bad) clients

- Pooled standard deviation:

$$S = \left(\frac{nS_g^2 + mS_b^2}{n+m} \right)^{\frac{1}{2}}$$

- Estimates of mean and standard dev. of scores for all clients μ_{ALL}, σ_{ALL} :

$$M = M_{ALL} = \frac{nM_g + mM_b}{n+m} \quad S_{ALL} = \left(\frac{nS_g^2 + mS_b^2 + n(M_g - M)^2 + m(M_b - M)^2}{(n+m)} \right)^{\frac{1}{2}}$$

Some results for normally distributed scores

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

$$D = \frac{\mu_g - \mu_b}{\sigma}$$

$$D = \frac{M_g - M_b}{S}$$

$$KS = \Phi\left(\frac{D}{2}\right) - \Phi\left(\frac{-D}{2}\right) = 2 \cdot \Phi\left(\frac{D}{2}\right) - 1$$

$$Gini = 2 \cdot \Phi\left(\frac{D}{\sqrt{2}}\right) - 1$$

$$Lift_q = \frac{1}{q} \Phi\left(\frac{\sigma_{ALL}}{\sigma} \cdot \Phi^{-1}(q) + p_G \cdot D\right)$$

$$Lift_q = \frac{1}{q} \Phi\left(\frac{S_{ALL}}{S} \cdot \Phi^{-1}(q) + p_G \cdot D\right)$$

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

Where $\Phi(\cdot)$ is the standardized normal distribution function, $\Phi_{\mu, \sigma^2}(\cdot)$ the normal distribution function with parameters μ , σ^2 and $\Phi^{-1}(\cdot)$ is the standard quantile function.

Some results for normally distributed scores

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

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

$$D^* = \frac{M_g - M_b}{\sqrt{S_g^2 + S_b^2}}$$

$$KS = \Phi\left(\frac{a}{b}\sigma_b \cdot D^* - \frac{1}{b}\sigma_g \sqrt{a^2 D^{*2} + 2b \cdot c}\right) - \Phi\left(\frac{a}{b}\sigma_g \cdot D^* - \frac{1}{b}\sigma_b \sqrt{a^2 D^{*2} + 2b \cdot c}\right)$$

where $a = \sqrt{\sigma_b^2 + \sigma_g^2}$, $b = \sigma_b^2 - \sigma_g^2$, $c = \ln\left(\frac{\sigma_g}{\sigma_b}\right)$

$$KS = \Phi\left(\frac{\sqrt{S_b^2 + S_g^2}}{S_b - S_g} S_b \cdot D^* - \frac{1}{S_b - S_g} S_g \sqrt{(S_b^2 + S_g^2) D^{*2} + 2 \cdot (S_b^2 - S_g^2) \ln\left(\frac{S_g}{S_b}\right)}\right) - \Phi\left(\frac{\sqrt{S_b^2 + S_g^2}}{S_b - S_g} S_g \cdot D^* - \frac{1}{S_b - S_g} S_b \sqrt{(S_b^2 + S_g^2) D^{*2} + 2 \cdot (S_b^2 - S_g^2) \ln\left(\frac{S_g}{S_b}\right)}\right)$$

Some results for normally distributed scores

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

$$Gini = 2 \cdot \Phi(D^*) - 1$$

$$Lift_q = \frac{1}{q} \Phi_{\mu_b, \sigma_b^2}(\mu_{ALL} + \sigma_{ALL} \cdot \Phi^{-1}(q)) = \frac{1}{q} \Phi\left(\frac{\sigma_{ALL} \cdot \Phi^{-1}(q) + \mu_{ALL} - \mu_b}{\sigma_b}\right)$$

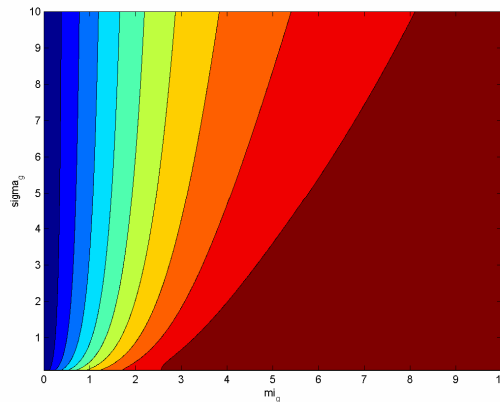
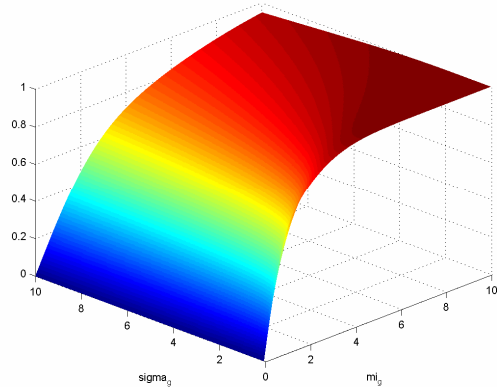
$$Lift_q = \frac{1}{q} \Phi\left(\frac{S_{ALL} \cdot \Phi^{-1}(q) + M - M_b}{S_b}\right)$$

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

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

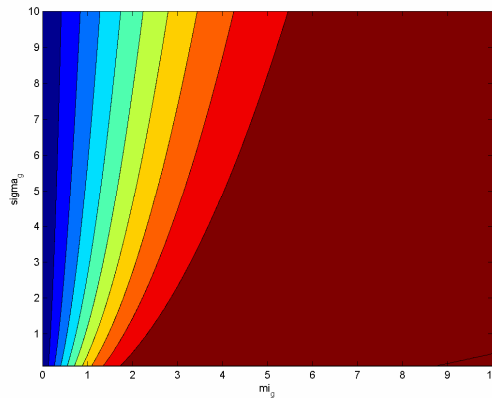
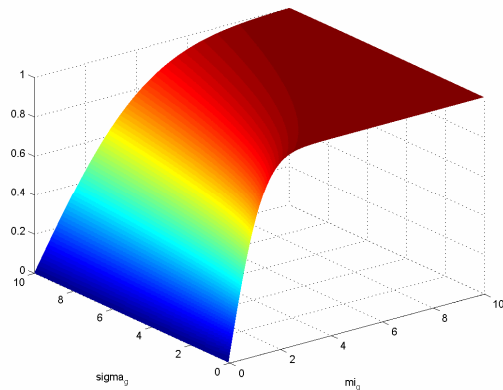
Some results for normally distributed scores

➤ **KS:** $\mu_b = 0, \sigma_b^2 = 1$

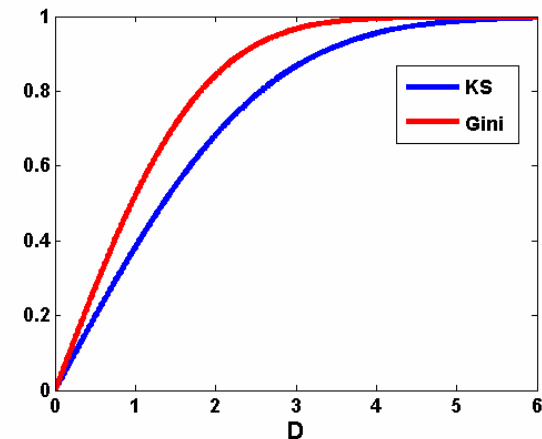


□ KS and the Gini react much more to change of μ_g and are almost unchanged in the direction of σ_g^2 .

➤ **Gini** $\mu_b = 0, \sigma_b^2 = 1$

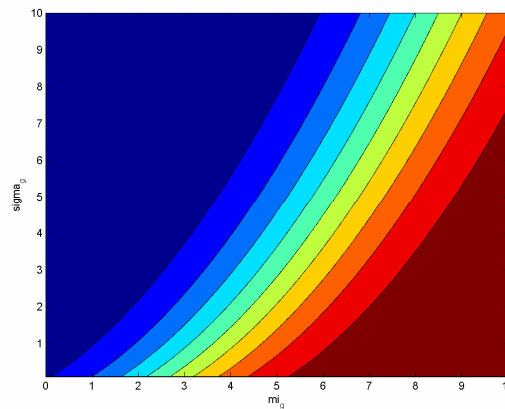
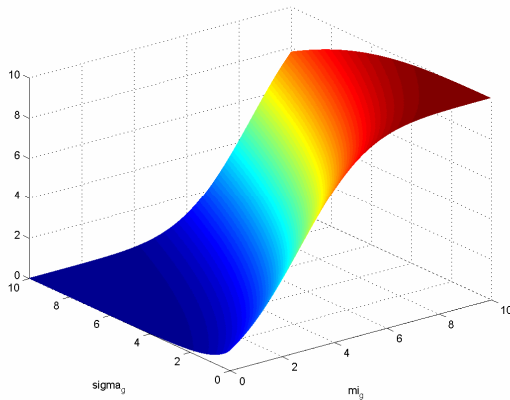


• **Gini > KS**



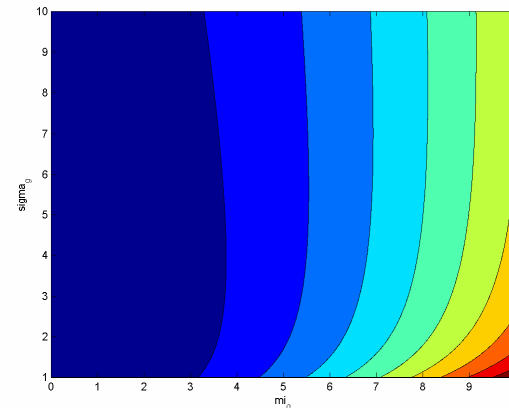
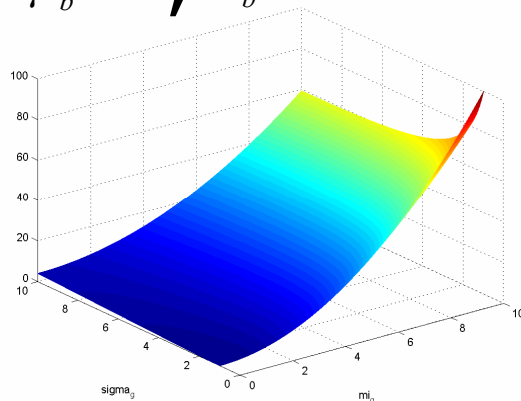
Some results for normally distributed scores

➤ **Lift_{10%}**: $\mu_b = 0, \sigma_b^2 = 1$



□ In case of Lift_{10%} it is evident strong dependence on μ_g and significantly higher dependence on σ_g^2 than in case of KS and Gini.

➤ **I_{val}**: $\mu_b = 0, \sigma_b^2 = 1$



□ Again strong dependence on μ_g . Furthermore value of I_{val} rises very quickly to infinity when σ_g^2 tends to zero.

ROC (Receiver operating characteristic)

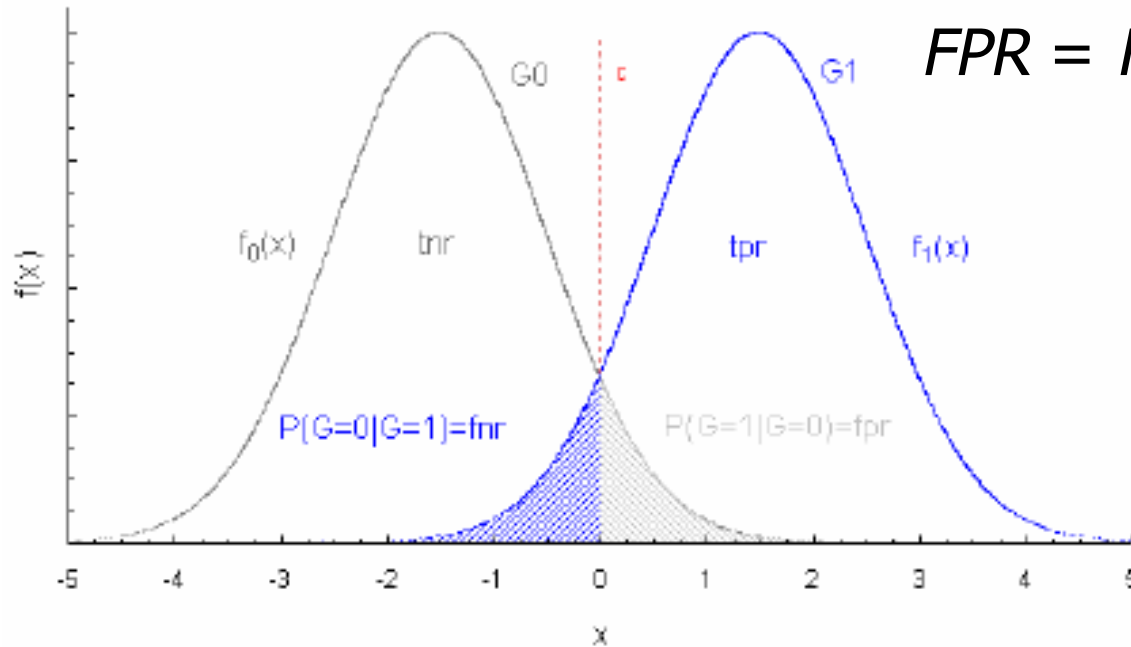
- *TN (true negative)* - počet správně klasifikovaných negativních případů
- *TP (true positive)* - počet správně klasifikovaných pozitivních případů
- *FP (false positive)* - počet nesprávně klasifikovaných negativních případů
- *FN (false negative)* - počet nesprávně klasifikovaných pozitivních případů

Skuteč.	Predikce		
	G0	G1	Celkem
G0	<i>TN</i>	<i>FP</i>	<i>N</i>
G1	<i>FN</i>	<i>TP</i>	<i>P</i>
Celkem	<i>Pneg</i>	<i>PPos</i>	<i>n</i>

ROC - TPR, FPR

$$TPR = TP / P = TP / (TP + FN)$$

$$FPR = FP / N = FP / (FP + TN)$$



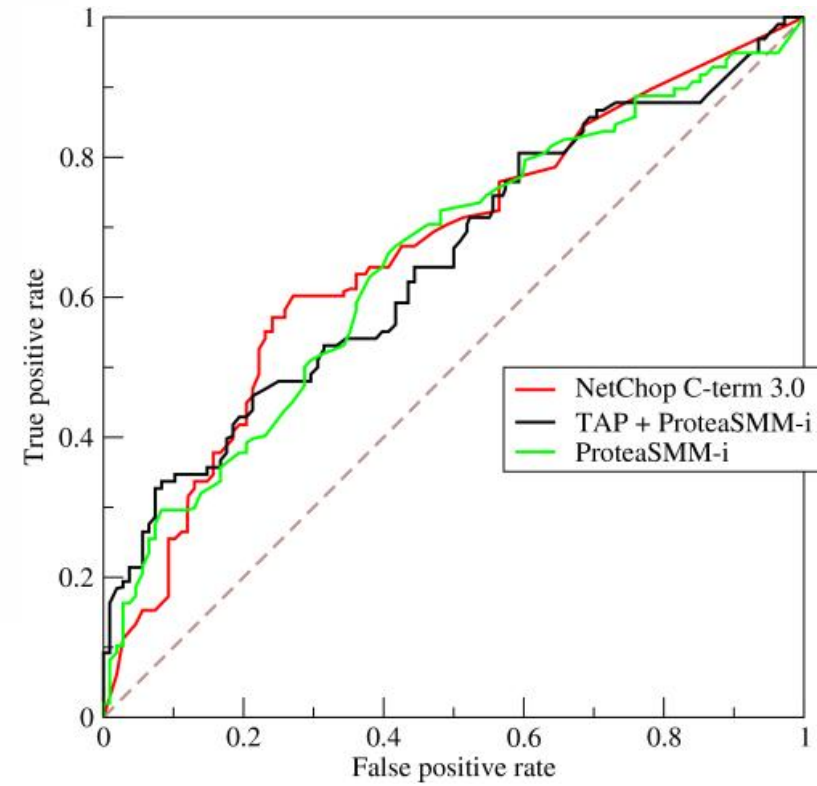
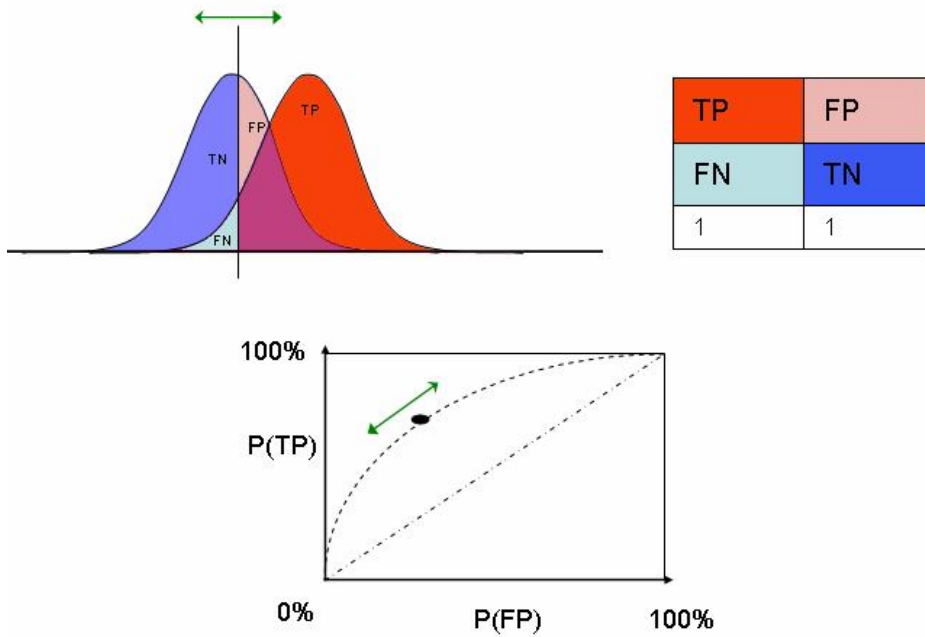
$$tpr(c) = P(X > c | G_1) = 1 - F_1(c)$$

$$tnr(c) = P(X < c | G_0) = F_0(c)$$

$$fpr(c) = P(X > c | G_0) = 1 - F_0(c)$$

$$fnr(c) = P(X < c | G_1) = F_1(c)$$

ROC

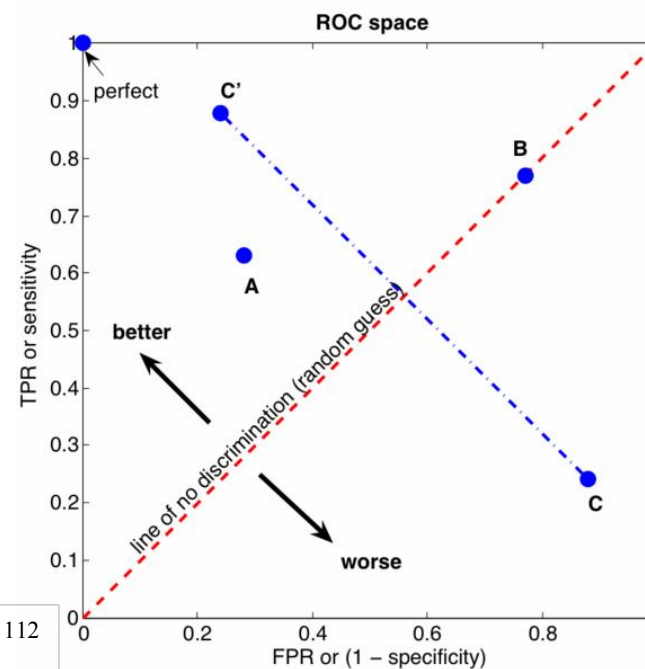


ROC - ACC

Accuracy:

$$ACC = (TP + TN) / (P + N)$$

A			B			C			C'		
TP=63	FP=28	91	TP=77	FP=77	154	TP=24	FP=88	112	TP=88	FP=24	112
FN=37	TN=72	109	FN=23	TN=23	46	FN=76	TN=12	88	FN=12	TN=76	88
100	100	200	100	100	200	100	100	200	100	100	200
TPR = 0.63			TPR = 0.77			TPR = 0.24			TPR = 0.88		
FPR = 0.28			FPR = 0.77			FPR = 0.88			FPR = 0.24		
ACC = 0.68			ACC = 0.50			ACC = 0.18			ACC = 0.82		



ROC – AUC, Gini

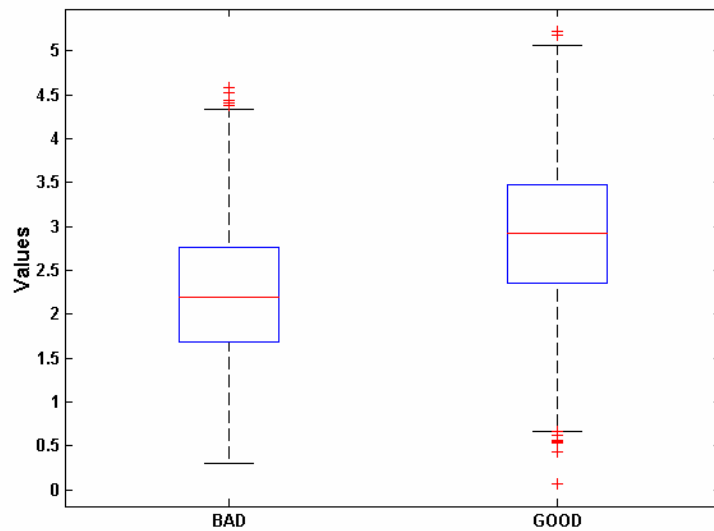


The AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. It can be shown that the area under the ROC curve is equivalent to the Mann-Whitney U, which tests for the median difference between scores obtained in the two groups considered if the groups are of continuous data. It is also equivalent to the Wilcoxon test of ranks. The AUC has been found to be related to the Gini coefficient(G) by the following formula

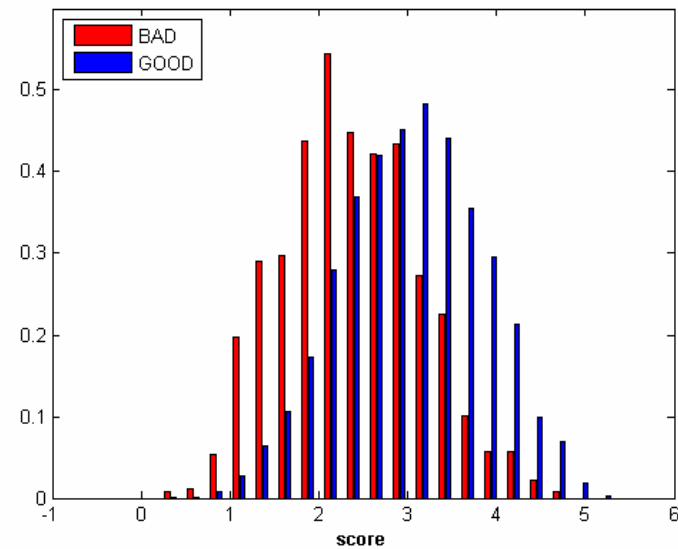
$$G + 1 = 2 \times AUC$$

Další evaluační grafy

Boxplot

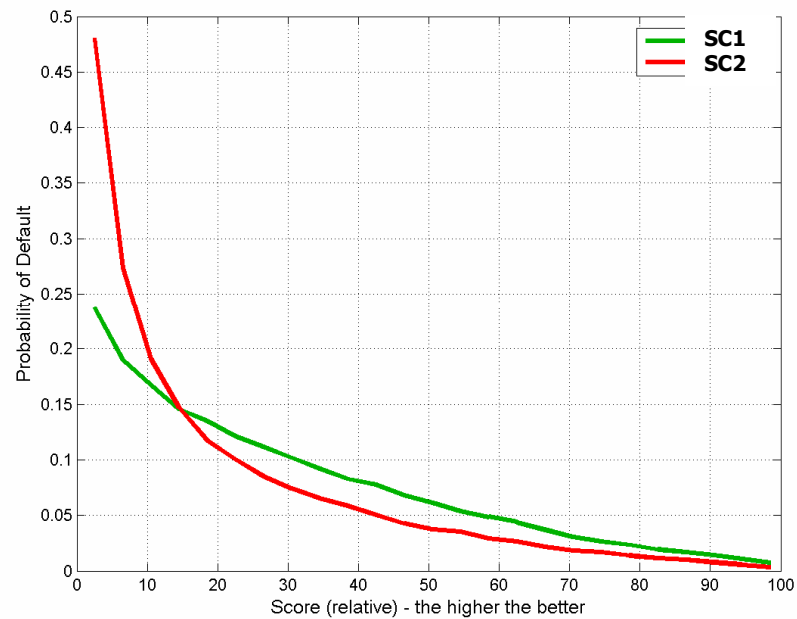


Histogram

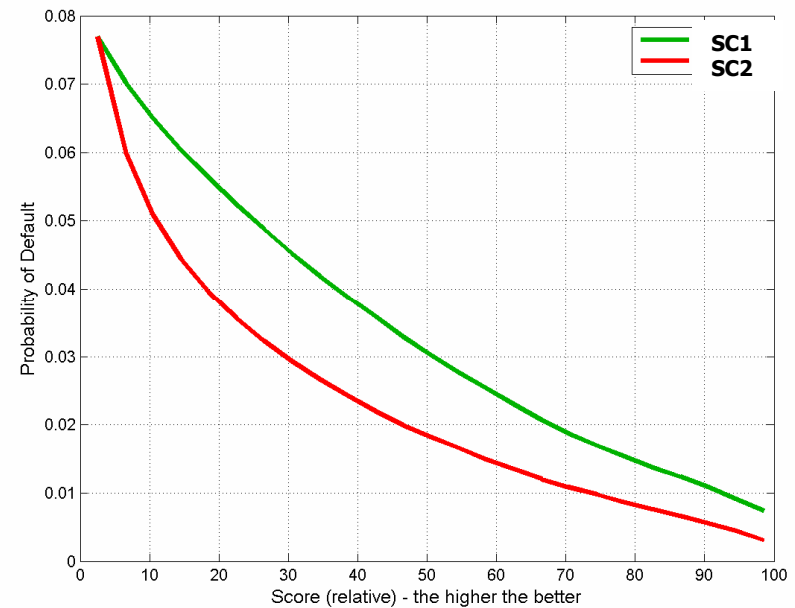


Další evaluační grafy

PD - absolutně

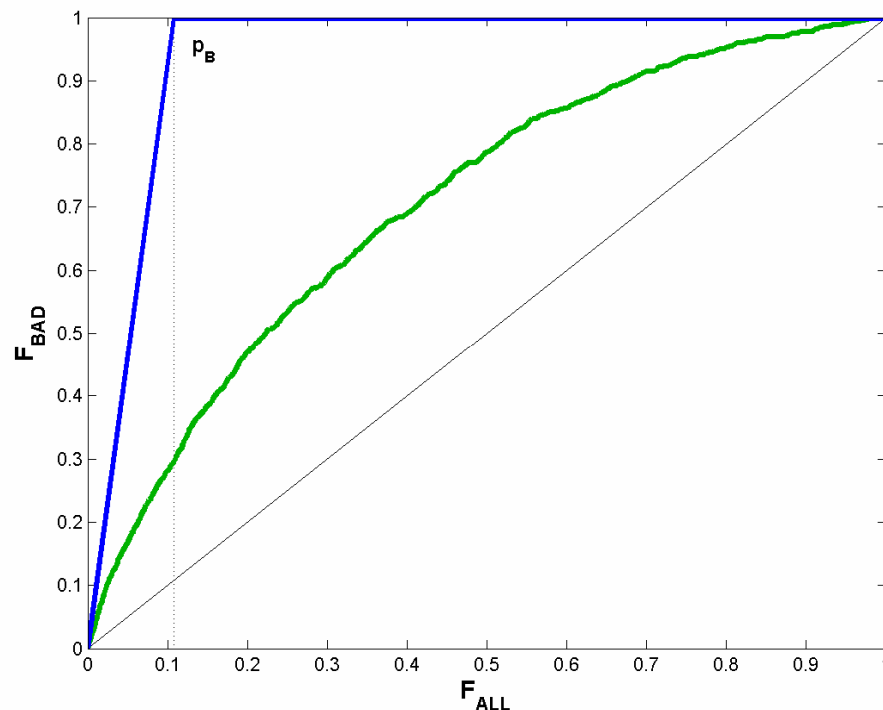


PD - kumulativně



Další evaluační grafy

Lift chart:



In this case we have the proportion of all clients (F_{ALL}) on the horizontal axis and the proportion of bad clients (F_{BAD}) on the vertical axis. The ideal model is now represented by polyline from $[0, 0]$ through $[p_B, 1]$ to $[1, 1]$. Advantage of this figure is that one can easily read the proportion of rejected bads vs. proportion of all rejected. For example we can see that if we want to reject 70% of bads, we have to reject about 40% of all applicants.

Postupy evaluace

➤ **evaluace na učících datech**

Evaluace na učících datech použitých k učicímu procesu není ke zjištění kvality modelu vhodná a má nízkou vypovídací schopnost, protože často může dojít k přeučení modelu. Odhad predikční kvality modelu na učících datech se nazývá resubstituční nebo interní odhad. Odhady ukazatelů kvality modelů provedených na učících datech jsou nadhodnocené, proto se místo nich používají testovací data, která se v rámci přípravy dat pro tyto účely vyčlení.

➤ **evaluace na testovacích datech**

Evaluace na testovacích datech již má patřičnou vypovídací schopnost, jelikož tato data nebyla použita k sestavení modelu. Na testovací data jsou kladeny určité požadavky. Soubor testovacích dat by měl obsahovat dostatečné množství dat a měl by reprezentovat či vystihovat charakteristiky učících dat. Empiricky doporučený poměr učících a testovacích dat je 75%, resp. 25% případů. Zajištění patřičné reprezentativnosti je realizováno pomocí náhodného stratifikovaného výběru.

Postupy evaluace

➤ **křížové ověřování** (*cross-validation*)

V případě nedostatečného počtu pozorování, kdy rozdělení datového souboru na učící a testovací data za účelem vyhodnocení modelu není možné, je vhodné použít metodu křížového ověřování. Výhodou této metody na rozdíl od dělení datového souboru je, že každý případ z dat je použit k sestavení modelu a každý případ je alespoň jednou použit k testování. Postup je následující:

- Soubor dat je náhodně rozdělen do n disjunktních podmnožin tak, že každá podmnožina obsahuje přibližně stejný počet záznamů. Výběry jsou stratifikovány podle tříd (příslušnosti k určité třídě), aby bylo zajištěno, že podíly jednotlivých tříd podmnožin jsou zhruba stejné jako v celém souboru.
- Z těchto n disjunktních podmnožin se vyčlení $n-1$ podmnožin pro sestavení modelu (konstrukční podmnožina) a zbývající podmnožina (validační podmnožina) je použita k jeho vyhodnocení. Model je tedy evaluován na podmnožině dat, ze kterých nebyl sestaven a na této množině dat je odhadována jeho predikční kvalita.
- Celý postup se zopakuje n -krát a dílčí odhady ukazatelů kvality se zprůměrnují. Velikost validační podmnožiny lze přibližně stanovit jako poměr počtu případů ku počtu validačních podmnožin.

Postupy evaluace

➤ **bootstrap metoda**

Metoda *bootstrap* zkoumá charakteristiky jednotlivých resamplovaných vzorků, které byly pořízeny z empirického výběru. Pokud původní výběr obsahuje m prvků, tak každý má naději objevit se v resamplovaném výběru. Při úplném resamplování o velikosti vzorku n jsou uvažovány všechny možné výběry a existuje tedy m^n možných výběrů. Úplné resamplování je teoreticky proveditelné, ale vyžádalo by si mnoho času. Alternativou je simulace *Monte Carlo*, pomocí níž se aproximuje úplné resamplování tak, že se provede B náhodných výběrů (obvykle se volí 500 – 10000 výběrů) s tím, že každý prvek je vždy nahrazen (vrácen zpět do osudí). Jsou-li dána data $X = \{X_1, \dots, X_n\}$ a je-li požadován odhad parametru θ , provede se z původních dat B výběrů a pro každý výběr je spočítán odhad parametru θ . *Bootstrap* odhad parametru je určen jako průměr dílčích odhadů. V případě evaluace modelů bude parametrem θ zvolený ukazatel predikční kvality.

➤ **jackknife**

Tato metoda je založena na sekvenční strategii odebírání a vracení prvků do výběru o velikosti n . Pro datový soubor, který obsahuje n prvků, procedura generuje n vzorků s počtem prvků $n-1$. Pro každý zmenšený výběr o velikosti $n-1$ je odhadnuta hodnota parametru. Dílčí odhady se následně zprůměrují podobně jako u metody *bootstrap*.